EOS 352 Continuum Dynamics
Convection in porous media

May 1, 2020

Overview

These notes cover the following

- the Boussinesq approximation for convection
- linearization and linear stability analysis
- Fourier series solution
- instability and the critical Rayleigh number
- A taster of bifurcations and nonlinear dynamics

A model for convection

Convection is the motion of a fluid driven by density differences, and hence by variations in the body force $\rho\mathbf{g}$, across space: lighter material wants to go up, denser material wants to go down. In convection, these density differences are usually caused by heating and cooling, and sometimes by chemical reactions. Here we will be interested in the thermal version of convection, which classically works by heating a fluid from underneath. That causes the fluid at the bottom of the domain to become warmer and less dense, giving it the tendency to want to rise. In order to do that, however, a symmetry often needs to be broken: we cannot have all the fluid rising up at once, as it somehow needs to be replaced by colder fluid descending from above, so we need the warm fluid to rise up in some parts of the domain, and the cold fluid to descend in others. This requires the formation of a pattern of motion, which often arises out

of an initial state that has near-perfect symmetry, with isodensity surfaces that are almost flat.

The way that this symmetry breaking happens is through an *instability*. The initial state does not have perfect symmetry, but small amounts of noise are always present. In practice, some parts of the bottom of the domain will be slightly warmer than others that are off to the side of the warmer fluid, and therefore have a greater tendency to rise than those that are slightly cooler. This can become self-reinforcing, with the rising of the warmer fluid drawing in other, initially cooler fluid sideways, which is heated up by the bottom of the domain as it flows, and sustains the upward motion in the location where the originally hotter fluid was located. The descent of the cooler fluid meanwhile draws in more cold fluid from above.

The purpose of these notes is to show mathematically how we can model this process, and also to show that it need not always occur: A fluid heated from below does not have to convect. If conduction is strong enough, any bits of warm fluid that try to rise up through the domain are cooled rapidly by conduction, increasing their density and causing them to sink again. By constructing a mathematical model, we can identify the conditions under which the tendency of warm fluid to rise is strong enough to cause self-sustaining convection, and conditions under which this is suppressed by conduction.

Beyond the specifics of convection, the purpose of these notes is also to draw together a number of themes that we have covered in this course, and to show how a modelling problem (can we predict how convection works in a particular setting) can be tackled from first principles using an array of tools that we have already learnt. The material covered here is therefore more challenging than what we have done so far, in the sense that it requires you to draw more fluently on the things you have learnt, and challenges you to adapt what you know to new situations. There are, as usual, numerous explanatory notes and exercises to provide context and test your understanding.

We base our model for convection on flow in a porous medium, rather than the alternative (and perhaps more common) example of a tank of pure, single-phase fluid heated from below. The reason is largely practical: the model is somewhat simpler to analyze for a porous medium, while retaining the essential features of convection in a pure liquid.

Neither is convection in a porous medium an oddity: it occurs in hydrothermal systems, and explains the formation of geysers, and of black smokers on the sea floor. Many mineral deposits in metamorphic rock are the result of hydrothermal alteration, where convecting hot water selectively dissolves and transports some minerals at depth, and then deposits them where it cools down as it reaches layers of rock at higher elevations. Hydrothermal power plants can also rely on convection in groundwater systems.

Convection is also more challenging than the continuum physics problem we have encountered so far because it couples conservation of mass, momentum and energy.

The relevant model in a porous medium is the following: first, we ensure we conserve mass and momentum. As described in the notes on porous media, this corresponds to

$$\frac{\partial[\rho_s(1-\phi)]}{\partial t} + \nabla \cdot [\rho_s(1-\phi)\mathbf{u}] = 0 \tag{1a}$$

$$\frac{\partial(\rho\phi)}{\partial t} + \nabla \cdot (\rho\phi\mathbf{u}) + \nabla \cdot \mathbf{q} = 0 \tag{1b}$$

$$\mathbf{q} = \frac{\rho k}{\mu}(\rho\mathbf{g} - \nabla p) \tag{1c}$$

where ρ_s is the density of the pure solid phase (the porous matrix), ρ is pure fluid density, ϕ is porosity. In order to close the problem, we need to specify the two densities ρ and ρ_s, as well as some combination of porosity ϕ and matrix velocity $\mathbf{u}$. Our aim here is to construct a *minimal* model of convection, with all the essential physics included but devoid of unnecessary complication.

Convection requires variations in the density of the convecting fluid with temperature, and we choose model for ρ that involves a simple, constant thermal expansion coefficient:

$$\rho = \rho_0(1 - \alpha(T - T_b)) \tag{1d}$$

where T is temperature and T_b is a reference temperature, which we will later equate with the temperature of the bottom of the porous medium. Because the matrix is not involved in convection, we also choose a rigid matrix, with constant ρ_s and ϕ, and put

$$\mathbf{u} = \mathbf{0} \tag{1e}$$

for the matrix velocity.

The constitutive relation for fluid density ρ requires us to solve a heat equation for temperature T. Since we are considering a mixture of matrix and fluid, both of which can store internal energy. If we take $V(t)$ to be a Lagrangian volume with respect to the solid matrix and assume for the time being a constant heat capacity c_s and c for solid and fluid, respectively, then the internal energy content of $V(t)$ is

$$\int_{V(t)} [\rho_s c_s(1-\phi) + \rho c\phi]\, T \,\mathrm{d}V.$$

The rate at which internal energy is removed from $V(t)$ is controlled by two transport processes: first, conduction, which we assume follows Fourier's law with some effective thermal conductivity κ that accounts for the potentially different conductivities of pure matrix material and pure fluid. Second, the fluid that moves relative to the solid matrix with flux $\mathbf{q}$ carries internal energy with it. Since $\mathbf{q}$ is the rate at which mass flows, and heat content per unit mass of fluid is cT, the rate of heat transport out of the volume $V(t)$ is

$$\int_{S(t)} -k\nabla T \cdot \hat{\mathbf{n}} \,\mathrm{d}S + \int_{S(t)} cT\mathbf{q} \cdot \hat{\mathbf{n}} \,\mathrm{d}S$$

and therefore, since there is no heat produced in the domain

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{V(t)}\left[\rho_s c_s(1-\phi)+\rho c\phi\right]T\,\mathrm{d}V=-\int_{S(t)}-\kappa\nabla T\cdot\hat{\mathbf{n}}\,\mathrm{d}S-\int_{S(t)}cT\mathbf{q}\cdot\hat{\mathbf{n}}\,\mathrm{d}S.$$

On applying Reynolds' transport theorem with zero matrix velocity $\mathbf{u}$ (which is to say, $V(t)$ is just a fixed volume V), we get

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{V(t)}\left[\rho_s c_s(1-\phi)+\rho c\phi\right]T=\int_{V(t)}\frac{\partial}{\partial t}\left\{\left[\rho_s c_s(1-\phi)+\rho c\phi\right]T\right\}\,\mathrm{d}V$$

and applying the divergence theorem to the surface integrals, assuming a constant thermal conductivity κ (to distinguish it from the permeability k), this results in

$$\int_{V(t)}\frac{\partial}{\partial t}\left\{\left[\rho_s c_s(1-\phi)+\rho c\phi\right]T\right\}+\nabla\cdot(cT\mathbf{q})-\kappa\nabla^2 T\,\mathrm{d}V=0.$$

As usual, because the Lagrangian volume is ultimately arbitrary, the integrand cannot be positive in any finite region, nor can it be negative, so

$$\frac{\partial}{\partial t}\left\{\left[\rho_s c_s(1-\phi)+\rho c\phi\right]T\right\}+\nabla\cdot(cT\mathbf{q})-\kappa\nabla^2 T=0. \tag{1f}$$

Note that the term $cT\mathbf{q}$ is effectively an advective flux of heat, with $\mathbf{q}$ taking the role usually played by $\rho\mathbf{u}$ in a single-phase fluid.

We can simply this equation and conservation of mass (1b) somewhat. Recall that c and ϕ are constant, so we can write

$$\frac{\partial(\rho c\phi T)}{\partial t}+\nabla\cdot(cT\mathbf{q})=c\left(\phi\frac{\partial(\rho T)}{\partial t}+\nabla\cdot(\mathbf{q}T)\right)=c\left(\phi T\frac{\partial\rho}{\partial t}+\phi\rho\frac{\partial T}{\partial t}+T\nabla\cdot\mathbf{q}+\mathbf{q}\cdot\nabla T\right) \tag{1g}$$

by the chain rule. But conservation of mass (1b) with constant porosity and vanishing matrix velocity becomes

$$\phi\frac{\partial\rho}{\partial t}+\nabla\cdot\mathbf{q}=0 \tag{1h}$$

and hence two terms on the right-hand side cancel to give

$$\frac{\partial(\rho c\phi T)}{\partial t}+\nabla\cdot(cT\mathbf{q})=\rho\phi c\frac{\partial T}{\partial t}+\mathbf{q}c\cdot\nabla T.$$

If we also use the fact that $\rho_s c_s(1-\phi)$ is constant, the heat equation (1f) becomes more simply

$$(\rho_s c_s(1-\phi)+\rho c\phi)\frac{\partial T}{\partial t}++\mathbf{q}c\cdot\nabla T-\kappa\nabla^2 T=0. \tag{1i}$$

The model as posed is not complete: in fact, it is still missing the actual driver for convection, which is that the fluid must be heated from beneath. We pick a very

simple, two-dimensional domain in the xz-plane here to illustrate how convection works: the strip defined by $0 < z < h$, where we assume that this strip is horizontal, so

$$\mathbf{g} = -g\mathbf{k}.$$

To be definite, we also assume that the solutions are periodic in x with a period a. This is somewhat artificial (as there is no definite distance a after which the flow pattern and temperature should repeat, but it avoids the need to specify what happens as $x \to \infty$. At the upper and lower boundaries, we need to set boundary conditions on temperature, which we do as

$$T = T_s \text{ at } z = h, \qquad T = T_b \text{ at } z = 0. \tag{1j}$$

That is, we assume that the top and the bottom of the porous layer are kept at fixed temperatures ('heat baths' in thermodynamics), and to cause convection, we expect we need

$$T_s < T_b;$$

the fluid is warmer at the bottom than at the top. There are other choices one could consider: for instance, one could imagine the upper boundary being a heat bath like the ocean, while the lower boundary could experience a fixed geothermal heat flux. The fixed temperature boundary conditions here lead to a simpler analysis while illustrating the basic physics involved in convection.

In addition to temperature boundary conditions, we need boundary conditions on the flow. A boundary condition suppressing flow at the lower boundary is the most obvious choice: this could be an aquifer overlying 'impermeable' bedrock, in which there is no pore water flow. This corresponds to

$$\mathbf{q} \cdot \hat{\mathbf{n}} = \frac{\rho k}{\mu}\left(\rho g + \frac{\partial p}{\partial z}\right) = 0 \qquad \text{at } z = 0. \tag{1k}$$

At the top of the domain, one possible choice is a fixed fluid pressure: that for instance would apply most obviously to an open boundary with an ocean kept at fixed temperature T_s. It turns out that that makes for a more complicated analysis later, and a confined aquifer — bounded by impermeable walls at top and bottom — is easier to deal with. That means we would like to impose the same boundary condition $\mathbf{q} \cdot \hat{\mathbf{n}} = 0$ at $z = h$. We cannot, however, do that exactly: the problem is that it would require the volume of fluid to remain constant, since the size of the domain does not change and we have prevent any fluid from leaving the domain by prescribing zero flux on the upper and lower boundaries, and imposed periodic boundary conditions at the sides, so any fluid that leaves the domain on the right re-enters on the left. A fixed fluid volume is, however, at odds with the fact that the density will decrease as the fluid is heated: in particular, if we integrate (1b) over the domain V given by $0 < z < h$, $0 < x < a$ with $\phi =$ constant and $\mathbf{u} = \mathbf{0}$, then we get

by the divergence theorem

$$\int_V \phi \frac{\partial \rho}{\partial t} \, \mathrm{d}V + \int_S \mathbf{q} \cdot \mathbf{n} \, \mathrm{d}S = 0.$$

But $\mathbf{q} \cdot \mathbf{n} = 0$ at $z = 0$ by (1k), while periodic boundary conditions at the side boundaries at $x = 0$ and $x = a$ mean the surface integrals over these side boundaries cancel, so

$$\int_V \phi \frac{\partial \rho}{\partial t} \, \mathrm{d}V + \int_0^a \mathbf{q}|_{z=h} \cdot \hat{\mathbf{n}} \, \mathrm{d}x = 0.$$

so $\mathbf{q} \cdot \hat{\mathbf{n}}$ at $z = h$ cannot vanish if ρ is allowed to change over time, even though we anticipate these density changes will be small.

A slightly awkward device that gets around this problem is to suppose that the top boundary is a thin layer of very impermeable material, and that there is fixed pressure p_s above that layer. We can then impose a model that relates the rate of mass loss through the top boundary to the difference between fluid pressure in the domain at the top boundary and the fixed pressure outside as

$$\mathbf{q} \cdot \hat{\mathbf{n}} = \frac{\rho k}{\mu} \left(-\rho g + \frac{\partial p}{\partial z} \right) = k_0 (p - p_s) \qquad \text{at } z = 0. \tag{1l}$$

where k_0 is a permeability-like constant for the upper boundary. We will make k_0 very small, so as to approximate the condition $\mathbf{q} \cdot \hat{\mathbf{n}} = 0$.

The Boussinesq approximation

The model above is more complicated than it needs to be, and we will non-dimensionalize it below in order to derive a simpler, approximate version. The basis for this simplification, known as the Boussinesq approximation, is the limit of a small thermal expansion coefficient α, and a small top layer expansion coefficient k_0.

Before we do so, we compact the model stated in the previous section, omitting the equation for conservation of matrix mass, since the latter is satisfied automatically. We also employ a trick to simplify our notation later: we expect that the pressure gradient ∇p in (1c) mostly balances the gravitational body force, which is given by $-\rho_0 g \mathbf{k}$ plus a much smaller correction: that correction $-\rho_0 \alpha (T - T_b) g \mathbf{k}$ is due to the fact that the density varies as temperature does. It is also the reason why flow occurs, and we expect that the total hydraulic gradient

$$-\rho g \mathbf{k} - \nabla p$$

is comparable in size to the small correction $-\rho_0 \alpha (T - T_b) g \mathbf{k}$, rather than to the much larger body force $-\rho_0 g \mathbf{k}$ that excludes temperature effects. To account for this, we define a *reduced pressure* p' through

$$p' = p - p_s - \rho_0 g (h - z),$$

that is, we remove from the actual pressure variable the hydrostatic pressure variation $\rho_0 g(h-z)$ that results from the reference density ρ_0. We also remove the constant 'outside' pressure p_s that appears in (1l) for convenience. For convenience, we also define a reduced temperature

$$T' = T - T_b.$$

The convection model then reduces to

$$\phi\frac{\partial \rho}{\partial t} + \nabla\cdot\mathbf{q} = 0 \qquad \text{for } 0<z<h, \tag{2a}$$

$$[\rho_s c_s(1-\phi) + \rho c\phi]\frac{\partial T'}{\partial t} + +\mathbf{q}c\cdot\nabla T' - \kappa\nabla^2 T' = 0 \qquad \text{for } 0<z<h, \tag{2b}$$

$$\mathbf{q} = \frac{\rho k}{\mu}\left(\rho_0\alpha T' g\mathbf{k} - \nabla p'\right) \qquad \text{for } 0<z<h, \tag{2c}$$

$$\rho = \rho_0(1-\alpha(T-T_b)) \qquad \text{for } 0<z<h, \tag{2d}$$

$$T' = 0 \qquad \text{at } z=0, \tag{2e}$$

$$T' = -(T_b - T_s) \qquad \text{at } z=h, \tag{2f}$$

$$\frac{\rho k}{\mu}\left(\rho_0\alpha T' g - \frac{\partial p'}{\partial z}\right) = 0 \qquad \text{at } z=0, \tag{2g}$$

$$\frac{\rho k}{\mu}\left(\rho_0\alpha T' g - \frac{\partial p'}{\partial z}\right) = k_0 p' \qquad \text{at } z=0, \tag{2h}$$

To non-dimensionalize, introduce scales $[x]$, $[t]$, $[q]$, $[\rho]$, $[T]$, $[p]$, and write

$$(x,z) = [x](x^*,z^*), \quad t=[t]t^*, \quad \mathbf{q}=[q]\mathbf{q}^*, \quad \rho=[\rho]\rho^*, \quad T'=[T]T^*, \quad p'=[p]p^*.$$

Since this is not the first time we non-dimensionalize a problem here, we do not display all the details here. Suffice it to say that (2) can be written in the following form

$$r\frac{\partial \rho^*}{\partial t^*} + \nabla^*\cdot\mathbf{q}^* = 0 \qquad \text{for } 0<z^*<1, \tag{3a}$$

$$Ra\left([1+r(\rho^*-1)]\frac{\partial T^*}{\partial t^*}\} + \mathbf{q}^*\cdot\nabla T^*\right) - \nabla^{*2}T^* = 0 \qquad \text{for } 0<z^*<1, \tag{3b}$$

$$\mathbf{q}^* = \rho^*\left(T^*\mathbf{k} - \nabla^* p^*\right) \qquad \text{for } 0<z^*<1, \tag{3c}$$

$$\rho^* = 1-\delta T^* \qquad \text{for } 0<z^*<1, \tag{3d}$$

$$T^* = 0 \qquad \text{at } z^*=0, \tag{3e}$$

$$T^* = -1 \qquad \text{at } z^*=1, \tag{3f}$$

$$\rho^*\left(T^* - \frac{\partial p^*}{\partial z^*}\right) = 0 \qquad \text{at } z^*=0, \tag{3g}$$

$$\rho^*\left(T^* - \frac{\partial p^*}{\partial z^*}\right) = \nu p^* \qquad \text{at } z^*=1, \tag{3h}$$

if we choose $[x] = h$, $[\rho] = \rho_0$, $[p] = [\rho]gh$, $[T] = T_b - T_s$, $[q] = [\rho]k\rho_0\alpha[T]g/\mu$, $[t] = (\rho_s c_s(1-\phi) + \rho c\phi)h/(c[q])$. In addition, the dimensionless parameters are

$$\delta = \alpha(T_b - T_s), \qquad \nu = \frac{k_0\mu h}{\rho_0\alpha(T_b - T_s)}, \qquad r = \frac{\rho_0 c\phi}{\rho_s c_s(1-\phi) + \rho_0 c\phi},$$

$$Ra = \frac{kg\rho_0^2 c\alpha(T_b - T_s)h}{\mu\kappa}. \tag{4}$$

Ra is the *Rayleigh number* for the porous medium, and measures the strength of advection relative to conduction in the heat equation. (In many other settings, this would be denoted as a *Péclet number*.)

Exercise 1 *Verify the non-dimensionalisation that leads to* (3)*, including the choice of scales and definition of the dimensionless parameters.*

The Boussinesq approximation is what we get if we assume that δ is very small (which we write as $\delta \ll 1$): while the fluid expands thermally, it does so quite weakly. The point here will be that the variations in density do affect the flow of the fluid through the body force $\rho^*\mathbf{k}$ in (3c), but that the fluid density can be treated as constant when we look at conservation of mass (3a) and of energy (3b). Recall also that we only introduced the leakage term $k_0(p - p_s)$ in (1l) only because this is necessary in principle to allow the fluid to expand in a domain of fixed size; demanding that that leakage term is insignificant will mean here that we also assume that $\nu \ll 1$.

With these assumptions in place, we can drop terms multiplied by δ and ν in (3). This primarily results in the simplification

$$\rho^* = 1;$$

in other words, as previously advertised, we treat the model as having a constant density everywhere than in the body force term. This leads to the following simplified model, which we will use from here on:

$$\begin{aligned}
\nabla\cdot\mathbf{q} &= 0 && \text{for } 0 < z < 1, && \text{(5a)}\\
Ra\left(\frac{\partial T}{\partial t} + \mathbf{q}\cdot\nabla T\right) - \nabla^2 T &= 0 && \text{for } 0 < z < 1, && \text{(5b)}\\
\mathbf{q} &= T\mathbf{k} - \nabla p && \text{for } 0 < z < 1, && \text{(5c)}\\
\rho &= 1 && \text{for } 0 < z < 1, && \text{(5d)}\\
T &= 0 && \text{at } z = 0,, && \text{(5e)}\\
T &= -1 && \text{at } z = 1, && \text{(5f)}\\
T - \frac{\partial p}{\partial z} &= 0 && \text{at } z = 0,\ 1. && \text{(5g)}
\end{aligned}$$

where the effect of variable density is represented purely by the term $-T\mathbf{k}$ in (5c) and (equivalently, since the left-hand side is the z-component of $\mathbf{q}$) by the term T in (5g)).

Observe that we have not only set $\rho = 1$ in (5), we have also quietly dropped the asterisks on the dimensionless variables. This is in fact very commonly done in practice, to simplify the notation and avoid carrying ever more 'decorations' on variable symbols. There is rarely any risk of confusing a dimensionless variable with its dimensional counterpart, since the absence of obvious dimensional variables usually marks out the scaled model equations from their original counterparts.

The fact that (5) also has the major advantage that it contains only a single dimensionless parameter, the Rayleigh number Ra, so model behaviour depends only on that one parameter, rather than the many original, dimensional parameters like ρ_0, α, T_s, T_b, k, μ, ϕ, g. Aside from the ability to lead to systematic approximations, this reduction in the size of the parameter space is the real power of non-dimensionalisation.

Steady state solution and linearization

The model (5) is *nonlinear*: equation (5b) includes the term $\mathbf{q} \cdot \nabla T$, which does not have the properties of a linear operator as previously defined. This puts an 'analytical' general solution (one you can write down with pencil and paper) of (5) out of reach, but it does not make further analysis impossible.

The first question we ask is whether the problem (5) has a steady state solution with no spatial structure in the lateral (x-) direction, meaning that the dependent variables $\mathbf{q}$, p and T are functions of z only. The answer, quite trivially, is 'yes'.

For future simplicity of notation, we add an overbar to the variable names $\mathbf{q}$, p and T to denote the steady state, so $\mathbf{q} = \bar{\mathbf{q}}(z)$, $p = \bar{p}(z)$ and $T = \bar{T}(z)$ in steady state. If $\bar{p}$ and $\bar{T}$ are functions of z only, then (5c) becomes

$$\mathbf{q} = \left(\bar{T} - \frac{\partial \bar{p}}{\partial z}\right)\mathbf{k}$$

and (5a) implies that $-(\bar{T} + \partial\bar{p}/\partial z)$ is a constant; from (5g), we conclude that that constant is zero

$$\bar{T} - \frac{\mathrm{d}\bar{p}}{\mathrm{d}z} = 0, \tag{6}$$

so there is no mass movement in steady state, as is to be expected with zero flux boundary condition at the top and the bottom surface:

$$\bar{\mathbf{q}} = 0,$$

The heat equation (5b) becomes

$$-\frac{\mathrm{d}^2\bar{T}}{\mathrm{d}z^2} = 0$$

subject to (5e) and (5f), giving

$$\bar{T}(z) = -z. \tag{7a}$$

Therefore (6) leads to

$$\bar{p}(z) = -\frac{z^2}{2} \tag{7b}$$

plus a constant whose value is immaterial, since the model (5) is unchanged if we add a constant to p.[1]

The next step is to asks what happens to small deviations from the steady state: if I nudge p, T and $\mathbf{q}$ slightly away from their steady state forms $\bar{p}$, $\bar{T}$ and $\bar{\mathbf{q}}$, will they evolve away from, or back towards, their original steady state? The procedure for doing this is to put

$$p(x,z,t) = \bar{p}(z) + \varepsilon p'(x,z,t), \qquad T(x,z,t) = \bar{T}(z) + \varepsilon T'(x,z,t),$$

$$\mathbf{q}(x,z,t) = \bar{\mathbf{q}}(z) + \varepsilon \mathbf{q}'(x,z,t),$$

and to solve for the evolution of the *perturbations* p', T' and $\mathbf{q}'$ in time. A word on notation: the primes on p', T' and $\mathbf{q}'$ do not denote differentiation, but are just a customary indication that these are perturbations away from a steady state.[2] The parameter ε is really just there formally to indicate that the perturbation is small in size: we are almost at the steady state, but not quite. It is not to be confused with the dimensionless groups derived earlier.

There will be quite a few equations that follow, all restating (5) in progressively modified form. This will probably seem like an overly complicated and arduous procedure, because we will lay out all the steps in detail. This may seem intimidating at first, but actually each individual step is fairly minor. Once you understand the procedure we are following, some of this detail will be redundant.

If we substitute in (5), we get

$$\nabla \cdot (\bar{\mathbf{q}} + \varepsilon \mathbf{q}') = 0 \quad \text{for } 0 < z < 1, \tag{8a}$$

$$Ra\left(\frac{\partial(\bar{T} + \varepsilon T')}{\partial t} + (\bar{\mathbf{q}} + \varepsilon \mathbf{q}') \cdot \nabla(\bar{T} + \varepsilon T')\right) - \nabla^2(\bar{T} + \varepsilon T') = 0 \quad \text{for } 0 < z < 1, \tag{8b}$$

[1] A variable that can be changed by adding a constant without affecting the equations it solves is also known as a *gauge variable*.

[2] Note that we have also recycled the prime decoration from a different earlier use, when p' was a reduced pressure, with the hydrostatic contribution $\rho_0 g z$ removed, and T' was similarly a reduced temperature, measured relative to a baseline temperature T_b. The meaning of those earlier, dimensional variables is entirely distinct from the T' and p' we use here. Bad practice, perhaps, but also an attempt to avoid what is sometimes called 'excessive notation'.

$$\bar{\mathbf{q}} + \varepsilon\mathbf{q}' = \left[(\bar{T} + \varepsilon T')\mathbf{k} - \nabla(\bar{p} + \varepsilon p')\right] \qquad \text{for } 0 < z < 1, \tag{8c}$$

$$\bar{T} + \varepsilon T' = 0 \qquad \text{at } z = 0, , \tag{8d}$$

$$\bar{T} + \varepsilon T' = -1 \qquad \text{at } z = 1, \tag{8e}$$

$$\bar{T} + \varepsilon T' - \frac{\partial(\bar{p} + \varepsilon p')}{\partial z} = 0 \qquad \text{at } z = 0,\ 1. \tag{8f}$$

The next step is to separate out terms that have factors of 1 (which is ε^0) from those that are multiplied by ε, ε^2, etc. This is trivial for *linear* terms, since we simply have, for instance,

$$\frac{\partial(\bar{T} + \varepsilon T')}{\partial t} = \frac{\partial \bar{T}}{\partial t} + \varepsilon \frac{\partial T'}{\partial t}$$

and we can similarly factorize ε for other linear terms. The only term for which this is not the case is the nonlinear advection term

$$(\bar{\mathbf{q}} + \varepsilon\mathbf{q}') \cdot \nabla(\bar{T} + \varepsilon T')$$

where we need to expand before we can separate terms with different powers of ε. Since we have a simple product of two terms, this expansion is straightforward, and we get

$$(\bar{\mathbf{q}} + \varepsilon\mathbf{q}') \cdot \nabla(\bar{T} + \varepsilon T') = \bar{\mathbf{q}} \cdot \nabla\bar{T} + \varepsilon\left(\bar{\mathbf{q}} \cdot \nabla T' + \mathbf{q}' \cdot \nabla\bar{T}\right) + \varepsilon^2 \mathbf{q}' \cdot \nabla T'.$$

If we were to apply the same procedure to other, more complex problems, we might have to resort to a Taylor expansion of nonlinear terms instead.

Ploughing ahead and separating out terms in (8), we get

$$\nabla \cdot \bar{\mathbf{q}} + \varepsilon \nabla \cdot \mathbf{q}' = 0 \quad \text{for } 0 < z < 1, \tag{9a}$$

$$Ra\left(\frac{\partial \bar{T}}{\partial t} + \bar{\mathbf{q}} \cdot \nabla\bar{T}\right) - \nabla^2\bar{T} + \varepsilon\left\{Ra\left(\frac{\partial T'}{\partial t} + \mathbf{q}' \cdot \nabla\bar{T} + \bar{\mathbf{q}} \cdot \nabla T')\right) - \nabla^2 T'\right\} + \varepsilon^2 Ra\mathbf{q}' \cdot \nabla T' = 0 \quad \text{for } 0 < z < 1, \tag{9b}$$

$$\bar{\mathbf{q}} + \varepsilon\mathbf{q}' = \left(\bar{T}\mathbf{k} - \nabla\bar{p}\right) + \varepsilon\left(T'\mathbf{k} - \nabla p'\right) \qquad \text{for } 0 < z < 1, \tag{9c}$$

$$\bar{T} + \varepsilon T' = 0 \qquad \text{at } z = 0, \tag{9d}$$

$$\bar{T} + \varepsilon T' = -1 \qquad \text{at } z = 1, \tag{9e}$$

$$\bar{T} - \frac{\partial \bar{p}}{\partial z} + \varepsilon\left(T' - \frac{\partial p'}{\partial z}\right) = 0 \qquad \text{at } z = 0,\ 1. \tag{9f}$$

Now look at the terms that do not have a coefficient of ε (these are typically called the 'zeroth order' or 'leading order' terms). If you look at the solutions for $\bar{\mathbf{q}}$, $\bar{p}$ and

$\bar{T}$, you will find that, in each equation, the zeroth order terms cancel exactly. That is, they satisfy

$$\nabla \cdot \bar{\mathbf{q}} = 0 \qquad \text{for } 0 < z < 1, \tag{10a}$$

$$Ra\left(\frac{\partial \bar{T}}{\partial t} + \bar{\mathbf{q}} \cdot \nabla \bar{T}\right) - \nabla^2 \bar{T} + = 0 \qquad \text{for } 0 < z < 1, \tag{10b}$$

$$\bar{\mathbf{q}} = \bar{T}\mathbf{k} - \nabla \bar{p} \qquad \text{for } 0 < z < 1, \tag{10c}$$

$$\bar{T} = 0 \qquad \text{at } z = 0, \tag{10d}$$

$$\bar{T} = -1 \qquad \text{at } z = 1, \tag{10e}$$

$$\bar{T} - \frac{\partial \bar{p}}{\partial z} = 0 \qquad \text{at } z = 0,\, 1. \tag{10f}$$

The fact that they do so is not an accident: the steady state solution $\bar{\mathbf{q}}$, $\bar{p}$ and $\bar{T}$ by construction satisfies the original problem (5), and therefore (10) (which is nothing more than (5) with $\mathbf{q}$, p and T replaced by $\bar{\mathbf{q}}$, $\bar{p}$ and $\bar{T}$.)

Given that, we lose the zeroth order terms in (9) and retain only those terms that have a coefficient ε or a higher order power of ε (specifically, because the nonlinear term $\nabla \cdot (T\mathbf{q})$ in (5) is a simple product, we get a power ε^2). If we divide all equations in (9) by ε, we therefore end up with

$$\nabla \cdot \mathbf{q}' = 0 \quad \text{for } 0 < z < 1, \tag{11a}$$

$$Ra\left(\frac{\partial T'}{\partial t} + \bar{\mathbf{q}} \cdot \nabla T' + \mathbf{q}' \cdot \nabla \bar{T}\right) - \nabla^2 T' + \varepsilon Ra \mathbf{q}' \cdot \nabla T' = 0 \quad \text{for } 0 < z < 1, \tag{11b}$$

$$\mathbf{q}' = T'\mathbf{k} - \nabla p' \qquad \text{for } 0 < z < 1, \tag{11c}$$

$$T' = 0 \qquad \text{at } z = 0,\, 1 \tag{11d}$$

$$T' - \frac{\partial p'}{\partial z} = 0 \qquad \text{at } z = 0,\, 1. \tag{11e}$$

All that we do now is insist that we can ignore higher order terms in ε in (11) on the basis that ε is small, which allows us to drop the nonlinear term $\mathbf{q}' \cdot \nabla T'$. If we also substitute for $\bar{\mathbf{q}} = \mathbf{0}$ and $\bar{T} = -z$, and we arrive at the following *linearized* model for the perturbations $\mathbf{q}'$, p' and T':

$$\nabla \cdot \mathbf{q}' = 0 \qquad \text{for } 0 < z < 1, \tag{12a}$$

$$Ra\left(\frac{\partial T'}{\partial t} - \mathbf{q}' \cdot \mathbf{k}\right) - \nabla^2 T' = 0 \qquad \text{for } 0 < z < 1, \tag{12b}$$

$$\mathbf{q}' = T'\mathbf{k} - \nabla p' \qquad \text{for } 0 < z < 1, \tag{12c}$$

$$T' = 0 \qquad \text{at } z = 0,\, 1 \tag{12d}$$

$$T' - \frac{\partial p'}{\partial z} = 0 \qquad \text{at } z = 0,\, 1. \tag{12e}$$

where we have made use of the fact that $\nabla\bar{T} = \nabla(-z) = -\mathbf{k}$.

Being linear and having constant coefficient,[3] this set of equations has a hope of being solveable by pencil-and-paper methods. That is, in a sense, the power of looking at small perturbations: we arrive at a model that is linear and solveable, from which we can therefore determine whether these perturbations grow. As we will see shortly, if growth occurs, it is *unbounded* in the linear model: the perturbations simply get bigger and bigger. This will eventually make the approximation scheme we have employed here invalid: dropping the nonlinear term was justified by assuming that T' and $\mathbf{q}'$ are comparable in size to unity, and ε is small. Once T' and $\mathbf{q}'$ get large enough that their product is comparable in size to ε^{-2}, neglecting the nonlinear term is no longer a valid approximation. The last section of these notes deal with nonlinear effects.

The value of the linearization is therefore in determining whether the small perturbations added to the steady state solution will grow or not, rather than in determining what size they grow to and whether inital growth eventually stops and gives rise to a steady convection pattern. That requires the solution of the full problem (5), which usually requires computational methods (though there are some advanced analytical methods that can be used under certain circumstances, described at the end of these notes).

We will focus on solving the linearized problem (12) in the bulk of these notes. This is called a *linear stability analysis*. Before we move onto that task, the note and exercises below give further context to the linearization procedure followed above, but are not strictly required to follow the main part of these notes.

Note 1 *The steps taken to arrive at the linearized convection model* (12) *probably obscure the essence of some of what is really going on here, so this note takes a look at an analogous, more abstract problem. An* autonomous dynamical system *is basically a set of coupled, first order linear differential equations, which we can write in the form*

$$\frac{\mathrm{d}y_i}{\mathrm{d}t} = F_i(y_1, y_2, \ldots, y_n), \qquad i = 1, \ldots, n \tag{13}$$

where each F_i is some known function.[4]

In more classical vector notation

$$\frac{\mathrm{d}\mathbf{y}}{\mathrm{d}t} = \mathbf{F}(\mathbf{y}).$$

Note that boldface letters here denote a vector like $\mathbf{y} = (y_1, y_2, \ldots, y_n)^{\mathrm{T}}$ of arbitrary dimension n-by-1, not necessarily a two- or three-dimensional vector that may have

[3] meaning, there are no coefficients depending on position (x, z) that multiply any of the terms, since Ra is a constant

[4] The 'autonomous' moniker in the label 'autonomous dynamical system' refers to the fact that none of the functions F_i depend on t.

an interpretation as physical vector. $\mathbf{F}$ *denotes a vector-valued function of a vector-valued argument* $\mathbf{y}$*, and is just a shorthand for the index notation in* (13)*. The components of* y_i *of the vector* $\mathbf{y}$ *are often known as* degrees of freedom*, especially if the dynamical system represents a mechanical system, in which the* y_i *describe the motion of some set of objects.*

Many problems can be cast as dynamical systems. For instance, Newton's second law for an object of mass m *with position* $x(t)$ *in one dimension, subject to a force* $f = f(x)$ *that depends only on its position* x*, satisfies*

$$m\frac{\mathrm{d}^2x}{\mathrm{d}t^2} = f(x).$$

This can be written as a system of equations

$$\frac{\mathrm{d}x}{\mathrm{d}t} = v, \qquad \frac{\mathrm{d}v}{\mathrm{d}t} = m^{-1}f(x),$$

which is of the form (13) *if we put* $n = 2$*, with* $y_1 = x$*,* $y_2 = v$*,* $F_1(y_1, y_2) = y_2 = v$*,* $F_2(y_1, y_2) = m^{-1}f(y_1) = m^{-1}f(x)$*. We will show later that the convection problem* (5) *can be thought of as an dynamical system, although with an infinite number of dimensions* n*.*

The study of dynamical systems has many uses, but one basic concept is the stability of steady states. A steady state solution, which we denote again by an overbar, solves

$$F_i(\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_n) = 0 \qquad \textit{for } i = 1, \ldots, n. \tag{14}$$

There being n *equations for* n *unknowns* $\bar{y}_1$*,* $\bar{y}_2$*,* $\ldots \bar{y}_n$*, the steady state solution is generally well-defined, if not necessarily unique.*

Whether the steady state is stable is determined by whether small perturbations grow, and the procedure involved is called a linear stability analysis. As in the convection problem, we put

$$y_1 = \bar{y}_1 + \varepsilon y_1', \qquad y_2 = \bar{y}_2 + \varepsilon y_2', \ldots, y_n = \bar{y}_n + \varepsilon y_n'$$

Substituting into (13) *and using the fact that the* $\bar{y}_i$ *are components of a steady-state solution, we get*

$$\varepsilon\frac{\mathrm{d}y_i'}{\mathrm{d}t} = F_i(\bar{y}_1 + \varepsilon y_1', \bar{y}_2 + \varepsilon y_2, \ldots, \bar{y}_n + \varepsilon y_n'), \qquad \textit{for } i = 1, 2, \ldots, n.$$

The right-hand side can be expanded up to linear order in ε *using a multivariable Taylor expansion,*

$$F(\bar{y}_1 + \varepsilon y_1', \bar{y}_2 + \varepsilon y_2, \ldots, \bar{y}_n + \varepsilon y_n') = F_i(\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_n) + \sum_{j=1}^{n} \left.\frac{\partial F_i}{\partial y_j}\right|_{(\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_n)} \varepsilon y_j' + O(\varepsilon^2), \tag{15}$$

where the notation $g|_{(\bar{y}_1,\bar{y}_2,\ldots,\bar{y}_n)}$ *indicates that the function* g *is evaluated at* $(\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_n)$ *and the notation* $O(\varepsilon^2)$ *denotes that the error in not expanding further as a Taylor series is comparable to* ε^2 *in size.*[5]

But from (14)*, the first term on the right-hand side is zero, and therefore*

$$\frac{\mathrm{d}y_i'}{\mathrm{d}t} = \sum_{j=1}^{n} J_{ij} y_j' \qquad \textit{for } i = 1, 2, \ldots, n, \tag{16}$$

where J_{ij} *is the constant matrix*

$$J_{ij} = \left.\frac{\partial F_i}{\partial y_j}\right|_{(\bar{y}_1,\bar{y}_2,\ldots,\bar{y}_n)},$$

known as the Jacobian *of* F_i*. Note that* (16) *is a linear equation, with two important attributes: first it is solveable in a form that we simply write down, and secondly, owing to linearity, we can simply add different solutions to* (16)*, and obtain a new solution.*

If the dynamical system is one-dimensional ($n = 1$*), then* (16) *is just*

$$\frac{\mathrm{d}y'}{\mathrm{d}t} = Jy'$$

where, because there is only one index $i = 1$*, we can drop that index altogether. By separation of variables,*

$$y' = c\exp(Jt).$$

It is then straightforward to see that that instability occurs (the perturbation y' *grows) if* $J > 0$*, and the solution is stable and perturbations shrink to zero over time if* $J < 0$.[6]

For $n > 1$*, a basic understanding of linear algebra is needed to make further prorgess. Solutions to* (16) *generally take the analogous form*

$$y_i' = c_i \exp(\sigma t). \tag{17}$$

where the coefficients sigma and c_i *satisfy*

$$\sigma c_i = \sum_{j=1}^{n} J_{ij} c_j \qquad \textit{for } i = 1, 2, \ldots, n.$$

[5]Technically, the notation $O(\varepsilon^2)$ means the following: Denote the omitted term by E. To say that $E = O(\varepsilon^2)$ indicates that, as $\varepsilon \to 0$, E/ε^2 remains bounded, so E has to go to zero at least as fast as ε^2.

[6]Stability for the marginal case $J = 0$ cannot be determined by the linearization procedure used here.

This follows from plain substitution of (17) *into* (16)*, and is better understood if written in classical matrix notation as*

$$\sigma \mathbf{c} = \mathbf{J}\mathbf{c} \tag{18}$$

where $\mathbf{c}$ *is a column vector and* $\mathbf{J}$ *is the Jacobian matrix, written in index-free form.*

Equation (18) *is an* eigenvalue problem*, and in general an* $n \times n$ *matrix will have* n *eigenvalues, though they may not all be distinct.*[7] *These eigenvalues are the roots of the* n*-th order* characteristic polynomial *in* σ *formed when expanding the determinant*

$$\det\left(\sigma \mathbf{I} - \mathbf{J}\right) = 0$$

where $\mathbf{I}$ *is the* $n \times n$ *identity matrix.*

In fact, in view of linearity, the solution (16) *consists of sums of such terms of the form* $y_i \exp(\sigma t)$*. If we give the* n *eigenvalues of* J_{ij} *their own label* σ_k*,* $k = 1, \ldots, m$*, then the general solution to* (16) *is*

$$y_i'(t) = \sum_{j=1}^{n} \alpha_j c_{ij} \exp(\sigma_j t) \tag{19}$$

when the eigenvalues are all distinct (see exercise 5 of the notes on Fourier series for the case of repeated eigenvalues, which may introduce polynomials in t *multiplying the exponential* $\exp(\sigma t)$ *for a repeated eigenvalue* σ*). Here,* c_{ij} *is the* i*th component of the eigenvector* $\mathbf{c}_j$ *associated with eigenvalue* σ_j*, and* α_j *is a coefficient that depends on the initial condition for the* y_i'*.*

The steady state solution is stable if all *eigenvalues of* $\mathbf{J}$ *have negative real part, which corresponds to exponential decay. This is straightforward to understand if the* σ*'s are negative real numbers, in which case all terms* $c_i \exp(-\sigma t)$ *decay away.*

If any of the roots of the characteristic polynomial are complex numbers then, because the coefficients of the polynomial are real, they occur in complex conjugate pairs. For each complex σ*, you can find another that is its complex conjugate. If, as is reasonable, we insist that solutions* y' *must be real, then for each complex eigenvalue* σ *and its associated eigenvector* $\mathbf{c}$*, the conjugate eigenvalue* $\bar{\sigma}$ *corresponds to a conjugate eigenvector* $\bar{\mathbf{c}}$*, and the sum of the two is real: take*

$$c_i \exp(\sigma t) + \bar{c}_i \exp(\bar{\sigma} t)$$

If we write c_i *in polar form* $c_i = C_i \exp(i\theta_i)$*, so* $\bar{c}_i = C_i \exp(-i\theta_i)$*, then*

$$\begin{aligned} c_i \exp(\sigma t) + \bar{c}_i \exp(\bar{\sigma} t) =& C_i \left[\exp(\mathrm{Re}(\sigma)t + i\mathrm{Im}(\sigma)t + i\theta_i) + \exp(\mathrm{Re}(\sigma)t - i\mathrm{Im}(\sigma)t - i\theta_i)\right] \\ =& C_i \exp(\mathrm{Re}(\sigma)t) \left[\exp[i(\mathrm{Im}(\sigma)t + \theta_i)] - \exp[-i(\mathrm{Im}(\sigma)t + \theta_i)]\right) \\ =& C_i \exp(\mathrm{Re}(\sigma)t) \cos(\mathrm{Im}(\sigma)t + \theta_i) \end{aligned}$$

[7] In the case of repeated eigenvalues, (17) may need to be replaced by a polynomial in t multiplying the exponential $\exp(\sigma t)$, see exercise 5 of the notes on Fourier series.

and decay of this oscillatory solution will occur if $\mathrm{Re}(\sigma) < 0$.

For instability to occur, by contrast, it suffices for just one *eigenvalue to have positive real part, since there is then a relevant solution that will grow (either as a constant times* $c_i \exp(\sigma t)$ *if* σ *is simply real, or* $C_i \exp(\mathrm{Re}(\sigma)T)\cos(\mathrm{Im}(\sigma)t + \theta_i)$ *if there is a complex conjugate eigenvalue pair.)*

Of course, as noted in the main text, unbounded exponential growth of the perturbation y_i' *implies that the linearization* (15) *eventually fails, since the quadratic term in the Taylor series is of the form*

$$\sum_{j=1}^{n}\sum_{k=1}^{n} \frac{1}{2}\frac{\partial^2 F_i}{\partial x_j \partial x_k}\bigg|_{(\bar{y}_1,\bar{y}_2,\ldots,\bar{y}_n)} \varepsilon^2 y_j' y_k',$$

which can be dismissed as small (the $O(\varepsilon^2)$ *omitted term in* (15)*) when* y_j' *and* y_k' *are comparable to unity, but becomes similar in size to the terms retained in* (15) *when* y_i' *and* y_j' *are comparable to* ε^{-1}

Exercise 2 *Identify which of the steps from equations* (5)–(12) *correspond to which steps in* (13)–(16).

Exercise 3 *In general, the solution of the eigenvalue problem* (18) *is not possible in 'closed form' (by pencil and paper), because doing so would require the solution of an* n*th order polynomial.* $n = 2$ *is an exception, because it only requires the solution of a quadratic. A concrete example is the following problem:*

$$\frac{\mathrm{d}S}{\mathrm{d}t} = c_1 S^\alpha + v_0 - c_2 S(h_0 - h), \tag{20a}$$

$$\frac{\mathrm{d}h}{\mathrm{d}t} = q_0 - c_3 S^\alpha, \tag{20b}$$

where α, c_1, c_2, c_3, v_0, h_0 *and* q_0 *are parameters (and therefore constant), with* $\alpha > 1$. *(For what it is worth, this is a simple model for the evolution of a glacier-dammed lake, where* h *is lake level and* S *is the cross-sectional area of a sub-ice channel that drains the lake, with* q_0 *being the rate of inflow of water from upstream into the lake).*

Find the steady state solution to (20)*, and linearize the dynamical system. Find its eigenvalues and eigenvectors. Where the eigenvalues are complex, show that they form a complex conjugate pair, and that the corresponding eigenvectors can also be written as complex conjugates of each other. Show that the steady state is unstable if*

$$q_0 > \frac{c_3 v_0}{c_1(\alpha - 1)}$$

and stable if the inequality is reversed. As a bonus exercise, you can try to solve the system (20) *numerically, with an initial condition* near *the steady state, to confirm the stability result: many scientific computing packages such as MATLAB or SciPy have numerical initial value solvers inbuilt into them that make this task easy:*

`ode45`

or similar in MATLAB,

`scipy.integrate.solve_ivp`

in Python. If you do solve the problem numerically, then you can also answer the following problem: when there is instability, what type of behaviour does the solution (S, h) *evolve towards?*

Fourier series solution

We can simplify (12) somewhat further by substituting for $\mathbf{q}'$, and using the boundary conditions on T' to simplify those on flux further:

$$\frac{\partial T'}{\partial z} - \nabla^2 p' = 0 \qquad \text{for } 0 < z < 1, \tag{21a}$$

$$Ra\left(\frac{\partial T'}{\partial t} - T' + \frac{\partial p'}{\partial z}\right) - \nabla^2 T' = 0 \qquad \text{for } 0 < z < 1, \tag{21b}$$

$$T' = 0 \qquad \text{at } z = 0,\ 1 \tag{21c}$$

$$\frac{\partial p'}{\partial z} = 0 \qquad \text{at } z = 0,\ 1. \tag{21d}$$

We previously specified periodic boundary conditions in x, with a somewhat arbitrary period a. This suggests that we should represent T' and p' as Fourier series in x:

$$T'(x, z, t) = \sum_{n=-\infty}^{\infty} T_n(z, t)\phi_n(x), \qquad p'(x, z, t) = \sum_{n=-\infty}^{\infty} p_n(z, t)\phi_n(x), \tag{22}$$

where

$$\phi_n(x) = \exp(ik_{xn}x), \qquad k_{xn} = \frac{2n\pi}{a}. \tag{23}$$

Note that we have added an additional subscript 'x' to the notation for the wavenumber for reasons that will become obvious shortly.

Substituting into (21) leads to

$$\sum_{n=-\infty}^{\infty}\left[\frac{\partial T_n}{\partial z}-\left(\frac{\partial^2 p_n}{\partial z^2}-k_{xn}^2 p_n\right)\right]\phi_n=0 \qquad \text{for } 0<z<1,$$

$$\sum_{n=-\infty}^{\infty}\left[\,Ra\left(\frac{\partial T_n}{\partial t}-T_n+\frac{\partial p_n}{\partial z}\right)-\left(\frac{\partial^2 T_n}{\partial z^2}-k_{xn}^2 T_n\right)\right]\phi_n=0 \qquad \text{for } 0<z<1,$$

$$\sum_{n=-\infty}^{\infty} T_n\phi_n=0 \qquad \text{at } z=0,\,1,$$

$$\sum_{n-\infty}^{\infty}\frac{\partial p_n}{\partial z}\phi_n=0 \qquad \text{at } z=0,\,1.$$

We can project onto individual Fourier modes by multiplying by $\overline{\phi_m(x)}$ and integrating over $0<x<a$: in the usual way, this simply amounts to dropping the summation signs, basis functions ϕ_n and (formally, as it does not make a practical difference) changing the dummy index n to the fixed value m: in other words, we are saying that in each equation, the coefficient of ϕ_m has to be equal to zero:

$$\frac{\partial T_m}{\partial z}-\left(\frac{\partial^2 p_m}{\partial z^2}-k_{xm}^2 p_m\right)=0 \qquad \text{for } 0<z<1, \tag{24a}$$

$$Ra\left(\frac{\partial T_m}{\partial t}-T_m+\frac{\partial p_m}{\partial z}\right)-\left(\frac{\partial^2 T_m}{\partial z^2}-k_{xm}^2 T_m\right)=0 \qquad \text{for } 0<z<1, \tag{24b}$$

$$T_m=0 \qquad \text{at } z=0,\,1 \tag{24c}$$

$$\frac{\partial p_m}{\partial z}=0 \qquad \text{at } z=0,\,1. \tag{24d}$$

Note 2 *Note that, when doing Fourier expansions, we do not use the summation convention, and sums are stated explicitly. In other words, when writing something like $k_{xm}^2 p_m$, there is no implied summation over m.*

The Fourier series representation in x has allowed us to reduce a partial differential equation with (x,z,t) as independent variables into one with (z,t) as independent variables. We would like to repeat that feat by also constructing something like a Fourier series representation in z and reduce the problem to a simple ordinary differential equation in time. This is not standard Fourier series territory, however, since the domain in z is not periodic

Key to further progress is to understand that functions T_m and p_m that are constrained to satisfy the boundary conditions (24c) and (24d) can still be represented by an infinite sum over all the sine and cosine functions that satisfy these boundary conditions. Take the representation of T_m first, which has to satisfy $T_m=0$ at $z=0$

and $z = 1$. These boundary conditions are also satisfied by $\sin(n\pi z)$ for $n = 1, 2, \ldots$, and we can write[8]

$$T_m(z,t) = \sum_{n=1}^{\infty} T_{mn}(t) \sin(n\pi z). \tag{25}$$

Similarly, p_m has to satisfy $\partial p_m/\partial z = 0$ at $z = 0$ and $z = 1$. These zero-derivative boundary conditions are also satisfied by $\cos(n\pi z)$, and we can write

$$p_m(z,t) = \sum_{n=0}^{\infty} p_{mn}(t) \cos(n\pi z). \tag{26}$$

Note that both series involve coefficients $n\pi$, and we will denote them as a vertical wavenumber k_{zn}, defined through

$$k_{zn} = n\pi.$$

Substituting this into (24) and differentiating the sine and cosine series defined above as required yields

$$\sum_{n=1}^{\infty} k_{zn} T_{mn} \cos(k_{zn} z) + \sum_{n=0}^{\infty} \left(k_{zn}^2 + k_{xm}^2\right) p_{mn} \cos(k_{zn} z) = 0 \quad \text{for } 0 < z < 1, \tag{27a}$$

$$Ra \left[\sum_{n=1}^{\infty} \left(\frac{\mathrm{d}T_{mn}}{\mathrm{d}t} - T_{mn}\right) \sin(k_{zn} z) - \sum_{n=0}^{\infty} k_{zn} p_{mn} \sin(k_{zn} z)\right] + \sum_{n=1}^{\infty} \left(k_{zn}^2 + k_{xm}^2\right) T_{mn} \sin(k_{zn} z) = 0 \quad \text{for } 0 < z < 1, \tag{27b}$$

$$\sum_{n=1}^{\infty} T_{mn} \sin(k_{zn} z) = 0 \qquad \text{at } z = 0,\ 1 \tag{27c}$$

$$-\sum_{n=1}^{\infty} k_{zn} p_{mn} \sin(k_{zn} z) = 0 \qquad \text{at } z = 0,\ 1. \tag{27d}$$

The boundary conditions (27c) and (27d) are satisfied automatically, since $\sin(k_{zn}) = \sin(n\pi z) = 0$ for $z = 0$ and $z = 1$: that was, in fact, precisely what the sine and cosine basis functions in (25) and (26) were selected for.

[8]The fact that *any* function T_m satisfying these boundary conditions can be written in this form is actually quite non-trivial to prove, just as it is not trivial to prove that a periodic function can be written as a standard Fourier series, and similar mathematical tools to standard Fourier series are needed to be prove this fact.

Fortuitously, the mass conservation equation for pore fluid (27a) and heat equation (27b) consist of sums over the same sine and cosine terms. As with an ordinary Fourier series, we can project onto individual modes by recognizing the following orthogonality results (see note 4):

$$\int_0^1 \cos(k_{zn}z)\cos(k_{zm}z)\,\mathrm{d}z = \frac{1}{2}\delta_{nm} \tag{28a}$$

$$\int_0^1 \sin(k_{zn}z)\sin(k_{zm}z)\,\mathrm{d}z = \frac{1}{2}\delta_{nm} \tag{28b}$$

As a result, we can multiply (27a) by $\cos(k_z z)$ and integrate over $0 < z < 1$, and likewise multply (27b) by $\sin(k_l z)$ integrate over $0 < z < 1$ to give

$$k_{zl}T_{ml} + \left(k_{xm}^2 + k_{zl}^2\right)p_{ml} = 0 \tag{29a}$$

$$Ra\left[\left(\frac{\mathrm{d}T_{ml}}{\mathrm{d}t} - T_{ml}\right) - k_{zl}p_{ml}\right] + \left(k_{zl}^2 + k_{xm}^2\right)T_{ml} = 0 \tag{29b}$$

You can view this as saying that the coefficients of each distinct cosine and sine term $\cos(k_{zn}z)$ and $\sin(k_{zn}z)$ have to add up to zero in (27a) and (27b).

Note 3 *Note that equations* (29) *are what we would have obtained if we had simply tried a single Fourier mode for T' and p', of the form*

$$T' = T_{mn}\exp(ik_{xm}x)\sin(k_{zn}z), \qquad p' = p_{mn}\exp(ik_{xm}x)\cos(k_{zn}z),$$

dispensing with the sums over m and n, and therefore with the elaborate projection procedure. The approach of substituting a single mode is often what you might want to try first when solving a linear equation like (21)*: it is what we tried when solving the original temperature wave problem, and it works here because the individual modes are* eigenfunctions *of the linear differential equation problem* (21) *(see the beginning of the next section for more detail on this). The approach of trying a single Fourier mode is however far from guaranteed to work: exercise 4 below examines a case where such a simple approach fails.*

Equations (29) apply for $l \geq 1$; for $l = 0$, we have to take account of the fact that the sum over T_{ml} ranges from over $l = 1, 2, \ldots$, and that there is therefore no term T_{m0}. The result of multiplying by $\cos(k_{z0}z) = 1$ and integrating with respect to z is that we simply obtain $p_{m0} = 0$ (there is no contribution to $p'(x, z, t)$ from a mode that varies purely in x but not in z). Eliminating p_{ml} between equations (29),

$$\frac{\mathrm{d}T_{ml}}{\mathrm{d}t} = \left[\frac{k_{xm}^2}{k_{xm}^2 + k_{zl}^2} - \frac{1}{Ra}\left(k_{zl}^2 + k_{xm}^2\right)\right]T_{ml}. \tag{30}$$

Therefore (replacing l by n for cosmetic reasons)

$$T_{mn}(t) = T_{mn}(0)\exp\left[\sigma(k_{xm}, k_{zn})t\right] \tag{31}$$

where

$$\sigma(k_{xm}, k_{zn}) = \frac{k_{xm}^2}{k_{xm}^2 + k_{zn}^2} - \frac{1}{Ra}\left(k_{zl}^2 + k_{xm}^2\right). \tag{32}$$

σ is known as the *growth rate* of the *mode* with wavenumbers (k_{xm}, k_{zn}).[9] The relationship between the growth rate and wavenumber is known as the *dispersion relation*.

The significance of the growth rate is easy to see: the Fourier coefficient $T_{mn}(t)$ grows with time if $\sigma(k_{xm}, k_{zn})$ is positive (or more generally, has positive real part), and shrinks if if $\sigma(k_{xm}, k_{zn})$ is negative (has negative real part). What is important here is the dependence of σ on wavenumbers. In order for the steady state solution $(\bar{T}, \bar{p})$ to be stable, the perturbations (T', p') have to shrink over time. Recall that

$$\begin{aligned} T'(x,z,t) &= \sum_{m=-\infty}^{\infty}\sum_{n=1}^{\infty} T_{mn}(t)\exp\left(ik_{xm}x\right)\sin\left(k_{zn}z\right) \\ &= \sum_{m=-\infty}^{\infty}\sum_{n=1}^{\infty} T_{mn}(0)\exp\left(ik_{xm}x\right)\sin\left(k_{zn}z\right)\exp\left[\sigma(k_{xm}, k_{zn})t\right]. \end{aligned} \tag{33}$$

The $T_{mn}(0)$ are of course determined by projection the initial condition $T(x, z, 0)$ (which we have to assume to be prescribed) onto the Fourier modes. Specifically

$$T_{mn}(0) = \frac{2}{a}\int_0^a \int_0^1 T(x,z,0)\overline{\phi_m(x)}\sin(k_{zn}z)\,\mathrm{d}z$$

From (33), we see that stability therefore requires that *all* modes (m, n) correspond to negative growth rates. Except in the physically implausible situation that the corresponding $T_{mn}(0)$ is exactly zero, a single mode with a positive growth rate will cause T' to grow and the solution $\bar{T} + \varepsilon T'$ to evolve away from, rather than back towards, the steady state $\bar{T}$. In other words, a single positive growth rate $\sigma(k_{xm}, k_{zn})$ signifies instability.

Note 4 *To show that the orthogonality results* (28) *hold, use the following*

$$\cos(A)\cos(B) = \frac{1}{2}\left[\cos(A+B) + \cos(A-B)\right] \tag{34a}$$

and

$$\sin(A)\sin(B) = \frac{1}{2}\left[\cos(A-B) - \cos(A+B)\right] \tag{34b}$$

[9] In a more general setting where a linearized model has a solution that is exponential in time but σ is complex, the real part of the coefficient σ would be known as the growth rate as explored in note 1.

Hence, with n and m both positive,

$$\begin{aligned}
\int_0^1 \cos(k_{zn}z)\cos(k_{zm}z)\,\mathrm{d}z =& \frac{1}{2}\left[\int_0^1 \cos[(k_{zn}+k_{zm})z]\,\mathrm{d}z + \int_0^1 \cos[(k_{zn}-k_{zm})z]\,\mathrm{d}z\right] \\
=& \frac{1}{2}\left[\int_0^1 \cos[(n+m)\pi z]\,\mathrm{d}z + \int_0^1 \cos[(n-m)\pi z]\,\mathrm{d}z\right] \\
=& \begin{cases} \frac{\sin[(n+m)\pi]-\sin(0)}{2(n+m)\pi} + \frac{\sin[(n-m)\pi]-\sin(0)}{2(n-m)\pi} & \text{if } n \neq m \\ \frac{\sin[(n+m)\pi]-\sin(0)}{2(n+m)\pi} + \frac{1}{2} & \text{if } n = m \end{cases} \\
=& \begin{cases} 0 & \text{if } n \neq m \\ \frac{1}{2} & \text{if } n = m \end{cases}
\end{aligned}$$

and

$$\begin{aligned}
\int_0^1 \sin(k_{zn}z)\sin(k_{zm}z)\,\mathrm{d}z =& \frac{1}{2}\left[\int_0^1 \cos[(k_{zn}-k_{zm})z]\,\mathrm{d}z - \int_0^1 \cos[(k_{zn}+k_{zm})z]\,\mathrm{d}z\right] \\
=& \frac{1}{2}\left[\int_0^1 \cos[(n-m)\pi z]\,\mathrm{d}z - \int_0^1 \cos[(n+m)\pi z]\,\mathrm{d}z\right] \\
=& \begin{cases} \frac{\sin[(n-m)\pi]-\sin(0)}{2(n-m)\pi} - \frac{\sin[(n+m)\pi]-\sin(0)}{2(n+m)\pi} + & \text{if } n \neq m \\ \frac{1}{2} - \frac{\sin[(n+m)\pi]-\sin(0)}{2(n+m)\pi} & \text{if } n = m \end{cases} \\
=& \begin{cases} 0 & \text{if } n \neq m \\ \frac{1}{2} & \text{if } n = m \end{cases}
\end{aligned}$$

Exercise 4 *The boundary conditions on our convection problem were chosen deliberately to make the linear stability analysis easier: in particular, the evolution of different modes T_{mn} decouples from each other: each satisfies a separate version of* (30)*. In this exercise and the next, we look at two alternative versions that lead to more complicated stability analyses, in which different modes couple with each other. Consider first the case in which the upper boundary is in contact with a fluid reservoir at constant pressure (such as the bottom of the ocean). This corresponds to keeping most of the model* (21)*, replacing only* (21d) *by*

$$\frac{\partial p'}{\partial z} = 0 \qquad \text{at } z = 0 \tag{35a}$$

$$p' = 0 \qquad \text{at } z = 1. \tag{35b}$$

In that case, we have to replace the expansion (26) *with*

$$p'(x,z,t) = \sum_{m=-\infty}^{\infty}\sum_{n=1}^{\infty} p_{mn}(t)\phi_m(x)\cos(k_{z,n-1/2}z) \tag{36}$$

where

$$k_{z,n-1/2} = \left(n - \frac{1}{2}\right)\pi.$$

Follow the steps leading from (21) *(but with the pressure boundary conditions replaced by* (35) *up to* (30)*, paying particular attention to the final projection step that led to* (29)*, using* (28)*. What replaces* (28)*? You may need to use the approach in note 4. Show that*

$$\sum_{l=1}^{\infty} \frac{(-1)^{l-n} l}{l^2 - (n-1//2)^2} k_{zl} T_{ml} + (k_{z(n-1/2)}^2 + k_{xm}^2) p_{mn} = 0$$

and

$$Ra\left[\frac{\mathrm{d}T_{mn}}{\mathrm{d}t} - T_{mn} - \sum_{q=1}^{\infty} \frac{(-1)^{n-q}(q-1/2)}{n^2 - (q-1/2)^2} k_{zl} p_{mq}\right] + (k_{zn}^2 + k_{xm}^2) T_{mn}.$$

Show that for fixed m*, we can combine the last two equations to get an evolution equation for the vector* T_{mn} *of Fourier coefficients analogous to* (16)

$$\frac{\mathrm{d}T_{mn}}{\mathrm{d}t} = \sum_{p=1}^{\infty} J(m)_{np} T_{mp}. \tag{37}$$

and find a formula for the components of the 'matrix'

$$J(m)_{np},$$

although this matrix has an infinite number of dimensions. If we were to look for an exponential solution $T_{mn}(t) = T_{mn}(0)\exp(\sigma t)$*,* (37) *becomes*

$$\sum_{p=1}^{\infty} \left(\sigma \delta_{np} - J(m)_{np}\right) T_{mn}(0) = 0,$$

which is equivalent to the eigenvalue problem (18)*.*

Note that we have written J *as a function of* m *since* m *does not play the role of an index in the matrix equation.* (37) *generalizes* (30) *to a matrix equation: effectively,* $J(m)_{np}$ *is a diagonal matrix in* (30)*, of the form*

$$J(m)_{np} = \left[\frac{k_{xm}^2}{k_{xm}^2 + k_{zn}^2} - \frac{1}{Ra}(k_{zn}^2 + k_{xm}^2)\right] \delta_{np}.$$

Exercise 5 *Next, consider also replacing the lower fixed-temperature boundary condition by a fixed geothermal heat flux. This amounts to replacing* (21c) *by*

$$\frac{\partial T'}{\partial z} = 0 \qquad \text{at } z = 0 \tag{38a}$$

$$p' = 0 \qquad \text{at } z = 1. \tag{38b}$$

Assume we do this in addition to replacing the impermeable upper boundary with an open boundary (replacing the boundary condition (21d) *by* (35)*), so we keep the new pressure expansion* (36)*, and additionally replace* (25) *by*

$$T'(x,z,t) = \sum_{m=-\infty}^{\infty} \sum_{n=1}^{\infty} T_{mn}(t)\phi_m(x)\sin(k_{z,n-1/2}z). \tag{39}$$

Again, re-do the steps that lead up to (30)*.*

Convection patterns

Before we delve into further detail about the growth rate, and how and when instability first occurs, we illustrate the solution we have derived graphically. Note that we have individual Fourier mode solutions that we can combine as

$$\begin{aligned} T_{mn}(t)\phi_m(x)\sin(k_{zn}z) + T_{-mn}(t)\phi_{-m}(x)\sin(k_{zn}z) = \\ 2\mathrm{Re}(T_{mn})\cos(k_m x)\sin(k_{zn}z) - 2\mathrm{Im}(T_{mn})\sin(k_{xm}x)\sin(k_{zn}z) \\ p_{mn}(t)\phi_m(x)\cos(k_{zn}z) + p_{-mn}(t)\phi_{-m}(x)\cos(k_{zn}z) = \\ 2\mathrm{Re}(p_{mn})\cos(k_{xm}+)\cos(k_{zn}z) - 2\mathrm{Im}(p_{mn})\sin(k_{mx}x)\cos(k_{zn}z) \end{aligned}$$

where

$$\mathrm{Re}(p_{mn}) = -\frac{k_{zn}}{k_{xm}^2 + k_{zn}^2}\mathrm{Re}(T_{mn}), \qquad \mathrm{Im}(p_{mn}) = -\frac{k_{zn}}{k_{xm}^2 + k_{zn}^2}\mathrm{Im}(T_{mn}).$$

Based on this, we can define cosine modes in x

$$\hat{T}_{c,mn} = \cos(k_{xm}x)\sin(k_{zn}z), \qquad \hat{p}_{c,mn} = -\frac{k_{zn}}{k_{xm}^2 + k_{zn}^2}\cos(k_{xm}x)\cos(k_{zn}z)$$

and similarly corresponding sine modes in x

$$\hat{T}_{s,mn} = \sin(k_{xm}x)\sin(k_{zn}z), \qquad \hat{p}_{s,mn} = -\frac{k_{zn}}{k_{xm}^2 + k_{zn}^2}\sin(k_{xm})\cos(k_{zn}z)$$

with corresponding flux modes

$$\hat{\mathbf{q}}_{c,mn} = \hat{T}_{c,mn}(x,z)\mathbf{k} - \nabla\hat{p}_{c,mn},$$

and equivalently for $\hat{\mathbf{q}}_{s,mn}$: the sine modes are simply phase shifted versions of the cosine modes.

Note that we have not included the time dependence $\exp(\sigma(k_{xm}, k_{zn})t)$ in the definition of the $\hat{T}_{mn}$ and $\hat{p}_{mn}$; these functions are purely spatial, and their dynamical significance is as so-called *eigenfunctions* of the linearized eigenvalue problem (21)

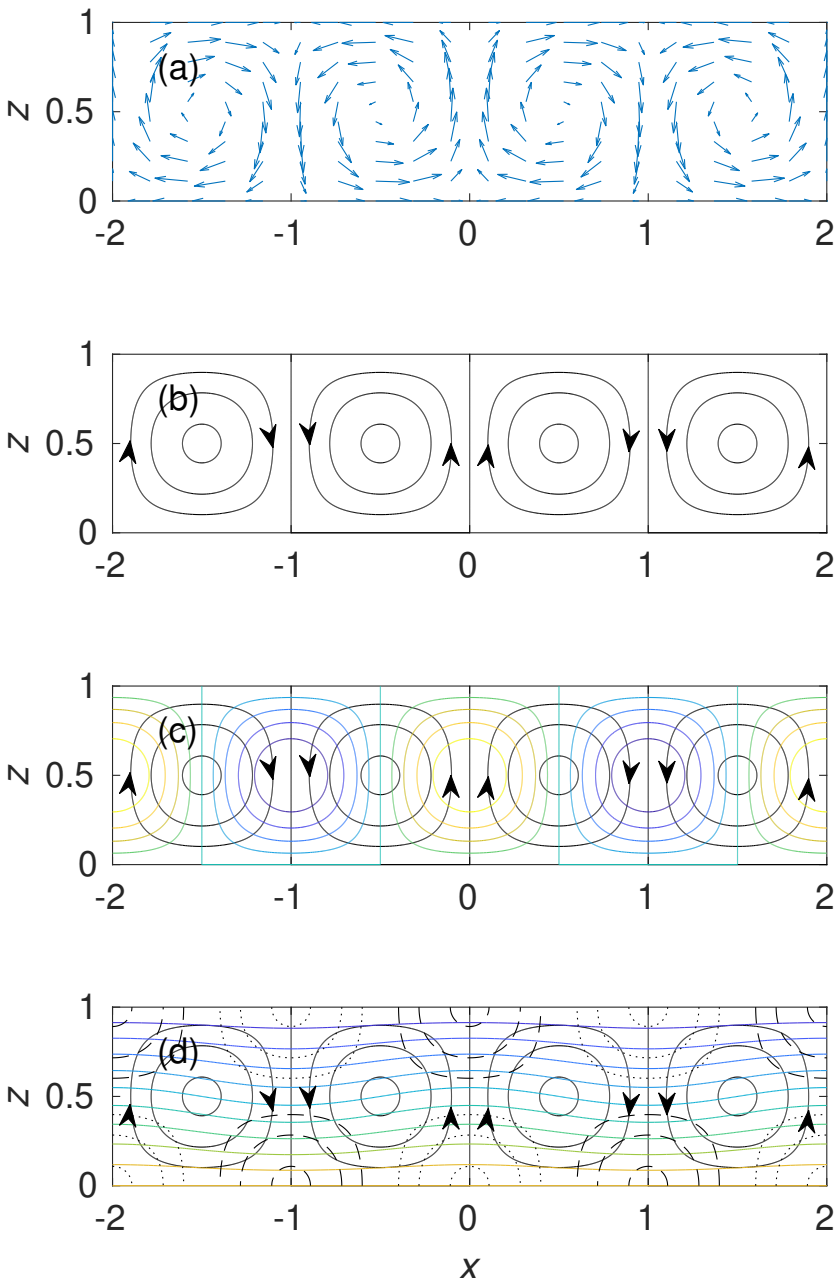

Figure 1: The solution $\hat{T}_{c,mn} = \cos(\pi x)\sin(\pi z)$, $\hat{p}_{c,mn} = -1/(2\pi)\cos(\pi x)cos(\pi z)$ for $m = n = 1$, visualized using the corresponding fluid flux solution $\hat{\mathbf{q}}_{c,mn}$ as (a) a vector shown using discrete vectors and (b) in the form of streamlines (see note 5). The flow is driven by the temperature perturbation, which are shown as a contour plot in (c), overlaid onto the streamlines. Warm (reddish) colours indicate high values of $\hat{T}_{c,mn}$ and cool (blue) colours low values of $\hat{T}_{c,mn}$. The actual temperature is the sum of the steady state temperature $\bar{T}(z) = -z$ and $\varepsilon T'$, which evolves in time. We show $\bar{T} + \varepsilon\hat{T}_{c,mn}$ as a contour plot in (d) with $\varepsilon = 0.05$. Additionally, $\hat{p}_{c,mn}$ is shown using dashed contour lines for positive $\hat{p}_{c,mn}$ where positive, dotted where negative.

with associated eigenvalue $\sigma_{mn} = \sigma(k_{xm}, k_{zn})$: They satisfy the differential eigenvalue problem that arises from (21) by replacing the time derivative $\partial T/\partial t$ by σT, as

$$\begin{aligned} \frac{\partial \hat{T}_{mn}}{\partial z} - \nabla^2 \hat{p}_{mn} &= 0 \qquad && \text{for } 0 < z < 1, \\ Ra\left(\sigma_{mn}\hat{T}_{mn} - \hat{T}_{mn} + \frac{\partial \hat{p}_{mn}}{\partial z}\right) - \nabla^2 \hat{T}_{mn} &= 0 \qquad && \text{for } 0 < z < 1, \\ \hat{T}_{mn} &= 0 \qquad && \text{at } z = 0,\, 1, \\ \frac{\partial \hat{p}_{mn}}{\partial z} &= 0 \qquad && \text{at } z = 0,\, 1. \end{aligned}$$

Moreover, the general solution to (21) can be represented as a sum over over these eigenfunctions multiplied by the corresponding exponential growth factor and a numerical coefficient, for instance

$$T'(x,z,t) = \sum_{m=0}^{\infty}\sum_{n=1}^{\infty} \alpha_{mn}\hat{T}_{c,mn}(x,z)\exp(\sigma_{mn}t) + \sum_{m=0}^{\infty}\sum_{n=1}^{\infty} \beta_{mn}\hat{T}_{s,mn}(x,z)\exp(\sigma_{mn}t),$$

$$p'(x,z,t) = \sum_{m=0}^{\infty}\sum_{n=1}^{\infty} \alpha_{mn}\hat{p}_{c,mn}(x,z)\exp(\sigma_{mn}t) + \sum_{m=0}^{\infty}\sum_{n=1}^{\infty} \beta_{mn}\hat{p}_{s,mn}(x,z)\exp(\sigma_{mn}t),$$

where any individual functions of the form $\hat{T}_{mn}(x,z)\exp(\sigma_{mn}t)$, $\hat{p}_{mn}(x,z)\exp(\sigma_{mn}t)$ by themselves solve (21). This superposition of eigenfunctions with an exponential time dependence mirrors the sum (19) over eigenvectors of J_{ij} in note 1.[10] In terms of the original coefficients T_{mn}, we have

$$\alpha_{mn} = 2\mathrm{Re}(T_{mn}(0)), \qquad \beta_{mn} = -2\mathrm{Im}(T_{mn}(0)).$$

As a result, we can regard these eigenfunctions as independent solutions to the convection problem (21), and visualize them individually. In figure 1, we set $a = 2$, and plot $\hat{T}_{c,1,1}$ (corresponding to $m = 1$, $n = 1$). Note that to avoid confusion, we add a comma between 'm' and 'n' in T_{mn} when giving numerical values, so $T_{-1,1}$ is T_{mn} with $m = -1$, $n = 1$, and ditto for $\hat{T}_{c,1,1}$ being $\hat{T}_{c,mn}$ with $m = 1$, $n = 1$.. The reason for choosing $a = 2$ and $m = n = 1$ for our main plot will become clearer in the next section.

The solution for fluid flux $\hat{\mathbf{q}}_{c,mn}$ with $m = n = 1$, as a typical vector field plot with flux shown using arrows, is displayed in panel (a). This is actually harder to

[10]The distinction between 'Fourier mode' and 'eigenfunction' is important: as exercise 4, a change in boundary conditions still allows us to use Fourier modes to solve the problem in principle, but individual Fourier modes are no longer eigenfunctions of the modified version of (21), and the way that the Fourier coefficients T_{mn} with different indices m and n evolve is coupled in that case, and we can no longer express the solution T' as a sum $\sum_m \sum_n T_{mn}\phi_m(x)\sin(k_{nz}z)\exp(\sigma_{mn}t)$, where each Fourier component has its own exponential dependence on t; the latter is the hallmark of an eigenfunction.

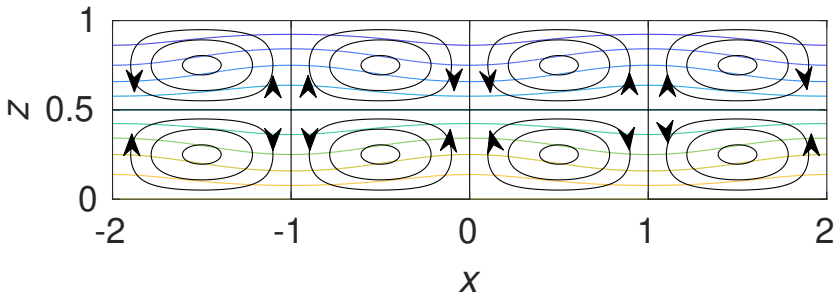

Figure 2: The solution $\hat{T}_{c,mn} = \cos(\pi x)\sin(2\pi z)$, $\hat{p}_{mn} = -2/(5\pi)\cos(\pi x)cos(2\pi z)$ for $m = 1$, $n = 2$, visualized using the corresponding fluid flux solution $\hat{\mathbf{q}}_{mn}$ in the form of streamlines, with the $\bar{T} + \varepsilon\hat{T}_{mn}$, shown as a contour plot in (d) with $\varepsilon = 0.05$, using the same colour scheme as in figure 1.

read than the associated streamline plot, which is shown in panel (b). The circulation pattern clearly consists of closed convection 'cells', which here have a height equal to the domain itself. These convection cells are driven by temperature variations across the domain. Panel (c) shows the corresponding contour lines of the corresponding temperature perturbation eigenfunction $\hat{T}_{c,mn}$, with warmer colours corresponding to higher temperatures. Unsurprisingly, the temperatures perturbations are warmest where the convection cells are upwelling in a vertical direction (the arrows on the streamlines point up) and the streamlines are most closely bunched (which turns out to correspond to the fastest velocities, see note 5).

The contours of the $\hat{T}_{c,mn}$ plotted in panel c are somewhat misleading since the actual temperature consists of a steady state temperature $\bar{T}$ with a small perturbation $\varepsilon T' = \varepsilon \sum_m \sum_n \alpha_{mn}(\hat{T}_{c,mn}(x,z) + \hat{T}_{s,mn}(x,z))\exp(\sigma_{mn}t)$ superimposed on it. Importantly $\bar{T} = -z$ decreases with z, and in general, the warmest temperatures are found at the bottom of the domain rather than in the centre of the upwelling as suggested by panel c In panel d, we plot $T = \bar{T} + \varepsilon\hat{T}_{c,mn}(x,z)\exp(\sigma_{mn}t)$, with a single mode $m = n = 1$ and $\varepsilon = 0.05$, as a snapshot of T evaluated at $t = 0$.

Clearly, upwelling fluid corresponds to elevated temperatures as expected. The sideways motion of the fluid required to complete the convection cells is not driven by temperature anomalies, but by pressure gradients: there is high fluid pressure at the top of each upwelling column of fluid (dashed contour lines in panel d), pushing that fluid sideways towards regions of downwelling, where pressures are lower (dotted contour lines in pane d). As the fluid travels horizontally the upper boundary, it loses heat to that colder boundary, and its temperature decreases. The resulting increase in density makes the fluid negatively buoyant, and it starts to descend again.

Figure 1 shows convection cells whose height is the height of the domain itself. It is possible to construct convection cells that are only an integer fraction of the

domain height, which results from picking k_{zn} with $n > 1$, Figure 2 shows an example using the same parameter values as figure 1, but plotting $\hat{T}_{mn}$ with $m = 1$, $n = 2$, and $T_{1,2} = T_{-1,2} = 1/2$ instead. Figure 2 shows the equivalent of panel d of figure 1 (streamlines of $\hat{\mathbf{q}}_{mn}$ and contours of $\bar{T} + \varepsilon\hat{T}_{mn}(x, z)$ with $\varepsilon = 0.05$). As we will show in the next section, the dispersion relation (32) does however not favour such double cells, with cells spanning the full height of the domain growing fastest.

Note 5 *This note explains a little more about how to construct streamline plots for the fluid flux vector field* $\mathbf{q}'$*. Note that temperature and pressure eigenfunctions are related through*

$$\hat{T}_{c,mn} = \cos(k_{xm}x)\sin(k_{zn}z), \qquad \hat{p}_{c,mn} = -\frac{k_{zn}}{k_{xm}^2 + k_{zn}^2}\cos(k_{xm}x)\sin(k_{zn}z)$$

where we have assumed that T_{mn} *is real (complex* T_{mn} *simply amounts to a phase shift, see note 2 of the notes on Fourier series).*

The corresponding fluid flux field defined by (12c) *takes the form*

$$\hat{\mathbf{q}}_{mn} = -\frac{k_{xm}k_{zn}}{k_{xm}^2 + k_{zn}^2}\sin(k_{xn}x)\cos(k_{zn}z)\mathbf{i} + \frac{2k_{xm}^2}{k_{xm}^2 + k_{zn}^2}\cos(k_{xm}x)\sin(k_{zn}z)\mathbf{k}.$$

It is easy to verify that this takes the form

$$\hat{\mathbf{q}}_{mn} = -\frac{\partial\Psi_{mn}}{\partial z}\mathbf{i} + \frac{\partial\Psi_{mn}}{\partial x}\mathbf{k} \tag{40}$$

if we define Ψ_{mn} *as*

$$\Psi_{mn}(x, z) = \frac{2k_{xm}}{k_{xm}^2 + k_{zn}^2}\sin(k_{xm}x)\sin(k_{zn}z). \tag{41}$$

Ψ_{mn} *is known as a* stream function*. The abstract representation* (40) *in terms of a stream function does not owe its existence to the particulars of the convection problem, but simply to the fact that flux* $\mathbf{q}' = \mathbf{q}$ *is divergence free, which implies that* $\mathbf{q}$ *can be written as the curl of a vector field, in this case of the form* $\Psi(x, z)\mathbf{j}$*: this is follows from a general result in advanced vector calculus known as the Helmholtz representation theorem.*

When we write $\mathbf{q}$ *in the form* (40)*, we can also show that the streamlines of* $\mathbf{q}$ *are isolines of* Ψ*. Recall that the streamlines of a vector field are curves* $\mathbf{r}(t) = X(t)\mathbf{i} + Z(t)\mathbf{k}$ *defined by treating* $\mathbf{q}$ *as the velocity of a particle at the location where the vector field* $\mathbf{q}(\mathbf{r}(t))$ *is evaluated, and tracing the motion of that particle:*

$$\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t} = \mathbf{q}(\mathbf{r}(t)).$$

Hence, using the chain rule and the definition of $\mathbf{q}$ *in terms of the stream function*

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\Psi(\mathbf{r}(t)) &= \nabla\Psi\cdot\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t}\\ &= \nabla\Psi\cdot\mathbf{q}\\ &= \left(\frac{\partial\Psi}{\partial x}\mathbf{i}+\frac{\partial\Psi}{\partial z}\mathbf{k}\right)\cdot\left(-\frac{\partial\Psi}{\partial z}\mathbf{i}+\frac{\partial\Psi}{\partial x}\mathbf{k}\right)\\ &= 0\end{aligned}$$

and Ψ *remains constant along streamlines. Consequently, the streamlines in figures 1 and 2 are plotted by plotting contour lines of* Ψ *defined in* (41).

Note also that the bunching of streamlines indicates faster flow: since

$$\mathbf{q} = \nabla\times(\Psi\mathbf{j}),$$

we equivalently have

$$\mathbf{q} = -\mathbf{j}\times\nabla\Psi.$$

so (as $\nabla\Psi$ *is in the* xz*-plane and therefore perpendicular to* $\mathbf{j}$*),*

$$|\mathbf{q}| = |\nabla\Psi|.$$

If we plot streamlines as contours of Ψ *with constant contour interval (which is done in figure 1, then proximity of contour lines to each other indicates the magnitude of gradient* $\nabla\Psi$*, and therefore of the magnitude of* $\mathbf{q}'$*, whose direction is obviously parallel to the streamlines.*

Critical Rayleigh number and wavelength selection

We can try to establish when positive growth rates will occur. Looking at the dispersion relation (32), it is clear that there are two competing terms: the first, positive one, and the second, negative one that is multiplied by a factor of Ra^{-1}. If we trace where these terms come from back to (29) and beyond to (21), it becomes clear that the first term on the right-hand side of (32) is associated with advection in the heat equation, while the second is associated with diffusion (that is to say, conduction) of heat.

This makes some sense: advection will cause rising, hot fluid to remain warm and therefore buoyant, increasing the tendency for it to continue rising. Advection should therefore promote instability. Diffusion, on the other hand, will tend to cool the rising fluid by transferring its heat content to surrounding, cooler fluid. This should suppress buoyancy and instability.

It is also immediately clear that the first (positive, and therefore destabilizing) term in (32) is bounded: it can never be larger than 1. By contrast, the second

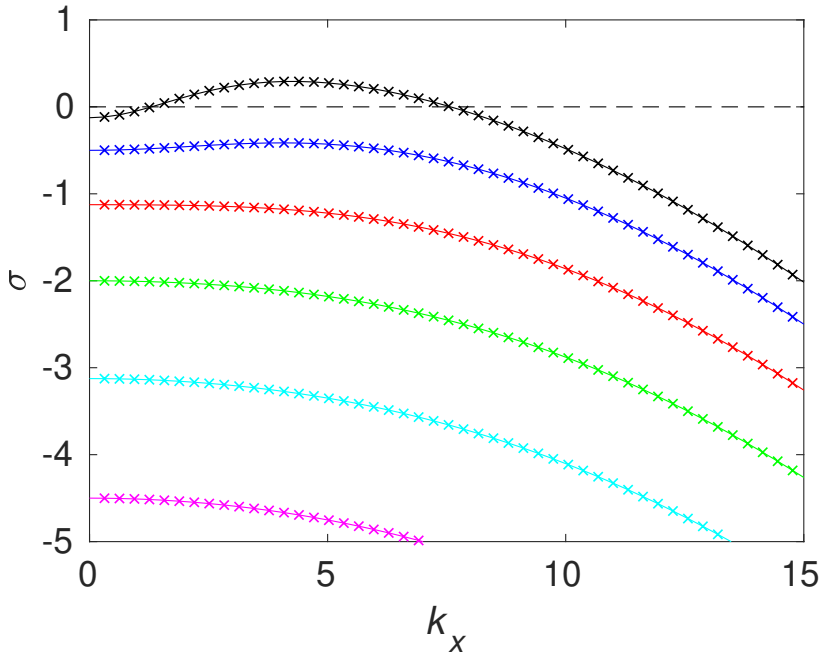

Figure 3: The growth rate σ as a function of k_x for $Ra = 8\pi^2$ for $k_z = \pi$ (black curve / markers), $k_z = 2\pi$ (blue), $k_z = 3\pi$) (red), $k_z = 4\pi$ (green), $k_z = 5\pi$ (cyan), $k_z = 6\pi$ (magenta) and $k_z = 7\pi$ (yellow). The cross-shaped markers represent discrete values of $k_{xn} = 2n\pi/a$ that we obtain with $a = 20$, $n = 1,2,\ldots$, while the solid curves treat k_x as a continuous variable, representing the limit of large a. The dashed black line is $\sigma = 0$, with any values of σ above that line indicating nodes (and corresponding wavenumbers (k_x, k_y) that grow unstably.

(negative, stabilizing) term becomes progressively larger as the wavenumbers k_{xm} and k_{zn} are increased, so σ is guaranteed to be negative for large wavenumbers. Recall that wavelengths are 2π divided by wavenumber. The second, stabilizing term therefore becomes dominant at short wavelengths: diffusion is particularly effective at smoothing out temperature variations, and therefore buoyancy, over short length scales. If growth occurs, it has to be at longer wavelengths.

In detail, we have to treat horizontal and vertical wavenumbers k_{xm} and k_{zn} somewhat differently. Recall that

$$k_{zn} = n\pi,$$

and we strictly have to look at k_{zn} as a discrete variable. It is immediately clear that σ is a decreasing function of k_{zn} (see figure 3, where curves corresponding to larger k_{zn} are lower down in the plot). The first, destabilizing term in (32) decreases as k_{zn} increases, while the magnitude of the second, stabilizing term increases with

increasing k_{zn}. The biggest growth rates are therefore obtained for

$$n = 1, \qquad k_{zn} = \pi.$$

In practice, this means that the corresponding convection cell span the height of the domain as in figure 1. Convection patterns with two ore more cells in the vertical as in figure 2 are in theory possible, but are not favoured by the instability mechanism: the associated growth rates are invariably slower than those for patterns with a single cell in the vertical.

The horizontal wavenumber k_{xm} can be treated much more plausibly as continuous (see figure 3, where the discrete crosses are wavenumbers k_{xm} that we obtain for $a = 20$, for larger a they would be even closer together). We have $k_{xm} = 2m\pi/a$, and therefore increments in k_{xm} come in units of $k_{x(m+1)} - k_{xm} = 2\pi/a$. We can make this as small as we like be picking a very wide domain width a.[11] The fact that increments in k_{xm} can be made very small means we can effectively treat k_{xm} as a continuous variable.

If we put $k_{zn} = k_{z1} = \pi$ because that maximizes σ with respect to k_{zn}, we find

$$\sigma(k_x, \pi) = \frac{k_x^2}{\pi^2 + k_x^2} - \frac{1}{Ra}(k_x^2 + \pi^2).$$

Treating k_x as continuous, we can differentiate with respect to k_x in order to find the maximum of σ

$$\begin{aligned}\left.\frac{\partial \sigma}{\partial k_x}\right|_{k_{zn}=\pi} =& 2k_x\left(\frac{1}{\pi^2 + k_x^2} - \frac{k_x^2}{(\pi^2 + k_x^2)^2} - \frac{1}{Ra}\right) \\ =& \frac{2k_x\left[Ra\pi^2 - (\pi^2 + k_x^2)^2\right]}{Ra(\pi^2 + k_x^2)^2}.\end{aligned}$$

Setting this to zero implies that the maximum of σ is attained either when

$$k_x = 0, \tag{42}$$

or when

$$Ra\pi^2 = (\pi^2 + k_x^2)^2, \qquad k_x^2 = Ra^{1/2}\pi - \pi^2. \tag{43}$$

Since we know that σ becomes negative for large k_x, one of these two must correspond to the global maximum of $\sigma(k_x, \pi)$.

Note that the value of σ at $k_x = 0$ is always negative,

$$\sigma(0, \pi) = -\frac{\pi^2}{Ra}, \tag{44}$$

[11] For infinitely wide domains that are not periodic in x, the relevant alternative to Fourier series is the *Fourier transform*, which works in a very similar manner but replaces the sum of the Fourier series with an integral over continuous k_x.

and never corresponds to instability. The second possible value of k_x at which σ can attain a maximum is given by (43), but only corresponds to a real k_x if

$$Ra \geq \pi^2. \tag{45}$$

In that case the value of σ at that point is

$$\sigma_{max} = 1 - \frac{2\pi}{Ra^{1/2}}. \tag{46}$$

When (45) is satisfied, note that σ_{max} is greater than the value than the value of σ at k_x given by (44). In other words, σ_{max} genuinely is the maximum of σ then. σ_{max} is positive (signifying growth) if and only if

$$Ra > 4\pi^2. \tag{47}$$

What we have is a *critical value* of the Rayleigh number $Ra_c = 4\pi^2$ at which instability starts: lower values of Ra correspond to the steady state solution $p = \bar{p}(z)$, $= \bar{T}(z)$ being stable, higher values correspond to it being unstable. This makes sense: remember that Ra controls the size of the second, stabilizing term in (32) relative to the first, stabilizing one. The more that term is suppressed, the less likely convection is to occur. If we look at the definition of Ra in (4),

$$Ra = \frac{kg\rho_0^2 c\alpha(T_b - T_s)h}{\mu\kappa},$$

it becomes obvious that convection can be triggered in number of ways: by increasing the dimensional height h of the domain, by increasing the temperature difference $T_b - T_s$ between bottom and top, or by using a fluid that is denser (larger ρ_0, has a greater thermal expansion coefficient α, larger heat capacity c or smaller viscosity μ. Convection is also favoured by a more permeable porous medium, or a smaller thermal conductivity κ.

When instability occurs, σ is maximized by a particular combination of wavenumbers

$$k_{z,max} = \pi, \qquad k_{x,max} = \left(Ra^{1/2}\pi - \pi^2\right)^{1/2}, \tag{48}$$

meaning, by a particular combination of wavelengths $2\pi/k_{z,max}$ and $2\pi/k_{x,max}$. In other words, the pattern of upwelling of warm fluid and downwelling of cold fluid that results from the instability occurs preferentially at specific spatial scales. This is called *wavelength selection*, and we can potentially learn by looking at these fastest growing wavelengths how the pattern that emerges from the instability depends on the physical parameters of the model.

As observed above, the value of vertical wavelength corresponding to maximum growth rate is always $2\pi/k_{z,max} = 2$. This corresponds to convection cells spanning the height of the domain as shown in figure 1. The value of the horizontal wavelength

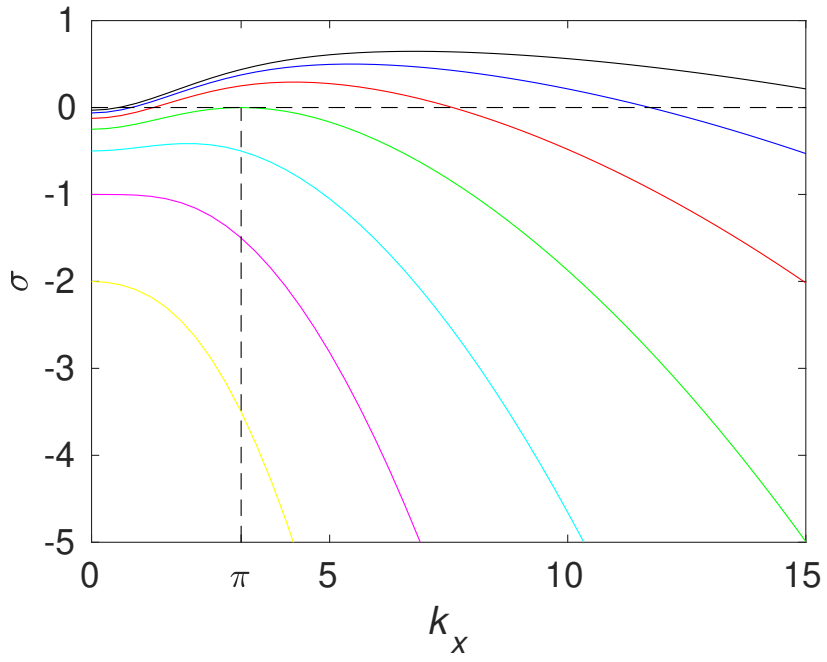

Figure 4: The growth rate σ as a function of k_x for $k_z = \pi$ at different values of $Ra = 32\pi^2$ (black curve / markers), $Ra = 16\pi^2$ (blue), $Ra = 8\pi^2$)(red, shown as a black curve in figure 3), $Ra = 4\pi^2$ (green), $Ra = 2\pi^2$ (cyan), $Ra = \pi^2$ (magenta) and $Ra = \pi^2/2$ (yellow). Note that the maximum value of σ attained for the critical $Ra = 4\pi^2$ is zero as described in the text, and is attained at $k_x = \pi$. Larger values of Ra lead to a range of k_x for which σ is positive, with a maximum attained at a value of k_x that increases with Ra.

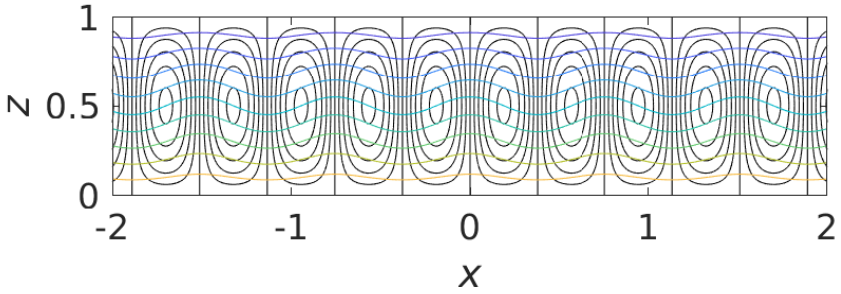

Figure 5: The fastest growing convection cell when $Ra = 16Ra_c$, using the same plotting scheme as in figure 1(d). Note that this convection pattern is much narrower then that shown in figure 1, reflecting the higher Rayleigh number.

$2\pi/k_x$ at corresponding to the fastest growth rate meanwhile decreases with Rayleigh number Ra (see also figure 4, where the peak in σ occurs further to the right in the plot for larger Ra): as the Rayleigh number is increased past its critical value Ra_c (for instance, by increasing the temperature difference $T_b - T_s$ between bottom and top of the tank), the convection pattern will consist of narrower cells (cells with larger wavenumbers).

Note that at $Ra = Ra_c$, and $\sigma_{max} = 0$ corresponds to $k_x = \pi$, or a wavelength of 2. This is the case illustrated in figure 1, which shows the 'first' Fourier mode to become unstable. Here each 'cell' is exactly as wide as it is tall, the cell being half a wavelength wide. Compare this with the fastest growing mode predicted when $Ra = 16Ra_c$, shown using the same plotting scheme in figure 5: the narrowing of the fastest-growing convection cells as Ra increases is obvious.

Boussinesq convection as a dynamical system

This section is about viewing the Boussinesq convection problem (8) as a dynamical system as defined in note 1, and is mostly intended to make the connection with such dynamical systems clearer, and also to provide something of an insight into possible means of numerical solution. Importantly we get to understand what happens when exponential growth starts to be curtailed by the nonlinear terms in (8). The material that follows is necessarily somewhat more advanced, though necessary if you want to get a fuller understanding of convection: what happens to the initially exponential growth of convection cells?

If you are still reading at this point, let me say the following: in this course, we have not made computational approaches anything more than peripheral, focusing on the development of theory. Purely theoretical development reaches the boundaries of what it can achieve sooner or later, unless you deliberately restrict yourself only to carefully selected problems that have an 'analytical' solution. You should aim to learn something about numerical methods to complement the theoretical content of this course; the description below barely scratches the surface, and also leads to a computational method known as a *spectral method* that, while powerful when applicable, can only be used in very simple domain geometries. You should aim to learn about other, more easily generalized methods such as finite difference, finite volume and finite element methods in a course on numerical partial differential equations.

Remember that ε was an arbitrary parameter, assumed to be small in order to look at small perturbations to the steady state. If we want to look at sizeable perturbations, we can therefore set $\varepsilon = 1$, although we can no longer ignore higher order

terms in ε. Simplifying (11) while retaining the nonlinear term leads to

$$\frac{\partial T}{\partial z} - \nabla^2 p' = 0 \quad \text{for } 0 < z < 1, \tag{49a}$$

$$Ra\left(\frac{\partial T'}{\partial t} - T' + \frac{\partial p'}{\partial z} + T'\frac{\partial T'}{\partial z} - \nabla p' \cdot \nabla T'\right) - \nabla^2 T' = 0 \quad \text{for } 0 < z < 1, \tag{49b}$$

$$T' = 0 \qquad \text{at } z = 0,\ 1, \tag{49c}$$

$$\frac{\partial p'}{\partial z} = 0 \qquad \text{at } z = 0,\ 1. \tag{49d}$$

The approach we follow below is this: in the analysis of the linearized problem (21), we represented T' and p' by Fourier series, and we were able to reduce the problem to separate ordinary differential equations of the form (30) for each Fourier coefficient $T_{mn}(t)$. The simplicity of that equation is actually surprising, given the apparently complicated nature of the initial partial differential equation problem. What happens if we retain the nonlinear terms in (49), which are omitted in (21)? What type of model do we obtain for the Fourier coefficients then?

To find out, substitute the double sums

$$T'(x,z,t) = \sum_{m=-\infty}^{\infty}\sum_{n=1}^{\infty} T_{mn}(t)\phi_m(x)\sin(k_{zn}z), \tag{50a}$$

$$p'(x,z,t) = \sum_{m=-\infty}^{\infty}\sum_{n=1}^{\infty} p_{mn}(t)\phi_m(x)\cos(k_{zn}z), \tag{50b}$$

into (49a) and (49b).[12] We get the two rather complicated-looking equations

$$\sum_{m=-\infty}^{\infty}\sum_{n=1}^{\infty}\left[k_{nz}T_{mn}\phi_m(x)\cos(k_{zn}z)+\left(k_{xm}^2+k_{zn}^2\right)p_{mn}\phi_m(x)\cos(k_{zn}z)\right]=0, \tag{51a}$$

$$\begin{aligned}
&Ra\sum_{m=-\infty}^{\infty}\sum_{n=1}^{\infty}\left(\frac{\mathrm{d}T_{mn}}{\mathrm{d}t}-T_{mn}-k_{zn}p_{mn}\right)\phi_m(x)\sin(k_{zn}z)\\
&+Ra\sum_{m=-\infty}^{\infty}\sum_{n=1}^{\infty}\sum_{q=-\infty}^{\infty}\sum_{r=1}^{\infty}T_{mn}T_{qr}k_{zr}\phi_m(x)\phi_q(x)\sin(k_{zn}z)\cos(k_{zr}z)\\
&+Ra\sum_{m=-\infty}^{\infty}\sum_{n=1}^{\infty}\sum_{q=-\infty}^{\infty}\sum_{r=1}^{\infty}T_{mn}p_{qr}k_{xm}k_{xq}\phi_m(x)\phi_q(x)\sin(k_{zn}z)\cos(k_{zr}z)\\
&+Ra\sum_{m=-\infty}^{\infty}\sum_{n=1}^{\infty}\sum_{q=-\infty}^{\infty}\sum_{r=1}^{\infty}T_{mn}p_{qr}k_{zn}k_{zr}\phi_m(x)\phi_q(x)\cos(k_{zn}z)\sin(k_{zr}z)\\
&+\sum_{m=\infty}^{\infty}\sum_{n=1}^{\infty}\left(k_{xm}^2+k_{zn}^2\right)T_{mn}\phi_m(x)\sin(k_{zn}z)=0.
\end{aligned} \tag{51b}$$

while the boundary conditions (49c) and (49d) are satisfied automatically. The equations above are the result of simple algebra, differentiation, and careful book-keeping in the sums.

Next, we take the usual projection steps again: we multiply both equations by $\overline{\phi_l(x)}$ and integrate over $0<x<a$, using the orthogonality of the basis functions ϕ_m defined in (23),

$$\frac{1}{a}\int_0^a \phi_m(x)\overline{\phi_n(x)}\,\mathrm{d}x=\delta_{mn},$$

In addition, we multiply (51a) by $\cos(k_{zj}z)$ and integrate with respect to z from 0 to 1. Using (28), we obtain (29a) once more

$$k_{zl}T_{lj}+\left(k_{xl}^2+k_{zj}^2\right)p_{lj}=0$$

so that (replacing l by m and j by n for cosmetic reasons)

$$p_{mn}=-\frac{k_{zn}}{k_{xm}^2+k_{zn}^2}T_{mn}. \tag{52}$$

[12] As in the linearized problem,it is easy to see that that no term $p_{m0}(t)\phi_m(x)$ is required, and if we changed the expression for p' to read

$$p'(x,z,t)=\sum_{m=-\infty}^{\infty}\sum_{n=0}^{\infty}p_{mn}(t)\phi_m(x)\cos(k_{zn}z),$$

we would obtain $p_{m0}=0$ from (52), see also the comment immediately below (29).

In addition to projection onto the mode ϕ_l (multiplying by $\overline{\phi_l(x)}$ and integrating with respect to x), we multiply (51b) by $\sin(k_{zj}z)$ and multiply with respect to z from 0 to 1. To deal with the nonlinear terms, we need the additional observation

$$\phi_m(x)\phi_q(x) = \phi_{m+q}(x).$$

and one further orthogonality result

$$\int_0^1 \sin(k_{zn}z)\sin(k_{zm}z)\cos(k_{zl}z)\,\mathrm{d}z = \frac{1}{4}\delta_{n,(m-l)} + \frac{1}{4}\delta_{m,(n-l)} - \frac{1}{4}\delta_{l,(n+m)}. \tag{53}$$

where, for clarity, we have separated the index arguments of the Kronecker delta by a comma (and we will do the same to other double indices where necessary below.)

Exercise 6 *Show that* (53) *holds, given* $k_{zq} = q\pi$ *for a generic index* q*, and that* n*,* m *and* l *in* (53) *are all positive. Use* (34)*.*

When we use these results in (51b), we construct something similar to the convolution theorem of the notes on Fourier series for the nonlinear terms. Overall, the procedure results in (replacing l by m and j by n again in the final answer, for cosmetic reasons)

$$\begin{aligned} &\frac{\mathrm{d}T_{mn}}{\mathrm{d}t} - T_{mn} - k_{zn}p_{mn} \\ +\mathrm{conv}_{mn}(k_{zr}T_{qr}, T_{qr}) + \mathrm{conv}_{mn}(k_{xq}p_{qr}, k_{xq}T_{qr}) &+ \mathrm{conv}_{mn}(k_{zr}T_{qr}, k_{zr}p_{qr}) \\ &+Ra^{-1}\left(k_{xm}^2 + k_{zn}^2\right)T_{mn} = 0. \end{aligned} \tag{54a}$$

where we define a convolution function conv_{mn} of two sets of Fourier coefficients C_{pq} and D_{pq} through

$$\mathrm{conv}_{mn}(C_{qr}, D_{qr}) = \frac{1}{2}\sum_{q=-\infty}^{\infty}\left(\sum_{r=1}^{n-1} C_{qr}D_{m-q,n-r} + \sum_{r=1}^{\infty} C_{qr}D_{m-q,n+r} - \sum_{r=n+1}^{\infty} C_{qr}D_{m-q,r-n}\right). \tag{55}$$

Again, it is worth emphasizing that in (54a), there is no summation over repeated indices implied, so for instance $k_{zr}T_{qr}$ is simply the rth wavenumber k_{zr} multiplied by the Fourier component T_{qr}, with no summation over r.

Exercise 7 *Carefully go through the projection steps to derive* (54a) *from* (51)*.*

Since each of the p_{kl} in (54a) is related to the corresponding T_{kl} through (52), equation (54a) can be re-written in the form

$$\frac{\mathrm{d}T_{mn}}{\mathrm{d}t} = F_{mn}(T_{1,1}, T_{1,2}, \ldots).$$

It therefore fits into the framework of dynamical systems in note 1.[13] What is more, the linearization procedure of note 1 about the steady state $T_{mn} \equiv 0$ leads precisely to (30).

The only caveat is that the dynamical system here does not have a finite number of *degrees of freedom*: in equation (13), each component y_i of the vector $\mathbf{y}$ is a degree of freedom of the dynamical system, which has a finite number (denoted by n in note 1) of such degrees of freedom (or of *dimensions*). The counterpart of $\mathbf{y}$ here, the collection of Fourier coefficients T_{mn}, is not finite-dimensional: that is because we have tried to represent a partial differential equation system (49) as a set of ordinary differential equations, and the price to be paid is that we have, in principle, an infinite number of them.

The formulation in equation (54a) does however lead to a plausible way of solving the nonlinear convection problem (49) computationally: we can *truncate* the Fourier series (50) for T' and p' as

$$T'(x,z,t) = \sum_{m=-N_1}^{N_1} \sum_{n=1}^{N_2} T_{mn}(t)\phi_m(x)\sin(k_{zn}z), \tag{56a}$$

$$p'(x,z,t) = \sum_{m=-N_1}^{N_1} \sum_{n=1}^{N_2} p_{mn}(t)\phi_m(x)\cos(k_{zn}z). \tag{56b}$$

We then solve (54a) combined with (52) only for T_{mn} and p_{mn} with indices $|m| \leq N_1$ and $n \leq N_2$, treating T_{kl} and p_{kl} with indices k and l outside these ranges in (54a) as zero. This reduces the problem to a finite number of ordinary differential equations. This method of solving (49) computationally is known as a *spectral method* (or more specifically, as a *Galerkin method*).

Using that approach, viewing (54a) as a finite collection of ordinary differential equations, allows us to solve for the temperature and pressure fields to the point where the perturbations are no longer small and unstable growth stops. One example is shown in figure 6, where we can see that a convection cell very similar to those shown in figures 1 and 5 is the final result of the instability.

The difference here is that the cell is effectively no longer growing. This is difficult to see in the plots in figure 6, but we can see the approach to steady state illustrated in the figure by plotting the magnitude of the temperature perturbation (denoted here by $||T||$) plotted against time; we define magnitude in a root mean square way:

$$||T'|| = \sqrt{\int_0^a \int_0^1 |T'(x,z,t)|^2 \,\mathrm{d}z\,\mathrm{d}x}$$

[13] Obviously you need to relabel the indices from a double to a single index to do this. There are many ways of doing this: it is analogous to identifying each entry in a matrix by a single integer label, rather than by two integers. The function *sub2ind* in MATLAB for instance does this for a finite dimensional, $N_1 \times N_2$ matrix. In that case, we can easily define a one-to-one mapping from (m,n) to a single index $p(m,n) = (n-1)N_1 + m$. m and n can be recovered as $m = ((p-1) \bmod N_1) + 1$ and $n = (p-m)/N_1 + 1$.

As a result of the integral (which is the way to average over space), $||T'||$ does not depend on position, but it does depend on time t. By an extension of Parseval's theorem (see the notes on Fourier series), this is easy to compute for our truncated Fourier series as

$$||T'|| = \sqrt{\frac{a}{2}\sum_{n=1}^{N_1}\sum_{m=1}^{N_2}|T_{mn}(t)|^2}. \tag{57}$$

For the same solution that is plotted as snapshots in figure 6, the evolution of $||T'||$ is plotted against time in figure 7

The important thing to be clear on here is that the steady state the system reaches is not the simple (or *trivial*), laterally uniform steady state solution $(\bar{T}, \bar{p})$ that we computed in (7) (which corresponds to $T' \equiv 0$), but one with spatial structure in the form of a convection cell of finite amplitude.

Note 6 *Computationally,* (54a) *is solved here using a so-called* backward Euler step. *Given a solution $y_i(t)$ at time t and a dynamical system $\mathrm{d}y_i/\,\mathrm{d}t = F_i(y_j)$, we approximate $y_i(t+\delta t)$ by solving*

$$y_i(t+\delta t) - y_i(t) = \delta t F_i(y_1(t+\delta t), \ldots, y_N(t+\delta t))$$

for $i = 1, \ldots, N$, which becomes more and more accurate the smaller we make the discrete time step δt. This is a nonlinear rootfinding problem for the updated solution $y_i(t+\delta t)$, solved here using a method know as Newton's method. *Basically, we are solving an equation of the form*

$$G_i(x_1, x_2, \ldots, x_N) = 0,$$

for a set of functions G_1, G_2, $\ldots G_N$, if we treat the unknowns x_i as representing $y_i(t+\delta t)$, and define $G_i(x_i) = x_i - y_i(t) - \delta t F_i(x_1, \ldots, x_N)$. Newton's method solves such rootfinding problems by iterating. *It starts with an initial guess $x_i^{(0)}$, and uses a set formula for finding an updated guess $x_i^{(1)}$, $x_i^{(2)}$ etc, where the superscript in round brackets indicates the number of updates (or* iterations*) that have been computed, rather than representing exponentiation. The formula that connects an update $x_i^{(n)}$ to the previous values $x_1^{(n-1)}$, $x_2^{(n-1)}$, $\ldots x_N^{(n-1)}$ is*

$$x_i^{(n)} = x_i^{(n-1)} - \sum_{j=1}^{n} J_{ij}^{-1} G_j\left(x_1^{(n-1)}, \ldots x_N^{(n-1)}\right),$$

where J_{ij} is the Jacobian matrix defined in the same sense as in note 1,

$$J_{ij} = \left.\frac{\partial G_i}{\partial x_j}\right|_{(x_i^{(n-1)}, x_2^{n-1}, \ldots, x_N^{(n-1)})}.$$

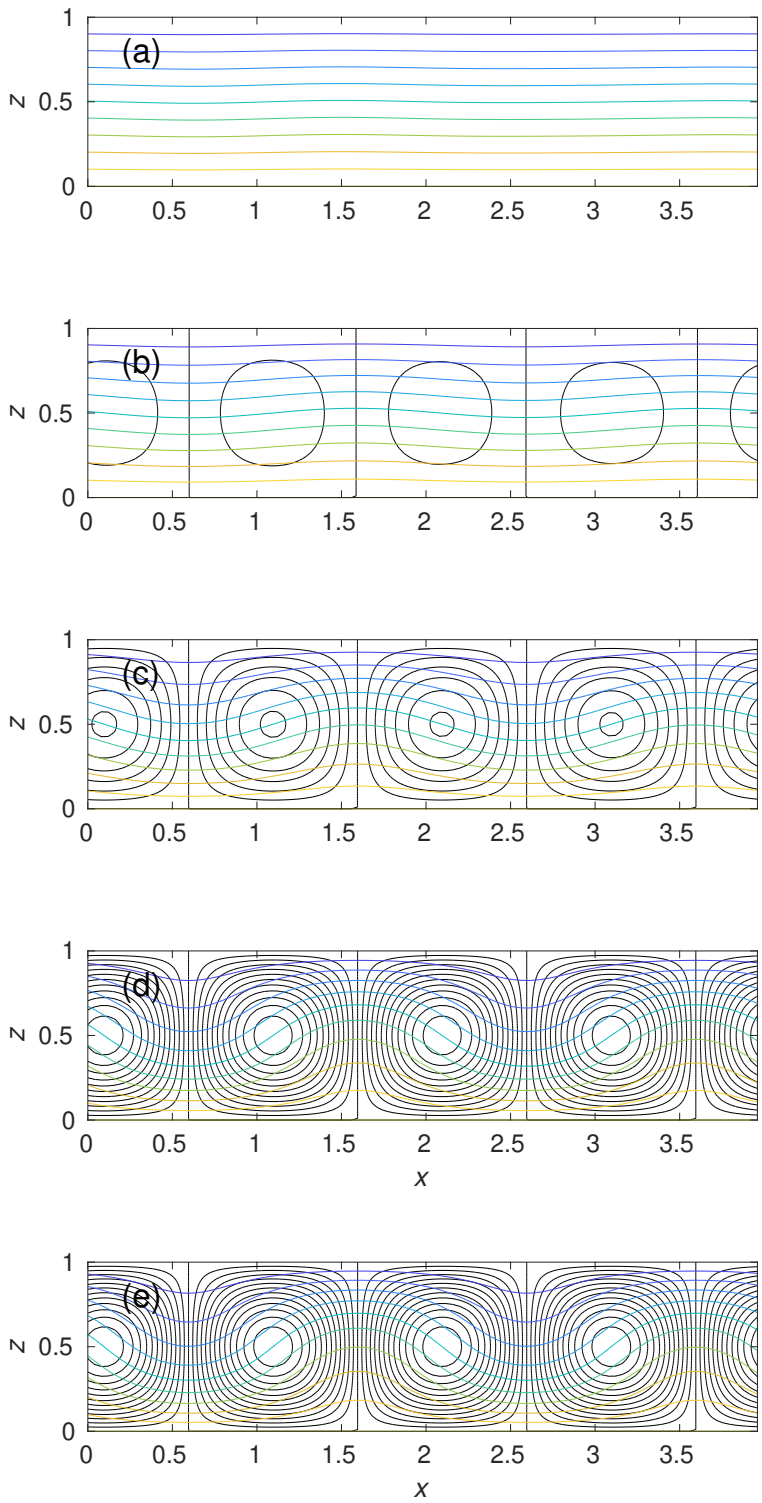

Figure 6: The evolvution of a convection pattern from random initial conditions, with $a = 4$ and $Ra = 9Ra_c/8 = 9\pi^2/5$, shown as streamlines (black) and temperature contours (colour) at times $t = 25$ (panel a), $t = 50$ (b), $t = 75$ (c) and $t = 100$ (d) and $t = 125$ (e). Contour interval for the streamfunction Ψ is 2.5×10^{-3}, and 0.1 for T in all panels

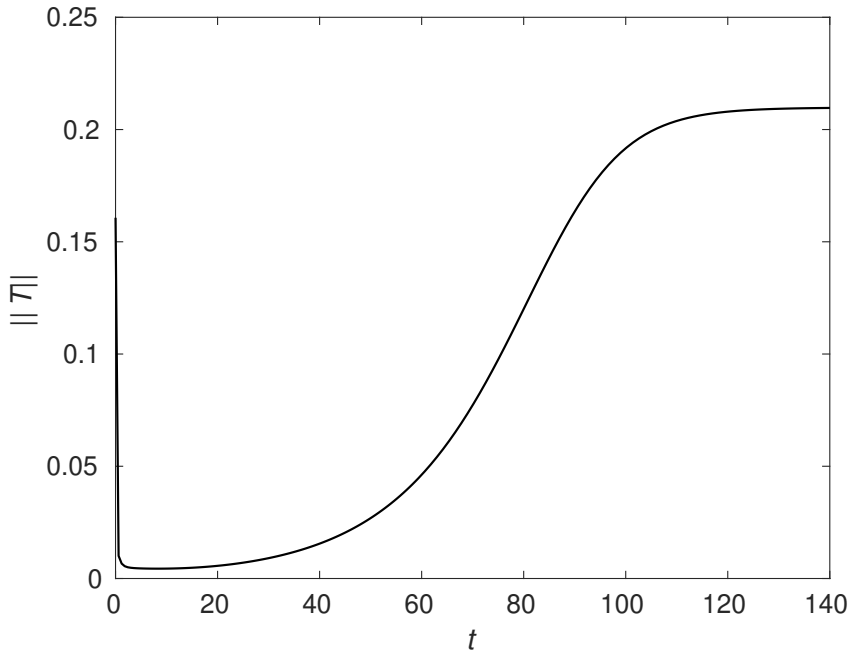

Figure 7: The evolution of $||T'||$ against t for the same solution as shown in figure 6. Early on, you will see a rapid decrease in $||T'||$: this is due to the decay in the modes predicted to be stable by the linear stability analysis, which is much faster than the growth of the unstable mode for the value of the Rayleigh number used here, which only exceeds the critical value Ra_c by a small amount. You can then see the exponential growth of that unstable mode, which eventually saturates in a way that might be reminiscent of solutions of the logistic equation $\mathrm{d}y/\,\mathrm{d}t = \lambda y(1 - y/y_0)$. That similarity is not accidental: exercise 13 shows that $||T'||^2$ approximately satisfies the logistic equation in time.

and J_{ij}^{-1} denotes the elements of the matrix inverse of J_{ij} (meaning, it emphatically does not denote the element-wise division $1/J_{ij}$, but is the (i,j)th component of the inverse $\mathbf{J}^{-1}$ of the matrix $\mathbf{J}$ defined by the components J_{ij}). In less fussy vector notation, we can also write

$$\mathbf{x}^{(n+1)} = \mathbf{x}^{(n)} - \mathbf{J}\left(\mathbf{x}^{(n)}\right)^{-1}\mathbf{F}\left(\mathbf{x}^{(n)}\right),$$

where we have also made explicitly clear that the Jacobian $\mathbf{J}$ is evaluated at the previous iterate $\mathbf{x}^{(n)}$

This scheme converges *if the initial guess $x_i^{(0)}$ is close enough to a true solution for all indices i, in the sense that in the limit $n \to \infty$, the iterates $x_i^{(n)}$ become the true solution. The iteration scheme relies on approximating the function G by a linear approximation scheme based on a Taylor expansion,*

$$G_i\left(x_1^{(n)},\ldots,x_N^{(n)}\right) \approx G_i\left(x_1^{(n-1)},\ldots,x_N^{(n-1)}\right) + \sum_{j=1}^{N} \left.\frac{\partial G_i}{\partial x_j}\right|_{(x_i^{(n-1)},x_2^{n-1)},\ldots,x_N^{(n-1)})} \left(x_j^{(n)} - x_j^{(n-1)}\right),$$

setting the right-hand side to zero. Newton's method will feature in any basic course on numerical analysis or scientific computing, and you should aim to take such a course.

Exercise 8 *Write a code to solve*

$$\frac{\mathrm{d}y_1}{\mathrm{d}t} = \lambda_1 y_1(1-y_2), \qquad \frac{\mathrm{d}y_2}{\mathrm{d}t} = \lambda_2 y_2(1-y_2)$$

using backward Euler steps and Newton's method. Experiment with the step size δt. Adapt the code to solve the dynamical system in exercise 3 by backward Euler steps. Compare the results you obtain with those computed using a MATLAB or Python ordinary differential equation solver. Next, adapt it to solve

$$\begin{aligned}\frac{\mathrm{d}y_1}{\mathrm{d}t} &= F_1(y_1,y_2;\mu) = \mu y_1 - y_1^3 + y_1 y_2\\ \frac{\mathrm{d}y_2}{\mathrm{d}t} &= F_2(y_1,y_2;\mu) = \mu - y_2 + c y_1^2.\end{aligned}$$

Note 7 *The reason for not using a regular ordinary differential equation solver to solve the dynamical system* (54a) *is that some time stepping routines used for ordinary differential equations may not be* stable *for dynamical systems obtained by* discretizing *partial differential equations (for instance, by truncating the double sum in* (56)*). The notion of stability involved is subtly different from that explored in note 1 in the sense that it refers to growth of perturbations associated purely with the transition from a continuous to a discrete system of equations, rather than an instability in the underlying dynamical system. You will need a course on numerical differential equations, scientific computing or numerical analysis to delve into this more deeply.*

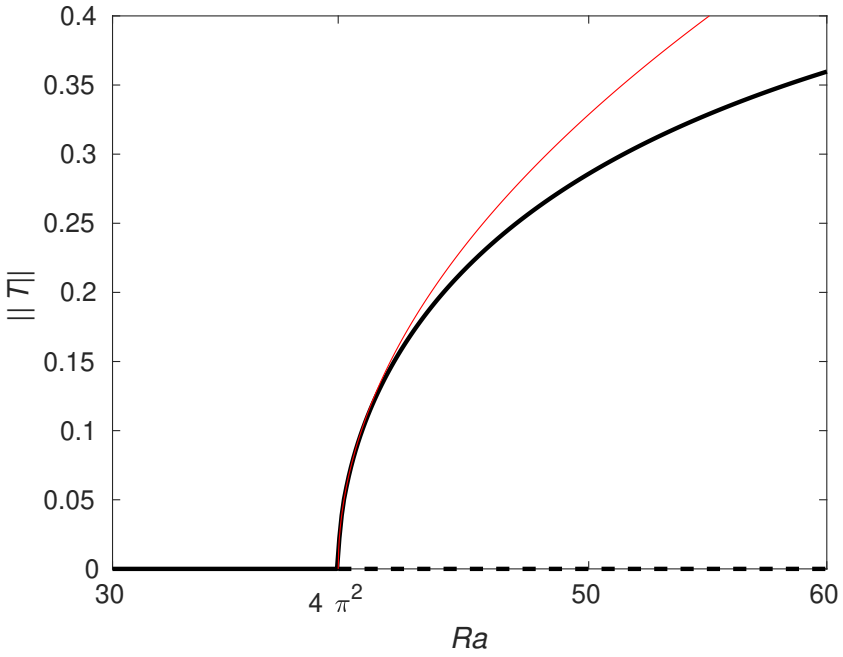

Figure 8: The magnitude $||T'||$ of steady state solutions against Ra. The solid black line corresponds to stable steady state solutions computed numerically, the dashed line to unstable steady state solutions. Note how the unstable and stable solutions meet at $Ra = Ra_c$. The red line shows $f(Ra) = \pi^{-2}(Ra - Ra_c)^{1/2}$; this is formula (92) for the behaviour of $||T'||$ near the bifurcation at $Ra = Ra_c$, which can be deduced analytically (see exercise 13

When solving partial differential equations numerically, it is nice to be able to illustrate the behaviour of a single realization as in figure 6, meaning, for a particular set of initial conditions and parameters.

That is often not the primary purpose, however. Yes, we may want to know what a convection pattern looks like. But what you might really want to know is how some aspect of the the convection pattern changes as we change the forcing of the system: in our case, that forcing can be taken for instance as the difference $T_b - T_s$ between bottom and top temperatures. In other words, we may ask how some aspect of the convection patterns depend on Ra, the only parameter in the system. The study of dynamical systems is really about exactly that: it addresses questions such as 'how do steady states change as we change a parameter value in the system'.[14]

Figure 8 shows how the amplitude of the temperature perturbation T' *in steady*

[14]The study of dynamical systems does much more besides, such as identifying recurring non-steady solutions, for instance periodic oscillations, or more complex, 'chaotic' solutions, and how they arise as a parameter value is changed.

state depends on Rayleigh number Ra. In other words, we compute lots of steady state solutions for different values of Ra and plot their magnitude against Ra. Naively, you might imagine this is done by solving the evolution problem (54a) against time for different values of Ra and waiting till a steady state is reached. That is actually a highly inefficient way of doing so, and computational approaches to dynamical systems offer much faster ways of achieving the same thing.[15]

What the figure makes clear is that, as Ra passes the critical value Ra_c, the onset of convection is gradual. The magnitude of steady state convection cells grows continuously from zero as Ra_c is passed (note that 'grow' is not meant in the sense of growing over time here: it refers to how the steady state depends on the parameter value Ra). In addition, these cells themselves are stable for $Ra > Ra_c$, while the 'trivial' steady state solution $T' \equiv 0$ flips from being stable for $Ra < Ra_c$ to unstable at $Ra > Ra_c$; the critical parameter value Ra_c at which this flip happens (and where the stable and unstable solution 'meet' in figure 8) is known as a *bifurcation*.

In many instances, looking at how a solution changes as a result of parameter changes is understood as a 'sensitivity analysis'. The latter phrase means the following: I suppose (or pretend) that I know the parameter values for a system I am interested in but admit that I do not know them exactly, so I ask how sensitive the solution is to changing those parameter values. Which is generally done in a linearized way, looking only at how much the solution changes if I change parameter values 'a little bit'. That is, I change the parameter values only by an amount that still allows a first order Taylor expansion in the parameter to determine how much the solution changes by. That is effectively what equation (60) in exercise 9 below does. What a sensitivity analysis misses, because it is linear, is the qualitative change in behaviour that occurs at a bifurcation. The next exercise explores this in more detail.

Exercise 9 *Bifurcations occur where the Jacobian of a dynamical system is singular (where the Jacobian matrix does not have an inverse), which is precisely where the growth rate σ goes to zero (since σ is an eigenvalue of the Jacobian matrix, and matrices with a zero eigenvalue are singular).*[16] *This occurs because of something called the* implicit function theorem*, which basically works as follows: consider a*

[15] Specifically, you can find steady states using so-called *continuation methods*, employing Newton's method to solve steady state versions of the dynamical system

$$F_i(\bar{y}_1, \ldots, \bar{y}_N; Ra) = 0$$

for $i = 1, \ldots, N$ while also changing the parameter Ra. In the notation here, we have explicitly written the parameter Ra as an argument of F_i to make clear that the dynamical system depends on that parameter.

[16] To be more precise, you will see in this exercise what is meant by one particular kind of bifurcation; there are others that are a little bit more complicated: for instance, a so-called Hopf bifurcation occurs when the Jacobian has a pair of purely imaginary eigenvalues rather than a zero eigenvalue, and corresponds to the appearance of an oscillatory solution; a course on dynamical systems theory will clarify this.

dynamical system in steady state in the form

$$F_i(y_1, \dots y_N; \mu) = 0, \qquad i = 1, \dots, N, \tag{58}$$

where μ is a parameter that the dynamical system depends on (like Ra in our convection problem). When including a parameter explicitly as an argument of the function $\mathbf{F}$ *defining a dynamical system, it is customary to separate the symbol for the parameter from the dynamical degrees of freedom y_i by a semicolon rather than a comma. Suppose that $y_i = \bar{y}_i(\mu)$ is the steady state solution corresponding to parameter value μ. If the Jacobian matrix*

$$J_{ij} = \left.\frac{\partial F_i}{\partial y_j}\right|_{(\bar{y}_1(\mu), \bar{y}_2(\mu), \dots, \bar{y}_N(\mu); \mu)}.$$

is non-singular and the parameter μ is changed by a sufficiently small amount from μ_0, then there is a 'locally unique' solution to (58) *whose dependence on μ we can compute by using chain rule, differenting*(58) *with respect to μ to give*

$$0 = \frac{\mathrm{d}}{\mathrm{d}\mu} F(\bar{y}_1(\mu), \dots \bar{y}_N(\mu); \mu) = \sum_{j=1}^{N} J_{ij} \frac{\mathrm{d}\bar{y}_j}{\mathrm{d}\mu} + \frac{\partial F_i}{\partial \mu} \tag{59}$$

for $i = 1, \dots, N$, where $\frac{\partial F_i}{\partial \mu}$ is also evaluated at $(\bar{y}_1(\mu), \dots, \bar{y}_N(\mu); \mu_0)$. Note that (59) *is a system of linear equations in the unknowns $\mathrm{d}\bar{y}_j/\,\mathrm{d}\mu$, and because J_{ij} is not singular, the system can be solved as*

$$\frac{\mathrm{d}\bar{y}_j}{\mathrm{d}\mu} = -\sum_{j=1}^{N} J_{ij}^{-1} \frac{\partial F_j}{\partial \mu}.$$

A first order Taylor expansion therefore gives

$$\bar{y}_i(\mu + \delta\mu) \approx \bar{y}_i(\mu) - \sum_{j=1}^{N} J_{ij}^{-1} \frac{\partial F_j}{\partial \mu} \delta\mu. \tag{60}$$

This scheme breaks down when J_{ij} is singular, and we have a bifurcation: there can be multiple solutions to the problem (58) *in the vicinity*[17] *of a point where J_{ij} is singular. The study of bifurcations focuses on identifying how many such steady state solutions there are, how they depend on changes in the paramter μ away from the critical value at which J_{ij} is singular, and which of these multiple solutions are stable and unstable.*

To make headway, you need to go to higher order in the Taylor expansion of F_i, which note 8 sketches out and exercise 13 applies to the convection problem. To get

[17] Or 'neighbourhood', if you are familiar with the basic concepts of topology.

a full understanding of how this works, you should also take a specialized course in dynamical systems.

Here we simply illustrate how bifurcations work by looking at a very simple problem with a single dynamical degree of freedom:

$$\frac{\mathrm{d}y}{\mathrm{d}t} = \mu y - y^3 \tag{61}$$

Equation (61) *fits into the dynamical systems framework by taking* $N = 1$ *and dropping the subscript* i *on* $\mathrm{d}y_i/\,\mathrm{d}t = F_i(y_1, \ldots; \mu)$ *since there is only one* y_i *with* $i = 1$*; also* $F(y, \mu) = \mu y - y^3$*. Identify where there is a bifurcation (i.e., at what critical value of* μ *does the 'Jacobian' become singular*[18]*). Find* all *real steady state solutions for* y *below and above the bifurcation, and determine if they are stable or not. Plot all the stable steady state magnitude* y *against the parameter* μ*. Also solve* (61) *analytically.* Hint: this is easiest if you multiply both sides by y and solve for y^2 as the dynamical degree of freedom instead. You should get a version of the logistic equation for y^2 when you do so; bear in mind however that y^2 must be positive.

Now repeat the exercise with

$$\frac{\mathrm{d}y}{\mathrm{d}t} = \mu y + y^3 \tag{62}$$

What changes?

The bifurcation you find here is known as a pitchfork bifurcation *(plotting* y *against* μ *should make clear why that name is chosen!), and this is actually a generically the same as the bifurcation that the convection problem undergoes at* $Ra = Ra_c$ *(see exercise 11 below. Which of the two cases you have worked through is analogous to the convection results we show in diagram 8?*

The most common quantity that we are likely to be interested in is not so much the amplitude $||T'||$, but the heat flux. The conductive heat flux across the system in the absence of convection is simply

$$k(T_b - T_s)/h.$$

Once convection commences, we expect to add to that rate of heat transfer: convection carries hot fluid upwards and brings it into closer contact with the cold upper surface. But how much does it add? Can we write heat flux — or more specifically, the heat flux averaged across the lower or upper surface, since the actual, local heat flux is no longer uniform — as a function of the temperature difference $T_b - T_s$ between bottom and top of the porous medium?

That obviously makes most sense if we do the computation when the system is in steady state, rather than during the transient evolution shown in figure 6. In

[18] The Jacobian of a one-by-one matrix $A_{ij} = a$ is simply a, if you wish.

dimensional terms, the spatially averaged heat flux out of the bottom of the domain (which we denote by Q_0) is

$$Q_0 = \frac{k(T_b - T_s)}{h} \times a^{-1} \int_0^a - \left.\frac{\partial T}{\partial z}\right|_{z=0} \mathrm{d}x = \frac{k(T_b - T_s)}{h} \left[1 - a^{-1} \int_0^a \left.\frac{\partial T'}{\partial z}\right|_{z=0} \mathrm{d}x\right]$$

where T, z and x continue to be dimensionless. Using the definition of T' in terms of the Fourier modes T_{mn}, this becomes

$$Q_0 = \frac{k(T_b - T_s)}{h} \left[1 - \sum_{m=1}^{N_2} k_{zm} T_{0m}\right],$$

The term in square brackets is determined the steady state solution of the dimensionless problem (54a), and therefore depends on Ra, so we can write

$$Q_0 = \frac{k(T_b - T_s)}{h} Nu(Ra).$$

where

$$Nu(Ra) = \left[1 - \sum_{m=1}^{N_2} k_{zm} T_{0m}\right] \tag{63}$$

determines, as a function of Ra, how much heat flux is enhanced by convection: it is the ratio of Q_0 to the purely conductive flux. Nu is, rather confusingly, referred to as the *Nusselt number*. Unlike the dimensionless groups you have met so far, also often named after long-dead male European or American scientists, Nu is not a dimensionless group that can be computed just by non-dimensionalizing an underlying system of equations, but a dimensionless ratio of heat fluxes that need to be computed by solving those equations.

Figure 9 shows the Nusselt number as a function of Ra for the steady state solutions also shown in figure 8. Below Ra_c, the Nusselt number is one, which is to be expected since heat transport is purely by conduction. Beyond Ra_c, the Nusselt number increases fairly linearly with Ra. The Rayleigh number Ra is of course itself linearly dependent on $T_b - T_s$ (see its definition in (4)), so Nu increasing with Ra implies a mean heat flux Q_0 increasing faster than linearly with temperature difference $T_b - T_s$. (There is a general lesson for thermal engineering here: typically, you want to suppress convection as well as reduce thermal conductivity k in insulation layers.)

Note that figures 8 and 9 show approximate forms for $||T'||$ as red lines. These approximate forms are highly accurate near the bifurcation at $Ra = Ra_c$, and that is no accident. The remaining exercises and note expand more on what happens at the bifurcation at $Ra = Ra_c$, expanding on exercise 9 above. In particular, we show that the local form of the dependence of $||T'||$ and of Nu on Ra near $Ra = Ra_c$ can be predicted without numerical computation — but as the work involved is also likely to make clear, the price for doing so is a great deal of algebra that may not be worthwhile if you can solve the problem efficiently on a computer.

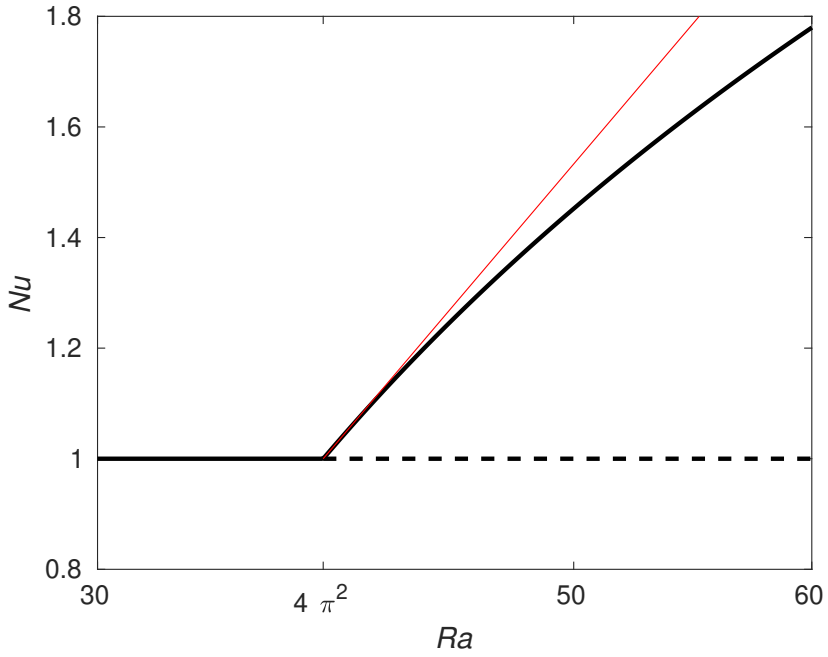

Figure 9: The Nusselt number Nu of steady state solutions against Ra. The solid black line corresponds to stable steady state solutions computed numerically, the dashed line to unstable steady state solutions. The red curve shows $f(Ra) = 1 + \pi^{-2}(Ra - Ra_c)/2$; this is formula (93), which can be computed analytically as the approximate form of the actual Nusselt number near the bifurcation

Exercise 10 *Consider the system*

$$\frac{\mathrm{d}y_1}{\mathrm{d}t} = F_1(y_1, y_2; \mu) = \mu y_1 - y_1^3 + y_1 y_2$$
$$\frac{\mathrm{d}y_2}{\mathrm{d}t} = F_2(y_1, y_2; \mu) = \mu - y_2 + c y_1^2.$$

Show that there is a bifurcation at $\mu = 0$, where the eigenvalues of the Jacobian are 0 and -1, with eigenvectors $(1,0)^{\mathrm{T}}$ and $(0,1)^{\mathrm{T}}$, respectively.

Show that the steady state solution can be written in the form $y_1 = |\mu|^{1/2} A_1$, $y_2 = |\mu| A_2$, with A_1 satisfying an equation of the form

$$S a A_1 - b A_2^3 = 0 \tag{64}$$

with $S = \mu/|\mu|$ and a, b independent of μ. Determine when there is one and when there are three roots of this equation. Also determine whether these solutions are stable or unstable.

Now modify the dynamical system to

$$\frac{\mathrm{d}y_1}{\mathrm{d}t} = F_1(y_1, y_2; \mu) = \mu y_1 - y_1^3 + y_1 y_2 + d_1 y_1^4 + d_2 y_1 y_2^2 \tag{65a}$$
$$\frac{\mathrm{d}y_2}{\mathrm{d}t} = F_2(y_1, y_2; \mu) = \mu - y_2 + c y_1^2 + e_1 y_1 y_2^2 + e_2 y_2^2. \tag{65b}$$

Suppose that μ is close to the critical value of 0, i.e., that $|\mu|$ is small. Scale y_1 and y_2 as before, $y_1 = |\mu|^{1/2} A_1$, $y_2 = |\mu| A_2$, and substitute for y_1 and y_2 in terms of A_1 and A_2 in (65). Show that, neglecting higher powers of $|\mu|$, you get the same equation for A_1 as before, (64).

Note 8 *Here we expand on exercises 9 and 10 to describe a more general theory of pitchfork bifurcations. This lays the theoretical ground work for showing that the onset of convection is indeed a pitchfork bifurcation ultimately described by a simple model analogous to (61): a single first-order ordinary differential equation whose right-hand side is the sum of a linear and cubic term. The actual demonstration of this fact follows in exercise 13 below. There are two ways in which you can tackle this note and the subsequent exercises: if you are on a more abstract bent, read the note first and do the exercise afterwards. Otherwise, try the exercise first, which is sufficiently self-contained (though therefore lacking in generality) to do that.*

Consider a dynamical system

$$\frac{\mathrm{d}\mathbf{y}}{\mathrm{d}t} = \mathbf{F}(\mathbf{y}; \mu),$$

(or, in index notation,

$$\frac{\mathrm{d}y_i}{\mathrm{d}t} = F(y_1, \ldots, y_N; \mu), \qquad i = 1, \ldots, N),$$

and let $\bar{\mathbf{y}}(\mu)$ *(or interchangeably,* $\bar{y}_i(\mu)$ *for* $i = 1, \ldots, N$*) denote the steady state solution for parameter value* μ*. Suppose that there is a critical parameter value* $\mu = \mu_0$ *at which the Jacobian* $\mathbf{J}$ *defined through*

$$J_{ij} = \left.\frac{\partial F_i}{\partial x_j}\right|_{(\bar{y}_1(\mu_0),\ldots,\bar{y}_N(\mu_0),\mu_0)}$$

is singular with a single vanishing eigenvalue.[19] *Remember that this generally signifies a bifurcation, where multiple steady state solutions can 'meet' as in exercise 9. Also assume that all the other eigenvalues of* J_{ij} *have negative real parts (if that was not satisfied, the steady state solution would automatically be unstable, and the behaviour of the dynamical system near that steady state solution would nor be of significant interest since the solution would definitely evolve away.)*

This note will describe how to find the behaviour of the solution $\mathbf{y}$ *near the bifurcation by using the* local *behaviour of* F*, as defined by the first few terms of its Taylor expansion around the steady state* $\bar{\mathbf{y}}(\mu_0)$ *at the critical value* $\mu = \mu_0$*.*

Our first step is somewhat technical, transforming from the $(y_1(t), \ldots, y_N(t))$ *to a new set of dynamical degrees of freedom* $(\alpha_1(t), \ldots, \alpha_N(t))$*. The degrees of freedom* α_i *we choose make the linearized problem much simpler by* diagonalizing *the Jacobian. Diagonalizing a matrix by using its eigenvectors is hopefully something you learnt about in a linear algebra course. The procedure below is one of many reasons why that procedure is an important tool, which we get to use in context here.*

To see how it works, let the eigenvectors of the matrix $\mathbf{J}$ *be* $\mathbf{e}_i$*,* $i = 1, \ldots, N$ *(since there are* N *eigenvectors), with corresponding eigenvalues* σ_i*, and order the eigenvector labels* i *such that* $\sigma_1 = 0$ *is the zero eigenvalue. For simplicity, assume that all the other eigenvalues are distinct, so the characteristic polynomial has no repeated roots. The eigenvectors are then linearly independent of each other, and can be arranged in an invertible matrix* $\mathbf{E}$ *defined through*

$$E_{ij} = e_{i,j}$$

where $e_{i,j}$ *is the* i*th component of* $\mathbf{e}_j$*; in other words,* $\mathbf{E}$ *is formed by lining the eigenvectors* $\mathbf{e}_i$ *up next to each other:*

$$\mathbf{E} = (\mathbf{e}_1, \ldots, \mathbf{e}_N).$$

Crucially, $\mathbf{E}$ *has an inverse* $\mathbf{E}^{-1}$*, because the eigenvectors* $\mathbf{e}_i$ *(the columns of* $\mathbf{E}$*) are linearly independent. The matrix* $\mathbf{E}$ *also has the property that*

$$\mathbf{JE} = (\mathbf{Je}_1, \mathbf{Je}_2, \ldots \mathbf{Je}_N) = (\sigma_1\mathbf{e}_1, \sigma_2\mathbf{e}_2, \ldots, \sigma_N\mathbf{e}_N)\,,$$

and it is straightforward to show that this equals

$$\mathbf{JE} = \mathbf{E}\mathbf{\Lambda} \tag{66}$$

[19] Meaning, the eigenvalue $\sigma = 0$ is not a repeated root of the characteristic polynomial.

where

$$\mathbf{\Lambda} = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_N) = \begin{pmatrix} \sigma_1 & 0 & \ldots & 0 \\ 0 & \sigma_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_N \end{pmatrix}$$

is the diagonal matrix created by the eigenvalues; suspending the summation convection, we could also write $\Lambda_{ij} = \sigma_i \delta_{ij}$ *(where there is no summation over* i *implied).*

A general solution vector $\mathbf{y}$ *can be expressed in terms of the eigenvectors* $\mathbf{e}_i$ *as a so-called* basis. *What we mean by that is that we can* always *write* $\mathbf{y}(t)$ *in the form*

$$\mathbf{y}(t) = \bar{\mathbf{y}} + \sum_{i=1}^{N} \alpha_i(t)\mathbf{e}_i.$$

which is the same as (19) *in note 1. Written this way, the steady state at* $\mu = \mu_0$ *is simply* $\alpha_i \equiv 0$. *In terms of the matrix* $\mathbf{E}$, *the relationship between* $\mathbf{y}$ *and the* α_i *can be expressed alternatively in the form*

$$\mathbf{y} - \bar{\mathbf{y}} = \mathbf{E}\boldsymbol{\alpha},$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)^{\mathrm{T}}$, *the superscript* T *denoting the usual transpose operation.*

To show that the coefficients α_i *exist and are unique for a given* $\mathbf{y}$, *it suffices to solve this equation for* $\boldsymbol{\alpha}$ *as*

$$\boldsymbol{\alpha} = \mathbf{E}^{-1}(\mathbf{y} - \bar{\mathbf{y}}).$$

In the same vein, we can re-write the dynamical system in the form

$$\frac{\mathrm{d}\boldsymbol{\alpha}}{\mathrm{d}t} = \frac{\mathrm{d}(\mathbf{E}^{-1}(\mathbf{y} - \bar{\mathbf{y}}))}{\mathrm{d}t} = \mathbf{E}^{-1}\frac{\mathrm{d}\mathbf{y}}{\mathrm{d}t} = \mathbf{E}^{-1}\mathbf{F}(\mathbf{y};\mu) = \mathbf{E}^{-1}\mathbf{F}(\bar{\mathbf{y}} + \mathbf{E}\boldsymbol{\alpha};\mu). \tag{67}$$

With this, solving for $\boldsymbol{\alpha}$ *is equivalent to solving for* $\mathbf{y}$ *since we can now transform back and forth between the two formulations.*

In order to tidy up notation going forward, we give the difference between the parameter μ *and its critical value* μ_0 *a new name*

$$\nu = \mu - \mu_0,$$

and define a new function $\mathbf{G}(\boldsymbol{\alpha};\nu)$ *to be equal to the right-hand side of* (67)*:*

$$\frac{\mathrm{d}\boldsymbol{\alpha}}{\mathrm{d}t} = \mathbf{G}(\boldsymbol{\alpha};\nu) \tag{68}$$

where

$$\mathbf{G}(\boldsymbol{\alpha};\nu) = \mathbf{E}^{-1}\mathbf{F}(\bar{\mathbf{y}} + \mathbf{E}\boldsymbol{\alpha};\mu). \tag{69}$$

Why go to all that trouble, you say? The reason is that the Jacobian of $\mathbf{G}$ *is much simpler than the Jacobian of* $\mathbf{F}$*. Define*

$$\tilde{J}_{ij} = \left.\frac{\partial G_i}{\partial \alpha_j}\right|_{(\alpha_1,\ldots,\alpha_N;\nu)=(0,\ldots,0;0)}.$$

By the chain rule

$$\begin{aligned}\tilde{J}_{ij} &= \sum_{k=1}^{N}\sum_{l=1}^{N} E^{-1}_{ik}\frac{\partial F_k}{\partial y_l}\frac{\partial y_l}{\partial \alpha_j} \\ &= \sum_{k=1}^{N}\sum_{l=1}^{N} E^{-1}_{ik} J_{kl} E_{li}\end{aligned}$$

or

$$\tilde{\mathbf{J}} = \mathbf{E}^{-1}\mathbf{J}\mathbf{E} = \mathbf{E}^{-1}\mathbf{E}\mathbf{\Lambda} = \mathbf{\Lambda}$$

since $\mathbf{JE} = \mathbf{\Lambda}$ *from* (66)*. In other words, the Jacobian after transforming to the* $(\alpha_1, \ldots, \alpha_N)$ *as the dynamical degrees of freedom is a diagonal matrix; if we linearize the problem, the different degrees of freedom* decouple *from each other. Note that this is automatically the case in the linearized convection problem, where the evolution of the Fourier coefficient* T_{mn} *only depends on* T_{mn} *with the same indices* m *and* n*, but not any other* T_{pq} *with* $p \neq m$ *or* $q \neq n$*; this is equation* (30)*). Similarly, the dynamical system in exercise 10 is already diagonalized in the same way.*

Explicitly, if we linearize around the steady state $\alpha_i = 0$ *for all* i *as in note 1, we get*

$$\frac{\mathrm{d}\alpha_i}{\mathrm{d}t} \approx \sum_{j=1}^{N} \tilde{J}_{ij}\alpha_j = \sigma_i \alpha_i$$

and therefore

$$\alpha_i(t) \approx \alpha_i(0)\exp(\sigma_i t).$$

Taking stock, we have not done a lot yet beyond making the dynamical system be simpler when linearized, but we can see more clearly some of the things that happen at a bifurcation. A special role here falls to α_1*: recall that this is the coefficient that corresponds to the zero eigenvalue* $\sigma_1 = 0$*, while all the other coefficients* α_i *with* $i > 0$ *corresponds to eigenvalues with negative real parts* $\mathrm{Re}(\sigma_i) < 0$*. The linearized model therefore predicts that all* α_i *with* $i > 1$ *decay over time, while* α_1 *remains unchanged.*

That is of course not accurate, and most importantly we cannot actually say from the linearized model how α_1 *changes over time: the best we can say is that it does not grow or shrink exponentially, but it might still evolve due to nonlinear terms that the linearization of the problem throws away.*

What we will do next is show how to construct an approximate solution for the dynamical system when μ *is close to* μ_0*, which is to say, when* ν *is small, paying*

special attention to the fact that we need non-linear terms to understand the evolution of α_1. We do so by Taylor expanding the right-hand side of (67) *around the steady state $(\alpha_1, \ldots, \alpha_N; \nu) = (0, \ldots, 0; 0)$. Remember this is what we did in note 1 to justify linearization. We still expect to look only at cases where the α_i are small (as we did when we linearized in note 1) but recognizing that a linear model is insufficient, we go further in the expansion here, initially up to third order*

$$\begin{aligned}\frac{\mathrm{d}\alpha_i}{\mathrm{d}t} =& G_i(\alpha_1, \ldots, \alpha_N, \nu) \qquad (70)\\ =& G_i(0, \ldots, 0; 0) + \sum_{j=1}^{N} \tilde{J}_{ij}\alpha_j + \sum_{j=1}^{N}\sum_{k=1}^{N} H_{ijk}\alpha_j\alpha_k + \sum_{j=1}^{N}\sum_{k=1}^{N}\sum_{l=1}^{N} K_{ijkl}\alpha_j\alpha_k\alpha_l \\ &+ \frac{\partial G_i}{\partial \nu}\nu + \sum_{j=1}^{N}\frac{\partial G_i}{\partial \alpha_j \partial \nu}\alpha_j\nu + \frac{1}{2!}\frac{\partial^2 G_i}{\partial \nu^2}\nu^2 + \sum_{j=1}^{N}\sum_{k=1}^{N}\frac{\partial G_i}{\partial \alpha_j \partial \alpha_k \partial \nu}\alpha_j\alpha_k\nu \\ &+ \sum_{j=1}^{N}\frac{1}{2}\frac{\partial^3 G_i}{\partial \alpha_j \partial \nu^2}\alpha_j\nu^2 + \frac{1}{3!}\frac{\partial^3 G_i}{\partial \nu^3}\nu^3 + O(\nu^4) \qquad (71)\end{aligned}$$

where

$$\begin{aligned}H_{ijk} =& \frac{1}{2!}\left.\frac{\partial^2 G_i}{\partial \alpha_j \alpha_k}\right|_{(\alpha_1,\ldots,\alpha_N;\nu)=(0,\ldots,0;0)},\\ K_{ijkl} =& \frac{1}{3!}\left.\frac{\partial^3 G_i}{\partial \alpha_j \alpha_k \alpha_l}\right|_{(\alpha_1,\ldots,\alpha_N;\nu)=(0,\ldots,0;0)}.\end{aligned}$$

and the derivatives of G_i with respect to ν are also evaluated at $(\alpha_1, \ldots, \alpha_N; \nu) = (0, \ldots, 0; 0)$

Next, we consider only the situation in which the following holds:

$$H_{111} = 0, \qquad \frac{\partial G_1}{\partial \mu} = 0. \qquad (72)$$

To understand the probably rather obscure-looking conditions (72)*, consider the expansion above for $i = 1$: this again refers to the equation for how α_1 evolves in time. Since the $G_i(0, \ldots, 0; 0) = 0$ at the steady state and using the fact that $\tilde{J}_{ij}$ is diagonal with $\alpha_1 = 0$, we get from* (71)

$$\begin{aligned}\frac{\mathrm{d}\alpha_1}{\mathrm{d}t} =& \sum_{j=2}^{N}(H_{11j} + H_{1j1})\alpha_1\alpha_j + \sum_{j=2}^{N}\sum_{k=2}^{N} H_{1jk}\alpha_j\alpha_k + \sum_{j=1}^{N}\sum_{k=1}^{N}\sum_{l=1}^{N} K_{1jkl}\alpha_j\alpha_k\alpha_l \\ &+ \sum_{j=1}^{N}\frac{\partial G_1}{\partial \alpha_j \partial \nu}\nu\alpha_j + \frac{1}{2!}\frac{\partial^2 G_1}{\partial \nu^2}\nu^2 + \sum_{j=1}^{N}\sum_{k=1}^{N}\frac{\partial G_1}{\partial \alpha_j \partial \alpha_k \partial \nu}\nu\alpha_j\alpha_k \\ &+ \sum_{j=1}^{N}\frac{1}{2}\frac{\partial^3 G_1}{\partial \alpha_j \partial \nu^2}\nu^2\alpha_j + \frac{1}{3!}\frac{\partial^3 G_1}{\partial \nu^3}\nu^3, \qquad (73)\end{aligned}$$

making use of the assumptions in (72) *(note that we have deliberately excluded* $H_{111}\alpha_1\alpha_1$ *from the sum over* H_{ijk} *by the restricting the limits of that sum, on account of* H_{111} *being zero from* (72)*). Convince yourself that* (73) *holds by substituting the known values of* $\tilde{J}_{1j}$ *and* K_{1jk} *into* (71)*.*

It will turn out that, if $H_{111} = 0$*, then* α_1 *does not couple back into the evolution of* α_1 *until we get to the third order term* $K_{1jkl}\alpha_j\alpha_k\alpha_l$*. The conditions in* (72) *are ultimately what is required to create a* pitchfork bifurcation *as exemplified by exercise 9; when* (72) *is not satisfied, other types of bifurcation will result (for instance, there are so-called* transcritical *and* saddle-node *bifurcations that can appear).*

Now we need to pay attention to how big the different components of the vector $\boldsymbol{\alpha}$ *are likely to become. In the absence of a zero eigenvalue of the Jacobian, the implicit function theorem construction in exercise 9 suggests that we should then have all the components* α_i *be comparable in size to* $\mu - \mu_0 = \nu$ *from* (60)*. With the zero eigenvalue, the coefficient* α_1 *is likely to be much larger than* ν*, however. In fact, as mentioned in the previous paragraph, the condition* (72) *ensures that the right-hand side of the equation for* $\mathrm{d}\alpha_1/\,\mathrm{d}t$ *contains* α_1 *only as a third power, and this should be comparable in size to the biggest of the other terms on the right-hand side. That turns out to mean that* α_1 *should scale as* $|\nu|^{1/2}$*.*

In addition, because of the zero eigenvalue at $\nu = 0$*,* α_1 *actually grows very slowly, in fact at a rate comparable to* ν*: effectively, if we were to linearize in* α_1 *in* (73)*, the leading linear term on the right-hand side would be*

$$\frac{\partial G_1}{\partial \alpha_j \partial \nu} \nu \, \alpha_j,$$

suggesting an eigenvalue of of $\partial G_1/\partial\alpha_j\partial\nu\nu$ *that is zero at the bifurcation* $\nu = 0$ *as required, and proportional to* ν *in its vicinity. It turns out that this is not quite accurate, as you will see below, but gives the right scaling: the eigenvalue is proportional to* ν*, though with a more complicated coefficient of proportionality than* $\partial G_1/\partial\alpha_j\partial\nu$*, and changes in* α_1 *over time happen on a slow time scale comparable with* ν^{-1}*.*

All of the above leads us to scale time and the coefficients α_i *as follows:*

$$t = |\nu|^{-1}T, \qquad \alpha_1 = |\nu|^{1/2}A_1, \qquad A_i = |\nu|A_i \qquad \text{for } i > 1.$$

If this does not make any sense to you yet, worry not: you should see below that we get a problem with a sensible solution by using these scalings (but note that the scalings here are the same as those you deduced in exercise 10, if you have worked through that exercise.) To deal with the pesky modulus signs above (necessary because we want to take a square root in one place, and we want to make sure that increasing T *corresponds to increasing* t*, that is, to moving forward in time, we also define*

$$S = \mathrm{sgn}(\nu) = \frac{\nu}{|\nu|}.$$

First, we can now set to work on (73). *We retain only the lowest power in* $|\nu|$ *on both sides of the equation, which turns out the* $|\nu|^{3/2}$ *on substituting* T, A_1, $A_2, \ldots, A_N$:

$$\frac{\mathrm{d}A_1}{\mathrm{d}T} = \sum_{j=2}^{N}(H_{11j} + H_{1j1})A_1A_j + K_{1111}A_1^3 + S\frac{\partial^2 G_1}{\partial\alpha_1\partial\nu}A_1. \tag{74}$$

Convince yourself that this is correct by doing the substitution and identifying the power of ν *multiplying each term.*

Next, we have to tackle the remaining equations in (71), *with* $i > 1$. *Again retaining only the lowest power in* $|\nu|$, *which is in this case* $|\nu|$ *itself,*

$$0 = \sigma_i A_i + H_{i11}A_1^2 + S\frac{\partial G_i}{\partial\nu}$$

We have made use of the fact that $\tilde{J}_{ij} = \sigma_i\delta_{ij}$ *to write the first term as stated; again you can see the value of having diagonalized the Jacobian here, because it allows you to solve easily for* A_i:

$$A_i = -\sigma_i^{-1}H_{i11}A_1^2 - \sigma_i^{-1}S\frac{\partial G_i}{\partial\nu}.$$

The time derivative disappears here because the approach to this quasi-steady solution is very rapid compared with the slow evolution of A_i *(this is the basis of something called a centre manifold reduction in dynamical systems theory. That is really how the analysis of bifurcations proceeds systematically, but is something you would have to learn about in a specialized course on dynamical systems).*

Substituting for A_i *in* (74) *leads to*

$$\frac{\mathrm{d}A_1}{\mathrm{d}T} = SaA_1 - bA_1^3 \tag{75}$$

where

$$a = \frac{\partial^2 G_1}{\partial\alpha_1\partial\nu} - \sum_{j=2}^{N}\sigma_j^{-1}(H_{11j} + H_{1j1})\frac{\partial G_j}{\partial\nu} \tag{76}$$

$$b = -K_{1111} + \sum_{j=2}^{N}\sigma_j^{-1}(H_{11j} + H_{1j1})H_{j11} \tag{77}$$

Equation (75) *is called the* Landau equation. *It contains all the information you need about the behaviour of the system near the bifurcation. The behaviour of* A_1 *now comes down exclusively to the signs of* Sa *and* b.

If you did exercise 9, you will already know how this works: if Sa *and* b *have the same sign, then there are three steady state solutions,* $A_1 = 0$ *and* $A_1 = \pm\sqrt{Sa/b}$, *while* Sa *and* b *of the opposite sign permit only the 'trivial' steady state solution* $A_1 = 0$. *In addition, it is easy to say that the sign of* Sa *determines whether* A_1

grows in magnitude when A_1 is small (which requires $Sa > 0$), while b determines whether A_1 will stop growing when A_1 is large (this requires $b > 0$). In other words, the sign of Sa determines whether the trivial solution is stable or not, while b determines whether, when the trivial solution is unstable, there is corresponding non-trivial but stable solution.

If that is too brief of an explanation, read on in the next note.

Note 9 *Consider again the Landau equation* (75) *in note 8,*

$$\frac{\mathrm{d}A_1}{\mathrm{d}T} = SaA_1 - bA_1^3.$$

In detail, recall that S is simply the sign of ν, $S = \nu/|\nu|$, so by flipping the sign in the parameter perturbation (making μ great or less than the critical value μ_0), we can create either one or three steady state solutions. That is the reason for the name 'pitchfork bifurcation', in which a single solution (the handle of the pitchfork) turns into three solutions (the points of the pitchfork).

In addition, we can look at the stability of these three solutions in the standard way (linearize around the steady state!)[20] *For the trivial steady state $A_1 = 0$, linearization of the Landau equation* (61) *gives*

$$\frac{\mathrm{d}A_1}{\mathrm{d}T} \approx SaA_1$$

and the trivial steady state solution is stable if Sa is negative, and unstable if Sa is positive. Also, changing the sign of ν changes the stability of the trivial steady state: if the trivial steady state is stable for $\nu < 0$, then it becomes unstable for $\nu > 0$ and vice versa. That is what we saw already for the convection problem, where the trivial steady state with no convection is stable below Ra_c and unstable above Ra_c

Conversely, if we linearize around the non-trivial steady state $A_1 = \pm\sqrt{Sa/b}$ as $A_1 = \pm\sqrt{Sa/b} + A_1'$, we get

$$\frac{\mathrm{d}A_1'}{\mathrm{d}T} = Sa - 2b\left(\pm\sqrt{Sa/b}\right)^2 A_1' = -SaA_1'.$$

The opposite conclusion applies now: the steady state is unstable if Sa is negative, and stable if Sa is positive. In other words, where there are three steady state solutions, the trivial solution $A_1 = 0$ and the nontrivial solution $A_1 = \pm\sqrt{Sa/b}$ have opposite stability properties.

[20] This may seem a little strange, but it does work, and it is not that hard to convince yourself that it should work, purely mathematically; the key is that perturbations in A_1 need to be made small. This contrasts with the linearizing the original dynamical system (68), for which we would merely have required that the α_i's should be small. If you recall that $A_1 = |nu|^{1/2} alpha_1$, you will see that to linearize (75) implies that you have to make the perturbation in α_1 extremely small, much smaller than $|\nu|^{1/2}$ where $\nu = \mu - \mu_0$ measures how far you are from the bifurcation point.

The distinction here is this: when the trivial soltution is unstable ($Sa > 0$), do the other two steady state solutions exist ($Sa/b > 0$, so $b > 0$) or not ($Sa/b < 0$, so $b < 0$).

In the first case, when $b > 0$, as the trivial steady state becomes unstable at the bifurcation, two non-trivial steady state solutions with small amplitudes

$$\alpha_1 = |\nu|^{1/2} A_1 = \pm\sqrt{\frac{S|\nu|a}{b}} = \sqrt{\frac{\nu a}{b}}$$

appear nearby (recall that $S = \nu/|\nu|$), both of them growing in amplitude as the square root of the difference between the parameter μ and its critical value μ_0, $|\nu|^{1/2} = |\mu - \mu_0|^{1/2}$; the three solutions 'meet' at $\alpha_1 = 0$ when ν goes to zero. This case is called a supercritical pitchfork bifurcation*, and is what occurs in the convection problem (see exercise 13 below). This is why in figure 8, we can see a gradual onset in the strength of convection.*

The opposite case $b < 0$ is that in which a trivial stable steady state solution coexists with a pair of non-trivial but unstable steady state solutions, both of which disappear as the parameter μ passes the critical value μ_0; in that case, the type of analysis we have developed in this note has little to say about what happens to the evolution of the solution away from the trivial steady state solution once it becomes unstable.

Exercise 11 *Consider again the system* (65) *in exercise 10. Go through the steps in example 8 to show that $\nu = \mu$, $\mathbf{E} = \mathbf{I}$, the $2-by-2$ identity matrix, $\mathbf{J}$ is diagonal, and hence $\boldsymbol{\alpha} = \mathbf{y}$, $\mathbf{G} = \mathbf{F}$, $\tilde{\mathbf{J}} = \mathbf{J}$. Compute the components of H_{ijk} and K_{ijkl}, most of which will be zero (so list only the non-zero ones). Show that the recipe for computing the Landau equation in note 8 leads to the same coefficients a and b as you found in equation 64.*

Exercise 12 *Consider the dynamical system*

$$\begin{aligned}\frac{\mathrm{d}y_1}{\mathrm{d}t} &= F_1(y_1, y_2; \mu) = \mu y_1 - y_1^3 + y_1(y_2 - y_1)\\ \frac{\mathrm{d}y_2}{\mathrm{d}t} &= F_2(y_1, y_2; \mu) = \mu - y_2 + y_1 + c y_1^2,\end{aligned}$$

which differs somewhat from that in exercise 10. Show that you still get a pitchfork bifurcation at μ, and the same Landau equation as in exercise 10.

Exercise 13 *Take* (54a) *combined with* (52)*, and recall that $k_{xm} = 2m\pi/a$, $k_{zn} = n\pi$. Let $a = 2$, the wavelength of the fastest growing mode at $Ra = Ra_c = 4\pi^2$. Now change Ra slightly past Ra_c, defining*

$$\nu = Ra - Ra_c.$$

For sufficiently small positive ν, the dispersion relation (32) *should make it clear that only the modes $T_{1,1}$ and $T_{-1,1}$ grow in a linearized version of the model. As per the standard symmetry of Fourier coefficients of real-valued functions, recall also that $T_{1,1}$ and $T_{-1,1}$ are not independent, but must be related through*

$$T_{1,1} = \overline{T_{-1,1}},$$

and in fact

$$T_{mn} = \overline{T_{-mn}}$$

for all m.

We can construct an approximate version of the model for such small ν by rescaling the T_{mn} coefficients as follows

$$T_{1,1} = |\nu|^{1/2}\Theta_{1,1}, \qquad T_{-1,1} = |\nu|^{1/2}\Theta_{-1,1}, \qquad T_{mn} = |\nu|\Theta_{mn} \qquad \text{if } |m| \neq 1 \text{ or } n \neq 1,$$

$$t = |\nu|^{-1}\tau \tag{78}$$

If you have already read note 8, the choice of scaling here may already make sense to you; if not, simply take it at face value and persevere: you will find it leads to a sensible approximate model below.

Because $T_{1,1}$ and $T_{-1,1}$ are much larger than the remaining T_{mn}, it is important to distinguish terms containing $T_{1,1}$ and $T_{-1,1}$ in (54a) *from terms containing only other Fourier coefficients T_{mn}. The only difficulty in doing so is to deal with the nonlinear terms $\mathrm{conv}_{mn}(k_z r T_{qr}, T_{qr})$, $\mathrm{conv}_{mn}(k_{zq}p_{qr}, k_{zq}T_{qr})$ and $\mathrm{conv}_{mn}(k_{zr}T_{qr}, k_{zr}p_{qr})$. Take the formula* (55) *and transform to a new set of coefficients c_{mn} and d_{mn} defined through*

$$C_{1,1} = |\nu|^{1/2}c_{1,1}, \qquad C_{-1,1} = |\nu|^{1/2}c_{-1,1},$$

$$C_{mn} = |\nu|c_{mn} \qquad \text{if either } |m| \neq 1 \text{ or } n \neq 1.$$

Ditto for components of D_{mn}, so $D_{1,1} = |\nu|^{1/2}d_{1,1}$, $D_{-1,1} = |\nu|^{1/2}d_{-1,1}$, $D_{mn} = |\nu|d_{mn}$ if $|m| \neq 1$ or $n \neq 1$. The scaling here reflects the scaling for the analogously-indexed coefficients T_{mn} in (78)*: for instance, if C_{mn} is $k_{zn}p_{mn}$ and p_{mn} is linked to T_{mn} through* (52)*, then $c_{mn} = -k_{zn}^2(k_{xm}^2 + k_{zn}^2)^{-1}\Theta_{mn}$, and analogously for the other arguments of the* conv_{mn} *function in* (54a)*.*

Show that if $m = 1$, $n = 1$,

$$\begin{aligned}\mathrm{conv}_{1,1}(C_{qr}, D_{qr}) =& \frac{1}{2}\sum_{q=-\infty}^{\infty}\left(\sum_{r=1}^{\infty} C_{qr}D_{1-q,1+r} - \sum_{r=2}^{\infty} C_{qr}D_{1-q,r-1}\right)\\ =& \frac{|\nu|^{3/2}}{2}\left(c_{-1,1}d_{2,2} + c_{1,1}d_{0,2} - c_{2,2}d_{-1,1} - c_{0,2}d_{1,1}\right)\\ &+ \frac{|\nu|^2}{2}\left(\sum_{(q,r)\in I_1} c_{qr}d_{1-q,1+r} - \sum_{(q,r)\in I_2} c_{qr}d_{1-q,r-1}\right)\end{aligned} \tag{79}$$

where it is easier to write the sums in the last expression in terms of index sets I_1 *and* I_2*. For* I_1*, this is the set of* (q, r) *for which* q *is an integer (positive, negative, or zero) and* r *is a natural number (a positive integer), but we miss out the combinations* $(-1, 1)$ *and* $(1, 1)$*. Likewise,* I_2 *is the set of* (q, r) *for which* q *is an integer,* r *is an integer greater than 1, and we miss out the combinations* $(0, 2)$ *and* $(2, 2)$*.*[21] The point here is that we are trying to keep track of the powers of $|\nu|$, eventually retaining only the lowest powers in a simplified model: *what the formula* (80) *shows is that a special role falls to the combination of terms* $(c_{-1,1}d_{2,2} + c_{1,1}d_{0,2} - c_{2,2}d_{-1,1} - c_{0,2}d_{1,1})$*, since they are much larger than the others. In fact, we can simply write*

$$\mathrm{conv}_{1,1}(C_{qr}, D_{qr}) = \frac{|\nu|^{3/2}}{2}\left(c_{-1,1}d_{2,2} + c_{1,1}d_{0,2} - c_{2,2}d_{-1,1} - c_{0,2}d_{1,1}\right) + O(|\nu|^2). \quad (80)$$

Similarly for $m = -1$*,* $n = 1$*, show that*

$$\mathrm{conv}_{-1,1}(C_{qr}, D_{qr}) = \frac{|\nu|^{3/2}}{2}\left(c_{1,1}d_{-2,2} + c_{-1,1}d_{0,2} - c_{-2,2}d_{1,1} - c_{0,2}d_{-1,1}\right) + O(|\nu|^2), \quad (81)$$

while for other values of m *and* n*, show that we have one special case* $m = 0$*,* $n = 2$*, for which*

$$\mathrm{conv}_{0,2}(C_{qr}, D_{qr}) = \frac{|\nu|}{2}\left(c_{1,1}d_{-1,1} + c_{-1,1}d_{1,1}\right) + O(|\nu|^{3/2}). \quad (82)$$

For other combinations of m *and* n*, show that*

$$\mathrm{conv}_{mn}(C_{qr}, D_{qr}) = O(|\nu|^{3/2}). \quad (83)$$

With this knowledge, you can re-write (54a)*, substituting for* p_{mn} *using* (52) *and subsequently transforming to the new variables* Θ_{mn} *and* τ *using their definition in* (78)*. Again only retain the lowest powers of* ν *that multiply terms not identically equal to zero, and expand* $1/Ra$ *as*

$$Ra^{-1} = (Ra_c + \nu)^{-1} = Ra_c^{-1} - \nu Ra_c^{-2} + O(\nu^2).$$

Show the following:

[21] Formally, we can write $I_1 = (\mathbb{Z}\times\mathbb{N})\backslash(\{(-1,1)\}\cup\{(1,1)\})$, $I_2 = (\mathbb{Z}\times(\mathbb{N}\backslash\{1\}))\backslash(\{(0,2)\}\cup\{(2,2)\})$. To demistify this a bit, what summing over $(q, r) \in I_1$ means is that the sum is over all distinct combinations of (q, r) in which q is an integer (positive or negative, meaning $q \in \mathbb{Z}$ where $\mathbb{Z}$ is the set of all integers) and r is a natural number (a positive integer, meaning $q \in \mathbb{N}$, where $\mathbb{N}$ is the set of natural numbers), but missing out (that is the meaning of the set exclusion symbol '\') both $(-1, 1)$ and $(1, 1)$.

1. *For $m = n = 1$,*

$$
\begin{aligned}
&|\nu|^{1/2}\left(-\frac{k_{x,1}^2}{k_{x,1}^2+k_{z,1}^2}\Theta_{1,1}+Ra_c^{-1}(k_{x,1}^2+k_{z,1}^2)\Theta_{1,1}\right)\\
&+|\nu|^{3/2}\Bigg[\frac{\mathrm{d}\Theta_{1,1}}{\mathrm{d}\tau}+\frac{1}{2}\left(k_{z,1}\Theta_{-1,1}\Theta_{2,2}+k_{z,1}\Theta_{1,1}\Theta_{0,2}-k_{z,2}\Theta_{2,2}\Theta_{-1,1}-k_{z,2}\Theta_{0,2}\Theta_{1,1}\right)\\
&\quad-\frac{1}{2}\Bigg(\frac{k_{x,-1}k_{z,1}k_{x,2}}{k_{x,-1}^2+k_{z,1}^2}\Theta_{-1,1}\Theta_{2,2}+\frac{k_{x,1}k_{z,1}k_{x,0}}{k_{x,1}^2+k_{z,1}^2}\Theta_{1,1}\Theta_{0,2}\\
&\quad-\frac{k_{z,2}k_{x,2}k_{x,-1}}{k_{x,2}^2+k_{z,2}^2}\Theta_{2,2}\Theta_{-1,1}-\frac{k_{x,0}k_{z,2}k_{x,1}}{k_{x,0}^2+k_{z,2}^2}\Theta_{0,2}\Theta_{1,1}\Bigg)\\
&\quad-\frac{1}{2}\Bigg(\frac{k_{z,1}k_{z,2}^2}{k_{x,2}^2+k_{z,2}^2}\Theta_{-1,1}\Theta_{2,2}+\frac{k_{z,1}k_{z,2}^2}{k_{x,0}^2+k_{z,2}^2}\Theta_{1,1}\Theta_{0,2}\\
&-\frac{k_{z,2}k_{z,1}^2}{k_{x,-1}^2+k_{z,1}^2}\Theta_{2,2}\Theta_{-1,1}-\frac{k_{z,2}k_{z,1}^2}{k_{x,1}^2+k_{z,1}^2}\Theta_{0,2}\Theta_{1,1}\Bigg)-SRa_c^{-2}\left(k_{x,1}^2+k_{z,1}^2\right)\Theta_{1,1}\Bigg]+O(\nu^2)=0.
\end{aligned}
\tag{84}
$$

where $S = \mathrm{sgn}(\nu) = \nu/|\nu|$.

2. *For $m = -1$, $n = 1$, show that we get the analogous*

$$
\begin{aligned}
&|\nu|^{1/2}\left(-\frac{k_{x,-1}^2}{k_{x,-1}^2+k_{z1}^2}\Theta_{-1,1}+Ra_c^{-1}(k_{x1}^2+k_{z1}^2)\Theta_{-1,1}\right)\\
&+|\nu|^{3/2}\Bigg[\frac{\mathrm{d}\Theta_{-1,1}}{\mathrm{d}\tau}+\frac{1}{2}\left(k_{z,1}\Theta_{1,1}\Theta_{2,2}+k_{z,1}\Theta_{-1,1}\Theta_{0,2}-k_{z,2}\Theta_{2,2}\Theta_{1,1}-k_{z,2}\Theta_{0,2}\Theta_{-1,1}\right)\\
&\quad-\frac{1}{2}\Bigg(\frac{k_{x,1}k_{z,1}k_{x,-2}}{k_{x,1}^2+k_{z,1}^2}\Theta_{1,1}\Theta_{-2,2}+\frac{k_{x,-1}k_{z,1}k_{x,0}}{k_{x,-1}^2+k_{z,1}^2}\Theta_{-1,1}\Theta_{0,2}\\
&\quad-\frac{k_{z,-2}k_{x,2}k_{x,1}}{k_{x,2}^2+k_{z,2}^2}\Theta_{-2,2}\Theta_{1,1}-\frac{k_{x,0}k_{z,2}k_{x,-1}}{k_{x,0}^2+k_{z,2}^2}\Theta_{0,2}\Theta_{-1,1}\Bigg)\\
&\quad-\frac{1}{2}\Bigg(\frac{k_{z,1}k_{z,2}^2}{k_{x,-2}^2+k_{z,2}^2}\Theta_{1,1}\Theta_{-2,2}+\frac{k_{z,1}k_{z,2}^2}{k_{x,0}^2+k_{z,2}^2}\Theta_{-1,1}\Theta_{0,2}\\
&-\frac{k_{z,2}k_{z,1}^2}{k_{x,1}^2+k_{z,1}^2}\Theta_{-2,2}\Theta_{1,1}-\frac{k_{z,2}k_{z,1}^2}{k_{x,-1}^2+k_{z,1}^2}\Theta_{0,2}\Theta_{-1,1}\Bigg)-SRa_c^{-2}\left(k_{x,-1}^2+k_{z,1}^2\right)\Theta_{-1,1}\Bigg]+O(\nu^2)=0.
\end{aligned}
\tag{85}
$$

3. For $m = 0$, $n = 2$, show that

$$|\nu|\Bigg[\frac{k_{x,0}^2}{k_{x,0}^2+k_{z,2}^2}\Theta_{0,2} + Ra_c^{-1}\left(k_{x,0}^2+k_{z,2}^2\right)\Theta_{0,2} + k_{z1}\Theta_{1,1}\Theta_{-1,1} - \frac{1}{2}\left(\frac{k_{x,-1}k_{z,1}k_{x,1}}{k_{x,-1}^2+k_{z,1}^2} + \frac{k_{x,1}k_{z,1}k_{x,-1}}{k_{x,1}^2+k_{z,1}^2}\right)\Theta_{1,1}\Theta_{-1,1} - \frac{1}{2}\left(\frac{k_{z,1}^3}{k_{x,-1}^2+k_{z,1}^2} + \frac{k_{z,1}^3}{k_{x,1}^2+k_{z,1}^2}\right)\Theta_{1,1}\Theta_{-1,1}\Bigg] + O(|\nu|^{3/2}) = 0 \tag{86}$$

4. For every other combination of m and n, show that

$$|\nu|\left[\frac{k_{xm}^2}{k_{xm}^2+k_{zn}^2}\Theta_{mn} + Ra_c^{-1}\left(k_{xm}^2+k_{zn}^2\right)\Theta_{mn}\right] + O(|\nu|^{3/2}) = 0. \tag{87}$$

Again, time to take stock — we now have somewhat simplified nonlinear equations for the different Fourier components Θ_{mn}. "Simplified" primarily means having converted the convolution sums (which are in principle infinite sums) to sums over only a handful of terms involving products of other Fourier coefficients. In particular, we see that the evolution of $\Theta_{1,1}$ and $\Theta_{-1,1}$ depends only on $\Theta_{1,1}$, $\Theta_{-1,1}$, $\Theta_{0,2}$, and on $\Theta_{2,2}$ and $\Theta_{-2,2}$ respectively, at 'leading order'. Likewise, the evolution of $\Theta_{0,2}$ depends only on $\Theta_{0,2}$, $\Theta_{1,1}$ and $\Theta_{-1,1}$, while we have eliminated dependence on anything but Θ_{mn} itself in (87).

In fact, we can tidy this up further. The first thing to notice is that we automatically have

$$-\frac{k_{x,1}^2}{k_{x,1}^2+k_{z,1}^2}\Theta_{1,1} + Ra_c^{-1}(k_{x,1}^2+k_{z,1}^2)\Theta_{1,1} = 0, \tag{88}$$

and likewise

$$-\frac{k_{x,-1}^2}{k_{x,-1}^2+k_{z1}^2}\Theta_{-1,1} + Ra_c^{-1}(k_{x1}^2+k_{z1}^2)\Theta_{-1,1} = 0. \tag{89}$$

These equations hold identically (regardless of the choice of $\Theta_{1,1}$ and $\Theta_{-1,1}$ because

$$-\frac{k_{x,1}^2}{k_{x,1}^2+k_{z,1}^2} + Ra_c^{-1}(k_{x,1}^2+k_{z,1}^2) = -\frac{k_{x,-1}^2}{k_{x,-1}^2+k_{z,1}^2} + Ra_c^{-1}(k_{x,-1}^2+k_{z,1}^2) = 0.$$

That is effectively the definition *of Ra_c: recall that the dispersion relation* (32) *is*

$$\sigma(k_{xm}, k_{zn}) = \frac{k_{xm}^2}{k_{xm}^2+k_{zn}^2} - \frac{1}{Ra}\left(k_{xm}^2+k_{zn}^2\right)$$

and the critical Rayleigh number is that for which the largest $\sigma(k_{xm}, k_{zn})$ (which here occurs at $|m| = n = 1$) is zero. With (88) *and* (89)*, the first, $O(|\nu|^{1/2})$ expressions in*

round brackets in (84) *and* (85) *vanish, and we are left with the next* $O(|\nu|^{3/2})$ *term as the leading term in the approximation in small* $|\nu|$*. This is no accident, but the very basis of the construction of the solution near the critical parameter value* Ra_c*: the linear approximation in* $\Theta_{1,1}$ *and* $\Theta_{-1,1}$ *vanishes at the bifurcation and we have to go to higher order.*

The second insight we can immediately glean from (87) *(which holds for* $(m,n) \neq (-1,1)$, $(1,1)$, $(0,2)$*), is that* $\Theta_{mn} = 0$ *if we exclude the* $O(|\nu|^{3/2})$ *correction term in* (87)*; what this really means is that these* Θ_{mn} *are small, of* $O(|\nu|^{1/2})$*, and the corresponding scaling in* (78) *should have been chosen accordingly. That is of no further consequence to us, however, as the knowledge that these* Θ_{mn} *are small allows us to drop the coupling with* $\Theta_{2,2}$ *in* (84) *and* (85)*. Combined with* (88) *and* (89)*, show that this leads to*

$$\frac{\mathrm{d}\Theta_{1,1}}{\mathrm{d}\tau} - SRa_c^{-2}\left(k_{x,1}^2 + k_{z,1}^2\right)\Theta_{1,1} - \frac{1}{2}\left(\frac{k_{x,1}^2 k_{z,2}}{k_{x,1}^2 + k_{z,1}^2}\right)\Theta_{0,2}\Theta_{1,1} + O(\nu^{1/2}) = 0. \qquad (90)$$

where you need to make use of the fact that $k_{x,0} = 0$*. Recognizing that* $k_{x,-1} = -k_{x,1}$ *show that the same equation holds for* $m = -1$*,* $n = 1$ *if we replace* $\Theta_{1,1}$ *by* $\Theta_{-1,1}$ *in* (90)*.*

This leaves us $\Theta_{0,2}$ *to deal with. Again using* $k_{x,0} = 0$ *and* $k_{x,-1} = -k_{x,1}$ *again in* (86)*, show that*

$$Ra_c^{-1}k_{z,2}^2\Theta_{0,2} + \frac{2k_{x,1}^2 k_{z,1}}{k_{x,1}^2 + k_{z,1}^2}\Theta_{1,1}\Theta_{-1,1} + O(|\nu|^{1/2}) = 0,$$

or

$$\Theta_{0,2} = -Ra_c\frac{2k_{x,1}^2 k_{z,1}}{k_{z,2}^2(k_{x,1}^2 + k_{z,1}^2)}\Theta_{1,1}\Theta_{-1,1} + O(|\nu|^{1/2}).$$

Recall that $\Theta_{-1,1} = \overline{\Theta_{1,1}}$ *and therefore*

$$\Theta_{1,1}\Theta_{-1,1} = |\Theta_{1,1}|^2 = |\Theta_{-1,1}|^2$$

Substituting in (90) *and omitting the 'reminder of the size of the terms we have omitted, show that*

$$\frac{\mathrm{d}\Theta_{1,1}}{\mathrm{d}\tau} = SRa_c^{-2}\left(k_{x,1}^2 + k_{z,1}^2\right)\Theta_{1,1} - Ra_c\frac{k_{x,1}^4 k_{z,1}}{k_{z,2}(k_{x,1}^2 + k_{z,1}^2)^2}|\Theta_{1,1}|^2\Theta_{1,1} \qquad (91)$$

This is the analogue of the Landau equation (75)*, in slightly more complicated form here because* $\Theta_{1,1}$ *can in general be complex:*

$$\frac{\mathrm{d}\Theta_{1,1}}{\mathrm{d}\tau} = Sa\Theta_{1,1} - b|\Theta_{1,1}|^2\Theta_{1,1}$$

with positive a and b. Show that for $\Theta_{-1,1} = \overline{\Theta_{1,1}}$, we obtain the same equation

$$\frac{\mathrm{d}\overline{\Theta_{1,1}}}{\mathrm{d}\tau} = Sa\overline{\Theta_{1,1}} - b|\Theta_{1,1}|^2\overline{\Theta_{1,1}}.$$

Show therefore that

$$\frac{\mathrm{d}|\Theta_{1,1}|^2}{\mathrm{d}\tau} = 2Sa|\Theta_{1,1}|^2 - 2b|\Theta_{1,1}|^4,$$

and that there is unique, stable steady state $|\Theta_{1,1}|^2 = 0$ if $S < 0$. For $S > 0$, show that there are two steady states $|\Theta_{1,1}|^2 = 0$ and $|\Theta_{1,1}|^2 = a/b$, the former being unstable and the latter stable.

Using these results, show using (57) *that for $Ra > Ra_c$*

$$||T'|| \approx \sqrt{\frac{a(k_{x,1}^2 + k_{z,1}^2)^3 k_{z,2}}{Ra_c^3 k_{x,1}^4 k_{z,1}}(Ra - Ra_c)}, \tag{92}$$

where we have reverted to "a" as denoting the periodicity of the domain.[22] *Hence $||T'||$ grows as the square root of $Ra - Ra_c$. Similarly, show using* (63) *that the Nusselt number grows linearly in $Ra - Ra_c$, as*

$$Nu(Ra) \approx 1 + \frac{2(k_{x,1}^2 + k_{z,1}^2)^2}{Ra_c^2 k_{x,1}^2} \tag{93}$$

[22] As before, given the limited number of letters available, it is not uncommon for variable names to be context-dependent; we used a to denote a coefficient in the Landau equation when this was unambiguous, but now revert to it denoting periodicity.