The Gravitational Instability and its Role in the Evolution of Protoplanetary and Protostellar Discs

Peter John Cossins
MMath (Exon) AMIMA ARAeS

Submitted in fulfillment of the requirements for the degree of
Doctor of Philosophy

July 2010

Theoretical Astrophysics Group
Department of Physics and Astronomy
University of Leicester

Principal Supervisor: Prof. Giuseppe Lodato
Università Degli Studi di Milano, Italy

Abstract

The Gravitational Instability and its Role in the Evolution of Protoplanetary and Protostellar Discs

Peter J Cossins MMath (Exon)

In this thesis I present numerical simulations of massive, cold, non-ionised self-gravitating accretion discs about a central massive object, and then use them to investigate structure formation and energy/angular momentum transport, the effects of different cooling regimes on the likelihood of bound condensates forming through direct gravitational fragmentation, and the potential for resolved sub-mm imaging of such systems. I also present a review of current theories of viscous and wave transport in astrophysical discs, observed properties of protostellar and protoplanetary discs and a numerical scheme suitable for conducting computational experiments on fluid discs.

I find that the structures excited in self-gravitating fluid discs self-regulate in such a manner that the density waves formed are very weak shocks, with the amplitude of the density perturbations forming the waves determined by the cooling regime. This self-regulation process ensures that for discs of $\lesssim 10\%$ of the central object mass the transport properties are determined principally by *local* effects, representing a crucial difference between collisional (fluid) and collisionless (stellar) discs as the latter cannot form shocks.

I further find that the effects of an opacity-based cooling function makes self-gravitating protoplanetary discs significantly more susceptible to fragment formation in certain opacity regimes at relatively high ($10^{-5} - 10^{-3}\ M_{\odot}\ \mathrm{yr}^{-1}$) accretion rates due to the dependence on temperature perturbations. Furthermore I find that fragment formation due to direct gravitational collapse is feasible in such discs only at radii $\gtrsim 50$ AU, and this radius increases with decreasing temperature if the background temperature falls below approximately 10K.

Finally I have used simple disc models in conjuction with a realistic telescope model to demonstrate that resolved images of spiral structure in massive, self-gravitating protostellar discs should be readily observable with ALMA, out to distances representative of local star-forming complexes.

Dedicated to my brother,

Thomas Michael Cossins (1983)

and to all those whose help and encouragement has allowed me to reach this point.

Acknowledgements

In grateful recognition of everyone whose influence has made me who I am today, and all those people who have helped me to reach this long-standing goal of (I hope!) indulging myself in a slightly more exclusive appellation than Mr!

In no particular order, I should like to thank my Mathematics teachers at King Edmund (Community) School, in particular Messrs. Lord and Pownall, the latter especially for his 'touch of the verbals' which scared everyone witless at the time but hammered home the mental arithmetic! From the peerless and life-changing University of Exeter I commend to history the efforts of my tutor, Dr. Ben Mestel, my Dissertation tutor, Dr. (now Prof.) Andrew Gilbert, and in particular Mr. Milo Dixon, whose General Relativity courses stoked the flames of an enduring interest in the minutiae of Theoretical Astrophysics. Geeky it undoubtedly is, but I still remember his derivation of $E = mc^2$ on the blackboard sending shivers down my spine.

Rolls-Royce deserve a special mention for a) biting the bullet and employing me as a young and idealistic graduate, b) teaching me a few hard but necessary lessons about life, c) bringing out my inner pragmatist (which has since proven invaluable), and d) providing me with an excellent grounding in the practicalities of fluid dynamics. That and for letting me loose on test with several million pounds-worth of engine, which was great fun! Thanks go in particular to Ken Hawkins, Colin Marshall, Steve Moore, Kevin Menzies, Jo Dorca-Luque, John Szeki, Nigel Cox, Andy Wood, Gary Way and Richard Dingley. Also to the Ops boys, Bojan Soldat, James "Jimmy" Bossard, Ben Mycock, Mark Campbell, Simon Cooper, Dave Wray, Jim Hardy, Dean Morgan and Ian Makin for sundry Christmas bashes and Glastonbury trips!

At Leicester University, I should like to thank the Theoretical Astrophysics Group, in particular my principal supervisor Giuseppe Lodato (now Associate Professor at the Universitá Degli Studi di Milano) without whom I should have been lost in the cosmic wilderness. Also to my current supervisor Sergei Nayakshin, Graham Wynn, for his in-depth knowledge of local watering holes, and Walter Dehnen for his assistance with recalcitrant code problems!

To the denizens of the Batcave I owe special thanks, having enlivened my life immeasurably over the last three-odd years, so thank you Patrick Deegan (now Awesome Dr. Deeg), Jalpesh Sachania (for sorting my Battlestar Galactica fix, and an excellent style file), Céline Combet (for her excellent skiing tips – "It's intuitive"!), fellow Batcoder Chris Nixon, Alex Hobbs, Lee Cullen, Fergus Wilson, Dave Cole,

Fabrizio Pedes, Kastytis Zubovas, Chris Power, Andreas Koch and Seung-Hoon Cha. Finally, as the human face of the department and worker of administrative miracles *extraordinaire*, many thanks must go to Lisa Brant for her unstinting efforts in the face of bureaucracy, booking of flights and exotic hotels at short notice and general maternal interest in our well-being – there will be Rosé coming your way once this is all over!

Likewise a special mention must go to the XROA rabble, including the lovely Lucy Heil (for, among many other things, doing all my worrying for me), "Little" Rich Owen, Amy Scott and the "Death by Teapot" collective. Also to James "Dukey" Duke for his enlightening tea-time conversations, and in particular to Matt Burleigh and Dave Baker, for a fantastic, once-in-a-lifetime trip to La Palma to go observing.

To my parents John and Margaret I owe an un-repayable debt of gratitude and love, and whose constant support and ever-present offer of a bolt hole at home has been immensely appreciated! Also to my brother Thomas, to whom I dedicate this my first book (albeit with a slightly limited readership...), and to my extended and ever-encouraging family; Jennifer, whose continuous enthusiasm for my research and insistence that the Universe is all part of God's Plan has made for many an interesting discussion, Richard, Tracey, Dan, Vicky, Emma, Wayne and their children (who make me feel old already watching them grow up so fast), Lizzie, Emily, Sophie, Lucy and Sam. To my uncle Colin I owe more than just a debt of gratitude, and I should like to take this opportunity to thank him for all his assistance over the years in all its forms. And how could I forget my Number One Aunt Annie, whose door is always open, and whose friends are specially primed to ask probing questions about my love life!

Finally to friends past and present, who have made it worthwhile both staying on the straight and narrow, and deviating from it when the opportunity arises. I thank the incomparable Mr. Robert Banks for his humour, derogatory remarks about my "essay" and endless supply of 'Robisms' – keep on Keepin it Gangsta Banks. Also thanks to Mimi Brousse de Gersigny (now finally Prideaux-Brune!) for being there without question and knowing me better than I know myself, Claire Craig for being wonderful for the last twenty odd years, and also Simon Grounds, Captain Al Reid, Stuart Valentine and many others too numerous to do justice to. Please forgive me for not mentioning you all by name, but know that my thanks go out to you all. Last but not least, a special mention must go to Cathrine Coutts *née* Eady, without whose encouragement, persuasion and belief in my ability I would never have got off my arse and applied for a PhD in the first place, and who is therefore to be credited in no small way with getting me to where I am today.

Contents

List of Figures

List of Tables

1

Theoretical Background

It is a very sad thing that nowadays there is so little useless information.

Oscar Wilde

1.1 Introduction

The formation of a disc of material about a central object is a common theme throughout the canon of astrophysics, occurring from the relatively small circumplanetary scale of Saturn's rings ($\sim 10^5$ km), through protostellar and protoplanetary discs on scales of hundreds of AU, up to the parsec scale of discs around Active Galactic Nuclei (AGN) and galactic discs on scales of several tens of kiloparsecs. The reason for this ubiquity is simple – infalling material almost always contains some angular momentum with respect to the central mass, and thus cannot fall directly on to it. However, unlike angular momentum, energy can be radiated away, and thus the material reduces to its lowest energy state – a (usually thin) circular disc around the central object.

Although the angular momentum cannot be removed, various processes exist that can redistribute it within any given system. Furthermore, even a vanishingly small amount of mass, if transported out to large radii can carry away much of the disc's angular momentum, allowing the remainder to orbit more closely to the central mass. Although energy is required to move the small amount of mass to large radii, a much greater amount of gravitational potential energy is liberated by the infall of the remaining material. Hence the accretion of mass on to the central object is generally speaking energetically favourable (Lynden-Bell & Pringle, 1974; Binney & Tremaine, 2008).

The omnipresence of discs, both actively accreting and otherwise, has led to them being implicated in a wide variety of astrophysical phenomena. At the largest scales, the majestic sweep of spiral arms belie the presence of gaseous and stellar discs in galaxies, as shown in Fig. 1.1. At the next scale down, the fuelling of AGN is widely expected to be due to an accretion disc/torus about the central supermassive black hole (SMBH) (Frank et al., 2002; Yu & Tremaine, 2002; Antonucci, 1993; Shlosman et al., 1990). Similarly, the observed discs of stars about Sgr A* (and indeed other galactic nuclei, see Vollmer et al. 2008; Shlosman & Begelman 1989) probably formed due to the fragmentation of such a disc that became gravitationally unstable (Hobbs & Nayakshin, 2009; Nayakshin et al., 2007), thereby preventing the SMBH at the centre of our own galaxy from being active during the current epoch.

On a smaller scale, the outbursts of various Cataclysmic Variable (CV) systems have been discussed in terms of accretion discs. Dwarf novae (DNe) and soft X-ray transients (SXTs) are two classes of low mass X-ray binaries (LMXBs) (themselves a subset of CV systems) where an explanation has been put forward in terms of

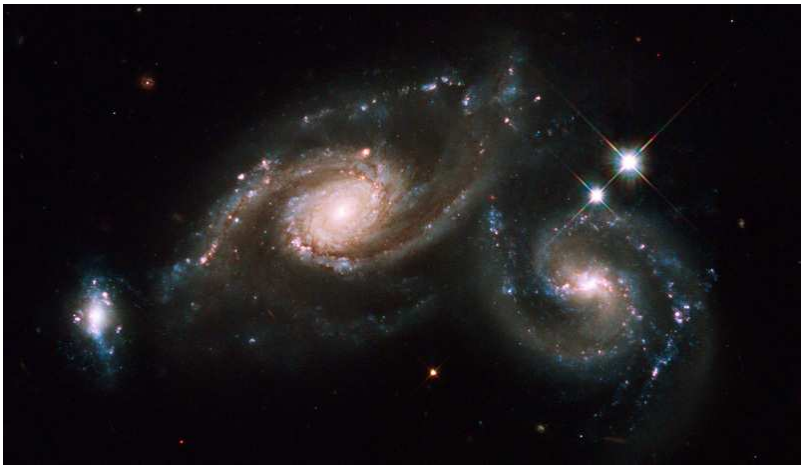

FIGURE 1.1: *Optical image of the Arp 274 triplet, showing the presence of spiral arms within the discs of the two major galaxies, which are not thought to be interacting – their apparent proximity simply being a chance alignment on the sky.*
Image credit: NASA, ESA, M. Livio and the Hubble Heritage Team (STScI/AURA)

---

the so-called thermal-viscous disc instability model (DIM) (Lewin & van der Klis, 2006; Frank et al., 2002; Lewin et al., 1997). Here a disc about a compact object (a white dwarf in the case of DNe and a neutron star or black hole in the case of SXTs) is fed from the secondary star, either via stellar winds or through Roche lobe overflow, at a rate which it is inherently unable to pass to the primary through steady state accretion. In the quiescent stage the disc accretes some mass on to the primary, but the remainder builds up until the disc becomes thermally unstable, and runaway heating occurs. This brings with it an associated increase in the disc viscosity and mass is accreted rapidly on to the primary, leading to the observed outbursts (Lewin & van der Klis, 2006). Although this is not the complete picture (see for instance Lasota 2008; Hameury & Lasota 2005; Lasota 2001), what remains clear is that the disc plays a crucial role in producing the observed effects. A cartoon of the generic disc-star system found in LMXBs is shown in Fig. 1.2.

Star formation is another area where discs play an important role. As cold gas begins to collapse into pre-stellar cores within molecular clouds, any small rotations become magnified by the collapse and result in the formation of a disc about the protostar. In the early stages of star formation the disc will grow in mass as in-falling gas from the protostellar envelope accretes on to it, and as such is likely to go through a self-gravitating phase (Bertin & Lodato, 2001a; Vorobyov & Basu, 2005; Hartmann, 2009a). In this phase the disc may undergo FU Orionis-type

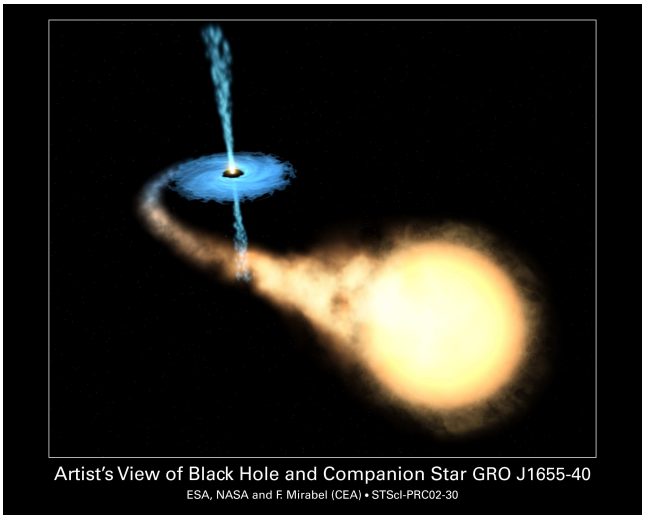

FIGURE 1.2: *An artist's impression of the LMXB GRO J1655-40, in which a few solar mass black hole accretes material from an evolved F type star through an accretion disc. In this case the disc emits a bi-polar jet, visible at radio wavelengths, and is therefore classified as a microquasar.*

outbursts as material is transported inward through the gravitational instability (Zhu et al., 2009), although it should be noted that this is not the only explanation for this phenomenon – envelope infall (Kenyon & Hartmann, 1991) and the disc instability model (Lodato & Clarke, 2004) have both been put forward as alternatives. Furthermore it is possible (though by no means certain) that companion Brown Dwarfs (BDs) and massive gas giant planets may form at this stage through gravitational fragmentation of the disc (Stamatellos & Whitworth, 2009a, 2008; Stamatellos et al., 2007a; Boss, 1997, 1998).

Finally at the smallest scales, the rings of Saturn are yet another example of disc formation, albeit a annular one. Exceptionally thin (the average thickness is thought to be $\lesssim 1$ km[1] although estimates vary), these consist primarily of ice particles rather than gas, and due to the huge success of the Cassini mission in gathering data about both the rings and their host planet, they are one of the better studied disc systems. Nonetheless, there is still considerable controversy regarding the lifetime of this particular 'disc'. Esposito (1986) and Griv & Gedalin (2006) have suggested that they may be as young as 100 million years, whereas others (Salmon et al., 2009; Daisaka et al., 2001) have suggested they may have a lifetime comparable to that of the solar system. It is however known that the E ring at least is being fed via outflows from cryovolcanism on Enceladus (Porco et al., 2006; Spahn et al., 2006)

[1]http://solarsystem.nasa.gov/planets/profile.cfm?Object=Saturn&Display=Rings

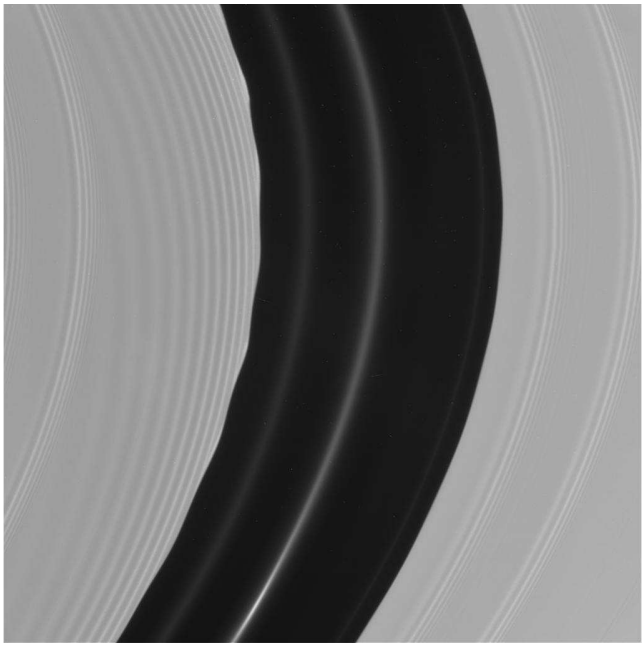

FIGURE 1.3: *This figure shows spiral density waves and gravitationally induced wakes in the Encke Gap of Saturn's A ring caused by the passage of the shepherding moonlet Pan, which orbits in the centre of the gap. This image was taken by the Cassini spacecraft, and is reproduced courtesy of NASA/JPL/Space Science Institute.*

---

and even on this scale the effects of gravity are notable through the sculpting of ring edges and the presence of spiral waves within the rings themselves, as illustrated in Figs. 1.3 and 1.4.

It is clear then that discs have a leading role to play in a wide range of astrophysical phenomena, and that similarly gravity is key to their formation and evolution. In this chapter I shall therefore give a more detailed introduction to a number of topics required to understand both discs and their interaction with gravity, especially as regards the evolution of gaseous, collisional systems. In Section 1.2 I derive and discuss the viscous thin disc approximation, and present general results for characteristic dimensions, timescales and accretion rates. In Sections 1.3 and 1.4 I shall present two of the instabilities to which discs are generally susceptible, namely the Magneto-Rotational Instability (MRI) and the Gravitational Instability (GI), the latter of which features heavily throughout this thesis. Hence I shall derive the dispersion relation for waves propagating in a self-gravitating thin disc, both for the standard axisymmetric case, and more generally. The application of theory to physical systems and observational verification I defer to the next chapter.

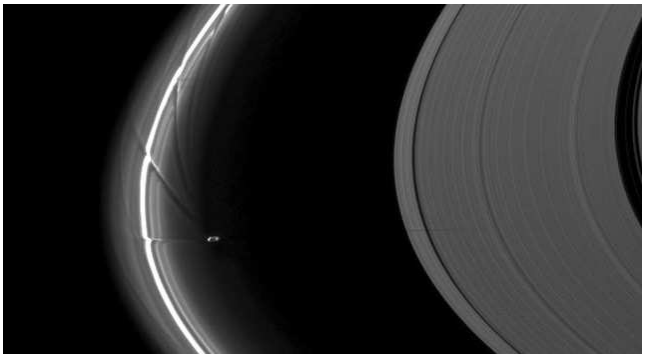

FIGURE 1.4: *This figure shows the presence of streamers in Saturn's F ring caused by gravitational interactions with the irregular moon Prometheus, seen just inside the ring. The shadow of the moon can be seen on the A ring to the right of the image. This image was taken by the Cassini spacecraft, and is reproduced courtesy of NASA/JPL/Space Science Institute.*

1.2 Disc Basics

In this section I shall introduce some of the concepts and derive some of the basic equations that govern the evolution of fluid discs about a single massive central object. Although it is intended to be reasonably comprehensive, further details and different approaches to the derivations may be found in Pringle (1981); Frank et al. (2002) and Lodato (2007) in particular.

At this early stage it is useful to introduce a few crucial assumptions, which together make the theoretical analysis of discs easier without significantly sacrificing generality. Firstly, we may reasonably assume that any discs we consider are thin, i.e. that their radial extent is much greater than their vertical extent. Although this may not be universally the case (discs around AGN for instance probably thicken and become more toroidal in shape throughout some of their radial range, Frank et al. 2002) it is an extremely useful and intuitively reasonable simplification, and one that is justified in the majority of cases. Introducing cylindrical polar co-ordinates (which I shall use throughout henceforth) with the central object at the origin, this is therefore formally equivalent to requiring that $H/R \ll 1$, where H is a representative scale height and R is the cylindrical radius.

Secondly, in order to describe an accretion disc, we require some mechanism to transport angular momentum and energy within the disc itself. As mentioned in the previous section, moving mass further into the potential liberates large quantities of gravitational potential energy, but doing so requires a process that converts this energy to another form. Viscous phenomena allow for the liberation of energy as

heat from ordered (rotational) motion, and are furthermore capable of transporting angular momentum through viscous torques, so it is reasonable to assume that there is some viscosity within the disc. At this stage however, I make no further assumptions about the *nature* of this viscosity, merely posit its existence.

Finally, for the time being it is useful to assume that our discs are axisymmetric, i.e. that all quantities are independent of the position angle θ, as this further simplifies matters. This assumption is however one I shall discard in due course, when considering the presence of spiral structures within the disc, which are clearly *not* axisymmetric.

Having introduced these assumptions, we can consider how they may be used to simplify the governing equations, which I shall introduce now.

1.2.1 Equations of Motion

Three equations are of particular importance to the evolution of astrophysical discs. Firstly, the continuity equation embodies conservation of mass, and is given by

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0, \tag{1.1}$$

where $\mathbf{v}$ is the fluid velocity, and ρ is the (volume) density. The Euler equation for inviscid flows encapsulates conservation of momentum, and is given by

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla)\mathbf{v} = -\frac{1}{\rho}\nabla P - \nabla \Phi, \tag{1.2}$$

where additionally we have the fluid pressure $P = \kappa \rho^{\gamma}$ (with κ as the adiabat and γ the adiabatic index), and gravitational potential Φ. This potential may be due to the central body alone, or in the case of a self-gravitating fluid, may be linked to the volume density through Poisson's equation,

$$\nabla^2 \Phi = 4\pi G \rho, \tag{1.3}$$

where as normal, G is the universal gravitation constant.

It is worth noting here that the Euler equation may be simplified somewhat by the introduction of the specific enthalpy h, linked to the specific internal energy

$u = \kappa\rho^{\gamma-1}/(\gamma-1)$ such that

$$h = u + \frac{P}{\rho} = \int \frac{1}{\rho}\,\mathrm{d}P = \frac{\kappa\gamma\rho^{\gamma-1}}{\gamma-1}, \tag{1.4}$$

where the central equality represents two equivalent definitions for the specific enthalpy. Using this, the Euler equation simplifies to become

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{v} = -\nabla(h+\Phi), \tag{1.5}$$

an expression that I shall call upon later in this chapter.

Furthermore, it should be noted that the Euler equation in either form is inadequate for describing the evolution of *viscous* discs. Introducing the viscosity to equation 1.2 we obtain the Navier-Stokes equation, such that

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{v} = -\frac{1}{\rho}\nabla P + \frac{1}{\rho}\nabla\cdot\sigma - \nabla\Phi, \tag{1.6}$$

where σ is an as yet undefined stress tensor of rank two.

1.2.1.1 Mass Conservation

Given the assumptions that discs are both thin and axisymmetric, we may take the vertical integral of the continuity equation (equation 1.1) and reduce the dependency of this equation from three positional variables (R, θ, z) to one (R). In this manner, equation 1.1 becomes

$$\int_{-\infty}^{\infty} \frac{\partial \rho}{\partial t} + \nabla\cdot(\rho\mathbf{v})\,\mathrm{d}z = c, \tag{1.7}$$

for some constant of integration c, and in the limit that the density goes to zero we see that this constant must itself be zero. Assuming that the velocity is independent of the vertical position z this simplifies to

$$\frac{\partial}{\partial t}\int_{-\infty}^{\infty} \rho\,\mathrm{d}z + \nabla\cdot\left(\int_{-\infty}^{\infty}\rho\,\mathrm{d}z\,\mathbf{v}\right) = 0. \tag{1.8}$$

Now we introduce a quantity that is of both mathematical use and observational importance; the surface or column density Σ, which is defined as

$$\Sigma(R) = \int_{-\infty}^{\infty} \rho(R)\,\mathrm{d}z. \tag{1.9}$$

Mathematically it clearly allows us to simplify the above equation, whilst observationally it represents a more readily measurable quantity than the volume density – where line of sight distances are hard to measure a column density is far more easily determined. Combining this mathematical trick with the assumption that the disc is axisymmetric (i.e.that $\mathbf{v} = \mathbf{v}(R)$) we obtain the final form of the continuity equation for a thin disc,

$$R\frac{\partial \Sigma}{\partial t} + \frac{\partial}{\partial R}\left(R\Sigma v_R\right) = 0, \tag{1.10}$$

with $v_R = v_R(R)$ the R-component of the velocity. From this equation it is therefore clear that at any given radius the rate at which the surface density evolves is determined by the radial velocity. Furthermore, in the case of an *accretion* disc, it is clear that this radial velocity should be negative, i.e. matter should be moving radially *inwards*.

1.2.1.2 Centrifugal Balance

Turning now to the Navier-Stokes equation, by considering the radial component we obtain the following;

$$\frac{\partial v_R}{\partial t} + v_R\frac{\partial v_R}{\partial R} + \frac{v_\theta}{R}\frac{\partial v_R}{\partial \theta} + v_z\frac{\partial v_R}{\partial z} - \frac{v_\theta^2}{R} = -\frac{1}{\rho}\frac{\partial P}{\partial R} - \frac{\partial \Phi}{\partial R} + \frac{1}{\rho}\,\nabla\cdot\sigma|_R\,, \tag{1.11}$$

where $v_\theta = v_\theta(R)$ is the azimuthal component of velocity and where the last term on the RHS represents the radial component of stress. At this point we have to make two further simplifying assumptions. Firstly, we assume that the radial component of velocity is much smaller than the azimuthal component, i.e. $v_R \ll v_\theta$. This is intuitively reasonable, as if the two were comparable the disc would be exceptionally short-lived, accreting on to the central object in less than the time required for an orbit. The second assumption is that the only non-vanishing components of the (symmetric) stress tensor are the shear terms $\sigma_{R\theta}, \sigma_{Rz}$ and $\sigma_{\theta z}$. (The diagonal components $T_{RR}, T_{\theta\theta}$ and T_{zz} represent normal stresses and thus apply only to the bulk viscosity of the fluid. As this is (usually) only important in shocks we assume these leading terms to be negligible in comparison to the shear (off-diagonal) terms.) Since the disc is thin, we expect the azimuthal velocity to be invariant under translations in z, and thus the shear terms involving z will also be zero. This leaves $\sigma_{R\theta} = \sigma_{\theta R}$ as the only non-zero component, and from our previous assumptions of thinness and axisymmetry we may infer that this term varies only with radius. Tak-

ing the R component of the stress tensor divergence, we then obtain $\nabla \cdot \sigma|_R \equiv 0$ (see Appendix A), and equation 1.11 can therefore be reduced to the following rather simpler form,

$$-\frac{v_\theta^2}{R} = -\frac{1}{\rho}\frac{\partial P}{\partial R} - \frac{\partial \Phi}{\partial R}. \tag{1.12}$$

Assuming a barotropic gas such that $P = P(\rho)$, we can introduce the sound speed c_s, which is defined as

$$c_\mathrm{s}^2 = \frac{\mathrm{d}P}{\mathrm{d}\rho}, \tag{1.13}$$

and the first term on the RHS of equation 1.12 becomes

$$\frac{1}{\rho}\frac{\partial P}{\partial R} = \frac{c_\mathrm{s}^2}{\rho}\frac{\partial \rho}{\partial R} \sim \frac{c_\mathrm{s}^2}{R}. \tag{1.14}$$

Now we invoke a requisite condition for the disc to be thin, namely that $c_\mathrm{s} \ll v_\theta$ (a condition which will be demonstrated in the following section), and thus we find that to a first approximation the radial component of the Navier-Stokes equation describes the centrifugal balance within the disc, such that

$$\frac{v_\theta^2}{R} \approx \frac{\partial \Phi}{\partial R}. \tag{1.15}$$

For a thin, non-self-gravitating disc where the gravitational potential is dominated by the central object of mass M, the potential is given by $\Phi = -GM/R$, and thus the tangential component of velocity v_θ is given by the Keplerian rotation speed, namely

$$v_\theta = R\Omega_K, \quad \text{where} \quad \Omega_K = \sqrt{\frac{GM}{R^3}}. \tag{1.16}$$

1.2.1.3 Vertical Hydrostatic Equilibrium

I shall now investigate the requirement that discs should be thin by assuming that the disc is in vertical hydrostatic equilibrium, i.e. that in the steady state the vertical component of velocity is zero. By considering the z-component of the Navier-Stokes equation (1.6) we obtain the following;

$$\rho\left[\frac{\partial v_z}{\partial t} + v_R\frac{\partial v_z}{\partial R} + \frac{v_\theta}{R}\frac{\partial v_z}{\partial \theta} + v_z\frac{\partial v_z}{\partial z}\right] = -\frac{\partial P}{\partial z} - \frac{\partial \Phi}{\partial z} + \nabla \cdot \sigma|_z\,. \tag{1.17}$$

where as before, $\nabla \cdot \sigma|_z$ is the z-component of the divergence of the stress tensor. Since the stress tensor is independent of z, from Appendix A we see that this term is identically zero, and since $v_z = 0$ by assumption, and the flow is steady, we find the LHS is also identically zero. For the non-self-gravitating case, where the potential is dominated by the central mass, we use the full (spherical) radius in the definition of the potential (i.e. $\Phi = GM/r$ where $r^2 = R^2 + z^2$), and therefore we find that the vertical structure of the disc is determined according to

$$\frac{1}{\rho}\frac{\partial P}{\partial z} = \frac{\partial}{\partial z}\left[\frac{GM}{(R^2+z^2)^{1/2}}\right]. \tag{1.18}$$

Given that the disc is assumed to be thin, i.e. $z \ll R$, we may expand the argument of the differential on the RHS as a Taylor series in z/R, and thus obtain the expression

$$\frac{c_\mathrm{s}^2}{\rho}\frac{\partial \rho}{\partial z} = -\frac{GMz}{R^3}, \tag{1.19}$$

where we have again used the definition of the sound speed given in equation 1.13. This is now a tractable equation for the density as a function of z, and where the sound speed is independent of height, the solution becomes a Gaussian, such that

$$\begin{aligned}\rho(z) &= \rho_0 \exp\left[-\frac{GMz^2}{2R^3c_\mathrm{s}^2}\right] \\ &= \rho_0 \exp\left[-\frac{z^2}{2H_\mathrm{nsg}^2}\right],\end{aligned} \tag{1.20}$$

where ρ_0 is the density of the fluid at the disc midplane, i.e. where $z = 0$. Here we have also set

$$H_\mathrm{nsg} = \sqrt{\frac{c_\mathrm{s}^2 R^3}{GM}} \tag{1.21}$$

to be a characteristic height scale, and thus the scale height H_nsg for a non-self-gravitating disc (where the gravitational potential is dominated by that of the central massive object) is given by

$$H_\mathrm{nsg} = \frac{c_\mathrm{s}}{\Omega_K}, \tag{1.22}$$

where Ω_K is the Keplerian angular frequency (equation 1.16).

The condition that a non-self-gravitating Keplerian disc should be thin (i.e. that

$H_{\rm nsg}/R \ll 1$) therefore amounts to requiring that

$$\frac{c_{\rm s}}{R\Omega_K} = \frac{c_{\rm s}}{v_\theta} = \frac{1}{\mathcal{M}} \ll 1, \tag{1.23}$$

i.e. that the disc Mach number $\mathcal{M} = v_\theta/c_{\rm s}$ is much greater than one. In the case where the flow is highly supersonic, we may readily assume that the contribution of the thermal pressure to the disc dynamics is small, and thus assumption made in the previous section that the pressure term has a negligible effect on the rotation curve is validated.

For the case of a self-gravitating disc, where the gravitational potential is dominated by the disc rather than the central object, we may use the assumption of vertical hydrostatic equilibrium to obtain a second, self-gravitating scale height $H_{\rm sg}$. Using a slightly different formulation for the surface density Σ such that

$$\Sigma(R, z) = \int_{-z}^{z} \rho(R, \zeta)\, \mathrm{d}\zeta \tag{1.24}$$

it can be shown that the hydrostatic equilibrium condition equation 1.17 becomes (Lodato, 2007)

$$\frac{c_{\rm s}^2}{\rho}\frac{\partial \rho}{\partial z} = -2\pi G\Sigma(z). \tag{1.25}$$

Two things should be noted here: First is that the derivation of the RHS of this equation is non-trivial, as it requires the Poisson equation (equation 1.3) to be integrated – see Bertin & Mark (1979). Secondly, although slightly different, the two definitions for the surface density Σ (equations 1.9 and 1.24) become equivalent in the limit as $z \to \infty$, and thus beyond the vertical extent of the disc these two definitions are equal and may therefore be used interchangeably.

Although there is no general solution to equation 1.25, for the isothermal case (i.e. where $c_{\rm s}$ is constant) the solution for ρ becomes

$$\rho(z) = \frac{\rho_0}{\cosh^2(z/H_{\rm sg})}, \tag{1.26}$$

where now the self-gravitating height scale is given by

$$H_{\rm sg} = \frac{c_{\rm s}^2}{\pi G\Sigma}, \tag{1.27}$$

and with ρ_0 the midplane density as before. Although only strictly valid for the

isothermal case, this value for $H_{\rm sg}$ is nonetheless taken generally to be a representative scale height in the case of self-gravitating discs.

Before moving on, it is instructive to consider these two height scales in a little more detail. In the case of a non-self-gravitating disc, we should expect $H_{\rm sg} \gg H_{\rm nsg}$, i.e. that at constant density and temperature in order to become self-gravitating the disc would have to be much thicker (and thus more massive) than it is. In a similar manner, in the self-gravitating case we should expect $H_{\rm nsg} \gg H_{\rm sg}$. We would therefore expect the transition from the non-self-gravitating to the self-gravitating case to occur when these two height scales are approximately equal, i.e. where $H_{\rm nsg} \approx H_{\rm sg}$. This leads to the condition that

$$\frac{c_{\rm s}\Omega_K}{\pi G\Sigma} \approx 1 \tag{1.28}$$

at the transition, a condition which we shall find to be of considerable importance in a later section.

1.2.1.4 Angular Momentum Conservation

The final component of equation 1.6 to consider is the azimuthal (θ) component, which as we shall see embodies the conservation of (angular) momentum. The full equation is given as

$$\rho\left[\frac{\partial v_\theta}{\partial t} + v_R\frac{\partial v_\theta}{\partial R} + \frac{v_\theta}{R}\frac{\partial v_\theta}{\partial \theta} + v_z\frac{\partial v_\theta}{\partial z} + \frac{v_R v_\theta}{R}\right] = -\frac{\partial P}{\partial \theta} - \frac{\partial \Phi}{\partial \theta} + \nabla\cdot\sigma|_\theta\,. \tag{1.29}$$

From Appendix A we see that the azimuthal component of the stress tensor divergence is

$$\nabla\cdot\sigma|_\theta = \left[\frac{\partial}{\partial R} + \frac{2}{R}\right]\sigma_{R\theta}, \tag{1.30}$$

and thus in the case of a steady, axisymmetric disc we obtain

$$\rho\left(v_R\frac{\partial v_\theta}{\partial R} + \frac{v_R v_\theta}{R}\right) = \left[\frac{\partial}{\partial R} + \frac{2}{R}\right]\sigma_{R\theta}. \tag{1.31}$$

Rewriting slightly and taking the vertical integral we obtain

$$v_R\frac{\partial}{\partial R}(Rv_\theta)\int_{-\infty}^{\infty}\rho\,{\rm d}z = \left[R\frac{\partial}{\partial R} + 2\right]\int_{-\infty}^{\infty}\sigma_{R\theta}\,{\rm d}z, \tag{1.32}$$

and hence using the (original) definition of the surface density Σ (equation 1.9) and setting T to be the vertically integrated stress tensor such that $\mathsf{T} = \int_{-\infty}^{\infty} \sigma \, \mathrm{d}z$, this becomes

$$\Sigma v_R \frac{\partial}{\partial R}(R v_\theta) = \left[R \frac{\partial}{\partial R} + 2 \right] T_{R\theta}. \tag{1.33}$$

Finally, to determine the conservation of angular momentum we integrate again over the azimuthal angle θ to obtain

$$\begin{aligned} 2\pi R \Sigma v_R \frac{\partial}{\partial R}(R v_\theta) &= 2\pi \left[R^2 \frac{\partial}{\partial R} + 2R \right] T_{R\theta} & (1.34) \\ &= 2\pi \frac{\partial}{\partial R}(R^2 T_{R\theta}), & (1.35) \end{aligned}$$

which, using primes to denote differentiation with respect to R we can rewrite in the standard form,

$$R \Sigma v_R = \frac{1}{(R v_\theta)'} \frac{\partial}{\partial R}(R^2 T_{R\theta}). \tag{1.36}$$

This equation can be seen to have units of mass flux, and as such links the mass accretion rate to the radial rate of change of the viscous torque, via a coupling constant connected to the shear rate at radius R, $(R v_\theta)'$.

I have now considered the three components of the Navier-Stokes equation, in conjunction with the continuity of mass equation, to derive a pair of governing equations for accretion discs in terms of the surface density and radial velocity, along with a self-consistent requirement that such discs be thin and highly supersonic. We can now use this pair of equations to determine the evolution of the surface density in terms of the (vertically integrated) viscous stress, and also a general form for the mass accretion rate.

1.2.1.5 Diffusion of the Surface Density

We note from the continuity and angular momentum equations (equations 1.1 and 1.36) that both contain the mass flux term $R\Sigma v_R$. As such, we may directly substitute equation 1.36 into equation 1.1, and obtain a non-linear diffusion equation for the surface density which links the temporal rate of change of Σ directly to the spatial rate of change of the mass flux;

$$\frac{\partial \Sigma}{\partial t} + \frac{1}{R} \frac{\partial}{\partial R} \left[\frac{1}{(R v_\theta)'} \frac{\partial}{\partial R}(R^2 T_{R\theta}) \right] = 0. \tag{1.37}$$

Although there is no general solution to this equation, it can be readily solved numerically, and in the case of constant stress it can be solved directly via Bessel functions – see for instance Pringle (1981); Frank et al. (2002); Lodato (2007). It is worth noting at this point that equation 1.36 can also be solved directly for the radial velocity, such that

$$v_R = \frac{1}{R\Sigma(Rv_\theta)'}\frac{\partial}{\partial R}(R^2 T_{R\theta}). \tag{1.38}$$

As equations 1.37 and 1.38 are in terms of the generic variables v_θ and $T_{R\theta}$, they represent a general description of viscously evolving discs, and as such make no assumptions about the form of the viscous stress other than that it should arise through shear. At this point therefore, it is instructive to consider the *form* (if not the exact mechanism) of the viscosity in a little more detail.

1.2.2 The Viscous Stress

The form of the viscous stress tensor and indeed the progenitor for the viscosity itself is a subject of much debate, and remains one of the great unsolved problems of disc physics. Consideration of the molecular viscosity of hydrogen (which will most likely dominate the gas fraction) shows that the time required to accrete even a very low mass (0.005 $M_\odot$) disc about a solar-mass star would be $\sim 10^2$ Hubble times (see for instance Lodato, 2007), and molecular viscosity can therefore be firmly discounted as a major driver of accretion. Nonetheless, we can relatively easily construct a plausible functional form for the viscous stress, and this allows us to look at the general properties of the viscosity in more detail. (It should be noted however that the following is neither exhaustive nor fully rigorous – for further details see Frank et al. (2002); Pringle (1981); Shakura & Sunyaev (1973) for example.)

We start by noting that for an azimuthal velocity profile $v_\theta(R)$ there is a shear rate $\dot{s}$ such that

$$\dot{s} = R\frac{\partial}{\partial R}\left(\frac{v_\theta}{R}\right) = R\Omega', \tag{1.39}$$

where we have used the fact that $v_\theta \equiv R\Omega$, with Ω as the angular frequency and where the prime denotes differentiation with respect to R as before. We can therefore construct a (vertically integrated) shear stress ς such that $\varsigma = \mu R\Omega'$, where μ is the vertically integrated dynamic viscosity of the gas. Furthermore, we can put this in terms of the kinematic viscosity $\nu = \mu/\Sigma$, and by noting that the shear stress ς is

simply the $R\theta$ component of the stress tensor (i.e. $\varsigma = T_{R\theta}$) we obtain

$$T_{R\theta} = \nu \Sigma R \Omega'. \tag{1.40}$$

This equation now allows us to consider the qualitative effects of viscosity on accretion discs, while the uncertainty regarding the exact form of the viscosity is conveniently reduced to the single variable ν. It is worth noting however, that in the case of rigid body rotation (where $\Omega = const$) the viscous stress vanishes, as would be intuitively expected – a viscous torque is only present when there is differential rotation between neighbouring annuli. As noted by Clarke & Pringle (2004), the viscous torque acts to transfer (angular) momentum *down* an angular velocity gradient, and hence for Keplerian rotation (where Ω decreases with radius) the torque acts to transport angular momentum outwards, as would be expected for an accreting disc. In any instance where the rotation rate *increases* with radius however, the torque will act so as to transport angular momentum inwards, and thus the net movement of material will be *outwards* under the action of viscous processes.

1.2.3 The α Prescription

The functional form of the viscosity may be further simplified by a basic consideration of what *might* give rise to viscous effects. In their seminal paper of 1973, Shakura & Sunyaev considered the properties of a turbulent viscosity, i.e. one arising not from the random thermal motions of particles (as is the case for hydrodynamic or molecular viscosity) but from random *turbulent* motions. Noting that (in the isotropic case) the maximum size of a turbulent cell is roughly the disc scale height H, and assuming that the turbulence is generally subsonic – supersonic turbulence would be highly dissipative, and would become subsonic on relatively short timescales – on dimensional grounds they constructed a viscosity of the form

$$\nu = \alpha c_{\mathrm{s}} H. \tag{1.41}$$

Here α is a dimensionless parameter, subject only to the condition that (in general) $\alpha \lesssim 1$ based on the conditions above. Assuming H is given by the non-self-gravitating scale height c_{s}/Ω_K, and noting that the local (vertically integrated) pressure $P \sim \Sigma c_{\mathrm{s}}^2$, we find that the stress tensor $T_{R\phi} \sim \alpha P$, i.e. that the stress is given in units of the local pressure, as would be expected.

Note however that, while useful, this parameterisation still tells us nothing about the nature of the viscous mechanism itself – we have simply concentrated all our uncertainties into the single, dimensionless parameter α. Furthermore, there is no reason to *expect* discs to be turbulent, as the Rayleigh stability criterion (that $(R^2\Omega)' > 0$) is readily satisfied (Pringle, 1981), unless there is a further mechanism to drive the turbulence. Nonetheless, this parameterisation has been used extensively, and remains very useful for forming a qualitative understanding of the viscous process.

1.2.4 Viscous Dissipation

Now that we have a functional form for the viscous stress, we can now consider the energetics of the disc in a little more detail. By a dimensional analysis of equation 1.35 we see that for an annulus of infinitesimal width δR, the viscous torque τ acting upon it is given by (Frank et al., 2002)

$$\tau = \frac{\partial}{\partial R}(2\pi R^2 T_{R\theta})\delta R. \tag{1.42}$$

Given that the annulus is rotating at a rate Ω, there is therefore a power W associated with the torque, such that

$$\begin{aligned} W &= \Omega \frac{\partial}{\partial R}(2\pi R^2 T_{R\theta})\delta R \\ &= \left[\frac{\partial}{\partial R}(2\pi\Omega R^2 T_{R\theta}) - 2\pi R^2 T_{R\theta}\Omega'\right]\delta R. \end{aligned} \tag{1.43}$$

Integrating across the radial range of the disc then gives the total power generated by the viscosity. Notable however is that the first term would integrate to give $[2\pi\Omega R^2 T_{R\theta}]_{R_{\rm in}}^{R_{\rm out}}$, where $R_{\rm out}$ and $R_{\rm in}$ are the outer and inner radii of the disc respectively. This term therefore represents a rate of viscous *convection* of energy through the disc, and is determined only by the boundary conditions. The second term in equation 1.43 is essentially the torque $(2\pi R^2 T_{R\theta})$ multiplied by the local shear rate $(\delta R\Omega')$, which represents the rate at which the torque is extracting energy from the fluid. This power $(2\pi R^2 T_{R\theta}\Omega'\delta R)$ is therefore the rate at which mechanical energy is being viscously dissipated as heat.

If the disc is to remain thin over long timescales, this heat must be lost. Given that the disc has two faces, the total area of the annulus available to radiate away

this energy is $4\pi R\delta R$, and thus the viscous dissipation rate $D(R)$, i.e. the rate at which energy is viscously liberated from the disc per unit surface area, is given by

$$D(R) = \frac{2\pi R^2 T_{R\phi}\Omega'\delta R}{4\pi R\delta R} = T_{R\phi}R\Omega'. \tag{1.44}$$

Finally, using the form of the stress tensor found in equation 1.40, this becomes

$$D(R) = \nu\Sigma(R\Omega')^2. \tag{1.45}$$

Here we note that the viscous dissipation rate is strictly non-negative, and, like the viscous stress, vanishes only in the case of rigid body rotation.

An interesting final point to note here is that the viscosity plays two roles in the energy budget of the disc. One is the expected dissipation, which occurs as a result of the local liberation of heat through the viscosity. The other however, is to transport rotational energy through the disc via convection, and this is dependent on external, non-local conditions. This distinction between local and non-local processes is important, and is one that will be revisited later.

1.2.5 Steady State Mass Accretion Rates

To return briefly to equation 1.36, we noted in Section 1.2.1.4 that it is expressed in units of a mass flux. By incorporating the correct factor of 2π from equation 1.35 and using the form of the viscous stress given in equation 1.40, we obtain the following equation for the steady mass accretion rate $\dot{M}$ on to the central object,

$$\dot{M} = -2\pi R v_R \Sigma = -\frac{2\pi}{(R^2\Omega)'}\frac{\partial}{\partial R}(\nu\Sigma R^3\Omega') \tag{1.46}$$

subject to the condition that v_R must be negative for accretion to occur. (Note that the first equality can also be obtained by direct integration of the continuity equation (1.10), from which the accretion rate arises as integration constant.) We may now integrate the outer equality in equation 1.46 with respect to R, to obtain

$$\dot{M}R^2\Omega + 2\pi\nu\Sigma R^3\Omega' = \dot{J}, \tag{1.47}$$

where $\dot{J}$ is the constant net angular momentum flux. A value for $\dot{J}$ can be obtained by considering the boundary between the star and the disc inner radius (see for instance Lodato 2007; Frank et al. 2002), but essentially, at large radii it becomes

negligible. In this case the inward advection of angular momentum $\dot{M}R^2\Omega$ balances the outward transport of angular momentum through viscous processes $2\pi\nu\Sigma R^3\Omega'$, and we therefore obtain the following expression for the mass accretion rate;

$$\dot{M} = -\frac{2\pi\nu\Sigma R\Omega'}{\Omega}. \tag{1.48}$$

Finally, as before we note that in the Keplerian case, $R\Omega' = -3\Omega/2$ and thus $\dot{M}$ becomes

$$\dot{M} = 3\pi\nu\Sigma, \tag{1.49}$$

the standard result for Keplerian discs.

1.2.6 Timescales

We have already considered the shear rate $R\Omega'$ for the disc and could therefore construct a timescale $t_{\rm shear} = 1/|R\Omega'|$ to characterise it. However, except in the above discussion this is not very instructive, especially (as we shall see below) the shear timescale is roughly equivalent to the dynamical timescale. There are nonetheless a number of other timescales characteristic of discs that are of importance, and it is to these that we now turn.

1.2.6.1 The Dynamical Timescale

The dynamical timescale $t_{\rm dyn}$ is the shortest timescale present in the disc, and is given by the reciprocal of the angular frequency;

$$t_{\rm dyn} = \frac{R}{v_\theta} = \Omega^{-1}. \tag{1.50}$$

This is clearly related to the orbital period, and indeed this is equal to $2\pi t_{\rm dyn}$.

1.2.6.2 The Vertical (Hydrostatic) Timescale

The timescale on which vertical hydrostatic equilibrium is established is also important, in that if it is long, the assumption of hydrostatic equilibrium is invalid. However, noting that in the non-self-gravitating case the characteristic height scale is $H_{\rm nsg}$ and the characteristic velocity is $c_{\rm s}$, we obtain a timescale

$$t_{\rm z} = \frac{H_{\rm nsg}}{c_{\rm s}} = \Omega_K^{-1} \tag{1.51}$$

using equation 1.22. In the case of Keplerian discs this is equal to the dynamical timescale, and even for non-Keplerian rotation, we still have $\Omega \sim \Omega_K$, and so the two timescales remain similar. Thus hydrostatic equilibrium is established on (approximately) the dynamical timescale, and therefore the thin disc assumption that $H \ll R$ is valid.

1.2.6.3 The Viscous Timescale

Using the definition of $T_{R\theta}$ (equation 1.40) in equation 1.37, we find that the surface density evolution is governed by the following equation;

$$\frac{\partial \Sigma}{\partial t} + \frac{1}{R}\frac{\partial}{\partial R}\left[\frac{1}{(R^2\Omega)'}\frac{\partial}{\partial R}(\nu \Sigma R^3 \Omega')\right]. \tag{1.52}$$

From a dimensional analysis of this equation, we see that the surface density Σ varies on a timescale

$$t_\nu \sim \frac{R^2}{\nu}. \tag{1.53}$$

By employing the α-prescription of Shakura & Sunyaev (equation 1.41) and using the definition of the disc height (equation 1.22) this can be re-written as

$$t_\nu \sim \frac{R^2\Omega^2}{\alpha c_s^2 \Omega} = \alpha^{-1}\mathcal{M}^{-2}\, t_{\rm dyn}. \tag{1.54}$$

Therefore (since $\alpha \lesssim 1$) the thin disc corollary that the disc flow is highly supersonic implies that the viscous timescale is much longer than the dynamical timescale, $t_\nu \gg t_{\rm dyn}$.

Additionally, by looking at the ratio of the viscous to dynamical timescales,

$$\frac{t_\nu}{t_{\rm dyn}} = \frac{R^2\Omega}{\nu}, \tag{1.55}$$

we see that this is simply the Reynolds number of the flow. Furthermore, it can also be seen to be the ratio of the specific angular momentum $R^2\Omega$ of the flow to that removed by the viscosity ν (which also has units of specific angular momentum), and thus ν can be thought of as the specific angular momentum transported through radius R by viscosity per dynamical time.

1.2.6.4 The Cooling Timescale

For a disc with an internal energy per unit surface area U, and a given cooling rate per unit surface area $\dot{U}$, we may construct a cooling timescale $t_{\rm cool}$ such that

$$t_{\rm cool} = \frac{U}{\dot{U}}. \tag{1.56}$$

While this may seem somewhat arbitrary, recall that we already know the viscous dissipation rate $D(R)$ of the disc from equation 1.45, and further that we can define the internal energy per unit surface as

$$U = \frac{\Sigma c_{\rm s}^2}{\gamma(\gamma-1)}, \tag{1.57}$$

where γ is the ratio of specific heats, in a direct analogy with the three dimensional case. Note that the internal energy per unit surface U is related to the *specific* internal energy u via $U = \Sigma u$. Given that in local thermal equilibrium the local cooling rate $\dot{U} = U/t_{\rm cool}$ should be equivalent to the local dissipation rate (the rate at which heat is liberated into the disc through viscosity), and assuming there are no *global* effects to redistribute the internal energy throughout the disc, then using equation 1.41 we find that

$$\frac{\Sigma c_{\rm s}^2}{\gamma(\gamma-1)t_{\rm cool}} = \alpha c_{\rm s}^2 \Sigma\Omega \left(\frac{R\Omega'}{\Omega}\right)^2. \tag{1.58}$$

Noting that $R\Omega'/\Omega = \mathrm{d}\ln\Omega/\mathrm{d}\ln R$ we obtain a relationship between the cooling time and the viscous α parameter, such that in thermal equilibrium

$$\alpha = \left(\frac{\mathrm{d}\ln\Omega}{\mathrm{d}\ln R}\right)^{-2} \frac{1}{\gamma(\gamma-1)\Omega t_{\rm cool}}. \tag{1.59}$$

In applying any of the above analysis to real astrophysical systems however, an understanding of the nature, origin and magnitude of the viscosity ν is crucial. In the next two sections I shall therefore first give a brief overview of the magneto-rotational instability, a process widely expected to provide an α-like viscosity in ionised discs, before moving on to a more in depth overview of the gravitational instability.

1.3 The Magneto-Rotational Instability

For ionised discs threaded by a (weak) magnetic field, one of the most likely sources of viscosity is the so-called magneto-rotational instability, or MRI. Although this was investigated in principle by Velikhov (1959) and Chandrasekhar (1960), who both considered the stability of magnetised Taylor-Couette flows, it was not until Balbus & Hawley (1991) that its astrophysical significance was recognised and it was codified into a means of providing viscosity in accretion discs.

Essentially, for a disc with an ionisation fraction of greater than $\sim 10^{-13}$ (corresponding to a temperature $T \gtrsim 1000K$) the magnetic field lines become coupled to the fluid flow (Blaes & Balbus, 1994; Gammie, 1996; Hartmann, 2009a). As the disc rotates, any field lines connecting neighbouring annuli will become stretched, due to the differential rotation. (Note that this connection between annuli may occur due to either the presence of a toroidal magnetic field, or more likely, due to perturbations within a poloidal field, see for instance Hawley et al., 1995.) The magnetic field then acts to oppose this shear, causing the inner annulus to slow down, losing angular momentum and thereby sinking further into the potential well, and the outer annulus to speed up, gaining angular momentum and moving outwards in radius. Clearly this is unstable, as any initial perturbations will grow at a rate determined by the local shear, and furthermore it is capable of transporting angular momentum (and therefore mass) and hence driving accretion. The onset of the MRI drives turbulent motions in the disc, as has been verified through numerical simulations with various initial magnetic field configurations (Hawley & Balbus, 1991, 1992; Hawley et al., 1995, 1996; Stone et al., 1996). As such, and due to the fact that the instability is determined by local parameters rather than by the global configuration, it is generally possible to consider the MRI-driven turbulence as an α-viscosity, (Balbus, 2003; Hartmann, 2009a) providing an $\alpha \sim 0.01 - 0.001$ (Winters et al., 2003; Sano et al., 2004; King et al., 2007), although there is considerable scatter both within and between different simulations.

One of the key features of the MRI is that for suitably ionised discs it should be active wherever the rotation rate decreases with radius (Balbus & Hawley, 1991), and where even a vanishingly small magnetic field is present. Indeed, too strong a magnetic field will lead to solid body rotation of the gas, suppressing the instability. To prevent this occurring, the magnetic pressure must be less than the thermal pressure (Hartmann, 2009a), or equivalently, the Alfvén speed must be less than the sound speed (Balbus & Hawley, 1991). Nonetheless, these conditions are likely to be

widely satisfied, at least in the inner parts of protostellar/protoplanetary discs. In the outer parts, it is likely that the disc will become layered - the outer layers, ionised through stellar X-ray or cosmic ray irradiation, supporting MRI-driven turbulence, with an inner "dead" zone (see for instance Gammie, 1996; Glassgold et al., 2000; Fromang et al., 2002). However, in regions of the disc where the ionisation fraction is low, the MRI will not operate, and therefore other processes must be present to drive accretion.

1.4 The Gravitational Instability

While the MRI is widely accepted as operating in even relatively weakly ionised discs, for those where the disc temperature $T \lesssim 1000K$ and cosmic/X-ray ionisation is insufficient, the magnetic field is not coupled strongly enough to the flow for the instability to operate. In the case where discs are sufficiently cold and massive however, they may become unstable to the effects of their own self-gravity, and as such this gravitational instability may provide the required angular momentum transport to drive accretion in regions where the external irradiation is too weak to trigger the MRI.

The study of fluid instabilities due to gravity has a heritage going back to the beginning of last century, when James Jeans published his seminal work "The Stability of a Spherical Nebula" (Jeans, 1902). Here he established that for a given temperature, there was a critical mass of gas above which gravity will overcome the thermal pressure support, and the mass becomes unstable to small perturbations. Lindblad (1927) speculated that the arms observed in the so-called "spiral nebulæ" were due to interactions between the orbits and gravitational potential of stars, but stopped short of suggesting that this was an instability. Further work on galactic spirals by Lin & Shu was based on the idea that the spirals were gravitationally induced density waves (Lin & Shu, 1964, 1966) and combined with Lindblad's idea that the structure should be secularly stable (and thus long-lived with respect to the orbital timescale) this became known as the Lin-Shu hypothesis. In essence, this states that spiral structure formed is a neutrally stable mode of a galactic disc (Binney & Tremaine, 2008). This can be generalised to any self-gravitating disc, and as such gravitationally induced spiral modes may equally be present in young protostellar discs, driving accretion on to the protostar, and possibly leading to the formation of low mass or brown dwarf companions, potentially even gas giant

planets.

In this section I shall therefore briefly consider the Jeans instability, before deriving and discussing the response of gaseous discs to gravitational perturbations.

1.4.1 The Jeans Instability

To demonstrate the Jeans instability we consider an infinite, static homogeneous medium of density ρ_0 and then introduce small perturbations $\rho_1, \mathbf{v}_1, h_1$ and Φ_1 in the density, velocity, enthalpy and potential respectively[2]. Using the fluid equations as defined in Section 1.2.1 and linearising, the continuity, Euler and Poisson equations respectively become

$$\frac{\partial \rho_1}{\partial t} + \nabla \cdot (\rho_0 \mathbf{v}_1) = 0, \tag{1.60}$$

$$\frac{\partial \mathbf{v}_1}{\partial t} + \nabla(h_1 + \Phi_1) = 0, \tag{1.61}$$

$$\nabla^2 \Phi_1 - 4\pi G \rho_1 = 0. \tag{1.62}$$

Similarly, using the definition of the specific enthalpy (again, as given in the Section 1.2.1) and linearising, one finds that

$$h_1 = c_\mathrm{s}^2 \frac{\rho_1}{\rho_0}, \tag{1.63}$$

where c_s is the unperturbed sound speed. Solving the system of equations 1.60 - 1.63 for ρ_1 yields the following single equation

$$\frac{\partial^2 \rho_1}{\partial t^2} - c_\mathrm{s}^2 \nabla^2 \rho_1 - 4\pi G \rho_0 \rho_1 = 0 \tag{1.64}$$

Finally, assuming a solution of the form

$$\rho_1 = \hat{\rho} e^{i(\omega t - \mathbf{k}\cdot\mathbf{r})}, \tag{1.65}$$

where $\mathbf{k}(\mathbf{r})$ is the wavevector at position $\mathbf{r}$ and $\omega = \omega(\mathbf{k})$ is the angular frequency, we can obtain the dispersion relation for (linear) perturbations in a uniform, static

[2]Note that this requires us to invoke the Jeans Swindle, whereby the unperturbed gravitational potential of the gas must be cancelled by some unspecified external potential for the derivation to be mathematically accurate. However, the equations in the perturbed quantities remain accurate – in this case the end justifies the means... See also Binney & Tremaine (2008).

homogeneous medium;

$$\omega^2 = c_{\rm s}^2|\mathbf{k}|^2 - 4\pi G\rho_0. \tag{1.66}$$

From the form of the solution given in equation 1.65 it is clear that the density perturbations will be oscillatory in time only if ω is real, and will grow or decay exponentially if it becomes complex. For stability we therefore require $\omega^2 \geq 0$ in equation 1.66 (with neutral stability in the case of equality), which translates into the requirement that

$$|\mathbf{k}|^2 \geq \frac{4\pi G\rho_0}{c_{\rm s}^2}, \tag{1.67}$$

or alternatively, that the wavelength $\lambda = 2\pi/|\mathbf{k}|$ is such that

$$\lambda \leq \lambda_J = \frac{c_{\rm s}\sqrt{\pi}}{\sqrt{G\rho_0}}, \tag{1.68}$$

where λ_J is known as the Jeans length. From this it is clear that the medium is stable to short wavelength (high wavenumber) perturbations (where $\lambda < \lambda_J$) as these are oscillatory, but for longer wavelengths (smaller wavenumbers) the medium becomes unstable. The form of the Jeans wavelength is intuitively reasonable, as it has the form of the speed at which the medium responds to perturbations (the local sound speed $c_{\rm s}$) times the local free fall timescale $1/\sqrt{G\rho_0}$ (see for instance Hartmann 2009a). If the perturbation wavelength is longer than this distance, information cannot be propagated far or fast enough to counter the collapse, and the medium becomes unstable.

A related quantity is the Jeans mass M_J, defined as the mass within a sphere of diameter λ_J;

$$M_J = \frac{4\pi}{3}\rho_0\left(\frac{\lambda_J}{2}\right)^3 = \frac{\pi^{5/2}}{6}\frac{c_{\rm s}^3}{G^{3/2}\rho^{1/2}}. \tag{1.69}$$

The Jeans mass plays an important role in star (and potentially planet) formation, as it sets an approximate maximum gravitationally stable mass within star (planet) forming regions. Although this is clearly not the whole story (for instance, rotation and the effects of magnetic fields and turbulence are discounted here), it has been suggested that this (thermal) Jeans mass may provide the characteristic mass present in the stellar initial mass function (IMF) (see for instance Bate & Bonnell (2005) and references therein, Larson 1992).

As with the Jeans length it is clear that increasing the density reduces the Jeans mass (making the medium more unstable) whereas increasing the temperature (and

therefore the sound speed) increases M_J, and stabilises the medium. In a similar manner, any given mass of homogeneous, isothermal gas may be rendered unstable to gravitational perturbations by either increasing its density or reducing its temperature. This result can in fact be generalised to other, less restrictive geometries, as will be seen in the next section where I consider the stability of gaseous discs to gravitational perturbations.

1.4.2 Spiral Waves in Discs

Before we can consider spiral instabilities in discs, we need to have a framework with which to describe them. For simplicity, we shall only consider waves in razor thin discs, and we can therefore consider perturbations in the radial (R) and azimuthal (θ) directions only. In order to describe a spiral, it is convenient to use the shape function, such that

$$m\theta + \int_0^R k(R')\,\mathrm{d}R' = c \mod 2\pi, \tag{1.70}$$

for some constant value c. Here m is the number of spiral arms, and given that these arms are equally spaced, we require $m \in \mathbb{N}_0$, that is for m to be a non-negative integer[3]. In turn this means we can define an azimuthal wavenumber, given by m/R. We allow the radial wavenumber k to be dependent on radius, but now with no restrictions on its value, so that $k \in \mathbb{R}$. There are therefore two major classes of waves; where the disc is rotating in the sense of increasing θ, $k < 0$ implies leading waves, with the wave tips pointing in the direction of rotation, whereas $k > 0$ implies trailing waves, where the wave tips point counter to the rotation direction.

Simple geometry then allows us to combine the radial and azimuthal wavelengths to determine the winding or opening angle i of the disc, which is given by

$$\tan i = \frac{m}{kR}. \tag{1.71}$$

This allows us to define the idea of *tightly wound* spiral waves, which therefore have the property that $m/(kR) \ll 1$, meaning that the radial wavenumber is very large (and thus the radial wavelength is very small) when compared with the azimuthal wavenumber. Indeed this property that $m/(kR) \ll 1$ can be used to carry out a Wentzel-Kramers-Brillouin (WKB) expansion of the perturbed fluid equations.

[3]Although we could also consistently require m to be a non-positive integer, conventionally $m \geq 0$, see Binney & Tremaine (2008).

Although this is not the approach I shall use here, it reproduces the dispersion relations that I derive in the following section (see for instance, Binney & Tremaine 2008).

1.4.3 Dispersions Relations for Fluid Discs

Having now established a framework for describing spiral waves, we can use this to determine the stability of a fluid disc to such spiral perturbations. As with the Jeans instability, in order to derive the relevant dispersion relation we assume an underlying, unperturbed solution to the Euler and continuity equations with velocity[4] $\mathbf{v}_0 = R\Omega\hat{\boldsymbol{\theta}}$, surface density Σ_0 and specific enthalpy h_0. Assuming as before a razor thin disc in the plane $z = 0$ with surface density Σ_0, the density becomes $\rho = \Sigma_0\delta(z)$ (where δ is the Dirac delta function), and thus the Poisson equation 1.3 is given by $\nabla^2\Phi_0 = 4\pi G\Sigma_0\delta(z)$. Introducing perturbed quantities as before, such that

$$\begin{aligned} \mathbf{v} &= \mathbf{v}_0 + \mathbf{v}_1, \qquad \mathbf{v}_1 = v_R\hat{\mathbf{R}} + v_\theta\hat{\boldsymbol{\theta}}, \\ \Sigma &= \Sigma_0 + \Sigma_1, \\ \Phi &= \Phi_0 + \Phi_1, \\ h &= h_0 + h_1, \end{aligned} \tag{1.72}$$

and linearising, the fluid equations equations 1.1 and 1.5 reduce to the following;

$$\frac{\partial\mathbf{v}_1}{\partial t} + (\mathbf{v}_0\cdot\nabla)\mathbf{v}_1 + (\mathbf{v}_1\cdot\nabla)\mathbf{v}_0 = -\nabla(h_1 + \Phi_1), \tag{1.73}$$

$$\frac{\partial\Sigma_1}{\partial t} + \nabla\cdot(\Sigma_1\mathbf{v}_0 + \mathbf{v}_1\Sigma_0) = 0, \tag{1.74}$$

where also, as before

$$h_1 = c_\mathrm{s}^2\frac{\Sigma_1}{\Sigma_0}. \tag{1.75}$$

Likewise, Poisson's equation becomes

$$\nabla^2\Phi_1 = 4\pi G\Sigma_1\delta(z). \tag{1.76}$$

We now assume solutions such that the perturbed quantities are proportional to $e^{i(\omega t - m\theta)}$, where ω is the angular frequency of the wave, and m is the number of

[4]Here and henceforth $\hat{\mathbf{R}}$ is the radial unit vector and $\hat{\boldsymbol{\theta}}$ is the azimuthal unit vector.

spiral arms, which implies that

$$\frac{\partial}{\partial t} \to i\omega, \qquad \frac{\partial}{\partial \theta} \to -im. \tag{1.77}$$

Note that for the time being we assume nothing about the R dependence of the solutions. Substituting these into equation 1.73 and solving for v_R and v_θ we obtain

$$\begin{aligned} v_R &= \frac{i}{\kappa^2 - (\omega - m\Omega)^2}\left[\frac{2m\Omega}{R} + (\omega - m\Omega)\frac{\partial}{\partial R}\right](h_1 + \Phi_1), \\ v_\theta &= \frac{1}{\kappa^2 - (\omega - m\Omega)^2}\left[-\frac{m}{R}(\omega - m\Omega) + \frac{\kappa^2}{2\Omega}\frac{\partial}{\partial R}\right](h_1 + \Phi_1), \end{aligned} \tag{1.78}$$

where as usual the epicyclic frequency κ is such that

$$\kappa^2 = 4\Omega^2\left(1 + \frac{R}{2\Omega}\frac{\mathrm{d}\Omega}{\mathrm{d}R}\right). \tag{1.79}$$

The continuity equation (equation 1.74) links these values to the perturbed surface density, such that

$$\Sigma_1 = \frac{i}{R(\omega - m\Omega)}\left(i\frac{\partial}{\partial R}(R\Sigma_0 v_\theta) + m\Sigma_0 v_R\right). \tag{1.80}$$

Substituting the values for v_R and v_θ into equation 1.80, we obtain (after, for a change, rather a *lot* of algebra) the following second order differential equation for the perturbed enthalpy h_1;

$$\left[\frac{\mathrm{d}^2}{\mathrm{d}R^2} + \mathcal{B}\frac{\mathrm{d}}{\mathrm{d}R} + \mathcal{C}\right](h_1 + \Phi_1) - \frac{\kappa^2 - (\omega - m\Omega)^2}{c_\mathrm{s}^2}h_1 = 0 \tag{1.81}$$

where

$$\begin{aligned} \mathcal{B} &= \frac{1}{R}\frac{\mathrm{d}\ln}{\mathrm{d}\ln R}\left(\frac{R\Sigma_0}{\kappa^2 - (\omega - m\Omega)^2}\right), \\ \mathcal{C} &= -\frac{m^2}{R^2} - \frac{4m\Omega\kappa}{R(\kappa^2 - (\omega - m\Omega)^2)}\frac{\partial}{\partial R}(\omega - m\Omega) + \frac{2m\Omega}{R^2\kappa(\omega - m\Omega)}\frac{\mathrm{d}\ln}{\mathrm{d}\ln R}\left(\frac{\kappa^2}{\Sigma_0\Omega}\right). \end{aligned}$$

A full derivation of this equation is given in Feldman & Lin (1973), and it is quoted verbatim in Bertin (2000); Bertin et al. (1989b); Lin & Lau (1979); Lau & Bertin (1978). Similar results are quoted in Griv (2007), who drops some curvature terms, Montenegro et al. (1999), who differ only in the form of the $\mathcal{C}$ coefficient, and

Goldreich & Tremaine (1979), who include an external forcing potential. Without at this stage going into any detail, it is clear from the form of $\mathcal{B}$ and $\mathcal{C}$ that any dispersion relation that arises from equation 1.81 must contain resonances, as both terms become singular wherever $\kappa = \pm(\omega - m\Omega)$, while $\mathcal{C}$ has an additional singularity wherever $\omega = m\Omega$.

1.4.3.1 The Cubic Dispersion Relation

The ODE given in equation 1.81 is clearly non-trivial to solve. However, if we make the assumption that the wave modes are determined by *local* effects only (and thus crucially, only by the *local* gravitational potential) we can make progress. With this is mind, we may assume that the radial dependence of the enthalpy, surface density potential and velocity has the form $e^{-\int^R k\,\mathrm{d}R}$ for some as-yet undefined radial wavenumber $k(R)$ as per the shape function in equation 1.70, and then once again

$$\frac{\mathrm{d}}{\mathrm{d}R} \to -ik. \tag{1.82}$$

Substituting this into equation 1.81 and using the definition of enthalpy (equation 1.75), we obtain the following relationship between the perturbed potential Φ_1 and the perturbed surface density Σ_1;

$$1 = \frac{c_\mathrm{s}^2 + \Phi_1\Sigma_0/\Sigma_1}{(\omega - m\Omega)^2 - \kappa^2}\left[k^2 + \frac{m^2}{R^2} + \frac{1}{\kappa^2 - (\omega - m\Omega)^2}\left(\frac{2m\Omega}{\kappa R}\right)^2\left|\frac{\mathrm{d}\ln\Omega}{\mathrm{d}\ln R}\right| + \frac{2m\Omega}{R^2\kappa(\omega - m\Omega)}\frac{\mathrm{d}\ln}{\mathrm{d}\ln R}\left(\frac{\kappa^2}{\Sigma_0\Omega}\right) + ik\mathcal{B}\right]. \tag{1.83}$$

A second equation linking the perturbed potential and the surface density can be obtained by solving Poisson's equation. By carrying out a WKB expansion to two orders in $\varepsilon = m/(kR)$ (assumed to be small), Bertin & Mark (1979) have established this second relationship, which may be summarised by noting that the in-phase surface density response Σ_a to an applied gravitational potential Φ_a is given by

$$\Phi_a = -\frac{2\pi G\Sigma_a}{|\mathbf{K}|}, \tag{1.84}$$

where $\mathbf{K}$ is the wavevector of the applied potential perturbations. A detailed discussion of this result is given in Binney & Tremaine (2008) for the case of plane waves in a razor thin sheet. (Note that this is strictly valid only in the tight winding limit,

due to the form of ε.)

In the case of a self-consistent response, the perturbed surface density Σ_1 must be equal to Σ_a, and similarly for the potential, $\Phi_1 = \Phi_a$. In terms of the wavevector, $\mathbf{K}$ must be equal to the total wavevector of the modes in the disc, such that

$$\mathbf{K} = k\hat{\mathbf{R}} + \frac{m}{R}\hat{\boldsymbol{\theta}}, \quad \Rightarrow \quad K = |\mathbf{K}| = \left(k^2 + \frac{m^2}{R^2}\right)^{1/2}. \tag{1.85}$$

By combining equations 1.83 and 1.84 to eliminate the perturbed potential Φ_1, and noting that Σ_1 cancels throughout, we obtain the local cubic dispersion relation for spiral waves in a self gravitating disc,

$$\kappa^2(1-\nu^2) = (2\pi G\Sigma|K| - c_{\rm s}^2K^2)\left(1 + \frac{\Gamma}{(1-\nu^2)K^2} + \frac{\Upsilon}{R\kappa\nu K^2}\right), \tag{1.86}$$

where

$$\nu = \frac{\omega - m\Omega}{\kappa}, \tag{1.87}$$

$$\Gamma = \left(\frac{2m\Omega}{R\kappa}\right)^2 \left|\frac{\mathrm{d}\ln\Omega}{\mathrm{d}\ln R}\right|, \quad \Upsilon = \frac{2m\Omega}{R\kappa}\frac{\mathrm{d}\ln}{\mathrm{d}\ln R}\left(\frac{\kappa^2}{\Sigma\Omega}\right), \tag{1.88}$$

and where we have dropped the subscripts on the unperturbed surface density for brevity. Note that in order to obtain equation 1.86 the term proportional to $ik\mathcal{B}$ in equation 1.83 has been dropped. Here I follow the method of Lau & Bertin (1978); Li et al. (1976), who argue that since this term is 90° out of phase with all the others, it does not affect the growth of the spiral modes.

1.4.3.2 Wave-Fluid Resonances

The terms proportional to Γ and Υ on the RHS of equation 1.86 are interesting for a number of reasons. Firstly, both are dependent on m/RK, and therefore must affect the openness of the spiral modes excited – this will be discussed in more detail in the next section. Perhaps more importantly however, both are inversely dependent on the dimensionless wave frequency ν as defined in equation 1.87, and both terms can become singular – as suggested previously these terms therefore define resonances within the disc.

The term proportional to Υ becomes singular in the limit that $\nu \to 0$, which from equation 1.87 implies that $\omega \to m\Omega$. Given that ω determines the angular frequency of waves with m-fold symmetry, it is instructive at this point to separate

ω into two factors, such that $\omega = m\Omega_\mathrm{p}$. Here Ω_p is the angular frequency of a single wave, known as the pattern speed as it represents the rate of rotation of the pattern as a whole. Clearly then, in the limit where $\nu \to 0$, $\Omega_\mathrm{p} \to \Omega$, and the pattern rotates with the local rotation speed of the disc. Hence the Υ term drives the co-rotation resonance in the disc. Notable however is that in the case of Keplerian rotation (where $\kappa = \Omega$) and where $\Sigma \propto \Omega$, Υ becomes identically zero and thus this resonant term is removed entirely. More detailed discussions of the co-rotation resonance are given in Bertin & Haass (1982), where it is suggested that the resonance has the effect of exciting density waves – a topic that I shall return to in Chapter 4.

The term proportional to Γ on the other hand becomes singular whenever $\nu \to \pm 1$, which defines the two Lindblad resonances. These are less important for fluid discs than for stellar discs (see for instance Binney & Tremaine 2008; Bertin 2000) and as such I shall not discuss them in any detail here. The Γ term itself however is related to the dimensionless J term found in Bertin (2000); Lau & Bertin (1978) and others, via

$$J^2 = \left(\frac{2\pi G\Sigma}{R\kappa^2}\right)^2 \left(\frac{2m\Omega}{\kappa}\right)^2 \left|\frac{\mathrm{d}\ln\Omega}{\mathrm{d}\ln R}\right| = \left(\frac{2\pi G\Sigma}{\kappa^2}\right)^2 \Gamma. \tag{1.89}$$

Since both Γ and Υ are dependent on m/R, for a fixed radial wavenumber k these two terms are effectively "openness" parameters – in the tight winding limit neither of these terms are important, especially away from resonances. It is clear however that the kinematics of an open disc, where $\Gamma/K^2, \Upsilon/K^2 \gtrsim 1$, will be qualitatively different to that of a tightly wound one.

1.4.3.3 The "Standard" Quadratic Dispersion Relation

Despite appearances, equation 1.86 is cubic in the total wavenumber K, and as such is still not trivial to solve. However, as stated above, in the tight winding limit ($m/kR \ll 1$) and away from either the Lindblad ($\nu = \pm 1$) and co-rotation ($\nu = 0$) resonances the effects of the Γ and Υ dependent terms becomes small. In this limit it is also clear that $\mathbf{K} \to k\hat{\mathbf{R}}$, and thus we obtain

$$\kappa^2(1-\nu^2) = 2\pi G\Sigma|k| - c_\mathrm{s}^2 k^2, \tag{1.90}$$

a considerable simplification. By rearranging, and employing the definition of the dimensionless frequency ν (equation 1.87) we recover the familiar "standard" quadratic

dispersion relation for spiral waves in fluid discs, such that

$$(\omega - m\Omega)^2 = \kappa^2 - 2\pi G\Sigma|k| + c_\mathrm{s}^2 k^2 \tag{1.91}$$

(see for instance, Binney & Tremaine 2008; Bertin 2000; Lin & Lau 1979). Since this equation is quadratic only in the radial wavenumber k, it is a rather more tractable form of the dispersion relation for tightly wound spirals, and as such this form will predominantly be the one used henceforth. Note that the azimuthal wavenumber m is still present however, in the LHS of equation 1.91.

1.4.4 Stability Criteria

A number of important results can be obtained by considering the properties of equation 1.91. As with the cubic dispersion relation, the disc remains stable to spiral perturbations as long as $(\omega - m\Omega)^2 = m^2(\Omega_\mathrm{p} - \Omega)^2 \geq 0$, with neutral stability in the case of equality. Note that to achieve this neutrally stable state away from co-rotation (a formal requirement for the dispersion relation to remain valid) m must be identically zero, and thus the disc must be axisymmetric. Although no longer representative of a spiral perturbation (except formally in the limit as $m/kR \to 0$) it transpires that considerable general insights can be gained by considering this specific case.

Since equation 1.91 is quadratic in $|k|$, it can readily be solved (with $m = 0$) to give

$$|k| = \frac{\pi G\Sigma}{c_\mathrm{s}^2}\left(1 - \sqrt{1 - \frac{c_\mathrm{s}^2\kappa^2}{\pi^2 G^2\Sigma^2}}\right). \tag{1.92}$$

Since this an equation for the *modulus* of k, rather than k itself, to have any physical meaning $|k| \geq 0$, leading to neutrally stable modes of wavenumber $\mathbf{k} = \pm|k|\hat{\mathbf{R}}$ (equivalent to rings at radial intervals of $2\pi/k$) propagating within the disc. For this to occur, we require that the second, dimensionless term under the square root must be less than unity. On the other hand, for the disc to be *stable* to such perturbations, there must be *no* positive value for $|k|$, and thus

$$Q = \frac{c_\mathrm{s}\kappa}{\pi G\Sigma} > 1, \tag{1.93}$$

where we have taken the positive square root, since all values are defined to be positive. This Q is the famous Toomre parameter (Toomre, 1964), although strictly

speaking he defined the equivalent stability condition for stellar discs, rather than the fluid discs considered here. Nevertheless, it is essentially a statement of the balance between the twin stabilising effects of rotation and (thermal) pressure, as characterised by κ and c_s respectively, and the destabilising effect of (self) gravity, characterised by the surface density. Clearly as the disc mass increases for a given temperature and rotation rate, there will come a point at which gravity dominates, leading to instability. Likewise increasing the temperature increases thermal pressure, stabilising the disc, and rotation acts to provide centrifugal support, and in the case of differential rotation, also to shear out any overdense regions.

Although strictly valid only for the axisymmetric case, the condition that $Q \gtrsim 1$ implies stability to gravitationally induced spiral modes is fairly general. Indeed, directly from equation 1.91 it can be shown that away from axisymmetry the stability condition becomes

$$Q^2 > 1 + \frac{m^2 c_\mathrm{s}^2}{\pi^2 G^2 \Sigma^2}(\Omega_\mathrm{p} - \Omega)^2. \tag{1.94}$$

Nonetheless, this correction remains small generally, and the $Q \gtrsim 1$ stability criterion has been verified by a wealth of numerical simulations of self-gravitating discs, for instance Mayer et al. (2003); Lodato & Rice (2004, 2005); Boley et al. (2007); Nayakshin et al. (2007); Boley & Durisen (2008); Stamatellos & Whitworth (2008); Forgan & Rice (2009).

A further analysis of the term outside the bracket in equation 1.92 shows it to have units of units of (length)$^{-1}$, and in fact this term corresponds to the reciprocal of the self-gravitating height scale,

$$H_\mathrm{sg} = \frac{c_\mathrm{s}^2}{\pi G \Sigma} \tag{1.95}$$

(for a direct derivation of this value see for instance Binney & Tremaine (2008); Lodato (2007)). Hence, from equation 1.92, we see that in the limit where $Q \to 1$, the wavenumber $|k| \to 1/H_\mathrm{sg}$. Thus we expect that a disc cooling towards instability to excite the $k = \pm 1/H_\mathrm{sg} = \pm \pi G \Sigma / c_\mathrm{s}^2$ mode first, i.e. for this to be the most unstable mode. This is intuitively reasonable, as the most unstable wavelength $\lambda_\mathrm{uns} = 2\pi H_\mathrm{sg}$ is therefore of the order of the disc thickness.

1.4.5 Finite Thickness Effects

Although all the results quoted above in the previous section are often used generally, and indeed have well defined height scales, it is important to note that the analysis has been carried out with respect to a *razor thin* disc, i.e. one of zero thickness lying solely in the plane $z = 0$. It is therefore instructive to briefly consider the effects of finite thickness on the disc, and in particular on the dispersion relation, equation 1.91.

Finite thickness acts to reduce the strength of the gravitational potential in the disc mid-plane, by allowing for the fact that the gravitating mass is now spread out over a vertical range $\approx 2H$ (where for now the exact definition of the height scale H remains undetermined). Allowing for a perturbed gravitational potential Φ_1 proportional to $e^{i(\omega t - \int^R k\,\mathrm{d}R) - |kz|}$, Toomre (1964) suggested that the effect should be to reduce the potential by a factor $e^{-|k|H}/(|k|H)$. Bertin (2000) and Vandervoort (1970a) used a correction $1/(1 + |k|H)$, which agrees to leading order in $|k|H$, and which gives the following quadratic dispersion relation corrected for finite thickness effects;

$$(\omega - m\Omega)^2 = \kappa^2 - \frac{2\pi G\Sigma|k|}{1 + |k|H} + c_\mathrm{s}^2 k^2. \tag{1.96}$$

The net effect of this correction term is to stabilise the disc slightly, as can be seen by noting that $1/(1 + |k|H) = 1 - |k|H$ to leading order in $|k|H$, and solving the resulting quadratic as in equation 1.92. In the finite thickness case the square root term gives the following criterion for stability;

$$Q^2 + \frac{H\kappa^2}{\pi G\Sigma} > 1. \tag{1.97}$$

Assuming that the self-gravitating scale height (equation 1.95) is a reasonable estimate for H this condition becomes

$$Q \gtrsim \frac{1}{\sqrt{2}}, \tag{1.98}$$

and thus a finite thickness disc can sustain a lower value of Q than the corresponding razor thin disc of equal mass and temperature. In a similar manner, by considering a finite thickness isothermal disc, Goldreich & Lynden-Bell (1965) obtained a stability criterion of $Q > 0.676$ (as quoted in Gammie 2001), commensurate with the above analysis.

1.5 Condensate Formation Through Fragmentation

It has long been recognised that the initiation and propagation of spiral density waves serves to increase either the velocity dispersion in the case of stellar discs, or the gas temperature in the case of fluid discs (Hohl, 1971; Paczynski, 1978). In both cases, this serves to stabilise the disc, shutting down the gravitational instability. However, in the case of fluid discs, radiative processes can cause the disc to cool down again, reinvigorating the instability, and it is this case that we shall consider henceforth.

Evidently then the long term evolution of self-gravitating fluid discs is strongly dependent on the rate at which they can cool, as compared to that at which heat is introduced to the disc. Considering the heating term first, we note that in the absence of other heat sources the presence of spiral density waves will heat fluid discs through both compression heating (although this should be balanced by the corresponding rarefaction once the wave has passed) and through shock heating. In either case, for a disc in which $Q \sim 1$, and where the most unstable wavelength is dominant (i.e. $|k|H_{\rm sg} \sim 1$) we may (from equation 1.91) assume that $\Omega_{\rm p} \sim \Omega$, and thus that heating occurs approximately on the dynamical timescale Ω^{-1} (Gammie, 2001). Comparing this with the as-yet unknown cooling timescale $t_{\rm cool}$, there are therefore three possible regimes, which we shall now consider in turn.

1.5.1 The Dynamic Steady State

Firstly, in the case where the cooling time is much greater than the heating timescale ($\Omega t_{\rm cool} \gg 1$), an initially marginally unstable disc will be heated rapidly by the gravitational instability until $Q \gg 1$. As the cooling time is long compared to the dynamical time, this state will persist for many rotation periods, and the disc can be considered stable except on secular timescales. Numerical studies of this regime are complicated by the presence of artificial numerical dissipation (discussed further in Section 3.4), which tends to dominate over any other physical form of heating when the disc is in the hot state. Nonetheless, from the point of view of investigating the gravitational instability, apart from the initial (transient) heating, investigation of this limit is not very edifying.

However, in the case where the heating and cooling timescales are similar

($\Omega t_{\rm cool} \sim 1$), we may expect an initially hot, gravitationally stable disc to cool slowly towards instability. The onset of instability pumps heat back into the disc, raising the disc temperature and either quenching the instability (leading to limit cycle behaviour) or enabling the disc to settle into a dynamic, marginally stable, quasi-steady state in which the stability parameter Q is maintained close to 1. This self-regulated state has been investigated both theoretically (Paczynski, 1978; Bertin & Lodato, 1999) and through numerical experiments (Gammie 2001; Johnson & Gammie 2003; Lodato & Rice 2004, 2005; Pickett et al. 2003; Mejía et al. 2005; Boley et al. 2006, to name but few). In essence, the spiral modes propagate through the disc (saturating at amplitudes determined by the non-linear rather than the linear regime) and provide enough heat to balance to radiative cooling. This quasi-steady state may then persist for many dynamical times, until eventually the surface density evolves significantly enough (on the viscous time) to vary Q.

Lodato & Rice (2004) have investigated this in detail, and find that where the disc to central object mass ratio $M_{\rm disc}/M_* \lesssim 0.25$, the gravitational instability at any point is well described by the local analysis described above. As such it acts as a pseudo-viscous process with an effective $\alpha \lesssim 0.06$, where α is the Shakura & Sunyaev viscosity parameter. Characterising the gravitational instability in this self-regulated regime, in terms of the modes that are excited and how heat is input to the disc, is a subject that will be investigated in detail in Chapter 4.

In the case where $M_{\rm disc}/M_* \gtrsim 0.25$ however, the long-range nature of the gravitational force is more apparent, and as such global effects start to become dynamically important (Lodato & Rice, 2005). Rather than enter a quasi-steady state, the disc enters a more limit cycle like state, whereby it oscillates between relatively quiescent states with only weak spirals, and those dominated by very strong transient two armed ($m = 2$) spiral modes, which drive rapid changes in the surface density associated with strong angular momentum transport. Similar results of transient episodes dominated by low m spiral modes in high mass discs have been observed by Sellwood & Carlberg (1984); Laughlin & Bodenheimer (1994).

1.5.2 The Rapid Cooling Limit

Finally, in the case where the cooling time is much shorter than the heating timescale ($\Omega t_{\rm cool} \ll 1$), an initially gravitationally stable disc will rapidly become unstable, leading to high amplitude perturbations in density. Once a certain maximum amplitude is reached, these over-densities effectively become Jeans unstable, and collapse

under their own gravity to form (potentially bound) condensates. Much work has gone into establishing the boundary at which the transition from the quasi-stable steady state (as detailed above) to the fragmenting regime occurs. Gammie (2001) argued in a similar manner to that given above that fragmentation should occur for $\Omega t_{\rm cool} \lesssim 1$, and used two-dimensional shearing sheet simulations to refine this to $\Omega t_{\rm cool} \lesssim 3$. In order to do this, he used a simple (and common) addition to the specific internal energy equation of the form

$$\left.\frac{\mathrm{d}u}{\mathrm{d}t}\right|_{\rm cooling} = -\frac{u}{t_{\rm cool}}, \tag{1.99}$$

and fixed cooling laws of the form $\Omega t_{\rm cool} = \beta$. A more complex opacity-linked cooling function based on the opacities of Bell & Lin (1994) was used by Johnson & Gammie (2003) (although still using a 2D shearing sheet simulation) to show that this fragmentation boundary could vary by up to an order of magnitude, dependent on the opacity regime. The details of this dependence, and the dependence of the fragmentation boundary on the cooling function will be discussed in greater detail in Chapter 5. Notwithstanding this, in the case where the ratio of dynamical to cooling times is held constant (i.e. $\Omega t_{\rm cool} = \beta$ for some value β) fully three-dimensional models have confirmed that fragmentation occurs when $\Omega t_{\rm cool} \sim 1 - 10$.

In doing so they have highlighted other, less clear dependencies of the fragmentation process. Rice et al. (2003a) for instance considered the fragmentation of a disc where $M_{\rm disc} = 0.1 M_*$, with a surface density profile $\Sigma \propto R^{-7/4}$, and found that fragmentation occurred for $\Omega t_{\rm cool} \lesssim 3$, in accordance with Gammie (2001). However, with a surface density profile $\Sigma \propto R^{-1}$ in an otherwise identical disc, Rice et al. (2005) found the fragmentation occurs for $\Omega t_{\rm cool} \leq 6$. In Chapter 4 I shall show that for the same disc parameters but with $\Sigma \propto R^{-3/2}$ the fragmentation boundary is such that fragments form for $\Omega t_{\rm cool} \leq 4.5$. As all these simulations were conducted with (essentially) the same code and the same numerical set-up, these results should therefore be directly comparable. It seems clear then that there is a weak dependence of the fragmentation boundary on the surface density, but as yet the exact nature of this dependence remains undetermined.

Furthermore, whether this translates into a direct disc mass dependence is less certain. Rice et al. (2005) show that with equal surface density profiles ($\Sigma \propto R^{-1}$), the fragmentation boundary remains fixed at $\Omega t_{\rm cool} \leq 6$ for disc masses in the range $M_{\rm disc}/M_* = 0.1 - 0.5$. However, the non-fragmenting runs at each disc mass

were only evolved for an extra outer rotation period relative to the non-fragmenting runs (presumably due to computational cost) and thus whether they *would* have fragmented if evolved significantly further is a moot point.

A clear link has been however demonstrated between the value of the adiabatic index γ and the fragmentation boundary. Both Lodato & Rice (2005) and Rice et al. (2005) show that increasing γ decreases the value of $\Omega t_{\rm cool}$ at which the disc can remain stable, i.e. it decreases the fragmentation boundary. This can be readily understood by considering equation 1.59, which in the case of a Keplerian disc can be given as

$$\alpha = \frac{4}{9\gamma(\gamma-1)\Omega t_{\rm cool}}. \tag{1.100}$$

What is found is that varying γ causes the values of $\Omega t_{\rm cool}$ at fragmentation to vary, but whilst maintaining a constant value of $\alpha \approx 0.06$ (Rice et al., 2005; Lodato & Rice, 2005). Given that these simulations are in the regime where the gravitational instability can be thought of as a predominantly local process ($M_{\rm disc}/M_* \lesssim 0.1$, Lodato & Rice 2004) these results therefore suggest that there is a maximum steady-state gravitationally-induced pseudo-viscous stress that the instability can provide before the onset of fragmentation.

It is worth noting that all these simulations have only considered discs that started with all the particles on circular orbits. Given that we may expect accretion events (particularly around AGN) to be somewhat chaotic, it is quite likely that discs with significant eccentricity will be common. The liberation of orbital energy as heat as the flow circularises may therefore be important in stabilising the disc against fragmentation, even in situations where it might otherwise be expected. However, Alexander et al. (2008b) have investigated this for constant eccentricity discs where the pericentres are co-linear at all radii and find that in this instance energy is *not* in general liberated quickly enough to prevent fragmentation. Nevertheless, due to tidal stripping as the clumps pass through pericentre the eccentricity may significantly affect the growth of the fragments that do form, and furthermore discs with a spread of eccentricities and/or pericentres may undergo substantially increased heating through shocks, and thus they may be stabilised to some degree through circularisation.

1.5.3 Other Drivers of Instability

Finally, note that we have only considered cooling as a means of driving discs towards instability, as this process is the one considered throughout the bulk of this thesis. From the nature of the stability parameter Q, it is however clear that other processes may cause discs to become gravitationally unstable. Another plausible possibility is that mass loading of the disc due to material falling on to it increases the surface density sufficiently to drive Q below 1, thus leading to instability. This situation is expected to occur during the early phases of star formation, where the disc is being fed via the infalling envelope.

Kratter et al. (2010) have investigated the effects of mass infall on to isothermal discs, and suggest that high infall rates can likewise lead to fragmentation, although potentially at rather higher Q values than would otherwise be expected. They further suggest that the gravitational stresses that the disc can support before the onset of fragmentation may be much higher, with a time-averaged $\alpha \approx 1$, and subsequently significantly increased accretion rates. Time-averaged values are of less use here however, as both Kratter et al. (2010) and Lodato & Rice (2005) found that for high mass discs the disc does *not* settle into a quasi-steady state, but its evolution is dominated by transient two-armed ($m = 2$) spirals. These global events are less well described by the local theory given above, and in turn drive large variations in the instantaneous Q and α values. This would lead in practice to sporadic bursts of accretion on to the central object, rather than the roughly constant accretion rate expected for the predominantly local steady state case.

Having now discussed the possibilities that exist for the gravitational stability (or otherwise) of discs in a largely theoretical sense, we are now in a position to apply this to a variety of realistic physical situations, and how this may be both constrained and verified by observations.

2

Observations and Implications

We are here and it is now. After that everything tends towards guesswork.

Terry Pratchett

2.1 Introduction

Having presented a theory of gravitational instability in discs in the previous chapter, I shall now turn to the astrophysical implications of this theory, what it may mean at various scales and how its ramifications may be observed or inferred. In this section I shall therefore consider the effects of the gravitational instability on a variety of astrophysically relevant discs, particularly as it affects the evolution of discs about young stellar objects (YSOs) and the potential for planet formation, but also briefly at the larger scales of galactic and AGN discs. I shall present details of certain observational techniques used to detect the presence of discs about YSOs, and what constraints such observations place on our theoretical understanding.

2.2 The Gravitational Instability in Galactic and AGN Discs

Perhaps the most obvious examples of spiral waves in discs, and indeed those which triggered the study of the subject initially, are to be found in galactic discs (see for instance Bertin et al., 1989a,b). As can be seen in Fig. 2.1 these show a wide range of morphologies, from the tightly-wound one-armed spiral seen in NGC4725, to the wider two-armed spiral in M81 and the barred spiral of NGC1300. Rather stranger is the structure observed in NGC4722 (Fig. 2.1, lower left), which shows both leading and trailing spiral arms.

Whilst these morphologies are all gravitational in origin, they are *not* however all due to the instability discussed in Chapter 1. The stranger cases of NGC4722 and NGC4725 are thought to be due to a minor merger (Buta et al., 2003) and tidal effects from another galaxy (Haynes, 1979; Wevers et al., 1984) respectively, and are therefore both due to interactions with an external body. The spirals in M81 and NGC1300 are both due to gravitational *self*-interaction, but NGC1300 is gravitationally bar-unstable, which in itself gives rise to well-defined two-armed spirals launched at co-rotation (Romero-Gómez et al., 2006, 2007; Athanassoula et al., 2009a,b). Spiral structures in galaxies might thus be the result of a number of factors, and in various cases the instability discussed in the previous chapter alone will not explain all of the observed features. Despite this caveat however, it remains generally applicable wherever gravitational self-interaction is the dominant dynamical process.

At this point is is instructive to note that while we have considered only a single phase fluid model for galactic structure, there is a corresponding instability due to self-gravity in the stellar (collisionless) component also. This obeys a dispersion relation similar to that given in equation 1.91, excepting that the stellar velocity dispersion takes the place of the sound speed (as may be expected, see for instance Binney & Tremaine 2008), and the self-gravity term is slightly reduced. This variation in the dispersion relation due to the fact that fluid discs are collisional while stellar discs are collisionless means that the gravitational instability behaves differently in these two components. This becomes particularly important in the non-linear regime, where large amplitude perturbations will lead to shocks in fluid discs which will not be present in the stellar component. Due to this difference, I shall not discuss collisionless discs in any further detail, and will concentrate instead on primarily fluid discs.

In contrast to galactic discs, those present around (active) galactic nuclei may be primarily gaseous (initially at least), with accretion from the disc on to the central supermassive black hole (SMBH) expected to drive the observed activity (see for instance Frank et al. 2002). However, about the SMBH at the centre of our own galaxy (Sgr A*) we observe a stellar disc[1] (Genzel et al., 2000, 2003; Levin & Beloborodov, 2003; Paumard et al., 2006; Bartko et al., 2009), the stars in which appear to have formed in situ (Bartko et al., 2010, 2009; Paumard et al., 2006). Both numerical simulations (Hobbs & Nayakshin, 2009; Bonnell & Rice, 2008; Nayakshin et al., 2007, 2006) and the observed "top heavy" initial mass function (IMF) (Bartko et al., 2010; Fatuzzo & Melia, 2009; Nayakshin & Sunyaev, 2005) strongly suggest that these stars formed in situ, the principal contender for this formation being direct gravitational fragmentation of a gaseous disc (Collin & Zahn, 2008; Nayakshin et al., 2007; Levin & Beloborodov, 2003; Shlosman & Begelman, 1989, 1987; Kolykhalov & Syunyaev, 1980).

If this was indeed the case, this therefore implies both that it is possible for AGN discs to become massive and cold enough to initiate the gravitational instability, and also that the cooling timescales are short enough for fragmentation to occur. As such, the presence of this *stellar* disc may be the best empirical confirmation

[1]There is some discussion as to whether there are one or two discs in the galactic centre, see for instance Bartko et al. 2009, 2008; Lu et al. 2009, 2006; Genzel et al. 2003. It should also be noted that there is a further population of stars (the so-called S stars) that orbit even more closely about Sgr A*, which probably do not share a common formation mechanism with the disc stars (Ghez et al., 2005, 2003; Eisenhauer et al., 2005).

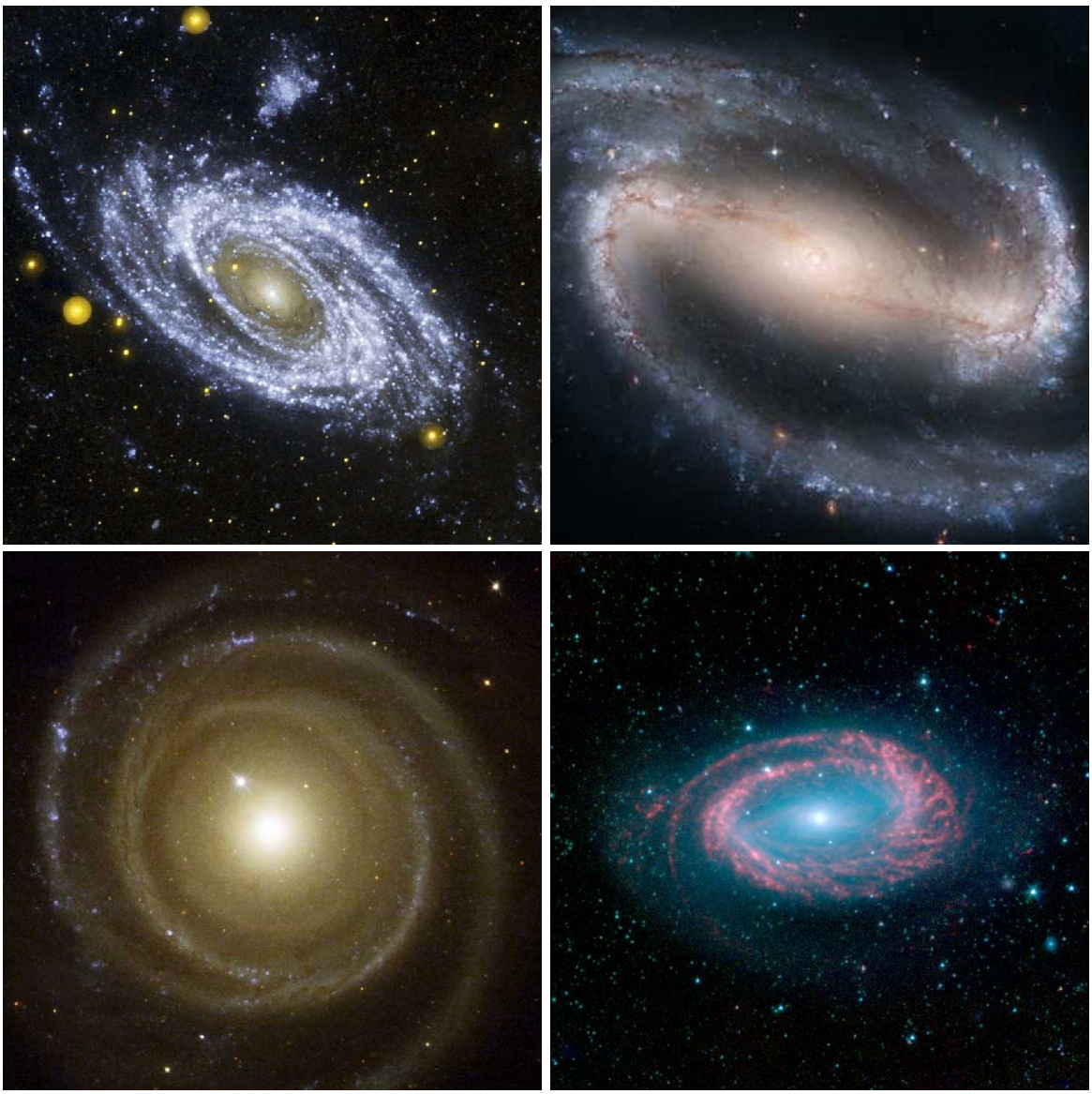

FIGURE 2.1: *Various morphologies of spiral density waves in galactic discs. Top left shows M81 imaged in UV, showing two principal arms with intermediate flocculent arms. Top right shows NGC1300 in optical, which also has a pronounced $m = 2$ mode, but also has a prominent bar. NGC4722 (bottom left, again in optical) shows both leading and trailing arms (the leading outer arms are thought to be due a relatively recent minor merger; Buta et al. 2003), and the Spitzer infrared image of NGC4725 (bottom right), which shows a one armed logarithmic spiral, or* spira mirabilis.
M81 image credit: NASA/JPL-Caltech/Harvard-Smithsonian CfA;
NGC1300 image credit: Hubble Heritage Team, ESA, NASA;
NGC4622 image credit: NASA and The Hubble Heritage Team (STScI/AURA);
NGC4725 image credit: NASA/JPL-Caltech/SST

currently available for the theory that bound condensates (in this case, stars) can form due to the gravitational instability of a *gaseous* disc. It should be noted however that alternative formation mechanisms have been put forward for these disc stars (for example the disruption of a stellar cluster by the tidal field of the SMBH, Levin & Beloborodov, 2003; Gerhard, 2001), and as such the exact origin of these stars remains unknown.

2.3 Circumstellar Discs

Moving now to a smaller scale, we can consider the effects of the gravitational instability in protostellar and protoplanetary discs. Before proceeding however, it is worth defining these two terms, and indeed some of the plethora of others that are applied to circumstellar discs, as there is some ambiguity regarding their use.

2.3.1 Protostellar/Protoplanetary Disc Nomenclature

In order to understand the naming conventions for circumstellar material it is necessary to consider such material in the context of star formation, and as such a generic cartoon of the star formation process is shown in Fig. 2.2.

As a rough guide to the nomenclature and evolutionary stages, an initial molecular cloud as shown in frame *a* collapses under its own gravity (frame *b*), leading to an overdense core (not shown). At this stage, the forming protostar/disc system may be tentatively referred to as a Class 0 object[2]. The infalling envelope then forms a disc and the core condenses into a protostar as shown in frame *c*, and this essentially corresponds to a Class I object. Discs present at these stages, where the envelope is still falling on to the disc, I shall refer to as *protostellar* discs (PSDs), and it is these discs that are most likely to be self-gravitating, due to the mass loading of the infalling material (Bertin & Lodato, 2001a; Vorobyov & Basu, 2006; Hartmann, 2009a). Mass accretion rates on to the protostar are measured to be at their highest at this stage, with episodic outbursts as characterised by FU Orionis-type objects a

[2]It should be noted at this point that I use the object 'Class' definitions in a fairly fluid manner, in common with many theoreticians. By implication however there are rather more rigorous *observational* definitions, determined by the slope of the spectral energy distribution between 2 and 14 μm, see for instance Lada, 1987; Andre & Montmerle, 1994. These more precise definitions are however hampered by the variation in SED with factors other than simply the evolutionary state of the object (such as source orientation, protostellar mass), hence the less formal associations given here.

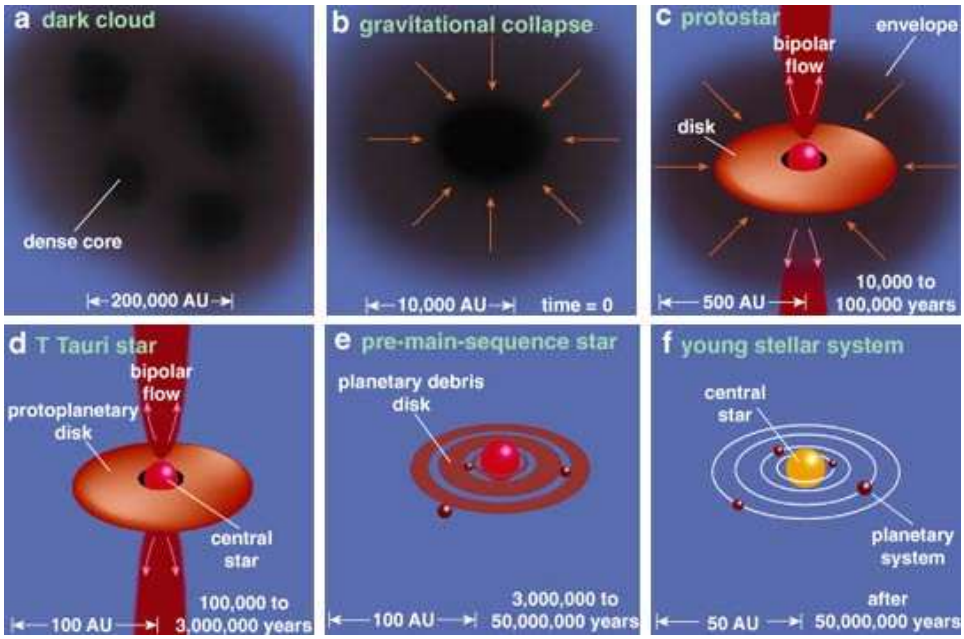

FIGURE 2.2: *Cartoon image of star formation illustrating the disc formation and evolution process, with approximate time and length scales. Image courtesy of Greene (2001).*

possibility.

Frame *d* of Fig. 2.2 shows the start of what I shall henceforth term the *protoplanetary* disc (PPD) stage, where the infall has ceased, the disc is now slowly processing mass on to the central pre-main-sequence (PMS) star and is also (possibly) forming planetesimals. Systems at this stage of their evolution may also be known as (Classic) T-Tauri or CTT systems[3], or Class II objects. Finally, as the disc is accreted on to the PMS star, planets form within the disc (frame *e*) which is steadily depleted, principally through accretion or X-ray/EUV photoevaporation (Alexander et al., 2006a,b; Ercolano et al., 2008, 2009; Owen et al., 2010), until all its gas is removed. At this stage the system can be termed a Weak Line T-Tauri (WTT) system or a Class III object, and the disc is known as a debris disc, consisting almost entirely of solid material. Thereafter, after ~ 10 million years, the star moves on to the main sequence, the disc has formed a planetary system (possibly with a Kuiper/asteroid belt and Oort cloud analogue) and the natal process of star formation is essentially over.

It should be noted at this point that this is only intended as a rough guide, and

[3]Once again it is worth noting that strictly speaking, the classical and weak line T Tauri objects are differentiated by the strength of optical emission lines such as Hα (Andre & Montmerle, 1994).

that as mentioned in the footnotes, some of the definitions are rather more rigorous than presented here. Nonetheless, throughout the following I shall use the prefixes 'protostellar' to refer to discs with an infalling envelope, and 'protoplanetary' to refer to gas-dominated, planet-forming discs once the envelope has been removed.

2.3.2 Energy and Angular Momentum Transport

As mentioned previously, due to the infall of material from the envelope on to the disc, protostellar discs are likely to undergo a self-gravitating phase. Observational evidence for this is currently both scarce and inconclusive, as the spiral patterns seen in the disc of GSS-39 in Ophiuchus are not robust at the 3σ detection level (Andrews et al., 2009) and those in IRAS 16293-2422B (Rodríguez et al., 2005) may equally plausibly be due to interaction with a companion. Nonetheless the theoretical arguments for such a phase to occur are good, and with the advent of the Atacama Large Millimeter/sub-mm Array (ALMA) in the near future more conclusive results should be forthcoming. This will be discussed in detail in Chapter 6.

Perhaps the most important effect of the gravitational instability operating in discs of protostellar material is to provide a means of transporting material on to the forming star. As mentioned in Chapter 1, the spiral density waves introduced to the disc by the gravitational instability carry both energy and (angular) momentum, in much the same way as sound waves carry energy and (linear) momentum. By this outward transport of angular momentum, matter can be accreted on to the protostar, and similarly gravitational energy is liberated as accretion luminosity. Although the details of this transport process will be discussed at greater length in Chapter 4, here I shall briefly consider the implications of this process for protostellar and protoplanetary discs.

For a steady-state self-gravitating disc of $\lesssim 10\%$ of the protostellar mass, with Q of order unity, an effective Shakura-Sunyaev α due to self-gravity of 0.05 and a temperature of $\sim 10^2 - 10^3$K, the expected mass accretion rate is $\sim 10^{-8}$ to a few times 10^{-7} $M_\odot$ yr^{-1}, in line with observed values (as will be discussed shortly). However, for discs more massive than 10% of the stellar mass (as may be expected in the protostellar case) the presence of strong *transient* spirals may be able to drive accretion rates much larger than this (Lodato & Rice, 2005), and infall on to the disc may also enable rather greater accretion rates (Harsono et al., 2010; Kratter et al., 2010).

In any case, it is expected that at low radii (less than a few AU) the gravitational

instability will become less efficient, as the requirement that $Q \sim 1$ becomes less plausible in the inner regions irradiated by the protostar unless the surface density becomes very large. In these inner regions, the stellar irradiation means that the MRI is likely to be active, and thus accretion on to the protostar will continue via this process (Armitage et al., 2001; Zhu et al., 2009, 2010). The exact nature of the transition between these two forms of transport is not well known however, and indeed it is not clear that there *is* a steady transition at all accretion rates. Indeed, observations of Herbig-Haro objects and knots of gas in the jets ejected from YSOs (Bally, 2007a,b) imply instead that in the protostellar phase accretion is more likely to be an unsteady process.

During the infall period of protostellar disc evolution, the mass infall rate of the envelope on to the disc of $\sim 10^{-5}\ M_{\odot}\ \mathrm{yr}^{-1}$ (Armitage et al., 2001) is likely to exceed the accretion rate on to the star by several orders of magnitude. The mismatch described above between the regions where the gravitational instability and the MRI can operate efficiently means that mass is transported inwards via gravitational instabilities, and then accumulates at low radii, where neither the gravitational instability nor the MRI is fully operative. Unsteady accretion is now possible in the manner of the so-called FU Orionis objects, where "quiescent" systems with low accretion rates rapidly enter an "outburst" phase – a period of high accretion on to the protostar where the mass accretion rate increases to $\sim 10^{-4} - 10^{-5}\ M_{\odot}\ \mathrm{yr}^{-1}$ (Herbig, 1977; Hartmann & Kenyon, 1996). The accumulated material heats up (due to the increase in density and through stellar irradiation) enough to become ionised, and is then accreted rapidly on to the protostar in an outburst via the MRI (Armitage et al., 2001; Gammie, 1999; Zhu et al., 2009, 2010) and possibly also through thermal instabilities (Bell & Lin, 1994). While the gravitational instability may be quenched or reduced during the outburst phase, once the disc cools it will once again transport material in towards the central star, and the process repeats until the reservoir of mass in the outer disc is depleted enough to inhibit the instability. (It should be noted however that this is not the only plausible mechanism for driving the outbursts – accretion of clumps formed through the gravitational instability alone (Vorobyov & Basu, 2006) and triggering the thermal instability by a massive planet further out in the disc (Lodato & Clarke, 2004) have also been suggested.)

Nonetheless, it is clear that the action of the gravitational instability in circumstellar discs leads to the transport of mass inwards from large radii. In this manner,

and in conjunction with the MRI and potentially also the thermal instability, it drives the evolution of the disc, particularly in the protostellar (Class I) phase.

2.3.3 Companion Formation in Protostellar Discs

A further significant effect of the gravitational instability operating in protostellar discs is the possible formation of companion brown dwarfs (BDs) or low mass stars (LMSs) through direct gravitational fragmentation of the spiral arms, in a precisely similar manner to the formation of stars in SMBH discs discussed above. While the early theoretical calculations of Matzner & Levin (2005) implied that this mechanism was probably not valid, later numerical models using radiative transfer to model the cooling suggest that it *is* a viable method of producing low mass companions (Stamatellos et al., 2007a; Stamatellos & Whitworth, 2009a; Walch et al., 2009, 2010). Later theoretical work has also tended to support this view (Rice et al., 2010), especially for higher mass stars (Kratter et al., 2008).

Although numerical simulations appear to be robustly supportive of the mechanism as a whole, its efficiency in terms of producing companions is less well constrained. Simulations starting from well-formed discs, both initially gravitationally stable (Rice et al., 2003b; Stamatellos et al., 2007a) and initially unstable (Stamatellos & Whitworth, 2009a,b), form multiple companions, ranging from low mass hydrogen burning stars to brown dwarfs, with a small number of planetary mass objects. In these simulations a large number of the brown dwarfs and all of the planetary mass objects are ejected into the field through dynamical interactions, with a significant fraction in brown dwarf-brown dwarf binaries. Furthermore, those brown dwarfs that do remain bound end up on wide orbits ($\sim$ 200 AU), in agreement with the so-called 'brown dwarf desert' – the apparent under-representation of sub-solar companions to solar-type stars, especially on orbits of less than 5 AU (Marcy & Butler, 2000; Klahr & Brandner, 2006).

As a caveat however, simulations modelling the formation of protostellar cores and the evolution of the discs around them from the initial collapse of molecular cloud cores show that the introduction of turbulence to the core has the effect of decreasing the likelihood of disc fragmentation (Walch et al., 2009, 2010). Similar results have been found by Begelman & Shlosman (2009), who find that in the case where *supersonic* turbulence is present (such that the average turbulent speed $v_{\rm turb} > c_{\rm s}$), this supports the disc against fragmentation where otherwise it would be expected to break up. This is intuitively reasonable, as turbulent motions support

the disc against collapse, and furthermore, dissipation of the turbulence imparts heat to the fluid. Indeed from $v_{\rm turb}$ it is possible to infer an effective 'turbulence temperature', in an analogous manner to the connection between the sound speed of a fluid and its thermodynamic temperature. With this in mind the decay timescale of the turbulence becomes important in determining the stability of the disc, in a manner analogous to the cooling timescale (Bertin & Lodato, 1999), and will therefore also impact the formation of stellar companions.

2.3.4 Protoplanetary Discs

A logical extension to the above mechanism for forming stars in protostellar discs is the formation of (principally gas giant) planets in self-gravitating protoplanetary discs. Although there is less evidence that such discs should be self-gravitating, this has been an active field of research since the idea was posited by Alan Boss in 1997, resurrecting an idea first proposed by Cameron (1978) and then seemingly passed over. As ever, the three strands of enquiry of theory, numerical experimentation and observation have produced varying levels of agreement on whether this mode of planet formation is feasible.

Many numerical studies have been carried out, using a variety of different methods (see Durisen et al. 2007 for a thorough review). The original simulations of Boss (1997, 1998, 2000) and others, such as Mayer et al. (2002) used either a simple ideal gas or a 'locally isothermal' equation of state, the latter case essentially maintaining an imposed radial temperature profile. In either case, fragmentation of the disc to produce gas giants was found within 20 AU, resulting in planets of approximately 5 Jupiter masses (Mayer et al., 2002, 2003). An improved ideal gas equation of state using a fixed (Pickett et al., 2003; Mejía et al., 2005) or radius-dependent (Gammie, 2001; Rice et al., 2003a) cooling time $t_{\rm cool}$, subject to the condition that the change in specific internal energy with time $\dot{u}$ is given by $\dot{u} = -u/t_{\rm cool}$, has also been extensively used, with the result that the discs become unstable to fragmentation over a wide radial range. More recently however, simulations with more realistic cooling functions and radiative transfer (Cai et al., 2006; Boley et al., 2006, 2007) have shown that giant planet formation through gravitational instability is unlikely to occur at low radii, but is still plausible at radii of $\gtrsim 10^2$ AU.

This conclusion is borne out by the theoretical analyses of Boley (2009); Clarke (2009); Rafikov (2009, 2005); Levin (2007, 2003) (and see also Chapter 5), who suggest that outside approximately 50 - 100 AU the fragmentation of self-gravitating

protoplanetary discs into bound clumps of order Jupiter's mass is still feasible. In addition, Kennedy & Kenyon (2008) have demonstrated that the timescale for planet formation via the core-accretion gas-capture model (which is widely accepted as the principal mode of planet formation at low radii, see Klahr & Brandner 2006; Lissauer 1993) beyond about 20 AU is longer than the expected lifetime of the disc, leaving the gravitational instability as the prime candidate. As there have been various observations of planetary mass objects at radii above this cut off (e.g. the HR 8799 system, (Marois et al., 2008), β-Pic b, (Lagrange et al., 2009), Fomalhaut b, (Kalas et al., 2008)) this may indicate the validity of this mechanism, although it should be noted that other processes such as planet scattering and migration could equally be responsible.

Direct planet formation is not the only way that gravitationally-induced spirals would affect protoplanetary discs however. The presence of spiral density fluctuations can act as a means of enhancing the density of the solid fraction to an extent well in excess of the equivalent increase in the gas fraction density (Rice et al., 2004, 2006), as the solids migrate to local pressure maxima. This could clearly have implications for the rate at which planetesimals grow by collisional agglomeration, and indeed may even lead to the solid fraction itself becoming Jeans unstable (Rice et al., 2004). However, the requirement that gravitationally-induced spirals must be present for this effect to come into play itself restricts this process to the outer, gravitationally unstable reaches of the disc (Clarke & Lodato, 2009) beyond a few tens of AU.

A final effect to consider, in contrast to the concentrating effects of the spiral structure mentioned above, is the possibility that previously existing planetesimals may be gravitationally scattered by the potential of the arms themselves. Both collisionless (pure N-body) and collisional (gas plus planetesimals) simulations have shown that the effects of the structures introduced by the disc's self-gravity is to drive eccentricity increases in the orbits of the planetesimals (Moore et al., 2008; Britsch et al., 2008; Walmswell et al., 2010, in prep.). The combination of these two effects, gravitational focusing on the one hand and scattering on the other, means that it is not clear what the dominant effect will be (although see Britsch et al. 2008 for a detailed discussion), or indeed whether or not such planetesimals would be retained within the system, scattered out of it, or lost on to the central star.

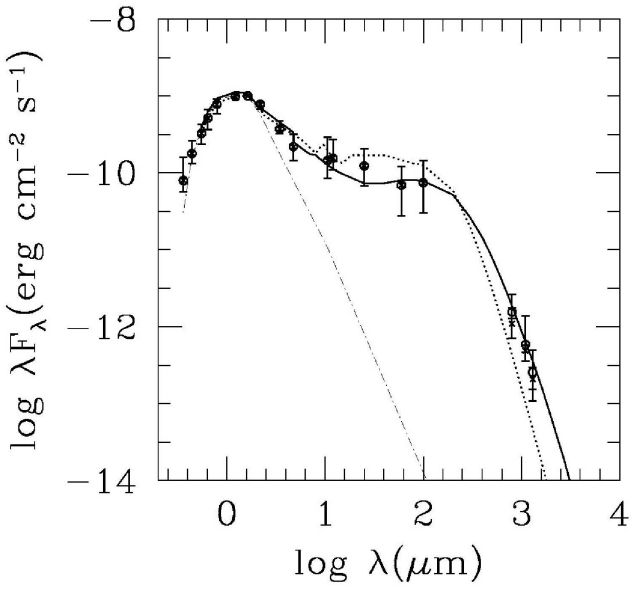

FIGURE 2.3: *Observed median spectral energy distribution (SED) for PMS stars in the Taurus-Auriga star forming region, with and without normalisation for optically thick emission (circles and crosses) respectively. The error bars show the quartile values. Note the stellar blackbody component (dot-dashed line). The solid and dotted lines show fits to the observations from two different dust models. Taken from D'Alessio et al. (2001), based on data from Kenyon & Hartmann (1995)*

2.4 Disc Observation Methods

There have now been many observations of circumstellar discs in a number of star-forming regions such as Orion (Eisner et al., 2008), Taurus-Auriga (Beckwith et al., 1990; Kitamura et al., 2002; Andrews & Williams, 2005, 2007b) and ρ-Ophiuchus (Andre & Montmerle, 1994; Andrews & Williams, 2007a), through a variety of different methods. In this section I shall give a brief overview of some of the methods of observing discs, and what information can be gleaned from them. In particular I shall consider what is probably the most general method of disc *detection*, the spectral energy distribution (SED), and then also I shall consider observations at wavelengths within the millimetre/sub-mm band.

2.4.1 The Spectral Energy Distribution

By considering the emitted spectrum across a range of wavelengths (the so-called spectral energy distribution or SED), it is possible to obtain a considerable amount

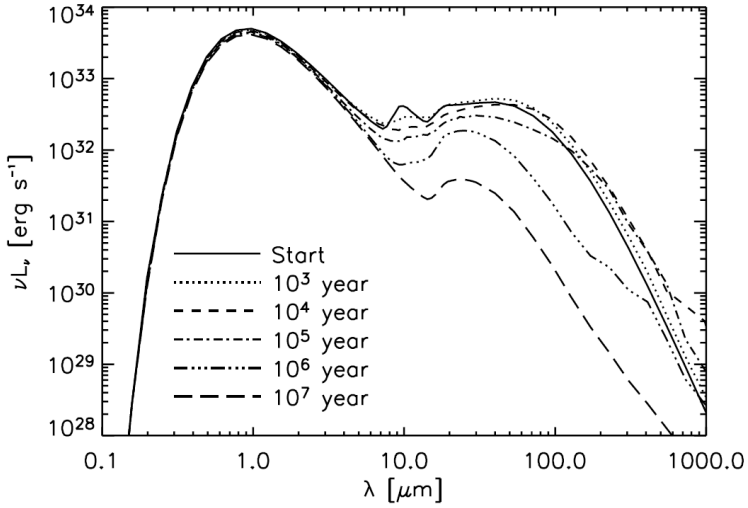

FIGURE 2.4: *Model results showing the clear reduction of the infrared excess due to grain growth within the disc with age. The initial blip at 10 microns is the silicate feature. Taken from Dullemond & Dominik (2005).*

---

of information about the emitting body. As suggested by Lynden-Bell & Pringle (1974), circumstellar discs were first inferred from the infrared excesses of young stars, as rather more flux was detected in the mid to far-infrared than would be expected from simply the blackbody emission of the stellar photosphere. As an example, the median SED of pre-main sequence stars in the Taurus-Auriga star-forming complex is shown in Fig. 2.3, where λF_λ plotted against wavelength, with F_λ being the observed flux at wavelength λ.

While the infrared excess above the stellar photospheric blackbody emission is clearly shown, it is clear that there is considerable scatter (the error bars show the inter-quartile range of the data), particularly at higher wavelengths. This variation can in fact be used to determine the age of the disc, as depletion due to both accretion on to the protostar and grain growth within the disc reduces its excess. Model results of the effects of grain growth are shown in Fig. 2.4, and a clear reduction in the IR excess is shown, until by $\sim$ 10 Myr the disc signature has largely vanished. This reduction in IR excess and consequent estimate for a general disc lifetime is shown to good effect in Fig. 2.5, which plots the percentage of sources with an IR excess in various star-forming clusters against the mean cluster age (from Haisch et al. 2001). This clearly indicates that after 6 Myr the disc fraction should fall to zero, and thus that the mean disc lifetime is approximately 6 Myr, a value commensurate with the

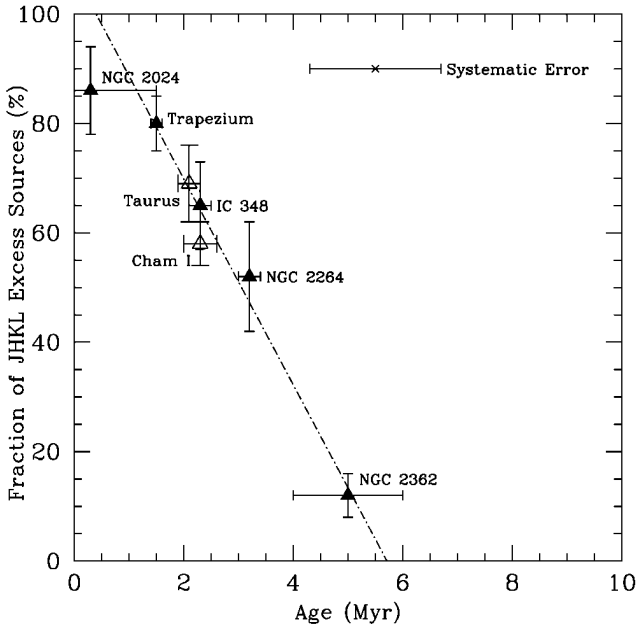

FIGURE 2.5: *Fraction of sources showing an infrared excess in a number of star-forming clusters against mean cluster age, suggesting a mean disc lifetime of approximately 6 Myrs. From Haisch et al. (2001).*

model estimate above.

Present in both Figs. 2.3 and 2.4 is the silicate feature at 10 microns associated with the dust fraction of the disc. The evolution of this feature can be used to trace grain growth and dust processing within the system, and thus by implication planetesimal growth. Clearly the above estimates for the disc lifetime additionally place constraints on the planet formation timescale – gas giant planets in particular must be fully formed by the time the disc disperses.

A further suggestion of Lynden-Bell & Pringle (1974) was a similar excess to the ultra-violet side of the photospheric emission, due to the direct accretion of matter on to the protostar. Although not shown in either of the above figures due to the ready absorption of UV by the interstellar medium (ISM), this excess is present in the SEDs of various objects such as BP Tau (Gullbring et al., 2000) and DL Tau (Kenyon & Hartmann, 1987), and has been shown to be well fit by models of an accretion shock caused by material falling on to the protostar, as shown in Fig. 2.6.

Note that the presence of the silicate feature and from the sensitivity to grain growth it is clear that the IR-ward part of the SED is primarily sensitive to the dust fraction within the disc, whereas the UV-ward part of the SED, which traces the

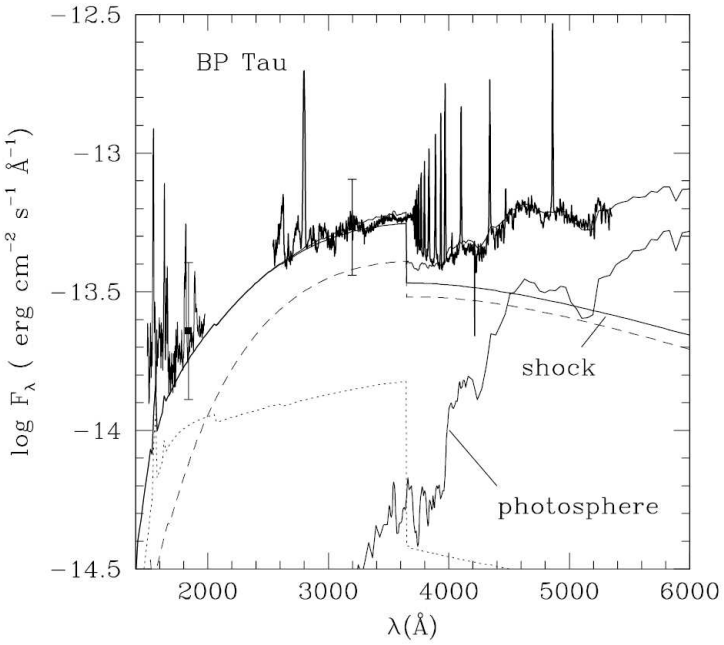

FIGURE 2.6: *Excess UV flux in BP Tau associated with accretion on to the protostar. The heavy solid line shows the observed flux, which is clearly in excess of the photospheric component, and the 'step' in the emission is due to the Balmer break at 3645.6Å (364.56 nm). Taken from Gullbring et al. (2000).*

hot accretion flows, is primarily sensitive to the gas fraction.

Finally, analysis of the SED may potentially also indicate whether or not the disc is self-gravitating (Lodato & Bertin, 2001; Bertin & Lodato, 1999). In contrast to a non-self-gravitating disc of a given mass, the outer parts of a self-gravitating disc of equal mass may be hotter due to the heating brought about by the gravitational instability, leading to an enhanced excess in the mid infrared. This effect may in particular be observationally significant for FU Orionis-type objects, where the disc is sufficiently massive (see Section 2.3.3 above) to affect the SED, but the system is not obscured by the infalling envelope (Lodato & Bertin, 2003).

2.4.2 Sub-Millimetre Observations

While broad wavelength SED observations are a very good indicator of the presence of circumstellar discs, considerably greater detail may be obtained by observing individual systems in narrower frequency bands. An important region of the spectrum for studying disc properties is the so-called millimetre/sub-millimetre range, which broadly speaking covers the spectrum from the far infrared (FIR) at $\sim 10\mu$m (30

THz) up to microwave wavelengths, on the order of 1 cm (30 GHz).

This range is particularly useful for studying discs because dust particles emit optically thin thermal radiation in this waveband. Dust emission is generally optically thick at wavelengths $\lesssim 10\mu$m (Eisner et al., 2008), meaning that it is hard to determine from the observed flux the *amount* of dust present. However, in the optically thin regime, the flux can be used as a tracer for the mass of dust in the disc (Hildebrand, 1983; Beckwith et al., 1990; Andrews & Williams, 2005), which can in turn be used to infer the total mass of material in the disc from an assumed gas-dust mass ratio, usually in the region of 100:1 (see for instance Rafikov, 2006). To a first approximation, this method can be demonstrated as follows:

Resolved observations of discs can provide details about the radial extent of discs, (see for instance Dutrey et al. 1996; Kitamura et al. 2002), and thus we can obtain estimates for the inner and outer radii, $R_{\rm in}$ and $R_{\rm out}$ respectively. For a vertically isothermal disc at an inclination angle i to the observer, the total luminosity at frequency ν is given by

$$L_\nu = 2\pi \cos i \int_{R_{\rm in}}^{R_{\rm out}} B_\nu(T) \left[1 - \exp\left(-\frac{\tau_\nu}{\cos i}\right)\right] R \, {\rm d}R, \tag{2.1}$$

where τ_ν is the *vertical* optical depth of the disc at frequency ν, T is the disc temperature and $B_\nu(T)$ is the Planck (black body) function, given by

$$B_\nu(T) = \frac{2h\nu^3}{c^2} \frac{1}{\exp(h\nu/kT) - 1}. \tag{2.2}$$

Here h is Planck's constant, c is the speed of light *in vacuo* and k is Boltzmann's constant. Furthermore, the total luminosity is linked to the observed flux at F_ν at frequency ν via

$$F_\nu = f_{ex} \frac{L_\nu}{4\pi D^2}, \tag{2.3}$$

where D is the distance to the object, and f_{ex} is a factor accounting for line-of-sight extinction between the source and the observer.

From equation 2.1 it is clear that in the optically thick regime $\tau_\nu \gg 1$ the effect of the exponential term is negligible, and the luminosity is determined solely by the black body emission at frequency ν, $B_\nu(T)$. Optically thick emission from the disc can therefore be used to infer the disc temperature T. However, at longer wavelengths (lower frequencies) where the disc is optically thin, we may use equation 2.1 in conjunction with this disc temperature to determine the optical depth.

Assuming the optical depth to be dominated by the dust fraction, it is related to the *dust* surface density $\Sigma_{\rm d}$ via

$$\tau_\nu = \Sigma_{\rm d}\kappa_\nu. \tag{2.4}$$

Now allowing for the mass ratio of gas to dust via a factor $f_{gd} \approx 100$, such that the total surface density $\Sigma = f_{gd}\Sigma_{\rm d}$, we can expand equation 2.1 to first order in the exponential and substitute into equation 2.3 to obtain

$$F_\nu = \frac{f_{ex}}{2f_{gd}D^2}\int_{R_{\rm in}}^{R_{\rm out}} B_\nu(T)\Sigma\kappa_\nu\, R\,{\rm d}R, \tag{2.5}$$

and thus the link between the surface density and observed optically thin flux is established. Finally, noting that

$$M_{\rm disc} = 2\pi\int_{R_{\rm in}}^{R_{\rm out}} \Sigma R^2\,{\rm d}R, \tag{2.6}$$

we may use the power-law equations for the disc temperature and surface density (equations 2.8 and 2.9) in conjunction with the long wavelength limit in the Planck function equation 2.2 and the assumption that $R_{\rm in} \ll R_{\rm out}$ to obtain the following estimate for the disc mass from the observed flux (Beckwith et al., 1990; Hartmann, 2009a);

$$F_\nu = \frac{f_{ex}}{f_{gd}}\frac{2k\nu^2\kappa_\nu}{c^2D^2}M_{\rm disc}T(R_{\rm out})\frac{2-p}{2-p-q}, \tag{2.7}$$

where $p+q \neq 2$, and q, p are power law indices for the temperature and surface density respectively – these will be explained further in the coming sections.

By considering the variation of the observed flux with wavelength, sub-mm observations can also be used to determine the evolution of the disc in terms of the grain growth within it, as the frequency dependence of the dust opacity is determined primarily by the size distribution of the grains themselves (Wilner et al., 2005; Andrews & Williams, 2005). (Note however that the shape, composition and spin rate of the grains will also have an effect, Rafikov 2006; Draine 2006). As the properties of dust in circumstellar discs have been found to be different to those of dust grains in the interstellar medium (ISM) (Testi et al., 2003; Acke et al., 2004; Natta et al., 2004; Rafikov, 2006), this in turn allows constraints to be placed on the growth and formation of planets/planetesimals within the disc.

Numerous sub-mm surveys have been undertaken of star forming regions in

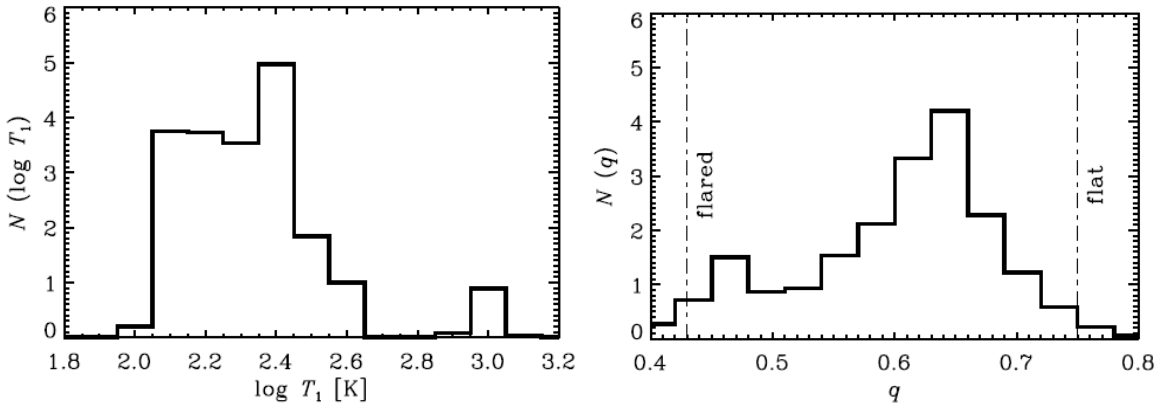

FIGURE 2.7: *Datum temperature T_1 at 1 AU (left) and the temperature power-law index q (right) distributions for circumstellar discs in Taurus. Taken from Andrews & Williams (2007b), which should be referred to for further details. Dotted lines in the right-hand panel show idealised q values for flared and flat discs.*

Taurus-Auriga (Beckwith et al., 1990; Dutrey et al., 1996; Kitamura et al., 2002; Andrews & Williams, 2005, 2007b), Ophiuchus (Andrews & Williams, 2007b,a) and the Orion Nebula Cluster (ONC) (Eisner & Carpenter, 2006; Eisner et al., 2008), which have provided details of a large number of circumstellar discs in these regions. In a similar manner, higher resolution sub-mm observations of single objects have been able to identify sub-structures within the discs of systems such as ϵ-Eridani (Greaves et al., 2005; Backman et al., 2009) and HL Tau (Greaves et al., 2008). In the former case this has been associated with a possible (though unconfirmed) gas giant planet ϵ-Eridani b (Marengo et al., 2006; Benedict et al., 2006; Janson et al., 2007), and in the latter case with the possible formation of a low mass companion. Although the use of sub-mm observations to obtain direct detections of massive planets on short period orbits has been postulated (Wolf & D'Angelo, 2005), the required resolution is not available with current facilities such as the Sub-Millimeter Array (SMA) and the Combined Array for Research in Millimeter Astronomy (CARMA). However, with the advent of the Atacama Large Millimeter/sub-millimeter Array (ALMA) which is currently in construction, such observations should be possible, and indeed using ALMA to detect the spiral patterns induced by self-gravity in circumstellar material is the subject of Chapter 6.

2.5 Observed Disc Properties

As mentioned above, sub-mm observations in particular can be used to determine a wealth of information about circumstellar discs. In this section I shall therefore outline some of the key observational results that can be used to constrain the theory presented in Chapter 1.

2.5.1 Temperature and Surface Density Profiles

The Taurus-Auriga star forming complex is one of the best studied regions in terms of circumstellar discs, and has been extensively observed in the sub-mm. Beckwith et al. (1990); Andrews & Williams (2007b) and others have used such observations to determine temperature profiles of discs around (primarily) sub-solar mass stars, and find that they are well matched by a power-law, such that

$$T(R) = T_1 \left(\frac{R}{1AU} \right)^{-q}, \tag{2.8}$$

where T_1 is the temperature at 1 AU from the central star, and $q \sim 0.5 - 0.7$ is the power law index – note that this is also the power-law index required by the mass estimates described above. Andrews & Williams (2007b), have further been able to show the distributions[4] of both these parameters, and these are shown in Fig. 2.7. From this they deduce that the median temperature at 1 AU is approximately 200K, and the median q value is approximately 0.62. Although specific to the Taurus-Auriga complex, the differences between this and other star-forming regions are relatively minor (Andrews & Williams, 2007a; Eisner & Carpenter, 2006), and thus these results are expected to be reasonably general.

In terms of the surface density, surveys of the Taurus-Auriga (Kitamura et al., 2002; Andrews & Williams, 2007b), and Ophiuchus (Andrews & Williams, 2007a; Andrews et al., 2009) complexes all show very similar results. Again assuming a power law dependence on radius, such that

$$\Sigma(R) = \Sigma_0 \left(\frac{R}{R_0} \right)^{-p}, \tag{2.9}$$

where $\Sigma_0 = \Sigma(R_0)$ is the surface density at some datum radius R_0, the power-law index p (again, the same as used for the mass estimate method given earlier) is

[4]For further details on how the distributions were generated, see Andrews & Williams 2007b.

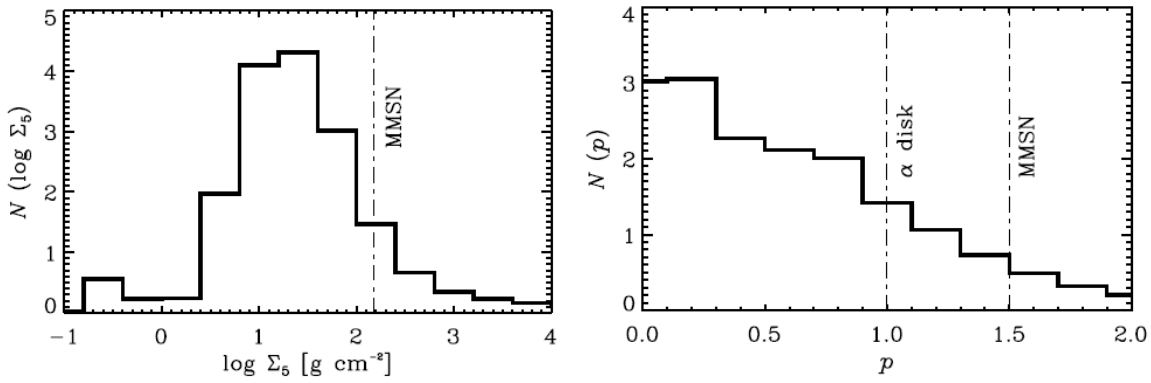

FIGURE 2.8: *Datum surface density Σ_5 at 5 AU (left) and the power-law index p (right) for circumstellar discs in Taurus, as given in Andrews & Williams (2007b). The dashed lines show the expected values for a Minimum Mass Solar Nebula (MMSN) and viscous disc with a constant α viscosity parameter (right).*

found to be in the range $p \sim 0.0 - 1.0$ for Taurus, with a median of approximately 0.5, and in the range $0.4 \lesssim p \lesssim 1.0$ for Ophiuchus, with median 0.9. For Taurus the median surface density at 5 AU is found to be ≈ 14 g cm^{-2} with a range from approximately 10 - 100 g cm^{-2} (Andrews & Williams, 2007b), and the distributions for this fiducial surface density and p are shown in Fig. 2.8. It should be noted that an explanation for the systematic difference in the values for p in between these two samples could be the use of a slightly more complex model for the surface density for the Ophiuchus data – in this case the radial dependence was given by a power law in the inner regions with an exponential tail at large radii (see Andrews et al. 2009; Isella et al. 2009 for further details). The reason for this change was that multi-wavelength observations do not support a sharp density gradient at the outer edge (McCaughrean & O'Dell, 1996; Piétu et al., 2005; Isella et al., 2007; Hughes et al., 2008), but show a variation in outer radius dependent on observation frequency and optical depth (Andrews et al., 2009). To illustrate this point, direct sub-mm observations in Taurus suggest outer radii of the order of 200 AU (Kitamura et al., 2002; Andrews & Williams, 2007b), while the characteristic radii (at which the transition from power law to an exponential decay occurs) of discs of the discs in Ophiuchus fall in the range 20-200 AU (Andrews et al., 2009; Isella et al., 2009), as shown in Fig. 2.9.

In either case, the observed values for p are lower than the theoretical estimate for the so-called Minimum Mass Solar Nebula (Weidenschilling, 1977), which suggests that within the early solar system $p \approx 1.5$, based on the current distribution of mass

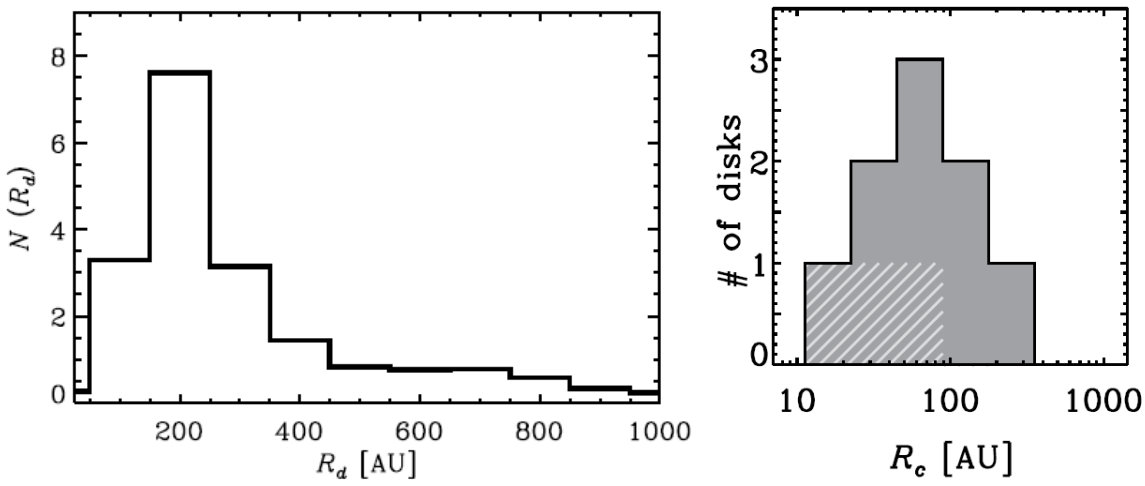

FIGURE 2.9: *Outer radii for Taurus, assuming a sharp disc outer edge (left, taken from Andrews & Williams 2007b) and characteristic radii for discs in Ophiuchus, assuming an exponential surface density decrease (right, Andrews et al. 2009).*

in the planets and asteroids. Clearly however, this assumes that there has been no change in the relative distribution of mass in the solar system since the dispersal of the gaseous component of the circumsolar disc.

2.5.2 Masses and Accretion Rates

Using the method described above, disc masses have been inferred for large samples of discs in the star forming regions of (at least) Taurus, Ophiuchus and the Orion Nebula Cluster (ONC) (Eisner & Carpenter, 2006; Andrews & Williams, 2007b,a; Eisner et al., 2008; Andrews et al., 2009), and are generally found to be low, such that $M_{\rm disc} \sim 10^{-2} - 10^{-3}\ M_{\odot}$. Andrews & Williams (2005) found the discs in Taurus to be approximately log-normally distributed with a mean of -2.31 ± 0.01 (giving a mean disc mass of 0.005 $M_{\odot}$ and variance 0.50 ± 0.02 dex. Relatively few high-mass discs ($M_{\rm disc} > 0.1\ M_{\odot}$) have been detected, although they have been observed; for instance CY Tau ($M_{\rm disc} = 0.129\ M_{\odot}$ Kitamura et al. 2002), GSS 39 in Ophiuchus ($M_{\rm disc} = 0.143\ M_{\odot}$ Andrews et al. 2009), WaOph 6 ($M_{\rm disc} \approx 0.17\ M_{\odot}$, Andrews & Williams 2007b) to name a few. It has been suggested however that there are fewer high mass discs in the ONC, possibly due to photoevaporation from the OB stars of the Trapezium cluster (Eisner et al., 2008). A representative distribution of disc masses in Taurus, taken from Andrews & Williams (2007b) is shown in Fig. 2.10. In any case, it has been suggested by Andrews & Williams (2007b);

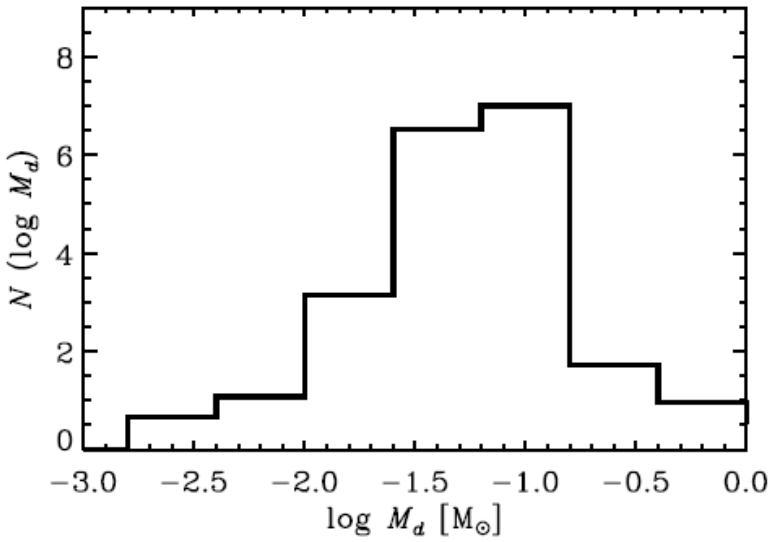

FIGURE 2.10: *Mass distribution for circumstellar discs in Taurus, taken from Andrews & Williams (2007b).*

Hartmann (2009b) that due to uncertainties in the dust opacity (see Section 2.5.3) disc masses obtained from sub-mm flux measurements may be systematically underestimating the true values.

As with the disc masses, there is clearly a spread in the stellar masses observed. Andrews & Williams (2007a) find values in the range $0.1 \lesssim M_*/\ M_\odot \lesssim 3.0$ in Ophiuchus, with a median stellar mass of around 0.3 $M_\odot$. Beckwith et al. (1990) find similar values in the Taurus-Auriga complex. However, while it is relatively easy to determine stellar masses for protoplanetary (Class II) systems, protostellar (Class I) systems are still embedded in the infalling envelope, and it becomes more difficult to isolate the stellar mass. As such, there is some uncertainty in a key parameter of interest, the disc-stellar mass ratio, for Class I objects.

The left-hand panel of Fig. 2.11 illustrates the combined distribution of this parameter in Ophiuchus and Taurus-Auriga for Class II objects as a dashed line, for comparison with the disc mass distribution for the same objects shown as a solid line. The mass *ratio* distribution is systematically shifted to the right, indicating the predominance of sub-solar mass stars. By comparison, the right-hand panel of Fig. 2.11 shows the disc mass distributions for Class II (solid) and Class I (dashed) objects, and here it is clear that discs about latter are generally more massive, indicating that the median disc to star mass ratio would be higher in Class I than Class II systems. Further taking into account the fact that we should expect stellar

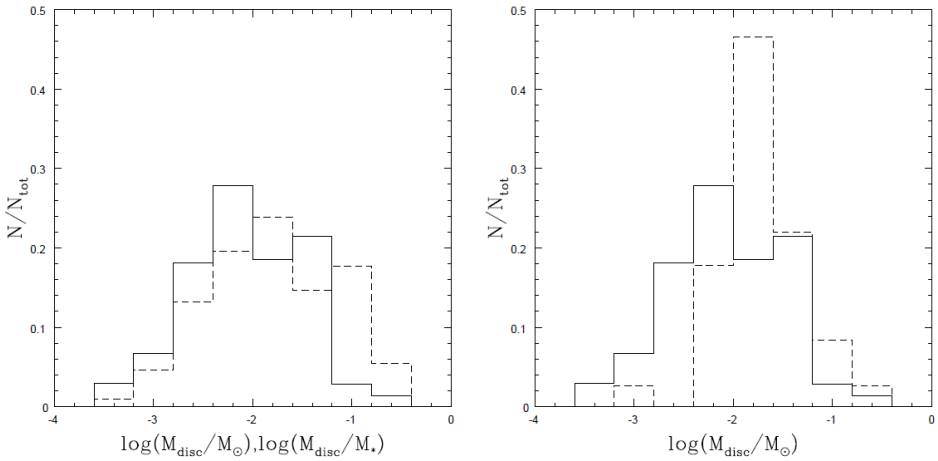

FIGURE 2.11: *Disc-star mass distributions in Taurus-Auriga and Ophiuchus. Left panel shows the distributions of disc masses (solid lines) and disc-star mass ratios (dashed line) for protoplanetary (Class II) objects. Right panel shows the disc mass distribution for protoplanetary (Class II) objects (solid line) and protostellar (Class I) objects (dashed line). Taken from Lodato et al. (2010, in prep.), courtesy of Sean Andrews, with data taken from Andrews & Williams (2005, 2007b)*

masses to be lower in Class I than in Class II systems as they are less evolved, the median mass ratio for Class I objects should be higher again, and it is therefore not unreasonable to expect disc to star mass ratios in the region of 0.1 for Class I objects.

As mentioned in Section 2.4.1, the mass accretion rate is generally measured from the UV continuum excess caused by material falling on to the stellar surface. Observations of circumstellar material in Taurus (Gullbring et al., 1998; Calvet et al., 2004), Orion (Calvet et al., 2004) and the Large Magellanic Cloud (Romaniello et al., 2004) have found typical accretion rates to be of the order of $10^{-7} - 10^{-8}$ $M_\odot$ yr^{-1} for Class II objects (classic T Tauri stars), although this can vary from around 10^{-9} $M_\odot$ yr^{-1} to 10^{-6} $M_\odot$ yr^{-1} (Hartmann, 2009a; Basri & Bertout, 1989; Hartigan et al., 1991, 1995) dependent on the age of the system and the stellar mass.

The age of the system generally has decreasing effect on the accretion rate, decreasing from $\sim 10^{-7}$ $M_\odot$ yr^{-1} at ~ 1 Myr to $\lesssim 10^{-9}$ $M_\odot$ yr^{-1} at $\gtrsim 5$ Myr, when the disc dissipates (Fedele et al., 2010; Sicilia-Aguilar et al., 2010; Hartmann, 2009a; Hartmann et al., 2006). At early times the accretion rate is expected to be unsteady, possibly undergoing repeated FU Orionis type outbursts, as illustrated in the left-hand panel of Fig. 2.12. The right hand panel shows the variation of the

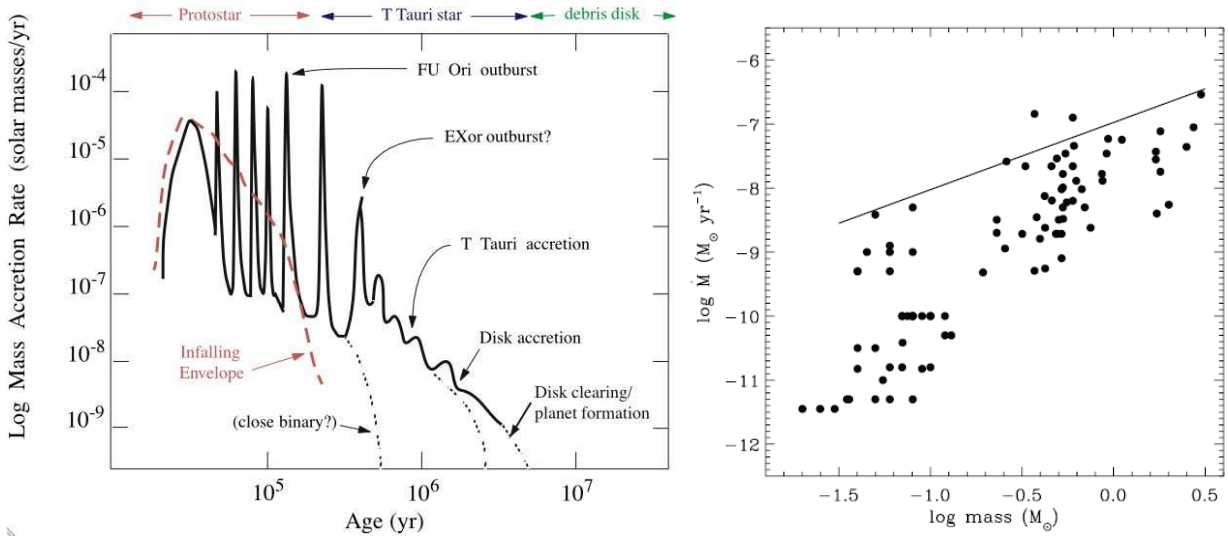

FIGURE 2.12: *Representative mass accretion rates as a function of system age (left), showing the various evolutionary phases (Hartmann, 2009a). The right-hand panel shows the variation of accretion rates with stellar mass, illustrating the approximate $\dot{M} \propto M_*^2$ relation. The solid line shows the rate at which sustained accretion over 1 Myr would imply an initially self-gravitating disc (Hartmann et al., 2006).*

accretion rate with the stellar mass, showing that most systems lie along a line such that $\dot{M} \sim M_*^2$, although the mechanism behind this relation is not currently well understood.

2.5.3 Dust Compositions

Sub-millimetre observations have also been able to place constraints on the grain size distribution within circumstellar discs, by determining the power-law dependence of the dust opacity κ_ν with frequency ν, where

$$\kappa_\nu = \kappa_0 \left(\frac{\nu}{\nu_0} \right)^\beta , \tag{2.10}$$

for some fiducial vales for the opacity and frequency κ_0 and ν_0 respectively. Grain growth within the disc is expected to lead to a reduction in the power-law index β with respect to the ISM value ($\beta \approx 1.7$ Hildebrand 1983), and indeed Andrews & Williams (2007b) find a median value for β of approximately 1.0 in

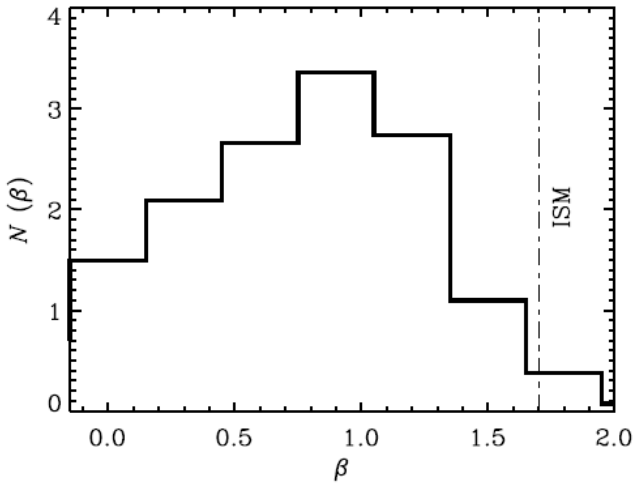

FIGURE 2.13: *Distribution of the power-lax index β of opacity with frequency, taken from Andrews & Williams (2007b). The dotted line shows the value of β for the interstellar medium.*

Taurus, albeit with considerable scatter – see Fig. 2.13. It should be noted that the value for κ_0 is not well constrained, with estimates varying by approximately an order of magnitude, from 0.1 g cm^{-2} (Beckwith et al., 1990) to approximately 0.016 g cm^{-2} (Kramer et al., 1998; Rafikov, 2009). This uncertainty in the absolute value of the dust mass opacity represents one of the primary sources of error in the mass estimates for circumstellar discs (Hartmann et al., 2006; Andrews & Williams, 2007a; Zhu et al., 2008; Hartmann, 2009a).

3

Smoothed Particle Hydrodynamics

Or: How I Learned to Stop Worrying and Love the Lagrangian*

* With apologies to Drs. Strangelove and Price

As far as the laws of mathematics refer to reality, they are not certain; and as far as they are certain, they do not refer to reality.

Albert Einstein

3.1 Introduction

The study of astrophysical phenomena presents a multitude of obstacles to the potential student. In addition to the usual obstacles of understanding the physical properties of the system in question, the sheer scale of astrophysical events renders laboratory experiments impossible in the vast majority of cases. To this end, it has been necessary to assemble a new, numerical laboratory in the form of computational simulations, and conduct experiments and analyses via this medium. The growth of computing power over the past 80 years, from the Colossus of Bletchley Park's Enigma code-cracking efforts in the 1940s, through ENIAC and Los Alamos National Laboratory's modelling of thermonuclear detonations in the 1950s, up to the supercomputers of today, has in turn allowed the computational domain to become a mainstay of astrophysical experimentation.

Two principal approaches to computational simulations have evolved to enable these numerical simulations. Eulerian methods use geometric grids, either fixed or adaptive (the so-called AMR or Adaptive Mesh Refinement codes), with the fluid parameters evaluated over the grid cells. Such codes formed the basis of the revolution in Computational Fluid Dynamics (CFD) that started in the late 1960s and early 70s, and as such they remain the most widely used approach. Applications of such codes cover a huge range, from industrial aerodynamics in the automotive and aerospace sectors, to stress calculations and solid mechanics for civil engineering and architecture, to chemical reaction modelling and protein folding in biomolecular models.

Lagrangian methods on the other hand dispense with fixed points in space and instead evolve the fluid equations in a co-moving frame. A common approach is to use discrete particles that are carried with the flow – hydrodynamic (and other) properties are then evaluated at the particle positions, and are calculated from a weighted average of the values on other local particles. In this manner each particle is essentially "smoothed" over a finite volume of fixed mass, and in this way these so-called Smoothed Particle Hydrodynamics or SPH codes are naturally adaptive with density. Although SPH was originally developed by the astrophysics community, it too has found uses and applications in a much wider range of fields. In engineering it has been applied to dam breaks and atomised oil lubrication flows, while a number of physics engines in computer games use SPH as a basis. The community has grown to the point where there is now a Europe-wide network of users called SPHERIC -

the SPH European Research Interest Community[1]. This aims to share advances in code development across the user community, and to prevent the re-invention of the wheel when it comes the solution of known problems.

Each of these approaches has advantages and disadvantages with respect to the other. Generally speaking, AMR codes have a higher resolution for a given number of grid cells than an SPH code with an equal number of particles. Furthermore, they can be made to adapt to any flow parameter (although this is not always trivial!), while SPH adapts primarily with density only. On the other hand, SPH naturally handles vacuum boundary conditions, whilst large grids are required with AMR codes to prevent the flow disappearing from the edge of the computational domain. As SPH is a Lagrangian method, advection of flow properties is inherent, whereas this presents problems for AMR codes, and which usually entails an unphysical increase in entropy. In a similar manner, SPH codes can be implemented in such a manner that they are inherently conservative of mass, momentum and energy, and similarly, unless it is explicitly added in shocks, they likewise conserve entropy. Nonetheless it is emphatically **not** true to say that either SPH or grid code methods are "better" than the other, simply that the more appropriate approach should be chosen for any given problem, and indeed greater confidence in the results will ensue if the two methods concur.

Having said that, throughout this chapter I shall however consider only the SPH approach, as it is this one that I have used to generate all the results discussed henceforth. Furthermore, as all the problems I have considered have been fully three-dimensional, throughout this chapter I shall consider only the derivation and discussion of SPH in 3D. This chapter is therefore structured as follows: In Section 3.2 I shall introduce the basic concepts of the SPH method, then in Section 3.3 this is used to re-write the fluid equations in a manner that can be solved numerically. Section 3.4 discusses the dissipative processes required for the correct implementation of artificial viscosity and the introduction of entropy in shock waves, and then in Section 3.5 I discuss how the SPH formalism may be made more adaptive still by the self-consistent inclusion of variable smoothing lengths. As many astrophysical problems are strongly influenced by gravitational forces I detail how these may be implemented in Section 3.6. In Section 3.7 I briefly summarise the methods used to find the nearest neighbours, and then in Section 3.8 I consider how the code is evolved forward in time, and various time-stepping criteria. Finally in Section 3.9 I

[1] http://wiki.manchester.ac.uk/spheric/index.php/SPHERIC_Home_Page

briefly outline the properties of the code I have used, point the reader in the direction of some standard numerical tests used for code evaluation and consider further extensions to the method.

3.2 SPH Basics

In the following section I shall discuss the derivation of the SPH formalism from first principles, showing how a continuous field can be mapped on to (and thus approximated by) a series of discrete particles, and the errors involved in this approximation. I then show how derivatives may be calculated, and discuss ways in which the particles may be suitably smoothed to represent the field.

3.2.1 Discrete Approximations to a Continuous Field

We start from the (mathematically) trivial identity

$$f(\mathbf{r}) = \int_V f(\mathbf{r}')\ \delta(\mathbf{r} - \mathbf{r}')\ \mathrm{d}\mathbf{r}', \tag{3.1}$$

where $f(\mathbf{r})$ is any (scalar) function defined on a three-dimensional co-ordinate system $\mathbf{r}$ ranging over a volume V. Similarly, $\delta(\mathbf{r})$ is the Dirac delta function, and $\mathbf{r}'$ is a dummy variable also ranging over V.

We may generalise the delta function to a so-called smoothing kernel W with a characteristic width h (known as the smoothing length) such that

$$\lim_{h\to 0} W(\mathbf{r}, h) = \delta(\mathbf{r}), \tag{3.2}$$

subject to the normalisation

$$\int_V W(\mathbf{r}, h)\ \mathrm{d}\mathbf{r}' = 1. \tag{3.3}$$

By expanding $W(\mathbf{r} - \mathbf{r}', h)$ as a Taylor series, it can be shown that for symmetric kernels $W(\mathbf{r} - \mathbf{r}', h) = W(\mathbf{r}' - \mathbf{r}, h)$, equation 3.1 becomes

$$f(\mathbf{r}) = \int_V f(\mathbf{r}')\ W(\mathbf{r} - \mathbf{r}', h)\ \mathrm{d}\mathbf{r}' + \mathcal{O}(h^2), \tag{3.4}$$

the second order accuracy arising from the vanishing of the kernel gradient at $\mathbf{r}' = \mathbf{r}$ (see for instance Price 2005; Benz 1990; Monaghan 1992). Note that more elaborate kernels accurate to $\mathcal{O}(h^4)$ can be constructed, but these suffer from the problem that $W(\mathbf{r}, h)$ can become negative in certain ranges (Price, 2005; Monaghan, 1992), thus potentially leading to negative density evaluations in certain pathological situations.

Nonetheless for a second order, symmetric kernel, for any finite density $\rho(\mathbf{r})$ within V, equation 3.4 is exactly equivalent to

$$f(\mathbf{r}) = \int_V \frac{f(\mathbf{r}')}{\rho(\mathbf{r}')} W(\mathbf{r} - \mathbf{r}', h)\ \rho(\mathbf{r}')\mathrm{d}\mathbf{r}' + \mathcal{O}(h^2). \tag{3.5}$$

Discretising this continuous field on to a series of particles of (potentially variable) mass $m = \rho(\mathbf{r}')\mathrm{d}\mathbf{r}'$, the original identity equation 3.1 becomes

$$f(\mathbf{r}) \approx \sum_i \frac{m_i}{\rho_i} f(\mathbf{r}_i)\ W(\mathbf{r} - \mathbf{r}_i, h), \tag{3.6}$$

where now $f(\mathbf{r}_i)$, m_i and $\rho_i = \rho(\mathbf{r}_i)$ are the scalar value, mass and density of the i^{th} particle, and i ranges over all particles within the smoothing kernel. Equation 3.6 therefore represents the discrete approximation to the continuous scalar field f at position $\mathbf{r}$ in the computational domain V, and is thus the basis of all SPH formalisms. Note that the position $\mathbf{r}$ at which the function f is approximated is completely general and is not restricted to particle positions, although in practice this is where the values are actually evaluated.

3.2.2 Spatial Derivatives and Vector Calculus

In order for the SPH discretisation of a field to be useful as a method of solving fluid flows, it is clear that the spatial derivatives of any given quantity must also have a suitable approximate form[2]. Here therefore, I summarise the SPH approximations for various vector calculus quantities.

3.2.2.1 Gradient of a Scalar Field

The approximation for the gradient of a scalar field can be derived by taking the spatial derivative of equation 3.1, and applying the smoothing kernel. Noting that

[2]Temporal derivatives are naturally also required, and these will be discussed in due course

$\nabla \equiv \partial/\partial\mathbf{r}$, we therefore see that

$$\nabla f(\mathbf{r}) = \frac{\partial}{\partial \mathbf{r}} \int_V f(\mathbf{r}')\ \delta(\mathbf{r}-\mathbf{r}')\ \mathrm{d}\mathbf{r}' \tag{3.7}$$

$$= \frac{\partial}{\partial \mathbf{r}} \int_V f(\mathbf{r}')\ W(\mathbf{r}-\mathbf{r}', h)\ \mathrm{d}\mathbf{r}' + \mathcal{O}(h^2), \tag{3.8}$$

in a similar manner to equation 3.4. Given that the only part to depend on $\mathbf{r}$ is the smoothing kernel W, and again introducing the density $\rho(\mathbf{r}')$ in both the numerator and the denominator we obtain

$$\nabla f(\mathbf{r}) = \int_V \frac{f(\mathbf{r}')}{\rho(\mathbf{r}')} \frac{\partial}{\partial \mathbf{r}} W(\mathbf{r}-\mathbf{r}', h)\rho(\mathbf{r}')\ \mathrm{d}\mathbf{r}' + \mathcal{O}(h^2). \tag{3.9}$$

Finally this may be discretised in the same way as before, to give

$$\nabla f(\mathbf{r}) \approx \sum_i \frac{m_i}{\rho_i} f(\mathbf{r}_i)\ \nabla W(\mathbf{r}-\mathbf{r}_i, h) \tag{3.10}$$

as an estimator for the gradient of a scalar field $f(\mathbf{r})$. Notable from the above result is that the gradient of a scalar field can be approximated by the values of the field itself along with the gradient of the kernel. Computationally this is very useful as at no point does ∇f have to be evaluated for any particle, whilst the gradient of the kernel will be known explicitly for any sensible choice of W.

3.2.2.2 Divergence of a Vector Field

Although equation 3.1 was given only for a scalar field, a similar identity may be given for a vector field $\mathbf{F}(\mathbf{R})$, namely

$$\mathbf{F}(\mathbf{r}) = \int_V \mathbf{F}(\mathbf{r}')\ \delta(\mathbf{r}-\mathbf{r}')\ \mathrm{d}\mathbf{r}', \tag{3.11}$$

Taking the divergence of this with respect to $\mathbf{r}$, and noting once again that the only term to depend on $\mathbf{r}$ is the smoothing kernel we find that the integral approximation becomes

$$\nabla \cdot \mathbf{F}(\mathbf{r}) = \int_V \mathbf{F}(\mathbf{r}') \cdot \nabla W(\mathbf{r}-\mathbf{r}', h)\mathrm{d}\mathbf{r}' + \mathcal{O}(h^2), \tag{3.12}$$

and thus as before this can be discretised to obtain the approximation

$$\nabla \cdot \mathbf{F}(\mathbf{r}) \approx \sum_i \frac{m_i}{\rho_i} \mathbf{F}(\mathbf{r}_i) \cdot \nabla W(\mathbf{r} - \mathbf{r}_i, h). \tag{3.13}$$

3.2.2.3 Curl of a Vector Field

By a precisely similar argument, it is possible to show that the curl of a vector $\mathbf{F}$, $\nabla \times \mathbf{F}$ can be approximated using

$$\nabla \times \mathbf{F}(\mathbf{r}) \approx \sum_i \frac{m_i}{\rho_i} \mathbf{F}(\mathbf{r}_i) \times \nabla W(\mathbf{r} - \mathbf{r}_i, h), \tag{3.14}$$

although this is relatively little used unless magnetohydrodynamic (MHD) effects are being taken into account.

3.2.3 Errors

The approximations given in equations 3.6, 3.10, 3.13 and 3.14 encompass both the $\mathcal{O}(h^2)$ errors of considering only the integral term, and also the errors inherent in the discretisation (which arise due to incomplete sampling of the smoothing kernel). In the former case we see that the $\mathcal{O}(h^2)$ errors are reduced by decreasing the smoothing length, while the discretisation (sampling) errors are minimised by increasing the number of particles within the smoothing kernel. Barring numerical stability issues (Read et al., 2010), this discrete approximation is therefore at its most accurate with large numbers of particles contained within a small smoothing length. However, this must be balanced against the need for computational speed and efficiency, and hence there is a compromise to be struck.

These errors are neatly illustrated by considering the approximations to a constant function $f(\mathbf{r}) \equiv 1$ and the zero function, which can be obtained by noting that with this definition of f, $\nabla f(\mathbf{r}) = 0$. The SPH approximations for one and zero therefore become

$$1 \approx \sum_i \frac{m_i}{\rho_i} W(\mathbf{r} - \mathbf{r}_i, h), \tag{3.15}$$

$$0 \approx \sum_i \frac{m_i}{\rho_i} \nabla W(\mathbf{r} - \mathbf{r}_i, h). \tag{3.16}$$

Since in neither case does the equation reduce to an identity, we see that there are

inherent errors in estimating even constant functions. Nonetheless, with suitable choices for the number of particles within the smoothing kernel and the smoothing length, these may be kept to an acceptable level. For a more detailed derivation and discussion of these errors, the reader is directed to Price (2005); Monaghan (1992); Benz (1990) and Read et al. (2010).

3.2.4 Improved Approximations for Spatial Gradients

Although the approximations given in equations 3.6, 3.10, 3.13 and 3.14 are those that arise most readily from the SPH approximation, it is possible to construct other estimators for the gradient of a scalar field. For instance, by noting that for any quantity $f(\mathbf{r}) \equiv 1.f(\mathbf{r})$, we see that

$$\nabla f(\mathbf{r}.1) = 1.\nabla f(\mathbf{r}) + f(\mathbf{r})\nabla 1 \tag{3.17}$$

and therefore that

$$\nabla f(\mathbf{r}) = \nabla f(\mathbf{r}) - f(\mathbf{r})\nabla 1. \tag{3.18}$$

Clearly, since $\nabla 1 = 0$ these forms should be identical. From equation 3.16 however, we see that the SPH approximation for $\nabla 1$ is non-zero, and thus using equation 3.18 we may define another estimate for $\nabla f(\mathbf{r})$ as

$$\nabla f(\mathbf{r}) = \sum_i \frac{m_i}{\rho_i} f(\mathbf{r}_i)\, \nabla W(\mathbf{r} - \mathbf{r}_i, h) - f(\mathbf{r}) \sum_i \frac{m_i}{\rho_i}\, \nabla W(\mathbf{r} - \mathbf{r}_i, h) \tag{3.19}$$

$$= \sum_i m_i \frac{f(\mathbf{r}_i) - f(\mathbf{r})}{\rho_i}\, \nabla W(\mathbf{r} - \mathbf{r}_i, h). \tag{3.20}$$

This approximation clearly has the advantage that it vanishes identically for constant functions.

A more general class of interpolants arises from considering the vector calculus identity

$$\nabla(f\rho^n) \equiv nf\rho^{n-1}\nabla\rho + \rho^n \nabla f, \tag{3.21}$$

valid for all $n \in \mathbb{R}$. This in turn leads to the following identity for ∇f

$$\nabla f \equiv \frac{1}{\rho^n}\left[\nabla(f\rho^n) - nf\rho^{n-1}\nabla\rho\right]. \tag{3.22}$$

Substituting ρ and $f\rho^n$ into equation 3.10, we obtain a general interpolant for ∇f,

such that

$$\nabla f(\mathbf{r}) = \frac{1}{\rho(\mathbf{r})^n} \sum_i m_i \left(f(\mathbf{r}_i)\rho(\mathbf{r}_i)^{n-1} - n f(\mathbf{r})\rho(\mathbf{r})^{n-1} \right) \nabla W(\mathbf{r} - \mathbf{r}_i, h). \tag{3.23}$$

Two instances of this general case turn out to be particularly useful, namely where $n = 1$ and $n = -1$. For the former case we obtain

$$\nabla f(\mathbf{r}) = \frac{1}{\rho(\mathbf{r})} \sum_i m_i \left(f(\mathbf{r}) - f(\mathbf{r}_i) \right) \nabla W(\mathbf{r} - \mathbf{r}_i, h). \tag{3.24}$$

This is very similar in form to that given in equation 3.20, with the exception that knowledge of the density at $\mathbf{r}$ is required *a priori*. Although no longer anti-symmetric in $f(\mathbf{r})$ and f_i, it is nonetheless exact for constant functions.

In the case where $n = -1$ we obtain

$$\nabla f(\mathbf{r}) = \rho(\mathbf{r}) \sum_i m_i \left(\frac{f(\mathbf{r})}{\rho(\mathbf{r})^2} + \frac{f(\mathbf{r}_i)}{\rho(\mathbf{r}_i)^2} \right) \nabla W(\mathbf{r} - \mathbf{r}_i, h). \tag{3.25}$$

While this form is no longer exact for constant functions, it is commonly used as an estimator for the pressure gradient $(\nabla P)/\rho$, as it is pairwise symmetric and as such ensures conservation of momentum. This is also the form of the gradient that arises naturally from a Lagrangian formulation of the fluid equations, as I shall show in Section 3.3.2.

3.2.5 Improved Divergence Estimates

In a similar manner to the gradient, improved estimates can be made for the divergence of a vector field. By noting that $\mathbf{F}(\mathbf{r}) = 1.\mathbf{F}(\mathbf{r})$, the estimate

$$\nabla \cdot \mathbf{F}(\mathbf{r}) \approx \sum_i \frac{m_i}{\rho_i} \left(\mathbf{F}(\mathbf{r}_i) - \mathbf{F}(\mathbf{r}) \right) \cdot \nabla W(\mathbf{r} - \mathbf{r}_i, h). \tag{3.26}$$

can be arrived at, which again becomes exact for constant functions. In a similar manner to the expansion given in equation 3.21, a general class of estimates can be arrived at by considering the identity

$$\nabla \cdot (\rho^n \mathbf{F}) = \rho^n \nabla \cdot \mathbf{F} + n\rho^{n-1} \mathbf{F} \cdot \nabla \rho, \tag{3.27}$$

the $n = 1, -1$ cases of which are given by

$$\nabla \cdot \mathbf{F}(\mathbf{r}) \approx \frac{1}{\rho(\mathbf{r})} \sum_i m_i \left(\mathbf{F}(\mathbf{r}_i) - \mathbf{F}(\mathbf{r})\right) \cdot \nabla W(\mathbf{r} - \mathbf{r}_i, h) \tag{3.28}$$

and

$$\nabla \cdot \mathbf{F}(\mathbf{r}) \approx \rho(\mathbf{r}) \sum_i m_i \left(\frac{\mathbf{F}(\mathbf{r}_i)}{\rho(\mathbf{r}_i)^2} + \frac{\mathbf{F}(\mathbf{r})}{\rho(\mathbf{r})^2}\right) \cdot \nabla W(\mathbf{r} - \mathbf{r}_i, h) \tag{3.29}$$

respectively. Once again these estimates have the advantages of being exact for constant functions in the former case and pairwise symmetric in $(\nabla \cdot \mathbf{F})/\rho$ in the latter case.

3.2.6 Smoothing Kernels

From the above it is clear that the choice of smoothing kernel is an important one. It must by definition obey the criteria set out in equations 3.2 and 3.3 in that it must tend to a δ-function as $h \to 0$ and it must be normalised so the area under the curve is unity. For the purposes of calculating the gradients of quantities it is also clear that it should have a continuous and well defined first derivative, and from a symmetry argument it should be spherically symmetric, and thus depend only on $r = |\mathbf{r} - \mathbf{r}'|$ and h.

One of the first choices for the smoothing kernel was the Gaussian function, such that

$$W(r, h) = \frac{1}{h^3 \pi^{3/2}} e^{-x^2}, \tag{3.30}$$

where $x = r/h$. However, this has the drawback that $W > 0$ for all r, and thus all particles within the computational domain contribute. The computational cost of such a kernel therefore scales as $\mathcal{O}(N^2)$, where N is the number of particles in the simulation. Given that (for purely hydrodynamical quantities) long range forces are negligible, it makes sense to restrict the kernel to those with compact support, i.e. make them subject to the condition that $W(r, h) = 0$ where $r/h > k$ for some constant k. This means that the computational cost scales as $\mathcal{O}(NN_{\text{neigh}})$, where N_{neigh} is the average number of particles within a sphere of radius $r = kh$ about any one particle.

For this reason, cubic spline kernels are often used (see Monaghan & Lattanzio

1985 for instance), where the kernel is defined as

$$W(r,h) = \frac{1}{\pi h^3} \begin{cases} 1 - \frac{3}{2}x^2 + \frac{3}{4}x^3 & 0 \leq x \leq 1; \\ \frac{1}{4}(2-x)^3 & 1 \leq x \leq 2; \\ 0 & x \geq 2, \end{cases} \tag{3.31}$$

where $x = r/h$ as in equation 3.30. Here only particles within $2h$ of the central particle contribute to the smoothing kernel, which is spherically symmetric and smoothly differentiable for all r. Although many other kernels are possible (see Monaghan 1992; Fulk & Quinn 1996; Price 2005; Read et al. 2010 for example) this is a commonly used kernel, and is the one present in the code I have used throughout.

Note from the above the gradient of the kernel is well defined for all values of x, such that

$$\nabla W(r,h) = \frac{\partial}{\partial r} W(r,h) \tag{3.32}$$

$$= \frac{1}{\pi h^4} \begin{cases} \frac{9}{4}x^2 - 3x & 0 \leq x \leq 1 \\ -\frac{3}{4}(2-x)^2 & 1 \leq x \leq 2 \\ 0 & x \geq 2 \end{cases} . \tag{3.33}$$

Finally it is worth noting that in general the form of the kernel makes little overall difference to the computational speed of the code. This is because most codes tabulate the values of both the kernel and its gradient rather than compute them directly, and thus the form of the kernel may be as simple or as complex as required, even (theoretically at least) to the extent of being non-analytic functions.

3.3 Fluid Equations

Given that the SPH formalism has now been put on a sound mathematical footing, in this section I shall use it to obtain approximations to the equations governing fluid motion, such that they can be used to construct a viable numerical algorithm for solving fluid flows. For the dual purposes of brevity and simplicity I shall here consider only the case of an inviscid compressible flow in the absence of body forces, although the inclusion of both gravity and (artificial) viscosity will be discussed in due course. First however, it is useful to summarise the principal equations of

motion in their standard conservative form.

The continuity (conservation of mass) equation is given by

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0 \tag{3.34}$$

where as normal, ρ is the density, t is time and $\mathbf{v}$ is velocity.

The Euler equation gives the equations of motion in the case of an inviscid fluid, and encapsulates the conservation of momentum. In the absence of external (body) forces it becomes

$$\frac{\partial \rho \mathbf{v}}{\partial t} + \nabla \cdot (\rho \mathbf{v} \otimes \mathbf{v}) + \nabla P = \mathbf{0}, \tag{3.35}$$

where P is the fluid pressure and $\otimes$ represents the outer or tensor product[3]. For compressible flows it is also necessary to take into account the energy equation, and as such the conservation of energy is embodied in the following equation;

$$\frac{\partial u}{\partial t} + \nabla \cdot [(u + P)\mathbf{v}] = 0, \tag{3.36}$$

where u is the specific internal energy and $v = |\mathbf{v}|$ is the magnitude of the velocity vector. Finally it is worth noting that these five equations (there are 3 components to the momentum equation) contain six unknowns (ρ, the three components of velocity v_x, v_y and v_z, P and u). Therefore in order to solve the system we require a further constraint; an equation of state is required. All the analysis I shall present henceforth uses the ideal gas equation of state, where

$$P = \kappa(s)\rho^\gamma, \tag{3.37}$$

$$= (\gamma - 1)u\rho, \tag{3.38}$$

where γ is the adiabatic index (the ratio of specific heats), which throughout has been set to 5/3, and $\kappa(s)$ is the adiabat, itself a function of the specific entropy s. In the case of isentropic flows, s and (thus κ) remains constant.

I shall now discuss the SPH formulation of each of the continuity, momentum and energy equations in turn. Note that again for the purposes of brevity I assume that the smoothing length is held constant (i.e. $\dot{h} = 0$, where the dot denotes the derivative with respect to time), and is equal for all particles. Individual, variable smoothing lengths will be discussed in due course. Furthermore, I assume through-

[3]The outer product of two vectors may be summarised as $\mathbf{A} \otimes \mathbf{B} = \mathbf{A}\mathbf{B}^T = A_i B_j$ (in indicial notation).

out that the mass of each particle is held constant, such that $m_i = const$, and again that all particles are of equal mass. Although it is possible to have individually varying particle masses, the code I use does not have this feature, and therefore I have not included a discussion of it here. Finally note that from here onwards, all the approximations are evaluated at specific particle positions, as this is how the SPH algorithm is implemented within particle-based codes.

3.3.1 Conservation of Mass

Using equation 3.6, we see that in the case of the density, the SPH approximation becomes very simple, namely that at particle j the density ρ_j becomes

$$\begin{aligned} \rho_j &= \sum_i m_i W(\mathbf{r}_j - \mathbf{r}_i, h), \\ &= \sum_i m_i W_{ji}, \end{aligned} \tag{3.39}$$

where we write $W_{ji} = W(\mathbf{r}_j - \mathbf{r}_i, h)$, and where by symmetry, $W_{ji} = W_{ij}$. Note that here and henceforth, as the SPH formalism is a discrete approximation to the underlying continuous medium, we assume equality between the estimator on the RHS and the SPH quantity on the LHS.

Taking the full time derivative of equation 3.39 we obtain

$$\frac{\mathrm{d}\rho_j}{\mathrm{d}t} = \sum_i m_i \left[\frac{\partial W_{ji}}{\partial \mathbf{r}_j} \cdot \frac{\mathrm{d}\mathbf{r}_j}{\mathrm{d}t} + \frac{\partial W_{ji}}{\partial \mathbf{r}_i} \cdot \frac{\mathrm{d}\mathbf{r}_i}{\mathrm{d}t} + \frac{\partial W_{ji}}{\partial h} \frac{\mathrm{d}h}{\mathrm{d}t} \right], \tag{3.40}$$

and noting that

$$\frac{\mathrm{d}\mathbf{r}_j}{\mathrm{d}t} = \mathbf{v}_j, \quad \frac{\mathrm{d}\mathbf{r}_i}{\mathrm{d}t} = \mathbf{v}_i, \quad \frac{\mathrm{d}h}{\mathrm{d}t} = 0,$$

we find that the time derivative of density becomes

$$\begin{aligned} \frac{\mathrm{d}\rho_j}{\mathrm{d}t} &= \sum_i m_i \left(\mathbf{v}_j \cdot \nabla_j W_{ji} + \mathbf{v}_i \cdot \nabla_i W_{ji} \right) \\ &= \sum_i m_i \, \mathbf{v}_{ji} \cdot \nabla_j W_{ji} \end{aligned} \tag{3.41}$$

where we use $\mathbf{v}_{ji} = \mathbf{v}_j - \mathbf{v}_i$, and where we note that the gradient of the kernel is antisymmetric, i.e. that

$$\nabla_i W_{ji} = -\nabla_j W_{ji}. \tag{3.42}$$

From equation 3.28, we note that the RHS of equation 3.41 is simply an estimator of $-\rho_j \nabla_j \cdot \mathbf{v}_j$. Hence equation 3.41 becomes

$$\frac{\mathrm{d}\rho_j}{\mathrm{d}t} = -\rho_j \nabla_j \cdot \mathbf{v}_j, \tag{3.43}$$

which is simply a reformulation of the continuity equation equation 3.34 using the Lagrangian time derivative

$$\frac{\mathrm{d}}{\mathrm{d}t} = \frac{\partial}{\partial t} + (\mathbf{v} \cdot \nabla), \tag{3.44}$$

in which the second term accounts for the advection of flow properties through the fluid. Therefore we see that the SPH estimate for density equation 3.39 is *automatically* conservative of mass (as long as equation 3.28 is used as an estimate for the divergence of velocity).

3.3.2 Conservation of Momentum

Although there are various ways of deriving the equations of motion consistently with the SPH framework, a particularly appealing one is to use the Lagrangian formalism. As long as the discrete Lagrangian functional preserves the fundamental symmetries of the underlying continuous one, this confers the inherent advantages that the resulting SPH equations of motion will automatically fulfil the requisite conservation laws (through Noether's Theorem) and also that the only approximations made are in the discretisation of the Lagrangian itself.

3.3.2.1 Linear Momentum

Defined as the total kinetic energy of the system minus the total internal energy (for purely hydrodynamical flows), the Lagrangian functional $\mathcal{L}$ for the fluid is

$$\mathcal{L}(\mathbf{r}, \mathbf{v}) = \int_V \frac{1}{2}\rho \mathbf{v} \cdot \mathbf{v} - \rho u \, \mathrm{d}\mathbf{r}, \tag{3.45}$$

where as before, u is the specific internal energy. For later simplicity, we note that through the equation of state (equation 3.38) the specific internal energy is a function of density and pressure $u = u(\rho, P)$, which in turn are functions of position. This gives $u = u(\mathbf{r})$. Now if we again make the discretisation $m_i = \rho \mathrm{d}\mathbf{r}$, the SPH

estimate of the Lagrangian becomes

$$\mathcal{L}(\mathbf{r},\mathbf{v}) = \sum_i m_i \left(\frac{1}{2}\mathbf{v}_i \cdot \mathbf{v}_i - u_i(\mathbf{r}_i)\right), \tag{3.46}$$

where i ranges over all particles.

The equations of motion for particle j are obtained from the Lagrangian through the Euler-Lagrange equations, as follows;

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial \mathcal{L}}{\partial \mathbf{v}_j}\right) - \frac{\partial \mathcal{L}}{\partial \mathbf{r}_j} = \mathbf{0}. \tag{3.47}$$

By considering each of the terms in this equation it is therefore possible to obtain an SPH approximation to the equations of motion that remains fully conservative. If we therefore consider the derivative of the Lagrangian with respect to the velocity at particle j, we find

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{v}_j} &= \frac{\partial}{\partial \mathbf{v}_j}\sum_i m_i \left(\frac{1}{2}\mathbf{v}_i \cdot \mathbf{v}_i - u_i(\mathbf{r}_i)\right), \\ &= m_j \mathbf{v}_j, \end{aligned} \tag{3.48}$$

noting that since the velocities are independent the differential is zero unless $i = j$.

Considering now the second term in the Euler-Lagrange equation 3.47 we find that

$$\frac{\partial \mathcal{L}}{\partial \mathbf{r}_j} = -\sum_i \left[\frac{\partial u_i}{\partial P_i}\frac{\partial P_i}{\partial \mathbf{r}_j} + \frac{\partial u_i}{\partial \rho_i}\frac{\partial \rho_i}{\partial \mathbf{r}_j}\right], \tag{3.49}$$

where we have used equation 3.38 to obtain the full derivative of the internal energy. In the isentropic (dissipationless) case we see that $\kappa(s)$ is constant, and thus the pressure is a function of density only, leading to

$$\frac{\partial \mathcal{L}}{\partial \mathbf{r}_j} = -\sum_i \left[\frac{\partial u_i}{\partial P_i}\frac{\mathrm{d}P_i}{\mathrm{d}\rho_i} + \frac{\partial u_i}{\partial \rho_i}\right]\frac{\partial \rho_i}{\partial \mathbf{r}_j}. \tag{3.50}$$

From the equation of state equation 3.38 we find that

$$\frac{\partial u_i}{\partial P_i}\frac{\mathrm{d}P_i}{\mathrm{d}\rho_i} + \frac{\partial u_i}{\partial \rho_i} = \frac{P_i}{\rho_i^2}, \tag{3.51}$$

and thus the derivative of the Lagrangian with respect to the position of particle j

becomes

$$\frac{\partial \mathcal{L}}{\partial \mathbf{r}_j} = -\sum_i \frac{P_i}{\rho_i^2}\frac{\partial \rho_i}{\partial \mathbf{r}_j}. \tag{3.52}$$

Using equation 3.39 we find that

$$\begin{aligned} \frac{\partial \rho_i}{\partial \mathbf{r}_j} &= \sum_k m_i \frac{\partial W_{ik}}{\partial \mathbf{r}_j} & (3.53)\\ &= \sum_k m_i \frac{\partial W_{ik}}{\partial r_{ik}}\frac{\partial r_{ik}}{\partial \mathbf{r}_j}, & (3.54) \end{aligned}$$

where we take $r_{ik} = |\mathbf{r}_{ik}|$, and use the fact that the kernel is spherically symmetric. By direct differentiation,

$$\frac{\partial r_{ik}}{\partial \mathbf{r}_j} = (\delta_{ij} - \delta_{kj})\hat{\mathbf{r}}_{ik}, \tag{3.55}$$

with $\hat{\mathbf{r}}_{ik} = \mathbf{r}_{ik}/r_{ik}$ the unit vector in the direction of $\mathbf{r}_{ik}$. Substituting this back into equation 3.54 we find that

$$\begin{aligned} \frac{\partial \rho_i}{\partial \mathbf{r}_j} &= \sum_k m_k \frac{\partial W_{ik}}{\partial r_{ik}}(\delta_{ij} - \delta_{kj})\hat{\mathbf{r}}_{ik} & (3.56)\\ &= \sum_k m_k \nabla_j W_{ik}\ (\delta_{ij} - \delta_{kj})\,, & (3.57) \end{aligned}$$

where in the second case we have used the fact that $\partial/\partial \mathbf{r}_j \equiv \nabla_j$.

With reference to equation 3.52 we find therefore that

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{r}_j} &= -\sum_i m_i \frac{P_i}{\rho_i^2}\sum_k m_k \nabla_j W_{ik}\,(\delta_{ij} - \delta_{kj}) & (3.58)\\ &= -m_j \frac{P_j}{\rho_j^2}\sum_k m_k \nabla_j W_{jk} - \sum_i m_i m_j \frac{P_i}{\rho_i^2}\nabla_j W_{ij} & (3.59)\\ &= -m_j \sum_i m_i \left(\frac{P_j}{\rho_j^2} + \frac{P_i}{\rho_i^2}\right)\nabla_j W_{ji}, & (3.60) \end{aligned}$$

where we have changed the summation index in the first term to i and used the fact that the gradient of the kernel is antisymmetric, i.e. that $\nabla_j W_{kj} = -\nabla_j W_{jk}$. Finally, by substituting equations 3.48 and 3.60 into equation 3.47 and dividing through by the common factor m_j, we find that the SPH equations of motion become

$$\frac{\mathrm{d}\mathbf{v}_j}{\mathrm{d}t} = -\sum_i m_i \left(\frac{P_j}{\rho_j^2} + \frac{P_i}{\rho_i^2}\right)\nabla_j W_{ji}. \tag{3.61}$$

Since this equation is pairwise symmetric in i, j, it is clear that the pressure force on particle j due to particle i is equal and opposite (due to the antisymmetry of the kernel gradient) to the force on particle i from particle j. In this manner, it is clear that this formulation of the equation of motion conserves linear momentum by construction.

3.3.2.2 Angular Momentum

To check that angular momentum $\mathbf{L} = \mathbf{r} \times m\mathbf{v}$ is conserved, we note that its derivative with respect to time should be zero. By using equation 3.61 we see that the time derivative of the angular momentum of particle j is given by

$$\frac{\mathrm{d}\mathbf{L}_j}{\mathrm{d}t} = m_j \mathbf{v}_j \times \mathbf{v}_j + m_j \mathbf{r}_j \times \frac{\mathrm{d}\mathbf{v}_j}{\mathrm{d}t} \tag{3.62}$$

$$= -m_j \sum_i m_i \left(\frac{P_j}{\rho_j^2} + \frac{P_i}{\rho_i^2} \right) \mathbf{r}_j \times \nabla_i W_{ij}, \tag{3.63}$$

since by definition $\mathbf{v}_j \times \mathbf{v}_j = \mathbf{0}$. The total time derivative of the angular momentum is therefore given by the sum over all particles j, such that

$$\frac{\mathrm{d}\mathbf{L}}{\mathrm{d}t} = -\sum_j \sum_i m_j m_i \left(\frac{P_j}{\rho_j^2} + \frac{P_i}{\rho_i^2} \right) \mathbf{r}_j \times \nabla_i W_{ij}. \tag{3.64}$$

Hence we see that by reversing the summation indices the entire sum is *anti*symmetric in i and j, i.e.

$$\frac{\mathrm{d}\mathbf{L}}{\mathrm{d}t} = -\sum_j \sum_i m_j m_i \left(\frac{P_j}{\rho_j^2} + \frac{P_i}{\rho_i^2} \right) \mathbf{r}_j \times \nabla_i W_{ij}, \tag{3.65}$$

$$= \sum_i \sum_j m_i m_j \left(\frac{P_i}{\rho_i^2} + \frac{P_j}{\rho_j^2} \right) \mathbf{r}_i \times \nabla_j W_{ji}, \tag{3.66}$$

which can only be the case where the total sum is zero. Hence the angular momentum is constant with time, and thus angular momentum is explicitly conserved.

3.3.3 Conservation of Energy

In the case of a purely hydrodynamical flow, the total energy $E = \rho u + \rho v^2/2$ is given by the sum of the kinetic and internal energies, such that the SPH estimator

becomes

$$E = \sum_i m_i \left(\frac{1}{2} \mathbf{v}_i \cdot \mathbf{v}_i + u_i \right). \tag{3.67}$$

Clearly, where energy is conserved the time derivative of the total energy should be zero. Taking the time derivative therefore, we find that

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \sum_i m_i \left(\mathbf{v}_i \cdot \frac{\mathrm{d}\mathbf{v}_i}{\mathrm{d}t} + \frac{\partial u_i}{\partial P_i} \frac{\mathrm{d}P_i}{\mathrm{d}t} + \frac{\partial u_i}{\partial \rho_i} \frac{\mathrm{d}\rho_i}{\mathrm{d}t} \right), \tag{3.68}$$

$$= \sum_i m_i \left(\mathbf{v}_i \cdot \frac{\mathrm{d}\mathbf{v}_i}{\mathrm{d}t} + \frac{P_i}{\rho_i^2} \frac{\mathrm{d}\rho_i}{\mathrm{d}t} \right), \tag{3.69}$$

where we have again used the fact that in the dissipationless case $P = P(\rho)$ and we can therefore amalgamate the latter two terms of the RHS of equation 3.68 using equation 3.51. Using also the equation of motion derived above in equation 3.61 and $\mathrm{d}\rho/\mathrm{d}t$ from the continuity equation 3.41 we therefore find that

$$\frac{\mathrm{d}E}{\mathrm{d}t} = -\sum_i m_i \mathbf{v}_i \cdot \sum_j m_j \left(\frac{P_i}{\rho_i^2} + \frac{P_j}{\rho_j^2} \right) \nabla_j W_{ji} + \sum_i m_i \frac{P_i}{\rho_i^2} \sum_j m_j (\mathbf{v}_i - \mathbf{v}_j) \cdot \nabla_j W_{ji} \tag{3.70}$$

$$= \sum_i \sum_j m_i m_j \left(\frac{P_j}{\rho_j^2} \mathbf{v}_i + \frac{P_i}{\rho_i^2} \mathbf{v}_j \right) \cdot \nabla_j W_{ji}, \tag{3.71}$$

where we have again used the fact that the kernel is antisymmetric to obtain equation 3.71. Now using the same argument we used to show that angular momentum is conserved, we note that equation 3.71 is antisymmetric under a reversal of i and j, and thus must be equal to zero. Hence we find that

$$\frac{\mathrm{d}E}{\mathrm{d}t} = 0, \tag{3.72}$$

and therefore that the total energy is also explicitly conserved.

A corollary of this is that the time derivative of the internal energy is given by the second term on the RHS of equation 3.69, such that

$$\frac{\mathrm{d}u_j}{\mathrm{d}t} = \frac{P_j}{\rho_j^2} \frac{\mathrm{d}\rho_j}{\mathrm{d}t}, \tag{3.73}$$

$$= \frac{P_j}{\rho_j^2} \sum_i m_i \left(\mathbf{v}_j - \mathbf{v}_i \right) \cdot \nabla_i W_{ji}, \tag{3.74}$$

and indeed this is how the internal energy is evolved within SPH codes.

It is worth noting that the formulation of SPH outlined above is therefore explicitly conservative of mass, momentum (in both forms) and energy. Hence, while there are inevitably errors inherent in the SPH discretisation of a continuous medium, these are the *only* errors that appear, at least in the case of a dissipationless hydrodynamical flow.

3.4 Dissipative Effects

So far we have assumed the fluid flow to be barotropic (i.e. $P = P(\rho)$), and polytropic, with the polytropic index set equal to the adiabatic index γ, the ratio of specific heats. This in turn means that the flow is isentropic, and therefore completely dissipationless. While this is an adequate approximation for many incompressible, inviscid and unshocked compressible flows, it presents serious problems when it comes to modelling transonic and supersonic flow regimes, as the conversion of mechanical (kinetic) energy into heat (internal) energy is not correctly captured. The problem occurs because at a shock front, flow properties such as the velocity, pressure, density and entropy change very rapidly, on the order of the mean free path of the gas particles. On large scales therefore these changes appear discontinuous, and flow solvers that do not resolve the mean free path (which is all of them) break down due to the apparently singular flow gradients.

There are two principal workarounds that allow numerical codes to solve shocked flows. One is to use a Riemann solver in a Gudonov-type code (see for instance Inutsuka 2002; Cha & Whitworth 2003), but I shall not go into any detail here as this is not the approach used in the code I have used. The alternative approach, used in the majority of SPH codes, is to broaden the shock across a small number of smoothing lengths. This ensures that the flow gradients do *not* become infinite, and gives the correct asymptotic behaviour away from the shock. This latter method is implemented by including an *artificial* dissipative term in the momentum and energy equations that is triggered only in the presence of shocks, and it is this method that I shall consider here.

3.4.1 Standard Artificial Viscosity Prescription

Due to the fact that by construction, shock capturing through a viscous process is an artificial one, there is considerable latitude in the way in which such an artificial viscosity may be implemented. This being said, it must obey the following general rules (von Neumann & Richtmyer, 1950; Rosswog, 2009):

- The flow equations should contain no discontinuities;
- The shock front should be of the order of a few times the smoothing length;
- The artificial viscosity should reduce to zero away from the shock front;
- The Rankine-Hugoniot equations should hold over length scales larger than that over which the shock is smoothed, i.e.

$$\rho_0 v_0 = \rho_1 v_1, \tag{3.75}$$

$$P_0 + \frac{\rho_0 v_0^2}{2} = P_1 + \frac{\rho_1 v_1^2}{2,} \tag{3.76}$$

$$\frac{P_0}{\rho_0} + u_0 + \frac{v_0^2}{2} = \frac{P_1}{\rho_1} + u_1 + \frac{v_1^2}{2}, \tag{3.77}$$

  where the subscripts 0 and 1 refer to pre- and post-shock regions respectively.
- The overall conservation of momentum and energy should not be adversely affected, while the entropy should rise from the pre- to post-shock regions.

By considering the SPH approximation to the momentum equation 3.61 where the force is based on pairwise addition of terms of the form P/ρ^2, on dimensional grounds it seems sensible to consider an artificial viscosity term Π of the form

$$\Pi \propto \frac{v^2}{\rho} \tag{3.78}$$

for some suitable velocity scale v. von Neumann & Richtmyer (1950) suggested a viscous term dependent on the squared velocity divergence (which gives an indication of the local expansion or contraction of the fluid), which translates into SPH form as

$$(\Pi_{ij})_{\mathrm{NR}} = \frac{\beta_{\mathrm{SPH}}\, h^2 |\nabla \cdot \mathbf{v}_{ij}|^2}{\bar{\rho}_{ij}}, \tag{3.79}$$

where h represents a characteristic length scale (in SPH this is equivalent to the smoothing length), $\bar{\rho}_{ij}$ is the average density of particles i and j and β_{SPH} is a

constant term of order unity. Noting that to first order

$$|\nabla \cdot \mathbf{v}_{ij}| = \frac{|\mathbf{v}_{ij}|}{|\mathbf{r}_{ij}|} \tag{3.80}$$

$$\approx \frac{|\mathbf{v}_{ij} \cdot \mathbf{r}_{ij}|}{|\mathbf{r}_{ij}|^2 + \epsilon h^2} \tag{3.81}$$

where as previously $\mathbf{r}_{ij} = \mathbf{r}_i - \mathbf{r}_j$ and where we have added the extra (small) term in the denominator to prevent it becoming singular, this von Neumann-Richtmyer term becomes

$$(\Pi_{ij})_{\mathrm{NR}} = \frac{\beta_{\mathrm{SPH}}\, \mu_{ij}^2}{\bar{\rho}_{ij}} \tag{3.82}$$

where

$$\mu_{ij} = \frac{h\mathbf{v}_{ij} \cdot \mathbf{r}_{ij}}{|\mathbf{r}_{ij}|^2 + \epsilon h^2}. \tag{3.83}$$

By considering the bulk and shear viscosities of a generalised fluid it is possible to obtain a second form of the artificial viscosity, and indeed this has been known for some time (Landshoff, 1930; Landau & Lifshitz, 1959). This form is linear in the velocity divergence and uses the average sound speed[4] $\bar{c}_{\mathrm{s},ij}$ as a second, characteristic velocity component, giving the overall form

$$(\Pi_{ij})_{\mathrm{b}} = -\frac{\alpha_{\mathrm{SPH}}\, \bar{c}_{\mathrm{s},ij}\, \mu_{ij}}{\bar{\rho}_{ij}}, \tag{3.84}$$

where μ_{ij} is as before, and α_{SPH} is a second constant of order unity. Note that the negative on the RHS arises from the requirements that the viscous force component must be non-negative (i.e. $\Pi_{ij} > 0$) and that it should be present only for convergent flows, where $\mathbf{v}_{ij} \cdot \mathbf{r}_{ij} < 0$. In fact these criteria also hold for the von Neumann-Richtmyer form of the viscosity, and therefore in both cases the viscosity is set to zero in expanding flow conditions.

These two forms have different and complementary numerical effects. At low Mach numbers ($\mathcal{M} \lesssim 5$) the linear form performs very well in shock tube tests (Monaghan, 1985), whereas for stronger shocks it fails to prevent inter-particle penetration (Lattanzio et al., 1985). This is an unphysical phenomenon in which the two streams pass through each other at the shock front, leading to the possibility of two particles occupying the same position with differing velocities – a multi-valued velocity field. This possibility can be prevented by using the quadratic form of

[4]Where as usual, the sound speed is defined as $c_{\mathrm{s}}^2 = \mathrm{d}P/\mathrm{d}\rho$.

von Neumann & Richtmyer as it provides a stronger viscosity for high Mach number, although conversely, on its own this decays too rapidly at low Mach numbers and fails to damp out the unphysical post-shock oscillations or "ringing" that occurs. The standard solution is therefore to use the sum of the two terms (Monaghan, 1989), resulting in a "standard" SPH viscous term of the form

$$\Pi_{ij} = \begin{cases} \dfrac{-\alpha_{\rm SPH}\, c_{{\rm s},ij}\mu_{ij} + \beta_{\rm SPH}\,\mu_{ij}^2}{\rho_{ij}} & \mathbf{v}_{ij}\cdot\mathbf{r}_{ij} < 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3.85}$$

Various numerical tests have showed that in general the constant values $\alpha_{\rm SPH} = 1$, $\beta_{\rm SPH} = 2$ and $\epsilon = 0.01$ in equation 3.83 give good results without significantly affecting non-shocked flows. However, throughout the simulations discussed in the later chapters of this thesis we have used values of $\alpha_{\rm SPH} = 0.1$ and $\beta_{\rm SPH} = 0.2$, which have been found to be adequate to accurately resolve (weak) shocks, while at the same time minimising the artificial heating which would have biased our simulation results – details can be found in Lodato & Rice (2004).

This general form of the viscosity can then be incorporated into the momentum equation to give the following form;

$$\frac{{\rm d}\mathbf{v}_j}{{\rm d}t} = -\sum_i m_i \left(\frac{P_j}{\rho_j^2} + \frac{P_i}{\rho_i^2} + \Pi_{ji}\right)\nabla_j W_{ji}. \tag{3.86}$$

Given that the artificial viscosity term is also pairwise symmetric in i, j (since both $\mathbf{r}_{ij}$ and $\mathbf{v}_{ij}$ are anti-symmetric in i and j) it is clear that this form of the equation of motion also conserves momentum exactly. Likewise it is clear that angular momentum is conserved, and furthermore, by a similar argument to that presented in Section 3.3.3 it is possible to show that in order to preserve energy conservation, the energy equation must be modified to include an extra dissipative term such that

$$\frac{{\rm d}u_j}{{\rm d}t} = \frac{P_j}{\rho_j^2}\sum_i m_i \mathbf{v}_{ji}\cdot\nabla_i W_{ji} + \frac{1}{2}\sum_i m_i \Pi_{ji}\mathbf{v}_{ji}\cdot\nabla_i W_{ji}. \tag{3.87}$$

In this manner it is therefore possible to include a dissipative term such that shocks can be accurately captured, albeit broadened across a few smoothing lengths. Since mass, momentum and energy are still explicitly conserved across the shock the Rankine-Hugoniot equations are automatically satisfied at distances greater than a few smoothing lengths from the shock. Furthermore, since (in theory at least) the

artificial viscosity is zero away from shocks, all the other initial criteria are satisfied also. However, there are various improvements that can be implemented, and these will now be briefly discussed.

3.4.2 More Advanced Viscosities

The thorn in the side of all viscosity prescriptions is the requirement that in the absence of shocks or other natural dissipative processes the artificial viscosity should reduce to zero, thereby requiring some means to discriminate between shocks and other flow features. Compounding the problem is the fact that careful consideration of the artificial viscosity given above (see for instance Lodato & Price 2010) shows that it provides both a bulk and a shear viscosity, while to resolve shocks only the bulk component is required. Any artificial viscosity in the form of equation 3.85 therefore necessarily introduces an unrequired shear viscosity, which can be problematic in situations where shear flows are important (such as discs), leading to spurious energy and angular momentum transport. Furthermore, since the shear force across any given particle varies with its smoothing length, it is clear that this shear component is resolution dependent. Generally speaking however this effect can be reduced by sensible choices for $\alpha_{\rm SPH}$ and $\beta_{\rm SPH}$ (Lodato & Rice, 2004) – this further explains the low values of $\alpha_{\rm SPH}$ and $\beta_{\rm SPH}$ mentioned earlier.

3.4.2.1 The Balsara Switch

An attempt to reduce the induced viscosity in shear flows was presented by Balsara (1995), in which the standard artificial viscosity term Π_{ij} is diminished by the factor $f_{ij} = |f_i + f_j|/2$, where

$$f_i = \frac{|\nabla \cdot \mathbf{v}_i|}{|\nabla \cdot \mathbf{v}_i| + |\nabla \times \mathbf{v}_i| + 0.0001 c_{{\rm s},i}/h}. \tag{3.88}$$

The inclusion of the vorticity (the curl of the flow field) allows this form of the viscosity to perform better in shearing and obliquely shocked flows (see for instance Steinmetz 1996), while remaining unaffected in the case of normal shocks. In a similar manner to the "standard" artificial viscosity term, this form also includes a small term $0.0001 c_{{\rm s},i}/h$ to prevent the viscosity from becoming singular.

3.4.2.2 The Morris & Monaghan Switch

Although the Balsara switch represents a considerable improvement over the standard form of artificial viscosity, problems still arise in the case of shocks in shearing flows, such as those found in accretion discs. For this reason, Morris & Monaghan (1997) introduced the idea of a time-variant viscosity such that Π_{ij} remains unchanged from the standard form, but where $\alpha_{\rm SPH} = \alpha(t)$, and where $\beta_{\rm SPH} = 2\alpha_{\rm SPH}$. The value of α is then evolved for each particle according to the following equation;

$$\frac{\mathrm{d}\alpha}{\mathrm{d}t} = -\frac{\alpha - \alpha_{\rm min}}{\tau} + S_{\mathbf{v}}. \tag{3.89}$$

Here $\alpha_{\rm min} \sim 0.1$ is some minimum value, justified by the requirement that *some* level of artificial viscosity is required to maintain particle order[5], $\tau \sim 0.1 - 0.2\ h/c_{\rm s}$ is a decay timescale (chosen so that the viscosity decays away over a few smoothing lengths) and $S_{\mathbf{v}} = \max(-\nabla \cdot \mathbf{v}, 0)$ is a source term, activated whenever the flow becomes convergent. Although this form of the source term is still non-zero for pure shear flows, this is counter-balanced to some extent by the decay term, and has been found to work well in many tests of the artificial viscosity (Dolag et al., 2005). Further variations on this theme have been effected, including incorporating the Balsara switch into the Π_{ij} term, and capping the maximum value to which α can rise by using a source term of the form $S_{\mathbf{v}} = \max((\alpha_{\rm max} - \alpha)\nabla \cdot \mathbf{v}, 0)$ (Rosswog et al., 2000). For a good general overview of the relative merits of a variety of artificial viscosity methods, see for instance Lombardi et al. (1999); Rosswog (2009); Cullen & Dehnen (2010, in prep.).

3.4.3 A Note on Entropy

All of the above methods have essentially been aiming to capture the same phenomenon, namely the increase in entropy found across a shock front, while simultaneously ensuring isentropic flow elsewhere. Furthermore all share the common feature that flow evolution proceeds via integration of the energy equation. However, an approach espoused by Springel & Hernquist (2002) is to consider evolving the entropy directly, thereby ensuring that the entropy can only *increase*.

In this manner, we recall that in terms of density ρ and specific entropy s, the

[5]Note that only very low levels of viscosity are required for this purpose – $\alpha_{\rm min} \sim 0.01$ should suffice (Cullen & Dehnen, 2010, in prep.).

equation of state is given by

$$P_i = \kappa_i(s_i)\rho_i^{\gamma}. \tag{3.90}$$

for some entropic function $\kappa(s)$. Similarly, the internal energy u_i may be obtained from ρ and s via

$$u_i = \frac{\kappa_i(s_i)}{\gamma - 1}\rho_i^{\gamma-1}. \tag{3.91}$$

In the case of isentropic flow, we have $\kappa(s) = const$, and thus by definition

$$\frac{\mathrm{d}\kappa_i}{\mathrm{d}t} = 0. \tag{3.92}$$

However, in the case where artificial viscosity is included, the time derivative of the entropic function becomes

$$\frac{\mathrm{d}\kappa_i}{\mathrm{d}t} = \frac{1}{2}\frac{\gamma - 1}{\rho_i^{\gamma-1}} \sum_j m_j \Pi_{ij} \mathbf{v}_{ij} \cdot \nabla_i W_{ij}. \tag{3.93}$$

By noting that

$$\nabla_i W_{ij} = |\nabla_i W_{ij}| \hat{\mathbf{r}}_{ij}, \tag{3.94}$$

and also that Π_{ij} is only non-zero for $\mathbf{v}_{ij} \cdot \mathbf{r}_{ij} < 0$, it is clear that the term on the RHS of equation 3.93 is strictly non-negative, and thus that entropy can only increase throughout the flow. Using this method of evolving the flow properties it is therefore possible to explicitly ensure that the entropy of any particle increases monotonically with time.

3.5 Variable Smoothing Lengths

Up to now, it has been assumed that the smoothing length h is held constant with time, and is moreover equal for all particles. In regions where the density (and thus the number of neighbours) is roughly constant, this maintains a constant (small) sampling error within the SPH smoothing kernel. This requirement of constant smoothing length is quite restrictive however, as it prevents the code adapting effectively to regions of higher or lower than average density (Steinmetz & Mueller, 1993). By allowing the smoothing length to vary both temporally and spatially, sampling errors can be minimised across regions of varying density, as either the number of neighbours or the mass within a smoothing kernel (and thus the resolution) may

be maintained. There are various simple ways of allowing variable effective smoothing lengths that have been introduced, for instance Benz (1990) suggested using a symmetrised smoothing length $h_{ij} = (h_i + h_j)/2$, such that the kernel becomes

$$W_{ij} = W\left(\mathbf{r}_{ij}, \frac{h_i + h_j}{2}\right). \tag{3.95}$$

An alternative method has been suggested by Hernquist & Katz (1989), in which the average kernel value is used rather than the average smoothing length, such that

$$W_{ij} = \frac{W(\mathbf{r}_{ij}, h_i) + W(\mathbf{r}_{ij}, h_j)}{2}. \tag{3.96}$$

With variable smoothing lengths it then becomes necessary to determine the value of h for each particle. A standard method of doing this is to link the smoothing length to the local density, such that

$$\rho_i h_i^3 = const. \tag{3.97}$$

Since this constant clearly has units of mass, it is frequently linked to the particle mass, giving the following prediction for the particle smoothing length;

$$h_i = \eta \left(\frac{m_i}{\rho_i}\right)^{1/3}, \tag{3.98}$$

where the coupling constant is generally in the range $1.2 < \eta < 1.5$ (Rosswog, 2009). By construction this method maintains a constant mass within the smoothing kernel. As each of the above formalisms remains pairwise symmetric, momentum remains fully conserved, and increased spatial resolution is achieved at relatively low computational cost. The latter method (using the averaged kernel value as in equation 3.96) has additional advantages in it is less problematic across shocks, and it couples better with tree methods for calculating self-gravity (Steinmetz & Mueller, 1993). Nonetheless, in both cases errors appear in either the entropy or energy equation, such that either

$$\frac{\mathrm{d}E}{\mathrm{d}t} \text{ or } \frac{\mathrm{d}\kappa(s)}{\mathrm{d}t} \sim \frac{\partial W}{\partial h}\frac{\partial h}{\partial t} \neq 0, \tag{3.99}$$

(Hernquist, 1993), and the relevant quantity is therefore not explicitly conserved.

It is however possible to construct SPH estimates that self-consistently account

for the variation in smoothing length, and therefore ensure exact energy conservation. In this case, the estimator for density equation 3.39 becomes

$$\rho_j = \sum_i m_i W(\mathbf{r}_{ji}, h_j), \tag{3.100}$$

noting that the smoothing length used in the kernel is that associated with particle j only, and thereby remains constant throughout the summation. By taking the (Lagrangian) time derivative, we obtain

$$\frac{\mathrm{d}\rho_j}{\mathrm{d}t} = \sum_i m_i \left(\mathbf{v}_{ji} \cdot \nabla_i W_{ji}(h_j) + \frac{\partial W_{ji}}{\partial h_j} \frac{\mathrm{d}h_j}{\mathrm{d}t} \right), \tag{3.101}$$

noting the extra terms compared to equation 3.41, and where now we set $W_{ji} = W(\mathbf{r}_{ji}, h_j)$. Noting that

$$\frac{\mathrm{d}h_j}{\mathrm{d}t} = \frac{\mathrm{d}h_j}{\mathrm{d}\rho_j} \frac{\mathrm{d}\rho_j}{\mathrm{d}t}, \tag{3.102}$$

and using equation 3.98 we see that

$$\frac{\mathrm{d}h_j}{\mathrm{d}\rho_j} = -\frac{h_j}{3\rho_j}. \tag{3.103}$$

Substituting this into equation 3.101 and gathering like terms, we find that the time derivative of the density becomes

$$\frac{\mathrm{d}\rho_j}{\mathrm{d}t} = \frac{1}{\Omega_j} \sum_i m_i \mathbf{v}_{ji} \cdot \nabla_i W_{ij}(h_j) \tag{3.104}$$

where

$$\Omega_j = 1 - \frac{\mathrm{d}h_j}{\mathrm{d}\rho_j} \sum_i m_i \frac{\partial W_{ji}(h_j)}{\partial h_j}, \tag{3.105}$$

$$= 1 + \frac{h_j}{3\rho_j} \sum_i m_i \frac{\partial W_{ji}(h_j)}{\partial h_j}, \tag{3.106}$$

and where $\partial W_{ji}/\partial h_j$ is known from the choice of kernel. Although it can be calculated directly from the kernel, in the case of the cubic spline kernel given in equation 3.31 it is generally evaluated by noting that

$$\frac{\partial W}{\partial h_j} = -x \nabla W - \frac{3}{h} W, \tag{3.107}$$

where W and ∇W are given by equations 3.31 and 3.33 respectively.

Similarly, there is a correction factor to the momentum equation to allow for the spatial variation in smoothing lengths. Recall from equation 3.52 that in order to calculate the spatial variation of the Lagrangian, we need to know the spatial derivative of the density. Allowing now for variable smoothing lengths and using equation 3.57, we therefore find that

$$\frac{\partial \rho_j}{\partial \mathbf{r}_i} = \sum_k m_k \left(\nabla_j W_{ji}(h_j) \left[\delta_{ji} - \delta_{jk} \right] + \frac{\partial W_{jk}(h_j)}{\partial h_j} \frac{\mathrm{d}h_j}{\mathrm{d}\rho_j} \frac{\partial \rho_j}{\partial \mathbf{r}_i} \right). \tag{3.108}$$

By gathering like terms, we find that the correction factor for the spatial derivative of the density is same as that for the temporal one, namely that

$$\frac{\partial \rho_j}{\partial \mathbf{r}_i} = \frac{1}{\Omega_j} \sum_k m_k \nabla_i W_{jk}(h_j) \left[\delta_{ji} - \delta_{jk} \right], \tag{3.109}$$

with the factor Ω_j defined as before in equation 3.106.

Following the same derivation as in Section 3.3.2, it is then easy to show that the acceleration due to hydrodynamic forces with spatially varying smoothing lengths is given by

$$\frac{\mathrm{d}\mathbf{v}_j}{\mathrm{d}t} = -\sum_i m_i \left(\frac{P_j}{\Omega_j \rho_j^2} \nabla_i W_{ji}(h_j) + \frac{P_i}{\Omega_i \rho_i^2} \nabla_j W_{ji}(h_i) \right). \tag{3.110}$$

Finally, from equation 3.73, we see that the evolution of the internal energy in the presence of variable smoothing lengths becomes

$$\frac{\mathrm{d}u_j}{\mathrm{d}t} = \frac{P_j}{\Omega_j \rho_j^2} \sum_i m_i \mathbf{v}_{ji} \cdot \nabla W_{ji}(h_j). \tag{3.111}$$

By an analogous process to that described in Section 3.3.3, it is possible to show that this equation for the evolution of the internal energy is also explicitly conservative of the total energy of the system, E. The three equations 3.100, 3.110 and 3.111 along with the relationship between the density and the smoothing length equation 3.98 therefore form a fully consistent, fully conservative SPH formalism with spatially varying smoothing lengths.

A problem exists however, in that in order to obtain the density, one needs to know the smoothing length (equation 3.100) and to obtain the smoothing length one needs to know the density (equation 3.98). In order to resolve this, this pair of equa-

tions can be solved iteratively by the Newton-Raphson method (Price & Monaghan, 2007). By rewriting equation 3.98, we can combine these two equations to reduce the problem to that of finding the root h_j of the equation $\zeta(h_j) = 0$, where

$$\zeta(h_j) = m_j \left(\frac{\eta}{h_j}\right)^3 - \sum_i m_i W(\mathbf{r}_{ji}, h_j). \tag{3.112}$$

Here the first term represents the density obtained from assuming a fixed mass within the smoothing kernel, while the second term is the standard SPH estimate for the density. From some initial estimate of the root h_j, the Newton-Raphson method gives a better estimate as being

$$h_{j,\text{new}} = h_j - \frac{\zeta(h_j)}{\zeta'(h_j)}, \tag{3.113}$$

where the prime denotes differentiation with respect to h. By using equation 3.106 we see that

$$\zeta'(h_j) = -\frac{3\rho_j \Omega_j}{h_j}, \tag{3.114}$$

and thus the updated value $h_{j,\text{new}}$ is given by

$$h_{j,\text{new}} = h_j \left(1 + \frac{\zeta(h_j)}{3\rho_j \Omega_j}\right). \tag{3.115}$$

This may be repeated until $|h_{j,\text{new}} - h_j|/h_j < \epsilon$ for some small value of ϵ, frequently set to 10^{-3}. Then in turn, a self consistent value of the density is then obtained from equation 3.98. As there is generally relatively little change in h_j and ρ_j between timesteps, the estimator for h_j is taken as the value from the previous timestep, and convergence usually occurs within a small number of iterations (Price & Monaghan, 2007). In pathological cases where the Newton-Raphson method does not converge, other, universally convergent but slower methods such as the bisection method may be used instead.

Although the inclusion of variable smoothing lengths through this method does inevitably increase the computational cost of the code, this is relatively small, and the conservation properties are recovered to within machine (and integrator) tolerance. Other tricks, such as predicting the change in the smoothing length using equation 3.102 can reduce the computational cost still further (see for instance, Price & Monaghan 2007).

3.6 Including Gravity

As many astrophysical situations are driven at some level by gravitational forces, it is important to be able to include this consistently within the SPH framework, and in such a manner that the inherent conservation properties of the algorithm are not compromised. While much work has been put into N-body simulations of discrete particles, within the SPH formalism we are aiming to model the gravitational force over a continuum, and thus it should be smoothed (or *softened* in SPH parlance) in a similar manner to that in which the discrete particle mass is smoothed into the density of a fluid continuum. In this section we therefore consider how this can be done in a consistent manner, and one in which as before momentum and energy are explicitly conserved.

3.6.1 Gravity in the Lagrangian

In an extension to the Lagrangian for the hydrodynamic equations of motion, it is possible to incorporate the effects of gravity by considering a Lagrangian of the form

$$\mathcal{L}(\mathbf{r},\mathbf{v}) = \sum_i m_i \left(\frac{1}{2}\mathbf{v}_i \cdot \mathbf{v}_i - u_i(\mathbf{r}_i)\right) - \Psi, \tag{3.116}$$

where Ψ is an as yet undefined measure of the total gravitational potential energy of the system. By comparison with equation 3.46 this is clearly just the hydrodynamic Lagrangian with an additional term

$$\mathcal{L}_{\text{grav}} = -\Psi \tag{3.117}$$

which describes the effects of gravity.

Now, as with the density, at position $\mathbf{r}_i$ we can obtain the local gravitational potential Φ_i, via a sum over all particles such that

$$\Phi_i = G\sum_j m_j \phi(\mathbf{r}_i - \mathbf{r}_j, \varepsilon_i), \tag{3.118}$$

where $\phi(\mathbf{r}_i - \mathbf{r}_j) = \phi(\mathbf{r}_{ij})$ is known as the (gravitational) softening kernel, G is the universal gravitational constant and where ε_i is the softening length associated with particle i. The softening kernel at this stage is fairly general, but it must have the following properties:

- $\phi(r,h) < 0$ for all r, h, as the local potential Φ must be strictly negative definite;
- $\nabla\phi(0,h) = 0$, such that the gravitational force exerted by any particle on itself is zero;
- $\lim_{r/h\to\infty} \phi(r,h) = -\frac{1}{r}$, i.e. the softening should reduce to zero at large inter-particle distances, and the Newtonian potential should be recovered.

Generally speaking, and throughout this thesis, it is assumed that the softening length is exactly equal to the smoothing length for all particles, i.e. $\varepsilon_i = h_i$ for all i. In a similar manner it is generally taken that the force should only be softened when $r < 2h$, so that force softening and density smoothing occur over exactly the same region.

Noting that the gravitational potential energy is just the mass times the gravitational potential, since the latter is defined over pairs of particles, by definition the *total* gravitational potential energy of the system is given by the sum over all *pairs* of particles, such that

$$\Psi = G\sum_i m_i \sum_{j\leq i} m_j \phi_{ij}(h_i) \tag{3.119}$$

$$= \frac{G}{2}\sum_i \sum_j m_i m_j \phi_{ij}(h_i) \tag{3.120}$$

Note that equation 3.119 sums over all pairs of particles, including the so-called self-interaction terms where $i = j$, and thus explains the factor of a half in equation 3.120. From this definition we therefore find that

$$\Psi = \frac{1}{2}\sum_i m_i \Phi_i, \tag{3.121}$$

and thus that the full Lagrangian in the presence of gravity becomes

$$\mathcal{L}(\mathbf{r},\mathbf{v}) = \sum_i m_i \left(\frac{1}{2}\mathbf{v}_i\cdot\mathbf{v}_i - u_i(\mathbf{r}_i) - \frac{1}{2}\Phi_i\right). \tag{3.122}$$

By considering only the gravitational term in the Lagrangian, we can as before use the Euler-Lagrange equations 3.47 to obtain the acceleration due to gravity,

which becomes

$$m_j \frac{\mathrm{d}v_j}{\mathrm{d}t} = \frac{\partial \mathcal{L}_{\mathrm{grav}}}{\partial \mathbf{r}_j}. \tag{3.123}$$

Using equations 3.117 and 3.120 we therefore find that the spatial derivative of the gravitational Lagrangian becomes

$$\frac{\partial \mathcal{L}_{\mathrm{grav}}}{\partial \mathbf{r}_j} = -\frac{G}{2} \sum_i \sum_k m_i m_k \frac{\partial \phi_{ik}(h_i)}{\partial \mathbf{r}_j} \tag{3.124}$$

$$= -\frac{G}{2} \sum_i \sum_k m_i m_k \left(\nabla_j \phi_{ik}(h_i) + \frac{\partial \phi_{ik}(h_i)}{\partial h_i} \frac{\partial h_i}{\partial \mathbf{r}_j} \right). \tag{3.125}$$

Here we see that in the case of fixed smoothing (and therefore softening) lengths, $\partial h_i / \partial \mathbf{r}_k = 0$, and thus we only require the first term to determine the effects of gravity. The second term is therefore a correction term to allow for spatial variation in h.

As before, using the method of equations 3.54 to 3.57 the spatial gradient becomes

$$\left. \frac{\partial \mathcal{L}_{\mathrm{grav}}}{\partial \mathbf{r}_j} \right|_h = -\frac{G}{2} \sum_i \sum_k m_i m_k \, \nabla_j \phi_{ik}(h_i) \left[\delta_{ij} - \delta_{kj} \right], \tag{3.126}$$

$$= -\frac{G}{2} m_j \sum_k m_k \nabla_j \phi_{jk}(h_j) + \frac{G}{2} \sum_i m_i m_j \nabla_j \phi_{ij}(h_i). \tag{3.127}$$

Now by changing the summation index of the first term to i, and noting again that the kernel is antisymmetric we obtain

$$\left. \frac{\partial \mathcal{L}_{\mathrm{grav}}}{\partial \mathbf{r}_j} \right|_h = -\frac{G}{2} m_j \sum_i m_i \left(\nabla_j \phi_{ji}(h_j) + \nabla_j \phi_{ji}(h_i) \right), \tag{3.128}$$

which therefore encapsulates the effects of gravity in the case of constant smoothing lengths.

If we now consider the second term in equation 3.125 and self-consistently correct for spatial variation in the smoothing length, we find that

$$\left. \frac{\partial \mathcal{L}_{\mathrm{grav}}}{\partial \mathbf{r}_j} \right|_{\mathrm{corr}} = -\frac{G}{2} \sum_i \sum_k m_i m_k \frac{\partial \phi_{ik}(h_i)}{\partial h_i} \frac{\mathrm{d}h_i}{\mathrm{d}\rho_i} \frac{\partial \rho_i}{\partial \mathbf{r}_j}. \tag{3.129}$$

By substituting equation 3.109 into the above we find that

$$\left.\frac{\partial \mathcal{L}_{\text{grav}}}{\partial \mathbf{r}_j}\right|_{\text{corr}} = -\frac{G}{2}\sum_i\sum_k m_i m_k \frac{\partial \phi_{ik}(h_i)}{\partial h_i}\frac{\mathrm{d}h_i}{\mathrm{d}\rho_i}\frac{1}{\Omega_i}\sum_l m_l \nabla_j W_{il}(h_l)[\delta_{ij}-\delta_{lj}] \quad (3.130)$$

$$\begin{aligned} &= -\frac{G}{2} m_j \sum_k m_k \frac{\partial \phi_{jk}}{\partial h_j}\frac{\mathrm{d}h_j}{\mathrm{d}\rho_j}\frac{1}{\Omega_j}\sum_l m_l \nabla_j W_{jl}(h_l) \\ &\quad + \frac{G}{2}\sum_i\sum_k \frac{\partial \phi_{ik}(h_i)}{\partial h_i}\frac{1}{\Omega_i} m_j \nabla_j W_{ij}(h_i). \end{aligned} \quad (3.131)$$

Now by changing the summation index of the second sum in the first term of equation 3.131 to i, defining a new quantity ξ_p such that

$$\xi_p = \frac{\mathrm{d}h_p}{\mathrm{d}\rho_p}\sum_q m_q \frac{\partial \phi_{pq}(h_p)}{\partial h_p}, \quad (3.132)$$

and using the antisymmetry property of the gradient of the smoothing kernel, we see that the correction term reduces to

$$\left.\frac{\partial \mathcal{L}_{\text{grav}}}{\partial \mathbf{r}_j}\right|_{\text{corr}} = -\frac{G}{2} m_j \sum_i m_i \left(\frac{\xi_j}{\Omega_j}\nabla_j W_{ji}(h_j) + \frac{\xi_i}{\Omega_i}\nabla_j W_{ji}(h_i)\right). \quad (3.133)$$

Finally, using equation 3.123 and by incorporating the effects of gravity into the equations of motion for a hydrodynamic flow with artificial viscosity (while self-consistently allowing for variable smoothing lengths) we find that the full equations of motion become

$$\begin{aligned} \frac{\mathrm{d}\mathbf{v}_j}{\mathrm{d}t} &= -\sum_i m_i \left(\frac{P_j}{\Omega_j \rho_j^2}\nabla_j W_{ji}(h_j) + \frac{P_i}{\Omega_i \rho_i^2}\nabla_j W_{ji}(h_i) + \Pi_{ji}\frac{\nabla_j W_{ji}(h_j) + \nabla_j W_{ji}(h_i)}{2}\right) \\ &\quad - \frac{G}{2}\sum_i m_i \left(\nabla_j \phi_{ji}(h_j) + \nabla_j \phi_{ji}(h_j)\right) \\ &\quad - \frac{G}{2}\sum_i m_i \left(\frac{\xi_j}{\Omega_j}\nabla_j W_{ji}(h_j) + \frac{\xi_i}{\Omega_i}\nabla_j W_{ji}(h_i)\right), \end{aligned} \quad (3.134)$$

with Ω_i and ξ_i defined as per equations 3.106 and 3.132 respectively[6]. As in the case for the pure hydrodynamic flow, the use of a Lagrangian in deriving these equations ensures the explicit conservation of both linear and angular momentum, which is also clear from the pairwise symmetry present in all terms in the above equation.

[6]Note that for consistency, the artificial viscosity term Π_{ji} uses the *average* value of the smoothing lengths $h_{ji} = (h_j + h_i)/2$ in its definition of μ_{ji} (equation 3.83).

3.6.2 Evolution of the Gravitational Potential

Clearly as particles move about within a gravitational potential, their potential energy (given in SPH terms by $m_j\Phi_j$) will also vary. Although the potential (and thus the potential energy) is obtained at any point by the sum over particles using equation 3.118, the time evolution of the potential energy is required to maintain energy conservation. Hence in a similar manner to Section 3.3.3 we must consider the total energy of the system, which including the gravitational potential energy becomes

$$E = \sum_j m_j \left(\frac{1}{2}\mathbf{v}_j \cdot \mathbf{v}_j + u_j + \frac{1}{2}\Phi_j \right). \tag{3.135}$$

As before, to ensure energy conservation we require that the time derivative of the total energy is zero, i.e. that

$$\sum_j m_j \left(\mathbf{v}_j \cdot \frac{\mathrm{d}\mathbf{v}_j}{\mathrm{d}t} + \frac{\mathrm{d}u_j}{\mathrm{d}t} + \frac{1}{2}\frac{\mathrm{d}\Phi_j}{\mathrm{d}t} \right) = 0. \tag{3.136}$$

By considering equation 3.118, we see that

$$\frac{\mathrm{d}\Phi_j}{\mathrm{d}t} = \frac{G}{2}\sum_i m_i \left(\nabla_j \phi_{ji}(h_j) \cdot \frac{\mathrm{d}\mathbf{r}_j}{\mathrm{d}t} + \nabla_i \phi_{ji}(h_j) \cdot \frac{\mathrm{d}\mathbf{r}_i}{\mathrm{d}t} + \frac{\partial \phi_{ji}}{\partial h_j}\frac{\mathrm{d}h_j}{\mathrm{d}\rho_j}\frac{\mathrm{d}\rho_j}{\mathrm{d}t} \right). \tag{3.137}$$

Recalling the definition of ξ_j from equation 3.132, and using equation 3.104 for the definition of $\mathrm{d}\rho_j/\mathrm{d}t$ with variable smoothing lengths, we obtain

$$\frac{\mathrm{d}\Phi_j}{\mathrm{d}t} = \frac{G}{2}\sum_i m_i \mathbf{v}_{ji} \cdot \left(\nabla_j \phi_{ji}(h_j) + \frac{\xi_j}{\Omega_j}\nabla_j W_{ji}(h_j) \right), \tag{3.138}$$

From Sections 3.3.3 and 3.5 we know that in the energy balance (equation 3.135), over all particles the hydrodynamic terms in the equations of motion (equation 3.110) exactly counteract the temporal rate of change of the internal energy,

$$\frac{dE_{\mathrm{hydro}}}{\mathrm{d}t} = \sum_j m_j \left(\mathbf{v_j} \cdot \left.\frac{\mathrm{d}\mathbf{v}_j}{\mathrm{d}t}\right|_{\mathrm{hydro}} + \frac{\mathrm{d}u_j}{\mathrm{d}t} \right) = 0, \tag{3.139}$$

and thus pure hydrodynamic flows are exactly conservative of energy. With the inclusion of gravity we therefore only need to show that over the whole system the gravitational terms in the equations of motion balance the time derivative of the

gravitational potential, i.e. that

$$\frac{dE_{\text{grav}}}{\mathrm{d}t} = \sum_j m_j \left(\mathbf{v_j} \cdot \left. \frac{\mathrm{d}\mathbf{v}_j}{\mathrm{d}t} \right|_{\text{grav}} + \frac{1}{2} \frac{\mathrm{d}\Phi_j}{\mathrm{d}t} \right) = 0 \tag{3.140}$$

in order to maintain exact conservation of energy in self-gravitating systems.

From equations 3.134 and 3.138 this gravitational energy balance becomes

$$\begin{aligned} \frac{dE_{\text{grav}}}{\mathrm{d}t} \;=\; & -\frac{G}{2} \sum_j \sum_i m_j m_i \mathbf{v}_j \cdot \Bigg(\nabla_j \phi_{ji}(h_j) + \nabla_j \phi_{ji}(h_i) + \\ & \qquad \frac{\xi_j}{\Omega_j} \nabla_j W_{ji}(h_j) + \frac{\xi_i}{\Omega_i} \nabla_j W_{ji}(h_i) \Bigg) \\ & + \frac{G}{2} \sum_j \sum_i m_j m_i \, \mathbf{v}_{ji} \cdot \left(\nabla_j \phi_{ji}(h_i) + \frac{\xi_j}{\Omega_j} \nabla_j W_{ji}(h_j) \right) \end{aligned} \tag{3.141}$$

Cancelling like terms, this reduces to

$$\begin{aligned} \frac{dE_{\text{grav}}}{\mathrm{d}t} \;=\; & -\frac{G}{2} \sum_j \sum_i m_j m_i \Bigg(\mathbf{v}_j \cdot \nabla_j \phi_{ji}(h_j) + \mathbf{v}_i \cdot \nabla_j \phi_{ji}(h_i) \\ & \qquad + \frac{\xi_i}{\Omega_i} \mathbf{v}_j \cdot \nabla_j W_{ji}(h_i) + \frac{\xi_j}{\Omega_j} \nabla_j W_{ji}(h_j) \Bigg) , \end{aligned} \tag{3.142}$$

and finally, noticing that the gradients of both the smoothing and the softening kernels are antisymmetric under a reversal of the summation indices i and j, we obtain the desired result that

$$\frac{dE_{\text{grav}}}{\mathrm{d}t} = 0. \tag{3.143}$$

Therefore, we see that gravity can be included into SPH in such a manner that the algorithm remains explicitly conservative of energy.

3.6.3 Gravitational Potentials and the Softening Kernel

Finally for this section, we need to consider the form of the gravitational softening kernel ϕ, and its relation to the smoothing kernel W. Recall that Poisson's equation links the gravitational potential $\Phi(\mathbf{r})$ to the density $\rho(\mathbf{r})$ at position $\mathbf{r}$, such that

$$\nabla^2 \Phi(\mathbf{r}) = 4\pi G \rho(\mathbf{r}). \tag{3.144}$$

Given that we implicitly assume each particle to be spherically symmetric, by using spherical polar co-ordinates and substituting equations 3.39 and 3.118 into equation 3.144 we find that (for a generalised radial co-ordinate r)

$$W(r,h) = \frac{1}{4\pi r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial\phi(r,h)}{\partial r}\right), \tag{3.145}$$

where we have neglected the spatial variation of h[7].

We can now integrate this, to link the derivative of the softening kernel $\partial\phi/\partial r$ (also known as the force kernel) to the smoothing kernel, such that

$$\frac{\partial\phi}{\partial r} = \frac{4\pi}{r^2}\int^r r'^2 W(r')\mathrm{d}r' + \frac{C_1}{r^2}, \tag{3.146}$$

with the integration constant C_1 subject to the condition that for $r \geq 2h$ we recover the standard Newtonian inverse square law, which using our definitions becomes $\partial\phi/\partial r = 1/r^2$. In a similar manner we can integrate this a step further (by parts), to give the full softening kernel, such that

$$\phi = 4\pi\left[-\frac{1}{r}\int^r r'^2 W(r')\mathrm{d}r' + \int^r r' W(r')\mathrm{d}r'\right] + \frac{C_1}{r^2} + \frac{C_2}{r}, \tag{3.147}$$

where the second integration constant allows the correct asymptotic behaviour (i.e. $\phi \to 0$ as $r \to \infty$) to be established.

With this in mind, for the cubic spline kernel defined in equation 3.31 the force kernel $\partial\phi/\partial r$ becomes

$$\frac{\partial\phi(r,h)}{\partial r} = \begin{cases} \dfrac{1}{h^2}\left(\dfrac{4}{3}x - \dfrac{6}{5}x^3 + \dfrac{1}{2}x^4\right) & 0 \leq x \leq 1, \\ \dfrac{1}{h^2}\left(\dfrac{8}{3}x - 3x^2 + \dfrac{6}{5}x^3 - \dfrac{1}{6}x^4 - \dfrac{1}{15x^2}\right) & 1 \leq x \leq 2, \\ \dfrac{1}{r^2} & x \geq 2, \end{cases} \tag{3.148}$$

where $x = r/h$ and where the integration constants have been absorbed to ensure piecewise continuity. Finally, we therefore find the full softening kernel consistent

[7]This is because the smoothing length essentially acts as a normalising constant in both the smoothing and the softening kernels, and for any given particle is held constant within Poisson's equation. Thus its spatial variation is immaterial here.

with the cubic spline smoothing kernel to be

$$\phi(r,h) = \begin{cases} \frac{1}{h}\left(\frac{2}{3}x^2 - \frac{3}{10}x^3 + \frac{1}{10}x^5 - \frac{7}{5}\right) & 0 \leq x \leq 1, \\ \frac{1}{h}\left(\frac{4}{3}x^2 - x^3 + \frac{3}{10}x^4 - \frac{1}{30}x^5 - \frac{8}{5} + \frac{1}{15x}\right) & 1 \leq x \leq 2, \\ -\frac{1}{r} & x \geq 2. \end{cases} \tag{3.149}$$

Using this definition of the softening kernel along with the cubic spline smoothing kernel equation 3.31, the equations of motion equation 3.134 and equation 3.138 for the evolution of the gravitational potential therefore allows gravity to be included in a manner that it is fully conservative, and is such that Poisson's equation is satisfied throughout.

3.7 Finding the Nearest Neighbours

Various methods exist for finding the nearest neighbours (i.e. those particles within the smoothing kernel of any given particle), with the simplest being a direct search over all particles. This becomes very expensive in the limit of large numbers of particles N however, as the computational cost scales as $\mathcal{O}(N^2)$. Other methods such as using an overlaid grid or a linked list of particle positions have been used (Hockney & Eastwood, 1981; Monaghan, 1985; Murray, 1996; Deegan, 2009). One of the more efficient methods however is to use a hierarchical tree structure, an approach that grew out the requirements of N-body codes to distinguish distant particles (where the gravitational forces could be evaluated via multipole expansions) from local particles (where direct N-body calculation of the forces was still required). These in general reduce the cost of neighbour-finding from $\mathcal{O}(N^2)$ to $\mathcal{O}(N \log N)$ (Barnes & Hut, 1986; Hernquist, 1987; Hernquist & Katz, 1989), although reductions to $\mathcal{O}(N)$ have been achieved (Dehnen, 2002, 2000).

Trees are essentially data structures which decompose the computational domain into a series of discrete volumes, the sum over which contains all the particles. The smallest volumes generally contain only a single particle, but this is not strictly necessary, and indeed may inhibit the efficiency of the code (Dehnen, 2009, private communication). By construction, particles which are near each other in space are near each other within the tree structure, and thus by looping over a relatively small part of the tree, the nearest neighbours may be found efficiently. There are various

different algorithms that perform this decomposition described in the literature, none of which are trivial, so I shall not attempt to go into any depth here. For further details, see for instance Press et al. (2007); Dehnen (2002); Steinmetz & Mueller (1993); Barnes & Hut (1986); Bentley (1975) and references therein.

A distinct advantage of using trees for neighbour finding within an SPH code is that they couple readily with pre-existing methods for evaluating the gravitational force between large numbers of particles. Rather than a direct summation (which is of $\mathcal{O}(N^2)$) over all particles to find the gravitational force at a specific location, particles at large distances can be effectively treated as a single body, and multipole expansions used to approximate the force. This approach has found success in various N-body codes as a means of reducing the computation time to $\mathcal{O}(N \log N)$ or lower (Barnes & Hut, 1986; Hernquist, 1987; Hernquist & Katz, 1989). Use of a tree algorithm therefore allows the process of neighbour finding to be coupled to that of finding the gravitational forces acting on a particle, with an attendant saving of computational expense.

3.8 Integration and Timestepping

So far we have obtained equations to evolve the density, the three components of velocity under the influence of pressure, gravitational and (artificial) viscous forces, the internal energy and the gravitational potential. Finally therefore it is time to consider *how* these equations are actually evolved, and to discuss the issues of temporal integration and time-stepping.

Generally speaking there are two principal methods used to perform the time evolution, and indeed the code I have used throughout includes the option to use either. They are the so-called Leapfrog integrator (also known as the kick-drift or Störmer-Verlet integrator) and the Runge-Kutta-Fehlberg method, and I shall now briefly consider both of these.

3.8.1 The Leapfrog Integrator

The leapfrog integrator is a second-order integrator, so-called because the position and the velocity are advanced half a timestep out of phase, with each update of position or velocity using the value of the velocity or position evaluated at the previous half timestep. In this manner, the positions and velocities "leap-frog" over

each other at every half timestep, giving rise to the name. The leapfrog method is widely used in N-body codes, since in the case where the acceleration is independent of the velocity, i.e. where $\mathbf{a} = \mathbf{a}(\mathbf{r})$ only, it is particularly simple to implement. In its "pure" form it is a time-reversible, symplectic integrator, which by definition is explicitly conservative of both energy and angular momentum (see for instance Springel (2005) and references therein).

In essence then, if the position, velocity and acceleration at time t_i are given by $\mathbf{r}_i$, $\mathbf{v}_i$ and $\mathbf{a}_i$ respectively, with a timestep δt the standard form of the leapfrog integrator gives the positions and velocities as

$$\begin{aligned} \mathbf{r}_{i+1} &= \mathbf{r}_i + \mathbf{v}_{i-1/2}\,\delta t, \\ \mathbf{v}_{i+1/2} &= \mathbf{v}_{i-1/2} + \mathbf{a}_i\,\delta t. \end{aligned} \tag{3.150}$$

Here it is clear that the positions and velocities are evaluated at half timesteps with respect to each other, and "leap-frog" over each other as they are evolved. In this form it is also clear that the integrator should be perfectly time-reversible.

A form that is often more readily applied is the equivalent definition at integer timesteps, which becomes

$$\begin{aligned} \mathbf{r}_{i+1} &= \mathbf{r}_i + \delta t \left(\mathbf{v}_i + \frac{\delta t}{2}\,\mathbf{a}_i \right), \\ \mathbf{v}_{i+1} &= \mathbf{v}_i + \frac{\delta t}{2}\left(\mathbf{a}_i + \mathbf{a}_{i+1}\right). \end{aligned} \tag{3.151}$$

Although it is now less obvious, these equations are still fully time reversible. Note further that the form of the increments on the RHS of each of the above equations is equivalent to an estimate of the relevant quantity at the following half timestep, noting in particular that

$$\mathbf{v}_i + \frac{\delta t}{2}\,\mathbf{a}_i = \mathbf{v}_{i+1/2}, \tag{3.152}$$

to first order.

Note however, that in both cases problems arise if the acceleration depends on the velocity, since from equation 3.151 we see that to calculate $\mathbf{v}_{i+1}$ we already need to know the acceleration $\mathbf{a}_{i+1}$, and the scheme becomes implicit. Since in SPH simulations both the pressure force and the artificial viscous force depend on the local velocity, it is clear that modifications are required before this integrator may be used. The standard way this correction is implemented (see for instance Springel et al. 2001; Wetzstein et al. 2009) is as follows:

- Firstly, predict the positions at time $t_{i+1/2}$ in a manner analogous to equation 3.152 via

$$\mathbf{r}_{i+1/2} = \mathbf{r}_i + \frac{\delta t}{2}\,\mathbf{v}_i. \tag{3.153}$$

- Secondly, use equation 3.152 to obtain the velocity at time $t_{i+1/2}$, and extrapolate other values (such as density, internal energy and gravitational potential) at the half timestep also. Hence calculate the acceleration at the half timestep, $\mathbf{a}_{i+1/2}$.

- Calculate the velocity at time t_{i+1} using

$$\mathbf{v}_{i+1} = \mathbf{v}_i + \delta t\,\mathbf{a}_{i+1/2}. \tag{3.154}$$

- Now update the positions to timestep t_{i+1} using

$$\mathbf{r}_{i+1} = \mathbf{r}_i + \frac{\delta t}{2}\left(\mathbf{v}_i + \mathbf{v}_{i+1}\right). \tag{3.155}$$

The process is now repeated as required.

Although the strict symmetry between the integration of positions and velocities has been lost by the inclusion of these predictor steps, this method still remains time-reversible. Furthermore, it is also now possible to include adaptive timestepping, which would have been problematic before precisely *because* of the symmetry between the integrations (Wetzstein et al., 2009). Generally speaking however, maintaining time-reversibility with adaptive timestepping is difficult (Springel et al., 2001; Quinn et al., 1997), though not impossible.

3.8.2 The Runge-Kutta-Fehlberg Integrator

Runge-Kutta methods for integrating systems of differential equations are well known, tried and trusted methods, which use multiple estimates of the derivative across a given timestep to arrive at accurate, generally high order estimates for the evolved quantities. Most common is the fourth order Runge-Kutta method, often simply abbreviated to RK4, which has been known and used for over a century (Kutta, 1901). Moreover, Fehlberg (1968, 1969) obtained a modified Runge-Kutta integrator (now known as a Runge-Kutta-Fehlberg integrator) which embedded a order $n+1$ method within an order n method. This allows the two methods to be compared to give an error estimate, and thus for the error to be controlled to some

given tolerance. The most common of these methods embeds a fifth order estimate within a fourth order scheme, and is therefore known as an RK45 integrator.

However, compared to the leapfrog integrator, which requires only one evaluation of the acceleration per timestep, the RK4 scheme requires four, and the RK45 method requires six. Therefore these methods, although correct to much higher order than the leapfrog, add significantly to the computation time required. (Note also that they are not necessarily more *accurate* either, as they are not explicitly conservative in the way that the leapfrog method is. See for instance Springel (2005); Wetzstein et al. (2009); Rosswog (2009) for a comparison of these integrators as applied to a simple Keplerian orbit evolved over many dynamical times.) The solution is to go to a lower order RKF method, where the implicit error control is still present but the number of derivative evaluations is reduced. A common choice for many SPH codes including VINE (Wetzstein et al., 2009; Nelson et al., 2009) and the one used in the code I have used, is the RK12 integrator developed by Fehlberg (1969) which proceeds as follows.

For a given variable $\mathbf{z}$, the evolution from $\mathbf{z}_i$ at time t_i to $\mathbf{z}_{i+1}$ at time $t_{i+1} = t_i + \delta t$ is given by

$$\mathbf{z}_{i+1} = \mathbf{z}_i + \left(\frac{1}{256} k_0 + \frac{255}{256} k_1 \right) \delta t, \tag{3.156}$$

where the values of k_0 and k_1 are provided by evaluating $\dot{\mathbf{z}}$ at various points, such that

$$\begin{aligned} k_0 &= \dot{\mathbf{z}}(t_i,\, \mathbf{z}_i), \\ k_1 &= \dot{\mathbf{z}}(t_i + \frac{\delta t}{2},\, \mathbf{z}_i + \frac{\delta t}{2} k_0) \end{aligned} \tag{3.157}$$

and where the dot as usual denotes differentiation with respect to time. Expansion via Taylor series shows that this is accurate to first order, with the choice of coefficients in equation 3.156 producing a leading order truncation error τ_{trunc} such that

$$\tau_{\mathrm{trunc}} = -\frac{1}{512} \delta t^2\, \ddot{\mathbf{z}}. \tag{3.158}$$

Using the values for k_0 and k_1 defined above, we can compute a further estimate $\mathbf{z}^*_{i+1}$ for $\mathbf{z}$ at time t_{i+1}, such that

$$\mathbf{z}^*_{i+1} = \mathbf{z}_i + \frac{\delta t}{2} \left(\frac{1}{512} k_0 + \frac{255}{256} k_1 + \frac{1}{512} k_2 \right), \tag{3.159}$$

with the additional value k_2 defined such that

$$\begin{aligned} k_2 &= \dot{\mathbf{z}}(t_i + \delta t,\, \mathbf{z}_i + \left(\frac{1}{256}k_0 + \frac{255}{256}k_1\right)\delta t), \\ &= \dot{\mathbf{z}}(t_{i+1}, \mathbf{z}_{i+1}). \end{aligned} \tag{3.160}$$

Again, by considering Taylor series expansions, this value $\mathbf{z}^*_{i+1}$ can be shown to be a second order estimate. One of the more appealing tricks of this method is that here k_2 is simply k_0 evaluated for the *next* timestep, and thus per timestep, only two derivative evaluations are required.

We now therefore have both a first and a second order estimate for the value of $\mathbf{z}$ at time t_{i+1}, with a known truncation error for the first order method. This can therefore be used for error control, to ensure that the timestep used is appropriate (see for instance, Press et al. 2007). However, in order for this error control to be valid, the *first* order scheme must be used for the evolution. To mitigate this, by construction this first order scheme has very small second order errors (equation 3.158), and so is effectively a quasi-second order integrator.

3.8.3 Timestepping Criteria

For either integrator, it is crucial that the timestep size is chosen correctly, both to ensure the accuracy of the evolution and to ensure numerical stability. In this section I shall briefly discuss the principal timestepping criteria in general use, and one specific to the code I have used.

3.8.3.1 CFL Criterion

By far the most general timestep criterion for gas-dynamical systems is the so-called *Courant-Friedrichs-Lewy* or CFL condition, given in it simplest form by

$$\delta t_{\mathrm{CFL}} \leq \frac{\delta x}{c}, \tag{3.161}$$

where δx is a characteristic length scale, and c is a characteristic speed (Anderson, 1995). For SPH simulations, these are both well defined; the smoothing length h provides the characteristic length, and sound speed c_{s} gives the characteristic speed

of the medium. The CFL condition for particle i then becomes

$$\delta t_{\mathrm{CFL}} \leq \frac{h}{c_{\mathrm{s}}}. \tag{3.162}$$

This has a ready physical interpretation in that it prevents spatial information transfer through the code at a rate greater than the local sound speed. In the presence of artificial viscosity this requires a slight alteration, and as such Bate et al. (1995) recommend using the following criterion;

$$\delta t_{\mathrm{CFL}} = \frac{0.3h}{c_{\mathrm{s}} + h|\nabla \cdot \mathbf{v}| + 1.2(\alpha_{\mathrm{SPH}} c_{\mathrm{s}} + \beta_{\mathrm{SPH}} h|\nabla \cdot \mathbf{v}|)}, \tag{3.163}$$

where the factors of 0.3 in the numerator and 1.2 in the denominator are empirically determined. The α_{SPH} and β_{SPH} terms are those used to determine the strength of the artificial viscosity (see Section 3.4), and it should be noted that the final term in the denominator is only included in the case where $|\nabla \cdot \mathbf{v}| < 0$. The extra $h|\nabla \cdot \mathbf{v}|$ term in the denominator accounts for the expansion or contraction of the flow, and thus explicitly allows for compressibility effects. There are variations on this theme (see for instance Deegan 2009; Monaghan 1992; Monaghan 1989) but the definition given above is the one present in the code I have used.

3.8.3.2 Force Condition

A further commonly used timestep condition is that based on the acceleration of the particle, known as the force condition. This is simple in form, and is given by

$$\delta t_{\mathrm{F}} = f_{\mathrm{F}} \sqrt{\frac{h}{|\mathbf{a}|}}, \tag{3.164}$$

where as before $\mathbf{a}$ is the particle acceleration, and $f_{\mathrm{F}} < 1$ is a tuning constant. Values for f_{F} vary from code to code but are generally in the range 0.25 - 0.5 (Wetzstein et al., 2009; Bate et al., 1995; Monaghan, 1989). The code I have used employs $f_{\mathrm{F}} = 0.3$.

3.8.3.3 Integrator Limits

Dependent on the choice of integrator, other timestep criteria may be required. In particular, if using the RKF method the timestep criterion associated with the error

correction must be incorporated. Using the method outlined above, this corresponds to a timestep of

$$\delta t_{\rm RKF} = \delta t_{\rm old} \sqrt{\frac{512\epsilon}{|\mathbf{z}_{RK2} - \mathbf{z}_{RK1}|}}, \tag{3.165}$$

where ϵ is the desired error tolerance (usually of the order of 10^{-4} - 10^{-5}) and the $\mathbf{z}_{RK1}, \mathbf{z}_{RK2}$ are the predictions for any quantity $\mathbf{z}$ from the first and second order methods within the integrator respectively. The $\delta t_{\rm old}$ term is simply the increment used for the previous step.

3.8.3.4 Generalised Timestep Criteria

A general class of additional timestep criteria may be obtained by dimensional analysis, in that for any time-varying quantity z we may define a characteristic timescale on which it varies as

$$\tau_z = \frac{z}{\dot{z}}, \tag{3.166}$$

where as usual $\dot{z}$ is the time derivative of z. To ensure that this timescale is properly resolved, we can therefore define a timestep condition such that

$$\delta t_z = f_z \frac{z}{\dot{z}}, \tag{3.167}$$

where $f_z < 1$ is a tuning factor. Although seldom required in general, a timestep criterion of this form was implemented into the code when looking at the effects of strongly varying cooling times in Chapter 5, and will be discussed in more detail there.

3.8.4 Setting the Timestep

There are therefore a variety of possible timestep choices, and thus to ensure that they are all satisfied, the timestep for each particle used is the minimum of all possibilities, i.e.

$$\delta t_i = \min(\delta t_{{\rm CFL},i}, \delta t_{{\rm F},i}, \delta t_{{\rm RKF},i}, \delta t_z). \tag{3.168}$$

Where there are only relatively small changes in the characteristic timescale, a standard choice is to use a global timestep $\delta t_{\rm glob}$, which is set by the minimum of the timesteps δt_i for the individual particles, such that

$$\delta t_{\rm glob} = \min_i (\delta t_i). \tag{3.169}$$

This has the advantage that all particles are evolved in lockstep, and thus there is no 'information lag' due to particles being on separate timesteps.

On the other hand, individual particle timestepping has the advantage of being much faster and thus more computationally efficient wherever there are large ranges in the timescales of the problem being investigated. It can however introduce instabilities into the integrator (Wetzstein et al., 2009), and can also lead to the phenomenon of low density particles on long timesteps drifting into regions of high density evolving on much shorted timesteps, leading to spurious entropy generation (Pearce, 2010, private communication). This latter effect is particularly noticeable in tests of Sedov blasts (see for instance Tasker et al. 2008), in which small entropy-driven bubbles lead to granularity in the post-shock region. This is a relatively uncommon phenomenon however, and occurs principally in the case of strongly shocked systems.

Integrator stability may be maximised (particularly for the leapfrog scheme) by using timesteps that are integer multiples of each other, and generally speaking, for a maximum timestep T, sub-timesteps will be given by $2^{-n}T$. The particle timesteps are therefore rounded *down* to the nearest relevant power of two in this case. This is the case in the code I have used for all the simulations presented in this thesis, which uses individual particle timesteps and is only weakly shocked throughout.

3.9 Summary

In this chapter I have derived the SPH algorithm from first principles, and then built it up in a series of steps to solve for pure hydrodynamical isentropic flows, dissipational flows, and finally dissipational flows under the influence of gravitational forces. Additionally I have shown that it is possible to self-consistently allow for spatially variable smoothing lengths, which allows the algorithm to be highly adaptive with the local fluid density, but to maintain exact conservation of mass, linear and angular momentum and energy, to within the integrator tolerance. In the case of isentropic flows, entropy is also conserved by construction. Furthermore I have also briefly detailed various methods of finding the nearest neighbours, and two means of evolving the fluid flow forward in time.

Since the problems I shall be considering in later chapters require only that dissipational flow in the presence of gravity to be modelled, this is all that I have covered here. However this is by no means the limit for the SPH for-

malism. Much effort has been put into including additional physics such as radiative transfer (Nayakshin et al. (2009); Petkova & Springel (2009); Forgan et al. (2009); Bisbas et al. (2009); Gritschneder et al. (2009); Pawlik & Schaye (2008) and Altay et al. (2008) to name but few of the recent efforts) and magnetic fields/MHD (see for instance Price (2010); Dolag & Stasyszyn (2009); Rosswog & Price (2007); Price & Monaghan (2005, 2004) and Price & Monaghan 2004), and this will no doubt continue as computing power steadily increases.

As with any numerical scheme however, SPH remains an approximation to reality, and as such reality checks are required in the form of standard tests. These act as calibration routines, to ensure the the results of any simulations are physically realistic, and can be relied upon. Many such tests exist, and there are far too many to do justice to here, but see for instance the astro code wiki[8], which has a number of cross-comparison tests with other codes, specifically aimed at disc-like models. As the code I use is a derivative of the one discussed in Price (2005) the discussion of numerical tests found here is particularly appropriate. A further suite of standard tests including Sod shocks and Sedov blasts among others, used for both code verification and comparison, is given in (Tasker et al., 2008).

3.9.1 Summary of Code Used

Finally I shall present here a brief overview of the code I have used throughout the remainder of this thesis:

- It is based throughout on the Lagrangian formulation detailed in Section 3.3.
- It incorporates spatially and temporally variable smoothing lengths.
- The initial code (used throughout Chapter 4) did *not* include the fully conservative correction terms (Ω_i – detailed in Section 3.5), but instead used averaged smoothing kernels and their gradients to allow for spatial variation in h. The later code (used in Chapters 5 and 6) *does* include the fully conservative corrections terms.
- The standard cubic spline kernel is used throughout (given by equation 3.31) using linear interpolation between 40,000 points equally spaced in x^2.

[8] http://www-theorie.physik.unizh.ch/astrosim/code/doku.php?id=home:home

- The internal energy is evolved (as opposed to the entropy) using an ideal gas equation of state with adiabatic index $\gamma = 5/3$ throughout.
- Energy is input to the gas through $P\mathrm{d}V$ work and through shocks (using artificial viscosity).
- The "standard" artificial viscosity (as given in equation 3.85) is used throughout, with $\alpha_{\rm SPH} = 0.1$, $\beta_{\rm SPH} = 0.2$.
- The gravitational softening length is set equal to the smoothing length throughout.
- The gravitational softening kernel is implemented as given in equation 3.149.
- The gravity force (via multipole expansion) and nearest neighbours are found using a binary tree, with efficiency $N \log N$.
- The target number of neighbours is set to 50 throughout, implemented using a fixed mass within the smoothing kernel - hence the actual number of neighbours is variable.
- Time integration is performed using a Runge-Kutta-Fehlberg integrator (found to have better energy conservation than the leapfrog for the types of simulations being run).
- All particles evolve on individual particle timesteps.
- All gas particles are of equal and constant mass.
- Sink particles are used to represent the central massive object, with gas particles accreted when they satisfy certain conditions (Bate et al., 1995). In this manner the central object can *increase* in mass and (spin) angular momentum.

4

Characterising the Gravitational Instability in Accretion Discs

We can lick gravity, but sometimes the paperwork is overwhelming.

Wernher von Braun

The material presented in this chapter has been published as

Characterising the gravitational instability in cooling accretion discs
P. Cossins, G. Lodato & C. J. Clarke, MNRAS, **393**, 1157-1173 (2009)

4.1 Introduction

In Chapter 1, we saw that for very weakly ionised discs where the MRI is ineffective and where the disc self-gravity is dynamically important, the gravitational instability is likely to be the primary driver of accretion. Where the cooling rate is sufficiently low to avoid fragmentation, spiral density waves propagate within the disc, extracting rotational and gravitational energy from the flow and returning it as heat into the disc as these waves steepen into shocks. In this manner the disc may be maintained in a marginally stable state where the Toomre parameter (equation 1.93) $Q \sim 1$. As the amplitude of these waves increases, so too does their energy density, increasing the reservoir of energy available to be returned to the disc as heat, and stabilising the disc against greater cooling rates. However, the larger amplitudes put the perturbations further into the non-linear regime, and eventually, for high enough cooling rates, the feedback process breaks down and fragmentation ensues. The quasi-steady marginal stability state therefore represents a restricted regime of dynamic thermal equilibrium, where the cooling is balanced by disc heating through gravitational instabilities.

In this chapter I seek to characterise the relationship between the strength of the cooling and the amplitude of the spiral density waves excited within the disc through self-gravity, while remaining within this dynamic thermal equilibrium state. To this end I use a Smoothed Particle Hydrodynamics (SPH) code to run global numerical simulations of self-gravitating gaseous discs. From these controlled numerical experiments, the amplitude of the density perturbations over a range of cooling times can be measured, down to the limit where the disc fragments. I then use Fourier analysis to characterise the mode spectra and pattern speeds associated with the structure formed within the disc, and to associate the dynamics of these spiral density waves with the thermodynamics of the disc self-regulation process.

4.2 Dynamics of Self-Gravitating Gaseous Discs

First however, I consider the dynamics of the spiral density waves in more detail, in particular with regard to the ability of such waves to transport energy and angular momentum. In non-axisymmetric discs, this is effected through torques arising due to the perturbed gravitational potential, usually described through a local, viscous model. It is however possible to obtain these torques directly from the energy density

of the waves themselves, and it is to these two approaches that I shall now turn.

4.2.1 The Stress Tensor

Recall that viscous accretion discs can be described by the α-formalism of Shakura & Sunyaev (1973), where (for an infinitesimally thin disc) the only non-vanishing component of the vertically integrated stress tensor T is the azimuthal shear term, given in equation 1.40 and repeated here for clarity;

$$T_{R\theta} = \nu\Sigma R\Omega' = \alpha\Sigma c_{\rm s}^2 \frac{{\rm d}\ln\Omega}{{\rm d}\ln R}. \tag{4.1}$$

As before, $\alpha \lesssim 1$ is a dimensionless parameter which measures and contains all the uncertainties concerning the viscosity. It is clear that the disc stress is linked to the local thermal pressure $\Sigma c_{\rm s}^2$, and this indicates that the α-formalism is fundamentally a local relationship. Furthermore, since α is not necessarily constant, this represents a completely general description for any purely local process. Also note that for Keplerian rotation, ${\rm d}\ln\Omega/{\rm d}\ln R = -3/2$, implying that the stress is negative, i.e. it acts to oppose rotation, and therefore allows for inward accretion flows as discussed in Chapter 1 and Clarke & Pringle (2004).

Viscous, local accretion disc theory identifies the origin of the stress with torques arising due to perturbations in a "turbulent" disc, where this turbulence may be driven by a variety of mechanisms such as the MRI or gravitational instabilities. Crucially for the gas dynamics, these perturbations manifest themselves as fluctuations in the mean flow velocity, the gravitational potential and the magnetic field threading the disc, see for instance Balbus & Papaloizou (1999); Lodato & Rice (2004); Lodato (2007). For non-magnetised self-gravitating discs such as are considered here, the stress tensor can therefore be broken down into a Reynolds stress term, associated with velocity fluctuations, and a gravitational stress term, associated with fluctuations in the gravitational potential. The Reynolds stress term $T_{R\theta}^{\rm Reyn}$ is such that

$$T_{R\theta}^{\rm Reyn} = \Sigma\langle\delta v_R \delta v_\theta\rangle, \tag{4.2}$$

where $\delta v_R, \delta v_\theta$ are the velocity fluctuations about the mean flow velocity in the R and θ directions respectively and the brackets indicate azimuthal averaging. Similarly,

the gravitational stress term $T_{R\theta}^{\rm grav}$ is given by Lynden-Bell & Kalnajs (1972) as

$$T_{R\theta}^{\rm grav} = \int \left\langle \frac{g_R g_\theta}{4\pi G} \right\rangle dz, \tag{4.3}$$

where again g_R, g_θ are the accelerations due to the perturbed gravitational potential of the disc in the R and θ directions respectively.

By comparison with equation 1.29, it is clear that the viscous torque per unit area $\dot{\mathbf{L}}_\alpha$ is related to the vertically integrated stress tensor T through

$$\dot{\mathbf{L}}_\alpha = R\nabla \cdot \mathsf{T}, \tag{4.4}$$

which in turn gives

$$\dot{\mathcal{L}}_\alpha = \frac{\partial}{\partial R}(R^2 T_{R\theta}) \tag{4.5}$$

as the only non-zero component of the torque (cf. equation 1.35). The power per unit surface $\dot{\mathcal{E}}_\alpha$ produced by this viscous torque is then given simply by (Frank et al. 2002; Pringle 1981, Section 1.2.4)

$$\dot{\mathcal{E}}_\alpha = \Omega \dot{\mathcal{L}}_\alpha, \tag{4.6}$$

where the subscript α indicates that this relation is expected for a viscous disc. Equation 4.6 therefore links the transport of angular momentum and the associated rate of work done by torques in the case of a local process, as historically modelled by the α-viscosity parameter.

4.2.2 Wave Energy and Angular Momentum Densities

In this section I shall consider the transport of energy and angular momentum through the propagation of spiral density waves. For clarity, it is useful at this point to reproduce the quadratic dispersion relation derived in Chapter 1, as it will be referred to repeatedly. It is given by

$$(\omega - m\Omega)^2 = c_{\rm s}^2 k^2 - 2\pi G\Sigma|k| + \kappa^2, \tag{4.7}$$

where as before, $c_{\rm s}$ is the fluid sound speed, Σ is the unperturbed surface density, κ and Ω are the epicyclic and angular frequencies of the fluid, ω is the wave angular frequency and k, m are the radial and azimuthal wavenumbers respectively.

However, returning to equation 1.83, away from resonances this quadratic dispersion relation (ignoring out of phase terms) can be written in the form

$$\frac{c_{\rm s}^2 k^2 + \Sigma\, \delta\Phi/\delta\Sigma}{(\omega - m\Omega)^2 - \kappa^2} = 1, \tag{4.8}$$

where $\delta\Phi$ and $\delta\Sigma$ are the perturbations to the gravitational potential and the surface density respectively. Using the self-consistent WKB solution to Poisson's equation given in equation 1.84, such that

$$\delta\Phi = -\frac{2\pi G\delta\Sigma}{|k|}, \tag{4.9}$$

these two equations can be combined to give an alternative form of the dispersion relation, $D(\delta\Sigma/\delta\Phi) = 0$, where

$$D = \frac{-k^2\Sigma}{(\omega - m\Omega)^2 - \kappa^2} + \frac{|k|}{2\pi G\delta\Sigma}. \tag{4.10}$$

From this form of the dispersion relation, one can now introduce a convenient quantity, the wave action surface density $\mathcal{A}$, defined as

$$\mathcal{A} = \frac{1}{4}\frac{\partial D}{\partial\omega}|\delta\Phi|^2 \tag{4.11}$$

(Fan & Lou, 1999; Bertin, 2000), which evaluates to

$$\mathcal{A} = \frac{m(\Omega_{\rm p} - \Omega)}{8\pi^2 G^2\Sigma}|\delta\Phi|^2, \tag{4.12}$$

(Toomre, 1969; Shu, 1970; Fan & Lou, 1999), and where as before I have introduced the pattern speed $\Omega_{\rm p}$ such that $\omega = m\Omega_{\rm p}$.

The wave energy surface density $\mathcal{E}_{\rm w}$ and the wave angular momentum surface density $\mathcal{L}_{\rm w}$ are obtained in a straightforward way from the wave action through the standard wave dynamics relations (Bertin, 2000; Shu, 1970)[1], such that

$$\mathcal{E}_{\rm w} = \omega\mathcal{A} = m\Omega_{\rm p}\mathcal{A}, \tag{4.13}$$

[1]Note that this system of equations is analogous to that found in quantum mechanics for an harmonic oscillator; from the quantum of action $\hbar$, the quantised energy E and angular momentum S are found via $E = \hbar\omega$ and $S = \hbar m$, where ω and m are the angular frequency and spin quantum number respectively.

$$\mathcal{L}_{\rm w} = m\mathcal{A}. \tag{4.14}$$

Combining the first of these relations with equations 4.12 and 4.9 one obtains

$$\mathcal{E}_{\rm w} = \frac{\Sigma v_{\rm p} \tilde{v}_{\rm p}}{2} \left(\frac{\delta\Sigma}{\Sigma} \right)^2, \tag{4.15}$$

where

$$v_{\rm p} = m\Omega_{\rm p}/k \tag{4.16}$$

$$\tilde{v}_{\rm p} = m(\Omega_{\rm p} - \Omega)/k \tag{4.17}$$

are the radial and Doppler-shifted radial phase speeds of the wave respectively. (Here and henceforth, a Doppler-shifted quantity refers to one measured in a frame co-moving with the fluid.) Note that this (density) wave energy surface density is analogous to the energy volume density of sound waves $\mathcal{E}_{\rm s}$, given by

$$\mathcal{E}_{\rm s} = \frac{1}{2}\rho c_{\rm s}^2 \left(\frac{\delta\rho}{\rho} \right)^2, \tag{4.18}$$

where now ρ and $\delta\rho$ are the unperturbed and perturbed volume densities respectively. Note also that that equations 4.15 and 4.17 together explain why self-induced density waves are launched at co-rotation – since the energy density changes sign at co-rotation, pairs of waves propagating away from co-rotation in opposite radial directions extract no net energy from the flow.

Looking again at equations 4.13 and 4.14, the relationship between the energy surface density and the angular momentum surface density in a wave is given by

$$\mathcal{E}_{\rm w} = \Omega_{\rm p}\mathcal{L}_{\rm w}. \tag{4.19}$$

In the case of quasi-stationary waves (where $\Omega_{\rm p}$ is constant) propagating in a disc in dynamic thermal equilibrium, the rate at which energy is lost per unit surface due to cooling must be balanced by the power surface density $\dot{\mathcal{E}}_{\rm w}$ dissipated by the waves. In order to maintain the amplitude of the wave, the instability has to keep extracting energy and angular momentum from the background flow. The flux of energy (angular momentum) carried by the wave is simply $\mathcal{E}_{\rm w}$ ($\mathcal{L}_{\rm w}$) times the local group velocity, which I shall consider shortly. Hence when a wave dissipates it adds energy and angular momentum to the flow in the ratio of $\mathcal{E}_{\rm w}$ to $\mathcal{L}_{\rm w}$, i.e. in the ratio

Ω_p. One can therefore conclude that

$$\dot{\mathcal{E}}_\mathrm{w} = \Omega_\mathrm{p}\dot{\mathcal{L}}_\mathrm{w} \tag{4.20}$$

and thus equation 4.20 is simply the wave analogue of equation 4.6. Comparing these two equations, a fundamental difference with respect to the viscous model can be seen – for a given torque $\dot{\mathcal{L}}$, waves extract energy from the flow at a rate proportional to the wave pattern speed Ω_p, whereas the rotation speed Ω is the underlying rate in the local (viscous) case.

Balbus & Papaloizou (1999) have similarly noted that in general, energy transport through the gravitational instability *cannot* be described purely in viscous terms, and indeed that this is only possible at co-rotation, when $\Omega_\mathrm{p} = \Omega$. This can be readily understood by noting that the wave energy surface density $\mathcal{E}_\mathrm{w}$ (equation 4.15) can be decomposed into two separate terms,

$$\mathcal{E}_\mathrm{w} = \Omega\mathcal{L}_\mathrm{w} + (\Omega_\mathrm{p} - \Omega)\mathcal{L}_\mathrm{w}, \tag{4.21}$$

$$= \frac{\Sigma}{2}\frac{m^2}{k^2}\Omega(\Omega_\mathrm{p} - \Omega)\left(\frac{\delta\Sigma}{\Sigma}\right)^2 + \frac{\Sigma}{2}\frac{m^2}{k^2}(\Omega_\mathrm{p} - \Omega)^2\left(\frac{\delta\Sigma}{\Sigma}\right)^2, \tag{4.22}$$

where as before,

$$\mathcal{L}_\mathrm{w} = \frac{\Sigma}{2}\frac{m^2}{k^2}(\Omega_\mathrm{p} - \Omega)\left(\frac{\delta\Sigma}{\Sigma}\right)^2 \tag{4.23}$$

is the wave angular momentum surface density. The first term on the RHS, equal to the wave angular momentum surface density $\mathcal{L}_\mathrm{w}$ times the rotation speed Ω, is a local energy transport term (cf. equation 4.6) and can therefore be represented using the α-formalism. The second term however, equal to the same angular momentum term times $\Omega_\mathrm{p} - \Omega$, is a non-local term. In fact the energy flux associated with this non-local transport term is is precisely that identified by Balbus & Papaloizou (1999) as an "anomalous flux", preventing self-gravitating discs from acting as pure α-discs. In the co-rotation limit as $\Omega_\mathrm{p} \to \Omega$, it is clear that transport by waves is exactly equivalent to viscous transport. However, away from co-rotation where $\Omega \neq \Omega_\mathrm{p}$, the second, non-local term becomes significant, and thus global transport becomes important within the disc.

A final quantity of importance to understanding wave transport is the (radial)

group velocity $v_{\rm g}$, which is defined as

$$v_{\rm g} = \frac{\partial \omega}{\partial k}. \tag{4.24}$$

From the dispersion relation equation 4.7, this becomes

$$v_{\rm g} = {\rm sgn}(k)\frac{|k|c_{\rm s}^2 - \pi G\Sigma}{\omega - m\Omega}, \tag{4.25}$$

where ${\rm sgn}(k)$ is simply the sign of the radial wavenumber k. With this one can now define the radial flux of energy and angular momentum $\mathcal{F}_{\mathcal{E}}$ and $\mathcal{F}_{\mathcal{A}}$ respectively as

$$\mathcal{F}_{\mathcal{E}} = v_{\rm g}\mathcal{E}_{\rm w}, \qquad \mathcal{F}_{\mathcal{A}} = v_{\rm g}\mathcal{A}_{\rm w}, \tag{4.26}$$

(Fan & Lou, 1999; Bertin, 2000). Note that the sign of the group velocity is dependent not only on the sign of k, but also whether the wave is inside or outside of co-rotation, and furthermore there is a dependence on whether the radial wavelength is greater or lesser than the most unstable wavelength $k_{\rm uns}$, which from Chapter 1 is given by

$$k_{\rm uns} = \frac{\pi G\Sigma}{c_{\rm s}^2} = \frac{1}{H_{\rm sg}}, \tag{4.27}$$

where $H_{\rm sg}$ is the self-gravitating scale height.

Laughlin et al. (1997) have suggested that once the spiral modes saturate in the non-linear regime, the dominant modes act as forcing terms for higher wavelength modes, leading to a cascade of energy through the spectrum. As such the spectral average of the excited radial modes is likely to be *greater* than the most unstable mode. With this in mind, for trailing waves ($k > 0$) launched at co-rotation, the fluxes of energy and angular momentum transport will on average be positive (i.e. outward) throughout the disc, as the average group velocity will depend only on whether the waves are inside or outside their co-rotation radius. Those waves inside co-rotation ($\Omega_{\rm p} < \Omega$) will therefore have both negative group velocity and negative energy density, whereas for those waves outside co-rotation the reverse will be true. As such the dissipation of such waves effects a net *outward* transport of energy, and by implication angular momentum.

If a wave dissipates at large radius (where $\Omega_{\rm p} \gg \Omega$) then the ratio in which energy and angular momentum are added to the disc (equation 4.20) is significantly greater than the equivalent ratio in the viscous case (equation 4.6). Consequently,

under such conditions the energy dissipated at large radii in a steady state disc with wave transport can significantly exceed that dissipated in an equivalent viscous disc – with this extra (gravitational) energy being extracted by the wave from deep in the potential and transported to large radii (Lodato & Bertin, 2001; Bertin & Lodato, 2001b). However, if waves instead dissipate close to co-rotation, the wave transport is dominated by the local term in equation 4.2.2; since energy and angular momentum transport are exchanged with the disc in roughly the same ratio as for a viscous process, then in this regime the α-formalism is a good approximation to the actual transport properties of the disc.

From equation 4.2.2 it is therefore possible to quantify a non-local transport fraction ξ from the ratio of the two terms on the RHS, such that

$$\xi = \left| \frac{\Omega - \Omega_{\rm p}}{\Omega} \right|, \tag{4.28}$$

where $\xi \approx 0$ implies locality of transport, and conversely $\xi \gg 0$ indicates significant non-local (i.e. global) transport effects. Thus in order to assess the importance of non-local effects, the relationship between the angular frequency Ω and the pattern speed $\Omega_{\rm p}$ must be known. In Section 4.5.5 I shall use the dispersion relation along with information extracted through Fourier analysis in order to estimate the pattern speed, and hence to evaluate ξ directly.

4.3 Simulating the Disc Thermodynamics

Realistically simulating the thermodynamics of accretion discs is a complex undertaking and as such has received much attention, from the opacity-based treatment employed by Johnson & Gammie (2003) through to the various convective and radiative transfer models of Boss (2004), Boley et al. (2007), Mayer et al. (2007), Stamatellos & Whitworth (2008) and Stamatellos & Whitworth (2009a), the latter two of which also account for heating from the central star.

However in this chapter, the aim is to investigate the relationship between the properties of the density perturbations and the rate at which the disc cools. This purpose is served most readily by *imposing* a known cooling rate against which the heating rate and subsequent disc structure may be easily correlated. It is therefore not necessary to consider the exact physics of the cooling regimes found in astrophysical discs, and hence I use a cooling law for the heat loss rate per unit mass $\dot{Q}^-$

such that

$$\dot{Q}^{-} = -\frac{u}{t_{\rm cool}}, \tag{4.29}$$

where u is the specific internal energy and where the details of the cooling function (and possibly of any additional *external* heating) are absorbed into the simple parameter $t_{\rm cool}$. As long as such a characteristic timescale can be defined, it is therefore possible to use this formalism to represent a wide range of cooling mechanisms. Within this chapter I use a fixed ratio between the local dynamical and cooling timescales, such that $\beta = \Omega t_{\rm cool}$ is constant. This form of cooling has been used extensively in simulations of discs in various contexts, for example Gammie (2001), Lodato & Rice (2005), Hobbs & Nayakshin (2009), and has proven useful in elucidating the properties of the gravitational instability in controlled numerical experiments.

In terms of the heating imparted by the gravitational instability, it was noted in the previous section that density waves extract energy from the disc. In addition to compression heating (which should in any case be roughly counterbalanced by the corresponding rarefaction once the wave has passed), in the case where the pattern speed differs from the rotation speed by more than the local sound speed the waves will steepen into shocks, liberating further heat into the disc. One may expect the rate at which energy is added to the disc to scale with the energy of the wave and the local dynamical timescale, and therefore the specific heating rate $\dot{Q}^{+}$ due to the instability can be expressed as

$$\dot{Q}^{+} = \frac{1}{\Sigma}\epsilon\Omega|\mathcal{E}_{\rm w}| = \epsilon\Omega c_{\rm s}^{2}\frac{\mathcal{M}\widetilde{\mathcal{M}}}{2}\left(\frac{\delta\Sigma}{\Sigma}\right)^{2}, \tag{4.30}$$

where the radial and Doppler shifted radial phase Mach numbers are defined to be $\mathcal{M} = |v_{\rm p}|/c_{\rm s}$ and $\widetilde{\mathcal{M}} = |\tilde{v}_{\rm p}|/c_{\rm s}$ respectively. Here I have also introduced a dimensionless proportionality factor ϵ, hereinafter referred to as the shock heating factor. If the relationship between the pattern speed and the angular speed is self-similar (i.e., it does not vary across the disc), one would expect ϵ to be constant, independent of radius.

Once the gravitational instability has been instigated and has subsequently saturated, the disc may be assumed to be in dynamic thermal equilibrium such that the rate at which energy is released through wave-driven shock heating is balanced by the imposed cooling rate, i.e. $\dot{Q}^{-} + \dot{Q}^{+} = 0$. Recalling that $u = c_{\rm s}^{2}/\gamma(\gamma-1)$, one can equate equations 4.30 and 4.29 and thereby determine the following relationship

between the amplitude of the density perturbations and strength of the cooling, as measured by the β parameter;

$$\left(\frac{\delta\Sigma}{\Sigma}\right)^2 = \frac{2}{\epsilon\beta}\frac{1}{\gamma(\gamma-1)}\left(\frac{1}{\mathcal{M}\widetilde{\mathcal{M}}}\right). \tag{4.31}$$

I shall therefore use global numerical simulations to test the above energy balance, and to investigate the relative magnitude of the local and and non-local transport terms.

4.4 Numerical Set-Up

4.4.1 The SPH Code

All of the simulations presented hereafter were performed using a 3D smoothed particle hydrodynamics (SPH) code, a Lagrangian hydrodynamics code capable of modelling self-gravity (see for example, Benz 1990; Monaghan 1992), full details of which are given in Chapter 3. It should be noted that the code used for this chapter did *not* employ the fully conservative form of the SPH formalism, as it did not take into account to so-called ∇h terms referred to in Section 3.5 but used instead an averaged smoothing kernel. However, as the variation in smoothing lengths across the disc was found to be reasonably smooth this is unlikely to have had a significant effect.

Note that within the SPH formalism (see Chapter 3), the integral of a physical quantity A over a given volume V is estimated by the sum over the individual particle values of this quantity, as below;

$$\int_V A\, dV \approx \sum_i \frac{m_i}{\rho_i} A_i, \tag{4.32}$$

where m_i is the particle mass, ρ_i is the particle volume density and i loops over all the particles within the volume V. In a similar manner, note that a volume-averaged value for A, which I shall call $\bar{A}$, can therefore be estimated via

$$\bar{A} \approx \sum_i \frac{A_i}{\rho_i} \Bigg/ \sum_i \frac{1}{\rho_i}, \tag{4.33}$$

where, as in all the simulations, all particles have equal masses.

The disc systems were modelled as a single point mass (on to which gas particles may accrete if they enter within a given sink radius, and satisfy certain boundness conditions – see Bate et al. 1995), orbited by 500,000 SPH gas particles; a set up common to many other SPH simulations of such systems, (e.g. Lodato & Rice 2004, 2005; Rice et al. 2003a; Clarke et al. 2007) but with increased resolution. The central object is free to move under the gravitational influence of the disc. In order to ensure the simulations were properly converged, resolution checks were undertaken with discs consisting of both 250,000 and 1,000,000 particles – these are discussed briefly in Appendix B.

As described in Section 4.3 I use a simple cooling model, implemented in the following manner

$$\frac{du_i}{dt} = -\frac{u_i}{t_{\mathrm{cool},i}}, \tag{4.34}$$

where the u_i and $t_{\mathrm{cool},i}$ are the specific internal energy and the cooling time associated with each particle respectively. Again as above the functional form of the cooling time is kept simple, such that $\Omega_i t_{\mathrm{cool},i} = \beta$, where Ω_i is the angular velocity of each particle, and where β is held constant throughout any particular simulation. All simulations have been run modelling the particles as a perfect gas, with the ratio of specific heats $\gamma = 5/3$, heat addition being allowed for via $P\mathrm{d}V$ work and shock heating and with the cooling implemented as specified above. Artificial viscosity has been included through the standard SPH formalism, with $\alpha_{\mathrm{SPH}} = 0.1$ and $\beta_{\mathrm{SPH}} = 0.2$. Note that these values are smaller than those commonly used in SPH simulations; I use these values to limit the transport induced by artificial viscosity. As shown in Lodato & Rice (2004), with this choice of parameters the transport of energy and angular momentum due to artificial viscosity is a factor of 10 smaller than that due to gravitational perturbations, while the weak shocks occurring in the simulations are still adequately resolved.

By using the cooling prescription outlined above, the rate at which the disc cools is governed by the dimensionless parameter β and the cooling is thus implemented scale free. The governing equations of the entire simulation can likewise be recast in dimensionless form. In common with the previous SPH simulations mentioned above, I define the unit mass to be that of the central object – the total disc mass and individual particle masses are therefore expressed as fractions of the central object mass. It is possible to self-consistently define an arbitrary (cylindrical) scale radius R_0, such that the radius R in code units is related to the physical radius r

via $r = RR_0$, and thus, with $G = 1$, the unit time is the dynamical time $t_{\rm dyn} = \Omega^{-1}$ at unit (code) radius $R = 1$.

4.4.2 Initial Conditions

All the simulations model a central point object of unit mass $M_* = 1$, surrounded by a gaseous disc of mass $M_{\rm disc}$. Although the bulk of the simulations have been conducted with a disc to central object mass ratio $q = M_{\rm disc}/M_*$ of 0.1, simulations were also run with $q = 0.05, 0.075$ and $q = 0.125$ to investigate the effects of the mass ratio on the non-local transport fraction ξ.

All the simulations run used an initial mass surface density profile $\Sigma \propto R^{-3/2}$, which implies that in the marginally stable state where $Q \approx 1$, the disc temperature profile should be approximately flat. Since the surface density evolves on the viscous time $t_{\rm visc} \gg t_{\rm dyn} = \Omega^{-1}$ this profile remains roughly unchanged throughout the simulations. The initial temperature profile is $c_{\rm s}^2 \propto R^{-1/2}$ and is such that the minimum value of the Toomre parameter $Q_{\rm min} = 2$ occurs at the outer edge of the disc. In this manner the disc is initially gravitationally stable throughout. Note that the disc is *not* initially in thermal equilibrium – heat is not input to the disc until gravitational instabilities are initiated.

Radially the disc extends from $R_{\rm in} = 0.25$ to $R_{\rm out} = 25.0$, as measured in the code units described above. The disc is initially in approximate vertical hydrostatic equilibrium with a Gaussian distribution of particles with (non-self-gravitating) scale height $H_{\rm nsg} = c_{\rm s}/\Omega$. The azimuthal velocities take into account both a pressure correction (see for instance Lodato, 2007) and the enclosed disc mass. In both cases, any variation from dynamical equilibrium is washed out on the dynamical timescale. Given the dimensions above, one outer dynamical timescale of the disc corresponds to 125 time units. To ensure that thermal equilibrium is reached and that the gravitational instability is saturated, all (non-fragmenting) simulations are followed for at least 10 outer cooling times. To this end I shall refer to the thermal time $t_{\rm therm}$ for each simulation as the cooling time evaluated at the initial outer edge of the disc, taken to be at $R = 25$ – thus $t_{\rm therm} = t_{\rm cool}(25) = 125\beta$ in code units.

4.4.3 Simulations Run

In all a total of ten distinct simulations were run for various values of the cooling parameter β and the disc to central object mass ratio q, as detailed in Table 4.1.

β	$q = M_{\rm disc}/M_*$	No. of Particles	Duration
4	0.10	500,000	4.0 $t_{\rm therm}$
5	0.10	500,000	10.0 $t_{\rm therm}$
6	0.10	500,000	10.0 $t_{\rm therm}$
7	0.10	500,000	10.0 $t_{\rm therm}$
8	0.10	500,000	10.0 $t_{\rm therm}$
9	0.10	500,000	10.0 $t_{\rm therm}$
10	0.10	500,000	10.0 $t_{\rm therm}$
5	0.050	500,000	10.0 $t_{\rm therm}$
5	0.075	500,000	10.0 $t_{\rm therm}$
5	0.100	500,000	10.0 $t_{\rm therm}$
5	0.125	500,000	10.0 $t_{\rm therm}$

TABLE 4.1: *Details of numerical simulations. Note that the duration is quoted in terms of the thermal time, equivalent to the cooling time at the outer radius* $\approx 125\beta$ *code units. The* $\beta = 4$ *case fragmented, and therefore did not run for as long as the other cases.*

Although previous investigations with $\Sigma \propto R^{-1}$ have found that the fragmentation boundary is at $\beta_{\rm frag} \approx 6$, (Rice et al., 2005, 2003a), I find that in the case where $\Sigma \propto R^{-3/2}$ the fragmentation boundary is slightly different, with $4 < \beta_{\rm frag} < 5$. (Note that this is discussed in more detail in Chapter 5.) The simulation where $\beta = 4$ therefore contains a fragment, and is included primarily for completeness. All results henceforth are given at the time quoted in the final column of Table 4.1 unless otherwise stated. The raw data are time-averaged over 500 unit times about these values to enhance the signal-to-noise ratio and to give the approximate steady-state values.

4.5 Simulation Results

Common to all the simulations is an initial phase in which the discs cool rapidly until the value of Q becomes approximately unity, at which point the gravitational instability is initiated and heat is liberated to balance the cooling. This stage is complete after approximately one thermal time, and from then on the discs settle into a quasi-steady state with $Q \approx 1$, characterised by the presence of spiral arms throughout almost the entire radial range. The quasi-static Q profiles to which the discs converge are shown in Fig. 4.1, with the cooling parameter β varying in the top panel, and the disc to central object mass ratio q varying in the bottom panel. Note that the data are plotted at the times given in Table 4.1. Throughout all the

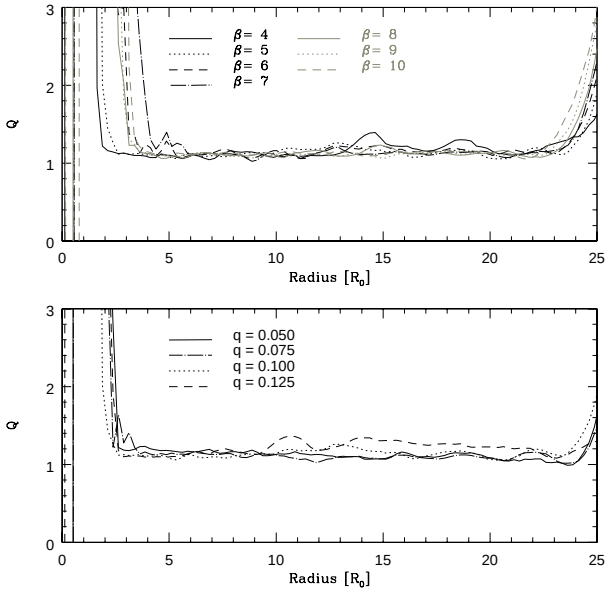

FIGURE 4.1: *Profiles of Q against radius for different values of the cooling parameter β (top) and mass ratio q (bottom) plotted at the times quoted in Table 4.1.*

simulations it can be seen that the discs self-regulate to the marginally stable $Q \approx 1$ condition over a large range of radii.

Once the disc has reached a quasi-steady state, the disc aspect ratio $H_{\rm nsg}/R$ also stabilises to the value predicted by the self-regulation condition $Q \approx 1$,

$$\frac{H_{\rm nsg}}{R} \approx \frac{\pi \Sigma(R) R^2}{M_*}, \tag{4.35}$$

which is shown as a function of radius in Fig. 4.2 for different values of β (top panel) and q (bottom panel).

4.5.1 Saturation Amplitude of the Instability

Once the gravitational instability has been initiated, for the simulation where $\beta \leq \beta_{\rm frag}$ (i.e. where $\beta = 4$) the amplitude of the perturbations required to balance the cooling rises to the point where the self-regulation mechanism breaks down, leading to the fragmentation of the disc into bound objects. In the cases where $\beta > \beta_{\rm frag}$ however, the amplitude increases on the dynamical timescale until the disc reaches dynamic thermal equilibrium, at which point the amplitude of the surface density

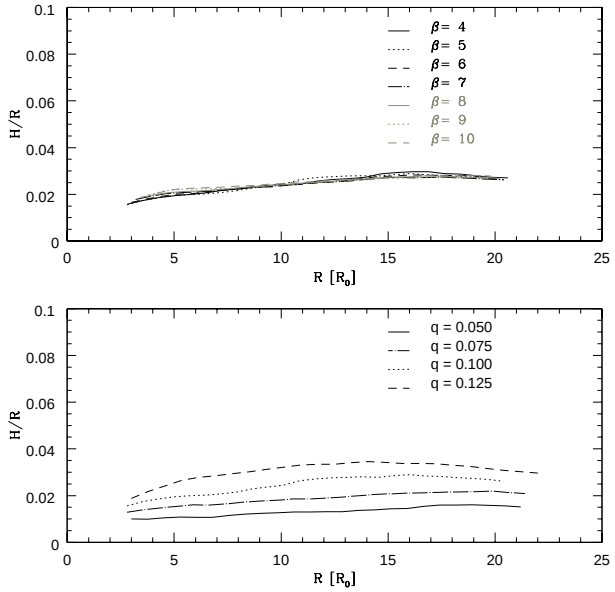

FIGURE 4.2: *Disc scale height over radius (H/R) plotted as a function of radius for varying β (top) and q (bottom). Note that in all cases $H/R \approx q/4$ as expected.*

fluctuations becomes constant, as the heating they provide balances the imposed cooling. This is observed in all simulations where $\beta \geq 5$.

From the simulations the prediction for the saturation amplitude provided by equation 4.31 can now be tested numerically. Fig. 4.3 shows images of the surface density of the disc for the two cases $\beta = 5$ and $\beta = 10$, respectively, where in both cases the mass ratio is $q = 0.1$. It can be seen that, while the overall disc structure remains essentially constant (as confirmed by a more detailed Fourier analysis, see below), the spiral wave amplitude as characterised by the surface density contrasts appears to decrease with increasing β. Noting that as the direction of rotation of the discs is anticlockwise, in all the simulations run the waves excited are trailing waves – they all point in opposition to the direction of rotation.

While SPH allows us to conduct a global 3D simulation of discs with relative ease, it does not readily permit the direct calculation of intrinsically two dimensional quantities, such as the surface density perturbation amplitude $\delta\Sigma/\Sigma$. Therefore, in order to calculate this quantity, I overlay a cylindrical grid on the disc such that each cell contains approximately $N_{\rm neigh}$ particles, where $N_{\rm neigh} \approx 50$ is the average number of neighbours within a smoothing kernel for the simulations. For each annulus of cells the average surface density $\bar{\Sigma}$ can therefore be calculated, and by comparing

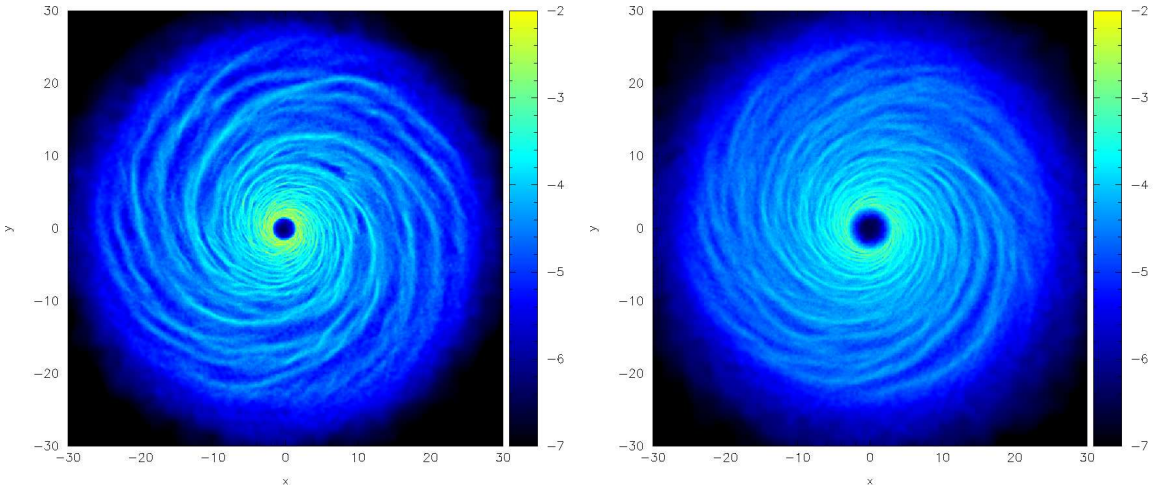

FIGURE 4.3: *Surface density structures for discs where the mass ratio $q = 0.1$, with $\beta = 5$ (left) and $\beta = 10$ (right). The logarithmic scales show surface density contours from 10^{-7} to 10^{-2} in code units. Note that the direction of rotation is anticlockwise, and that the plots are given at the times quoted in Table 4.1.*

---

this to the value calculated for each cell within this annulus an annulus averaged RMS value is evaluated for the perturbation amplitude $\delta\Sigma/\bar{\Sigma}$. This is shown as a function of radius R and the cooling parameter β in Fig. 4.4.

From Fig. 4.4, it is clear that there is an increasing trend in $\delta\Sigma/\bar{\Sigma}$ with decreasing β and that furthermore, away from the disc boundaries the saturation amplitude is approximately constant with radius. The low values for the perturbation amplitude at small radii ($R \lesssim 5$) are probably due to the increased number of particles per grid cell smoothing out the underlying variation. The strength of the surface density perturbation can be characterised by simply averaging $\delta\Sigma/\bar{\Sigma}$ over the self-regulated portion of the disc, which I define as $5 \leq R \leq 25$ (cf. Fig. 4.1). Fig. 4.5 shows the relation between the azimuthally and radially averaged amplitude, denoted by $\langle\delta\Sigma/\bar{\Sigma}\rangle$, and the cooling parameter β. Each point represents a single simulation, while the curve shows the best fit to the data using the inverse square root dependence predicted by equation 4.31. From the simulations I therefore obtain the following empirical relationship, for the case where $q = 0.1$;

$$\left\langle \frac{\delta\Sigma}{\bar{\Sigma}} \right\rangle \approx \frac{1.0}{\sqrt{\beta}}. \tag{4.36}$$

In a similar manner variation in $\delta\Sigma/\bar{\Sigma}$ with q can also be obtained, and this is shown in Fig. 4.6. It is clear that the strength of the perturbation tends to increase

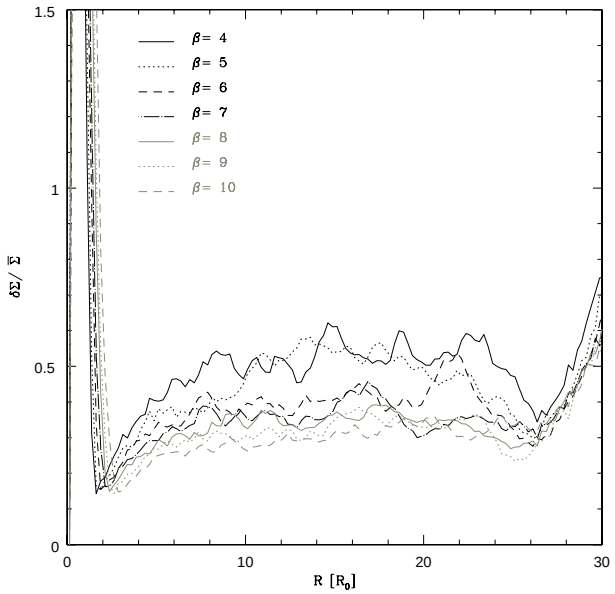

FIGURE 4.4: *Variation of the relative mass surface density perturbation amplitude* $\delta\Sigma/\bar{\Sigma}$ *with radius for various values of the cooling parameter* β*. All data plotted at the times shown in Table 4.1.*

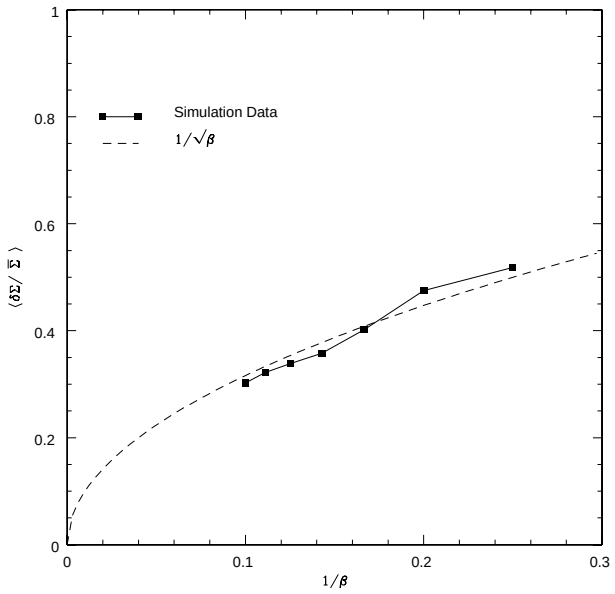

FIGURE 4.5: *Variation of the radially and azimuthally averaged relative surface density perturbation amplitude* $\delta\Sigma/\bar{\Sigma}$ *with the inverse cooling parameter* $1/\beta$*. The radial average is calculated over the range* $5 \leq R \leq 24$*.*

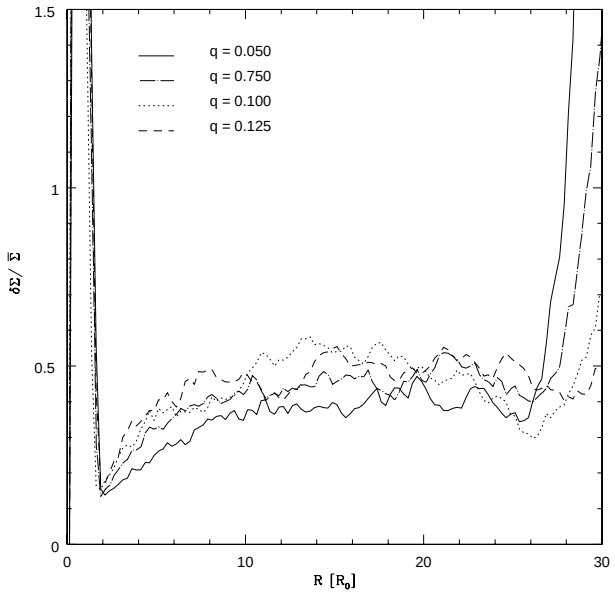

FIGURE 4.6: *Variation of the relative mass surface density perturbation amplitude $\delta\Sigma/\bar{\Sigma}$ with radius for various values of the disc to central object mass ratio q.*

with the mass ratio q, although this dependence on q is rather less than linear.

4.5.2 Fourier Analysis: Azimuthal Structure

From the simulations I have found empirically that the perturbation strength $\left\langle \delta\Sigma/\bar{\Sigma} \right\rangle$ follows a $\beta^{-1/2}$ relationship, as predicted by equation 4.31. However, this equation also shows a dependence on the wave modes excited within the disc through the action of the gravitational instability, via the phase Mach numbers. To elucidate this relationship further I have therefore conducted a Fourier analysis of the wave modes in the disc, a full description of which may be found in Appendix C. In this section I therefore consider the effects of both the cooling (via β) and the disc to central object mass ratio q on the excitation of the azimuthal m wavenumbers – the next section will describe the excitation of the radial wavenumbers.

In general, whatever the imposed cooling regime, for a given mass ratio of $q = 0.1$ the distribution of the azimuthal wavenumbers determined by the gravitational instability remains approximately constant, with the dominant mode at around $m \approx 5$. The mode distributions at five radii throughout the disc for the cases where $\beta = 4, 6, 8$ and 10 are shown in Fig. 4.7. It is clear that the spectral

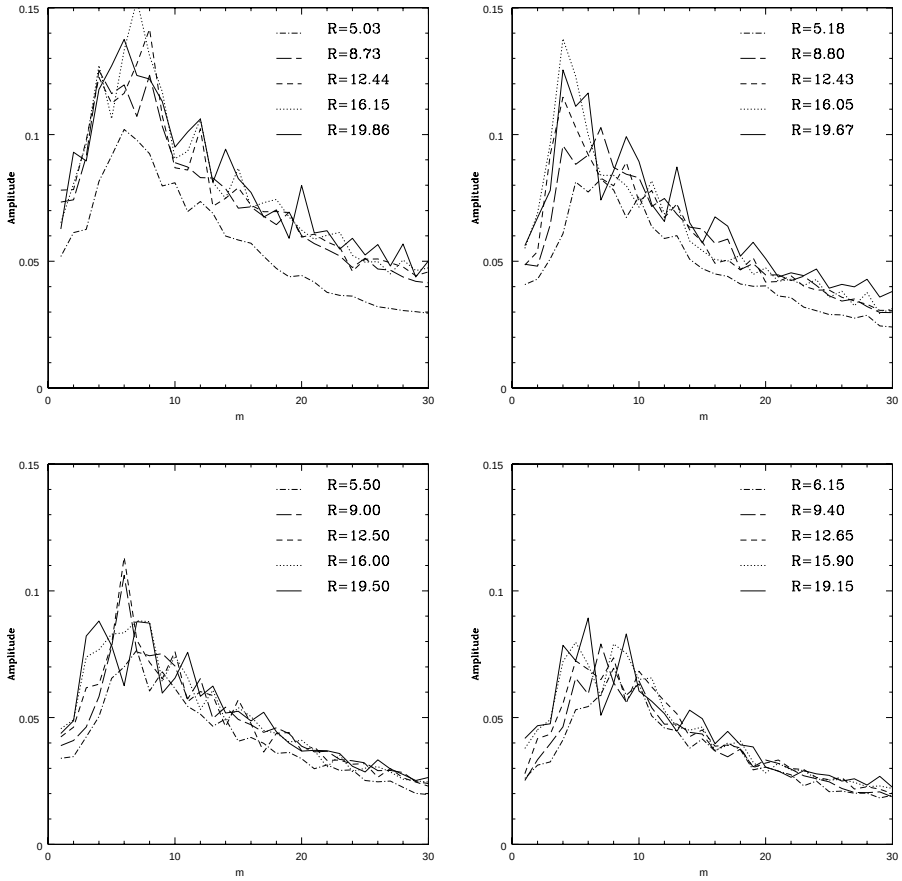

FIGURE 4.7: *Azimuthal mode amplitudes excited at various radii where $\beta = 4$ (top left), $\beta = 6$ (top right), $\beta = 8$ (bottom left) and $\beta = 10$ (bottom right).*

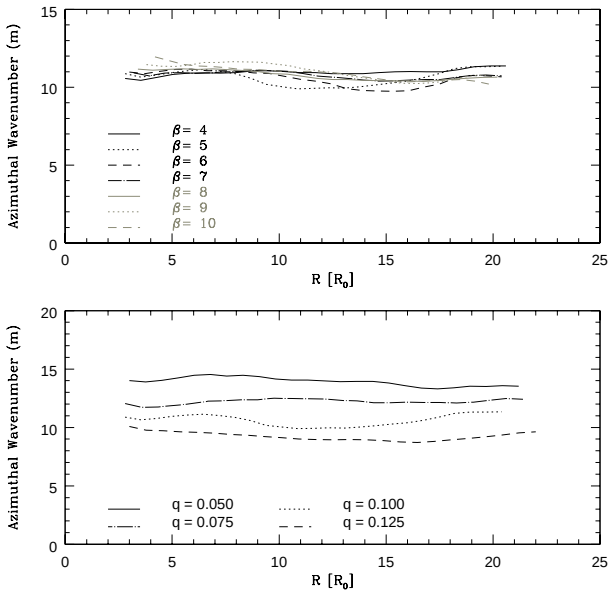

FIGURE 4.8: *Variation of the average azimuthal wavenumber excited as a function of wavenumber for $\beta = 4 - 10$ where $q = 0.1$ (top) and as q varies with $\beta = 5$ (bottom).*

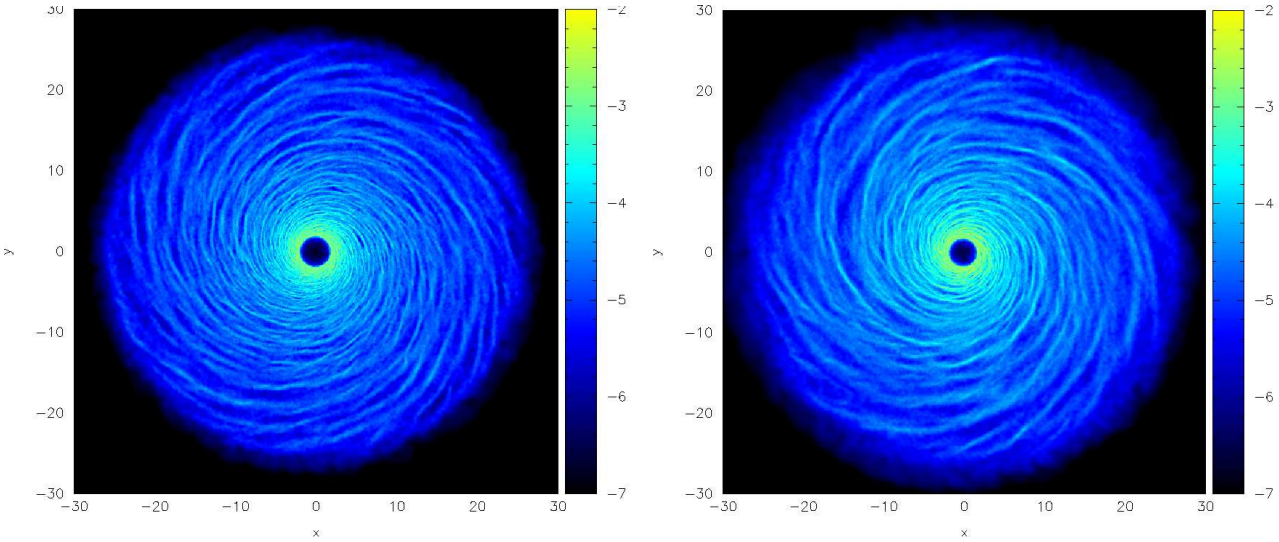

FIGURE 4.9: *Surface density structures for discs with $q = 0.05$ (left) and $q = 0.125$ (right), with $\beta = 5$. The logarithmic scale shows mass surface density contours from 10^{-7} to 10^{-2} in code units. These therefore form a direct comparison with Fig. 4.3, where $q = 0.1$, $\beta = 5$. Once again the direction of rotation of the discs is anticlockwise.*

distribution of the modes shows little variation with either radius or with the imposed cooling, except that the amplitude of the modes decreases as β increases, as expected in view of the decreased perturbation amplitudes seen in Figs. 4.3, 4.4 and 4.5. Additionally, the top panel of Fig. 4.8 shows the variation of the (power-weighted) average wavenumber against radius for all the values of β that have been tested, and note that although some small variations are seen, they are uncorrelated with the imposed cooling.

The bottom panel of Fig. 4.8 shows the variation in the average azimuthal modes excited as the disc to central object mass ratio q varies, while the cooling is held constant at $\beta = 5$. It can be seen from the plot that variation of this parameter does have a marked effect on the power spectrum of the waves – the average mode number varies inversely with the mass ratio, from $m_{\rm av} \approx 15$ where $q = 0.05$ to $m_{\rm av} \approx 10$ where $q = 0.125$. This variation is also clearly seen in Fig. 4.9, where a large number of flocculent arms are present in the disc with $q = 0.05$, and fewer, rather more well-defined spiral arms appear in the disc where $q = 0.125$ (cf. the similar result obtained in Lodato & Rice 2004). By comparison, the left panel of Fig. 4.3 shows a disc with $\beta = 5$ and $q = 0.1$, and the pattern of spiral arms present is intermediate to those shown in Fig. 4.9.

4.5.3 Fourier Analysis: Radial Structure

I now consider the radial wavenumbers k of the waves excited by the gravitational instability. In contrast to the azimuthal modes, it is clear from Figs. 4.3 and 4.9 that there is significant variation in the radial wavenumber k with radius, and Fig. 4.9 suggests that there is an additional variation with the disc to central object mass ratio.

Fig. 4.10 shows the variation in the power spectrum of different radial wavenumbers for the cases where $\beta = 4, 6, 8$ and 10 and $q = 0.1$, at the same radii as the azimuthal wavenumbers shown in Fig. 4.7. As with the azimuthal modes, there is little overall change in the spectral distribution of the modes with varying β excepting that the amplitudes of the modes decrease as the cooling weakens. Conversely however, these plots show a significant variation with radius, in that the peak wavenumber decreases with increasing radius, and thus the dominant wavelength similarly increases with radius.

Fig. 4.11 on the other hand shows the power spectrum as a function of $kH_{\rm sg}$, where the wavenumber k is normalised to the expected most unstable wavenumber,

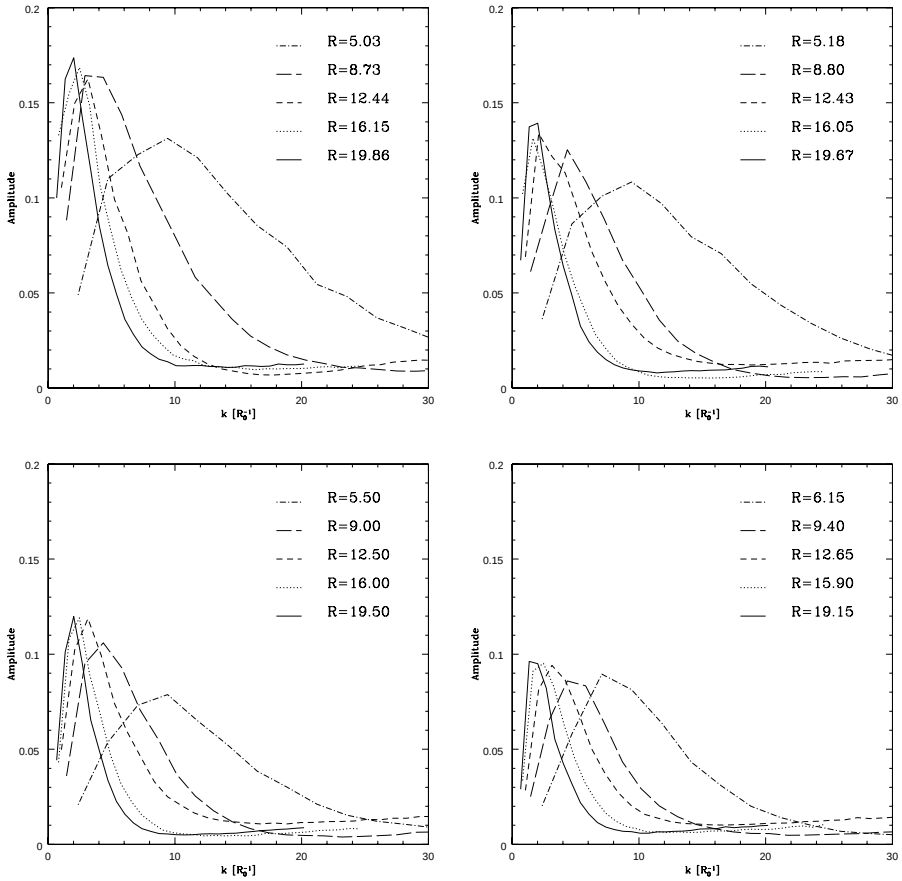

FIGURE 4.10: *Radial mode amplitudes excited at various radii where $\beta = 4$ (top left), $\beta = 6$ (top right), $\beta = 8$ (bottom left) and $\beta = 10$ (bottom right).*

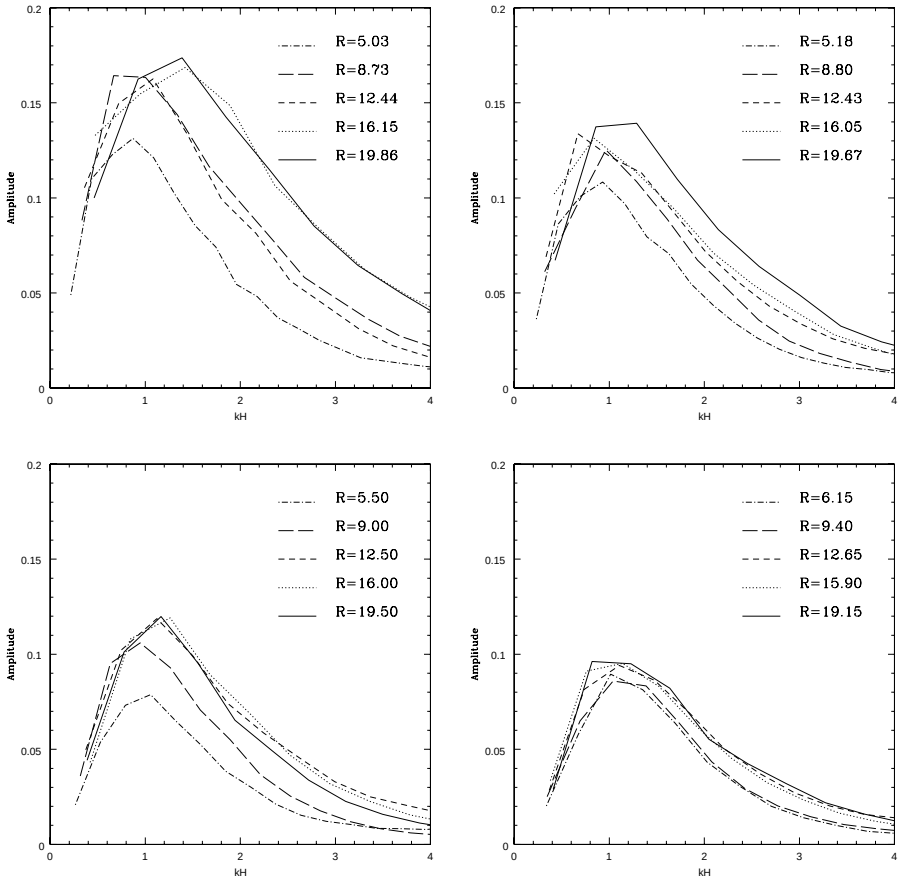

FIGURE 4.11: *Mode amplitudes plotted against the product of the peak radial modenumber and the disc scale height H excited for various radii where $\beta = 4$ (top left), $\beta = 6$ (top right), $\beta = 8$ (bottom left) and $\beta = 10$ (bottom right).*

$k_{\rm uns} = H_{\rm sg}^{-1}$ (see Section 2.1). This confirms the expectations from the linear WKB approach, as these plots show a clear peak at $kH_{\rm sg} \approx 1$. Note also that the distribution of wavenumbers is not symmetric about $kH_{\rm sg} = 1$, implying that in accordance with the results of Laughlin et al. (1997) and the assumptions in Section 4.2.2, there *is* a cascade of energy from the most unstable wavelength to shorter wavelengths. Taken in conjunction with the fact that $m \neq 0$, this peak at $kH_{\rm sg} \approx 1$ also suggests that throughout the disc the waves that are excited are close to co-rotation. Fig. 4.12 (top panel) shows the average radial wavenumber as a function of radius for all the values of β considered, again confirming the trends already discussed and also further showing that, excepting the variation in amplitude discussed above, the structure excited by the simulation is essentially independent of the cooling imposed. It also hows that the simulation to simulation scatter is very small.

The bottom panel of Fig. 4.12 shows that, as with the azimuthal wavenumbers, there is clear variation in the power-weighted average radial wavenumber $k_{\rm av}$ with the disc to central object mass ratio for a given β (in this case $\beta = 5$); increasing the mass ratio decreases the average wavenumber in approximately inverse proportion. The dashed grey line in the bottom panel of Fig. 4.12 plots a sample $R^{-3/2}$ curve, indicating that the average wavenumber follows a power-law distribution with radius, such that $k_{\rm av} \sim R^{-3/2}$, which remains constant with varying mass ratio. This is easily understood by noting that since the sound speed $c_{\rm s}$ is approximately constant by construction, equation 4.27 indicates that $k \sim \Sigma \sim R^{-3/2}$.

4.5.4 Mach Number of the Spiral Modes

Returning briefly to the dispersion relation given in equation 4.7, we note that this is only strictly valid for infinitesimally thin discs. As the simulations are fully three dimensional, a correction to the self-gravity term is required to account for this, as discussed in Chapter 1 and Bertin (2000). For clarity this is repeated from equation 1.96 below,

$$m^2(\Omega_{\rm p} - \Omega)^2 = c_{\rm s}^2 k^2 - \frac{2\pi G\Sigma|k|}{1 + |k|H_{\rm sg}} + \Omega^2, \tag{4.37}$$

where I have also used the fact that the discs are approximately Keplerian, and thus $\kappa \approx \Omega$. Recall that the reduction factor of $1/(1 + |k|H_{\rm sg})$ arises from the vertical dilution of the gravitational potential due to the finite thickness $H_{\rm sg}$ of the disc (Bertin, 2000; Binney & Tremaine, 2008; Vandervoort, 1970b). In contrast

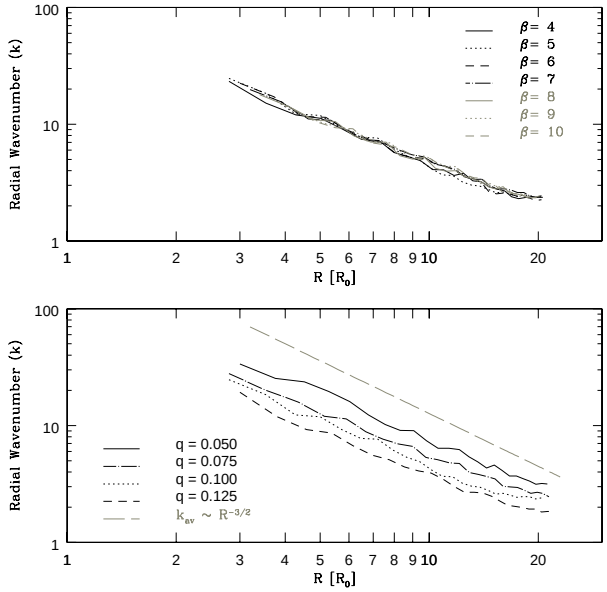

FIGURE 4.12: *Variation of the average radial wavenumber as a function of radius for* $\beta = 4 - 10$ *and* $q = 0.1$ *(top) and as* q *varies with* $\beta = 5$ *(bottom).*

to Chapter 1, here I have explicitly set the scale height to be the self-gravitating scale height $H_{\rm sg}$, for consistency with the other parameters calculated via the Fourier analysis. Using this finite-thickness dispersion relation and the averaged values for k and m it is possible to calculate a spectrally averaged Doppler-shifted angular speed $|\Omega_{\rm p} - \Omega|$, noting that the sign of $\Omega_{\rm p} - \Omega$ cannot be determined from equation 4.37. Since the average radial wave-number is generally very close to the most unstable one, and since the disc is almost exactly marginally stable, the resultant average pattern speed turns out to be always very close to co-rotation. I shall quantify the deviation of the pattern speed from co-rotation later in Section 4.5.5.

Furthermore, the radial and Doppler-shifted radial phase Mach numbers can be calculated, and these are shown in Fig. 4.13. The upper panel shows the wave radial phase Mach number $\mathcal{M}$ (thick lines) and the Doppler shifted radial phase Mach number $\widetilde{\mathcal{M}}$ (thin lines) as functions of radius for various values of β with $q = 0.1$. Similarly, the lower panel of Fig. 4.13 shows the variation of these Mach numbers with the mass ratio q for $\beta = 5$. Immediately it is clear that both quantities are independent of the cooling rate as measured by β with very little scatter. Moreover, the Doppler-shifted phase Mach number is very close to unity. In a similar manner this quantity remains unchanged with variations in the mass ratio, although the

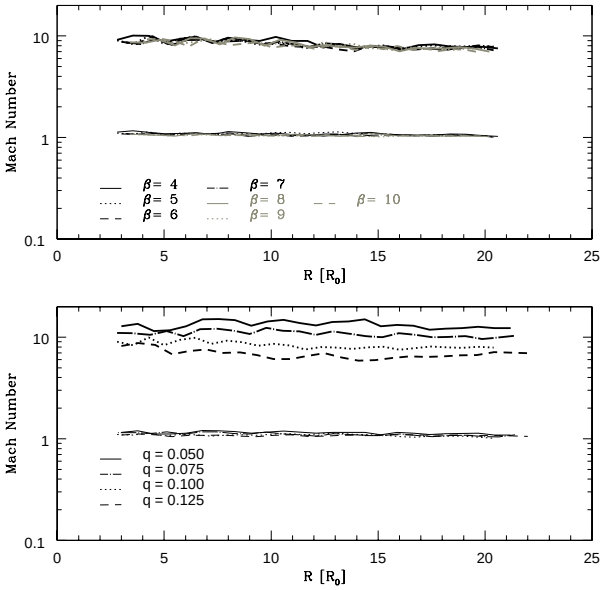

FIGURE 4.13: *Wave phase Mach number $\mathcal{M}$ (thick lines) and the Doppler-shifted phase Mach number $\widetilde{\mathcal{M}}$ (thin lines) as a function of radius for various values of β with $q = 0.1$ (top) and as q varies with $\beta = 5$ (bottom).*

phase Mach number decreases with increasing q.

The above results essentially imply that the wave structure is determined by the requirement that the normal component of the flow into the shock is almost exactly sonic – a natural criterion for a quasi-steady system due to the dissipative nature of shocks. For waves with winding angle i, and Doppler-shifted radial phase speed $\tilde{v}_\mathrm{p}$, a sonic normal component of velocity into the shock implies $\tilde{v}_\mathrm{p} \cos i = c_\mathrm{s}$, leading to

$$\widetilde{\mathcal{M}} = \frac{1}{\cos i}. \tag{4.38}$$

Hence, in the limit of tightly wound waves where $\cos i \approx 1$, one would expect that $\widetilde{\mathcal{M}} \approx 1$, as indeed is found in Fig. 4.13. For completeness, Fig. 4.14 shows the winding angle i as a function of radius for varying β (top) and mass ratio q, (bottom), using the definition $\tan i = m/kR$. In all cases, $i \lesssim 15°$, so the waves are reasonably tightly wound throughout. Again there is no significant variation with cooling, but the structure becomes more open as the mass ratio increases, as expected from Figs. 4.3 and 4.9.

One can also use equation 4.31 to estimate the amount of energy dissipated by

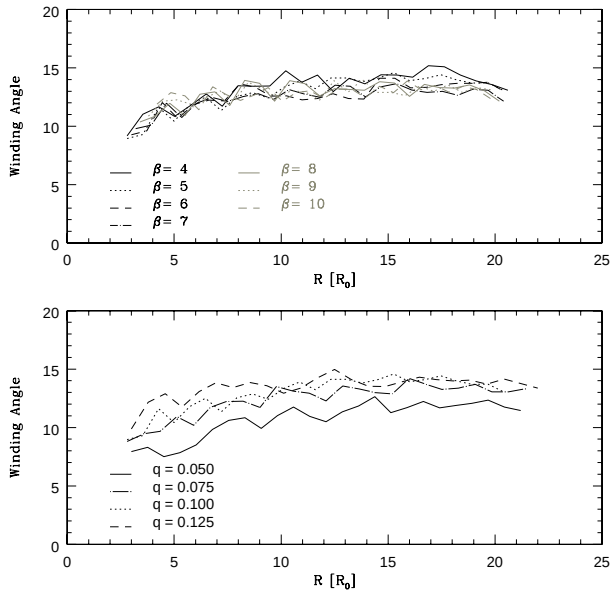

FIGURE 4.14: *Wave winding angle i as a function of radius plotted against varying cooling (top) and mass ratio (bottom).*

---

the weak spiral shocks per dynamical time as characterised by ϵ, which is shown in Fig. 4.15 and is found to be roughly 20% of the available wave energy. Through the constancy of the Doppler-shifted phase Mach number it can be seen that the shock structure that forms in the disc is indeed self-similar and thus the heating factor ϵ is also largely independent of the applied cooling, the mass ratio and the radial position. Note that the larger values for ϵ generated at low radii ($R \lesssim 5$) are probably due to the inaccuracies in calculating $\delta\Sigma/\bar{\Sigma}$ in this region rather than a breakdown in self-similarity.

4.5.5 The Locality of Transport Induced by Self-Gravity

In the previous subsection I noted that the (spectrally averaged) pattern speed of the waves $\Omega_{\rm p}$ is always very close to the angular velocity of the flow Ω, thereby indicating that the waves are close to the co-rotation resonance as suggested earlier by the results of the radial mode decomposition. One can estimate more quantitatively how close to co-rotation the spiral waves lie by calculating the quantity ξ, as given in equation 4.28. This is shown in Fig. 4.16 as a function of radius for all the values of β and mass ratio simulated.

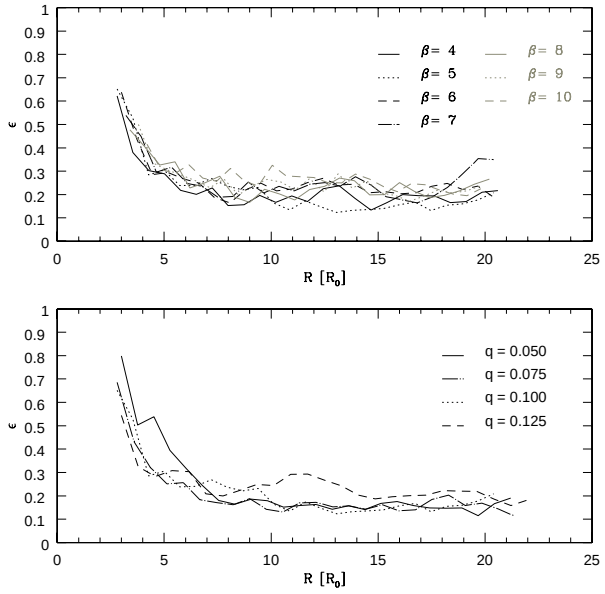

FIGURE 4.15: *The heating factor ϵ as a function of radius for various values of β with $q = 0.1$ (top) and as q varies with $\beta = 5$ (bottom).*

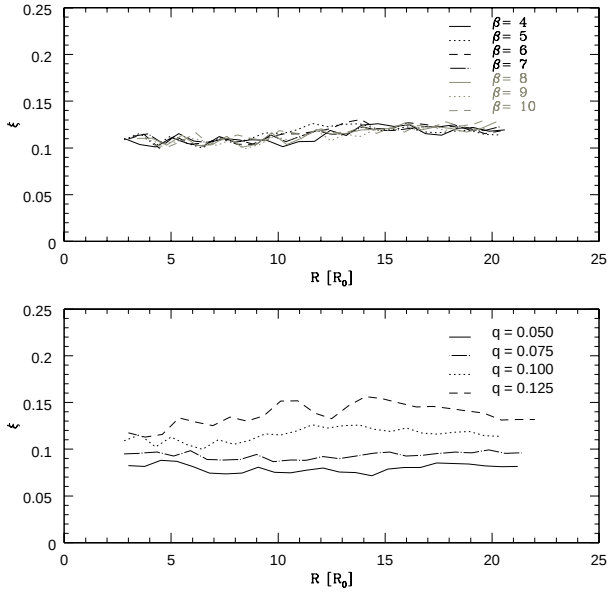

FIGURE 4.16: *Non-local transport fraction ξ as a function of radius for various values of β with $q = 0.1$ (top) and as q varies with $\beta = 5$ (bottom).*

For the $q = 0.1$ case, this shows that varying the cooling has no significant effect on the transport properties of the disc, with $\xi \approx 0.1$ throughout the radial range, albeit with some scatter. This means that in this configuration the disc is dominated by local transport processes, and is as such reasonably well described by the viscous α prescription of Shakura & Sunyaev – global effects, although not negligible, are smaller than local effects by an order of magnitude.

By varying the mass ratio, it is clear that the strength of non-local effects increases with q, rising to $\xi \approx 15\%$ for the case where $q = 0.125$. This confirms the results of Lodato & Rice (2004), who found similarly that non-local effects (characterised by strong transient $m = 2$ spirals) become increasingly important as the disc mass ratio rises, although for the parameter range considered here the disc remains dominated by local effects. This non-local behaviour can be elucidated further by noting that from the definitions of ξ and $\tilde{v}_\mathrm{p}$ (equations 4.28 and 4.17) the non-local transport fraction becomes

$$\xi = \widetilde{\mathcal{M}}\left(\frac{kH_\mathrm{nsg}}{m}\right) = \widetilde{\mathcal{M}}\left(\frac{kH_\mathrm{sg}}{mQ}\right) \tag{4.39}$$

where I have also used the fact that $QH_\mathrm{nsg} = H_\mathrm{sg}$. With $kH_\mathrm{sg} \approx 1$, $Q \approx 1$, this reduces to

$$\xi \approx \frac{\widetilde{\mathcal{M}}}{m} = \frac{1}{m\cos i}, \tag{4.40}$$

using equation 4.38, and thus the non-locality of the transport is directly linked to the openness of the structure induced in the disc through self-gravity. As was noted earlier, for larger disc masses the spiral structure tends to become more open and dominated by lower m modes; equation 4.40 therefore implies that such discs will be more subject to non-local effects than lower mass discs.

4.6 Discussion and Conclusions

In this chapter I have undertaken 3D global numerical simulations of gaseous, non-magnetised discs, evolving under the influence of a massive central object and their own self-gravity. The fluid disc was modelled as an ideal gas with $\gamma = 5/3$, together with a simple cooling prescription based on a local cooling timescale. I have used these simulations to investigate the structure that forms once the discs have settled into a quasi-steady marginally stable state as a function of both the imposed cooling

and the disc to central object mass ratio.

I have found that the amplitude of spiral arms induced in self-gravitating discs, as characterised by the RMS surface density perturbations, can be described straightforwardly through the empirical relationship

$$\left\langle \frac{\delta\Sigma}{\Sigma} \right\rangle \approx \frac{1.0}{\sqrt{\beta}}, \tag{4.41}$$

(where β is the ratio between the local cooling and dynamical timescales), with only a weak dependence on the disc to central object mass ratio. This is in fact closely linked to the result that the Doppler-shifted Mach number is very close to unity – by considering the entropy change ΔS across an adiabatic shock where the Mach number $M \approx 1$, it can be shown that $\Delta S \sim (M^2 - 1)^2$. Thermal equilibrium in these discs is established between cooling, at a rate inversely proportional to β, and the irreversible conversion of mechanical energy into heat, at a rate proportional to the entropy jump ΔS at the shock front – therefore $\beta \sim (M^2 - 1)^{-2}$. Standard shock relations show that the density perturbation $\delta\rho/\rho \sim (M^2 - 1)$, and hence simply from considering the properties of weak adiabatic shocks one can arrive at the relationship $\delta\rho/\rho \sim \beta^{-1/2}$. (Additional details of this argument can be found in Appendix D.) Also I find that the heating factor ϵ – that fraction of the available wave energy that is liberated as heat back into the disc – remains essentially invariant at $\approx 20\%$ with both the imposed cooling regime and the mass ratio of the disc to the central object.

As expected, the simulations show that the dominant radial wavenumber is approximately equal to the reciprocal of the local self-gravitating scale height of the disc throughout the radial range, $k \approx \pi G\Sigma/c_{\rm s}^2$. I therefore find that the radial spacing of the arms is dependent only on the surface density and temperature profiles of the disc. Likewise although further work is required to understand the relationship fully, the azimuthal disc structure is dependent on the disc to central object mass ratio, with more massive discs being characterised by more open structures than their lower mass counterparts for a given central object mass.

These numerical results bear out the theoretical analyses of Balbus & Papaloizou (1999) and Gammie (2001), who suggest that discs in the $Q \approx 1$ marginally stable state may be modelled as predominantly local. Simulations of self-gravitating discs with radiative transfer by Boley et al. (2006) also found that close to co-rotation, angular momentum transport was well modelled by a local α-prescription even when

global modes were present. Balbus & Papaloizou (1999) further predicted that non-local transport from an "anomalous flux" proportional to $\Omega - \Omega_\mathrm{p}$ would become significant far from co-rotation, a result that I have derived analytically using the WKB approximation for tightly wound waves. The WKB dispersion relation has then been used along with empirically determined information on the dominant wavenumbers to make an estimation of $|\Omega_\mathrm{p} - \Omega|$. I find that, at least for low mass discs, this is a small fraction of Ω (less than 15% for discs with $q \leq 0.125$, regardless of the efficacy of the cooling). The results on the magnitude of the non-local transport fraction $\xi = |\Omega_\mathrm{p} - \Omega|/\Omega$ can furthermore be readily understood in terms of the empirical constancy of the Doppler-shifted radial phase Mach number, $\widetilde{\mathcal{M}}$. I conclude that the importance of such non-local effects in gaseous self-gravitating discs is set by the self-adjustment of the pattern speed to ensure that the flow speed normal to the arms is approximately sonic. I have then demonstrated that this condition implies that $\xi \approx m^{-1} \sec i$, where i is the opening angle of the spiral structure. Since the structure within the disc becomes more open as the disc to central object mass ratio increases, this also implies that the importance of non-local transport scales with q.

Note that in collisionless systems such as stellar discs, this self-regulation process for the pattern speed breaks down as shocks cannot form. Hence it is possible to excite global modes in such discs, and thus non-local transport of energy and angular momentum may be more significant dynamically. The results that I present here are therefore restricted to the case of predominantly collisional, gaseous discs, and provide a theoretical underpinning for the findings of Lodato & Rice (2004, 2005) on how the importance of global transport depends on the disc to central object mass ratio in gaseous discs. In particular, note that in cases like those described here where the disc mass is a small fraction of the central object mass (as could be the case for relatively evolved self-gravitating protostellar discs), the effects of self-gravity are expected to be well described as a pseudo-viscous process.

One of the most important applications of this study is that the amplitude of spiral modes in gaseous discs can be related to the cooling regime. With ALMA coming online in the relatively near future, promising milli-arcsecond resolution in the millimetre/sub-mm range, it is possible that observations of spiral structure in protostellar discs may become technically feasible (see Chapter 6). Such observations may therefore provide empirical confirmation of this suggestion, and may further be used to provide an estimator for the strength of the cooling in other systems.

5

Opacity and Gravitational Stability in Protoplanetary Discs

I am sorry to say that there is too much point to the wisecrack that life is extinct on other planets because their scientists were more advanced than ours.

John Fitzgerald Kennedy

The material presented in this has chapter been published as

The effects of opacity on gravitational stability in protoplanetary discs
P. Cossins, G. Lodato & C. J. Clarke, MNRAS, **401**, 2587 - 2598, (2010)

5.1 Introduction

In this chapter I shall consider the effect that varying the functional form of the imposed cooling has on the stability of self-gravitating protoplanetary discs to fragmentation into bound objects, a process that has been put forward as a potential (giant) planet formation mechanism (see Chapter 2). Within this framework I shall then consider how this relates to the effects of dust opacity on the stability of such discs.

The formation of planets within protoplanetary discs is a subject that attracts considerable interest, with two main competing schools of thought. The core accretion-gas capture model (Lissauer, 1993; Lissauer & Stevenson, 2007; Klahr, 2008) posits hierarchical growth, with the collisional coagulation of dust grains initially leading to centimetre-sized particles, and thence on to planetesimals and rocky planets. Once a critical mass is reached, it is then possible to accrete a gaseous envelope and hence form giant Jupiter-like planets. Various observations have successfully confirmed this mode of planet formation, for example Marcy et al. (2005), Dodson-Robinson & Bodenheimer (2009).

However, this model cannot explain all the available observations. Kennedy & Kenyon (2008) show that beyond approximately 20 AU, the timescales for giant planet formation via core accretion exceed the expected disc lifetime of approximately 10 Myr, implying that no planets should be detected in this region. However, recent observations of HR8799 with the Keck and Gemini telescopes have produced direct images of giant planets (5 - 13 $M_{\rm J}$) orbiting at radii of up to $\sim$ 70 AU (Marois et al., 2008). Similar observations of other systems (e.g. β Pic b (Lagrange et al., 2009) and Fomalhaut (Kalas et al., 2008)) and theoretical work on the formation of 2MASS1207b (Lodato et al., 2005) have suggested that there is another mechanism for planet formation at work, and this is thought to be the effect of gravitational instabilities within the protoplanetary discs themselves.

As discussed in detail in Chapter 2, in protoplanetary discs where the self-gravity of the gas is dynamically important, direct gravitational collapse of locally Jeans-unstable over-densities within the disc (Boss, 1997, 1998; Durisen et al., 2007) would produce giant planets very rapidly, on the local dynamical timescale. Although the likelihood of this occurring in protoplanetary discs is uncertain, and is indeed something I shall discuss in detail in this chapter, it has a good pedigree – a similar process of gravitational instability leading to local collapse is a strong candidate for the formation of stellar discs in galactic centres (Levin & Beloborodov, 2003;

Nayakshin & Cuadra, 2005; Nayakshin et al., 2007), and may also lead to the formation of brown dwarfs and other low mass stellar companions (Stamatellos et al., 2007a) in protostellar discs.

Recall from Chapters 1 and 4 that the emergence of the gravitational instability within a disc is governed by the parameter Q (Toomre, 1964), which for a gaseous Keplerian disc is given by

$$Q = \frac{c_{\rm s}\Omega}{\pi G\Sigma}. \tag{5.1}$$

Once the gravitational instability is initiated, heat is input to the disc on the dynamical timescale through the passage of spiral compression/shock waves (see for instance Chapter 4). Various numerical studies using both 2D and 3D models of self-gravitating discs have produced the result that, in order to induce fragmentation, the disc must be able to cool on a timescale faster than a few times the local dynamical time, $t_{\rm dyn} = \Omega^{-1}$ (Gammie, 2001; Rice et al., 2005, for instance). This condition is likely to occur only at relatively large radii ($\sim$ 100 AU), on the assumption that stellar or external irradiation of the disc is negligible (Rafikov, 2009; Stamatellos & Whitworth, 2009a).

As mentioned in Chapter 1, many of these models have used a cooling rate prescribed by using a fixed ratio between the local cooling ($t_{\rm cool}$) and dynamical (Ω^{-1}) times, such that

$$\Omega t_{\rm cool} = \beta \tag{5.2}$$

for some constant β throughout the radial extent of the disc, indeed this was used throughout the work presented in the previous chapter. Various authors as noted above have found that fragmentation occurs whenever $\Omega t_{\rm cool} \approx 3 - 7$. By using a more realistic cooling framework based on the optical depth, Johnson & Gammie (2003) found that the fragmentation boundary (defined hereafter as the ratio of the cooling to dynamical timescales, $\Omega t_{\rm cool}$ at fragmentation) may in fact be over an order of magnitude greater than this, leading to an enhanced tendency towards fragmentation. This variation in $\Omega t_{\rm cool}$ they ascribed to the implicit dependence of the cooling function on the disc opacity, and hence on temperature.

From the opacity tables of Bell & Lin (1994) it is clear that the opacity is a strong function of temperature in certain regimes, and by modelling protoplanetary discs as optically thick in the Rosseland mean sense, $\Omega t_{\rm cool}$ shows power law dependencies on both the local temperature and density. In cases where this dependence is strong, it is therefore possible that small temperature fluctuations may push the local value of

$\Omega t_{\rm cool}$ below the fragmentation boundary, even when the *average* value is significantly above it.

In this chapter I shall therefore seek to investigate and clarify the exact relationship between the fragmentation boundary and the temperature dependence of $\Omega t_{\rm cool}$, using a Smoothed Particle Hydrodynamics (SPH) code to conduct global, 3D numerical simulations of discs where the cooling time follows a power-law dependence on the local temperature. From Chapter 4 and various other studies, it can be seen that in a quasi-steady state the gravitational instability may be modelled approximately pseudo-viscously (Lodato & Rice 2005; Clarke 2009; Rafikov 2009), assuming the disc to star mass ratio does not rise above $q \sim 0.1$. I shall therefore use the α-prescription of Shakura & Sunyaev (1973) and the assumption of local thermal equilibrium, where

$$\Omega t_{\rm cool} = \frac{4}{9\gamma(\gamma-1)\alpha} \tag{5.3}$$

(Chapter 1; Pringle, 1981; Lodato & Rice, 2004; Lodato, 2007) to construct an analytical model of the opacity regimes present within a marginally gravitationally stable disc. From this one can therefore predict analytically if and where such discs will become prone to fragmentation, and also compare these results to more complex simulations where radiative transfer is modelled, such as Boley (2009); Stamatellos & Whitworth (2009a).

The structure of this chapter is therefore as follows. In Section 2 I shall discuss some of the theoretical results relevant to protoplanetary discs, and introduce a simplified cooling function derived from the various opacity regimes. I further consider the effects these cooling prescriptions may be expected to have on the susceptibility of protoplanetary discs to fragmentation. In Section 3 I briefly outline the numerical modelling techniques used in the simulations and detail the initial conditions. In Section 4 I shall present the results from these simulations, before proceeding to collate these with analytical predictions in Section 5. Finally in Section 6 I shall discuss the ramifications of this work and the conclusions that may be drawn from it.

5.2 Theoretical Results

In this section I shall derive analytical results for the dependence of the cooling timescale $t_{\rm cool}$ on temperature and density, such as might be expected in a quasi-gravitationally stable protoplanetary disc environment. I will also consider from an analytical perspective the effects that a (specifically) temperature dependent cooling time will have on the stability of such a disc to fragmentation.

5.2.1 $\Omega t_{\rm cool}$ in the Optically Thick Regime

As in the case of Gammie (2001), one may start from the following basic equations:

$$t_{\rm cool} = \frac{u\Sigma}{\Lambda}, \tag{5.4}$$

$$\tau \approx \rho H_{\rm nsg}\kappa, \tag{5.5}$$

$$\Sigma = 2\rho H_{\rm nsg}, \tag{5.6}$$

$$c_{\rm s}^2 = \frac{\gamma\mathcal{R}T}{\mu}, \tag{5.7}$$

where u is the specific internal energy, Λ is the cooling rate per unit area, τ is the optical depth, ρ is the (volume) density, $H_{\rm nsg} = c_{\rm s}/\Omega$ is the disc scale height, κ is the opacity, γ is the ratio of specific heats, $\mathcal{R} = k/m_{\rm H}$ is the universal gas constant (k being the Boltzmann constant and $m_{\rm H}$ the mass of a hydrogen atom), T is the local mid-plane temperature and μ is the mean molecular weight of the gas.

In the case where the disc is optically thick (in terms of the Rosseland mean), then the cooling rate per surface area Λ is given as

$$\Lambda = \frac{16\sigma T^4}{3\tau}, \tag{5.8}$$

where σ is the Stefan-Boltzmann constant. Note that this is strictly valid only in the case where energy is transported radiatively within the disc — convective transport or stratification within the disc will alter this relationship (see for example Rafikov 2007). For the purely radiative case, the vertical temperature structure of the disc is therefore accounted for via this formalism, and is characterised by the midplane temperature T and the optical depth τ. In order to prevent divergence of this cooling function at low optical depths and to interpolate smoothly into the optically thin regime, others including Johnson & Gammie (2003); Rice & Armitage (2009) have

used a cooling function of the form

$$\Lambda = \frac{16\sigma T^4}{3}\left(\tau + \frac{1}{\tau}\right)^{-1}, \tag{5.9}$$

which becomes directly proportional to the optical depth in the optically thin limit. Note that in general however, I find that discs only become optically thin at large radii, and that this correction is therefore only really relevant to the case where the cooling is dominated by ices.

Furthermore, note that for systems where the stellar mass dominates over that of the disc, the density ρ may be approximated by

$$\rho \approx \frac{M_*}{2\pi R^3 Q} \tag{5.10}$$

where M_* is the mass of the central star and R the radial distance from the central star, and therefore $\Omega^2 = 2\pi G\rho Q$ in the case of Keplerian rotation, where G is the universal gravitation constant. Recalling also that $c_s^2 = u\gamma(\gamma - 1)$, equations 5.4 – 5.10 may be rearranged to show that in the optically thick case, the ratio of cooling to dynamical times should be

$$\Omega t_{\rm cool} = \frac{3\mathcal{R}^2}{8\sigma\sqrt{2\pi G}}\frac{\gamma}{\gamma - 1}\frac{\kappa}{\mu^2}\, Q^{-1/2}\, \rho^{3/2}\, T^{-2}. \tag{5.11}$$

Bell & Lin (1994) found that the Rosseland mean opacity can be reasonably well approximated by power-law dependencies on temperature and density[1], such that

$$\kappa = \kappa_0\, \rho^a\, T^b. \tag{5.12}$$

Specific values of a, b and κ_0 apply for each opacity regime, such that the value of κ varies continuously over the regime boundaries. Using these approximations, one finds that the $\Omega t_{\rm cool}$ value for the various opacity regimes can be given by

$$\Omega t_{\rm cool} = \frac{3\mathcal{R}^2}{8\sigma\sqrt{2\pi G}}\frac{\gamma\kappa_0}{\mu^2(\gamma - 1)}\, Q^{-1/2}\, \rho^{a+3/2}\, T^{b-2}. \tag{5.13}$$

[1]It should be noted that various studies have produced more accurate estimates of the Rosseland mean opacity (for instance Ferguson et al., 2005; Marigo & Aringer, 2009), but they are correspondingly more complex – these power law estimates are adequate for the purposes required here.

Opacity Regime	κ_0 (cm^2 g^{-1})	**a**	**b**	**Max. Temp.** (K)	**Dep. of** $\Omega t_{\rm cool}$
Ices	2×10^{-4}	0	2	166.81	$\rho^{3/2}$
Sublimation of Ices	2×10^{16}	0	-7	202.68	$\rho^{3/2}\ T^{-9}$
Dust Grains	1×10^{-1}	0	1/2	2286.7 $\rho^{2/49}$	$\rho^{3/2}\ T^{-5/2}$
Dust Sublimation	2×10^{81}	1	-24	2029.7 $\rho^{1/81}$	$\rho^{5/2}\ T^{-26}$
Molecules	1×10^{-8}	2/3	3	10000 $\rho^{1/21}$	$\rho^{13/6}\ T^{1}$
Hydrogen scattering	1×10^{-36}	1/3	10	31195 $\rho^{4/75}$	$\rho^{11/6}\ T^{8}$
Bound-Free/Free-Free	1.5×10^{20}	1	-5/2	$1.7939 \times 10^{8}\ \rho^{2/5}$	$\rho^{5/2}\ T^{-9/2}$
Electron scattering	0.348	0	0	———	$\rho^{3/2}\ T^{-2}$

TABLE 5.1: *Details of the various optical regimes by type, showing the transition temperatures and the functional dependence of* $\Omega t_{\rm cool}$ *on the temperature and density in the optically thick regime. Note that all values are quoted in cgs units. The final column gives the functional dependence of* $\Omega t_{\rm cool}$ *on density and temperature. See Bell & Lin (1994) for further details.*

For each opacity regime, the constant κ_0, the exponents a and b, the transition temperatures between the regimes and the functional dependence of $\Omega t_{\rm cool}$ on temperature and density are given in Table 5.1. It should be noted that for the purposes of these tables the density should be measured in cgs units.

5.2.2 Effects of Temperature Dependence on Fragmentation

I shall now specifically consider the effects of temperature fluctuations on the stability of a disc to fragmentation, using a simplified cooling prescription derived from a consideration of equation 5.13.

In the previous section it was noted that the ratio of the local cooling and dynamical times $\Omega t_{\rm cool}$ has a direct dependence on the local mid-plane temperature T. Given that (from Table 5.1) this dependence is generally much stronger than that on density, it is physically reasonable to consider a simplified cooling function where only the effects of temperature are included, and where the cooling time is *defined* via the relationship

$$\Omega t_{\rm cool} = \beta \left(\frac{T}{\bar{T}} \right)^{-n}, \tag{5.14}$$

for some general value of the cooling exponent n and cooling parameter β. Here $\bar{T}$ is the azimuthally averaged mid-plane temperature T once the disc has settled to thermal equilibrium, and thus one finds that when this state is reached, the average cooling timescale is expected to reduce to $\langle \Omega t_{\rm cool} \rangle = \beta$, with a fragmentation boundary β_n associated with each value of n. In particular, with $n = 0$, at fragmentation

$\Omega t_{\rm cool} = \langle \Omega t_{\rm cool} \rangle = \beta_0$, which Gammie (2001), Rice et al. (2005) and others have found to be in the range $3 - 7$.

In the case of temperature dependent cooling (where $n \neq 0$), if the equilibrium value of the cooling parameter $\beta > \beta_0$, the disc may still fragment due to temperature fluctuations leading to a short term (relative to the cooling timescale) decrease in the instantaneous value of β to less than the threshold value. For a power-law index n, in order to calculate the value β_n of the equilibrium cooling parameter below which fragmentation occurs, I make the assumption that fragmentation takes place wherever the instantaneous value of $\Omega t_{\rm cool}$ is held at or below the critical value β_0 for longer than a dynamical time, *independent* of the mechanism by which the cooling is effected. Considering temperature fluctuations of the form $T = \bar{T} + \delta T$, at the fragmentation boundary one therefore finds that

$$\beta_0 = \beta_n \left(1 + \frac{\delta T}{\bar{T}} \right)^{-n}. \tag{5.15}$$

In Chapter 4 it was found that for the case where $M_{\rm disc}/M_* = 0.1$, on average the strength of the surface density perturbations $\delta\Sigma/\bar{\Sigma}$ can be linked to the strength of the cooling through the following relationship,

$$\left\langle \frac{\delta\Sigma}{\bar{\Sigma}} \right\rangle \approx \frac{1}{\langle \Omega t_{\rm cool} \rangle^{1/2}}, \tag{5.16}$$

where angle brackets denote the RMS value. In a similar manner one may expect an equivalent relationship to exist for temperature, given by

$$\left\langle \frac{\delta T}{\bar{T}} \right\rangle = \frac{k}{\langle \Omega t_{\rm cool} \rangle^{1/2}}, \tag{5.17}$$

where k is to be defined empirically. At fragmentation therefore this becomes

$$\left\langle \frac{\delta T}{\bar{T}} \right\rangle = \frac{k}{\beta_n^{1/2}}, \tag{5.18}$$

noting that by construction for a given index n, at fragmentation $\langle \Omega t_{\rm cool} \rangle = \beta_n$. Combining this with equation 5.15 I find that in the case where the cooling is allowed to vary with temperature as per equation 5.14, the fragmentation boundary

β_n satisfies the following equation;

$$\beta_0 = \beta_n \left(1 + \frac{k}{\beta_n^{1/2}}\right)^{-n}. \tag{5.19}$$

This implicit equation can therefore be solved to find the value of the fragmentation boundary β_n for all $n \gtrsim -2$ (below this β_n becomes undefined), as shown later in Table 5.4.

5.3 Numerical Set Up

5.3.1 The SPH Code

All of the simulations presented hereafter were performed using a 3D smoothed particle hydrodynamics (SPH) code, a Lagrangian hydrodynamics code capable of modelling self-gravity (see for example, Benz 1990, Monaghan 1992). As discussed in detail in Chapter 3 and Springel & Hernquist (2002); Price & Monaghan (2007) the code self-consistently incorporates the so-called ∇h terms to ensure energy conservation. As previously, all particles evolve according to individual time-steps governed by the Courant condition, a force condition (Monaghan, 1992) and an integrator limit (Bate et al., 1995), however for this study I use an additional condition that ensures the local timestep is always less than some fraction of the local cooling time.

As with the simulations run in Chapter 4, and in a manner common to many SPH simulations of such discs (for instance Rice et al., 2003a; Lodato & Rice, 2004, 2005; Clarke et al., 2007) the system is modelled as a single point mass (on to which gas particles may accrete if they enter within a given sink radius and satisfy certain boundness conditions — see Bate et al. 1995) orbited by 500,000 SPH gas particles. The central object is free to move under the gravitational influence of the disc.

Again, as in Chapter 4 and in common with many other simulations where cooling is being investigated (for example Gammie, 2001; Lodato & Rice, 2005) I use a simple implementation of the following form;

$$\frac{\mathrm{d}u_i}{\mathrm{d}t} = -\frac{u_i}{t_{\mathrm{cool},i}}, \tag{5.20}$$

where u_i and $t_{\mathrm{cool},i}$ are the specific internal energy and cooling time associated with each particle respectively. The cooling time is allowed to vary with the particle

temperature T_i in such a manner that

$$\Omega_i t_{\mathrm{cool},i} = \hat{\beta} \left(\frac{T_i}{\bar{T}} \right)^{-n}, \tag{5.21}$$

where Ω_i is the angular velocity of the particle, $\bar{T}$ is the equilibrium temperature, and $\hat{\beta}$ and n are input values held constant throughout any given simulation. Given that $T \propto c_{\mathrm{s}}^2$, equation 5.1 shows that for a given value of the surface density Σ this is equivalent to

$$\Omega_i t_{\mathrm{cool},i} = \hat{\beta} \left(\frac{Q_i}{\bar{Q}} \right)^{-2n}, \tag{5.22}$$

where again Q_i is the value of the Q parameter evaluated at each particle, and $\bar{Q}$ is the expected equilibrium value of Q, which I take to be 1 throughout. Note that *a priori* it is not known exactly what the equilibrium value of Q will be once the gravitational instability has saturated. Indeed as will be seen this turns out to be slightly greater than unity, but still such that $Q \approx 1$. The *effective* value of the cooling parameter is given by

$$\beta = \hat{\beta} Q^{-2n}, \tag{5.23}$$

where Q is the actual value to which the simulations settle. Since relatively large values of n are being considered, β can vary significantly from the input value $\hat{\beta}$ for even small changes in Q.

Finally I calculate the equivalent surface density Σ_i (and thus Q_i) at the radial location of each particle R_i by dividing up the disc into (cylindrical) annuli, calculating the surface density for each annulus, and then interpolating radially to obtain $\Sigma_i(R_i)$. To prevent boundary effects, for simulations where $n > 1.0$ the temperature dependent effects are limited to an annulus $15 \leq R \leq 20$ (in code units — note that initially $R_{\mathrm{in}} = 0.25$ and $R_{\mathrm{out}} = 25.0$). At other radii the cooling rate is fixed such that $\Omega t_{\mathrm{cool}} = 8$, a value chosen to suppress fragmentation in regions outside the annulus of interest (see for instance Alexander et al., 2008a).

All the simulations were run with the particles modelled as a perfect gas, with the ratio of specific heats $\gamma = 5/3$. Heat addition is allowed for via $P\mathrm{d}V$ work and shock heating. Artificial viscosity has been included through the standard SPH formalism, with $\alpha_{\mathrm{SPH}} = 0.1$ and $\beta_{\mathrm{SPH}} = 0.2$ — although these values are smaller than those commonly used in SPH simulations, this limits the transport and heating induced by artificial viscosity. As noted earlier in Chapter 4 and shown in Lodato & Rice (2004), with this choice of parameters the transport of energy and

angular momentum due to artificial viscosity is a factor of 10 smaller than that due to gravitational perturbations, whilst the weak shocks appearing in the simulations are still well resolved.

By using the cooling prescription outlined above in equation 5.22, the rate at which the disc cools is governed by the dimensionless parameters Q, $\hat{\beta}$ and n, and the cooling is thereby implemented scale free. The governing equations of the entire simulation can therefore likewise be recast in dimensionless form. As with the simulations detailed in Chapter 4, I define the unit mass to be that of the central star, assign an arbitrary scale radius R_0 such that the physical radius r is related to the code unit radius R via $r = RR_0$, and then define the unit time to be the dynamical time $t_{\rm dyn} = \Omega^{-1}$ at radius $R = 1$, where furthermore the gravitation constant G is also taken to be of unit value.

5.3.2 Initial Conditions

The initial conditions used for these simulations are identical to those described in Chapter 4 – all the simulations model a central object of mass M_*, surrounded by a gaseous disc of mass $M_{\rm disc} = 0.1M_*$. I use an initial surface density profile $\Sigma \propto R^{-3/2}$, which implies that in the marginally stable state where $Q \approx 1$, the disc temperature profile should be approximately flat for a Keplerian rotation curve, and since the surface density evolves on the viscous time $t_{\rm visc} \gg t_{\rm dyn} = \Omega^{-1}$ this profile remains roughly unchanged throughout the simulations. Radially the disc extends from $R_{\rm in} = 0.25$ to $R_{\rm out} = 25.0$, as measured in the code units described above, and as before the disc is initially in approximate hydrostatic equilibrium in a Gaussian distribution of particles with scale height $H_{\rm nsg} = c_{\rm s}/\Omega$. The azimuthal velocities take into account both a pressure correction (Lodato, 2007) and the enclosed disc mass, and in both cases, any variations from dynamical equilibrium are washed out on the dynamical timescale.

The initial temperature profile is $c_{\rm s}^2 \propto R^{-1/2}$ and is such that the minimum value of the Toomre parameter $Q_{\rm min} = 2$ occurs at the outer edge of the disc – in this manner the disc is initially gravitationally stable throughout. As before the disc is *not* initially in thermal equilibrium – heat is not input to the disc until gravitational instabilities are initiated.

Exponent (n)	**Input cooling parameter** ($\hat{\beta}$)
0.0	3, 4, 4.5, 5, 6
0.5	4, 4.5, 5, 5.5, 6
1.0	3, 4, 5, 6, 7, 8, 9, 10
1.5	7, 8, 9, 10, 11
2.0	10, 11, 12, 13, 14, 15, 16, 17, 18
3.0	20, 22.5, 25, 27.5, 30, 32.5, 35, 37.5, 40

TABLE 5.2: *Table of simulations run for various values of the cooling exponent n and rate β. Note that since many of these simulations were run concurrently, there is a degree of overlap in the β values used.*

5.3.3 Simulations Run

Since the simulations use a slightly different surface density profile to that used by previous authors ($\Sigma \propto R^{-3/2}$, cf. $\Sigma \propto R^{-1}$ in Rice et al. 2005, $\Sigma \propto R^{-7/4}$ in Rice et al. 2003a) I initially ran five simulations at various values of β with the cooling exponent n set equal to zero to find the fragmentation boundary in the case where the cooling is independent of temperature. Thereafter, simulations were run at various β values as n was incremented up to $n = 3$ to ascertain the fragmentation boundary in each case. A summary of the simulations run is given in Table 5.2.

5.4 Simulation Results

5.4.1 Detecting Fragmentation

First of all it is useful to explain how fragmentation has been detected in these simulations. Throughout all the numerical simulations run, the maximum density over all particles has been tracked as a function of elapsed time. In the case of a non-fragmenting disc, the maximum always occurs at the inner edge of the disc (as would be expected), and is relatively stable over time. However, once a fragment forms, this maximum density (now corresponding to the radius at which the fragment forms) rises exponentially, on its own dynamical timescale. An example is shown in Fig. 5.1, and the various changes in gradient correspond to various fragments at different radii (and thus with differing growth rates) achieving peak density. A similar increase in the central density of proto-fragments is observed in Stamatellos & Whitworth (2009b), although the timescales differ due to the use of different equations of state.

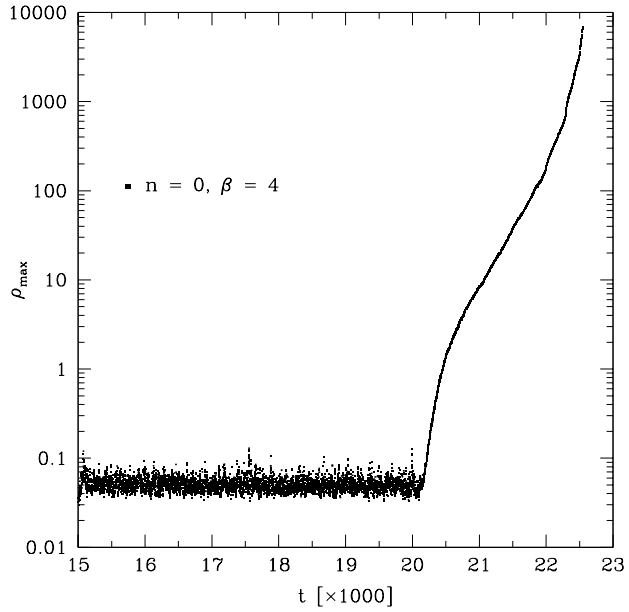

FIGURE 5.1: *Maximum density plot showing the characteristic rise due to fragment formation, seen here for the simulation where $\beta = 4.0$, $n = 0$ (where the cooling is independent of temperature). There is clear evidence of fragment formation at $t \approx 20,000$, with both density and time being shown in code units.*

This rise in the maximum density has therefore been used throughout as a tracer of fragment formation, and the evolution has been followed until the fragments are at least four orders of magnitude greater than the original peak density.

5.4.2 Averaging Techniques

Throughout the following analysis, I define the average value of a (strictly positive) quantity, denoted by an overbar, as the *geometric* mean of the particle quantities. The reason for this is that in the "gravito-turbulent" equilibrium state, properties such as the temperature, density and Q value are log-normally distributed. This is shown for example in Fig. 5.2, where the temperature data from the simulation match a predicted log-normal distribution to within one percent. (Note the reduced radial range to diminish the effect of the inherent gradual decrease in temperature with radius.) The geometric mean being precisely equivalent to the exponential of the arithmetic mean of the logged values, this process recovers the mean value of the normal distribution of $\ln T$.

Similarly, to calculate the perturbation strengths (e.g $\delta A/\bar{A}$ for some quantity

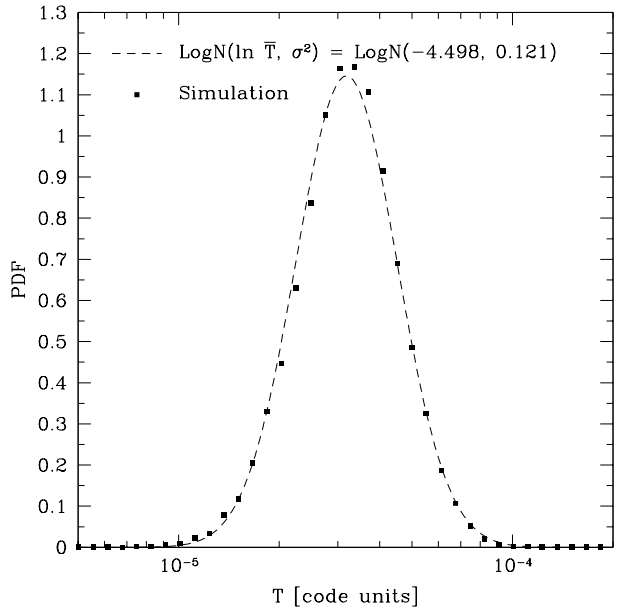

FIGURE 5.2: *Distribution of particle temperatures for* $16.25 \leq R \leq 18.75$*, and a predicted log-normal distribution based on the same data. The two are equal to within approximately 1%.*

A) note that

$$\frac{\delta A}{\bar{A}} \approx \frac{\mathrm{d}A}{A} = \mathrm{d}\ln A. \tag{5.24}$$

The RMS value of $\delta A/\bar{A}$ is then equivalent to the standard deviation of $\ln A$, which again can be recovered directly from the log-normal distribution. Referring again to Fig. 5.2 one can therefore see that $\bar{T} = 10^{-4.498} = 3.177 \times 10^{-5}$ (in code units), and that $\delta T/\bar{T} = \sigma = 0.348$.

5.4.3 Equilibrium States

First of all, the exact value of the fragmentation boundary was determined for the case where $n = 0$ (and thus where $\beta = \hat{\beta}$), which I denote by β_0. As seen in Table 5.2, simulations were run at various values $3.0 \leq \beta \leq 6.0$, and the boundary was found to lie between 4.0 and 4.5. I therefore take the critical value as being the midpoint, such that $\beta_0 = 4.25$.

Continuing with the $n = 0$ case, I find throughout that the value of Q to which the simulations settle is slightly above unity. The steady state values (time averaged over 1000 timesteps) are shown for various β values in Fig. 5.3, and the average Q value is found to be approximately 1.091, where I have averaged over both β and

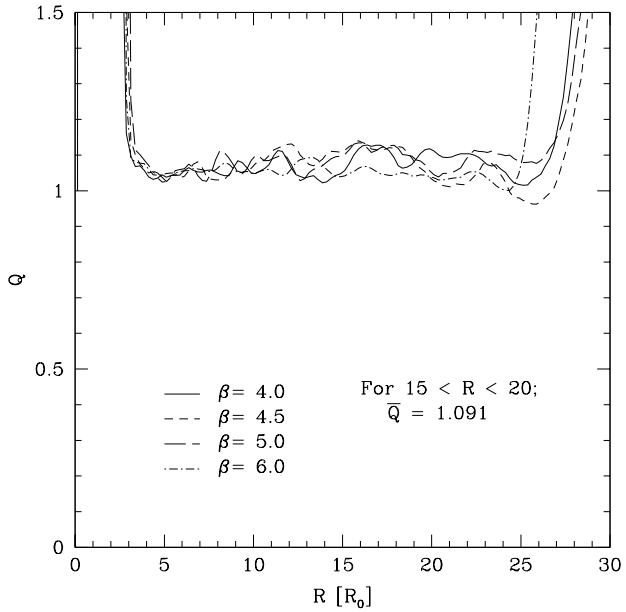

FIGURE 5.3: *Plot of Q against radius for various values of β in the temperature independent case $n = 0$. For the fragmenting cases ($\beta < 4.25$) the values shown are from immediately prior to fragmentation.*

---

radius (where $15 \leq R \leq 20$, for comparison with simulations with higher n). Note further that there is scatter of $\sim 10\%$ about this average, and (although not shown) this is equally true of the simulations where $n > 0$.

Due to this deviation of Q from unity, for large n the effective value of the cooling parameter β at any given radius may be substantially different from the numerical input value $\hat{\beta} = \beta Q^{2n}$ (see equation 5.23) that is used to characterise the cooling law. In order to determine the fragmentation boundary with any accuracy, the true value of β should be considered rather than the input value $\hat{\beta}$.

5.4.4 Cooling Strength and Temperature Fluctuations

In order to characterise the fragmentation boundary, it is necessary to validate the assumption encompassed by equation 5.17, that the temperature perturbation strength is correlated to that of the applied cooling. Using the method outlined above in section 5.4.2, for each simulation one can calculate azimuthally averaged RMS values for the strength of the temperature fluctuations, which I denote by $\left\langle \delta T/\bar{T} \right\rangle$. Where $n = 0$, these temperature perturbations are plotted as a function of radius for various values of β in Fig. 5.4, and it is clear that there is a sys-

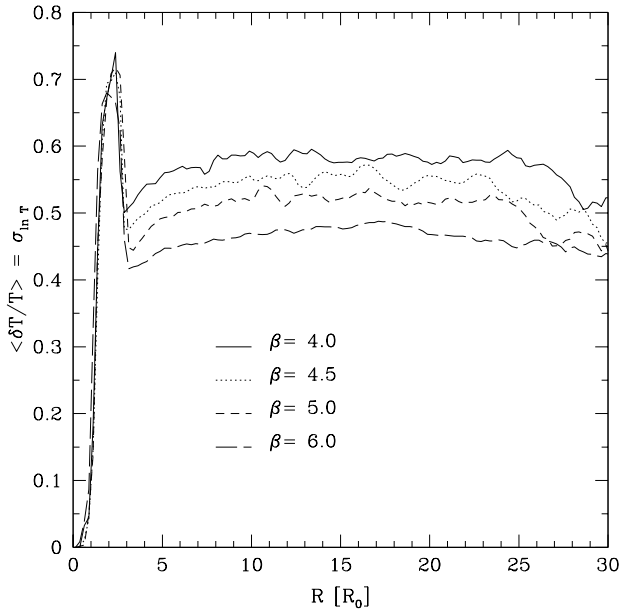

FIGURE 5.4: *Plot showing the strength of temperature perturbations within the disc as a function of radius and β for the temperature independent case, where $n = 0$.*

---

tematic decrease in the perturbation strength with increasing β, and also that the perturbation strength is almost constant with radius across the self-regulating region ($5 \lesssim R \lesssim 25$) of the disc. Using equation 5.17 one can therefore calculate an empirical value for k, and hence averaging both radially (for $15 \leq R \leq 20$ as before) and over the available values of β I find $k = 1.170$ where $n = 0$.

Furthermore, note that in the temperature dependent case (where $n \neq 0$), by construction the average value $\langle \Omega t_{\rm cool} \rangle$ is simply the effective value of the cooling strength, β. The value of k can therefore be calculated for cases where $n \neq 0$, and I find that again k remains constant both with the index n and with radius. Hence I take the value of k to be 1.170, as in the $n = 0$ case, and empirically one may therefore say that on average

$$\left\langle \frac{\delta T}{\bar{T}} \right\rangle = \frac{1.170}{\sqrt{\beta}}, \tag{5.25}$$

for all n.

Exponent (n)	**Effective cooling rate** ($\boldsymbol{\beta}$)		$\boldsymbol{\beta_n}$
	Fragmenting	Non-Fragmenting	
0.0	4.000	4.500	4.250
0.5	4.825	5.263	5.044
1.0	5.915	6.654	6.284
1.5	6.949	7.644	7.296
2.0	8.458	9.022	8.740
3.0	10.051	11.056	10.554

TABLE 5.3: *Table showing the fragmentation boundaries obtained from the simulations. The central columns show respectively the highest fragmenting and lowest non-fragmenting values of β simulated, with β_n being the midpoint of these. Throughout, β is calculated using equation 5.23.*

5.4.5 The Fragmentation Boundary

It is now possible to predict empirically the fragmentation boundary in the case where $n \neq 0$, and to compare this directly with the results of the simulations. Table 5.3 shows the fragmentation boundary β_n as obtained from the simulations, where once again it is taken as the average of the highest fragmenting and lowest non-fragmenting values of β simulated. I find that as expected, there is indeed a rise in the fragmentation boundary as the dependence of the cooling on temperature increases. This variation of the fragmentation boundary is shown against the cooling exponent n in Fig. 5.5, (where the error bars show the upper and lower bounds from Table 5.3) along with predicted values generated using the following empirically defined implicit relationship

$$\beta_0 = \beta_n \left(1 + \frac{1.170}{\sqrt{\beta_n}}\right)^{-n}, \tag{5.26}$$

where I have used $\beta_0 = 4.25$. Clear from this plot is the fact that the predictions are a very good match to the data obtained from the simulations, and the theoretical model in which the increased tendency for fragmentation is due to the effects of temperature fluctuations on the cooling rate is therefore valid. The transition zone shown is bounded by curves corresponding to predictions using $\beta_0 = 4.00$ and 4.50, the upper and lower bounds for β_0 obtained from the simulations.

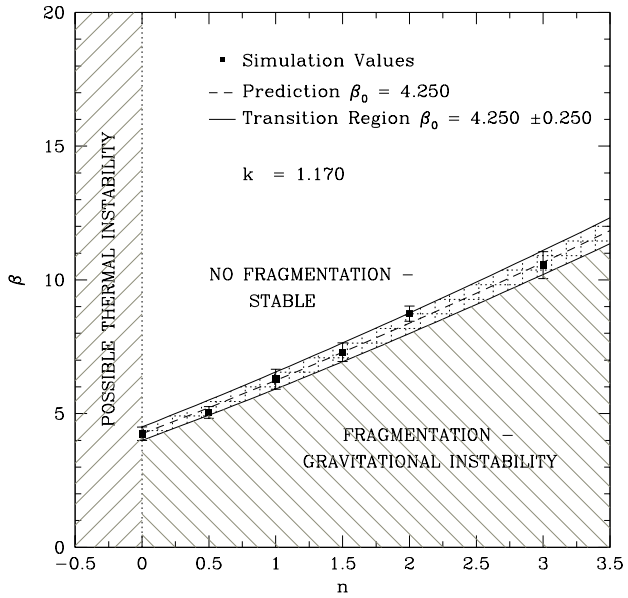

FIGURE 5.5: *Plot of β_n at fragmentation for various values of n. The error bars correspond to the greatest non-fragmenting and smallest fragmenting values of β found in the simulations, and the cross-hatched transition region represents uncertainty in the exact value of β_0. Note also that where $n < 0$ discs may become thermally unstable.*

5.4.6 Statistical Analysis

The effects of temperature perturbations on the fragmentation boundary can be neatly illustrated statistically, if one assumes that the distribution of temperatures about the geometric mean $\ln \bar{T}$ is log-normal (as found in the simulations). Using standard notation one can therefore say that

$$\ln T \sim N(\ln \bar{T}, \sigma^2), \tag{5.27}$$

with standard deviation σ. By taking logs of equation 5.14 it can be seen that

$$\ln \Omega t_{\rm cool} = \ln \beta - n \ln T + n \ln \bar{T}. \tag{5.28}$$

A standard property of the normal distribution is that for a normally distributed random variable $X \sim N(\mu, \sigma^2)$, the distribution of $aX+b$ is given by $N(a\mu+b, a^2\sigma^2)$. Hence from equation 5.28 the distribution of $\ln \Omega t_{\rm cool}$ at fragmentation is such that

$$\ln \Omega t_{\rm cool} \sim N(\ln \beta_n, n^2\sigma^2), \tag{5.29}$$

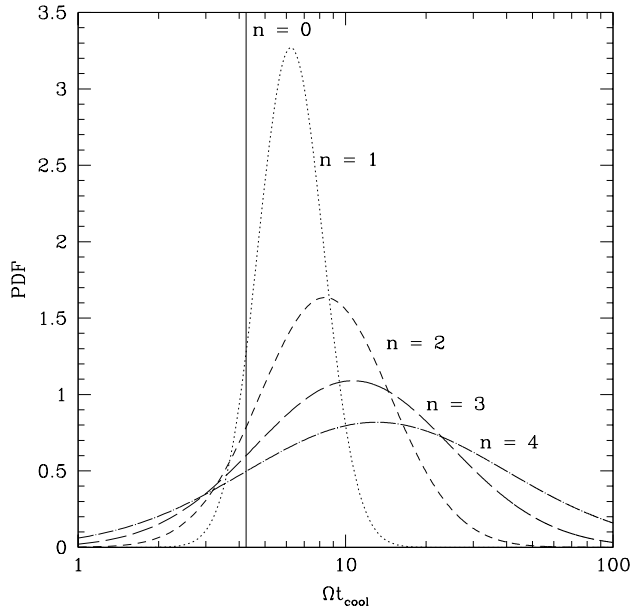

FIGURE 5.6: *Variation in the distribution of* $\ln \Omega t_{\rm cool}$ *as a function of* n*, clearly showing the increasing width of the distribution with increasing* n*. Note that in the case where* $n = 0$ *the distribution reduces to a* δ*-function.*

---

i.e., the distribution of $\ln \Omega t_{\rm cool}$ is centred around $\ln \beta_n$ for all n, reducing to a δ-function in the limit where n becomes zero and becoming more spread out as n becomes large. Thus in order to counteract the increased width of the distribution, and thus the increased fraction of the gas that is below the fragmentation threshold, the average must rise. This is clearly illustrated in Fig. 5.6, for values of n between 0 and 4, and where β_n is given in each case by equation 5.26 with $\beta_0 = 4.25$.

5.5 Opacity-Based Analytic Disc Models

Having quantified the effects of a temperature-dependent cooling law on the fragmentation boundary of protoplanetary discs, it is now possible to use the known cooling laws for each opacity regime (as given by equation 5.13) to determine the dominant cooling mechanisms throughout the radial range. This can therefore also be used to re-evaluate the regions of such discs that are unstable to fragmentation, in a similar manner to the analysis undertaken by Clarke (2009).

In order to do this in a physically realistic manner one must also take into account the effects of the magneto-rotational instability (MRI), which operates when the

disc becomes sufficiently ionised. Considering only thermal ionisation, I assume that the MRI becomes active when the disc temperature rises above $1000K$ (Clarke, 2009). Although estimates of the viscosity provided through this instability vary (see King et al. (2007) for a summary), numerical simulations suggest it should be in the range $0.001 \lesssim \alpha_{\rm MRI} \lesssim 0.01$ (Winters et al., 2003; Sano et al., 2004). Additionally, (Hartmann et al., 1998) suggest that $\alpha \sim 0.01$ is consistent with observations and I therefore assume that the MRI is the dominant instability in the disc, providing an α of 0.01 wherever $T > 1000K$ and the α delivered by the gravitational instability falls below 0.01.

To obtain the disc temperature, note that equations 5.3, 5.7, 5.10, 5.11 and 5.12 self-consistently allow the disc properties to be evaluated for any given stellar mass M_*, mass accretion rate $\dot{M}$ and radius R, when combined with the relation

$$\dot{M} = \frac{3\alpha c_{\rm s}^3}{GQ} \tag{5.30}$$

(see for instance Clarke 2009; Rafikov 2009; Rice & Armitage 2009). The dependence of the disc temperature T on Q, M_*, R and $\dot{M}$ can therefore be obtained, such that

$$T = \left[\frac{32\sigma}{9\kappa_0}\left(\frac{2\pi\mu}{G\gamma\mathcal{R}}\right)^{\frac{1}{2}}\left(\frac{M_*}{2\pi}\right)^{-(a+\frac{3}{2})} Q^{a+1} R^{3a+\frac{9}{2}} \dot{M}^{-1}\right]^{\frac{2}{2b-7}}. \tag{5.31}$$

Finally, in order to prevent the temperature becoming too low, I assume a fiducial background temperature for the interstellar medium (ISM) of $10K$ (D'Alessio et al., 1998; Hartmann et al., 1998). In the case where the temperature evaluated from equation 5.31 falls below this background temperature, I no longer assume that equation 5.3 holds (as there is additional heating from the background as well as from the gravitational instability) and T is set to $10K$.

Since there is a strong dependence on temperature in certain opacity regimes (see Table 5.1) it is important that the equation of state adequately captures the correct behaviour of both the ratio of specific heats γ and the mean molecular weight μ, as variation in these can have significant effects on the system overall. To implement the equation of state I have followed the lead of Black & Bodenheimer (1975); Stamatellos et al. (2007b) and made the assumption that the gas phase of the disc contains only hydrogen and helium, in the ratio 70 : 30. This assumption is valid because although the metallicity of the disc is important for the opacity (and thus

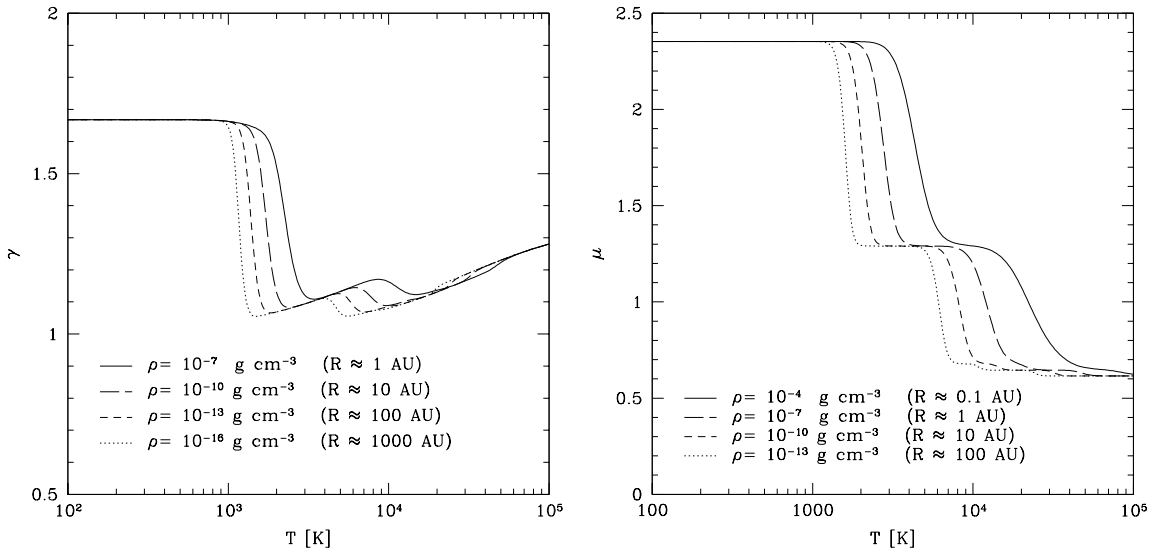

FIGURE 5.7: *Plot showing the variation of the ratio of specific heats γ (left) and the mean molecular weight μ (right) as functions of temperature and density. The density corresponds to the quoted radii for a $Q = 1$ disc about a 1 $M_\odot$ star. Note that the inverse function for temperature in terms of γ, $T(\gamma, \rho)$ is multi-valued.*

---

the cooling), it makes very little contribution to the equation of state. Furthermore, the ratio of ortho- to para-hydrogen is assumed to be held constant at 3 : 1. Following on from the analysis of Black & Bodenheimer (1975), Stamatellos et al. (2007b) produced tabulated values of ρ, T, γ and μ for this equation of state and it is these values that I have used throughout[2]. The variations of γ and μ with both temperature and density are shown in Fig. 5.7 – see also Stamatellos et al. (2007b) and Forgan et al. (2009).

With this tabulated equation of state one can now solve the system of equations for $\Omega t_{\rm cool}$ for any given values of Q, R, $\dot{M}$ and M_* for each opacity regime. For simplicity I assume that the system is marginally gravitationally stable throughout, such that $Q = 1$. Furthermore, since the dependence of $\Omega t_{\rm cool}$ on temperature is known for each of the opacity regimes, one can use equation 5.26 (with $\beta_0 = 4.25$) to predict the (average) value of $\Omega t_{\rm cool}$ at which fragmentation would be expected, the results of which are shown in Table 5.4. Note that since the value of β_n depends only on the relative size of the perturbations in temperature and not on either the mean temperature itself or the value of Q, no variation in β_n is expected with varying Q,

[2]I am indebted to Duncan Forgan of the Royal Observatory, Edinburgh for providing these equation of state tables, as they saved many an hour of prospective labour!

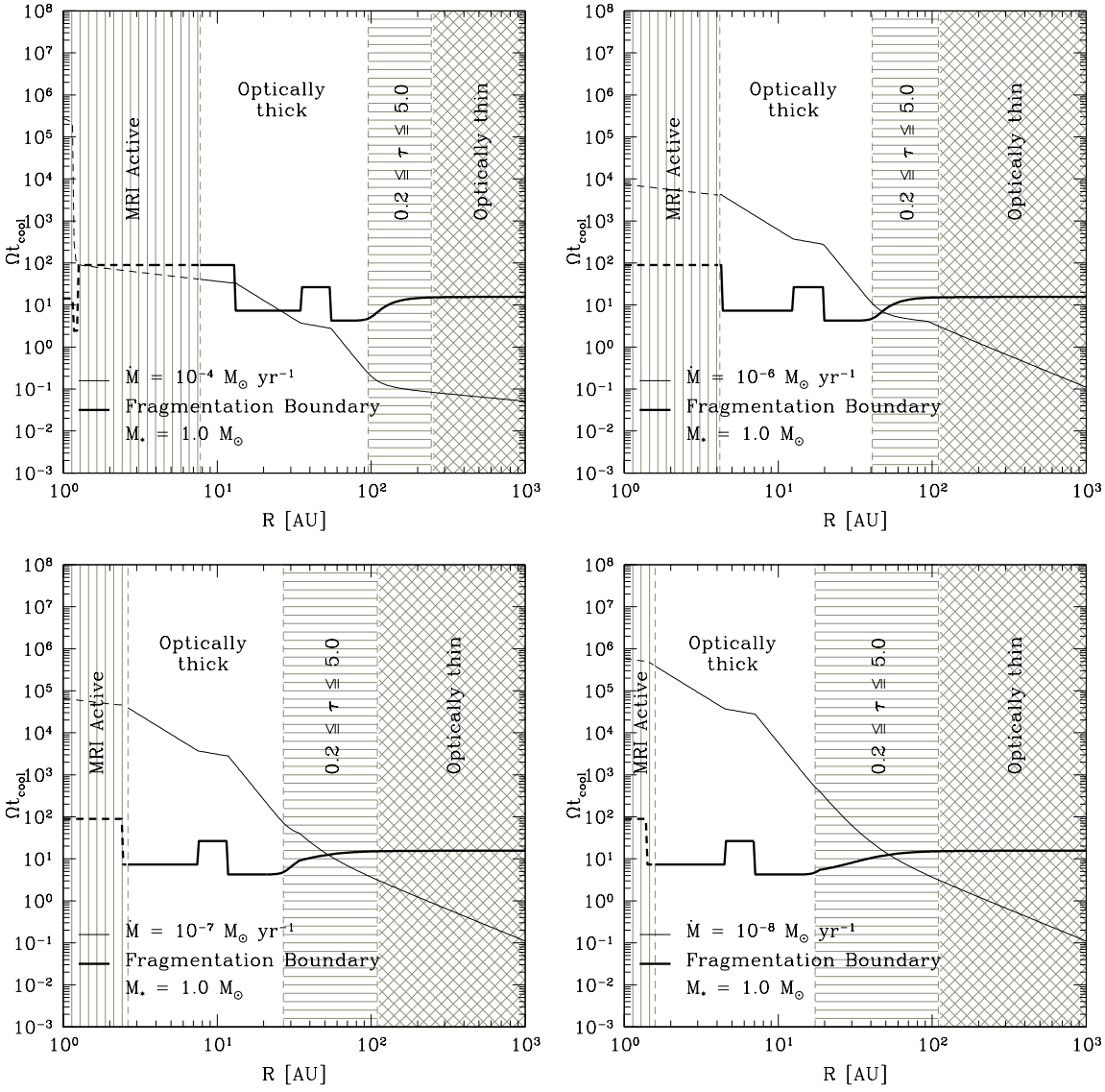

FIGURE 5.8: *Value of $\Omega t_{\rm cool}$ as a function of radius for accretion rates of 10^{-4} (top left), 10^{-6} (top right), 10^{-7} (bottom left) and 10^{-8} (bottom right) $M_\odot$ yr^{-1}, for a disc about a $1M_\odot$ star. The unshaded regions are optically thick ($\tau > 5$), the horizontally shaded areas are transitional ($0.2 < \tau \leq 5$) and the cross-hatched regions are optically thin ($\tau < 0.2$). The vertically shaded areas denote regions of the disc that are MRI active. The disc is stable against fragmentation wherever the value of $\Omega t_{\rm cool}$ is greater than the fragmentation boundary (shown by the heavy solid line). The dotted lines show the values that $\Omega t_{\rm cool}$ and the fragmentation boundary would take if the MRI were not active.*

Opacity Regime	**Dependence of $\Omega t_{\rm cool}$**		$\boldsymbol{\beta_n}$
	on T	**on Q**	
Ices	– none –	Q^{-2}	4.250
*Ices**	T^{-5}	$Q^{2/3}$	*15.570*
Ice Sublimation	T^{-9}	$Q^{-8/7}$	26.688
Dust Grains	$T^{-3/2}$	$Q^{-3/2}$	7.292
Dust Sublimation	T^{-26}	$Q^{-61/55}$	88.296
Molecules	T^{1}	Q^{-6}	2.427
Hydrogen scattering	T^{8}	$Q^{-9/13}$	undefined
Bound-Free & Free-Free	$T^{-9/2}$	$Q^{-3/2}$	14.297
Electron scattering	T^{-2}	$Q^{-10/7}$	8.380

TABLE 5.4: *Predictions for the fragmentation boundary β_n for each opacity regime in the optically thick case. The italicised case gives the prediction in the optically thin limit for ices, the only regime in these models where the disc becomes optically thin. Note that for large positive exponents (such as for hydrogen scattering) the value of β_n becomes undefined. Note also that where the temperature exponent n is positive the regime may become susceptible to thermally instabilities.*

whereas the value of $\Omega t_{\rm cool}$ will vary with both. Using equations 5.31 and 5.13 the Q dependence of $\Omega t_{\rm cool}$ is found to be

$$\Omega t_{\rm cool} \propto Q^{-1+3(a+1)/(2b-7)}, \tag{5.32}$$

and thus except where $b \approx 3.5$ (such as in the regime where molecular line cooling dominates the opacity) the effects of Q variation are small. Nonetheless, in all optically thick cases, the effect of an increase in Q is to decrease the value of $\Omega t_{\rm cool}$, as can be seen from Table 5.4.

In Fig. 5.8 I therefore show the variation in $\Omega t_{\rm cool}$ for a disc about a $1 M_\odot$ protostar as a function of radius at mass accretion rates of 10^{-4}, 10^{-6}, 10^{-7} and 10^{-8} $M_\odot$ yr^{-1}. (For completeness, the various opacity regimes are shown in Fig. 5.9 for an accretion rate of 10^{-4} $M_\odot$ yr^{-1} – all other accretion rates are qualitatively similar.) From the lower two panels (where the accretion rates are 10^{-7} and 10^{-8} $M_\odot$ yr^{-1} for the left and right panels respectively), note that at low accretion rates the fragmentation boundary becomes fixed at approximately $50 AU$, and that this is unaffected by the transition to the optically thin regime. This is down to the fact that the temperature becomes limited below by the background ISM temperature of $10K$, and is therefore decoupled from the mass accretion rate.

As the accretion rate rises to $\sim 10^{-4}$ $M_\odot$ yr^{-1} however, the disc becomes unstable

to fragmentation at a wide range of radii due to the increase in the fragmentation boundary caused by the temperature dependence. Although an island of stability exists between approximately $10-25AU$ (where cooling is dominated by dust grains), all other radii become unstable.

Note also that at low radii the disc becomes MRI active. This occurs at radii from $\sim 1-8AU$ dependent on $\dot{M}$, which corresponds roughly to the transition to the dust sublimation opacity regime. For accretion rates of $\dot{M} \lesssim 10^{-4}\ M_{\odot}\ \mathrm{yr}^{-1}$ Fig. 5.8 suggests that the disc will be stable against fragmentation when the MRI is active, as in the absence of the MRI the value of Ωt_{cool} would be above the fragmentation boundary. However, where $\dot{M} \approx 10^{-4}\ M_{\odot}\ \mathrm{yr}^{-1}$ the picture is less clear, as the disc is MRI active whilst simultaneously being unstable to fragmentation. However, Fromang et al. (2004) have suggested that where both instabilities operate the interaction causes the gravitationally-induced stress to weaken by a factor of two or so, which may stabilise the region against fragmentation.

Nonetheless, throughout the range of mass accretion rates investigated here there are *no* purely self-gravitating solutions at low radii, as the MRI is always active. It is however clear that for radii of $\sim 5-50AU$ the susceptibility to fragmentation of a disc depends strongly on its steady state accretion rate, and that beyond approximately $50AU$, with a $10K$ background ISM temperature discs are always unstable to fragmentation.

Finally it is useful to see how the fragmentation and MRI boundaries vary as a function of both R and $\dot{M}$, and this is shown in Fig. 5.10 assuming that as before the central protostar has mass $M_{*} = 1M_{\odot}$. Here I have also included the fragmentation boundary in the case where the effects of temperature perturbations are ignored, i.e. where $\beta = 4.25$ at fragmentation for all opacity regimes, which allows for comparison with the work of Clarke (2009); Levin (2007, 2003).

Fig. 5.10 shows clearly that by including the effects of temperature perturbations, the mass accretion rate at which fragmentation occurs is reduced, with an increased effect as the dependence of Ωt_{cool} on temperature increases. As before note that there is now a region with $\dot{M} \approx 10^{-4}\ M_{\odot}\ \mathrm{yr}^{-1}$ and $R \lesssim 10AU$ where both the MRI is active and the disc is unstable to fragmentation. For accretion rates of $\sim 10^{-5} - 10^{-3}\ M_{\odot}\ \mathrm{yr}^{-1}$ there are limited radial ranges where a marginally gravitationally stable state exists, with regions that are unstable to fragmentation at both higher and lower radii.

Fig. 5.10 also shows how the stability of the disc to fragmentation varies with

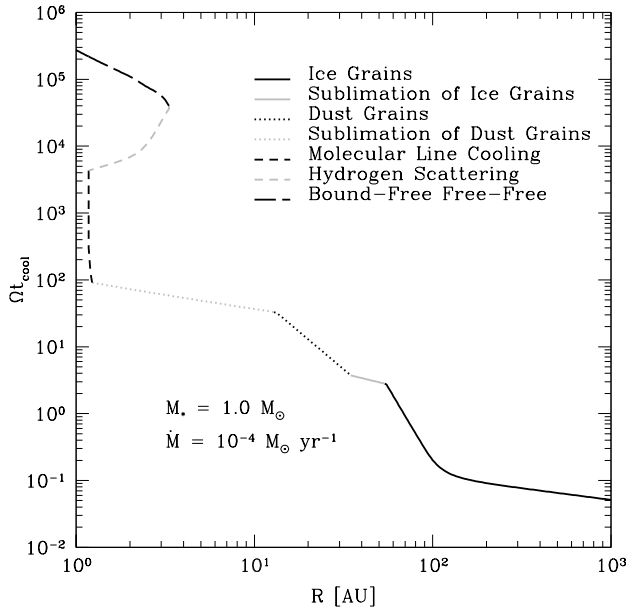

FIGURE 5.9: *Plot of* $\Omega t_{\rm cool}$ *for a mass accretion rate of* $10^{-4}\ M_\odot\ {\rm yr}^{-1}$ *indicating the effects of the various opacity regimes.*

---

the background ISM temperature. For low mass accretion rates it can be seen that as the background temperature decreases, the disc actually becomes stable out to larger radii. This can be explained as follows: In the optically thin case where the cooling is dominated by ices (the regime in which this phenomenon is found) the value of $\Omega t_{\rm cool}$ is given by

$$\Omega t_{\rm cool} = \frac{3\mathcal{R}\sqrt{2\pi G}}{8\sigma\kappa_0}\frac{1}{\mu(\gamma-1)}Q^{1/2}\rho^{1/2}T^{-5} \tag{5.33}$$

$$= \frac{3\mathcal{R}\sqrt{GM_*}}{8\sigma\kappa_0}\frac{1}{\mu(\gamma-1)}R^{-3/2}T^{-5}, \tag{5.34}$$

where I have used equation 5.10 to eliminate ρ in equation 5.34. Hence at a fixed radius $R = R_{\rm frag}$, increasing the temperature decreases $\Omega t_{\rm cool}$ and thereby *destabilises* the disc. Eventually, for some $T = T_{\rm frag}$, $\Omega t_{\rm cool}$ reaches a value of 15.570 (from Table 5.4) and the disc becomes unstable to fragmentation.

From equation 5.34 it can be seen that on the fragmentation boundary (where by construction $\Omega t_{\rm cool} = 15.570$ is constant), $T_{\rm frag} \propto R_{\rm frag}^{-3/10}$. Now assuming that the temperature at which fragmentation occurs is at or above the background temperature (i.e. $T_{\rm frag} \geq T_{\rm min}$) then equation 5.30 holds, and therefore the accretion rate

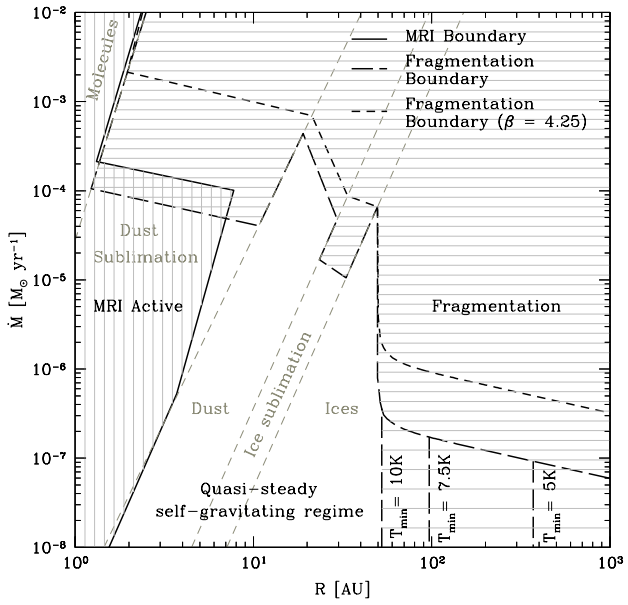

FIGURE 5.10: *Plot showing the regions expected to be marginally gravitationally stable (unshaded), unstable to fragmentation (horizontal shading) and unstable to the magneto-rotational instability (vertical shading) in a disc about a $1M_{\odot}$ protostar. The cross-hatched regions show where where the disc is unstable to both the MRI and fragmentation. The more widely spaced horizontally shaded region to the lower right would become unstable to fragmentation if the minimum temperature limit of 10K was removed, and the fragmentation boundary moves to the right as the minimum temperature is decreased. The short dashed line corresponds to the fragmentation boundary if a fixed value of $\beta = 4.25$ is used (cf. Clarke 2009; Levin 2003, 2007).*

---

at fragmentation $\dot{M}_{\rm frag}$ is given by $\dot{M}_{\rm frag} \propto T_{\rm frag}^{3/2}$. Hence the radius at which fragmentation occurs increases with decreasing accretion rate such that $R_{\rm frag} \propto \dot{M}_{\rm frag}^{-20/9}$, and therefore decreasing the background temperature decreases the accretion rate at which the disc becomes unstable to fragmentation, and likewise increases the radius at which this occurs.

Note however that once $T_{\rm frag}$ is below the background temperature, (i.e. when $T_{\rm frag} < T_{\rm min}$) the disc temperature becomes decoupled from the accretion rate, and hence all accretion rates below $\dot{M}_{\rm min} = \dot{M}_{\rm frag}(T_{\rm min})$ are unstable to fragmentation for radii $R \geq R_{\rm frag}$.

5.6 Discussion and Conclusions

In summary, I have found from controlled numerical experiments with an imposed temperature-dependent cooling law that the effect of temperature dependence is to

increase the value of $\Omega t_{\rm cool}$ at which the disc will fragment into bound objects. Furthermore, this tendency to fragment is greater the more strongly the cooling function depends on the local disc temperature. In this respect, this confirms the results of Johnson & Gammie (2003), who likewise noted a markedly increased tendency towards fragmentation in certain opacity regimes. This result has been attributed to uncertainty in the value of Q in the self-regulated state (Clarke, 2009), equivalent to uncertainty in the equilibrium temperature in this model.

However, these results show that this is only one of two mechanisms that affect the fragmentation boundary, and one that I have been able to account for *a posteriori* by using the effective values of β rather than those input to the simulations. The other effect is due to the strength of the intrinsic temperature perturbations about the mean. In the case where the cooling law is dependent on temperature, perturbations about the equilibrium temperature will mean that some fraction of the gas has a *lower* value of $\Omega t_{\rm cool}$ than average. Once this fraction reaches a critical value, the disc will become unstable to fragmentation. As the dependence of the cooling on these temperature perturbations increases, at a given average value of $\Omega t_{\rm cool}$ the percentage of gas that lies below the critical value also increases, and thus the average must increase to avoid fragmentation.

I therefore find that the effect of allowing the cooling function to depend on the local temperature is to make the disc more unstable to fragmentation, and this variation has been quantified in equation 5.26. Combining this with predictions of the temperature dependence of protoplanetary discs using opacity-based cooling functions, I find that the fragmentation boundary can be increased by over an order of magnitude in terms of $\Omega t_{\rm cool}$, in close agreement with Johnson & Gammie (2003). Furthermore I have also found that the RMS strength of the temperature perturbations can be correlated to the average cooling strength (see equation 5.25), in a very similar manner to that found for the surface density fluctuations in Chapter 4.

Using these predicted values in analytic models of marginally-gravitationally stable $Q = 1$ discs with a representative equation of state, I have found that the susceptibility of such discs to fragmentation into bound objects is also sensitive to the steady state mass accretion rate, as shown in Fig. 5.10. Others have noted that in the optically thick limit where the opacity is dominated by ices, $\Omega t_{\rm cool}$ is *independent* of temperature, and thus the cooling rate is determined only by the local density, itself a function of radius (Matzner & Levin, 2005; Rafikov, 2005; Clarke, 2009). It has therefore been suggested that once the cooling becomes dominated by ices

fragmentation beyond some radius on the order of 100 AU becomes inevitable, and indeed I find that with a background ISM temperature of $10K$, fragmentation occurs at $\sim 50AU$ for all accretion rates below $\sim 10^{-5}\ M_{\odot}\ \mathrm{yr}^{-1}$.

However, if this minimum temperature condition is relaxed, the change in cooling due to entering the optically thin regime has the effect of stabilising the disc out to large radii. (The fact that allowing it to become cooler actually *stabilises* the disc is due to the fact that in this regime Ωt_{cool} increases with decreasing temperature, and thus a hot disc has a shorter cooling time than a cold one.) For Class II / Classic T Tauri objects embedded in a cold medium with accretion rates below a few times $10^{-7}\ M_{\odot}\ \mathrm{yr}^{-1}$, it is therefore possible that extended discs well beyond 100 AU may be stable against fragmentation (they may well be stable against gravitational instabilities altogether), and indeed discs with radii of at least 200 AU have been observed (see for example Eisner et al., 2008, Chapter 2). Nonetheless, discs with accretion rates at the higher end of the scale ($\dot{M} \approx 10^{-6}\ M_{\odot}\ \mathrm{yr}^{-1}$, Hartmann 2009a) will still be unstable to fragmentation at radii beyond $\sim 50AU$. It should be borne in mind however that in the outer regions of discs where the surface density is low, non-thermal ionisation (from cosmic rays, X-rays etc) can trigger the MRI, and this may provide an alternative mechanism for preventing fragmentation, as shown in Clarke (2009).

Fig. 5.10 also shows another important result, that for accretion rates between $10^{-8} - 10^{-2}\ M_{\odot}\ \mathrm{yr}^{-1}$ discs cannot exist in a non-fragmenting purely self-gravitating state at radii $\lesssim 2 - 5AU$. In this regime discs are either MRI active ($\dot{M} \lesssim 10^{-4}\ M_{\odot}\ \mathrm{yr}^{-1}$) or unstable to fragmentation ($\dot{M} \gtrsim 10^{-4}\ M_{\odot}\ \mathrm{yr}^{-1}$). In a narrow band of accretion rates $\sim 10^{-4}\ M_{\odot}\ \mathrm{yr}^{-1}$ I have found that it is possible for discs to be both MRI active and unstable to fragmentation, although the exact interaction of these two instabilities is uncertain (see Fromang et al. 2004). It is therefore the case that for steady-state protoplanetary discs the gravitational instability alone cannot drive accretion directly on to the protostar – either the MRI or the thermal instability must act at low radii, as has been proposed for FU Orionis outbursts (Armitage et al. 2001; Zhu et al. 2009, 2010).

Finally, these results agree with the generally accepted view that planet formation through gravitationally-induced fragmentation is unlikely to occur at radii less than 50 - 100 AU (Matzner & Levin, 2005; Rafikov, 2005; Whitworth & Stamatellos, 2006; Clarke, 2009; Rafikov, 2009), although this critical radius varies with both the mass accretion rate and the background ISM temperature. Within this radius the

core accretion model remains likely to be the dominant mode of planet formation. Outside this radius however, the fragmentation of spiral arms will produce gaseous planets, a result which matches that of Boley (2009) using a grid-based hydrodynamical model with radiative transfer – fragmentation was noted at $\sim 100AU$ about a $1M_{\odot}$ protostar. This result is further corroborated by Stamatellos & Whitworth (2008) whose radiative transfer SPH code suggested a massive disc about a $0.7M_{\odot}$ protostar would rapidly fragment into planetary mass objects or brown dwarf companions beyond approximately $100AU$. Although the mass accretion rate on to the central object is not stated in either case, I find that these figures are nonetheless in general agreement with the analytical predictions presented here.

6

Imaging self-gravitating circumstellar discs with ALMA

You can observe a lot by just watching.

Yogi Berra

The material presented in this chapter has been submitted as

Resolved images of self-gravitating circumstellar discs with ALMA
P. Cossins, G. Lodato & L. Testi, MNRAS, *accepted*

6.1 Introduction

In this chapter I shall present simulated observations of massive self-gravitating circumstellar discs using the Atacama Large Millimeter/sub-millimeter Array (ALMA), with the aim of demonstrating the feasibility of such observations for objects at a variety of realistic distances.

Circumstellar discs play an important role in the formation and evolution of both stars and planets, and as such have been the object of much study and observation in recent years. As discussed in Chapter 2, millimetre wavelength surveys of star forming regions in Taurus (Beckwith et al., 1990; Kitamura et al., 2002; Andrews & Williams, 2005), Orion (Eisner et al., 2008) and ρ Ophiuchus (Andre & Montmerle, 1994; Andrews & Williams, 2007b) have provided extensive evidence for discs of circumstellar material, while optical images of HH30 (Burrows et al., 1996) and the Fomalhaut system (Kalas et al., 2008) have been provided by the Hubble Space Telescope (HST). However, relatively few systems have been imaged with high enough resolution to determine the disc structure on scales of less than a few tens of AU.

This may be about to change however, with ALMA due to come on line in the near future. ALMA is an international collaboration between the European Southern Observatory (ESO), the US National Radio Astronomy Observatory (NRAO) and the National Astronomical Observatory of Japan (NAOJ), and will consist of up to 64 12m antennas covering a frequency range from 31 - 950 GHz (approximately 315 μm - 10cm). With a minimum beam diameter of approximately 5 milli-arcseconds at $\sim$ 900 GHz, ALMA should ideally provide resolution down to $\sim 2AU$ for discs observed in Orion ($\sim$ 410 parsecs), with sub-AU resolution for systems in Taurus-Auriga ($\sim$ 140 pc). It will therefore allow observations of disc sub-structure in unprecedented detail. The discovery of structures within discs could have important implications for our understanding of the evolution of both the discs themselves (Rice & Armitage, 2009) and the protostars they orbit.

Gaseous circumstellar discs are expected to be "turbulent" in some form or another (Ebert, 1994; Gammie, 1996), with the internal stresses that this induces being the driver for angular momentum and energy transport, and thus accretion. This turbulence, whatever its origin, will lead to sub-structure being present at all scales within the disc, with those on smaller scales (for instance due to the magneto-rotational instability) remaining undetectable, while larger scale structure induced by the gravitational instability should be resolvable with ALMA.

As discussed in Chapter 2 it is expected that protostellar discs about Class 0/Class I objects will go through a self-gravitating phase (Bertin & Lodato, 2001a; Vorobyov & Basu, 2005; Hartmann, 2009a) as the infall rate from the gaseous envelope is much greater than the accretion rate on to the protostar (Vorobyov & Basu, 2006). In this phase large amplitude spiral structures will form, driving accretion on to the protostar (Lin & Pringle, 1987; Laughlin & Bodenheimer, 1994; Armitage et al., 2001; Lodato & Rice, 2004, 2005) and potentially leading to the formation of brown dwarf (BD) or planetary-mass companions (Rice et al., 2003b; Stamatellos et al., 2007a; Clarke, 2009). There are already possible detections of spiral structures in the discs of GSS 39 in Ophiuchus (Andrews et al., 2009) and in IRAS 16293-2422B (Rodríguez et al., 2005), although these are not yet definitive.

Furthermore, in Chapter 5 the possibility that protoplanetary discs around Classic T-Tauri (Class II) stars may also exhibit spiral patterns due to the presence of gravitational instabilities was investigated (see also Chapter 5; Bertin & Lodato, 2001a; Boley et al., 2006; Vorobyov & Basu, 2008), and it was found that at radii greater than approximately $50AU$, planet formation through direct fragmentation of these spiral over-densities into bound objects is possible (Boss, 1997, 1998; Clarke, 2009; Rafikov, 2009). While Wolf & D'Angelo (2005) have indicated that the forming giant (proto-)planets themselves may be detectable using ALMA, the observability of the large-scale spiral structure itself within protostellar and protoplanetary discs, though implied (Testi & Leurini, 2008), has remained undemonstrated.

In this chapter I therefore present a simple self-gravitating disc simulation, and from it I derive mock observations of disc systems at the resolutions and sensitivities that should be possible with ALMA. Hence in Section 6.2 I briefly detail the simulation used and in Section 6.3 I shall discuss how the mock observations are generated from it, taking into account telescope effects and sensitivities. Then in Section 6.4 I present the observations for various system and telescope parameters, and finally in Section 6.5 I discuss the significance of the results.

6.2 Simulations of Structure Formation

The simulation I have used to generate the mock observations was performed in the manner detailed in Chapters 4 and 5, using a 3D smoothed particle hydrodynamics (SPH) code, a Lagrangian hydrodynamics code capable of modelling self-gravity (see for example, Benz 1990, Monaghan 1992), the particulars of which are given

in Chapter 3. As previously, the system consisted of a single point mass orbited by 500,000 SPH gas particles, with the central object free to move under the gravitational influence of the disc. All particles were evolved according to individual time-steps governed by the Courant condition, a force condition (Monaghan, 1992) and an integrator limit (Bate et al., 1995).

In the manner of the simulations presented in Chapter 4, I allowed the disc to cool towards gravitational instability by implementing a simple cooling law of the form

$$\frac{\mathrm{d}u_i}{\mathrm{d}t} = -\frac{u_i}{t_{\mathrm{cool},i}}, \tag{6.1}$$

where u_i and $t_{\mathrm{cool},i}$ are the specific internal energy and cooling time associated with the i^{th} particle. The cooling time t_{cool} was then determined through the simple prescription that

$$\Omega t_{\mathrm{cool}} = \beta \tag{6.2}$$

where Ω is the angular frequency and β is a fixed parameter throughout each simulation. Although this is clearly an *ad hoc* cooling function, it can be used as a simple parameterisation in order to conduct controlled numerical experiments. In Chapter 4 I found that for a given disc mass, the spiral structures formed through the gravitational instability (as characterised by the radial and azimuthal wavenumbers of the excited modes) are *independent* of the cooling, but that the strength of the modes (characterised by the relative RMS amplitude of the surface density perturbations $\langle \delta\Sigma/\Sigma \rangle$) varies such that

$$\left\langle \frac{\delta\Sigma}{\Sigma} \right\rangle \approx \frac{1}{\sqrt{\Omega t_{\mathrm{cool}}}}. \tag{6.3}$$

Therefore, notwithstanding the simplicity of the cooling prescription, one may reasonably assume that spiral structures formed in physical systems with characteristic masses corresponding to those in the simulation will be qualitatively similar to those formed in the simulation, with the uncertainty lying primarily in the amplitudes of the density perturbations.

Furthermore, in Chapter 5 I have characterised the behaviour of Ωt_{cool} with radius, and shown that for accretion rates $\sim 10^{-7}\ M_{\odot}\ \mathrm{yr}^{-1}$ (typical for Classic T Tauri objects) at radii of $20 - 50AU$ the value of Ωt_{cool} is $\sim 10^1 - 10^2$, decreasing with increasing radius. Hence although the cooling formalism given in equation 6.2 is very simple, it produces spiral structure in the correct modes and at approximately

the correct amplitudes for the outer radii of discs, as long as one chooses β in the range $\gtrsim 10$. It is therefore useful as a means of generating test cases to investigate whether such structures would actually be observable, particularly in the outer parts of discs.

6.2.1 Simulation Details

The simulation used to generate the mock observations described here consists of a 0.2 $M_{\odot}$ disc about a 1.0 $M_{\odot}$ mass star, extending out to approximately 25 AU. Although this is a relatively high mass ratio, it is within observed bounds (Andrews & Williams, 2005, 2007b), and is plausible for the early (Class I) stages of intermediate mass protostellar evolution, when the disc is expected to be self-gravitating (Bertin & Lodato, 2001a; Vorobyov & Basu, 2005; Hartmann, 2009a).

The initial conditions consist of a disc of gas particles on circular orbits, distributed with a surface density profile $\Sigma \propto R^{-p}$ with $p = 1.5$, as for the Minimum Mass Solar Nebula (MMSN, Weidenschilling 1977). Note that observational constraints for p range from $0.4 \lesssim p \lesssim 1.0$ for protoplanetary discs in Ophiuchus (Andrews et al., 2009) to $0.1 \lesssim p \lesssim 1.7$ in Taurus (Andrews & Williams, 2007a), so this value is within observed limits.

As per the discussion on Chapters 4 and 5, the disc is initially in approximate vertical hydrostatic equilibrium with a Gaussian distribution of particles and non-self-gravitating scale height $H_{\rm nsg} = c_{\rm s}/\Omega$, where $c_{\rm s}$ is the sound speed. The azimuthal velocities take into account both a pressure correction (Lodato, 2007) and the enclosed disc mass and any variation from dynamical equilibrium is washed out on the dynamical timescale. The initial temperature profile is such that $c_{\rm s}^2 \propto R^{-1/2}$, with the minimum value of the Toomre parameter $Q_{\rm min} = 2$ occurring at the outer edge of the disc. In this manner the disc is initially gravitationally stable throughout its radial range.

Note that the SPH code and the initial conditions used are exactly the same as were used in Lodato & Rice (2004, 2005) and in Chapters 4 and 5, excepting the fact that here I have used a mass ratio of 0.2 – further details may be found in the preceding chapters. Finally, in terms of cooling, I have set $\beta = 7$ and use the cooling formalism described in equation 6.1 above. This is approximately in the expected range, and is low enough to avoid spurious numerical heating effects due to artificial viscosity (Lodato & Rice, 2004) – again, see Chapters 4 and 5 for further details.

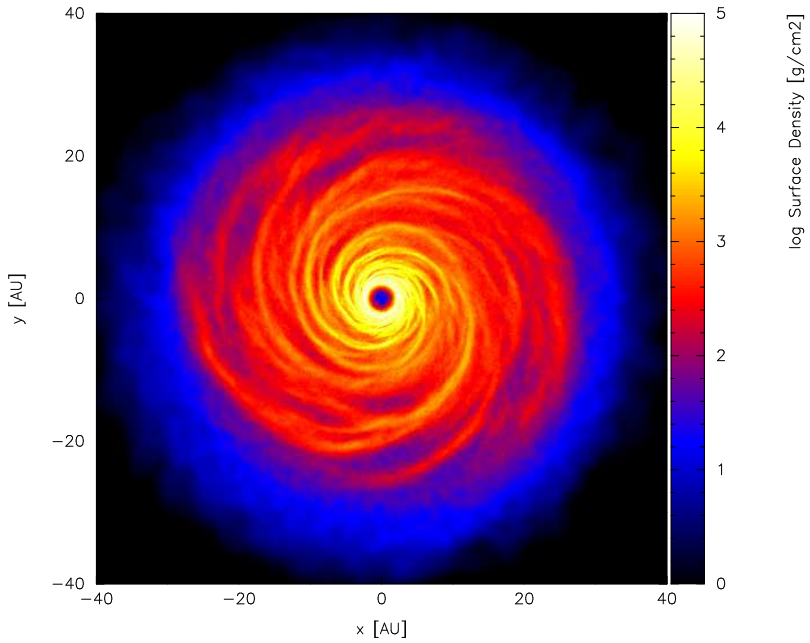

FIGURE 6.1: *Simulated surface density perturbations in a 0.2 $M_{\odot}$ disc about a 1.0 $M_{\odot}$ protostar. The gravitationally induced spiral waves that impart heat to the disc are clearly visible.*

6.2.2 Disc Evolution

Although the thermal profile of the disc initially ensures it is gravitationally stable at all radii, it is however not in thermal equilibrium. As the simulation evolves, the disc therefore cools towards gravitational instability, which is initiated when $Q \approx 1$, after approximately 1000 years. The disc then settles into a marginally stable, quasi-steady dynamic thermal equilibrium state, characterised by the presence of spiral density waves that propagate across the face of the disc, and which provide heat (through shocks) to balance the imposed cooling, as described in Chapter 4. These spiral waves are clearly seen in the surface density, as shown in Fig. 6.1. The simulation was run for approximately 10,000 years (equivalent to ~ 10 cooling times at the disc outer edge), suggesting that during the self-gravitating period of protostellar evolution (expected to last a few $\times 10^5$ years during and immediately after the infall phase) circumstellar discs should indeed be able to reach this quasi-steady state.

Once the disc has reached the self-regulated dynamic thermal equilibrium state, the disc temperature settles to approximately 20 - 40 K, with the minimum at the disc outer edge (shown in Fig. 6.2). Between roughly 10 and 25 AU the temperature falls off as R^{-1}, in reasonable agreement with observations (Andrews & Williams,

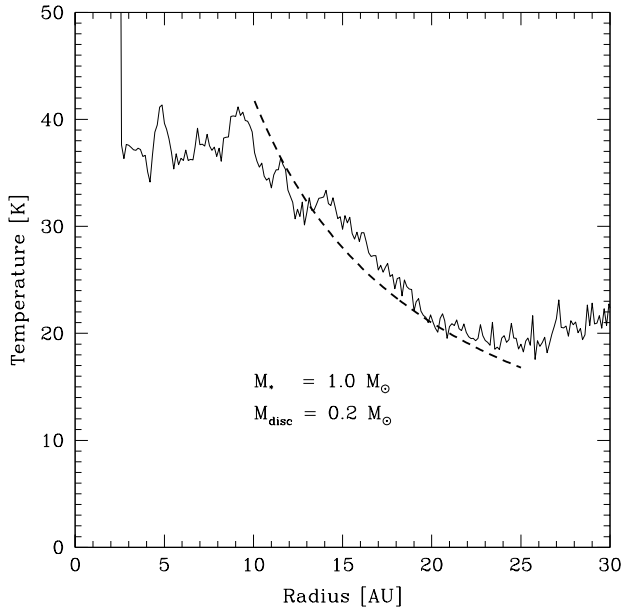

FIGURE 6.2: *Azimuthally averaged temperature profile of the disc once it has settled into the dynamic thermal equilibrium state. The dashed line shows an R^{-1} profile.*

2007a). This is in contrast to the previous simulations, where the discs were roughly isothermal with radius, and is due to changes in both the surface density and Q profiles due to the increased disc mass – non-local transport is of greater importance in this simulation than those in previous chapters. Note also that the slight temperature *rise* outside of $R \approx 25$ AU is due to very low density, higher temperature gas external to the main bulk of the disc.

6.3 Generation of Mock Observations

Having run the hydrodynamic simulations, it is then necessary to use the disc parameters to create flux maps of the objects as they would appear using ALMA. For simplicity, throughout the following I assume that the disc is face-on to the observer – as the observed flux varies with $\cos i$ where i is the inclination angle of the disc (such that $i = 0°$ is face-on) this will be reasonably accurate for all discs within $\sim 10°$ of face-on. Since i is randomly distributed this choice may be idealised, but it is clearly justified as it will produce the most unambiguous signal, and would be an obvious selection criterion for an observation proposal.

In order to calculate the emission I use the individual particle densities $\rho_{\rm i}$ gener-

ated from the SPH simulations, and calculate the particles' absorption coefficients $\alpha_{\nu,\mathrm{i}}$ at frequency ν using

$$\alpha_{\nu,\mathrm{i}} = \rho_{\mathrm{i}}\kappa_{\nu}, \tag{6.4}$$

where κ_{ν} is the opacity of the disc at frequency ν.

For a face-on disc the optical depth τ_{ν} at frequency ν is defined in the following manner

$$\tau_{\nu} = \int_{-\infty}^{\infty} \alpha_{\nu}(z)\,\mathrm{d}z, \tag{6.5}$$

and hence for a disc of SPH particles, one may evaluate the vertical optical depth at any point on the disc face by approximating this integral as a sum over all the relevant interacting particles. I therefore use the following approximation for the optical depth of the disc as seen by a distant observer

$$\tau_{\nu} \approx \sum_{\mathrm{i}} \rho_{\mathrm{i}}\kappa_{\nu}w_{\mathrm{i}} \tag{6.6}$$

where w_{i} is a weighting function related to the particle's mass, density, and the SPH smoothing kernel (see Price 2007) and where i loops over all particles along a given line of sight through the disc.

Assuming that radiation from the disc is in thermal equilibrium with itself and thus that the disc emits as a black body, the source function S_{ν} at frequency ν is given by the Planck function,

$$S_{\nu} = B_{\nu}(T) = \frac{2h\nu^3}{c^2}\frac{1}{e^{h\nu/kT}-1}, \tag{6.7}$$

where h is Planck's constant, c is the speed of light *in vacuo*, k is Boltzmann's constant and T is the temperature of the emitter. For simplicity, I assume that the disc is vertically isothermal with temperature T, (which I obtain via a vertical average from the simulations) meaning that at each point in the disc the source function is constant with height, and is given by $B_{\nu}(T)$. Modelling the disc in this manner, as a geometrically thin structure with no vertical temperature gradient, is obviously only a relatively crude approximation. Nevertheless, as the interest lies in predicting the emission in the (sub-)millimetre range, this approximation is justified since the bulk of the emission will indeed come from the roughly isothermal layer located about the disc midplane.

For such a constant source function, the specific intensity (surface brightness) I_{ν}

at frequency ν and optical depth τ_ν is given by

$$I_\nu = B_\nu(T)\left(1 - e^{-\tau_\nu}\right). \tag{6.8}$$

Note that this is essentially the same method used in reverse to infer the disc mass from the sub-mm surface brightness, as discussed in detail in Chapter 2 (see also Beckwith et al., 1990; Andrews et al., 2009). Furthermore, from equation 6.8 it is clear that once the disc becomes optically thick its emission is determined solely by its temperature, whereas in optically thin regions the emission is exponentially dependent on τ_ν. Optically thick structures forming in optically thin parts of the disc are therefore likely to show greater variation in intensity (and thus be more readily observable) than structure in purely optically thick regions.

6.3.1 Dust Opacities

A critical parameter in the above calculation is the mass opacity of the disc κ_ν. At temperatures below $\approx 160K$ the Rosseland mean opacity becomes dominated by ices (see Chapter 5; Bell & Lin, 1994), but the specific value of the opacity at a given frequency κ_ν is determined by various factors, including the grain size distribution (Miyake & Nakagawa, 1993) and the spin rate of the grains (Rafikov, 2006).

To determine the value of κ_ν I use the power-law model of Beckwith et al. (1990), such that

$$\kappa_\nu = \kappa_{12}\left(\frac{\nu}{\nu_{12}}\right)^\beta, \tag{6.9}$$

where $\nu_{12} = 10^{12}$ Hz, and where κ_{12} is the value of the opacity at this fiducial frequency. As discussed in Chapter 2 the normalisation constant κ_{12} is relatively poorly constrained, with values ranging from 0.1 cm^2 g^{-1} (Beckwith et al., 1990) to approximately 0.016 (Kramer et al., 1998; Rafikov, 2009). For protoplanetary and protostellar discs the power-law index β can be determined from the disc SED in the mm/sub-mm range, and is found to be in the region $0.3 \lesssim \beta \lesssim 1.5$ (Kitamura et al., 2002; Testi et al., 2003; Ricci et al., 2010). Note that this range is below that expected for the interstellar medium $\beta \approx 1.7 - 2$ (Finkbeiner et al., 1999; Chakrabarti & McKee, 2008; Hartmann, 2009a), which is attributed to grain processing within the disc.

For the results presented here I use a fiducial dust opacity $\kappa_{12} = 0.025$ cm^2 g^{-1}, and in accordance with Rafikov (2009) I have assumed throughout that $\beta = 1.0$,

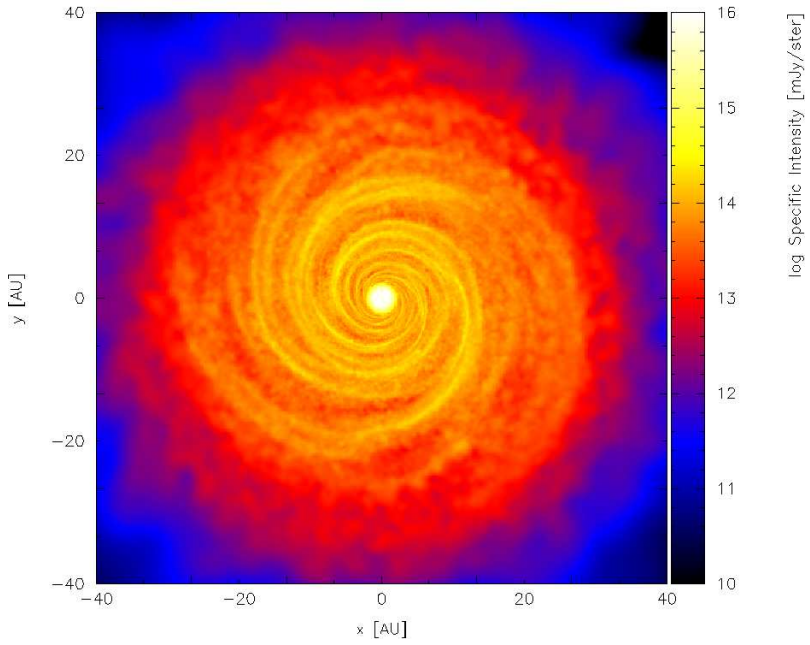

FIGURE 6.3: *Specific intensity of the disc emission at 345 GHz, showing that the underlying spiral structure remains clearly visible.*

such that

$$\kappa_\nu = 0.025 \left(\frac{\nu}{10^{12}\ \mathrm{Hz}} \right)\ \mathrm{cm}^2\ \mathrm{g}^{-1}. \tag{6.10}$$

Using this dust opacity and the method given in Section 6.3, the specific intensity can be generated across the face of the disc, and this is shown for reference in Fig. 6.3, where the observing frequency is set to 345 GHz. While the spiral structure is clearly visible, it is possible now to see additional features in the emission as compared to the surface density in Fig. 6.1, with the leading and trailing edges of the arms clearly delineated in many cases.

6.3.2 The ALMA Simulator

Note that up to now all plots have been generated using SPLASH (Price, 2007), with certain changes made to accommodate the computation of optical depths and specific intensities. However, in order to produce realistic simulated ALMA observations of circumstellar discs from the models, a change of software was required, and as such the ALMA simulator `simdata` in CASA[1] was used. Note that this software is

[1]CASA (Common Astronomy Software Applications) is being primarily developed by ALMA and the NRAO as the main off-line data reduction package for ALMA and EVLA, the Expanded

Frequency	**Atmospheric Conditions**		**Expected Sensitivity**	**Resolution at 50pc**		**Resolution at 140pc**	
(GHz)	pwv (mm)	τ_0	1σ (μJy/beam)	(mas)	(AU)	(mas)	(AU)
45	2.3	0.05	3	120	6	120	16.8
100	2.3	0.03	4	50	2.5	50	7
220	2.3	0.1	10	60	3	20	2.8
345	1.2	0.2	20	40	2	25	3.5
680	0.5	0.6	60	50	2.5	35	4.9
870	0.5	0.7	100	60	3	45	6.3

TABLE 6.1: *Atmospheric conditions, expected sensitivities and angular and linear resolutions as a function of frequency for the simulated observations, assuming distances of 50 pc (corresponding to the distance of the TW Hya association) and 140 pc (roughly corresponding to the Taurus-Auriga, Ophiuchus and Chamaeleon star forming regions). The amount of precipitable water vapour (pwv, in mm) has been chosen according to the current ALMA dynamical scheduling expectations. Note that at the higher frequencies the varying resolutions at different distances is due to the fact that I have considered different ALMA configurations in order to obtain an optimal compromise between resolution and sensitivity. On the other hand, at low frequency, the resolution is dictated by the largest available array configuration, and thus remains constant with distance.*

rather specialised, and as such the simulated ALMA plots presented hereafter were generated by a collaborator (Leonardo Testi of ESO), using the specific intensity maps (as illustrated in Fig. 6.3) as inputs.

The version of the simulator used (2.4) allows the use of the latest version of the planned ALMA antenna configurations, the expected receiver noise based on technical specifications and the contribution due to the atmosphere, itself based on input values for the atmospheric temperature $T_{\rm atm}$ and the optical depth at the frequency of the simulations. In all the simulations presented here it was assumed that $T_{\rm atm} = 265$ K, and the optical depth was computed using the ATM atmosphere models of Pardo et al. (2002) with typical Chajnantor[2] conditions and an amount of precipitable water vapour (pwv) as expected for dynamical scheduling of the observations. In Table 6.1 I therefore show the water content and optical depth of the atmosphere, the expected theoretical noise and the angular and linear resolution of the simulations for discs at 50 and 140 pc for each frequency.

Note that simulations were run for the lowest (45 GHz) and highest (870 GHz) planned ALMA bands – contrary to the intermediate frequencies, these frequency

Very Large Array. `http://casa.nrao.edu`

[2]ALMA is located on the Chajnantor Plateau in the Chilean Andes, at an altitude of approximately 5,000m (16,500 feet)

bands will not be available when ALMA first comes on-line. High frequencies receivers are planned to be introduced during the early years of ALMA science operations, while the low frequency band is under discussion for the longer-term development programme.

The array configuration used varied for each frequency, as the aim was to use the configuration that offered the best compromise between surface brightness sensitivity and angular resolution. Note that at low frequencies the angular resolution is limited by the largest available array configuration. In all cases the observations are based on aperture synthesis simulations with a transit duration of 2 hrs.

6.4 Results

Using the disc simulation shown in Figs. 6.1 and 6.3, I have investigated the emission at the various frequencies shown in Table 6.1. Based on the requirement that the optical depth should vary between optically thick in the spiral arms and optically thin in the inter-arm regions (to maximise the contrast in emission across these regions) I find that the results in the 220 - 345 GHz range provide the greatest resolution of the spiral structure. Furthermore, in this frequency range the spatial resolution is roughly 1 AU at both TW Hydrae and Taurus-Auriga distances, while telescope sensitivity is not a limiting factor. To illustrate the effects of varying the observing frequency, the optical depth at 45, 345 and 870 GHz is shown in Fig. 6.4 for comparison; also clear from this figure is the expected increase in optical depth with frequency due to increasing opacity.

6.4.1 Simulated ALMA Images

Having generated the specific intensity maps by the method described in Section 6.3, they have then been used to simulate ALMA observations as described in Section 6.3.2. In Figs. 6.5 and 6.6 I therefore show the results of these simulations for discs at a distance of 50 pc (TW Hya) and at 140 pc (Chamaeleon, Ophiuchus, Taurus-Auriga). Note that the simulations only include the effects of thermal noise from receivers and the atmosphere, but do not take into account calibration uncertainties and residual phase noise after calibration. These effects are likely to be most important at high frequencies and long baselines, so the simulated maps at 680 and 870 GHz, especially for the 140pc case, represent observations carried out in ideal

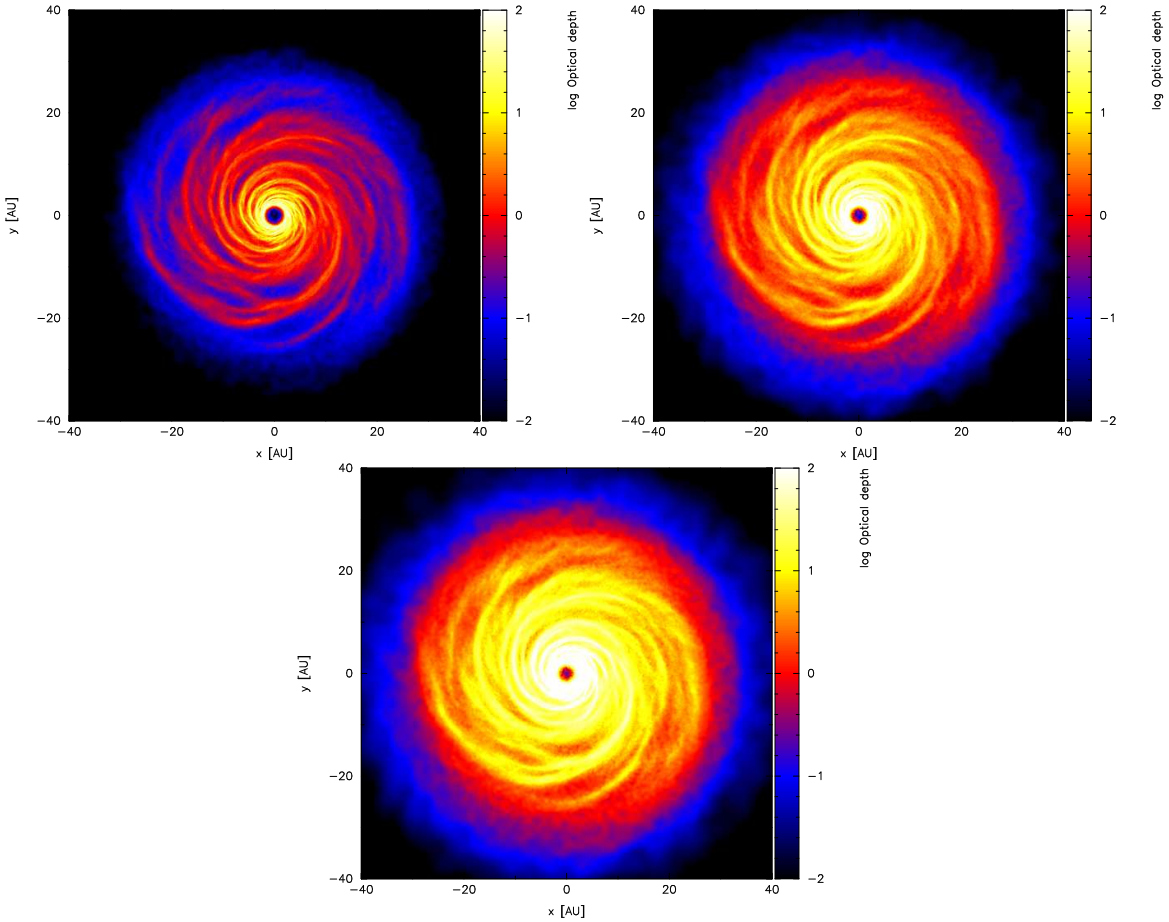

FIGURE 6.4: *Comparison of the (logarithmic) optical depth across the disc face at frequencies of 45 GHz (left), 345 GHz (centre) and 870 GHz (right). The greatest contrast between optically thick and optically thin regions across the arm/inter-arm regions in the bulk of the disc is to be seen in the 345 GHz case.*

conditions and with excellent calibrations.

The simulations show that the predicted spiral structure is readily detectable at all but the lowest frequencies at the 50 pc distance of the TW Hya association. In the case of star-forming regions at 140pc the situation is less clear cut – at low frequencies ($\lesssim$ 100 GHz) even ALMA will probably not provide the angular resolution required to image the spiral structure clearly, whereas at the highest frequencies, as noted above, the simulations are probably over optimistic. Nevertheless, the simulations show that at 220 and 345 GHz (ALMA Bands 6 and 7), the predicted structures should remain clearly detectable.

Finally, in Fig. 6.7 I show the predicted observability of a disc at 410 pc (equivalent to the Orion Nebula Cluster distance) imaged at 345 GHz. While the non-axisymmetric nature of the disc is clear, the spiral structure *per se* is not well resolved, and thus one can infer that even with ALMA such structures will not be conclusively detectable.

6.5 Discussion

In this chapter, I have used a 3D, global SPH simulation of a massive (0.2 $M_{\odot}$) compact ($R_{\rm out}$ = 25 AU) self-gravitating disc about a young star (1.0 $M_{\odot}$) to demonstrate that the spiral modes excited by the gravitational instability should be detectable in face-on circumstellar discs using ALMA. At distances comparable to the TW Hydrae association ($\sim$ 50 pc), such spiral density waves are readily apparent with observation times of 2 hours, whereas at Taurus distances of $\sim$ 140 pc a careful choice of the observing frequency and excellent observing conditions may be required for significant detections. These results suggest that structure in such discs in Orion ($\sim$ 410 pc) will most likely not be resolvable, although disc asymmetries may remain detectable, and improved resolution may be possible with longer observations.

In order to generate these predicted observations I have used temperature and density maps provided by numerical simulations, together with an empirical relationship for the dust opacity of circumstellar material to obtain the optical depth at each part of the disc face. In a precisely similar manner to that used to obtain disc properties from sub-mm observations (Chapter 2, Beckwith et al., 1990) I have then been able to determine the specific intensity of the disc emission across the disc face, which has then been used as input for the ALMA simulator in CASA to produce

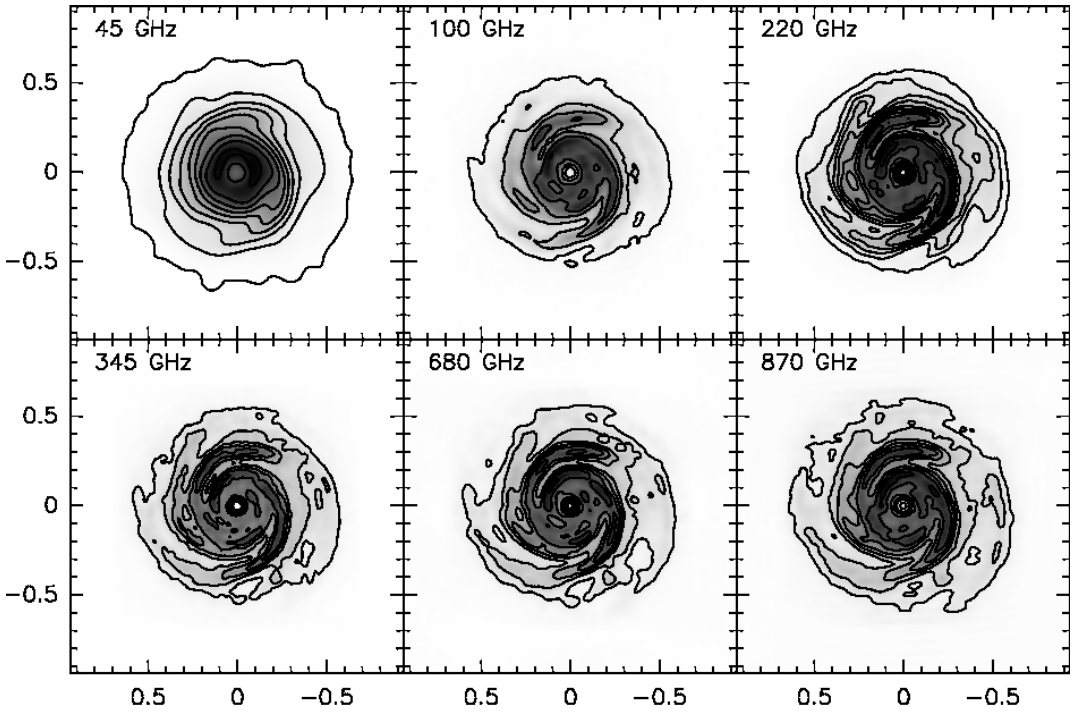

FIGURE 6.5: *Simulated aperture synthesis ALMA images for a disc at 50 pc, with a transit duration of 2 hours. From left to right I show simulations computed for an observing frequency of 45, 100, 220 (top) and 345, 670 and 870 GHz (bottom). Axis scales are in arcseconds. Contours start at 0.01 and are spaced by 0.08 mJy/beam at 45 GHz, start at 0.08 and are spaced by 0.2 mJy/beam at 100 GHz, start at 0.5 and are spaced by 0.5 mJy/beam at 220 GHz, start at 0.8 and are spaced by 0.8 mJy/beam at 345 GHz, start at 2 and are spaced by 2 mJy/beam at 680 GHz, start at 4 and are spaced by 4 mJy/beam at 870 GHz.*

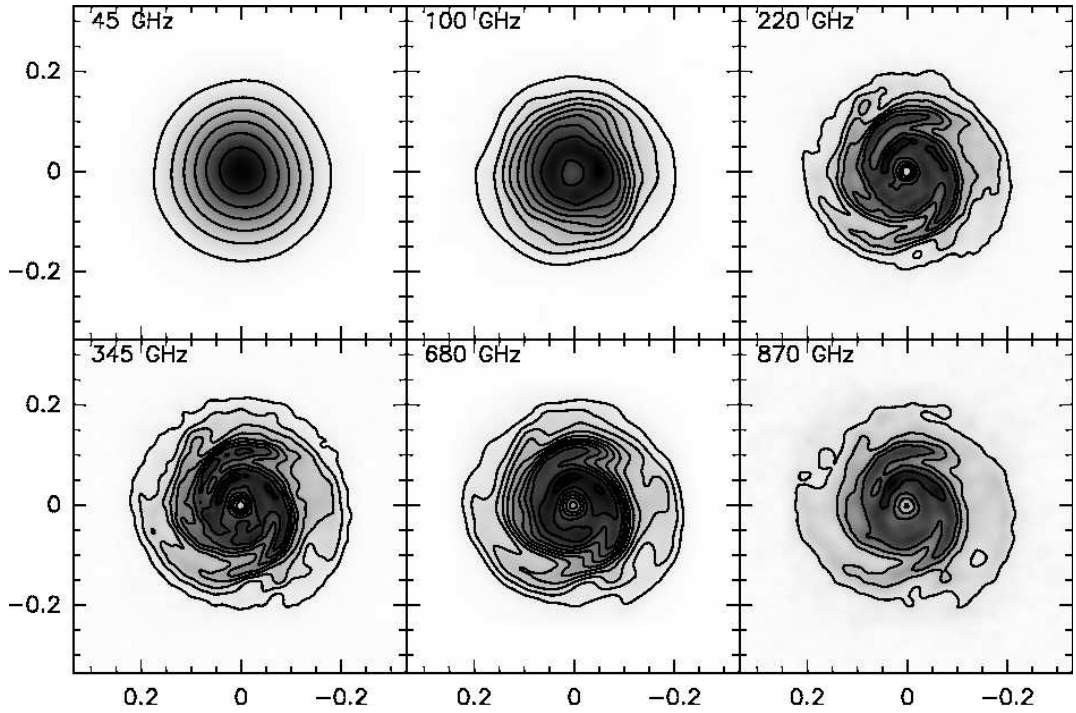

FIGURE 6.6: *As for Fig. 6.5, but for a disc at 140 pc. Contours start at 0.08 and are spaced by 0.08 mJy/beam at 45 GHz, start at 0.08 and are spaced by 0.08 mJy/beam at 100 GHz, start at 0.1 and are spaced by 0.1 mJy/beam at 220 GHz, start at 0.2 and are spaced by 0.2 mJy/beam at 345 GHz, start at 2 and are spaced by 1 mJy/beam at 680 GHz, start at 1 and are spaced by 1 mJy/beam at 870 GHz.*

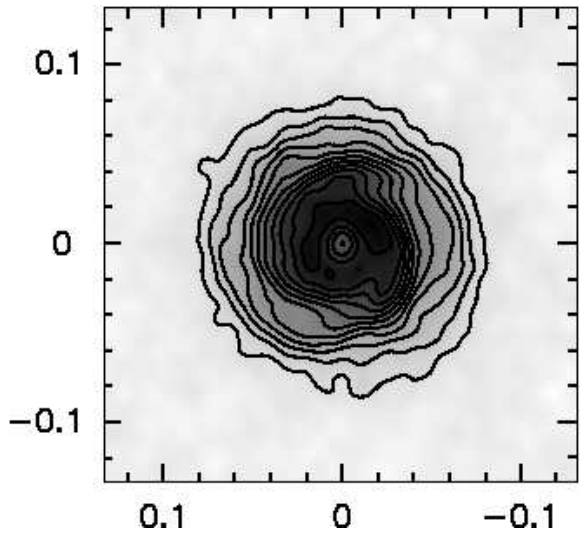

FIGURE 6.7: *Simulated 345 GHz aperture synthesis image of a disc at 410 pc, with a transit duration of 2 hours. Axes are marked in arc-seconds, and contours start at and are spaced by 60 mJy/beam. Clear asymmetries are present, but the underlying spiral structure is not well resolved.*

simulated observations. I find that observation frequencies of 220 – 345 GHz (870 μm – 1.3 mm) are ideal for this kind of observation, as the spiral arms are found to be optically thick, whereas the inter-arm regions are optically thin, maximising the contrast in emission between the regions. It should be noted however that this 'ideal' frequency range is dependent on the assumptions made about the dust opacity, in which there is considerable scatter.

There are certain limitations to the model which should be noted however. The stellar and disc masses used are both at the upper end of the expected distributions (Andrews & Williams, 2005, 2007b; Beckwith et al., 1990), although in the early stages of star formation (Class I objects) such high disc masses are not unreasonable due to the infalling envelope. Of necessity, in order to be self-gravitating the disc is cold ($20 - 40$ K), which implies both a relatively low ambient temperature (again, not unreasonable since giant molecular clouds tend to have temperatures of $\sim 10K$, Myers et al. 1983; Myers & Benson 1983) and for heating from the protostar to be negligible. While this latter assumption is clearly unlikely to be valid for the surface layers of the disc that are irradiated directly by the star, the disc midplane (which dominates the emission) is likely to be cold enough to justify this assumption (Andrews & Williams, 2005; Dullemond et al., 2007; D'Alessio et al., 1998). In a similar manner, this assumption of a colder inner layer allows us to ignore the effects of the magneto-rotational instability, as it will be below the ionisation threshold required for this instability to operate (Gammie, 1996). (Note that ionisation through stellar and cosmic ray irradiation is also neglected.)

Given that the disc is quite compact, and many discs are observed to extend out to much larger radii ($\sim 10^2 - 10^3$ AU, Chapter 2, Andrews et al. 2009; Eisner et al. 2008; Andrews & Williams 2007b; Kitamura et al. 2002) the colder regions outside of ~ 50 AU are if anything more likely to show evidence of gravitational instability than at the radii simulated, and indeed any spiral structures forming at large radii will be more easily resolvable. In this sense therefore these predictions may even be conservative in terms of the maximum distance at which spiral structures may be detectable.

Likewise the cooling prescription, although simplistic, is valid for regions at large radii where the temperature is below the ice sublimation temperature, and is therefore not an unreasonable simplification. It should be noted that the ratio of the cooling time to the dynamical time $\beta = \Omega t_{\rm cool}$ determines the amplitude of the spiral perturbation and hence the contrast in the simulated images. The value

of $\beta = 7$ adopted in this paper is in the right range for discs at a few tens of AU. However, a larger value of β (that is, less efficient cooling) would provide a relatively smaller contrast in the ALMA images.

Finally, note that in this paper I have considered the contribution to the sub-millimetre emission due solely by the disc. In the earliest phases of star formation, the system might show substantial emission on larger scales, produced by the infalling envelope feeding the disc. This larger scale contribution has been neglected in the present paper.

As noted in the introduction to this chapter there are already possible detections of spiral structures in the discs of GSS 39 in Ophiuchus (Andrews et al., 2009) and in IRAS 16293-2422B (Rodríguez et al., 2005). However the structures in GSS 39 are not robust at the 3σ level (Andrews et al., 2009), and those in IRAS 16293-2422B, whilst appearing to be genuine, may plausibly be due to interactions with a companion. Confirmed, unambiguous observations of gravitationally induced spiral structures within protostellar discs would be valuable for a number of reasons. As the gravitational instability is expected to operate during the early phases of star formation, processing the infalling envelope and allowing rapid accretion on to the protostar, such detections would validate this mechanism for growing the masses of protostars. Furthermore, it may enable models of brown dwarf/low mass stellar companion formation through disc fragmentation (Stamatellos et al., 2007a; Stamatellos & Whitworth, 2009a; Clarke, 2009) to be validated, as the presence (or otherwise) and amplitudes of spiral arms would allow constraints to be placed on the numbers and masses of such companions that may be expected.

In a similar vein, detections of spiral features may enable us to determine the dominant mode of planet formation at various radii, about which there is much debate, with the standard core-accretion model being favoured at low radii (Lissauer, 1993; Bodenheimer et al., 2000; Klahr, 2008; Boley, 2009) and the direct fragmentation of gravitationally induced spirals a candidate mechanism at radii above ~ 50 AU (Chapter 5; Boss, 1997; Boley et al., 2006; Rafikov, 2009). As the presence of a peak in the 1.3 cm emission around HL Tau has been put forward as a promising candidate for a planetary mass companion formed through gravitational instability, detections of the spiral wave progenitors of such companions would provide significant backing for this mechanism.

Furthermore, the presence of large amplitude spiral density perturbations may be important for the formation and growth of planetesimals, both due to the concentra-

tion of the dust fraction within the arms (Rice et al., 2004, 2006; Clarke & Lodato, 2009) and further due to the possible scattering of planetesimals by the spiral potential (Britsch et al., 2008). In either case, observations of the spiral arms themselves would place constraints on, and therefore allow us to discriminate between, the two planet formation modes at large radii.

7
Conclusions

There is a theory which states that if ever anybody discovers exactly what the Universe is for and why it is here, it will instantly disappear and be replaced by something even more bizarre and inexplicable. There is another theory which states that this has already happened.

Douglas Adams

In this thesis, I have presented theoretical (Chapter 1) and observational (Chapter 2) evidence for the presence and properties of accretion discs in a variety of astrophysical situations, and considered the implications of the gravitational instability operating within them. By using the numerical scheme presented in Chapter 3 to study them computationally, I have then investigated how this instability may be characterised in fluid discs in terms of structure formation and transport properties (Chapter 4), how the form of the cooling law, and in particular the opacity, affects the tendency of protoplanetary discs to form bound fragments (Chapter 5), and finally I have considered how future observations of the structure formed through self-gravity may be observed using ALMA (Chapter 6). In this concluding chapter I shall therefore give a brief overview of the key findings presented so far, before considering some of the unanswered questions that remain and the direction that future research may take as a result.

7.1 Summary

In Chapter 4 I considered the evolution of massive, non-ionised self-gravitating fluid discs with a prescribed radius-dependent cooling function from an initially stable state to one in which the gravitational instability had saturated, leading to a marginally stable quasi-steady state characterised by $Q \approx 1$ and the presence of spiral structures throughout the radial range.

By conducting a Fourier analysis of the modes excited by the gravitational instability, I found that these spiral structures represent very weak shocks, in that the pattern speed self-adjusts in such a manner that the normal flow into the waves is almost exactly sonic. As heat is primarily added to the disc through these shocks this is intuitively reasonable – strong shocks would lead to a very dissipative disc, in turn leading to significant evolution on relatively short timescales. Moreover this represents a significant difference in the properties of collisional (fluid) discs compared to their collisionless (stellar) counterparts – as the latter cannot form shocks this form of the self-regulation process cannot occur.

Furthermore, I found that for discs in the self-regulated marginally stable state, for a given surface density profile the wave modes that are excited depend on the disc to central object mass ratio, and are *independent* of the cooling rate. In this way more massive discs lead to the formation of more open structures, while light discs show tightly wound spirals. This in turn leads to the wave pattern speed being

further from co-rotation in the case of more massive discs, an effect I have found to be linked to the increased importance of non-local transport of energy and angular momentum within the disc. This has the corollary that low mass discs are dominated by primarily local effects, a result in agreement with Lodato & Rice (2004, 2005). Additionally, although the theoretical predictions for spiral structure in fluid discs are formally only valid in the limit of tightly-wound waves away from co-rotation, I have found very good agreement between the theoretical analysis and the results from simulations (a similar result is noted in Binney & Tremaine 2008) as long as finite thickness effects are accounted for.

A final key result is that although the imposed cooling does not affect the spiral modes excited by the instability, I have found that it *does* affect the amplitude of the density perturbations induced by the waves, a result confirmed by two separate theoretical arguments. In this manner the energy and angular momentum carried by the waves is increased with stronger cooling, itself leading to a rise in the transport of these properties by the gravitational instability.

In Chapter 5 I conducted very similar numerical experiments on identical discs, with the exception that the cooling prescription included a power-law dependence on the local temperature, motivated by a consideration of the various opacity regimes in protoplanetary discs. I found that using a cooling law of this form increased the susceptibility of gravitationally unstable discs to fragmentation into bound objects by almost two orders of magnitude, a result that agrees well with a previous 2D analysis by Johnson & Gammie (2003). The primary reason for this was found to be down to the effects of the distribution of local temperature perturbations about the mean value, and the stronger the dependence of cooling on the local temperature, the greater the effects of these perturbations on the disc stability.

By using this result in a simple 1D analysis of the stability of protoplanetary discs to both gravitational fragmentation and the MRI, I have shown that the quasi-steady, marginally gravitationally stable state can exist in such discs out to approximately 50 AU for accretion rates of $\lesssim 10^{-5}\ M_{\odot}\ \mathrm{yr}^{-1}$, beyond which the disc becomes unstable to fragmentation. Notable however is that as the local background ISM temperature drops, for accretion rates below $\sim 10^{-7}\ M_{\odot}\ \mathrm{yr}^{-1}$ I have found the disc to remain stable to fragmentation out to much larger radii, due to the transition to the optically thin regime. Furthermore, at radii of less than a few AU I find that there are *no* purely marginally stable self-gravitating solutions, with discs either becoming unstable to fragmentation at accretion rates $\gtrsim 10^{-4}\ M_{\odot}\ \mathrm{yr}^{-1}$ or

becoming MRI active for lower values of $\dot{M}$. Finally in Chapter 6 I show that for high mass protostellar discs about solar-mass stars, observations in which the spirals arms induced by the gravitational instability are well resolved should be feasible using next-generation sub-millimetre telescopes such as ALMA. Although such objects may not be resolvable at distances comparable to the ONC (410 pc), at distances representative of the TW Hydrae (50 pc) and Taurus-Auriga (140 pc) star-forming complexes self-gravitating face-on protostellar discs should be readily detectable at frequencies in the central bands of the ALMA observing range.

7.2 Discussion and Open Questions

Whilst a number of new results have been presented here, the work of this thesis leaves certain questions unanswered, and indeed suggests a number of new ones. A key unsolved problem is the determination of the azimuthal wavenumber m – while a clear dependence on the disc mass was shown in Chapter 4, both the form of this dependence and a theoretical underpinning for it remain elusive. It is readily shown from the quadratic and the cubic dispersion relations (equations 1.91 and 1.86) that varying neither the disc mass nor the surface density (at constant $Q, kH_{\rm sg} \approx 1$) should have any effect on the wave modes that are excited. However, the derivation for these dispersion relations explicitly assumes that the waves are determined by the *local* gravitational potential, and therefore if global effects are involved in determining the azimuthal wavenumber then this would not be captured by the current analysis. Furthermore, given that the results presented here show that the excited modes are close to co-rotation throughout the disc, it may be that the effects of this resonance should be taken into account by using the cubic dispersion relation, which is also more appropriate for the more open spirals found at higher disc to star mass ratios. Another obvious complicating factor is that the waves saturate in the non-linear regime, thereby providing motivation for the study of how dynamically important such non-linear effects are in determining the exact structure formation.

In Chapter 4 I noted that the fragmentation boundary shows a dependence on the radial profile of the surface density, an effect which is not currently understood. Although this effect is small, the changes in the surface density profile driven by the action of the gravitational instability itself (Lodato & Rice, 2005; Rice & Armitage, 2009; Rice et al., 2010) may increase its importance, and may therefore lead to fragments forming in regions where stability would otherwise be expected. It may

be that this is also due to the quadratic dispersion relation being an incomplete descriptor for the dynamics involved, again due to the dependence of the co-rotation term on the surface density.

On the other hand, this may turn out to be an artefact of the boundary conditions, another effect that should potentially be quantified. Notable in many fragmenting simulations close to the fragmentation boundary (Rice et al., 2005, also many of my own simulations) is that the condensates form in the outer parts of the disc, very soon after the wavefront of structure formation has reached the outer edge of the disc. It seems possible that reflection or interaction of the density waves with the outer edge leads to constructive interfere with pre-existing waves just inside the outer boundary, which may then result in unduly large amplitude over-densities, and thus fragmentation in regions where the disc should otherwise be stable. Partial evidence for this comes from the work of Clarke et al. (2007), who found that relaxed discs where structure had been allowed to saturate at values of the cooling parameter β well above the fragmentation limit which were then subject to a steady decline in β, were stable to fragmentation at much lower levels of the cooling parameter than would be the case if the simulation was run throughout with the stronger cooling.

Hence, simulations with much larger radial ranges may therefore produce interesting results, as they would allow for comparisons to be made with equivalent simulations with smaller dynamic ranges, thereby isolating boundary effects, but furthermore the more open structures formed at large radii may be similar to those produced at low radii but with higher mass discs, and thus such simulations may help to determine the exact effects of the mass ratio on the disc dynamics. Finally, in terms of the observability modelling carried out in Chapter 6, the inclusion of the vertical temperature gradient and the use of a variety of fiducial dust opacities may allow further constraints to be placed on the detectability of gravitationally induced spiral structure in protostellar discs. Clearly a sensible follow-on from the current work would be to submit a proposal to observe a suitable object (e.g. IRAS 16293-2422B) once ALMA comes online.

While most of these suggestions for future work may appear to be finessing an already idealised problem that is now reasonably well understood, I think it is important that such controlled numerical experiments are undertaken rigorously, and in a manner in which all the various processes that affect the structure formation, transport properties and fragmentation of gravitationally unstable discs may be

fully investigated and understood in isolation from each other, insofar as this is feasible. Having deconstructed and understood the individual physical processes in this way, it would then be possible to recombine them, building up a steadily more complete picture of the evolution of circumstellar material under the influence of the gravitational instability whilst remaining aware of the limitations of simplified numerical approximations to a complex physical problem.

This is not to say of course that the net should not be cast wider. An immediate extension of the results from Chapter 5 would be to use a cooling function tied directly to the expected opacity, in a similar manner to that used by Johnson & Gammie (2003), thereby extending their work to a fully three-dimensional, global case. The effects of stellar irradiation could be included relatively straightforwardly, as indeed could a background temperature, and although the results would then no longer be independent of scale, it would no doubt be enlightening to try and reproduce the one dimensional results of Chapter 5, that discs become more stable at large radii due to a lower background temperature, in a fully three dimensional simulation.

Looking further ahead, inclusion of magnetic effects would allow the interaction the gravitational and magneto-rotational instabilities to be investigated, which would have relevance for models of FU Orionis-type outbursts. More prosaically, reproduction of the current results at much higher resolutions (say a factor of a thousand or more in particle number) would significantly reduce computational noise within the simulations, and may therefore uncover effects that are currently hidden, such as the possible non-linear effects discussed above. Finally, direct comparison with either a different SPH code, and/or with a grid code of similar resolution may prove worthwhile, either through highlighting any unsuspected numerical issues and thereby enabling the development of improved models or (for preference!) through direct confirmation of the current results.

In conclusion therefore, I have presented numerical simulations of structure formation and evolution in massive, cold, non-ionised protostellar and protoplanetary discs, and developed both theoretical and empirical models to characterise this structure in terms of its ability to transport energy and angular momentum, and also its susceptibility to form bound fragments, dependent on the cooling regime and local background temperature. In this manner I have demonstrated that the gravitational instability is an important factor in understanding the dynamics and evolution of self-gravitating circumstellar material. Using these numerical models I have also

been able to demonstrate that resolved sub-millimetre observations of structures within protostellar discs are feasible using ALMA, and therefore should be available to the scientific community with the next few years.

8

Appendices

A little inaccuracy sometimes saves tons of explanation

H H Munro (Saki)

8.1 Appendix A: Divergence of a Tensor

Extending the concept of divergence to tensors is not trivial, as it requires recourse to differential (Riemannian) geometry, itself not a trivial matter! Without going into detail however, we may define the divergence of a (contravariant) rank 2 tensor $\mathsf{T} = \mathsf{T}^{\alpha\beta}$ to be the contraction of the covariant derivative of the tensor (which is in fact a direct extension of the definition of the divergence of a vector).

From the definition of the covariant derivative – see for instance Misner et al. (1973) – we obtain the covariant derivative $\mathsf{T}^{\alpha\beta}{}_{;\gamma}$ of $\mathsf{T}^{\alpha\beta}$ to be

$$\mathsf{T}^{\alpha\beta}{}_{;\gamma} = \frac{\partial \mathsf{T}^{\alpha\beta}}{\partial x^\gamma} + \Gamma^\alpha_{\delta\gamma}\mathsf{T}^{\delta\beta} + \Gamma^\beta_{\delta\gamma}\mathsf{T}^{\alpha\delta}, \tag{8.1}$$

where the x^γ represent the three spatial co-ordinates in any required curvi-linear co-ordinate system, and $\Gamma^{\alpha\beta}_\gamma$ are the Christoffel symbols of the second kind in the same co-ordinate system.

By taking the contraction of this derivative over β and γ we obtain the divergence of T, which therefore becomes

$$\begin{aligned} \nabla \cdot \mathsf{T} &= \mathsf{T}^{\alpha\beta}{}_{;\beta} && (8.2)\\ &= \frac{\partial \mathsf{T}^{\alpha\beta}}{\partial x^\beta} + \Gamma^\alpha_{\delta\beta}\mathsf{T}^{\delta\beta} + \Gamma^\beta_{\delta\beta}\mathsf{T}^{\alpha\delta}. && (8.3) \end{aligned}$$

In cylindrical polars, the only non-zero components of the Christoffel symbols become

$$\Gamma^\theta_{R\theta} = \Gamma^\theta_{\theta R} = \frac{1}{R} \quad \text{and} \quad \Gamma^R_{\theta\theta} = -\frac{1}{R}, \tag{8.4}$$

and thus the divergence of a (contravariant) tensor becomes (in component form)

$$\begin{aligned} &\frac{\partial T^{RR}}{\partial R} + \frac{\partial T^{R\theta}}{\partial \theta} + \frac{\partial T^{Rz}}{\partial z} + \frac{1}{R}T^{RR} - \frac{1}{R}T^{\theta\theta} && R-\text{component}, && (8.5)\\ &\frac{\partial T^{\theta R}}{\partial R} + \frac{\partial T^{\theta\theta}}{\partial \theta} + \frac{\partial T^{\theta z}}{\partial z} + \frac{1}{R}T^{\theta R} + \frac{1}{R}T^{R\theta} && \theta-\text{component}, && (8.6)\\ &\frac{\partial T^{zR}}{\partial R} + \frac{\partial T^{z\theta}}{\partial \theta} + \frac{\partial T^{zz}}{\partial z} + \frac{1}{R}T^{zR} && z-\text{component}. && (8.7) \end{aligned}$$

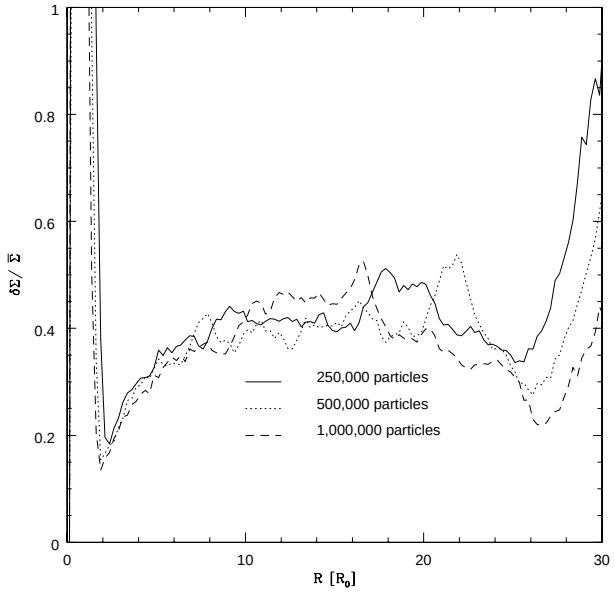

FIGURE 8.1: *Variation of the radial RMS surface density perturbation amplitude as a function of simulation resolution.*

8.2 Appendix B: Resolution and Convergence Tests

In this appendix I shall briefly outline the tests that were undertaken to ensure the convergence of these results.

Three simulations were run, all with the cooling parameter $\beta = 6$ and mass ratio $q = 0.1$, using discs of 250,000, 500,000 and 1,000,000 particles. These were otherwise identical to the simulations that were used for this paper, as described in full in Section 4.4. The three values that are of most significance to the results presented here are the RMS surface density perturbation amplitude $\delta\Sigma/\bar{\Sigma}$ and the average radial and azimuthal wavenumbers $k_{\rm av}$ and $m_{\rm av}$ respectively. Fig. 8.1 shows how $\delta\Sigma/\bar{\Sigma}$ varies with resolution, and although there is considerable scatter it is clear that there is no systematic variation with resolution. A similar result (with even less scatter) is also obtained when one conducts a Fourier analysis of the simulations – there is no systematic variation with resolution. I therefore conclude that the simulations are converged, and that the resolution when using 500,000 particles is satisfactory.

8.3 Appendix C: Fourier Decomposition Methods

In this appendix I detail how the Fourier mode analysis was conducted, using the SPH particle positions as the input values. For simplicity, I begin by discussing how the radial k mode amplitudes were computed, as this had practical implications on how the azimuthal m mode analysis was performed.

8.3.1 Radial Mode Analysis

Calculating the radial Fourier mode amplitudes within a disc presents certain problems, since even a cursory glance at Fig. 4.3 reveals that the radial wavenumber k varies significantly with radius. Also, unlike the azimuthal wavenumbers, the disc is neither uniform nor periodic in radius, and therefore the underlying Fourier distribution corresponding to the disc surface density profile also has to be taken into account. The following method addresses both of these problems while keeping the signal to noise ratio as high as possible.

The disc to be analysed is divided into numerous overlapping annuli of width ΔR, which varies with the central radius of the annulus, and into a number of sectors of fixed angular width $\Delta\theta$. The ΔR values are chosen such that each annulus is of sufficient radial extent to resolve the greatest radial wavelength present at that radius, likewise each sector must be narrow enough to ensure the wave crests are distinct and not smeared out across a wide range in R. In this manner, the smaller the winding angle $i = \tan^{-1}|m/kR|$ of the waves the wider the sectors can be for a given resolution.

Since the radial wavenumber profile depends on the disc to central object mass ratio, the radial extent ΔR of the annuli varies likewise, in order to capture all the relevant modes. The values used in these analyses are summarised in Table 8.1. Note that the widths of the annuli increase linearly across the disc, from the initial to the final widths quoted.

To calculate the underlying Fourier distribution due to the unperturbed surface density profile, the Fourier transform was taken over the whole of each annulus. This thereby smears out all the waves and takes the *average* distribution, and is evaluated according to the following relation;

$$A_{\rm k} = \frac{1}{N_{\rm ann}} \left| \sum_{j=1}^{N_{\rm ann}} e^{-ikR_j} \right|, \tag{8.8}$$

$M_{\mathbf{disc}}/M_*$	Annuli	Initial width	Final width	No. of Sectors
0.050	25	2	8	60
0.075	25	2	8	60
0.100	25	2	10	60
0.125	25	2	10	60

TABLE 8.1: *Details of the Fourier analyses for the various disc to central object mass ratios analysed.*

where $A_{\rm k}$ is the k mode amplitude corresponding to the underlying disc distribution, $N_{\rm ann}$ is the number of particles per annulus, k is the radial wavenumber and the R_j are the radii of the individual particles.

The Fourier distribution of the waves overlaid on the disc are calculated by taking an equivalent transformation over each sector within the annulus, such that

$$A_{\rm k,n} = \frac{1}{N_{\rm sect}} \left| \sum_{j=1}^{N_{\rm sect}} e^{-ikR_j} \right|, \tag{8.9}$$

where $A_{\rm k,n}$ is the k mode amplitude of the waves and disc evaluated in the nth sector, and $N_{\rm sect}$ is the number of particles in that sector.

Finally the Fourier distribution due solely to the waves in each sector is given by the difference between equations 8.9 and 8.8. Since each sector should be statistically similar to the others one may then average over all the sectors $N_{\rm sectors}$ (which in this case does not smear the wave component out, but reduces computational noise), to give the average radial Fourier mode amplitudes of the waves $\langle A_k \rangle$, where

$$\langle A_k \rangle = \frac{1}{N_{\rm sectors}} \sum_{n=1}^{N_{\rm sectors}} (A_{\rm k,n} - A_{\rm k}). \tag{8.10}$$

8.3.2 Azimuthal Mode Analysis

For the azimuthal m wavenumbers the analysis is more straightforward. The disc is initially divided into annuli of fixed width ΔR in such a manner that each of these annuli is narrow enough to ensure the wave crests occupy only a small range in θ. In contrast to the radial modes, these annuli therefore need to become narrower with decreasing winding angle to maintain resolution. I found $\Delta R = 0.2$ (in code

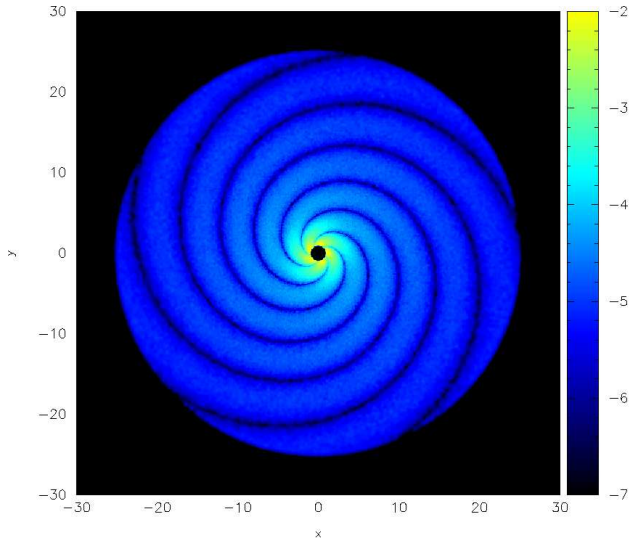

FIGURE 8.2: *Test case for Fourier analysis showing the imposed structure. The colour scale shows logarithmic surface density.*

---

units) to be sufficient for the purposes of this analysis. The azimuthal wavenumber amplitudes A_m within each annulus are then computed via

$$A_m = \frac{1}{N_{\rm ann}} \left| \sum_{j=1}^{N_{\rm ann}} e^{-im\theta_j} \right|, \tag{8.11}$$

where the θ_j are the azimuthal angles of the individual particles, $N_{\rm ann}$ the number of particles in each annulus and m is the radial wavenumber of the wave, corresponding to the number of arms in the spiral.

However, to ensure that the azimuthal m-mode amplitudes are specified at the same radii as the radial k-modes, an average value is taken of the m-mode amplitudes over all annuli where the central radius falls within that annulus in which the k-modes are determined.

8.3.3 Analysis Checks

To ensure that the results of the Fourier analysis are accurate, I ran the following test case. A disc with an underlying surface density profile $\Sigma \propto R^{-3/2}$ and an analytically superimposed structure was created with five spiral arms, such that $m = 5$ throughout the entire disc and with the radial wavelength increasing linearly from $\lambda_{\rm min} = 2$ to $\lambda_{\rm max} = 7.6$. This gives a total of five full wavelengths across the

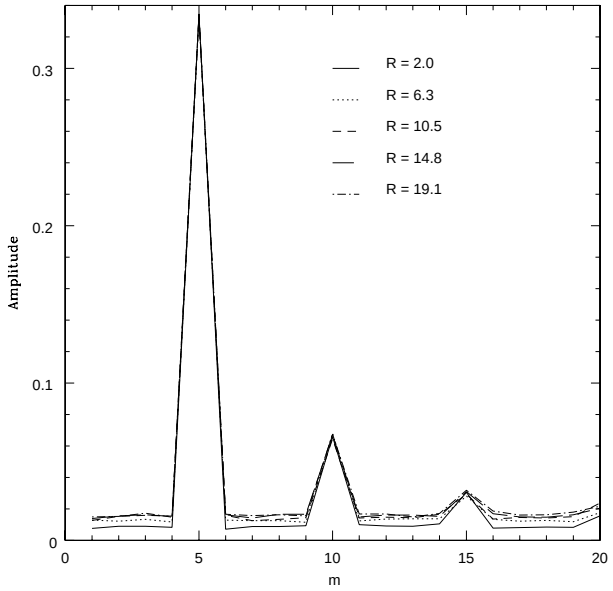

FIGURE 8.3: *Results of the Fourier decomposition of test disc shown in Fig. 8.2 in terms of the azimuthal wavenumber, m.*

face of the disc, which extends from $R = 1$ to $R = 25$. The surface density of the disc, clearly indicating the imposed structure, is shown in Fig. 8.2. The Fourier analysis was then conducted using the annulus widths quoted in Table 8.2, where as described above the width of the annuli increased linearly from the minimum to the maximum quoted value.

The results from the azimuthal Fourier decomposition are shown in Fig. 8.3, and show that the azimuthal wavenumber is resolved extremely well. The fundamental frequency $m = 5$ is clearly dominant, with no other modes except higher harmonics present at any significant amplitude. Note that the results for the azimuthal modes

Annuli	**Initial width**	**Final width**	**Sectors**
10	2.0	7.6	60
10	1.5	6.5	60
10	3.0	8.5	60
10	4.5	10.0	60
10	6.0	12.5	60

TABLE 8.2: *Details of the radial Fourier analyses for the test case*

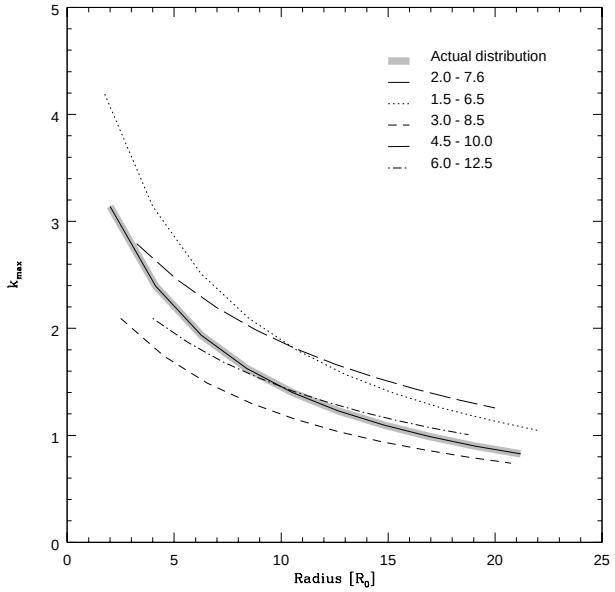

FIGURE 8.4: *Results of the Fourier decomposition of the test disc shown in Fig. 8.2 showing the peak radial wavenumber $k_{\max}$ as a function of radius. Various analyses are shown using the annuli given in Table 8.2.*

show no sensitivity to the annuli used for the analysis, and are quoted for the first case in Table 8.2 where the annuli width correspond exactly to the radial wavelengths.

The results of the Fourier decomposition for the radial wavenumbers are shown in Fig. 8.4, which shows both the actual distribution of wavenumbers (as calculated directly from the known distribution of wavelengths) and the distributions derived from analyses using the annuli given in Table 8.2. Note that the analysis using annuli that fit the wavelengths exactly correspondingly reproduces the exact result. Clearly there is scatter within the results for the radial wavenumbers, which arises from the fact that if the actual wavelength is not an integer divisor of the annulus over which the analysis is being conducted, more than one wavenumber appears to be excited. Note however that the scatter is never more than a factor of 1.5 above or below the true value, which is deemed to be accurate enough for the purposes of this analysis.

8.3.4 Resolution Limits

Throughout this paper, I have used discs of 500,000 particles when undertaking the Fourier analysis. From the fundamental SPH resolution limit of the local smoothing length h the maximum resolvable azimuthal and radial wavenumbers $m_{\rm max}$ and $k_{\rm max}$ can be evaluated as a function of radius throughout the disc, such that

$$m_{\rm max} = \frac{2\pi R}{h} = 2\pi \left(\frac{R}{H_{\rm sg}}\right)\left(\frac{H_{\rm sg}}{h}\right), \tag{8.12}$$

$$k_{\rm max} = \frac{2\pi}{h} = \frac{2\pi}{H_{\rm sg}}\left(\frac{H_{\rm sg}}{h}\right). \tag{8.13}$$

Since the approximate expected values for radial and azimuthal wavenumbers are such that $kH_{\rm sg} \approx 1$ and $mH_{\rm sg}/R \approx 1$ respectively, equations 8.12 and 8.13 show that the accuracy of the Fourier analysis is closely tied to the vertical resolution of the disc through $H_{\rm sg}/h$. Using the average smoothing length at each radius, these resolution limits are shown in Fig. 8.5. Here I have used data from the simulation where $\beta = 10$, as this gives the most conservative limits of all the experiments.

The vertical resolution of the disc as indicated by $H_{\rm sg}/h$ is shown in Fig. 8.6 for simulations using 500,000 particles. I find that the disc height is covered by approximately two smoothing lengths throughout, and thus is adequately resolved. For the Fourier analysis we therefore see that the expected peak wavenumbers are resolved by a factor of approximately 4π throughout the radial range. For the radial wavenumbers, $k < 10$ is the primary regime of interest, which is well resolved until at least $R = 25$, again adequate for the analyses I have undertaken. I conclude therefore that throughout the radial ranges of interest, the Fourier analyses I have presented are well resolved.

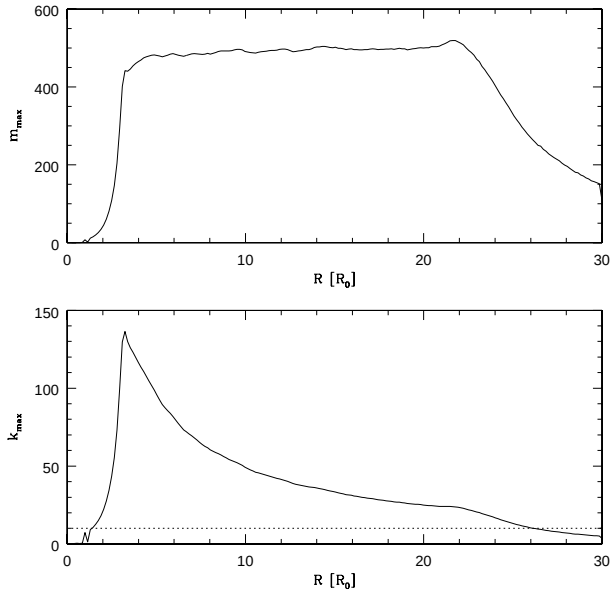

FIGURE 8.5: *Resolution limits for the Fourier analysis in terms of the azimuthal wavenumbers (top) and radial wavenumbers (bottom). The dashed line in the bottom plot indicates* $k_{\max} = 10$. *The simulation parameters are* $\beta = 10$, $q = 0.1$.

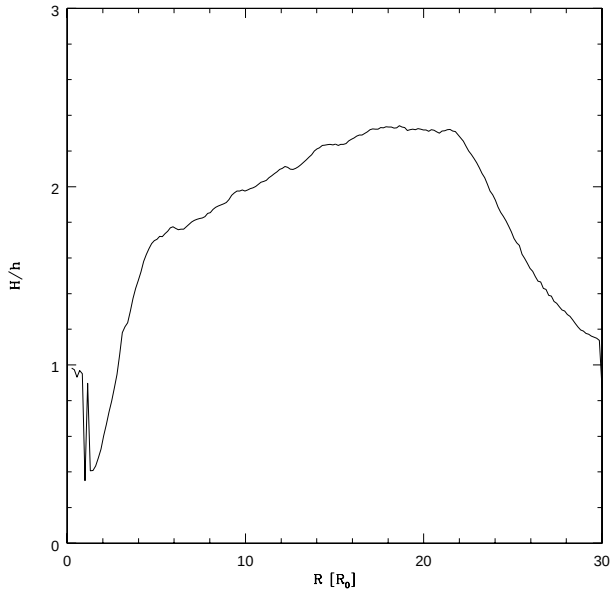

FIGURE 8.6: *Ratio of the disc scale thickness* H *to the average smoothing length* h *as a function of radius. Again the simulation parameters are* $\beta = 10$, $q = 0.1$.

8.4 Appendix D: The Entropy Argument

In this appendix I shall flesh out the argument presented earlier, that the relationship between the cooling and the density amplitude can be obtained by considering the entropy added across a weak shock. For simplicity, I shall use standard three-dimensional shock results throughout this appendix, as mentioned in the discussion section of this chapter. Upstream (unshocked) quantities are denoted by the subscript 0, and downstream (shocked) quantities are denoted by a corresponding subscript 1.

Firstly, consider the density change across a shock, which is given by standard normal shock relations as

$$\frac{\rho_1}{\rho_0} = \frac{(\gamma+1)\mathcal{M}^2}{2+(\gamma-1)\mathcal{M}^2}, \tag{8.14}$$

where $\mathcal{M}$ is the Mach number of the flow normal to the shock. In terms of the relative density perturbation $\delta\rho/\bar{\rho}$, note that $\delta\rho = \rho_1 - \rho_0$, and I shall assume that for weak shocks the average density $\bar{\rho}$ can be given by the unshocked quantity ρ_0. In this manner, the relative density perturbation becomes

$$\frac{\delta\rho}{\bar{\rho}} \approx \frac{\rho_1}{\rho_0} - 1 \tag{8.15}$$

$$\approx \frac{\mathcal{M}^2-1}{1+\left(\frac{\gamma-1}{2}\right)\mathcal{M}^2}, \tag{8.16}$$

which in the limit of very weak shocks $\mathcal{M} \to 1$ is given by

$$\frac{\delta\rho}{\bar{\rho}} \approx \frac{2}{\gamma+1}(\mathcal{M}^2-1) \sim (\mathcal{M}^2-1). \tag{8.17}$$

Consider now the entropy S, which can be given as

$$S = \ln\kappa, \quad \text{where} \quad \kappa = \frac{P}{\rho^\gamma}, \tag{8.18}$$

and κ is the adiabat. Across a shock, the entropy change ΔS is therefore given as

$$\Delta S = \ln\left(\frac{P_1}{P_0}\right) - \gamma\ln\left(\frac{\rho_1}{\rho_0}\right). \tag{8.19}$$

From above, and again using standard shock relations for pressure this becomes

$$\Delta S = \ln\left(1 + \frac{2\gamma}{\gamma+1}(\mathcal{M}^2 - 1)\right) - \gamma \ln\left(1 + \frac{\mathcal{M}^2 - 1}{1 + \left(\frac{\gamma-1}{2}\right)\mathcal{M}^2}\right), \tag{8.20}$$

which can be simplified using the Taylor expansion for natural logs $\ln(1+x) = x - x^2/2 + x^3/3 - ...$ to give (to first order)

$$\Delta S \approx \gamma(\mathcal{M}^2 - 1)\left[\frac{2}{\gamma+1} - \frac{1}{1 + \left(\frac{\gamma-1}{2}\right)\mathcal{M}^2}\right]. \tag{8.21}$$

In the limit of weak shocks as above, this then reduces to

$$\Delta S \approx (\mathcal{M}^2 - 1)^2\left[\frac{\gamma-1}{\gamma+1}\right] \sim (\mathcal{M}^2 - 1)^2. \tag{8.22}$$

Finally, note that from the first law of thermodynamics, the heat (irreversibly) added to the disc ΔQ^+ is related to the the entropy via

$$\Delta Q^+ = T\Delta S \tag{8.23}$$

$$= \frac{u(\gamma-1)}{\mathcal{R}}\Delta S, \tag{8.24}$$

where as before u is the specific internal energy and $\mathcal{R}$ is a specific gas constant.

In thermal equilibrium this is balanced by the applied cooling $\Delta Q^- = -u/\beta$, to give

$$\beta = \frac{\mathcal{R}}{(\gamma-1)\Delta S}, \tag{8.25}$$

and thus

$$\beta \sim \frac{1}{(\mathcal{M}^2 - 1)^2}. \tag{8.26}$$

Finally, by comparing this with the result for the relative density perturbation given in equation 8.17, one finds that

$$\frac{\delta\rho}{\bar{\rho}} \sim \frac{1}{\sqrt{\beta}}, \tag{8.27}$$

thereby confirming the results found earlier.

References

Acke, B., van den Ancker, M. E., Dullemond, C. P., van Boekel, R., & Waters, L. B. F. M.: 2004, *A&A* **422**, 621

Alexander, R. D., Armitage, P. J., & Cuadra, J.: 2008a, *MNRAS* **389**, 1655

Alexander, R. D., Armitage, P. J., Cuadra, J., & Begelman, M. C.: 2008b, *ApJ* **674**, 927

Alexander, R. D., Clarke, C. J., & Pringle, J. E.: 2006a, *MNRAS* **369**, 216

Alexander, R. D., Clarke, C. J., & Pringle, J. E.: 2006b, *MNRAS* **369**, 229

Altay, G., Croft, R. A. C., & Pelupessy, I.: 2008, *MNRAS* **386**, 1931

Anderson, J. D.: 1995, *Computational Fluid Dynamics: The Basics With Applications*, McGraw-Hill

Andre, P. & Montmerle, T.: 1994, *ApJ* **420**, 837

Andrews, S. M. & Williams, J. P.: 2005, *ApJ* **631**, 1134

Andrews, S. M. & Williams, J. P.: 2007a, *ApJ* **671**, 1800

Andrews, S. M. & Williams, J. P.: 2007b, *ApJ* **659**, 705

Andrews, S. M., Wilner, D. J., Hughes, A. M., Qi, C., & Dullemond, C. P.: 2009, *ApJ* **700**, 1502

Antonucci, R.: 1993, *ARA&A* **31**, 473

Armitage, P. J., Livio, M., & Pringle, J. E.: 2001, *MNRAS* **324**, 705

Athanassoula, E., Romero-Gómez, M., Bosma, A., & Masdemont, J. J.: 2009a, *MNRAS* **400**, 1706

Athanassoula, E., Romero-Gómez, M., & Masdemont, J. J.: 2009b, *MNRAS* **394**, 67

Backman, D., Marengo, M., Stapelfeldt, K., Su, K., Wilner, D., Dowell, C. D., Watson, D., Stansberry, J., Rieke, G., Megeath, T., Fazio, G., & Werner, M.: 2009, *ApJ* **690**, 1522

Balbus, S. A.: 2003, *ARA&A* **41**, 555

Balbus, S. A. & Hawley, J. F.: 1991, *ApJ* **376**, 214

Balbus, S. A. & Papaloizou, J. C. B.: 1999, *ApJ* **521**, 650

Bally, J.: 2007a, in B. G. Elmegreen & J. Palous (ed.), *IAU Symposium*, Vol. 237 of *IAU Symposium*, pp 165–171

Bally, J.: 2007b, *APSS* **311**, 15

Balsara, D. S.: 1995, *Journal of Computational Physics* **121**, 357

Barnes, J. & Hut, P.: 1986, *Nature* **324**, 446

Bartko, H., Eisenhauer, F., Fritz, T., Genzel, R., Gillessen, S., Martins, F., Ott, T., Paumard, T., Pfuhl, O., & Trippe, S.: 2008, *Journal of Physics Conference Series* **131(1)**, 012010

Bartko, H., Martins, F., Fritz, T. K., Genzel, R., Levin, Y., Perets, H. B., Paumard, T., Nayakshin, S., Gerhard, O., Alexander, T., Dodds-Eden, K., Eisenhauer, F., Gillessen, S., Mascetti, L., Ott, T., Perrin, G., Pfuhl, O., Reid, M. J., Rouan, D., Sternberg, A., & Trippe, S.: 2009, *ApJ* **697**, 1741

Bartko, H., Martins, F., Trippe, S., Fritz, T. K., Genzel, R., Ott, T., Eisenhauer, F., Gillessen, S., Paumard, T., Alexander, T., Dodds-Eden, K., Gerhard, O., Levin, Y., Mascetti, L., Nayakshin, S., Perets, H. B., Perrin, G., Pfuhl, O., Reid, M. J., Rouan, D., Zilka, M., & Sternberg, A.: 2010, *ApJ* **708**, 834

Basri, G. & Bertout, C.: 1989, *ApJ* **341**, 340

Bate, M. R. & Bonnell, I. A.: 2005, *MNRAS* **356**, 1201

Bate, M. R., Bonnell, I. A., & Price, N. M.: 1995, *MNRAS* **277**, 362

Beckwith, S. V. W., Sargent, A. I., Chini, R. S., & Guesten, R.: 1990, *AJ* **99**, 924

Begelman, M. C. & Shlosman, I.: 2009, *ApJ* **702**, L5

Bell, K. R. & Lin, D. N. C.: 1994, *ApJ* **427**, 987

Benedict, G. F., McArthur, B. E., Gatewood, G., Nelan, E., Cochran, W. D., Hatzes, A., Endl, M., Wittenmyer, R., Baliunas, S. L., Walker, G. A. H., Yang, S., Kürster, M., Els, S., & Paulson, D. B.: 2006, *AJ* **132**, 2206

Bentley, J. L.: 1975, *Commun. ACM* **18(9)**, 509

Benz, W.: 1990, in J. R. Buchler (ed.), *Numerical Modelling of Nonlinear Stellar Pulsations Problems and Prospects*, p. 269

Bertin, G.: 2000, *Dynamics of Galaxies*, Cambridge University Press

Bertin, G. & Haass, J.: 1982, *A&A* **108**, 265

Bertin, G., Lin, C. C., Lowe, S. A., & Thurstans, R. P.: 1989a, *ApJ* **338**, 78

Bertin, G., Lin, C. C., Lowe, S. A., & Thurstans, R. P.: 1989b, *ApJ* **338**, 104

Bertin, G. & Lodato, G.: 1999, *A&A* **350**, 694

Bertin, G. & Lodato, G.: 2001a, *A&A* **370**, 342

Bertin, G. & Lodato, G.: 2001b, *A&A* **370**, 342

Bertin, G. & Mark, J.: 1979, *SIAM J. Appl. Math., Vol. 36, p. 407 - 420* **36**, 407

Binney, J. & Tremaine, S.: 2008, *Galactic Dynamics: Second Edition*, Princeton University Press

Bisbas, T. G., Wünsch, R., Whitworth, A. P., & Hubber, D. A.: 2009, *A&A* **497**, 649

Black, D. C. & Bodenheimer, P.: 1975, *ApJ* **199**, 619

Blaes, O. M. & Balbus, S. A.: 1994, *ApJ* **421**, 163

Bodenheimer, P., Hubickyj, O., & Lissauer, J. J.: 2000, *Icarus* **143**, 2

Boley, A. C.: 2009, *ApJL* **695**, L53

Boley, A. C. & Durisen, R. H.: 2008, *ApJ* **685**, 1193

Boley, A. C., Durisen, R. H., Nordlund, Å., & Lord, J.: 2007, *ApJ* **665**, 1254

Boley, A. C., Mejía, A. C., Durisen, R. H., Cai, K., Pickett, M. K., & D'Alessio, P.: 2006, *ApJ* **651**, 517

Bonnell, I. A. & Rice, W. K. M.: 2008, *Science* **321**, 1060

Boss, A. P.: 1997, *Science* **276**, 1836

Boss, A. P.: 1998, *ApJ* **503**, 923

Boss, A. P.: 2000, *ApJ* **536**, L101

Boss, A. P.: 2004, *ApJ* **610**, 456

Britsch, M., Clarke, C. J., & Lodato, G.: 2008, *MNRAS* **385**, 1067

Burrows, C. J., Stapelfeldt, K. R., Watson, A. M., Krist, J. E., Ballester, G. E., Clarke, J. T., Crisp, D., Gallagher, III, J. S., Griffiths, R. E., Hester, J. J., Hoessel, J. G., Holtzman, J. A., Mould, J. R., Scowen, P. A., Trauger, J. T., & Westphal, J. A.: 1996, *ApJ* **473**, 437

Buta, R. J., Byrd, G. G., & Freeman, T.: 2003, *ApJ* **125**, 634

Cai, K., Durisen, R. H., Michael, S., Boley, A. C., Mejía, A. C., Pickett, M. K., & D'Alessio, P.: 2006, *ApJ* **636**, L149

Calvet, N., Muzerolle, J., Briceño, C., Hernández, J., Hartmann, L., Saucedo, J. L., & Gordon, K. D.: 2004, *AJ* **128**, 1294

Cameron, A. G. W.: 1978, *Moon and Planets* **18**, 5

Cha, S. & Whitworth, A. P.: 2003, *MNRAS* **340**, 73

Chakrabarti, S. & McKee, C. F.: 2008, *ApJ* **683**, 693

Chandrasekhar, S.: 1960, *Proceedings of the National Academy of Science* **46**, 253

Clarke, C. J.: 2009, *MNRAS* **396**, 1066

Clarke, C. J., Harper-Clark, E., & Lodato, G.: 2007, *MNRAS* **381**, 1543

Clarke, C. J. & Lodato, G.: 2009, *MNRAS* **398**, L6

Clarke, C. J. & Pringle, J. E.: 2004, *MNRAS* **351**, 1187

Collin, S. & Zahn, J.: 2008, *A&A* **477**, 419

Cullen, L. & Dehnen, W.: 2010, *arXiv:1006.1524*

Daisaka, H., Tanaka, H., & Ida, S.: 2001, *Icarus* **154**, 296

D'Alessio, P., Calvet, N., & Hartmann, L.: 2001, *ApJ* **553**, 321

D'Alessio, P., Canto, J., Calvet, N., & Lizano, S.: 1998, *ApJ* **500**, 411

Deegan, P.: 2009, *Ph.D. thesis*, Univ. Leicester, (2009)

Dehnen, W.: 2000, *ApJ* **536**, L39

Dehnen, W.: 2002, *Journal of Computational Physics* **179**, 27

Dehnen, W.: 2009, Private Communication

Dodson-Robinson, S. E. & Bodenheimer, P.: 2009, *ApJ* **695**, L159

Dolag, K. & Stasyszyn, F.: 2009, *MNRAS* **398**, 1678

Dolag, K., Vazza, F., Brunetti, G., & Tormen, G.: 2005, *MNRAS* **364**, 753

Draine, B. T.: 2006, *ApJ* **636**, 1114

Dullemond, C. P. & Dominik, C.: 2005, *A&A* **434**, 971

Dullemond, C. P., Hollenbach, D., Kamp, I., & D'Alessio, P.: 2007, in B. Reipurth, D. Jewitt, & K. Keil (eds.), *Protostars and Planets V*, pp 555–572

Durisen, R. H., Boss, A. P., Mayer, L., Nelson, A. F., Quinn, T., & Rice, W. K. M.: 2007, in *Protostars and Planets V*, pp 607–622

Dutrey, A., Guilloteau, S., Duvert, G., Prato, L., Simon, M., Schuster, K., & Menard, F.: 1996, *A&A* **309**, 493

Ebert, R.: 1994, *A&A* **286**, 997

Eisenhauer, F., Genzel, R., Alexander, T., Abuter, R., Paumard, T., Ott, T., Gilbert, A., Gillessen, S., Horrobin, M., Trippe, S., Bonnet, H., Dumas, C., Hubin, N., Kaufer, A., Kissler-Patig, M., Monnet, G., Ströbele, S., Szeifert, T., Eckart, A., Schödel, R., & Zucker, S.: 2005, *ApJ* **628**, 246

Eisner, J. A. & Carpenter, J. M.: 2006, *ApJ* **641**, 1162

Eisner, J. A., Plambeck, R. L., Carpenter, J. M., Corder, S. A., Qi, C., & Wilner, D.: 2008, *ApJ* **683**, 304

Ercolano, B., Clarke, C. J., & Drake, J. J.: 2009, *ApJ* **699**, 1639

Ercolano, B., Drake, J. J., Raymond, J. C., & Clarke, C. C.: 2008, *ApJ* **688**, 398

Esposito, L. W.: 1986, *Icarus* **67**, 345

Fan, Z. & Lou, Y.-Q.: 1999, *MNRAS* **307**, 645

Fatuzzo, M. & Melia, F.: 2009, *Publications of the Astronomical Society of the Pacific* **121**, 585

Fedele, D., van den Ancker, M. E., Henning, T., Jayawardhana, R., & Oliveira, J. M.: 2010, *A&A* **510**, A72+

Fehlberg, E.: 1968, *NASA Technical Report,* NASA-TR-R-287

Fehlberg, E.: 1969, *NASA Technical Report,* NASA-TR-R-315

Feldman, S. I. & Lin, C. C.: 1973, *Studies in Applied Mathematics* **52**, 1

Ferguson, J. W., Alexander, D. R., Allard, F., Barman, T., Bodnarik, J. G., Hauschildt, P. H., Heffner-Wong, A., & Tamanai, A.: 2005, *ApJ* **623**, 585

Finkbeiner, D. P., Davis, M., & Schlegel, D. J.: 1999, *ApJ* **524**, 867

Forgan, D. & Rice, K.: 2009, *MNRAS* **400**, 2022

Forgan, D., Rice, K., Stamatellos, D., & Whitworth, A.: 2009, *MNRAS* **394**, 882

Frank, J., King, A., & Raine, D. J.: 2002, *Accretion Power in Astrophysics: Third Edition*, Cambridge University Press

Fromang, S., Balbus, S. A., Terquem, C., & De Villiers, J.-P.: 2004, *ApJ* **616**, 364

Fromang, S., Terquem, C., & Balbus, S. A.: 2002, *MNRAS* **329**, 18

Fulk, D. A. & Quinn, D. W.: 1996, *Journal of Computational Physics* **126(1)**, 165

Gammie, C. F.: 1996, *ApJ* **457**, 355

Gammie, C. F.: 1999, Vol. 160 of *Astronomical Society of the Pacific Conference Series*, p. 122

Gammie, C. F.: 2001, *ApJ* **553**, 174

Genzel, R., Pichon, C., Eckart, A., Gerhard, O. E., & Ott, T.: 2000, *MNRAS* **317**, 348

Genzel, R., Schödel, R., Ott, T., Eisenhauer, F., Hofmann, R., Lehnert, M., Eckart, A., Alexander, T., Sternberg, A., Lenzen, R., Clénet, Y., Lacombe, F., Rouan, D., Renzini, A., & Tacconi-Garman, L. E.: 2003, *ApJ* **594**, 812

Gerhard, O.: 2001, *ApJ* **546**, L39

Ghez, A. M., Duchêne, G., Matthews, K., Hornstein, S. D., Tanner, A., Larkin, J., Morris, M., Becklin, E. E., Salim, S., Kremenek, T., Thompson, D., Soifer, B. T., Neugebauer, G., & McLean, I.: 2003, *ApJ* **586**, L127

Ghez, A. M., Salim, S., Hornstein, S. D., Tanner, A., Lu, J. R., Morris, M., Becklin, E. E., & Duchêne, G.: 2005, *ApJ* **620**, 744

Glassgold, A. E., Feigelson, E. D., & Montmerle, T.: 2000, *Protostars and Planets IV* p. 429

Goldreich, P. & Lynden-Bell, D.: 1965, *MNRAS* **130**, 125

Goldreich, P. & Tremaine, S.: 1979, *ApJ* **233**, 857

Greaves, J. S., Holland, W. S., Wyatt, M. C., Dent, W. R. F., Robson, E. I., Coulson, I. M., Jenness, T., Moriarty-Schieven, G. H., Davis, G. R., Butner, H. M., Gear, W. K., Dominik, C., & Walker, H. J.: 2005, *ApJ* **619**, L187

Greaves, J. S., Richards, A. M. S., Rice, W. K. M., & Muxlow, T. W. B.: 2008, *MNRAS* **391**, L74

Greene, T.: 2001, *American Scientist* **89,4**, 4

Gritschneder, M., Naab, T., Burkert, A., Walch, S., Heitsch, F., & Wetzstein, M.: 2009, *MNRAS* **393**, 21

Griv, E.: 2007, *Planetary and Space Science* **55**, 203

Griv, E. & Gedalin, M.: 2006, *Planetary and Space Science* **54**, 794

Gullbring, E., Calvet, N., Muzerolle, J., & Hartmann, L.: 2000, *ApJ* **544**, 927

Gullbring, E., Hartmann, L., Briceno, C., & Calvet, N.: 1998, *ApJ* **492**, 323

Haisch, Jr., K. E., Lada, E. A., & Lada, C. J.: 2001, *ApJ* **553**, L153

Hameury, J. & Lasota, J.: 2005, *A&A* **443**, 283

Harsono, D., Alexander, R. D., & Levin, Y.: 2010, *MNRAS, submitted*

Hartigan, P., Edwards, S., & Ghandour, L.: 1995, *ApJ* **452**, 736

Hartigan, P., Kenyon, S. J., Hartmann, L., Strom, S. E., Edwards, S., Welty, A. D., & Stauffer, J.: 1991, *ApJ* **382**, 617

Hartmann, L.: 2009a, *Accretion Processes in Star Formation: Second Edition*, Cambridge University Press

Hartmann, L.: 2009b, in *Dynamics of Discs and Planets, http://www.newton.ac.uk/programmes/DDP/seminars/081710004.ppt*

Hartmann, L., Calvet, N., Gullbring, E., & D'Alessio, P.: 1998, *ApJ* **495**, 385

Hartmann, L., D'Alessio, P., Calvet, N., & Muzerolle, J.: 2006, *ApJ* **648**, 484

Hartmann, L. & Kenyon, S. J.: 1996, *ARA&A* **34**, 207

Hawley, J. F. & Balbus, S. A.: 1991, *ApJ* **376**, 223

Hawley, J. F. & Balbus, S. A.: 1992, *ApJ* **400**, 595

Hawley, J. F., Gammie, C. F., & Balbus, S. A.: 1995, *ApJ* **440**, 742

Hawley, J. F., Gammie, C. F., & Balbus, S. A.: 1996, *ApJ* **464**, 690

Haynes, M. P.: 1979, *ApJ* **84**, 1830

Herbig, G. H.: 1977, *ApJ* **217**, 693

Hernquist, L.: 1987, *ApJS* **64**, 715

Hernquist, L.: 1993, *ApJ* **404**, 717

Hernquist, L. & Katz, N.: 1989, *ApJS* **70**, 419

Hildebrand, R. H.: 1983, *Royal Astronomy Society Quarterly Journal* **24**, 267

Hobbs, A. & Nayakshin, S.: 2009, *MNRAS* **394**, 191

Hockney, R. W. & Eastwood, J. W.: 1981, *Computer Simulation Using Particles*

Hohl, F.: 1971, *ApJ* **168**, 343

Hughes, A. M., Wilner, D. J., Qi, C., & Hogerheijde, M. R.: 2008, *ApJ* **678**, 1119

Inutsuka, S.: 2002, *Journal of Computational Physics* **179**, 238

Isella, A., Carpenter, J. M., & Sargent, A. I.: 2009, *ApJ* **701**, 260

Isella, A., Testi, L., Natta, A., Neri, R., Wilner, D., & Qi, C.: 2007, *A&A* **469**, 213

Janson, M., Brandner, W., Henning, T., Lenzen, R., McArthur, B., Benedict, G. F., Reffert, S., Nielsen, E., Close, L., Biller, B., Kellner, S., Günther, E., Hatzes, A., Masciadri, E., Geissler, K., & Hartung, M.: 2007, *AJ* **133**, 2442

Jeans, J. H.: 1902, *Philosophical Transactions of the Royal Society Series A* **199**, 1

Johnson, B. M. & Gammie, C. F.: 2003, *ApJ* **597**, 131

Kalas, P., Graham, J. R., Chiang, E., Fitzgerald, M. P., Clampin, M., Kite, E. S., Stapelfeldt, K., Marois, C., & Krist, J.: 2008, *Science* **322**, 1345

Kennedy, G. M. & Kenyon, S. J.: 2008, *ApJ* **673**, 502

Kenyon, S. J. & Hartmann, L.: 1987, *ApJ* **323**, 714

Kenyon, S. J. & Hartmann, L.: 1995, *ApJ* **101**, 117

Kenyon, S. J. & Hartmann, L. W.: 1991, *ApJ* **383**, 664

King, A. R., Pringle, J. E., & Livio, M.: 2007, *MNRAS* **376**, 1740

Kitamura, Y., Momose, M., Yokogawa, S., Kawabe, R., Tamura, M., & Ida, S.: 2002, *ApJ* **581**, 357

Klahr, H.: 2008, *New Astronomy Review* **52**, 78

Klahr, H. & Brandner, W.: 2006, *Planet Formation*

Kolykhalov, P. I. & Syunyaev, R. A.: 1980, *Soviet Astronomy Letters* **6**, 357

Kramer, C., Alves, J., Lada, C., Lada, E., Sievers, A., Ungerechts, H., & Walmsley, M.: 1998, *A&A* **329**, L33

Kratter, K. M., Matzner, C. D., & Krumholz, M. R.: 2008, *ApJ* **681**, 375

Kratter, K. M., Matzner, C. D., Krumholz, M. R., & Klein, R. I.: 2010, *ApJ* **708**, 1585

Kutta, W.: 1901, *Zeitschrift fr. Mathematische Physik* **46**, 435

Lada, C. J.: 1987, in M. Peimbert & J. Jugaku (ed.), *Star Forming Regions*, Vol. 115 of *IAU Symposium*, pp 1–17

Lagrange, A.-M., Gratadour, D., Chauvin, G., Fusco, T., Ehrenreich, D., Mouillet, D., Rousset, G., Rouan, D., Allard, F., Gendron, É., Charton, J., Mugnier, L., Rabou, P., Montri, J., & Lacombe, F.: 2009, *A&A* **493**, L21

Landau, L. D. & Lifshitz, E. M.: 1959, *Fluid mechanics*

Landshoff, R.: 1930, *Los Alamos Laboratory Report*

Larson, R. B.: 1992, *MNRAS* **256**, 641

Lasota, J.: 2008, *New Astronomy Review* **51**, 752

Lasota, J.-P.: 2001, *New Astronomy Review* **45**, 449

Lattanzio, J. C., Monaghan, J. J., Pongracic, H., & Schwarz, M. P.: 1985, *MNRAS* **215**, 125

Lau, Y. Y. & Bertin, G.: 1978, *ApJ* **226**, 508

Laughlin, G. & Bodenheimer, P.: 1994, *ApJ* **436**, 335

Laughlin, G., Korchagin, V., & Adams, F. C.: 1997, *ApJ* **477**, 410

Levin, Y.: 2003, *arXiv:astro-ph/0307084*

Levin, Y.: 2007, *MNRAS* **374**, 515

Levin, Y. & Beloborodov, A. M.: 2003, *ApJL* **590**, L33

Lewin, W. H. G. & van der Klis, M.: 2006, *Compact stellar X-ray sources*

Lewin, W. H. G., van Paradijs, J., & van den Heuvel, E. P. J.: 1997, *X-ray Binaries*

Li, C., Han, N., & Lin, C.: 1976, *Scientia Sinica* **19**, 665

Lin, C. C. & Lau, Y. Y.: 1979, *Studies in Applied Mathematics* **60**, 97

Lin, C. C. & Shu, F. H.: 1964, *ApJ* **140**, 646

Lin, C. C. & Shu, F. H.: 1966, *Proceedings of the National Academy of Science* **55**, 229

Lin, D. N. C. & Pringle, J. E.: 1987, *MNRAS* **225**, 607

Lindblad, B.: 1927, *MNRAS* **87**, 420

Lissauer, J. J.: 1993, *ARA&A* **31**, 129

Lissauer, J. J. & Stevenson, D. J.: 2007, in B. Reipurth, D. Jewitt, & K. Keil (eds.), *Protostars and Planets V*, pp 591–606

Lodato, G.: 2007, *Nuovo Cimento Rivista Serie* **30**, 293

Lodato, G. & Bertin, G.: 2001, *A&A* **375**, 455

Lodato, G. & Bertin, G.: 2003, *A&A* **408**, 1015

Lodato, G. & Clarke, C. J.: 2004, *MNRAS* **353**, 841

Lodato, G., Cossins, P. J., Clarke, C., & Testi, L.: 2010, in Bertin. G, De Luca. F, Lodato. G, Pozzoli. R, Rome. M (ed.), *Proceeding of the International Symposium "Plasmas in the Laboratory and in the Universe: Interactions, Patterns and Turbulence"*, AIP Conf. Ser.

Lodato, G., Delgado-Donate, E., & Clarke, C. J.: 2005, *MNRAS* **364**, L91

Lodato, G. & Price, D. J.: 2010, *MNRAS* **405**, 1212

Lodato, G. & Rice, W. K. M.: 2004, *MNRAS* **351**, 630

Lodato, G. & Rice, W. K. M.: 2005, *MNRAS* **358**, 1489

Lombardi, J. C., Sills, A., Rasio, F. A., & Shapiro, S. L.: 1999, *Journal of Computational Physics* **152**, 687

Lu, J. R., Ghez, A. M., Hornstein, S. D., Morris, M., Matthews, K., Thompson, D. J., & Becklin, E. E.: 2006, *Journal of Physics Conference Series* **54**, 279

Lu, J. R., Ghez, A. M., Hornstein, S. D., Morris, M. R., Becklin, E. E., & Matthews, K.: 2009, *ApJ* **690**, 1463

Lynden-Bell, D. & Kalnajs, A. J.: 1972, *MNRAS* **157**, 1

Lynden-Bell, D. & Pringle, J. E.: 1974, *MNRAS* **168**, 603

Marcy, G., Butler, R. P., Fischer, D., Vogt, S., Wright, J. T., Tinney, C. G., & Jones, H. R. A.: 2005, *Progress of Theoretical Physics Supplement* **158**, 24

Marcy, G. W. & Butler, R. P.: 2000, *Publications of the Astronomical Society of the Pacific* **112**, 137

Marengo, M., Megeath, S. T., Fazio, G. G., Stapelfeldt, K. R., Werner, M. W., & Backman, D. E.: 2006, *ApJ* **647**, 1437

Marigo, P. & Aringer, B.: 2009, *A&A* **508**, 1539

Marois, C., Macintosh, B., Barman, T., Zuckerman, B., Song, I., Patience, J., Lafrenière, D., & Doyon, R.: 2008, *Science* **322**, 1348

Matzner, C. D. & Levin, Y.: 2005, *ApJ* **628**, 817

Mayer, L., Lufkin, G., Quinn, T., & Wadsley, J.: 2007, *ApJl* **661**, L77

Mayer, L., Quinn, T., Wadsley, J., & Stadel, J.: 2002, *Science* **298**, 1756

Mayer, L., Quinn, T., Wadsley, J., & Stadel, J.: 2003, in D. Deming & S. Seager (ed.), *Scientific Frontiers in Research on Extrasolar Planets*, Vol. 294 of *Astronomical Society of the Pacific Conference Series*, pp 281–286

McCaughrean, M. J. & O'Dell, C. R.: 1996, *AJ* **111**, 1977

Mejía, A. C., Durisen, R. H., Pickett, M. K., & Cai, K.: 2005, *ApJ* **619**, 1098

Misner, C. W., Thorne, K. S., & Wheeler, J. A.: 1973, *Gravitation*

Miyake, K. & Nakagawa, Y.: 1993, *Icarus* **106**, 20

Monaghan, J.: 1989, *Journal of Computational Physics* **82(1)**, 1

Monaghan, J. J.: 1985, *Computer Physics Reports* **3(2)**, 71

Monaghan, J. J.: 1992, *ARA&A* **30**, 543

Monaghan, J. J. & Lattanzio, J. C.: 1985, *A&A* **149**, 135

Montenegro, L. E., Yuan, C., & Elmegreen, B. G.: 1999, *ApJ* **520**, 592

Moore, A. J., Quillen, A. C., & Edgar, R. G.: 2008, *arXiv:0809.2855*

Morris, J. P. & Monaghan, J. J.: 1997, *Journal of Computational Physics* **136(1)**, 41

Murray, J. R.: 1996, *MNRAS* **279**, 402

Myers, P. C. & Benson, P. J.: 1983, *ApJ* **266**, 309

Myers, P. C., Linke, R. A., & Benson, P. J.: 1983, *ApJ* **264**, 517

Natta, A., Testi, L., Neri, R., Shepherd, D. S., & Wilner, D. J.: 2004, *A&A* **416**, 179

Nayakshin, S., Cha, S., & Hobbs, A.: 2009, *MNRAS* **397**, 1314

Nayakshin, S. & Cuadra, J.: 2005, *A&A* **437**, 437

Nayakshin, S., Cuadra, J., & Springel, V.: 2007, *MNRAS* **379**, 21

Nayakshin, S., Dehnen, W., Cuadra, J., & Genzel, R.: 2006, *MNRAS* **366**, 1410

Nayakshin, S. & Sunyaev, R.: 2005, *MNRAS* **364**, L23

Nelson, A. F., Wetzstein, M., & Naab, T.: 2009, *ApJS* **184**, 326

Owen, J. E., Ercolano, B., Clarke, C. J., & Alexander, R. D.: 2010, *MNRAS* **401**, 1415

Paczynski, B.: 1978, *Acta Astronomica* **28**, 91

Pardo, J., Cernicharo, J., & Serabyn, E.: 2002, in J. Vernin, Z. Benkhaldoun, & C. Muñoz-Tuñón (ed.), *Astronomical Site Evaluation in the Visible and Radio Range*, Vol. 266 of *Astronomical Society of the Pacific Conference Series*, p. 188

Paumard, T., Genzel, R., Martins, F., Nayakshin, S., Beloborodov, A. M., Levin, Y., Trippe, S., Eisenhauer, F., Ott, T., Gillessen, S., Abuter, R., Cuadra, J., Alexander, T., & Sternberg, A.: 2006, *ApJ* **643**, 1011

Pawlik, A. H. & Schaye, J.: 2008, *MNRAS* **389**, 651

Pearce, F. R.: 2010, Private Communication

Petkova, M. & Springel, V.: 2009, *MNRAS* **396**, 1383

Pickett, B. K., Mejía, A. C., Durisen, R. H., Cassen, P. M., Berry, D. K., & Link, R. P.: 2003, *ApJ* **590**, 1060

Piétu, V., Guilloteau, S., & Dutrey, A.: 2005, *A&A* **443**, 945

Porco, C. C., Helfenstein, P., Thomas, P. C., Ingersoll, A. P., Wisdom, J., West, R., Neukum, G., Denk, T., Wagner, R., Roatsch, T., Kieffer, S., Turtle, E., McEwen, A., Johnson, T. V., Rathbun, J., Veverka, J., Wilson, D., Perry, J., Spitale, J., Brahic, A., Burns, J. A., Del Genio, A. D., Dones, L., Murray, C. D., & Squyres, S.: 2006, *Science* **311**, 1393

Press, W., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P.: 2007, *Numerical Recipes – the Art of Scientific Computing (3rd Edition)*, Cambridge University Press

Price, D.: 2005, *arXiv:0507.472*

Price, D. J.: 2007, *Publications of the Astronomical Society of Australia* **24**, 159

Price, D. J.: 2010, *MNRAS* **401**, 1475

Price, D. J. & Monaghan, J. J.: 2004, *MNRAS* **348**, 139

Price, D. J. & Monaghan, J. J.: 2005, *MNRAS* **364**, 384

Price, D. J. & Monaghan, J. J.: 2007, *MNRAS* **374**, 1347

Pringle, J. E.: 1981, *ARA&A* **19**, 137

Quinn, T., Katz, N., Stadel, J., & Lake, G.: 1997, *arXiv:astro-ph/9710043*

Rafikov, R. R.: 2005, *ApJ* **621**, L69

Rafikov, R. R.: 2006, *ApJ* **646**, 288

Rafikov, R. R.: 2007, *ApJ* **662**, 642

Rafikov, R. R.: 2009, *ApJ* **704**, 281

Read, J. I., Hayfield, T., & Agertz, O.: 2010, *MNRAS* **405**, 1513

Ricci, L., Testi, L., Natta, A., Neri, R., Cabrit, S., & Herczeg, G. J.: 2010, *A&A* **512**, A15+

Rice, W. K. M. & Armitage, P. J.: 2009, *MNRAS* **396**, 2228

Rice, W. K. M., Armitage, P. J., Bate, M. R., & Bonnell, I. A.: 2003a, *MNRAS* **339**, 1025

Rice, W. K. M., Armitage, P. J., Bonnell, I. A., Bate, M. R., Jeffers, S. V., & Vine, S. G.: 2003b, *MNRAS* **346**, L36

Rice, W. K. M., Lodato, G., & Armitage, P. J.: 2005, *MNRAS* **364**, L56

Rice, W. K. M., Lodato, G., Pringle, J. E., Armitage, P. J., & Bonnell, I. A.: 2004, *MNRAS* **355**, 543

Rice, W. K. M., Lodato, G., Pringle, J. E., Armitage, P. J., & Bonnell, I. A.: 2006, *MNRAS* **372**, L9

Rice, W. K. M., Mayo, J. H., & Armitage, P. J.: 2010, *MNRAS* **402**, 1740

Rodríguez, L. F., Loinard, L., D'Alessio, P., Wilner, D. J., & Ho, P. T. P.: 2005, *ApJ* **621**, L133

Romaniello, M., Robberto, M., & Panagia, N.: 2004, *ApJ* **608**, 220

Romero-Gómez, M., Athanassoula, E., Masdemont, J. J., & García-Gómez, C.: 2007, *A&A* **472**, 63

Romero-Gómez, M., Masdemont, J. J., Athanassoula, E., & García-Gómez, C.: 2006, *A&A* **453**, 39

Rosswog, S.: 2009, *New Astronomy Review* **53**, 78

Rosswog, S., Davies, M. B., Thielemann, F., & Piran, T.: 2000, *A&A* **360**, 171

Rosswog, S. & Price, D.: 2007, *MNRAS* **379**, 915

Salmon, J., Charnoz, S., Crida, A., & Brahic, A.: 2009, *SF2A-2009: Proceedings of the Annual meeting of the French Society of Astronomy and Astrophysics* p. 249

Sano, T., Inutsuka, S.-i., Turner, N. J., & Stone, J. M.: 2004, *ApJ* **605**, 321

Sellwood, J. A. & Carlberg, R. G.: 1984, *ApJ* **282**, 61

Shakura, N. I. & Sunyaev, R. A.: 1973, *A&A* **24**, 337

Shlosman, I. & Begelman, M. C.: 1987, *Nature* **329**, 810

Shlosman, I. & Begelman, M. C.: 1989, *ApJ* **341**, 685

Shlosman, I., Begelman, M. C., & Frank, J.: 1990, *Nature* **345**, 679

Shu, F. H.: 1970, *ApJ* **160**, 99

Sicilia-Aguilar, A., Henning, T., & Hartmann, L. W.: 2010, *ApJ* **710**, 597

Spahn, F., Schmidt, J., Albers, N., Hörning, M., Makuch, M., Seiß, M., Kempf, S., Srama, R., Dikarev, V., Helfert, S., Moragas-Klostermeyer, G., Krivov, A. V., Sremčević, M., Tuzzolino, A. J., Economou, T., & Grün, E.: 2006, *Science* **311**, 1416

Springel, V.: 2005, *MNRAS* **364**, 1105

Springel, V. & Hernquist, L.: 2002, *MNRAS* **333**, 649

Springel, V., Yoshida, N., & White, S. D. M.: 2001, *New Astronomy* **6**, 79

Stamatellos, D., Hubber, D. A., & Whitworth, A. P.: 2007a, *MNRAS* **382**, L30

Stamatellos, D. & Whitworth, A. P.: 2008, *A&A* **480**, 879

Stamatellos, D. & Whitworth, A. P.: 2009a, *MNRAS* **392**, 413

Stamatellos, D. & Whitworth, A. P.: 2009b, *MNRAS* **400**, 1563

Stamatellos, D., Whitworth, A. P., Bisbas, T., & Goodwin, S.: 2007b, *A&A* **475**, 37

Steinmetz, M.: 1996, *MNRAS* **278**, 1005

Steinmetz, M. & Mueller, E.: 1993, *A&A* **268**, 391

Stone, J. M., Hawley, J. F., Gammie, C. F., & Balbus, S. A.: 1996, *ApJ* **463**, 656

Tasker, E. J., Brunino, R., Mitchell, N. L., Michielsen, D., Hopton, S., Pearce, F. R., Bryan, G. L., & Theuns, T.: 2008, *MNRAS* **390**, 1267

Testi, L. & Leurini, S.: 2008, *New Astronomy Review* **52**, 105

Testi, L., Natta, A., Shepherd, D. S., & Wilner, D. J.: 2003, *A&A* **403**, 323

Toomre, A.: 1964, *ApJ* **139**, 1217

Toomre, A.: 1969, *ApJ* **158**, 899

Vandervoort, P. O.: 1970a, *ApJ* **161**, 87

Vandervoort, P. O.: 1970b, *ApJ* **161**, 67

Velikhov, E. P.: 1959, *Soviet Physics - JETP* **36**, 995

Vollmer, B., Beckert, T., & Davies, R. I.: 2008, *A&A* **491**, 441

von Neumann, J. & Richtmyer, R. D.: 1950, *Journal of Applied Physics* **21**, 232

Vorobyov, E. I. & Basu, S.: 2005, *ApJL* **633**, L137

Vorobyov, E. I. & Basu, S.: 2006, *ApJ* **650**, 956

Vorobyov, E. I. & Basu, S.: 2008, *ApJ* **676**, L139

Walch, S., Burkert, A., Whitworth, A., Naab, T., & Gritschneder, M.: 2009, *MNRAS* **400**, 13

Walch, S., Naab, T., Whitworth, A., Burkert, A., & Gritschneder, M.: 2010, *MNRAS* **402**, 2253

Walmswell, J., Clarke, C., & Cossins, P.: 2010, *In prep*

Weidenschilling, S. J.: 1977, *Ap&SS* **51**, 153

Wetzstein, M., Nelson, A. F., Naab, T., & Burkert, A.: 2009, *ApJS* **184**, 298

Wevers, B. M. H. R., Appleton, P. N., Davies, R. D., & Hart, L.: 1984, *A&A* **140**, 125

Whitworth, A. P. & Stamatellos, D.: 2006, *A&A* **458**, 817

Wilner, D. J., D'Alessio, P., Calvet, N., Claussen, M. J., & Hartmann, L.: 2005, *ApJ* **626**, L109

Winters, W. F., Balbus, S. A., & Hawley, J. F.: 2003, *ApJ* **589**, 543

Wolf, S. & D'Angelo, G.: 2005, *ApJ* **619**, 1114

Yu, Q. & Tremaine, S.: 2002, *MNRAS* **335**, 965

Zhu, Z., Hartmann, L., Calvet, N., Hernandez, J., Tannirkulam, A., & D'Alessio, P.: 2008, *ApJ* **684**, 1281

Zhu, Z., Hartmann, L., Gammie, C., & McKinney, J. C.: 2009, *ApJ* **701**, 620

Zhu, Z., Hartmann, L., Gammie, C. F., Book, L. G., Simon, J. B., & Engelhard, E.: 2010, *ApJ* **713**, 1134