Probing Self-Gravitating Protostellar Discs using Smoothed Particle Hydrodynamics and Radiative Transfer

Duncan Hugh Forgan, MPhys (Hons)

Institute for Astronomy
School of Physics and Astronomy

University of Edinburgh
Doctor of Philosophy

September 2010

Abstract

Stars are likely to form with non-zero initial angular momentum, and will consequently possess a substantial gaseous protostellar disc in the early phases of their evolution. At this early stage, the disc mass is expected to be comparable to the mass of the protostar. **The disc's self-gravity therefore plays an important role in the subsequent evolution of the system**, regulating the accretion of matter onto the protostar, as well as being potentially capable of forming low mass stars and massive planets by disc fragmentation. The protostellar disc may later evolve into a protoplanetary disc, providing the feedstock for planet formation. Therefore, if the current stellar populations and exoplanetary systems are to be understood, an understanding of the evolution of protostellar discs is crucial, especially their earliest self-gravitating phases. I have used various methods of numerical simulation to probe the physics of self-gravitating protostellar discs and their constituents.

When constructing a model for self-gravitating protostellar discs, including detailed thermodynamics and radiative transfer is essential. I have developed two distinct numerical techniques for incorporating radiative transfer into Smoothed Particle Hydrodynamics (SPH) simulations. The first allows the modelling of frequency-averaged radiative transfer during the SPH simulation, in effect approximating radiative SPH (RSPH) with only a marginal increase in runtime (around 6%). The second takes the output from SPH simulations, and creates synthetic, wavelength-dependent telescope images and spectra of SPH systems. This allows the direct construction of observables from SPH simulations, providing, for the first time, a direct connection between the output of SPH simulations and observations.

I have used these numerical methods to analyse, in detail, the local angular momentum transport induced by self-gravity in protostellar discs, testing the robustness of the "pseudo-viscous" analytical approximation for local disc stresses. I confirm that semi-analytical disc modellers are justified in using the pseudo-viscous approximation in some cases, but I also outline the limits in which non-local transport effects causes the approximation to fail.

Also, I have investigated the evolution of protostellar discs when perturbed by a secondary companion, in particular identifying whether such events will in general trigger a) a disc fragmentation event, or b) a stellar outburst event. For case a), I found no significant evidence that perturbation by a companion improves the possibility of disc fragmentation in compact discs - in case b), I found that stellar outburst events do indeed occur, but they are unlikely to be seen by observers due to their rare occurrence, as well as due to self-obscuration effects.

Declaration

This thesis contains no material that has been accepted for the award of any other degree or diploma. To the best of the author's knowledge, this thesis contains no material previously published or written by another author, except where due reference is made in the text of the thesis. All articles published by the author during this PhD are listed in Appendix E. All work presented is primarily that of the author.

Science is a collaborative pursuit, and the science presented in this thesis very much so. The collaborators listed below have all provided important support in obtaining the presented results.

Dr. Ken Rice
Dr. Dimitris Stamatellos
Prof. Anthony Whitworth
Dr. Giuseppe Lodato
Dr. Peter Cossins
Vladimir Bozhilov

Minor alterations have been made to results previously published in peer-reviewed academic journals to maintain consistency of style.

Duncan Forgan

15th September 2010

Acknowledgements

It is said that Descartes devised the Cartesian co-ordinate system while in bed one night, watching the motion of a fly on his ceiling. Such attention to one's surroundings is the basis of all human discovery. For encouraging me to study my surroundings, I dedicate this tome to my parents. Without their love and support, none of my ambitions could ever have borne fruit. My brother also deserves credit, for continuing to provide healthy sibling rivalry for the last twenty-six years. I am eternally grateful to Heather for many reasons, above all for keeping my feet on the ground while allowing my head to reach the clouds.

Being a PhD student is not a solitary experience, and would be much less enjoyable without the camaraderie and trench humour of my colleagues. In particular, I would like to thank my office-mates, Alina and Brendan (and formerly Rachel), for the banter and general *bonhomie*, and Julia for our faithful observance of coffee-time.

I would also like to take the opportunity to thank my supervisor, Dr. Ken Rice, for giving first-rate supervision of my work. In particular, I am grateful to him for allowing me freedom in defining my research aims and methods. This independence gave me the confidence to attempt scientific endeavour boldly, but not without the due care and precision required. I must also mention the other members of staff at the Institute for Astronomy for my undergraduate education and postgraduate tutelage. Output from Dr. Daniel Price's SPLASH visualisation software (Price, 2007) appears frequently in this book - for developing this software (and consequently saving me a great deal of time), I am most appreciative.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Although I consider this work unworthy to be put before you, yet I am fully confident that you will be kind enough to accept it, seeing that I could not give you a more valuable gift than the means of being able in a very short space of time to grasp all that I, over so many years and with so much affliction and peril, have learned and understood.

Niccolò Machiavelli, in correspondence with the Magnificent Lorenzo dè Medici

1.1 Discs in Astrophysics

Discs are ubiquitous in astrophysics. They are second only to the Greeks' favoured shape - the sphere - in their simplicity and success in describing the contents of the Universe. Astronomers have successfully invoked the existence of disc structures to explain phenomena at every step of the cosmic distance ladder - from planets to stars, galaxies and beyond. They have become critical to our understanding of the heavens, and are responsible for some of the most wonderful displays of Nature.

The concept of astrophysical discs finds its beginning in the Scientific Revolution of the Renaissance. Galileo Galilei, one of the fathers of modern experimental physics and astronomy, observed Saturn with the newly-invented refracting telescope in 1610 (four hundred years ago to the time of press). The puzzling image he obtained of the planet led him to conclude the planet had "ears". The disappearance of these ears two years later puzzled him further. It is said that he exclaimed, "Has Saturn swallowed its children?", referencing the actions of the planet's

namesake, the Greco-Roman Titan who consumed his children to prevent their attempts to overthrow him. It would be several decades before Christiaan Huygens, with vastly improved instruments, would correctly deduce that the "ears" were in fact a thin *ring*. Their mysterious disappearance in 1612 was a consequence of the ring system aligning with Earth such that it was edge-on. Giovanni Cassini would later show that the ring was not a single object, and was in fact divided into several rings.

Four centuries later, and his legacy lives on in NASA's *Cassini* probe, which affords us a much better picture of the rings. As proven by James Clerk Maxwell in 1859, the rings are not solid, and are instead collections of particles of varying sizes. They range from a few metres across down to a few micrometres, consisting mostly of ice with some silicates (Nicholson et al., 2008). The rings themselves are extremely thin - while thousands of kilometres in radius, they are only a few metres thick. The total mass in the ring is comparable to the typical mass of Saturn's larger moons, and hence the rings are subject to gravitational perturbations from the other satellites. In some cases, the rings exist thanks to the gravity of the moons, which act as "shepherds" (as shown in Figure 1.1).

Figure 1.1: A *Cassini* visible light image of Saturn's A and F rings. The thin F ring is shepherded by the moon Prometheus (both in the right hand side of the image). *Cassini* was approximately 200,000 miles from Prometheus when this image was taken. Image credit: NASA/JPL/Space Science Institute.

The use of disc geometries as a modelling tool would prove to be essential to the understanding of the Milky Way itself. In the 11th Century, Islamic astronomers confirmed through parallax measurements that the Milky Way was relatively distant compared to the stars. Six hundred years later, Galileo's observations confirmed it was composed of a large collection of stars. It would be this insight that would lead many to hypothesise that the Milky Way's true shape was a flattened disc. The Sun's position inside the disc would result in the thin band of

stars that we see in the night sky.

The early 20th Century saw Jacobus Kapteyn and Harlow Shapley continue the work of Herschel, who attempted to elucidate this structure in more detail. While all agreed the Milky Way was essentially disc-like, they differed on the location of the Sun within it. Kapteyn favoured a heliocentric Milky Way, whereas Shapley favoured a model where the Sun was far from the centre. The Great Debate that ensued - which included the nature of observed spiral nebulae or "island universes" - would be settled by a better understanding of interstellar extinction by dust, and the use of Cepheid variables to calculate the distance to the spiral nebulae. These developments would later vindicate Shapley's rejection of the heliocentric hypothesis. They would also confirm that the "island universes" (or *galaxies* as we now call them, such as M51 in Figure 1.2) were indeed analogues of the Milky Way. The later discovery that some of these galaxies were "active", emitting large amounts of radiation from a very small region at the centre of the galaxy, would require yet another disc structure to explain.

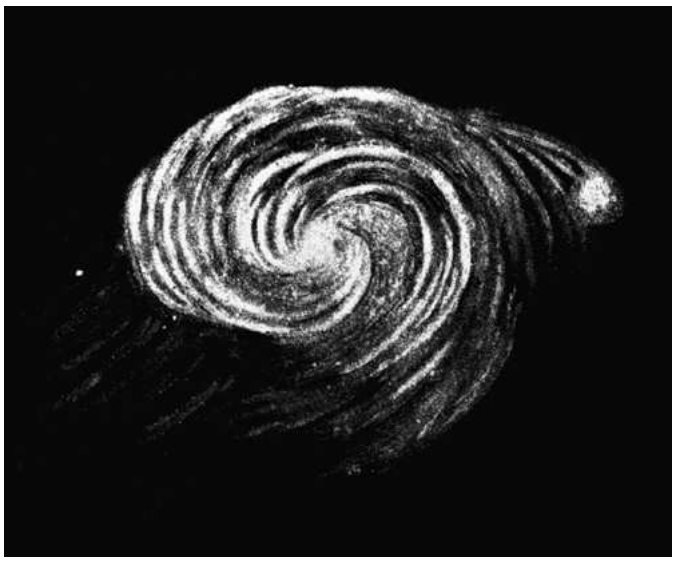

Figure 1.2: A sketch of M51 (the Whirlpool galaxy), made by Lord Rosse in 1845. M51 consists of two interacting galaxies (designated M51A and M51B). Image credit: Lord Rosse (copyright expired).

Producing such overwhelming power from a small region requires a very efficient engine; one of the most efficient engines in Nature is the *black hole*. An object so dense that light cannot escape from its gravitational pull, the black hole's existence as a compact object of stellar mass had been generally agreed as an explanation for X-Ray binary systems. These systems contain a black hole with a stellar companion. The black hole strips material from the companion, heating it to high temperatures and emitting X-Ray radiation in the process. The angular momentum of the material in these binaries would require an *accretion disc* to form, which regulates the flow of matter onto the black hole.

The black hole's high efficiency in converting matter to energy made it an obvious contender for the central engine of active galaxies. However, the mass of these central engines would need to be millions of times higher than the stellar black holes - so-called *supermassive black holes*. Astonishing theoretical success (and strong observational evidence) has since made clear that supermassive black holes appear to exist at the centre of many, if not all galaxies. Infrared observations of our own Galactic Centre (in the direction of Sagittarius A*) has shown a population of massive stars with orbits consistent with the presence of a supermassive black hole (Figure 1.3). Surrounded by an accretion disc comparable in size to the Solar System, these black holes (if actively accreting) can shine in much the same fashion as their stellar counterparts, blasting incredible energy through jets and outshining their host galaxy at the peak of their powers.

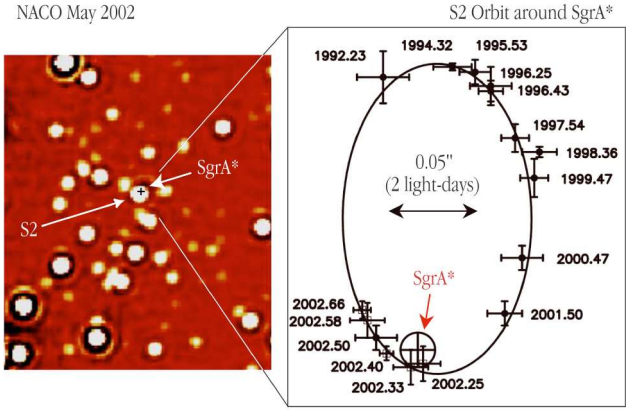

Figure 1.3: The Galactic Centre, imaged in the infrared using the NACO instrument on the VLT YEPUN telescope at ESO. Orbits of stars around the Galactic Centre indicate that there is a very compact object at the location of the object Sgr A*, too massive and too dense to avoid catastrophic gravitational collapse. It is most likely that this object is a supermassive black hole. Image credit: ESO.

We have seen how discs feature in many aspects of astronomy, but I have deliberately avoided until now the discs most relevant to this thesis - the *circumstellar discs*.

1.2 Discs in Star and Planet Formation

The larger role of discs in the formation of Solar Systems was first considered in the 18th Century, most noticeably by Laplace, who restated and developed the concepts involved. By considering the locations of the planets in orbit around the Sun, the presence of a disc structure in the Solar System becomes obvious. The planets are (more or less) orbiting in the same plane (the *ecliptic*), and in the same direction. It is reasonable therefore to assume the progenitor

of the planets also rotated in a single plane with a single direction. The simplest object that satisfies these criteria is a disc.

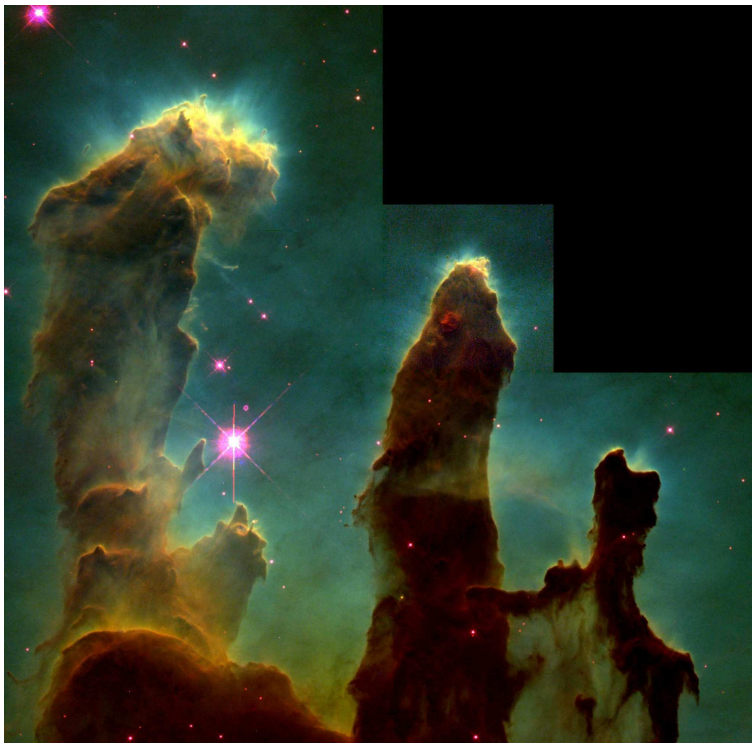

Figure 1.4: The Eagle Nebula, and the so-called "Pillars of Creation", a known site of active star formation. This image was taken by the Hubble Space Telescope, using the WFPC2 instrument. Image Credit: NASA.

To understand the origin of this disc, let us now consider the formation of the Sun itself. Stars are born in giant molecular clouds (GMCs), turbulent structures of gas and dust (like the Eagle Nebula shown in Figure 1.4). Regions of the cloud become gravitationally unstable and then begin to collapse to form stars. As the collapse proceeds, conservation of angular momentum requires that rotation around some axis must be preserved. While some of this angular momentum will end up in the protostar formed immediately after the collapse, most of it will be distributed into a disc of material that surrounds the star. This disc is known as the circumstellar or protostellar disc.

We can now connect the appearance of the Solar System to our understanding of star formation. The circumstellar disc will be of similar composition to the GMC (i.e., it will consist of similar ratios of dust and gas). This dust and gas can be assembled into planets, or it can be fed onto the star, in much the same way that black holes are fed from accretion discs. The circumstellar disc therefore plays two critical roles in star and planet formation:

1. By virtue of its structure and dynamics, it regulates the growth of the star in its early phases.

2. Any material leftover from the star's formation becomes the feedstock for planet formation, constraining the properties of the resulting planetary system (although in practice star and planet formation may occur at the same time).

The properties of circumstellar discs - their dynamics, their various instabilities and turbulent behaviour, their thermodynamic and chemical properties - leave an indelible imprint on the objects that subsequently form. Such discs link the fields of star and planet formation, and are key agents in defining the properties of not only our Solar System, but that of extrasolar systems and stellar phenomena in general.

1.3 Motivation for this Thesis

It has become clear that *self-gravity* is important in protostellar discs. While this was already known for other disc structures such as galaxies (exhibiting gravitational instability through grand design spiral structures), the properties of self-gravitating protostellar discs have only begun to be explored more recently. Current star formation theory suggests that the initial protostellar disc mass will be comparable to the initial protostar mass. The disc's self-gravitating phase must therefore be studied, especially as a potential agent for redistributing mass and angular momentum in the disc, and allowing the protostar to accrete.

Thermodynamics plays a crucial role in the evolution of self-gravitating discs. Radiative physics dictates disc stability and spiral structure, which in turn constrains how angular momentum is transported through the disc. It also governs *disc fragmentation*, a process by which discs can condense locally to produce bound objects, thought to be a possible route by which giant planets and very low-mass stars can be created. Therefore, any attempt to model or simulate self-gravitating discs must consider radiative transfer effects, which until recently has not been fully incorporated into theoretical studies.

My approach to studying protostellar discs has been primarily numerical. I have developed a hybrid algorithm to model the effects of frequency averaged radiative transfer in Smoothed Particle Hydrodynamics (SPH). I discuss the construction of SPH and how the hybrid algorithm is incorporated in Chapter 3. I have since applied this algorithm to SPH studies of two problems in self-gravitating disc evolution.

Firstly, what is the nature of angular momentum transport in fully three dimensional, radiative self-gravitating discs? I answer this question in Chapter 4, focusing in particular on the "pseudo-viscous" approximations used by semi-analytic disc modellers. Secondly, I investigate the behaviour of self-gravitating discs under perturbation from a binary companion. This work (described in Chapter 5) is split into two parts - the first studying the possibility of fragmentation under such perturbations, and the second studying the observational effects of these encounters, and whether these events may be construed as "outburst phenomena".

While these studies are important from a theoretical standpoint, they are of limited utility to observers attempting to characterise observational features in discs. This problem is especially

topical with the advent of the next generation of astronomical instruments, which may have the potential to detect features such as spiral structure in self-gravitating discs. To be able to contribute to the scientific discussion regarding future observations, SPH simulations require a means by which they can be converted into synthetic telescope images. I have developed such an algorithm using Monte Carlo Radiative Transfer (MCRT) techniques. MCRT is typically implemented in grids and adaptive meshes - I have developed numerical techniques that allow MCRT to be implemented directly into SPH density fields. By doing this, I hope to provide connections between SPH simulators and observers, allowing SPH to make more informed predictions about the observable features of self-gravitating discs. I describe my algorithms in detail in Chapter 6.

The thesis is constructed in the following fashion: having briefly discussed the nebular hypothesis in this introduction, I present a more detailed summary of the current theory of star and planet formation in Chapter 2 (with a particular emphasis on the role that discs play in both processes). I outline the numerical framework of SPH and describe my radiative transfer algorithm in Chapter 3; in Chapter 4, I investigate angular momentum transport in isolated, self-gravitating discs; and in Chapter 5 I discuss the evolution of the physical and observational properties of self-gravitating discs under perturbations from stellar companions. Finally, in Chapter 6, I will discuss the principles of MCRT, and my method for allowing it to function in SPH without the use of grids. I include Appendices for derivations that, while relevant, detract from the flow of the main narrative. For the interested reader, Appendix D details the astrobiological research I have conducted during this studentship, which again does not directly contribute to the narrative of the thesis. I also include a list of articles published during this studentship (and a glossary) for reference.

Chapter 2

A Chronology of Star and Planet Formation

The greatest happiness for the thinking man is to have fathomed the fathomable, and to quietly revere the unfathomable.
Johann Wolfgang von Goethe, *Sprüche in Prosa*

2.1 Author's Note

In this chapter, I will attempt to place the circumstellar disc in context, as an object which connects the processes of star and planet formation. The terminology regarding these discs is not rigorously defined, with contradictions appearing frequently across the literature. I will attempt to be self-consistent and implement the following nomenclature:

- I will use *protostellar disc* to describe the disc during the epoch of star formation, from the moment of the disc's birth until the disc has lost a significant fraction of its gas mass.
- The term *protoplanetary disc* will be used to describe the disc in its post-protostellar phase, during the epoch of planet formation (which may be contemporaneous with star formation).
- *Debris disc* will describe the disc in its gas-poor, post-planet formation phase.
- *Circumstellar disc* will be a catch-all term to describe the disc at any stage in its evolution.

While I will attempt to capture a flavour of the richness of the current theories of star and planet formation, to maintain brevity I will emphasise areas where circumstellar discs are most active. I also consign some derivations to the Appendices to preserve the pace of the text. Before we can begin the chronology, we must take stock of the fundamental physical processes at play[1].

2.2 Fundamental Physics I: Hydrodynamics

Gases form the majority of the feedstock for star formation, as well as an important fraction of the feedstock for planet formation. Therefore, the understanding of *fluid dynamics* is essential to any description of star and planet formation. We must be able to describe the evolution of a fluid, given the initial conditions that describe the fluid's velocity field $\mathbf{v}$, its pressure P and its density ρ. I will simply list the equations here - the interested reader can consult Appendix A for derivations. Firstly, conservation of mass leads to the equation of continuity:

$$\frac{\partial \rho}{\partial t} + \nabla.(\rho \mathbf{v}) = 0. \tag{2.1}$$

Newton's second law $\mathbf{F} = m\frac{d\mathbf{v}}{dt}$ when applied to inviscid fluids gives Euler's equation:

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}.\nabla)\mathbf{v} = -\frac{1}{\rho}\nabla P. \tag{2.2}$$

If the fluid is viscous (and incompressible, i.e, $\rho = const.$) then this becomes the Navier-Stokes equation:

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}.\nabla)\mathbf{v} = -\frac{1}{\rho}\nabla P + \nu\nabla^2\mathbf{v}. \tag{2.3}$$

where ν is the kinematic viscosity. The internal energy equation is derived from the First Law of Thermodynamics:

$$\frac{du}{dt} = T\frac{ds}{dt} + \frac{P}{\rho^2}\frac{d\rho}{dt} = T\frac{ds}{dt} - \frac{P}{\rho}(\nabla.\mathbf{v}). \tag{2.4}$$

The sound speed of the medium c_s is given by

$$c_s^2 = \frac{dP}{d\rho}. \tag{2.5}$$

For an adiabatic ideal gas,

$$c_s^2 = \sqrt{\frac{\gamma P}{\rho}}, \tag{2.6}$$

where

[1] As magnetic fields are not considered in this thesis, I leave their discussion for the most part in the Appendices

$$\gamma = \frac{C_P}{C_V} = \left(\frac{dQ}{dT}\right)_P \left(\frac{dQ}{dT}\right)_V^{-1}. \tag{2.7}$$

In this instance, the internal energy can be simply calculated:

$$u = \frac{c_s^2}{\gamma(\gamma - 1)}. \tag{2.8}$$

Turbulence Let us now compare the typical strengths of viscous and inertial forces in the fluid. Typical inertial forces will take the form:

$$|\mathbf{F_{inertial}}| = \rho \bar{v}^2 \ell^2, \tag{2.9}$$

which we can justify from dimensional analysis ($\bar{v}$ is the mean fluid speed, and ℓ is some characteristic length scale). The dimensions of ν are $[\text{length}]^2[\text{time}]^{-1}$, allowing us to construct the typical viscous force as

$$|\mathbf{F_{viscous}}| = \rho \nu \bar{v} \ell. \tag{2.10}$$

Taking the ratio of these two quantities gives the dimensionless parameter known as the *Reynolds number*, *Re*:

$$Re = \frac{|\mathbf{F_{inertial}}|}{|\mathbf{F_{viscous}}|} = \frac{\bar{v}\ell}{\nu}. \tag{2.11}$$

We can now characterise the fluid using the three variables that compose *Re*: the kinematic viscosity ν, the mean fluid velocity $\bar{v}$, and a typical length scale ℓ (for example, a fluid in a box has a characteristic length equal to the side of the box). Because *Re* is dimensionless, it can be shown that two flows which are *similar* (that is, the coordinates and velocities of each flow are the same relative to some scaling factor) will have the same Reynolds number (Reynolds, 1883).

The Reynolds number is important for stability in flows. Flows with a small Reynolds number will be steady (or *laminar*). Such flows will be stable against perturbation - any perturbations in the flow will decrease as a function of time. As the Reynolds number increases past some critical value, perturbations to the flow will no longer decrease with time, becoming unstable, and resulting in *turbulent* flow. We can understand therefore that fluids with low viscosity will be more susceptible to turbulence, as the ability of viscous forces to combat inertial forces and damp perturbations will be weaker in general.

The precise value of the critical Reynolds number is a function of the boundary conditions of the fluid and the problem under consideration. Canonical values are typically of the order $Re_{crit} \sim 10 - 100$ for standard laboratory flows such as the Couette flow between shearing parallel plates (Landau & Lifshitz, 1959).

2.3 Fundamental Physics II: Gravity

As one of the four fundamental forces of nature, influencing all massive objects in the Universe, gravity must be accounted for in our study of star and planet formation. After all, it is the action of gravity that collapses molecular clouds, and allows stars and planets to coalesce. Einstein's theory of general relativity (GR) is still the most widely accepted theory of gravitation to date, despite the remaining questions surrounding Solar System phenomena such as the Pioneer anomalies (Anderson et al., 1998), and extragalactic phenomena such as the ubiquitous dark matter originally inferred by Zwicky (1933). While GR is required to explain aspects of orbital dynamics on Solar System scales (such as the precession of Mercury's perihelion), we will find that Newtonian gravity will be sufficient for the purposes of this thesis. Therefore, we can adopt the Universal Law of Gravitation to find the force field per unit mass $\mathbf{g}(\mathbf{r})$:

$$\mathbf{g}(\mathbf{r}) = G \int \frac{\rho(\mathbf{r}')(\mathbf{r} - \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^3} \, d^3\mathbf{r}', \tag{2.12}$$

where $\rho(\mathbf{r})$ is the density field and G is Newton's gravitational constant. We can determine a field equation for the gravitational potential Φ, through the relation

$$\mathbf{g} = -\nabla\Phi, \tag{2.13}$$

Which leads to Poisson's Equation (see Appendix B for derivation):

$$\nabla^2\Phi(\mathbf{r}) = 4\pi G\rho(\mathbf{r}). \tag{2.14}$$

Having solved for the gravitational potential, we can then incorporate gravity into hydrodynamics through the Navier-Stokes equation:

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}.\nabla)\mathbf{v} = -\frac{1}{\rho}\nabla P + \nu\nabla^2\mathbf{v} - \nabla\Phi. \tag{2.15}$$

2.4 Fundamental Physics III: Radiative Transfer

As the title of this thesis would suggest, the physics of radiation transport will prove to be particularly important in the physics of protostellar discs. Radiative transfer is a complex, multi-dimensional process, which requires explanation at length. Therefore, to preserve balance in this thesis, I will introduce only the basic concepts here - I will leave certain topics to be discussed in Chapters 3 and 6 (where they become more relevant to the narrative). For this section, I will draw on the descriptions given by Chandrasekhar (1960), Shu (1991) and Castor (2004).

Specific Intensity Let us begin by defining the variables that determine an electromagnetic (EM) radiation field. The EM radiation field is composed of individual quanta known as *photons*, each with an energy $E_\gamma = h\nu$. Consider a *photon packet* or beam of radiant energy

dE_ν, composed of photons in a specified infinitesimal frequency interval $(\nu, \nu + d\nu)$. The packet moves through a surface element (with area dA and normal vector $\mathbf{dA}$), along a vector $\mathbf{n}$ with a non-zero beam size or solid angle $d\Omega$. The packet traverses the surface in time dt. If $\mathbf{dA}.\mathbf{n} = dA\cos\theta$, then we can define the *specific intensity* of the radiation field I_ν:

$$dE_\nu = I_\nu \cos\theta \, d\nu \, dA \, d\Omega \, dt. \tag{2.16}$$

The radiation field will generally interact with the medium through which it travels. The medium can absorb, scatter and emit photons, with all three processes modifying I_ν as a function of space, time and incident radiation. Therefore, in general

$$I_\nu = I_\nu(x, y, z, n_x, n_y, n_z, t). \tag{2.17}$$

The enormity of the problem now becomes clear - to fully understand the radiation field, we must be able to constrain it as a function of seven independent coordinates (three in space for position, two angle coordinates for direction, and as a function of time and frequency), incorporating the detailed interactions between photons and subatomic particles/atoms/molecules. I will discuss these interactions in more detail in Chapter 6. For the time being, it will be sufficient to discuss the radiation in terms of continuous beams.

Absorption (and Scattering) Firstly, let us discuss the weakening of the radiation field by the medium. If the specific intensity changes from I_ν to $I_\nu + dI_\nu$ after traversing a distance ds through the medium, then we define the mass absorption coefficient χ_ν thus:

$$dI_\nu = -\chi_\nu \rho I_\nu ds \tag{2.18}$$

We have not yet specified whether this coefficient applies to scattering or absorption. If we assume that the medium does not absorb at all, then our beam will have photons scattered away from it at a rate given by

$$dI_\nu = \chi_\nu \rho \, ds I_\nu \cos\theta \, d\nu \, dA \, d\Omega = \chi_\nu I_\nu \, dm \, d\nu \, d\Omega, \tag{2.19}$$

where $dm = \rho \cos\theta \, ds \, dA$ represents the intervening mass of the medium that the photon has traversed in distance ds. This assumes that the scattering is isotropic, which is not typically the case. To account for the angular distribution of scattered radiation, we introduce the *phase function* $P(\cos\Theta)$ to indicate the rate of energy loss by scattering into a specific direction at an angle of Θ to the incident radiation direction:

$$\chi_\nu I_\nu \, P(\cos\Theta) \frac{d\Omega'}{4\pi} \, dm \, d\nu \, d\Omega, \tag{2.20}$$

where $d\Omega'$ denotes the solid angle of the scattered radiation. To find the energy loss in all directions, the above expression must be integrated over $d\Omega'$. For the two above equations to

be consistent (in the absence of absorption), we require that the phase function be normalised thus:

$$\int P(\cos\Theta)\,\frac{d\Omega'}{4\pi} = 1. \tag{2.21}$$

We can reuse this apparatus in the general case where both scattering and absorption exist, modifying the above constraint such that

$$\int P(\cos\Theta)\,\frac{d\Omega'}{4\pi} = a \leq 1. \tag{2.22}$$

We can now identify a as the *albedo*, the fraction of light lost due to scattering. Conversely, $(1-a)$ is the fraction of light lost to absorption processes. If the albedo is equal to unity, then we can say the medium is a *perfect scatterer*, and the mass absorption coefficient is equal to the *scattering cross section* σ_ν. If the albedo is equal to zero, then the medium is a perfect absorber, and the mass absorption coefficient is equal to the *opacity* κ_ν.

Emission The medium can contribute photons to the beam as well as remove them. We may define an emission coefficient j_ν such that a mass element dm emits radiation into a solid angle $d\Omega$, in a frequency range $(\nu, \nu + d\nu)$ with a total radiant energy of

$$dE_{\nu,emit} = j_\nu\, dm\, d\nu\, d\Omega\, dt \tag{2.23}$$

in the time interval dt. Scattering of light into the domain considered will also contribute to the emission coefficient. From the results of the previous section, we should be able to quantify this contribution. If our beam proceeds along the direction given by (θ, ϕ), then scattered radiation from the direction (θ', ϕ') will contribute an amount given by

$$\chi_\nu dm\, d\nu\, d\Omega P(\cos\Theta) I_\nu(\theta', \phi')\frac{\sin\theta' d\theta' d\phi'}{4\pi}, \tag{2.24}$$

where Θ now denotes the angle between the directions given by (θ, ϕ) and (θ', ϕ'). To obtain the full contribution from all directions, we must integrate this quantity over (θ', ϕ'), in general a difficult task. The situation becomes significantly simpler if we assume local thermodynamic equilibrium (LTE) - if this is the case, then Kirchhoff's Law dictates a simple relation between the emission and absorption coefficients, defined by the local temperature T:

$$j_\nu = \chi_\nu\, B_\nu(T), \tag{2.25}$$

where

$$B_\nu(T) = \frac{2h\nu^3}{c^2}\frac{1}{\exp(\frac{h\nu}{kT}) - 1} \tag{2.26}$$

is the Planck function, whose form arises from the quantisation of the beam into photons with energy $E_\gamma = h\nu$.

The Equation of Radiative Transfer Now that the three principal radiative processes are characterised, we can construct the fundamental equation which evolves I_ν, for a medium with mass absorption coefficient χ_ν and emission coefficient j_ν. If we consider a cylinder with cross-section dA and length ds, then we can define the energy change in time dt as our beam of radiation moves through it as

$$\Delta E = \frac{dI_\nu}{ds}(ds\, d\nu\, dA\, d\Omega\, dt). \tag{2.27}$$

The energy change can either result from emission in the cylinder, or from absorption/scattering. Therefore

$$\Delta E = (j_\nu \rho - \chi_\nu \rho I_\nu)(ds\, d\nu\, dA\, d\Omega\, dt). \tag{2.28}$$

By comparing and contrasting the above two equations, we obtain

$$\frac{dI_\nu}{ds} = j_\nu \rho - \chi_\nu \rho I_\nu. \tag{2.29}$$

We can further define the *optical depth* τ_ν and the *source function* S_ν

$$\tau_\nu = \int_0^s \rho \chi_\nu \, ds', \tag{2.30}$$

$$d\tau_\nu = \rho \chi_\nu ds, \tag{2.31}$$

$$S_\nu = \frac{j_\nu}{\chi_\nu}, \tag{2.32}$$

to produce a neater version of the equation of radiative transfer:

$$\frac{dI_\nu}{d\tau_\nu} = S_\nu - I_\nu. \tag{2.33}$$

This equation is straightforward to solve:

$$I_\nu(\tau_\nu) = I_\nu(0)e^{-\tau_\nu} + \int_0^{\tau_\nu} S(\tau'_\nu)e^{-\tau'_\nu}\, d\tau'_\nu. \tag{2.34}$$

In the limit where $j_\nu = 0$, the integrand is also zero, and we see that I_ν will decay exponentially with increasing optical depth. τ is defined along the line of sight, and will in general be a non-trivial function of space, orientation and frequency in astrophysical fluids. Also, because the source function will in general depend on I_ν, we can see that the equation of radiative transfer is an *integro-differential* equation, i.e. the equation depends both on derivatives and integrals. The exception is LTE, where $S_\nu = B_\nu$.

Moments of the Radiation While I_ν can describe the radiation field completely, its multivariate nature is typically too unwieldy to use directly. Instead it is often used by proxy through the use of its (angle) moments. The first moment is the energy density of the radiation $\mathcal{U}_\nu$:

$$\mathcal{U}_\nu = \frac{1}{c}\int_0^{4\pi} I_\nu \, d\Omega. \tag{2.35}$$

The next is the vector flux $\mathbf{F}_\nu$, which we define along a direction vector $\mathbf{n}$:

$$\mathbf{F}_\nu = \int_0^{4\pi} \mathbf{n} I_\nu \, d\Omega. \tag{2.36}$$

Intuitively, it is clear that for a surface element with normal vector $d\mathbf{A}$, that $\mathbf{F}_\nu.d\mathbf{A}$ gives the radiative flux through that surface element, with the sign indicating its direction. Note that $\mathbf{F}_\nu$ has no c term. The last moment we will consider is the radiative pressure tensor $\mathcal{P}_\nu$:

$$\mathcal{P}_\nu = \frac{1}{c}\int_0^{4\pi} \mathbf{n}\mathbf{n} I_\nu \, d\Omega, \tag{2.37}$$

which allows us to characterise the important hydrodynamic effects of radiation. Other quantities are often used in lieu of these moments to deal with the factors of 4π and c:

$$J_\nu = \frac{c}{4\pi}\mathcal{U}_\nu, \tag{2.38}$$

$$\mathbf{H}_\nu = \frac{1}{4\pi}\mathbf{F}_\nu, \tag{2.39}$$

$$\mathcal{K}_\nu = \frac{c}{4\pi}\mathcal{P}_\nu, \tag{2.40}$$

We will use the former set of moments rather than the latter.

Radiative Hydrodynamics To model a fully radiative fluid, we now require equations that govern the moments of the radiation, as well as modifying the existing equations of hydrodynamics. It is useful to begin by defining a continuity equation for I_ν (cf Castor 2004). In a perfect vacuum, the optical depth $\tau = 0$ and the equation of radiative transfer guarantees our beam of intensity I_ν will remain constant as it moves along its (unit) direction vector $\mathbf{n}$, in some time interval Δt:

$$I_\nu(\mathbf{r} + (c\Delta t)\mathbf{n}, t + \Delta t) = I_\nu(\mathbf{r}, t). \tag{2.41}$$

Taylor expanding the left hand side around $(\mathbf{r}, t)$ gives

$$I_\nu(\mathbf{r} + (c\Delta t)\mathbf{n}, t + \Delta t) = I_\nu(\mathbf{r}, t) + \left(\frac{\partial I_\nu}{\partial t}\Delta t + (c\Delta t)\mathbf{n}.\nabla I_\nu\right) + O(\Delta t^2). \tag{2.42}$$

Equating the right hand sides of the above two equations (and dividing out by $c\Delta t$) gives

$$\frac{1}{c}\frac{\partial I_\nu}{\partial t} + \mathbf{n}.\nabla I_\nu = 0. \tag{2.43}$$

This can be easily modified to incorporate the presence of radiation sources and sinks (giving a more general version of the radiative transfer equation):

$$\frac{1}{c}\frac{\partial I_\nu}{\partial t} + \mathbf{n}.\nabla I_\nu = \rho j_\nu - \rho\chi_\nu I_\nu, \tag{2.44}$$

where we could recover our previous 1D radiative transfer equation by setting $\frac{\partial I_\nu}{\partial t} = 0$ and aligning the axis of integration with $\mathbf{n}$. By averaging over solid angle, we can also recover continuity equations for the moments of I_ν (assuming the radiation field to be isotropic). We can now construct a complete set of equations for radiative hydrodynamics (RHD). I will now list this set of equations (averaged over frequency), where we must modify the momentum and energy equations for the matter to incorporate radiative gains and losses:

$$\frac{\partial \mathcal{U}}{\partial t} + \nabla.\mathbf{F} + \nabla\mathbf{v} : \mathcal{P} = 4\pi\rho j - \kappa_E c\rho\mathcal{U}, \tag{2.45}$$

$$\frac{1}{c}\frac{\partial \mathbf{F}}{\partial t} + c\nabla.\mathcal{P} = -\chi\rho F, \tag{2.46}$$

$$\frac{\partial \rho}{\partial t} + \nabla.(\rho\mathbf{v}) = 0, \tag{2.47}$$

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}.\nabla)\mathbf{v} = -\frac{1}{\rho}\left(\nabla P + \frac{\chi\rho\mathbf{F}}{c}\right) - \nabla\Phi, \tag{2.48}$$

$$\frac{du}{dt} = T\frac{ds}{dt} - \frac{P}{\rho}(\nabla.\mathbf{v}) - 4\pi\rho j + \kappa_E c\rho\mathcal{U}, \tag{2.49}$$

where we have added a tensor product term to the radiative energy density equation (which is due to the fluid's interaction with the radiation field, allowing it to advect radiation away along its velocity vector). We denote the energy mean absorptive opacity by κ_E, and the flux mean absorptive opacity by χ.

Flux-limited Diffusion If we assume the photon mean free path is small compared with the typical length scales we are interested in, then we can simplify the equations of RHD immensely. In this approximation the medium is *optically thick* ($\chi \gg 1$), and the photons can be modelled as *diffusing* through the medium through a series of collisions. Let us rearrange equation (2.44) for I_ν:

$$I_\nu = \frac{j_\nu}{\chi_\nu} - \frac{1}{\chi_\nu}\left(\frac{1}{c}\frac{\partial I_\nu}{\partial t} + \mathbf{n}.\nabla I_\nu\right). \tag{2.50}$$

As we are in the limit of large χ, we can consider the second term to be a small correction to the first - hence to first approximation

$$I_\nu \approx \frac{j_\nu}{\chi_\nu} = S_\nu, \tag{2.51}$$

i.e. the radiation is in equilibrium with the medium. If we can assign a temperature T to the medium (which will be the case in this thesis), then we know that the source function is simply the Planck function $B_\nu(T)$, which allows us to substitute for the emissivity j in our RHD equations. We can further simplify (if the radiation field is isotropic) through the Eddington approximation

$$\mathcal{P} = \frac{1}{3}\mathcal{U}\mathbf{I}. \tag{2.52}$$

where $\mathbf{I}$ is the identity matrix. In the steady state, $\frac{\partial \mathbf{F}}{\partial t} = 0$ and we can now obtain from equation (2.46):

$$\mathbf{F} = -\frac{c}{3\chi\rho}\nabla\mathcal{U} \tag{2.53}$$

The radiation flux is now represented by a diffusion equation, with the diffusion constant D given by the prefactor to the gradient term. This approximation fails if the medium is optically thin (i.e. the photon mean free paths become too large). However, the utility of this solution has inspired attempts to increase the range of optical depths over which this approximation remains valid, by the use of *flux-limited diffusion* (FLD), where D is replaced by (Levermore & Pomraning, 1981; Bodenheimer et al., 1990):

$$D' = \frac{c\lambda}{\chi\rho}, \tag{2.54}$$

Where λ is the dimensionless flux limiter. This modifies the approximation for $\mathcal{P}$ to

$$\mathcal{P} = \frac{1}{2}\mathcal{U}\left((1-f)\mathbf{I} + (3f-1)\left(\frac{\nabla\mathcal{U}}{|\nabla\mathcal{U}|}\right)^2\right), \tag{2.55}$$

with the Eddington factor f

$$f = \lambda + \lambda^2 R^2, \tag{2.56}$$

and

$$R = \frac{|\nabla\mathcal{U}|}{\chi\rho\mathcal{U}}. \tag{2.57}$$

All that is left to do is choose a suitable flux limiter, (e.g. Levermore & Pomraning 1981; Bodenheimer et al. 1990):

$$\lambda(R) = \frac{2+R}{6+3R+R^2}. \tag{2.58}$$

Therefore, the equations of RHD in the flux-limited diffusion approximation are:

$$\frac{\partial \mathcal{U}_\nu}{\partial t} + \nabla.\mathbf{F} + \nabla\mathbf{v} : \mathcal{P} = 4\pi\rho\kappa_P B - \kappa_E c\rho\mathcal{U}, \tag{2.59}$$

$$\frac{1}{c}\frac{\partial \mathbf{F}}{\partial t} + c\nabla.\mathcal{P} = -\chi\rho F, \tag{2.60}$$

$$\frac{\partial \rho}{\partial t} + \nabla.(\rho \mathbf{v}) = 0, \tag{2.61}$$

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}.\nabla)\mathbf{v} = -\frac{1}{\rho}\left(\nabla P + \frac{\chi \rho \mathbf{F}}{c}\right) - \nabla \Phi, \tag{2.62}$$

$$\frac{du}{dt} = T\frac{ds}{dt} - \frac{P}{\rho}(\nabla.\mathbf{v}) - 4\pi\rho\kappa_P B + \kappa_E c \rho \mathcal{U}, \tag{2.63}$$

Where the Planck opacity κ_P is defined

$$\kappa_P(T) = \frac{\int \kappa_\nu B_\nu(T) d\nu}{\int B_\nu(T) d\nu} \tag{2.64}$$

2.5 The Giant Molecular Clouds

Armed with the fundamental physics required, we can now investigate the beginning of all star systems - the giant molecular clouds (GMCs). These are turbulent clouds of molecular gas, with typical temperatures of order 10 K. In the Solar neighbourhood, GMCs have masses consistent with a distribution given by (Williams & McKee, 1997; McKee & Ostriker, 2007):

$$\frac{dN_c}{d\ln M_c} = \mathcal{N}\left(\frac{\mathcal{M}}{M}\right)^\alpha \quad (M < \mathcal{M}), \tag{2.65}$$

with $\mathcal{N} = 63$ clouds, $\alpha = 0.6$ and the cutoff mass $\mathcal{M} = 6 \times 10^6 M_\odot$. The expression $\mathcal{N}/\alpha$ is equivalent to the number of clouds that are eliminated from the distribution by cutting it off at $M = \mathcal{M}$. Most of the mass is in molecular gas of various species - the total mass of a cloud is typically inferred from observing the $J = 1 - 0$ transition in either ^{12}CO or ^{13}CO. Larson (1981) describes three laws that give a useful general description of GMCs:

1. They display supersonic turbulence with velocity dispersions σ that scale with the cloud size R, for example Solomon et al. (1987)'s result:

$$\sigma = (0.72 \pm 0.07)\left(\frac{R}{1\,\mathrm{pc}}\right)^{0.5\pm0.05} \mathrm{km\,s^{-1}}. \tag{2.66}$$

2. They are gravitationally bound.

3. They share similar column densities (in hydrogen).

The turbulence in GMCs results in hierarchical structures that exists at a range of scales, from the size of the cloud down to scales where gravity matches local pressure support. Stars can therefore be formed in large groups (*clusters*) as a result of the collapse of large star forming clumps, or in much smaller numbers as the result of collapsing prestellar cores (Ward-Thompson et al., 2007). The similarities between the apparently universal core mass function (CMF) and the initial stellar mass function (IMF) (e.g. Smith et al. 2009) suggest that the star formation

process is strongly governed by the initial mass distribution in the cloud (rather than the alternative scenario where stellar accretion is halted by some internal or external process). With this in mind, we should derive the requirements for a gaseous region to collapse under its own gravity, through the Jeans instability.

2.5.1 Conditions for Cloud Collapse: The Jeans Instability

Let us begin with the fundamental equations of hydrodynamics in the presence of gravity- the continuity equation:

$$\frac{\partial \rho}{\partial t} + \nabla.(\rho \mathbf{v}) = 0, \tag{2.67}$$

Euler's equation

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}.\nabla)\mathbf{v} = -\frac{1}{\rho}\nabla P - \nabla \mathbf{\Phi}, \tag{2.68}$$

and Poisson's Equation:

$$\nabla^2 \mathbf{\Phi} = 4\pi G\rho, \tag{2.69}$$

where $\rho(\mathbf{r}, t)$ is the density, $P(\mathbf{r}, t)$ is the pressure, $\mathbf{v}(\mathbf{r}, t)$ is the fluid velocity and $\Phi(\mathbf{r}, t)$ is the gravitational potential field. We will also assume a simple equation of state to link ρ and p:

$$P = c_s^2 \rho, \tag{2.70}$$

where c_s is the sound speed. Now let us consider a fluid in equilibrium (indicated by variables with a "0" subscript), which is then given a small perturbation:

$$\rho(\mathbf{r}, t) = \rho_0(\mathbf{r}, t) + \epsilon\rho_1(\mathbf{r}, t) \ ; \ P(\mathbf{r}, t) = P_0(\mathbf{r}, t) + \epsilon P_1(\mathbf{r}, t), \tag{2.71}$$

$$\mathbf{v}(\mathbf{r}, t) = \mathbf{v}_0(\mathbf{r}, t) + \epsilon\mathbf{v}_1(\mathbf{r}, t) \ ; \ \mathbf{\Phi}(\mathbf{r}, t) = \mathbf{\Phi}_0(\mathbf{r}, t) + \epsilon\mathbf{\Phi}_1(\mathbf{r}, t), \tag{2.72}$$

where $\epsilon \ll 1$. We can quickly see that when substituting back into the equations above, the terms not involving ϵ will cancel. If we also ignore terms of $O(\epsilon^2)$, this then gives

$$\frac{\partial \rho_1}{\partial t} + \nabla.(\rho_0 \mathbf{v_1}) + \nabla.(\rho_1 \mathbf{v_0}) = 0, \tag{2.73}$$

$$\frac{\partial \mathbf{v}_1}{\partial t} + (\mathbf{v_0}.\nabla)\mathbf{v_1} + (\mathbf{v_1}.\nabla)\mathbf{v_0} = -\frac{\rho_1}{\rho_0^2}\nabla P_0 - \frac{1}{\rho_0}\nabla P_1 - \nabla \mathbf{\Phi_1}, \tag{2.74}$$

$$\nabla^2(\mathbf{\Phi}_0) = 4\pi G\rho_0, \tag{2.75}$$

$$\nabla^2(\mathbf{\Phi}_1) = 4\pi G\rho_1. \tag{2.76}$$

To obtain this result we have used the specific enthalpy of the gas (for zero entropy change $ds = 0$):

$$h(\mathbf{r}, t) = \int \frac{dP}{\rho}, \tag{2.77}$$

which gives the perturbed value as

$$h_1 = c_s^2 \frac{\rho_1}{\rho_0}. \tag{2.78}$$

Consider a homogeneous equilibrium state where $\rho_0 = const.$, $\mathbf{v}_0 = 0$ and $\Phi_0 = 0$. We get away with the last constraint by assuming that the medium is infinite, and that the Poisson equation only governs perturbations in the medium (Jeans, 1928) - this is sometimes referred to as the *Jeans swindle*. This provides useful simplification:

$$\frac{\partial \rho_1}{\partial t} + \rho_0 \nabla.(\mathbf{v_1}) = 0, \tag{2.79}$$

$$\frac{\partial \mathbf{v}_1}{\partial t} = -\frac{1}{\rho_0} \nabla P_1 - \nabla \mathbf{\Phi_1}. \tag{2.80}$$

This can be combined into one equation by taking the time derivative of equation (2.79) and the divergence of (2.80) and substituting for $\frac{\partial}{\partial t}(\nabla.\mathbf{v})$:

$$\frac{\partial^2 \rho_1}{\partial t^2} - c_s^2 \nabla^2 \rho_1 - 4\pi G \rho_1 \rho_0 = 0, \tag{2.81}$$

where we have also used the perturbed version of Poisson's equation. A possible solution is

$$\rho_1(\mathbf{r}, t) = \rho_0 e^{i(\mathbf{k}.\mathbf{r} - \omega t)}. \tag{2.82}$$

Substituting this gives a dispersion relation for the perturbation, constraining the possible values of ω and the wavenumber $\mathbf{k}$:

$$\omega^2 = c_s^2 k^2 - 4\pi G \rho_0. \tag{2.83}$$

For large wavenumbers (or low ρ_0), the second term can be neglected and the dispersion relation reduces to that of a sound wave. If the wavenumber is sufficiently small, then $\omega^2 < 0$, and ω is imaginary. The wave is no longer stable: $\rho_1(\mathbf{r}, t) \propto e^{\pm|\omega|t}$, and the wave can either grow or decay exponentially with time. $\omega = 0$ therefore presents a condition for stability - alternatively,

$$k^2 < \frac{4\pi G \rho_0}{c_s^2}. \tag{2.84}$$

By converting the wavenumber into a wavelength (using $\lambda = 2\pi/|k|$), we can derive the *Jeans length*. If the perturbation's physical extent λ satisfies

$$\lambda^2 > \lambda_J^2 = \frac{\pi c_s^2}{G \rho_0}, \tag{2.85}$$

then we can expect the perturbation to begin collapsing under its own gravity. We can convert this into a *Jeans Mass* assuming a sphere of uniform density ρ_0 and radius λ_J:

$$M_J = \frac{4}{3}\pi\lambda_J^3\rho_0. \tag{2.86}$$

2.5.2 The Collapse of Rotating Clouds to form Discs

Once the Jeans criterion is satisfied, a section of the cloud with a characteristic radius equal to λ_J will begin to collapse. The collapse will occur on roughly the *free-fall timescale* t_{ff} (assuming a sphere with uniform density ρ, see Appendix B for derivation):

$$t_{ff} = \sqrt{\frac{3\pi}{32G\rho}}. \tag{2.87}$$

The gas will be initially optically thin and hence isothermal, remaining at temperatures of order 10 K. This lasts until the central density increases to the point that the cloud can become optically thick, allowing the temperature to rise to the point where the contraction can be decelerated, forming the initial protostellar core. Once the gas heats sufficiently to dissociate the molecular hydrogen, the collapse restarts to produce a second, denser core (Masunaga & Inutsuka, 2000; Bate, 2010; Machida & Matsumoto, 2010). We carry out a simulation of a non-rotating collapse in Chapter 3 which illustrates the details of this process.

Of course, in reality the cloud collapse occurs within a wider context of turbulence, magnetic fields, radiative transfer, angular momentum and long range tidal forces from the rest of the GMC. In particular, the addition of angular momentum has an indelible influence on the products of the collapse process. While magnetic fields can act to remove angular momentum, this is only effective when neutral and ionised gas are well coupled. Once decoupling occurs, the cloud will collapse with a near constant angular momentum. If we define the cloud's axis of rotation along the z-axis, it is straightforward to consider the properties of the rotating core. The angular momentum will provide centrifugal support perpendicular to the axis of rotation (i.e. in the x-y plane). By doing this, the *axial ratio* of the core is changed - the core's extent along the x and y-axis will be greater than its extent along the z-axis. The core becomes increasingly oblate for increased initial angular momentum. If the angular momentum of the cloud is sufficient, the cloud collapse will not form an isolated protostellar core, but will instead form a smaller protostellar core with an extended distribution of matter around it (e.g. Machida & Matsumoto 2010). This extended distribution will eventually form the protostellar disc.

We can quantify the influence of angular momentum by considering a simple approximation (Lin & Pringle, 1990). Let us begin with an isolated spherical cloud, with mass M_c, initial radius R_c equal to the local Jeans length, which is in rigid body rotation (i.e., the angular velocity is constant with radius), with angular velocity Ω_c. If we assume that none of the cloud's mass is lost during the collapse, then we can equate the pre-collapse and post-collapse angular momentum as:

$$M_c R_c^2 \Omega_c = M_c r_{disc}^2 \Omega_{disc}, \tag{2.88}$$

where r_{disc} is the maximum radius of the disc, and Ω_{disc} is the angular velocity of the disc. It is reasonable to assume that the disc is Keplerian (see section 2.6.4), giving

$$\Omega_{disc} = \sqrt{\frac{GM_c}{r_{disc}^3}}. \tag{2.89}$$

Substituting this expression into equation (2.88) allows us to rearrange for r_{disc}:

$$r_{disc} = \frac{\Omega_c^2 R_c^4}{GM_c}. \tag{2.90}$$

It is common practice to define the cloud's initial conditions in terms of two dimensionless variables:

$$\alpha_c = \frac{E_{therm}}{E_{grav}} \ , \ \beta_c = \frac{E_{rot}}{E_{grav}}, \tag{2.91}$$

where E_{therm}, E_{rot} and E_{grav} are the cloud's initial thermal, rotational and gravitational potential energy respectively. For our rigidly rotating sphere, we can define the rotational energy by calculating the moment of inertia I:

$$E_{rot} = \frac{1}{2} I \Omega_c^2 = \frac{1}{5} M_c R_c^2 \Omega_c^2 = \beta_c E_{grav} \tag{2.92}$$

The gravitational potential energy:

$$E_{grav} = \frac{3}{5} \frac{GM_c^2}{R_c}. \tag{2.93}$$

We can now express Ω_c in terms of β_c:

$$\Omega_c^2 = \frac{GM_c \beta_c}{R_c^3}. \tag{2.94}$$

This gives a much more intuitive expression for r_{disc}:

$$r_{disc} = \frac{3GM_c \beta_c R_c^4}{GM_c R_c^3} = 3\beta_c R_c. \tag{2.95}$$

Typical values for β_c vary from 2×10^{-4} to 1.4, with a median value of 0.03 (Goodman et al., 1993). Under this approximation, we can expect initial disc radii of order 10% of the cloud radius. Of course, this is a very simple approximation - other physical processes can provide support against gravitational collapse, as we have already discussed, and Keplerian rotation may not apply if the core is not sufficiently centrally condensed. This can result in β_c becoming a function of radius and other local properties, and resulting in an infall pattern that depends on initial radius R_c. The infall pattern can be considered as a series of concentric shells falling onto different disc radii, building a density profile which depends on $\beta_c(R_c)$. Semi-analytic methods (Lin & Pringle, 1990; Rice et al., 2010) or numerical simulations (Bate, 2010; Machida

& Matsumoto, 2010) are required to fully describe the formation of the star-disc system in this case.

2.6 The Structure of Protostellar Discs

I will now set out the theoretical concepts used to model protostellar discs. Much of this framework is taken from the more general *accretion disc* theory, which has been applied to discs of all size scales. While the theory is predominantly scale-invariant, the application of certain physical processes (especially radiative transfer) will limit the resulting equations to specific length and mass scales.

2.6.1 The Thin Disc Approximation

Fundamentally, protostellar discs are expected to be "thin" in a loose sense of the word, i.e. their vertical extent will be less than their radial extent. More rigorously, we can define a vertical length scale (the disc scale height H), and compare it to the disc radius (in cylindrical polar coordinates) r. If the *aspect ratio*

$$\frac{H}{r} \ll 1, \tag{2.96}$$

then the disc can be said to be *geometrically thin*. Protostellar discs typically have $H/r \sim 0.1$, so satisfy this criterion reasonably well. However, the aspect ratio is generally a function of radius in realistic protostellar discs, so care should be taken in using the thin disc approximation.

2.6.2 Conservation of Mass and Angular Momentum

We can construct a continuity equation for the disc by considering an annulus at radius r, with extent Δr (cf Pringle 1981). The annulus has surface density $\Sigma(r,t)$ (calculated by vertically integrating the density $\rho(r,t)$), and vertically averaged radial velocity $v_r(r,t)$. The mass inside the annulus is

$$M_{ann} = 2\pi r \Delta r \Sigma \tag{2.97}$$

We can equate the rate of change of mass in the annulus to the net flow of mass out of the annulus, i.e.

$$2\pi r \Delta r \frac{\partial \Sigma}{\partial t} = -\left[2\pi(r+\Delta r)\Sigma(r+\Delta r)v_r(r+\Delta r) - 2\pi r \Sigma(r) v_r(r)\right] \tag{2.98}$$

For small Δr, we can Taylor expand $\Sigma v_r r$ and rearrange to obtain

$$\frac{\partial \Sigma}{\partial t} + \frac{1}{r}\frac{\partial}{\partial r}(r\Sigma v_r) = 0. \tag{2.99}$$

We could also have obtained this equation by vertically averaging the standard continuity equation assuming axisymmetry. We can repeat the process for the angular momentum in the annulus $J_{ann} = \Sigma r v_\phi$:

$$\frac{\partial}{\partial t}(\Sigma r v_\phi) + \frac{1}{r}\frac{\partial}{\partial r}(r^2 \Sigma v_r v_\phi) = \frac{\mathcal{G}}{r}, \tag{2.100}$$

Where we define $\mathcal{G}$ as the net effect of torques from neighbouring annuli. To determine this torque, we can consider the Navier-Stokes Equation for a viscous gravitating fluid, again vertically integrated:

$$\Sigma\left(\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}.\nabla)\mathbf{v}\right) = -(\nabla P - \nabla.T) - \rho\nabla\Phi, \tag{2.101}$$

Where T_{ij} is the vertically integrated viscous stress tensor. If we assume only shearing viscosity applies, then the only surviving term in the tensor is

$$T_{r\phi} = \nu\Sigma r\frac{d\Omega}{dr} = \nu\Sigma r\Omega', \tag{2.102}$$

where ν is the vertically averaged kinematic viscosity. We can construct this from the dynamic viscosity η:

$$\Sigma\nu = \int_{-\infty}^{\infty} \eta\, dz. \tag{2.103}$$

$T_{r\phi}$ details the viscous force per unit length around the circumference. The net effect of viscous torques $\mathcal{G}$ is therefore

$$\mathcal{G}\Delta r = \frac{1}{2\pi} 2\pi \left[(r+\Delta r)^2\, T_{r\phi}|_{r+\Delta r} - r^2\, T_{r\phi}|_r\right] = \frac{\partial}{\partial r}(r^2 T_{r\phi}) \tag{2.104}$$

And the conservation of angular momentum gives

$$\frac{\partial}{\partial t}(\Sigma r v_\phi) + \frac{1}{r}\frac{\partial}{\partial r}(\Sigma r v_\phi r v_r) = \frac{1}{r}\frac{\partial}{\partial r}\left(\nu\Sigma r^3\Omega'\right). \tag{2.105}$$

This clearly shows the action of viscous forces providing torques to extract angular momentum from the disc. To evaluate Φ, we use Poisson's equation. For an infinitesimally thin disc, it takes the form

$$\nabla^2\Phi = 4\pi G\Sigma\delta(z), \tag{2.106}$$

Where $\delta(z)$ is the Dirac δ-function.

2.6.3 Vertical Structure

Consider the vertical component of the Navier-Stokes equation. If the thin disc approximation holds, then v_z should be small enough to neglect the left hand side. Viscous forces should also

disappear as they only act in the (r, ϕ) direction. This only leaves the pressure gradient and the gravitational forces to balance:

$$\frac{\partial P}{\partial z} = -\rho \frac{\partial \Phi}{\partial z}. \tag{2.107}$$

In other words, we expect the disc to be in *hydrostatic equilibrium*. If the disc is non-self gravitating, the vertical component of acceleration is

$$a_z = \frac{GM}{r_{sphere}^2} \frac{z}{r_{sphere}} = \frac{GMz}{r^3}, \tag{2.108}$$

where r_{sphere} is the radial separation in spherical polar co-ordinates, and hence the approximation $r = r_{sphere}$ is good for small z. If the gas is barotropic, that is

$$P = P(\rho);\ c_s^2 = \frac{dP}{d\rho}, \tag{2.109}$$

then we can rewrite the condition for hydrostatic equilibrium as the following differential equation

$$\frac{\partial \rho}{\partial z} = -\frac{GMz}{c_s^2 r^3} \rho. \tag{2.110}$$

This has a simple exponential solution:

$$\rho = \rho_0 \exp\left(\frac{-GMz^2}{2r^3 c_s^2}\right) \tag{2.111}$$

We can now define the characteristic vertical length scale in the non self-gravitating case: the disc scale height H:

$$H = \frac{c_s r^{3/2}}{(GM)^{1/2}} = \frac{c_s}{\Omega_K}, \tag{2.112}$$

where we have substituted for the Keplerian angular velocity Ω_K (see section 2.6.4). This gives the expression

$$\rho = \rho_0 \exp\left(\frac{-z^2}{2H^2}\right). \tag{2.113}$$

Let us now consider the self-gravitating case:

$$\frac{c_s^2}{\rho} \frac{d\rho}{dz} = -\frac{d\Phi}{dz}, \tag{2.114}$$

By taking the spatial derivative, we can substitute for $\nabla^2 \Phi$ using equation (2.106). Assuming the sound speed to be independent of z allows us to derive a result for the isothermal slab:

$$c_s^2 \frac{d}{dz}\left(\frac{1}{\rho}\frac{d\rho}{dz}\right) = 4\pi G \Sigma \delta(z). \tag{2.115}$$

The solution for this equation can be found by normalising ρ and z (Spitzer, 1942):

$$z = \left(\frac{c_s^2}{4\pi G\rho_0}\right)\xi, \tag{2.116}$$

$$\rho = \rho_0\Lambda(\xi). \tag{2.117}$$

This gives the hydrostatic equilibrium condition as

$$\frac{d^2\Lambda}{d\xi^2} - \frac{1}{\Lambda}\left(\frac{d\Lambda}{d\xi}\right)^2 - \Lambda = 0. \tag{2.118}$$

A solution to this equation is

$$\Lambda = \frac{1}{\cosh^2(\xi/2)} \rightarrow \rho = \frac{\rho_0}{\cosh^2(z/H_{sg})}, \tag{2.119}$$

where we have defined the self-gravitating disc thickness as

$$H_{sg} = \frac{c_s^2}{\pi G\Sigma}. \tag{2.120}$$

2.6.4 Radial Structure

Let us return to the Navier-Stokes equation (without averaging), and consider its radial component (substituting the correct expression for the gradient of a vector in cylindrical polar coordinates):

$$\frac{\partial v_r}{\partial t} + v_r\frac{\partial v_r}{\partial r} + \frac{v_\phi}{r}\left(\frac{\partial v_r}{\partial \phi} - v_\phi\right) + v_z\frac{\partial v_r}{\partial z} = -\frac{1}{\rho}\frac{\partial P}{\partial r} - \frac{\partial \Phi}{\partial r}. \tag{2.121}$$

Note that we can discount the viscous stress tensor as shear viscosity should have no radial component. If we rearrange the thin disc criterion, we can ascertain (assuming the disc is rotationally stable and the radial velocity is small):

$$v_r \ll c_s \ll v_\phi, \tag{2.122}$$

which allows us to discount the first two terms on the left hand side, as well as the last term. We can rearrange the first term on the right hand side to obtain

$$-\frac{1}{\rho}\frac{\partial P}{\partial r} = -\frac{c_s^2}{\rho}\frac{\partial \rho}{\partial r} \tag{2.123}$$

Which we can also discount in the thin disc case. Finally, if the disc is axisymmetric, $\frac{\partial v_r}{\partial \phi} = 0$, and we are left with

$$\frac{v_\phi^2}{r} = \frac{\partial \Phi}{\partial r} \tag{2.124}$$

This is intuitively sensible, as it indicates that the disc is in centrifugal balance. If the disc is not self-gravitating, the potential is essentially determined by the central star

$$\Phi = -\frac{GM_*}{r} \rightarrow \frac{\partial \Phi}{\partial r} = \frac{GM_*}{r^2} \tag{2.125}$$

and we obtain the standard Keplerian velocity

$$v_\phi = v_K = \sqrt{\frac{GM_*}{r}} \rightarrow \Omega_K = \frac{v_K}{r} = \sqrt{\frac{GM_*}{r^3}} \tag{2.126}$$

2.6.5 Viscous Evolution

Armed with equations (2.99), (2.105) and (2.106), we can compute the evolution of Σ. We can solve for v_r:

$$v_r = \frac{1}{r\Sigma(r^2\Omega)'}\frac{\partial}{\partial r}(\nu\Sigma r^3\Omega'), \tag{2.127}$$

which gives the continuity equation to be:

$$\frac{\partial \Sigma}{\partial t} = -\frac{1}{r}\frac{\partial}{\partial r}\left(\frac{1}{(r^2\Omega)'}\frac{\partial}{\partial r}(\nu\Sigma r^3\Omega')\right). \tag{2.128}$$

As we will see, ν is in general a function of radius and Σ in realistic accretion discs. Indeed, we expect that Ω will become sensitive to Σ in the self gravitating case. This makes equation (2.128) particularly non-linear, and solvable in general only by numerical methods.

Arguably the most important component in this equation is the viscosity. We can understand this by taking the simplified case where Ω and ν are independent of Σ, and the disc is Keplerian:

$$\frac{\partial \Sigma}{\partial t} = \frac{3}{r}\frac{\partial}{\partial r}\left(r^{1/2}\frac{\partial}{\partial r}(\nu\Sigma r^{1/2})\right). \tag{2.129}$$

As I will show in Chapter 4, the fundamental stumbling block in protostellar accretion disc theory is the determination of the viscosity.

2.6.6 Temperature Structure and Observational Properties

Finally, we should discuss a critical component of accretion disc theory - its predictions for observation. If we assume the disc achieves a steady state, then equation (2.99) can be rewritten in terms of the accretion rate through the disc $\dot{M}(r)$, i.e. the rate of change of mass in an annulus at radius r:

$$\dot{M} = -2\pi r v_r \Sigma. \tag{2.130}$$

The conservation of angular momentum gives

$$\dot{M}\Omega r^2 - 2\pi\nu\Sigma\left|\frac{d\Omega}{dr}\right| r^3 = \dot{J}, \tag{2.131}$$

where $\dot{J}$ is the net flux of angular momentum. We can see again that this net flux consists of advected angular momentum (the first term on the left hand side) and the outward flux due

to viscosity (the second term). If we demand that the disc is Keplerian, and at the disc's inner edge r_{in} the angular velocity profile flattens (as it comes into corotation with the central star), then the second term disappears and we obtain the boundary condition (Pringle, 1981; Lodato, 2007):

$$\dot{J} = \dot{M}(\Omega r^2)_{in} = \dot{M}\sqrt{GMr_{in}}. \tag{2.132}$$

As our disc is in a steady state, $\dot{J}$ is constant across all radii, so we can substitute and rearrange to obtain:

$$2\left|\frac{d\ln\Omega}{d\ln r}\right|\pi\nu\Sigma = \dot{M}\left(1-\sqrt{\frac{r_{in}}{r}}\right). \tag{2.133}$$

At large enough distance from r_{in}, the square root is negligible and we obtain a simple expression for $\dot{M}$ in terms of ν and Σ. If we wish to determine the power per unit surface dissipated by viscous stresses $D(r)$, we can now approximate for large radii using the accretion rate:

$$D(r) = \nu\Sigma\left(r\frac{d\Omega}{dr}\right)^2 \approx \left|\frac{d\ln\Omega}{d\ln r}\right|\frac{\dot{M}}{2\pi}\Omega^2. \tag{2.134}$$

We can construct a spectrum if we assume that the disc is optically thick, and therefore this dissipation occurs as blackbody radiation at the disc photosphere. If this is true, then the flux is equal to $2\sigma_{SB}T_{eff}^4$ (as the disc can emit from both its upper and lower surfaces). We can rearrange to achieve the effective temperature of the radiation $T_{eff}(r)$:

$$T_{eff}(r) = \left(\left|\frac{d\ln\Omega}{d\ln r}\right|\frac{\dot{M}}{4\pi\sigma_{SB}}\Omega^2\right)^{1/4} = \left(\frac{3GM\dot{M}}{8\pi\sigma_{SB}r^3}\right)^{1/4}, \tag{2.135}$$

where we have substituted for the Keplerian angular velocity Ω_K in the last line. We can then construct a spectrum by a superposition of blackbody spectra from a series of annuli between the inner and outer radii r_{in} and r_{out} respectively. To obtain the flux at a given wavelength F_λ, we integrate

$$F_\lambda = \frac{\cos i}{D^2}\int_{r_{in}}^{r_{out}} 2\pi r B_\lambda(T_{eff}(r))\,dr. \tag{2.136}$$

We pre-multiply the integral to account for the distance D to the disc and its inclination to the observer, i. The spectral energy distribution (SED) is a "stretched" blackbody spectrum (see Figure 2.1), with an extended wavelength range according to the effective temperature at the inner and outer edges:

$$\lambda_{min} = \frac{hc}{kT_{in}}, \quad \lambda_{max} = \frac{hc}{kT_{out}}. \tag{2.137}$$

The co-adding of various spectra results in a power law slope at intermediate wavelengths, with an index related to the temperature profile of the disc.

This picture is satisfactory if the system's SED is dominated by accretion, and the dust and gas temperatures are well coupled. Stellar irradiation will in general nullify this assumption by

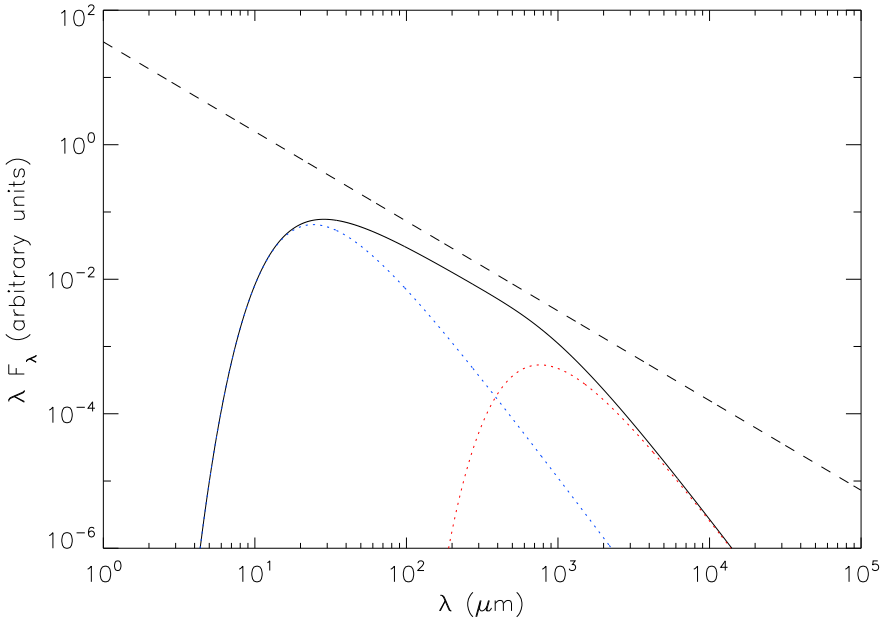

Figure 2.1: An example of a typical disc SED (Class II, as this model has no envelope). Note the elongated power law slope between 10 and 1000 μm (a result of co-adding blackbody spectra with various effective temperatures). Blackbody spectra corresponding to the inner and outer disc temperature are plotted in blue and red respectively.

heating the surface layer of dust. This can result in optically thin emission from the surface layer at much higher effective temperatures than the midplane, producing emission line features such as the silicate feature at around 12μm (cf. Dullemond et al. 2007).

It is common practice to classify protostellar SEDs, in an attempt to characterise their current evolutionary state (from the earliest phases of collapse to the end of the accretion-dominated phase described above). The classification is summarised by Andre et al. (2000) thus:

- *Class 0* - These sources are very faint at wavelengths below $\lambda \sim 10\mu$m, with much of their luminosity in the submillimetre. This is the embedded phase of the protostar, where the mass of the envelope is comparable to the protostar mass, resulting in high extinction at shorter wavelengths.

- *Class I* - These sources have SEDs with positive power law slope between $2 - 20\mu$m, i.e.

$$\lambda F_\lambda \propto \lambda^p \tag{2.138}$$

  with $p > 0$. During this phase the protostar is embedded in a much less massive envelope, and now possesses a protostellar disc.

- *Class II* - As the source's SED evolves towards $-1.5 < p < 0$, the source progresses from Class I to Class II. These are often known as classical T Tauri stars (CTTS) (due to their resemblance to the archetype T Tauri). These protostars now have significant circumstellar discs, and are strongly accreting. Flaring of the disc (increased H at larger radii) will also contribute to boosting the infrared flux through reprocessing of stellar radiation (Kenyon & Hartmann, 1987).

- *Class III* - Once the disc is depleted, the accretion rate slows and the SED steepens to $p < -1.5$ (often known as weak lined T Tauri stars, or WLTTS). This is presumed to happen on relatively short timescales, as intermediate objects (where depletion has not occurred across the entire extent of the disc) are typically rare (Dullemond et al., 2007).

This classification system is not without its ambiguities - depending on inclination, the central star can experience varying amounts of obscuration by the disc, changing the SED from Class II to Class I as the inclination is increased (White et al., 2007). If the protostellar envelope is deformed by magnetic fields or jets, this can also affect the classification of the system.

2.7 Instability in Protostellar Discs

We will now deal with the stability of discs under various perturbations. As reflects their importance to this thesis, I will tend to focus on the treatment of some more than others, in particular on the gravitational instability. As a consequence, this is by no means an exhaustive list of the instabilities that accretion discs can suffer from, but it does reflect the key instabilities we should be mindful of for self-gravitating protostellar discs.

2.7.1 Rotational Instability

We should start with the most conceptually simple instability, which can occur as a result of the disc's rotation. Let us consider a simple axisymmetric disc, composed of an incompressible fluid, which rotates subject to both pressure and gravitational forces. For centrifugal balance, we require:

$$r\Omega^2(r) = \frac{d\Phi}{dr} + \frac{1}{\rho}\frac{dP}{dr} \tag{2.139}$$

Consider two rings of fluid of equal mass, at radii r_1 and r_2 ($r_2 > r_1$). If we interchange the fluid in these rings, the axisymmetry demands the rings' specific angular momenta $j = r^2\Omega(r)$ be conserved. If we rewrite the above equation for ring 1 in terms of j, we get the net acceleration on the ring as a result of the interchange to be

$$\frac{j_1^2}{r_2^3} - \left(\left.\frac{d\Phi}{dr}\right|_{r_2} + \frac{1}{\rho}\left.\frac{dP}{dr}\right|_{r_2} \right). \tag{2.140}$$

But we know the equilibrium condition at the position of ring 2:

$$\frac{j_2^2}{r_2^3} = \left.\frac{d\Phi}{dr}\right|_{r_2} + \frac{1}{\rho}\left.\frac{dP}{dr}\right|_{r_2}. \tag{2.141}$$

Eliminating Φ and P gives the net acceleration on ring 1 to be:

$$\frac{1}{r_2^3}(j_1^2 - j_2^2). \tag{2.142}$$

If $j_1 > j_2$, then the acceleration is positive and continues to force the ring outward. Conversely, if $j_1 < j_2$, the net acceleration is restorative, and attempts to return ring 1 to its original position. We therefore require the gradient of the square of angular momentum to increase with radius, i.e.

$$\frac{d}{dr}\left[(r^2\Omega)^2\right] > 0 \tag{2.143}$$

for rotational stability. This is known as *Rayleigh's criterion* for rotational stability. It is trivial to show that for Keplerian discs

$$\frac{d}{dr}\left[(r^2\Omega_K)^2\right] = GM, \tag{2.144}$$

and are therefore rotationally stable at all radii.

2.7.2 Magnetorotational Instability

While this instability does not feature significantly in this thesis (as magnetic fields are not considered), it does play an important role in discs that are even weakly magnetised (and we will be forced to consider its absence in the context of our simulation results in Chapters 4 and 5). A simple description of the instability will suffice for our purposes - the interested reader should consult Balbus & Hawley (1991) for an in-depth discussion.

Consider a magnetised disc threaded by a vertical magnetic field, with some rotation profile $\Omega(r)$. In ideal magnetohydrodynamics (MHD, see Appendix A) the magnetic field lines are tethered elastically to the fluid. If we displace a fluid element outwards from r_1 to r_2, then the magnetic field will act on the element to:

1. Return the element back to r_1 by resisting field stretching.

2. Attempt to enforce a rigid rotation profile $\Omega(r) = const.$ (by resisting shearing)

The first action is restorative and stabilising, but the second action is not in general. In fact, in the right circumstances it can lead to instability. In the case of zero magnetic field, if $\Omega(r)$ decreases with r, then a displaced fluid element will equilibrate with the flow at r_2, decreasing its angular velocity. If a magnetic field is present, then the field will prevent this, forcing the fluid element to rotate too fast by the second action above (see Figure 2.2). This allows outward

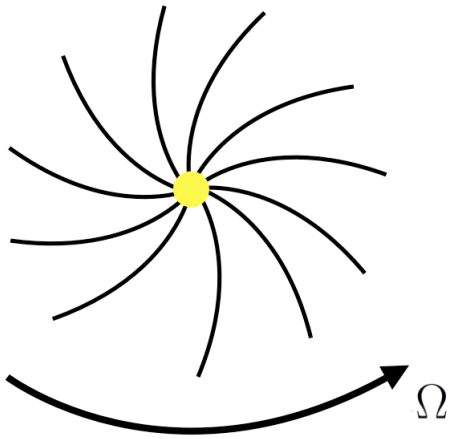

Figure 2.2: A schematic of the magnetic field lines in a differentially rotating disc. The field lines are tethered to the fluid, and are hence curved into the trailing structures seen above as the inner fluid rotates faster than the outer fluid. The field will attempt to resist this curvature by forcing the outer material to rotate faster.

angular momentum transport (and inward mass transport). Interestingly, the stability criterion is independent of the field strength (Balbus & Hawley, 1991):

$$\frac{d\Omega^2}{dr} \geq 0, \tag{2.145}$$

and the result applies without the application of the thin disc approximation. However, the critical wavelength at which the stability activates is directly proportional to the field strength, where even a small seed field strength can generate magnetorotational instability (MRI). This will prove to be important in the inner regions of the discs we study, where the temperatures are high enough for the gas to be sufficiently ionised, allowing MRI to develop. This is of importance for theories concerning the generation of FU Orionis outbursts (Armitage et al., 2001; Zhu et al., 2009b), which is discussed further in Chapter 5.

2.7.3 Thermal Instability

If our gas discs are to be modelled correctly, complex radiative physics resulting from a variety of different chemical and physical processes must be incorporated. The cooling functions we will generate in such discs (i.e., the energy loss rate as a function of density and temperature) will be undoubtedly non-trivial. If these functions satisfy certain criteria, then it should be clear that the disc can become thermally unstable, rapidly undergoing transitions from one phase of gas to another on macroscopic scales.

Let us consider a generalised *net* cooling rate function for a gas, $\Lambda(\rho, T)$, which we define as the local cooling rate minus the local heating rate (negative values of Λ indicate the gas will undergo net heating). The condition for thermal equilibrium is simply

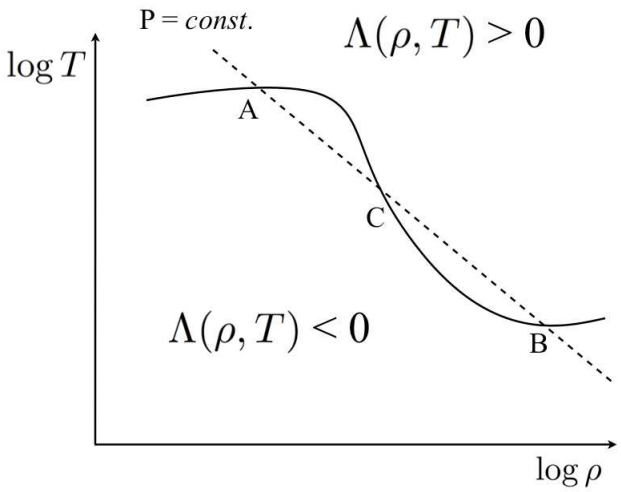

Figure 2.3: Demonstrating the thermal instability using a non-trivial cooling function $\Lambda(\rho, T)$. The solid line represents the densities and temperatures at which $\Lambda = 0$, and hence no heating or cooling occurs. The dotted line represents a locus of constant pressure in $\log \rho - \log T$ space. The gas is thermally stable to perturbations at points A and B in the diagram, but not at point C.

$$\Lambda(\rho, T) = 0. \tag{2.146}$$

This condition can be represented by a contour or line in (ρ, T) space (although it is more useful to use $(\log \rho, \log T)$ given the many orders of magnitude available to astrophysical gases). A cartoon of this line can be seen in Figure 2.3. The steps exhibited by the curve are somewhat representative of reality due to the quantum interactions which give rise to discrete energy levels in the atoms/molecules (such as rotational/vibrational excitations and ionisation). If we consider a gas at constant ρ while we increase T, then these energy levels become excited, and are able to subsequently cool radiatively to return to their original energy state. The strong gradient of Λ through this jump is due to the sensitive temperature dependence of the level populations, which is proportional to the Boltzmann factor $e^{-E/kT}$, where E is the energy in excess of the activation energy for the transition. As T is increased to values where $kT \sim E$, the fraction of excited atoms/molecules increases rapidly, giving the gas more participating cooling agents and increasing the value of Λ over a short range of temperatures, giving rise to the steep jumps.

If we now imagine that we hold the *pressure* of the gas constant, then for an ideal gas this corresponds to the locus $\rho T = const.$[2]. This locus can be represented by a line in $(\log \rho, \log T)$ space (dotted line in Figure 2.3). We have selected the value of P to represent three interesting limiting cases where initially $\Lambda = 0$.

In the first limit (case A), we can displace the gas along the constant pressure locus upwards

[2]We neglect changes to molecular weight and ionisation for illustrative ease

in temperature (and downwards in density), giving a positive Λ which activates cooling to restore the gas to equilibrium. Equally, we can displace the gas downwards in temperature and upwards in density to give a negative Λ which heats the gas back to equilibrium. We can repeat the process for case B and find that again the gas returns to equilibrium. Perturbations at these locations are thermally stable.

In case C, we find very different behaviour. Displacing the gas upward in temperature (and downwards in density) results in a negative Λ, which continues to heat the gas until it reaches the equilibrium point at case A. If we displace the gas downward in temperature, then the gas will begin to cool until it reaches the equilibrium point at case B. Case C is thermally unstable - the slightest perturbation to equilibrium forces the gas to undergo a phase transition to another state. This allows us to predict density-temperature configurations that should not be observable due to their thermal instability.

We therefore see that increasing temperature requires an increased Λ (and decreasing temperature requires a decreased Λ) for stability. More rigorously

$$\left(\frac{\partial \Lambda}{\partial T}\right)_P > 0. \tag{2.147}$$

Equally in terms of density

$$\left(\frac{\partial \Lambda}{\partial \rho}\right)_P < 0. \tag{2.148}$$

We have admittedly constructed a very simple case for constant pressure cases only - we have neglected the effects of the changing level populations on the form of Λ, and we have also implicitly assumed it to be optically thin by assigning local values of T to Λ. In the case where the gas is optically thick, then radiative transfer can act to reduce the instability by reducing the gradients in Λ (Field, 1965). Equally, radiative transfer may encourage thermal instability by heating neighbouring gas towards a regime where the cooling function satisfies equation (2.147), which may not have been possible otherwise. For these reasons, predicting the outcome of thermal instability in non-axisymmetric phenomena such as discs with strong spiral structure (see next section) is typically the province of numerical simulations. Despite these difficulties, thermal instability has been proposed as a viable alternative to the MRI theories of FU Orionis outburst phenomena (Bell & Lin 1994 and references within, see Chapter 5).

2.7.4 Gravitational Instability

Arguably of the most importance to self-gravitating discs, perturbations can grow under their own self-gravity to destabilise the disc. Under differential rotation, such perturbations will grow into spiral density waves. Therefore to characterise the growth of gravitational instability, we must calculate under what conditions a spiral density wave can grow and sustain its amplitude in the disc. In Appendix C I elaborate on the description of spiral structure in the disc in terms

of the radial wavenumber k and azimuthal wavenumber m, and derive a dispersion relation which connects them (for thin discs):

$$m^2(\Omega - \Omega_p)^2 = c_s^2 k^2 - 2\pi G\Sigma\,|k| + \kappa^2, \tag{2.149}$$

where κ is the epicyclic frequency, and Ω_p is the pattern speed of the wave. By subjecting the disc to an axisymmetric disturbance ($m = 0$) and substituting $\omega = m\Omega_p$, then equation (2.149) gives

$$\omega^2 = c_s^2 k^2 - 2\pi G\Sigma\,|k| + \kappa^2. \tag{2.150}$$

The right hand side contains all real terms. This means that the left hand side should be real also. However, if $\omega^2 < 0$, then ω is imaginary, and the disturbance's amplitude will grow exponentially: $e^{-i\omega t} \rightarrow e^{\pm|\omega|t}$. Therefore, we can define a stability criterion at $\omega = 0$:

$$c_s^2 k^2 - 2\pi G\Sigma\,|k| + \kappa^2 = 0. \tag{2.151}$$

This is a quadratic equation for k. If the disc is to be stable to any axisymmetric disturbance, then there must be no solution for a positive value of $|k|$. This is equivalent to the requirement that the roots of this equation be imaginary. This constrains the discriminant of the quadratic thus:

$$4\pi^2 G^2\Sigma^2 - 4c_s^2\kappa^2 < 0. \tag{2.152}$$

Rearranging and taking square roots gives the stability condition

$$Q = \frac{c_s\kappa}{\pi G\Sigma} > 1. \tag{2.153}$$

This is known as *Toomre's stability criterion*, and Q is known as the Toomre parameter (Toomre 1964, although similar results were derived by Safronov some years earlier). This indicates the susceptibility of discs to axisymmetric disturbances - differential rotation will ensure the growth of spiral arms which are distinctly non-axisymmetric. Numerical work has shown that $Q > 1.5 - 1.7$ for stability against non-axisymmetric disturbances (Durisen et al., 2007).

Marginal Instability and Self-Regulation The Toomre parameter is more fundamental than its derivation might admit. It is composed of the three fundamental physical processes at work in self-gravitating gaseous discs: angular momentum (through the epicyclic frequency), thermodynamics (through the sound speed) and self-gravity (through Σ). Q represents the balance between these three processes.

"Hot" discs, that is those with $Q \gg 1$, are able to cool (in particular through radiative processes), reducing their sound speed and decreasing Q towards the instability limit. As they reach the boundary, the disc becomes "cool" (i.e. $Q \lesssim 2$, and unstable to non-axisymmetric disturbances), and spiral waves develop. As the amplitude of the waves increases, they begin

to heat the discs through shocks[3]. This heating will drive the sound speed up, and increase Q to above the instability limit, where the spiral activity will then cease.

The crossover between the global heating and global cooling behaviour is the instability boundary, and represents a locus of stable equilibrium. The disc can return to this state if it is perturbed away from it in either direction (i.e, by heating or cooling), provided that the perturbations are small. Therefore, self-gravitating discs can sustain spiral waves of steady amplitude over long timescales, where their value of Q is *self-regulated* to be close to unity. This self-regulated state is referred to as *Marginal Instability* (Paczynski, 1978).

Marginally unstable discs can generate a range of m mode spirals. Their steady production generates turbulence in the disc (often referred to as *gravito-turbulence*). The failure of hydrodynamical viscosity to produce observed accretion discs has led authors to find other sources of "viscosity", including gravito-turbulence. By doing this, an effective viscosity can be attributed to the gravitational instability, and the viscous equations described above can be used to evolve the disc. This provides a straightforward route to constructing semi-analytic models of self-gravitating discs (Gammie, 2001; Clarke, 2009; Rice & Armitage, 2009; Rice et al., 2010). The efficacy of this approach will be discussed in more detail in Chapter 4.

2.8 Disc Depletion and the End of the Self-Gravitating Phase

Whether by the steady accretion of the protostellar disc material by the central star or by other mass loss processes, the disc's mass will eventually be depleted. The disc's self-gravity will eventually become negligible, as the ratio of disc mass to star mass becomes smaller and smaller. The duration of the self-gravitating phase is determined primarily by the duration of the infall phase between Class I and Class II (in which the disc accretes from its surrounding envelope still extant after the collapse of the molecular cloud), and the extent to which stellar irradiation acts on the gas.

Stellar irradiation will heat the surface layers of gas in the disc. If the sound speed of the gas exceeds the escape velocity, then the surface gas will be liberated from the disc, giving rise to *photoevaporation*. This typically shortens the disc lifetime by encouraging viscous forces to "fill in the gaps" (Clarke et al., 2001; Alexander et al., 2006a,b). In extreme cases (e.g., where the system is near a massive star), the disc's shape is distinctively distorted, giving rise to the so-called "proplyds" observed in Orion (Hollenbach & Adams, 2004).

Typically, photons in the far and extreme ultraviolet (FUV and EUV respectively) are responsible for photoevaporation - X rays are not typically considered a significant photoevaporation source. EUV photons heat the surface layer to $T \sim 10^4$ K, whereas FUV photons tend to

[3]Spiral waves also provide compressive heating, but this heating is usually coupled with rarefaction and cooling, so while there is local heating, the global net heating effect is close to zero.

penetrate deeper, heating the lower layers up to $T \sim 10^3$ K. By comparing local sound speeds and escape velocities, we can derive a critical radius at which photoevaporation can begin, r_g:

$$r_g = \frac{2GM\mu m_H}{kT} \tag{2.154}$$

This radius typically corresponds to a few AU. If the disc radius is larger than r_g, then the disc can either evaporate inwards or outwards, depending on the evaporation timescale:

$$t_{vap} = \frac{\Sigma}{\dot{\Sigma}} \tag{2.155}$$

Hollenbach et al (1994) show that typically $\dot{\Sigma} \propto r^{-2.5}$. Therefore discs that are steeper than $\Sigma \propto r^{-2.5}$ will have an evaporation timescale that decreases with increasing r, resulting in evaporation from r_g outward. If the disc is shallower than $\Sigma \propto r^{-2.5}$ (as is typically the case for most circumstellar discs), the evaporation will proceed from r_g inward, resulting in an inner hole.

The combination of viscous processes and photoevaporation gives a typical gas disc timescale of $1 - 10$ Myr (Vorobyov, 2010), describing the transition from massive, gaseous circumstellar disc to lower mass protoplanetary disc to gas-free debris disc. Any solids strongly coupled to the gas (such as dust grains) will also be removed at this time, leaving only intermediate and large solid bodies in the system. If gaseous planet formation is to take place in such a star system, then this process sets the time limit by which they must be formed. A discussion of current planet formation theory is given below.

2.9 Planet Formation in Protoplanetary Discs

The current science of planet formation covers a rich panoply of dynamical, chemical and radiative processes which govern the transition from a protoplanetary disc of dust and gas to a planetary system with planets, moons, comets, asteroids and other debris. To avoid going off on a tangent, I will give only a *prècis* of the current theory, with the focus firmly on the role of the disc in the process of planet formation.

Before we begin, we should carefully define what we mean by "planet". It is useful to quote IAU Resolution 5A (2006) for planets **in our Solar System**:

"The IAU therefore resolves that planets and other bodies in our Solar System, except satellites, be defined into three distinct categories in the following way:

1. *A 'planet' is a celestial body that (a) is in orbit around the Sun, (b) has sufficient mass for its self-gravity to overcome rigid body forces so that it assumes a hydrostatic equilibrium (nearly round) shape, and (c) has cleared the neighbourhood around its orbit.*

2. *A 'dwarf planet' is a celestial body that (a) is in orbit around the Sun, (b) has sufficient mass for its self-gravity to overcome rigid body forces so that it assumes a hydrostatic*

equilibrium (nearly round) shape, (c) has not cleared the neighbourhood around its orbit, and (d) is not a satellite.

3. *All other objects, except satellites, orbiting the Sun shall be referred to collectively as 'Small Solar-System Bodies'."*

This definition is satisfactory for our Solar System, but does not deal with the so-called *brown dwarfs*, low mass stellar objects which perform deuterium fusion at their core, and by this action not categorisable as a planet *per se*. We will therefore add to the above definition that planets must not fuse deuterium in their core, in effect establishing a maximum planetary mass at around $12 - 14M_J$, where M_J indicates the mass of Jupiter.

As any child knows, planets in the Solar System come in roughly two flavours - the rocky *terrestrial planets* (e.g. Mercury, Venus, Earth and Mars) and the more massive *gas giants* (Jupiter, Saturn, Uranus and Neptune), with objects such as Pluto falling into the dwarf planet category. More pedantic children will tell you that we should consider Uranus and Neptune separately as *ice giants* due to their composition - this is a useful distinction, and recognising it may give us an important diagnostic of the Solar protoplanetary disc at early times.

It is widely accepted that the terrestrial planets formed by the process of *core accretion* (which is discussed in more detail below), although the exact details are complicated by the chaotic dynamics of the many body problem. The origin of the giant planets is on less firm ground, although in the Solar System core accretion again is the most commonly accepted theory (despite its issues which I discuss below). However, we cannot base our theories on the Solar System alone - the discovery of the extrasolar planets (or exoplanets) has increased the available dataset by almost two orders of magnitude (given the latest data from the Kepler space telescope, see Borucki 2010). We now know that the Solar System is not exactly the archetype for planetary systems in our Galaxy - we see giant planets orbiting at semi-major axes of $a < 0.1$ AU, planets with masses intermediate to the terrestrial and giant planets (Borucki, 2010), planets on highly eccentric orbits (e.g. HD 80606 b with $e = 0.934$, Naef et al. 2001; Moutou et al. 2009), orbiting retrograde (e.g. WASP-8b, Queloz et al. 2010), multiple systems of planets near the deuterium burning limit at large separations from their parent star (such as HR 8799, Marois et al. 2008) and so on. Any theory of planet formation must be able to explain all the systems we see, not just the Solar System. I will now describe the core accretion model, along with its healthiest competitor, the *disc instability* model.

2.9.1 The Core Accretion Theory of Planet Formation

Consider a two-phase protoplanetary disc still consisting mostly of gas, with a dust component of mass equal to some fixed fraction of the total gas mass - usually around 1% for interstellar grains, although this need not be the minimum value (Thi et al., 2010). The gas orbits at near Keplerian velocities, with some reduction in speed due to the radial pressure gradient the gas creates. The dust in the disc will orbit at Keplerian velocities, not being directly affected by

the pressure gradient. The two phases therefore share a velocity difference, and interact by drag forces.

The smallest (sub)micron-sized dust grains are tightly coupled to the gas, and orbit in train with it (Whipple, 1973; Weidenschilling, 1977). At values of semi-major axis a beyond the *snow line*, where the disc is cool enough for water ice to condense, the grains can begin to coagulate by adhering to the ice particles. As the grains grow, the drag forces begin to grow, and the dust experiences a torque that directs it towards the pressure maxima in the disc. As protoplanetary discs typically have negative radial and vertical pressure gradients, this results in two important effects: inward radial migration of dust resulting in enhanced accretion onto the central star, and vertical settling of dust into the midplane. The effects of drag peak when the dust agglomerates to intermediate sizes (centimetres to metres), coinciding with the regime where collisions at typical velocities will begin to grind the aggregates back to lower grain size rather than grow them (cf Stewart & Leinhardt 2009).

This hampering of growth combined with the rapid inward migration should result in the destruction of grains before they can exceed grain sizes of a few metres. This is one of the greatest weaknesses of core accretion theory, and has attracted a great deal of effort in its resolution.

Solutions to it typically invoke local pressure maxima preventing the dust from migrating radially. Examples of this are disc spiral arms (Rice et al., 2004; Clarke & Lodato, 2009) or vortices in the disc at smaller scales (Mamatsashvili & Rice, 2009; Lyra et al., 2009). These solutions have the added advantage of increasing the local dust density and reducing the relative velocity of the grains, potentially increasing the likelihood of sticking collisions.

Whatever the route of accretion, the grains eventually aggregate into *planetesimals* a few kilometres in size. The planetesimals then undergo further collisions to form *planetary embryos*. The accretion rate is sensitive to the ratio between the escape velocity of the embryo and its relative velocity to the planetesimal disc. The initial growth rate is slow and decreases with mass, taking typically several Myr to develop embryos with masses equivalent to the dwarf planets.

Once the escape velocity is significantly larger than the planetesimal's relative velocity, the growth rate begins to increase with mass, and runaway growth begins, forming *protoplanets*. As multiple embryos will typically grow in a planetesimal disc, the embryos will tend to form in regular spacings of semi-major axis with masses regulated relative to each other, a process referred to as *oligarchic growth*.

In the case of the giant planets, the protoplanets are the seed that forms the planet's core. These protoplanets can then begin to accrete gas from the disc, forming an envelope. The envelope must thermally relax to become bound to the protoplanet, ensuring that this process is initially quasistatic. This continues until the pressure support of the envelope is finally superseded by gravity, and the protoplanet rapidly accretes a substantial amount of gas to form the planet's atmosphere.

During this period of runaway growth, the protoplanet's accretion rate is limited by the available feedstock provided by the disc. The volume from which the protoplanet can successfully accrete is determined by the Hill Radius, R_h, which defines the region in which its gravity dominates the local potential:

$$R_h = a \left(\frac{M_p}{3M_*} \right)^{1/3}. \tag{2.156}$$

The protoplanet will also begin to exert a tidal interaction on the gas disc, reducing the accretion rate. Eventually, the planet is formed once its feedstock is completely depleted - either because it has accreted all the gas within its Hill sphere, or because the disc has lost its gas through processes such as photoevaporation (which occurs on a typical timescale of $1 - 10$ Myr). The restriction imposed by the gas depletion timescale ensures that core accretion is unlikely to form gas giants at large radii before the gas disappears, due to the low surface density of material.

2.9.2 The Disc Instability Theory of Planet Formation

We have seen that the star formation process occurs as a result of gravitational instability - is it not conceivable that giant planets may form through a similar route (Kuiper, 1951; Cameron, 1978)? This idea has been recently revived by Boss (1997), and is the subject of a great deal of study. We have set out the process by which discs can become gravitationally unstable in previous sections. It should be clear that if the instability is sufficiently strong, the disc will *fragment* into clumps, some of which may then become bound, self-gravitating objects. This picture is particularly attractive because such a process would take place on the orbital timescale of the disc, i.e. within around 10^{-3} Myr, much shorter than under core accretion.

In section 2.7.4 we derived the Toomre criterion for gravitational instability in (Keplerian) discs:

$$Q = \frac{c_s \Omega}{\pi G \Sigma} \lesssim 2 \tag{2.157}$$

However, we have also shown that discs can thermally regulate their gravitational instability, modulating the local sound speed to keep the disc marginally stable and non-fragmenting at $Q \sim 2$. Therefore, the Toomre criterion is a necessary but insufficient condition for disc fragmentation - it only gives information about how the local physical processes of radiative transfer, hydrodynamics and self-gravity compare to each other at any one instant. We require a second criterion which informs us of how these processes change relative to each other as a function of time.

We can outline the form of the second criterion heuristically. Consider a gravitationally unstable disc with $Q < 2$. The development of non-axisymmetric instabilities creates spiral structure in the disc, which grow on the dynamical timescale $t_{dyn} \sim \Omega^{-1}$. The disc will be heated by shocks, and cooled through radiative processes. If the cooling is inefficient, then the shock heating will allow the disc to increase Q back to a marginally stable state. If the

cooling is efficient, shock heating will be unable to raise Q and shut off the instability. Thermal regulation and marginal instability will fail, and unstable perturbations (such as clumps) can then grow in the disc.

Efficient cooling also allows these clumps to become bound objects by radiating away their thermal energy, until its magnitude is less than the gravitational potential energy of the clump. Without efficient cooling, the clumps will not become bound and will be sheared out as a result of the differential rotation in the disc.

We can express this criterion as a ratio of cooling time t_{cool} to the dynamical timescale (Gammie, 2001; Rice et al., 2003):

$$t_{cool}\Omega = \beta < \beta_{crit}(\gamma) \tag{2.158}$$

The exact value of β_{crit} depends on the local equation of state, but ranges between 3 and 6 depending on the value of the ratio of specific heats, γ (Rice et al., 2005). This dependence can be couched in terms of a maximum rate of angular momentum transport, which we write in terms of the Shakura-Sunyaev parameter α:

$$\alpha = \left(\frac{d\ln\Omega}{d\ln r}\right)^{-2} \frac{1}{\gamma(\gamma-1)\beta}. \tag{2.159}$$

A full discussion of angular momentum transport in discs (including a derivation of the above equation) can be found in Chapter 4. When we substitute empirically determined values of β_{crit} for various values of γ into the above equation, we find that the critical $\alpha \approx 0.06$ in all cases. This is confirmed by Cossins et al. (2009)'s empirical result that the amplitude of the spiral arms is inversely correlated to the cooling time. If the cooling time is too short, then the torques induced by self-gravity become too large to be sustainable, and fragmentation is therefore inevitable.

This result only holds in the steady state - if the disc undergoes transient episodes of strong spiral activity on timescales shorter than the dynamical timescale, then α can (briefly) exceed 0.06. Interestingly, if fragmentation occurs as a result of unsustainable torques, these may be generated by forcing the disc to maintain a high accretion rate from its surroundings, even if the cooling time is long (e.g. Kratter et al. 2010; Vorobyov & Basu 2010).

What can we say about planets formed by disc instability? If we note that the disc's most unstable wavelength is $\lambda = 2\pi H$ (Lodato, 2007), then we can estimate a fragment mass of order $M_{frag} = \Sigma\lambda^2$. If $Q \sim 1$, we can substitute for Σ to obtain (for a Keplerian disc)

$$M_{frag} = 4\pi M_* \left(\frac{H}{r}\right)^3 \tag{2.160}$$

Typical protostellar discs have an aspect ratio $\frac{H}{r} \approx 0.1$. For a solar mass star, we therefore derive a typical fragment mass $M_{frag} \approx 13 M_J$. This rules out disc instability as a means of directly forming terrestrial planets or low-mass gas giants - indeed, being so close to the deuterium burning limit it appears to almost rule out forming planets of any kind at all.

However, it may provide an excellent means of forming brown dwarfs (Stamatellos et al., 2007a; Stamatellos & Whitworth, 2009).

And what of the cores of these objects? Initially they will not have a solid core of any sort - in this scenario, the core forms after the planet, mainly by differentiation in the planet itself as a result of the drag forces discussed in the previous section. This occurs at a speed $v_d = gt_f$, where g is the gravitational acceleration inside the planet and t_f is the friction time, which is a function of the volume density of the solids ρ_d and their radius r_d (Epstein, 1924):

$$t_f = \frac{\rho_d}{\rho_g} \frac{r_d}{c_s} \tag{2.161}$$

The timescale for this sedimentation to occur is short compared to the thermal relaxation timescale of the gas (Helled et al., 2006), even with moderate turbulence within the protoplanet. Indeed, this sedimentation is expected to occur in the spiral arms of the gravitationally unstable disc, boosting the fragment's initial metallicity (Rice et al., 2004; Boley & Durisen, 2010). For these reasons, it is reasonable to expect a $1M_J$ planet to be able to form a core of around $6M_\oplus$ (D'Angelo et al., 2010). While around 2 to 4 times less than what core accretion would predict, this is still a tenable core mass.

2.9.3 Disc-Planet Interactions

Regardless of however it has come into existence, once a planet has formed it will in general still be interacting with the remainder of the protoplanetary disc (whether it is still gaseous or has evolved to the gas-free, dusty debris disc phase). The planet exerts gravitational torques on the disc, imparting angular momentum to the material outwith its orbit and depleting the angular momentum of the material within. Conservation of angular momentum demands that any torque the planet exerts on the disc must have a corresponding torque exerted by the disc on the planet (with opposite sign). Therefore, the planet loses angular momentum to the material outwith its orbit, and gain it from the material within the orbit. These torques will not in general balance, and the planet will undergo modification of its eccentricity (Goldreich & Sari, 2003) and potentially *orbital migration*. In typical cases, the net torque the planet experiences is negative, and the planet will migrate inwards.

For low mass planets (whose Hill radius is less than the disc scale height), the planet undergoes *type I* migration, which can be rapid compared to the thermal relaxation timescale of the planet's atmosphere. This poses problems for core accretion models in that giant planets at larger radii must somehow evade this rapid migration to remain at their current radius (and avoid their demise by accretion onto the central star). Disc instability models fare better, as the migration mode switches to the more sedate *type II* migration, when the planet is sufficiently massive that its Hill radius is comparable to the local scale height of the disc. The planet can then establish an annular gap, depleting the surface density of gas significantly near the planet (cf Lin & Papaloizou 1986). If the gap is "clean" and empty of material, the planet is

locked in the gap and does not migrate. However, the disc is also subject to viscous torques, which will act to spread the disc and fill the gap. This allows migration to commence, but at a much slower timescale controlled by viscous diffusion. If the disc is depleted for any other reason (e.g. at the disc's inner edge) this will also result in stopping the planet's migration. This provides an appealing explanation for the Jupiter mass exoplanets found at $a < 0.1$ AU, whose *in situ* formation would otherwise stump both core accretion and disc instability theory. An intermediate migration type, sometimes referred to as Type III, has also been identified (Masset & Papaloizou, 2003) where a partial gap is opened in a massive disc, and corotation torques induce a rapid migration phase.

The basic picture of disc-planet interaction was initially established for laminar discs with an isothermal equation of state, assumptions which have since been relaxed. More general equations of state (Paardekooper & Papaloizou, 2008) have shown that altering the local entropy gradient significantly affects the migration, potentially altering its direction. If the disc is turbulent, low mass planets can undergo stochastic migration (Nelson & Papaloizou, 2004) where the net torque can change sign rapidly, slowing or even reversing Type I migration. These results underline the important role of non-trivial disc physics in determining the fate of newly formed planets.

2.9.4 Testing Planet Formation Theory

The primary test of any theory is observation. Looking at the current dataset, we should consider core accretion and disc instability in turn, and the general trends that they both predict.

The positive correlation between stellar metallicity and frequency of giant planets (Udry & Santos, 2007) tends to favour core accretion theory if stellar metallicity and density of dust grains are also positively correlated (Kornet et al., 2005). Disc instability would appear to favour a negative correlation in general due to the effects of metallicity on radiative efficiency (Cai et al., 2006) although this may be complicated by grain collection in spiral arms.

An important result to still be determined is an empirical relationship between stellar mass and planet frequency. Core accretion would predict lower frequency of giant planets for lower mass stars, although Neptune mass planets may be more common (Laughlin et al., 2004). Disc instability does not predict any strong correlation, as the conditions for disc fragmentation are relatively insensitive to the stellar mass (Boss, 2006). In general, statistical bias towards more massive, low-a planets makes the process of establishing correlations difficult, if not impossible. We do not have sufficient information on low mass planets, planets around low mass stars, or planets at high a.

Alternatively, we can investigate the two theories in the wider context of planetary science. Current evolutionary models predict that Jupiter and Saturn have core masses of around $10 - 20 M_{\oplus}$ (Podolak et al., 1995; Saumon & Guillot, 2004). Core accretion readily explains the construction of such massive cores, whereas disc instability has more difficulty in doing so (not

least because forming giant planets at $5 - 10$ AU by gravitational instability seems highly unlikely).

Comparing the respective failures of both theories to consistently explain the entire exoplanet dataset, it may be sensible to consider planet formation as a two-mode process, where both core accretion and disc instability act in different regimes to produce the distributions of masses and orbital parameters we see (Boley, 2009). This would therefore suggest that core accretion forms terrestrial planets and gaseous objects at $a \lesssim 50 - 100$ AU, whereas disc instability forms massive gaseous objects at $a \gtrsim 50 - 100$ AU, with the exact boundaries of each regime being blurred by dynamical effects such as migration and planet-planet scattering. On top of this, gravitational instability may prove important in concentrating dust in disc spiral arms, providing the required reduction in aggregation timescales for core accretion to function, or to provide a means of two-phase disc fragmentation (Rice et al., 2006).

While there have been some significant advances in planet formation modelling, there are still gaps in our knowledge surrounding the initial conditions for the exoplanets. We must fully understand the typical masses and lifetimes of protoplanetary discs, as well as their temperature and composition if we are to explain the distributions of planetary parameters. This requires a well defined causal chain that takes us from the protoplanetary disc back to the protostellar disc, through its self-gravitating phase and connecting finally to the molecular cloud from which the star-disc system originates. I have structured this chapter in chronological order primarily to emphasise this need, and to illustrate the fundamental role that discs play in both star and planet formation.

Chapter 3

Smoothed Particle Hydrodynamics, and a Hybrid Method of Radiative Transfer

Everything should be made as simple as possible, but no simpler.
Albert Einstein, *On the Method of Theoretical Physics*

3.1 Author's Note

The first half of this chapter is pedagogical in content, and introduces the mathematical and numerical apparatus on which SPH is built. The second half of this chapter describes a radiative transfer algorithm I developed for SPH along with several collaborators (Forgan et al., 2009). For derivations involving index notation, Einstein's Summation Convention (ESC) is assumed.

3.2 Introducing Smoothed Particle Hydrodynamics

Motivation In the previous chapter, we have seen the importance of fluid dynamics in astrophysics. Understanding the flow of gas under the influence of hydrodynamics, gravity and radiative physics is critical to understanding the formation of practically every celestial object. While the equations of hydrodynamics may be directly soluble in certain circumstances, in general astrophysical flows are too complex to entertain an analytical solution. Therefore,

at the frontier of human knowledge, *numerical hydrodynamic simulations* are important for elucidating the varying roles of the fundamental physical forces in astronomy - for example, understanding the modes of low-mass star formation in molecular clouds requires a description of how turbulent gas clouds evolve under gravity, hydrodynamics and radiative physics to produce observed density and velocity structures. Complex multi-scale, multi-physics evolution of this nature can only be described numerically.

Numerical hydrodynamics can be broadly classified into two types[1]. Historically, the first to be developed were grid-based, finite difference methods, which function by dividing the spatial domain into grid cells, calculating the equations of hydrodynamics at cell interfaces. Grid-based methods are well established in hydrodynamic simulations, and have a good pedigree in simulating astrophysical fluids. However, before the development of adaptive mesh refinement (AMR) techniques (e.g. Winkler 1984), they were often subject to resolution issues, where the established Eulerian grid (which is fixed in space) is unable to simulate the evolution of non-axisymmetric overdensities. Even with AMR or the use of Lagrangian grids (which are allowed to move and deform with the flow), grid-based codes are still subject to boundary condition and advection problems.

The second type of simulation is particle-based. Instead of a spatial discretisation, particle-based methods discretise the fluid's material content directly, solving the equations of hydrodynamics between particle pairs or ensembles. This method is Lagrangian by construction, and does not require boundaries to be imposed on the simulation. The first particle-based simulation was espoused by Lucy (1977) to investigate the fission hypothesis in the formation of binary star systems. This method was further developed and formalised (Gingold & Monaghan, 1978; Monaghan, 1982; Gingold & Monaghan, 1982), and was later dubbed Smoothed Particle Hydrodynamics (SPH). It has gone on to become an important simulation technique for astrophysical fluids at all size scales, due to its extreme adaptability and its ability to intuitively incorporate new physics.

The motivation for this particle-based method is rooted in Monte Carlo theory (see Hammersley & Handscomb 1964). Consider the function $A(\mathbf{r})$, defined in some volume V:

$$A(\mathbf{r}) = \int_V W(\mathbf{r}-\mathbf{r}')\xi(\mathbf{r}')\rho(\mathbf{r}')dV'. \tag{3.1}$$

If a set of N points are distributed randomly in the volume at positions $\{\mathbf{r}_j\}$ (such that the probability of a point being found in a volume element dV' at $\mathbf{r}'$ is equal to $\rho(\mathbf{r}')dV'$), then Monte Carlo theory indicates that the approximation

$$\tilde{A}(\mathbf{r}) = \frac{1}{N}\sum_j W(\mathbf{r}-\mathbf{r_j})\xi(\mathbf{r}_j) \tag{3.2}$$

will converge on the true $A(\mathbf{r})$ as $N \to \infty$. If we constrain W thusly:

[1]Strictly speaking, spectral methods constitute a third class of hydrodynamic simulation. Considering the two main types alone is more useful to the narrative of this thesis, as there is typically a dichotomy both in methods and use, which I address somewhat in Chapter 6.

$$\int_V W(\mathbf{r}-\mathbf{r}')dV' = 1, \tag{3.3}$$

$$W = 0 \,\text{for}\, |\mathbf{r}-\mathbf{r}'| > \Delta, \tag{3.4}$$

then it can be shown that

$$A(\mathbf{r}) \to \xi(\mathbf{r})\rho(\mathbf{r}) \tag{3.5}$$

and correspondingly

$$\tilde{A}(\mathbf{r}) \to \xi(\mathbf{r})\rho(\mathbf{r}) \tag{3.6}$$

as $N \to \infty$ and $\Delta \to 0$. $A(\mathbf{r})$ can be viewed as an interpolated function, with W representing an interpolating kernel. If we take A to be a state variable in a fluid (such as pressure or internal energy), then its functional form over all space can be estimated by calculating its value at a set of points which are placed according to the density of the fluid, and interpolating as shown above.

This is the crux of the SPH method. The N points are viewed as particles which move with the fluid. Each particle has a mass m_j, density ρ_j, position $\mathbf{r}_j$ and velocity $\mathbf{v}_j$. From this set of disordered points $\{m_j, \rho_j, \mathbf{r}_j, \mathbf{v}_j\}$ (and an equation of state), all the properties of the fluid can be estimated, and evolved using the standard equations of hydrodynamics. A state variable $A(\mathbf{r})$ is represented in SPH as

$$A_{SPH}(\mathbf{r}) = \sum_j m_j \frac{A_j}{\rho_j} W\left(\mathbf{r}-\mathbf{r}_j, h\right) \tag{3.7}$$

Where A_j is the value of A at the position of particle j, W is the *smoothing kernel*, and h is the *smoothing length* (this corresponds to Δ in the previous formalism). I drop the SPH subscript here - the context should now make it clear when an SPH summation is being performed. The smoothing kernel satisfies

$$\int W(\mathbf{r}-\mathbf{r}', h)d\mathbf{r}' = 1 \tag{3.8}$$

$$W = 0 \,\text{for}\, |\mathbf{r}-\mathbf{r}'| > 2h, \tag{3.9}$$

$$\lim_{h\to 0} W(\mathbf{r}-\mathbf{r}', h)d\mathbf{r}' = \delta\left(\mathbf{r}-\mathbf{r}'\right) \tag{3.10}$$

where we employ the Dirac δ-function in the last equation. In practice, the interpolating sum is evaluated over a list of nearest neighbours only, defined by the smoothing length. This is borne out by the second criterion on W: particles outside the so called *smoothing volume* or *neighbour sphere* (a sphere of radius $2h$) do not contribute to the sum. Kernels which have this

feature are said to have *compact support*. The smoothing kernel is an analytical, differentiable function, whose derivatives are known. Therefore, derivatives of A are easy to obtain by using W:

$$\nabla A(\mathbf{r}) = \sum_j m_j \frac{A_j}{\rho_j} \nabla W\left(\mathbf{r} - \mathbf{r}_j\right) \tag{3.11}$$

In practice, more stable derivatives are calculated using symmetrised forms (see following sections).

Properties of The Smoothing Kernel and its derivatives The smoothing kernel can be any function which satisfies the above conditions, and is differentiable. The most obvious is a Gaussian kernel (Gingold & Monaghan, 1978):

$$W(x,h) = \frac{1}{h\sqrt{\pi}} e^{-\left(\frac{x^2}{h^2}\right)} \tag{3.12}$$

Monaghan (1992) posits a "first golden rule of SPH" thus: To understand the physical interpretation of SPH equations, assume a Gaussian kernel. Other kernels are typically used, including those based on splines (Monaghan & Lattanzio, 1985). For example, in 3D simulations, a typical cubic spline kernel takes the form:

$$W(\mathbf{q} = \frac{\mathbf{r}}{h}, h) = \frac{1}{\pi h^3} \begin{cases} 1 - \frac{3}{2}q^2 + \frac{3}{4}q^3 & 0 \leq q \leq 1 \\ \frac{1}{4}(2-q)^3 & 1 \leq q \leq 2 \\ 0 & q \geq 2 \end{cases} \tag{3.13}$$

This kernel is efficacious as it has compact support, and its second derivative is continuous. Therefore, this kernel is insensitive to disorder and the error in interpolation. However, we should also be mindful of the properties of the first derivative of the kernel: we can think of it as a "potential well" for particles in the neighbour sphere (this is borne out by the forms of the SPH equations of motion derived in the following sections).

Figure 3.1 shows W and dW/dq for the kernel described in equation (3.13). Particles will tend to "collect" in the minimum at $r/h = 2/3$, causing artificial clumping of the particle distribution with this preferred spacing. Problems such as this can be sidestepped by choosing a kernel without such a minimum, e.g. using a quintic spline rather than a cubic spline, although this typically requires a larger number of neighbours in the calculation for stability.

Before we proceed with outlining the equations of SPH, we should look at the derivatives of the kernel itself (see Rosswog 2009). Provided that the kernel is spherically symmetric, i.e.

$$W_{ij} = W(|\mathbf{r}_i - \mathbf{r}_j|) = W(|\mathbf{r}_{ij}|) = W_{ji} \tag{3.14}$$

Then we can derive some useful identities regarding its derivatives, and will save us some work

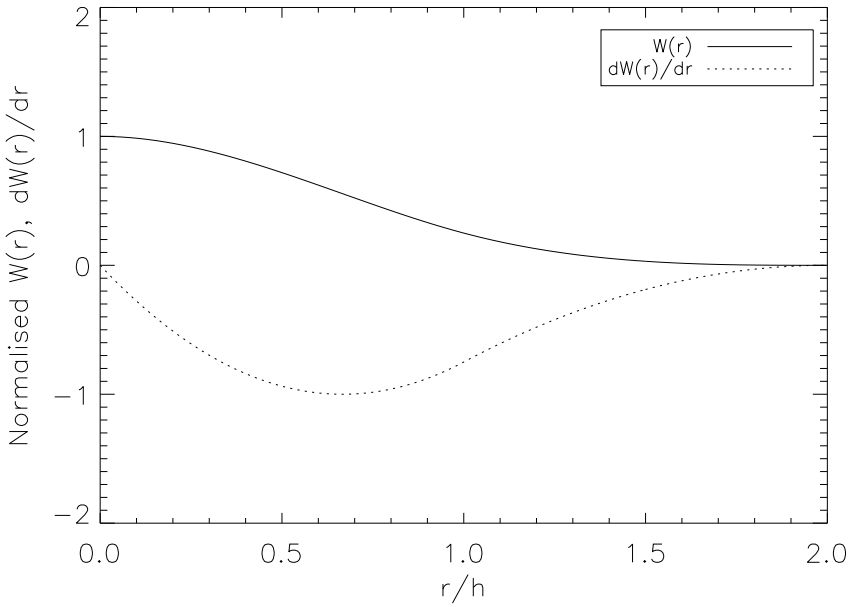

Figure 3.1: The functional form of the cubic spline kernel described in equation 3.13 (solid line), and its first derivative (dotted line). Note the minimum in the derivative at $r/h = 2/3$.

in later derivations[2]. If we begin with the separation vector $\mathbf{r}_{ij}$ - its derivative with respect to $\mathbf{r}_a$ is:

$$\frac{\partial \mathbf{r}_{ij}}{\partial \mathbf{r}_a} = \hat{\mathbf{r}}_{ij}(\delta_{ia} - \delta_{ja}) \tag{3.15}$$

Where we have defined the unit separation vector $\hat{\mathbf{r}}_{\mathbf{ij}}$. The time derivative is (using the chain rule):

$$\frac{d\mathbf{r}_{ij}}{dt} = \frac{\partial \mathbf{r}_{ij}}{\partial x_i}\frac{dx_i}{dt} + \frac{\partial \mathbf{r}_{ij}}{\partial y_i}\frac{dy_i}{dt} + \frac{\partial \mathbf{r}_{ij}}{\partial z_i}\frac{dz_i}{dt} + \frac{\partial \mathbf{r}_{ij}}{\partial x_j}\frac{dx_j}{dt} + \frac{\partial \mathbf{r}_{ij}}{\partial y_j}\frac{dy_j}{dt} + \frac{\partial \mathbf{r}_{ij}}{\partial z_j}\frac{dz_j}{dt} \tag{3.16}$$

This can be better written as

$$\frac{d\mathbf{r}_{ij}}{dt} = \nabla_i \mathbf{r}_{ij}.\mathbf{v}_i + \nabla_j \mathbf{r}_{ij}.\mathbf{v}_j \tag{3.17}$$

We can use the property that $\nabla_j \mathbf{r}_{ij} = -\nabla_i \mathbf{r}_{ij}$ (which follows from the symmetry of the separation vector):

$$\frac{d\mathbf{r}_{ij}}{dt} = \nabla_i \mathbf{r}_{ij}.\mathbf{v}_i - \nabla_i \mathbf{r}_{ij}.\mathbf{v}_j \tag{3.18}$$

Defining the relative velocity vector $\mathbf{v}_{ij} = \mathbf{v}_i - \mathbf{v}_j$, we can arrive at a neat result:

$$\frac{d\mathbf{r}_{ij}}{dt} = \nabla_i \mathbf{r}_{ij}.\mathbf{v}_{ij} = \hat{\mathbf{r}}_{ij}.\mathbf{v}_{ij} \tag{3.19}$$

[2]For $W_{ij} = W_{ji}$, we require that h is unchanged. This is often achieved between particle pairs by using the mean smoothing length: $\bar{h} = \frac{h_i + h_j}{2}$.

We can now apply equations (3.15) and (3.19) to constructing derivatives for the kernel. Beginning with the spatial derivative:

$$\nabla_a W_{ij} = \frac{\partial W_{ij}}{\partial \mathbf{r}_a} = \frac{\partial W_{ij}}{\partial \mathbf{r}_{ij}} \frac{\partial \mathbf{r}_{ij}}{\partial \mathbf{r}_a} = \frac{\partial W_{ij}}{\partial \mathbf{r}_{ij}} \hat{\mathbf{r}}_{ij} (\delta_{ia} - \delta_{ja}) = \nabla_j W_{ij} (\delta_{ia} - \delta_{ja}) \tag{3.20}$$

We can use this result to confirm an important property of the kernel:

$$\nabla_j W_{ij} = \nabla_i W_{ji} (\delta_{ij} - \delta_{jj}) = -\nabla_i W_{ij} \tag{3.21}$$

We shall require this in later sections. Also, we can calculate

$$\frac{dW_{ij}}{dt} = \frac{\partial W_{ij}}{\partial \mathbf{r}_{ij}} \frac{d\mathbf{r}_{ij}}{dt} = \frac{\partial W_{ij}}{\partial \mathbf{r}_{ij}} \hat{\mathbf{r}}_{ij} . \mathbf{v}_{ij} = \mathbf{v}_{ij} . \nabla_i W_{ij} \tag{3.22}$$

3.2.1 Deriving SPH I: Discretising the Equations of Hydrodynamics

I will now describe how we can construct the SPH equations of motion assuming the fundamental equations of hydrodynamics - the continuity equation, the momentum equation and the internal energy equation.

The Continuity Equation

Consider an SPH particle a, and its neighbours, denoted by b. The masses m of all particles are assumed to be constant. The separation vector is given by $\mathbf{r}_{ab}$, the relative velocity is $\mathbf{v}_{ab}$, and $W(\mathbf{r}_{ab}, h) = W_{ab}$. The density at particle a is given by substituting $A = \rho$ in equation (3.7):

$$\rho_a(\mathbf{r}) = \sum_b m_b W_{ab} \tag{3.23}$$

This in effect is the continuity equation of hydrodynamics. We can construct a more analogous representation by calculating the time derivative of both sides (using the kernel derivative dW/dt):

$$\frac{d\rho_a}{dt} = \sum_b m_b \mathbf{v}_{ab} . \nabla_a W_{ab} \tag{3.24}$$

Where ∇_a is the gradient taken at the coordinates of particle a. Note that time derivatives in SPH denote derivatives following the motion (see Appendix A & Glossary). As SPH is a Lagrangian formalism, performing these derivatives at the location of the particles automatically incorporates this.

The Momentum Equation

Euler's Equation (without gravity) is:

$$\frac{d\mathbf{v}}{dt} = -\frac{1}{\rho}\nabla P \tag{3.25}$$

Where P is the pressure. We could approximate the pressure gradient by

$$\rho_a \nabla P_a = \sum_b m_b (P_b - P_a) \nabla_a W_{ab} \tag{3.26}$$

This will vanish when the pressure is constant, as required, but it does not exactly conserve linear and angular momentum. If two isolated particles have slightly different pressures, then the particles will disappear to infinity, showing the inherent instability in this expression. It is better to construct a symmetrised form in this case. We can rewrite $\frac{\nabla P}{\rho}$ using the product rule:

$$\frac{\nabla P}{\rho} = \nabla \left(\frac{P}{\rho} \right) + \frac{P}{\rho^2} \nabla \rho \tag{3.27}$$

These two terms can then be converted into SPH interpolants, giving

$$\frac{d\mathbf{v_a}}{dt} = -\frac{\nabla P_a}{\rho_a} = -\sum_b m_b \left(\frac{P_b}{\rho_b^2} + \frac{P_a}{\rho_a^2} \right) \nabla_a W_{ab} \tag{3.28}$$

This is a symmetrised form as the force generated by this pressure gradient will produce a antisymmetric central force between pairs of particles a and b. Swapping a and b in this expression gives $\frac{d\mathbf{v_b}}{dt} = -\frac{d\mathbf{v_a}}{dt}$, guaranteeing the conservation of linear and angular momentum. This guaranteed conservation of momentum is one of the greatest strengths of SPH.

The Internal Energy Equation

Excluding radiative processes, the rate of change of internal energy per unit mass is

$$\frac{du}{dt} = -\left(\frac{P}{\rho} \right) \nabla . \mathbf{v} \tag{3.29}$$

In the simplest form, this gives

$$\frac{du_a}{dt} = \left(\frac{P_a}{\rho_a^2} \right) \sum_b m_b \mathbf{v}_{ab} . \nabla_a W_{ab} \tag{3.30}$$

Producing a symmetrised form takes more work here. Rewriting the right hand side of the internal energy equation (again using the product rule) gives

$$\frac{du}{dt} = -\nabla \left(\frac{P\mathbf{v}}{\rho} \right) + \mathbf{v} . \nabla \left(\frac{P}{\rho} \right) \tag{3.31}$$

This gives a second version of the SPH internal energy equation:

$$\frac{du_a}{dt} = \sum_b m_b \left(\frac{P_b}{\rho_b^2}\right) \mathbf{v}_{ab}.\nabla_a W_{ab} \tag{3.32}$$

A symmetrised form can be found by averaging equations (3.30) and (3.32):

$$\frac{du_a}{dt} = \frac{1}{2}\sum_b m_b \left(\frac{P_b}{\rho_b^2} + \frac{P_a}{\rho_a^2}\right) \mathbf{v}_{ab}.\nabla_a W_{ab} \tag{3.33}$$

3.2.2 Shocks and Artificial Viscosity

SPH implicitly assumes that the properties of the fluid it simulates varies smoothly on the smallest resolvable scale (the local smoothing length). However, the evolution of most astrophysical fluids will guarantee that this assumption will eventually become invalid. If any region of the fluid undergoes a shock, then the shock front ensures that the fluid is discontinuous on scales smaller than the smoothing length, and therefore will not be resolved. There are two options to combat this problem:

1. If the smoothing length can be reduced sufficiently, then the discontinuity across the shock front could be resolved.

2. If the shock front can be smoothed out over a region larger than h, then the simulation will be able to resolve it.

Option 1 is in general impractical, so we are forced to consider option 2. The most straightforward approach to smooth shock fronts is to introduce an artificial viscosity (AV) (Von Neumann & Richtmyer, 1950). This is often considered outdated by many finite difference simulators, who correctly note that adding extra viscosity will also add extra dissipative heating to the simulation. Other methods of handling the shock front are possible, for example using Godunov schemes which use the exact solution to the Riemann problem to limit the pre- and post-shock variables (e.g. Cha & Whitworth 2003). While these methods remove the need for artificial viscosity and excess dissipation, they are limited in their ability to handle the introduction of new physics, such as a more advanced equation of state than the ideal gas law (Price, 2005).

The most common formulation of the artificial viscosity can be found in Monaghan (1992), and is used in this thesis. The momentum equation is thusly modified:

$$\frac{d\mathbf{v_a}}{dt} = -\sum_b m_b \left(\frac{P_b}{\rho_b^2} + \frac{P_a}{\rho_a^2} + \Pi_{ab}\right) \nabla_a W_{ab} \tag{3.34}$$

with

$$\Pi_{ab} = \begin{cases} \frac{-\alpha \bar{c}_{ab}\mu_{ab} + \beta\mu_{ab}^2}{\bar{\rho}_{ab}} & \mathbf{v}_{ab}.\mathbf{r}_{ab} \leq 0 \\ 0 & \mathbf{v}_{ab}.\mathbf{r}_{ab} > 0 \end{cases} \tag{3.35}$$

and

$$\mu_{ab} = \frac{h\mathbf{v}_{ab}.\mathbf{r}_{ab}}{r_{ab}^2 + \eta^2}, \tag{3.36}$$

where $\bar{c}_{ab}$ is the average sound speed and $\bar{\rho}_{ab}$ is the average density. We must similarly modify the internal energy equation:

$$\frac{du_a}{dt} = \frac{1}{2}\sum_b m_b \left(\frac{P_b}{\rho_b^2} + \frac{P_a}{\rho_a^2} + \frac{1}{2}\Pi_{ab}\right)\mathbf{v}_{ab}.\nabla_a W_{ab} \tag{3.37}$$

The viscosity is inactive for diverging elements. The linear μ term produces a shear and bulk viscosity, and is used for low Mach number shocks. The quadratic term handles shocks which are highly supersonic, and is most like the Von Neumann-Richtmyer viscosity which finite difference simulators would find familiar. α and β are pre-specified constants, usually $\alpha = 1$, and $\beta = 2$ in astrophysical situations (Monaghan, 1992), although depending on particle resolution other values may be used (however it is recommended that the ratio of the two should remain constant: $\beta = 2\alpha$).

There have been a variety of suggestions for improved viscosity switches in SPH - while not used here, I will list some for the sake of interest. Balsara (1989) outlines a possible switch using the "Balsara factor", calculated for each SPH particle

$$B_i = \frac{|\nabla.\mathbf{v}|_i}{|\nabla \times \mathbf{v}|_i + |\nabla.\mathbf{v}|_i} \tag{3.38}$$

Which then modifies Π:

$$\Pi_{ab} \rightarrow \frac{B_a + B_b}{2}\Pi_{ab} \tag{3.39}$$

The upshot is that in an exactly Keplerian disc

$$\mathbf{v} = (v_r, v_\phi, v_z) = (0, \sqrt{GM/r}, 0) \tag{3.40}$$

The velocity divergence is exactly zero, giving $B_i = 0$, switching off AV except for true convergent flows. Another option is to use the time dependent viscosity (TDV) of Morris and Monaghan (1997), where the α parameter for each particle evolves with time according to

$$\frac{d\alpha_i}{dt} = -\frac{\alpha_i - \alpha^*}{\tau_i} + S_i \tag{3.41}$$

Where the e-folding time τ_i is

$$\tau_i = \frac{h_i}{C_1 c_{s,i}} \tag{3.42}$$

and the source term S_i is

$$S_i = \max(-(\nabla.\mathbf{v})_i, 0) \tag{3.43}$$

When the flow begins to converge, α_i increases. Once the convergence has ended, α_i decays back to a steady state value α^* (with a decay timescale controlled by the tunable parameter

C_1, which Morris and Monaghan suggest be set at 0.2, which keeps the viscosity at a minimum value at a distance of 5 smoothing lengths from the shock).

It must be noted that these switches have their faults also. Cartwright & Stamatellos (2010) show that even in Keplerian flow, the Balsara switch is non-zero due to Poisson noise in the SPH system, preventing positive and negative contributions to the divergence from cancelling to zero. They also show that the flow produces artificial alignments that prove problematic for the TDV formalism. By proposing a switch which directly tests for non-Keplerian flow around an SPH particle, Cartwright and Stamatellos were able to reduce the magnitude of AV significantly. As Keplerian velocities are perpendicular to their radius, we expect

$$\mathbf{v}.\mathbf{r} = 0 \tag{3.44}$$

This is a simple test that can be carried out not only for the SPH particle itself, but for its neighbours also. If the scalar product of a particle exceeds some tolerance ϵ, then it is considered to be "anomalous". The proportion of anomalous particles to the total number of neighbours gives a viscosity switch akin to the Balsara switch. While this only detects non-Keplerian flow, the divergence condition in the traditional AV handles the detection of convergence, so the combination of both works well in Keplerian discs. We should however note that self-gravitating discs (which deviate from Keplerian flow as the disc mass to star mass ratio is increased) will find this Keplerian switch less beneficial.

3.2.3 The Smoothing Length

Until now, I have largely avoided discussing the role of the smoothing length in the SPH formalism. This is in part due to some subtleties as to its usage, which deserve their own section to discuss. The first such issue arises when attributing a value to the smoothing length. We are presented with two possibilities:

1. The "gather" method. A smoothing length is assigned to a location in space, not to a particle. The location's smoothing length is such that a fixed number of neighbours to it are enclosed in the smoothing volume (see left panel of Figure 3.2).

2. The "scatter" method. Smoothing lengths are assigned to individual particles. At any location, the density is computed from all the particles which have smoothing volumes that encompass the location (see right panel of Figure 3.2).

The second issue arises when discussing how the smoothing length should evolve over time. Originally, computational power limited SPH calculations, forcing a global smoothing length to be specified for all particles. This fixes the maximum resolution of the simulation - in cases where the density will increase by several orders of magnitude (such as molecular cloud collapse), this is very unsatisfactory - fixing a low initial resolution precludes the correct evolution of the

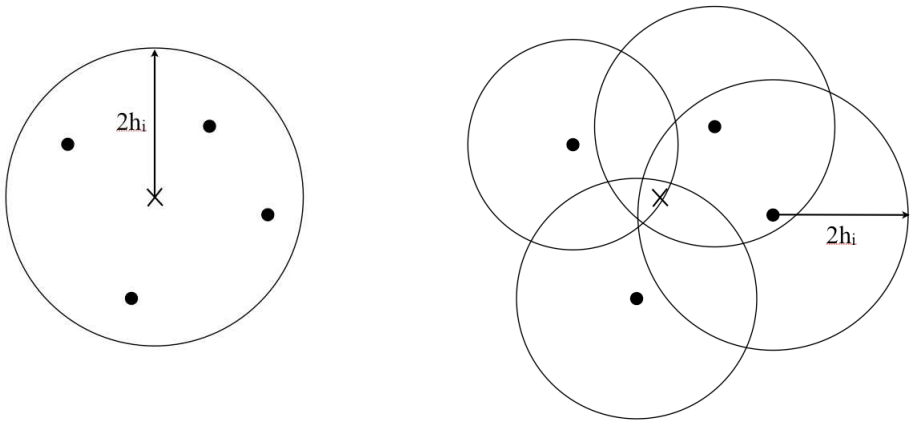

Figure 3.2: Depictions of the "gather" and "scatter" interpretations (left and right panels respectively). Figures adapted from Hernquist & Katz (1989).

cloud collapse after a short period of time; fixing a high initial resolution requires many more particles, and will be much slower than necessary.

It is better instead to allow the local smoothing length of each particle to change according to the local density:

$$h_a = \eta \left(\frac{m_a}{\rho_a} \right)^{1/3} \tag{3.45}$$

where $\eta \sim 1.2 - 1.5$ (Rosswog 2009 and references within). By reducing the smoothing length in overdense regions, the effective resolution is boosted. This is the true adaptive power of SPH - when this is combined with SPH's Lagrangian nature (ensuring it can simulate advecting non-trivial flows with ease), it is clear why it has become a popular choice for astrophysical simulators.

However, allowing the smoothing length to evolve has its price. The equations of motion derived above have assumed that h is constant. Re-deriving the equations with this assumption relaxed produces new terms - the so-called "grad-h" terms (Nelson & Papaloizou, 1994; Rosswog, 2009). The impact of these terms is not intuitively clear, and is typically a function of the problem being studied (and the resolution used). Note that the code used in this thesis does not account for these terms.

3.2.4 Deriving SPH II: A Variational Formulation

We have seen how SPH can be developed by a direct discretisation of the equations of hydrodynamics. However, we must also note that this approach requires extra "guidance" to select the appropriate symmetrised forms to ensure the conservation of momentum, etc. For completeness I will also show how we can derive a self-consistent, conservative SPH from a Lagrangian, plus the first law of thermodynamics, and the SPH summation prescription given in equation (3.7).

For simplicity, we will assume gravity to be absent, and we will neglect the "grad-h" terms (the interested reader can find a derivation including the grad-h terms in Rosswog 2009).

We shall begin with the Lagrangian for a perfect fluid:

$$\mathcal{L} = \int \left(\frac{1}{2}\rho v^2 - u(\rho, s) \right) dV \tag{3.46}$$

Where ρ is the density, $\mathbf{v}$ is the fluid velocity, u is the internal energy and s is the entropy. We use the SPH formalism to obtain

$$\mathcal{L} = \sum_b m_b \left(\frac{1}{2} v_b^2 - u_b \right) \tag{3.47}$$

This must satisfy the Euler-Lagrange Equations:

$$\frac{d}{dt}\left(\frac{\partial \mathcal{L}}{\partial \mathbf{v}_a} \right) - \frac{\partial \mathcal{L}}{\partial \mathbf{r}_a} = 0, \tag{3.48}$$

where the derivatives are taken with respect to particle a. Considering the first term:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{v}_a} = \frac{\partial}{\partial \mathbf{v}_a}\left[\sum_b m_b \left(\frac{1}{2} v_b^2 - u_b \right) \right] = m_a \mathbf{v}_a \tag{3.49}$$

And therefore its time derivative gives the force on particle a:

$$\frac{d}{dt}\left(\frac{\partial \mathcal{L}}{\partial \mathbf{v}_a} \right) = m_a \frac{d\mathbf{v}_a}{dt} \tag{3.50}$$

The second term becomes

$$\frac{\partial \mathcal{L}}{\partial \mathbf{r}_a} = \frac{\partial}{\partial \mathbf{r}_a}\left[\sum_b m_b \left(\frac{1}{2} v_b^2 - u_b \right) \right] = -\sum_b m_b \left(\frac{\partial u_b}{\partial \rho_b} \right)_s \frac{\partial \rho_b}{\partial \mathbf{r}_a} \tag{3.51}$$

To continue this derivation, we must use the First Law of Thermodynamics (for quantities per unit mass, assuming $dQ = 0$):

$$du = \frac{P}{\rho^2} d\rho \tag{3.52}$$

We can simplify equation (3.51):

$$\frac{\partial \mathcal{L}}{\partial \mathbf{r}_a} = -\sum_b m_b \left(\frac{P_b}{\rho_b^2} \right) \frac{\partial \rho_b}{\partial \mathbf{r}_a} \tag{3.53}$$

We now use equation (3.23) to evaluate $\frac{\partial \rho_b}{\partial \mathbf{r}_a}$ (cf. Price 2005):

$$\frac{\partial \rho_b}{\partial \mathbf{r}_a} = \sum_c m_c \nabla_a W_{bc} \tag{3.54}$$

We will have to utilise the properties of the kernel to their fullest if we are to proceed. From the results of the previous section,

$$\nabla_a W_{bc} = \nabla_b W_{cb}(\delta_{ba} - \delta_{ca}) \tag{3.55}$$

This then substitutes to give

$$\frac{\partial \mathcal{L}}{\partial \mathbf{r}_a} = -\sum_b m_b \left(\frac{P_b}{\rho_b^2}\right) \sum_c m_c \nabla_b W_{cb}(\delta_{ba} - \delta_{ca}) \tag{3.56}$$

Using the Kronecker δ to sift this expression produces

$$\frac{\partial \mathcal{L}}{\partial \mathbf{r}_a} = -m_a \left(\frac{P_a}{\rho_a^2}\right) \sum_c m_c \nabla_a W_{ca} - \sum_b m_b \left(\frac{P_b}{\rho_b^2}\right) m_a \nabla_b W_{ab} \tag{3.57}$$

Re-labelling the summation over c (and noting that $\nabla_a W_{ba} = \nabla_a W_{ab}$) produces

$$\frac{\partial \mathcal{L}}{\partial \mathbf{r}_a} = -m_a \left(\frac{P_a}{\rho_a^2}\right) \sum_b m_b \nabla_a W_{ab} - \sum_b m_b \left(\frac{P_b}{\rho_b^2}\right) m_a \nabla_b W_{ab} \tag{3.58}$$

Finally, we use $\nabla_b W_{ab} = -\nabla_a W_{ab}$ to give

$$\frac{\partial \mathcal{L}}{\partial \mathbf{r}_a} = \frac{d}{dt}\left(\frac{\partial \mathcal{L}}{\partial \mathbf{v}_a}\right) = m_a \frac{d\mathbf{v}_a}{dt} = -m_a \sum_b \left(\frac{P_a}{\rho_a^2} + \frac{P_b}{\rho_b^2}\right) m_b \nabla_a W_{ab} \tag{3.59}$$

Which is precisely the momentum equation we have been looking for. We can derive the internal energy equation using the First Law of Thermodynamics:

$$\frac{du}{dt} = \frac{P}{\rho^2}\frac{d\rho}{dt} \tag{3.60}$$

Where the time derivative of ρ is:

$$\frac{d\rho}{dt} = \sum_b m_b \frac{dW_{ab}}{dt} \tag{3.61}$$

We can return to our results regarding kernel derivatives to obtain

$$\frac{d\rho}{dt} = \sum_b m_b \mathbf{v}_{ab}.\nabla_a W_{ab} \tag{3.62}$$

Giving the internal energy equation as

$$\frac{du}{dt} = \frac{P_a}{\rho_a^2} \sum_b m_b \mathbf{v}_{ab}.\nabla_a W_{ab} \tag{3.63}$$

3.2.5 Timesteps

All hydrodynamic simulations have characteristic timescales on which the fluid evolves - the dynamical timescale, the viscous timescale, etc. To ensure the fluid behaves correctly, the *timestep* (the minimum time interval on which the equations of motion are integrated) must be smaller than the smallest timescale of interest to correctly capture the appropriate physics.

For hydrodynamic fluids in the absence of external forces, this is referred to the *Courant-Friedrich-Lewys* (CFL) condition (Courant et al., 1928). Similar conditions also exist in the presence of external forces such as gravity. Numerically, the combined CFL condition for SPH is (Monaghan, 1989, 1992):

$$\delta t = \vartheta \min\left(\delta t_f, \delta t_{cv}\right) \tag{3.64}$$

Where δt_f describes the time constraints given by the local force per unit mass $\mathbf{f}$:

$$\delta t_f = \min_a \left(\frac{h_a}{\mathbf{f}_a}\right)^{1/2} \tag{3.65}$$

and δt_{cv} folds in both the "standard" CFL Condition and the constraints posed by the artificial viscosity:

$$\delta t_{cv} = \min_a \frac{h_a}{c_a + 0.6(\alpha c_a + \beta \max(\mu_{ab}))} \tag{3.66}$$

Where α and β are the artificial viscosity parameters described in section 3.2.2, c is the local sound speed, and μ_{ab} is defined in equation (3.36). Typically, the normalisation constant $\vartheta \sim 0.25$.

The efficiency of the code can be enhanced by allowing particles to have individual timesteps (Ewell, 1988; Hernquist & Katz, 1989). In this scenario, the key requirement is that the particle's timesteps synchronise regularly throughout the evolution. This is done by restricting how the timesteps can vary. An initial maximum timestep t_{max} is chosen for all particles. For a particle i, any subsequent timestep is selected such that

$$t_i = \frac{t_{max}}{2^{n_i}} \tag{3.67}$$

Where the *time bin* $n_i \geq 0$. In this fashion, particles with $n_i = 1$ will synchronise with the main timestep t_{max} every 2 t_i timesteps. The system will therefore occupy a range of time bins from $n_i = 0$ to some maximum value $n_i = n_{max}$. The system must be evolved from this maximum occupied time bin, i.e. the minimum time step, given by

$$t_{min} = \frac{t_{max}}{2^{n_{max}}} \tag{3.68}$$

Particles in the maximum bin will be evolved at every timestep t_{min}. As the maximum bin synchronises with the lower bins, the particles in these bins will also be evolved. The procedure therefore ensures that particles that do not require a strict timestep do not receive one, and are updated less frequently than those which do.

3.2.6 Gravity in SPH

Divergent Forces and Kernel Softening

The incorporation of long-range forces such as gravity into SPH are straightforward to implement, and are fully Lagrangian in nature. Simple particle-particle equations can be used to calculate the gravitational force from Newton's Law of Gravitation. For particle i,

$$F_i = \sum_{j=1}^{N} \frac{G m_i m_j}{|\mathbf{r}_i - \mathbf{r}_j|^2} \tag{3.69}$$

However, the gravitational force diverges if the particle separation is small. N-Body codes have the same problem, and usually posit a *softening length* ϵ:

$$F_i = \sum_{j=1}^{N} \frac{G m_i m_j}{|\mathbf{r}_i - \mathbf{r}_j|^2 + \epsilon^2} \tag{3.70}$$

This limits the maximum gravitational force the two particles can exert on each other. A similar approach can be used for SPH (and is in fact used for pointmasses, see 3.2.7), but a more elegant solution involves the smoothing kernel. Consider two particles interacting by gravity, and particle j is within particle i's smoothing volume. We can consider the mass of particle i distributed in its volume with a density profile set by the kernel:

$$\rho(\mathbf{r}) = m_i W(|\mathbf{r} - \mathbf{r}_i|, h_i) \tag{3.71}$$

The mass of particle i enclosed within a radius r therefore is

$$m_i(< r) = 4\pi \int_0^r r'^2 \rho(r') dr' = 4\pi m_i \int_0^r r'^2 W(r' - r_i, h_i) dr' \tag{3.72}$$

If we replace m_i with $m_i(< r_j)$ in equation (3.69), then only the mass of particle i that lies between $\mathbf{r_i}$ and $\mathbf{r}_j$ will contribute to the gravitational force. When $\mathbf{r}_i = \mathbf{r}_j$, $m_i(< r_j) = 0$ and the gravitational force vanishes, removing the unwanted divergence at $|\mathbf{r}_i - \mathbf{r}_j| = 0$. If the particles do not overlap (i.e. $\mathbf{r}_{ij} > 2h_i$), then $m_i(< r_j) = m_i$ and we recover the original form of equation (3.69).

The Use of Tree Structures

Calculating gravitational forces using the simple particle-particle approach is an $O(N^2)$ calculation, and hence can become very computationally expensive when N becomes large. Most modern SPH codes use *tree structures* to optimise this calculation[3]. Tree structures organise the particles into hierarchical cells. When the gravitational force on particle i is calculated, the influence of nearby particles are calculated as above; more distant particles are approximated by calculating the gravitational force due to the appropriate hierarchical cell that the particles occupy. Which level in the hierarchy is used is decided

[3]Tree structures can also be used for neighbour finding algorithms, providing a twofold boost in efficiency.

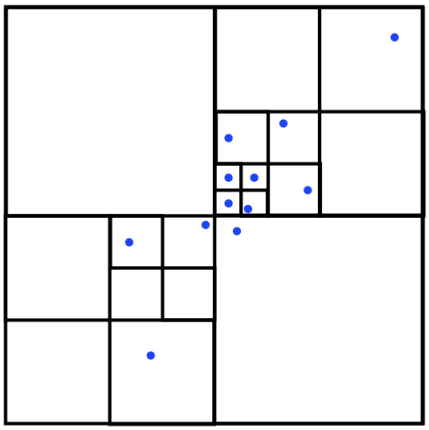

Figure 3.3: An example of a quadtree (the two dimensional equivalent of the octree). In this schematic, the tree cells subdivide into four children cells until no more than one particle is in a given cell.

according to some criterion of tolerance. By switching the majority of the force calculation from particle-particle interactions to particle-cell (multipole) interactions, the process becomes an $O(N \log N)$ process. As Hernquist & Katz (1989) note, this approximation is well-motivated:

"...the dynamics of the Earth-Moon system is relatively insensitive to the detailed mass distribution of the two bodies... [numerical] errors will always be present in particle simulations from round off, truncation and discreteness effects. Given these, it is not necessary to compute the force-field to arbitrarily high *precision."*

There are many varieties of tree structure that can be used. As an example, I will discuss the *octree* favoured by Barnes & Hut (1986) (as it will become important in Chapter 6). The entire simulation is placed in a box aligned with the Cartesian axes (this is known as the *root cell*). This root cell is then subdivided into eight children cells. If a cell contains a sufficient number of particles, the cell is then subdivided into a further eight grandchildren cells. The subdivision process continues until the particle occupancy of a cell is less than a specified value. These end cells are often referred to as *leaf cells*. This is an effective spatial indexing tool - subdivision only occurs where it is required (see Figure 3.3).
The gravitational force on a single particle using the octree is calculated thus:

1. "Walk" through the tree, beginning at the root cell and descending through the children.
2. At each level in the hierarchy, the cells at that level have some size s. If the distance from the particle to the cell is d, then we check if

$$\frac{s}{d} \leq \theta, \tag{3.73}$$

where θ is our tolerance criterion. If the above inequality is true, then the gravitational force from the particles in that cell are calculated as a single particle-cell interaction.

3. If the above inequality is false, then we continue descending down the tree. If we reach the bottom of a branch, and find a leaf cell for which this inequality is false, then we calculate the individual particle-particle forces for all particles in the leaf.

When calculating a cell's contribution to the gravitational force, the cell (with side s and mass M calculated from the sum of the occupying particles $M = \sum_i m_i$) is presumed to have a uniform density ρ_0:

$$M = \rho_0 s^3 \tag{3.74}$$

A mass element in the cube has position $\tilde{\mathbf{r}} = (\tilde{x}, \tilde{y}, \tilde{z})$ and mass $\rho_0 \, d\tilde{x} \, d\tilde{y} \, d\tilde{z}$. For any position outside the cube (denoted by $\mathbf{r} = (x, y, z)$), the potential contribution is therefore

$$\Phi(\mathbf{r}) = G\rho_0 \int_{-s/2}^{s/2} d\tilde{x} \int_{-s/2}^{s/2} d\tilde{y} \int_{-s/2}^{s/2} d\tilde{z} \frac{1}{|\mathbf{r} - \tilde{\mathbf{r}}|} \tag{3.75}$$

The inverse separation of $\mathbf{r}$ and $\tilde{\mathbf{r}}$ is typically approximated using the Legendre Polynomials $P_l(\mu)$ (cf. Jackson 1975), where

$$\mu = \cos\gamma = \frac{\mathbf{r}.\tilde{\mathbf{r}}}{r\tilde{r}} \tag{3.76}$$

and γ is the angle between the vectors $\mathbf{r}$ and $\tilde{\mathbf{r}}$. The expansion is given by

$$\frac{1}{|\mathbf{r} - \tilde{\mathbf{r}}|} = \frac{1}{r}\left[P_0\mu + \frac{\tilde{r}}{r}P_1(\mu) + \left(\frac{\tilde{r}}{r}\right)^2 P_2(\mu) + ..\right] \tag{3.77}$$

I list the first five Legendre Polynomials required here:

$$\begin{aligned} P_0(\mu) &= 1 \\ P_1(\mu) &= \mu \\ P_2(\mu) &= 3/2(\mu^2 - 1/3) \\ P_3(\mu) &= 5/2(\mu^3 - 3/5\mu) \\ P_4(\mu) &= 35/8(\mu^4 - 6/7\mu^2 + 3/35) \end{aligned} \tag{3.78}$$

The first term in equation (3.77) gives the first contribution to the potential, Φ_0:

$$\Phi_0(\mathbf{r}) = G\rho_0 \int_{-s/2}^{s/2} d\tilde{x} \int_{-s/2}^{s/2} d\tilde{y} \int_{-s/2}^{s/2} d\tilde{z} \, \frac{1}{r} = \frac{G\rho_0 s^3}{r} = \frac{GM}{r} \tag{3.79}$$

which is the intuitive first approximation (and the dominant term for separations much larger than the cell size $r \gg s$). The next term (originating from P_1) gives

$$\Phi_1(\mathbf{r}) = \frac{G\rho_0}{r^3} \int_{-s/2}^{s/2} d\tilde{x} \int_{-s/2}^{s/2} d\tilde{y} \int_{-s/2}^{s/2} d\tilde{z} \, \mathbf{r}.\tilde{\mathbf{r}} \tag{3.80}$$

Expanding the scalar product gives

$$\mathbf{r}.\tilde{\mathbf{r}} = x\tilde{x} + y\tilde{y} + z\tilde{z} \tag{3.81}$$

The integral is therefore split into three terms. Considering the first term:

$$\int_{-s/2}^{s/2}\int_{-s/2}^{s/2}\int_{-s/2}^{s/2} x\tilde{x}\, d\tilde{x}\, d\tilde{y}\, d\tilde{z} \tag{3.82}$$

Performing the integral over $d\tilde{x}$ will give zero, as the integration limits are symmetric about zero, and the integrand is odd. The corresponding integrals for $d\tilde{y}$ and $d\tilde{z}$ are also zero, therefore

$$\Phi_1 = 0 \tag{3.83}$$

In a similar vein, it can be shown that $\Phi_2 = \Phi_3 = 0$, and that the next correction term is the quadropole moment

$$\Phi_4 = \frac{7}{960}\frac{Ms^4}{r^9}\left[3r^4 - 5(x^4 + y^4 + z^4)\right] \tag{3.84}$$

The interested reader is directed to the appendices of Barnes & Hut (1989) for a full discussion of these calculations, and the higher order corrections to Φ.

3.2.7 Pointmass Creation

Despite SPH's Lagrangian nature, and its ability to use individual timesteps, simulations where dense compact objects form (e.g. protostars in a star formation region) will inevitably have to be halted due to the increasingly high densities reached. If the simulation uses an isothermal equation of state, then a switch can be put in place to change the equation of state to an "optically thick" variant, slowing the collapse (the polytropic equation of state is a good example, Bonnell & Bate 1994). However, if the simulation uses a more complex equation of state (such as the simulations I will discuss in this thesis, see section 3.4.5) then this technique becomes unfeasible. Imposing a minimum smoothing length on the simulation can also help - this also amounts to changing the equation of state, in a more implicit sense. Again, with complex equations of state in use, this becomes difficult to implement.

Bate et al. (1995) propose a solution which avoids these difficulties, which was previously implemented in finite difference codes (e.g. Boss & Black 1982) - the creation of *pointmasses*. These are a special kind of particle which are introduced into SPH simulations at high density regions of bound gas, removing the particles with extremely small timesteps and hence removing the halting problem. The pointmass represents a bound object such as a protostar or protoplanet, and is endowed with the same mass, spin[4] and angular momentum as the SPH particles it replaces. The pointmass can then accrete SPH particles from its surroundings if the

[4]Note that the pointmass spin does not typically contribute to the SPH simulation - it is recorded to test that the global angular momentum is conserved.

SPH particle passes within its *accretion radius* (specified before the simulation begins), and if the angular momentum of the SPH particle about the pointmass is low enough that it cannot form a circular orbit at the accretion radius.

For dynamic creation of pointmasses, care must be taken to ensure that a dense region is bound, and will continue to collapse. Once a particle surpasses a preset density (say 10^5 times the original mean density), then a series of tests begin. The particle is firstly tested to ensure that ~ 50 neighbours exist within one accretion radius. If this is true, then the particle and its neighbours are tested as an ensemble. The ratio of thermal to gravitational energy α is calculated, along with the ratio of rotational to gravitational energy β. The particles can only become a pointmass if:

1. $\alpha \leq 1/2$
2. $\alpha + \beta \leq 1$
3. The total energy of the ensemble is negative (to check the ensemble is bound)
4. The divergence of the particle accelerations is less than zero (to check the ensemble is collapsing).

If any of the above criteria are not satisfied, a pointmass will not be created. If all are satisfied, the pointmass is created, with the ensemble's total mass, spin and angular momentum. The SPH particles are then removed, and play no further part in the simulation.

The presence of a pointmass in the simulation requires several corrections to the calculation of fluid variables such as pressure. For example, particles which are separated by a pointmass will not experience pressure forces from each other. Also, artificial shear viscosity can be enhanced near the accretion radius due to the lack of neighbours inside the radius. These problems are addressed in Bate et al. (1995), and the interested reader is referred to them for further detail.

3.3 Radiative Transfer in SPH

We come now to the meat of this chapter, and to the problem of including radiative effects in SPH. As we saw in section 2.4, the radiative transfer problem is multidimensional, with a variety of interdependent processes interacting on very short timescales. Fluid elements in a radiation field can either emit, absorb or scatter. Even the apparently simple problem of calculating emission at a position $\mathbf{r}$ (which depends *prima facie* on local thermodynamic variables only) can rapidly become obfuscated if we consider stimulated emission as a result of radiation incident from some location $\mathbf{r}'$. This is clearly a non-local phenomenon, which increases the distance between causally connected parts of a simulation by several orders of magnitude (in comparison to the non-radiative case). Scattering and absorption are also in general non-local processes. The construction of SPH is such that the distance between causally connected parts in any

simulation is of order the smoothing length (for one timestep). Implementing full radiative transfer in SPH is currently impossible (although future computing power may eventually allow it). We are then faced with a compromise. In attempting to model the radiation field, we must balance the two competing requirements of any simulation:

1. The simulation must be sufficiently accurate.
2. The simulation must run at sufficient speed to produce useful results (for a given particle number of interest).

The inclusion of radiative transfer in SPH has been the subject of much study over the years, beginning with the very first SPH codes (Lucy, 1977). The first approximation to radiative cooling is to add a term to the internal energy equation of the form

$$\dot{u}_{rad} = -\frac{u}{t_{cool}} \tag{3.85}$$

Where the calculation of the cooling time t_{cool} varies. The simplest is

$$t_{cool} = \beta \tag{3.86}$$

where β is some constant. More sophisticated versions (especially those used for disc simulations) have related the cooling time to the local dynamical timescale:

$$t_{cool} = \beta\Omega^{-1} \tag{3.87}$$

These versions are particularly popular for controlled numerical experiments in grid-based codes as well as SPH, as they explicitly control the relationship between the local radiative timescale and the local dynamical timescale, which is important for studies of gravitational instability (Gammie, 2001; Rice et al., 2003; Lodato & Rice, 2004; Pickett et al., 2003; Cossins et al., 2009). The specification of β gives a high level of control over the simulation, allowing the experimenter to derive useful criteria that can then be tested in more "realistic" simulations. An example of such a realistic simulation gives a non-trivial form for the cooling time:

$$t_{cool} = t_{cool}(T, \tau), \tag{3.88}$$

which requires a more complete description of the radiative processes active in the simulation. Despite the advantages of the cooling time prescription, it can only deal with energy loss by local emission of radiation, and it does not allow radiation to be exchanged between neighbouring particles. An alternative solution has been to implement *flux-limited diffusion* (FLD), a frequency averaged solution to RHD where the optically thick radiation field is modelled in LTE as a simple thermal conduction process (the details of which are explained in section 3.4.2). This is a common solution with an excellent pedigree in SPH - its first use was by Lucy (1977) in the earliest implementation of the algorithm. However, FLD alone in SPH is incapable of modelling energy loss. We will discuss this problem in more detail in later sections.

3.4 A Hybrid Method of Radiative Transfer

I will now begin to elaborate my efforts in constructing a radiative transfer algorithm for SPH. I draw on the two classes of radiative model described above to synthesise a hybrid method which incorporates the strengths of both. We will begin with the description of the radiative cooling.

3.4.1 The Polytropic Cooling Approximation

This approximation (originally conceived by Stamatellos et al. 2007b) uses an SPH particle's density ρ_i, temperature T_i, and gravitational potential ψ_i to estimate a mean optical depth for the particle. The approximation is achieved as follows: assume the particle is embedded in a spherically symmetric polytropic "pseudocloud". The properties of the cloud are calculable analytically (using the Lane Emden equation), given the particle's (dimensionless) radius ξ from the centre: $R = \xi R_0$. Therefore, by appropriate selection of the central values of density and temperature ρ_c, T_c, the particle's own values can be recovered:

$$\rho_i = \rho_c \theta^n(\xi) \tag{3.89}$$

$$T_i = T_c \theta(\xi) \tag{3.90}$$

$$\psi_i = -4\pi G \rho_c R_0^2 \phi(\xi) \tag{3.91}$$

where θ is the solution to the Lane-Emden equation for a polytrope of index n, and

$$\phi(\xi) = -\xi_B \frac{d\theta}{d\xi}(\xi_B) + \theta(\xi) \tag{3.92}$$

(where ξ_B is the boundary of the polytrope) and R_0 satisfies

$$R_0 = \left[\frac{-\psi_i \theta^n(\xi)}{4\pi G \rho_i \phi(\xi)} \right]^{1/2}. \tag{3.93}$$

This provides the tools to calculate a column density from any given (dimensionless) radius to the boundary of the cloud:

$$\Sigma_i(\xi) = \int_{\xi'=\xi}^{\xi'=\xi_B} \rho_c \theta^n(\xi') R_0 \, d\xi' \tag{3.94}$$

$$\Sigma_i(\xi) = \left[\frac{-\psi_i \rho_i}{4\pi G \phi(\xi) \theta^n(\xi)} \right]^{1/2} \int_{\xi'=\xi}^{\xi'=\xi_B} \theta^n(\xi') \, d\xi' \tag{3.95}$$

However, it is assumed that the value of ξ for the particle is unknown. Instead, a value for the column density is arrived at by performing a *mass weighted average* over all possible values of ξ up to the polytrope's boundary:

$$\overline{\Sigma}_i = \left[-\xi_B^2 \frac{d\theta}{d\xi}(\xi_B) \right]^{-1} \int_{\xi'=0}^{\xi'=\xi_B} \Sigma_i(\xi') \, \theta^n(\xi') \xi'^2 \, d\xi' \tag{3.96}$$

The total (dimensionless) mass of the polytrope is $\left[-\xi_B^2 \frac{d\theta}{d\xi}(\xi_B)\right]$, and $\theta^n(\xi)\xi^2 d\xi$ is the dimensionless mass element between $[\xi, \xi + d\xi]$. In real terms, $\overline{\Sigma}_i$ becomes a simple algebraic quantity

$$\overline{\Sigma}_i = \zeta_n \left[\frac{-\psi_i \rho_i}{4\pi G}\right]^{1/2} \tag{3.97}$$

with the integral folded into the constant ζ_n, which is dependent only on the polytropic index n:

$$\zeta_n = \left[-\xi_B^2 \frac{d\theta}{d\xi}(\xi_B)\right]^{-1} \int_{\xi=0}^{\xi=\xi_B} \int_{\xi'=\xi}^{\xi'=\xi_B} \theta^n(\xi') d\xi' \left[\frac{\theta^n(\xi)}{\phi(\xi)}\right]^{1/2} \xi^2 d\xi. \tag{3.98}$$

Stamatellos et al. (2007b) show that this constant is insensitive to the value of n. They select $n = 2$ for their work, as this would give a polytropic exponent of 3/2, in keeping with polytropic exponents of protostars in quasistatic equilibrium. For this thesis, we will assume $n = 2$, (and hence $\zeta_2 = 0.368$) except where otherwise stated. The simple expression for the column density illustrates its ability to capture the effects of the local environment (through the presence of ρ) and the effects of the system's global geometry (through the gravitational potential ψ).

In the same vein, a mass weighted optical depth can be calculated. The optical depth from any radius to the edge of the pseudocloud is

$$\tau_i(\xi) = \int_{\xi'=\xi}^{\xi'=\xi_B} \kappa_i \left(\rho_c \theta^n(\xi'), T_c \theta(\xi')\right) \rho_c \theta^n(\xi') R_0 \, d\xi' \tag{3.99}$$

Substituting for ρ_c, T_c and R_0 gives

$$\tau_i(\xi) = \left[\frac{-\psi_i \rho_i \theta^n(\xi)}{4\pi G \phi(\xi)}\right]^{1/2} \int_{\xi'=\xi}^{\xi'=\xi_B} \kappa_i \left(\rho_i \left[\frac{\theta(\xi')}{\theta(\xi)}\right]^n, T_i \left[\frac{\theta(\xi')}{\theta(\xi)}\right]\right) \left[\frac{\theta(\xi')}{\theta(\xi)}\right]^n d\xi' \tag{3.100}$$

Taking a mass weighted average then gives the rather messy

$$\overline{\tau}_i = \left[-\xi_B^2 \frac{d\theta}{d\xi}(\xi_B)\right]^{-1} \left[\frac{-\psi_i \rho_i}{4\pi G}\right]^{1/2} \times$$
$$\int_{\xi=0}^{\xi=\xi_B} \int_{\xi'=\xi}^{\xi'=\xi_B} \kappa \left(\rho_i \left[\frac{\theta(\xi')}{\theta(\xi)}\right]^n, T_i \left[\frac{\theta(\xi')}{\theta(\xi)}\right]\right) \theta^n(\xi') d\xi' \left[\frac{\theta^n(\xi)}{\phi(\xi)}\right]^{1/2} \xi^2 \, d\xi \tag{3.101}$$

This appears to be a complicated function to calculate during a simulation: however, using the previous result for $\overline{\Sigma}$, a mass weighted opacity can be defined:

$$\overline{\kappa} = \frac{\overline{\tau}}{\overline{\Sigma}} \tag{3.102}$$

which can be evaluated in advance, and stored for later interpolation. Hence, for a given (ρ, T):

$$\overline{\kappa}(\rho, T) = \left[-\zeta_n \xi_B^2 \frac{d\theta}{d\xi}(\xi_B)\right]^{-1} \times$$

$$\int_{\xi=0}^{\xi=\xi_B} \int_{\xi'=\xi}^{\xi'=\xi_B} \kappa \left(\rho \left[\frac{\theta(\xi')}{\theta(\xi)} \right]^n, T \left[\frac{\theta(\xi')}{\theta(\xi)} \right] \right) \theta^n(\xi') d\xi' \left[\frac{\theta^n(\xi)}{\phi(\xi)} \right]^{1/2} \xi^2 \, d\xi \tag{3.103}$$

The interpretation of this result is important: embedding the particle at some position in the polytrope ensures that the environment immediately surrounding the particle has an important effect on its optical depth, and hence its emission. This allows (for example) insulation of hot particles by cooler surroundings. It is vital at this juncture to appreciate the meaning of this: *the formalism is attempting to compensate for absorption* of escaping radiation by increasing the effective opacity. This effect can be seen in Figure 3.6 in the following section.

The net cooling term for SPH particle i is then

$$\dot{u}_{i,cool} = \frac{4\sigma_{SB} \left(T_0^4(\mathbf{r_i}) - T_i^4 \right)}{\overline{\Sigma}_i^2 \overline{\kappa}_i(\rho_i, T_i) + \kappa_i^{-1}(\rho_i, T_i)} \tag{3.104}$$

The addition of T_0 allows for external heating from a background radiation field (which can be configured to include irradiation from stellar objects). Note that both the particle's opacity and mass weighted opacity are required to interpolate between the optically thick and optically thin regimes. The first term in the denominator becomes dominant in the optically thick case (where the particle's environment will absorb much of the cooling radiation it emits, reducing the energy loss), and the second term becomes dominant in the optically thin case (where the effects of the environment are less important, so the standard opacity is used). Strictly speaking, the first term should be a mass weighted average of the Rosseland-mean opacity, and the second term should use the Planck-mean opacity, but in the case of this work the Rosseland-mean and Planck-mean opacities are taken to be equal.

The construction of $\dot{u}_{i,cool}$ allows the code to move smoothly from optically thin to optically thick regimes, and also identifies an optimum regime where the optical depth is of order unity, where the particle can emit radiation most efficiently (i.e. the photosphere).

The method is very efficient, having little impact on the total simulation time and performing very well in several tests of its ability (Stamatellos et al., 2007b). Unfortunately, it does suffer from some key limitations:

1. Assuming a spherical pseudocloud will place restrictions on how well the code models different geometries: configurations that lack spherical symmetry will not be modelled as accurately as those that are spherically symmetric, although its general accuracy has been shown to be good (Stamatellos et al., 2007b).

2. Although the formalism accounts for the surroundings of the particle when modelling its emission, it does not deal with the "absorbed" radiation (e.g. by distributing it amongst its nearest neighbours). This makes it less capable of capturing all the physics of an optically thick regime.

3.4.2 Flux-Limited Diffusion

The modelling of energy exchange is implemented using the flux-limited diffusion formalism. In SPH this is usually couched in terms of heat conduction, with the appropriate equation (without sources or sinks of heat) is

$$\dot{u}_{cond} = \frac{1}{\rho}\nabla(k\nabla T) \tag{3.105}$$

where c_p is the heat capacity at constant pressure, k is the thermal conductivity and T is the temperature. The second derivative of T (in 1D) for particle a in the SPH formalism is

$$\left(\frac{d^2T}{dx^2}\right)_a = \sum_b m_b T_b \left(\frac{d^2W(x_a - x_b, h)}{dx^2}\right) \tag{3.106}$$

However, this derivative is not a suitable choice for our purposes, not least because it is sensitive to particle disorder, but because the second derivative of the kernel changes sign (Monaghan, 2005), changing the direction of heat flow, violating the Zeroth Law of Thermodynamics. Cleary & Monaghan (1999) show that an integral approximation provides a more stable form. If we begin with

$$I = \int (k(\mathbf{r}) + k(\mathbf{r}'))(T(\mathbf{r}) - T(\mathbf{r}'))F(|\mathbf{r} - \mathbf{r}'|)d\mathbf{r}' \tag{3.107}$$

where $\mathbf{q}F(|\mathbf{q}|) = \nabla W(\mathbf{q}, h)$, then we can Taylor expand $k(\mathbf{r}')$ and $T(\mathbf{r}')$ around $\mathbf{r}$. To first order, this is

$$I = \nabla.(k\nabla T) + O(h^2) \tag{3.108}$$

We can then simply convert the integral into an SPH sum:

$$I = \sum_b \frac{m_b}{\rho_i \rho_b}(k_i + k_b)(T_i - T_b)\frac{\mathbf{r_{ib}}.\nabla\mathbf{W}}{|r_{ib}|^2} \tag{3.109}$$

Cleary and Monaghan modify this expression further, motivated by analysing finite difference cases to replace $(k_i + k_b)$ with $\frac{4k_ik_b}{k_i+k_b}$, to give the commonly used form (Whitehouse & Bate, 2004; Mayer et al., 2007; Forgan et al., 2009):

$$\dot{u}_{i,diff} = \sum_b \frac{4m_b}{\rho_i \rho_b}\frac{k_i k_b}{k_i + k_b}(T_i - T_b)\frac{\mathbf{r_{ib}}.\nabla\mathbf{W}}{|r_{ib}|^2}, \tag{3.110}$$

where b describes the nearest neighbours, W is the smoothing kernel, r_{ib} is the separation vector between particles i and b (where $i \neq b$), and k_i describes the thermal conductivity of the particle. The gradient of the kernel is always negative, so if $T_i > T_b$, the summand will be negative (i.e., energy will flow from particle i to particle b, in accordance with the laws of thermodynamics). If the system's energy budget is defined entirely by diffusion, the particles will exchange energy amongst themselves in order to reduce temperature gradients, ensuring the long term evolution of the system towards a single, equilibrium temperature. This "washing

out" of temperature gradients is of critical importance. When simulating protoplanetary discs, the temperature profile (both radially and vertically) can define the regions of the disc where possible fragmentation can occur, and hence the regions where giant planets may form (Boss, 1997). Any process which affects these profiles will influence where these regions are located.

It should be noted at this point that all energy changes due to these diffusion terms are *pairwise* (i.e., any energy loss by one particle will be matched by gain in its counterpart). This means that the total energy change over the entire system due to diffusion must be zero:

$$\sum_i \dot{u}_{i,diff} = 0. \tag{3.111}$$

This is an important feature, which allows it to be used in the hybrid method, as will be shown later. The thermal conductivity is

$$k_i = \frac{16\sigma_{SB}}{\rho_i \kappa_i} \lambda_i T_i^3 \tag{3.112}$$

where κ_i is the opacity, σ_{SB} is the Stefan-Boltzmann constant and λ_i is the flux limiter. I use the expression for λ_i given in section 2.4 (Levermore & Pomraning, 1981; Bodenheimer et al., 1990):

$$\lambda_i(R_i) = \frac{2 + R_i}{6 + 3R_i + R_i^2}, \tag{3.113}$$

where R_i is a function of the radiation energy density at the particle's position, $u_r(\mathbf{r_i})$:

$$R_i = \frac{|\nabla u_r(\mathbf{r_i})|}{u_r(\mathbf{r_i})\rho_i \kappa_i}. \tag{3.114}$$

Studying the expression for R_i, there are two clear limiting cases:

- When the region is very optically thick, ρ and κ become large (and the radiation field becomes uniform), and hence $R_i \to 0$. In this limit, the flux limiter $\lambda_i \to 1/3$, in accordance with the diffusion approximation.
- In the very optically thin limit, R_i becomes very large, $\lambda_i \to 1/R$ and k_i tends to a constant value.

This approximation is valid in the optically thick regime (and to lower optical depths with the use of the flux limiter, which limits energy exchange as the mean free path of the radiation becomes prohibitively large). Limitations of this method are:

1. It does not model radiation well at very low optical depths
2. It does not allow the system to lose energy (i.e. it does not model radiative cooling).

3.4.3 Cooling and Diffusion Together: The Hybrid Method

Comparing the limitations of the above two methods, it should be clear that a union of these two procedures should be complementary: polytropic cooling handles the important energy loss from the system (which flux-limited diffusion cannot), and flux-limited diffusion handles exchange of heat between fluid elements (which polytropic cooling cannot)[5]. Indeed, polytropic cooling's inability to model transport (and flux-limited diffusion's inability to model energy loss) allow the two methods to work together correctly, modelling all aspects of the system's radiative energy budget without encroaching on each other. The energy equation simply becomes

$$\dot{u}_{i,total} = \dot{u}_{i,hydro} + \dot{u}_{i,cool} + \dot{u}_{i,diff} \tag{3.115}$$

Where $\dot{u}_{i,hydro}$ describes the energy change due to the hydrodynamics of the system, e.g. compressive PdV heating. We must make one alteration to $\dot{u}_{i,diff}$ to ensure that the hybrid method works correctly at low optical depths. In this free-streaming limit, the diffusion must be turned off so that the polytropic cooling alone gives the correct solution. This is achieved using a timestep switch - if the diffusion timescale for any pair of particles (i, b)

$$t_{diff}(i,b) = \frac{u_i + u_b}{2\dot{u}_{diff}(i,b)} \tag{3.116}$$

(where $\dot{u}_{diff}(i,b)$ is the diffusion term between particles i and b, i.e. $\dot{u}_{i,diff} = \sum_b \dot{u}_{diff}(i,b)$) is shorter than the local timestep, then the diffusion term is turned off. This is correctly symmetrised such that if diffusion is turned off for particle i, it is turned off for particle b also, preserving its pairwise nature. This timestep switch is used both in the tests and in the simulations conducted in this thesis.

The true advantage to using this hybrid method is in its simplicity:

- By construction, the hybrid method is fully three-dimensional, and capable of handling arbitrary particle geometries.
- There is no requirement to grid the system.
- The algorithm is continuous over a wide range of optical depths, so there are no requirements to match separate atmospheres at some boundary.
- As no extra boundary conditions are required, there are no extra parameters to be specified, so the simulation's results are only dependent on the traditional SPH parameters (timestep, particle number, smoothing length etc).

However, it still suffers from some disadvantages:

[5] One small alteration worthy of note is to how the column density is calculated for polytropic cooling. To ensure that absorption is as fully accounted for as possible, the column density used in the hybrid method is not mass-averaged, and is instead taken to be the central value $\Sigma_i(0)$. The effect of this is equivalent to changing ζ_n from 0.368 to 0.386, a change of less than 5%.

1. This method is still insufficient to model frequency dependent radiative transfer

2. As with polytropic cooling, the hybrid method is better suited to modelling the cooling of spherical geometries.

3.4.4 Updating Energy: A Semi-Implicit Scheme

The use of an explicit scheme to update energy can result in very short time steps. To avoid this, a modified version of the implicit scheme adopted by Stamatellos et al. (2007b) is used. This models each particle's approach to its equilibrium temperature T_{eq}, which satisfies

$$\dot{u}_{i,hydro} + \frac{4\sigma_{SB}\left(T_0^4(\mathbf{r_i}) - T_{eq,i}^4\right)}{\overline{\Sigma}_i^2\overline{\kappa}_i(\rho_i, T_{eq,i}) + \kappa_i^{-1}(\rho_i, T_{eq,i})} + \dot{u}_{i,diff} = 0 \tag{3.117}$$

From this, the equilibrium internal energy $u_{eq,i} = u(\rho_i, T_{eq,i})$ can be calculated, and hence the thermalisation timescale:

$$t_{therm} = \frac{u_{eq,i} - u_i}{\dot{u}_{i,total}} \tag{3.118}$$

With knowledge of how quickly each particle can be thermalised, the particle's energy can be updated thus:

$$u_i(t + \Delta t) = u_i(t) \exp\left[\frac{-\Delta t}{t_{therm}}\right] + u_{eq,i}\left(1 - \exp\left[\frac{-\Delta t}{t_{therm}}\right]\right) \tag{3.119}$$

For particles that will thermalise very quickly ($t_{therm} \ll \Delta t$), which would result in very short timesteps, this equation reduces to

$$u_i(t + \Delta t) \approx u_{eq,i}, \tag{3.120}$$

and the particle rapidly reaches equilibrium. If thermalisation happens on a long timescale ($t_{therm} \gg \Delta t$), then the equation becomes

$$u_i(t + \Delta t) \approx u_i(t) + (u_{eq,i} - u_i(t))\frac{\Delta t}{t_{therm}}. \tag{3.121}$$

3.4.5 Modelling the Properties of Dust and Gas - Equations of State and the Opacity Law

SPH alone only evolves a limited set of variables in the system, and it needs more data to calculate temperatures and opacities. To calculate the pressure, the ideal gas approximation can be used in these circumstances, as well as barotropic equations of state:

$$P = P\left(\rho\left(\mathbf{r}, t\right)\right). \tag{3.122}$$

However, the chemistry of the system can have important effects on the internal energy. These effects are expressed in the energy equation

$$u = u(\rho, T). \tag{3.123}$$

Most simulations now account for the phases of the gas - they assume the gas is mainly hydrogen & helium (the contribution due to metals being relatively small), and then calculate fractions of dissociation and ionisation. Boley et al. (2007b) set out a procedure for calculating internal energy contributions from molecular hydrogen.

First of all, the two types of molecular hydrogen must be considered: ortho-hydrogen (where the molecule's proton spins are parallel) and para-hydrogen (where the molecule's proton spins are antiparallel). The partition function for para-hydrogen is

$$z_p = \sum_{even\, j} (2j+1) \exp\left(\frac{-j(j+1)\theta_{rot}}{T}\right), \tag{3.124}$$

and for ortho-hydrogen

$$z_o = \sum_{odd\, j} (2j+1) \exp\left(\frac{-j(j+1)\theta_{rot}}{T}\right), \tag{3.125}$$

where $\theta_{rot} = 85.4K$. The combined partition function for rotating molecules depends on the relative abundances on the two species. If the species exist in equilibrium, the partition function is

$$z_{rot} = z_o + z_p. \tag{3.126}$$

If the species cannot be converted from one to the other, the ratio of ortho to para-hydrogen $(b:a)$ becomes fixed. In this situation, the partition function then becomes

$$z_{rot} = (z_p)^{\frac{a}{a+b}} \left(z_o \exp\left(\frac{2\theta_{rot}}{T}\right)\right)^{\frac{b}{a+b}}. \tag{3.127}$$

The vibrational states are approximated by a harmonic oscillator with partition function

$$z_{vib} = \frac{1}{1 - \exp\left(-\frac{\theta_{vib}}{T}\right)}, \tag{3.128}$$

where $\theta_{vib} = 5987\,K$. The specific internal energy can then be calculated from the partition functions as

$$u_{H_2} = \frac{R}{\mu} T^2 \frac{\partial \ln z}{\partial T}, \tag{3.129}$$

Combining the above equations, this can be expressed as

$$u_{H_2} = \frac{R}{2}\left[\frac{3}{2}T + \frac{T^2}{z_{rot}}\frac{\partial z_{rot}}{\partial T} + \theta_{vib}\frac{\exp\left(-\frac{\theta_{vib}}{T}\right)}{1 - \exp\left(-\frac{\theta_{vib}}{T}\right)}\right]. \tag{3.130}$$

To calculate the contributions due to atomic and ionised hydrogen and helium, the degrees of dissociation and ionisation of both elements need to be defined. For hydrogen, the degree of dissociation y and ionisation x are

$$y = \frac{n_{H^0}}{n_{H^0} + n_{H_2}} \tag{3.131}$$

$$x = \frac{n_{H^+}}{n_{H^+} + n_{H^0}} \tag{3.132}$$

For Helium, the degrees of single and double ionisation respectively are

$$z_1 = \frac{n_{He^+}}{n_{He^+} + n_{He^0}} \tag{3.133}$$

$$z_2 = \frac{n_{He^{++}}}{n_{He^+} + n_{He^{++}}} \tag{3.134}$$

y is calculated using the expression given by Larson (1968) (based on the data of Tsuji 1966):

$$\frac{y^2}{1-y} = \frac{2.11}{\rho X} \exp\left(\frac{-52490}{T}\right) \tag{3.135}$$

x, z_1 and z_2 are calculated from standard Saha equations:

$$\frac{x^2}{1-x} = \frac{m_H}{\rho X} \frac{(2\pi m_e kT)^{3/2}}{\left(2\pi\bar{h}\right)^3} \exp\left(\frac{-13.6\,\mathrm{eV}}{kT}\right) \tag{3.136}$$

$$\frac{z_1}{1-z_1} = \frac{4m_H}{\rho\left(X + z_1 Y/4\right)} \frac{(2\pi m_e kT)^{3/2}}{\left(2\pi\bar{h}\right)^3} \exp\left(\frac{-24.6\,\mathrm{eV}}{kT}\right) \tag{3.137}$$

$$\frac{z_2}{1-z_2} = \frac{m_H}{\rho\left(X + Y/4 + z_2 Y/4\right)} \frac{(2\pi m_e kT)^{3/2}}{\left(2\pi\bar{h}\right)^3} \exp\left(\frac{-54.4\,\mathrm{eV}}{kT}\right) \tag{3.138}$$

With the underlying assumption that ionisation of a particular state can only begin once its predecessor is complete: i.e. H^0 ionisation cannot occur before H_2 dissociation is complete, etc. By specifying the mass fractions of Hydrogen and Helium in the gas (X, Y), it is relatively straightforward to calculate the other internal energy contributions due to atomic hydrogen, ionised hydrogen, and Helium in its various ionisation states (see Stamatellos et al. 2007b):

$$u_{H^0} = Xy(1+x)\frac{3k_B T}{2m_H} \tag{3.139}$$

$$u_{H_2\,DISS} = Xy\frac{D_{H_2 DISS}}{2m_H} \tag{3.140}$$

$$u_{H\,ION} = Xxy\frac{I_{H\,ION}}{m_H} \tag{3.141}$$

$$u_{He^0} = Y(1+z_1+z_2)\frac{3k_B T}{8m_H} \tag{3.142}$$

$$u_{He\,ION} = Yz_1(1-z_2)\frac{I_{He\,ION}}{2m_H} \tag{3.143}$$

$$u_{He^+\,ION} = Y z_1 z_2 \frac{I_{He^+\,ION}}{2m_H} \tag{3.144}$$

where in the above $D_{H_2 DISS} = 4.5\,eV$ is the dissociation energy of H_2, $I_{H\,ION} = 13.6\,eV$, $I_{He\,ION} = 24.6\,eV$ and $I_{He^+\,ION} = 54.4\,eV$ are the respective ionisation energies of each species. This thesis assumes $X = 0.7$, $Y = 0.3$ (with no metallicity).

The final equation for $u(\rho, T)$ is:

$$u(\rho,T) = u_{H_2} + u_{H^0} + u_{H_2\,DISS} + u_{H\,ION} + u_{He} + u_{He\,ION} + u_{He^+\,ION}. \tag{3.145}$$

As can be seen, the equation of state rapidly becomes complicated, even in the relatively simple circumstances of zero metallicity. Instead of attempting to calculate these equations at every timestep, values of $u(\rho, T)$ are calculated for a range of values of ρ and T, which are then stored in a dense look up table, which is read from file by the simulation. The dependence of u upon temperature can be seen in Figure 3.4. As can be seen, the energy moves upwards in steps between 10^3 and 10^4 K, reflecting the energy required to dissociate and ionise the various species.

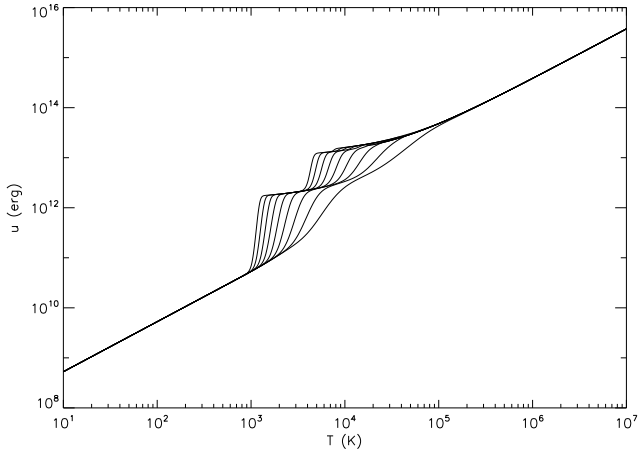

Figure 3.4: Internal energy u as a function of T for various densities. Curves are plotted for $\rho = 10^{-18}\,g\,cm^{-3}$ to $\rho = 1\,g\,cm^{-3}$

The added advantage of using this look-up table system is that other variables dependent on ρ and T can be calculated in advance, and stored for later use. The two key variables that are stored are the mean molecular weight

$$\mu = \left[(1 + y + 2xy)\frac{X}{2} + (1 + z_1 + z_1 z_2)\frac{Y}{4}\right]^{-1} \tag{3.146}$$

and the opacity, κ. Opacity is generally a function of incident photon frequency: in this case, the Rosseland frequency-averaged opacity is used. Its dependence on ρ and T is parametrised according to the opacity law described in Bell & Lin (1994):

Table 3.1: Opacity Law parameters (see also Bell & Lin 1994).

No.	Dominant component	$\kappa_0 \left(cm^{2+3a}\, g^{-(a+1)} K^{-b}\right)$	a	b
1	Ice grains	2×10^{-4}	0	2
2	Evaporation of ice grains	2×10^{16}	0	-7
3	Metal grains	0.1	0	1/2
4	Evaporation of metal grains	2×10^{81}	1	-24
5	Molecules	10^{-8}	2/3	3
6	H^- absorption	10^{-36}	1/3	10
7	bound-free / free-free	1.5×10^{20}	1	$-5/2$
8	Electron scattering	0.348	0	0

$$\kappa(\rho, T) = \kappa_0 \rho^a T^b \tag{3.147}$$

where κ_0, a and b are selected according to which density/temperature regime the particle is in (see Table 3.1).

The opacity law can be seen in Figure 3.5. Note the so-called "opacity gap", where dust begins to evaporate, and can no longer provide an effective opacity source. The mass averaged opacity (equation 3.103) can be seen in Figure 3.6, where the opacity gap can be seen to be much shallower (which is due to polytropic cooling's compensation for absorption by nearest neighbours, see previous section).

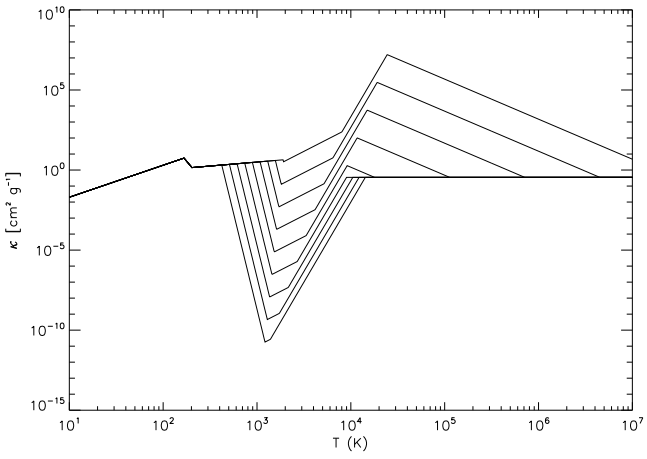

Figure 3.5: Rosseland mean opacity as a function of temperature for a series of different densities. Curves are plotted for $\rho = 10^{-18}\, g\, cm^{-3}$ to $\rho = 1\, g\, cm^{-3}$.

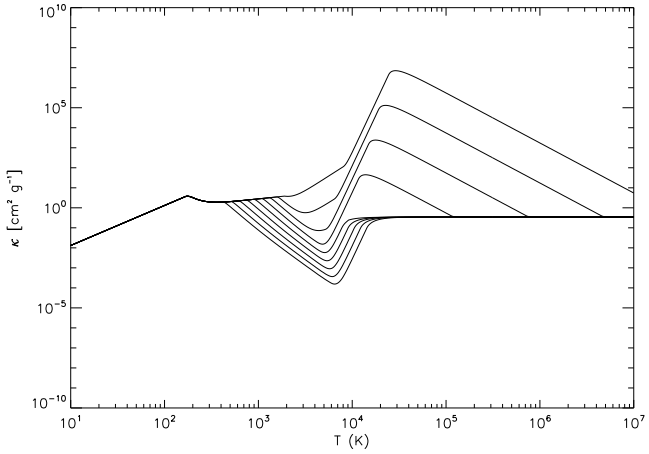

Figure 3.6: The mass averaged opacity as a function of temperature for a series of different densities. Curves are plotted for $\rho = 10^{-18}\, g\, cm^{-3}$ to $\rho = 1\, g\, cm^{-3}$.

3.5 Testing the Hybrid Method

The code used to perform these tests is based on the SPH code developed by Bate et al. (1995). It uses variable individual smoothing lengths h_i so that the number of nearest neighbours for any particle is 50 ± 20. It uses individual particle timesteps to allow dense regions to be simulated with greater time resolution while preventing oversimulation of less dense regions. A binary tree is employed to calculate neighbour lists and calculate gravity forces. The standard artificial viscosity is also used. To ensure that potential fragmentation is resolved, the minimum Jeans mass resolvable (one neighbour group of SPH particles) must be sufficiently small (Bate & Burkert, 1997):

$$M_{\rm min} = 2N_{\rm neigh}m_{\rm i} = 2M_{\rm tot}\frac{N_{\rm neigh}}{N_{\rm tot}}. \tag{3.148}$$

All simulations are sufficiently populated to satisfy this for Jeans masses of $30\, M_\oplus$ or less. These conditions are sufficient for the cloud simulations performed.

For the disc simulations performed, the Toomre length becomes important in the regions that are unstable, and places stricter resolution conditions. As the disc simulation is Keplerian, the following relation can be used between the Jeans length and the Toomre length (Nelson, 2006):

$$\lambda_T = \sqrt{\frac{2Q}{f}}\lambda_J, \tag{3.149}$$

where $f \sim 1$ represents the conversion factor between surface and volume densities. As the disc is marginally unstable ($Q \sim 1$), the Toomre length can be simply calculated. Converting this (assuming a homogeneous sphere) into a Toomre Mass, it is calculated that the disc simulation can resolve Toomre masses of $\sim 85\, M_\oplus$ or more.

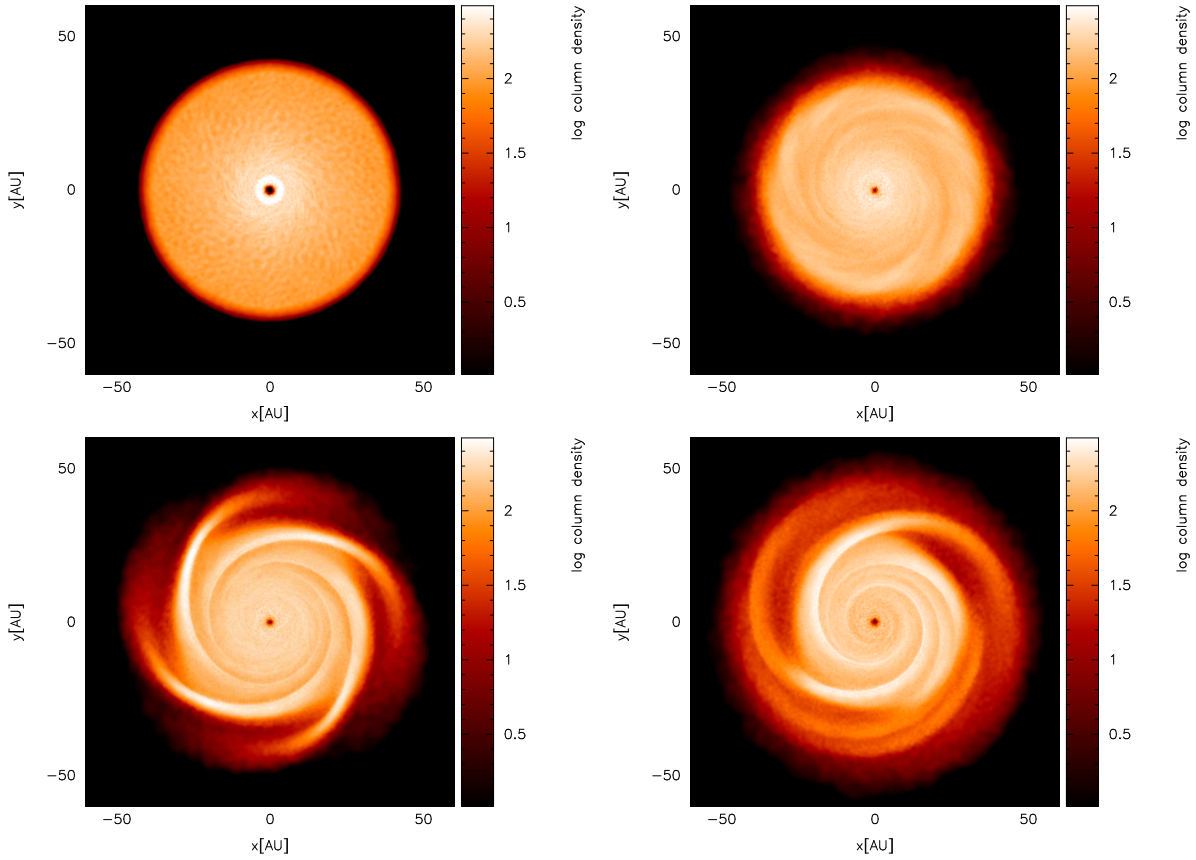

Figure 3.7: Surface density snapshots of the Boley disc at various times under the hybrid method . The images are taken at the following times: 9.72 years (top left), 506 years (top right), 992 years (bottom left), 1906 years (bottom right).

3.5.1 The Evolution of a Protoplanetary Disc

As a means of comparison with previous results, the conditions used for this test are those proposed by Mejía et al, and used in a series of papers describing radiative transfer in protoplanetary discs (Pickett et al., 2003; Mejia et al., 2005; Boley et al., 2006; Cai et al., 2008). The model is a $0.07\,M_{\odot}$ Keplerian disc which extends to 40 AU, orbiting a star of $0.5\,M_{\odot}$. Initially, the surface density profile is $\Sigma \sim r^{-1/2}$, with a temperature profile of $T \sim r^{-1}$. The disc is modelled using 2.5×10^5 SPH particles, with one sink particle representing the star. The disc is immersed in a radiation field of $T_0(\mathbf{r}) = 3K$; the effects of disc irradiation by the central star are not included.

The properties of the evolved disc using both the hybrid method and the polytropic cooling approximation alone are shown in 3.8. For both methods, several key phases are identified: the initial settling phase, during which the disc adjusts its outer radius by axisymmetric evolution, and ring formation and contraction occurs; the "burst" phase where non-axisymmetric

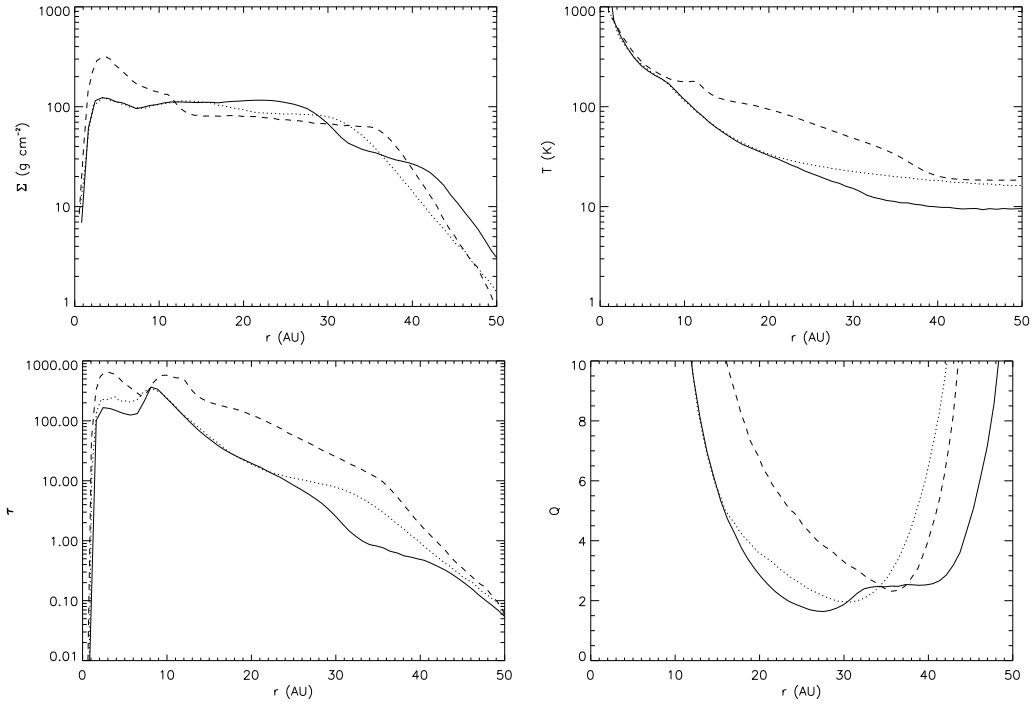

Figure 3.8: Azimuthally averaged radial profiles of the Boley disc at $t = 1906$ years: the solid lines are the results obtained using the hybrid method, the dotted lines are the results obtained using the polytropic cooling approximation alone, and the dashed lines are the disc at $t = 9.72$ years using the hybrid method. The top left panel shows the surface density of the disc; the top right panel shows the midplane temperature of the disc; the bottom left panel shows the optical depth from the midplane to the disc surface; the bottom right panel shows the Toomre instability parameter.

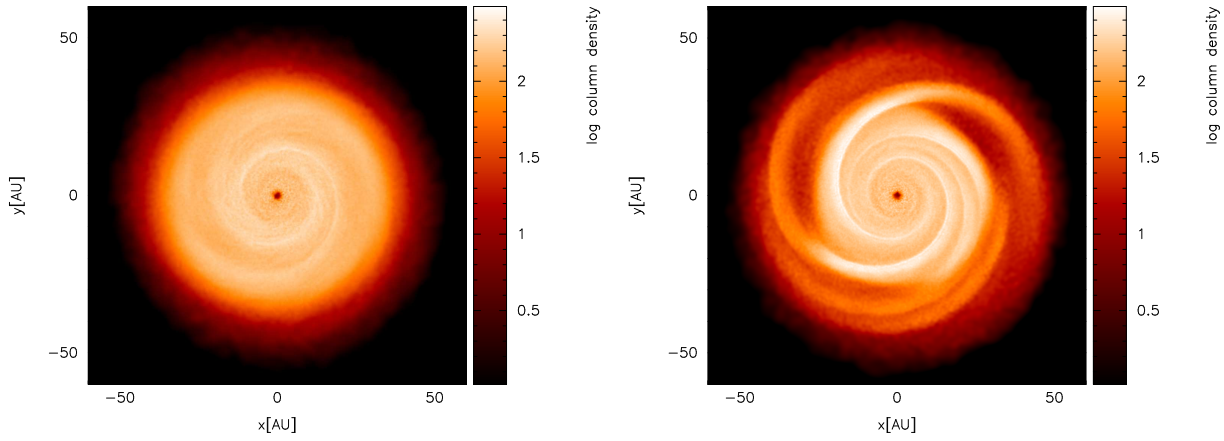

Figure 3.9: Comparing the hybrid method and polytropic cooling for the Mejía disc at $t = 1906$ years. The left panel shows the evolved disc under polytropic cooling alone; the right panel shows the evolved disc under the hybrid method.

instabilities, in the form of spiral waves, begin to grow; and the later asymptotic phase, where the disc's radial extent is more firmly established, and the gravitational instability is regulated (cf. Boley et al. 2006). As the evolution of this asymptotic phase continues, the low-m modes begin to dominate. In terms of timescale, the settling phase lasts until $t \sim 500$ years, the burst phase until $t \sim 1200$ years, where the asymptotic phase then begins, resulting in a quasi-equilibrium state.

Comparing the hybrid method against the results of using the polytropic cooling approximation alone, there are significant differences. The hybrid method transports more mass radially outward, which can be seen in the surface density of the disc (Figure 3.8, top left panel). This has several important consequences. It allows the optical depth to be reduced in the region $r \sim 20 - 40$ AU, which allows an increase in radiative cooling. This in turn allows the outer disc to be cooler, and for the outer regions of the disc ($r > 20$ AU) to become less stable (as can be seen in the other panels of Figure 3.8). Snapshots of the disc under both methods can be seen in Figure 3.9. Note the stronger spiral structure in the disc under the hybrid method, with instabilities extending to larger radii. All these differences are critical if the formation of giant planets by gravitational instability is to be effectively tested by simulation.

Comparing the hybrid method to the results of Boley et al. (2006, 2007a), the two are qualitatively consistent. Each has a burst phase and an asymptotic phase, and each has a two-component surface density profile (approximately flat at lower radii, with a cut off at larger radii). There are also some quantitative consistencies. The optical depth from the midplane to the surface in the hybrid method reaches unity at $R \sim 27$ AU, which is coincident with the region of the disc that is most unstable (i.e., the Toomre Q parameter is at a global minimum), and is in keeping with the work of Boley et al. It can also be seen (by comparing the surface density profiles of the hybrid method and polytropic cooling) that there appears to be a surplus of matter within $R \sim 20 - 27$ AU, and a slight deficit at $R \sim 27 - 40$ AU, indicating that $R \sim 27$ AU may be the location where mass transport switches from inward to outward, again consistent with the results of Boley et al. However, there are some important differences to be considered. The burst phase of the hybrid method is noticeably weaker, and the disc undergoes less radial spreading. This also means that in the asymptotic phase, the unstable region is much narrower in radius. The first component of the surface density profile also appears to be flatter at lower radii for the hybrid method.

It should be noted at this point that there are mitigating factors at work. The equation of state and opacity law used in this work is different from that of Boley et al; also, they fix the star at the centre of their grid, while the star used in our simulation is allowed to move. The differences in the equation of state and the opacity law will have a stronger effect in the hotter inner regions of the disc, perhaps explaining the differences in surface density profile, and the lack of radial spreading. It should also be noted that the inner disc stays somewhat hotter than expected (for both polytropic cooling alone, and for the hybrid method). This may be due to

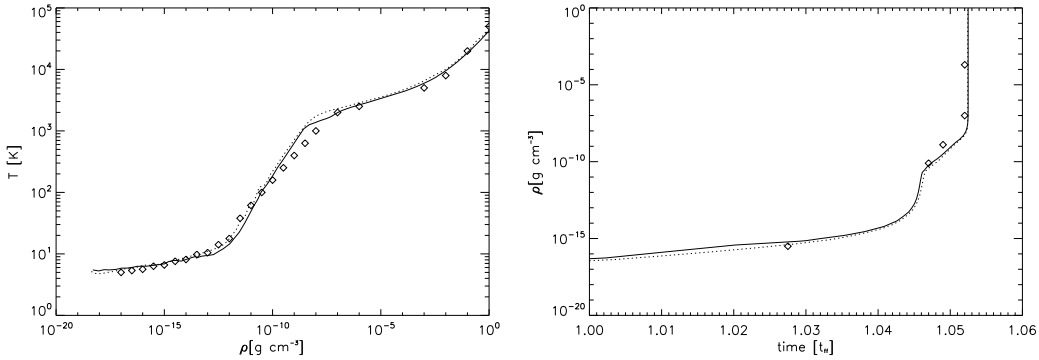

Figure 3.10: Evolution of the central density of the Masunaga Cloud - the left panel shows the evolution of central temperature with increasing central density, the right panel shows the time evolution of the central density. The solid lines represent the hybrid method, the dotted lines represent polytropic cooling only, and the diamonds represent the data of Masunaga & Inutsuka (2000).

SPH viscosity. As the distance to the centre decreases, the magnitude of the SPH viscosity increases, and may become significant (relative to the effective gravitational viscosity). This phenomenon will be investigated in more detail in Chapter 4.

Although exciting spiral waves in all three cases (polytropic cooling alone, the hybrid method and the work of Boley et al), the instability in the disc does not lead to fragmentation. Also, the disc is only Toomre unstable at larger radii, which does not bode well for *in situ* formation of Jovian objects at $R \leq 20$ AU (at least under these conditions).

3.5.2 The Collapse of a $1\,M_{\odot}$ Cloud

The collapse of a non-rotating molecular cloud was then simulated. The spherical, uniform density cloud contains $1\,M_{\odot}$ of material (populated by 5×10^5 SPH particles), and has a radius of 10^4 AU (which gives a density of $\rho_0 = 1.41 \times 10^{-19}\,\mathrm{g\,cm^{-3}}$), and is immersed in a background radiation field of $T_0(\mathbf{r}) = 5K$. These conditions were initially investigated by Masunaga & Inutsuka (2000) by solving the full radiative transfer in 3D (with the hydrodynamics solved in 1D), and were revisited by Stamatellos et al. (2007b). These conditions therefore represent not only a solid test of the code's ability to match Masunaga & Inutsuka's data, but also allow us to compare with the results of Stamatellos et al to identify the effects of adding flux-limited diffusion.

In the initial phase, the collapse is isothermal. The temperature remains at approximately 5 K through seven orders of magnitude in density (see Figure 3.10, left panel), until the central density reaches $\rho \sim 10^{-12}\,\mathrm{g\,cm^{-3}}$. The cloud then becomes optically thick, and the temperature starts to rise. As the temperature reaches $T \sim 100$ K, the rotational degrees of freedom of molecular hydrogen are activated, slowing the temperature increase slightly (this

can be seen in the small bump in the left panel of Figure 3.10). The increased heating in the centre eventually decelerates the contraction at around $\rho = 10^{-9}\,\mathrm{g\,cm^{-3}}$, and the first core is formed. The contraction and heating of this core proceed until the central temperature is around $T \sim 2000$ K. The H_2 present begins to dissociate, using some of the available compressive energy due to the contraction. This allows a second collapse, which can continue until most of the H_2 is dissociated. After this the contraction decelerates again at around $\rho = 10^{-3}\,\mathrm{g\,cm^{-3}}$, and the second core forms.

The dotted line in the left hand panel of Figure 3.10 shows the evolution of the Masunaga Cloud using polytropic cooling alone. Both methods approximate the data of Masunaga & Inutsuka (2000) well (diamonds in Figure 3.10). However, there are two key differences: the hybrid run stays isothermal to slightly higher densities (where the extra loss of energy along temperature gradients due to diffusion keeps the cooling efficient enough to allow this), and the slight bump at $\rho_0 \sim 10^{-9}\,\mathrm{g\,cm^{-3}}$ (again diffusion allowing the centre to cool more efficiently). This demonstrates that the polytropic cooling method alone provides a good approximation of the energy exchange between neighbouring particles (in this highly symmetric case) by correctly modelling the net radiative losses; the addition of flux-limited diffusion constitutes only a small additional exchange of energy.

The time evolution of the cloud (Figure 3.10, right panel) follows closely the evolution described by Stamatellos et al. (2007b) and Masunaga & Inutsuka (2000). As with Stamatellos et al, there are discrepancies with Masunaga and Inutsuka's data due to the use of different opacities, and slight variations in initial conditions. By synchronising the simulations at a central density of $\rho = 4.34 \times 10^{-13}\,\mathrm{g\,cm^{-3}}$ (Stamatellos et al., 2007b), good agreement is obtained.

3.5.3 The Spiegel Test

As a final test, the thermal relaxation of a static, spherical cloud with a well-defined temperature perturbation allows comparison of the hybrid method with analytic results. The cloud is uniform in density, with $\rho = 10^{-19}\mathrm{g\,cm^{-3}}$, and a radius of $R = 10^4$ AU. The equilibrium temperature is taken to be $T_0 = 10$ K, and an initial temperature perturbation which satisfies

$$T(r) = T_0 + \Delta T_0 \frac{\sin kr}{kr}, \tag{3.150}$$

where $\Delta T_0 = 0.15$ K is the amplitude, and $k = \frac{\pi}{2500\,\mathrm{AU}}$ is the characteristic wavenumber (Spiegel, 1957; Masunaga et al., 1998). At a later time t, this perturbation evolves according to (Masunaga et al., 1998)

$$T(r,t) = T_0 + \Delta T_0 \frac{\sin kr}{kr} e^{-\omega(k)t}. \tag{3.151}$$

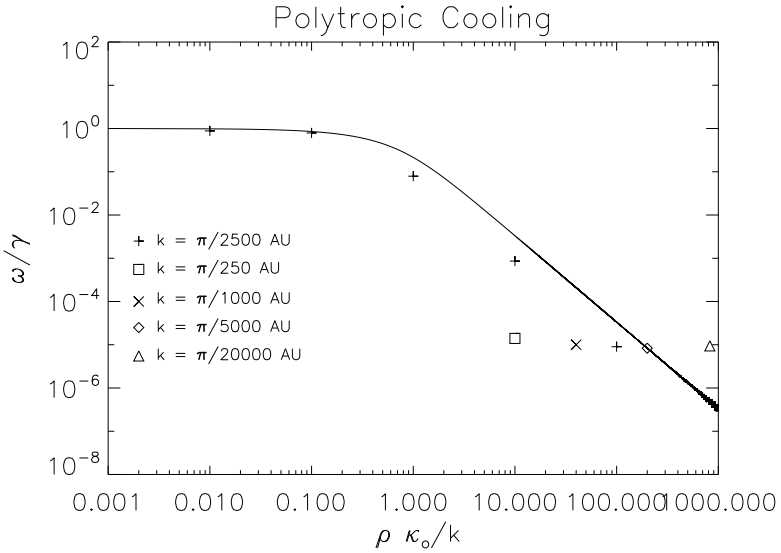

Figure 3.11: The dispersion relation ω for the Spiegel Test using polytropic cooling.

In Equation (3.151),

$$\omega(k) = \gamma \left[1 - \frac{\kappa_0}{k} cot^{-1}\left(\frac{\kappa_0}{k}\right)\right] \tag{3.152}$$

and

$$\gamma = \frac{16\sigma_{SB}\kappa_0 T_0^3}{\rho c_v}. \tag{3.153}$$

Here κ_0 is the opacity at equilibrium and c_v is the heat capacity of the material. This test was also performed by Stamatellos et al. (2007b), and hence provides an extra means of comparing polytropic cooling and the hybrid method.

The key analytical result is the dispersion relation $\omega(k)$, which is shown as the solid lines in Figures 3.11 and 3.12. The points in each panel are obtained by fitting the analytical curve of equation (3.151) to the 2×10^5 SPH particles in the simulation. The best-fit curve gives a value for ω/γ. This is repeated at several instants in the simulation to obtain a mean value. We carry out the test for various values of κ_0/k to compare to the analytical curve. We also repeat the test for different perturbation wavenumbers k. Equation (3.152) is normalised in k, so two points with the same κ_0/k but different k should still approximate the same curve.

Figure 3.11 shows the results using polytropic cooling only; Figure 3.12 shows the results using the hybrid method. Considering the $k = \pi/2500$ AU simulations, both methods correctly model the optically thin regime (low κ_0/k), as they are essentially the same (diffusion being turned off by the timestep switch described previously). As the optical depth increases, the hybrid method approximates the curve better, as it can model the local radiation transport between particle neighbours that occurs in the optically thick limit. The hybrid method is also

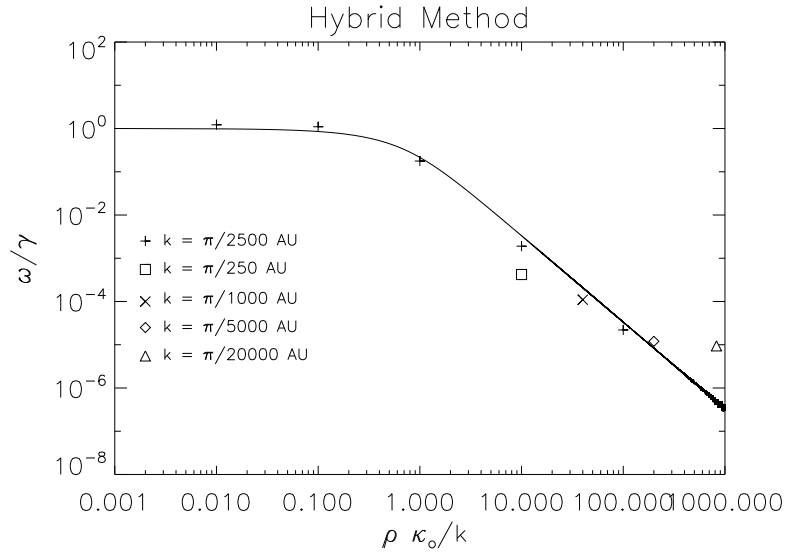

Figure 3.12: The dispersion relation ω for the Spiegel Test using the hybrid method.

more robust to changes in k: it can be seen that the various simulations still cluster closely to the analytical curve. Changes in k result in changes in the local temperature gradient, information that polytropic cooling does not incorporate into its formalism. Therefore, its cooling rate remains a function of κ_0 only, and will not change with k as it should (note the horizontal line the points adopt around $\omega/\gamma = 10^{-5}$). The hybrid's explicit incorporation of temperature gradients allows it to adjust correctly, maintaining the correct curve over various values of k. The exception is the very small $k = \pi/20000$ AU, which fails for both methods as it does not incorporate a full wavelength into the cloud, invalidating the approximations made by Spiegel to obtain equation (3.151).

For extra comparison, the temperature profiles of the cloud for polytropic cooling and the hybrid method are shown in Figures 3.13 & 3.14. In the optically thin case (Figure 3.13), the two panels are basically identical, since flux-limited diffusion is not active in this limit; both illustrate the decaying sinusoidal function described in equation (3.151). In the optically thick case (Figure 3.14), the curve for polytropic cooling begins to spread, filling the regions between the troughs/peaks and 10 K. The same panel for the hybrid method shows less spreading, retaining a more robust sinusoidal pattern.

3.6 Conclusions

I have presented a new means of modelling radiative transfer in SPH by fusing two well tested methods, polytropic cooling and flux-limited diffusion, in order that they may complement each other, and perform the functions that the other cannot. By this fusion, the physics of three-dimensional frequency-averaged radiative transfer in SPH is captured without the need

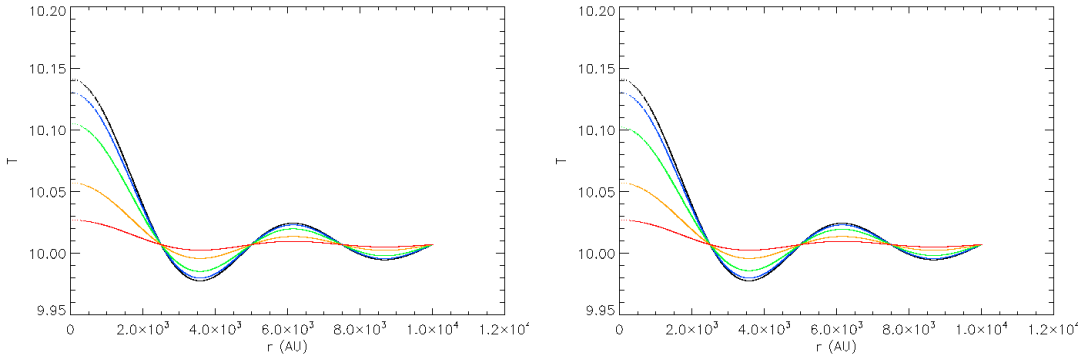

Figure 3.13: Temperature profiles for the thermal relaxation of an optically thin sphere ($\kappa_0/k = 0.01$, $k = \pi/2500$ AU). The left panel shows the data for polytropic cooling only, the right hand panel shows the data for the hybrid method.

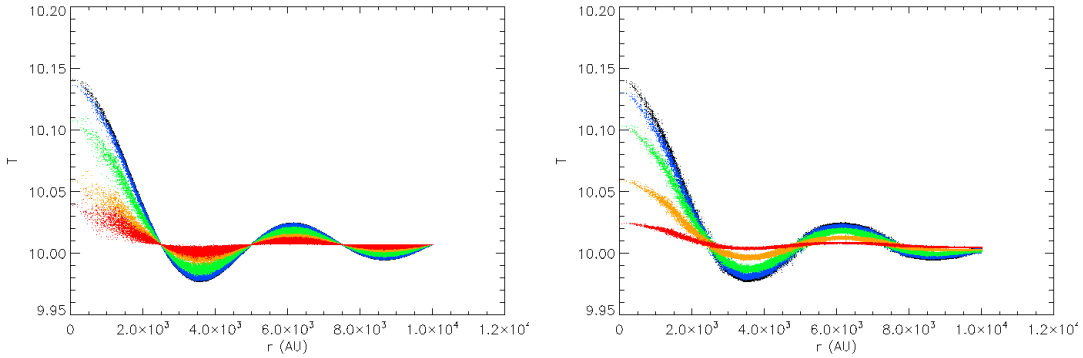

Figure 3.14: Temperature profiles for the thermal relaxation of an optically thick sphere ($\kappa_0/k = 100$, $k = \pi/2500$ AU). The left panel shows the data for polytropic cooling only, the right hand panel shows the data for the hybrid method.

for complex boundary conditions, photosphere mapping or extra parameters. Temperatures and opacities are obtained using a non-trivial equation of state which captures the effects of H_2 dissociation, H^0 ionisation, He^0 and He^+ ionisation, ice evaporation, dust sublimation, molecular absorption, bound-free and free-free transitions and electron scattering. This data is tabulated pre-simulation for interpolation by the code.

The algorithm is fast, with only a 6% increase in CPU time incurred in comparison to standard SPH simulations performed with a barotropic equation of state. It has shown itself to be accurate in the tests outlined in the previous section: the evolution of a protoplanetary disc, with parameters proposed by Mejía, Boley, Cai et al (Pickett et al., 2003; Mejia et al., 2005) from a uniform state through ring formation and contraction to instability; the complex thermal history of a collapsing molecular cloud (as studied by Masunaga & Inutsuka 2000); and the smoothing of temperature fluctuations in a homogeneous, static sphere (Spiegel, 1957; Masunaga et al., 1998). However, the scheme is still approximate, and can only partially describe radiative effects that occur over midrange distances (unlike the scheme proposed by Boley et al. 2007a, albeit in the vertical direction only).

Comparisons with simulations using polytropic cooling alone have shown that the hybrid method is in effect only a small correction to the polytropic cooling method, which however can become important in some problems where temperature gradients and system geometries become complex (such as the Spiegel test, and the protoplanetary disc simulations described in this thesis).

In subsequent chapters I will describe how I have applied this algorithm to two different scenarios. In Chapter 4 I study angular momentum transport and spiral structure in isolated self-gravitating discs, and in Chapter 5 I will discuss the effects of stellar encounters on self-gravitating discs.

Chapter 4

Angular Momentum Transport in Self-Gravitating Protostellar Discs

Never confuse Motion with Action.
Marlene Dietrich, quoting Ernest Hemingway, as documented in *Papa Hemingway*

4.1 Author's Note

This chapter is based on work detailed in Forgan et al. (2010). I include a more in-depth discussion of angular momentum transport to provide the appropriate context for understanding the work.

4.2 The Problem of Viscosity

As noted in section 2.6.5, accretion disc theory can completely determine the evolution of protostellar discs, provided that the viscosity can be evaluated. An obvious first choice is the standard kinetic viscosity between colliding gas molecules. The viscosity will be of order the typical collision velocity (i.e. the local sound speed c_s) multiplied by the collisional mean free path ℓ:

$$\nu = c_s \ell = \frac{c_s}{n\sigma_{coll}}, \tag{4.1}$$

where n is the number density of molecules, and σ_{coll} is the collisional cross-section. We can convert n into readily attainable disc variables by noting $n = \frac{\rho}{\mu m_p}$, and using the thin disc approximation $\rho = \frac{\Sigma}{2H}$. This gives

$$\nu = \left(\frac{2\mu m_p}{\Sigma \sigma_{coll}}\right) c_s H, \tag{4.2}$$

where μ is the mean molecular weight and m_p is the proton mass. We need to ascertain whether or not this viscosity is responsible for the accretion rates in observed protostellar discs. We can do this by considering the *viscous timescale* t_{visc}, and compare this with the dynamical timescale $t_{dyn} = \Omega^{-1}$. We can estimate this using the Reynolds number:

$$Re = \frac{|\mathbf{F}_{inertial}|}{|\mathbf{F}_{viscous}|} = \frac{|\mathrm{d}\mathbf{v}/\mathrm{d}t_{dyn}|}{|\mathrm{d}\mathbf{v}/\mathrm{d}t_{visc}|} = \frac{\bar{v}\ell}{\nu}. \tag{4.3}$$

From this, we can see that in fact the ratio of t_{visc}/t_{dyn} is equal to the Reynolds Number. If we take a typical length scale as the disc radius r, and a typical velocity $v = r\Omega$, we can see that

$$\frac{t_{visc}}{t_{dyn}} = \frac{r^2\Omega}{\nu}. \tag{4.4}$$

Substituting $\Omega = c_s/H$ gives

$$\frac{t_{visc}}{t_{dyn}} = \left(\frac{\Sigma\sigma_{coll}}{2\mu m_p}\right)\left(\frac{H}{r}\right)^{-2}. \tag{4.5}$$

Lodato (2007) shows that, given typical values for protostellar discs, $\frac{t_{visc}}{t_{dyn}} \approx 10^{11}$. The dynamical timescale of discs ranges from $1-1000$ years (depending on radius and disc mass). We therefore recover typical disc accretion rates of order $\dot{M} = 3\pi\nu\Sigma \sim 10^{-14}\, M_\odot yr^{-1}$. This is around ten orders of magnitude lower than observed disc accretion rates of up to $\sim 10^{-6}\, M_\odot yr^{-1}$ (Kenyon et al., 1990). We must therefore conclude that simple kinetic viscosity is insufficient to reproduce observed angular momentum transport in protostellar discs.

4.3 Turbulence, and the α Prescription for Viscosity

Although kinetic viscosity fails to reproduce plausible accretion rates for protostellar discs, we should consider the extremely large value of Re it implies. Flows with a very large Reynolds number can be expected to be turbulent. Turbulent flows can achieve a higher effective viscosity than laminar flows, as they can exchange angular momentum through the mixing of fluid elements over relatively large length scales, much larger than the collisional mean free path. We now appear to have a viable means by which we can produce an effective viscosity required to replicate accretion rates observed in the Universe. Our problem now rests in characterising this turbulent viscosity. A successful approach taken by Shakura & Sunyaev (1973) and many subsequent authors involves developing a dimensionless parameter α to define the viscosity. If

we consider the kinematic viscosity ν, dimensional analysis suggests that we should multiply α by a length scale and a velocity scale. The length scale should represent the largest eddies in the disc, and the velocity scale should represent the typical turbulent velocity. If the turbulence is isotropic, then we can expect the eddies to be limited to the disc thickness H. We also expect the turbulence to be subsonic (as supersonic turbulence would be dissipated through shocks). Therefore, a good approximation would be

$$\nu = \alpha c_s H. \tag{4.6}$$

In terms of the viscous stress tensor component $T_{r\phi}$ using cylindrical co-ordinates[1], this is

$$T_{r\phi} = \frac{d\ln\Omega}{d\ln r}\alpha\Sigma c_s^2. \tag{4.7}$$

We can make sense of the above equation by noting that the stress tensor has the same dimensions as pressure (Σc_s^2), ensuring its divergence has the correct units to be used in the Navier-Stokes equation. The log derivative of Ω is dimensionless, and equal to $-3/2$ in Keplerian discs.

4.4 Self-Gravity as a source of "Viscous" Transport

The α-prescription allows the theorist a certain freedom in choosing which turbulence is responsible for generating the viscosity. Magnetohydrodynamic (MHD) turbulence has been evoked successfully by many authors (Balbus & Hawley, 1991; Balbus & Papaloizou, 1999; Papaloizou & Nelson, 2003). However, this approach is not appropriate if the disc ionisation is very weak (as is the case with most protostellar discs in their early phases).

Systems with a high disc-to-star mass ratio, q, could undergo gravitational instability and generate self-regulated gravito-turbulence (as outlined in section 2.7.4). We could therefore derive an effective gravito-turbulent viscosity, driving angular momentum transport in two ways: the "direct" action of gravitational stresses in the disc (produced as a result of shear in the gravitational field), and the "indirect" action of the gravito-turbulence producing correlated perturbations and Reynolds stresses. I will now discuss the form of these tensor components in more detail.

4.4.1 The Gravitational Stress Tensor

If we are to use self-gravity as a source of angular momentum transport, we should determine the form of the viscous stress tensor from gravitational forces (Morgan & Bondi, 1970; Lynden-Bell & Kalnajs, 1972; Shu, 1991). Working in index notation, we begin with Poisson's equation in terms of the gravitational acceleration $\mathbf{g}$:

[1] We assume that only shear viscosity acts here

$$\frac{\partial g_j}{\partial x_j} = -4\pi G\rho. \tag{4.8}$$

We can obtain a quantity of the same dimension as $\underline{\nabla}.T$ using ρg_i:

$$\rho g_i = -\frac{1}{4\pi G}\frac{\partial g_j}{\partial x_j} g_i. \tag{4.9}$$

We can manipulate this further using the product rule:

$$\rho g_i = -\frac{1}{4\pi G}\left(\frac{\partial}{\partial x_j}(g_j g_i) - g_j\frac{\partial g_i}{\partial x_j}\right). \tag{4.10}$$

Now consider the second term on the right hand side. It can be transformed by considering

$$g_j\frac{\partial g_i}{\partial x_j} = g_j\frac{\partial^2\Phi}{\partial x_j\partial x_i} = g_j\frac{\partial g_j}{\partial x_i}. \tag{4.11}$$

We would like to to convert the $\frac{\partial}{\partial x_i}$ term into a $\frac{\partial}{\partial x_j}$ term (this will allow us to obtain the stress tensor's correct form by factoring out the derivative). Using the product rule (and relabelling the dummy index j to m) gives:

$$g_j\frac{\partial g_j}{\partial x_i} = \frac{1}{2}\frac{\partial}{\partial x_i}(g_m g_m). \tag{4.12}$$

The Kronecker δ can now be used to remove the $\frac{\partial}{\partial x_i}$ dependence:

$$g_j\frac{\partial g_j}{\partial x_i} = \frac{1}{2}\frac{\partial}{\partial x_j}(g_m g_m \delta_{ij}). \tag{4.13}$$

Substituting back into equation (4.9) gives

$$\rho g_i = -\frac{1}{4\pi G}\frac{\partial}{\partial x_j}\left(g_j g_i - \frac{1}{2}g_m g_m \delta_{ij}\right) = \frac{\partial T_{ij}^{\mathrm{grav}}}{\partial x_j}. \tag{4.14}$$

This gives the gravitational stress tensor to be

$$T_{ij}^{\mathrm{grav}} = -\frac{g_i g_j}{4\pi G} - \delta_{ij}\frac{|\mathbf{g}|^2}{8\pi G}. \tag{4.15}$$

If we wish to use this tensor in our self-gravitating discs, then we must use the only non-zero component, $T_{r\phi}$, vertically averaged. The Kronecker δ is zero, and we obtain

$$T_{\mathrm{r}\phi}^{\mathrm{grav}} = -\int \frac{g_r g_\phi}{4\pi G}\,\mathrm{d}z \tag{4.16}$$

We can also express this as the dimensionless α_{grav}:

$$\alpha_{\mathrm{grav}} = \frac{-1}{\Sigma c_s^2 \frac{d\ln\Omega}{d\ln r}}\left(\int \frac{g_r g_\phi}{4\pi G}\,\mathrm{d}z\right). \tag{4.17}$$

4.4.2 The Reynolds Stress Tensor

The generation of turbulence will produce a separate contribution to the local stress, as fluctuations in the local velocity field correlate. This contribution is described by the Reynolds stress tensor, which we obtain by decomposing the Navier-Stokes Equation into a mean component and a fluctuating component (often referred to as a Reynolds decomposition). This requires the following substitutions for the fluid variables:

$$\begin{aligned} \mathbf{v} &\rightarrow \bar{\mathbf{v}} + \delta\mathbf{v} \\ P &\rightarrow \bar{P} + \delta P \end{aligned} \tag{4.18}$$

For simplicity, we shall assume the fluid is incompressible, hence $\rho = const$. As we anticipate the production of a tensor from this derivation, we shall again work in index notation. The continuity equation for incompressible fluids is

$$\frac{\partial(\bar{v}_i + \delta v_i)}{\partial x_i} = 0, \tag{4.19}$$

and the Navier-Stokes equation is

$$\frac{\partial(\bar{v}_i + \delta v_i)}{\partial t} + (\bar{v}_j + \delta v_j)\frac{\partial(\bar{v}_i + \delta v_i)}{\partial x_j} = -\frac{1}{\rho}\frac{\partial(\bar{P} + \delta P)}{\partial x_i} + \nu\frac{\partial^2(\bar{v}_i + \delta v_i)}{\partial x_j^2}. \tag{4.20}$$

Note that

$$v_j\frac{\partial v_i}{\partial x_j} = \frac{\partial(v_i v_j)}{\partial x_j} - v_i\frac{\partial v_j}{\partial x_j} = \frac{\partial(v_i v_j)}{\partial x_j}, \tag{4.21}$$

where we have used the continuity equation to remove the second term. The Navier-Stokes equation then becomes:

$$\frac{\partial(\bar{v}_i + \delta v_i)}{\partial t} + \frac{\partial(\bar{v}_j + \delta v_j)(\bar{v}_i + \delta v_i)}{\partial x_j} = -\frac{1}{\rho}\frac{\partial(\bar{P} + \delta P)}{\partial x_i} + \nu\frac{\partial^2(\bar{v}_i + \delta v_i)}{\partial x_j^2} \tag{4.22}$$

Let us now take the time average of this equation. The mean components will be unchanged by time averaging: $\langle\bar{a}\rangle = \bar{a}$. The time averaged value of any fluctuation $\langle deltaa\rangle = 0$ by construction, but in general $\langle\delta a\,\delta b\rangle \neq 0$ as we do not have information as to how they correlate. Consider again the second term on the left hand side. Expanding the brackets gives:

$$\frac{\partial}{\partial x_j}\left(\bar{v}_i\bar{v}_j + \bar{v}_i\delta v_j + \delta v_i\bar{v}_j + \delta v_i\delta v_j\right). \tag{4.23}$$

As $\langle\bar{a}\delta b\rangle = \bar{a}\langle\delta b\rangle = 0$, we can remove the middle two terms. The time averaged Navier-Stokes equation therefore is

$$\frac{\partial\bar{v}_i}{\partial t} + \frac{\partial(\bar{v}_i\bar{v}_j)}{\partial x_j} + \frac{\partial(\langle\delta v_i\delta v_j\rangle)}{\partial x_j} = -\frac{1}{\rho}\frac{\partial\bar{P}}{\partial x_i} + \nu\frac{\partial^2\bar{v}_i}{\partial x_j^2}. \tag{4.24}$$

The product rule allows the second term on the left hand side to be broken into two terms, with only one term non-zero (thanks to the continuity equation). We can rearrange the equation above to recover the original form of the Navier-Stokes equation:

$$\frac{\partial \bar{v}_i}{\partial t} + \bar{v}_j \frac{\partial \bar{v}_i}{\partial x_j} = -\frac{1}{\rho}\frac{\partial \bar{P}}{\partial x_i} + \nu \frac{\partial^2 \bar{v}_i}{\partial x_j^2} + \frac{1}{\rho}\frac{\partial}{\partial x_j}\left(\rho \langle \delta v_i \delta v_j \rangle\right) \tag{4.25}$$

Revealing the form of the Reynolds stress tensor:

$$T_{ij}^{\mathrm{Reyn}} = \rho \langle \delta v_i \delta v_j \rangle. \tag{4.26}$$

The form of the tensor is linked to the correlations in velocity perturbations induced in the fluid as a result of turbulence, as we might expect intuitively before beginning any derivation. As before, we only expect the (r, ϕ) component to act in the cases we will consider. For our discs, the vertically averaged Reynolds stress tensor therefore becomes

$$T_{\mathrm{r}\phi}^{\mathrm{Reyn}} = \Sigma \delta v_r \delta v_\phi, \tag{4.27}$$

where we now use $(\delta v_r, \delta v_\phi)$ to describe vertically averaged velocity perturbations in the r and ϕ directions respectively. Equivalently, in terms of α_{Reyn}:

$$\alpha_{\mathrm{Reyn}} = \frac{1}{\Sigma c_s^2 \frac{d \ln \Omega}{d \ln r}} \left(\Sigma \delta v_r \delta v_\phi\right). \tag{4.28}$$

4.5 An Analytic Approximation for α in Self-Gravitating Discs

While these tensor components can be calculated numerically for three dimensional gravito-hydrodynamic simulations, it is generally not possible to calculate them for semi-analytic models of self-gravitating discs. The alternative in the latter case is to propose an approximation to calculate the value of α. We assume that the disc has achieved a steady state (i.e. a constant accretion rate) and is in thermal equilibrium, i.e. the heating and cooling processes in the disc are in balance. We assume that gravitational instability provides "pseudo-viscous" angular momentum transport, with a magnitude defined by

$$\alpha_{\mathrm{tot}} = \alpha_{\mathrm{grav}} + \alpha_{\mathrm{Reyn}}. \tag{4.29}$$

We further stipulate that the *local* heating and cooling must be in balance. Heating in the disc will occur as a result of viscous dissipation. The dissipation rate per unit surface Q^+ can be calculated by multiplying the stress tensor by the rate of shearing:

$$Q^+ = T_{\mathrm{r}\phi} \left| r \frac{d\Omega}{dr} \right| = \alpha \Sigma c_s^2 \Omega \left| \frac{d \ln \Omega}{d \ln r} \right|^2, \tag{4.30}$$

The radiative cooling per unit surface Q^- can be parametrised in terms of the local cooling time, t_{cool}, giving

$$Q^- = \frac{U}{t_{\rm cool}} = \frac{\Sigma c_{\rm s}^2}{\gamma(\gamma-1)t_{\rm cool}}, \tag{4.31}$$

where U is the internal energy per unit surface, and γ is the ratio of specific heats. Equating Q^+ and Q^- and rearranging gives the following expression for α (Pringle, 1981; Gammie, 2001):

$$\alpha_{\rm cool} = \left(\frac{d\ln\Omega}{d\ln r}\right)^{-2} \frac{1}{\gamma(\gamma-1)t_{\rm cool}\Omega}. \tag{4.32}$$

This approximation has been of tremendous use to a variety of semi-analytic works (Clarke, 2009; Rice & Armitage, 2009; Rice et al., 2010), as it depends only on local disc properties, and allows the use of the viscous equations of motion described in section 2.6, where

$$\nu = \alpha_{\rm cool} c_s H. \tag{4.33}$$

The principal issue for semi-analytic modellers is the inclusion of radiative transfer and an appropriate equation of state to correctly model $t_{\rm cool}$ and γ. The calculation of these properties is relatively simple if the thin disc approximation holds: the cooling time can be estimated from the optical depth, which is estimated from the midplane density and midplane temperature (as is γ). If the disc is geometrically thick, then midplane variables will not in general be representative of the disc properties, and semi-analytic models will require modification to reflect this. Equation (4.32) is not valid if the local heating and cooling is not in balance - it is conceivable that a disc may be in thermal equilibrium globally, but the local heating and cooling may be out of balance, thanks to non-local effects. I discuss this problem in the next section.

4.6 How this Approximation Can Fail - Non-local Transport

If the local pseudo-viscous approximation is successful, the value of α obtained from calculating the disc stresses directly, $\alpha_{\rm total}$

$$\alpha_{\rm total} = \alpha_{\rm Reyn} + \alpha_{\rm grav} = \frac{1}{\Sigma c_s^2 \frac{d\ln\Omega}{d\ln r}} \left(\Sigma \delta v_r \delta v_\phi - \int \frac{g_r g_\phi}{4\pi G}\,{\rm d}z\right) \tag{4.34}$$

shall be well approximated by $\alpha_{\rm cool}$ defined in equation (4.32). But, we have specifically mentioned the requirement that the transport depends on locally defined quantities only. The gravitational potential is in general not a locally defined quantity - if density structures are significant on global scales (e.g. strong amplitude spiral waves), then gravitational stresses will be non-local in origin, and the approximation will fail. To quantify this failure, we must compare the transport of angular momentum and energy in spiral density waves to viscous transport.

The viscous dissipation rate in axisymmetric discs is (cf. Pringle 1981)

$$\dot{E}_{visc} = \Omega \dot{L}_{visc} = \Omega r \nabla . T = r \frac{\partial}{\partial r}(T_{\mathrm{r}\phi}), \tag{4.35}$$

where $\dot{L}_{visc}$ represents the viscous torque. We now wish to produce a similar equation linking the torques induced by spiral density waves to the wave energy dissipation rate. To do this, we shall introduce the wave action density $\mathcal{A}$ (Shu, 1970; Cossins et al., 2009)

$$\mathcal{A} = \frac{m(\Omega_p - \Omega)|\delta\Phi|^2}{8\pi G^2 \Sigma}, \tag{4.36}$$

where the perturbed potential $\delta\Phi$ is given by Poisson's equation for thin discs (equation 2.106):

$$\delta\Phi = \frac{2\pi G \delta\Sigma}{|k|}. \tag{4.37}$$

The wave energy E_{wave} and the wave angular momentum L_{wave} are derived from the wave dynamics relations (Shu 1970; Bertin 2000 and references within):

$$E_{wave} = m\Omega_p \mathcal{A}, \tag{4.38}$$

$$L_{wave} = m\mathcal{A}, \tag{4.39}$$

where Ω_p is the pattern speed of the wave. From this it is clear that (if the pattern speed is constant), the wave energy dissipation rate is

$$\dot{E}_{wave} = \Omega_p \dot{L}_{wave}. \tag{4.40}$$

We now see the (dis)connection between viscous and wave transport. If the viscous torque is generated as a result of the spiral structures induced by gravito-turbulence, then we should expect $\dot{L}_{visc} = \dot{L}_{wave}$. If the density wave is launched at corotation (i.e. $\Omega \approx \Omega_p$), then the viscous dissipation rate and the wave energy dissipation rate are equal, and our approximation is safe. If the pattern speed is significantly different from the rotation speed, then the viscous dissipation may be much smaller than the wave dissipation (especially if the wave dissipates at large radii, where it is much more likely that $\Omega_p > \Omega$. This is confirmed by Balbus & Papaloizou (1999), who show that "anomalous flux" emerges if the pattern speed is significantly different from the rotation speed. We can see this more clearly by considering the full form of E_{wave}:

$$E_{wave} = \frac{m^2 \Sigma}{2k^2} \left(\frac{\delta\Sigma}{\Sigma} \right)^2 \Omega_p (\Omega_p - \Omega). \tag{4.41}$$

Rewriting $\Omega_p = (\Omega_p - \Omega) + \Omega$ allows us to decompose E_{wave} into two terms:

$$E_{wave} = \frac{m^2 \Sigma}{2k^2} \left(\frac{\delta\Sigma}{\Sigma} \right)^2 (\Omega(\Omega_p - \Omega) + (\Omega_p - \Omega)^2). \tag{4.42}$$

The first term is a local transport term (due to the factor of Ω) and the second term is a non-local transport term (equivalent to the "anomalous flux" of Balbus & Papaloizou 1999). We can

see straight away that as $|\Omega_p - \Omega|$ increases relative to Ω, the importance of non-local transport increases also. Equivalently, the non-local transport fraction ξ must deviate significantly from zero (Cossins et al., 2009), where

$$\xi = \frac{|\Omega - \Omega_p|}{\Omega}. \tag{4.43}$$

If the pseudo-viscous approximation is shown to fail for a given set of disc parameters, then the current breed of semi-analytic models will also fail to correctly reproduce the properties of that disc. It is therefore important to know the parameter space in which it is appropriate to use the pseudo-viscous approximation for gravito-turbulent transport, as well as the parameter space in which we expect the approximation to fail. Full, 3D hydrodynamic simulations with radiative transfer are the only way to conclusively survey these parameters, as the gravitational instability is sensitive to the radiative physics (as we have seen in previous sections).

4.7 Testing Angular Momentum Transport using SPH

I will now describe how we can test the nature of angular momentum transport using 3D SPH simulations with our hybrid method of radiative transfer. We will run a series of disc simulations, investigating a wide parameter space of disc and stellar masses, and analyse in detail the dimensionless disc stresses $\alpha_{\rm grav}$ and $\alpha_{\rm Reyn}$, comparing these to the expected analytical value $\alpha_{\rm cool}$. We will also use ξ as defined in the previous section to probe the locality of transport, as well as measuring how quasi-steady the discs are. By doing this we can delineate the regimes in which semi-analytic models can correctly function using the pseudo-viscous approximation.

4.7.1 Initial Disc Conditions

The gas discs used in this work were initialised with $500,000$ SPH particles located between $r_{\rm in} = 10$ AU and $r_{\rm out} = 50$ AU, distributed such that the initial surface density profile was $\Sigma \propto r^{-3/2}$ and with an initial sound speed profile of $c_s \propto r^{-1/4}$. We are primarily interested in considering quasi-steady self-gravitating systems, rather than systems that could fragment to form bound companions. These initial conditions (in particular the small outer disc radii) were therefore motivated by recent work suggesting that massive discs will fragment at radii beyond $\sim 60-70$ AU (Rafikov, 2005; Stamatellos et al., 2007a; Stamatellos & Whitworth, 2008; Clarke, 2009; Rice & Armitage, 2009). This result is consistent with observations that massive discs tend to have outer radii less than 100 AU (Rodríguez et al., 2005) and with observations suggesting the presence of a protoplanet at ~ 65 AU in the disc around HL Tau (Greaves et al., 2008). A summary of the disc parameters investigated can be found in Table 4.1. The simulations were selected to evaluate the α-approximation's ability to function under

Table 4.1: Summary of the disc parameters investigated in this work.

Simulation	M_* ($M_\odot$)	$q_{\rm init} = M_d/M_*$	M_d ($M_\odot$)
1	1.0	0.25	0.25
2	1.0	0.5	0.5
3	1.0	1.0	1.0
4	1.0	1.5	1.5
5	0.5	0.25	0.125
6	2.0	0.25	0.5
7	5.0	0.25	1.25
8	0.5	1.0	0.5
9	2.0	1.0	2.0

1. increasing disc-to-star mass ratio, q, and
2. increasing stellar mass, M_*

As we are interested in q, which will evolve as the star accretes from the disc, we should be rigorous and also define $q_{\rm init}$ as the value of q at the start of the simulation.

4.7.2 Resolution - Fragmentation and Artificial Viscosity

There are several resolution requirements that must be discussed. The first is the standard Jeans criterion (Bate & Burkert, 1997). As some of the discs used in this work are very massive compared to the mass of the parent star, the possibility of fragmentation exists. To ensure that potential fragmentation is resolved, the minimum Jeans mass resolvable (one neighbour group of SPH particles, around 50 in the case of the code used) must be sufficiently small:

$$M_{\rm min} = 2N_{\rm neigh}m_{\rm i} = 2M_{\rm tot}\frac{N_{\rm neigh}}{N_{\rm tot}}. \tag{4.44}$$

The minimum Jeans mass resolvable ranges between $50M_\oplus$ for the most massive disc and $4M_\oplus$ for the least massive. As it is expected that fragment masses will be typically several orders of magnitude higher than these values (Kratter et al., 2010), this establishes that the simulations would comfortably resolve disc fragmentation if it were to occur.

Perhaps more important are the resolution issues raised by artificial viscosity. It is a required component of the SPH code, but it introduces extra dissipation into the simulation. If this extra dissipation exceeds the dissipation produced by the gravitational instability, then the simulation can no longer provide useful data on angular momentum transport. We must therefore quantify the artificial viscosity in the disc, so we can identify where in the disc the artificial viscosity is likely to be lower than the effective viscosity generated by the gravitational instabilities. The linear term for the artificial viscosity can be expressed as (Artymowicz & Lubow, 1994; Murray, 1996; Lodato & Price, 2010)

$$\nu_{\rm art} = \frac{1}{10}\alpha_{\rm SPH} c_{\rm s} h, \tag{4.45}$$

where $c_{\rm s}$ is the local sound speed, h is the local SPH smoothing length, and $\alpha_{\rm SPH}$ is the linear viscosity coefficient used by the SPH code (taken to be 0.1 for this work). We can define an effective α parameter associated with the artificial viscosity by using equation (4.6) (Lodato & Rice, 2004)

$$\nu_{\rm art} = \alpha_{\rm art} c_{\rm s} H, \tag{4.46}$$

and hence combining equations (4.45) and (4.46) gives (Artymowicz & Lubow, 1994; Murray, 1996; Lodato & Price, 2010)

$$\alpha_{\rm art} = \frac{1}{10}\alpha_{\rm SPH}\frac{h}{H}. \tag{4.47}$$

This shows that where the vertical structure is not well resolved (i.e., $\frac{h}{H}$ is large), artificial viscosity will dominate. In the simulations presented here, this is likely to be the case inside ~ 10 AU, so any data inside this region can not be used. We therefore did not initially populate the region inside 10 AU and although particles will move inside 10 AU during the course of the simulations, we only consider results outside this radius.

4.8 Results and Discussion

All of the simulations presented here were evolved for 27 outer rotation periods (ORPs) [1]. This ensures that all our simulations have sufficient time to settle into quasi-steady states. In fact, the duration of these simulations ($\sim 10^4$ years) is roughly 10% of the main infall phase, during which we expect protostellar discs to be self-gravitating, and therefore we capture a significant fraction of the self-gravitating history of such discs.

We consider two free parameters, the *disc mass* $M_{\rm d}$, and the *disc-to-star mass ratio*, $q = M_{\rm d}/M_*$. Both q and the local sound speed determine whether a disc is self-gravitating or not. The sound speed is determined by the local radiative physics, in particular the optical depth to the midplane. The optical depth is a function of the disc surface density, Σ, which in turn is related to the disc mass, $M_{\rm d}$. It can then be seen that the values of both q and $M_{\rm d}$ will affect the disc's evolution under self-gravity.

Secondly, there is the issue of how to calculate $\alpha_{\rm cool}$. The hybrid method of radiative transfer allows the calculation of $t_{\rm cool}$ for each SPH particle (see Chapter 3), and therefore each particle has its own $\alpha_{\rm cool}$. However, equation (4.32) shows that particles with short cooling times (e.g., those at higher elevation from the midplane) can skew attempts to create azimuthally averaged radial profiles. Therefore, when comparing $\alpha_{\rm cool}$ with $\alpha_{\rm total}$, two quantities are considered:

[1] Outer rotation periods are defined as the rotation period at the initial outer radius of the disc, $r_{\rm out} = 50$ AU, with 1 ORP equal to 354 years.

$\alpha_{\rm cool}$, using the midplane values of $t_{\rm cool}$, Ω and γ, and $\alpha_{\rm cool}$ calculated using vertically averaged values of $\bar{t}_{\rm cool}$, $\bar{\Omega}$, and $\bar{\gamma}$. We calculate $\bar{t}_{\rm cool}$ by first averaging the specific internal energy u and its rate of change $\dot{u}$ separately, giving

$$\bar{t}_{\rm cool} = \frac{\bar{u}}{\bar{\dot{u}}}. \tag{4.48}$$

This distinction between midplane and vertically averaged values is important. Using the midplane values of $t_{\rm cool}$ allows us to determine the validity of recent 1D semi-analytic models, such as Clarke (2009) and Rice & Armitage (2009), that calculate transport properties based on the midplane temperature. The vertically averaged quantities, however, give a more accurate estimate of the rate at which the disc loses energy and allows us to establish if local heating and cooling is in balance. This will then determine if the local α-approximation is still appropriate, even if using midplane values is not.

4.8.1 The Influence of Disc Mass

To study the effect of increasing disc mass on angular momentum transport, Simulations 1, 2, 3 & 4, which share the same stellar mass ($M_* = 1 M_\odot$) are analysed together. These discs have initial masses of 0.25, 0.5, 1.0 and 1.5 $M_\odot$ respectively.

General Evolution

Despite all four simulations beginning with a wide range of disc masses, their surface density profiles do not differ greatly between $r \sim 20 - 60$ AU, as can be seen in Figure 4.2. The higher mass discs ($q_{\rm init} = 1$ & $q_{\rm init} = 1.5$) are in general much denser between $r \sim 10 - 20$ AU, indicating mass build-up in the inner regions as suggested and seen by other authors (Armitage et al., 2001; Zhu et al., 2009a; Rice et al., 2010). The lower-mass discs ($q_{\rm init} = 0.25$ & $q_{\rm init} = 0.5$) undergo a period of quiescent settling lasting approximately 2000 years, adjusting themselves by accretion onto the central star, spreading in radius (see Figure 4.1) and by cooling towards marginal instability, ultimately settling into quasi-steady, self-regulated states (Lodato & Rice, 2004).

The higher mass discs ($q_{\rm init} = 1$ & $q_{\rm init} = 1.5$) undergo several transient burst events, marked by persistently strong $m = 2$ spiral activity (see Figure 4.1). They also adjust their q more rapidly compared to the two lower mass discs, with reductions between 20-30% over approximately 10 ORPs. This is due to significant accretion, with the central star accreting a total of $0.23 M_\odot$ for $q_{\rm init} = 1$ and $0.38 M_\odot$ for $q_{\rm init} = 1.5$, and is consistent with the suggestion (Rice & Armitage, 2009; Clarke, 2009) that the mass accretion rate has a very strong dependence on surface density or, equivalently, disc mass. The discs with $q_{\rm init} > 0.5$ also spread to a much larger radius than the $q_{\rm init} \leq 0.5$ discs (which is clear in Figure 4.1), with significant fractions of mass outside 60 AU. All the discs in Figure 4.2 are stable against fragmentation, with $\beta = t_{\rm cool}\Omega \gg 3$ ($\alpha_{\rm cool} < 0.06$) at all radii (Gammie, 2001; Rice et al., 2003). The values of β as

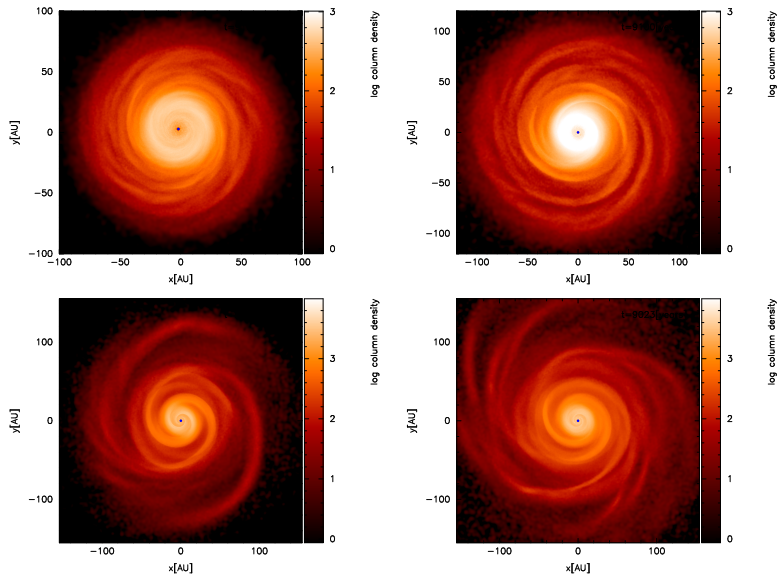

Figure 4.1: Images showing the surface density structure of Simulations 1 (top left), 2 (top right), 3 (bottom left) & 4 (bottom right) after 27 ORPs. The stellar mass in each case is 1 $M_{\odot}$, and the initial disc masses of 0.25 $M_{\odot}$, 0.5 $M_{\odot}$, 1 $M_{\odot}$ and 1.5 $M_{\odot}$ respectively. The axis ranges are shown in each figure and it is clear that the more massive discs exhibit higher amplitude spiral structures, in particular the $m = 2$ mode.

a function of opacity regime are also in good agreement with those predicted by Cossins et al. (2010).

Considering the azimuthal Fourier modes of the higher mass discs (Figure 4.3) confirms previous results regarding mode strength and disc mass ratio (Lodato & Rice, 2004, 2005; Cossins et al., 2009). The lower mass ratio discs have power distributed over a range of modes (up to $m \sim 8$) with the $m = 2$ mode (and its harmonics) becoming dominant as q increases, indicating the possibility of global transport in the discs.

The α Approximation

What we really want to establish is whether or not these discs obey the local viscous approximation. If they do, then the effective α parameter for these discs can be approximated using equation (4.32). Figure 4.4 shows the azimuthally averaged, radial α profiles for the 4 simulations in which $M_* = 1M_{\odot}$. The radial profiles in each case are also time averaged over the final 13 ORPs. In each panel, the solid line is $\alpha_{\rm total}$ computed using Equation (4.34), the dashed line the midplane $\alpha_{\rm cool}$ and the dotted line a vertically averaged $\alpha_{\rm cool}$.

In the low-mass case ($q_{\rm init} = 0.25$), it can be seen (Figure 4.4) that $\alpha_{\rm cool}$ calculated from both the midplane cooling time (dashed line) and the vertically averaged cooling time (dotted line) approximates well $\alpha_{\rm total}$, computed directly from the Reynolds and gravitational stresses. That $\alpha_{\rm total}$ increases with radius beyond $15-20$ AU is also consistent with numerical and semi-

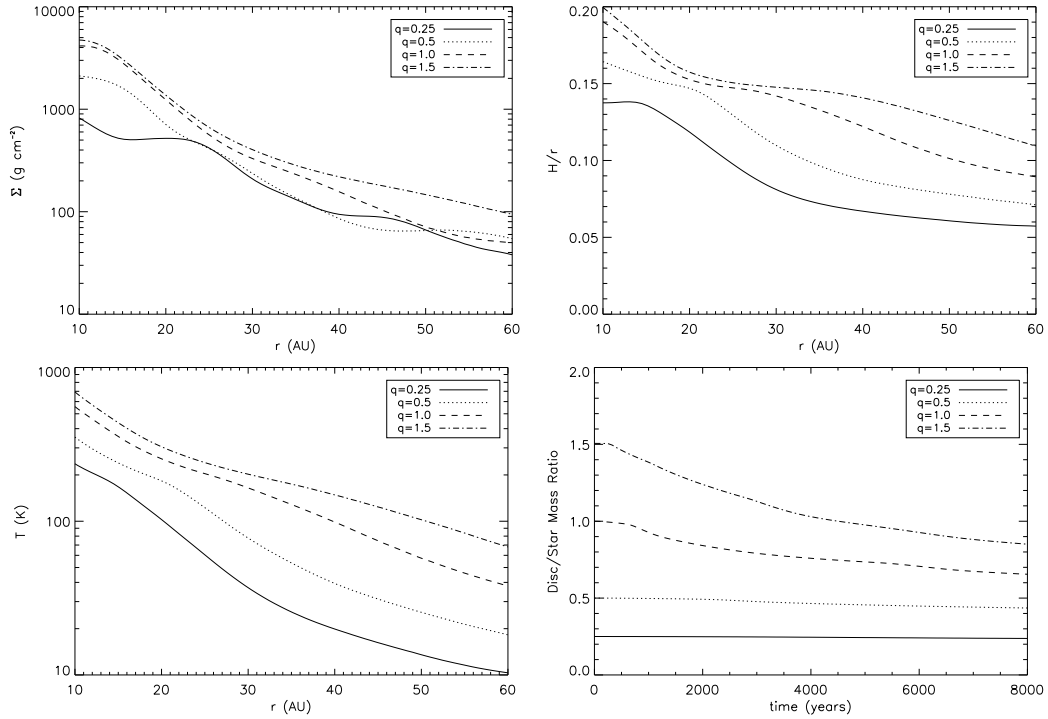

Figure 4.2: Azimuthally averaged radial profiles from the $M_* = 1M_\odot$ simulations (Simulation 1 (solid line), Simulation 2 (dotted lines), Simulation 3 (dashed lines) and Simulation 4 (dot-dashed lines)) after 27 ORPs. The figures show the time average of each variable (taken from the last 13 ORPs, to give the discs time to settle into quasi-steady states). The top left panel shows the surface density profile, the top right shows the aspect ratio, the bottom left shows the midplane temperature, and the right hand panel shows the disc-to-star mass ratio, q, as a function of time. Artificial viscosity dominates inside 10 AU, so data from inside this region is ignored.

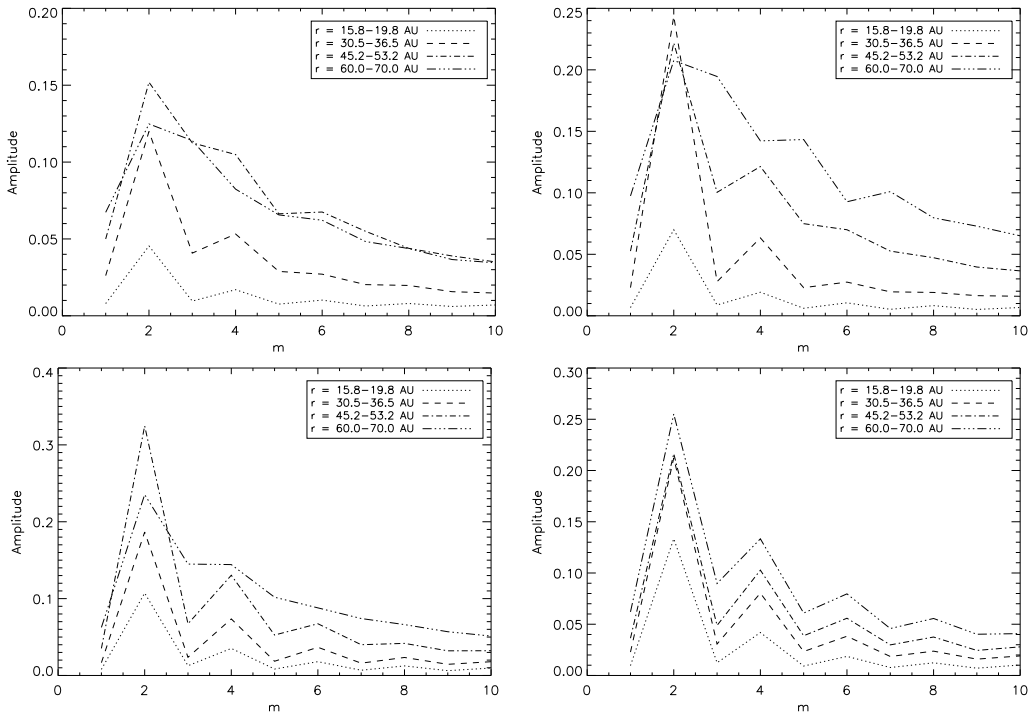

Figure 4.3: Azimuthal mode amplitudes for the $M_* = 1M_\odot$ simulations (Simulation 1 top left, Simulation 2 top right, Simulation 3 bottom left, Simulation 4 bottom right). The figures show the time average of the modes (taken from the last 13 ORPs). These figures illustrate how the $m = 2$ mode becomes more dominant as the disc-to-star mass ratio, q, increases, indicating the presence of large-scale, global spiral density waves.

analytic calculations that use the local approximation for calculating the effective gravitational viscosity (Zhu et al., 2009b; Rice & Armitage, 2009; Clarke, 2009). The same is true for $q_{\rm init} = 0.5$, but it can be seen that this approximation fails for the higher-mass discs in Simulations 3 and 4, with the profile for $\alpha_{\rm total}$ being quite different to that for the midplane $\alpha_{\rm cool}$. The vertically averaged $\alpha_{\rm cool}$ is a slightly better match to $\alpha_{\rm total}$, however, the radial profiles are quite different with $\alpha_{\rm total}$ being flatter than $\alpha_{\rm cool}$. This shows that, for the higher-mass discs, the local torque - in a time-averaged sense - is different to what would be expected if the effective viscous dissipation rate matched the local cooling rate and suggests the presence of non-local energy transport (Cossins et al., 2009). That $\alpha_{\rm total}$ exceeds the vertically averaged $\alpha_{\rm cool}$ at small radii ($r \lesssim 40$ AU), and is less than the vertically averaged $\alpha_{\rm cool}$ at larger radii ($r \gtrsim 40$ AU) suggests that energy is being transported, via global wave modes, from the inner to the outer disc.

Note that both of the high-mass simulations have disc aspect ratios above 0.1 across their entire disc radius, suggested to be a critical value by Lodato & Rice (2004) for deviations from local transport. Kratter et al. (2008) have suggested that there should be two self-gravitating α parametrisations, one for when high-m modes dominate and another for when low-m modes dominate. Our results would suggest that there is some merit in this suggestion with the local approximation being appropriate when $q_{\rm init} < 0.5$, changing to an approximately radially independent α when $q_{\rm init} > 0.5$. Fixing the value of α in the latter case appears difficult although our results may suggest that the value derived from the local approximation at $r \sim 40$ AU may be suitable.

The increase of α with decreasing radius inside 20 AU is a result of the numerical viscosity α_{art} (the triple-dot dashed lines in Figure 4.4) dominating in these inner regions, illustrating why we do not consider the region inside 10 AU. The dash-dot lines in Figure 4.4 show the effective gravitational α computed using only the gravitational stresses (i.e., $\alpha_{\rm grav} = (d\ln\Omega/d\ln r)^{-1} T^{\rm grav}_{r\phi}/\Sigma c_s^2$). This illustrates that in the inner disc, due to the dominance of the numerical viscosity (triple-dot dashed lines), the Reynolds stresses dominate over the gravitational stresses. If we were able to reduce the numerical viscosity significantly we would expect, as suggested by Zhu et al. (2009b) and Rice & Armitage (2009), that the effective gravitational α in the $q < 0.5$ simulations would continue decreasing to very small values in the inner disc, potentially leading to a pile-up of mass and periodic FU Orionis-like outbursts if the temperature in the these inner regions becomes high enough for MRI to operate (Armitage et al., 2001; Zhu et al., 2009b).

Are the discs quasi-steady?

Although the mismatch between the $\alpha_{\rm total}$ profiles and the $\alpha_{\rm cool}$ profiles in the higher-mass simulations (see Figure 4.4) suggests the presence of non-local transport, it does not tell us whether these simulations reach quasi-steady states or not. To identify how quasi-steady the discs are, the discs' temperature profiles and Toomre instability profiles are averaged over the

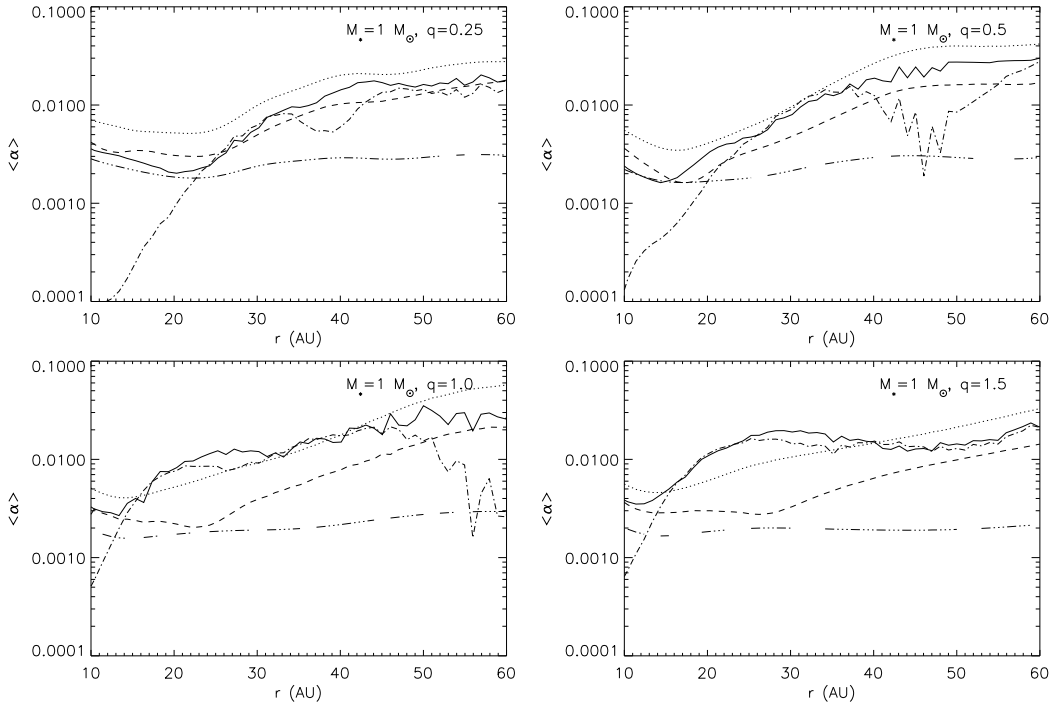

Figure 4.4: Azimuthally averaged α parameter, time-averaged over the last 13 ORPs of the simulations (Simulation 1 top left, Simulation 2 top right, Simulation 3 bottom left and Simulation 4 bottom right). The solid line indicates the α calculated from Reynolds and gravitational stresses, the dashed line indicates $\alpha_{\rm cool}$ calculated using the midplane cooling time, while the dotted line indicates $\alpha_{\rm cool}$ calculated from the vertically averaged cooling time. For illustrative purposes, we also show the stress tensor component due to gravitational instability $\alpha_{\rm grav}$, indicated by the dot-dashed line, and the the stress tensor component due to the artificial viscosity $\alpha_{\rm art}$.

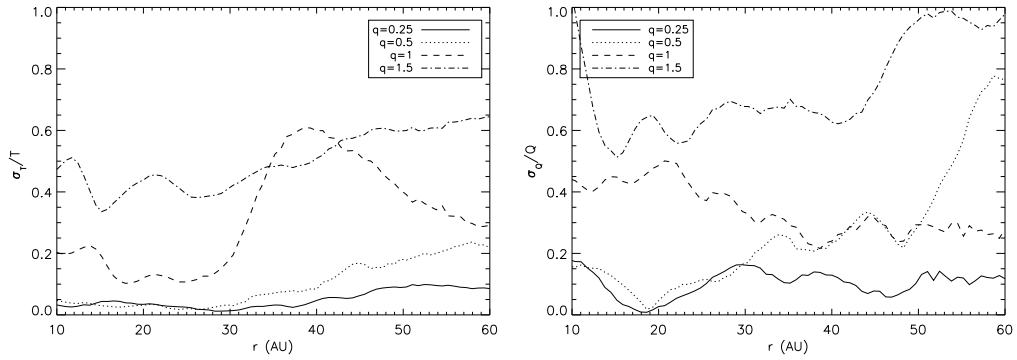

Figure 4.5: Variation in the mean temperature profile (left panel) and the mean Toomre instability profile (right panel) for the $M_* = 1M_\odot$ simulations (Simulation 1 (solid lines), Simulation 2 (dotted lines), Simulation 3 (dashed lines) and Simulation 4 (dot-dashed lines)), averaged over the last 13 ORPs.

final 13 ORPs. The standard deviation about this mean is then measured, and the normalised quantities $\sigma_{\rm T}/T$ and $\sigma_{\rm Q}/Q$ are calculated for each radius (Figure 4.5). This shows the deviation of the disc from quasi-steady, thermal equilibrium (through $\sigma_{\rm T}/T$) and its deviation from a marginally-stable, self-regulated state (through $\sigma_{\rm Q}/Q$).

Simulation 1 ($q_{\rm init} = 0.25$, solid line in Figure 4.5) shows the lowest temperature deviation, maintaining thermal balance to within around 5% except in the outer regions, where this is mainly due to the reduced value of T. A deviation of 1 K from a mean of 10 K will be more significant than from a mean of 100 K. This is also true for $q_{\rm init} = 0.5$ (dotted line in Figure 4.5), although the amplitude increases further at larger radii. The lower-mass simulations ($q_{\rm init} < 0.5$) are therefore not only local, but also settle into long-lived, quasi-steady states.

The temperature profiles for the high-mass ($q_{\rm init} > 0.5$) discs (dashed and dash-dot lines in Figure 4.5) show significant variation (varying by as much as 60% in the worst case), illustrating that these discs not only have non-local transport, but also do not attain well-defined, long-lived quasi-steady states. This implies that in these discs - at any given location - there will be periods when the dissipation rate exceeds the local cooling rate (causing the temperature to rise) followed by a period when the cooling rate dominates. This is presumably inherently linked to the global nature of the energy transport in these simulations. Energy is being transported non-locally, and is hence not being generated and dissipated at the same location, and therefore it is not possible for the local heating and cooling rates to balance at all locations in the disc.

Figure 4.5 also shows deviations from uniform Q, with again the lower-mass discs showing the lowest deviation in the inner 50 AU, averaging around 10%. Simulations 3 & 4 ($q_{\rm init} > 0.5$) again vary much more significantly, peaking at around 40%. These results show that for $q_{\rm init} > 0.5$ a disc is unable to settle into a long-lived, marginally-stable, self-gravitating state.

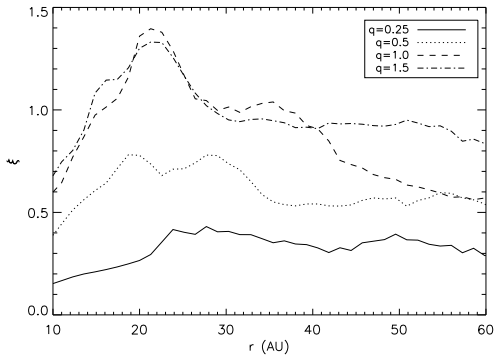

Figure 4.6: The non-local transport fraction, ξ, for the $M_* = 1M_\odot$ simulations (Simulation 1 (solid lines), Simulation 2 (dotted lines), Simulation 3 (dashed lines) and Simulation 4 (dot-dashed lines)), averaged over the last 13 ORPs.

Is the transport non-local?

Although the above suggests that there is non-local transport in the higher-mass discs, we have not yet convincingly shown that this is indeed the case. We must therefore compare the pattern speed of the dominant spiral mode Ω_p with the rotation speed Ω through the use of $\xi = \frac{|\Omega - \Omega_p|}{\Omega}$. We can calculate the numerator from the dispersion relation for finite thickness Keplerian discs (Bertin 2000; Cossins et al. 2009, see also Appendix C)

$$m^2 \left(\Omega_p - \Omega\right)^2 = c_s^2 k^2 - \frac{2\pi G \Sigma |k|}{1 + |k|H} + \Omega^2. \tag{4.49}$$

The factor of $1+|k|H$ is required as the disc thickness dilutes the vertical gravitational potential. In order to determine the dominant modes, the radial and azimuthal wavenumbers (k, m) are spectrally averaged for each radius (i.e., the average is weighted by the squared amplitude in each mode), and hence Ω_p is calculated for each radius, which allows the calculation of $\xi(r)$, shown in Figure 4.6 (where we have averaged ξ over the last 13 ORPs). As can be seen, ξ increases with increasing disc mass, exceeding 1 for $q_{\rm init} \geq 1$, illustrating that non-local transport becomes important as the disc-to-star mass ratio exceeds 0.5. The most massive disc ($q_{\rm init} = 1.5$) undergoes rapid evolution to adjust its q towards 0.85 with a flat surface density profile, ensuring that ξ is also flat out to larger radii (exceeding the $q_{\rm init} = 1$ disc outside 40 AU). The peak values of ξ at around $20 - 30$ AU are consistent with the peak deviations of $\alpha_{\rm total}$ from $\alpha_{\rm cool}$, lending weight to the conclusion that non-local effects transport energy from the inner disc to the outer disc.

4.8.2 The Influence of Stellar Mass

To disentangle the influences of disc mass and disc-to-star mass ratio, two sets of simulations are to be analysed together. The first set of discs have $q_{\rm init} = 0.25$ (Simulations 1, 5, 6 & 7), but have different stellar mass. The previous section showed that the α-approximation holds

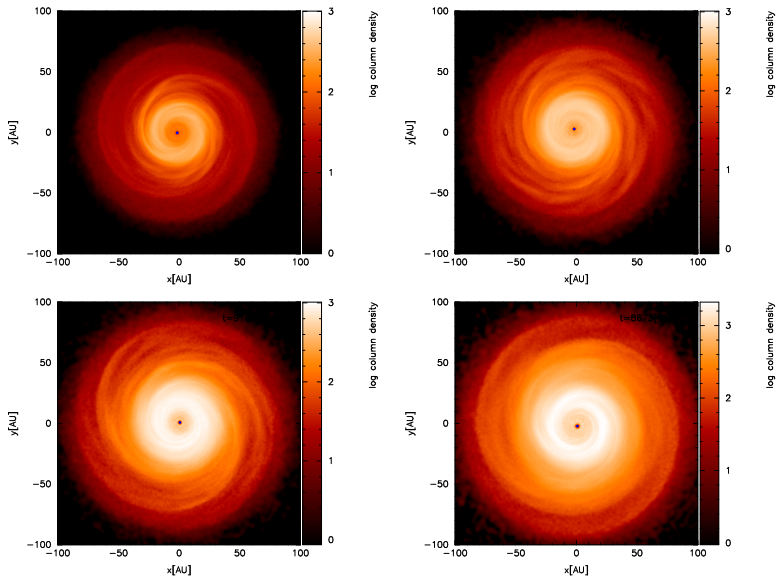

Figure 4.7: Images showing the surface density structure of Simulations 5 (top left), 1 (top right), 6 (bottom left) & 7 (bottom right) after 27 ORPs. The discs shown have initial mass ratios of $q = 0.25$, with star masses of 0.5 $M_\odot$, 1 $M_\odot$, 2 $M_\odot$ and 5 $M_\odot$ respectively.

well for Simulation 1. If disc-to-star mass ratio is the key property that determines the nature of angular momentum transport (and not the local sound speed), then the α-approximation should be equally effective for all simulations in this first set.

The second set will analyse the discs with $q_{\rm init} = 1$ (Simulations 3, 8 & 9). If q is key to the nature of angular momentum transport, then it should be expected that non-local transport should be exhibited by all the discs in the second set.

The $q_{\rm init} = 0.25$ discs

General Evolution As with the previous set of simulations, the discs undergo an initial settling phase, and become marginally-stable after a period of cooling. The low initial value of q is relatively unchanged in all simulations, with the most massive disc changing mass by less than 20% (see Figure 4.8, bottom right panel). All four simulations share similar aspect ratios - this follows from the result that the aspect ratio H/r is proportional to q during marginal instability (c.f., Lodato 2007). For this to be possible, the surface density profiles must therefore increase with disc mass, as can be seen in the top left panel. However, the radial dependence of the surface density is roughly the same for all discs. This in turn ensures the more massive discs are hotter (bottom left panel), with similar radial temperature profiles for all four simulations.

The α Approximation Repeating a similar analysis of α as was done for the $M_* = 1M_\odot$ discs, it can be seen (Figure 4.9) that the α-approximation holds with increasing stellar mass, confirming that the key parameter is the disc-to-star mass ratio, q, which is held constant

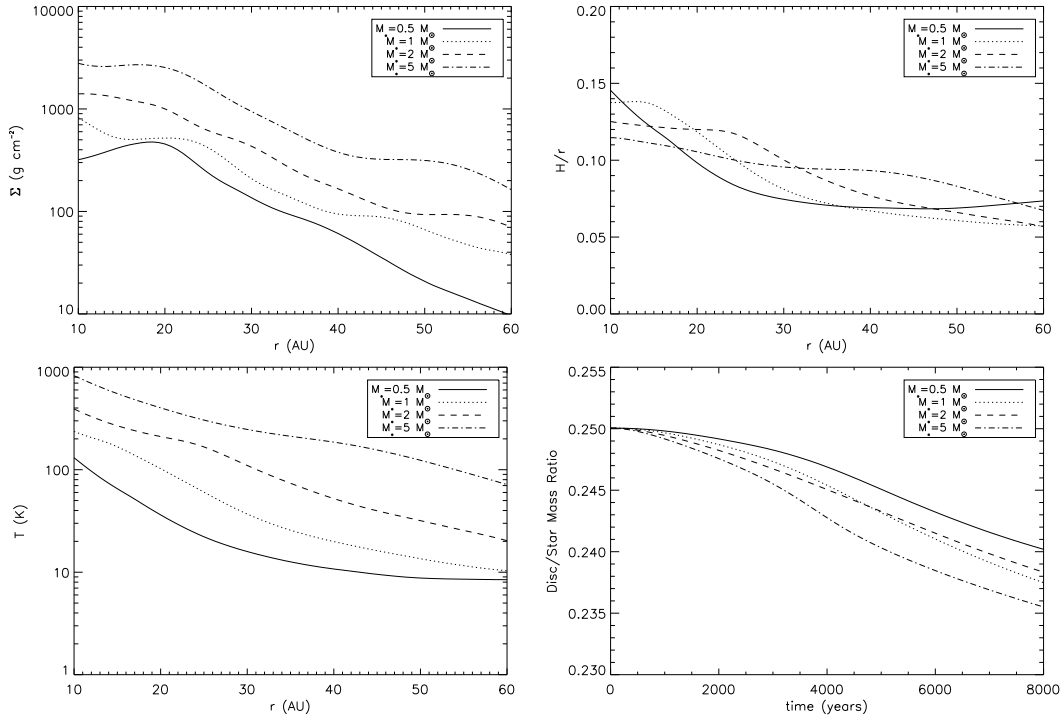

Figure 4.8: Azimuthally averaged radial profiles from the $q_{\rm init} = 0.25$ simulations (Simulation 5 (solid line), Simulation 1 (dotted lines), Simulation 6 (dashed lines) and Simulation 7 (dot-dashed lines)) after 27 ORPs. The figures show the time average of each variable (taken from the last 13 ORPs). The top left panel shows the surface density profile, the top right shows the aspect ratio, the bottom left shows the midplane temperature, and the right hand panel shows the disc-to-star mass ratio, q, as a function of time. Artificial viscosity dominates inside 10 AU, so data from inside this region is ignored.

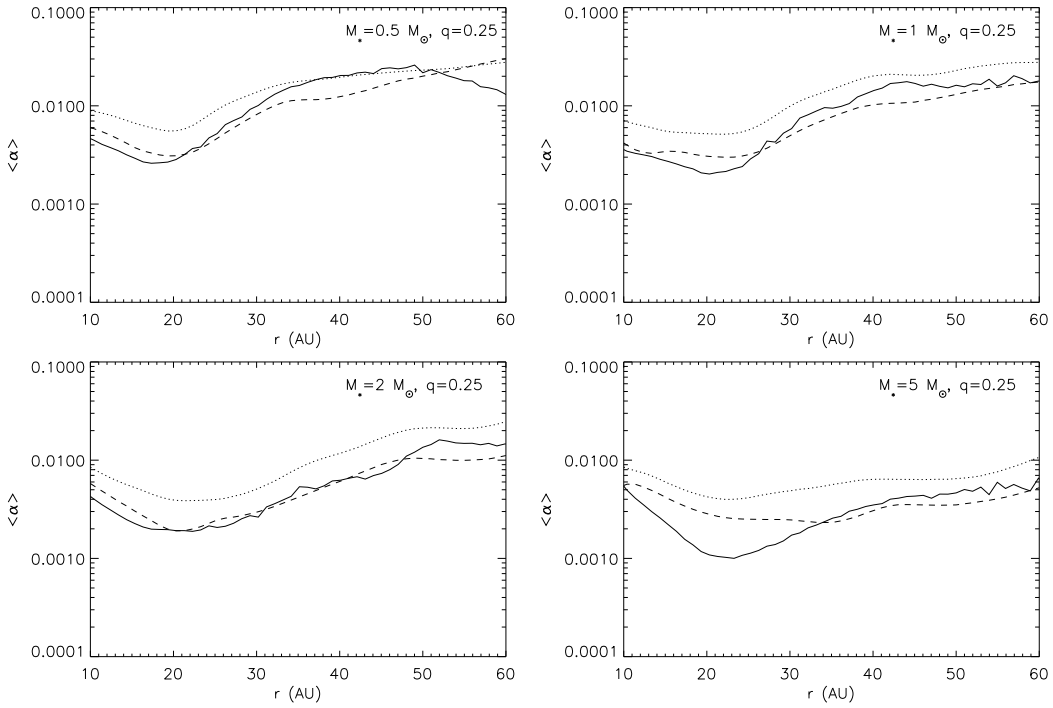

Figure 4.9: The α parameter for the $q_{\rm init} = 0.25$ simulations (Simulation 5 top left, Simulation 1 top right, Simulation 6 bottom left & Simulation 7 bottom right), averaged over the last 13 ORPs of the simulations. The black line indicates the α calculated from the Reynolds and gravitational stresses ($\alpha_{\rm total}$), the dashed line indicates $\alpha_{\rm cool}$ calculated using the midplane cooling time at that radius, and the dotted line indicates the $\alpha_{\rm cool}$ calculated from the vertically averaged cooling time.

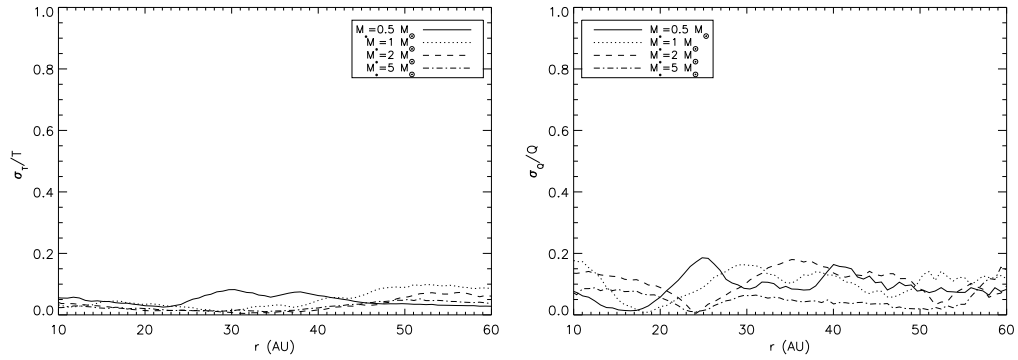

Figure 4.10: Variation in the mean temperature profile (left) and the mean Toomre instability profile (right) for the $q = 0.25$ simulations (Simulation 5 (solid line), Simulation 1 (dotted lines), Simulation 6 (dashed lines) and Simulation 7 (dot-dashed lines)), averaged over the last 13 ORPs.

here. A local approximation therefore appears to be suitable for systems in which $q_{\rm init} < 0.5$. Simulation 7 in which $M_* = 5M_\odot$ suggests that there may be some dependence on the stellar mass as the calculated $\alpha_{\rm total}$ is somewhat lower than the expected $\alpha_{\rm cool}$ inside 30 AU. The aspect ratio of this disc is, however, quite flat with $H/r > 0.1$ for a much wider radial range than in the other simulations. The region where the aspect ratio exceeds 0.1 corresponds with the region where $\alpha_{\rm total}$ deviates from the expected values, consistent with previous analysis (Lodato & Rice, 2004) suggesting that the local approximation is suitable when $H/r < 0.1$.

The local and quasi-steady assumptions Figure 4.10 also shows that, for $q_{\rm init} = 0.25$, the temperature fluctuates by less than 10% and Q fluctuates by 10% - 20%, over the final 13 ORPs. This illustrates that all these discs settle into quasi-steady states that are marginally stable. The non-local transport fraction (Figure 4.11) also remains low. The seemingly high ξ for $M_* = 0.5M_\odot$ is due to its slightly elevated mass ratio in comparison to the other discs (Figure 4.8, bottom right panel). This, coupled with its comparatively lower sound speed and lower surface density (with the scale height kept constant) will boost the non-local transport fraction to a higher value than expected *ab initio*. However, its maximum value is still below that of the $q_{\rm init} = 0.5$ disc studied in this analysis (Simulation 2), so this is not inconsistent with expectations.

The $q_{\rm init} = 1$ discs

General Evolution Figure 4.12 shows the profiles of the $q_{\rm init} = 1$ discs, averaged over the final 13 ORPs. The initial stellar masses are $M_* = 0.5M_\odot$, $M_* = 1M_\odot$, and $M_* = 2M_\odot$. The discs grow hotter with increasing disc mass (with a flatter temperature profile), while maintaining a similar surface density profile. This results in the higher disc mass simulations obtaining a flatter aspect ratio (top right panel).

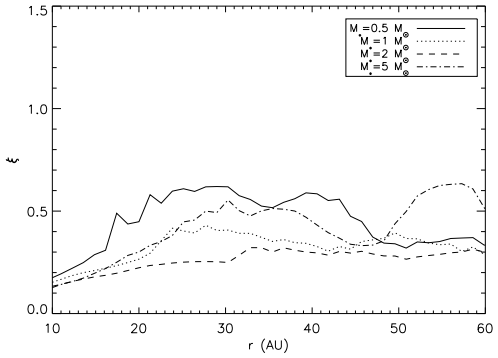

Figure 4.11: The non-local transport fraction for the $q_{\rm init} = 0.25$ simulations (Simulation 5 (solid line), Simulation 1 (dotted lines), Simulation 6 (dashed lines) and Simulation 7 (dot-dashed lines)), averaged over the last 13 ORPs.

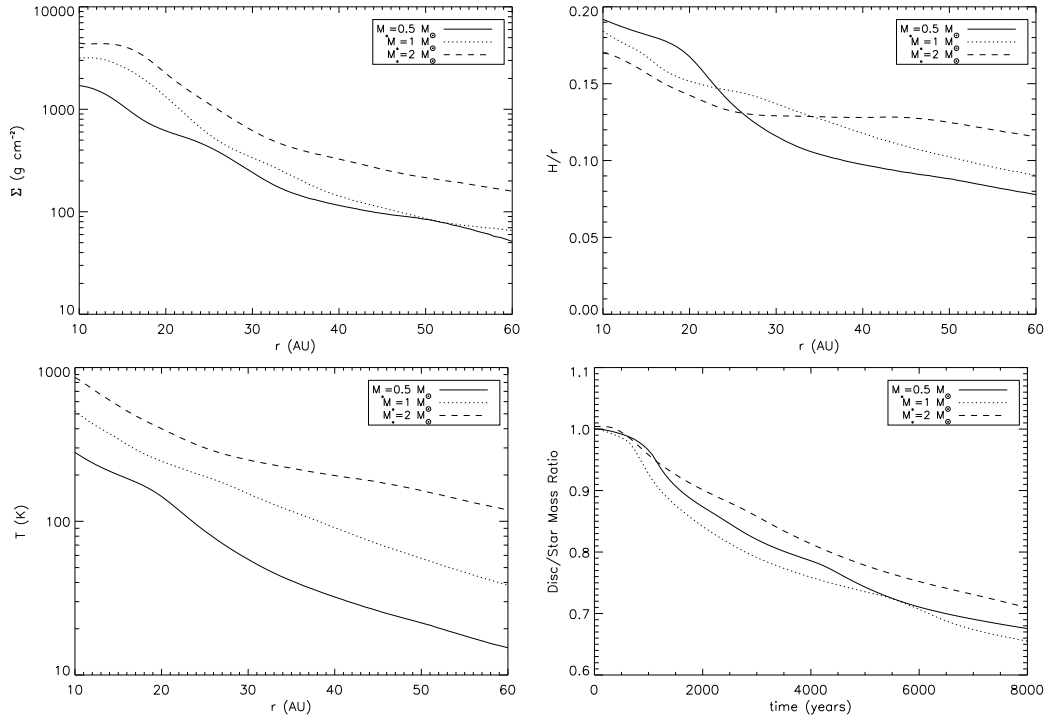

Figure 4.12: Azimuthally averaged radial profiles from the $q_{\rm init} = 1$ simulations (Simulation 8 (solid line), Simulation 3 (dotted lines), and Simulation 9 (dashed lines)). The figures show the time average of each variable, taken from the last 13 ORPs. The top left panel shows the surface density profile, the top right shows the aspect ratio, the bottom left shows the midplane temperature, and the right hand panel shows the disc mass ratio q as a function of time. Artificial viscosity dominates inside 10 AU, so data from inside this region is ignored.

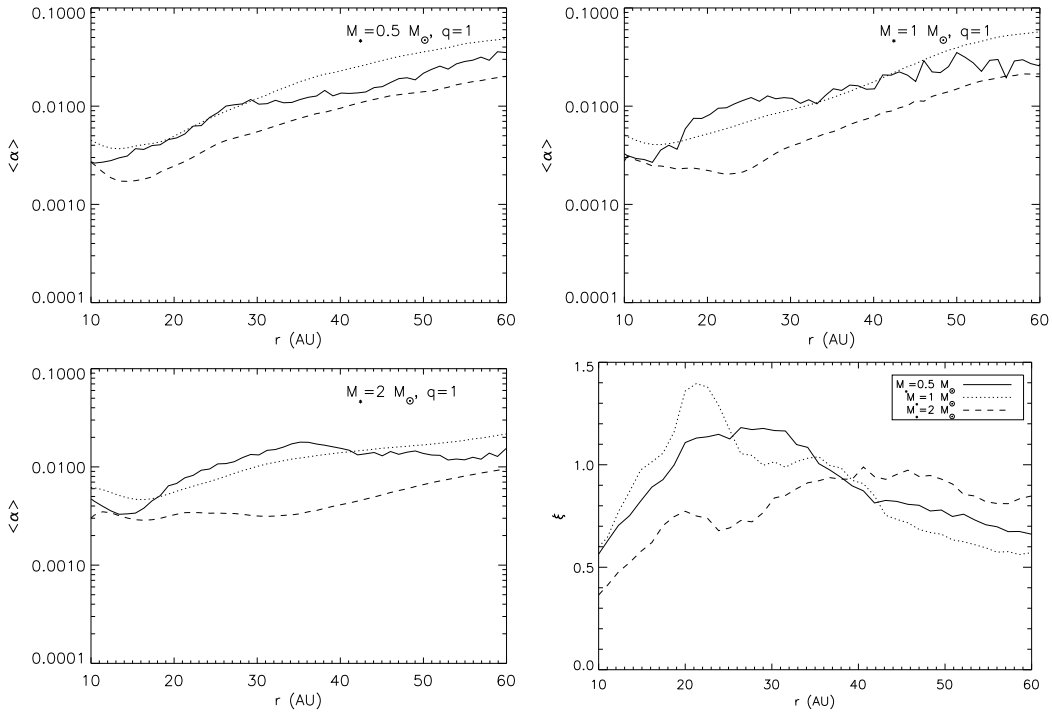

Figure 4.13: The α parameter (Simulation 8 top left, Simulation 3 top right, & Simulation 9 bottom right), averaged over the last 13 ORPs of the simulations. The black line indicates the alpha calculated from Reynolds and gravitational stresses, the dashed line indicates the alpha calculated by the midplane cooling time at that radius, and the dotted line indicates the alpha calculated from the vertically averaged cooling time. The bottom right panel shows the non-local transport fraction for the $q_{\rm init} = 1$ simulations (Simulation 8 (solid line), Simulation 3 (dotted lines) and Simulation 9 (dashed lines)), averaged over the last 13 ORPs.

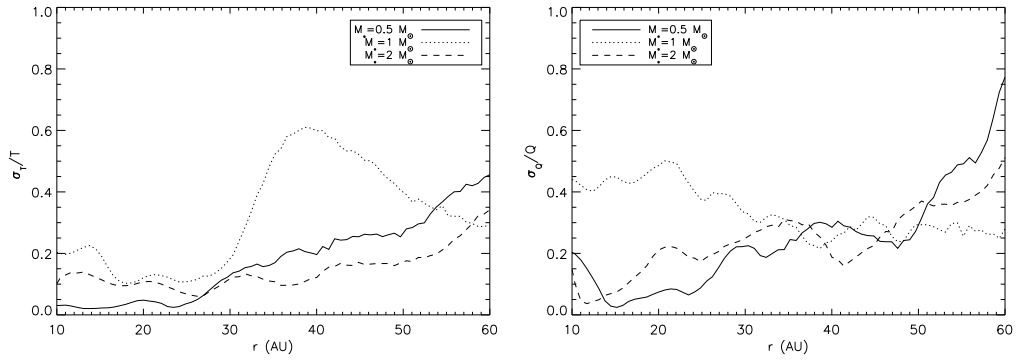

Figure 4.14: Variation in the mean temperature profile (left) and the mean Toomre instability profile (right) for the $q = 1$ simulations (Simulation 8 (solid line), Simulation 3 (dotted lines), and Simulation 9 (dashed lines)), averaged over the last 13 ORPs.

The α Approximation Figure 4.13 shows that all the discs have similar qualitative α profiles, with $\alpha_{\rm total}$ being different to what would be expected if the local approximation were appropriate ($\alpha_{\rm cool}$). The value of the enhancement appears to increase with increasing disc mass, showing that while q dictates whether or not a disc deviates from the local approximation, the disc mass $M_{\rm d}$ controls the strength of this deviation (through its influence on Σ and ultimately the disc thickness). All three discs have aspect ratios in excess of 0.1 for most of their radial extent, again consistent with previous predictions for non-locality (Lodato & Rice, 2004).

The local and quasi-steady assumptions Figure 4.14 also shows that the quasi-steady approximation also appears to be violated (Figure 4.14). The temperature fluctuates at values of $\sim 20\%$ and higher, with similar fluctuations in Q. The non-local transport fraction (bottom right panel in Figure 4.13) in all three cases is ~ 1 or larger showing that the transport is very non-local.

4.9 Conclusions

This work has studied in detail whether a local, viscous approximation can accurately model the angular momentum transport in realistic, radiative, self-gravitating protostellar discs. For the viscous approximation to hold, the angular momentum transport must be local. If the analytical results of Gammie (2001) and others also hold (which calculate the stresses using the assumption that the dissipation rate matches the local cooling rate), the discs must also be in approximate thermodynamic equilibrium.

A series of simulations using SPH with radiative transfer were carried out, and the effective viscosity generated by the gravitational instability was calculated directly from the Reynolds and gravitational stresses in the simulated discs. This was then compared with the expected viscosity, based on the assumption of local thermodynamic equilibrium, and the results analysed

as a function of increasing disc-to-star mass ratio, and increasing stellar mass.

The results show that if the discs have an initial disc-to-star mass ratio $q_{\rm init} < 0.5$, and are geometrically thin ($H/r \leq 0.1$), the local viscous approximation performs well in calculating the angular momentum transport. Such discs are shown to have a low non-local transport fraction (Cossins et al., 2009), moderate azimuthal Fourier mode amplitudes up to $m \sim 8$ (with increased power at $m = 2$), and maintain a strictly self-regulated, quasi-steady state (Lodato & Rice, 2004, 2005). It has also been demonstrated that increasing stellar mass (while keeping q constant) does not significantly affect the efficacy of the viscous approximation, holding over at least an order of magnitude in stellar mass. There is, however, some suggestion that there is some dependence on stellar mass with the $M_* = 5M_\odot$ simulation showing some evidence for non-local transport corresponding to regions of the disc where $H/r > 0.1$.

However, if the disc-to-star mass ratio $q_{\rm init} > 0.5$, the azimuthal $m = 2$ spiral modes begin to dominate. The strength of these global spiral waves introduces strong non-local torques, and are also subject to transient burst events. The disc stresses calculated show that locally, in a time averaged sense, the amount of energy released through cooling does not match the thermal energy generated by the instability. It is likely that this excess energy is transported by the low-m mode global waves to larger ($r \geq 40$ AU) radii where it can be lost through radiative cooling. This is a clear indication of global effects and is confirmed by their high non-local transport fractions (Cossins et al., 2009). Together, these violate the assumptions made to satisfy the viscous approximation.

In summary, semi-analytic models are justified in using the viscous approximation to model realistic self-gravitating protostellar discs, provided that the parameter space studied does not include discs that are too massive, or geometrically thick. The current semi-analytic models (Clarke, 2009; Rice & Armitage, 2009) in which the midplane cooling time is used to determine the effective gravitational α will, however, certainly underestimate the value of the effective viscosity in massive, geometrically thick discs.

Chapter 5

Protoplanetary Discs and Stellar Encounters

The unquiet republic of the maze of planets, struggling fierce towards heaven's free wilderness.
Percy Bysshe Shelley, *Prometheus Unbound*

5.1 Author's Note

This chapter describes work previously published in Forgan & Rice (2009) and Forgan & Rice (2010c). To maintain consistency (and to prevent repetition), these papers have been rewritten into a single narrative.

5.2 The Importance of Stellar Encounters

Stars are not typically born in isolation. The turbulent structures in GMCs from which they form lead to the formation of multiple stars in clusters (Lada & Lada, 2003). Multiple systems such as binaries are also common (Duquennoy & Mayor, 1991). In these regions of increased stellar density, the possibility that stars will come close enough to each other to induce gravitational perturbations becomes much more likely. Thies et al. (2005) show that a disc in a typical open cluster of 1000 stars (within a characteristic radius of 0.5 pc) will undergo encounters within 500 AU every 10 Myr. This is of order the lifetime of the gaseous component in circumstellar discs, indicating the possibility of perturbations to the disc at early times in its evolution.

We will consider two possible consequences of such encounters, which are of importance both for theory and observation.

5.2.1 Encounters and Disc Fragmentation

To recap the results of section 2.9.2, a self-gravitating (Keplerian) disc will fragment if the following two conditions are satisfied:

$$Q = \frac{c_s \Omega}{\pi G \Sigma} \leq 2 \tag{5.1}$$

$$t_{\rm cool}\Omega = \beta < \beta_{crit} \quad \Leftrightarrow \quad \alpha \gtrsim 0.06, \tag{5.2}$$

where α is the Shakura-Sunyaev viscosity parameter discussed at length in Chapter 4. In isolated quasi-steady discs, Σ and Ω can be considered to be constants. Therefore, the only viable route to fragmentation is through reducing c_s and t_{cool}. However, if the disc can be perturbed out of equilibrium, then Σ and Ω can be modified and the routes to fragmentation multiply. Stellar encounters can significantly perturb protostellar discs, generating non-axisymmetric features such as tidal arms. Being able to increase Σ locally in this fashion may provide a trigger for local gravitational collapse, if the local optical depth remains small (giving a short cooling time). This cooling time criterion demands an accurate treatment of the local disc thermodynamics. Compression of the disc by a companion should occur on the dynamical timescale, which is shorter than the thermal timescale. Hence, the disc's subsequent evolution will be adiabatic, and fragmentation will not occur unless the cooling time can be reduced significantly by the compression, i.e.

$$\left(\frac{dt_{cool}}{d\Sigma} \right)_S < 0. \tag{5.3}$$

This may be possible if the subsequent temperature increase pushes the gas into the "opacity gap" (where ice and dust evaporate, see Figure 3.5), but the decrease in opacity reduces in magnitude as the density of the gas is increased, so fragmentation still appears to be unlikely (Johnson & Gammie, 2003). However, while high values of β prohibit fragmentation, given an amenable temperature and opacity regime β can be rapidly decreased towards β_{crit} in one dynamical timescale. This implies that inside the opacity gap, initial values of β can be as high as ~ 90 and still cool rapidly to produce fragmentation (Cossins et al., 2010). Also, Clarke et al. (2007) have shown that decreasing β slowly toward β_{crit} (as would be the case for isolated discs) stabilises the disc against fragmentation, so triggered fragmentation may be more successful.

The differing physical processes at work (coupled with its deeply non-axisymmetric nature) have prevented fully analytic studies of encounter-driven fragmentation. Simulations have shown encounters will destabilise and fragment isothermal extended discs (e.g. Boffin et al. 1998; Watkins et al. 1998a,b; Lin 1998), whereas encounters with compact, adiabatic discs have been shown to be stable against fragmentation, suppressing instability through compressive and shock heating (e.g. Lodato et al. 2007). Until recently, most analyses of the

problem have not considered the full effects of radiative transfer, or have used simple β cooling parametrisations (Lodato et al., 2007). The local thermodynamics (and the thermal history) are crucial to the fragmentation problem, and therefore any conclusive study must incorporate these if it is to succeed. I detail the results of such a study in section 5.4.

5.2.2 Encounters and Outburst Phenomena

Conservation of angular momentum will in general ensure that, within typical free-fall times of $\sim 10^5$ yr, molecular cloud collapse produces a protostar with protostellar disc. These formation rates are consistent with the observed statistics of protostellar objects, for example those in Taurus (Kenyon et al., 1990) and with numerical simulations (Bate et al., 2003; Stamatellos et al., 2007a; Bate, 2009). If these formation timescales are converted to average mass accretion rates, then it appears that the standard picture will form a star at the rate of $10^{-5} M_{\odot}\, yr^{-1}$, whereas current observations suggests an average infall rate of $10^{-6} M_{\odot}\, yr^{-1}$ or lower. This leads to the realisation that accretion rates in protostars are not constant - which has been confirmed observationally (Herbig, 1977; Armitage et al., 2001; Zhu et al., 2009a) - and that short periods of increased accretion, accompanied by periods of mass pile-up where the infalling matter is not accreted, can solve the apparent inconsistencies between observation and theory.

This also provides an explanation for the FU Orionis (FU Ori) outburst objects, Class 0 to Class II protostellar objects which undergo a characteristic rapid rise in luminosity spanning up to five magnitudes, and then decay on timescales of a few hundred years (Hartmann & Kenyon, 1996). Maximum accretion rates for FU Ori objects are typically $10^{-4} M_{\odot}\, yr^{-1}$ or more (Herbig, 1977; Hartmann & Kenyon, 1996), which show a strong increase over typical infall rates (Kenyon et al., 1993; Furlan et al., 2008). Observations are beginning to reveal that outburst phenomena appear to belong to different classes, e.g. FU Ori (FUors), EX Lupi (EXors) (Herbig, 2007), and others.

The precise origin of FU Ori outbursts is not known, although almost all theories involve a protostellar disc, as these discs are expected to be present around the majority of early-type stars. The thermal instability model (Lin et al., 1985; Bell & Lin, 1994) relies on the inner disc being hot enough to ionise hydrogen (at temperatures of order 10^4 K). The opacity is very sensitive to temperature changes in this regime, and is thermally unstable (see Figure 3.5). Decreasing the local molecular weight will increase the local sound speed, and increase the local effective viscosity (as $\nu = \alpha c_s H$). The increase in viscosity results in a burst of mass accretion, which persists until the disc cools sufficiently for hydrogen to return to its neutral state, and the instability ceases. This model explains many features of FU Ori, including the expected repetition of outbursts, but does not fully explain their rapid rise times.

If the disc is self-gravitating, we cannot ignore gravitational instability as a source of variable accretion (Vorobyov & Basu, 2005, 2006, 2008; Boley & Durisen, 2008). This is particularly true for more massive discs (see Chapter 4), with the potential for shocks to drive thermal instability as a result. The concept of instabilities triggering instabilities has also been extended to the

Table 5.1: Summary of the orbital parameters investigated in this chapter.

Sim.	$M_d/M_\odot$	$\Sigma \propto r^{-x}$	$M_2/M_\odot$	Calc. R_{peri} (AU)	Actual R_{peri} (AU)	e	Dir'n	$i°$
1	0.1	1	0.1	40	28	1	Pro	0°
2	0.1	1	0.1	30	25	1	Pro	0°
3	0.1	1	0.1	50	34	1	Pro	0°
4	0.1	1	0.1	100	100	1	Pro	0°
5	0.2	1	0.1	50	36	1	Pro	0°
6	0.1	1	0.1	50	40	1	Retro	0°
7	0.1	1	0.1	30	30	1	-	90°
8	0.1	1	0.1	30	33	7	Pro	0°
9	0.1	1	0.5	40	10	1	Pro	0°
10	0.1	1.5	0.1	40	35	1	Pro	0°

case of gravitational instability triggering the magnetorotational instability (MRI) (Armitage et al., 2001; Zhu et al., 2009a), where the gravitational stresses decrease with proximity to the star (e.g. Figure 4.4). This decrease in stress requires the local surface density to increase to maintain a steady state accretion rate (as $\dot{M} \sim \alpha\Sigma$). This "pile-up" of material will increase in temperature to maintain equilibrium. This eventually results in ionisation, which activates MRI and increases the local stress (as MRI stresses are typically several orders of magnitude higher than gravitational stresses, cf. Chapter 4). As with the thermal instability case, this increased stress results in a burst of accretion.

All these theories agree that triggering one or several disc instability events is crucial to providing the observed accretion rates that cause the outburst. I aim not to identify the correct theory of FU Ori objects, but to return to what was once considered a potential cause of FU Ori: the encounter of a primary star plus protostellar disc with a discless secondary (Kenyon et al., 1988; Bonnell & Bastien, 1992). The possibility of a stellar encounter imparting angular momentum to the disc would boost the local stresses and potentially enhance the accretion rate, with a repeatability linked to the period of the secondary's orbit. I propose that stellar encounters could produce outbursts that share many observational features with FUors and EXors (as was suggested by Pfalzner (2008), combining treecode simulations with cluster dynamics), and yet may belong to a different class. Similar theories also exist for massive planets and brown dwarfs (Clarke & Syer, 1996; Lodato & Clarke, 2004), but I will only consider stellar companions in this work.

5.3 The Simulations

Originally, it was intended that the disc conditions used by Lodato et al. (2007) would be used to facilitate a direct comparison of techniques. However, these initial conditions are scale-free,

which poses problems when radiative transfer is simulated. The Lodato et al. (2007) parameters were therefore modified to ensure that the disc is marginally stable. In essence, this required decreasing the star mass, and increasing the radial extent of the disc. The disc extends from $r_{in} = 1$ AU to $r_{out} = 40$ AU. The disc has a mass of $0.1\,M_{\odot}$, with a central primary star of mass $0.5\,M_{\odot}$. The initial surface density profile was chosen to be $\Sigma \propto r^{-1}$, with a sound speed profile of $c_s \propto r^{-\frac{1}{4}}$. Two variants of the disc were also run: one with a disc mass of $0.2\,M_{\odot}$, and one with $\Sigma \propto r^{-3/2}$.

The discs were evolved in isolation for several Outer Rotation Periods (ORPs). This allows the disc to approach an equilibrium state and become marginally stable (Lodato & Rice, 2004; Forgan et al., 2010), and to develop steady-state spiral structures (with the exception of the $\Sigma \propto r^{-3/2}$ disc, see Figure 5.1). As the discs are either stable or marginally stable, the discs will require external stimuli if fragmentation is to occur.

5.3.1 The Stimulus: Adding a Companion

With the disc evolved into a quasi-steady state, a secondary star was then added (at a separation of 100 AU to the primary, in order to prevent any non-linear perturbations in the disc by the secondary's sudden appearance). The secondary was added with a variety of orbital parameters, comprising a suite of 10 simulations (see Table 5.1 for details). Of the encounters simulated, eight were coplanar, in prograde motion; one was coplanar, with retrograde motion; and one was inclined at an angle of 90°. Eight simulations were conducted using the standard 0.1 $M_{\odot}$ disc, one was conducted using the 0.2 $M_{\odot}$ disc, and one using the $\Sigma \propto r^{-3/2}$ disc. This provides data on how disc perturbations vary with the disc mass, surface density profile, the secondary's periastron, eccentricity, inclination, and the relative angular momenta of the secondary and the disc.

5.3.2 Resolving Disc Fragmentation

As I have mentioned several times in previous chapters, SPH simulations of disc fragmentation must correctly resolve the fragments. Bate & Burkert (1997) show that SPH correctly reproduces fragmentation if the minimum resolvable mass,

$$M_{min} = 2N_{neigh}m_i = 2M_{tot}\frac{N_{neigh}}{N_{tot}}, \tag{5.4}$$

is less than the Jeans mass of interest (m_i is the mass of a single SPH particle, N_{tot} is the total number of particles, N_{neigh} is the typical number of nearest neighbours for a single SPH particle, and M_{tot} is the total mass). The Jeans mass can be estimated using the fact that the most unstable wavelength to gravitational instability is the disc's scale height, H (Lodato, 2007; Cossins et al., 2009). This gives a characteristic Jeans Mass $M_J = \Sigma H^2 = M_{tot}\left(\frac{H}{R}\right)^2$, and hence

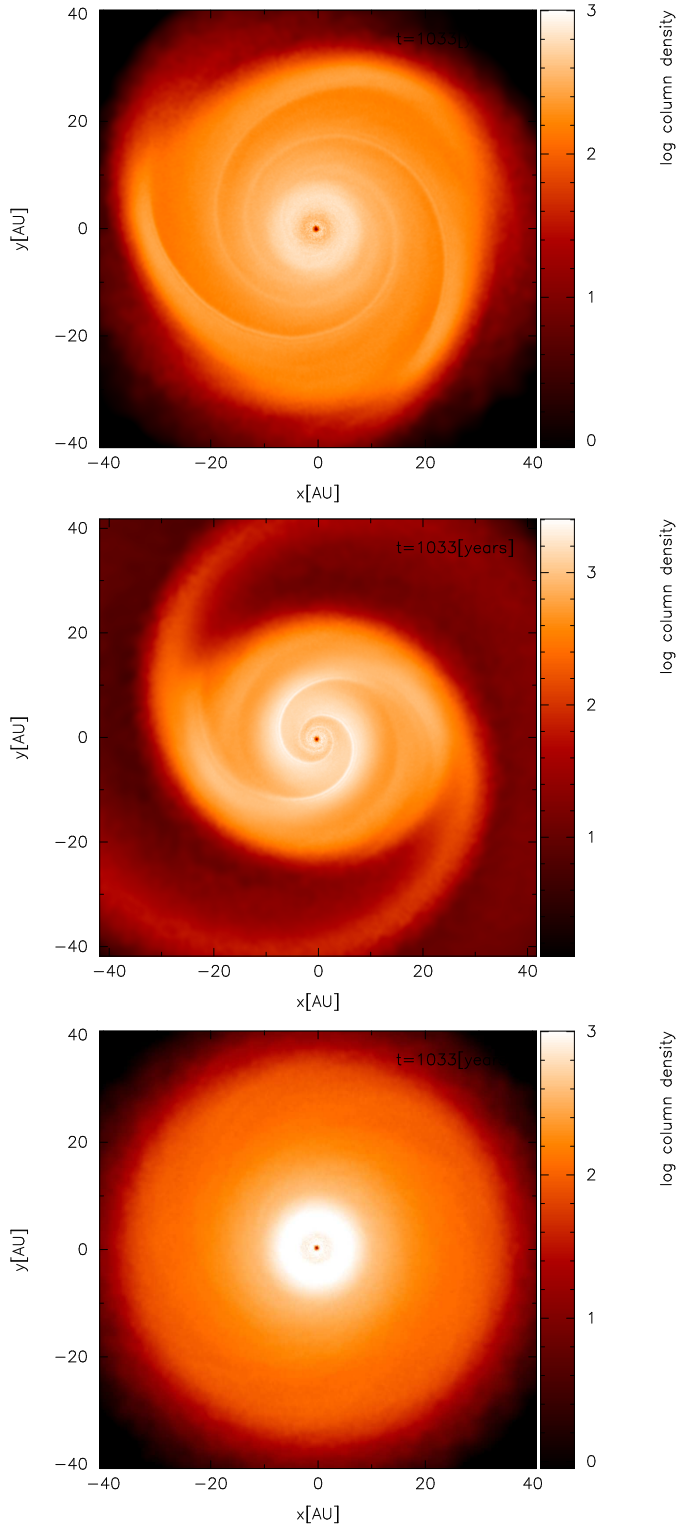

Figure 5.1: Snapshots of the three discs used after several ORPs: the $0.1\,M_{\odot}$ disc (top), the $0.2\,M_{\odot}$ disc (middle) and the $\Sigma \propto r^{-3/2}$ disc (bottom). Note each plot has a specific colour bar.

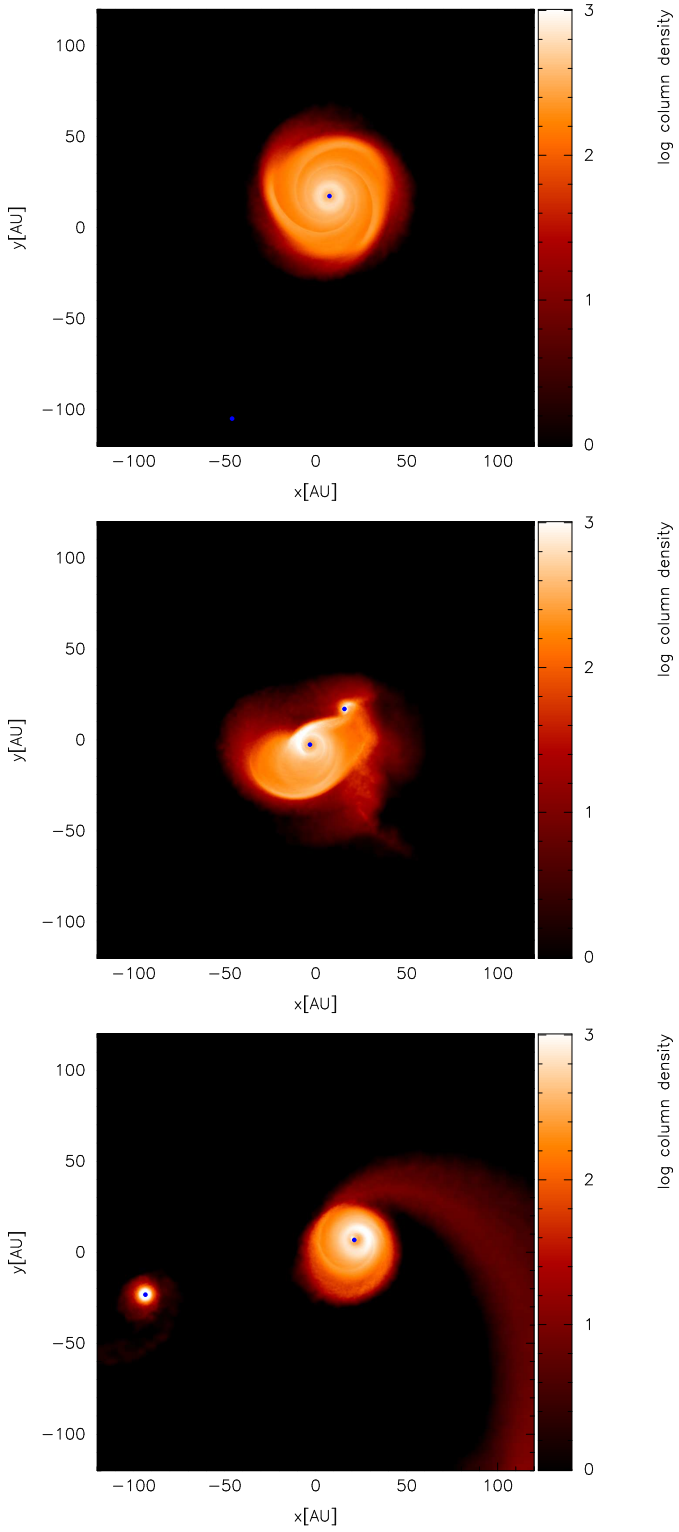

Figure 5.2: Images of the Simulation 1 disc before, during, and after the encounter.

$$\frac{M_J}{M_{min}} \approx \frac{1}{2}\left(\frac{H}{R}\right)^2 \frac{N_{tot}}{N_{neigh}} \approx \frac{1}{2}\left(\frac{M_d}{M_*}\right)^2 \frac{N_{tot}}{N_{neigh}} \tag{5.5}$$

All the simulations use 500,000 particles, and have $N_{neigh} = 50$. Using $M_d = 0.1M_\odot$, and $M_* = 0.5M_\odot$, this gives $M_J \approx 200M_{min}$. With the Jeans mass of interest being two orders of magnitude above the minimum resolved mass, this shows that all simulations carried out in this work easily satisfy the necessary resolution conditions.

5.3.3 Resolving Mass Accretion

To correctly model the accretion of matter onto both masses, the available feedstock for accretion must be identified. The primary can accrete from the inner regions of the disc, which are the densest (and hence best represented by SPH particles). In the innermost annulus (say up to 0.5 AU from the disc's inner edge), there is 0.015 $M_\odot$ of material (given a disc of mass 0.1 $M_\odot$ and a surface density profile $\Sigma \propto r^{-1}$). These simulations use 500,000 SPH particles: this mass therefore corresponds to around 80,000 particles. The minimum mass element that SPH can resolve is typically one nearest neighbour group (e.g. Bate & Burkert 1997), which in this work corresponds to 50 particles. Therefore, the accretion feedstock for the primary constitutes around 1600 nearest neighbour groups, comfortably above the minimum mass resolution. The problem of increasing numerical viscosity in the inner regions is unfortunately insurmountable (Clarke, 2009; Lodato & Price, 2010; Forgan et al., 2010). The artificially high viscosity in the inner regions will prevent mass piling up (e.g. Rice & Armitage 2009; Rice et al. 2010), so it should be expected that the primary's peak accretion rate will be underestimated (although the total mass accreted may not be affected, as the accretion will begin earlier, and for a longer duration). Comparing the contributions to the viscosity parameter α (Shakura & Sunyaev, 1973) from Reynolds stresses and from gravity shows that artificial viscosity dominates these discs within the inner 10 AU (see Chapter 4).

The secondary can accrete from matter it encounters along its trajectory: the matter must come sufficiently close to become bound to the secondary before accretion is possible. This defines the secondary's feedstock locale (approximately) as a semi-annulus in the disc, centred on periastron, with upper and lower radial boundaries based on the secondary's gravitational influence. It is a semi-annulus because only disc material in the correct orbital phase will come into close proximity with the secondary. Material on the opposite side of the disc during the encounter is typically teased into a tidal tail that is not accreted by either body.

In more rigorous terms, a particle i must satisfy the following for capture by the secondary:

$$E = E_{kin} + E_{pot} = \frac{1}{2}m_i v_{i2}^2 - \frac{GM_2 m_i}{r_{i2}} < 0, \tag{5.6}$$

where M_2 is the secondary mass, and (r_{i2}, v_{i2}) are the position and velocity of the gas element relative to the secondary respectively. Assuming that the majority of capture occurs around periastron, then

$$v_{i2} \approx |v_{i1} - v_{2,peri}| \tag{5.7}$$

$$r_{i2} \approx |r_{i1} - r_{2,peri}|, \tag{5.8}$$

where r_{i1} is the separation between the gas element and the primary, and v_{i1} is the velocity of the gas relative to the primary. If the disc is Keplerian (and no radial motion is assumed), then

$$v_{i1} = r_{i1}\Omega_{i1} = \sqrt{\frac{GM_1}{r_{i1}}} = \sqrt{\frac{GM_1}{r_{i2} + r_{2,peri}}}, \tag{5.9}$$

and

$$v_{2,peri} = \sqrt{\frac{2GM_1}{r_{2,peri}}} \tag{5.10}$$

This defines the semi-annulus in the primary disc, where the secondary exerts sufficient influence to potentially capture disc material. The upper and lower limits of this semi-annulus can be found numerically by solving

$$\frac{1}{2}\left(\sqrt{\frac{GM_1}{r_{i2} + r_{2,peri}}} - \sqrt{\frac{2GM_1}{r_{2,peri}}}\right)^2 - \frac{GM_2}{r_{i2}} = 0. \tag{5.11}$$

for r_{i2}. Having defined the semi-annulus, it is a simple matter to calculate its mass, given a surface density profile for the disc. In the case of a $\Sigma \propto r^{-1}$ disc with mass 0.1 $M_\odot$ (and an encounter according to the parameters of Simulation 1) this gives an available mass of $\sim$ 0.0057 $M_\odot$. Again, the simulations use 500,000 SPH particles, and hence this corresponds to a particle number of $\sim$ 30,000. Therefore, the accretion feedstock for the secondary constitutes 600 nearest neighbour groups, again well above the minimum mass resolution.

5.4 Results I - Do Stellar Encounters Stimulate Fragmentation?

The results for each simulation are discussed below. A summary of the simulation parameters can be found in Table 5.1. Qualitatively speaking, the results show features common to all simulations. Therefore, the results for simulation 1 (the reference simulation) will be discussed in detail - the other simulations will be more briefly described, focusing on their unique and differing features.

5.4.1 Simulation 1 - The Reference Simulation

Images of the reference simulation can be seen in Figure 5.2: the azimuthally-averaged radial profiles at the three instants are shown in Figures 5.4, 5.5, 5.6, & 5.7. As the secondary passes through the disc, it imparts significant energy to the disc, causing a strong temperature spike

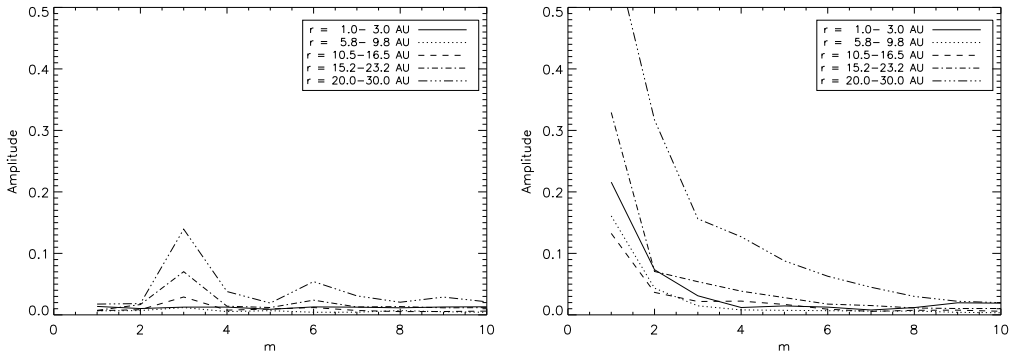

Figure 5.3: Fourier m modes of the Simulation 1 disc before and after the encounter. Modes are calculated from within 30 AU to avoid the region containing the tidal arm. However, its indirect gravitational influence is apparent in the strong $m = 1$ mode generated in the disc at all radii.

at the location of the secondary, and the scale height of the disc to be enlarged. The disc exerts significant tidal forces on the secondary as it reaches periastron, and angular momentum is transferred to the disc. There are three consequences to this:

1. The orbital parameters are altered, such that although the initial velocity of the secondary is consistent with $r_{peri} = 40$ AU, the actual periastron is lower (this can be seen in the Σ spike in Figure 5.4).

2. The secondary can capture disc matter to form a secondary disc ($M \sim 0.006 M_{\odot}$, around 5% of the initial disc), as well as drawing out tidal tails (see Figure 5.2), which depletes the primary disc. In total, around 20% of the initial disc mass is lost.

3. The disc is pushed out of equilibrium, and angular momentum transport is required to stabilise the disc: this results in a steeper surface density profile (and a flatter Q profile). This is a common feature to all the simulations carried out (although the strength of these effects varies with the orbital parameters used).

The disc spiral structure, which is initially well organised into azimuthal $m = 3$ and its associated harmonic modes, becomes dominated by the strong $m = 1$ contribution of the tidal arm after the encounter (Figure 5.3). The disc is gravitationally stable at all radii, with a higher surface density and scale height, and an unchanged temperature profile.

5.4.2 Simulation 2 - A Low Periastron Encounter

With a low periastron radius in prograde motion, the disc exerts stronger tidal forces, resulting in the capture of the secondary into a highly eccentric orbit. The secondary captures a similar disc to Simulation 1, but the total disc mass lost is slightly higher (due to more massive tidal tails being induced). This increased loss results in a greater steepening of the surface density profile, as the primary disc must make a larger readjustment to retain stability.

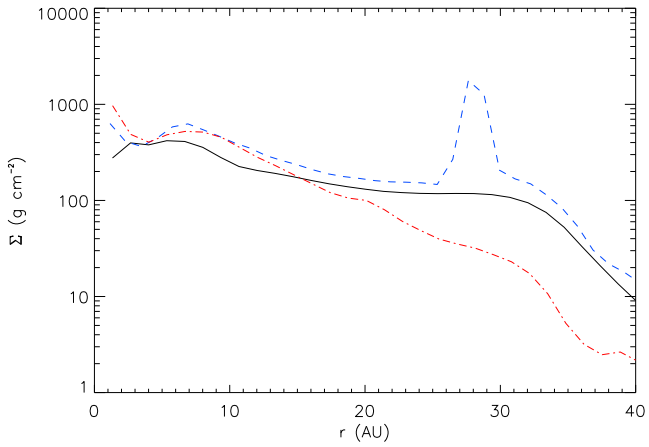

Figure 5.4: Surface density profile of the Simulation 1 disc before the encounter (solid line), at periastron (dashed line), and after the encounter (dot-dashed line).

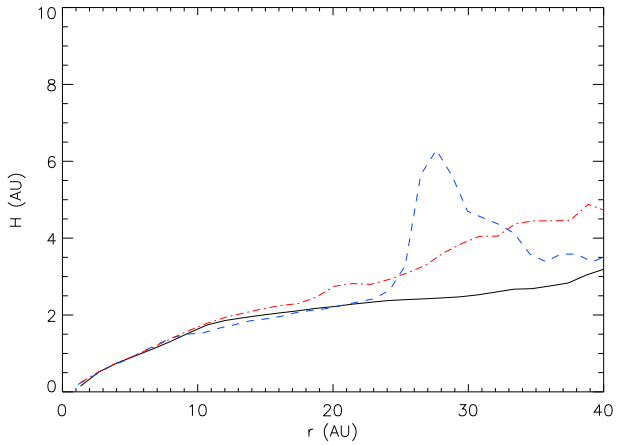

Figure 5.5: Scale height of the Simulation 1 disc before the encounter (solid line), at periastron (dashed line), and after the encounter (dot-dashed line).

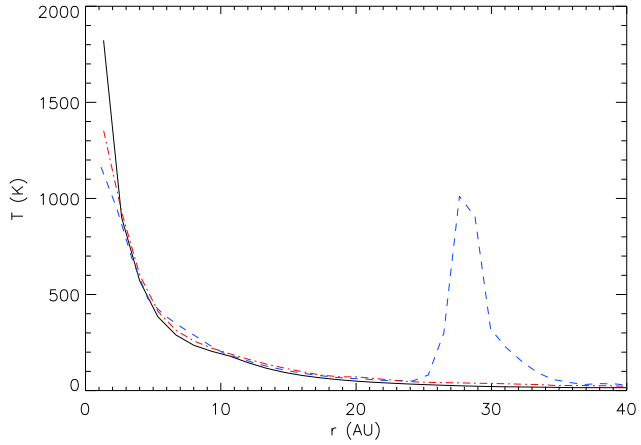

Figure 5.6: Midplane temperature profile of the Simulation 1 disc before the encounter (solid line), at periastron (dashed line), and after the encounter (dot-dashed line).

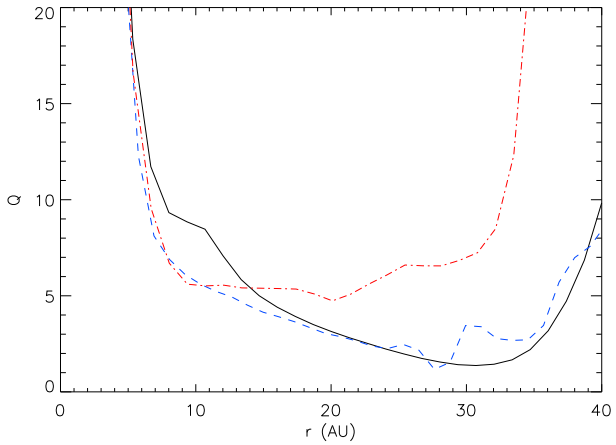

Figure 5.7: Toomre Q profile of the Simulation 1 disc before the encounter (solid line), at periastron (dashed line), and after the encounter (dot-dashed line).

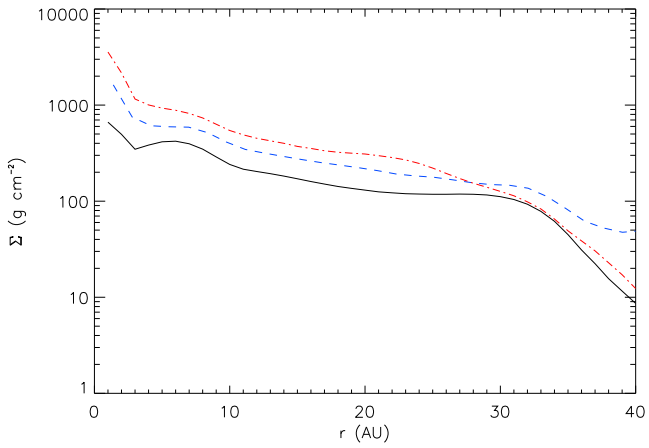

Figure 5.8: Surface density profile of the Simulation 4 disc before the encounter (solid line), at periastron (dashed line), and after the encounter (dot-dashed line).

5.4.3 Simulation 3 - A High Periastron Encounter

A high periastron encounter, although having the same qualitative features as Simulations 1 & 2, has a reduced quantitative effect. This is of course consistent with the gravitational influence of an object decreasing with separation. The surface density profile is steepened, but the effect is only noticeable at large radii, which gives a Q profile that remains flat out to around 35 AU (slightly further than Simulation 1).

5.4.4 Simulation 4 - A Distant Periastron Encounter

When the periastron is very large, the effects of the encounter are reduced significantly. However, this simulation is worth discussing. The distance of the secondary from the disc ensures that the disc heating is minimal, which allows the outer regions of the disc to remain marginally unstable. Also, the surface density profile is slightly modified in the inner regions (Figure 5.8), without affecting the scale height or the temperature profile. This results in a disc that is marginally unstable over a larger range of radii (Figure 5.9). Although the disc does not fragment in this simulation, it has been modified by the encounter to become more amenable to fragmentation. This suggests that there is a range of periastra for which encounters are most effective at destabilising the disc, which will depend on the disc's (and the secondary's) properties.

5.4.5 Simulation 5 - A Higher Disc Mass Encounter

The previous simulations have shown that the effect of an encounter is to steepen the disc's surface density profile, and to flatten the Q profile. Will this behaviour hold for higher disc masses? To investigate this, a prograde coplanar encounter was simulated using the 0.2 $M_{\odot}$

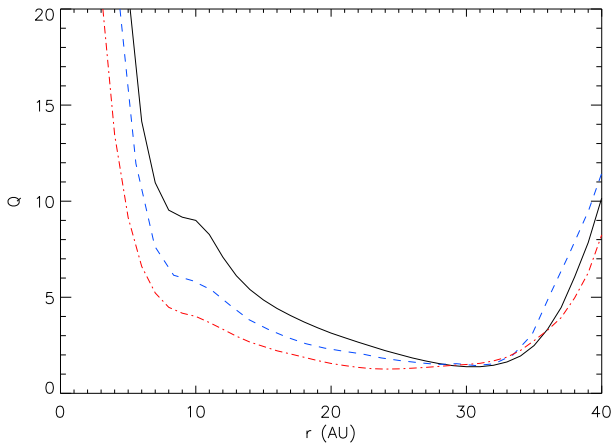

Figure 5.9: Toomre Q profile of the Simulation 4 disc before the encounter (solid line), at periastron (dashed line), and after the encounter (dot-dashed line).

disc described in section 5.3. Figure 5.10 shows that the surface density profile does steepen, but not a great deal: this is primarily due to significant mass depletion (and relatively reduced mass redistribution) - the secondary captures a disc of 0.01 $M_{\odot}$, removing matter from the outer regions to steepen the profile. At periastron, a significant increase of the surface density (and scale height) can be seen; the increased density at periastron (in comparison with Simulation 1) prevents the temperature increase from reaching the same magnitude. The Q profile for the disc (Figure 5.11) was initially rather flat; during the encounter, the secondary induces a small region to become very unstable around periastron, but due to the mass stripping, the majority of the disc stabilises: the region of instability decreases from around $10 - 40$ AU to around $10 - 20$ AU.

5.4.6 Simulation 6 - A Retrograde Encounter

In prograde encounters, the ability of the disc spiral structure to couple with the perturber is much improved in comparison with the retrograde encounter, and hence angular momentum transfer between the perturber and the disc in a retrograde encounter is relatively smaller (Hall et al., 1996). In general, this is reflected in the results obtained. However, depending on the relative phases of the spiral density wave and the secondary, the secondary can encourage the spirals to wind more tightly, allowing compressive heating in the inner regions. This is evident in the temperature spike at ~ 10 AU in Figure 5.12. Despite this, the inner regions of the disc appear to be less stable than they were initially. This dependence on the relative orbital phases shows that the influence of encounters on disc dynamics are indeed more complex than a simple study of orbital parameters (such as this work) may initially suggest.

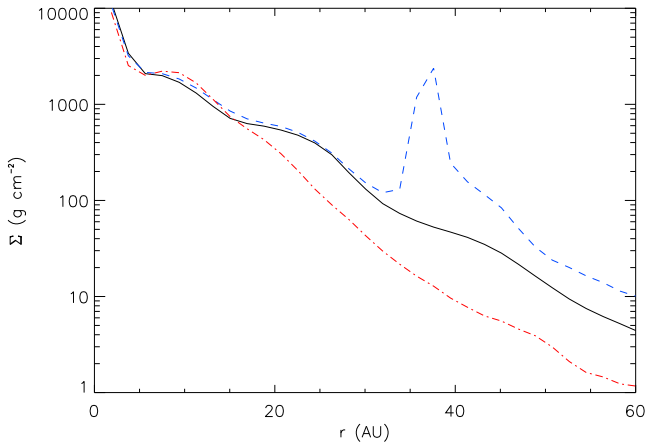

Figure 5.10: Surface density profile of the Simulation 5 disc before the encounter (solid line), at periastron (dashed line), and after the encounter (dot-dashed line).

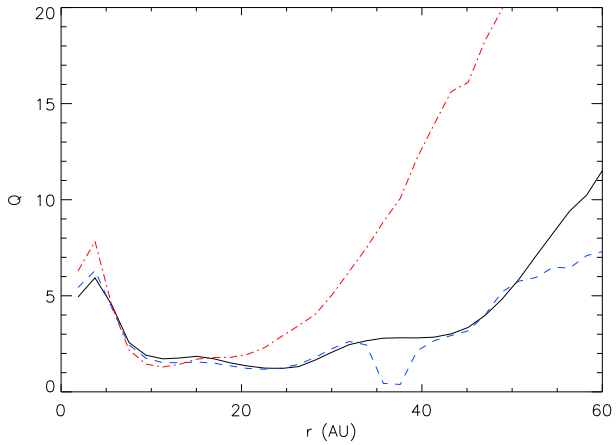

Figure 5.11: Toomre Q profile of the Simulation 5 disc before the encounter (solid line), at periastron (dashed line), and after the encounter (dot-dashed line).

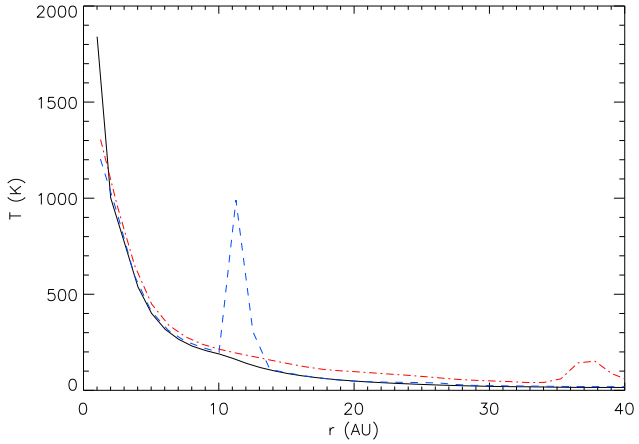

Figure 5.12: Midplane temperature profile of the Simulation 6 disc before the encounter (solid line), at periastron (dashed line), and after the encounter (dot-dashed line).

5.4.7 Simulation 7 - A High Inclination Encounter

As a secondary on a high inclination orbit has a short disc interaction timescale, it should be expected that the strength of the perturbation is reduced in comparison to a coplanar orbit. Simulation 7 seems to confirm this. Although the outer regions of the disc appear to increase in mass (and become less stable), the inner regions of the disc are almost unaffected: the Q and Σ profiles remain similar in the region $10 - 30$ AU, while the temperature profile is unaffected at virtually all radii. It could be argued that the transport of mass outward out of optically thick regions (allowing more efficient cooling) is favourable to fragmentation, however none was seen.

5.4.8 Simulation 8 - A Hyperbolic Encounter

As with the high inclination encounter, a hyperbolic encounter has a shorter interaction timescale than a parabolic encounter, and so should exert less influence on the disc dynamics. Figure 5.13 does indeed show that the perturbation to Q is weak in comparison with Simulation 1: however, the impact of the high-velocity encounter has resulted in significant mass stripping, causing a decreased Q in the outer regions: however, the velocity dispersion of the outflow ensures that the gas cannot form bound objects.

5.4.9 Simulation 9 - A High Mass Secondary Encounter

For the last of the simulations that emulate Lodato et al. (2007), the effect of increasing the secondary mass is studied here. With the primary and secondary now of equal mass, the disc begins to feel the influence of the encounter much earlier than in previous simulations. A substantial tidal tail is formed at an early time, and becomes significantly concentrated. This provides significant torques at much earlier times than the other simulations, which results in

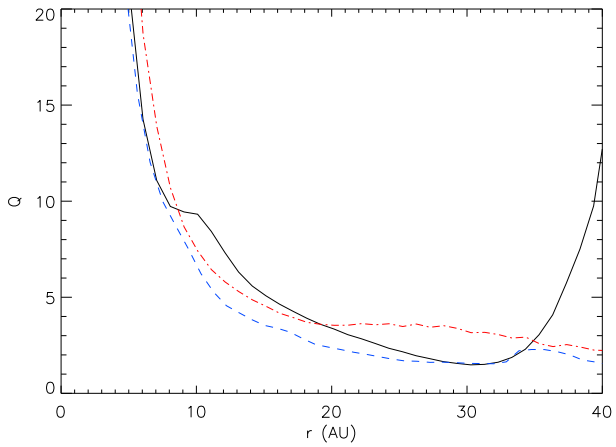

Figure 5.13: Toomre Q profile of the Simulation 8 disc before the encounter (solid line), at periastron (dashed line), and after the encounter (dot-dashed line).

a far smaller periastron ($\sim$ 10 AU), and at greater speed, which in turn heats the inner disc significantly (see Figure 5.14). Very efficient mass-stripping causes the disc to decrease in mass by more than half: a large fraction has been swept into the tidal tail at larger radii, and into a secondary disc. The significant mass loss and heating ensures the primary disc itself cannot fragment - it is now only a few AU in radius, and extremely hot.

These results stand in contrast to pseudo-viscous particle simulations conducted by Clarke & Pringle (1993) (hereafter referring to model B, with equal mass stars undergoing a parabolic encounter that does not penetrate the disc). Considering the proportions of matter in each component (primary disc, secondary disc and neither), Clarke & Pringle (1993) find that 60% remains bound to the primary, whereas only 34% remains in Simulation 9. The majority of the matter in Simulation 9 is unbound (61%), three times more than Clarke & Pringle (1993)'s model B. As a consequence, Simulation 9's secondary disc only contains around 5% of the original matter, compared to 23% in model B. To understand these differences, it must be realised that the encounters being compared are qualitatively different in two ways - firstly, the secondary in Simulation 9 penetrates the disc, whereas in model B it does not; secondly, model B does not account for disc self-gravity, which proves important in Simulation 9 at early times, where the tidal tail generated significantly affects the gravitational potential.

The results of Simulation 9 would suggest that high mass encounters are even less likely to promote fragmentation than low mass encounters (at least with periastra within the disc). It is possible, however, that a distant encounter with a high mass secondary would be more successful at producing fragments (cf. Simulation 4), although with a high mass secondary's larger perturbative potential, the periastron would most likely need to be larger than is considered here, and an interesting avenue for further work.

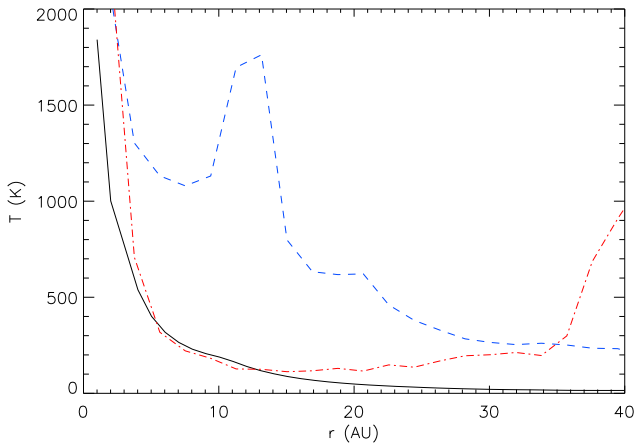

Figure 5.14: Midplane temperature profile of the Simulation 9 disc before the encounter (solid line), at periastron (dashed line), and after the encounter (dot-dashed line).

5.4.10 Simulation 10 - A Steep Disc Profile Encounter

Having studied the results of the previous simulations, it seems that the result of most encounters is to steepen the surface density profile of the disc. This behaviour then begs the question: what if the profile is steeper initially? Does the encounter result in an even steeper profile, or does there exist an asymptotic profile to which a disc will tend (given enough encounters)? To answer this question, an additional simulation was run with a disc exhibiting a $\Sigma \propto r^{-3/2}$ profile. The disc was run in isolation (as with the other discs used in this work) for the same timescale: this does mean however that the disc is stable over most radii when the encounter is begun (although Q is tending towards instability at larger radii). Again, this simulation was run for the purpose of studying the effects of encounters on the surface density profile, not fragmentation, so the need for a marginally stable disc is not so important here.

As can be seen in Figure 5.15, the effect of the encounter still steepens the profile (although comparing with Simulation 1 (Figure 5.4, the magnitude of disc readjustment is indeed smaller). As the outer regions of the disc are lower in density than in Simulation 1, the scale height is more sensitive to the encounter, and therefore see a greater increase as the secondary passes through the disc. However, the low density also ensures that compressive heating is less effective: therefore a lower temperature increase is seen.

This additional simulation has shown that surface density steepening appears to be common to all discs, regardless of their initial profile. However, the steeper the disc initially, the less effect the encounter has: this does suggest that there may be a profile steep enough that an encounter does not affect it significantly - but, this putative profile is most likely too steep to be physically appropriate.

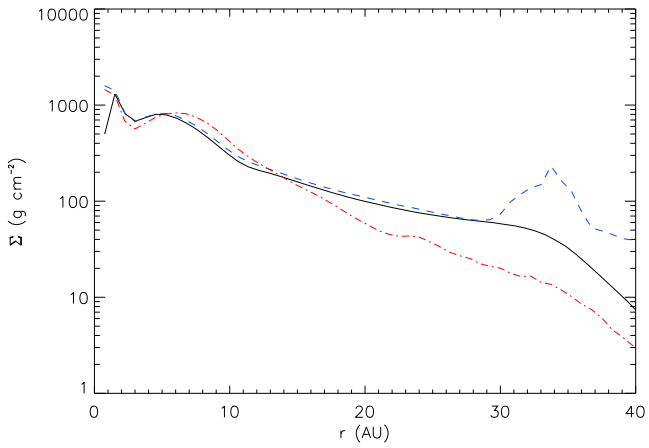

Figure 5.15: Surface density profile of the Simulation 10 disc before the encounter (solid line), at periastron (dashed line), and after the encounter (dot-dashed line).

5.5 Discussion I - Disc/Orbital Parameters and the Potential for Fragmentation

Having displayed results from a number of simulations, the broad trends uncovered by this work will now be discussed.

5.5.1 The Influence of Disc Mass, Disc Profile and Secondary Mass

Changing the mass involved in the encounter (and its distribution in the system) will of course affect the tidal forces experienced by all participants in the encounter. Increasing the disc mass does not in general indicate a more unstable disc after an encounter. The inner regions of a more massive disc can remain more gravitationally unstable, but the inefficiency of cooling in these regions precludes the possibility of fragmentation (Gammie, 2001; Rice et al., 2003; Rafikov, 2005; Rice & Armitage, 2009).

Encounters in general steepen the surface density profile: this appears to be true regardless of the initial profile, but the magnitude of profile steepening does seem to decrease if the initial profile is steep. They also flatten the Q profile as a secondary effect, but this does not appear to be true if the Q profile is initially flat.

As far as secondary mass is concerned, the results show that its primary influence is increased tidal forces subjected on the disc (as well as increased compressive and shock heating at periastron). Therefore, for disc-penetrating encounters, a larger secondary has a deleterious effect on potential fragmentation in comparison to a smaller secondary.

5.5.2 The Influence of Periastron Radius

A crude expectation may be that reducing periastron radius will destabilise the disc more, and induce fragmentation more readily. This is incorrect, primarily due to radiative effects. The ability of the disc to cool has been shown to be an important factor. If the periastron is too low, the disc is heated too strongly by the secondary's motion through it, and the disc is readjusted into a much more stable state due to mass stripping and angular momentum redistribution. If the periastron is too high, the disc will not feel the effects of the secondary's motion, effectively giving no stimulus at all. However, if the periastron is within some range of values (determined by the properties of the primary, its disc and the secondary), then the result of the encounter is to *increase* the region of instability in the disc. An interesting question that can now be asked is: would repeated encounters lead to fragmentation? The results indicate that repeated encounters would in general increase outward angular momentum transport (and inward mass transport), pushing the surface density profile to an equilibrium value that is too steep for gravitational instability to act, as the available mass at large radii is insufficient (Clarke, 2009; Rice & Armitage, 2009).

5.5.3 The Influence of Angular Momentum Alignment (and Inclination)

As has been previously discovered (Hall et al., 1996; Lodato et al., 2007), the most effective encounters are prograde in nature, and coplanar to the disc. However, the results of this work show that retrograde encounters may be more complex than initially thought. The effect of spiral winding by the secondary can encourage compressive heating in the inner regions, and the angular momentum transport in this case allows a disc that is more unstable in the inner regions. However, there is no indication that fragmentation would be better achieved using anything other than prograde encounters.

Encounters outside of the disc plane (Simulation 7) can drag matter to larger vertical distance, increasing the disc scale height: but, the interaction time is insufficient for the encounter to encourage significant change.

5.5.4 The Possibility of Binary Capture

It is difficult to make an encounter precisely parabolic for numerical reasons. With the exception of Simulation 8, the encounters simulated in this work have all been almost parabolic: that is, the total energy of the secondary is close to zero, and that any decrease in this total energy will result in the secondary becoming more bound to the primary-disc system. The efficiency of binary capture can be estimated from the change in total energy experienced as a result of the

Table 5.2: Orbital modification as a result of the encounter, for a selection of simulations. The energy is given in code units, where one unit $= 1.76 \times 10^{46}$ ergs.

Simulation	E_i	E_f	Δe	Secondary Disc?	Tidal Tail?
2 (30 AU)	-3.2×10^{-4}	-6.7×10^{-4}	-0.225	Yes	Yes
1 (40 AU)	-2.4×10^{-4}	-4.6×10^{-4}	-0.199	Yes	Yes
3 (50 AU)	-1.9×10^{-4}	-3.7×10^{-4}	-0.183	Yes	Yes
4 (100 AU)	-9.5×10^{-5}	-1.2×10^{-4}	-0.009	Negligible	Negligible
6 (Retro)	-4.5×10^{-4}	-8.5×10^{-4}	-0.175	Negligible	Yes
8 (Hyperbolic)	2.4×10^{-3}	2.2×10^{-3}	+0.16	No	Yes

encounter, as well as the modification of the orbital eccentricity[1]. Table 5.2 shows the initial total energy of the secondary (E_i), the total energy of the secondary after the encounter (E_f), and the change in orbital eccentricity (Δe).

The total energy decreases by a factor of 2 in most cases. This appears to be due to the angular momentum transfer between the secondary and the disc it accretes. As the secondary leaves the primary-disc system, in order to extract material orbiting in the same locale (i.e. to allow it to achieve escape velocity) it must impart dynamical energy to the gas. Also, the process of exciting a tidal tail (from matter at the opposite orbital phase) requires angular momentum from the secondary, although this appears to be less important. Encounters that do not create a sufficient disc (e.g. distant or high-velocity encounters) do not experience a significant orbital modification. The energy change in disc-penetrating encounters is always greater than or equal to the binding energy of the disc exterior to periastron, as was found by Hall et al. (1996). However, where they find the energy change is roughly twice this binding energy for prograde encounters (and five times for retrograde), no such relationship exists in these simulations. The values range from 120% of the binding energy in the prograde Simulation 2 to three times the binding energy in the retrograde Simulation 6. This difference can be ascribed to the disc's self-gravity: Hall et al. (1996) use a massless disc, and hence do not account for the secondary's response to the disrupted disc, which plays a crucial role in this work.

To assess the likelihood of capture, we can convert these energy changes into an equivalent velocities at infinity (i.e. assuming the energy change is entirely kinetic). This gives velocity changes ranging from 0.46 km s^{-1} for Simulation 4 to 1.75 km s^{-1} for Simulation 1 (and 1.87 km s^{-1} for the retrograde Simulation 6). Given that the typical velocity dispersion in open clusters is of order 1 km s^{-1} (see section 5.7), these results show that disc-penetrating encounters will be effective at capturing binaries (although they will be rare events, again see section 5.7). Non-penetrating encounters are much more likely, and also much less effective at binary capture. Capture efficiency increases with decreasing distance, provided that the velocity of the orbit is low enough that the secondary captures a non-negligible disc of its own. If the orbit is initially

[1]These eccentricities are calculated at the initial and final time of each simulation assuming two bodies in the system: i) the primary and its disc, and ii) the secondary and its disc (if any) - the tidal tail is not considered.

strongly unbound (such as Simulation 8), then no secondary disc can be captured, and the binary formation efficiency is effectively zero. Furthermore, a high velocity encounter causes severe disc stripping, which diminishes the possibility of a future encounter resulting in capture. These results are consistent with the work of Clarke & Pringle (1991), who study the upper limit of capture efficiency for hyperbolic orbits, and find that the capture rate in those cases is indeed low.

5.6 Results II - Do Stellar Encounters Produce Outburst Behaviour?

We will now turn to the accretion rates of the pointmasses in the simulations, and compare and contrast with the expected accretion rates of outburst phenomena such as FU Ori. We will only use a subset of the simulations described in section 5.4 for this analysis.

5.6.1 Simulation 1

The accretion rate of the primary and secondary can be seen in the left panel of Figure 5.16, with corresponding accretion luminosities in Figure 5.17. The accretion luminosity is calculated using

$$L_{acc} = \frac{1}{2}\frac{GM\dot{M}}{R} \tag{5.12}$$

where R indicates the accretion radius of the object (taken for both objects to be 0.1 AU in these simulations). Particles are accreted if they pass within the accretion radius of the object, and are gravitationally bound. R is held constant throughout the simulation, only M and $\dot{M}$ change. As the luminosity varies linearly with the accretion rate, the accretion luminosity show peaks and troughs similar to those of the accretion rate (with the important modification that the primary's larger mass makes it more luminous).

As a guide, Figure 5.17 has several important events delineated by vertical lines:

- t $\sim$ 1209 yr - the secondary begins its ingress to the disc, establishing a non-zero accretion rate.
- t $\sim$ 1256 yr - the secondary reaches periastron. The disc's temperature also peaks at this time.
- t $\sim$ 1279 yr - The secondary reaches its peak accretion rate: the matter that forms the secondary disc (located in the outer tidal tail generated at the location of the secondary) is in the process of infall onto its new parent star, not taking a well defined spheroidal shape until it has begun to exit the disc. The disc's luminosity has returned to near pre-encounter levels.

- t $\sim$ 1352 yr - the secondary begins its egress from the disc, and its gravitational influence diminishes. The primary disc must now begin readjustment - an increase in outward angular momentum transport gives a corresponding increase in inward mass flux, boosting the primary's accretion rate. The rate has a rise time of approximately 25 years.
- t $>$ 1352 yr - the luminosities decay with timescales of hundreds of years.

It must be emphasised that there are two separate accretion events, with distinct characteristics: firstly, the secondary's accretion undergoes enhancement for a period of around 150 years, reaching a maximum accretion rate of $5 \times 10^{-4}\, M_{\odot} yr^{-1}$ at periastron. Its accretion rate curve is a superposition of three characteristic features:

(i) A smooth symmetric curve, width approximately 110 years. This is linked to the secondary's passage through the smooth component of the disc.

(ii) A series of small spikes, widths of 1 year or less. These are due to the secondary's passage through overdensities caused by the disc's spiral structure.

(iii) A major spike, width approximately 10 years. This is caused by the formation of the secondary disc.

The second accretion event involves the primary, whose accretion curve consists of a steady increase to values of $\sim 10^{-5}\, M_{\odot} yr^{-1}$ as the secondary leaves the disc, with a rise time of order $\sim$ 10 years. This is followed by a slow decay with timescales in excess of 300 years. The accretion curves of the subsequent simulations discussed share some or all of these characteristic features.

The right hand panel of Figure 5.16 shows the accretion rate in the disc at 25 AU throughout the encounter. Similar spiking behaviour is also observed. The first peak occurs at periastron (where the secondary is around 3 AU away from the disc location studied). After this initial spike, there is a steady flow of matter around ten times the usual value until the secondary leaves the disc, resulting in a second spike due to the readjustment of the disc and the subsequent boost in primary accretion rate. After this readjustment is made, the disc returns to a steady state, with accretion occurring at a rate reduced by around a factor of 2 relative to its initial value.

It should again be noted that these simulations are subject to numerical viscosity in their inner regions, preventing mass build up. The artificially high viscosity in the inner regions can be thought of as acting like MRI (cf. Armitage et al. 2001; Zhu et al. 2009a), which activates at the "wrong" temperatures, facilitating mass accretion without pile up, and hence underestimating accretion rates during the outburst. These simulations can then be thought of as presenting the lower bounds of the accretion rate (in the limit where MRI is overactive).
The tidal interactions between the disc and secondary cause the secondary to be captured on an eccentric orbit: this facilitates study of the repetition of this outburst event over several orbits. Figure 5.18 shows the evolution of the accretion rate over three periastra passages; it can be seen that the magnitude of the outburst reduces significantly with each passage. The ability

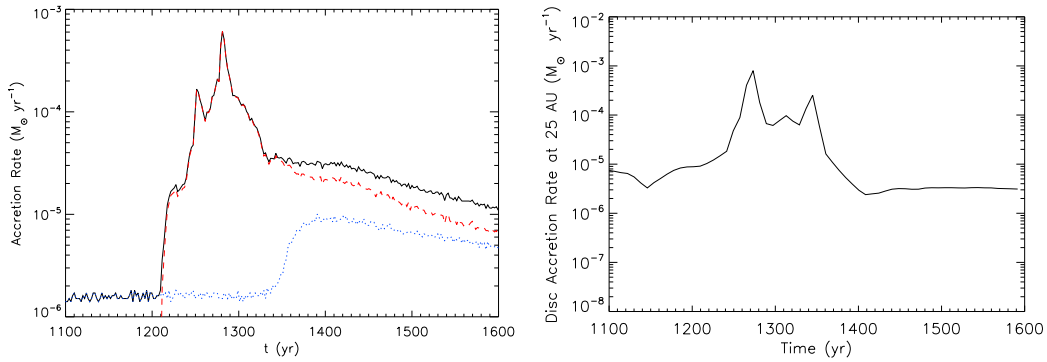

Figure 5.16: Left: Accretion rates for the primary and secondary in Simulation 1. The black line denotes total mass accretion, the blue line indicates the primary accretion, the red line indicates secondary accretion. Right: The calculated accretion rates in the disc at 25 AU. Note the similar spiking behaviour as the stellar accretion, with a differing time offset related to the proximity of the secondary.

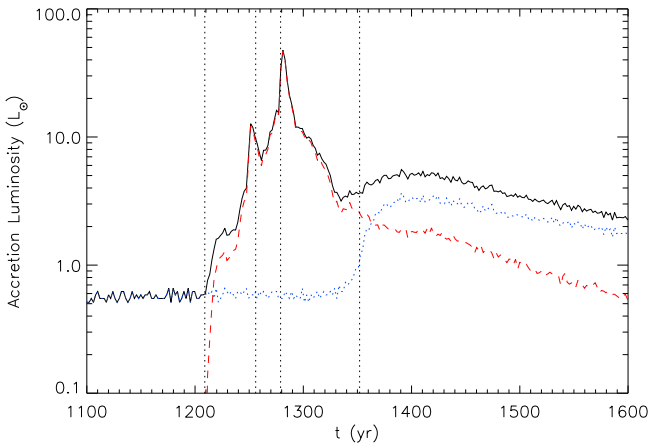

Figure 5.17: Accretion luminosities for the primary and secondary (and disc luminosity) in Simulation 1. The black line denotes total luminosity, the blue line indicates the primary accretion, the red line indicates secondary accretion, and the green line indicates the disc luminosity.

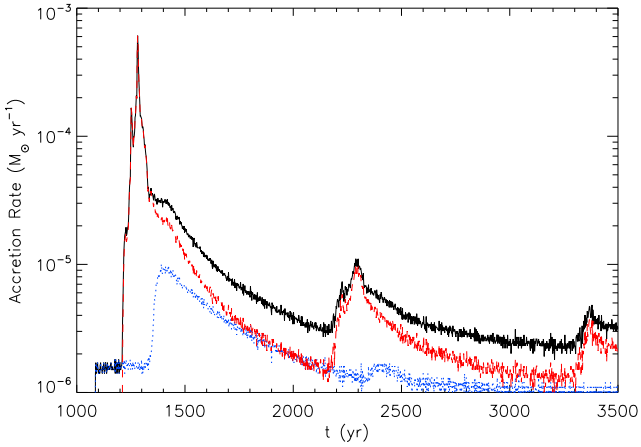

Figure 5.18: Long term evolution of the accretion rates for the primary and secondary in Simulation 1. The black line denotes total mass accretion, the blue line indicates the primary accretion, the red line indicates secondary accretion.

of the secondary to strip mass from the outer disc diminishes with each passage, partially due to the depletion of the outer disc itself, as well as the secondary disc's influence in regulating mass flow. Without mass stripping, the secondary cannot accrete, and the disc is not obliged to significantly readjust its mass distribution, preventing the primary accretion event. This would suggest that these outbursts are not easily repeatable.

5.6.2 Simulation 2 - Close Periastron

Can repeatability be attained by reducing the secondary periastron? It is not unreasonable to assume that a closer approach allows the secondary to accrete from a more plentiful mass supply, potentially improving its ability to create repeatable outbursts. The secondary reaches periastron at t $\sim$ 1155 yr, resulting in the first spike in Figure 5.19. Feature (ii) is less prominent, as the secondary passes through weaker spiral structure in the inner regions. Again, the peak accretion rate occurs when the secondary disc begins its infall (feature (iii)), giving a similar order of magnitude increase in accretion rate as Simulation 1.

The second encounter is of similar magnitude, with an accretion peak at t $\sim$ 1691 yr. Note that the accretion rate peak is smoother than the first: the spiral structure in the primary disc has been almost completely erased, so the secondary sees a smooth distribution of mass along its trajectory. Also, the secondary disc can maintain its shape during the encounter, reducing the infall onto the secondary. These facts combined eliminate the possibility of seeing features (ii) and (iii) in the second peak: all that remains is the smooth component (feature (i)). However, this magnitude of outburst cannot be maintained through subsequent orbits, decaying in similar fashion to Simulation 1. Repeatability therefore seems to be limited for outbursts of this type.

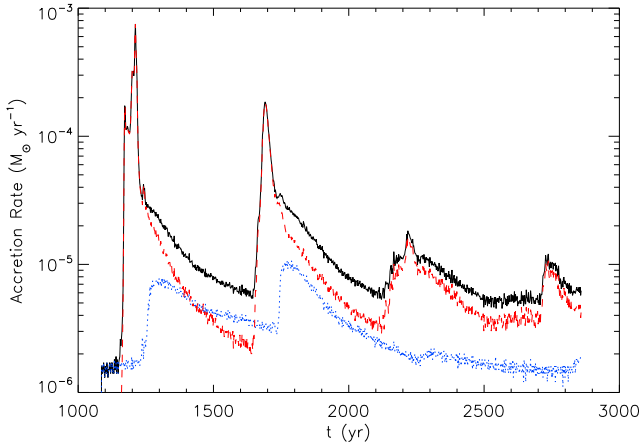

Figure 5.19: Accretion rates for the primary and secondary in Simulation 2. The black line denotes total mass accretion, the blue line indicates the primary accretion, the red line indicates secondary accretion.

5.6.3 Simulation 5 - Increasing Disc Mass

As the important factor is enhanced accretion, will adding more mass to the disc result in a larger accretion event? The results of Simulation 5 (where the disc is double the mass of Simulation 1) indicate some important differences (Figure 5.20). The disc's lack of high-m spiral modes compared to the lower mass disc (see Figure 5.1) prevents feature (ii) from being presented here. Periastron occurs at t $\sim$ 1284 yr, where no significant accretion peak can be seen. The peak accretion rate is again seen when secondary disc infall occurs at t $\sim$ 1338 yr. Overall, the features of this outburst event are similar in magnitude and duration to those seen in the simulations with a less massive disc, in particular the peak accretion rate of the secondary and its decay timescale. This would suggest that the behaviour of the secondary during the outburst is relatively insensitive to disc mass (possibly because the secondary is already operating at peak accretion efficiency at lower disc masses). With that said, the pre-outburst and post-outburst accretion rates of the primary are slightly higher, and its decay timescale is slightly shorter. Depending on the orientation of this system to the observer, this will have important implications for observation (see Discussion).

5.6.4 Simulation 8 - A Hyperbolic Encounter

The previous sections of this chapter have indicated that accretion efficiency of the secondary is linked to the velocity of the secondary relative to the disc. If the secondary moves through the disc too quickly, it may be expected that the accretion event will be reduced in magnitude in comparison to the other simulations shown. Simulation 8 was run to discover the effects of increased orbital velocity, by specifying a hyperbolic encounter ($e = 7$). Figure 5.21 indicates that the accretion efficiency is indeed reduced. The peak accretion rate is lower, and the

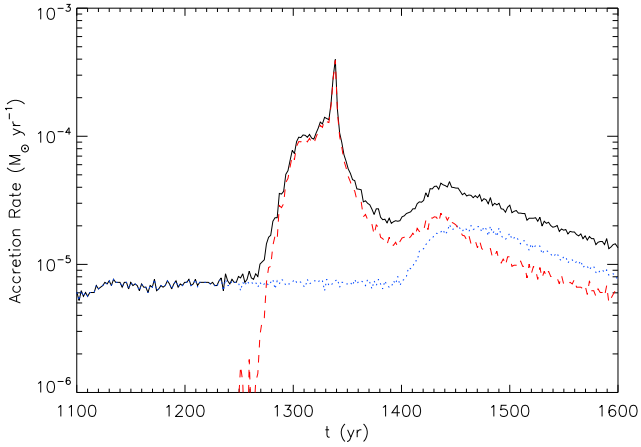

Figure 5.20: Accretion rates for the primary and secondary in Simulation 5. The black line denotes total mass accretion, the blue line indicates the primary accretion, the red line indicates secondary accretion.

duration of the accretion event is also reduced, lasting around 50 years. There is no evidence of feature (ii): the high velocity of the secondary's motion (and the limited time resolution of the data) prevent the detection of these peaks. There is no peak associated with the secondary's periastron (t $\sim$ 1110 yr).

The reduced peak is again associated with the infall of matter onto the secondary after the encounter (feature (iii)), but no disc is formed. The high-velocity encounter essentially destroys the disc, throwing significant amounts of material to large distances. It is this halo of matter which the secondary accretes from, but not efficiently or for any length of time, as the velocity dispersion of the material is quite large.

The primary's accretion behaviour remains similar to the other simulations, despite the dispersive action of the secondary. Indeed, the act of removing disc material may help in the detection of such an event (see following sections).

5.7 Discussion II - The Potential for Observation of Encounter-Driven Outbursts

5.7.1 Frequency of Occurrence

For outbursts from stellar encounters to be observed and correctly classified, their occurrence must be sufficiently frequent in an observer's field of view. To estimate the frequency of stellar encounters in a star cluster, the collision rate calculations of Clarke & Pringle (1991) are employed. They calculate the collision rate Γ_{hit} of a star-disc system with discless companions in a star cluster with stellar number density n_0, and a Gaussian velocity distribution, characterised

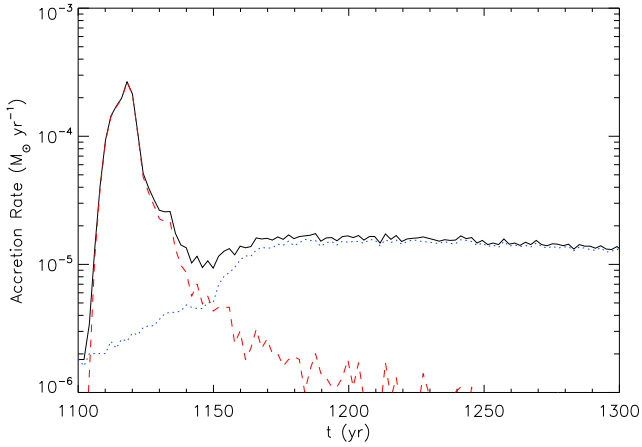

Figure 5.21: Accretion rates for the primary and secondary in Simulation 8. The black line denotes total mass accretion, the blue line indicates the primary accretion, the red line indicates secondary accretion.

in one dimension by the velocity dispersion V_*:

$$\Gamma_{hit} = \Gamma_0 \left(1 + \frac{V_*^2 R_{disc}}{GM_*}\right) \tag{5.13}$$

where

$$\Gamma_0 = \frac{4\sqrt{\pi} n_0 G M_* R_{disc}}{V_*} \tag{5.14}$$

The discs used in these simulations have $R_{disc} = 40$ AU, and $M_* = 0.5 M_\odot$. To mimic the core of an open cluster, the free parameters are selected to be $n_0 = 100\, pc^{-3}$, $V_* = 1\, km\, s^{-1}$ (Binney & Tremaine, 1987; Clarke & Pringle, 1991). This yields $\Gamma_0 = 3.01 \times 10^{-4}\, Myr^{-1}$, and $\Gamma_{hit} = 3.28 \times 10^{-4}\, Myr^{-1}$. Assuming that the average protostellar disc has a lifetime of ~ 1 Myr, Γ_{hit} is the probability that one disc will undergo an encounter in its lifetime. Therefore, out of roughly 3000 stars with discs, one will experience an encounter in its lifetime. This would imply that in an average open cluster, there will be at least one encounter-driven accretion event per Myr. Assuming that these encounters are randomly distributed in inclination, it should be expected that only around 6% of these encounters will be coplanar[2]. Finally, the duration of these encounters must be taken into account: the primary's accretion rate remains enhanced for time intervals of $\Delta t = 500$ years, which improves the chances of detection somewhat. If the probability of detecting an event is defined as

$$P_{obs} = N_{events} \Delta t = \Gamma_{hit} N_{stars} f_{inc} \Delta t, \tag{5.15}$$

[2]For an encounter to be coplanar, the inclination must be lower than the disc's opening angle, defined by its aspect ratio $\frac{H}{r} = \frac{c_s}{\Omega r}$. This can be justified by rewriting equation 5.11 for an inclined secondary

with $f_{inc} = 0.06$, and $\Delta t = 500$ yr, then for $N_{stars} = 3000$, this yields $P_{obs} \sim 10^{-4}$. This shows that encounters of this type should not be frequently observed. This is of course an oversimplification: it does not account for a distribution of disc radii or masses, nor the subclustering that exists in bound systems. It also does not reflect the young ages of most FU Ori systems (in general less than half the disc's lifetime). However, it provides a sufficient order-of-magnitude estimation to illustrate the rarity of these events.

5.7.2 The Problems of Obscuration

The accretion luminosities shown in the previous sections are intrinsic luminosities: they do not account for the effects of optical depth. Consider the case where the disc is face-on to an observer: the primary star resides in a gap at the centre of the disc, and so the optical depth to the observer is relatively low. The secondary, however, penetrates the disc, accumulating matter and heating the surroundings. The Rosseland mean optical depth of the secondary at periastron reaches $\tau \sim 300$. In this optically thick regime, the secondary is strongly obscured: any detectable emission will be reprocessed and emitted by the disc at longer wavelengths.

The effect of obscuration increases with inclination to the observer. The optical depth of the primary (and secondary) to the observer can increase to $\tau \sim 10^7$ at edge-on, completely shielding the accretion event from observers. This would suggest that encounters have to occur within a restricted range of inclinations to the observer in order to be directly observable.

5.7.3 Indirect Observational Signatures

If the secondary's accretion event is mostly screened by the disc, then what observational signatures can be identified? The disc SED for Simulation 1 was calculated by assuming that the disc emits a blackbody spectrum at each annulus, and integrating the contributions:

$$F_\lambda = \int 2\pi r B_\lambda(T_{eff}(r))dr \tag{5.16}$$

This expression requires the disc to be optically thick. This is true out to radii of ~ 37 AU, effectively the disc's outer radius, so this expression is justified[3]. The effective temperature of the spectrum is calculated numerically by identifying the location of the disc's photosphere. The resulting SEDs for Simulation 1 before, during and after the encounter can be seen in Figure 5.22. The effect of the encounter is to steepen the surface density profile, which in turn steepens the optical depth profile. This allows the cooler outer regions to radiate more efficiently, while the inner regions experience stronger screening. The result is a boost in the disc's flux at longer wavelengths during the encounter, increasing the flux at $850\mu m$ (for example) by a factor of 10. Add to this the obscuration of the secondary (especially its shorter wavelengths), and it appears that the outburst event is best searched for in the far-IR to sub-mm regimes.

[3]We have confirmed this by using more complicated expressions for optically thin and optically thick discs (see equations 6 and 7 of Armitage et al. 1999) and have found the results to be indistinguishable.

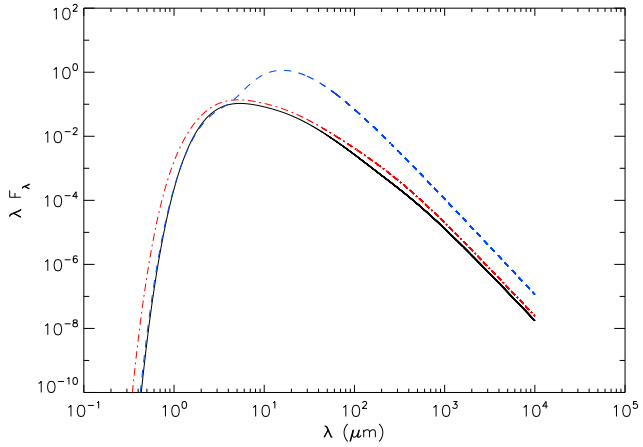

Figure 5.22: The SED of Simulation 1 before (black line), during (blue line) and after the encounter (red line). The general effect of the encounter is to boost sub-millimetre and infrared flux for a short time.

However, the effects of stellar irradiation are not included in the simulation. Can a substantial increase in reprocessed emission and scattered light be expected from the encounter? Stellar irradiation (and reprocessed emission) can be estimated as a function of temperature, assuming radiative equilibrium (Ida & Lin, 2004):

$$T(r) \approx 280 \left(\frac{r}{1\,AU}\right)^{-\frac{1}{2}} \left(\frac{L_*}{L_\odot}\right)^{\frac{1}{4}} K \tag{5.17}$$

In the initial stage, the primary is the sole contributor: the primary's initial accretion luminosity (plus an intrinsic luminosity of a similar mass main-sequence star) gives

$$T(r) \approx 300 \left(\frac{r}{1\,AU}\right)^{-\frac{1}{2}} K \tag{5.18}$$

By comparison, the disc reaches temperatures of over 1000 K in the inner 3 AU initially (from dynamical heating sustained by marginal instability), so it is expected that reprocessed emission as a result of accretion should not be significant initially. During the encounter, the secondary's accretion luminosity increases to values of $\sim 40 L_\odot$. This implies that stellar irradiation should be responsible for temperatures of around 700 K at 1 AU from the secondary. Compressive heating increases the temperature in this region to a similar value: this suggests that reprocessed starlight will play a significant role, and should also be considered in terms of observations.

5.8 Conclusions

I have detailed the results of SPH simulations with hybrid radiative transfer (Chapter 3) of the encounters between a protostellar disc system and a secondary star. I will now summarise the results of the two separate studies of fragmentation and outbursts.

5.8.1 Encounters and Fragmentation

In general: *encounters do not induce fragmentation in compact discs* (confirming the results of Lodato et al. 2007), although there appears to be a subset of orbital parameters that modify the disc to make it more unstable over a larger range of radii. As to whether this is an indication that there are specific orbital parameters for an encounter that cause fragmentation (e.g. a highly eccentric elliptical coplanar orbit which has a distant non-penetrating periastron), more work needs to be done. Equally, the results presented here show that factors not incorporated in this parameter study (such as the relative orbital phases of the primary, the disc spiral structure and the secondary) have a non-negligible role to play in the resulting disc dynamics, indicating more work is required to fully understand their effects. It has been shown that angular momentum transfer between the secondary and the disc is significant: however, calculations indicate that only around 1 in 10,000 discs will experience encounters of this nature in 1 Myr (Clarke & Pringle, 1991), so this cannot be a typical trigger for angular momentum transport in discs.

The most important conclusion to draw from this work is the same as was found by Lodato et al. (2007) - the key parameter that determines disc fragmentation is the cooling rate (Gammie, 2001; Rice et al., 2003; Mejia et al., 2005; Rice et al., 2005). This is independent of whether discs become unstable in isolation, or are driven to it by external stimuli such as stellar encounters.

5.8.2 Encounters and Outbursts

I have investigated the possibility of a stellar encounter (where one participant has a proto-stellar disc) being the progenitor of an outburst phenomenon. The outbursts described here have several key features in common, originating from two distinct accretion events (for the primary and secondary respectively), independent of the secondary's orbital parameters. The secondary's accretion rate grows and fluctuates as it traverses the disc's spiral structure: the peak occurs when the secondary disc is being formed from infalling stripped primary disc material, and the accretion rate then decays over several hundred years. The primary accretion rate (perhaps the easiest to observe) increases slowly with a rise timescale of tens of years as the primary disc readjusts to mass stripping, and decays on a longer timescale than the secondary.

It has been established that these stellar encounters can enhance accretion rates to levels corresponding to FU Ori and EX Lupi outbursts, and that they have distinct observational signatures, that although low in probability to detect, are nonetheless possible in principle to see. But, it is important to emphasise that these encounters cannot be responsible for *all* FU Ori phenomena (or EX Lupi for that matter), for the following reasons:

1. This type of stellar encounter is too infrequent to explain the catalogue of outbursts currently observed,

2. Such encounters cannot maintain the rapid periodicity or repeatability required of some outbursts without significant decay of the outburst strength,

3. This origin would predict the detection of a companion (perhaps in the infrared or sub-millimetre) for most outburst hosts. This is not the case; out of at least twenty FUors, only seven have a confirmed companion (Pfalzner, 2008).

Despite this, outbursts from stellar encounters can mimic FU Ori well, with the correct general behavioural trends. They may also potentially have the same triggering mechanism - mass pile up leading to MRI activation (Armitage et al., 2001; Zhu et al., 2009a), although higher resolution simulations (which resolve the inner region and reduce the artificial viscosity) are required to confirm this. Taking the fraction of FUors with a companion as a guide, it is estimated that at most 30% of outbursts detected could be due to an encounter or binary (with the actual figure presumably much smaller). The encounter-driven outbursts identified here represent a subtly different type of object, that although not currently detected, may be detected in future large-scale surveys of star-forming regions at far-infrared and sub-mm wavelengths.

Chapter 6

Native Synthetic Imaging of Smoothed Particle Hydrodynamics Simulations

Hail, holy light! offspring of heaven first-born.
John Milton, *Paradise Lost*

6.1 Author's Note

This chapter includes work published by myself as first author (Forgan & Rice, 2010a).

6.2 Monte Carlo Radiative Transfer as an Imaging Technique

This thesis has been (so far) primarily concerned with theoretical studies and numerical simulations of self-gravitating discs. While these studies are necessary and useful, it is difficult for observers to fully utilise the fruits of this research. The ease with which SPH simulations can be visualised (using software such as SPLASH, Price 2007) is not replicated in real astronomical observations. For example, a 2D surface density "image" of an SPH simulation such as those presented in Chapters 4 and 5 will generally not resemble observations from any telescope.

The symbiotic relationship between theory and observation is crucial for continued progress in astrophysics - without theory, observers cannot correctly interpret their data; without observations, theorists cannot reject conjectures or constrain models. With the publication of the USA's Astronomy and Astrophysics Decadal Survey (Committee For A Decadal Survey Of Astronomy And Astrophysics, 2010) heralding the next generation of telescopes, along with the recent and impending activation of instruments such as *Herschel*, ALMA, eMERLIN and many others, it is clear that for theorists to remain relevant in future science projects, they must cultivate means by which they can provide observers with interpretations and predictions.

One way in which SPH simulators can provide predictions is to synthesise telescope images of their simulations. This is essentially a radiative transfer problem, something we have discussed at length in this thesis. However, the preceding analyses have largely ignored the complications introduced by the quantum nature of light, in particular the scattering of photons in the medium. If the medium is optically thick, this scattering is well-approximated by the diffusion framework discussed in Chapter 3. Once the medium transitions from optically thick to optically thin, the individual scattering events can radically influence the resulting evolution of the radiation field. In particular, the imaging of an astrophysical system will be strongly affected by such scattering. An example of this is scattered starlight from the surface of circumstellar discs.

The scattering process depends strongly on the relationship between the frequency of the photon and the scatterer's physical extent. For this reason, we cannot use the hybrid radiative transfer method I described in Chapter 3 as an imaging tool, as it is frequency-averaged. Instead, we must develop a tool which correctly reproduces optically thin scattering of photons in our simulation. This is essential if we are to be able to correctly produce synthetic images of SPH simulations.

Let us begin with the solution to the radiative transfer equation (assuming the source function of the medium is zero):

$$I_\nu(\tau_\nu) = I_\nu(0)e^{-\tau_\nu}. \tag{6.1}$$

This gives the classical solution for a beam of intensity I_ν, composed of a large number of monochromatic photons (i.e. photons with frequencies in the range $(\nu, \nu + d\nu)$). From the solution for I_ν, we can use frequentist arguments to deduce that the probability of a photon in the beam travelling an optical depth τ_ν without interacting with the medium is:

$$P(\tau_\nu) = e^{-\tau_\nu}. \tag{6.2}$$

This does not allow us to predict the behaviour of an individual photon, but instead we can follow the evolution of a statistical ensemble of photons. Such a reliance on probabilities requires us to use so-called *Monte Carlo* methods to track the behaviour of photons stochastically (by utilising random sampling techniques). Specifically, this is referred to as Monte Carlo Radiative Transfer (MCRT). The method can be briefly summarised as:

1. Emit *photon packets* (a bundle of monochromatic photons) from some source in the medium.

2. Calculate the optical depth the packets experience as they proceed along their direction vector.

3. Compare this optical depth with a randomly sampled value $\tau \equiv \tau_{random}$. We sample τ_{random} from the probability distribution given in equation (6.2) to reproduce the correct scattering/absorption behaviour.

4. Once the photon packet's optical depth exceeds τ_{random}, either scatter or absorb the photon according to the local albedo. The direction the packet is scattered into will be randomly sampled, again to reproduce the local statistical properties of scattering in the medium. If the packet is absorbed, either its total intensity is reduced, or the whole packet is absorbed and immediately re-emitted to conserve energy.

5. Repeat this process of scattering and absorption until the photon packet either exits the medium or is completely absorbed. Packets that escape the medium can then be binned on a pixelated image plane (analogous to the pixels of a CCD in real telescopes).

MCRT methods have been in use for several decades, for both continuum radiative transfer and for non-LTE line transfer (House, 1969; Bernes, 1979; Pinte et al., 2006), even extending to the general relativistic regime (Dolence et al., 2009). It has been widely used in studying circumstellar discs, in particular for producing spectral energy distributions (SEDs) from various disc models (Wood et al., 2002; Woitke et al., 2010). This modelling reveals the spectral signature of important physical features such as inner disc gaps/holes, flaring and the effects of grain size.

These methods have traditionally used grids to represent the medium under study - I will outline in this chapter a means of directly implementing MCRT on an SPH density field. Before discussing this, I will describe in greater detail the standard MCRT algorithm.

6.3 Aside: Pitfalls in Generating Random Numbers

These methods receive the epithet "Monte Carlo" because of their reliance on probability and the generation of random numbers (much like *Le Grand Casino* in Monte Carlo, from which the moniker is derived). It must be noted that the random numbers generated by computers are not truly "random" - the output of a Geiger counter may be random, but no computer can truly simulate such a process. However, there are algorithms that generate an excellent substitute - the so-called *pseudo-random* numbers, which generate very long sequences of numbers which satisfy tests for randomness. However, the user must always be aware that the algorithms can only maintain their "randomness" for a finite sequence length - once above this critical length, the sequence of numbers attains a periodicity which could prove dangerous for Monte Carlo

studies. I use the *ran2* routine detailed in Numerical Recipes for FORTRAN (Chapter 7, Press et al. 1992). In the following sections I will use Y to denote a random number sampled from a uniform distribution between $[0, 1]$.

6.4 The Emission of the Photon Packet

We begin with the creation of the photon packet itself. These packets represent a large number of monochromatic photons. We will see that the accuracy of the simulation is related to how many packets are emitted in the simulation. If the system has a total luminosity L_{tot}, distributed amongst a set of sources $\{i\}$, then we distribute N_γ photon packets amongst the sources accordingly:

$$N_i = \frac{L_i}{L_{tot}} N_\gamma. \tag{6.3}$$

All packets will share the same total energy under this emission scheme[1]. More luminous sources emit more photon packets. As N_i is a discrete quantity, faint sources may not emit photon packets at all if N_γ is too small. This is the first potential source of error associated with N_γ, and there are others to follow.

Each photon packet must be assigned a direction vector and intensity. Typically, sources emit isotropically, and hence the direction vector $\mathbf{n} = (n_x, n_y, n_z)$ is selected at random, i.e

$$\begin{aligned} n_x &= \sin\theta\cos\phi \\ n_y &= \sin\theta\sin\phi \\ n_z &= \cos\theta, \end{aligned} \tag{6.4}$$

Where the direction angles are generated from two samplings of Y:

$$\cos\theta = 2Y - 1 \tag{6.5}$$

$$\phi = 2\pi Y. \tag{6.6}$$

This can be altered for anisotropic emission if necessary, but isotropic emission is sufficient for our purposes. The frequency of the photon is also sampled to reproduce the correct distribution - in our case, it is satisfactory to use the Planck function:

$$B_\nu(T) = \frac{2h\nu^3}{c^2} \frac{1}{\exp(\frac{h\nu}{kT}) - 1} \tag{6.7}$$

Sampling from B_ν is done straightforwardly by the *accept-reject* method. The algorithm is described below:

- The frequency is sampled uniformly in the frequency range $[\nu_{min}, \nu_{max}]$, giving a value ν_{test}.

[1] If the packets vary in frequency, then this is equivalent to varying the total number of photons in the packet $N_{packet} = E_{packet}/h\nu_{packet}$.

- The Planck function is calculated for this frequency $B_{\nu_{test}}$.
- This value is compared against the quantity YB_{max}, where Y is another uniform sampling between $[0, 1]$ and B_{max} is the maximum value of B_ν in the available frequency range.
- If $YB_{max} > B_{\nu_{test}}$, then ν_{test} is rejected as the photon's frequency, and the algorithm returns to part 1.
- If $YB_{max} < B_{\nu_{test}}$, then ν_{test} is accepted as the photon's frequency.

Frequencies with larger values of B will be accepted more often, assuring the form of the probability distribution is recovered. This is a simple way to sample randomly from complicated functions, and is useful when they cannot be easily integrated. However, the shape of the function may demand that many rejections must be made, which can be computationally expensive. If at all possible, other sampling methods should be used, such as the method described in the next section.

6.5 The Location of Interaction Events

We must now deal with how to sample τ_{random}, and locate where an individual photon packet will interact with the medium. Using the *probability integral transform*, we can construct a random variable Y using a random variable X, and its cumulative distribution function (CDF) $F(X)$:

$$Y = F(X). \tag{6.8}$$

Under this definition, Y is uniformly distributed between $[0, 1]$. Taking the inverse of this statement allows us to sample for X:

$$X = F^{-1}(Y). \tag{6.9}$$

We now have a means to generate X by using a simple uniformly distributed variable Y, where for our case $X = \tau$. The CDF for τ is obtained by integrating equation (6.2):

$$F(\tau) = \int_0^\tau e^{-\tau'} d\tau' = 1 - e^{-\tau}. \tag{6.10}$$

Replacing $F(\tau)$ by Y, we can rearrange to find τ:

$$\tau_{random} = -\log(1 - Y). \tag{6.11}$$

In practice, $(1 - Y)$ is also uniformly distributed in the range $[0, 1]$, so this equation is often rewritten

$$\tau_{random} = -\log(Y) \tag{6.12}$$

to save computational expense. τ_{random} defines the distance L the photon packet can travel before the interaction event occurs:

$$\tau_{random} = \int_0^L \rho(\mathbf{r})\kappa_\nu \, d\ell. \tag{6.13}$$

This line integral is sensitive to the direction of travel of the photon and its point of origin. The non-trivial structure of $\rho(\mathbf{r})$ in numerical simulations makes this calculation essentially impossible to perform analytically. Most of the computational effort in MCRT is devoted to a numerical solution of this equation to obtain L.

In the simplest case (a Cartesian grid), the calculation for each cell is relatively straightforward. The density in each cell is constant, so the optical depth across any cell is $\tau_{cell} = \rho_{cell}\kappa_{cell}s$, where s is the path length travelled through the cell by the photon. The calculation of s depends on the direction vector of the photon $\mathbf{n} = (n_x, n_y, n_z)$ and the entry point of the photon into the cell (x, y, z). It is straightforward to calculate the distance along the vector to the next x cell face x_{face}:

$$s_x = \frac{x_{face} - x}{n_x}, \tag{6.14}$$

and similarly for y and z. The path length s is the smallest of these three distances:

$$s = \min(s_x, s_y, s_z). \tag{6.15}$$

This process is then repeated for each cell, with the total optical depth τ_{tot} being the sum of the optical depths in each cell. The process ends when the total optical depth exceeds τ_{random}. As the photon will typically scatter inside a cell, the path length in the last cell is calculated by

$$s = \frac{\tau_{random} - \tau_{tot}}{\rho_{cell}\kappa_{cell}}. \tag{6.16}$$

We should now discuss what happens once the photon interacts with the medium.

6.6 A Detailed Description of Scattering

We have already noted that a photon can undergo one of two types of interaction with the medium: it can be absorbed, or it can be scattered. The albedo determines which type of interaction occurs. Absorption is more straightforward: the photon packet is composed of many photons, so a fraction of the packet is absorbed and the other fraction is scattered. I will now elucidate the more non-trivial framework of scattering below.

6.6.1 Polarisation and Stoke's Representation

Photons can be modelled as electric and magnetic waves oscillating perpendicular to each other (with each oscillating perpendicular to the propagation direction $\mathbf{n}$). As the magnetic wave

is constrained to always be perpendicular to the electric wave, we can discuss the orientation of the waves purely in terms of the electric field, which can be described as a two component vector **E**. Whenever photons undergo scattering of any form, we expect them in general to become *polarised*, that is the two components of **E** should become correlated. This polarisation will in general affect subsequent scatterings of the photon, so we must be able to describe this mathematically. For the sake of computation, we should also attempt to characterise polarisation in as economical a fashion as is possible. I will now demonstrate Stoke's representation, which describes the polarisation state in terms of four parameters, (often known as the *Stokes parameters* or collectively as the *Stokes vector*). This derivation is a modified version of that in Chandrasekhar (1960).

Consider a beam of elliptically polarised light. By definition, **E** will sweep out an ellipse in the plane transverse to the beam's direction. The components E_l and E_r (defined along directions l and r) should satisfy:

$$E_l = E_l^0 \sin(\omega t - \epsilon_l) \tag{6.17}$$

$$E_r = E_r^0 \sin(\omega t - \epsilon_r). \tag{6.18}$$

Note that l and r need not be aligned with the principal axes of the ellipse: our only requirement is that the ratio of amplitudes E_l/E_r is constant, and that the difference in phase $\epsilon_l - \epsilon_r$ is also constant. In the equations above, we will simplify further and set $(E_r^0, E_l^0, \epsilon_l, \epsilon_r, \omega)$ as constants. Now consider the principal axes of the ellipse, which lie at an angle η to l (see Figure 6.1). Defining an orthogonal basis that is aligned with the principal axes, we can write a more simplified decomposition of **E**:

$$E_\eta = E^0 \cos\beta \sin\omega t \tag{6.19}$$

$$E_{\eta+\pi/2} = E^0 \sin\beta \cos\omega t, \tag{6.20}$$

where we have defined β such that $\tan\beta$ gives the ratio of the ellipse axes. Further, the sign of β is positive or negative when the polarisation is right or left handed respectively[2]. The intensity of the beam I is related to the amplitudes in each direction:

$$I = (E^0)^2 = (E_l^0)^2 + (E_r^0)^2 = I_l + I_r. \tag{6.21}$$

We can immediately see the advantages in defining β as we have - it falls out of the above expression as $\cos^2\beta + \sin^2\beta = 1$. We can use the η components of **E** to constrain the values of the constants in equations (6.17) and (6.18). Performing a counter clockwise rotation by our angle η, we can transform E_η to E_l and $E_{\eta+\pi/2}$ to E_r:

[2]Handedness refers to the direction of rotation of **E**. Conventions differ depending on the scientific discipline: in physics and astronomy, right-handed polarisation refers to clockwise rotation, and left-handed to anti-clockwise rotation from the observer's point of view.

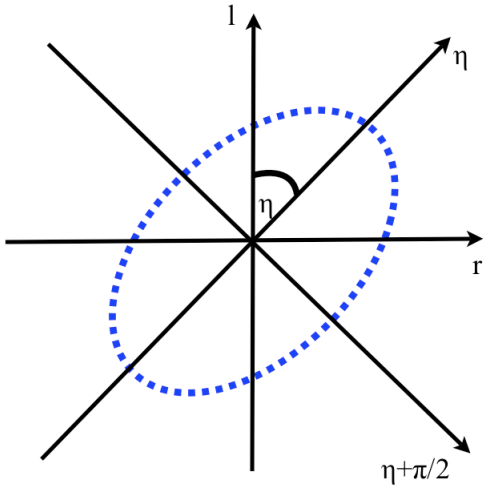

Figure 6.1: An example of elliptical polarisation. The electric field vector sweeps out the ellipse drawn in blue, with an orientation to some axes (l, r) given by the angle η between one axis (in this case l) and the semi-major axis of the ellipse.

$$\begin{pmatrix} E_l \\ E_r \end{pmatrix} = \begin{pmatrix} \cos\eta & -\sin\eta \\ \sin\eta & \cos\eta \end{pmatrix} \begin{pmatrix} E_\eta \\ E_{\eta+\pi/2} \end{pmatrix}. \tag{6.22}$$

This gives

$$E_l = E^0(\cos\beta\cos\eta\sin\omega t - \sin\beta\sin\eta\cos\omega t) \tag{6.23}$$

$$E_r = E^0(\cos\beta\sin\eta\sin\omega t + \sin\beta\cos\eta\cos\omega t). \tag{6.24}$$

We can use the double angle formulae to expand equation (6.17), and compare it with equation (6.23) to find terms with common factors of $\sin\omega t$ and $\cos\omega t$ respectively, giving the equalities

$$E^0\cos\epsilon_l = E_l^0\cos\beta\cos\eta \tag{6.25}$$

$$E^0\sin\epsilon_l = E_l^0\sin\beta\sin\eta. \tag{6.26}$$

Dividing one equation by the other gives

$$\tan\epsilon_l = \tan\beta\tan\eta. \tag{6.27}$$

We can compare equations (6.18) and (6.24) in a similar manner to obtain

$$\tan\epsilon_r = -\tan\beta\cot\eta. \tag{6.28}$$

We have been able to define these without defining the relationship between E^0 and E_l^0. We can rearrange equations (6.25) and (6.26), and equate them:

$$\frac{E_l^0}{E^0} = \frac{\cos\beta\cos\eta}{\cos\epsilon_l} = \frac{\sin\beta\sin\eta}{\sin\epsilon_l}. \tag{6.29}$$

Squaring both sides and adding gives

$$\left(\frac{E_l^0}{E^0}\right)^2 (\cos^2\epsilon_l + \sin^2\epsilon_l) = \left(\frac{E_l^0}{E^0}\right)^2 = \cos^2\beta\cos^2\eta + \sin^2\beta\sin^2\eta, \tag{6.30}$$

and similarly for E_r

$$\left(\frac{E_r^0}{E^0}\right)^2 = \cos^2\beta\sin^2\eta + \sin^2\beta\cos^2\eta. \tag{6.31}$$

The intensity of the beam along the l and r directions is therefore:

$$I_l = (E_l^0)^2 = I(\cos^2\beta\cos^2\eta + \sin^2\beta\sin^2\eta) \tag{6.32}$$

$$I_r = (E_r^0)^2 = I(\cos^2\beta\sin^2\eta + \sin^2\beta\cos^2\eta). \tag{6.33}$$

We can now construct the Stokes parameters (I, Q, U, V):

$$\begin{aligned} I &= (E_l^0)^2 + (E_r^0)^2 = I_l + I_r \\ Q &= (E_l^0)^2 - (E_r^0)^2 = I_l - I_r \\ U &= 2E_l^0 E_r^0\cos(\epsilon_l - \epsilon_r) \\ V &= 2E_l^0 E_r^0\sin(\epsilon_l - \epsilon_r) \end{aligned} \tag{6.34}$$

We can express (Q, U, V) in terms of our original variables:

$$Q = I(\cos^2\beta\cos^2\eta + \sin^2\beta\sin^2\eta - \cos^2\beta\sin^2\eta - \sin^2\beta\cos^2\eta) = I\cos 2\beta\cos 2\eta. \tag{6.35}$$

We can use equation (6.29) (and its equivalent for E_r) to calculate the trigonometric identities to obtain U and V:

$$U = I\cos 2\beta\sin 2\eta \tag{6.36}$$

$$V = I\sin 2\beta \tag{6.37}$$

Note that these relations hold in the case that $E_l^0/E_r^0 = const.$ and $\epsilon_l - \epsilon_r = const.$ (I, Q, U, V) are often written as the components of the four-dimensional Stokes vector $\mathbf{S}$. It can be shown (Chandrasekhar, 1960) that any two beams with the same Stokes parameters are equivalent (in the sense that they cannot be distinguished by optical analyses), and that the resultant Stokes vector obtained by adding two separate beams is simply the sum of the Stokes vectors of the two beams. This also allows the construction of unpolarised light by adding beams of equal intensity but opposite polarisation.

6.6.2 Transformation of the Stokes Parameters

We will find that to correctly evolve the Stokes parameters of a photon as it is scattered in the medium, we must be able to define these on the correct plane - therefore we must be able to transform our Stokes Vector $\mathbf{S} = (I, Q, U, V)$ by rotation of the polarisation axes (defined by η). It should be reasonably obvious that I is invariant under this transformation, as is $V = I \sin 2\beta$. For the other Stokes parameters, a rotation by an angle ψ gives

$$Q' = I \cos 2\beta \cos 2(\eta - \psi) \tag{6.38}$$

$$U' = I \cos 2\beta \sin 2(\eta - \psi). \tag{6.39}$$

Expanding the double angle expressions gives

$$Q' = I \cos 2\beta \cos 2\eta \cos 2\psi + I \cos 2\beta \sin 2\eta \sin 2\psi = Q \cos 2\psi + U \sin 2\psi \tag{6.40}$$

$$U' = I \cos 2\beta \sin 2\eta \cos 2\psi - I \cos 2\beta \cos 2\eta \sin 2\psi = -Q \sin 2\psi + U \cos 2\psi. \tag{6.41}$$

We can use the above expressions to construct the *Mueller matrix* R to transform the Stokes Vector:

$$\mathbf{S}' = R(\psi)\mathbf{S}. \tag{6.42}$$

The matrix is populated by reading off the terms in the above equations:

$$R = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos 2\psi & \sin 2\psi & 0 \\ 0 & -\sin 2\psi & \cos 2\psi & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{6.43}$$

This matrix will now allow us to rotate the Stokes vector in and out of the frame of the scatterer (with respect to the incident photon direction).

6.6.3 The Phase Matrix

We have constructed a matrix for the rotation of the polarisation axes, so it would be sensible to continue in this vein and construct a matrix for the scattering process itself. If we return to the results of section 2.4, a completely unpolarised beam will scatter into a direction Θ according to the phase function $P(\cos\Theta)$. We should therefore aim to construct a matrix that will decompose into the phase function in the event that $Q = U = V = 0$. We could completely populate this 4×4 matrix to determine the probability of scattering a photon into Θ (this in fact is necessary when magnetic fields are involved), but for our purposes we need only specify four parameters $\{M_1, M_2, M_3, M_4\}$, and the albedo a (Pinte et al., 2006; Forgan & Rice, 2010a):

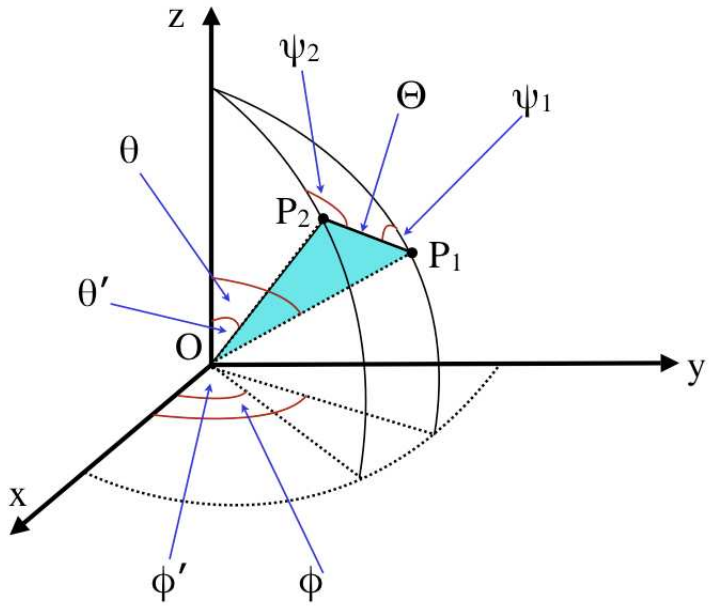

Figure 6.2: The scattering of a photon from (θ, ϕ) to (θ', ϕ'). The plane of scattering OP_1P2 is shaded in light blue - the scattering matrix acts on the Stokes parameters assuming that the polarisation axes are defined in the plane of scattering - therefore S must be rotated in and out of this plane when the calculation is performed.

$$M(\Theta) = a \begin{bmatrix} M_1 & M_2 & 0 & 0 \\ M_2 & M_1 & 0 & 0 \\ 0 & 0 & M_3 & -M_4 \\ 0 & 0 & M_4 & M_3 \end{bmatrix}. \tag{6.44}$$

Given Θ, we can now update the Stokes parameters

$$\mathbf{S}' = M(\Theta)\mathbf{S}, \tag{6.45}$$

but only if the Stokes vector is defined relative to the *plane of scattering*. This plane is defined by the direction of the incident photon and its subsequent direction when scattered. If we have defined the Stokes vector relative to some other origin, then we must perform two rotations of S: the first (ψ_1) will rotate it into the plane of scattering so that the matrix can be correctly used, and the second will be used to rotate the resulting vector back into its original frame (ψ_2):

$$\mathbf{S}' = R(\pi - \psi_2)M(\Theta)R(-\psi_1)\mathbf{S}. \tag{6.46}$$

The correct values for ψ_1 and ψ_2 can be seen from Figure 6.2. The plane of scattering is defined by OP_1P_2 (shaded in light blue), whereas S is originally defined relative to the meridian OP_1z.

The angle between these two planes gives ψ_1 above. Once the scattering matrix has been used, we must rotate back to the meridian OP_2z using ψ_2. In practice, we use the complement of ψ_2 as it is easier to calculate.

This allows a complete description of the photon packet before and after scattering: the only requirement is to define the elements of M. In the case of dust scattering, we use a single peaked Henyey-Greenstein function for the phase function M_1 (see White 1979). The other parameters are defined as a function of M_1:

$$M_1 = \frac{1-g^2}{(1+g^2-2g\cos\Theta)^{3/2}}$$
$$M_2 = -p_l M_1 \frac{1-\cos^2\Theta}{1+\cos^2\Theta}$$
$$M_3 = M_1 \frac{2\cos\Theta}{1+\cos^2\Theta}$$
$$M_4 = -p_c M_1 \frac{1-\cos^2\Theta_f}{1+\cos^2\Theta_f}. \tag{6.47}$$

g is the scattering asymmetry parameter: $g = 0$ defines isotropic scattering, and $g = 1$ defines forward throwing scattering. p_l and p_c are the peak linear and circular polarisation respectively; and

$$\Theta_f = \Theta(1 + 3.13\, s\, e^{-7\Theta/\pi}), \tag{6.48}$$

where $s = 1$ (White, 1979). If the incident radiation is unpolarised, then we recover the phase function form as required:

$$I' = M_1 I. \tag{6.49}$$

We can sample the scattering angle Θ from the CDF:

$$F(\Theta) = \frac{\int_0^\Theta M_1 \sin\Theta' d\Theta'}{\int_0^\pi M_1 \sin\Theta' d\Theta'}. \tag{6.50}$$

We can integrate this by substitution of $\mu = -\cos\Theta$ and rearrange to achieve a generating function for Θ (substituting $F(\Theta)$ with our uniform variable Y):

$$\cos\Theta = \frac{1+g^2-\left[\frac{1-g^2}{1-g+2Y}\right]^2}{2g}. \tag{6.51}$$

We must also calculate the azimuthal angle ϕ. If the incident radiation is unpolarised, then the azimuthal angle is uniformly distributed around 2π. If the photon has a non-zero linear polarisation P, where

$$P = \frac{\sqrt{Q^2+U^2}}{I}, \tag{6.52}$$

then the CDF for ϕ is modified thus (Pinte et al., 2006):

$$F_{\Theta}(\phi) = \frac{1}{2\pi}\left(\phi - \left(\frac{M_1 - M_2}{M_1 + M_2}\right)\frac{P}{2}\sin 2\phi\right). \tag{6.53}$$

In general, whether the photon is absorbed or scattered, the total number of photons is conserved by forcing absorbed photons to be immediately re-emitted. As MCRT deals with photon packets, energy can be conserved by reducing the energy of each photon packet after a scattering/absorption event by a factor equal to the local albedo. This equates with the concept of a fraction of the photons in the packet being absorbed by the medium, and the complementary fraction being scattered. Other methods can also be used, e.g. "killing" a photon if the local albedo is less than a randomly sampled value, which saves computing emission from low-intensity packets.

6.7 Imaging

When photons exit the medium, they are captured on an image plane, oriented at user-specified angles θ_{im}, ϕ_{im} to the system, at a fixed distance d (see Figure 6.3). They are then binned by their (x, y) coordinates on this plane to provide a pixelated image, averaged over solid angle (analogous to imaging in a CCD). If spectra are of interest to the user, then these can also be obtained by binning in λ (or indeed, an entire datacube can be obtained by binning in all three). "Classic" MCRT methods do not specify a single viewing angle, and instead bin the photons over several lines of sight, which allow the construction of a series of image planes from one simulation. However, such "multi-plane" simulations will have to run many more photons than a "one-plane" simulation, to maintain a comparably low level of random error associated with each pixel[3]. For this reason the code used in this work adopts the "one-plane" imaging scheme.

We calculate the photon's location on the image plane by deprojecting its exit coordinates (x, y, z) onto the 2D plane. For a plane oriented at angles (θ_{im}, ϕ_{im}) to the system, we must perform two rotation transformations, firstly in ϕ and secondly in θ. The resulting location of the photon on the image plane is then given by

$$x_{im} = z\sin\theta - y\cos\theta\sin\phi - x\cos\theta\cos\phi \tag{6.54}$$

$$y_{im} = y\cos\phi - x\sin\phi. \tag{6.55}$$

The pixel bin for the photon is calculated from (x_{im}, y_{im}), and the photon's intensity is added to the bin. As the photons are evolved stochastically, the images themselves are subject to random sampling errors. For any given "pixel", the error is

$$\sigma_{ij} = \frac{F_{ij}}{\sqrt{N_{ij}}}. \tag{6.56}$$

[3]This is also a factor when comparing one-plane simulations with different pixel resolutions.

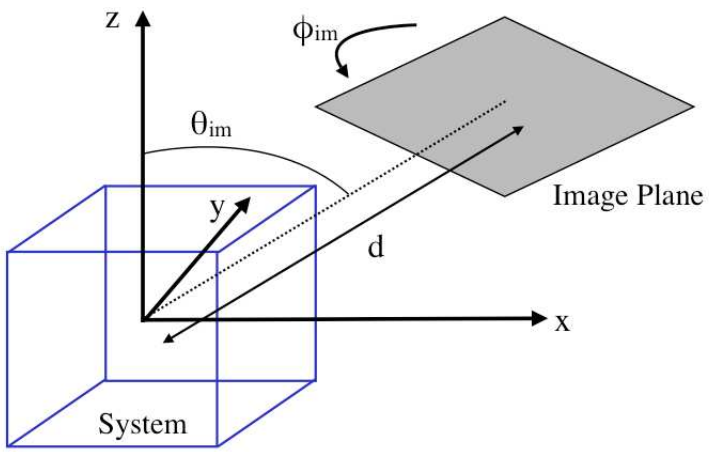

Figure 6.3: Defining an image plane .

Where F_{ij} is the flux in that pixel, and N_{ij} is the number of photons in that pixel. Therefore, in order to achieve acceptable signal-to-noise in these images, a large number of photons must be used. In order to ameliorate this, a "peeling off" weighting scheme can be used (cf. Yusef-Zadeh et al. 1984), where at each scattering event, flux is "peeled off" and sent to the observer (with a weight proportional to the probability that a photon scattered at that location would reach the observer).

Two types of photons are emitted: direct photons and scattered photons. Direct photons do not scatter in the medium: they are emitted towards the observer, and are weighted

$$W_{direct} = \frac{e^{-\tau}}{4\pi d^2}, \tag{6.57}$$

where τ is the optical depth between the emission location and the observer, and d is the distance. The scattered photons are forced to scatter at least once in the medium as follows: emission occurs in a random direction, and the optical depth to the edge of the grid, τ_{edge}, is calculated. Then, a random optical depth is sampled such that the photon must scatter before exit:

$$\tau = -\log\left[1 - Y(1 - e^{-\tau_{edge}})\right] \tag{6.58}$$

The above equation maintains the correct probability distribution for τ. The photon is then allowed to scatter normally until it exits the medium. Each scattering event causes some of the flux in the photon packet to be "peeled off" and received by the observer. This portion is described by the weight

$$W_{scatt} = a^N \left(1 - e^{-\tau_{edge,1}}\right) e^{-\tau_{edge,2}} P(\cos\Theta). \tag{6.59}$$

N is the number of scattering events so far, $\tau_{edge,1}$ is the value of the optical depth to the grid's edge at the point of emission, $\tau_{edge,2}$ is the value at the current position, and $P(\cos\Theta)$ is the scattering phase function. The total intensity received is then the sum of the two weights, W_{scatt} and W_{direct}, over all photons emitted.

6.8 Radiative Equilibrium

We have assumed until now that the temperature structure of the system is known. However, this need not be the case for MCRT to work - in fact, MCRT can be used to calculate the temperature structure, employing the detailed scattering and absorption to create an SED which reflects the complex geometry and chemical composition of the system under study.

The principal method for calculating SEDs from MCRT are so-called *radiative equilibrium methods*. These methods assume that the system is in local thermodynamic equilibrium - the energy absorbed in a grid cell is equal to the energy emitted. By doing so, the temperature structure can be relaxed to an equilibrium solution. There are several ways to do this - I will describe the method proposed by Bjorkman & Wood (2001), as it does not require iteration (as most radiative equilibrium methods do), and it exactly conserves energy.

Recall that the luminosity of the system L_{tot} is discretised into N_γ photon packets. The energy in each packet is given by:

$$E_\gamma = \frac{L_{tot}\Delta t}{N_\gamma}, \tag{6.60}$$

where Δt is some time interval. Therefore, if a cell i absorbs N_i photons, the amount of energy absorbed is simply

$$E_i^{abs} = N_i E_\gamma. \tag{6.61}$$

The energy emitted by a cell in LTE is given by

$$E_i^{emit} = 4\pi\Delta t \int dV_i \int \rho_i \kappa_\nu B_\nu(T_i) d\nu. \tag{6.62}$$

By using the Planck opacity

$$\kappa_P(T) = \frac{\int \kappa_\nu B_\nu(T) d\nu}{\int B_\nu(T) d\nu}, \tag{6.63}$$

this becomes

$$E_i^{emit} = 4\pi\Delta t m_i \kappa_P(T_i) \int B_\nu(T_i) d\nu, \tag{6.64}$$

where we have integrated over volume to obtain the mass in the cell m_i. The integral of the Planck function is equal to $B = \sigma T^4/\pi$. Equating the energy emitted with the energy absorbed then gives

$$N_i \frac{L_{tot}}{N_\gamma} = 4 m_i \kappa_P(T_i) \sigma T_i^4. \tag{6.65}$$

We can rearrange for T to obtain an implicit form:

$$\sigma T_i^4 = \frac{N_i L_{tot}}{4 N_\gamma \kappa_P(T_i) m_i}. \tag{6.66}$$

This equation must be solved every time a photon is absorbed by the cell to calculate the cell's new temperature. Simple iteration can achieve this for dust scattering, as the Planck opacity is a slowly increasing function of temperature.

To conserve energy, we immediately re-emit the photon after it has been absorbed. However, we should be cognisant of the cell's slight increase in temperature ΔT. It will have previously emitted photons with an emissivity given by

$$j_\nu = \kappa_\nu B_\nu(T_i). \tag{6.67}$$

After absorbing the photon, it will now emit according to

$$j'_\nu = \kappa_\nu B_\nu(T_i + \Delta T). \tag{6.68}$$

This change in emissivity changes the frequency distribution from which photons are emitted from the cell. Therefore, photons emitted previously will not have been emitted with the correct frequency distribution. We need to correct for this by considering the change in j. If the increase in temperature is small enough (which will be true if the energy per photon E_γ is small, i.e. N_γ is large), then we can approximate the change in j by Taylor expansion:

$$\Delta j_\nu = j'_\nu - j_\nu = \kappa_\nu B_\nu(T_i + \Delta T) - \kappa_\nu B_\nu(T_i) \approx \kappa_\nu \Delta T \frac{dB_\nu}{dT}. \tag{6.69}$$

This gives the shape of the probability distribution function from which to re-emit the photon packet. This conserves energy and corrects the frequency distribution. The correct normalised PDF is given by:

$$P(\nu) = \frac{\kappa_\nu}{C} \left(\frac{dB_\nu}{dT} \right)_{T=T_i}, \tag{6.70}$$

where

$$C = \int \kappa_\nu \left(\frac{dB_\nu}{dT} \right) d\nu. \tag{6.71}$$

When the simulation begins, the temperature throughout is zero. Photon packets are emitted from some source, and absorb and scatter in the medium according to the albedo. The temperatures of the cells begin to rise as absorptions occur: as N_i becomes large, the temperature of the cells converge on the solution, producing the correct emergent SED.

6.9 A History of MCRT in SPH

The most conceptually straightforward method to perform MCRT in an SPH system is to smooth the distribution onto a grid (Oxley & Woolfson, 2003; Kurosawa et al., 2004; Stamatellos & Whitworth, 2005; Bisbas et al., 2009; Acreman et al., 2010). However, the choice of grid geometry will influence the final gridded field, and adaptive mesh refinement is almost always required. Gridding SPH systems in this fashion must be done carefully to avoid creating axis-aligned features not present in the original SPH density field.

The alternative is to utilise ray-tracing techniques directly on the SPH fields, allowing the full power of the SPH formalism to be applied to the calculation of optical depths. This has been successfully achieved using a variety of tracing methods for photoionisation studies (Kessel-Deynet & Burkert, 2000; Dale et al., 2007; Altay et al., 2008; Pawlik & Schaye, 2008)

However, these methods currently only calculate optical depths along the full extent of the ray for the purpose of photoionisation, etc. For imaging circumstellar discs, we must include scattering and polarisation. This requires us to produce a ray tracing method that calculates the optical depth through an arbitrary distance in the SPH field, so we can determine the scattering location accurately. The ray tracing method I will describe in this chapter is motivated by the techniques used by Altay et al. (2008) in their SPHRAY code.

6.10 Photon Emission in an SPH density field

The SPH simulations used in this work produce temperatures for each SPH particle self-consistently, so the system is assumed to already be in temperature equilibrium. It is also assumed that the dust is thermally coupled to the gas, and that $T_{dust} = T_{radiation}$. This is suitable for most purposes, except where significant stellar irradiation dominates the radiation field (not modelled in the simulations described in this thesis). However, as has been already said, radiative equilibrium techniques can be used to sidestep this issue: an implementation of radiative equilibrium in SPH fields is discussed in a later section.

It is reasonable to (initially) assume all objects emit according to a blackbody spectrum, which then gives:

$$L_{star} = 4\pi^2 R_s^2 \int_{\nu_{min}}^{\nu_{max}} B_\nu(T_s) d\nu \tag{6.72}$$

$$L_{gas} = M_{gas} \int \left[\int_{\nu_{min}}^{\nu_{max}} \epsilon_\nu(\mathbf{r}) d\nu \right] d\mathbf{r}, \tag{6.73}$$

for source and diffuse emission respectively (given the frequency range $[\nu_{min}, \nu_{max}]$ of interest), where $\epsilon_\nu(\mathbf{r})$ is the emissivity interpolated from the SPH particle field, i.e.

$$\epsilon_\nu(\mathbf{r}) = 4\pi \sum_j \frac{\kappa_\nu B_\nu(T_j) m_j}{\rho_j} W(\mathbf{r} - \mathbf{r}_j, h), \tag{6.74}$$

where the sum over j indicates a sum over nearest neighbours, κ_ν is the absorptive opacity, T_j is the temperature of each SPH particle, m_j and ρ_j are the masses and densities respectively, and W is the smoothing kernel (see Chapter 3 for more detail on the SPH formalism).

There is an extra subtlety regarding the frequency distribution of photons emitted from the gas. The frequency distribution will depend on the local emissivity, which as shown above is an interpolative sum. While in theory the full form of equation (6.73) should be used to calculate gas luminosity (and the frequency distribution of the photons it emits), this chapter approximates the sum using the contribution from individual particles only, i.e.

$$L_{gas} = \sum_i L_i = \sum_i 4\pi M_i \int_{\nu_{min}}^{\nu_{max}} \kappa_\nu B_\nu(T_i) d\nu. \tag{6.75}$$

This saves computational expense, and is sufficiently accurate for the examples in this thesis, bearing in mind a) that the approximation breaks down only in the inner regions of the disc, which are already under-resolved for other reasons (see section 6.14.2) and b) the effective resolution of the images is too low for this effect to be significant.

The luminosity of each object (whether pointmass or SPH particle) defines how many photon packets are emitted from that object using

$$N_{\gamma,object} = N_{\gamma,tot} \left(\frac{L_{object}}{L_{tot}} \right) \tag{6.76}$$

6.11 Optical Depths in an SPH density field

As has been said previously, the key component of any MCRT code (and the source of the greatest computational burden) is the calculation of optical depths. This calculation requires a specification of the density field at all locations, which can be given by the SPH formalism. To recap, the discretisation of a scalar field $A(\mathbf{r})$ using an SPH particle distribution is performed using the interpolation

$$A(\mathbf{r}) = \sum_j \frac{A_j m_j}{\rho_j} W(\mathbf{r} - \mathbf{r}_j, h) \tag{6.77}$$

where A_j is the value of A at the location of particle j, m_j is the mass of particle j, ρ_j is the density of particle j, and W is the smoothing kernel. The summation is typically over N_{neigh} nearest neighbour particles to the location $\mathbf{r}$. The *smoothing length* of particle j, h_j, is defined such that a sphere of radius $2h_j$ will contain the N_{neigh} nearest neighbour particles to j. For example, to calculate density, substitute $A = \rho$:

$$\rho(\mathbf{r}) = \sum_j m_j W(\mathbf{r} - \mathbf{r}_j, h). \tag{6.78}$$

The sphere that contains the N_{neigh} nearest neighbours (i.e. a sphere of radius $2h_j$) is referred to as the *smoothing volume*. As was discussed in section 3.2.3, there are two means by which to

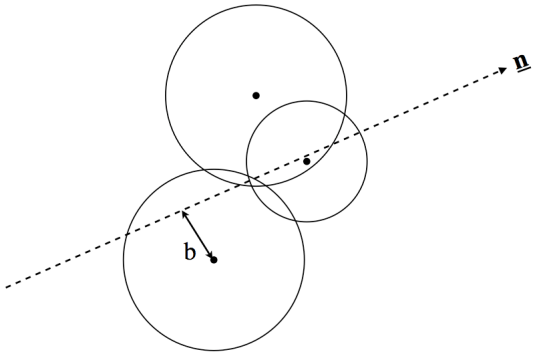

Figure 6.4: Illustrating the "scatter" method of raytracing. Particles only contribute to the density along the ray if their smoothing volume intersects it.

estimate density: the first is the so-called "gather" method, where the smoothing length $h = h_i$ is defined for the location r_i, and

$$\rho(\mathbf{r}_i) = \sum_j m_j W(\mathbf{r}_i - \mathbf{r}_j, h_i), \tag{6.79}$$

where the index j refers to all particles which are contained within a radius $2h_i$ of the location r_i. The second method (which is used in this work and in SPHRAY, Altay et al. 2008) is the "scatter" method. The smoothing length $h = h_j$ is used - the density at any one location is calculated by adding the contributions from N particles whose smoothing volume intersects the location:

$$\rho(\mathbf{r}_i) = \sum_{j=1}^{N} m_j W(\mathbf{r}_i - \mathbf{r}_j, h_j). \tag{6.80}$$

In the context of ray tracing, the density along the ray is affected only by particles with smoothing volumes that intersect it (see Figure 6.4). By determining which particles intersect the ray, the rest of the particle distribution can be ignored for the purposes of calculating optical depth, reducing computational expense (whereas with the gather method, the ensemble of particles contributing to the calculation changes significantly with position, and requires the inclusion of a larger subset of SPH particles to perform the calculation).

If we denote the ray's path as the vector $\mathbf{n}$ (with length n and infinitesimal vector segment $\mathbf{dn}$) we can express the column density Σ along the ray as the following line integral:

$$\Sigma = \int_0^L \rho(\mathbf{n}).\mathbf{dn} = \int_0^L \sum_{j=1}^{N} [m_j W(|\mathbf{n} - \mathbf{r}_j|, h_j)]\, dn, \tag{6.81}$$

which can be rearranged to give

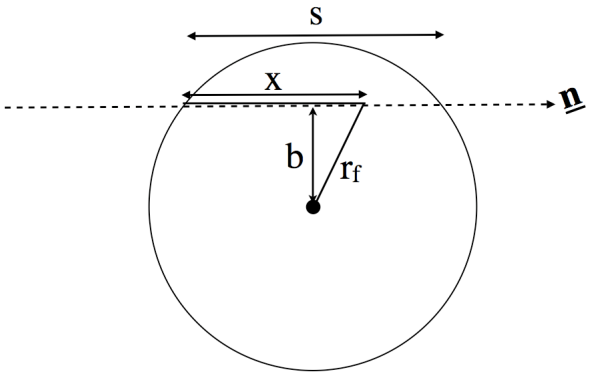

Figure 6.5: Optical Depths through a single smoothing volume .

$$\Sigma = \sum_{j=1}^{N} \left[\int_0^L m_j W(|\mathbf{n} - \mathbf{r}_j|, h_j) dn \right]. \tag{6.82}$$

The integral is now decomposed into N integrals, where N is the number of particles intersected by the ray. Each integral is defined by the impact parameter b (see Figure 6.4). The calculation itself can be performed for a smoothing volume of $h = 1$, and scaled upwards (this is due to the construction of the smoothing kernel). The entire optical depth calculation has been decomposed into the repetition of a single algorithm for calculating the optical depth through a single smoothing volume. This calculation will now be expounded.

The Optical Depth Calculation for A Single Particle

Consider Figure 6.5. The ray (with direction vector $\mathbf{n}$) intersects the sphere with impact parameter b. If the ray penetrates a distance x into the sphere (out of a total possible distance defined by the chord s), then the integral can be defined analytically, given the functional form of W. Let us define

$$\tilde{r} = r/h \tag{6.83}$$

$$\tilde{b} = b/h \tag{6.84}$$

$$\tilde{x} = x/h \tag{6.85}$$

$$\tilde{r}_f = r_f/h, \tag{6.86}$$

where the final scattering radius r_f can be calculated simply by Pythagoras' Theorem:

$$r_f = \sqrt{(s/2 - x)^2 + b^2}. \tag{6.87}$$

The integral in equation (6.82) becomes:

$$I = \begin{cases} \int_{\tilde{r}_f}^{2} W(\tilde{r}')d\tilde{r}' & x < s/2 \\ \int_{\tilde{b}}^{2} W(\tilde{r}')d\tilde{r}' + \int_{\tilde{b}}^{\tilde{r}_f} W(\tilde{r}')d\tilde{r}' & x > s/2 \end{cases} \tag{6.88}$$

The column density through a smoothing volume (for any impact parameter and any distance into the sphere) can now be calculated; multiplying by an opacity then provides an optical depth. Typically, W is constructed using cubic splines (Monaghan, 1992). The kernel used in this work is

$$W(\tilde{r}) = \begin{cases} 1 - \frac{3}{2}\tilde{r}^2 + \frac{3}{4}\tilde{r}^3 & \tilde{r} < 1 \\ \frac{1}{4}(2 - \tilde{r})^3 & 1 < \tilde{r} < 2 \\ 0 & \tilde{r} > 2 \end{cases} \tag{6.89}$$

This provides compact support (i.e. it reduces to zero outside the smoothing volume), and is simple to integrate.

6.12 Optimising the Code

The optical depth calculation for one sphere must now be scaled up to many spheres. This requires the construction of a scheme for efficiently calculating ray/sphere intersections. To this end, the code creates a data object called a *raylist*, which stores (in order of intersection) all particles that the ray (given its origin and direction vector) will intersect. Once the list is created, the optical depth can be calculated quickly using equation (6.80).

The construction of the raylist must be computationally efficient for the code to be effective. The procedure is similar to that implemented by Altay et al. (2008) in SPHRAY; the code constructs an octree (as described in section 3.2.6) to spatially index the particles efficiently (as there may be density changes over several orders of magnitude). The cells either contain child cells, or particles (the *leaf cells*). The tree is constrained to have a maximum number of particles in each leaf. All cells have an associated Axis Aligned Bounding Box (AABB), which is the minimum box size, aligned to the three cartesian axes, to contain all the smoothing volumes of the particles in the cell (see Figure 6.6).

This allows the determination of intersections between the ray and the cells (or more correctly, their AABBs). Starting with the root cell, each child cell is tested for intersection, constituting a walk through the tree. If a leaf cell is intersected by the ray, then the particles in the leaf are tested for intersection (by calculating their impact parameters). This ensures that only a minimum fraction of the particles in the system need testing for intersection. This illustrates the necessity of AABBs; tree nodes may contain a particle, but not its entire smoothing volume. Thus, calculating intersections between a ray and tree nodes may miss contributions to the density field from smoothing volumes that cross node intersections.

Tests for intersections between rays and AABBs are carried out using the ray slopes algorithm (Eisemann et al., 2007). The essence of the algorithm involves a series of 2D tests of the

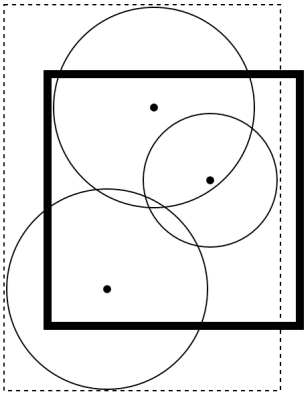

Figure 6.6: Schematic of an Axis Aligned Bounding Box (AABB). For a given set of occupants of a tree cell (solid line), the AABB (dashed line) is set to the minimum dimensions required to completely contain their smoothing volumes.

slope of the ray against the appropriate face of an AABB. For example, consider a ray with direction vector $\mathbf{n} = (n_x, n_y, n_z)$ and origin $\mathbf{o} = (o_x, o_y, o_z)$, being tested for intersection with an AABB, which we can define using its minimum and maximum extent in each coordinate: $\mathbf{b}_{min} = (x_0, y_0, z_0)$, and $\mathbf{b}_{max} = (x_1, y_1, z_1)$. For the ray to intersect the AABB, it must intersect three 2D faces constructed by deprojecting the AABB into the $x-y$, $x-z$ and $y-z$ planes respectively (the AABB can of course construct six, but we can use the direction of the ray to eliminate the testing of the other three). For example, in the $x-y$ plane (for a ray with positive values for all elements of $\mathbf{n}$), we construct its 2D slope as

$$S_{xy}(\mathbf{n}) = \frac{n_y}{n_x} \tag{6.90}$$

For this ray to intersect our 2D plane, defined by a box with vertices at (x_0, y_0), (x_0, y_1), (x_1, y_0), (x_1, y_1), (see Figure 6.7), we require the ray's slope to be larger than a minimum value given by drawing a line $\mathbf{a}$ from $\mathbf{o}$ to (x_1, y_0), and smaller than a maximum value given by drawing a line $\mathbf{b}$ from $\mathbf{o}$ to (x_0, y_1). These lines can be parametrised in a similar fashion to $\mathbf{n}$. The two criteria to be satisfied for the ray to intersect the plane are therefore

$$S_{xy}(\mathbf{n}) > S_{xy}(\mathbf{a}), \quad S_{xy}(\mathbf{n}) < S_{xy}(\mathbf{b}) \tag{6.91}$$

This is repeated similarly for the $x-z$ and $y-z$ planes. If any of the three intersection tests fail, then the ray will not intersect the AABB, and the calculation will end. The algorithm is very simple - its only complication is its need to classify the ray's direction vector into one of 26 permutations of positive, negative and zero components, which need only be done once for every ray. Eisemann et al. (2007) show in tests that it out-performs other commonly used methods, such as using Plücker coordinates (Mahovsky & Wyvill, 2004).

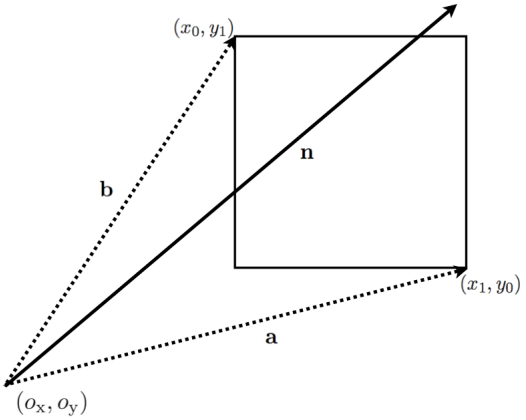

Figure 6.7: Demonstrating the ray slopes algorithm for ray-AABB intersection testing. The ray (indicated by the vector **n** emanating from origin (o_x, o_y)) successfully passes the first of three intersection tests (one for each of the Cartesian planes) as its 2D slope is greater than the slope of vector **a** and less than that of the slope of vector **b** (which all have the same origin). The ray must pass similar tests in the other two planes for a successful intersection with the AABB.

Determining the Scattering Location

In general, the scattering location will occur inside a smoothing volume, and possibly at a location where the density depends on the contributions from several particles. Therefore, when attempting to determine the scattering location, it is important to define four classes of particle:

(i) Particles that do not intersect the ray (*unlisted*)

(ii) Particles that intersect, but do not contain the location of emission (*distant-listed*)

(iii) Particles that intersect, and contain the location of emission in front of them (*front-listed*)

(iv) Particles that intersect, and contain the location of emission behind them (*back-listed*)

The classes are illustrated in Figure 6.8. Particles of class (i) obviously do not affect the calculation - particles of class (ii) are accounted for simply. Particles of classes (iii) and (iv) will have differing effects on the optical depth calculation, and will require separate treatments.

The scattering location is determined by iteration: firstly, the optical depth is calculated particle by particle using the raylist until the optical depth exceeds the randomly selected optical depth $\tau_{scatter}$ at particle k. Then, the optical depth is calculated from the beginning of the sphere for particle $(k-1)$ (ensuring that all potential contributors before and after this location are accounted for), iterating over distance until the answer converges on $\tau_{scatter}$. As the optical depth always increases with distance, convergence can be achieved with simple algorithms and relatively little computation. This code uses a recursive bisector algorithm to perform the iteration. Starting from the path length between the beginning of sphere $(k-1)$ to

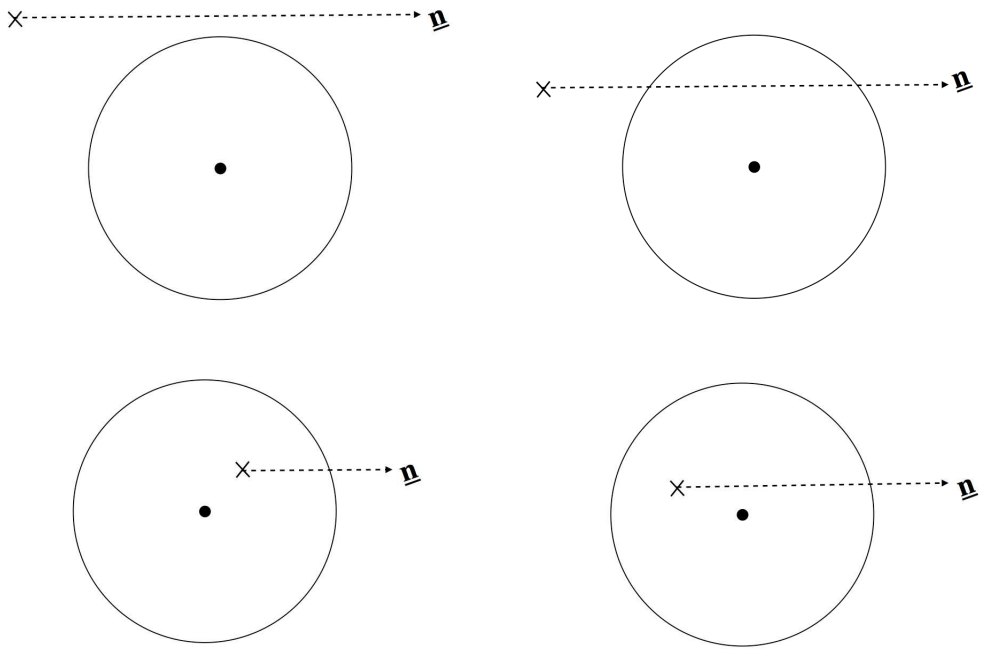

Figure 6.8: The four classes of SPH particle in raytracing: unlisted (top left), distant-listed (top right), front-listed (bottom left) and back-listed (bottom right). The "x" denotes the emission location of the photon.

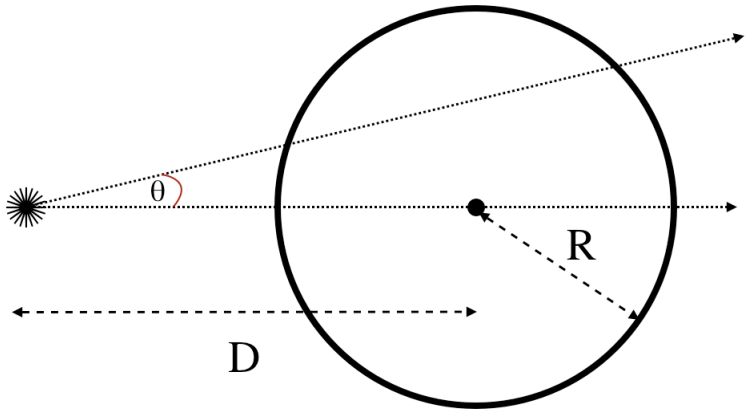

Figure 6.9: Schematic of the raytracing experiment. The optical depth of the ray as a function of θ can be calculated analytically to compare with the code's output.

the end of sphere k, this value is halved recursively until the correct optical depth is obtained (to within some tolerance) or until the path length reaches a minimum value (defined as a fraction of the smallest smoothing length in the simulation).

6.13 Tests and Applications

6.13.1 Comparison with Analytic Results

To confirm the raytracing component of the code was working correctly, a simple test case was devised (Figure 6.9). Consider a uniform density sphere, with radius R, density ρ_0 and total opacity χ. A point source is located a distance D from the centre of the sphere. It emits rays at an angle θ from the vector connecting the source and the sphere's centre. The optical depth $\tau(\theta)$ therefore has an analytic solution:

$$\tau(\theta) = 2\rho_0\chi\sqrt{R^2 - D^2\sin^2(\theta)} \tag{6.92}$$

Three SPH snapshots were generated, each containing a uniform density sphere (mass $1\,M_\odot$, $R = 2133$ AU). Three were generated to check convergence: snapshot 1 used 10^5 SPH particles; snapshot 2 used 5×10^5; snapshot 3 used 10^6. A point source was placed at a distance $D = 4000$ AU, and the optical depth along the ray (assuming $\chi = 1$) was calculated as a function of θ. The numerical results are compared with the analytical result (scaled such that the maximum optical depth is 1) in Figure 6.10. The column density is subject to the underlying random noise (at a level of around 5%) associated with generating an SPH snapshot (which has not undergone any settling). However, the results vary little with increasing particle number, showing that the column density has converged even for the relatively low particle number of 10^5.

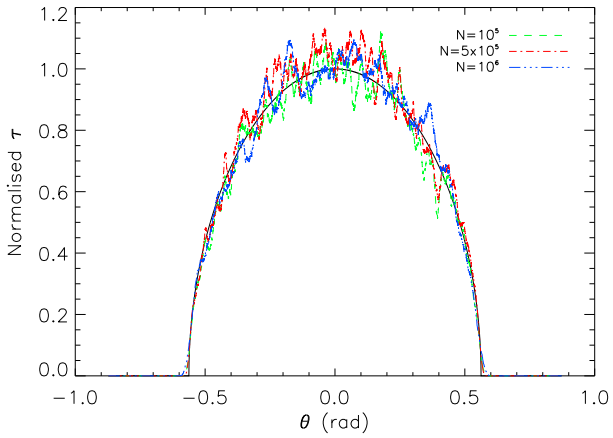

Figure 6.10: Results of the raytracing experiment for three uniform density spheres (with underlying particle noise). The black solid line indicates the analytical result; the coloured dashed lines show simulation results for increasing particle number.

To confirm the role of noise in the SPH snapshot, a fourth sphere with 10^5 particles was generated without noise. This was done by creating a cubic lattice of SPH particles and extracting the sphere from it. The results can be seen in Figure 6.11. Comparing the optical depth in the noisy sphere (green line) with the optical depth in the noise-free sphere (blue line) shows a dramatic improvement. The noise-free sphere has a more uniform density, and is therefore more able to reproduce the idealised conditions of the experiment, matching the analytic curve (black line) much more accurately.

This result is something of a double-edged sword - on the one hand, it clearly shows the dangers of using noisy SPH simulations to represent more idealised objects such as homogeneous spheres for raytracing. Conversely, it shows the sensitivity of the raytracing method - it can probe the underlying noise very well. If the SPH system to be imaged has been "relaxed" such that its noise content is reduced, then this sensitivity will reveal the "true" density field just as accurately.

6.13.2 A Low Mass Companion for HL Tau?

Greaves et al. (2008) imaged HL Tau using the Very Large Array (VLA) with a resolution of 0.08" at a wavelength of 1.3cm (corresponding to a spatial resolution of 10 AU at HL Tau's distance of $\sim$ 140 pc). They discovered excess emission at $\sim$ 65 AU, which they identified as a candidate protoplanet in the earliest stages of formation, a possible example of protoplanetary disc fragmentation forming a bound object (Boss, 1997). To lend weight to their hypothesis, they conducted SPH simulations (containing 250,000 particles) of an unstable star-disc system with similar parameters to HL Tau, in which a clump forms at $\sim$ 75 AU, with a similar mass as that deduced for the candidate (see Figure 6.12). The simulation uses an SPH code

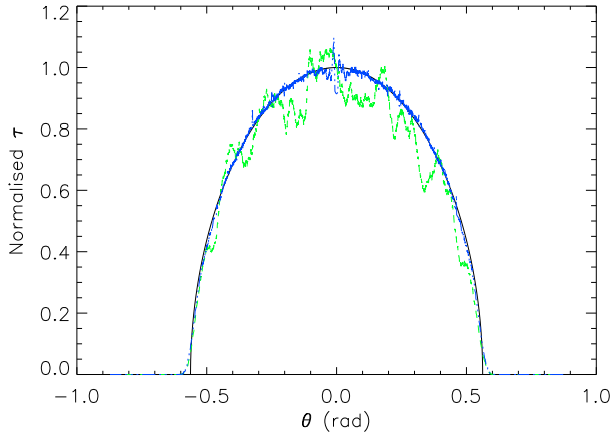

Figure 6.11: Comparing raytracing results for a $N = 10^5$ sphere with Poisson noise (green line) and without Poisson noise (blue line).The black line indicates the analytical curve.

based on the work of Bate et al. (1995). It employs individual particle timesteps, and models hydrodynamics and gravity, with radiative cooling as prescribed by Stamatellos et al. (2007b). If the simulation is an accurate approximation of HL Tau, an interesting question to answer is: what can a telescope like ALMA be expected to see in HL Tau? Will the candidate protoplanet be resolvable at millimetre wavelengths?

The SPH simulation from Greaves et al. (2008) was imaged to answer this question, using the same dust scattering parameters as used in Wood et al. (2002)[4]. The wavelength and resolution parameters were set to be representative of ALMA [5]: the wavelength range is $[0.01, 0.1]$ cm, with an effective resolution of 0.01". The star is $0.5\,M_{\odot}$, and the disc is $0.3\,M_{\odot}$, with an initial surface density profile of $\Sigma \propto r^{-1}$. A dust to gas ratio of 0.005 (corresponding to 50% solar metallicity) was assumed throughout. Dust sublimation is prescribed by enforcing any SPH particle with temperature greater than 1600 K to have zero dust mass. Temperatures were calculated using the equation of state outlined in section 3.4.5. The star's emission was modelled by a blackbody source with radius $2R_{\odot}$ and temperature 2000K.

Figure 6.12 shows that the clump, which reaches temperatures of around 1500 K at its centre, is detectable by its emission, although somewhat fainter than the main star-disc system. In fact, it should be expected that the clump is fainter still, if dust sublimation is more appropriately modelled as opposed to a one temperature cut-off. The $m = 2$ spiral mode attached to the clump is also apparent; however with a flux of around 0.001 mJy/beam, ALMA will find it challenging to image it, highlighting the need for resolution of order ~ 1 AU at 100 pc *and* sufficient sensitivity in the detection of disc spiral arms. The integrated emission is around 0.5 Jy, which is within a factor of two of the observed SCUBA measurements of HL Tau [6]. The remaining discrepancy is most likely due to uncertainty in the selected disc and star parameters

[4]Note that the examples shown here approximate $\chi_\nu = \kappa_\nu$
[5]http://www.eso.org/sci/facilities/alma/
[6]http://www.jach.hawaii.edu/JCMT/continuum/calibration/sens/calibrators.html

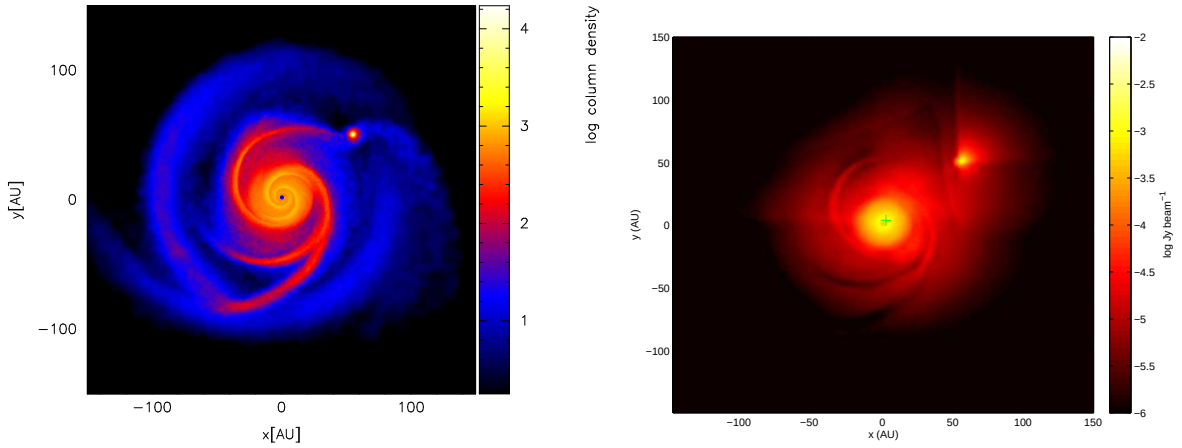

Figure 6.12: Left: Surface density plot of the HL Tau simulation to be imaged. Right: image of the HL Tau simulation, integrated over the wavelength range $[0.01, 0.1]$ cm. The green cross indicates the pixel with maximum intensity.

(as well as the dust data). To summarise, the clump is indeed detectable with ALMA, but a telescope with better sensitivity is required to detect the spiral arms correctly.

6.13.3 Observational Features of Stellar Encounters

We now return to the simulations of Chapter 5 to investigate what observational features ALMA may detect in self-gravitating discs as a result of a stellar encounter. We image Simulation 1 (see Table 5.1) at two epochs - the initial pre-encounter phase, where the disc remains in a marginally stable, self-gravitating state; and during periastron, where the disc is heated strongly by the secondary's motion through it.

The primary is $0.5\,M_{\odot}$ (modelled by a blackbody source with radius $2R_{\odot}$ and temperature 2000 K), and the secondary is $0.1\,M_{\odot}$ (modelled by a blackbody source with radius $0.2R_{\odot}$ and temperature 1000 K). The disc (of mass $0.1\,M_{\odot}$) has a dust to gas ratio of 0.01, i.e metallicity equal to solar. The wavelength and resolution parameters were set to be representative of ALMA: the wavelength range is $[0.01, 0.1]$ cm, with an effective resolution of 0.01".

Disc asymmetry plays an important role in the imaging of this system (Figures 6.13 and 6.14). In Figure 6.13, the disc displays a non-zero ellipticity due to its spiral structure, with shadowing indicating the increased surface density of the disc corresponding to the spirals themselves. Again, however, these shadows are only apparent if the sensitivity of the telescope is 1 μJy/beam, well beyond the ALMA's current continuum sensitivity estimates. This ellipticity is enhanced during the encounter (Figure 6.14), with the semi major axis of the ellipse aligned with the orbital vector of the primary and secondary. The erasure of the strong spiral structure during the encounter results in a detectable tidal arm with flux around 1 mJy/beam. The stellar emission is dwarfed by the disc at these wavelengths. The inner regions of the disc are

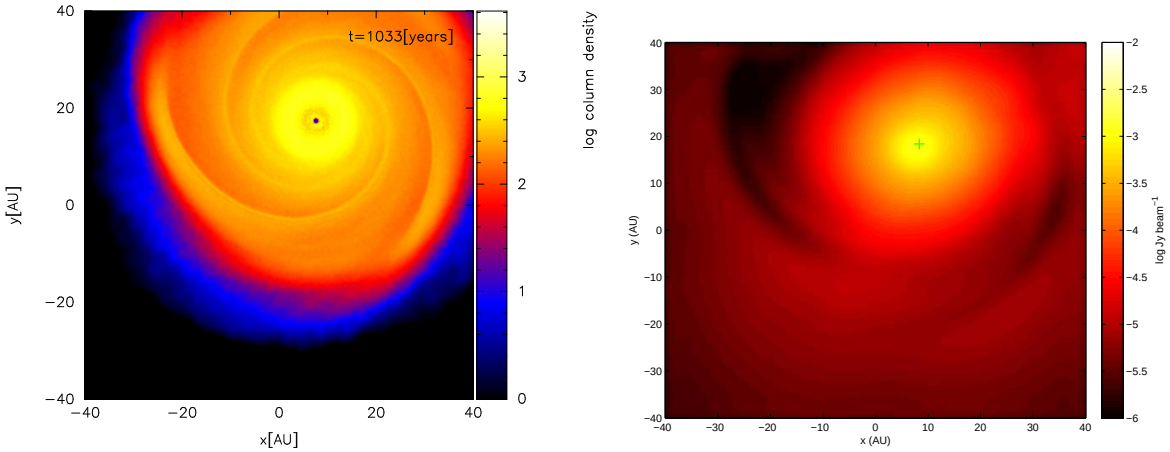

Figure 6.13: Left: Surface density plot of the disc pre-outburst. The secondary is out of frame. Right: image of the simulation, integrated over the wavelength range $[0.01, 0.1]$ cm. The green cross indicates the pixel with maximum intensity

not hot enough for dust sublimation to open a resolvable gap (the intrinsic inner gap is very much below the resolution limit). The resolution of spiral structure in a self-gravitating disc will not be possible with ALMA unless higher sensitivities are achieved, but it *will* be possible to detect enhanced emission from a tidal spiral wave generated as the result of a stellar encounter.

6.14 Discussion

6.14.1 Runtime Scaling

As the SPH systems being imaged by this code will be in general disordered and impossible to render analytically, a true runtime scaling is not feasible. However, an example scaling assuming simple geometry can be calculated.

In general, the runtime T goes as

$$T \sim N_\gamma N_{steps}, \tag{6.93}$$

where N_γ is the number of photons emitted by the code, and N_{steps} represents the computational expense required to track one photon from emission to capture. This can hence be written

$$T \sim N_\gamma < N_{ray} > N_{scatt}, \tag{6.94}$$

where $< N_{ray} >$ is the mean number of particles intersected by any ray, and N_{scatt} indicates how many times a photon will scatter before it exits. Generally, $N_{scatt} \sim < \tau >^2$, and

$$< N_{ray} > \sim N_p \, P_{intersect}, \tag{6.95}$$

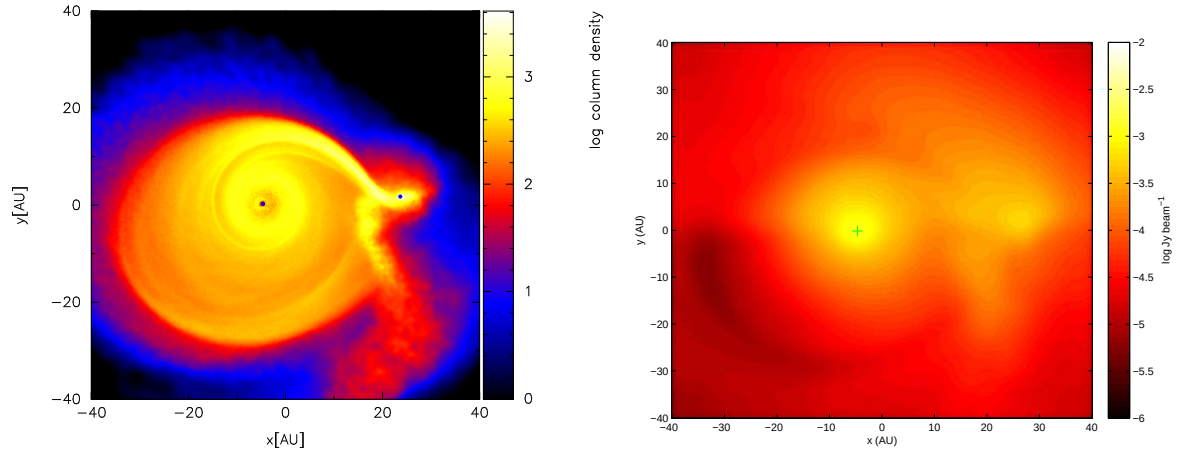

Figure 6.14: Left: Surface density plot of the disc at periastron. Right: image of the simulation, integrated over the wavelength range $[0.01, 0.1]$ cm. The green cross indicates the pixel with maximum intensity

where N_p is the number of SPH particles, and $P_{intersect}$ is the probability that any one particle is intersected by the ray. This can be estimated for simple geometries: assuming the SPH particle distribution is a sphere, then

$$P_{intersect} \sim 1 - e^{-Rn\sigma_s} \tag{6.96}$$

where R is the sphere's radius, $n = N_p/V \sim N_p/R^3$ is the number density of the sphere, and $\sigma_s \sim 4\pi < h >^2$ is the average cross-section of the SPH particles. In a homogeneous sphere,

$$< h > \sim \left(\frac{1}{\rho}\right)^{\frac{1}{3}} \sim \frac{R}{N_p^{\frac{1}{3}}}. \tag{6.97}$$

Combining these results gives

$$T \sim N_\gamma < \tau >^2 N_p \left(1 - e^{-N_p^{\frac{1}{3}}}\right) \tag{6.98}$$

This runtime scaling shares common features with grid-based MCRT codes, in particular the dependence on optical depth and number of particles/cells. For illustration, the simulations used in this work typically took approximately 100 CPU hours to create an image, with $N_p = 500,000$, $N_\gamma = 10^8$ in systems with significant optically thick components.

6.14.2 Resolution

When imaging these simulations, it must be demonstrable that the local density field is sufficiently resolved to allow a satisfactory calculation of the optical depth (more specifically the optical depth to the photosphere, as this is where most of the received emission will originate). In terms of the smoothing length, h, and the mean free path λ_{mfp}, this condition is

$$h < \lambda_{mfp} = \frac{1}{\rho\kappa}. \tag{6.99}$$

As the value of h is inversely proportional to the local number density of SPH particles, the resolution of an MCRT image of an SPH density field is closely linked to the total number of particles in the simulation, an intuitive result. This condition is satisfied in the outer disc easily, as the mean free path is of order the system scale height, (which is resolved by several smoothing lengths). However, within the inner 5-10 AU of the disc, two separate issues arise:

1. The smoothing length exceeds the local mean free path. This is partially compensated by SPH's smoothing prescription - the density field is defined continuously across all space, and therefore affords a kind of sub-h resolution. However, this smoothing erases information about fluctuations in the density field below this length scale, and this will lead to an overestimation of the escaping flux.

2. The disc is not well resolved several scale heights above the midplane, where the photosphere is located. This under-resolving will result in a decrease in scattering events at the photosphere, affecting the scattered light component of the flux.

These two issues show that inside $\sim$ 10 AU, the flux escaping from the disc will be subject to resolution-based error, as well as the error associated with artificial viscosity dominating in the inner regions (Clarke, 2009; Lodato & Price, 2010; Forgan et al., 2010). Users concerned with the inner regions of these discs could adopt a particle splitting prescription to boost the resolution (e.g. Kitsionas & Whitworth 2002). However, as this work is concerned mainly with millimetre emission from the cooler, well-resolved outer disc, this emission will be adequately resolved, and the issues described above will not play a significant role.

6.14.3 Radiative Equilibrium: A Proof of Concept

The work described here has directly utilised the temperature structures produced in the SPH simulations, but such outputs are not a requirement for imaging them using MCRT. The temperature for each fluid element can be calculated with MCRT using radiative equilibrium methods, as was described in earlier sections. In essence, these methods involve absorbed photons incrementing the local temperature field at the absorption point $T(\mathbf{r}_j)$ by an amount $\Delta T(\mathbf{r}_j)$, and then being re-emitted. In grid-based methods, the procedure of increasing the local temperature is simple - the increment ΔT is added to the grid cell the photon is absorbed in. In SPH fields, the situation is somewhat more complex. A scalar field X at a location $\mathbf{r}_j$ is defined as:

$$X(\mathbf{r}_j) = \sum_i \frac{X_i m_i}{\rho_i} W(\mathbf{r}_i - \mathbf{r}_j, h_i) \tag{6.100}$$

(Where we have used the "scatter" interpretation). We can substitute for the temperature increase ΔT:

$$\Delta T(\mathbf{r}_j) = \sum_i \frac{\Delta T_{ij} m_i}{\rho_i} W(\mathbf{r}_i - \mathbf{r}_j, h_i) \tag{6.101}$$

The left hand side of equation (6.101) is known, and can be calculated using the traditional radiative equilibrium methods. The desired quantities are the ΔT_{ij}, that is the individual increases in temperature each SPH particle must be assigned. Equation (6.101) therefore acts as a constraint on the ΔT_{ij}, but is insufficient to close the system. A method of closure is presented here as proof that the radiative equilibrium framework can indeed be incorporated - its application is left for future work. Closure can be achieved by specifying the following *ansatz*:

$$\Delta T_{ij} = A_{ij} \Delta T g(r_{ij}) \tag{6.102}$$

Where A_{ij} is some real number, r_{ij} is the separation of particle i from the location j and $g(r_{ij})$ is some function that satisfies:

$$g(0) = 1, \tag{6.103}$$

$$g(r_{ij} > 2h_i) = 0, \tag{6.104}$$

i.e., particles with smoothing volumes that do not intersect the location j do not contribute. This will satisfy equation (6.101) provided

$$\sum_i \frac{m_i}{\rho_i} A_{ij} g(r_{ij}) W(r_{ij}, h_i) = 1 \tag{6.105}$$

This has recast the problem from solving for ΔT_{ij} to the prefactors A_{ij} and function $g(r_{ij})$. This may appear to be a sideways step: however, giving a specific example is illuminating. To obtain a suitable function $g(r_{ij})$, consider the decay of flux F_0 along a path of optical depth τ:

$$F = F_0 e^{-\tau} \tag{6.106}$$

Using the Stefan-Boltzmann law, $F = \sigma T^4$ yields

$$\sigma(\Delta T_{ij})^4 = \sigma(\Delta T)^4 e^{-\tau_{ij}} \tag{6.107}$$

Inspecting the above equation suggests the following form for g:

$$g(r_{ij}) = e^{-\tau_{ij}/4} \tag{6.108}$$

Assuming that $A_{ij} = A$ for all (i, j), the normalisation condition becomes

$$A = \frac{1}{\sum_i \frac{m_i}{\rho_i} e^{-\tau_{ij}/4} W(r_{ij}, h_i)} \tag{6.109}$$

The system has now been closed with an *ansatz* that is physically motivated, and reproduces expected behaviour (particles with a reduced optical depth to the location will receive a greater temperature boost than others in more optically thick regions).

Other radiative equilibrium methods can also be recast in SPH form, such as that of Lucy (1999), which computes grid cell temperatures by counting the path lengths of any photons that pass through it during one iteration, retrieving the temperature structure rapidly within a few iterations. By replacing grid cells with particle smoothing volumes, it can be seen that this method is indeed amenable to SPH fields also.

6.15 Conclusions

I have outlined a means for applying Monte Carlo Radiative Transfer (MCRT) techniques directly to Smoothed Particle Hydrodynamics (SPH) density fields. In doing so, it gives a means for synthetic telescope images to be made, allowing the flexibility of SPH simulations to be retained in calculating optical depths, scattering and polarisation. Non-trivial features in the imaging of these simulations are well-traced, and the resulting images give new opportunities for theoretical input to observed astrophysical phenomena. In particular, it can show the resolutions and sensitivities required to observe these features. For example, ALMA will be able to resolve spiral structure (or the observational features it produces) if its predicted sensitivities can be improved by at least a factor of 100. In the case of tidal arms produced as a result of stellar encounters, the odds of detection improve significantly.

The inherently graphical nature of the MCRT method allows the code to be greatly optimised, whether by standard parallelisation methods or by the use of Graphical Processing Units (GPUs) to handle the ray intersections. While used for imaging in this work, the method is not restricted to this alone. As the algorithm only modifies how the optical depth is calculated (in order to work in SPH fields), it can be applied to any circumstances traditional gridded MCRT has been applied to. This includes radiative equilibrium simulations where the temperature structure is calculated from stellar emission, or moving beyond continuum emission to perform line radiative transfer calculations. In effect this algorithm places SPH on a par with grids or meshes for MCRT techniques.

Chapter 7

Conclusions

A good traveller has no fixed plans, and is not intent upon arriving.
Lao Tzu, *Tao Te Ching*

7.1 The Content of the Thesis

In this thesis, I have discussed my research into the dynamics and evolution of self-gravitating protostellar discs. This research has been conducted using Smoothed Particle Hydrodynamics (SPH) simulations. As the local thermodynamics and radiative physics is key to the evolution of these discs, I have developed two radiative transfer algorithms for SPH to improve its realism and predictive power respectively.

The first algorithm is a hybrid method which allows frequency-averaged radiative transfer to be modelled during the runtime of an SPH simulation with minimal loss of computational speed (around 6%). I refer to it as a hybrid method as it fuses two pre-existing radiative algorithms - polytropic cooling (Stamatellos et al., 2007b) and flux-limited diffusion (e.g. Whitehouse & Bate 2004) - to capture the individual strengths of both. Its flexibility and simplicity lends it to the study of many astrophysical phenomena beyond circumstellar discs. It is currently being incorporated into the SEREN[1] SPH code, in development at Cardiff University and the University of Sheffield. SPH with the hybrid method has allowed me to study two important problems in self-gravitating disc evolution.

[1] http://dhubber.staff.shef.ac.uk/seren/seren.html

I have investigated the nature of angular momentum transport in protostellar discs by self-gravity, which occurs as a result of the gravito-turbulence generated by gravitational instability. In particular, I investigated whether its phenomenology is well represented by an effective turbulent viscosity (the so-called "pseudo-viscous" approximation). I show that this approximation is valid provided that the disc-to-star mass ratio is less than 0.5, and the disc aspect ratio (the ratio of disc scale height to disc radius) is less than 0.1. I further show that beyond these limits, the approximation fails in the face of significant non-local transport by global spiral waves. Perhaps most importantly, I show that massive compact discs can remain stable over significantly long timescales (around 10% of the embedded phase of protostellar evolution).

Secondly, I have considered self-gravitating discs under perturbation from a stellar companion. I have shown that these events will not in general trigger disc fragmentation (provided the discs are relatively compact, i.e. their outer radii is around 50 AU or less). Instead, the local heating induced by the companion stabilises the disc against fragmentation, and the stripping of mass from the outer regions by the secondary results in a steeper surface density profile, which also inhibits fragmentation. As a result of the secondary's accretion, its orbital parameters are modified, raising the possibility of binary capture by this mechanism.

Disc-penetrating encounters result in several bursts of accretion (one each for the disc, primary and secondary). This accretion burst is several orders of magnitude above steady state accretion rates; subsequently the luminosity of the system also increases by several orders of magnitude. This is consistent with outburst phenomena such as FU Orionis (FU Ori) objects, thought to be an important component of pre-main sequence evolution. I show that these encounter-driven outbursts do not share the same characteristics as FU Ori, and therefore constitute a different class of object which will be difficult to see.

The second radiative transfer algorithm I have developed aims to increase the effectiveness of SPH in making observational predictions. It uses Monte Carlo Radiative Transfer (MCRT) techniques to simulate in detail the frequency-dependent absorption and scattering of photons in a medium. Traditionally, MCRT techniques require the SPH simulation to be smoothed onto a grid before use. I have demonstrated a means by which grids are no longer required, and the native resolution of SPH simulations can therefore be retained during the imaging process. It utilises algorithms designed for computer graphics to decrease its runtime and memory load, and has been shown to be effective at probing the level of Poisson noise in an SPH system during raytracing tests.

Astrophysical systems modelled in SPH can therefore be "observed" virtually at the behest of the user, using a simulacrum of whatever telescope they wish. As astronomy is a multi-instrument, multi-wavelength science, this flexibility will prove to be crucial (especially with a future generation of telescopes coming online in the next decade or so). I have used the code to produce images of the encounter simulations described above using a virtual ALMA-type instrument, which has proved instructive in terms of what observational features ALMA will be able to resolve. In particular, I have shown that while encounter-specific phenomena such

as tidal arms will be detectable in objects at distances of ~ 140 pc with ALMA, the typical spiral waves produced by self-gravitating discs in isolation will be difficult to detect without an improvement in sensitivity.

7.2 Implications of the Research Conducted

There are some broad trends which emerge from the work carried out in this thesis, which I will now briefly summarise.

Firstly, the fragmentation of self-gravitating discs into bound objects is in general difficult to achieve inside radii of around $50 - 70$ AU (for typical disc parameters). Having run a significant number of disc simulations with outer radii of around 50 AU (both in isolation and under external perturbation), almost all have failed to fragment. In Chapter 4 I show discs with masses greater than the star mass undergoing stable, quasi-steady evolution without fragmentation. The few discs that do fragment (which are not presented in this thesis) are most likely the result of the initial conditions being incorrectly set up (as the fragmentation occurs very rapidly). These findings are in accord with the current consensus in self-gravitating disc evolution. The semi-analytic models of Rafikov (2005), Rice & Armitage (2009), Clarke (2009) and Rice et al. (2010) forbid fragmentation in the inner disc, as do both grid-based and particle-based numerical simulations, such as those of Boley et al. (2006), Stamatellos et al. (2007a), Lodato et al. (2007), and many others. There is a caveat regarding the accretion of matter into the disc from the surrounding envelope, which could provide the necessary stresses to fragment the disc. However, relatively large accretion rates are required, and again this is only expected to occur in the outer regions of the disc (Kratter et al., 2010; Vorobyov & Basu, 2010).

With fragmentation restricted to the outer regions of protostellar discs, and the expected fragment masses to be relatively large (near the deuterium burning limit which delineates planets from brown dwarfs), this constrains objects formed by disc fragmentation to giant planets and brown dwarfs (Stamatellos & Whitworth, 2009). Terrestrial planets are almost certainly not formed from disc fragments, but rather cores grown by coagulation in a manner resembling the core accretion model (Boley, 2009). Equally, protostellar discs in the self-gravitating phase must leave their imprints on the planetary system that forms, either by the enhancement of grain growth in spiral arms (Rice et al., 2004), for which chondrules might be forensic evidence (Boley & Durisen, 2008), or by orbital migration through competing disc torques. Both these phenomena depend sensitively on the local chemo-thermodynamics (Clarke & Lodato, 2009; Paardekooper & Papaloizou, 2008), hence the importance of studies of self-gravitating disc dynamics such as those carried out in this thesis.

The self-gravitating phase is also a critical phase in pre-main sequence stellar evolution. It is vital to understanding the growth of stars and their arrival on the Main Sequence, as the disc itself appears to form before the star (Machida & Matsumoto, 2010; Bate, 2010).

The subsequent accretion of matter from the disc onto the protostar will define its end state. These factors are all influenced by the angular momentum transport induced by self-gravity, which I have shown evolves the disc towards a quasi-steady state and uniform accretion rate, provided that the disc-to-star mass ratio is low. If the disc-to-star mass ratio is high, then high accretion variability and strong global spiral waves are manifest. However, compact discs can still evolve towards the quasi-steady state by rapid stellar accretion and viscous spreading, preventing catastrophic instability. The self-gravitating phase is therefore longer than has been assumed in the past, and plays a much more important role than was previously considered in the formation of both stars and planets. In this respect, the work of this thesis adds strength to the current theoretical foundations underpinning both types of object.

On a wider note, the development of a synthetic imaging system which uses the native resolution of an SPH simulation has important implications for those attempting to confirm the existence of self-gravitating discs. By giving as accurate a prediction as possible regarding which observational features to look for at which wavelength, observers can tailor their campaigns to match these predictions. The algorithm's use extends much further than disc phenomena, being applicable to any SPH simulation. While dust scattering is modelled in this thesis, the code can be easily modified to follow electron scattering (for example), which would give users the opportunity to image phenomena emitting in the continuum at higher temperatures (such as active galaxies).

7.3 Limitations and Opportunities for Future Research

Numerical simulation suffers from fundamental limitations which are often only surpassed with improved computing resources. However, there are some options for future research that are not at the mercy of Moore's Law of technological advancement, and may have important consequences for the theories discussed in the above sections. I will discuss some of these below.

The hybrid method, while being effective at many radiative transfer problems, does not yet fully incorporate radiative feedback (adopting instead a non-uniform background temperature as a surrogate). Being able to simulate non-axisymmetric radiative pressure from the central star will be important for studies of disc stability (e.g. Cai et al. 2008) as well as the star formation process itself (Bate, 2010). Non-axisymmetric radiative feedback requires the code to calculate optical depths along arbitrary lines of sight. The raytracing framework developed for the MCRT code of Chapter 6 could be incorporated into the hybrid method to perform this function.

If we wish to model grain growth in spiral arms as discussed above, then further modifications to the code are required to model the dust. An obvious first goal would be the idealised, non-self-gravitating, super-particle model adopted by Rice et al. (2004), which models the drag of

the gas upon the dust. This would allow the testing of Clarke & Lodato (2009)'s predictions regarding the disc radii at which dust aggregation would be most efficient. In the long term, a full two-fluid model, which incorporates mixing forces and the back-reaction of the dust upon the gas (such as that of Barrière-Fouchet et al. 2005) will be necessary to model the disc evolution correctly, and to ascertain how likely gravitational instability and fragmentation can be enhanced by the interplay of these forces.

Despite such improvements to the code, resolution remains an unavoidable issue with numerical simulations. For example, a comprehensive test of accretion variability and FU Orionis outbursts requires the artificial viscosity (AV) in the inner regions of discs to be significantly lower than the effective viscosity produced by the gravitational instability. For example, in the simulations described in Chapter 4, AV dominates at radii less than 10-20 AU. To correctly resolve the inner regions, AV must be reduced much further. To reduce AV until it dominates only within the inner 1 AU using the SPH code used in this thesis, the particle number must be increased from half a million to 5 quadrillion! This is clearly unfeasible, but there are alternative routes to reducing AV, such as that of Cartwright & Stamatellos (2010). By combining these improvements with a prescription for MRI turbulence, it may be possible to achieve a full 3D simulation of the "instability cascade" model for FU Orionis outbursts (e.g. Armitage et al. 2001; Zhu et al. 2009a), where self-gravity "piles up" material in the inner regions, activates MRI and rapidly drains the inner disc onto the central star. Equally, it may be possible to resolve an instability cascade where thermal instability is activated instead (Lin et al., 1985; Bell & Lin, 1994).

As regards the synthetic imaging codes, there remains some important pieces of code development to be done. The code can currently only handle continuum radiation, which while important, limits the utility of the code somewhat. As the code uses the same principles and algorithms as any gridded MCRT code, it is certainly possible that line radiative transfer could be implemented. Also, the code is well suited for massive parallelisation and the use of Graphical Processing Units (GPUs) to vastly speed up its computation. This would allow the number of photons used in a single image to increase dramatically, adding either improved spatial or spectral resolution depending on the user's requirements.

While these code developments are important, this does not preclude the use of the codes in their current state to investigate further science questions. For example, it is clear that the role of the envelope in the evolution of self-gravitating discs is important, and may affect whether the stable evolution towards the quasi-steady state seen in Chapter 4 is preserved when the disc is able to replenish itself by accreting from its surrounding environment.

Further to this, studying the relationship between the angular momentum of the progenitor molecular cloud and the resulting disc structure it forms will help to indicate how common the disc fragmentation process is. The early results of simulations I have conducted (which I intend to continue in future research) suggest that discs grown from the collapse of a homogeneous, uniformly rotating cloud can remain stable with high disc-to-star mass ratios for *even longer*

than in the isolated cases I have previously described, and that a critical angular momentum for fragmentation does exist, and is relatively insensitive to cloud mass. However, these results are preliminary, and must be investigated in more detail. I also intend to expand this research to large scale simulations of star formation in turbulent clouds, to look at the role of disc stresses induced by external gravitational forces in the fragmentation process.

7.4 Closing Remarks

It is an exciting time for circumstellar disc researchers. The advent of instruments such as ALMA and *Herschel* will bring new observational constraints to current star formation theory, while the *Kepler* space telescope is soon to expand the crop of known exoplanets and test the current models of planet formation. Future instruments such as the *James Webb Space Telescope* (JWST) and the *European Extremely Large Telescope* (E-ELT) will provide a step-change in both the quality and quantity of observational data. Even missions such as WFIRST, designed for weak gravitational lensing on cosmic scales, will add to the knowledge of exoplanetary systems by doubling as a microlensing instrument.

On the theoretical front, a coherent picture is emerging of the earliest stages of star formation. This picture is being continually strengthened by the latest numerical simulations and semi-analytic models as more physics is successfully incorporated. Planet formation theory, thought to be stymied by the difficulties surrounding core accretion, has been re-energised by discussions of disc instability and grain growth working in tandem to overstep the problems of aggregating intermediate size objects. To properly understand the influence of these processes, sophisticated modelling and simulations are required to elucidate the complex interactions between gas and dust in a radiative, gravito-hydrodynamic setting.

In short, circumstellar disc theory is part of a rapidly progressing front of scientific knowledge, and is making important contributions to the wider science of star and planet formation, with which it is intimately linked. The work in this thesis underlines the importance of self-gravitating disc theory in dictating the evolution of both our Solar System and star systems throughout the Universe.

References

Acreman D. M., Harries T. J., Rundle D. A., 2010, MNRAS, 403, 1143

Alexander R. D., Clarke C. J., Pringle J. E., 2006a, MNRAS, 369, 216

—, 2006b, MNRAS, 369, 229

Altay G., Croft R. A. C., Pelupessy I., 2008, MNRAS, 386, 1931

Anderson J., Laing P., Lau E., Liu A., Nieto M., Turyshev S., 1998, Physical Review Letters, 81, 2858

Andre P., Ward-Thompson D., Barsony M., 2000, in Protostars and Planets IV, Mannings, ed., University of Arizona Press

Annis J., 1999, J. Br. Interplanet. Soc., 52, 19

Armitage P. J., 2007, ApJ, 665, 1381

Armitage P. J., Clarke C. J., Tout C. A., 1999, Monthly Notices of the Royal Astronomical Society, 304, 425

Armitage P. J., Livio M., Pringle J. E., 2001, MNRAS, 324, 705

Artymowicz P., Lubow S. H., 1994, ApJ, 421, 651

Balbus S. A., Hawley J. F., 1991, ApJ, 376, 214

Balbus S. A., Papaloizou J., 1999, ApJ, 521, 650

Balsara D. S., 1989, PhD thesis, University of Illinois

Barnes J. E., Hut P., 1986, Nature, 324, 446

—, 1989, ApJs, 70, 389

Barrière-Fouchet L., Gonzalez J.-F., Murray J. R., Humble R. J., Maddison S. T., 2005, A&A, 443, 185

Bate M. R., 2009, MNRAS, 392, 1363

—, 2010, MNRAS, 404, L79

Bate M. R., Bonnell I. A., Bromm V., 2003, MNRAS, 339, 577

Bate M. R., Bonnell I. A., Price N. M., 1995, MNRAS, 277, 362

Bate M. R., Burkert A., 1997, MNRAS, 288, 1060

Batygin K., Laughlin G., 2008, ApJ, 683, 1207

Bell K. R., Lin D. N. C., 1994, ApJ, 427, 987

Bernes C., 1979, A&A, 73, 67

Bertin G., 2000, Dynamics of Galaxies. Cambridge University Press

Binney J., Tremaine S., 1987, Galactic Dynamics, Binney J., Tremaine S., eds. Princeton University Press

Bisbas T. G., Wünsch R., Whitworth A. P., Hubber D. A., 2009, A&A, 497, 649

Bjorkman J. E., Wood K., 2001, ApJ, 554, 615

Bodenheimer P., Yorke H. W., Rozyczka M., Tohline J. E., 1990, ApJ, 355, 651

Boffin H. M. J., Watkins S. J., Bhattal A. S., Francis N., Whitworth A. P., 1998, MNRAS, 300, 1189

Boley A. C., 2009, ApJ, 695, L53

Boley A. C., Durisen R., 2008, ApJ, 685, 1193

Boley A. C., Durisen R., Nordlund A., Lord J., 2007a, ApJ, 665, 1254

Boley A. C., Durisen R. H., 2010, ApJ, submitted

Boley A. C., Hartquist T. W., Durisen R. H., Michael S., 2007b, ApJ, 656, L89

Boley A. C., Mejia A. C., Durisen R., Cai K., Pickett M. K., D'Alessio P., 2006, ApJ, 651, 517

Bonnell I. A., Bastien P., 1992, ApJ, 401, L31

Bonnell I. A., Bate M. R., 1994, MNRAS, 271, 999

Borucki W. J., 2010, ArXiv eprint:1006.2799

Boss A. P., 1997, Science, 276, 1836

—, 2006, ApJ, 643, 501

Boss A. P., Black D. C., 1982, ApJ, 258, 270

Bounama C., von Bloh W., Franck S., 2007, Astrobiology, 7, 745

Bozhilov V., Forgan D. H., 2010, International Journal of Astrobiology, 9, 175

Brin G. D., 1983, QJRAS, 24, 283

Cai K., Durisen R., Boley A. C., Pickett M. K., Mejia A. C., 2008, ApJ, 673, 1138

Cai K., Durisen R. H., Michael S., Boley A. C., Mejía A. C., Pickett M. K., D'Alessio P., 2006, ApJ, 636, L149

Cameron A. G. W., 1978, The Moon and the Planets, 18, 5

Canfield D. E., 2005, Annual Review of Earth and Planetary Sciences, 33, 1

Carr M. H., Belton M. J., Chapman C. R., Davies M. E., Geissler P., Greenberg R., McEwen A. S., Tufts B. R., Greeley R., Sullivan R., Head J. W., Pappalardo R. T., Klaasen K. P., Johnson T. V., Kaufman J., Senske D., Moore J., Neukum G., Schubert G., Burns J. A., Thomas P., Veverka J., 1998, Nature, 391, 363

Carter B., 2008, International Journal of Astrobiology, 7, 177

Cartwright A., Stamatellos D., 2010, A&A, 516, A99

Castor J. I., 2004, Radiation Hydrodynamics. Cambridge University Press

Cavicchioli R., 2002, Astrobiology, 2, 281

Cha S.-H., Whitworth A. P., 2003, MNRAS, 340, 73

Chandrasekhar S., 1960, Radiative Transfer. New York: Dover

Cirkovic M. M., 2008, J. Br. Interplanet. Soc., 61, 246

—, 2009, Serbian Astronomical Journal, 178, 1

Clarke C. J., 2009, MNRAS, 396, 1066

Clarke C. J., Gendrin A., Sotomayor M., 2001, MNRAS, 328, 485

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., 1991, MNRAS, 249, 584

—, 1993, Royal Astronomical Society, 261, 190

Clarke C. J., Syer D., 1996, MNRAS, 278, L23

Cleary P. W., Monaghan J. J., 1999, Journal of Computational Physics, 148, 227

Committee For A Decadal Survey Of Astronomy And Astrophysics, 2010, New Worlds, New Horizons in Astronomy and Astrophysics. The National Academies Press, Washington, D.C.

Cossins P., Lodato G., Clarke C. J., 2009, MNRAS, 393, 1157

—, 2010, MNRAS, 401, 2587

Courant R., Friedrichs K., Lewy H., 1928, Mathematische Annalen, 100, 32

Cresswell P., Nelson R. P., 2006, A&A, 450, 833

Crick F., 1973, Icarus, 19, 341

Dale J. E., Ercolano B., Clarke C. J., 2007, MNRAS, 382, 1759

D'Angelo G., Durisen R. H., Lissauer J. J., 2010, Giant Planet Formation, Seager S., ed. University of Arizona Press

Davies P., 2010, The Eerie Silence: Are We Alone in the Universe? Allen Lane

Dawkins R., 1990, The Selfish Gene. Oxford University Press

Diaz B., Schulze-Makuch D., 2006, Astrobiology, 6, 332

Dolence J. C., Gammie C. F., Mościbrodzka M., Leung P. K., 2009, ApJs, 184, 387

Dullemond C. P., Hollenbach D., Kamp I., D'Alessio P., 2007, in Protostars and Planets V, Reipurth B., Jewitt D., Keil K., eds., University of Arizona Press

Duquennoy A., Mayor M., 1991, A&A, 248, 485

Durisen R., Boss A. P., Mayer L., Nelson A. F., Quinn T., Rice W. K. M., 2007, in Protostars and Planets V, Reipurth B., Jewitt D., Keil K., eds., University of Arizona Press

Dyson F. J., 1960, Science, 131, 1667

Eisemann M., Grosch T., Müller S., Magnor M., 2007, Journal of Graphics, GPU, and Game Tools, 12, 35

Ewell M., 1988, PhD thesis, Princeton Univ.

Field G. B., 1965, ApJ, 142, 531

Ford E. B., Rasio F. A., 2008, ApJ, 686, 621

Forgan D. H., 2009, International Journal of Astrobiology, 8, 121

—, 2010, Journal of Cosmology, 5, 811

Forgan D. H., Nichol R. C., 2010, International Journal of Astrobiology, in press

Forgan D. H., Rice K., 2009, MNRAS, 400, 2022

—, 2010a, MNRAS, 406, 2549

—, 2010b, International Journal of Astrobiology, 9, 73

—, 2010c, MNRAS, 402, 1349

Forgan D. H., Rice K., Cossins P., Lodato G., 2010, MNRAS, in press

Forgan D. H., Rice K., Stamatellos D., Whitworth A. P., 2009, MNRAS, 394, 882

Formisano V., Atreya S., Encrenaz T., Ignatiev N., Giuranna M., 2004, Science, 306, 1758

Fu S., 2007, Arxiv e-prints 0712.2108

Furlan E., McClure M., Calvet N., Hartmann L., DAlessio P., Forrest W. J., Watson D. M., Uchida K. I., Sargent B., Green J. D., Herter T. L., 2008, ApJs, 176, 184

Gammie C., 2001, ApJ, 553, 174

Gingold R., Monaghan J., 1978, MNRAS, 184, 481

—, 1982, Journal of Computational Physics, 46, 429

Goldreich P., Sari R., 2003, ApJ, 585, 1024

Goodman A. A., Benson P. J., Fuller G. A., Myers P. C., 1993, ApJ, 406, 528

Greaves J., Richards A., Rice W. K. M., Muxlow T., 2008, MNRAS, 391, L74

Hall S. M., Clarke C. J., Pringle J. E., 1996, MNRAS, 278, 303

Hammersley J. M., Handscomb D. C., 1964, Monte Carlo Methods. Taylor & Francis

Hart M. H., 1979, Icarus, 37, 351

Hartmann L., Kenyon S. J., 1996, ARA&A, 34, 207

Helled R., Podolak M., Kovetz A., 2006, Icarus, 185, 64

Herbig G. H., 1977, ApJ, 217, 693

—, 2007, The Astronomical Journal, 133, 2679

Hernquist L., Katz N., 1989, ApJs, 70, 419

Hollenbach D., Adams F. C., 2004, Debris Disks and the Formation of Planets: A Symposium in Memory of Fred Gillett, 324

Horner J., Jones B. W., 2008a, International Journal of Astrobiology, 7, 251

—, 2008b, International Journal of Astrobiology, 8, 75

Horner J., Jones B. W., Chambers J., 2009, International Journal of Astrobiology, 9, 1

House L., 1969, Journal of Quantitative Spectroscopy and Radiative Transfer, 9, 1579

Ida S., Lin D. N. C., 2004, ApJ, 604, 388

—, 2008, ApJ, 673, 487

Jaakkola S., El-Showk S., Annila A., 2008, eprint arXiv:0807.0892

Jaakkola S., Sharma V., Annila A., 2009, eprint arXiv:0906.0254

Jackson J. D., 1975, Classical Electrodynamics, 2nd edn. New York, Wiley

Jeans J. H., 1928, Astronomy and Cosmogony. Cambridge University Press

Johnson B. M., Gammie C. F., 2003, ApJ, 597, 131

Kaila V. R., Annila A., 2008, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 464, 3055

Kaltenegger L., Selsis F., 2010, EAS Publications Series, 41, 485

Kardashev N. S., 1964, Soviet Astronomy, 8, 217

Kasting J. F., 1993, Icarus, 101, 108

Kenyon S. J., Calvet N., Hartmann L., 1993, ApJ, 414, 676

Kenyon S. J., Hartmann L., 1987, ApJ, 323, 714

Kenyon S. J., Hartmann L., Hewett R., 1988, ApJ, 325, 231

Kenyon S. J., Hartmann L., Strom K. M., Strom S. E., 1990, The Astronomical Journal, 99, 869

Kessel-Deynet O., Burkert A., 2000, MNRAS, 315, 713

Kipping D. M., Fossey S. J., Campanella G., 2009, MNRAS, 400, 398

Kitsionas S., Whitworth A. P., 2002, MNRAS, 330, 129

Kornet K., Bodenheimer P., Róyczka M., Stepinski T. F., 2005, A&A, 430, 1133

Krasnopolsky V., Maillard J., Owen T., 2004, Icarus, 172, 537

Kratter K. M., Matzner C. D., Krumholz M. R., 2008, ApJ, 681, 375

Kratter K. M., Murray-Clay R. A., Youdin A. N., 2010, ApJ, 710, 1375

Kuiper G. P., 1951, Proceedings of the National Academy of Sciences, 37, 1

Kurosawa R., Harries T. J., Bate M. R., Symington N. H., 2004, MNRAS, 351, 1134

Lada C. J., Lada E. A., 2003, ARA&A, 41, 57

Landau L. D., Lifshitz E. M., 1959, Fluid Mechanics. Oxford Pergamon Press

Larson R., 1968, PhD thesis, California Institute of Technology

Laughlin G., Bodenheimer P., Adams F. C., 2004, ApJ, 612, L73

Learned J. G., Pakvasa S., Simmons W. A., Tata X., 1994, QJRAS, 35, 321

Levermore C. D., Pomraning G. C., 1981, ApJ, 248, 321

Lin D. N. C., 1998, Science, 281, 2025

Lin D. N. C., Faulkner J., Papaloizou J., 1985, MNRAS, 212, 105

Lin D. N. C., Papaloizou J., 1986, ApJ, 309, 846

Lin D. N. C., Pringle J. E., 1990, ApJ, 358, 515

Lineweaver C. H., 2001, Icarus, 151, 307

Lineweaver C. H., Fenner Y., Gibson B. K., 2004, Science, 303, 59

Lodato G., 2007, Rivista Del Nuovo Cimento, 30, 293

Lodato G., Clarke C. J., 2004, MNRAS, 353, 841

Lodato G., Meru F., Clarke C. J., Rice W. K. M., 2007, MNRAS, 374, 590

Lodato G., Price D. J., 2010, MNRAS, 405, 1212

Lodato G., Rice W. K. M., 2004, MNRAS, 351, 630

—, 2005, MNRAS, 358, 1489

Loeb A., Zaldarriaga M., 2007, Journal of Cosmology and Astroparticle Physics, 2007, 020

Lucy L., 1999, A&A, 344, 282

Lucy L. B., 1977, The Astronomical Journal, 82, 1013

Lynden-Bell D., Kalnajs A. J., 1972, MNRAS, 157, 1

Lyra W., Johansen A., Zsom A., Klahr H., Piskunov N., 2009, A&A, 497, 869

Maccone C., 2009, Acta Astronautica, in press

Machida M. N., Matsumoto T., 2010, MNRAS, submitted

Mahovsky J., Wyvill B., 2004, Journal of Graphics Tools, 9, 35

Mamatsashvili G. R., Rice W. K. M., 2009, MNRAS, 394, 2153

Marois C., Macintosh B., Barman T., Zuckerman B., Song I., Patience J., Lafrenière D., Doyon R., 2008, Science, 322, 1348

Masset F. S., Papaloizou J. C. B., 2003, ApJ, 588, 494

Masunaga H., Inutsuka S., 2000, ApJ, 531, 350

Masunaga H., Miyama S. M., Inutsuka S., 1998, ApJ, 495, 346

Mayer L., Lufkin G., Quinn T., Wadsley J., 2007, ApJ, 661, L77

Mayor M., Queloz D., 1995, Nature, 378, 355

McKee C. F., Ostriker E. C., 2007, ARA&A, 45, 565

Mejia A. C., Durisen R., Pickett M. K., Cai K., 2005, ApJ, 619, 1098

Miller G. E., Scalo J. M., 1979, ApJs, 41, 513

Monaghan J., 1989, Journal of Computational Physics, 82, 1

Monaghan J., Lattanzio J., 1985, A&A, 149, 135

Monaghan J. J., 1982, SIAM Journal on Scientific and Statistical Computing, 3, 422

—, 1992, ARA&A, 30, 543

—, 2005, Reports on Progress in Physics, 68, 1703

Morgan T., Bondi H., 1970, Proceedings of the Royal Society of London. Series A, 320, 277

Morris S. C., 2006, Current Biology, 16, R826

Moutou C., Hébrard G., Bouchy F., Eggenberger A., Boisse I., Bonfils X., Gravallon D., Ehrenreich D., Forveille T., Delfosse X., Desort M., Lagrange A.-M., Lovis C., Mayor M., Pepe F., Perrier C., Pont F., Queloz D., Santos N. C., Ségransan D., Udry S., Vidal-Madjar A., 2009, A&A, 498, L5

Murray J. R., 1996, MNRAS, 279, 402

Naef D., Latham D. W., Mayor M., Mazeh T., Beuzit J. L., Drukier G. A., Perrier-Bellet C., Queloz D., Sivan J. P., Torres G., Udry S., Zucker S., 2001, A&A, 375, L27

Nelson A. F., 2006, MNRAS, 373, 1039

Nelson R. P., Papaloizou J. C. B., 1994, MNRAS, 270, 1

—, 2004, Extrasolar Planets: Today and Tomorrow, 321

Nicholson P., Hedman M., Clark R., Showalter M., Cruikshank D., Cuzzi J., Filacchione G., Capaccioni F., Cerroni P., Hansen G., 2008, Icarus, 193, 182

Ollivier M., Chazelas B., Bordé P., 2008, Physica Scripta, T130, 14

Ostlie D. A., Carroll B. W., 1996, An Introduction to Modern Stellar Astrophysics, Ostlie D. A., Carroll B. W., eds. Pearson Education

Oxley S., Woolfson M. M., 2003, MNRAS, 343, 900

Paardekooper S.-J., Papaloizou J. C. B., 2008, A&A, 485, 877

Paczynski B., 1978, Acta Astronomica, 28, 91

Papaloizou J. C. B., Nelson R. P., 2003, MNRAS, 339, 983

Parkinson C. D., Liang M.-C., Hartman H., Hansen C. J., Tinetti G., Meadows V., Kirschvink J. L., Yung Y. L., 2007, A&A, 463, 353

Pawlik A. H., Schaye J., 2008, MNRAS, 389, 651

Penny A. J., 2004, eprint arXiv:astro-ph/0408473

Pfalzner S., 2008, A&A, 492, 735

Pickett B. K., Mejia A. C., Durisen R., Cassen P. M., Berry D. K., Link R. P., 2003, ApJ, 590, 1060

Pinte C., Ménard F., Duchêne G., Bastien P., 2006, A&A, 459, 797

Podolak M., Weizman A., Marley M., 1995, Planetary and Space Science, 43, 1517

Press W. H., Teukolsky S. A., Vetterling W. T., Flannery B. P., 1992, Numerical Recipes in FORTRAN 90. Cambridge University Press

Prialnik D., 2000, An Introduction to the Theory of Stellar Structure and Evolution, Prialnik D., ed. Cambridge University Press

Price D., 2005, PhD thesis, University of Cambridge

Price D. J., 2007, PASA, 24, 159

Pringle J. E., 1981, ARA&A, 19, 137

Queloz D., Anderson D., Collier Cameron A., Gillon M., Hebb L., Hellier C., Maxted P., Pepe F., Pollacco D., Ségransan D., Smalley B., Triaud A. H. M. J., Udry S., West R., 2010, eprint arXiv:1006.5089

Rafikov R., 2005, ApJ, 621, 69

Raup D. M., Sepkoski J. J., 1982, Science, 215, 1501

Raymond S. N., Armitage P. J., Gorelick N., 2009, ApJ, 699, L88

Raymond S. N., Mandell A. M., Sigurdsson S., 2006, Science, 313, 1413

Reynolds O., 1883, Philosophical Transactions of the Royal Society of London, 174, 935

Rice W. K. M., Armitage P. J., 2009, MNRAS, 396, 2228

Rice W. K. M., Armitage P. J., Bate M. R., Bonnell I. A., 2003, MNRAS, 339, 1025

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

—, 2006, MNRAS, 372, L9

Rice W. K. M., Mayo J. H., Armitage P. J., 2010, MNRAS, 402, 1740

Rocha-Pinto H. J., Maciel W. J., Scalo J., Flynn C., 2000a, A&A, 358, 850

—, 2000b, A&A, 358, 869

Rodríguez L. F., Loinard L., D'Alessio P., Wilner D. J., Ho P. T. P., 2005, ApJ, 621, L133

Rolleston W. R. J., Smartt S. J., Dufton P. L., Ryans R. S. I., 2000, A&A, 363, 537

Rosswog S., 2009, New Astronomy Reviews, 53, 78

Sartoretti P., Schneider J., 1999, A&A, 134, 553

Saumon D., Guillot T., 2004, ApJ, 609, 1170

Schröder K.-P., Connon Smith R., 2008, MNRAS, 386, 155

Shakura N. I., Sunyaev R. A., 1973, A&A, 24

Sharma V., Annila A., 2007, Biophysical Chemistry, 127, 123

Shu F., 1991, Physics of Astrophysics, Vol. II: Gas Dynamics. University Science Books, New York

Shu F. H., 1970, ApJ, 160, 99

Silagadze Z. K., 2008, Acta Physica Polonica B, 39, 2943

Smith R., Clark P., Bonnell I. A., 2009, MNRAS, 396, 830

Solomon P. M., Rivolo A. R., Barrett J., Yahil A., 1987, ApJ, 319, 730

Spencer J., Grinspoon D., 2007, Nature, 445, 376

Spiegel D. S., Menou K., Scharf C. A., 2008, ApJ, 681, 1609

Spiegel E. A., 1957, ApJ, 126, 202

Spitzer L., 1942, ApJ, 95, 329

Stamatellos D., Hubber D. A., Whitworth A. P., 2007a, MNRAS, 382, L30

Stamatellos D., Whitworth A. P., 2005, A&A, 439, 153

—, 2008, A&A, 480, 879

—, 2009, MNRAS, 392, 413

Stamatellos D., Whitworth A. P., Bisbas T., Goodwin S., 2007b, A&A, 475, 37

Stewart S. T., Leinhardt Z. M., 2009, ApJ, 691, L133

Stofan E. R., Elachi C., Lunine J. I., Lorenz R. D., Stiles B., Mitchell K. L., Ostro S., Soderblom L., Wood C., Zebker H., Wall S., Janssen M., Kirk R., Lopes R., Paganelli F., Radebaugh J., Wye L., Anderson Y., Allison M., Boehmer R., Callahan P., Encrenaz P., Flamini E., Francescetti G., Gim Y., Hamilton G., Hensley S., Johnson W. T. K., Kelleher K., Muhleman D., Paillou P., Picardi G., Posa F., Roth L., Seu R., Shaffer S., Vetrella S., West R., 2007, Nature, 445, 61

Tarter J., 2004, New Astronomy Reviews, 48, 1543

Thi W.-F., Mathews G., Ménard F., Woitke P., Meeus G., Riviere-Marichalar P., Pinte C., Howard C. D., Roberge A., Sandell G., Pascucci I., Riaz B., Grady C. A., Dent W. R. F., Kamp I., Duchêne G., Augereau J.-C., Pantin E., Vandenbussche B., Tilling I., Williams J. P., Eiroa C., Barrado D., Alacid J. M., Andrews S., Ardila D. R., Aresu G., Brittain S., Ciardi D. R., Danchi W., Fedele D., de Gregorio-Monsalvo I., Heras A., Huelamo N., Krivov A., Lebreton J., Liseau R., Martin-Zaidi C., Mendigutía I., Montesinos B., Mora A., Morales-Calderon M., Nomura H., Phillips N., Podio L., Poelman D. R., Ramsay S., Rice K., Solano E., Walker H., White G. J., Wright G., 2010, A&A, 518, L125

Thies I., Kroupa P., Theis C., 2005, MNRAS, 364, 961

Toomre A., 1964, ApJ, 139, 1217

Tsuji T., 1966, PASJ, 18, 127

Udry S., Bonfils X., Delfosse X., Forveille T., Mayor M., Perrier C., Bouchy F., Lovis C., Pepe F., Queloz D., Bertaux J.-L., 2007, A&A, 469, L43

Udry S., Santos N. C., 2007, ARA&A, 45, 397

Von Neumann J., Richtmyer R. D., 1950, Journal of Applied Physics, 21, 232

Vorobyov E. I., 2010, ApJ, 713, 1059

Vorobyov E. I., Basu S., 2005, ApJ, 633, L137

—, 2006, ApJ, 650, 956

—, 2008, ApJ, 676, L139

—, 2010, ApJ, 714, L133

Vukotic B., Cirkovic M. M., 2007, Serbian Astronomical Journal, 175, 45

—, 2008, Serbian Astronomical Journal, 176, 71

Walters C., 1980, Icarus, 41, 193

Waltham D., 2004, Astrobiology, 4, 460

Ward P., Brownlee D., 2000, Rare Earth : Why Complex Life is Uncommon in the Universe, Ward P., Brownlee D., eds. Springer

Ward-Thompson D., André P., Crutcher R., Johnstone D., Onishi T., Wilson C., 2007, in Protostars and Planets V, Reipurth B., Jewitt D., Keil K., eds., University of Arizona Press

Watkins S. J., Bhattal A. S., Boffin H. M. J., Francis N., Whitworth A. P., 1998a, MNRAS, 300, 1205

—, 1998b, MNRAS, 300, 1214

Weidenschilling S. J., 1977, MNRAS, 180, 57

Whipple F. L., 1973, in Evolutionary and Physical Properties of Meteoroids, IAU Colloquium 13., Hemenway C., Millman P., Cook A., eds., NASA SP

White R. J., Greene T. P., Doppmann G. W., Covey K. R., Hillenbrand L. A., 2007, in Protostars and Planets V, Reipurth B., Jewitt D., Keil K., eds., University of Arizona Press

White R. L., 1979, ApJ, 229, 954

Whitehouse S. C., Bate M. R., 2004, MNRAS, 353, 1078

Williams D. M., Pollard D., 2002, International Journal of Astrobiology, 1, 61

Williams J. P., McKee C. F., 1997, ApJ, 476, 166

Winkler K., 1984, Journal of Quantitative Spectroscopy and Radiative Transfer, 31, 473

Woitke P., Pinte C., Tilling I., Ménard F., Kamp I., Thi W.-F., Duchêne G., Augereau J.-C., 2010, MNRAS, 405, L26

Wood K., Wolff M. J., Bjorkman J. E., Whitney B., 2002, ApJ, 564, 887

Würtz P., Annila A., 2008, Journal of Biophysics, 2008, 654

Wyatt M. C., Clarke C. J., Greaves J. S., 2007, MNRAS, 380, 1737

Yusef-Zadeh F., Morris M., White R. L., 1984, ApJ, 278, 186

Zhu Z., Hartmann L., Gammie C., 2009a, ApJ, 694, 1045

Zhu Z., Hartmann L., Gammie C., McKinney J. C., 2009b, ApJ, 701, 620

Zwicky F., 1933, Helvetica Physica Acta, 6, 110

Appendix A

Deriving the Equations of Hydrodynamics

In this Appendix I will illustrate the construction of the equations of hydrodynamics (initially without the presence of gravity).

A.1 The Equation of Continuity

The *equation of continuity* expresses the conservation of matter in the fluid. Consider a volume V, with a mass of fluid inside given by $\int \rho \, dV$. We can bound this volume by some surface S - for an infinitesimal surface element, there exists a normal vector $\mathbf{dA}$ (with a magnitude dA equal to the area of the surface element). The mass of fluid flowing out of this surface in unit time is given by $\rho \mathbf{v}.d\mathbf{A}$, with the sign of this quantity indicating if the flow is outward (positive) or inward (negative). We must integrate over the whole surface to obtain the total mass flux out of the volume:

$$\oint_S \rho \mathbf{v}.\mathbf{dA}. \tag{A.1}$$

The decrease of mass in the volume is simply

$$-\frac{\partial}{\partial t} \int_V \rho \, dV. \tag{A.2}$$

To ensure conservation of mass, these terms must be equated:

$$\oint_S \rho \mathbf{v}.d\mathbf{A} = -\frac{\partial}{\partial t}\int_V \rho\, dV. \tag{A.3}$$

The divergence theorem can convert the surface integral into a volume integral, and we can add to the right hand side to the left hand side:

$$\int_V \left(\frac{\partial \rho}{\partial t} + \nabla.(\rho \mathbf{v})\right) dV = 0. \tag{A.4}$$

As we have not specified any geometry for the volume in this integral, the integrand must always be zero. This gives us the equation of continuity:

$$\frac{\partial \rho}{\partial t} + \nabla.(\rho \mathbf{v}) = 0. \tag{A.5}$$

A.2 Euler's Equation

We must now find an equation for the fluid velocity. Let us consider the same volume as in the previous section. The force acting on this volume is given by the integral

$$\mathbf{F} = -\oint P\, d\mathbf{A} = -\int \nabla P\, dV, \tag{A.6}$$

where we have used the divergence theorem for the (scalar) pressure P. We now understand that a force $-\nabla P$ acts on the fluid (per unit volume). Newton's second law allows us to derive an equation of motion:

$$\mathbf{F} = \rho \frac{d\mathbf{v}}{dt} = -\nabla P. \tag{A.7}$$

We must now acknowledge a common subtlety of fluid dynamics related to derivatives. The above derivative is a *Lagrangian or connective derivative*, which is said to *follow the motion* - i.e. the derivative denotes the rate of change of velocity for a given fluid element as it moves in the medium (as opposed to the alternative *Eulerian derivative* where the rate of change is measured at a fixed location and does not directly refer to any individual fluid element). We must therefore decompose the Lagrangian derivative into its two components - the velocity change at a fixed location, and the difference in velocities between the fluid element's initial and final positions:

$$\frac{d\mathbf{v}}{dt} = \frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}.\nabla)\mathbf{v}. \tag{A.8}$$

Substituting this result into our equation of motion gives *Euler's Equation*:

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}.\nabla)\mathbf{v} = -\frac{1}{\rho}\nabla P. \tag{A.9}$$

If an external force $\mathbf{F}$ is exerted on the fluid, this equation is simply modified:

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}.\nabla)\mathbf{v} = -\frac{1}{\rho}\left(\nabla P + \mathbf{F}\right). \tag{A.10}$$

A.3 The Energy Equation

The law of conservation of energy must be observed by the fluid - we should therefore use this constraint to identify how the fluid's energy evolves as a function of space and time. The First Law of Thermodynamics states

$$dU = TdS - PdV, \tag{A.11}$$

where U is the internal energy of the fluid, T is the temperature of the fluid and S is the entropy. If we wish to deal in quantities per unit mass, then we obtain

$$du = Tds - Pd\left(\frac{1}{\rho}\right) = Tds + \frac{P}{\rho^2}d\rho. \tag{A.12}$$

Taking the derivative with respect to t gives the internal energy equation:

$$\frac{du}{dt} = T\frac{ds}{dt} + \frac{P}{\rho^2}\frac{d\rho}{dt} = T\frac{ds}{dt} - \frac{P}{\rho}(\nabla.\mathbf{v}). \tag{A.13}$$

Where we have used the continuity equation in the last line[1]. If the fluid is adiabatic, then $ads = 0$, and the rate of change of u is given entirely by the second term. The total energy of the fluid is

$$E = \rho e = \frac{1}{2}\rho v^2 + \rho u. \tag{A.14}$$

By inspection, it is clear that E is completely determined by the continuity equation, Euler's equation and the internal energy equation.

The *sound speed* of the medium c_s is:

$$c_s^2 = \frac{dP}{d\rho}, \tag{A.15}$$

and the *ratio of specific heats*, γ:

$$\gamma = \frac{C_P}{C_V} = \left(\frac{dQ}{dT}\right)_P \left(\frac{dQ}{dT}\right)_V^{-1}. \tag{A.16}$$

For an adiabatic ideal gas, the speed of sound can be given by

$$c_s = \sqrt{\frac{\gamma P}{\rho}}, \tag{A.17}$$

and the relation $P = K\rho^\gamma$ holds. If this is true, then we can derive an expression for the internal energy u as a function of c_s and γ. The First Law of Thermodynamics in this case is

[1] We have assumed the fluid is incompressible. Also, note that the derivatives are Eulerian in this case

$$du = \frac{P}{\rho^2} d\rho. \tag{A.18}$$

Substituting for P and integrating gives

$$u = \int_0^\rho K\rho^{\gamma-2} d\rho = \frac{K\rho^{\gamma-1}}{\gamma - 1}. \tag{A.19}$$

The sound speed is $c_s^2 = \gamma K \rho^{\gamma-1}$, so we can substitute to obtain

$$u = \frac{c_s^2}{\gamma(\gamma - 1)}. \tag{A.20}$$

A.4 Viscosity and the Navier-Stokes Equation

We must now consider the properties of viscous fluids (that is, fluids which dissipate energy and propagate momentum as a result of internal friction). Before we can do so, we must discuss explicitly the momentum density of the fluid, given by $\rho\mathbf{v}$. We will begin by discussing momentum flux for fluids without viscosity.

Momentum Flux for Inviscid Fluids We can derive an equation of motion for momentum flux in the inviscid case by considering Euler's equation and the equation of continuity. We will use index notation (in anticipation of a tensor to follow), assuming Einstein's Summation Convention. The product rule gives us:

$$\frac{\partial}{\partial t}(\rho v_i) = \rho \frac{\partial v_i}{\partial t} + \frac{\partial \rho}{\partial t} v_i. \tag{A.21}$$

We will now substitute using equations (A.5) and (A.9) in index notation:

$$\frac{\partial \rho}{\partial t} = -\frac{\partial(\rho v_k)}{\partial v_k}, \tag{A.22}$$

$$\frac{\partial v_i}{\partial t} = -v_k \frac{\partial v_i}{\partial x_k} - \frac{1}{\rho}\frac{\partial P}{\partial x_i}. \tag{A.23}$$

This gives

$$\frac{\partial}{\partial t}(\rho v_i) = -\rho v_k \frac{\partial v_i}{\partial x_k} - \frac{\partial P}{\partial x_i} - v_i \frac{\partial(\rho v_k)}{\partial v_k}. \tag{A.24}$$

By use of the Kronecker delta δ_{ik} on $\frac{\partial P}{\partial x_i}$, we can obtain an expression of the form

$$\frac{\partial}{\partial t}(\rho v_i) = -\frac{\partial \Pi_{ik}}{\partial x_k}, \tag{A.25}$$

where we have defined the *momentum flux density tensor* Π:

$$\Pi_{ik} = P\delta_{ik} + \rho v_i v_k. \tag{A.26}$$

We can confirm its function by specifying a normal vector $\mathbf{n}$, and taking a scalar product

$$\Pi_{ik} n_k = P n_i + \rho v_i v_k n_k, \tag{A.27}$$

or equivalently $P\mathbf{n} + \rho\mathbf{v}(\mathbf{v}.\mathbf{n})$, describing the momentum flux along the normal vector.

Momentum Flux for Viscous Fluids We must now consider the form of Π_{ik} for a viscous fluid. We can simply add a tensor to this equation to do this - however we have a choice of two. We can choose the viscous stress tensor σ'_{ik}

$$\Pi_{ik} = P\delta_{ik} + \rho v_i v_k - \sigma'_{ik}, \tag{A.28}$$

or the more convenient stress tensor σ_{ik}:

$$\sigma_{ik} = \sigma'_{ik} - P\delta_{ik}, \tag{A.29}$$

Giving the simplified

$$\Pi_{ik} = \rho v_i v_k - \sigma_{ik}. \tag{A.30}$$

The full stress tensor describes any momentum flux not associated with direct transfer of momentum "in carriage" by the motion of fluid elements. The simplest course of action is to now proceed to articulate the form of the viscous stress tensor σ'_{ik}, associated with internal friction in the fluid. We can intuit that the tensor must be related to derivatives of the velocity as a function of space (as friction requires neighbouring fluid elements to move with differing velocities to each other). In the limit of small velocity gradients, we can assume the first order derivatives $\frac{\partial v_i}{\partial x_k}$ to be sufficient, and that only linear terms are necessary. Also, for the same reasons as we have just described, we require σ'_{ik} to vanish in the absence of velocity gradients (as friction will be zero under these circumstances).

For example, if we set the fluid rotating uniformly with angular velocity $\mathbf{\Omega}$, for σ'_{ik} to vanish

$$\mathbf{v} = \mathbf{\Omega} \times \mathbf{r} = const. \tag{A.31}$$

The components of $\mathbf{v}$ are

$$\begin{aligned} v_i &= \Omega_j r_k - \Omega_k r_j \\ v_j &= \Omega_k r_i - \Omega_i r_k \\ v_k &= \Omega_i r_j - \Omega_j r_i \end{aligned} \tag{A.32}$$

We can see that the sum

$$\frac{\partial v_i}{\partial x_k} + \frac{\partial v_k}{\partial x_i} \tag{A.33}$$

satisfies all the conditions above (including the condition that Π should remain symmetric). Let us therefore reveal the most general rank 2 tensor that we can propose satisfying all the above conditions:

$$\sigma'_{ik} = \eta\left(\frac{\partial v_i}{\partial x_k} + \frac{\partial v_k}{\partial x_i} - \frac{2}{3}\delta_{ik}\frac{\partial v_l}{\partial x_l}\right) + \zeta\delta_{ik}\frac{\partial v_l}{\partial x_l}. \tag{A.34}$$

We select this form to ensure that the bracketed term disappears for $i = k$. η and ζ are *coefficients of viscosity*: η is usually referred to as the *dynamic viscosity*. We can define the *kinematic viscosity* ν as

$$\nu = \frac{\eta}{\rho}. \tag{A.35}$$

Euler's Equation with Viscosity - The Navier-Stokes Equation We can now update Euler's equation to include the viscous component (in index notation):

$$\frac{\partial v_i}{\partial t} + v_k\frac{\partial v_i}{\partial x_k} = -\frac{1}{\rho}\frac{\partial P}{\partial x_i} + \frac{\partial \sigma'_{ik}}{\partial x_k}. \tag{A.36}$$

We shall assume that the coefficients of viscosity are slowly varying in the fluid, and are therefore considerable as constants. We can now arrive at the *Navier-Stokes equation*

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}.\nabla)\mathbf{v} = -\frac{1}{\rho}\nabla P + \frac{\eta}{\rho}\nabla^2\mathbf{v} + \frac{1}{\rho}(\zeta + 1/3\eta)\nabla(\nabla.\mathbf{v}). \tag{A.37}$$

If we regard the fluid as *incompressible* (i.e. $\rho = const.$), then the continuity equation gives $\nabla.\mathbf{v} = 0$, simplifying the Navier-Stokes equation to a function of one coefficient of viscosity (where we now substitute for ν):

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}.\nabla)\mathbf{v} = -\frac{1}{\rho}\nabla P + \nu\nabla^2\mathbf{v}. \tag{A.38}$$

A.5 Magneto-hydrodynamics

While magnetic fields will not figure largely in the numerical work and resulting science covered in this thesis, it is important to discuss their effects to provide context. For an inviscid non-gravitating fluid, the equations of *ideal MHD* (where the fluid has resistivity $R = 0$) are:

$$\frac{\partial \rho}{\partial t} + \nabla.(\rho\mathbf{v}) = 0, \tag{A.39}$$

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}.\nabla)\mathbf{v} = -\frac{1}{\rho}\left(\nabla P + \mathbf{j}\times\mathbf{B}\right), \tag{A.40}$$

Alongside Maxwell's equations to evolve the magnetic field $\mathbf{B}$ and the current density $\mathbf{j}$,

$$\nabla.\mathbf{B} = 0, \tag{A.41}$$

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{j}, \tag{A.42}$$

$$\nabla \times (\mathbf{v} \times \mathbf{B}) = -\frac{\partial \mathbf{B}}{\partial t}, \tag{A.43}$$

where we have eliminated $\mathbf{E}$ from the last equation by using Ohm's Law in the limit of zero resistivity:

$$\mathbf{j}R = \mathbf{E} + \mathbf{v} \times \mathbf{B} = 0. \tag{A.44}$$

It behooves us to study the phenomenology of the magnetic field in a fluid, as it will impact significantly on an important type of instability, (see section 2.7.2). We have added a force of form $\mathbf{j} \times \mathbf{B}$ to Euler's equation. We can expand this to attain insight into how this force will act:

$$\mathbf{j} \times \mathbf{B} = \frac{1}{\mu_0}(\nabla \times \mathbf{B}) \times \mathbf{B}. \tag{A.45}$$

The usual relation for a vector triple product gives:

$$\mathbf{j} \times \mathbf{B} = \frac{1}{\mu_0}\left(\mathbf{B}(\mathbf{B}.\nabla) - \nabla.(\mathbf{B}.\mathbf{B})\right) = \frac{1}{\mu_0}(\mathbf{B}.\nabla)\mathbf{B} - \nabla\left(\frac{B^2}{\mu_0}\right). \tag{A.46}$$

The force is now split into two terms. The second term takes the form of a magnetic pressure, acting along the gradient of the magnetic field. The first term takes the form of a *tension*, which is non-zero when the magnetic field lines are curved. In ideal MHD, the magnetic field lines are tethered elastically to the fluid, and therefore this tension is also felt by the fluid. If the fluid is perturbed by other forces, the magnetic field will act to remove induced curvature in the field line (in accordance with Lenz's Law), dragging the fluid with it. This will prove to be important for stability studies in magnetised discs (section 2.7.2).

Appendix B

Derivations Regarding Gravity

Poisson's Equation from Newton's Universal Law of Gravitation

To obtain the form of Poisson's equation for the gravitational field, we must solve the expression

$$\nabla^2\phi(\mathbf{r}) = \nabla.\mathbf{g}(\mathbf{r}) = G\int \rho(\mathbf{r}')\nabla.\left(\frac{(\mathbf{r}-\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|^3}\right) d^3\mathbf{r}. \tag{B.1}$$

Consider the expression $\nabla.\left(\frac{\mathbf{r}}{|\mathbf{r}|^3}\right) = \nabla.\left(\frac{\hat{\mathbf{r}}}{|\mathbf{r}|^2}\right)$. As the expression is a function of $\hat{\mathbf{r}}$ only, the divergence in spherical polar coordinates has only one term:

$$\nabla.\left(\frac{\hat{\mathbf{r}}}{|\mathbf{r}|^2}\right) = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{1}{r^2}\right) = 0. \tag{B.2}$$

But, we can see that this solution is discontinuous and infinite when $r = 0$. To resolve this discontinuity, we can integrate over some volume (using the divergence theorem):

$$\int_V \nabla.\left(\frac{\hat{\mathbf{r}}}{|\mathbf{r}|^2}\right) dV = \oint_S \frac{\hat{\mathbf{r}}}{|\mathbf{r}|^2}.d\mathbf{A}. \tag{B.3}$$

If we assume the volume is spherical, then the surface integral is easily solved in spherical polar coordinates:

$$\int_0^{\pi}\int_0^{2\pi} \frac{1}{r^2}r^2\sin\theta\, d\theta\, d\phi = 4\pi. \tag{B.4}$$

The integral is finite, but the integrand is discontinuous at $r = 0$. To solve this, we employ the Dirac delta function, which is defined:

$$\delta(\mathbf{x}) = \begin{cases} \infty & \mathbf{x} = 0 \\ 0 & \text{otherwise} \end{cases} \tag{B.5}$$

We now have the self-consistent solution

$$\nabla . \left(\frac{\hat{\mathbf{r}}}{|\mathbf{r}|^2} \right) = 4\pi\delta(\mathbf{r}), \tag{B.6}$$

which we can substitute into our expression for $\nabla.\mathbf{g}$:

$$\nabla.\mathbf{g} = G \int \rho(\mathbf{r}')4\pi\delta(\mathbf{r} - \mathbf{r}')\, d^3\mathbf{r}'. \tag{B.7}$$

The sifting property of the δ function allows us to solve the integral simply, retrieving *Poisson's Equation* for the gravitational field:

$$\nabla.\mathbf{g} = \nabla^2\Phi(\mathbf{r}) = 4\pi G\rho(\mathbf{r}). \tag{B.8}$$

The Free Fall Time for a Uniform Sphere

Our aim is to calculate the free-fall time for a sphere of uniform density ρ and radius r. We begin this calculation with Kepler's Third Law:

$$P^2 = \frac{4\pi^2}{GM}a^3 \tag{B.9}$$

where P is the period of the orbit, M is the mass of the primary and a is the semi-major axis of the orbit. This applies to elliptic orbits only - we wish to detail the *radial* motion of a mass m toward the centre of mass M. We can do this by using a *degenerate ellipse*, defined as an ellipse with eccentricity $e = 1$ and semi-minor axis $b = a(1 - e^2) = 0$, which is equivalent to a line segment. As this line segment constitutes the entire orbit, we should set $a = r/2$.

As we are modelling free-fall, the mass m will experience only half the orbit before reaching the centre of mass. Therefore, the free-fall time is half the orbital timescale for the degenerate ellipse:

$$t_{ff} = P/2 = \frac{1}{2}\sqrt{\frac{\pi^2 r^3}{2GM}} \tag{B.10}$$

To complete the calculation, we substitute $M = 4/3\pi r^3\rho$ (this is appropriate as the gravitational acceleration felt by m depends only on the mass within r). This gives the free-fall timescale as

$$t_{ff} = \sqrt{\frac{3\pi}{32G\rho}}. \tag{B.11}$$

Appendix C

Spiral Structure in Discs and the Dispersion Relation

Spiral structure is an important component of self-gravitating discs. Indeed, without spiral arms, self-gravitating protostellar discs would not be able to achieve a quasi-steady state, allowing them to evolve into non self-gravitating discs, which in turn sets the scene for planet formation.

The study of spiral structure in discs was originally motivated by the pursuit of understanding spiral arms in galaxies. A persistent problem to the astronomers of the early 20th Century, the presence of spiral arms was often erroneously ascribed to the interstellar magnetic fields that pervaded the galaxy. It was Lindblad who eventually recognised that the spiral arms were dynamical in origin, a result of the interaction between the stars' orbits and the gravitational field they generate. While the qualitative insight was correct, the quantitative framework would have to wait until Lin and Shu's density wave hypothesis (Binney & Tremaine, 1987). By regarding the arm as a density wave, much of the mathematics of wave mechanics could be brought to bear on the subject. They also hypothesised that the spirals were *long lasting*, a somewhat controversial statement. These two statements are usually gathered together, and referred to as the Lin-Shu hypothesis. While originally intended for spiral galaxies, the formalism is scale-free - I will use their results to elucidate how spiral structure can be characterised in protostellar discs.

The WKB Approximation

We wish to characterise the spiral wave's *radial* and *azimuthal* structure. This is in general a complicated process - as the spiral structure is determined by the disc's self-gravity, which is a long range force and will have some non-local behaviour, the most sensible approach is to determine these numerically.

However, if we make some limiting assumptions about the waves, then the analysis becomes much easier. These assumptions are referred to as the *WKB*[1] or *tight-winding* approximation. By assuming the waves are tightly wound, the long-range coupling effects disappear, and analytic solutions become available.

Let us consider the equation for a spiral arm in a disc, in terms of azimuthal angle ϕ and some function g:

$$\phi + g(r,t) = const. \tag{C.1}$$

If the disc has $m > 0$ identical spiral arms, the pattern will be invariant to rotations of $2\pi/m$. We can then define m spiral arms as

$$m\phi + f(r,t) = const. \,(\mathrm{mod}\, 2\pi), \tag{C.2}$$

where $f(r,t) = mg(r,t)$. The radial separation between two arms Δr is determined by

$$|f(r+\Delta r, t) - f(r,t)| = 2\pi. \tag{C.3}$$

By assuming the arms are tightly wound, then

$$\left| f(r,t) + \frac{\partial f}{\partial r}\Delta r - f(r,t) \right| = 2\pi. \tag{C.4}$$

Δr becomes the radial wavelength λ

$$\lambda(r,t) = \frac{2\pi}{\left|\frac{\partial f(r,t)}{\partial r}\right|}. \tag{C.5}$$

It is useful for our purposes to introduce the radial wavenumber k:

$$k(r,t) = \frac{\partial f(r,t)}{\partial r}. \tag{C.6}$$

While $\lambda = 2\pi/|k|$ is constrained to be always positive, k can be either positive or negative. This indicates whether the arms *lead* ($k < 0$) or *trail* ($k > 0$). We define the pitch angle of the arm as

$$\cot i = \left| \frac{kr}{m} \right|. \tag{C.7}$$

For the WKB approximation to hold, $\cot i \gg 1$.

[1] It receives this name thanks to Wentzel, Kramers and Brillouin, for their work on a similar approximation in quantum mechanics.

The Dispersion Relation

If the WKB approximation holds, we can determine the dispersion relation for an infinitesimally thin gaseous disc responding to a tightly wound spiral perturbation (see also Binney & Tremaine 1987). Let us begin with Euler's Equation for the radial and azimuthal velocity components of the disc:

$$\frac{\partial v_r}{\partial t} + v_r\frac{\partial v_r}{\partial r} + \frac{v_\phi}{r}\frac{\partial v_r}{\partial \phi} - \frac{v_\phi^2}{r} = -\frac{\partial \Phi}{\partial r} - \frac{1}{\Sigma}\frac{\partial P}{\partial r}, \tag{C.8}$$

$$\frac{\partial v_\phi}{\partial t} + v_r\frac{\partial v_\phi}{\partial r} + \frac{v_\phi}{r}\frac{\partial v_\phi}{\partial \phi} + \frac{v_r v_\phi}{r} = -\frac{\partial \Phi}{\partial \phi} - \frac{1}{\Sigma r}\frac{\partial P}{\partial \phi}. \tag{C.9}$$

We will assume a simple equation of state for the gas

$$P = K\Sigma^\gamma, \tag{C.10}$$

and simplify the above equations by using the specific enthalpy $h = \int \frac{dP}{\rho}$. This gives

$$h = \frac{\gamma}{\gamma - 1}K\Sigma^{\gamma-1}, \tag{C.11}$$

which reduces the right hand side of equation (C.8) to

$$-\frac{\partial \Phi}{\partial r} - \gamma K\Sigma^{\gamma-2}\frac{\partial \Sigma}{\partial r} = \frac{\partial}{\partial r}(\Phi + h), \tag{C.12}$$

and similarly for equation (C.9). Let us now add a perturbation to the disc, writing our variables in the form

$$v_r = v_{r0} + v_{r1}, \quad v_\phi = v_{\phi_0} + v_{\phi_1}, \tag{C.13}$$

$$h = h_0 + h_1, \quad \Phi = \Phi_0 + \Phi_1, \quad \Sigma = \Sigma_0 + \Sigma_1. \tag{C.14}$$

The unperturbed disc should be radially stable and axisymmetric, i.e. $v_{r0} = 0$, $d\Phi_0/d\phi = dh_0/d\phi = 0$. We will also assume that the local sound speed is much smaller than the rotation speed, allowing us to approximate

$$v_{\phi 0} = \sqrt{r\frac{d\Phi_0}{dr}} = r\Omega. \tag{C.15}$$

This can be seen by considering equation (C.8) in the unperturbed case, neglecting the pressure term. We can now arrive at the perturbed version of equations (C.8) and (C.9):

$$\frac{\partial v_{r1}}{\partial t} + \Omega\frac{\partial v_{r1}}{\partial \phi} - 2\Omega v_{\phi 1} = -\frac{\partial}{\partial r}(\Phi_1 + h_1), \tag{C.16}$$

$$\frac{\partial v_{\phi 1}}{\partial t} + v_{r1}\frac{d(\Omega r)}{dr} + v_{r1}\Omega + \Omega\frac{\partial v_{\phi 1}}{\partial \phi} = -\frac{1}{r}\frac{\partial}{\partial \phi}(\Phi_1 + h_1). \tag{C.17}$$

We shall introduce a new variable $B(r)$[2] and the epicyclic frequency $\kappa(r)$ to assist us:

$$B(r) = -\frac{1}{2}\left(\frac{d(\Omega r)}{dr} + \Omega\right) = -\Omega - \frac{1}{2}\frac{d\Omega}{dr}r, \tag{C.18}$$

$$\kappa^2(r) = -4B\Omega. \tag{C.19}$$

In a Keplerian disc, $\kappa = \Omega$. To get to the dispersion relation, we shall assume a wave solution to the perturbed Euler's equations, i.e.

$$v_{r1} = Re\left[v_{ra}e^{i(m\phi-\omega t)}\right], \tag{C.20}$$

and similarly for the perturbed variables $v_{\phi 1}$,Φ_1, Σ_1 and h_1. Substituting these solutions (and retaining linear order terms in v_{r1}, etc) gives a set of simultaneous equations for v_{ra} and $v_{\phi a}$, which reduce to:

$$v_{ra} = -\frac{i}{\kappa^2 - (m\Omega-\omega)^2}\left((m\Omega-\omega)\frac{d}{dr}(\Phi_a + h_a) + \frac{2m\Omega}{r}(\Phi_a + h_a)\right), \tag{C.21}$$

$$v_{\phi a} = \frac{1}{\kappa^2 - (m\Omega-\omega)^2}\left(-2B\frac{d}{dr}(\Phi_a + h_a) + \frac{m(m\Omega-\omega)}{r}(\Phi_a + h_a)\right). \tag{C.22}$$

We must also constrain Φ_a and h_a. Firstly, the enthalpy:

$$h_a = c_s^2\frac{\Sigma_a}{\Sigma_0}. \tag{C.23}$$

The surface density must itself obey the continuity equation. We should be careful however to include the extra terms that deal with the non-axisymmetry induced by the perturbation:

$$\frac{\partial\Sigma_1}{\partial t} + \frac{1}{r}\frac{\partial}{\partial r}(r\Sigma_0 v_{r1}) + \Omega\frac{\partial\Sigma_1}{\partial\phi} + \frac{\Sigma_0}{r}\frac{\partial v_{\phi 1}}{\partial\phi} = 0. \tag{C.24}$$

Assuming a wave solution gives

$$i(m\Omega-\omega)\Sigma_a + \frac{1}{r}\frac{d}{dr}(r\Sigma_0 v_{ra}) + \frac{im\Sigma_0}{r}v_{\phi a} = 0. \tag{C.25}$$

We now have four constraints for our set of five variables $(v_{ra}, v_{\phi a}, h_a, \Sigma_a, \Phi_a)$. We obtain the fifth by using Poisson's equation to link the potential to the surface density. Under the WKB approximation, the potential for a tightly wound spiral wave is of the form:

$$\Phi_a(r) = F(r)e^{i\int k dr}, \tag{C.26}$$

where we have already defined the radial wavenumber k. This holds when $|kr| \gg 1$, and the fractional error of the solution is $O(|kr|^{-1})$. The complex exponential in Φ appears in Σ_a thanks to Poisson's equation, and subsequently in h_a. Therefore the same fractional error is

[2]Note that this is equivalent to Oort's Constant when considering the Sun's orbit in the Galaxy.

also propagated, allowing us to discard the final terms in equations (C.21) and (C.22) without increasing the existing error. Equally, we may rewrite

$$\frac{d(\Phi_a + h_a)}{dr} = ik(\Phi_a + h_a), \tag{C.27}$$

without further increase of error. Equations (C.21) and (C.22) then become

$$v_{ra} = \frac{(m\Omega - \omega)k(\Phi_a + h_a)}{\kappa^2 - (m\Omega - \omega)^2}, \tag{C.28}$$

$$v_{\phi_a} = \frac{2ikB(\Phi_a + h_a)}{\kappa^2 - (m\Omega - \omega)^2}. \tag{C.29}$$

And we can simplify the continuity equation in a similar manner:

$$(m\Omega - \omega)\Sigma_a + k\Sigma_0 v_{ra} + \frac{m\Sigma_0}{r} v_{\phi a} = 0. \tag{C.30}$$

The second term dominates the third term as $|kr| \gg 1$ and we have established the two velocity components to be similar in magnitude. We then have

$$(m\Omega - \omega)\Sigma_a + k\Sigma_0 v_{ra} = 0. \tag{C.31}$$

We can substitute for v_{ra}, h_a and Φ_a having solved Poisson's Equation (Binney & Tremaine, 1987):

$$\Phi_a = -\frac{2\pi G\Sigma_a}{|k|}. \tag{C.32}$$

After some arrangement, the dispersion relation finally appears:

$$m^2(\Omega - \Omega_p)^2 = c_s^2 k^2 - 2\pi G\Sigma\,|k| + \kappa^2, \tag{C.33}$$

where Ω_p is the pattern speed of the wave (i.e. the wave is of the form $\propto e^{im(\phi - \Omega_p t)}$). If the disc has finite thickness, the solution for the potential must be modified, giving the dispersion relation to be (Bertin, 2000; Cossins et al., 2009)

$$m^2(\Omega - \Omega_p)^2 = c_s^2 k^2 - \frac{2\pi G\Sigma\,|k|}{1 + |k|H} + \kappa^2, \tag{C.34}$$

where H is the scale height of the disc.

Appendix D

Monte Carlo Realisation Techniques, and a Numerical Testbed for SETI

Let's think the unthinkable, let's do the undoable, let's prepare to grapple with the ineffable itself, and see if we may not eff it after all.
Douglas Adams, *Dirk Gently's Holistic Detective Agency*

D.1 Author's Note

I include this appendix for the interested reader. Its substance does not directly contribute to the main narrative of this thesis, but does describe astrobiological research I conducted during the course of my studentship. The content of this appendix has appeared in various articles, of which I was first author (with one exception) (Forgan, 2009; Forgan & Rice, 2010b; Forgan, 2010; Bozhilov & Forgan, 2010; Forgan & Nichol, 2010). The numerical method described in Forgan (2009) was significantly updated in Forgan & Rice (2010b). For the sake of brevity, I will only discuss the updated version.

D.2 Introduction

Astrobiology is undergoing something of a phase transition. In the past two decades, it has progressed from broad discussions hamstrung by limited data, such as Fermi's first formulation of his classic Paradox, towards an era of burgeoning empiricism, driven largely by two separate observational astrophysical programs.

The first is the search for planets around other stars, begun in earnest with the first detection of a planet around the solar type star 51 Pegasi (Mayor & Queloz, 1995). These discoveries have since led to a growing program devoted to exoplanet detection, with several methods of identifying planets populating diverse parameters. With close to 500 exoplanets detected at the time of writing, astronomers are receiving their first glimpse at data on the available niches for life in the Galaxy - while at the same jettisoning assumptions collected over the centuries when the only known planets to exist were those in our Solar System.

Secondly, missions in the Solar System itself have increased our knowledge of our neighbours. For example, Cassini has illustrated the commonalities between the Saturnian regular satellites and Earth - the presence of hydrocarbon lakes and watery matrices on Titan (e.g. Stofan et al. 2007), or geysers on Enceladus (Parkinson et al., 2007; Spencer & Grinspoon, 2007). Discoveries such as these, coupled with our understanding of the adaptability of extremophiles, erode the concept of the stellar habitable zone (Hart, 1979; Kasting, 1993) as the only locale in a star system where amenable conditions for life exist. Given the fact that microorganisms have been discovered even in the most life-neutralising environments, from miles beneath the earth, at the bottom of the ocean, within radioactive waste, and in below zero temperatures within snow packs and ice, the notion of "habitability" as a discrete quantity for a celestial body is increasing difficult to apply even to the Earth (Spiegel et al., 2008). How do we apply these new discoveries to the science of the Search for Extraterrestrial Intelligence (SETI)?

D.3 A History of Numerical Techniques in SETI

D.3.1 Drake's Equation

The science of SETI has always suffered from a lack of quantitative substance (purely resulting from its reliance on one-sample statistics) relative to its sister astronomical sciences. In 1961, Frank Drake took the first steps to quantifying the field by developing the now-famous Drake Equation, a simple algebraic expression which provides an estimate for the number of communicating civilisations in the Milky Way. Unfortunately, its simplistic nature leaves it open to frequent re-expression, hence there are in fact many variants of the equation, and no clear canonical form. For the purpose of this appendix, the following form will be used (Walters, 1980):

$$N = R_* f_g f_p n_e f_l f_i f_c L \tag{D.1}$$

With the symbols having the following meanings:

N = The number of Galactic civilisations who can communicate with Earth
R_* = The mean star formation rate of the Milky Way
f_g = The fraction of stars that could support habitable planets
f_p = The fraction of stars that host planetary systems
n_e = The number of planets in each system which are potentially habitable
f_l = The fraction of habitable planets where life originates and becomes complex
f_i = The fraction of life-bearing planets which bear intelligence
f_c = The fraction of intelligence bearing planets where technology can develop
L = The mean lifetime of a technological civilisation within the detection window

The equation itself does suffer from some key weaknesses: it relies strongly on mean estimations of variables such as the star formation rate; it is unable to incorporate the effects of the physico-chemical history of the galaxy, or the time-dependence of its terms. Indeed, it is criticised for its polarising effect on "contact optimists" and "contact pessimists", who ascribe very different values to the parameters, and return values of N between 10^{-5} and 10^{6} (!).

D.3.2 Fermi's Paradox

A decade before, attempting to analyse the problem from a different angle, Enrico Fermi used order of magnitude estimates for the timescales required for an Earthlike civilisation to arise and colonise the galaxy, arriving at the conclusion that the Milky Way should be teeming with intelligence, and that they should be seen all over the sky. This lead him to pose the Fermi Paradox, by asking, "Where are they?". The power of this question, along with the enormous chain of events required for intelligent observers to exist on Earth to pose it, has lead many to the conclusion that the conditions for life to flourish are rare, possibly even unique to Earth (Ward & Brownlee, 2000). The inference by Lineweaver (2001) that the median age of terrestrial planets in the Milky Way is 1.8 ± 0.9 Gyr older than Earth would suggest that a significant number of Earthlike civilisations have had enough time to evolve, and hence be detectable: the absence of such detection lends weight to the so-called "rare Earth" hypothesis. However, there have been many posited solutions to the Fermi Paradox that allow ETI to be prevalent, such as:

- They are already here, in hiding
- Contact with Earth is forbidden for ethical reasons
- They were here, but they are now extinct
- They will be here, if Mankind can survive long enough

Thorough reviews of the Paradox can be found in Brin (1983) and Cirkovic (2009). Some of these answers are inherently sociological, and are difficult to model. Others are dependent on the evolution of the galaxy and its stars, and are much more straightforward to verify. As a whole, astrobiologists are at a tremendous advantage in comparison with Drake and Fermi: the development of astronomy over the last fifty years - in particular the discovery of the first extra solar planet (Mayor & Queloz, 1995) and some hundreds thereafter, as well as the concepts of habitable zones, from planetary (Spiegel et al., 2008), to stellar (Hart, 1979; Kasting, 1993) and galactic (Lineweaver et al., 2004) - have allowed a more quantifiable analysis of the problem. However, the key issue still affecting SETI (and astrobiology as a whole) is that there is no consensus as to how to assign values to the key *biological* parameters involved in the Drake Equation and Fermi Paradox.

D.4 The Numerical Method - Constructing a Synthetic Galaxy

I will now outline the numerical techniques I have developed for SETI research. The method involves Monte Carlo Realisation (MCR) - the overall procedure can be summarised as:

1. Randomly generate a galaxy of N_* stars, with parameters that share the same distribution as observations

2. Randomly generate planetary systems for these stars

3. Assign life to some of the planets depending on their parameters (e.g. distance from the habitable zone)

4. For each life-bearing planet, follow life's evolution into intelligence using stochastic equations

This will produce one Monte Carlo Realisation (MCR) of the Milky Way in its entirety. The concept of using MCR techniques in astrobiology is itself not new - Vukotic & Cirkovic (2007, 2008) used similar techniques coupled with the concept of "global regulation mechanisms" Annis (1999) to challenge Carter's classic argument against SETI. By allowing galactic gamma ray bursts (GRBs) to impede life across the entire Galaxy at the same instant in history, their numerical work demonstrates that the astrophysical timescale (e.g. the lifetime of a Main Sequence star such as the Sun) and the biological timescale required for the formation of intelligent life become correlated, undermining Carter's principal assumption that they are in fact uncorrelated (Cirkovic, 2009). The age of the Galaxy is therefore the incorrect timescale to adopt for Fermis Paradox, and should instead be the time from the last "resetting event". Being much shorter than the galaxy crossing timescale, the Paradox is then resolved.

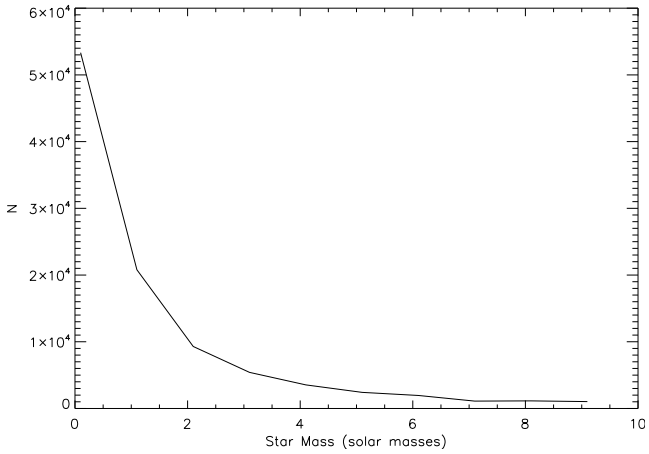

Figure D.1: The stellar IMF used in this work (Miller & Scalo, 1979). This is an example produced as part of a single MCR.

In order to provide error estimates, the Monte Carlo Realisation procedure must be repeated many times, so as to produce many MCRs, and to hence produce results with a well-defined sample mean and sample standard deviation. The procedure relies on generating parameters in three categories: stellar, planetary and biological.

Stellar Properties

The study of stars in the Milky Way has been extensive, and their properties are well-constrained. Assuming the stars concerned are all main sequence objects allows much of their characteristics to be determined by their mass. Stellar masses are randomly sampled to reproduce the observed initial mass function (IMF) of the Milky Way, which is taken from Miller & Scalo (1979) (see Figure D.1).

Stellar radii can then be calculated using (Prialnik, 2000):

$$\frac{R_*}{R_\odot} = \left[\frac{M_*}{M_\odot}\right]^{\frac{n-1}{n+3}} \tag{D.2}$$

where $n = 4$ if the primary fusion mechanism is the p-p chain ($M_* \leq 1.1\,M_\odot$), and $n = 16$ if the primary fusion mechanism is the CNO cycle ($M_* > 1.1 M_\odot$). The luminosity is calculated using a simple mass-luminosity relation:

$$\frac{L_*}{L_\odot} = \left[\frac{M_*}{M_\odot}\right]^4 \tag{D.3}$$

The stars' effective temperature can be calculated, assuming a blackbody:

$$T_* = \left[\frac{L_*}{4\pi R_*^2 \sigma_{SB}}\right]^{1/4} \tag{D.4}$$

As stars evolve along the main sequence, their luminosity increases (Schröder & Connon Smith, 2008). As the luminosity increases, the location of the stellar habitable zone must move further away from the star (Hart, 1979; Kasting, 1993). This implies that any planets with biospheres initially in the habitable zone can leave the habitable zone on a timescale τ_{HZ}, which is a function of the host star's initial luminosity and the planet's distance from it. Together with the main sequence lifetime τ_{MS}, it defines a maximum lifetime for any biosphere:

$$\tau_{max} = MIN(\tau_{MS}, \tau_{HZ}) \tag{D.5}$$

Where τ_{MS} is given by

$$\frac{t_{MS}}{t_{MS,\odot}} = \left[\frac{M_*}{M_\odot}\right]^{-3} \tag{D.6}$$

The luminosity evolution of the stars are approximated by extrapolating the simulated solar luminosity data of Schröder & Connon Smith (2008) to all main sequence stars[1]:

$$L(t) = \left(0.7 + 0.144\left(\frac{t}{Gyr}\right)\right)\frac{L_*}{L_\odot} \tag{D.7}$$

Reproducing Galactic Structure The star's age is sampled to reproduce the star formation history of the Milky Way (Rocha-Pinto et al. 2000a, see Figure D.2). The stars' galactic cylindrical polar coordinates (r, z) are randomly sampled to reproduce the thick and thin stellar discs (Ostlie & Carroll, 1996):

$$\rho(r, z) = n_0 e^{-r_{gal}/r_H}\left(e^{-z_{gal}/z_{thin}} + 0.02e^{-z_{gal}/z_{thick}}\right) \tag{D.8}$$

The metallicity gradient of the Milky Way is also emulated:

$$Z_* = -z_{grad}\log\left(\frac{r_{gal}}{r_{gal,\odot}}\right) \tag{D.9}$$

In truth, there are many differing measurements of the abundance gradient in the Galaxy (Rolleston et al., 2000), dependent on the metals studied. This reflects the different synthesis processes at work for differing elements. This can be (crudely) reproduced by allowing z_{grad} to have a distribution of values - in this case, a Gaussian distribution, with sample mean and sample standard deviation defined by the measurements of Rolleston et al. (2000).

Finally, measures have been taken to correlate the age and metallicity of the stars. The Age Metallicity Relation of Rocha-Pinto et al. (2000b) (with its errors) defines upper and lower bounds to the age of a star (given its metallicity).

[1] This may seem a weak assumption, but the stars of interest in these simulations will be close to solar type, so the approximation is reasonable in this first instance.

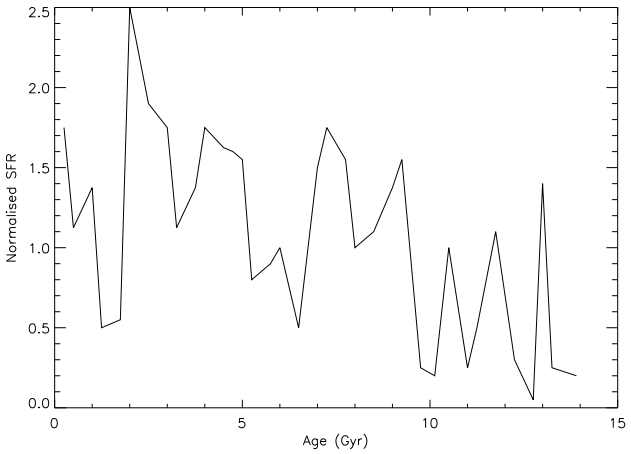

Figure D.2: The star formation history used in this work (Rocha-Pinto et al., 2000a).

Planetary Properties

Current exoplanet data, while impressive, is still incomplete. This introduces significant bias into the results of any simulation (Forgan, 2009), and currently precludes the simulation of Earthlike planets if using observations alone. To bypass this problem, the empirical data used by Forgan (2009) is replaced by theoretical relations: this allows the simulation of planetary objects down to lunar masses.

The probability of a star hosting planets is a function of its metallicity: this code uses the distribution as described by Wyatt et al. (2007):

$$P(z) = 0.03 \times 10^{\frac{Z}{Z_{\odot}}} \tag{D.10}$$

The planetary initial mass function is approximated by a simple power law:

$$P(M_P) = (M_P)^{-1} \tag{D.11}$$

which operates over the mass range of $[M_{moon}, 25\, M_{Jup}]$. To correctly reproduce the distribution of planetary radii, two different radii distribution functions are used. Jovian planets reproduce the data of Armitage (2007), which accounts for the effects of Type II planetary migration. For terrestrial planets, the data of Ida & Lin (2008) (Fig. 1c) is emulated: a simple parametrisation allows the trend for low mass objects to be recovered (see Figure D.3). It should be noted that in essence this is swapping one weakness for another: while the bias of empirical data is lost, the uncertainty of current planet formation models is gained.

Also, as the mass function can simulate Moon-mass objects, any object with a mass less than Pluto's that resides within another planet's Hill sphere (that is, it resides within the gravita-

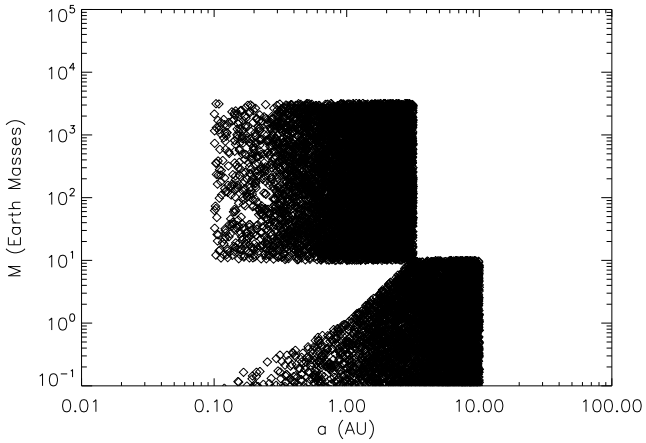

Figure D.3: The simulated mass-radius relation for a sample of planets.

tional influence of said planet) is considered a moon of that planet[2]. This property will become important for the simulation of the Rare Earth Hypothesis.

Biological Parameters

Life Parameters The model now enters the realm of essentially pure conjecture: all the available data for these parameters is derived from observations of a single biosphere, and hence there is little that can be done to constrain these parameters (at least without making wide-ranging assumptions about the mechanisms of life as a whole). The model implicitly assumes the "hard step" scenario of evolution (e.g. Carter 2008), i.e. life must achieve essential evolutionary goals in order to become intelligent organisms with the ability to construct sufficiently complex technological artifacts. The key biological parameters are listed as follows:

- N_{stages}: The number of stages life must evolve through to become intelligent
- τ_i: The time required for each of these stages to be reached
- τ_{int}: The time required for an intelligent technological civilisation to form from life's creation
- $P_{annihilate}$: The probability that a reset event will cause complete annihilation

The intelligence timescale τ_{int} is calculated by the following stochastic process: if life does evolve on a planet, N_{stages} is randomly sampled. The resetting events N_{resets} are placed uniformly throughout each of the stages. If $N_{resets} > N_{stages}$, then any given stage may suffer several reset events.

[2]This does not guarantee the observational rule of thumb that the probability of a terrestrial planet having a substantial moon is around 0.25. However, this is a reasonable first approximation, with more detailed studies of this issue requiring future observational input (e.g. Sartoretti & Schneider 1999; Kipping et al. 2009).

The effect of a reset is to reduce biodiversity by a fraction x: if x is greater than 1, the planet is sterilised. Appealing to the Central Limit Theorem, x is selected from a Gaussian Distribution with mean 0.5, and standard deviation 0.25. This means that on average, 5% of resets will result in annihilation. The average number of resets increases with proximity to the Galactic Centre - this is parametrised by:

$$\mu_{resets} = \mu_{Earth} e^{-\left(\frac{r_{gal}}{r_{gal,\odot}}\right)} \tag{D.12}$$

Where $\mu_{Earth} = 5$ (Raup & Sepkoski, 1982). This provides the mean of a Gaussian distribution from which N_{resets} is sampled.

The following procedure is used to calculate τ_{int}:

1. τ_i is sampled
2. If a reset occurs, test if that reset results in annihilation: if annihilation occurs, life is exterminated, and the process ends; otherwise, i decreases by 1.
3. τ_{int} is increased by some fraction of τ_i (or simply by τ_i if no resets occur)
4. Increase i by 1
5. If $i > N_{stages}$, then end the process; otherwise, return to 1.

This procedure continues until either a) life has reached the final stage (an intelligent technologically capable civilisation has evolved), or b) the intelligence timescale becomes greater than the maximum lifetime of the planet's habitability: $\tau_{int} > \tau_{max}$. Once a civilisation has formed, it is assumed that detectable signals or signal leakage begins to be emitted. This emission will continue until either (a) the civilisation destroys itself (see next section) or (b) the parent star moves off the main sequence (see next section). This in itself is a conservative estimate of the length of time a civilisation may be detected, as civilisations may in fact migrate between the stars, or use stellar engineering to prolong their parent star's life.

Civilisation Parameters Once a technologically capable civilisation has formed, it must move through a "fledgling phase": it is susceptible to some catastrophic event caused partially or fully by its own actions (e.g. war, plague, catastrophic climate change, botched macro-engineering projects). This is described by the parameter $P_{destroy}$: the probability that a fledgling civilisation will destroy itself. If a civilisation can survive this phase, it becomes sufficiently advanced to prevent such self-destruction events, and becomes stable, on a timescale τ_{adv}. If a civilisation is destroyed, then it will survive some fraction of τ_{adv} before destruction.

What advanced civilisations can then do is at the behest of the user: civilisations may colonise all planets within their solar system, resulting in signals appearing on all planets in that

system. Probes may be sent into the galaxy at large, which could define an explorable volume of the galaxy for a given advanced civilisation. Civilisations may even attempt to generate new biospheres on neighbouring planets - the "directed panspermia" model (Crick, 1973).

The key civilisation parameters are:

- τ_{adv}: The timescale for a civilisation to move from "fledgling" to "advanced".
- $P_{destroy}$: The probability that a fledgling civilisation will destroy itself.
- L_{signal}: The lifetime of any signal or leakage from a civilisation

The signal lifetime of a self-destroying civilisation is:

$$L_{signal} = \zeta \tau_{adv} \tag{D.13}$$

Where ζ is a uniformly sampled number between 0 and 1. If the civilisation becomes advanced, this becomes

$$L_{signal} = \tau_{max} - \tau_{int} \tag{D.14}$$

(i.e. civilisations exist until their parent star leaves the Main Sequence). For a planet colonised by an advanced civilisation, this is

$$L_{signal} = \tau_{max} - \tau_{int} - \tau_{adv} \tag{D.15}$$

At the end of any MCR run, each planet will have been assigned a habitation index based on its biological history.

$$I_{inhabit} = \begin{cases} -1 & \text{Biosphere which has been annihilated} \\ 0 & \text{Planet is lifeless} \\ 0.5 & \text{Planet has microbial life} \\ 1 & \text{Planet has primitive animal life} \\ 2 & \text{Planet has intelligent life} \\ 3 & \text{Planet had intelligent life, but it destroyed itself} \\ 4 & \text{Planet has an advanced civilisation} \\ 5 & \text{Planet has been colonised by an advanced civilisation} \end{cases} \tag{D.16}$$

If the planet has habitation index 0.5, 1 or 2, the biological process has been ended by the destruction of the parent star. Planets with $I_{inhabit} \geq 0.5$ may contain biomarkers in their atmosphere (e.g. ozone or water spectral features) which could be detected. Planets with an index of 2 or higher will emit artificial signals or signal leakage. Planets with an index of 4 or 5 may display evidence of a postbiological civilisation or of large scale macro-engineering projects, e.g. Dyson spheres (Dyson, 1960). Signals from these systems may even be consistent

with those expected from Kardashev Type II civilisations (those which can harness all the energy of their parent star, see Kardashev 1964), and hence could produce characteristic stellar spectral signatures that Earth astronomers could detect.

The Connectivity of Civilisation Pairs

As the code produces data pertaining to each individual civilisation, it is possible to study the entire dataset for each run, and identify the potential for communication between all possible pairs of civilisations. For N galactic civilisations, there are $N(N-1)/2$ pairs of civilisations. For each intelligent civilisation pair (ICP), the following outputs can be calculated:

1. Their physical separation in kiloparsecs (dx),
2. The available window of communication dt (that is, the maximum time interval where both civilisations exist and are able to communicate),
3. The space time interval $ds^2 = c^2dt^2 - dx^2$. This quantity determines whether a signal travelling at lightspeed can traverse the distance between two civilisations within the communication window (assuming the intervening space to be Minkowskian). If $ds^2 < 0$, then the signal will fail to reach its destination before the window closes. If $ds^2 = 0$, then the signal will reach its destination at the same instant the window closes. If $ds^2 > 0$, then the signal will reach its destination within the window, and it is therefore possible for communication between the two civilisations to be established.
4. The "contact factor"

$$f_{contact} = \frac{cdt}{2dx} \tag{D.17}$$

which counts how many "conversations" (pairs of signals) can travel between the two civilisations.

These outputs give extra information on the distribution of civilisations in the Galaxy, and their potential connectedness by signals travelling at lightspeed.

I have set out the numerical methods used in this appendix - I will now describe the research I have undertaken with them.

D.5 Implications for the Rare Earth Hypothesis

The attributes of the planet Earth are of critical importance to the existence and survival of life upon it. In fact, it may be so finely tuned that few planets in the Galaxy share its life-friendly characteristics. From this premise, it is almost inevitable to reach the conclusion that intelligent life (at least any that is predicated on evolution from complex metazoans) is

also rare - perhaps unique to the planet Earth.

These ideas have been encapsulated in what is known as the Rare Earth Hypothesis (Ward & Brownlee, 2000). It can be summarised thus:

1. Simple life may be commonplace in the Universe. The existence of extremophilic organisms in what were originally considered to be inhospitable regions (hydrothermal vents, acidic pools, toxic waste, deep in the Earth's crust) has shown the hardiness of simple life (Cavicchioli, 2002; Diaz & Schulze-Makuch, 2006). Indeed, these habitats are believed to be duplicated elsewhere in the Solar System - e.g. Mars (Formisano et al., 2004; Krasnopolsky et al., 2004), Europa (Carr et al., 1998), Titan (Stofan et al., 2007), Enceladus (Parkinson et al., 2007; Spencer & Grinspoon, 2007) - so it is still possible that "alien" life may be closer to home than once thought.

2. However, although simple life is resilient and adaptable, the evolution of complex animal life is extremely difficult. For this to be achieved, there are certain criteria (hereafter referred to as *the Earth Criteria*) that must be satisfied, in order for animals to thrive.

A (non-exhaustive) list of the Earth Criteria follows:

- A planet within a critical range of orbital radii - the "stellar habitable zone" (Hart, 1979; Kasting, 1993)
- A star within a critical mass range (large enough to push the habitable zone outside the planet tidal locking radius, and small enough to provide sufficient energy while avoiding UV exposure)
- A star located in a critical region of the Galaxy - the "galactic habitable zone" (Lineweaver et al., 2004)
- A planet within a critical mass range to maintain a suitable atmosphere
- A planet with a stable low eccentricity orbit (to avoid extreme temperature changes). This also requires a relatively large moon to provide axial stability (Waltham, 2004).
- A planet with sufficient raw materials to generate amino acids and proteins
- A planet with suitable atmospheric composition, in particular the production of atmospheric oxygen - initially produced by cyanobacteria in Earth's early history (Canfield, 2005)
- A planet with plate tectonic activity to regulate atmospheric composition and the balance of carbon (Bounama et al., 2007)
- The presence of Jupiter to control the rate of comet and asteroid impacts - although new details are emerging on the exact role of Jovian planets in this process (Horner & Jones, 2008a,b; Horner et al., 2009).

The weakness of this hypothesis rests in the (usually implicit) assumption that all the Earth Criteria are independent of each other. Taking Jupiter as an example: asking whether Jupiter exists or otherwise in the Solar System is not meaningful, as planet formation is a complex, non-linear process: every planet in the Solar System owes its formation to its surrounding environment, and therefore its planetary neighbours, through the dynamics of migration (Raymond et al., 2006; Paardekooper & Papaloizou, 2008) planet-planet scattering (Ford & Rasio, 2008; Raymond et al., 2009), resonances (Cresswell & Nelson, 2006), and other secular phenomena (e.g. Batygin & Laughlin 2008). Without Jupiter, the Earth as it is today may not have formed at all.

We shall now put our code to use, testing the effects of some of the Earth Criteria on the distribution of life and intelligence. While we know that some Earth Criteria are not independent of each other, as a first approximation we can test the importance of how stringent the conditions need to be for life to form.

D.5.1 Inputs

Two separate hypotheses were tested with this model. Each was subjected to 30 Monte Carlo Realisations (MCRs), with each realisation containing $N_* = 10^9$. This is of course two orders of magnitude short of the Milky Way's stellar content, but computational constraints prevented increasing N_* further. The interested will be able to multiply subsequent results by 100 to obtain an estimate of Milky Way figures. In any case, absolute numbers are less relevant to the issue at hand: this study focuses on comparing two hypotheses, and comparing *relative trends* (which is a more reliable route in studies of this nature).

The Baseline Hypothesis

This basic hypothesis requires only that a planet must be in the stellar habitable zone for life to form upon it. If the planet's surface temperature lies between $[0, 100]^\circ C$, then microbial life can form upon it. Complex animal life will only form if the planet's surface temperature lies between $[4, 50]^\circ C$ (Ward & Brownlee, 2000). This hypothesis was tested to provide a comparison with the results of the Rare Earth Hypothesis.

The Rare Earth Hypothesis

This hypothesis builds on the baseline by also requiring that *animal* life will only form on a planet if the following four conditions are met:

1. The planet's mass is between $[0.5, 2.0]M_\oplus$,
2. The star's mass is between $[0.5, 1.5]M_\odot$,
3. The planet has at least one moon, (for axial stability and tides)

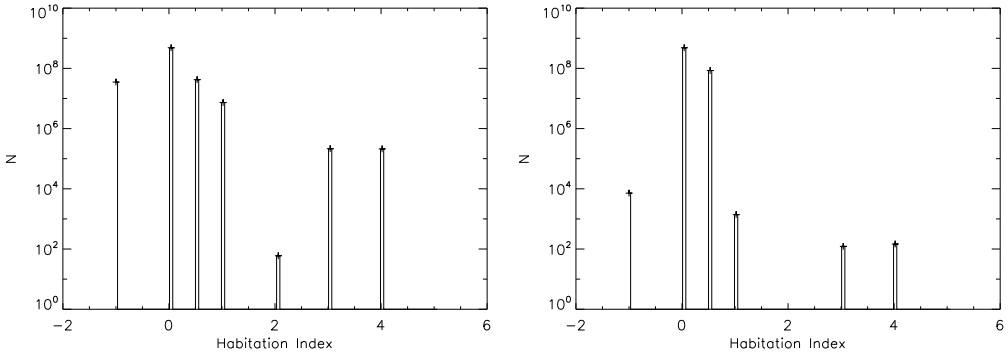

Figure D.4: The habitation index for the Baseline Hypothesis (left) and the Rare Earth Hypothesis (right).

4. The star system has at least one planet with mass $> 10M_{\oplus}$ in an outer orbit (for shepherding asteroids)

D.5.2 Results & Discussion

The Distributions of Life and Intelligence

The properties of the Rare Earth hypothesis galaxy can now be compared against those of the Baseline hypothesis. Comparing the habitation index for both hypotheses (Figure D.4), it can be seen that (by construction) microbial life ($I_{inhabit} = 0.5$) is unaffected by the Rare Earth hypothesis, whereas the prevalence of animal life ($I_{inhabit} = 1$) is reduced by a factor of 10^4 against the baseline. This reduction is thus propagated into the intelligent biospheres ($I_{inhabit} > 2$). However, despite some quite stringent conditions on the planetary system architecture (conditions 3 and 4, section D.5.1) the number of intelligent biospheres numbers in the hundreds: the implications for SETI are discussed in the next section.

As stellar mass is a key condition to the Rare Earth Hypothesis, it should be expected that the two hypotheses' distributions diverge, and this is indeed the case: Figure D.5 shows the distribution of stellar mass for both hypotheses. The IMF is modified by the effects of the habitable zone (and the distribution of exoplanet semi-major axis) to give the characteristic bump between 1 and 2 solar masses. Comparing the hypotheses shows that although the Baseline hypothesis favours lower mass stars for intelligent biospheres (for their increased longevity), the Rare Earth Hypothesis must discard the substantial number of stars that are less than $0.5M_{\odot}$.

This bias towards lower mass should be reflected in the distribution in semimajor axis (as lower mass stars have closer, more stationary habitable zones). Figure D.6 shows that this is true for the Baseline Hypothesis (with intelligent biospheres dropping off as $R > 1.5\,AU$), and

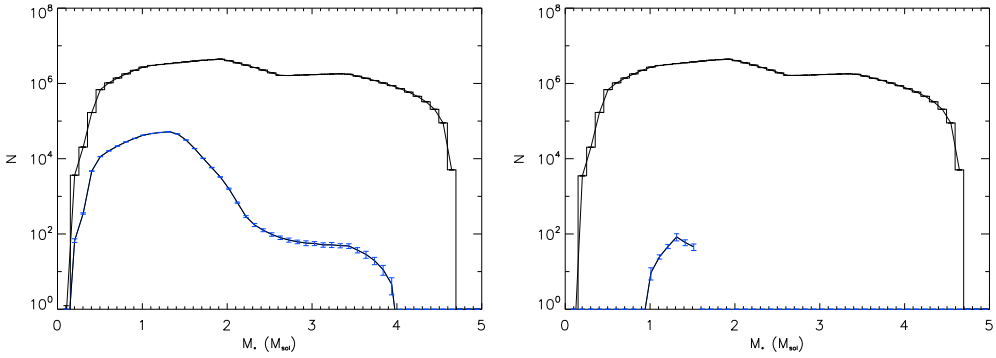

Figure D.5: Distribution of Stellar Mass for the Baseline Hypothesis (left) and the Rare Earth Hypothesis (right). The black lines indicate all biospheres, the blue lines indicate all intelligent biospheres.

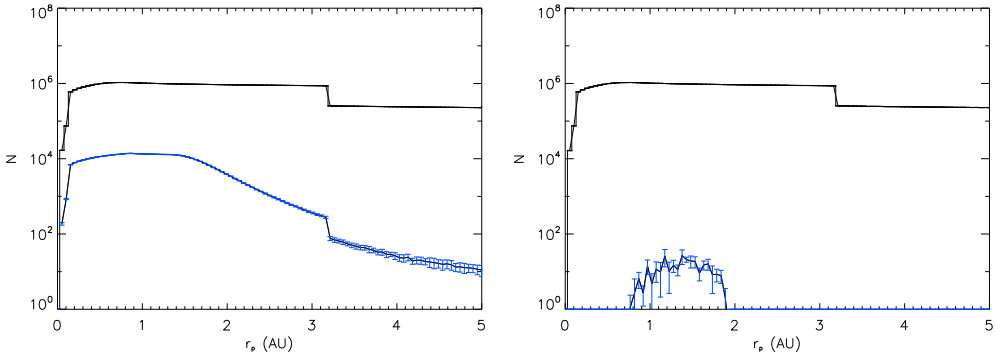

Figure D.6: Distribution of planet semimajor axis for the Baseline Hypothesis (left) and the Rare Earth Hypothesis (right). The black lines indicate all biospheres, the blue lines indicate all intelligent biospheres.

doubly true for the Rare Earth Hypothesis, selecting a narrow radial range between $[0.8, 1.9]$ AU.

The most striking difference can be seen in the distribution of galactocentric radius (Figure D.7). While the Galactic Habitable Zone (GHZ) can be identified in the Baseline Hypothesis (with a small contingent at lower radii, which presumably exists due to the lack of modelling of the central supermassive black hole (SMBH) and hypervelocity stars in the inner regions), the Rare Earth Hypothesis appears to have no GHZ. This is unexpected: the four conditions of the Rare Earth Hypothesis tested here do not affect where intelligent systems should lie; why then does the GHZ not appear (with reduced numbers)?

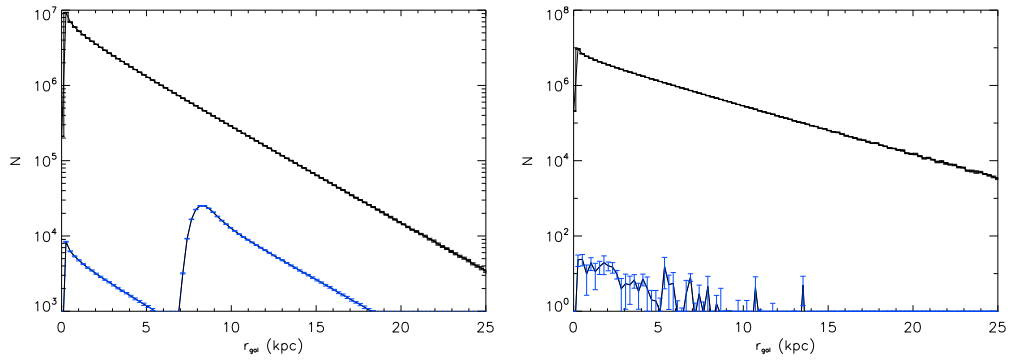

Figure D.7: Distribution of galactocentric radius for the Baseline Hypothesis (left) and the Rare Earth Hypothesis (right). The black lines indicate all biospheres, the blue lines indicate all intelligent biospheres.

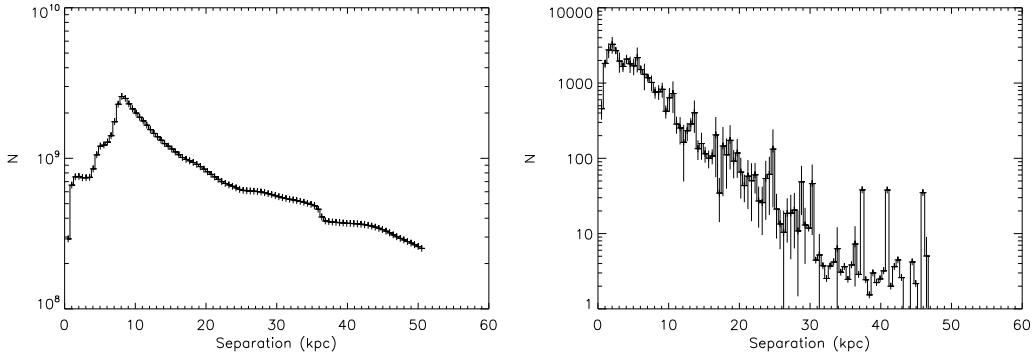

Figure D.8: Separations of ICPs for the Baseline Hypothesis (left) and the Rare Earth Hypothesis (right).

Communication and Connectivity

The *prima facie* conclusion (having studied the results of the previous section), is that if the Rare Earth Hypothesis is correct, and intelligent civilisations are infrequent, then the potential for communication is also low. This expectation can be tested by calculating the interaction variables discussed previously. As the focus has now shifted from individual intelligent civilisations to intelligent civilisation pairs (ICPs), the numbers duly increase from N to $\frac{N(N-1)}{2}$. Figure D.8 shows the distribution of ICP separation dx for both hypotheses. The baseline hypothesis exhibits a sharp peak at around 8 kpc (the location of the GHZ), accompanied by a long decay. This distribution is reminiscent of the lognormal distribution expected if the tools of the Statistical Drake Equation were applied (Maccone, 2009). The distribution reaches its mode in steps: these steps are presumably sensitive to the local galactic spiral structure. The Rare Earth Hypothesis has no apparent GHZ, so the distribution (though reduced in magnitude) peaks at a much lower 3 kpc.

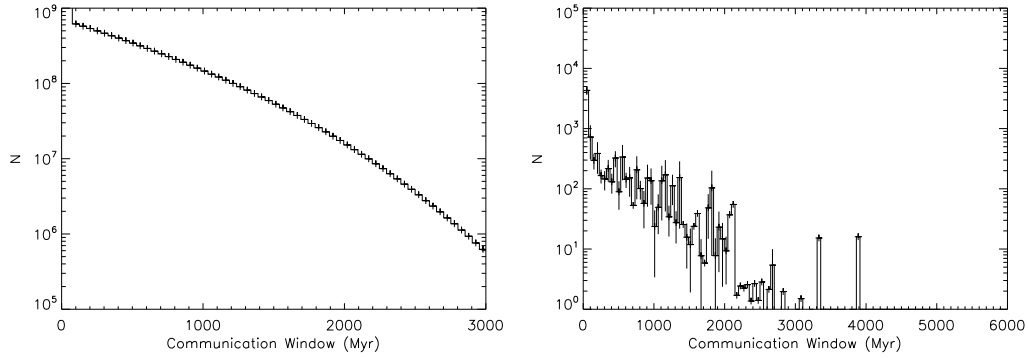

Figure D.9: Communications Window (maximum time interval for communication) for ICPs in the Baseline Hypothesis (left) and the Rare Earth Hypothesis (right).

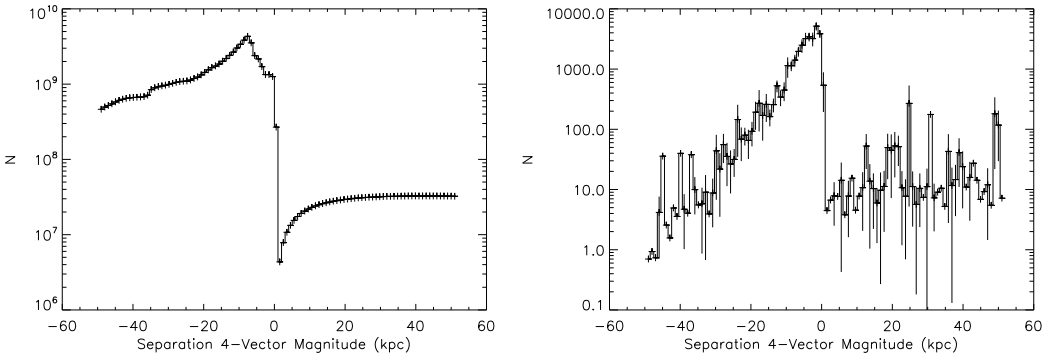

Figure D.10: Space-time interval (ds^2) between ICPs for the Baseline Hypothesis (left) and the Rare Earth Hypothesis (right). Unconnected ICPs have $ds^2 < 0$, connected civilisations have $ds^2 \geq 0$.

Does this reduced separation imply increased connectivity? The answer depends on the communication window for the ICP. The longer the window (i.e. the larger overlap in history where both civilisations exist), the longer the separation can be while allowing the ICP to be connected. The Baseline Hypothesis favours shorter communication windows (Figure D.9), which reduces the connectivity. Apart from small fluctuations at larger values, the Rare Earth Hypothesis agrees. When considering the space-time interval ds^2 (Figure D.10), the reduced connectivity becomes apparent. ICPs that are unconnected (negative values) are much more frequent than connected civilisations (positive or zero values). This does not spell the end for SETI, however, when the contact factor (i.e. the number of conversations) is considered: although few ICPs enjoy the privilege of contacting each other, those that do can expect a great deal of conversation (Figure D.11), where each hypothesis agrees that a select few will enjoy potentially thousands of exchanges with other civilisations.

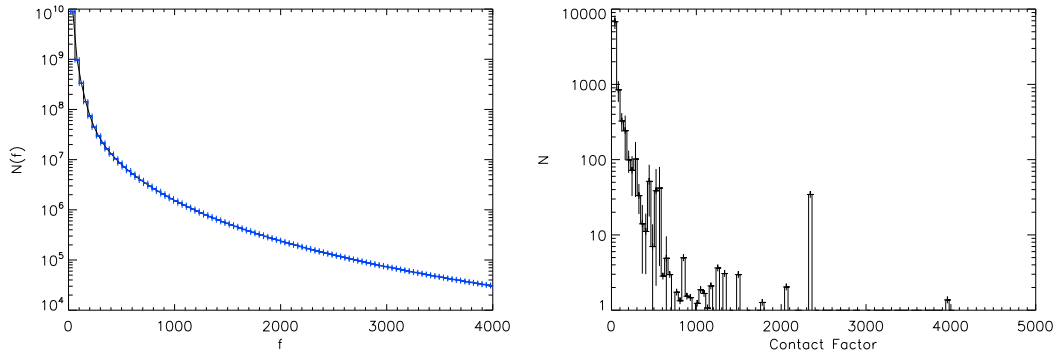

Figure D.11: Contact Factor (number of exchanged signal pairs) for ICPs for the Baseline Hypothesis (left) and the Rare Earth Hypothesis (right).

D.5.3 Conclusions

I have described numerical testing of the Rare Earth Hypothesis (Ward & Brownlee, 2000) using Monte Carlo Realisation techniques. By comparing the results to a baseline hypothesis, the influences of the criteria for a planet to be officially designated as "an Earthlike planet" can be studied. In this work, the criteria were limited to planet mass, star mass, the presence of a moon, and the presence of a Jupiter type object in a more distant orbit. It is shown that these criteria alone greatly reduce the number of intelligent civilisations in the Galaxy (compared to the baseline). As expected, the stellar mass criterion results in a narrow range of planet semi-major axes where intelligent biospheres exist. Interestingly, the Galactic Habitable Zone, apparent in the Baseline Hypothesis, was not visible in the Rare Earth Hypothesis.

This result is important for civilisation connectivity: reducing the civilisation separation means that, for a given time interval of communication, civilisations under the Rare Earth Hypothesis are able to exchange more signals than civilisations under more contact-optimistic hypotheses. The implications for SETI are somewhat mixed: while Earth may be much more likely to be a disconnected than connected civilisation, if it is connected, it can expect substantial conversation from other civilisations (while the Sun remains in the Main Sequence). Therefore, the Rare Earth Hypothesis (in the formulation described in this work) is a "soft" or "exclusive" hypothesis (using the nomenclature of Brin (1983); Cirkovic (2009)), in that it is not a completely contact-pessimistic hypothesis, but one that is contact optimistic for a small subset of civilisations in the Galaxy.

D.6 Intelligent Civilisations and the Second Law of Thermodynamics

According to recent theory, the Second Law of Thermodynamics could prove to be essential for understanding biological evolution. The mathematical analysis of the importance of the

Second Law in evolution is stated in Kaila & Annila (2008):

"The second law of thermodynamics is a powerful imperative that has acquired several expressions during the past centuries. Connections between two of its most prominent forms, i.e. the evolutionary principle by natural selection and the principle of least action, are examined. Although no fundamentally new findings are provided, it is illuminating to see how the two principles rationalising natural motions reconcile to one law."

The equation of evolution including entropy can be used Jaakkola et al. (2008) to explain the differences in the genome, as a consequence of the second law of thermodynamics. Others take the problem to a macroscopic scale and demonstrate that the same technique can explain the species-area relationship, one of eco-biology's key problems (Würtz & Annila, 2008). Indeed, it is possible that the universal criterion for evolutionary selection is the entropy principle (Sharma & Annila, 2007; Jaakkola et al., 2009). Taking into account the reach of the law of entropy over various scales, we now summarise what we name The Entropy Hypothesis (Bozhilov & Forgan, 2010):

The intelligent technological civilisations are a typical (although not guaranteed) consequence of the biological evolution of complex life forms, provided the necessary conditions are met. This is due to the efficiency of technological civilisations at increasing the entropy of their planetary system on very short timescales, satisfying the 2nd Law of Thermodynamics. The destruction of such a technological civilisation, which may be inherent in their evolution, will in general be the most effective way for biological evolution to fulfil the law of Entropy. So, whenever the conditions for evolution of complex life forms towards intelligence are met, an intelligent technological civilisation will appear, constantly evolve technically until it is self-destructed, colonised by another civilisation, or starts colonising space itself, thus ensuring the increase of entropy on even larger scales.

What is the basis for this hypothesis? Life and living organisms in a closed ecosystem decrease entropy. However, the organisms that survive are those that absorb the free energy as effectively as possible (Sharma & Annila, 2007; Jaakkola et al., 2009; Fu, 2007). Still, the entropy can be increased globally, if there is a way to alter ecosystems on large size scales and short timescales.

For the approx. 10^5 years since the emergence of *Homo sapiens*, Mankind has developed technology that can affect the Earth globally (in particular through the construction of buildings and deforestation destroying habitats). Much of Earth's surface has been altered in a short amount of (cosmic) time. If our technology continues to evolve and/or a technological breakdown occurs, vast amounts of the planet could easily be destroyed or contaminated.

From technology, therefore, a sociological pressure is derived: intelligent life is capable of self-destruction. This is true for Mankind, and we assume it to be true for other technological civilisations; we argue (Bozhilov & Forgan, 2010) this is a general effect of the second law of thermodynamics on macroscopic scales.

As an efficient entropy generator, we could suppose technology is connected to and would evolve shortly after the development of an intelligent species on a given planet, giving a natural mechanism for rapidly increasing the entropy on a planetary scale, as an extension of the law of entropy guiding natural selection. This would suggest that the intelligent technological civilisations could no longer be thought of as an exceptional or even rare event in biological evolution. That is not the only avenue which evolution can take, but its effect on increasing the entropy (we argue) may favour it over other potential avenues. Thus, we have theoretical expectations on the routes evolution can take on other biospheres we might find in space, other than Earth. On planets that are Earthlike, we expect that while the initial biochemistry of life may be very different from Earths, the selection pressures introduced by the environment will be similar, possibly resulting in convergent evolution (e.g. Morris 2006).

Furthermore, we can try to elaborate a definition of intelligence in the framework of the second law of thermodynamics. Human evolution might be regarded as survival of the fittest replicators, where by replicators we denote organisms, devices or even concepts (e.g. memes, Dawkins 1990) that can reproduce themselves and are subject in some form to natural selection.

The primary (biochemical) replicators are the genes, or more specifically RNA and DNA. As has been argued by previous authors (ibid.) the replicators that survive during natural selection are namely the ones that absorb free energy most effectively, i.e. natural selection is directly interrelated with the entropy principle. We can identify technology and culture as crucial milestones in the development of Man as a sentient species. These can be thought of as replicators that have been artificially synthesised by humans, or (to take the neo-Darwinian view) the genes themselves. Thus, given the entropy hypothesis of biological evolution, we can make a tentative definition of intelligence:

Intelligence is the process by which replicators artificially synthesise a radically new and fundamentally different type of replicator.

As replicators are subject to natural selection (and therefore the entropy principle by extension), this definition encourages us to regard intelligence as a standard effect in evolution, which arises in order to ensure the 2nd law of thermodynamics holds on macroscopic scales. This definition can also be used in other disciplines such as sociology, economics and biology, allowing us to gain new insights and deepen our understanding of intelligence as a natural paradigm (see also Kaila & Annila 2008).

Intelligence defined in this way is not restricted to biological replicators alone. Other replicators also have the potential to satisfy our criterion for intelligence, such as machines, provided these replicators synthesise others without guided supervision (e.g. without intervention of external intelligent observers). Indeed, it suggests that for machines to become truly intelligent, they must be sufficiently developed to become true replicators subject to the laws of natural selection.

Once an intelligent civilisation has arisen, it can either self-destruct, or alternatively, provided it survives long enough and develops the necessary technology, it can begin colonising other nearby planets, thus still increasing entropy at a maximum possible rate, by changing and/or destroying another worlds. However, to simplify this analysis, colonisation effects will not be considered in this work.

In brief, in the frame of current results, it seems plausible that the law of entropy could be a primary cause for the development of intelligent species and a key factor for the advent of technological civilisation, regarded as a natural mechanism, assuring the quickest possible entropy rate increase in a given planetary ecosystem.

D.6.1 Inputs

The Baseline Hypothesis

We use the same baseline hypothesis as for the Rare Earth Hypothesis study. If the planet's surface temperature lies between $[0, 100]^{\circ}C$, then microbial life can form upon it. Complex animal life will only form if the planet's surface temperature lies between $[4, 50]^{\circ}C$ (Ward & Brownlee, 2000).

The Entropy Hypothesis

We now introduce the effect predicted by the entropy hypothesis:

$$P_{destroy} = 1.65 \times 10^{-3} e^{\frac{t_{adv}}{0.056}} \tag{D.18}$$

This is done so that $P_{destroy}$ ranges from 0.01 to 0.9 across all possible t_{adv} values (where t_{adv} is the timescale for a civilisation to move beyond its fledgling stage, and escape self-destruction). t_{adv} is selected from a Gaussian with mean 0.25 Gyr, and standard deviation of 0.1 Gyr. This approach represents the sociological pressure, marking the possibility of destruction (the preferred maximum entropy state) becoming more probable with technological advance. By comparison, the baseline sets $P_{destroy} = 0.5$ for all civilisations, reflecting ignorance as to the sociological issues of each individual civilisation.

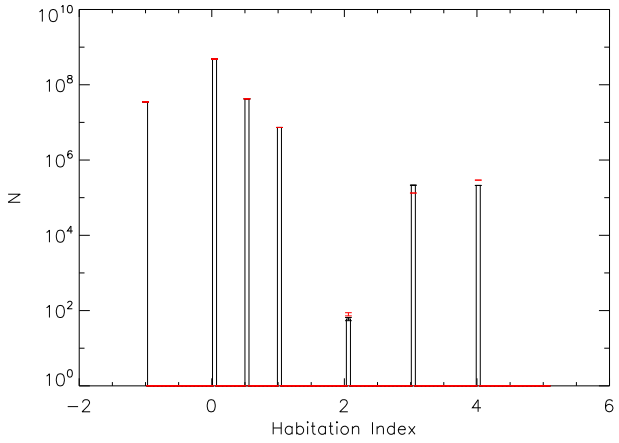

Figure D.12: Habitation index for the Entropy (red) and The Baseline (black) Hypothesis.

D.6.2 Results

The breakdown of habitation index for both hypotheses is shown together in Figure D.12. No significant change is observed: most of the planets in the simulation never develop intelligent life ($-1 \leq I_{inhabit} \leq 1$). The planets with index $I_{inhabit} = 2$ i.e. the planets with young technological civilisations or fledgling civilisations, are relatively the same number (around 102). There is a small change in the number of fledgling civilisations which destroy them-selves ($I_{inhabit} = 3$). The Entropy hypothesis produces a smaller number of self-destructed young ETIs, which gives rise to a higher number of advanced ($I_{inhabit} = 4$) technological civilisations instead. As expected, the Entropy hypothesis as characterised in section 3.2 affects only the latter stages of evolution and technological development (which seems to be more prevalent against the baseline). This is in concordance with the expectation that the ETIs evolution is entropy-driven, which results in a innovate or die sociological pressure.

The Entropy Hypothesis does not speculate on the location of life in the Galaxy, so it is expected to match the results of the baseline. This is indeed the case: there is no change in the galactocentric radius of the planets for both hypotheses, as can be seen in Figure D.13. Both hypotheses reproduce as expected the Galactic Habitable Zone (Lineweaver et al., 2004) at around 8 kpc, demonstrating the balance between Galactic chemical gradients and potentially sterilising astrophysical phenomena.

We can define the signal history of the Galaxy as the total number of communicating civilisations as a function of time. Will this history identify the influence of entropy's sociological pressure on ETIs? Figure D.14 shows that the Entropy hypothesis gives a slight enhancement against the baseline; the increase in advanced civilisations means their signal lifetime is longer, enhancing their number N over a more significant period of cosmic time. However, this

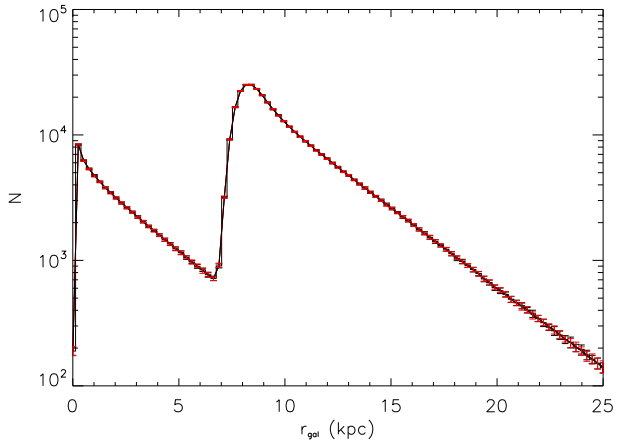

Figure D.13: Distribution of galactocentric radius for the Entropy (red) and Baseline (black) hypothesis.

enhancement is within the error bars of the baseline measurement. This would suggest that confirming the entropy hypothesis by observing N alone would be hazardous if not impossible (not least because N is currently measured as 1).

If the increased signal lifetime is responsible for this slightly enhanced signal history, then this should be quantified. The lifetime of the emitted signals for each hypothesis is shown in Figure D.15. Although there is a significant increase in the number of signals, this occurs only to signals with lifetime around 1-3 Gyr. No significant change is observed for longer times.

It is difficult to place any constraints on SETI based on studies of this nature: analysing the connectedness of these civilisations is the most concrete means at our disposal. Figure D.16 displays the results of the contact factor. Most of the simulated planets, even if they host ETI, are disconnected: either the distance is too great, or the time interval in which the pair co-exists is too short. The Entropy hypothesis tends to produce civilisations that are more connected, with a higher contact factor (this is to be expected if the Entropy hypothesis produces more advanced civilisations on average than the baseline).

Note that the contact factor is dependent on the space separation between each inhabited planet and the time interval for each civilisation pair. The entropy hypothesis yields changes only to the time interval, but not to the physical separation, i.e. the distance between the host planets of alien civilisations. Also, our estimates of connectivity make no assumptions about the methods by which communication is established. Traditionally SETI has favoured radio emission, but this particular method may restrict connectivity if the civilisations are human-like (Forgan & Nichol, 2010). It is plausible that civilisations (including ours) will employ other communication techniques, e.g. EM radiation at other wavelengths, or more

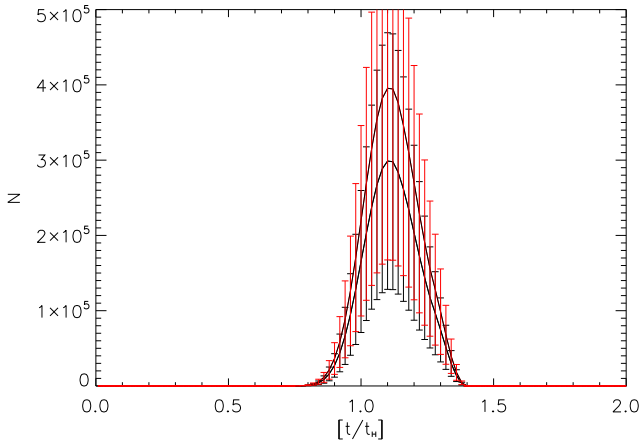

Figure D.14: The signal history of the Milky way, normalised by the Hubble time t_H. $t = t_H$ represents the present day. The baseline hypothesis is shown in black and the entropy hypothesis in red, along with the proper error bars.

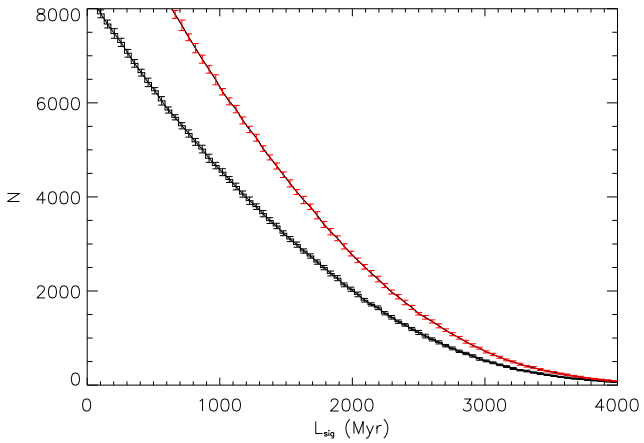

Figure D.15: The lifetime of emitted (leaked) ETI signals, relative to the number of signals. The Baseline hypothesis is shown in black and the Entropy hypothesis in red. The count of the longest long-living signals (over than 4 Gyr), which are important for communication over very long distances (naturally of most interest to SETI researchers) remain unchanged. A considerable increase in number occurs in the signals with shorter lifetimes (1-2 Gyr).

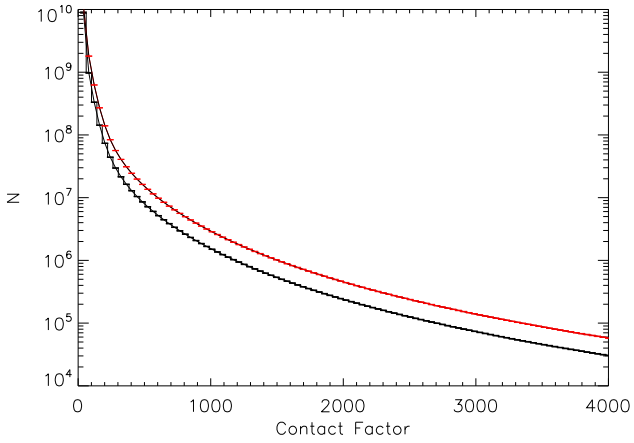

Figure D.16: Results of the contact factor (measuring the number of successfully exchanged signals) are plotted. The baseline hypothesis is shown in black, the entropy hypothesis in red. For most of the signals, no change observed. The significant increase in successfully transmitted communication occurs only to a small number of pairs at high contact factors

exotic communications methods based on less attenuating phenomena like neutrinos (Learned et al., 1994; Silagadze, 2008). However, increasing connectivity by this method will require more free energy, perhaps discouraging civilisations to communicate when resources become scarce (Cirkovic, 2008).

D.6.3 Discussion

The entropy principle is connected with the actions and the sociological behaviour of ETI, so it becomes decisive only at a later evolutionary stage e.g. after technology is developed. Note that in this analysis we implicitly suppose that intelligent species always discover technology if they survive long enough. The actual behaviour of technological civilisations might differ from our expectations.

In fact, intelligent civilisations may behave more like viruses, expanding and optimising their free energy consumption to attain a steady state if there are sufficient populations of hosts (i.e. biospheres) available (Starling & Forgan, in prep.). There is a well-defined period in which the technological civilisations are highly susceptible to destruction (the "fledgling period" in our numerical simulations where civilisations progress from $I_{inhabit} = 2$ to $I_{inhabit} = 3$ or $I_{inhabit} = 4$). Nevertheless, the tendency towards expansion (i.e. consuming free energy available) will ultimately lead the civilisation first to harvest the energy of its planet, then the energy of the host star and finally - maybe even the energy of entire galaxy. Thus, by the use of the entropy principle, we motivate theoretically the Kardashev scale for classifying technological civilisations based on their energy consumption (Kardashev, 1964).

The numerical results exhibit a clear tendency of favouring the evolution of increasingly advanced technology (see Figure D.12). The Entropy hypothesis speculates this could be due to the inherited effect of the entropy principle: the ETIs are bound, if possible and under the right environmental conditions, to discover and evolve technology as a new source of increasing complexity, until either self-destruction or progression to a stable, advanced state occurs. Thus, we state that intelligent species may be regarded as the most effective way of securing the fast and constant increase of the rate of entropy, considering a given planetary ecosystem.

However, given the observational data available, there is no current means of proving or disproving the entropy hypothesis. The potential for future radio telescopes such as the Square Kilometer Array (SKA) to map out artificial signals is limited to a region less than 300 light years from the Earth (Loeb & Zaldarriaga, 2007), and the combined decay of signal leakage from the Earth suggests that even the SKA will only be efficient at detecting purposeful "beacons" (Forgan & Nichol, 2010). Even in the distant future, when a more detailed Galactic census is available to us, we may not be able to distinguish the effects of entropy-driven evolution from a more neutral hypothesis.

Still, if intelligence could be reasoned as a normal stage in evolution, as expected by the entropy hypothesis, there might well be a larger number of emitted or successfully exchanged signals between the different ETIs. The results demonstrate that although there is indeed a minor increase in expected civilisations in the present day (Figure D.14), it will be difficult to observe this increase (or the increased connectivity suggested in Figures D.15 and D.16). While the entropy mechanism provides an appealing explanation of the emergence of intelligent species, current SETI-type observations will not provide sufficient evidence to be able to establish its veracity.

Conversely, the entropy hypothesis provides an (often-cited) answer to the Fermi paradox - we do not see alien life, because ETIs tend to have a shorter lifetime than we might naively expect. In addition, by incorporating the second law of thermodynamics into recent astrobiological analysis, we may be able to uncover more details on the origin of civilisation, and its interaction with Earth's ecosystem.

D.7 The Efficacy of Single Waveband SETI with the SKA

D.7.1 Introduction

It is generally believed that communications from an extraterrestrial intelligence will come in two possible forms. The first would be a civilisation more advanced than ours that has the technology and power to broadcast signals across the Galaxy, specifically for others to detect (this is a "beacon"). The second would be technologically younger civilisations, like ours, that are just beginning to use advanced communications, which then leak into space for others to

eavesdrop on. As outlined by Penny (2004), radio frequencies are believed to be the most natural place to look for both types of these signals and therefore, a majority of past, present and future searches for extraterrestrial intelligence are in the radio (or microwave) range of the electromagnetic spectrum.

The most ambitious search so far for extraterrestrial beacons was the Phoenix Project, which ran for nearly ten years (from 1995 to 2004), and observed 800 stars (out to 240 light years from earth) with Arecibo, Parkes and the Green Bank radio telescopes (over a frequency range of 1.2 to 3 GHz). Alternatively, the BETA project used a 26m radio telescope to perform an all-sky, narrow-band, microwave search for extraterrestrial beacons in the so-called "water hole" between 1400-1720 MHz. This range corresponds to a gap in the radio noise spectrum coming from space located between the hydrogen line and the strongest hydroxyl line (water), and it is believed an intelligent race would pick this region of the electromagnetic spectrum to broadcast their beacons. For both experiments, no unusual radio signals were detected and limits were placed on the possible strength of the beacons being broadcast.

A new era for the search for extraterrestrial Intelligence (or SETI) has begun with the construction of the Allen Telescope Array (ATA), built to continue the search for "beacons" across the Galaxy. The first 42 dishes of the ATA have already begun scientific operations, with a further 300 dishes planned. The ATA is targeting $\sim 250,000$ stars (including stars with known exoplanets) in the "Water Hole" looking for alien beacons, while also doing a deep blind survey (20 square degrees) towards the Galactic centre looking for strong beacons from billions of stars in that direction.

In addition to the dedicated ATA, other radio telescopes are being built across the world, which could also facilitate SETI. For example, the Low-Frequency Array (LOFAR) in Europe has just started exploring the Megahertz regime of the radio spectrum in search of the red-shifted 21cm line of neutral hydrogen from the Epoch of Re-ionisation. Ironically, LOFAR will encounter significant Radio Frequency Interference (RFI) from modern human activity (radar, mobile phones, radio and TV stations, etc.), and dedicated SETI observations have tended to avoid these radio frequencies because of this problem (working at Gigahertz frequencies instead).

Recently, Loeb & Zaldarriaga (2007) have highlighted the irony of this approach and noted that the increased sensitivity of new radio arrays like LOFAR could eavesdrop on the RFI produced by distant extraterrestrial civilisations. The attraction of this idea is simple: First, the human race is not broadcasting a radio beacon (at Gigahertz frequencies) for other civilisation to detect as the power required is prohibitive. Secondly, we are radiating significant radiation from everyday activities in the Megahertz region of the electromagnetic spectrum, so we know of at least one (advanced) civilisation that could, in theory, be detected at these radio frequencies (humans). Finally, SETI could easily "piggy-back" on the existing operations of LOFAR and other radio arrays.

In their paper, Loeb and Zaldarriaga focused on the radio emission potentially produced by

human-like military radar, which is one of the most powerful sources of radio "leakage" into space from Earth. The radar employed by the US Ballistic Missile Defense System (BMDS) can generate isotropic radiation with a total power of 2 gigawatts, or two orders of magnitude higher if beamed. Likewise, over-the-horizon radar, which bounces signals off the ionosphere, can reach similar power output (Tarter, 2004). Using such signals as a blueprint for possible extraterrestrial radio emission, Loeb and Zaldarriaga estimated that LOFAR could detect civilisations like ours out to a distance of 50 parsecs with a month of observation (see Figure 1 of their paper). This volume of the Galaxy contains $\sim 10^5$ stars and several possible rocky exoplanets, e.g., Gliese 581c (Udry et al., 2007). Beyond LOFAR, plans are already underway for the construction of a Square Kilometer Array (SKA) radio telescope, which Loeb & Zaldarriaga show could see human-like radio signals to ~ 200 pcs (in one month of observation).

However, we should think carefully about what we might expect the SKA to see. While humans are still leaking radio emission into the Galaxy, the extent of this emission has diminished. Technological improvements have reduced the transmission power required to broadcast, and the dawn of the digital age has begun to supersede traditional radio entirely. These events have occurred in just over 100 years, putting us on the path to becoming a "radio quiet" civilisation. If the Biological Copernican Principle is true (i.e. humans are not atypical as intelligent species), then what happens if *all* civilisations rapidly become radio quiet?

In this section, we expand on the ideas of Loeb & Zaldarriaga (2007) to assess the likelihood of eavesdropping on the radio emission from other advanced civilisations. To achieve this, we use the latest statistical model for likely habitable planets in the Galaxy (see Forgan 2009 for details) combined with an estimation for the likely life-time of radio leakage into space. Together, these provide a probability for being close enough to an advanced civilisation, which is broadcasting, for the SKA to detect it.

D.7.2 Numerical Methods

We again turn to the connectivity of civilisations, as described in section D.4. To study the effect of our SKA constraint on connectivity, two sets of calculations are required to compare relative trends. Both calculations will use the same ensemble of civilisations generated using a relatively optimistic hypothesis of life, merely that life will arise on any planet within the stellar habitable zone (i.e., the zone around each star where the temperature is right for liquid water to be on the surface of any planet). This ensemble corresponds to the results of the Baseline Hypothesis (described above).

A total of 30 MCRs were run for this hypothesis, giving a mean signal number of $N \sim 5\times10^5$. This provides an *ab initio* optimistic data set: these simulations have a substantial population of civilisations existing at a similar time in cosmic history (Figure D.17), offering us the best chance to communicate with ETIs, either by accident (eavesdropping) or using beacons.

The first calculation of connectivity will use "unconstrained" values of dx and dt to calculate the contact factor (i.e. there is no maximum distance over which civilisations can communicate,

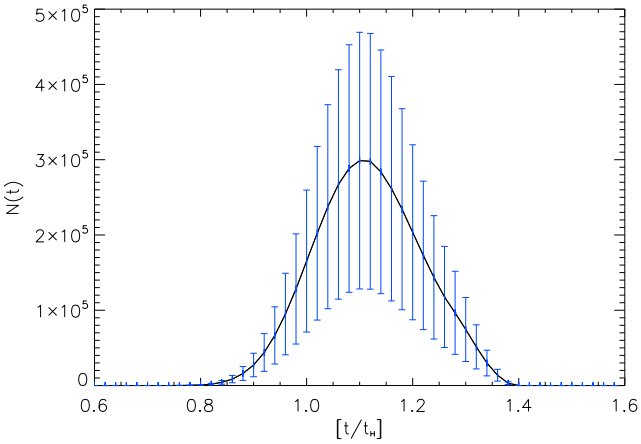

Figure D.17: The signal history of the Baseline Hypothesis. The time axis is scaled in units of the Hubble Time (i.e. $t = 1t_h$ indicates the present day). The black curve represents the mean from 30 individual Monte Carlo Realisations (MCRs) - the blue error bars indicate the standard deviation.

and the time limit in which civilisations can communicate is the maximum available time, i.e. the communications interval). These civilisations therefore use a variety of different instruments and techniques to conduct their search at various wavelengths and energy scales.

The second calculations will use "constrained" values of dx and dt, to emulate the constraints imposed by observing human-like radiation with an SKA-like instrument. More rigorously, for this "eavesdropping" to take place, we require $dx \leq 100$ pc and $dt \leq 100$ yr. Therefore, civilisations that are separated by more than 100 pc can not be seen (in a month of observation), and we impose a limit of 100 years on the timescale for "leaking" radio emission into space. This latter constraint is based upon likely human trends towards more directed (energy efficient) modes of communication (e.g. high-speed internet via optical fibres). We thus assume humans will cease using radio communications in ~ 100 years and will no longer be visible in the radio regime of the electromagnetic spectrum.

D.7.3 Results

In the unconstrained case, it can be seen that communication is relatively straightforward. Figure D.18 shows that a substantial number of civilisation pairs can have a significant number of conversations (represented by a large f value) before the communication interval ends. While shorter contacts (low f) are in general favoured over longer contacts (high f), there remains plenty of opportunities for ETI to communicate.

In the constrained case, all communication disappears (i.e. the maximum value of $f = 0$). The combined space-time constraints placed on the civilisations by advancing technology and the spatial limitations of the SKA are extremely strong. Civilisations are typically separated by distances much greater than 100 parsecs, and are therefore unable to communicate. Civilisations

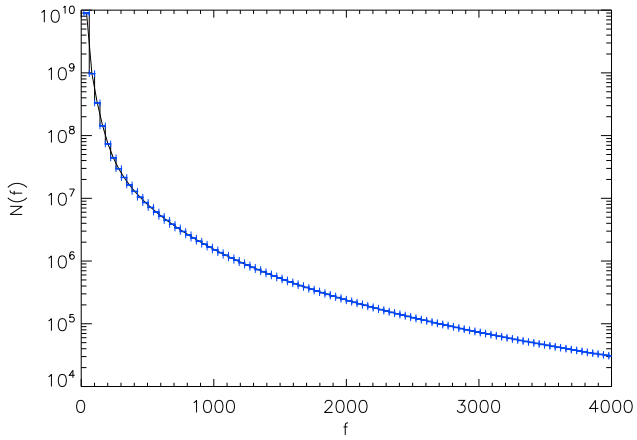

Figure D.18: Distribution of contact factor for the civilisations in the Baseline Hypothesis, assuming they are radio loud for their entire existence, and there are no constraints on the separation of communicating civilisations (except the constraints imposed by the speed of light and the length of the communication interval). The black curve represents the mean from 30 individual Monte Carlo Realisations (MCRs) - the blue error bars indicate the standard deviation.

separated by a distance of 100 pc require a minimum communication interval of

$$dt = \frac{2\,dx}{c} = 652\,\text{years} \tag{D.19}$$

in order to establish a dialogue. Constraining the communications interval to be 100 years long *at its maximum* more or less completely removes the opportunity for communication. Remember that the communication interval refers to the overlapping time interval in which both civilisations can communicate - it can be as short as a year depending on circumstances and the development of the civilisations that constitute the ICP.

This analysis does not consider the possibility that if only one message is sent and received (rather than two), ICPs will renew their efforts to continue communications by other methods, despite now being radio quiet. In our formalism, this corresponds to $f = 0.5$ (i.e. instead of a pair of signals being exchanged, only one signal is sent and received). If this occurs, the communications interval is effectively extended beyond 100 years if $f = 0.5$ for $dt \leq 100$ yr. This extension will only marginally affect the results - most ICPs will still never be able to make the requisite half a conversation to instigate communication. The fact remains that the probability of an exchange being established is drastically reduced by the constraints of working with SKA-like instruments in an increasingly radio-quiet Galaxy.

D.7.4 Discussion and Conclusions

We have used Monte Carlo Realisation techniques to study the connectivity of intelligent civilisations when a) the civilisations are radio loud or are detectable by other means for their entire

existence, with no spatial limits on their detectability, and b) the civilisations are human-like in nature, and are limited in their use of communication tools (e.g., radar) and astronomical instruments, similar in capability to the Square Kilometre Array (SKA): We hypothesis that they become radio quiet (and hence undetectable) in 100 years, and can only detect other civilisations within 100 pc of each other (using an SKA-like telescope).

Our results show that civilisations in case a) enjoy significant connectivity, allowing thousands of exchanges of signals travelling at light-speed. In case b), connectivity decreases to virtually zero, showing that a SKA eavesdropping experiment would struggle to detect any human-like ETI in the Galaxy.

We can justify these results by considering the *civilisation formation rate density* $\dot{n}$, defined as the number of civilisations forming in a given volume in a given time interval. For the SKA to be able to detect human-like civilisations, then $\dot{n}$ must satisfy:

$$\dot{n} \geq 2(100\,pc)^{-3}(100yr)^{-1} = 2 \times 10^{-8} pc^{-3} yr^{-1} \tag{D.20}$$

If we assume that civilisations form across the entire Galaxy at a constant rate over its lifetime ($\sim$ 10 Gyr), then we can calculate the total number of civilisations that the Galaxy must produce:

$$N \geq \dot{n}\pi R^2 z \Delta t \sim 10^{14} \tag{D.21}$$

Where we have taken canonical values for the Galactic Radius R and height z. If we attempt to model the civilisation formation rate we obtain from our simulations, then an appropriate model is a Gaussian, with a standard deviation σ_t, and peak formation time t_0 , forming in the Galactic Habitable Zone, an annulus at radius $R = 7$ kpc with width $\Delta R = 2$kpc and height $z = 300$ pc, (cf. Lineweaver et al., 2004):

$$n(t) = \frac{N_0}{V} \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left(-\frac{(t-t_0)^2}{2\sigma_t^2}\right) \tag{D.22}$$

Figure D.17 confirms that $t_0 \approx 1.1 t_H$ and $\sigma_t \approx 0.3 t_H$, where $t_H = 13.7$ Gyr is the current age of the Universe. We normalise the curve using our peak signal $N_{peak} = 3 \times 10^5$, giving $N_0 = N_{peak}\sqrt{2\pi\sigma_t^2}$. This gives a peak formation rate of

$$\dot{n} \approx 5 \times 10^{-15} pc^{-3} yr^{-1} \tag{D.23}$$

Modelling in this manner allows us to calculate exactly how populous the Galaxy must be for SKA-based communication. For a civilisation formation rate density that satisfies equation (D.20) at the present day, we require that

$$N_{peak} \approx 10^{12} \tag{D.24}$$

This is considerably larger than the most positive estimates taken from the Drake Equation. Coupled with our experimental results, it shows that human-like civilisations will find it difficult

to see each other using SKA-like instruments alone. If we assume that galaxies with a peak formation rate matching that of equation (D.20) will have communicating civilisations with a probability of 1, then we can estimate that our model predicts a probability of communication of $\sim 10^{-7}$. These results are based on an assumed observation duration of 1 month. If we extend this to ~ 10 years, then the SKA's detection radius extends to ~ 1000 pc. Equation (D.20) then becomes

$$\dot{n} \geq 2(1000\,pc)^{-3}(100yr)^{-1} = 2 \times 10^{-11} pc^{-3} yr^{-1} \tag{D.25}$$

And the probability of detection is increased to $\sim 10^{-4}$, indicating that for 10^5 civilisations, a handful of contacts may become possible with this detection radius. However, 10 years of constant observation with the SKA is unlikely, given the plethora of demands for SKA time from many astronomical fields.

While the SKA remains an important instrument for SETI researchers, its abilities are limited to detecting civilisations that are at a less advanced state than Mankind, i.e. they must develop radio technology that remains radio-loud for a significant period of time. This strengthens the argument for a multi-wavelength approach to SETI, as radio-quiet civilisations may be optically loud, or detectable at some other energy scale (cf. Learned et al. 1994; Silagadze 2008).

Our calculations suggest that accidental communications, through the eavesdropping of radiation, is highly unlikely and therefore would require civilisations to actively beam signals to us to maximise our chances of detecting their existence. This requires a lot of energy and thus demands civilisations much more advanced than humans, and also better placed to manage the "costs" of such work (Cirkovic, 2008). The beam would most likely be collimated to boost its strength, hence the civilisation must find a target for the beam. Therefore, they would also need to know we are here, demanding more sophisticated astronomical instruments, e.g. radio telescopes far in excess of the SKA, or optical instruments that can characterise habitable planets beyond even the capabilities of the putative DARWIN TPF flotilla (Ollivier et al., 2008), or the European Extremely Large Telescope (Kaltenegger & Selsis, 2010).

These conclusions are not new, and it has been appreciated for some time that the discovery of radiation from civilisations at our present-day level and with short timescales would be unlikely. Our work simply updates these arguments using the latest predictions for the possible distribution of planets harbouring life in our Galaxy, and the probable sensitivities of the newest radio arrays. Such a conclusion could also be seen as being rather pessimistic for SETI but, like others, we would like to stress that we are guaranteed to find nothing if we give up looking. What these results stress strongly is that SETI must be a multi-wavelength endeavour, conducted with broader horizons (Davies, 2010) and a better understanding of our own limitations.

D.8 Summary and Future Improvements

This appendix has detailed my research in numerical methods in SETI. It has been used to study the Rare Earth Hypothesis (Forgan & Rice, 2010b), the hypothetical effects of entropy and sociological pressures (Bozhilov & Forgan, 2010), and to assess the efficacy of the SKA given certain assumptions about the evolution of radio technology (Forgan & Nichol, 2010). While this research shows interesting features when compared relative to some baseline model, the methods are still subject to the same core problems as all SETI research - we remain (apparently) alone in the Universe, with only the history of Earth to guide us, when the combined histories of all the intelligent worlds is likely to be far more diverse and unimaginable.

Numerical modelling of this type is generally a shadow of the entity it attempts to model, in this case the Milky Way and its constituent stars, planets and other objects. Several suggestions are listed here.

1. A more accurate Galactic model, taking into better account its chemical diversity, stellar clustering and the inner regions (specifically the central supermassive black hole and the hypervelocity stars orbiting it).

2. An improved planetary architecture model, better equipped to deal with moons and the planet mass-semimajor axis distribution. Also missing is the modelling of orbital eccentricity and inclination, potentially of great importance in issues of habitability (e.g. Williams & Pollard 2002; Spiegel et al. 2008).

3. Improved modelling of the connectivity of civilisations (potentially extending to the modelling of interstellar colonisation and face-to-face contact).

Along with improvements in the code's architecture (e.g. parallelisation and the potential use of distributed computing services like Condor, which are well suited to low-memory, long runtime tasks), these numerical methods have the potential to grow and develop into powerful tools for testing astrobiological theory. With a commensurate improvement in observational data, numerical techniques such as MCR may prove to be crucial in determining whether our status as an intelligent technological civilisation is unique in the Universe.

Appendix E

Articles Published in the Course of this PhD

For reference, I list articles I have authored during my PhD. Up to date publication information can be obtained at my homepage (current link http://www.roe.ac.uk/∼dhf), or by searching "Duncan Hugh Forgan" online.

1. Forgan D.H., "A Numerical Testbed for Hypotheses of Extraterrestrial Life and Intelligence", IJA, (2009), 8, pp 121-131

2. Forgan D.H., Rice W.K.M., Stamatellos D., Whitworth A.P., "Introducing a Hybrid Method of Radiative Transfer for Smoothed Particle Hydrodynamics", MNRAS, (2009), 394, pp 882-891

3. Forgan D.H., Rice W.K.M., "Stellar Encounters: A Stimulus for Disc Fragmentation?", MNRAS, (2009), 400, pp 2022-2031

4. Forgan D.H., Rice W.K.M., "Stellar Encounters in the Context of Outburst Phenomena", MNRAS, (2010), 402, pp 1349-1356

5. Forgan D.H., Rice W.K.M., "Numerical Testing of the Rare Earth Hypothesis using Monte Carlo Realisation Techniques", IJA, (2010), 9, pp 73-80

6. Forgan D.H., "Numerical Astrophysics, Numerical Astrobiology, and the Search for Extraterrestrial Life", JOC, (2010), 5, pp 811-817

7. Bozhilov, V., Forgan D.H., "The Entropy Principle, and the Influence of Sociological Pressures on SETI", IJA, (2010), 9, pp 175-181

8. Forgan D.H., Rice W.K.M., "Native Synthetic Imaging of Smoothed Particle Hydrodynamics density fields using gridless Monte Carlo Radiative Transfer", MNRAS, (2010), 406, pp 2549-2558

9. Forgan D.H., Nichol R.C., "A Failure of Serendipity: the Square Kilometre Array will struggle to eavesdrop on Human-like Extraterrestrial Intelligence", IJA, (2010), in press

10. Forgan D.H., Rice W.K.M., Cossins P.J., Lodato G., "The Nature of Angular Momentum Transport in Radiative Self-Gravitating Protostellar Discs", MNRAS, (2010), in press

Appendix F

Glossary

- **AABB** - Axis Aligned Bounding Box. A cubic construct aligned with the Cartesian axes, often used in computer graphics.
- **Accretion Disc** - A general term for a body of low aspect ratio rotating around a central object. Circumstellar discs constitute a subset of the accretion discs.
- **Adiabatic** - Describes a thermodynamic process where heating and cooling are due to pressure changes with no heat transfer.
- **Advection** - The transport of properties in a fluid due to the fluid's bulk motion.
- **Albedo** - Describes the ratio of scattered light to total incident light in a medium.
- **AMR** - Adaptive Mesh Refinement. A process by which grid based codes increase local resolution when required by increasing the number of grid cells used.
- **Ansatz** - From the German for "approach" or "attempt", it is commonly used in mathematical sciences to describe an educated guess which is later shown to be accurate and/or consistent.
- **Apastron** - In orbital dynamics, the orbital angle at which the orbiting body is at its maximum distance from the central body. The suffix is often changed to describe the system in question (e.g. aphelion for objects in the Solar System, apgalacticon for objects in the Galaxy).
- **Aspect Ratio** - In protostellar discs, the ratio of vertical scale height to horizontal scale height.

- **AU** - Astronomical Unit. A unit of distance equal to the semimajor axis of Earth's orbit.
- **Axial Ratio** - The ratio of major to minor axis of a spheroid (assuming that two of the three possible axes are of equal length). Oblate spheroids have a major axis greater than the minor axis, prolate spheroids have a minor axis greater than the major axis. Triaxial spheroids have three axes of differing lengths.
- **AV** - Artificial Viscosity, used in hydrodynamic simulations to improve the stability of the code during shocks and discontinuities.
- **Barotropic** - Describes a fluid whose pressure is a function of the density only.
- **Brown Dwarf** - Objects intermediate in mass between the stars and planets. They are capable of fusing deuterium at their cores, but cannot fuse hydrogen as stars on the Main Sequence do.
- **CDF** - Cumulative Distribution Function $F(x)$, the probability that a random variate X will be less than some value x.
- **Cepheid Variable** - A class of pulsating star whose luminosity can be determined accurately based on the period of its pulsations. It is this property which has allowed it to become a "standard candle", giving a *bona fide* measurement of cosmic distances that are otherwise difficult to establish.
- **CFL** - Courant-Friedrich-Lewys condition. Describes the maximum timestep a simulation should use to correctly model information exchange between fluid elements, e.g. sound waves.
- **Chondrules** - Round grains composed primarily of silicates. They are formed by melting before being accreted by primitive asteroids.
- **Circumstellar/Protostellar/Protoplanetary/Debris Disc** - A disc of material around a star. The choice of word indicates the evolutionary phase of the disc: I detail the nomenclature at the beginning of chapter 2.
- **CMF** - Core Mass Function, describes the mass distribution of prestellar cores in star forming regions.
- **Corotation** - In discs, it describes a radius where some phenomenon (such as a spiral wave) rotates with the same angular velocity as the bulk at that radius.
- **Cross Section** - In scattering physics, the effective area presented by a particle to some impinging object (often larger than the geometric area of the particle due to an interacting force).
- **CTTS** - Classical T Tauri Star, describing young stars with circumstellar discs that are accreting at a significant rate.

- **Dark Matter** - A posited exotic substance which is influenced only by gravitational forces, used as an explanation for phenomena such as anomalous galactic rotation curves and mass differentiation in colliding shocking clusters (such as the "Bullet" Cluster).
- **Diffusion Equation** - A partial differential equation that describes the diffusion of a substance or property in a medium.
- **Dispersion Relation** - In wave mechanics, a relation that exists between (for example) frequency and wavenumber.
- **Eccentricity** - A measure of an orbit's deviation from a perfect circle. Zero eccentricity is perfectly circular, and an eccentricity greater than one leads to an unbound orbit.
- **EM** - Electromagnetism/Electromagnetic.
- **Enthalpy** - A thermodynamic potential which calculates the heat transfer during a quasi-static, isobaric process (in a closed system).
- **Entropy** - A fundamental property of thermodynamic systems which characterises the level of disorder.
- **Epicyclic Frequency** - The *radial* frequency of oscillation of a perturbed body (as opposed to its azimuthal frequency, usually denoted Ω). Keplerian discs have a epicyclic frequency equal to their azimuthal frequency.
- **ESC** - Einstein's Summation Convention. When using index notation, repeated indices are "dummy indices" to be summed over. ESC is used to reduce algebraic "clutter" by removing the need for summation symbols.
- **Euler-Lagrange Equation** - A set of differential equations which, in mechanics and field theory, are used to derive equations of motion in generalised coordinate systems.
- **Eulerian Derivative** - Describes a derivative with respect to a fixed spatial coordinate (as opposed to the Lagrangian derivative).
- **Euler's Equation** - The equation of motion for an inviscid fluid.
- **EUV** - Extreme Ultraviolet. In disc photoevaporation, this describes radiation at energies between $[13.6, 100]$ eV (or wavelengths of approximately 90 nm to 10 nm). An EUV photon will typically ionise a hydrogen atom (as the ionisation energy of hydrogen is equal to the Rydberg constant at 13.6 eV).
- **Exoplanet/Extrasolar planet** - Refers to planets outside the Solar System.
- **EXors** - EX Lupi, refers to a specific class of stellar outburst phenomenon.
- **Extremophile** - Organisms which thrive in extreme environments, e.g. thermophiles, which survive in high temperature environments.

- **FLD** - Flux Limited Diffusion, an approximation in radiative transfer which describes the photons' passage through a medium using a diffusion equation.
- **Frequentist** - Refers to the school of probability where probabilities are assigned based on a notional infinite repetition of a particular event. For example, to define the probability that tossing a fair coin gives tails, the frequentist definition of probability states that repeating the coin toss ad infinitum will give tails on half of the events, and therefore the probability is 0.5. This is in contrast to the *Bayesian* school, which defines probabilities based on a "degree of belief" which is supported by available prior information.
- **FU Ori/ FUor** - FU Orionis, refers to the outburst phenomena of the same name.
- **FUV** - Far Ultraviolet. In disc photoevaporation, this describes radiation at energies between $[6, 13.6]$ eV (or wavelengths of approximately 200 nm to 90 nm). Unlike EUV radiation (above), FUV photons will not typically ionise hydrogen.
- **Gas Giant** - A giant planet with a substantial gas content. Typically of mass greater than 10 Earth Masses.
- **Geometrically thick/thin** - Describes discs with a high/low aspect ratio.
- **GMC** - Giant Molecular Cloud, a large cloud of turbulent molecular gas which is capable of forming hundreds to thousands of stars.
- **GR** - Einstein's theory of General Relativity. While still to be reconciled with physics at the quantum level, it is currently the standard theory of gravity.
- **Gravito-turbulence** - A phenomenon generated by marginally unstable discs, where the action of spiral waves produced by the gravitational instability generates turbulence in the disc.
- **Grid-based (hydrodynamics)** - A simulation method whereby the fluid is discretised into grid cells, and the hydrodynamic equations are solved at cell interfaces or inside cells.
- **Gyr** - Gigayear, or 1 billion years
- **Heliocentric** - Describes a theory which places the Sun at the centre of a system.
- **Hill Radius** - Given a massive object, the Hill Radius is the maximum distance at which that object still dominates the local gravitational potential.
- **Hydrostatic Equilibrium** - A state of force balance between gravity and pressure gradients in a fluid.
- **Ice Giant** - Giant planets with a substantial ice content (such as Uranus or Neptune).
- **Isobaric** - A surface of constant pressure in a fluid.

- **Isothermal** - Describes thermodynamic processes where the temperature remains constant.
- **Isotropic** - Describes objects or systems that are uniform in all directions.
- **Inviscid** - Describes a fluid with no viscosity.
- **IMF** - Initial Mass Function, describes the mass distribution of stars in a system such as a cluster or a galaxy.
- **Keplerian Discs** - Discs with an angular velocity profile that follows Kepler's 3rd Law.
- **Kirchhoff's Law** - Systems in thermal equilibrium will have an emissivity equal to their absorption coefficient multiplied by the Planck Function.
- **Lagrangian** - A function that "encodes" the dynamics of a system. Substituting it into the Euler-Lagrange Equations produces equations of motion for the system.
- **Lagrangian Derivative** - also *convective derivative* or *derivative following the motion*. Defines a derivative taken along a path with a given velocity. In fluid mechanics, it describes a derivative relating to a specific fluid element moving in the aforementioned path.
- **Laminar** - Describes a smooth, undisrupted fluid flow which has a low Reynolds number.
- **Lane-Emden Equation** - Poisson's equation for spherically symmetric, polytropic, self-gravitating fluids.
- **Legendre Polynomials** - Solutions to Legendre's equation, they appear when solving certain partial differential equations in spherical polar coordinates.
- **Lenz's Law** - A statement of Newton's Third Law in electromagnetism. In short, any induced current flows in a direction that opposes the phenomena causing it.
- **LTE** - Local Thermodynamic Equilibrium, a state of balance where there are no local temperature, pressure or chemical gradients.
- **Marginal Instability** - A state self-gravitating discs attain when the heating generated by gravitational instability matches the radiative cooling, resulting in a self-regulated, quasi-steady gravito-turbulent state.
- **Metazoan** - describes multicellular, animal life based on eukaryotic cells (i.e., cells with nuclei).
- **Monochromatic** - In radiation, describes a beam of a single frequency (sometimes a small range of frequencies).

- **Monte Carlo** - Refers to computational algorithms which utilise stochastic processes (such as sampling from random number generators).
- **MCR** - Monte Carlo Realisation. Refers to Monte Carlo simulations which use a number of runs or realisations to average out random errors.
- **MCRT** - Monte Carlo Radiative Transfer. This method uses stochastic equations to model the progress of photons in a medium.
- **MRI** - Magnetorotational Instability - occurs in rotating, magnetised fluids.
- **Myr** - Megayear, or 1 million years.
- **Navier-Stokes Equation** - The equation of motion for a viscous fluid.
- **Neighbour Sphere** - See Smoothing Volume.
- **Octree** - a data structure which spatially indexes a system by decomposing it into hierarchical octants (see section 3.2.6).
- **Optical Depth** - A measure of how transparent a medium is to radiation.
- **Optically thick/thin** - Describes a medium which is opaque/transparent to radiation.
- **ORP** - Outer Rotation Period. Refers to the rotation period at the outer edge of a disc.
- **Panspermia** - Theories of life's origin which suggest bacteria can be transmitted between planets, "seeding" them with life.
- **Parallax** - Describes the motion of stars on the sky that occurs as a result of Earth's motion around the Sun.
- **Pattern Speed** - In spiral waves in discs, describes the apparent angular speed at which the spiral moves through the disc.
- **pc** - parsec, a unit of distance. Objects at distances of 1 parsec have a **par**allax of one arc**sec**ond.
- **PDF** - Probability Distribution Function/Probability Density Function. Concerning a random variable X, it describes the probability that a sampling of X will yield a result in a given range of values.
- **Periastron** - In orbital dynamics, the orbital angle at which the orbiting body is at its minimum distance to the central body. The suffix is often changed to describe the system in question (e.g. perihelion for objects in the Solar System, perigalacticon for objects in the Galaxy).

- **Phase Function / Phase Matrix** - Describes the probability that a scattered photon will scatter into a given solid angle at a given orientation. The phase function (which acts on the intensity I) becomes a phase matrix when it acts on all four Stokes parameters.
- **Photoevaporation** - The process by which radiation liberates gas from an object by imparting sufficient thermal energy.
- **Photon** - The basic unit of EM radiation, a massless quantum of energy which exhibits both wave and particle properties.
- **Polarisation** - A property of waves which defines the orientation of their oscillations. In electromagnetism, it measures the orientation of the electric field relative to the direction of travel.
- **Polytrope** - Describes gas which obeys relations of the form $P = CV^{-n}$.
- **Pointmass** - A non-gaseous particle in a hydrodynamics simulation which plays the role of accreting objects such as protostars or protoplanets.
- **Poisson Noise** - Fluctuations of an observed property around some mean value, where the variance of the fluctuation is equal to the square root of the mean.
- **Poisson's Equation** - A standard form of a partial differential equation: $\nabla^2\phi = f$, often used for gravitational calculations.
- **Prestellar Core** - Condensations of gas in star-forming regions that, while gravitationally bound, have not yet formed any stars.
- **Prograde** - In orbital dynamics, it describes an orbit with orbital angular momentum with the same sign as the spin angular momentum of the central body.
- **Protoplanet** - Describes planetary bodies in an early phase of their evolution.
- **Protostar** - Describes stellar bodies in an early phase of their evolution.
- **Quasistatic** - Describes a system in a state close to equilibrium. The system evolves on a relatively slow timescale, and at any given instant the system is close to equilibrium.
- **Quantum/Quanta** - Refers to a particle which is a discrete, indivisible unit of mass/energy. Electromagnetic energy is discretised into quanta known as photons (see above).
- **Retrograde** - In orbital dynamics, it describes an orbit with orbital angular momentum of opposite sign to the spin angular momentum of the central body.
- **Reynolds Number** - A dimensionless parameter which compares the inertial and viscous forces in a fluid. High Reynolds Numbers are a symptom of turbulence in the fluid.

- **RHD** - Radiative Hydrodynamics.
- **SED** - Spectral Energy Distribution. Describes how the radiant energy from an astronomical object varies with photon wavelength.
- **Smoothing Kernel/Interpolating Kernel** - In SPH, the function that interpolates between particles to estimate the fluid properties at arbitrary positions.
- **Smoothing Length** - A scaling factor used in the smoothing kernel to facilitate the interpolation process. Varying the smoothing length in effect varies the local resolution of the SPH simulation.
- **Smoothing Volume** - A sphere, with radius equal to twice the smoothing length. In SPH, the smoothing volume contains the nearest neighbours of the particle, and defines the causal region in which the particle experiences hydrodynamic forces.
- **Solid Angle** - A two-dimensional angular area subtended by objects on the sky. Its maximum value is 4π, and is dimensionless.
- **Source Function** - Defined as the ratio of emissivity over absorption coefficient. In thermal equilibrium, it is equal to the Planck blackbody function.
- **SPH** - Smoothed Particle Hydrodynamics. A Lagrangian method, it simulates fluids using particles, which act as buoys or sensors in the fluid.
- **Strain** - Sometimes *deformation*, a dimensionless measure of change in the metric properties of a body, i.e. the distances between the constituents of the body.
- **Stress** - Measures the force per unit area on the internal structure of a deformable body. It is calculated by measuring the total force acting on a defined surface within the body, and has units of pressure.
- **Stokes Vector** - In radiative transfer, a vector constructed from four parameters $S = (I, Q, U, V)$ which define the radiation field.
- **Tensor** - A generalisation of geometric properties such as scalars, vectors and matrices, allowing higher-order constructions.
- **TDV** - Time Dependent Viscosity, an algorithm to improve the implementation of artificial viscosity by varying its scale as a function of time.
- **Terrestrial Planet** - Broadly defined as a planet composed primarily of rocky material. Typically of mass less than 10 Earth Masses.
- **Turbulence** - A fluid flow which is "rough", stochastic and disrupted, mixing on scales larger than the local mean free path, with a high Reynolds number.

- **Viscosity** - In a fluid, this is a dissipative process, originating from internal friction or stresses. High viscosity fluids tend to be laminar, and low viscosity fluids are more likely to be turbulent.
- **WLTTS** - Weak Lined T Tauri Stars. Stars with circumstellar discs that are no longer accreting strongly.