A Study of Protoplanetary Disk Dynamics using Accelerated Hydrodynamics Simulations on Graphics Processing Units

by

Jeffrey Fung

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Astronomy and Astrophysics
University of Toronto

Abstract

A Study of Protoplanetary Disk Dynamics using Accelerated Hydrodynamics Simulations on Graphics Processing Units

Jeffrey Fung
Doctor of Philosophy
Graduate Department of Astronomy and Astrophysics
University of Toronto
2015

This thesis focuses on the dynamical interaction between the gaseous component of a protoplanetary disk and the solid bodies within. We identify and characterize new dynamical behaviors of solid bodies ranging from micron-size dust grains to Jupiter-size planets, using hydrodynamics simulations accelerated by graphics processing units (GPUs). Chapter 1 outlines the relevant physics and explains our research motivation. Chapter 2 gives a detail description of our GPU hydrodynamics code `PEnGUIn`. Our benchmark shows that, running on a GTX-Titan graphics card, `PEnGUIn` can update 25 million grid cells per second in a three-dimensional (3D) calculation. Chapter 3 combines `PEnGUIn` simulations and semi-analytic calculations to demonstrate the existence of a new disk instability, called the irradiation instability. We find that when the star exerts a sufficiently strong radiation pressure, the interplay between dust grains, gas, and radiation is unstable to linear perturbations, and, in extreme cases, can result in "clumping", local surface density enhancements beyond 10 times the initial value. In Chapter 4 we consider disk gaps opened by giant planets. We determine how the average surface density inside the gap, $\Sigma_{\rm gap}$, depends on planet-to-star mass ratio q, Shakura-Sunyaev viscosity parameter α, and disk height-to-radius aspect ratio h/r. We derive an analytical scaling that predicts $\Sigma_{\rm gap} \propto q^{-2}\alpha^{1}(h/r)^{5}$, and show that it compares well to results determined numerically with both `PEnGUIn` and `ZEUS90`, a modified version of the publicly available code `ZEUS`. In the end, we turn our attention to Earth-size planets which exchange mass and angular momentum with the disk without significantly modifying the local disk structure. Most work done on this topic has been under the assumption of an infinitely thin 2D disk, and so a precise description in 3D has been lacking. 3D simulations with `PEnGUIn` described in Chapter 5 reveal that vertical motion plays an important role in the 3D flow field around an embedded planet, and has a direct impact on both planet accretion and migration. In particular, the size of the planet's atmosphere is much smaller than anticipated, and the corotation torque on the planet deviates significantly from 2D predictions.

Dedication

To my parents, who never put a limit on me.

Acknowledgements

This thesis was funded by the University of Toronto, a Discovery Grant by the Natural Sciences and Engineering Research Council of Canada, Queen Elizabeth II and Walter John Helm Graduate Scholarships in Science and Technology, and Ontario Graduate Scholarships.

The work presented in this thesis would not have been possible without the help from the mentors and collaborators I have had the fortune to meet. I would like to thank my thesis advisor, Pawel Artymowicz, for introducing me to the field of astrophysics years before I even entered graduate school, setting a path of great potential ahead of me, and sharing incredible insights whenever I managed to track him down.

I am especially indebted to Eugene Chiang, who not only inspired our work on disk gaps, but has since been providing countless advice, guidance, and consultation in every stage of my career, carrying me to where I did not believe I could reach.

I am grateful to Yanqin Wu, for her mentorship, and the tireless reading and editing of our manuscripts; and Ji-Ming Shi, whose expertise in computational hydrodynamics greatly elevated our work.

I am also grateful to Bob Abraham, Chris Matzner, and Norm Murray, for their helpful advice and support.

And finally, thank you, Etsuko, for depending on me, and letting me depend on you.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Stretched over a radius of more than 100 AU, rotating around newly born stars, protoplanetary disks are vast, fertile grounds for fetal planetary systems. Planets are built literally from dust inside these disks over hundreds of thousands of years, before they become the planets we observe today, including, of course, our own planet Earth. The ongoing dynamical processes in protoplanetary disks heavily impact their structure, such as their density and pressure profiles, which in turn influence how planets are formed. This thesis covers a range of topics relating to the dynamical evolution of protoplanetary disks. We are specifically interested in the interaction between the gaseous disks and the embedded solid bodies, such as micron-size dust grains, Earth-size planets, and giant planets like Jupiter.

Throughout this work we make use of both numerical simulations and analytic models to provide support for our theories. The numerical work is performed using an application of a new branch of computing technology: general-purpose computing on graphics processing units (GPUs). In this introductory chapter, we will first give a description of protoplanetary disks, and then provide a summary on our current understanding of the dynamical behavior of embedded dust grains and planets. In the end, we will also explain our motivation for utilizing the GPU technology for scientific research.

1.1 General Properties of Protoplanetary Disks

Through gravitational collapse and thermal cooling, dense molecular gas clouds are formed into stars, and the residual gas from this formation process is left as rotationally supported disks, orbiting around the newly born stars. These protoplanetary disks have a typical lifetime of $10^6 \sim 10^7$ years (Haisch et al., 2001), before being dispersed by photoevaporation. Because this is a short time compared to the stellar evolution timescale, newly born stars, or "young stellar objects" (YSOs), make for prime targets for observing these disks. A common type of these objects are T Tauri stars, which are bright, variable stars still in the process of contracting and reaching the stellar main sequence. The structure of a protoplanetary disk is in many ways tied to the properties of its host stars, both in terms of the gravity that holds the star-disk system together, and the star's radiation that provides thermal energy for the disk.

The rotation of a protoplanetary disk closely follows Keplerian motion, with a speed of $v_\mathrm{k} = \sqrt{GM_*/r}$, where G is the gravitational constant, M_* is the mass of the host star, and r is the distance from the star. This is modified only by the disks' self gravity, and thermal pressure. In general we omit to consider self-gravity due to its complexity, and this is justified because the total disk mass is expected to be $\sim 1\%$ of its host stars, estimated

by a minimum mass solar nebula (MMSN; Hayashi 1981).

Thermal pressure, on the other hand, is a major influence on the structure of a protoplanetary disk. In hydrostatic equilibrium, thermal pressure not only modifies the rotation curve, but also determines the thickness of the disk. There are two main sources of heat: irradiation from the host stars, and friction due to disk viscosity. If the disk is mainly heated by stellar irradiation, it is called a passive disk; otherwise, it is active. Our study focuses on passive disks, and we typically use a locally isothermal equation of state, where the temperature of the disk has a fixed radial profile. This assumes the stellar irradiation is fixed for a given radius, and that the variation in temperature from the disk surface to midplane is small. This is a major simplification commonly employed when thermal dynamics is not treated explicitly. In general, one expects a sharp change in temperature from the hot, optically thin disk surface to the cold, optically thick midplane. On the other hand, since most of the disk mass resides near the midplane, from a dynamical standpoint, it is justified to assume that most mass at a given radius share a constant temperature. Given this equation of state, a useful quantity for describing the disk structure is the sound speed c_s:

$$c_s^2 = \frac{k_B T}{m_\mu}\,, \tag{1.1}$$

where k_B is the Boltzmann constant, T the local disk temperature, and m_μ the mean molecular mass. In hydrostatic equilibrium, the rotation curve is modified by a factor of $\sqrt{1 + (c_s/v_k)^2(\mathrm{d}\ln\Sigma/\mathrm{d}\ln r)}$, where Σ is the disk surface density. Because Σ must in general be decreasing for increasing r, the disk rotates at a sub-Keplerian speed. The vertical structure of a disk is described by the balance between the vertical component of the host star's gravity and the vertical pressure gradient. Since we already assumed the disk is vertically (locally) isothermal, the disk density ρ is:

$$\rho = \rho_0 \exp\left[-\frac{r^2}{h^2}\left(1 - \frac{r}{\sqrt{r^2+z^2}}\right)\right], \tag{1.2}$$

where z is the vertical displacement from the midplane, ρ_0 is the density at the midplane, and $h = rc_s/v_k$ is the disk scale height, which characterizes the thickness of the disk. For a thin disk, this equation can be simplified to $\rho = \rho_0 e^{-\frac{z^2}{2h^2}}$ by taking the limit $z \ll r$, and ρ_0 is related to Σ by the simple relation $\rho_0 = \Sigma/(h\sqrt{2\pi})$.

We can estimate the aspect ratio h/r for a solar-mass host star at 1 AU by estimating the temperature to be 300 K, which is the equilibrium temperature today, and μ to be 2.3 proton mass, appropriate for a molecular gas of cosmic composition. This gives $h/r \sim 0.035$, showing that protoplanetary disks are indeed geometrically thin. The temperature profile of a passive disk was derived by Chiang & Goldreich (1997), who showed that the disk has a flaring shape, where h/r follows a power-law in r, with the power varying between 0.28 to 0.5. So, h/r increases gradually to ~ 0.1 at $r \sim 10^2$ AU. Going back to the disk's rotation, plugging in our value of h/r, or equivalently c_s/v_k, we find that the rotation is only sub-Keplerian by a margin of $((h/r)^2/2)v_k \sim 10^{-4}v_k$. While this difference is small, it is important when considering the coupling between gas and dust, which we will discuss in more detail in Section 1.2.

Protoplanetary disks are also accretion disks that are constantly feeding gas to their host stars. This process requires gas in the disk to lose angular momentum, which is done through the outward angular momentum transport caused by disk turbulence. Many sources of turbulence has been identified, including the magnetorotational instability (Balbus & Hawley, 1991), gravitational instability (Lin & Pringle, 1987; Gammie, 2001), and vertical shear instability (Urpin & Brandenburg, 1998; Nelson et al., 2013). Each of them involves a different and elaborate physical mechanism. Rather than directly including these mechanisms into our analyses, a simple method to parameterize the effect of turbulence is to solve Navier-Stokes equations with a finite kinematic viscosity ν, which produces a Reynolds stress in the fluid that we can control by varying the value of ν. Shakura & Sunyaev (1973) proposed that, knowing the largest eddy size in a turbulent disk cannot exceed h, and the typical speed of the tur-

bulent motion should be related to the sound speed c_s, one can write a more refined parameterization: $\nu = \alpha h c_s$, where α is a dimensionless parameter. Observations of protoplanetary disks have shown that $\alpha \approx 10^{-4} \sim 10^{-2}$ (Kitamura et al., 2002; Andrews & Williams, 2007; Andrews et al., 2009). An α value much less than unity implies that the global disk structure undergoes viscous evolution over a timescale much longer than the dynamical time, $\Omega_k^{-1} = r/v_k$. However, for situations where there the local disk structure is considered, viscosity can still play an important role, as we will see in Chapters 4 and 5.

1.2 Dust Grains in Protoplanetary Disks

Protoplanetary disks have an initial chemical composition that they inherit from the interstellar medium that formed the star-disk systems. In addition to the hydrogen and helium gas that make up the bulk of the mass, about 1% is in the form of micron-size dust grains. These grains are mostly made of silicates, such as $MgSi0_4$, and ice, such as water ice. Although they are a small fraction of the disk mass, they are the dominant source of disk opacity. This is primarily because they give very a large sum of surface area. For example, if all of the grains has a size of $s = 10^{-4}$ cm, and a density of $\rho_d = 3$ gcm^{-3}, like the rocks we find on Earth, the total opacity of the gas-dust mixture would be ~ 10 cm^{2}g^{-1}, which, given a gas density of $\sim 10^{-9}$ gcm^{-3} around 1AU in a MMSN, results in a vertical optical depth of $\tau = 10^3$ at 1AU.

This order-of-magnitude estimate only takes into account the dust grains' geometric cross-section. In reality, the cross-section will depend on the grains' shapes, chemical compositions, size distribution, and the wavelength of the incoming radiation. For example, the Rosseland mean opacity in protoplanetary disks with embedded spherical homogeneous dust grains was calculated by Semenov et al. (2003), who found an opacity of $1 \sim 10$ cm^{2}g^{-1} for a temperature range between 100 to 1000 K. Nonetheless, we can see that a protoplanetary disk at these distances to the star is optically thick. It only becomes optically thin at distances around 10^2 AU, as both the surface density and temperature of the disk decreases outward.

While dust grains process star light into the disk's thermal energy, the optically thick inner disk also emits observable thermal radiation. Again, because of their large total surface area, this thermal radiation is bright and, because it is in the infrared, can be clearly distinguished from the star's radiation. In fact, the existence of a protoplanetary disk is often inferred from the observed infrared emission from dust. Because of the prominent role they play in both the thermal properties and observations of protoplanetary disks, it is critical that we also understand their behavior dynamically.

When dust grains absorb photons, they are also pushed, in the form of radiation pressure. Without taking into account photon-scattering by these grains, the force of radiation pressure is always outward and purely radial. As a result, it partially cancels the attractive force of the star's gravity, and the equilibrium orbital speed for the grains becomes sub-Keplerian. Recalling that the gaseous component also has a sub-Keplerian orbital speed due to thermal pressure, the difference between the gas and dust speeds results in a drag on the dust, since the gas component is significantly more massive. Consider first, the case where radiation pressure is insignificant, such as when the star's luminosity-to-mass ratio is low. Then the grains will attempt to travel at the Keplerian speed, and experience a head-wind as they do so, forcing them to slow down to the speed of the gas. The non-dimensional stopping time:

$$t_s = \sqrt{\frac{\pi}{8}} \frac{s \rho_d v_k}{r \rho_g c_s}, \tag{1.3}$$

describes the time its takes for the dust grains to slow down compared to the dynamical time, where ρ_d is the density of an individual dust grain, and ρ_g is the gas density. t_s is derived using the Epstein drag law, appropriate

when the the mean free path of gas molecules is much longer than the size of the grains. Plugging in values for silicate grains in a MMSN, we find centimeter-size grains has $t_s \sim 1$ at 1AU, indicating that grains larger than this size are expected to be decoupled from the gas. If $t_s \ll 1$, which is generally true for micron-size bodies, the grains will orbit at the same sub-Keplerian speed as the gas. However, this means the centrifugal force on them will be weakened, and unable to balance the star's gravity. Consequently, the dust grains will migrate inward.

The opposite happens when radiation pressure is strong, such that the dust grains travel at a speed even slower than the gas. This time the grains will experience a back-wind, accelerating it to the speed of the gas. The centrifugal force plus the force of radiation pressure will then over-power gravity, pushing the grains outward. The detail analysis of this migration mechanism was done by Takeuchi & Artymowicz (2001). In a steady state, the dust grains will settle at a radius where the equilibrium orbital speed reduced by radiation pressure is exactly equal to the sub-Keplerian speed of the gas.

The above description establishes an excellent framework for understanding the radial migration of dust grains. Since then, further development includes Krauss & Wurm (2005) who examined the effects of photophoresis, and Dominik & Dullemond (2011) who also studied radiation pressure and gas drag on dust grains, but taking into account light extinction and the gas accretion flow. While all of these 1D studies contributed to our understanding, the picture is incomplete without generalizing into 2D, where the shapes of the dust grains' orbits can be taken into account. This is because dust grains do not necessarily migrate monotonically over the time of one orbit. When they experience a radial perturbation, either from radiation pressure or gas drag, they will oscillate radially at the epicyclic frequency due to the conservation of angular momentum. This generates a time-variation in the amount of radiation pressure they receive, which in turn can have a profound impact on the evolution of their orbits. In Chapter 3, we will see that the dynamics of dust grains under the effects of radiation pressure is more complicated than described here. We will show that asymmetric modes can grow into large scale instabilities, even if both the irradiation from the star and the initial disk are axisymmetric.

1.3 Disk-Planet Interaction

Unlike dust grains, planets interacts with the disk almost solely through gravity. Therefore, for the purpose of disk-planet interaction, they can be modeled as point masses with a gravitational potential determined only by their masses. The calculation of the linear response in a disk triggered by an external potential was pioneered by Goldreich & Tremaine (1979). By considering the initial response as a small perturbation, one can linearize the equations for the conservation of mass and momentum, and obtain a solution for the perturbation. This linear perturbation theory is a powerful method that allows us to obtain near-exact analytic solutions to situations that can be considered "linear", where the change in the disk due to the external force is small, and even provide insight into the more general nonlinear situations.

For a planet on a fixed circular orbit embedded in an infinitely thin (2D) disk, one can write its gravitational potential as a series in azimuthal modes:

$$\Phi(r, \phi, t) = \sum_{m=0}^{\infty} \Phi_m(r) e^{i(m\phi - \Omega_p t)}, \tag{1.4}$$

where m is the azimuthal mode number, and Ω_p is the orbital frequency of the planet. Each Φ_m describes the periodic component in Φ that has an m number of cycles from 0 to 2π along the azimuth ϕ. There are two important types of resonances associated with each m^{th} component. The first type is Lindblad resonance (LR).

They are positioned at the radius where:

$$\kappa = \pm m\left(\Omega - \Omega_{\rm p}\right), \tag{1.5}$$

where κ is the epicyclic frequency of the disk, and Ω the orbital frequency. The second is corotation resonance (CR), which is located where $\Omega = \Omega_{\rm p}$. In a Keplerian disk, we have $\kappa = \Omega$, and two Lindblad resonances for every m, one located at an orbit lower than the planet's, called the "inner" Lindblad resonance (ILR), and one at a higher orbit, called the "outer" Lindblad resonance (OLR). The corotation resonance is uniquely located at the planet's orbit. Goldreich & Tremaine (1979) showed that in the regions inward of the ILR and outward of the OLR, the linear equations have solutions in the form of traveling waves, and they carry a conserved amount of angular momentum flux. In the region between the two LRs, where the CR is located, the waves are evanescent. This implies that waves are excited at the LRs, where a transfer of angular momentum occurs between the disk and the planet.

In the following, we will give an overview on the evaluation of the Lindblad and corotation torques, referring to the torques on the planet as a result of the angular momentum transfer at LRs and CRs respectively. For simplicity, we will restrict ourselves to an isothermal disk.

1.3.1 Lindblad Torque

An expression for the Lindblad torque on the planet, derived by Goldreich & Tremaine (1979), is:

$$T_{{\rm LR},\,m} = \frac{m\pi^2\Sigma}{r({\rm d}D/{\rm d}r)}\left(r\frac{{\rm d}\Phi_m}{{\rm d}r} + \frac{2\Omega}{\Omega - \Omega_{\rm p}}\Phi_m\right)^2, \tag{1.6}$$

where $D = \kappa^2 - m^2(\Omega - \Omega_{\rm p})^2$. This equation is evaluated at the location of a Lindblad resonance. Applying the properties of a Keplerian disk, it is straightforward to show that for the ILRs, where $\kappa = m(\Omega - \Omega_{\rm p})$, ${\rm d}D/{\rm d}r$ is positive, and it is negative at the OLRs, where $\kappa = -m(\Omega - \Omega_{\rm p})$. Therefore the inner Lindblad torques are positive, while the outer Lindblad torques are negative. Also, for a given m, the OLR is always a little closer to the planet than the ILR, because κ changes more rapidly inward of the planet's orbit than outward. As a result, the terms proportion to the planet's potential are evaluated closer to the planet for the OLRs, giving the OLRs a stronger torque on the planet. We can therefore conclude that the while the ILRs and OLRs generate torques of opposite signs, the outer torque exceeds the inner one in magnitude, so the net Lindblad torque, also called the differential Lindblad torque, referring to the partial cancellation between the inner the outer parts, is negative.

By Newton's third law, Equation 1.6 tells us that disk material at the OLRs will gain angular momentum from the planet, and rise to higher orbits, while those at the ILRs will lose angular momentum and move to lower orbits. In either case, they move away from the planet. Therefore, Lindblad torques act like a repulsive force that clears disk material near the planet's orbit. The result is that in the vicinity around $r = a$, where a is the planet's radial position, material is constantly removed, creating a gap in the disk. These gaps are of profound significance since they are one of the most prominent observable features of disk-planet interaction. When we study the gap formation process in Chapter 4, we will consider the balance between the rate at which Lindblad torques transfer angular momentum, and the viscous diffusion rate that brings angular momentum back to the locations of the resonances.

A careful inspection of Equation 1.6 reveals a puzzling problem: in the limit of a large m, the position of the Lindblad resonances can be approximated as $r_{\rm LR} = a(1 \pm 2/(3m))$, and ${\rm d}D/{\rm d}r$ becomes proportional to m^2. Consequently, the magnitude of $T_{{\rm LR},\,m}$ scales as $m({\rm d}D/{\rm d}r)^{-1}(\Phi_m/(r_{\rm LR} - a))^2 \propto m^5$, meaning the series diverges. This problem is resolved byArtymowicz (1993b), who retained the terms explicitly dependent on m in the wave

equation, which Goldreich & Tremaine (1979) dropped in favor of the WKB approximation, and found that the location of wave-launching is slightly shifted away from the location of the resonance. The effective LRs, where waves are launched, are positioned at:

$$r_{\mathrm{LR,\ eff}} = a \pm \frac{2h_0}{3\xi}\sqrt{1+\xi^2}\,, \tag{1.7}$$

where $\xi = m(h_0/a)$, and h_0 is the disk scale height at $r = a$. In the limit where $\xi \ll 1$, we recover $r_{\mathrm{LR,\ eff}} \sim r_{\mathrm{LR}}$, but when $\xi \gg 1$, we get $|r_{\mathrm{LR,\ eff}} - a| \sim 2h_0/3$, so the effective LRs are separated from the planet by at least a distance of $2h/3$. The resulting $T_{\mathrm{LR},\ m}$ peaks around $\xi \sim 0.5$, and has an exponential cut-off when $\xi \gg 1$.

Paardekooper & Papaloizou (2008) generalized the problem further by including the softening parameter r_{s}. This parameter modifies the planet's potential from that of a point mass, $\Phi = -GM_{\mathrm{p}}/|\mathbf{r}-\mathbf{r}_{\mathrm{p}}|$, to $-GM_{\mathrm{p}}/\sqrt{(\mathbf{r}-\mathbf{r}_{\mathrm{p}})^2 + r_{\mathrm{s}}^2}$, where M_{p} is the planet's mass, and $\mathbf{r}_{\mathrm{p}}$ is planet's position vector. For 2D analyses, r_{s} serves the purpose of creating a simple, mock-up representation of a vertically averaged 3D potential. Müller et al. (2012) investigated the use of r_{s} extensively, and found that the choice of $r_{\mathrm{s}} \lesssim h$ is generally appropriate. The expression for the differential (net) Lindblad torque in a 2D isothermal disk, given by Paardekooper & Papaloizou (2008), is:

$$T_{\mathrm{LR}}^{\mathrm{2D}} = -(2.5 - 0.1\beta)\left(\frac{r_{\mathrm{s}}}{0.4h_0}\right)^{0.71} T_0\,, \tag{1.8}$$

where $T_0 = \Sigma_0\Omega_{\mathrm{p}}^2 a^4 q^2 (h_0/a)^{-2}$, and $q = M_{\mathrm{p}}/M_*$ is the mass ratio between the planet and the host star; Σ_0 and $\beta = -\mathrm{d}\ln\Sigma/\mathrm{d}\ln r$ are the disk surface density and negative dimensionless surface density gradient at the location of the planet respectively. While this expression is highly useful for 2D modeling of disk-planet interaction, the most physically relevant calculation of the Lindblad torque was done by Tanaka et al. (2002), who solved the linear equations in 3D. Their expression is:

$$T_{\mathrm{LR}}^{\mathrm{3D}} = -(2.34 - 0.1\beta)T_0\,. \tag{1.9}$$

Because β is of order unity ($\beta = 1.5$ in a MMSN), we expect T_{LR} to be negative and the planet to migrate inward.

The magnitude of the differential Lindblad torque can be converted into a migration time scale in the form:

$$\frac{a}{\dot{a}} = \frac{M_{\mathrm{p}}\Omega_{\mathrm{p}}a^2}{2|T_{\mathrm{LR}}|}\,. \tag{1.10}$$

For an Earth-size planet orbiting a solar-mass star at 1AU, and plugging in MMSN values for the disk, we get $a/\dot{a} \sim 10^5$ years, indicating the time it takes before the planet loses all of its angular momentum. This is the so called "type I" migration regime. Comparing this migration time to the lifetime of the disk, which is $10^6 \sim 10^7$ years as mentioned before, we find T_{LR} leads to the destruction of planets well before the disk disperses. Fortunately, we have yet to take into account the corotation torque, which may be the saving grace for planet migration theory.

1.3.2 Corotation Torque

The corotation resonance behaves very differently from Lindblad resonances. First, waves are evanescent around it, so it does not transfer angular momentum via wave excitation. Second, while the planet does excite an *initial* linear response around the corotation resonance, the gas orbits there are are non-linear in nature, i.e. they are strongly modified by the planet's gravity, and cannot be adequately approximated as circular orbits. This second point is apparent if we take a different perspective from the linear perturbation theory. In the restricted three-body

problem, test particles near the orbit of the secondary (the planet) move in "horseshoe" orbits. That is, in the corotating frame of the planet, they librate around the L3, L4, and L5 Lagrange points, without ever crossing the azimuth of the planet's position. We refer to the region where the gas exhibits horseshoe motion the "co-orbital" region.

To give a summary on the past research done on understanding the corotation torque, we first focus on the regime where the planet is sufficiently small, so that its co-orbital region does not overlap with the effective Lindblad resonances, which are minimally separated from the planet by $2h_0/3$, as mentioned previously. In this case, the linear corotation torque ($T_{\mathrm{CR,\,lin}}$) generated by an isolated corotation resonance can be calculated in much the same way as the Lindblad torque, by solving the linear equations and summing up contributions from all m modes. The expressions given by Tanaka et al. (2002) for 2D and 3D disks are:

$$T^{\mathrm{2D}}_{\mathrm{CR,\,lin}} = (1.36\zeta)T_0\,, \tag{1.11}$$

$$T^{\mathrm{3D}}_{\mathrm{CR,\,lin}} = (0.02 + 0.64\zeta)T_0\,, \tag{1.12}$$

where $\zeta = -\mathrm{d}\ln(\Omega/\Sigma)/\mathrm{d}\ln r$ is the negative dimensionless disk vortensity gradient, noting that for a Keplerian disk, the vortensity, or specific vorticity, is $|\nabla \times \mathbf{v}|/\Sigma \propto \Omega/\Sigma$. These values correspond to the corotation torque at the time when the planet is first introduced to the disk, so the disk orbits are still well approximated as circular, even in the vicinity of the planet. In the more general, nonlinear case where we have horseshoe orbits, Ward (1991) derived the so called "horseshoe drag", which we will refer to simply as the corotation torque T_{CR}. It describes the rate of angular momentum exchange between the horseshoe orbits and the planet, assuming a 2D disk. Its expression is:

$$T^{\mathrm{2D}}_{\mathrm{CR}} = \left(\frac{3}{4}\frac{w^4}{a^4q^2(h_0/a)^{-2}}\zeta\right)T_0\,. \tag{1.13}$$

where w is the half-width of the co-orbital region, or equivalently the half-width of the widest horseshoe orbit.

The dependence on ζ in both the linear and nonlinear expressions can be understood based on the notation that vortensity is a conserved quantity along a 2D streamline. For a fluid element to maintain its vortensity when it travels radially, it expands or contracts, adjusting its density to compensate for the change in Ω. When the initial vortensity profile is not constant, this results in a modification of the surface density in the co-orbital region. For example, if Σ has an initially constant profile, so the vortensity is decreasing outward, the horseshoe orbits moving from the inner disk ($r < a$) to the outer disk ($r > a$) would lead to a region of under-density downstream; the opposite happens for the horseshoe orbits on the other side, creating a region of over-density in the inner disk downstream. A similar picture applies to the linear case as well, since linear modes also oscillate between the inner and outer disks. We note here, that vortensity is not conserved in the general 3D case, so this result only applies when little vertical motion is present.

To evaluate T_{CR}, we must first find w. Masset et al. (2006) suggested that by equating $T^{\mathrm{2D}}_{\mathrm{CR}}$ and $T^{\mathrm{2D}}_{\mathrm{CR,\,lin}}$, one can obtain a good estimate for w, which is $\sim 1.16a\sqrt{q/(h_0/a)}$. Paardekooper & Papaloizou (2008) proposed that, including the effects of a softening parameter $r_{\mathrm{s}} \sim h_0$, w can be estimated as $\sim 0.82a\sqrt{q/(r_{\mathrm{s}}/a)}$, which gives a similar answer to that suggested by Masset et al. (2006) if $r_{\mathrm{s}} = 0.5h_0$. Note that the corotation torque scales with w^4, so if the expression by Paardekooper & Papaloizou (2008) holds, the corotation torque will also scale with r_{s}^{-2}. This was verified by Paardekooper & Papaloizou (2009a). Because r_{s} is an arbitrary parameter introduced for convenience, a result that is strongly dependent on it indicates large uncertainty. In fact, this strong dependence on r_{s} most likely implies the 2D approximation is not suitable for the evaluation of the corotation torque. However, a description of the 3D corotation torque is lacking in the literature. Chapter 5 describes our efforts in resolving

this issue by analyzing the co-orbital flow around an embedded planet using high-resolution 3D simulations.

For more massive planets, where their co-orbital regions overlap with their effective Lindblad resonances, w becomes insensitive to r_s, and scales with the planet's Hill radius: $w \sim 2.5 r_H$ where $r_H = a(q/3)^{1/3}$ (Masset et al., 2006; Peplinski, 2008). This limit is reached when $r_H/h_0 \gtrsim 1$. We can infer from these results that the dynamics in the co-orbital region is only well approximated as 2D for massive planets, but not for small ones. We note that for a solar-mass star and a MMSN, $r_H/h_0 = 1$ occurs around $q \sim 10^{-4}$, or about a few Neptune-mass.

Regardless of the uncertainties, the corotation torque appears to have, at least in the simple case of an isothermal disk, a positive value. This is encouraging, since it can partially cancel with the differential Lindblad torque, and possibly save planets from the destructive type I migration. However, there is one more aspect to the corotation torque that ultimately diminishes its influence. As mentioned before, the corotation torque stems from the disk's effort to flatten the vortensity profile. This is achieved after the streamlines in the co-orbital region becomes thoroughly mixed, and then there would be no more corotation torque since ζ would become zero. This is referred to as "torque saturation". Ward (1991) suggested that one way to have a sustained corotation is to have a sufficiently fast viscous diffusion rate in the disk, so that the initial vortensity profile can be restored. The saturation process is controlled by the parameter:

$$p = \frac{2}{3}\sqrt{\frac{w^3 \Omega_p}{2\pi a \nu}}, \tag{1.14}$$

which is a ratio between the libration time, or the synodic period of the widest horseshoe orbit, and the viscous diffusion time across the co-orbital region (Casoli & Masset, 2009; Paardekooper et al., 2011). When $p \gg 1$, viscosity is weak, and saturation is expected to occur. This result is once again subjected to the 2D approximation, where the co-orbital flow is necessarily separated from the rest from the disk, since orbits cannot cross in 2D. In 3D, the vertical dimension provides an additional degree of freedom to the fluid's motion, and so the co-orbital flow is allowed to mix into the disk via meridional motion and vice-versa. Such mixing would effectively serve the same purpose as viscous diffusion, leading to a sustained corotation torque. This is another topic that we look into in Chapter 5.

1.4 Using Graphics Processing Units for Scientific Computing

Following the rapid development in computing technology in recent decades, using numerical methods to solve complex, nonlinear systems of equations has become an important part of scientific research, . In astrophysics, it is complementary to astronomical observations by providing the connection between the information we gather from celestial objects and the physical mechanisms that drive their evolution. In this area of work, the amount of computational resources available in many ways determines the level of detail and complexity in the research itself. The demand for high-performance computing (HPC) has led to the construction of massive CPU (central processing unit) clusters with typically tens of thousands, up to millions, of computational cores. For example, the SciNet General Purpose Cluster (GPC) located at the University of Torontos SciNet HPC facility has 30912 cores, and a theoretical peak performance of 313 TFLOPs (10^{12} floating-point-operation-per-second). It was Canada's fastest supercomputer at its inception, and is the 3rd fastest now. While these numbers are impressive, such a massive amount of resource is not meant to serve a single purpose. In fact, thousands of researchers in all fields of science across the country share this tremendous power. Due to limits in availability and wallclock time (48 hours in the case of the SciNet GPC), the amount of resource per capita should not be over-estimated.

In the past decade, a new branch of technology has grown to challenge the traditional role of CPU in HPC. It is called general-purpose computing on graphics processing units (GPUs). Beginning as devices that handle

computation for computer graphics only, GPUs have evolved to become powerful computing devices in their own right, thanks to an ever increasing demand from the gaming industry. The graphics card that we use, the GTX-Titan, was developed by Nvidia and released in 2013. This card packs 2688 cores, and has a peak performance of 1.5 TFLOPs. At a release price of 999 USD, these cards are a cheap alternative to using CPU clusters. For example, with a budget of 5000 dollars, one can purchase at least three GTX-Titans along with a functioning desktop computer, plug in the cards, and enjoy a speed of 4.5 TFLOPs (equivalent to 1.4% of SciNet GPC) all to oneself. At a cost-to-speed ratio of $\sim$ 1000 dollars per TFLOPs, this one node of GPU machinery is about 100 times more cost effective than SciNet GPC, which costed of order 10 million dollars to build, not to mention the cost of maintenance and electricity. Also, the ownership that comes with a private GPU machine is particularly valuable during the development stage of a research project, when one performs numerous test cases to find the optimal parameters and settings.

The difference between GPUs and CPUs can be traced back to their origins. GPUs began as parallel devices, because their original purpose was to continuously update all of the pixels on a display. So it is not surprising that the ideal structure for them is to have many cores, each having fast access to a separate, small set of memory, i.e. a pixel. CPUs, on the other hand, are the main computing device on a computer, and so they are responsible for a handful of main tasks at any given time. To finish those tasks as quickly as possible, their development has been focused on building the fastest core possible.

Even though a CPU core is typically much faster than a GPU core, what GPU lacks in speed, it makes up in number. If a GPU is fully utilized, its speed can be 10 or even 100 times that of a CPU. This is GPU's strength, but also its weakness, because it means a GPU is only fast if the task is massively parallelized, like updating pixels on a display. Fortunately, hydrodynamics simulation is one of such tasks that can be massively parallelized. In the following chapter, I will describe a new GPU hydrodynamics code called `PEnGUIn` that we developed over my graduate career. `PEnGUIn` is used extensively in all of the research topics described in this thesis, and has proven to be fast, robust, and reliable.

Chapter 2

PEnGUIn: A GPU Hydrodynamics Code

2.1 Introduction

Motivated by the advantages GPU has over traditional CPU clusters, as discussed in Section 1.4, we developed the GPU hydrodynamics code `PEnGUIn` (**P**iecewise Parabolic Hydro-code **En**hanced with **G**raphics Processing **U**nit **I**mplementatio**n**). Although `PEnGUIn` is not a translation of any existing CPU hydrodynamics code, its main solver employs the piecewise parabolic method (PPM; Colella & Woodward 1984) in the Lagrangian frame, and so it does resemble any existing code that uses the same method, such as `VH-1` (Blondin & Lufkin, 1993). The main difference is, of course, `PEnGUIn` is written in the GPU language CUDA C, and is formulated under a GPU algorithm that optimizes the use of the thousands of cores available in a graphics card.

In this chapter, we will first give a description of the numerical method we use (Section 2.2), which is a combination of that documented by Colella & Woodward (1984) and Blondin & Lufkin (1993), with some modifications in the flattening procedure (Section 2.2.3) and an original module for viscosity implementation (Section 2.2.4). In our description, we will mainly focus on the case of a Cartesian grid, but we have generalized our grid geometry to cylindrical and spherical coordinates as well. After that, we will discuss `PEnGUIn`'s GPU algorithm, including the division of labor among computational cores, and the management of data in GPU's hierarchical memory structure (Section 2.3). We will measure `PEnGUIn`'s speed, both when running on a single GPU, and on multiple GPUs simultaneously (Section 2.4). Finally, we will present the results from a number of hydrodynamics tests, which characterize `PEnGUIn`'s performance under a variety of circumstances, such as in the presence of shock waves, mixing flow, viscous diffusion, and the gravitational field of a planet (Section 2.5).

2.2 Numerical Method

In its most basic form, `PEnGUIn` solves Euler equations in the Lagrangian frame:

$$\frac{\mathrm{D}\rho}{\mathrm{D}t} + \rho(\nabla \cdot \mathbf{u}) = 0\,, \tag{2.1}$$

$$\frac{\mathrm{D}u}{\mathrm{D}t} + \frac{\nabla p}{\rho} = \mathbf{f}\,, \tag{2.2}$$

$$\frac{\mathrm{D}e}{\mathrm{D}t} + \frac{\nabla \cdot (\mathbf{u}p)}{\rho} = \mathbf{u} \cdot \mathbf{f}\,, \tag{2.3}$$

where $\mathrm{D}/\mathrm{D}t$ is the Lagrangian derivative; ρ is the density of the fluid, $\mathbf{u}$ the velocity, e the total specific energy, p the pressure, and $\mathbf{f}$ is the specific force acting on the fluid. These variables have the following relations:

$$e = \frac{p}{(\gamma-1)\rho} + \frac{|\mathbf{u}|^2}{2}\,, \tag{2.4}$$

$$p = p_0\left(\frac{\rho}{\rho_0}\right)^{\gamma}\,, \tag{2.5}$$

where γ is the ratio of specific heats, also known as the adiabatic index; p_0 and ρ_0 are constants. In Navier-Stokes equations, viscous stress in the fluid produces the external force:

$$\mathbf{f} = \frac{1}{\rho}\boldsymbol{\nabla}\cdot\mathbb{T}\,, \tag{2.6}$$

where $\mathbb{T}$ is the viscous stress tensor. To solve Equations 2.1 to 2.3, we first split the problem dimensionally: we treat the 3D problem as 3 alternating 1D problems. While this dimensionally-split approach greatly reduces the complexity of the problem, it does introduce some limitations to the code. For instance, flow symmetry not parallel to the coordinate axes will not be well-maintained. It is therefore particularly important for the users of a dimensionally-split code to choose a suitable coordinate system for the problems they wish to solve.

For the effective 1D problem, we solve the Riemann problem between individual grid cells in Lagrangian frame through the PPM reconstruction of cell boundary quantities, and remap the Lagrangian quantities back to the static grid via an advection scheme. In Section 2.2.1 we give a description of our Riemann solver; in Section 2.2.2, we explain the remapping scheme; in Section 2.2.3, we discuss the PPM reconstruction method; and finally, in Section 2.2.4, we describe our implementation of viscosity in the hydrodynamics problem.

2.2.1 Riemann Solver

The PPM method for solving hydrodynamics equations in Lagrangian frame is discussed in detail in Section 2 of Colella & Woodward (1984). Here we reiterate the central notions of the method. We use subscript "i" to denote the discrete spatial index, and superscript "n" as the discrete temporal index. A 2^{nd} order 1D finite-volume approximation of Equations 2.1 to 2.3 is:

$$\rho_i^{n+1} = \rho_i^n \frac{\Delta V_i^n}{\Delta V_i^{n+1}}\,, \tag{2.7}$$

$$u_i^{n+1} = u_i^n + \frac{\Delta t}{2}(f_i^n + f_i^{n+1}) - \frac{\Delta t}{2\rho_i^n \Delta V_i^n}(A_{i+1/2}^{n+1/2} + A_{i-1/2}^{n+1/2})(p_{i+1/2}^{n+1/2} - p_{i-1/2}^{n+1/2})\,, \tag{2.8}$$

$$e_i^{n+1} = e_i^n + \frac{\Delta t}{2}(u_i^n f_i^n + u_i^{n+1} f_i^{n+1}) - \frac{\Delta t}{\rho_i^n \Delta V_i^n}(A_{i+1/2}^{n+1/2} u_{i+1/2}^{n+1/2} p_{i+1/2}^{n+1/2} - A_{i-1/2}^{n+1/2} u_{i-1/2}^{n+1/2} p_{i-1/2}^{n+1/2})\,, \tag{2.9}$$

where Δt is the length of a timestep, ΔV is the volume of a grid cell; A is a geometric factor that equals to unity for Cartesian coordinates, and depends on the curvature of the coordinate in general. ρ_i, u_i, and e_i are averages over the i^{th} cell. The superscript "n+1/2" means the quantity is evaluated at half a timestep, and the subscript "i+1/2" refers to the cell boundary between the cells "i" and "i+1". Since this is in Lagrangian frame, the cell boundaries also moves over one timestep:

$$x_{i+1/2}^{n+1} = x_{i+1/2}^n + \Delta t\, u_{i+1/2}^{n+1/2}\,. \tag{2.10}$$

Δt is evaluated following the Courant-Friedrichs-Lewy condition: $\Delta t < C_{\mathrm{CFL}} \min(\Delta x_i/[|u_i| + c_{\mathrm{s},i}])$ for all i, where $\Delta x_i = x_{i+1/2} - x_{i-1/2}$ is the cell size, and $C_{\mathrm{CFL}} < 1$ is the Courant number. This condition is necessary in order to

prevent fluid from entering cells beyond its immediate neighbors within one timestep. We choose $C_{\rm CFL} = 0.5$ so that the Riemann problem we construct in this section at the cell boundaries will not interfere with one another.

Now we describe the Riemann problem we solve in order to advance cell quantities in time. It is evident from Equations 2.7 to 2.10 that, in addition to information from the previous timestep, $u_{i+1/2}^{n+1/2}$ and $p_{i+1/2}^{n+1/2}$ are the two necessary quantities to evaluate. We solve for them by constructing a Riemann problem, where the left (denoted by subscript "L") state is the spatial average over a region in the $i^{\rm th}$ cell, defined by how far sound waves launched from the boundary can reach over a time of Δt, and the right state (denoted by subscript "R") state is the equivalent spatial average in the $(i+1)^{\rm th}$ cell:

$$a_{\rm i+1/2,L} = \frac{1}{\Delta t\ c_{\rm s,i}} \int_{1-\Delta t\ c_{\rm s,i}}^{1} a_i(\xi)\ {\rm d}\xi\,, \tag{2.11}$$

$$a_{\rm i+1/2,R} = \frac{1}{\Delta t\ c_{\rm s,i+1}} \int_{0}^{\Delta t\ c_{\rm s,i+1}} a_{i+1}(\xi)\ {\rm d}\xi\,, \tag{2.12}$$

where a refers to density ρ and pressure p, $c_{\rm s} = \sqrt{\gamma p/\rho}$ is the adiabatic sound speed, and $\xi \equiv (x - x_{i-1/2})/\Delta x_i$. For the velocity u, it is additionally modified by the external force:

$$u_{\rm i+1/2,L} = \frac{1}{\Delta t\ c_{\rm s,i}} \int_{1-\Delta t\ c_{\rm s,i}}^{1} u_i(\xi)\ {\rm d}\xi + \frac{\Delta t}{2} f_i^n\,, \tag{2.13}$$

$$u_{\rm i+1/2,R} = \frac{1}{\Delta t\ c_{\rm s,i+1}} \int_{0}^{\Delta t\ c_{\rm s,i+1}} u_{i+1}(\xi)\ {\rm d}\xi + \frac{\Delta t}{2} f_{i+1}^n\,. \tag{2.14}$$

The continuous spatial function $a(\xi)$ is described by a piecewise parabolic interpolation:

$$a(\xi) = a_{\rm l} + \xi[\Delta a + (1-\xi)a_6]\,. \tag{2.15}$$

The evaluation of the coefficients $a_{\rm l}$, Δa, and a_6 follows the reconstruction method described by Blondin & Lufkin (1993). We give a discussion on their method in Section 2.2.3.

To solve this Riemann problem in Lagrangian frame, we rewrite the hydrodynamics equations in matrix form:

$$\frac{\rm D}{{\rm D}t}\begin{bmatrix}\rho\\ u\\ e\end{bmatrix} + \begin{bmatrix}0 & \rho & 0\\ \frac{p}{\rho^2} & -(\gamma-1)u & \gamma-1\\ \frac{up}{\rho^2} & \frac{p}{\rho}-(\gamma-1)u^2 & (\gamma-1)u\end{bmatrix}\frac{\partial}{\partial x}\begin{bmatrix}\rho\\ u\\ e\end{bmatrix} = 0\,. \tag{2.16}$$

The two eigenvalues of this set of equations, $\pm c_{\rm s}$, describe the two characteristics of the system: one left-ward, and one right-ward propagating sound wave. Therefore $u_{i+1/2}^{n+1/2}$ and $p_{i+1/2}^{n+1/2}$ are superpositions of the left-ward propagating wave from the right state, and the right-ward wave from the left state. We express this using a discrete form of Equation 2.2, where ${\rm d}x/{\rm d}t = \pm c_{\rm s}$ for the left ($+c_{\rm s}$) and right ($-c_{\rm s}$) states:

$$u - u_{\rm L} + \frac{p - p_{\rm L}}{\sqrt{\gamma p_{\rm L}\rho_{\rm L}\left(1 + \frac{\gamma+1}{2\gamma}\left(\frac{p}{p_{\rm L}} - 1\right)\right)}} = 0\,, \tag{2.17}$$

$$u_{\rm R} - u - \frac{p_{\rm R} - p}{\sqrt{\gamma p_{\rm R}\rho_{\rm R}\left(1 + \frac{\gamma+1}{2\gamma}\left(\frac{p}{p_{\rm R}} - 1\right)\right)}} = 0\,, \tag{2.18}$$

where we have omitted the superscript $n+1/2$ and subscript $i+1/2$ for clarity. Generally, these two equations can be solved iteratively for $p_{i+1/2}^{n+1/2}$ and $u_{i+1/2}^{n+1/2}$. A special case is when $\gamma = 1$, corresponding to an isothermal equation

of state. This simplifies Equations 2.17 and 2.18 into a quadratic form:

$$p + \frac{(u_\mathrm{R} - u_\mathrm{L})\sqrt{\rho_\mathrm{L}\rho_\mathrm{R}}}{\sqrt{\rho_\mathrm{L}} + \sqrt{\rho_\mathrm{R}}}\sqrt{p} - \frac{p_\mathrm{L}\sqrt{\rho_\mathrm{R}} + p_\mathrm{R}\sqrt{\rho_\mathrm{L}}}{\sqrt{\rho_\mathrm{L}} + \sqrt{\rho_\mathrm{R}}} = 0\,, \tag{2.19}$$

which can be solved exactly. Plugging these values back into Equations 2.7 to 2.9, we obtain the updated Lagrangian values for our conservative quantities. However, as shown by Equation 2.10, the cell boundaries have also shifted in this process. Therefore, the next step is to remap the updated values back to the starting cell locations.

2.2.2 Remapping Scheme

We described in the previous section our method of solving the hydrodynamics problem in Lagrangian frame, and in this section we document our method of remapping that solution to our Eulerian grid, which we denote with the subscript "E". In other words, we evaluate the cell average of the conservative quantities ρ, u, and e at $x_{\mathrm{i,E}}$, given their values at the shifted Lagrangian cells as described by Equations 2.7 to 2.10. For ρ, we evaluate the cell-volume average:

$$\rho_{\mathrm{i,E}} = \frac{1}{\Delta V_{\mathrm{i,E}}}\left(\rho_i \Delta V_i + f_i^\rho \Delta V_{\mathrm{i,remap}} - f_{i+1}^\rho \Delta V_{\mathrm{i+1,remap}}\right), \tag{2.20}$$

where f_i^ρ is the amount of mass entering the i^{th} Eulerian grid cell, and $\Delta V_{\mathrm{i,remap}}$ is the volume enclosed by the boundaries x_i and $x_{\mathrm{i,E}}$. Similar to Equations 2.11 and 2.12, f_i^ρ is a spatial average of the piecewise parabolic function $\rho_i(\xi)$ (Equation 2.15):

$$f_i^\rho = \begin{cases} \frac{1}{y}\int_{1-y}^{1} \rho_{\mathrm{i}-1}(\xi)\,\mathrm{d}\xi & \text{if} x_i > x_{\mathrm{i,E}} \\ \frac{1}{y}\int_{0}^{y} \rho_i(\xi)\,\mathrm{d}\xi & \text{if} x_i < x_{\mathrm{i,E}} \end{cases} \tag{2.21}$$

where

$$y = \begin{cases} \frac{x_i - x_{\mathrm{i,E}}}{x_i - x_{\mathrm{i}-1}} & \text{if } x_i > x_{\mathrm{i,E}} \\ \frac{x_{\mathrm{i,E}} - x_i}{x_{i+1} - x_i} & \text{if } x_i < x_{\mathrm{i,E}} \end{cases} \tag{2.22}$$

For the other conservative quantities such as u and e, we evaluate the their cell-mass averages:

$$a_{\mathrm{i,E}} = \frac{1}{\rho_{\mathrm{i,E}}\Delta V_{\mathrm{i,E}}}\left(a_i \rho_i \Delta V_i + f_i^a f_i^\rho \Delta V_{\mathrm{i,remap}} - f_{i+1}^a f_{i+1}^\rho \Delta V_{\mathrm{i+1,remap}}\right), \tag{2.23}$$

where f_i^a is evaluated for the quantities u or e, in the same manner as how f_i^ρ is evaluated for ρ. Finally, we need to recover pressure p from the remapped energy e. Normally this is done simply by using Equation 2.4. However, there is a problem with this approach, first identified by Blondin & Lufkin (1993). When e is dominated by kinetic energy, this can lead to significant error because we only have the cell average values for u, not u^2. It is not sufficient to simply compute the cell average for u^2, because in multi-dimensional problems, kinetic energy depends on velocities in other dimensions as well, whereas our remapping procedure is only in one dimension. A correct treatment in multi-dimensional space would be too complex and computationally intensive. Therefore, we evade this problem by additionally computing the cell average for the internal energy $\frac{p}{(\gamma-1)\rho}$, so that in cases where the flow is highly supersonic, $|\mathbf{u}| > 10c_\mathrm{s}$, we can recover p from the internal energy instead. A trade-off with this approach is that energy conservation will be less accurate in highly supersonic scenarios.

Finally, we note that remapping should be applied on the appropriate conservative quantities specific to the problem at hand. For example, we remap angular, instead of linear, momentum when we perform calculations on rotating disks, such as those in Sections 2.5.5 and 2.5.6.

2.2.3 Reconstruction Method

We now discuss how we compute the interpolation coefficients a_l, Δa, and a_6 in Equation 2.15. a_l, the value at the cell boundary, is obtained through a quartic polynomial interpolation of the following integral:

$$A_{i-1/2} = \sum_{\mathrm{k<i}} a_\mathrm{k} \Delta V_\mathrm{k} \,, \tag{2.24}$$

through the points $x_{i-5/2}$, $x_{i-3/2}$, $x_{i-1/2}$, $x_{i+1/2}$, and $x_{i+3/2}$. After we obtain $A(x)$ as a continuous spatial function, a_l is simply:

$$a_\mathrm{l} = \left[\left(\frac{\partial V}{\partial x} \right)^{-1} \frac{\mathrm{d}A}{\mathrm{d}x} \right]_{x=x_{i-1/2}} \,, \tag{2.25}$$

and $\Delta a = a_{i+1,\mathrm{l}} - a_{i,\mathrm{l}}$ is the difference in a across one cell. a_6 is related to a_l and Δa by:

$$a_6 = 6(a - a_\mathrm{l} - \Delta a/2) \,, \tag{2.26}$$

The interpolation used to calculate a_l follows the method by Blondin & Lufkin (1993); we refer the reader to their paper for the derivation because the technicality involved is beyond the scope of this chapter. We remark that their method was improved upon the one given by Colella & Woodward (1984), who performed the quartic interpolation on volume, that is, A is described as a function of $\int \mathrm{d}V$. Blondin & Lufkin (1993) showed that in curvilinear coordinates, this is inferior to interpolating directly on the coordinate, writing A as a function of x.

This interpolation method is 3$^{\mathrm{rd}}$ order in general (4$^{\mathrm{th}}$ for uniform cell sizes). This is often referred to as "3$^{\mathrm{rd}}$ order in space", implying that in the limit where $\partial/\partial t = 0$, the error due to spatial reconstruction alone is of order Δx^3. In practice, we do not expect 3$^{\mathrm{rd}}$ order convergence rate for any meaningful time-dependent problems, because Equations 2.7 to 2.9 are 2$^{\mathrm{nd}}$ order in general. Instead, this high-order reconstruction method ensures that the error from the reconstruction procedure is small compared to the error from time-evolution, increasing the overall accuracy of the code.

The quartic interpolation needs to be modified in two ways before being used. First, we include the conditions for monotonicity also derived by Blondin & Lufkin (1993). Monotonicity is important because new extrema arising from interpolation would lead to spurious high frequency oscillations in the numerical solution. Second, we include a flattening procedure that lowers the order of the interpolation when a discontinuity is detected. Without the flattening procedure, the discontinuity would be fit with an oscillatory polynomial, generating the artificial "post-shock oscillations". The flattening procedure was introduced by Colella & Woodward (1984), and `PEnGUIn` uses a modified version of it. We define a flattening parameter F, where $F = 0$ equals no flattening, and $F = 1$ reduces the interpolation to zeroth-order:

$$a_{\mathrm{l,flat}} = Fa + (1 - F)a_\mathrm{l} \,. \tag{2.27}$$

A discontinuity is present when 1) there is a rapid change in a conservative quantity and 2) the change is large compared to the value itself. These conditions are satisfied when the following ratios are large: $(a_{i+1} - a_{i-1})/(a_{i+2} - a_{i-2})$, and $|a_{i+1} - a_{i-1}|/a_i$. We therefore introduce the quantity F^a

$$F_i^a = 4 \frac{|a_{i+1} - a_{i-1}|}{a_i} \left(\frac{a_{i+1} - a_{i-1}}{a_{i+2} - a_{i-2}} - \frac{3}{4} \right) \,, \tag{2.28}$$

such that the flattening parameter F_i is given by:

$$F_i = \max(F^p_{i-1}, F^p_i, F^p_{i+1}, F^\rho_{i-1}, F^\rho_i, F^\rho_{i+1}), \tag{2.29}$$

which ensures the same flattening parameter is applied to the entire span over which the discontinuity is resolved. F_i is additionally restricted to be no less than 0 or larger than 1. F^a is devised such that a factor of 2 jump in either ρ or p over a distance of less than 2 cells results in $F^a = 1$. $F^a = 0$, i.e., the flow is considered smooth, when $(a_{i+1} - a_{i-1})/(a_{i+2} - a_{i-2}) < 3/4$, indicating a gradual change over the span of 5 cells. The test in Section 2.5.2 demonstrates the effectiveness of this procedure.

2.2.4 Viscosity Implementation

For a viscous fluid, $\mathbf{f}$ in Equation 2.2 includes the divergence of the viscous stress tensor. The Newtonian viscous stress tensor $\mathbb{T}$ in Cartesian coordinates can be expressed as:

$$\mathbb{T}_{kl} = \nu\rho\left(\frac{\partial u_k}{\partial x_l} + \frac{\partial u_l}{\partial x_k}\right), \tag{2.30}$$

where ν is the kinematic viscosity. We treat this viscous term with an operator splitting method: after the hydrodynamics step, we additional modify the velocity of the fluid following:

$$\mathbf{u} = \mathbf{u}_{\text{hydro}} + \frac{\Delta t}{\rho}\nabla \cdot \mathbb{T}, \tag{2.31}$$

where $\mathbf{u}_{\text{hydro}}$ is the velocity given by the hydrodynamics solver at the end of a timestep. To evaluate $\nabla \cdot \mathbb{T}$, we once again simplify based on the assumption of dimensional-splitting. When calculating along a given direction, we only include terms containing velocity in that direction. In other words, for the k^{th} direction, we let $u_l = 0$ if $l \neq k$. In Cartesian coordinate, this corresponds to:

$$(\nabla \cdot \mathbb{T})_k = 2\left(\nu\rho\frac{\partial^2 u_k}{\partial x_k^2} + \frac{\partial(\nu\rho)}{\partial x_k}\frac{\partial u_k}{\partial x_k}\right) + \sum_{k\neq l}\left(\nu\rho\frac{\partial^2 u_k}{\partial x_l^2} + \frac{\partial(\nu\rho)}{\partial x_l}\frac{\partial u_k}{\partial x_l}\right). \tag{2.32}$$

The derivatives are computed using a parabolic interpolation through the points x_{i-1}, x_i, and x_{i+1}. For the study of protoplanetary disks, the most important term in $\nabla \cdot \mathbb{T}$ is the one relating to the Keplerian shear that leads to viscous angular momentum transport. This term is maintained in our dimensionally-split approach.

Our treatment of the viscous term is 1^{st} order in time and 2^{nd} in space, which is less accurate than our hydrodynamics method. This is acceptable for the problems we are interested in, because the viscous force is typically much weaker than both gravity and pressure forces in protoplanetary disks, so the error incurred from this crude method will be insignificant.

2.3 GPU Algorithm

A GPU program's speed is largely determined by two aspects: parallelization and memory management. GPUs can simultaneously launch a massive number of parallel computational threads. For instance, the GTX Titan that we use houses 14 multiprocessors, each capable of launching 2048 computational threads, making a total of 28672 threads on just one card. To achieve maximum "occupancy" - fully utilizing all GPU cores - the number of parallel operations at any given time should be at least of order 10^4, and ideally much larger to ensure the amount

of work is evenly distributed among cores. The base unit for PEnGUIn's paralleization is a single grid cell; that is, every cell is assigned a unique thread responsible for updating quantities belong to that cell. When threads require information from neighboring cells, they can communicate via memory banks in "blocks", referring to a set of threads that are launched and terminated together. This leads to how memory is managed in PEnGUIn.

GPU has a hierarchical dynamic memory structure: there is global memory (GM), which can be accessed by all threads at any time; shared memory (SM), which can only be accessed by threads belong to the same block; and register space (RS), which is private to each individual thread. These levels have a varying degree of capacity and read/write speed: in terms of capacity, GM $\gg$ SM $\gg$ RS, but in terms of speed, GM $\ll$ SM $\ll$ RS. Additionally, only GM lasts for the duration of the program (e.g. one simulation), while SM and RS only last for the lifetime of the program kernel (i.e. one timestep). How data is stored, transferred, and read in these different levels of memory structure is a particularly important aspect of code optimization, because typically, computation on GPUs is sufficiently fast that the code's speed is in fact limited by memory bandwidth.

While the computational grid must be stored in the global memory, computation itself occurs within the register space. Since our code's speed is bandwidth-limited, optimized performance is achieved by minimizing the frequency of access to both global and shared memory. There are at least two instances of global memory access per timestep: we must download the grid information from GM to RS in the beginning of a timestep, update the cell variables, and then upload the information back from RS to GM. However, this simple procedure is not applicable to us because the Lagrangian PPM method described in Section 2.2 requires each cell to have access to information in 6 neighboring cells on each side, 12 in total, far exceeding the amount of memory available in the register space. To accommodate this, we utilize SM as a buffer between GM and RS. For all variables shared between neighboring cells, we store them in the SM. During computation, we create temporary variables used for computation in the RS . This way, memory in the RS can be recycled between intermediate steps, while the necessary information is kept safe in the SM.

One example of how we manage memory is our program for solving the Riemann problem described in Section 2.2.1. In essence, it is a process of using the variables ρ_i^n, p_i^n, and u_i^n to obtain the final products $p_{i+1/2}^{n+1/2}$, and $u_{i+1/2}^{n+1/2}$. Therefore, intermediate variables such as the left and right states of each Riemann problem, are not needed between neighboring cells. So, we keep them in the register space during computation, and delete them after the solutions $p_{i+1/2}^{n+1/2}$ and $u_{i+1/2}^{n+1/2}$ are obtained. In this process, we access the shared memory only 5 times: downloading ρ_i^n, p_i^n, and u_i^n; and uploading $p_{i+1/2}^{n+1/2}$ and $u_{i+1/2}^{n+1/2}$.

Recall that shared memory is accessible only by threads belong to the same block. This further complicates the matter because the amount of shared memory allocated per block cannot exceed a certain limit, which is 48KB for our hardware, which, for most purposes, is a very small amount comparing to the entire computational grid. We therefore divide the grid into a number of subgrids, each having a dimension of $L_\mathrm{b} \times L_\mathrm{b} \times L_\mathrm{b}$ (in 3D simulations), where L_b is the number of threads (grid cells) in a block. These subgrids are then split into L_b^2 blocks during computation.

We construct each block as a self-contained 1D grid containing boundary conditions, as opposed to having the blocks communicating and exchanging information with each other during runtime. This is because a GPU cannot launch an arbitrary number of blocks at once: it is limited by the number and capability of the GPU's multiprocessors. For our hardware. each card can launch at most 28 blocks at a time, so an excess number of blocks would be launched in a serial manner. As a result, blocks generally do not exist simultaneously even during the runtime of a kernel call, and each block must operate without knowledge of one another, meaning that each of them must include boundary cells that are overlapping with neighboring blocks. The actual computational domain in a block is therefore $N_\mathrm{b} = L_\mathrm{b} - 12$. PEnGUIn's speed strongly depends on the choice of N_b, as we will see in the

next section.

2.4 Code Speed

We mentioned in the previous section that memory management is an important aspect to achieving an optimal speed, and an important parameter in our algorithm is N_b, the size of a block minus the boundary cells. We want to choose a large N_b to minimize the ratio between the number of boundary cells and the total number of cells. If N_b is too large, however, there would be a shortage of shared memory. While this would not prevent the code from running, the GPU would have to use slower memory in place of shared memory. Consequently, there exists an optimal N_b that maximizes the code's speed. Figure 2.1 plots the code's speed, measured in the number of cells updated per second, in 3D simulations with different N_b's and total grid sizes, running on a single GTX-Titan card. It shows that $N_b \sim 100$ gives the best performance; a N_b much larger or smaller will lead to about a factor of 2 speed reduction. Figure 2.1 also illustrates a speed difference between an adiabatic equation of state and an isothermal one. This difference arises from two factors: 1) adiabatic calculations keep track of the total and internal energy separately, while isothermal ones only require the internal energy; and 2) the adiabatic Riemann solver is iterative, while the isothermal one is exact. We find that isothermal calculations are $\sim 20\%$ faster. Overall, `PEnGUIn` can reach speeds above 2.5×10^7 cells per second in 3D. This speed is 1.5 times faster for 2D simulations, and 3 times for 1D.

Figure 2.2 gives a breakdown of how much time is used to perform different tasks in a timestep. "Hydrodynamics solver" refers to the computation involved in the Riemann solver and remapping procedure, including the time spent on data transfer between SM and RS during computation. "Kernel launching and global memory access" refers to the setting up of the block structure and shared memory allocation for the hydrodynamics solver, and the time spent on GM access at the beginning and the end of a timestep. "Viscosity module" refers to the additional computation required for the operator-split viscosity treatment described in Section 2.2.4. "Timestep evaluation" refers to the searching of the entire grid to find the largest timestep that satisfies the Courant criterion. The amount of computation involved in timestep evaluation is insignificant compared to the hydrodynamics solver, but it does take a significant amount of time to complete because it requires initial access to GM. It is therefore not surprising that it takes about half the time compared to "Kernel launching and global memory access", which accesses GM twice per timestep. The same can be said for the viscosity module. Even though it is more computationally intensive, its speed is still limited by GM access, once to download grid information, and a second time to upload updated velocities. Consequently, all of the categories except "hydrodynamics solver" are limited by GM access, while "hydrodynamics solver" includes time consumption by both computation and SM access.

Finally, we can reach an even higher speed by running `PEnGUIn` on multiple GPUs in parallel. Advancement in the GPU technology has allowed GPUs on a single node to transfer data directly between one another via Peripheral Component Interconnect Express (PCIe), completely bypassing the CPU. This has significantly reduced the latency in multi-GPU applications. `PEnGUIn` is well-suited to run on parallel GPUs since our computational grid is already divided into subgrids, and each of them can be treated independently on different GPUs. The only additional task is to exchange boundary conditions between timesteps. Figure 2.3 shows the speedup factor `PEnGUIn` achieves on a 3-GPU node compared to a single GPU. In this setup, we simulate a 2D grid with $N_b = 100$, and so each subgrid contains N_b^2 grid cells. As we increase the number of subgrids, the efficiency of the multi-GPU system increases, because the amount of boundary cells exchanged between GPUs is proportional to the square root of the total number of grid cells for a 2D grid, so the "boundary-to-volume" ratio goes down

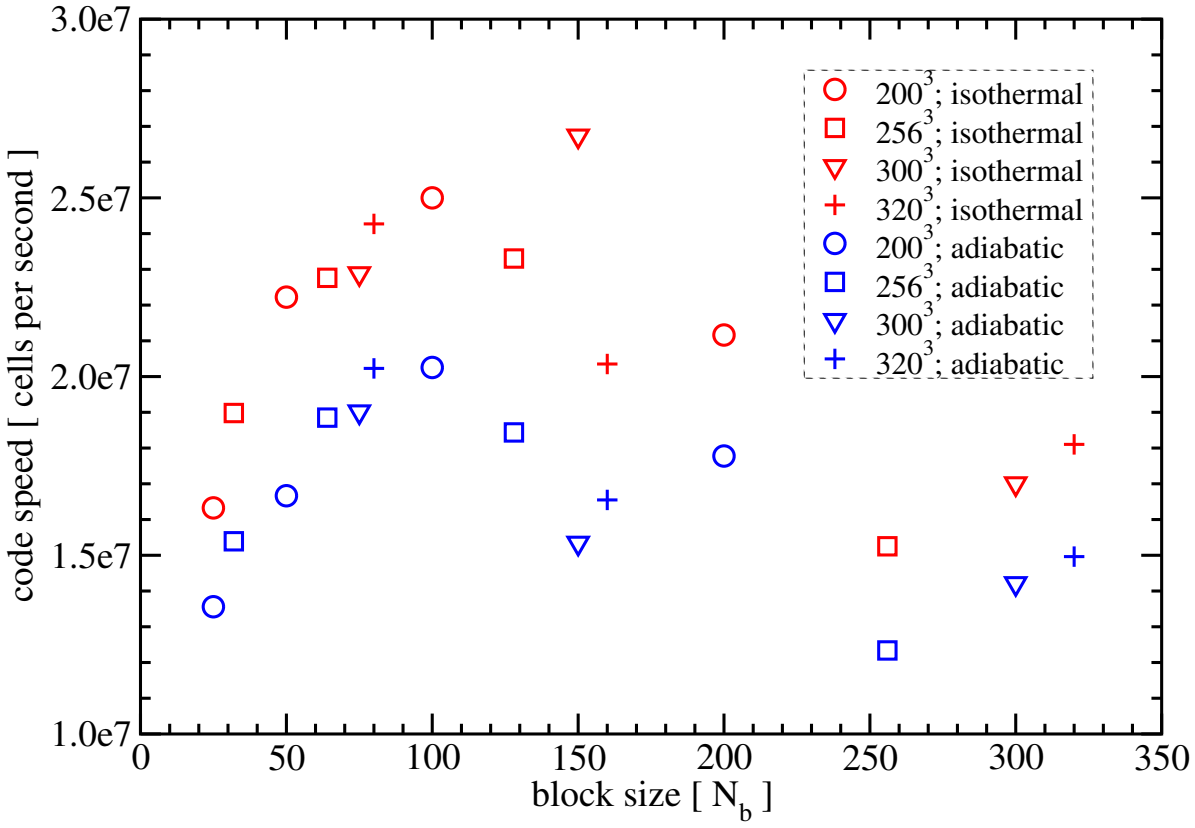

Figure 2.1 PEnGUIn's speed on a single GTX-Titan graphics card for a range of N_b. All simulations are performed in double precision, in 3D. Each symbol corresponds to a different total number of grid cells: circles have $200^3 = 8 \times 10^6$ cells, squares have $256^3 \sim 1.7 \times 10^7$ cells, triangles have $300^3 = 2.7 \times 10^7$ cells, and pluses have $320^3 \sim 3.3 \times 10^7$ cells. Red symbols are isothermal simulations, while blue ones are adiabatic. We find $N_b \sim 100$ to be an optimal choice for PEnGUIn.

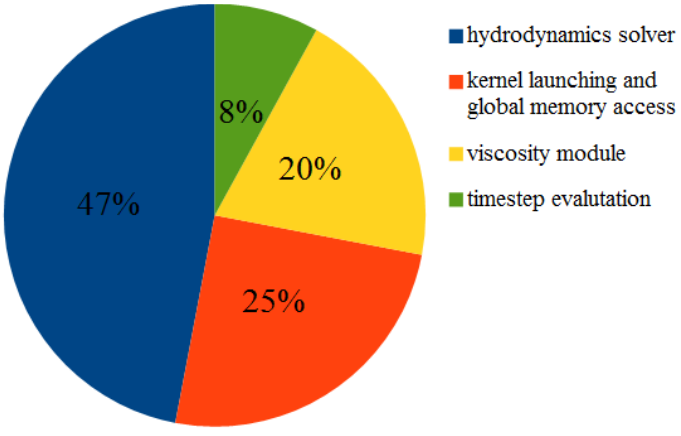

Figure 2.2 The distribution of time consumption by different tasks in PEnGUIn. See text in Section 2.4 for the definition of each category.

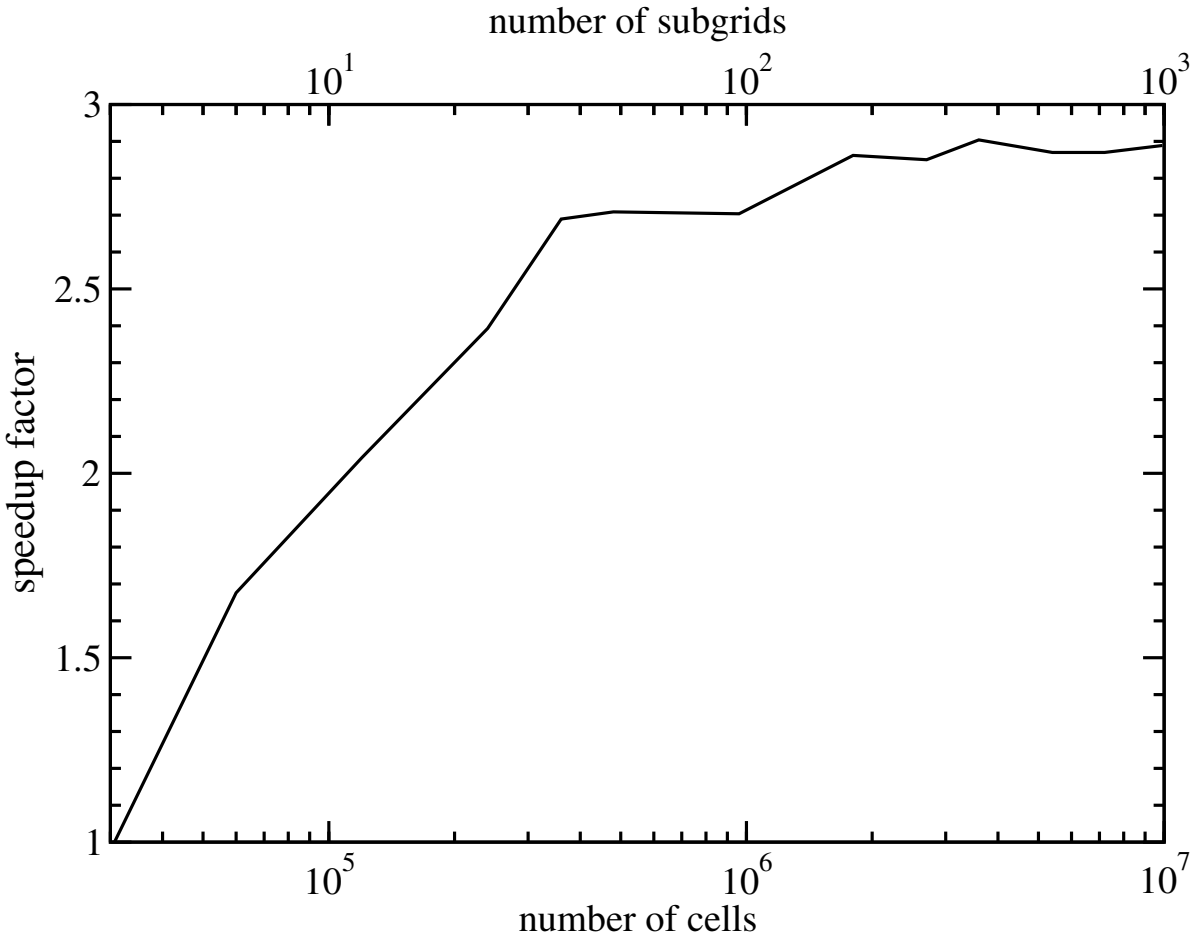

Figure 2.3 Speedup factor for a 3-GPU node compared to a single GPU as a function of the total number of grid cells. The speedup factor increases with the number of grid cells, and converges to a value close to 2.9, corresponding to a 97% efficiency, when the number of grid cells exceeds 2×10^6. The upper x-axis shows the corresponding number of subgrids, where each subgrid contains 10^4 cells.

as we increase the number of cells. When we simulate over 2×10^6 cells, `PEnGUIn` achieves a speedup factor of ~ 2.9, corresponding to an efficiency of $\sim 97\%$ for our 3-GPU node.

2.5 Tests

2.5.1 Sod Shock Tube

This standard 1D test was first devised by Sod (1978), and has since served as a standard test for hydrodynamics code. This test has the advantage of allowing one to directly compare simulation results to an semi-analytic solution for a nonlinear problem, and simultaneously testing the code's ability to resolve a propagating shock wave, contact discontinuity, and rarefaction wave. The initial condition for this test is a 1D Riemann problem. The left state ($x < 0$) has $\rho = 1$ and $p = 1$; and the right state ($x > 0$) has $\rho = 0.125$ and $p = 0.1$. We use an adiabatic index $\gamma = 1.4$. The velocity is initially zero everywhere. The semi-analytic solution to the Riemann problem can be found in standard textbooks (e.g. Toro, 2009).

Figure 2.4 plots the results at $t = 0.2$ for 3 simulations with different resolutions, together with the analytic

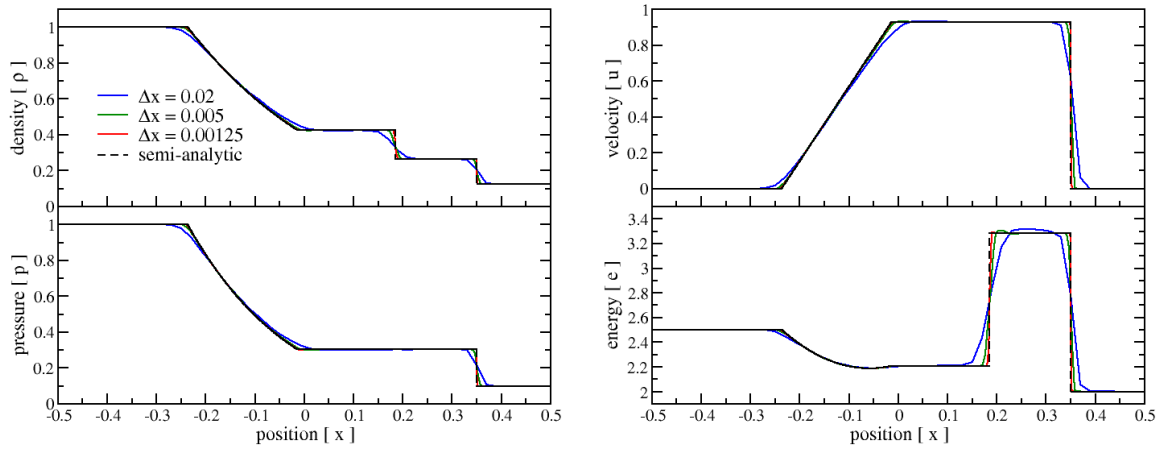

Figure 2.4 Simulation results at $t = 0.2$ for the Sod shock tube test performed with 3 different resolutions. The blue curve corresponds to 50 cells across the domain $-0.5 < x < 0.5$; green is 200 cells, and red is 800. The black-dashed curve is the semi-analytic solution shown for comparison. Note that the red and black curves are nearly on top of one another. The contact discontinuity propagating rightward is located at $x = 0.19$, the shock wave is at $x = 0.35$, and the rarefaction wave is between $x = -0.24$ and -0.014.

solution for comparison. In all cases we find `PEnGUIn` conserving mass to machine accuracy at all times, and correctly capturing the behavior of the system: a contact discontinuity moving rightward, with a shock front propagating ahead of it, and a rarefaction wave behind. The discontinuities do suffer a varying degree of spreading, which is lessened at higher resolutions. As we increase resolution, we find our results converge to the semi-analytic solution correctly.

2.5.2 Strong Shock

This is another 1D test similar to the Sod shock tube test in the previous section. This time we set up a strong pressure discontinuity in an initially stationary fluid with a constant density, so the initial conditions are $\rho = 1$, $u = 0$, $p = 1000$ if $x < 0$, and $p = 0.01$ if $x > 0$, with $\gamma = 1.4$. While it is similar to the previous test, this test magnifies possible numerical artifacts created near shocks and discontinuities. Figure 2.5 shows the simulation results along with the semi-analytic solution at $t = 0.012$. Once again, we find good agreement between our simulated results and the expected behavior of the system, and as we increase resolution, the simulations converge to the correct solution. However, this test does prove to be more demanding compared to the previous one, as we need to double our resolution in order to find a similar level of convergence.

Because of the presence of a very strong shock, this test also clearly illustrates the importance of the flattening procedure described in Section 2.2.3. Figure 2.6 shows the difference between two high-resolution runs: one lowers the order of interpolation at discontinuities, while the other one employs the quartic interpolation at all times. When flattening is not used, it is clear that some artificial fluctuations are introduced near the rarefaction wave and around the contact discontinuity. These artifacts do not weaken as resolution increases. With flattening, they are mostly eliminated.

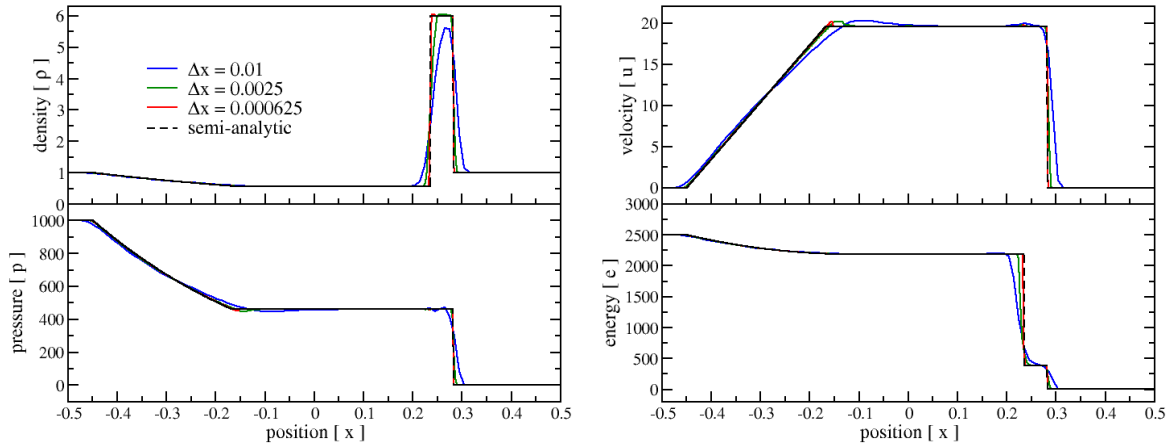

Figure 2.5 Simulation results at $t = 0.012$ for the strong shock test performed with 3 different resolutions. The blue curve corresponds to 100 cells across the domain $-0.5 < x < 0.5$; green is 400 cells, and red is 1600. They have doubled resolutions compared to the simulations in Figure 2.4. The black-dashed curve is the semi-analytic solution shown for comparison. A close inspection of the blue and green curves shows that the shock wave at $x = 0.28$ is better resolved than the contact discontinuity at $x = 0.23$.

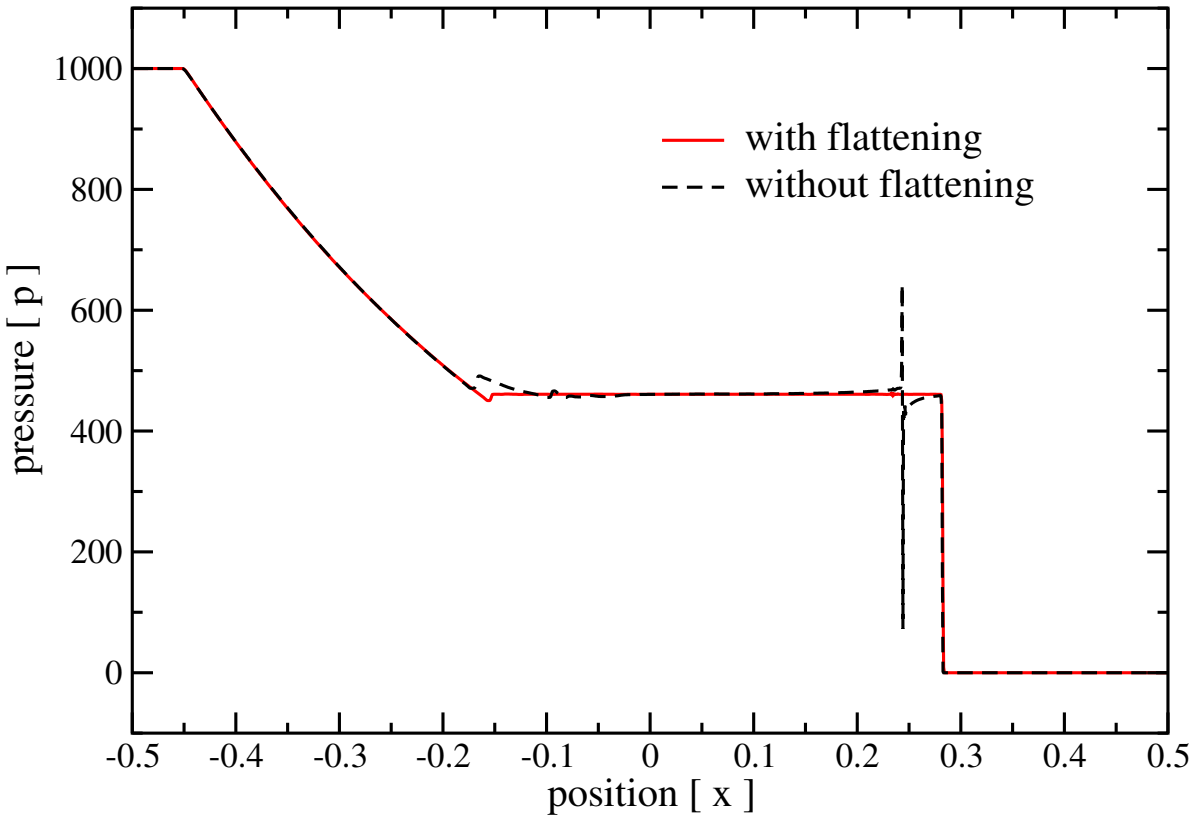

Figure 2.6 Simulation results at $t = 0.012$ for the strong shock test performed with and without the flattening procedure described in Section 2.2.3. The red curve is the same as the one in Figure 2.5, and the dashed black curve is from a simulation of the same resolution, but without any flattening. The error in the dashed black curve near the rarefaction wave ($x \sim 0.16$) and the contact discontinuity ($x \sim 0.23$) are clearly visible. Flattening is able to largely remove these errors, as shown by the red curve.

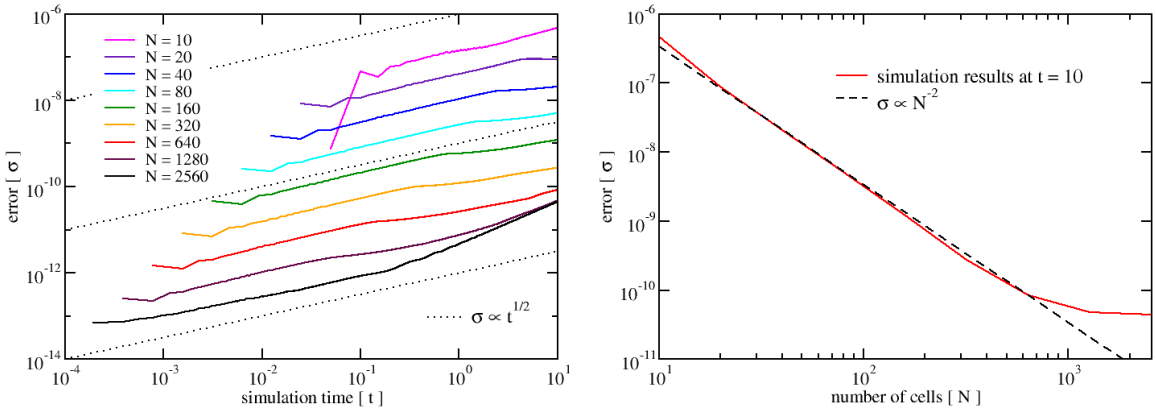

Figure 2.7 The error σ in our linear wave test as a function of simulation time on the left panel, and a function of the number of grid cells on the right. σ is described by Equation 2.33. On the left panel, each solid curve represents a simulation of a different resolution, and the dotted black lines indicate constant power-law slopes of $1/2$. On the right, the solid red curve shows how the error, measured at $t = 10$, reduces as we increase resolution, and the dashed black line has a constant power-law slope of -2. Note that the time for one wave cycle is 1.

2.5.3 Linear Wave

Unlike the previous two tests, this test does not include any discontinuity, and instead seek to evaluate the accuracy of the code when simulating a smooth, continuous flow, by simulating a 1D linear wave embedded in a stationary background fluid. This test is valuable in that it allows us to directly measure the order of the code, which is not possible with the previous two tests since the code is effectively 1$^{\text{st}}$ order at discontinuities. The fluid in this system is described as $\rho_{\text{lin}} = 1 + A \sin 2\pi(x - c_s t)$, $u_{\text{lin}} = A \sin 2\pi(x - c_s t)$, and $e_{\text{lin}} = (1 + \gamma A \sin 2\pi(x - c_s t))/(\gamma^2 - \gamma)$, where $A \ll 1$ is the amplitude of the linear wave. We set $A = 10^{-6}$ and the adiabatic index $\gamma = 1.4$. Note that given our set up, $c_s = 1$ is the background sound speed. The simulation domain is $x = \{0, 1\}$ with periodic boundary conditions, and we keep track of the absolute error σ:

$$\sigma = \sqrt{\frac{1}{N} \sum_i \left([\rho_i - \rho_{\text{lin}}]^2 + [u_i - u_{\text{lin}}]^2 + [e_i - e_{\text{lin}}]^2 \right)}, \tag{2.33}$$

where N is the number of grid cell. We run 9 simulations of varying resolution over 10 wave cycles, and the results are shown in Figure 2.7. It plots σ both as a function of time on the left panel, and as a function of resolution on the right. Our results indicate that σ grows in time as $t^{1/2}$, and reduces with increasing resolution as N^{-2}, which suggests the code, as expected, is 2$^{\text{nd}}$ order for continuous flow. We find that the error is mainly in the amplitude of the wave, while the phase is precisely maintained. This means the trends we observe in σ also applies to numerical diffusion. When $N > 1000$, σ converges to a single value. We believe this is not a behavior of the code: it is the level in the perturbation where the linear assumption breaks down, and so it is an error in ρ_{lin}, u_{lin}, and e_{lin}, rather than the simulations themselves.

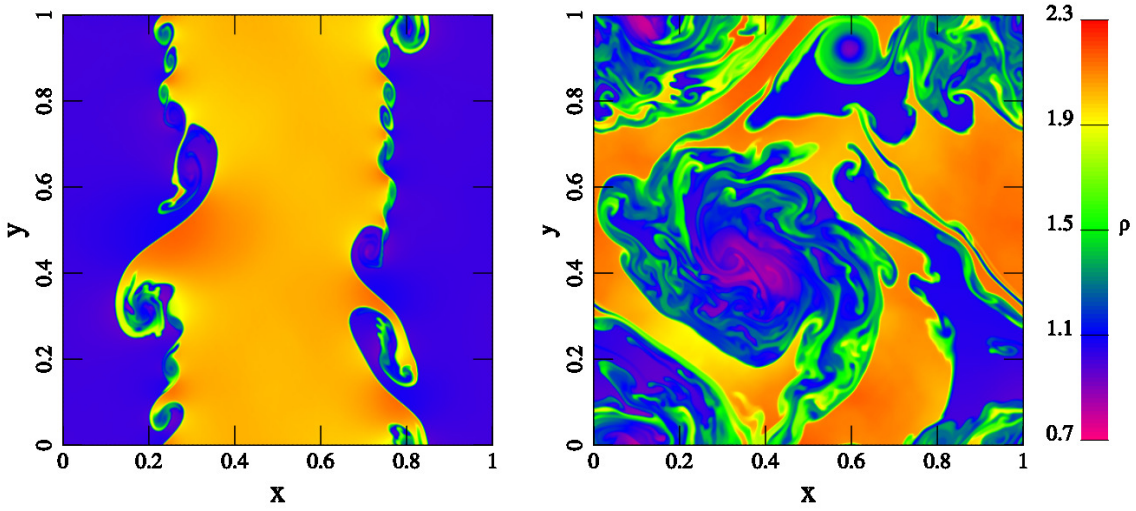

Figure 2.8 Snapshots of our simulated Kelvin-Helmholtz instability, showing the fluid density in a color-scale. The left panel is a snapshot at $t = 1.5$, when the vortex roll-up at the shear interface first begin to manifest. The right panel is at $t = 5$, when a fully non-linear turbulence has been established.

2.5.4 Kelvin-Helmholtz Instability

Now that we have established `PEnGUIn`'s accuracy and behavior in 1D problems, we move on to a 2D problem that tests its ability to resolve mixing flow. In this test, we set up an initial condition where Kelvin-Helmholtz instability, an instability triggered by the shear of two fluids and results in turbulent flow, is expected to operate. The setup is as follows: we simulate a 1 by 1 x-y domain with periodic boundary conditions, and within this domain we have $\rho = 2$, $u_y = 0.5$ if $0.25 < x < 0.75$; and $\rho = 1$, $u_y = -0.5$ otherwise. The transverse velocity is $u_x = 0$ everywhere, and the pressure has a uniform value of $p = 2.5$ with $\gamma = 1.4$. We aid the instability by adding a small initial velocity perturbation to this setup, so that $\mathbf{u}_{\text{perturbed}} = \mathbf{u} + 10^{-4}\sin(2\pi y)(\hat{\mathbf{x}} + \hat{\mathbf{y}})$.

Figure 2.8 plots two simulation snapshots of the density distribution at different times. This simulation has a resolution of 500×500 cells. The left panel shows a snapshot at $t = 1.5$. `PEnGUIn` correctly captures the vortex roll-up expected to appear at the interface of the shearing fluids. Vortex cores of sizes larger than ~ 10 cells are well resolved. By $t = 5$, the system has evolved to a turbulent state. We find the division between high and low density fluid remains sharp, and the vortices remain strong, showing little signs of numerical diffusion. As mentioned in Section 2.2, one of `PEnGUIn`'s weakness is its dimensionally-split approach that produces more numerical diffusion to velocities not parallel to the coordinate axes. This test demonstrates that this weakness is not a severe one: highly complex flows, such as vortices and turbulent motions seen here, are adequately generated and resolved.

2.5.5 Viscous Ring

For the study of protoplanetary disks, it is particularly crucial that `PEnGUIn` can accurately simulate the behavior of a viscous disk orbiting around a central mass. The case that we investigate here is the spreading of an initially

thin ring. One can approximate the profile of such a ring as it evolves in time if one can neglect gas pressure. An infinitesimally thin ring is described by:

$$\Sigma_{\rm ring}(r, t=0) = \frac{M_{\rm ring}}{2\pi r_0^2}\delta(r-r_0)\,, \tag{2.34}$$

where $M_{\rm ring}$ is the total mass of the ring, r_0 is its radial location, and δ is Dirac's delta function. This ring, neglecting gas pressure, rotates at the Keplerian speed $v_\phi = \sqrt{GM_*/r}$, where GM_*, the gravitational constant times the mass of the central object, is set to 1. Then the net force acting on this ring is $\boldsymbol{f} = -GM_*/r^2 + (\boldsymbol{\nabla}\cdot\mathbb{T})/\rho$, and the viscous term can be approximated as:

$$\begin{aligned}(\boldsymbol{\nabla}\cdot\mathbb{T})_\phi &\approx \nu\Sigma\left(\frac{\partial^2 v_\phi}{\partial r^2} + \frac{1}{r}\frac{\partial v_\phi}{\partial r} - \frac{v_\phi}{r^2}\right) + \frac{\partial(\nu\Sigma)}{\partial r}\left(\frac{\partial v_\phi}{\partial r} - \frac{v_\phi}{r}\right),\\ &= -\sqrt{\frac{GM_*}{r^3}}\left(\frac{3}{4}\frac{\nu\Sigma}{r} + \frac{3}{2}\frac{\partial(\nu\Sigma)}{\partial r}\right),\end{aligned} \tag{2.35}$$

where we have assumed the Keplerian shear is the dominant factor in the evaluation of the viscous stress, and dropped all other terms. This shows that at the edges of the ring, where $\partial\Sigma/\partial r$ is a dominating factor, the viscous torque is negative at the inner edge, and positive at the outer, so the fluid is expected to lose angular momentum at the inner edge and fall toward the center, while the outer edge will get gain angular momentum and rise to a higher orbit, resulting in the spreading of the ring. Inserting this into Equation 2.2, and solving together with Equation 2.1, one can find that $\Sigma_{\rm ring}$ evolves as:

$$\Sigma_{\rm ring}(x,\tau) = \frac{M_{\rm ring}}{\pi r_0^2}\frac{e^{-(1+x^2)/\tau}}{x^{\frac{1}{4}}\tau}I_{1/4}\left(\frac{2x}{\tau}\right), \tag{2.36}$$

where $x = r/r_0$ and $\tau = (12\nu/r_0^2)t$. We will use this expression to compare with our simulation.

For our numerical setup, we use a 2D cylindrical grid spanning the full 2π azimuthal and 0.5 to 1.5 in radius. The grid size is 500×500 cells. We set up a background disk that has a constant surface density $\Sigma_0 = 1$. Our unit for distance is $r_0 = 1$, and so $\Omega_0^{-1} \equiv (GM_*/r_0^3)^{-1/2} = 1$ is our unit for time. On top of this background disk, we add a thin ring that we expect to spread due to the effect of viscous diffusion. Since one cannot set up an infinitesimally thin ring in a grid with a finite cell size, The ring is initialized as $\Sigma_{\rm ring}(\tau = 0.03^2)$, which is well approximated by a Gaussian profile with a width of $0.03r_0$. The equation of state is isothermal ($\gamma = 1$), and we choose a low sound speed, $c_{\rm s} = 0.01\ r_0\Omega_0$ to ensure a good match between our simulation and Equation 2.36. The initial velocity profile is $\boldsymbol{u} = (v_{\rm r}, v_\phi)$, where:

$$v_{\rm r} = -\frac{3\nu}{r}\left(\frac{{\rm d}\ln\Sigma}{{\rm d}\ln r} + \frac{1}{2}\right), \tag{2.37}$$

$$v_\phi = \sqrt{\frac{GM_*}{r} + c_{\rm s}^2\frac{{\rm d}\ln\Sigma}{{\rm d}\ln r}}\,, \tag{2.38}$$

and $\Sigma = \Sigma_0 + \Sigma_{\rm ring}(\tau = 0.03^2)$. We set the kinematic viscosity $\nu = 10^{-4}\ r_0^2\Omega_0$.

In Figure 2.9, we plot our simulated profiles at $t = 2$, 6, and 18, which corresponds to $\tau = 0.0033$, 0.0081, and 0.0225[1]. We find excellent agreement between our simulated profiles and Equation 2.36, which lends confidence to our viscosity implementation. We note that at $t = 2$, our simulated profile has a slightly lower peak than our

[1]Note that $t = 0$ corresponds to $\tau = 0.03^2$ for our setup.

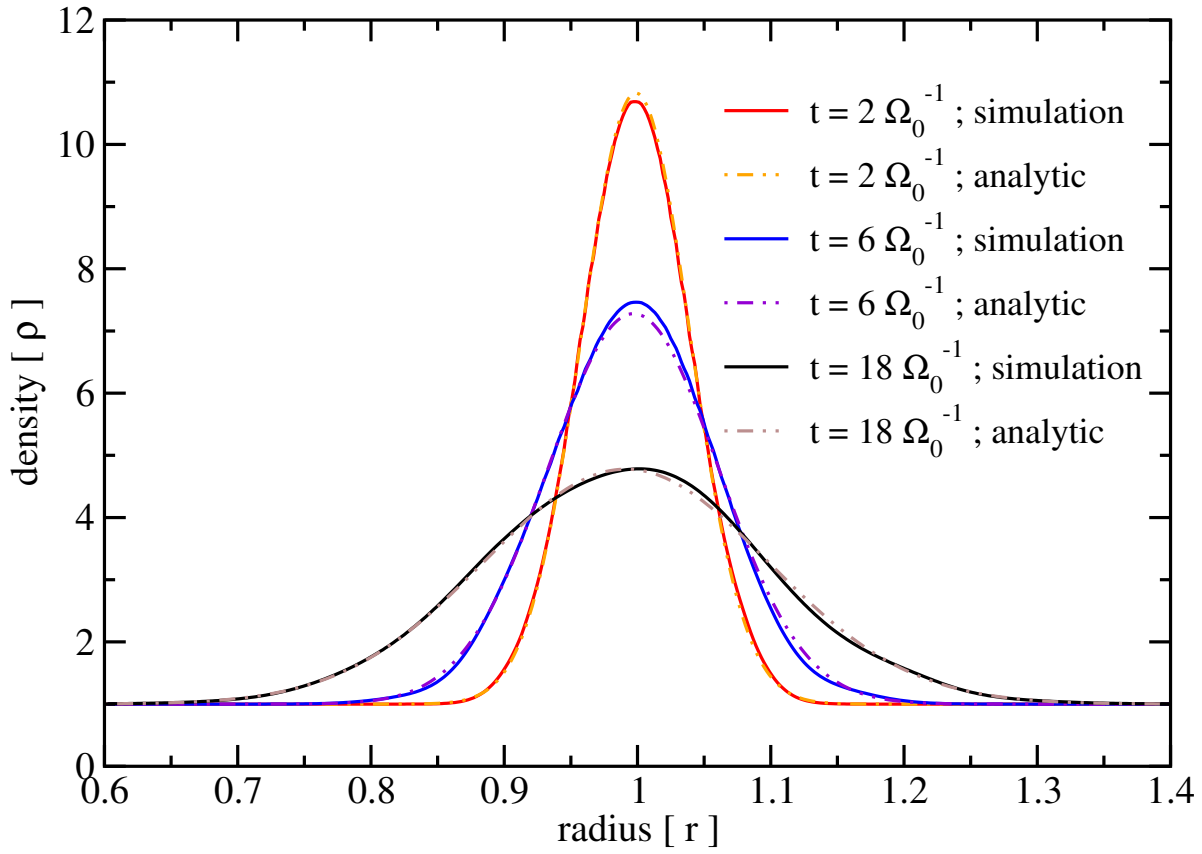

Figure 2.9 Viscous diffusion for a ring orbiting around a point mass. The solid curves are surface density profiles extracted at different simulation times; the dot-dot-dashed curves are the corresponding analytic profiles described by Equation 2.36. The agreement between the two demonstrates the capability of the viscosity implementation in PEnGUIn.

analytic prediction, which is consistent with the fact that Equation 2.36 did not take into account the additional spreading due to the radial component of the viscous term. At a later time, however, this is reversed and our profiles have a higher peak instead. We speculate this is due to the gas-pressure-modified rotation curve that we use. At the inner edge, the fluid rotates at a speed slightly above Keplerian, while the outer edge is slightly sub-Keplerian. So, the fluid at the inner (outer) edge has a little more (less) angular momentum than assumed by Equation 2.36; as a result, the ring spreads a little slower than predicted.

2.5.6 Planetary Torque

For our final test, we tackle a significantly more complex and interesting problem: disk-planet interaction. Like the previous test, we have an isothermal disk orbiting a central mass, but instead of adding a ring, we add a secondary mass (a planet) that orbits the central mass (a star) in a fixed, circular orbit. The total gravitational potential is therefore:

$$\Phi = -\frac{GM_*}{\sqrt{r^2 + r_1^2 + 2rr_1\cos(\phi - \phi_\mathrm{p})}} - \frac{GM_\mathrm{p}}{\sqrt{r^2 + r_2^2 - 2rr_2\cos(\phi - \phi_\mathrm{p}) + r_\mathrm{s}^2}}, \tag{2.39}$$

where M_* and $M_\mathrm{p} = qM_*$ are the masses of the star and the planet, respectively; $r_1 = qa/(1+q)$ and $r_2 = a/(1+q)$ are their radial positions, with a is the total (fixed) separation; $\phi_\mathrm{p} - \pi$ and ϕ_p are their angular positions; and r_s is the softening length of the planet's potential. The mass ratio q is 10^{-5}. We set $G(M_* + M_\mathrm{p}) = 1$ and $r_\mathrm{p} = 1$, so that the planet's orbital frequency $\Omega_\mathrm{p} = 1$, and period $P_\mathrm{p} = 2\pi$.

The isothermal sound speed c_s is set to $c_\mathrm{s} = 0.03\, a\Omega_p$, so the aspect ratio at the planet's position is $h_0/a = 0.03$. We set $r_\mathrm{s} = 0.5h_0$. We also include a low level of viscosity, $\nu = 10^{-8}\, r_0^2\Omega_0$. The initial velocities are set in the same way as Equation 2.37 and 2.38. We simulate a domain from $0.7a$ to $1.3a$ radially, and the grid size is $500(r) \times 1000(\phi)$ with uniform grid cells.

From Section 1.3, we expect this planet to experience a differential Lindblad torque, and a corotation torque that saturates over time. Figure 2.10 plots the net torque on the planet as a function of time, for four different disks, each with a different initial surface density profile: $\Sigma = \Sigma_0(r/a)^{-\beta}$, where $\Sigma_0 = 1$ and β takes on the values 0, 0.5, 1.0, and 1.5.

Our results are in excellent agreement with existing 2D models of disk-planet interaction. The damped oscillation seen in Figure 2.10 has a period of 2 t_lib, where $t_\mathrm{lib} \sim 60\, P_\mathrm{p}$ is the libration time. This value of t_lib corresponds to a horseshoe half-width of $\sim 0.022\, a$, similar to $0.82a\sqrt{q/(r_\mathrm{s}/a)} = 0.021\, a$ given by Paardekooper & Papaloizou (2008). The corotation torque is saturated after a few libration time, so we measure the torque on the planet at $t = 500P_\mathrm{p}$ to determine the differential Lindblad torque acting on them. Once again, our measurement agrees closely with the linear calculations by Paardekooper & Papaloizou (2008), stated in Equation 1.8, to within 4%. The small difference may be related to the nonlinearity of the problem.

2.6 Conclusions

In this chapter we have shown that `PEnGUIn` is capable of utilizing GPUs to achieve very high speeds, up to 20 (25) million grid cells per second running adiabatic (isothermal) calculations on a single GPU, and can reach close to triple that performance on a three-GPU computer. Through a series of hydrodynamics tests, `PEnGUIn` also shows excellent accuracy and consistency, proving to be capable of tackling complex problems of our interest, such as disk-planet interaction. The significance of `PEnGUIn` is not only in that it is a fast, powerful tool for

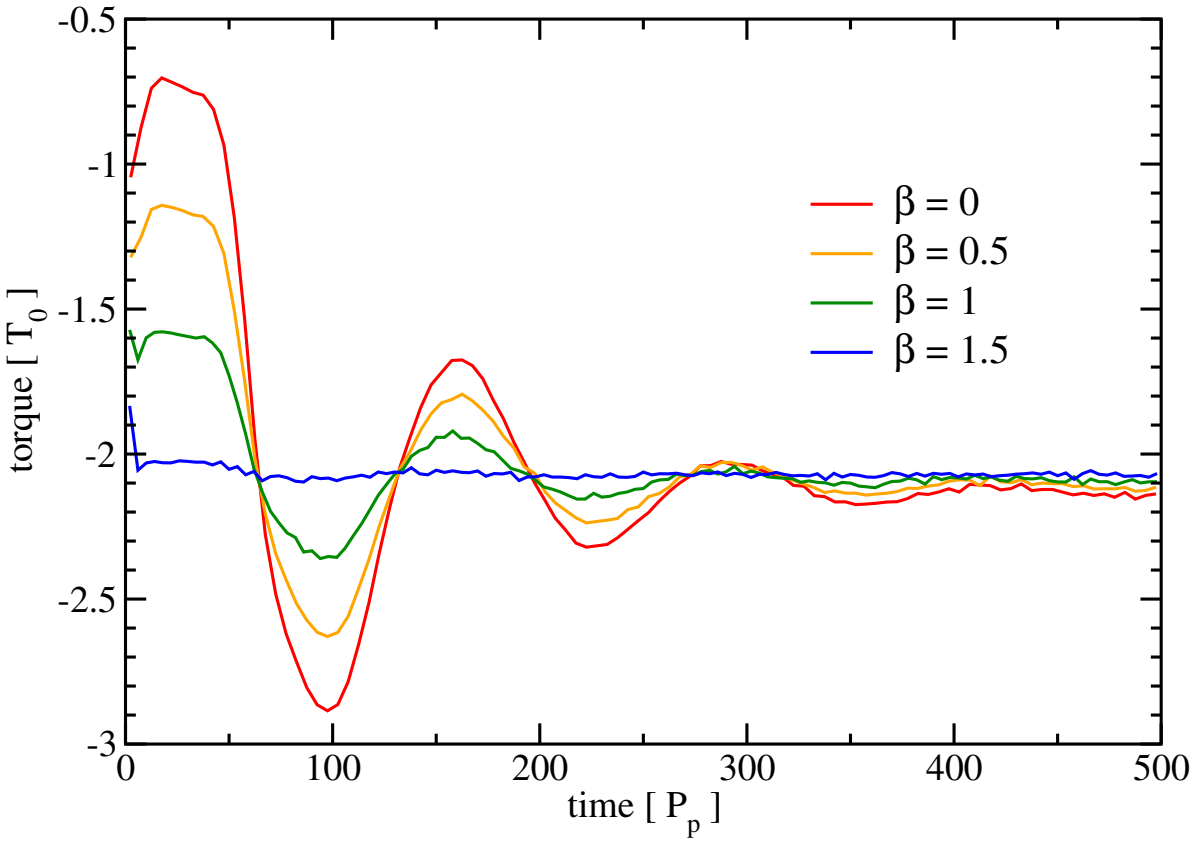

Figure 2.10 Net torque on the planet in units of $T_0 = \Sigma_0 \Omega_p^2 a^4 q^2 (h_0/a)^{-2}$ as a function of time. The data points are time-averaged over one orbital period, P_p, of the planet. The libration time is about $60 P_p$, which corresponds to a half of the period of the oscillations.

our research, but also that it realizes the vast potential of GPU computing. GPUs as personal high-performance computing devices is now certainly a realistic option.

Armed with `PEnGUIn`, we will now delve into protoplanetary disk dynamics. In the following chapters, we will perform both 2D and 3D simulations of protoplanetary disks, accounting for a variety for external forces, such as radiation pressure in Chapter 3, and a combination of a planet's gravitational force and disk viscous stress in Chapters 4 and 5.

Chapter 3

Irradiation Instability by Embedded Dust Grains

A version of this chapter has been published in The Astrophysical Journal as "Irradiation Instability at the Inner Edges of Accretion Disks", Fung, J., and Artymowicz, P., volume 790, issue 1, article id. 78, 2014. Reproduced by permission of the AAS.

3.1 Introduction

Accretion disks are susceptible to a wide range of instabilities, generating turbulence and creating complex, sometimes extreme, structures, such as the formation of planets in protoplanetary disks. Section 1.1 already gave a few examples of these instabilities, and we can add to the list with purely hydrodynamical instabilities that are triggered by specific disk structures, such as a narrow ring, where the Papaloizou-Pringle instability can operate (Papaloizou & Pringle, 1984, 1985, 1987; Goldreich et al., 1986), and disks with a local vortensity extremum, which is favored by the Rossby wave instability (RWI) (Lovelace et al., 1999). The instability we consider in this chapter is in some ways similar to these instabilities, because it is also biased towards disks with narrow features, such as inner disk edges, but it is not purely hydrodynamical. We consider the dynamical effect of radiation pressure from a central source, exerted on a disk consists of a tightly coupled gas-dust mixture.

Radiation pressure is a force generally present in all types of accretion disks. Its effect on accretion disks has been studied in many different aspects, including driving disk winds in active galactic nuclei (AGN) (e.g. Higginbottom et al., 2014), shaping particle size distributions in debris disks (Thebault et al., 2014), and influencing the motions of the inner rims of transitional disks (Chiang & Murray-Clay, 2007; Dominik & Dullemond, 2011). We demonstrate in this chapter that radiation pressure can also cause a disk instability of its own kind. In the following, we give a brief introduction to this instability before launching into the formal theoretical work.

The strength of radiation pressure compared to gravity is measured by the number β:

$$\beta = \frac{\kappa_{\rm opa} L}{4\pi c G M} , \tag{3.1}$$

where L is the central object's luminosity, M is its mass, and $\kappa_{\rm opa}$ is the opacity of the disk material; c and G are the speed of light and gravitational constant respectively. The key to this instability is shadowing. As the front part of the disk gets pushed by radiation pressure, it also casts a shadow that reduces the amount of radiation

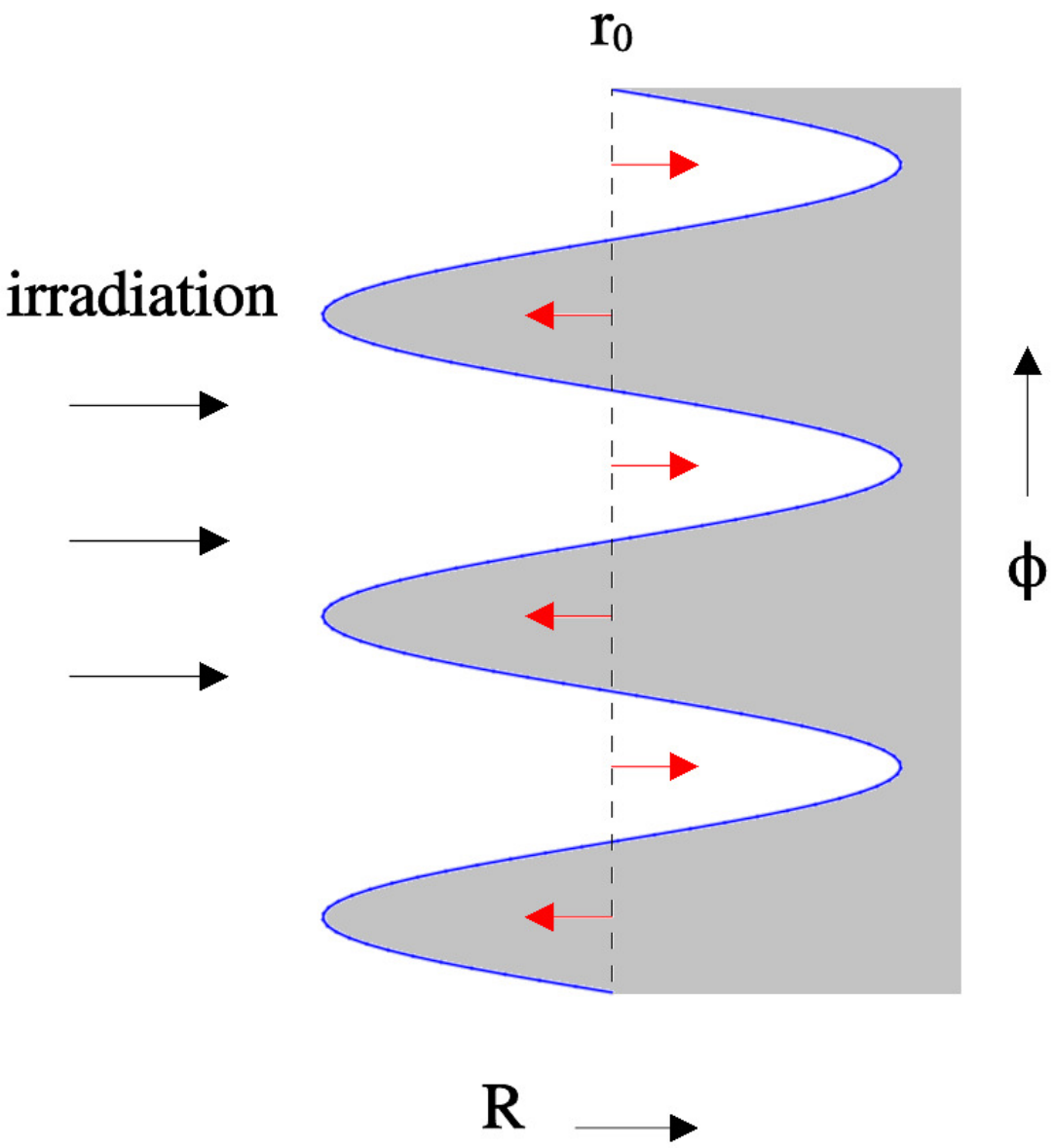

Figure 3.1 Simple illustration describing IRI. The blue curve denotes the orbit of a perturbed disk element oscillating around its guiding center, denoted by the dashed black line at r_0. The shaded area is where the disk sees the shadow cast by the perturbed element. The red arrows show the directions of radial forcing on the background disk relative to the average amount of radiation pressure received along r_0. These arrows are inward when they are in the shadow of the element, and outward when they are not. One can see that the background disk near r_0 is forced in the direction of amplifying the initial perturbation.

pressure on the material further out in the disk. In a 1D, radial picture, since radiation pressure always diminishes outward, the inner part of a disk always feels a stronger push than the outer part, and the net effect is therefore radial compression. In other words, any two concentric disk annuli would feel an attraction between them due to the combined effects of radiation pressure and shadowing.

This 1D scenario does not easily extend to a 2D disk however, because radiation pressure from a central source does not exert any azimuthal force. By the conservation of angular momentum, when a disk element is perturbed radially, it will oscillate at some epicyclic frequency. Figure 3.1 illustrates what effect this oscillating element has on the disk. One can see that disk material near the orbit of the perturbed element will experience a variation in shadowing along the azimuth. This variation creates a forcing that induces the unperturbed material to follow the motion of the perturbed element. The result is a global collective motion that is capable of growing on its own. We term this phenomenon the "irradiation instability" (IRI), since it relies on irradiation by the central object.

Because a larger β allows for a more rapid radial motion, its value is crucial for the survival of this collective motion against disk shear. In most systems, dust grains provide the largest contribution to β. In circumstellar disks, micron-size grains can have $\beta > 1$ for F-type stars, and up to $\beta \sim 10^1$ for A-type stars (e.g., Equation 10 of Kirchschlager & Wolf (2013)). Given that the gas-to-dust ratio is typically $\sim 10^2$, β of a perfectly coupled gas-dust mixture may be of order a few percent. We note that the coupling between gas and dust is expected to be strong for the small grains that contributes most to disk opacity (Section 1.2). Additionally, dust settling can enhance β in the midplane by reducing the local gas-to-dust ratio, while the radial migration of dust results in size segregation (Thebault et al., 2014), which can also enhance β at local radii. In other systems where radiation pressure can drive significant mass loss, such as AGN accretion disks, one would even expect β to exceed unity.

This chapter aims to provide a basic understanding of IRI, of both the conditions that trigger it, and its consequences. In Section 3.2, we present a theoretical foundation for IRI and derive its instability criterion. Section 3.3 contains our disk model. Section 3.4 describes our semi-analytic and numerical methods. Section 3.5 reports the modal growth rate as a function of β and the sound speed c_s of the disk, and gives a discussion on the nonlinear evolution of IRI. Section 3.6 concludes with an outlook for future work.

3.2 The Linear Theory

We follow the method of Goldreich & Tremaine (1979), using similar notation, to derive the linear response of a 2D disk stirred by radiation pressure. We start with the continuity equation and the conservation of momentum:

$$\frac{\partial \Sigma}{\partial t} + \nabla \cdot (\Sigma \mathbf{v}) = 0, \tag{3.2}$$

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla)\mathbf{v} = -\nabla \eta - \frac{GM\,(1 - \beta e^{-\tau})}{r^2}\hat{\mathbf{r}}, \tag{3.3}$$

where Σ is the surface density of the disk; $\mathbf{v}$ is the 2D velocity field; η is the specific enthalpy such that $\nabla\eta = \nabla P/\Sigma$, where P is the vertically averaged gas pressure; and τ is the optical depth of the disk. We denote the Keplerian orbital frequency as Ω_k, and the sound speed c_s is defined by the ideal gas law $P = c_s^2\Sigma$. τ depends on the density distribution by the following equation:

$$\tau = \int_0^r \kappa_{\rm opa}\rho \mathrm{d}r' \ , \tag{3.4}$$

where ρ is the density of the disk. Near the midplane, $\rho \propto \Sigma/h$, where $h = c_s/\Omega_k$ is the scale height of the disk. Note that with Equation 3.3 we have neglected the scattering of light into the azimuthal direction.

Σ, η, and $\mathbf{v}$ can be separated into a background quantity (without any subscript) and a perturbed quantity (denoted by the subscript "m"). We assume the background disk to be axisymmetric and in hydrostatic equilibrium so that $\mathbf{v} =(0, r\Omega)$, where

$$\Omega = \sqrt{\Omega_k^2 (1 - \beta e^{-\tau}) + \frac{1}{r}\frac{\mathrm{d}\eta}{\mathrm{d}r}} \,, \tag{3.5}$$

and the components of the perturbed velocity are denoted as $\mathbf{v}_m \equiv(u, \upsilon)$. To simplify the notation, we also define the background and perturbed radiation force as F and F_m:

$$F = r\Omega_k^2 \beta e^{-\tau}, \tag{3.6}$$

$$F_m = -F \int_0^r \frac{\eta_m}{c_s^2}\frac{\mathrm{d}\tau}{\mathrm{d}r'}\mathrm{d}r' \,. \tag{3.7}$$

For a small perturbation, it follows from Equations 3.2 and 3.3 that the perturbed quantities are governed by the following linearized equations:

$$\frac{\partial \Sigma_m}{\partial t} + \nabla(\Sigma \mathbf{v}_m) + \nabla(\Sigma_m \mathbf{v}) = 0 \,, \tag{3.8}$$

$$\frac{\partial \mathbf{v}_m}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{v}_m + (\mathbf{v}_m\cdot\nabla)\mathbf{v} = -\nabla\eta_m + F_m\hat{\mathbf{r}} \,. \tag{3.9}$$

Without a loss of generality, we can assume a form of the solution for the perturbed quantities Σ_m, η_m, u and υ:

$$X_m(r, \theta, t) = X(r)e^{i(m\theta - \omega t)} \,, \tag{3.10}$$

for some complex function $X(r)$ and complex number ω, while m is the azimuthal mode number. Substituting this form into Equation 3.9, we find

$$u = -\frac{i}{D}\left[\frac{2m\Omega}{r}\eta_m + \Omega_m\left(\frac{\partial \eta_m}{\partial r} - F_m\right)\right] \,, \tag{3.11}$$

$$\upsilon = \frac{1}{D}\left[\frac{m\Omega_m}{r}\eta_m + 2\left(\Omega + \frac{r}{2}\frac{\mathrm{d}\Omega}{\mathrm{d}r}\right)\left(\frac{\partial \eta_m}{\partial r} - F_m\right)\right] \,. \tag{3.12}$$

The pattern rotation frequency Ω_m and the coefficient D are defined as

$$\Omega_m \equiv m\Omega - \omega \,, \tag{3.13}$$

$$D \equiv \kappa^2 - \Omega_m^2 \,, \tag{3.14}$$

$$\kappa^2 = \frac{1}{r^3}\frac{\mathrm{d}\left[r^4\Omega^2\right]}{\mathrm{d}r} \,, \tag{3.15}$$

where κ is the epicyclic frequency of the unperturbed orbit. To solve for Σ_m, or equivalently η_m, we substitute Equation 3.11 and 3.12 into Equation 3.8, giving

$$\frac{\partial^2 \eta_m}{\partial r^2} + a(r)\frac{\partial \eta_m}{\partial r} + b(r)\eta_m + c(r)\int_0^r \frac{\eta_m}{c_s^2}\frac{\mathrm{d}\tau}{\mathrm{d}r'}\mathrm{d}r' = 0, \tag{3.16}$$

where

$$a \equiv \frac{\partial}{\partial r} \ln\left(\frac{r\Sigma}{D}\right),$$
$$b \equiv \frac{2m\Omega}{r\Omega_m}\frac{\partial}{\partial r}\ln\left(\frac{\Sigma\Omega_0}{D}\right) - \frac{m^2}{r^2} + \frac{1}{c_s^2}\left(F\frac{\mathrm{d}\tau}{\mathrm{d}r} - D\right),$$
$$c \equiv F\left(\frac{\partial}{\partial r}\ln\left(\frac{r\Sigma F}{D}\right) - \frac{2m\Omega}{r\Omega_m}\right).$$

We arrive at a second-order integro-differential equation for η_m.

3.2.1 Instability Criterion

A local criterion for axisymmetric instability can be derived from Equation 3.16. We apply the WKB approximation and write $\eta_m \sim e^{i\int_0^r k_r \mathrm{d}r'}$, where $k_r \gg \frac{1}{r}$ is the radial wave number. We then separate the real and imaginary part of the equation. Finally, setting $m = 0$, the dispersion relation can be written as:

$$\omega^2 = \kappa^2 + k_r^2 c_s^2 - \Omega_k^2 \beta e^{-\tau}\left(\frac{\mathrm{d}\tau}{\mathrm{d}\ln r} + \tilde{\tau}_m \frac{\mathrm{d}\ln[r\mathscr{R}]}{\mathrm{d}\ln r}\right), \tag{3.17}$$

where

$$\mathscr{R} \equiv \frac{\Sigma\Omega_k\beta e^{-\tau}}{\kappa^2}, \tag{3.18}$$
$$\tilde{\tau}_m \equiv \tau_m\left(\frac{\Sigma_m}{\Sigma}\right)^{-1} = \frac{c_s^2}{\eta_m}\int_0^r \frac{\eta_m}{c_s^2}\frac{\mathrm{d}\tau}{\mathrm{d}r'}\mathrm{d}r'. \tag{3.19}$$

$\mathscr{R}$ has the same units as the inverse of vortensity, but is a quantity that depends on radiation pressure. $\tilde{\tau}_m$ is the ratio between the perturbed optical depth τ_m and the relative surface density perturbation Σ_m/Σ. The local disk is unstable if a solution for k_r exists given $\omega^2 = 0$, which denotes the line of neutral stability. Setting $\omega^2 = 0$, the condition for $k_r^2 > 0$ is

$$\beta e^{-\tau}\left(\frac{\kappa}{\Omega_k}\right)^{-2}\left(\frac{\mathrm{d}\tau}{\mathrm{d}\ln r} + \tilde{\tau}_m\frac{\mathrm{d}\ln[r\mathscr{R}]}{\mathrm{d}\ln r}\right) > 1. \tag{3.20}$$

It is important to note that κ contains dependencies on both radiation and gas pressure. In the interest of specifically studying IRI, we consider the case when the rotation curve is solely modified by radiation pressure. Then κ can be expressed as

$$\left(\frac{\kappa}{\Omega_k}\right)^2 = 1 - \beta e^{-\tau}\frac{\mathrm{d}\ln[r\beta]}{\mathrm{d}\ln r} + \beta e^{-\tau}\frac{\mathrm{d}\tau}{\mathrm{d}\ln r}. \tag{3.21}$$

Plugging Equation 3.21 into Equation 3.17, the condition for instability becomes

$$q_\beta \equiv \beta e^{-\tau}\left(\frac{\mathrm{d}\ln[r\beta]}{\mathrm{d}\ln r} + \tilde{\tau}_m\frac{\mathrm{d}\ln[r\mathscr{R}]}{\mathrm{d}\ln r}\right) > 1. \tag{3.22}$$

To complete our derivation, we need to evaluate $\tilde{\tau}_m$. We begin by integrating Equation 3.19 by parts

$$\tilde{\tau}_m = \tau - ik_r\frac{c_s^2}{\eta_m}\int_0^r \frac{\eta_m}{c_s^2}\tau\mathrm{d}r'. \tag{3.23}$$

If in the disk there exists a "transition region" where the disk sharply transitions from being radially transparent

to opaque, then one can show that inside this region, the second term on the right-hand side of Equation 3.23 has a magnitude of order $k_r \Delta r$, where Δr is the width of the transition region. This allows us to approximate $\tilde{\tau}_m \sim \tau$ in the limit $\frac{1}{\Delta r} \gg k_r$. Moreover, even when $\frac{1}{\Delta r} \sim k_r$, we expect $\tilde{\tau}_m \sim \tau$ to remain accurate to within an order of unity. In Section 3.5.3 we evaluate $\tilde{\tau}_m$ explicitly and find out to what extend this holds true.

While Equation 3.20 is the more general form, Equation 3.22 does reveal surprising behavior: it contains no explicit dependence on $\frac{\mathrm{d}\tau}{\mathrm{d}r}$, as it is completely canceled by the stabilizing effect of κ^2. Replacing it is a term containing $\frac{\mathrm{d}\beta}{\mathrm{d}r}$, whose effect is to lower κ^2 to the point of triggering a form of irradiation-induced Rayleigh instability. While it does contribute to the instability of the disk, we do not consider it the true trigger of IRI. Rather, we focus on the second term inside the bracket. First, it implies that a disk is unstable to IRI if it has a positive gradient in $\mathscr{R}$, which can be created by a gradient in Σ and/or β. Second, this gradient must be located where $\tilde{\tau}_m e^{-\tau} \sim \tau e^{-\tau}$ is reasonably large, which is precisely the transition region. This is consistent with our picture that IRI is driven by shadowing. Because of the uncertainty in $\tilde{\tau}_m$, as well as the other assumptions stated in the beginning of this section, Equations 3.20 and 3.22 should be taken as order-of-magnitude guidelines rather than rigid conditions.

3.2.2 Corotating Modes

If a linear mode exists, its corotation radius can be found by solving Equation 3.16 for $\Omega_m = 0$. Similar to Section 3.2.1, we apply the WKB approximation, and then the real part of Equation 3.16 evaluated at the corotation radius can be rewritten as:

$$\Omega_m = 0 = 2m\Omega \frac{\frac{h^2}{r^2}\frac{\mathrm{d}\ln\mathscr{F}}{\mathrm{d}\ln r} - \beta e^{-\tau}\tilde{\tau}_m}{\frac{\kappa^2}{\Omega_k^2} + |k|^2 h^2 - \beta e^{-\tau}\left(\frac{\mathrm{d}\tau}{\mathrm{d}\ln r} + \tilde{\tau}_m \frac{\mathrm{d}\ln[r\mathscr{R}]}{\mathrm{d}\ln r}\right)}\,, \tag{3.24}$$

where $|k|^2 = k_r^2 + m^2/r^2$, and $\mathscr{F} \equiv \Sigma\Omega/\kappa^2$ is a quantity inversely proportional to the vortensity of the disk. The corotation radius is therefore located at where the following condition is satisfied:

$$\frac{\mathrm{d}\ln\mathscr{F}}{\mathrm{d}\ln r} = \left(\frac{h}{r}\right)^{-2} \beta e^{-\tau}\tilde{\tau}_m\,. \tag{3.25}$$

For barotropic flow and $\beta = 0$, this condition becomes identical to that described in Section 2.2 of Lovelace et al. (1999) for RWI. The usefulness of Equation 3.25 is limited because without a full solution, the exact value of $\tilde{\tau}_m$ is unknown. However, allowing that $\tilde{\tau}_m \sim \tau$, it does provide an insight: since the the right-hand side of Equation 3.25 is always positive, if $\mathscr{F}$ contains a local maximum, the corotation radius will always be located at a lower orbit than where this maximum is. In our disk model described in the following section, $\mathscr{F}$ does contain a local maximum within the transition region, so we expect the corotation radius to be smaller for disks with a larger value of $\left(\frac{h}{r}\right)^{-2}\beta$. This prediction is tested in Section 3.5.1.

3.3 Disk Model

For simplicity we do not consider any spatial variation in the composition of the disk, therefore β and κ_{opa} are constants. With this simplification, Equation 3.22 says that the disk is most unstable if $\mathscr{R}$ has a large positive gradient near $\tau = 1$. We create this condition with a disk that contains a sharp inner edge. At this edge, Σ increases by orders of magnitude across a small radial range, while τ rises from a small value to above unity. Our

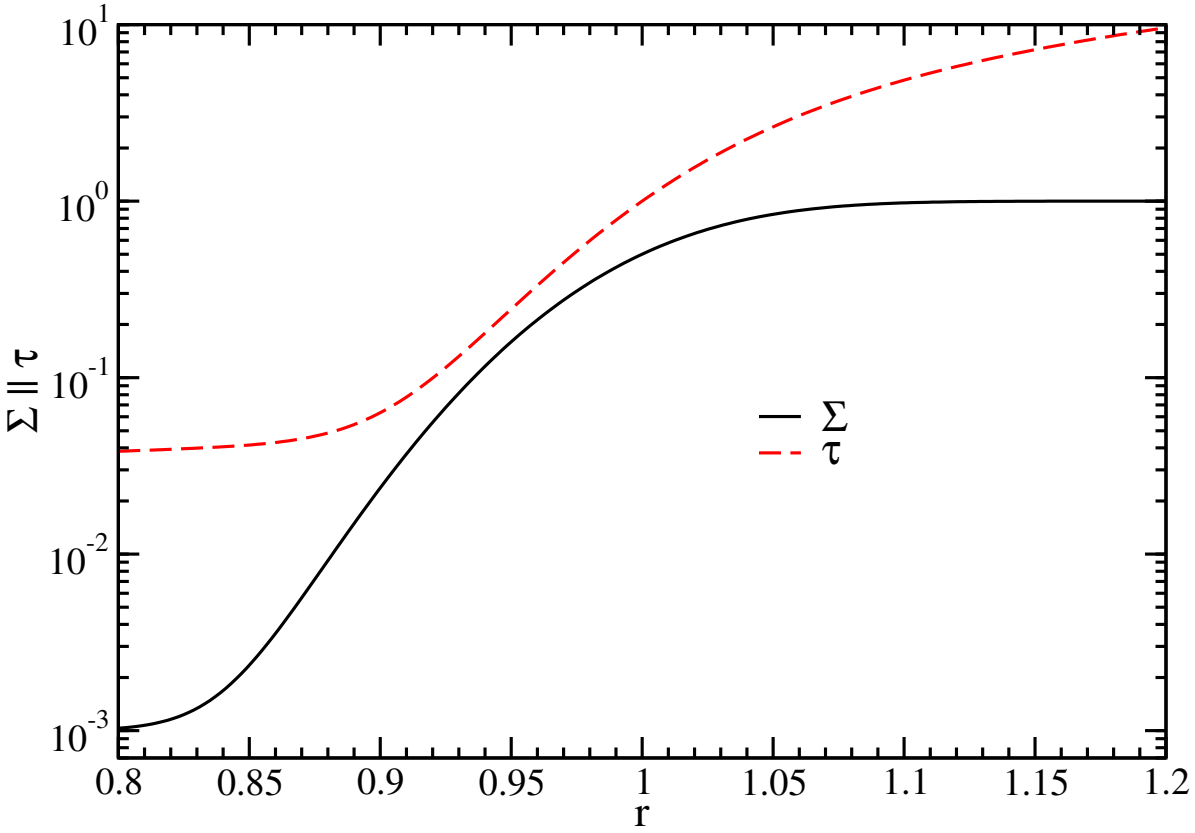

Figure 3.2 Black solid line plotting the surface density profile described by Equation 3.26 and red dashed line plotting the optical depth profile.

prescription for such a disk is:

$$\Sigma(r) = \frac{1}{2}\left(\Sigma_d + \Sigma_c\right) + \frac{1}{2}\mathrm{erf}\left(\frac{r - r_0}{\sqrt{2\Delta r^2}}\right)\left(\Sigma_d - \Sigma_c\right), \tag{3.26}$$

where Σ_d is the surface density of the disk, Σ_c is the surface density inside the cavity , r_0 is the radius at which the inner edge is located, and Δr is the width of this edge. We set $\Sigma_d = 1$ and $\Sigma_c = 0.001$ for a density contrast of 10^3. We also set $r_0 = 1$ and $GM = 1$ so that the dynamical time $t_{\rm dyn}$ at the edge is $\Omega_k^{-1} = 1$. For the sharpness of the edge, we set $\Delta r = 0.05$. The motivation for this choice is that Δr is unlikely to be shorter than h, which for protoplanetary disks has a typical value of $0.05r$. $\kappa_{\rm opa}$ is chosen such that $\tau(r_0) = 1$. If we move this $\tau = 1$ point to a much smaller/larger radius, the disk edge will be become optically thick/thin, and thus one would expect the instability to weaken or even disappear. Figure 3.2 plots both the Σ and τ profile. To complete the equation set, we adopt an isothermal equation of state so that $c_{\rm s}$ is a constant.

This leaves two free parameters in our model: β and $c_{\rm s}$. We perform a parameter study over the range $\beta = \{0, 0.3\}$ and $c_{\rm s} = \{0.02, 0.06\}$. Note that for $h(r_0) \gtrsim \Delta r$, corresponding to $c_{\rm s} \gtrsim 0.05$, the disk edge may become hydrodynamically unstable. We deliberately include this limit in our parameter space both as a sanity check and to investigate how IRI can be differentiated from other forms of instabilities.

3.4 Two Independent Approaches

For our given disk model, we aim to find out for the IRI (1) how the modal growth rate varies as a function of β and c_s, and (2) what are the properties of its nonlinear phase. Two independent approaches are used: a numerical method using hydrodynamical simulations and a semi-analytic method that solves the linearized problem (Equation 3.16). These two methods not only serve as verifications for each other, but are also complementary since a full simulation gives us an insight into the nonlinear phase, while the semi-analytic method is not subjected to limitations such as resolution and numerical noise.

3.4.1 Hydrodynamical Simulation

We numerically simulate the 2D disk described in Section 3.3, using our GPU-based, Lagrangian, dimensionally-split, shock-capturing hydrodynamics code `PEnGUIn`, described in detail in Chapter 2. We implement an additional module to compute τ using piecewise parabolic interpolation to match the order of `PEnGUIn`.

Our simulations have a domain spanning 0.5 to 2.0 in radial (in units where the disk edge is located at $r_0 = 1$) and the full 0 to 2π in azimuth. Moving the inner boundary to 0.7 or the outer boundary to 1.5 has a negligible effect on the growth of linear modes. We opt for a larger domain to accommodate the more violent nonlinear evolution.

The resolution is 1024 (r) by 3072 (ϕ). Azimuthal grid spacing is uniform everywhere, but radial grid spacing is uniform only between 0.5 and 1.3; from 1.3 to 2.0 it is logarithmic. This takes advantage of `PEnGUIn`'s ability to utilize non-uniform grids to enhance the resolution around the disk edge. The resulting grid size at r_0 is about 0.001 (r) by 0.002 (ϕ). This gives at least 10 cells per h for even the smallest h we consider. Figure 3.3 shows how our simulations converge with resolution.

In each simulation, we extract the amplitudes of azimuthal modes as functions of time, resolving up to $m = 50$:

$$A_m(t) = \frac{1}{2\pi}\left|\int_0^{2\pi}\int_{1.0}^{1.1}\Sigma(t)e^{im\phi}\mathrm{d}r\,\mathrm{d}\phi\right| , \tag{3.27}$$

where we have chosen to integrate over the radial range $r = \{1.0, 1.1\}$. Instantaneous values of A_m are not the focus; rather, we seek a distinct period of exponential growth where we can measure its growth rate, i.e., the imaginary part of ω. Figure 3.4 shows one example of how modal growth behaves in these simulations. For a disk of a given set of parameters, the highest growth rate characterizes its timescale for instability.

We use a boundary condition fixed to the initial values described by Equations 3.5 and 3.26, with zero radial velocity. To reduce noise in A_m, we also include wave-killing zones in $r = \{0.5, 0.6\}$ for the inner boundary and $r = \{1.6, 2.0\}$ for the outer. Within these zones, we include an artificial damping term:

$$\frac{\partial X}{\partial t} = (X(t=0) - X)\,\frac{2c_s|r - r_{\rm kill}|}{d_{\rm kill}^2} , \tag{3.28}$$

where X includes all disk variables Σ, P, and $\mathbf{v}$; $r_{\rm kill}$ is the starting radius of the wave-killing zone, which is 0.6 for the inner boundary and 1.6 for the outer; and $d_{\rm kill}$ is the width of these zones, which equals 0.1 for the inner boundary and 0.4 for the outer. In the end we are able to resolve A_m as small as 10^{-10}, such as shown in Figure 3.4.

Simulations are terminated soon after the instability becomes fully nonlinear: up to 100 orbits, or 628 $t_{\rm dyn}$. For very slowly growing modes, numerical noise severely hampers the precision of growth rate measurements. Consequently, this method is only capable of measuring growth rates $\gtrsim 0.01 t_{\rm dyn}^{-1}$. The computational time for

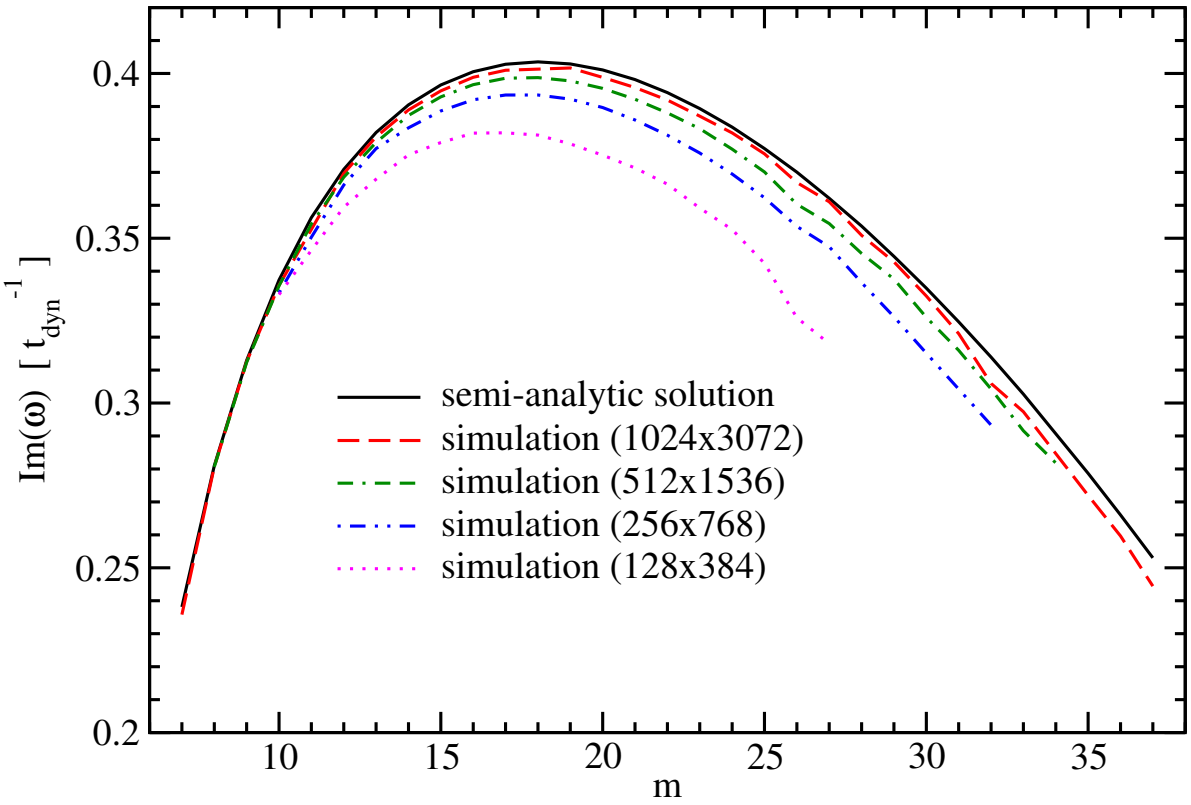

Figure 3.3 Growth rates of azimuthal modes with $(\beta, c_s) = (0.2, 0.02)$. At 1024 (r) by 3072 ($\phi$), the growth rates extracted from simulation match those found by the semi-analytic method to $\sim 1\%$. For this particular case, the fastest growing mode is $m = 18$, with a growth rate of $\mathrm{Im}(\omega) = 4.0 \times 10^{-1} t_{\mathrm{dyn}}^{-1}$. See Section 3.5.1 for further discussions on how these results vary with β and c_s.

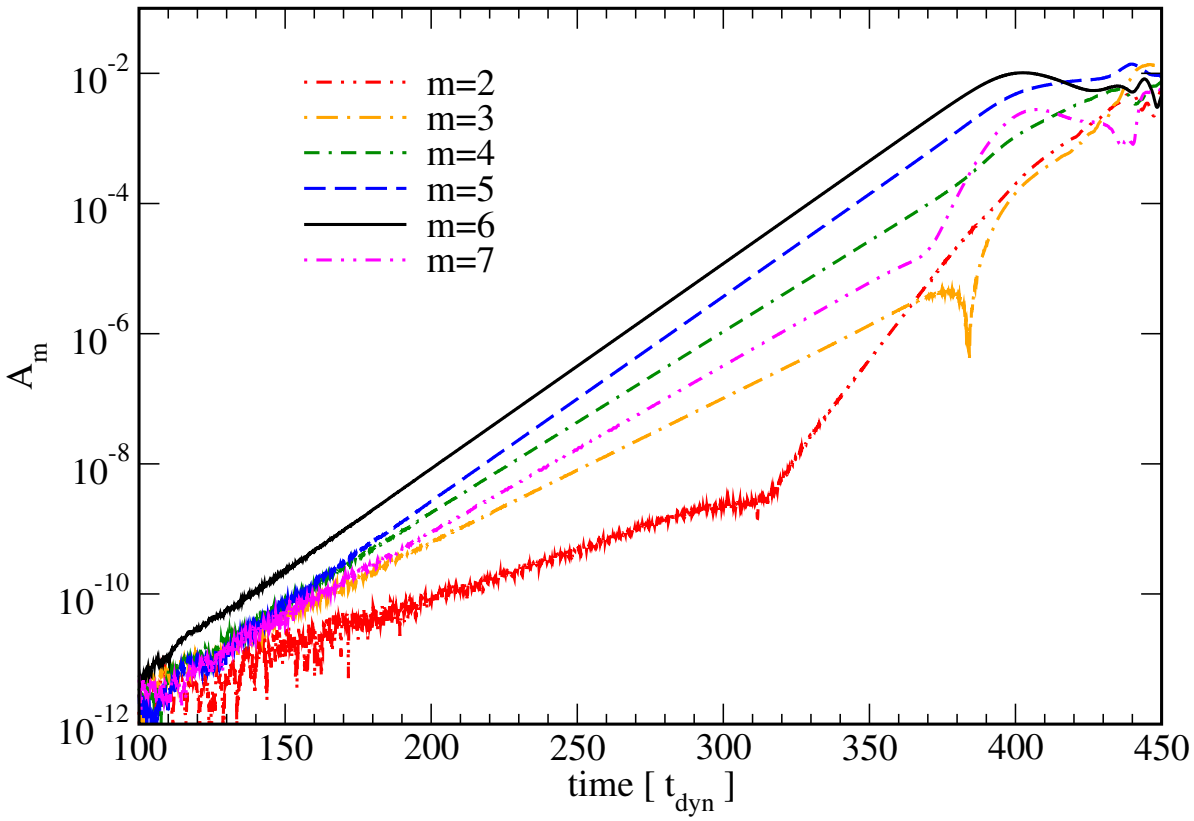

Figure 3.4 Temporal evolution of A_m (see Equation 3.27) with $(\beta, c_s) = (0.05, 0.05)$. A well-defined exponentially growing phase can be seen around $t = 200 \sim 300$. Beyond $t = 300$ the modes begin to exhibit higher-order coupling.

Table 3.1. Semi-analytic Results

β	c_s	Im(ω) (t_{dyn}^{-1})	m[a]	r_{cor}[b](r_0)
0	0.06	7.7×10^{-2}	4	1.046
0	0.05	2.5×10^{-3}	1	1.052
0.1	0.05	7.2×10^{-2}	6	0.979
0.15	0.04	1.6×10^{-2}	7	0.956
0.15	0.03	1.4×10^{-1}	8	0.943
0.2	0.02	4.0×10^{-1}	18	0.938

[a]We only report the properties of the fastest growing mode.

[b]r_{cor} denotes the corotation radius.

PEnGUIn is about 12 minutes per orbit on a single GTX-Titan graphics card.

3.4.2 Semi-analytic Method

Equation 3.16 constitutes an eigenvalue problem, where η_m is the eigenfunction and ω is the eigenvalue. To solve this problem, we develop a code that directly integrates the differential equations, iterates for the correct boundary conditions, and optimizes to find the eigenvalues. The complexity of this code is mainly to overcome the difficulty imposed by the integral in Equation 3.16, which effectively raises the order of the differential equation. The details are documented in Appendix A.

Despite the fact that it solves the linearized equations, our semi-analytic method in fact requires a much longer computational time than simulations using PEnGUIn. Due to limited resources, initially we only apply it to five sets of parameters. Table 3.1 contains a list of these sets. One major advantage of this method is that it does not have a limit to how slow of a growth rate can be detected, so we also apply it to all cases where simulations do not detect any modal growth. Among them, we find a positive growth rate for one case. [1] It is also listed in Table 3.1, making a total of six sets of parameters.

3.5 Results

The sets of parameters we consider are $\beta = \{0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$ and $c_s = \{0.02, 0.03, 0.04, 0.05, 0.06\}$. All 30 combinations of these values are simulated, but only a select few are solved with our semi-analytic method (see Table 3.1). The growth rates found by our two independent approaches agree to $\sim 1\%$. Figure 3.3 gives one example of this agreement. Also, the shapes of the modes extracted from simulations are nearly identical to the ones solved semi-analytically. Comparing Figure 3.5 to 3.6, results from the two methods are only distinguishable near the outer boundary, where the simulated ones show some artificial damping due to the wave-killing zones imposed. Because of the excellent agreement we are able to combine the results of the two approaches to give a detailed picture for the IRI linear modes, complemented by the nonlinear evolution provided by simulations.

[1]This case has $(\beta, c_s) = (0, 0.05)$. Since $\beta = 0$, the modal growth is purely hydrodynamical and unrelated to IRI. See Section 3.5.1 for further discussions.

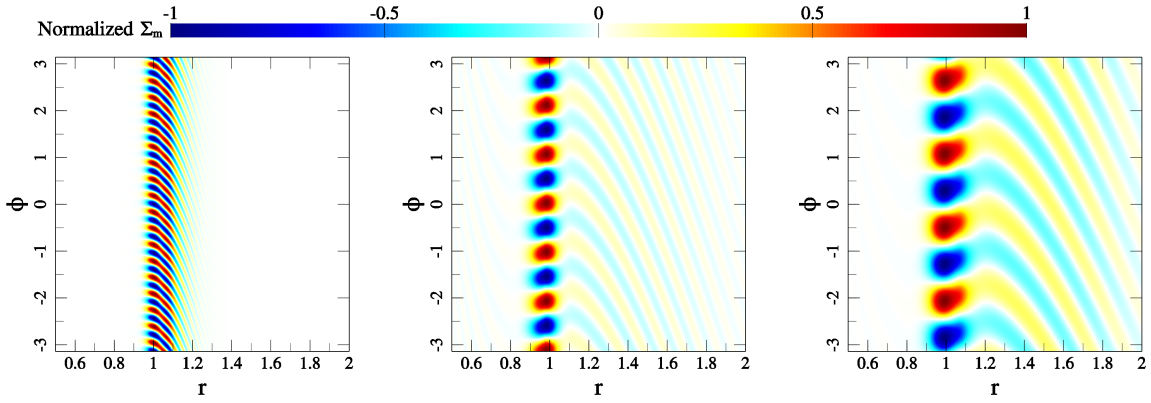

Figure 3.5 Fastest growing modes extracted from simulations through Fourier decomposition. Color shows the surface density normalized to the peak of each mode. On the left is an $m = 18$ mode from $(\beta, c_s) = (0.2, 0.02)$; in the middle is $m = 6$ from $(\beta, c_s) = (0.1, 0.05)$; and on the right is $m = 4$ from $(\beta, c_s) = (0, 0.06)$.

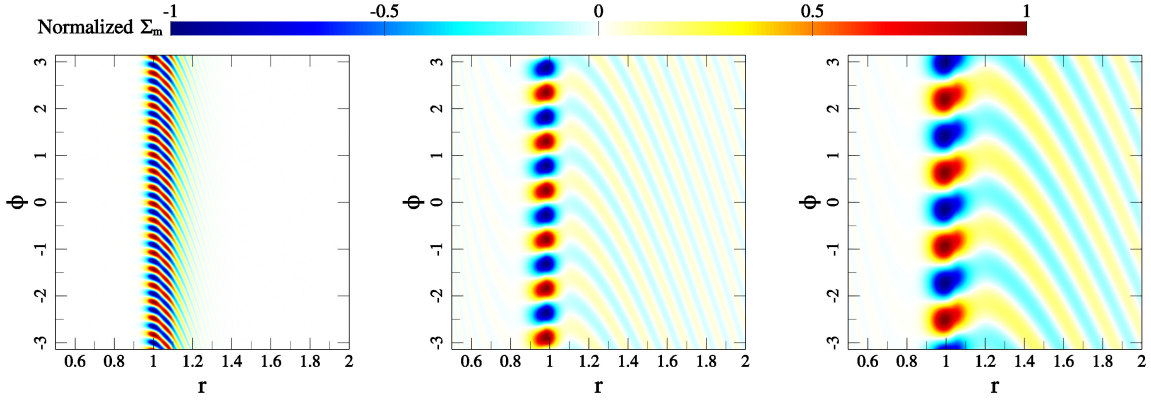

Figure 3.6 The fastest growing modes directly computed using our semi-analytic method for the same parameters listed in Figure 3.5. Note the near-perfect agreement with Figure 3.5.

3.5.1 Linear Modes

We find clear growth of asymmetric modes for all cases with β larger than a certain threshold value that is weakly dependent on c_s across our parameter space. For most of our chosen c_s values, modal growth is only detected when $\beta \geq 0.1$, except for $c_s \sim 0.02$, where this threshold rises to $\beta \geq 0.15$. From a simple perspective, we expect the disk to be more unstable for a larger β and smaller c_s, because β measures the strength of radiation pressure while c_s is a source of resistance to external forcing. In general, we do find the growth rate to increase with β and decrease with c_s, but with obvious exceptions.

In Figure 3.7 we divide our parameter space into three regions: regions I and II where modal growth is driven by radiation pressure, and region III where it is mainly driven by hydrodynamical effects. In regions I and II, growth rate scales roughly linearly with β for any given c_s, a trend that can be more easily seen in Figure 3.8. This is consistent with our expectations.

Figure 3.9, however, reveals a more complicated aspect of IRI. Disregarding the $\beta = 0$ data points that belong to region III, the growth rate is very close to constant over the range $0.04 \leq c_s \leq 0.06$ for any given β. Once c_s goes below 0.04 it shows different trends depending on the value of β: the growth rate increases as c_s decreases for $\beta \geq 0.2$, but for a smaller β the trend flattens or even begins to drop. This complex behavior may relate to how sound waves and IRI modes couple. While sound waves have a length scale h, IRI modes are mainly restricted by the sharpness of the transition region, which has a length scale Δr. Our results suggest that the coupling is weak when $h \sim \Delta r$, and becomes much stronger as h decreases, allowing the transition region to accommodate the full wavelength of the longest wave.

Region III is where radiation pressure becomes a smaller effect than gas pressure. For $c_s \geq 0.05$, or equivalently, $h(r_0) \geq \Delta r$, we detect modal growth even in the absence of any radiation pressure. In fact, when $\beta = 0$, our disk model is similar to the "homentropic step jump" model used by Li et al. (2000) to study RWI (see their Figure 2). Their Figure 11 shows that for a pressure jump with a width $\Delta r = 0.05$, RWI modes will develop if $c_s \gtrsim 0.06$ [2]. Our results are consistent with their findings.

The division between IRI and RWI is clear to us because the two mechanisms appear to destructively interfere with each other. For $c_s = 0.06$, there is a clear drop in growth rate from $\beta = 0$ to $\beta = 0.05$ before it rises again (see Figure 3.8). Similarly for $c_s = 0.05$, we do not detect any modal growth at $\beta = 0.05$ even though it is detected at both $\beta = 0$ and $\beta = 0.1$. One clue to this behavior is that we find c_s and β to have opposing effects on the epicyclic frequency κ. In Figure 3.10 we see that gas pressure lowers κ near $r = r_0$, while β raises it. Two effects roughly cancel when $\beta = 0.05$. It is unclear whether this is coincidental or not.

This dividing line may not remain at $\beta = 0.05$ for a different value of Δr or c_s. If we create a sharper edge by reducing Δr, both IRI and RWI are expected to be enhanced and it is unclear to us whether this dividing line will move to a higher or lower β. Additionally, a high c_s can push κ^2 below zero and trigger Rayleigh instability which further complicates the matter. For our disk model this limit is at $c_s \sim 0.07$. Since the focus of this work is to characterize IRI, we defer the thorough investigation on the interactions between IRI and other forms of instability to a future study.

Other than the growth rate, we also find other general trends about the linear modes. For an increasing β or decreasing c_s, the azimuthal mode number m of the fastest growing mode increases. The dependence on β is particularly pronounced. In the most extreme case, the fastest growing mode is $m = 47$ when $(\beta, c_s) = (0.3, 0.02)$. In contrast, for all cases with $\beta = 0.1$, $m = 5 \sim 6$ is the fastest growing mode. Figure 3.3 shows an intermediate

[2] There are a few small differences between the disk model used by Li et al. (2000) and ours. For example, they use an adiabatic equation of state with an adiabatic index of 5/3, and their pressure jump is modeled with a different formula (compare their Equation 3 to our Equation 3.26). We consider these differences insignificant.

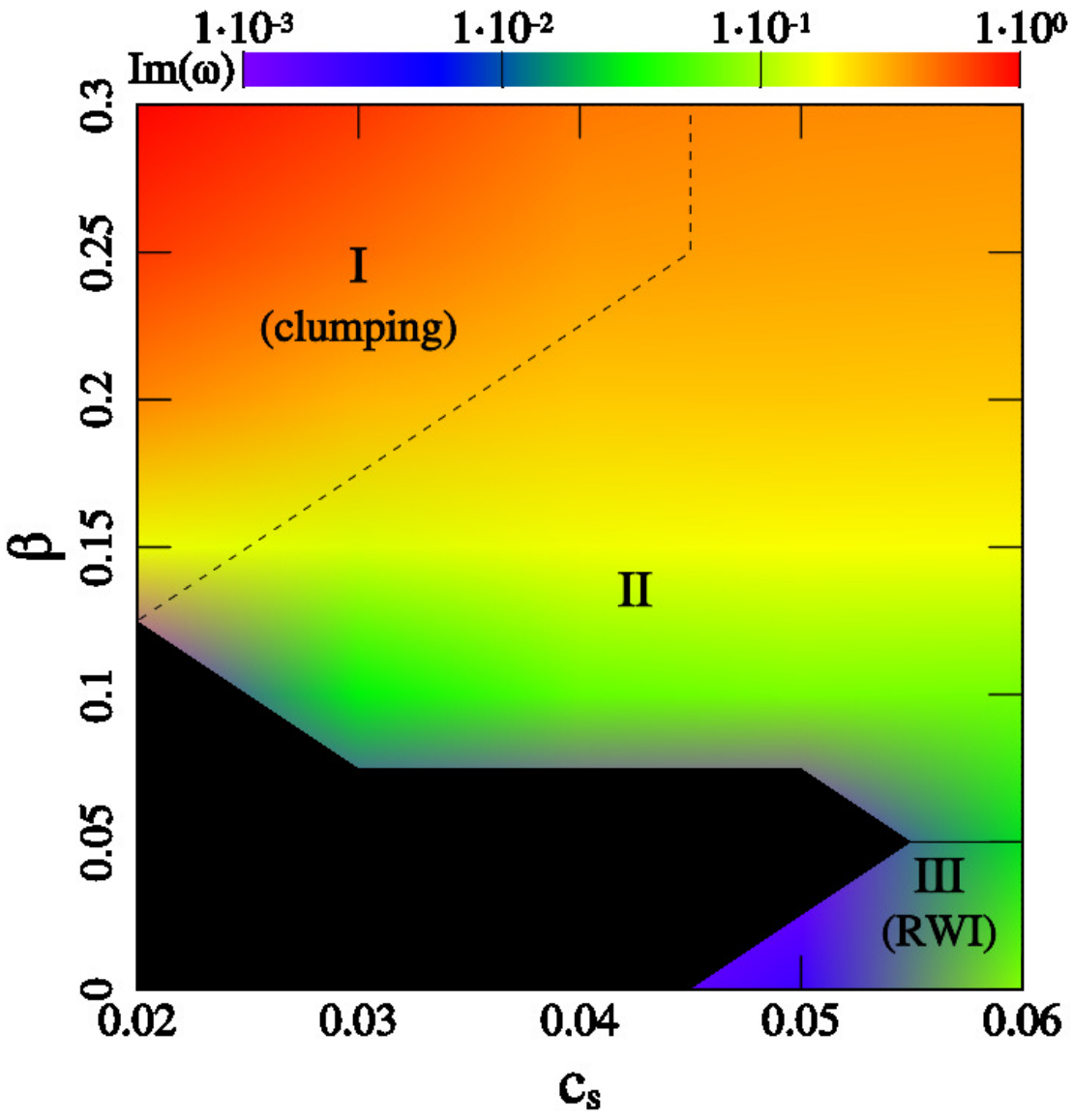

Figure 3.7 Growth rate of the fastest growing mode as a function of β and c_s. The black region is where a positive growth rate is not found with both of our approaches. regions I and II are where IRI operates, while region III sees the purely hydrodynamical RWI. In the nonlinear phase, clumping occurs in region I, where local surface density is enhanced by at least a factor of two, often much higher.

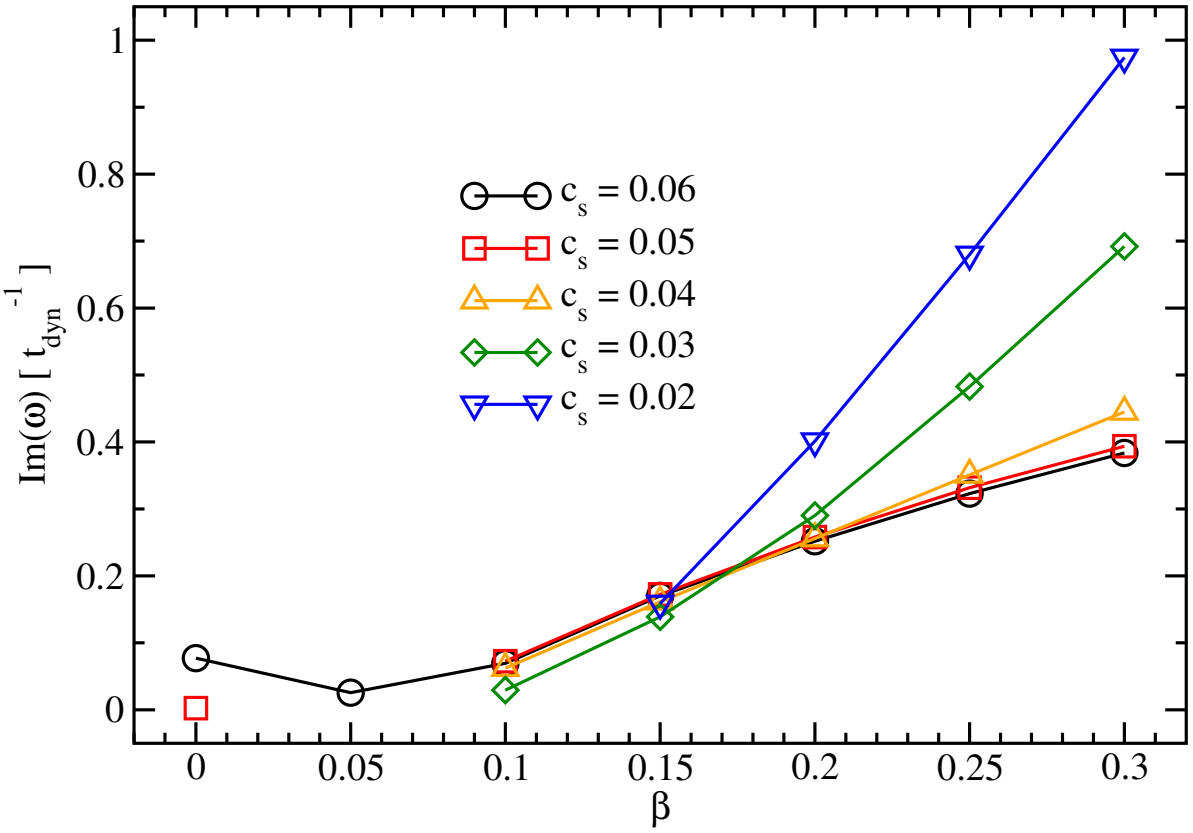

Figure 3.8 Growth rate of the fastest growing mode as a function of β. The $(\beta, c_s) = (0, 0.05)$ point is disconnected because no modal growth is detected at $(\beta, c_s) = (0.05, 0.05)$.

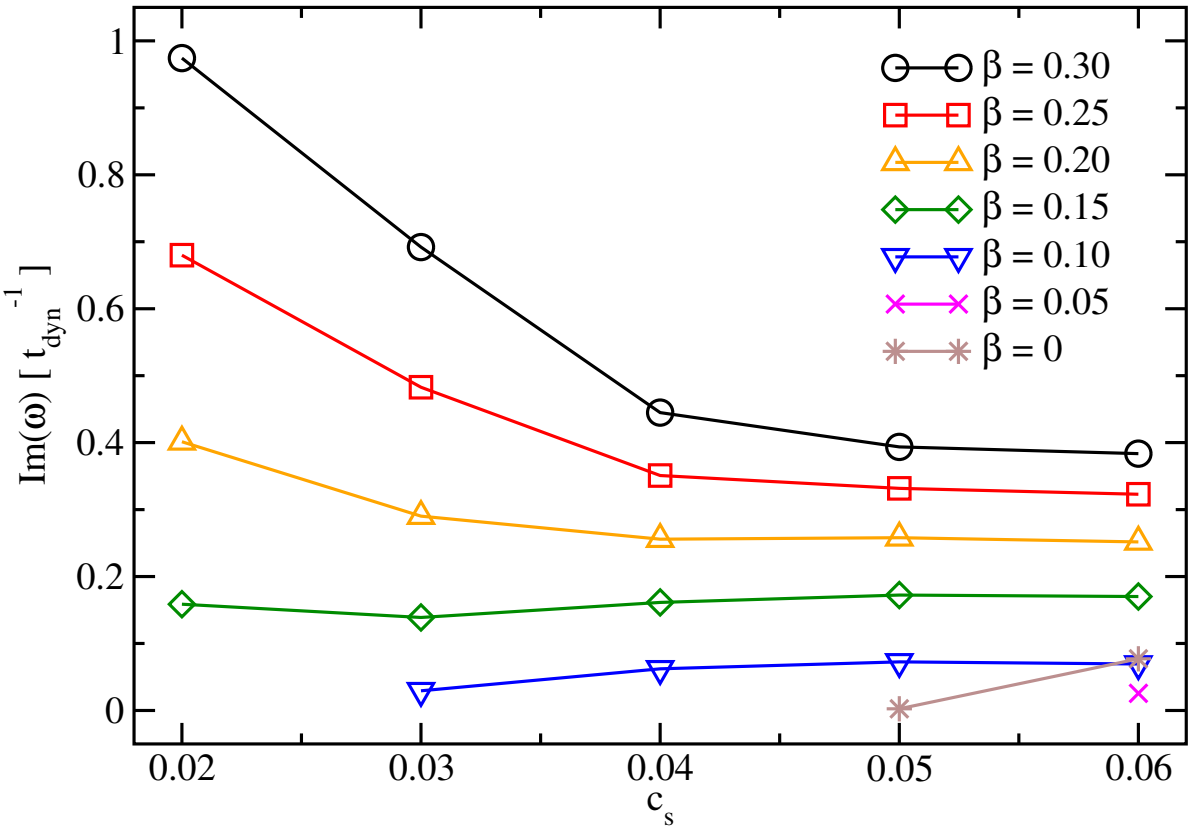

Figure 3.9 Growth rate of the fastest growing mode as a function of c_s.

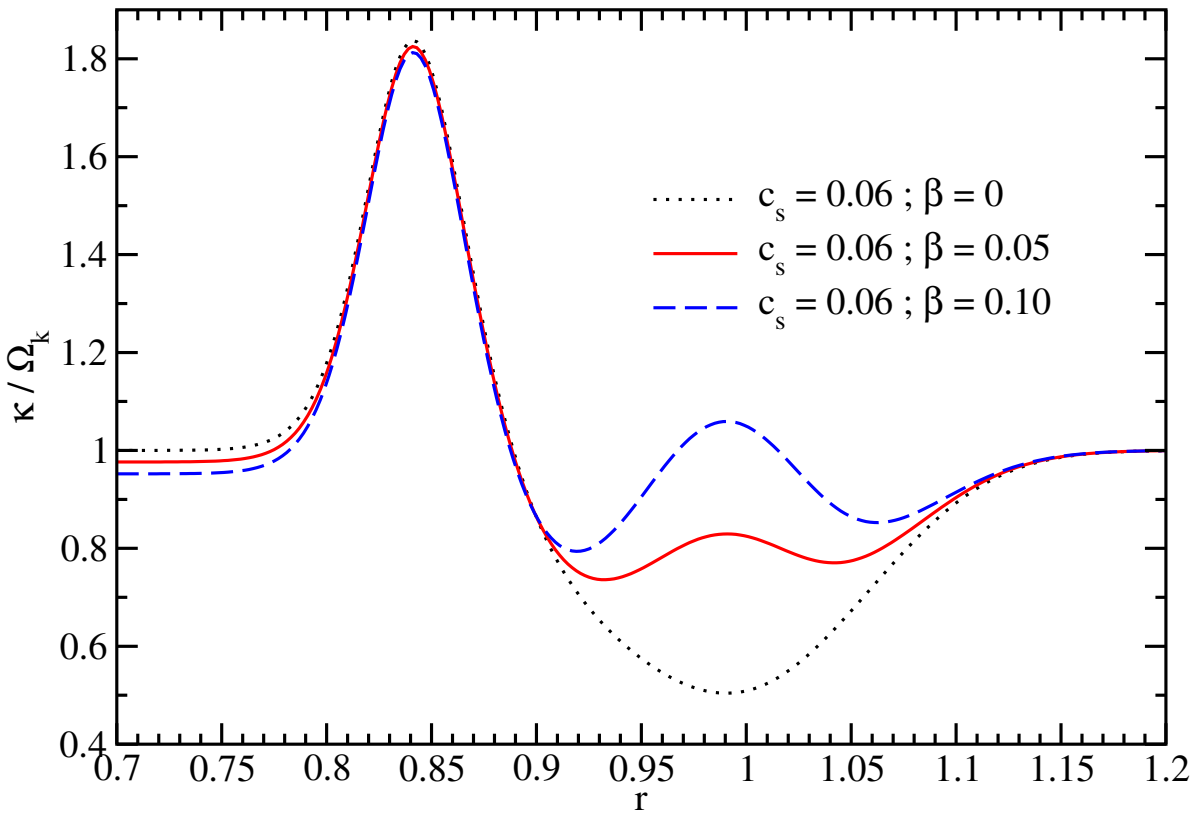

Figure 3.10 κ vs. r for three sets of parameters. The black dotted curve shows κ modified by gas pressure only. As β increases, the local minimum near $r = 1$ is flattened (red solid curve) and reversed (blue dashed curve).

case where the fastest growing mode is $m = 18$. See Table 3.1 for more examples. Similarly, we find that the radial extent of the fastest growing mode becomes more confined as β increases and c_s decreases, as can be seen in Figures 3.5 and 3.6. It is therefore empirically apparent that a higher β encourages the growth of a shorter wavelength mode. While our theory does not make any predictions about which mode grows the fastest, in hindsight this result is not surprising because the radial motion of an IRI mode must be driven by radiation pressure, so a stronger perturbing force should generate a faster radial motion, and therefore a higher frequency wave.

Another trend is that for an increasing β or decreasing c_s, the corotation radius decreases. This is in accordance with our prediction in Section 3.2.2. The dependence is weak but noticeable (see Table 3.1). Curiously, Figure 3.5 and 3.6 show that the peak location of each mode is relatively insensitive to variations in both β and c_s. Consequently, these peaks generally do not coincide with their corotation radii.

For the bulk of this work we do not explicitly vary Δr as a free parameter. Since our choice of $\Delta r = 0.05$ is arbitrary, it is useful to find out to what extent our results would change for a different Δr. Setting $c_s = 0.04$ and $\Delta r = 0.1$, we find the threshold for modal growth becomes $\beta \geq 0.25$, roughly twice as large as when $\Delta r = 0.05$. This suggests that the threshold value scales roughly linearly with Δr for the parameter space we considered.

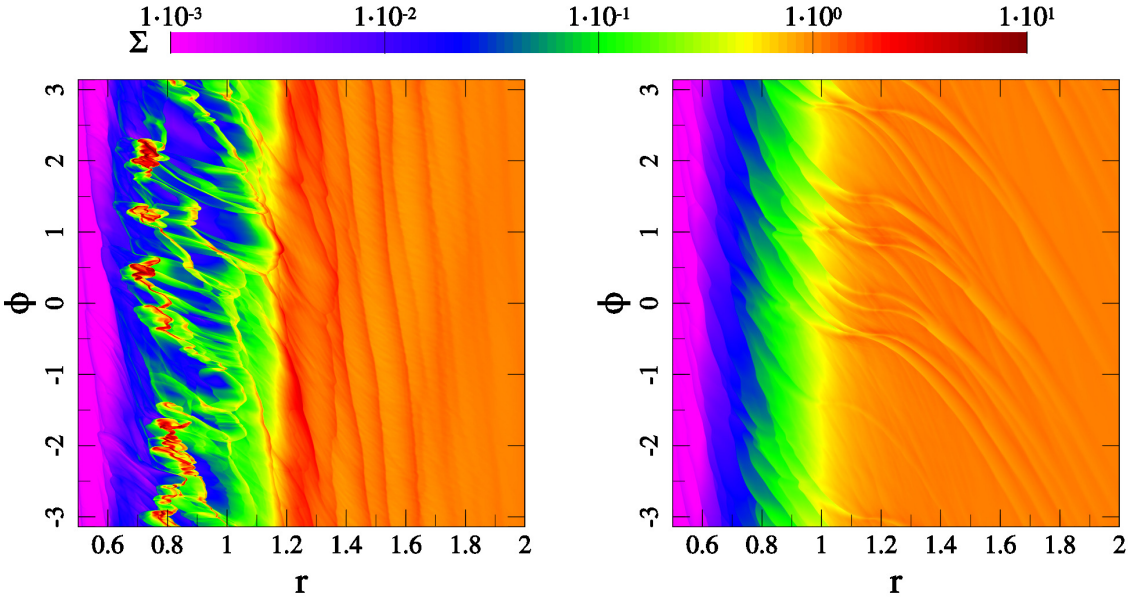

Figure 3.11 Snapshots of our simulations for $(\beta, c_s) = (0.2, 0.02)$ on the left and $(\beta, c_s) = (0.1, 0.05)$ on the right, taken at $t = 100$ orbits. Surface density is shown in logarithmic scale. The simulation on the left, belonging to region I of Figure 3.7, shows very high local surface density, an effect we describe as "clumping". On the right, belonging to region II of Figure 3.7, shows 6 vortices with different orbital frequencies but all lining up near $r = 1.1 \sim 1.2$. Each of these vortices launches two pairs of spiral arms.

3.5.2 Nonlinear Evolution

Nonlinear evolution is what separates region I from region II [3] in Figure 3.7. Figure 3.11 shows the simulation snapshots for the same two sets of parameters in the left and middle panel of Figure 3.5 and 3.6. The left panel, which belongs to region I with $(\beta, c_s) = (0.2, 0.02)$, shows local regions of very high surface density, exceeding 10 times Σ_d, the initial surface density of the disk defined in Equation 3.26. This type of "clumping", which we define as a detection of $\Sigma > 2\Sigma_d$ anywhere in the disk, is characteristic of region I. Note that clumping is not a necessary product of IRI, since region II is also driven by IRI. The transition from region II to I is rapid, in the sense that as we move toward the upper left corner of Figure 3.7, the highest local surface density quickly rises to a few tens of Σ_d.

To compare the two regions in detail, we use the right panel of Figure 3.11 as a typical case for region II. It shows a significant widening of the edge by a factor of ~ 3. Vortices are formed along the edge and they create mild local enhancements in density. Their structure is complex as they typically launch two sets of spiral arms instead of one. The number of vortices is initially equal to the mode number of the fastest growing linear mode, but as they interact with each other, they can occasionally merge. It is unclear how many will remain in the long run since our simulations only last for 100 orbits at most.

In comparison to region II, the clumping in region I creates a much different, almost violent, nonlinear evolution. The clumps are very sharp features, with jumps over three orders of magnitude in density while their sizes are merely $\sim 0.1 r_0$. They are constantly formed and destroyed by disk shear over a dynamical timescale. The destroyed clumps form high density streams that are also visible in the left panel of Figure 3.11. Another consequence of this clumping is that by concentrating a large amount of matter in a small region, radiation is

[3] Region III is omitted from the discussion to maintain focus on IRI. We refer the reader to the literature, e.g. Li et al. (2001); Meheut et al. (2012); Lin (2013), for detailed studies on the nonlinear evolution of RWI.

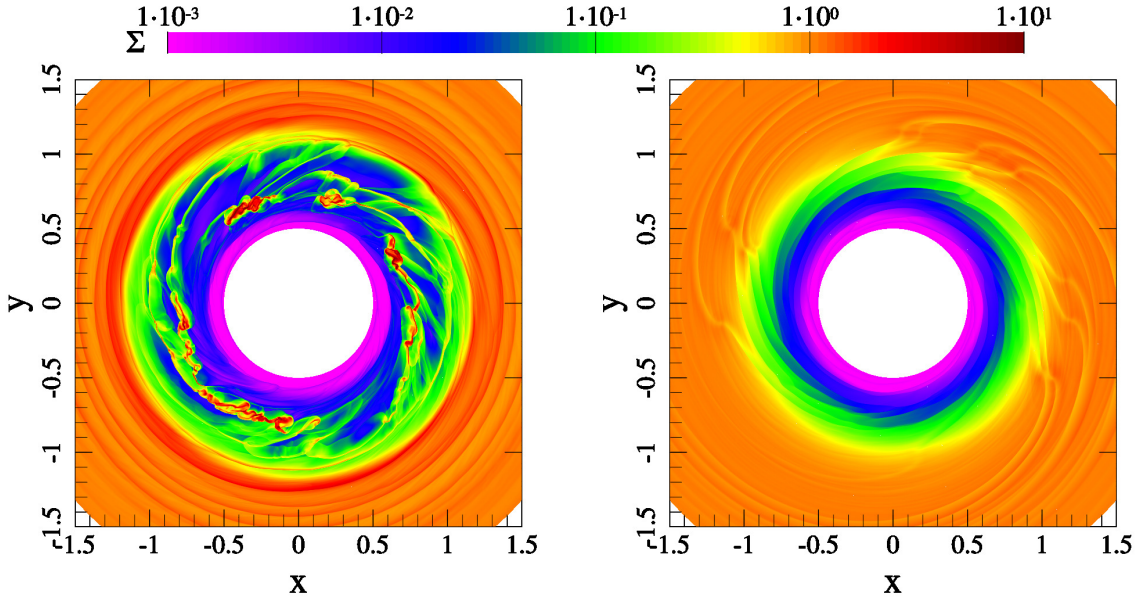

Figure 3.12 Cartesian view of Figure 3.11.

able to penetrate further into the disk and push the edge of the disk to a higher orbit. In the same figure one can see that the edge of the disk is shifted to $\sim 1.2 r_0$. We speculate that the difference between regions I and II is due to the influence of gas pressure. Higher gas pressure results in vortex formation more similar to the purely hydrodynamical RWI, and as gas pressure becomes weaker compared to radiation, sharper features are created.

3.5.3 $\tilde{\tau}_m$ and the Instability Criterion Revisited

In our derivation for the instability criterion in Section 3.2.1, we propose the crude assumption $\tilde{\tau}_m \sim \tau$. Using our semi-analytic method to obtain solutions for η_m, we are able to evaluate $\tilde{\tau}_m$ explicitly. For all IRI modes we have solved semi-analytically, we find $\tilde{\tau}_m/\tau > 1$ within the region $r = \{r_0 - \Delta r, r_0\}$, but it is never an order of magnitude above unity. For example, Figure 3.13 plots $\tilde{\tau}_m/\tau$ for the $m = 18$ mode with $(\beta, c_s) = (0.2, 0.02)$. It shows that within $0.93 \leq r \leq 1.02$, The real part of $\tilde{\tau}_m/\tau$ is within 1 to 6, while the imaginary part is close to vanishing. Using this empirical result we can rewrite Equation 3.22 as:

$$q_\beta \approx \beta e^{-\tau} \left(\frac{\mathrm{d} \ln [r\beta]}{\mathrm{d} \ln r} + f\tau \frac{\mathrm{d} \ln [r\mathscr{R}]}{\mathrm{d} \ln r} \right) , \tag{3.29}$$

where we substitute $\tilde{\tau}_m$ for $f\tau$, and $f > 1$ is a number of order unity. Choosing $f = 3$, Figure 3.14 plots q_β for a few different β. This choice of f puts the threshold for instability ($q_\beta > 1$) at around $\beta = 0.1$, similar to our empirical results.

q_β becomes negative as soon as $\tau > 1$ ($r > r_0 = 1$ for our disk model), because the exponential factor in $\mathscr{R}$ quickly drives its gradient negative. This implies that the instability must originate from the $\tau < 1$ region. Along the same line, we find the corotation radii of IRI modes to be within r_0, as shown in Table 3.1.

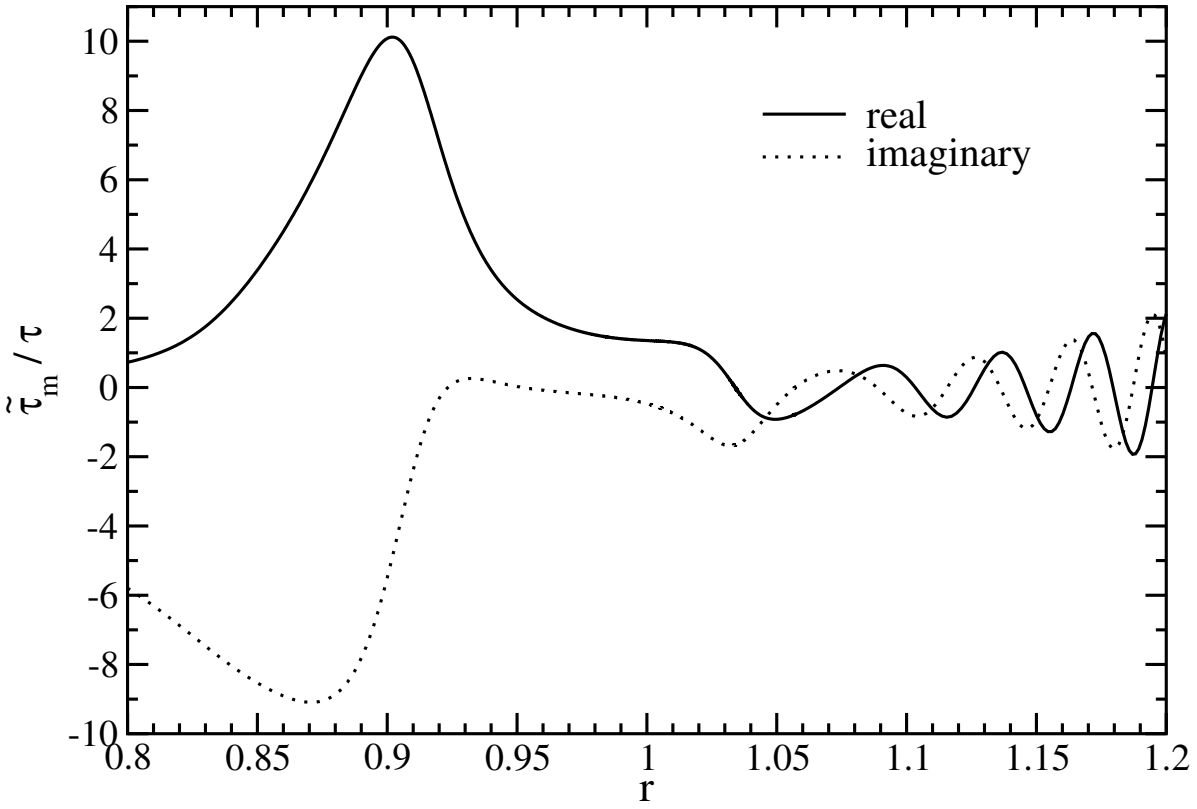

Figure 3.13 $\tilde{\tau}_m/\tau$ for the $m = 18$ mode with $(\beta, c_s) = (0.2, 0.02)$. Inside the transition region, $r \approx \{0.95, 1.0\}$, the approximation $\tilde{\tau}_m \sim \tau$ is accurate to within order unity.

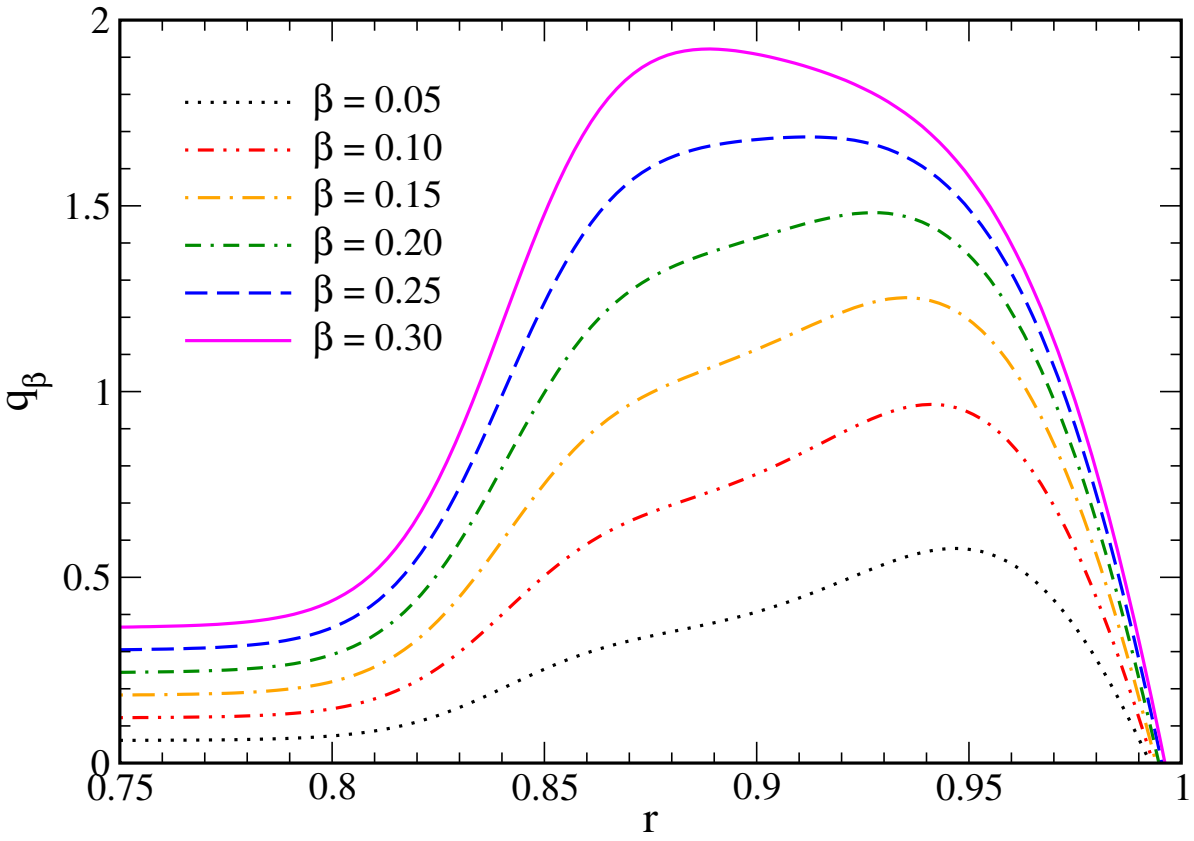

Figure 3.14 q_β from Equation 3.29 for different values of β. We choose $f = 3$ to best match our empirical results.

3.6 Conclusions and Discussions

We demonstrated that IRI can operate at an inner disk edge where there is a transition from being radially transparent to opaque. A local criterion for axisymmetric instability was derived (Equation 3.22). For our given disk model we computed the linear modal growth rates for β varying from 0 to 0.3, and c_s from 0.02 to 0.06. We found growth rates ranging from 10^{-2} to $10^{0}\ t_{\rm dyn}^{-1}$ (Figure 3.7). The fastest rates were found for the largest β and smallest c_s. We empirically determined that the threshold for IRI is $\beta \sim 0.1$ when $\Delta r = 0.05$, with a weak dependence on c_s. For a wider edge, $\Delta r = 0.1$, this threshold rises to $\beta \sim 0.25$. We note that this implies the threshold can be lowered by reducing Δr; however, at the same time c_s must also be lowered for IRI to dominate over other forms of instability that may be triggered by the sharpness of the edge, such as RWI and Rayleigh instability. We employed two independent approaches to obtain the growth rates of the linear modes: simulating the disks numerically using `PEnGUIn`, and solving the linearized equations semi-analytically. Their excellent agreement lends confidence in our results. Moreover, we discovered a parameter space, labeled region I in Figure 3.7, where "clumping" occurs. There one can find over 10 times the local surface density enhancements in the nonlinear evolution of IRI.

3.6.1 Connection to Physical Disks

Our disk model is inspired by transitional disks (e.g. Calvet et al., 2005; Espaillat et al., 2007, 2008; Andrews et al., 2011). The inner edges of these disks are currently unresolved by observation, but theoretical work has shown that the sharpness of disk edges created by X-ray photoevaporation (e.g. Owen et al., 2010) is similar to that described by our Equation 3.26 with $\Delta r = 0.05$ (compare our Figure 3.2 to Figure 2 of Owen et al. (2013)). If a transitional disk undergoes IRI, the asymmetric structure at the inner edge will create an azimuthal variation in shadowing. Flaherty & Muzerolle (2010) showed that this can lead to a significant variation in disk emission. Indeed, some variability in the infrared emission of transitional disks has been reported by Muzerolle et al. (2009), Flaherty et al. (2011), and Espaillat et al. (2011).

On the other hand, IRI is by no means limited to circumstellar disks. AGN accretion disks, for example, can be subjected to IRI if there are any sharp jumps in density and/or opacity, such as the inner edges of the board-line regions. IRI can potentially generate the stochastic asymmetry, which is used to explain the variability in the double-peaked Balmer emission lines in radio-loud AGNs (Flohic & Eracleous, 2008). We note that the dynamics in AGN accretion disks are considerably more complicated since they do not have a point-like light source.

3.6.2 Implications of "Clumping"

The "clumping" found in a part of our parameter space (Figure 3.7) opens new possibilities for IRI. For instance, very high density regions in protoplanetary disks may be favorable environments for the formation of planetary cores. The density of individual clumps may even become high enough to trigger gravitational instability at the inner edges of massive disks. One should be cautious to interpret the enhancement factors reported as realistic, however, since it is only one disk model that we have studied.

The clumping also leads to a possibility of preventing inward dust migration. Dominik & Dullemond (2011) demonstrated that while radiation pressure can initially push dust outward and form a dust wall, the wall eventually succumbs to the global accretion flow and migrates inward. If this wall becomes unstable due to IRI, clumping can occur, effectively creating "leakage" within the wall, allowing radiation to push dust further back. The true behavior of these dust walls is important to understand disks where inner clearings have been observed, such as

transitional disks. Dynamical interactions between radiation, dust, and gas must be considered for this kind of study.

3.6.3 Outlook

There are three main aspects of our model that we feel would benefit greatly from a more realistic treatment. First, our model ignores the vertical dimension. A notable difference from 2D to 3D is that the location of the inner edge of a disk, defined as the $\tau = 1$ point, would become a function of height, spreading over a distance of $\sim h$. One possible consequence is that IRI would generate a vertical circulation at the inner edge, which would dilute the opacity in the midplane and allow radiation pressure to penetrate further into the disk. Additionally, in a flared disk, radiation pressure is exerted on the photosphere of the entire disk rather than just the inner edge. On the other hand, because of dust settling, we expect the value of β in the photosphere to be smaller than the midplane, making it even more difficult to reach the $\beta \approx 0.1$ threshold. Nonetheless, for disks around exceptionally luminous stars, IRI can potentially operate at all radii.

Second, we assume a perfect coupling between gas and dust. In a more realistic approach, dust should be allowed to migrate with respect to gas. One expects dust to gather near the initial $\tau = 1$ point, because where it is optically thin, dust migrates outward due to the effect of radiation pressure, and in the optically thick disk, dust migrates inward due to gas drag. This behavior of dust is described in Section 3 of Takeuchi & Artymowicz (2001). The buildup of a dust wall is almost certain to trigger IRI due to its large β gradient.

Lastly, we lack a realistic treatment for radiative transfer. As the disk crosses from being radially transparent to opaque, the midplane of the disk also transitions from being heated directly by irradiation, to passively by the irradiated atmosphere. Consequently the midplane temperature should be decreasing across the disk edge. This is not captured by our globally isothermal assumption. Additionally, the clumps we find in some of our nonlinear results are sufficiently dense that they are optically thick. With our isothermal treatment, they remain the same temperature as their surroundings, while in truth these clumps should be capable of shielding themselves from irradiation and creating a non-trivial internal temperature structure. Whether this is an effect that aids or inhibits their formation and survival requires future investigation.

Chapter 4

Gap Opening by Giant Planets

A version of this chapter has been published in The Astrophysical Journal as "How Empty are Disk Gaps Opened by Giant Planets?", Fung, J., Shi, J., and Chiang, E., volume 782, issue 2, article id. 88, 2014. Reproduced by permission of the AAS.

4.1 Introduction

Observational studies of giant planet formation will begin in earnest once we detect planets still embedded in their natal gas disks. Directly imaging young gas giants is made easier by their ability to clear material away from their orbits. Planetary (Lindblad) torques open gaps while viscous torques fill them back in (Section 1.3.1 and references therein): a balance between these torques sets the equilibrium surface density near the planet. In the previous chapter we examined how the inner edges of "transitional" disks may be susceptible to the irradiation instability, but we have yet discussed how the edges are formed in the first place. It has been proposed that the optically thin cavities in these disks are hints of gap clearing by planets (e.g., Kraus & Ireland, 2012; Debes et al., 2013; Quanz et al., 2013). Transition disk holes are surprisingly large; one gap opened by a single planet would be too narrow to explain the observed cavity sizes that range up to ~100 AU, and so a given system might have to contain multiple super-Jovian planets to clear a wide enough swath (e.g., Zhu et al., 2011; Dodson-Robinson & Salyk, 2011). Even so, the holes are so optically thin that planets alone seem incapable of torquing material strongly enough to compete with viscous diffusion — at least for typically assumed parameters — and appeals are made to opacity reductions through grain growth or dust filtration at the outer gap edge (Zhu et al., 2012; Dong et al., 2012). Part of the motivation of our study is to expand the parameter space explored and see how empty a gap can be cleared.

Determining gap surface densities and corresponding optical depths is relevant not only for observations but also for theory: material that corotates with the planet (executing quasi-horseshoe orbits) can backreact gravitationally on the planet and influence its dynamical evolution. The corotation torque takes its place among the litany of resonant planet-disk interactions that can alter orbital eccentricities and semimajor axes (Section 1.3.2; also see Kley & Nelson 2012 for a review). The delicate balance of forces within the gap, including the thermodynamic behavior of matter there, may determine how low-mass and giant planets survive the threats of Types I and II orbital migration (Ward 1997; see also section 2.2 of Kley & Nelson 2012 and references therein).

Despite its importance, $\Sigma_{\rm gap}$ — the surface density averaged over the bottom of the gap — remains poorly understood. Notwithstanding the huge number of simulations of planet-disk interactions published in the past

two decades,[1] a systematic parameter study has yet to be performed that determines $\Sigma_{\rm gap}$ as a function of planet-to-star mass ratio $q \equiv M_p/M_*$; Shakura-Sunyaev viscosity parameter α; and disk height-to-radius aspect ratio (equivalently, disk temperature) h/r. Crida et al. (2006) examined how these parameters influence gap shape and width, but not gap depth — i.e., they studied the onset of the gap, but not the bottom of the gap. The numerics can be challenging. Measuring $\Sigma_{\rm gap}$ accurately requires global simulations that (i) resolve the disk well in at least azimuth ϕ and radius r; (ii) resolve large density contrasts; (iii) converge with time to a steady state; (iv) model how the planet accretes from the ambient flow; and (v) possess well-separated radial boundaries that maintain a steady mass accretion rate across the entire domain, i.e., the simulation presumably should enforce $\dot{M}(r) =$ constant $\neq 0$, as befits real accretion disks. Feature (v) is captured by only a minority of studies, and feature (iv) can only be mocked up in a parameterized way (e.g., Lubow & D'Angelo, 2006; Zhu et al., 2011).

This chapter aims to provide an empirical relation for $\Sigma_{\rm gap}(q, \alpha, h/r)$ for a single non-accreting giant planet on a fixed circular orbit embedded in a 2D, locally isothermal, steadily accreting disk. We restrict our study to disk gas only, and ignore how dust and gas flows might differ. We utilize two independent codes: `ZEUS` (Stone & Norman, 1992), and `PEnGUIn`, a new Lagrangian PPM (piecewise parabolic method)-based code that we have implemented on multiple GPUs (graphics processing units). To the extent possible, results from one code will be validated against the other.

4.1.1 An Analytic Scaling Relation

Although our study is primarily numerical, we derive here an approximate analytic relation for $\Sigma_{\rm gap}$ that we will use to put our numerical results in context. We admit at the outset that our derivation can hardly be called such, as our reasoning will ignore many details and make assumptions not carefully justified. But as the rest of this chapter will show, the simple relation we now present will yield results surprisingly close to those of detailed numerical simulations.

First examine the outer disk, exterior to the planet's orbit. The outer Lindblad torque exerted by the planet transmits angular momentum outward at a rate:

$$T_{\rm L} \sim q^2 \left(\frac{r}{h}\right)^3 \Sigma_{\rm gap}\, \Omega^2 r^4 \tag{4.1}$$

where Ω is the disk angular frequency (Goldreich & Tremaine, 1980). This is the total torque from linear perturbations to one side of the planet, integrated over all resonances up to the torque cut-off at azimuthal wavenumber $m \sim r/h$ (see, e.g., equation 2 of Crida et al. 2006, and references therein). Note that we have used $\Sigma_{\rm gap}$, the surface density averaged over the bottom of the gap, in our evaluation of the integrated Lindblad torque. Our justification for this choice is that the integrated torque is dominated by resonances at the torque cut-off, i.e., at distances $\sim h$ from the planet (see figures 2 and 3 of Goldreich & Tremaine 1980), and bottoms of gaps opened by giant planets in gas disks typically extend this far (see, e.g., the simulated gaps of Crida et al. 2006, or our Figure 4.2). One caveat is that the linear theory from which equation (4.1) originates is formally valid only for low-mass planets for which $q \lesssim (h/r)^3$; the highest-mass planets we simulate will violate this condition, and indeed for such super-Jovian objects our numerical simulations will reveal deviations from the simple-minded scaling law we derive in this section.

In steady state, the outward transmission of angular momentum by the (outer) Lindblad torque must be bal-

[1] Much of the literature is marked by a peculiar insistence on plotting surface density Σ on a linear scale. The practice is unhelpful since surface density contrasts in and out of gaps can span orders of magnitude — indeed they must if they are to reproduce the enormous optical depth contrasts inferred from observations of transition disks (e.g., Dong et al. 2012).

anced by the angular momentum transmitted inward by the viscous torque:[2]

$$T_{\rm v} \sim \Sigma_0 \, \nu \, \Omega \, r^2 \tag{4.2}$$

where $\nu = \alpha(h/r)^2 \Omega r^2$ is the kinematic viscosity (e.g., Frank et al., 2002). Here we have used Σ_0, the surface density at the gap periphery — or more conveniently, the surface density at the planet's location if the planet were massless — to evaluate $T_{\rm v}$. The viscous torque depends on the gradient of Σ, and this gradient is larger outside the flat-bottomed gap than inside it (see, e.g., the gap profile shown in Figure 4.2; by "gap periphery" we mean a location like $r \approx 1.4$, where the gradient might reasonably be approximated as Σ_0/r, which is what equation 4.2 essentially assumes). Conscripting Σ_0 in this way is a gross simplification, but alternatives would require that we actually compute the precise shape of the gap, which is what we are trying to avoid with our order-of-magnitude derivation.

Usually one thinks of viscous torques as transmitting angular momentum outward, but in the gap edge of the outer disk, the direction of viscous transport is inward because the surface density there has a sharp and positive gradient ($d\Sigma/dr > 0$).

Setting $T_{\rm L} = T_{\rm v}$ yields

$$\frac{\Sigma_{\rm gap}}{\Sigma_0} \sim \frac{\alpha(h/r)^5}{q^2} \, . \tag{4.3}$$

Exactly the same scaling relation applies to the inner disk, interior to the planet's orbit. The signs in the inner disk are reversed from those in the outer disk: the inner Lindblad torque transmits angular momentum inward, while the local viscous torque transmits angular momentum outward.

As we were completing our numerical tests of equation (4.3) and preparing our manuscript for publication, we became aware of the study by Duffell & MacFadyen (2013) who found the same scaling relation on purely empirical grounds (although these authors did not explicitly vary the Mach number r/h). Duffell & MacFadyen (2013) concentrated on the low-mass $q \lesssim 10^{-4}$ regime. Our study complements theirs by studying the high-mass $q \gtrsim 10^{-4}$ regime; we will see to what extent equation (4.3) also holds true for giant planets. See also our Section 4.4.3 where we discuss to what extent the short derivation given in this subsection captures the whole story.

This chapter is organized as follows: Section 4.2 contains our numerical methods and simulation parameters. Section 3 presents our results for $\Sigma_{\rm gap}(q, \alpha, h/r)$. Section 4 concludes and charts directions for future work.

4.2 Numerical Methods

We numerically simulate a planet on a fixed circular orbit embedded in a co-planar, viscously accreting disk. Our two independent codes, `PEnGUIn` and `ZEUS`, solve the usual continuity and momentum equations in two dimensions, similar to Equations 3.2 and 3.3; we give here the equations in the inertial, barycentric frame:

$$\frac{D\Sigma}{Dt} + \Sigma \, (\nabla \cdot \mathbf{v}) = 0 \, , \tag{4.4}$$

$$\frac{D\mathbf{v}}{Dt} = -\frac{1}{\Sigma} \nabla P + \frac{1}{\Sigma} \nabla \cdot \mathbb{T} - \nabla \Phi \, , \tag{4.5}$$

where D/Dt is the Lagrangian derivative, Σ is the surface density, P is the vertically averaged pressure, $\mathbf{v}$ is the velocity, $\mathbb{T}$ is the Newtonian viscous stress tensor, and Φ is the gravitational potential of the central star and planet

[2] There is also a so-called "pressure" torque of comparable magnitude to the Lindblad and viscous torques (Crida et al., 2006), but we neglect this third torque for our order-of-magnitude derivation.

(but not the disk). In polar coordinates, $\mathbf{v} = (v_r, \Omega r)$; in component form, equation (4.5) reads:

$$\frac{Dv_r}{Dt} = -\frac{1}{\Sigma}\frac{\partial P}{\partial r} + \frac{2}{\Sigma r}\frac{\partial}{\partial r}\left(\nu\Sigma r\frac{\partial v_r}{\partial r}\right) + \frac{1}{\Sigma r}\frac{\partial}{\partial \phi}\left[\nu\Sigma\left(r\frac{\partial \Omega}{\partial r} + \frac{1}{r}\frac{\partial v_r}{\partial \phi}\right)\right] - \frac{\partial \Phi}{\partial r}\,, \tag{4.6}$$

$$\frac{D(r\Omega)}{Dt} = -\frac{1}{\Sigma r}\frac{\partial P}{\partial \phi} + \frac{2}{\Sigma r}\frac{\partial}{\partial \phi}\left(\nu\Sigma\frac{\partial \Omega}{\partial \phi}\right) + \frac{1}{\Sigma r^2}\frac{\partial}{\partial r}\left[\nu\Sigma r^2\left(r\frac{\partial \Omega}{\partial r} + \frac{1}{r}\frac{\partial v_r}{\partial \phi}\right)\right] - \frac{1}{r}\frac{\partial \Phi}{\partial \phi}\,. \tag{4.7}$$

Here $\nu = \alpha c_s h$ is the kinematic viscosity following Shakura & Sunyaev (1973), with c_s equal to the sound speed. We complete the equation set with a locally isothermal equation of state $P = \Sigma c_s^2$, with $c_s \propto r^{-1/2}$ so that the disk aspect ratio h/r = constant.

In the center-of-mass frame,

$$\Phi = -\frac{GM_*}{\sqrt{r^2 + r_1^2 + 2rr_1\cos(\phi - \phi_p)}} - \frac{GM_p}{\sqrt{r^2 + r_2^2 - 2rr_2\cos(\phi - \phi_p) + r_s^2}}, \tag{4.8}$$

where M_* and $M_p = qM_*$ are the masses of the star and the planet, respectively; $r_1 = qr_p/(1+q)$ and $r_2 = r_p/(1+q)$ are their radial positions, with r_p the total (fixed) separation; $\phi_p - \pi$ and ϕ_p are their angular positions; and r_s is the softening length of the planet's potential. (Note this expression is identical to Equation 2.39, which we restate here for completeness.) We set $G(M_* + M_p) = 1$ and $r_p = 1$ so that the planet's orbital frequency $\Omega_p = 1$ and period $P_p = 2\pi$.

We use our GPU-based, Lagrangian, dimensionally-split, shock-capturing hydrodynamics code `PEnGUIn` (Chapter 2) to simulate disk gaps. Technically, `PEnGUIn`'s reference frame is a barycentric frame that rotates at Ω_p; thus the planet's position is fixed in time, but the Coriolis force is not computed as an explicit source term. Rather, it is absorbed into the conservative form of the angular momentum equation (Kley, 1998).

The hardware used are three GTX-Titan graphics cards all connected to a single node. Running in double precision on all three cards simultaneously, `PEnGUIn` takes 12 seconds to run per planetary orbit for $(q, \alpha, h/r) = (10^{-3}, 10^{-2}, 0.05)$.

4.2.1 `ZEUS90`: Code Description

For comparison with `PEnGUIn`, we also carried out simulations with `ZEUS90`: a modern version of `ZEUS` (Stone & Norman, 1992; Hawley & Stone, 1995) written in FORTRAN 90. It is a three-dimensional, operator-split, time-explicit, Eulerian finite-differencing magnetohydrodynamics code, widely used to simulate a variety of systems, including magnetorotationally-unstable circumbinary disks (Shi et al., 2012) and warped disks (Sorathia et al., 2013). For our application, we suppress the vertical dimension and magnetic fields; implement the Navier-Stokes viscosity module; and add a planetary potential having a specified time dependence. The von Neumann-Richtmyer artificial viscosity, commonly used to capture shock waves, is switched off in the presence of an explicit viscosity. The reference frame for `ZEUS90` is a non-rotating frame centered on the star, and so ordinarily there is an extra term in Φ due to the indirect potential: $GM_p r\cos(\phi - \phi_p)/r_p^2$. However, we find in practice that the indirect term results in a "wobbling" of the disk that is difficult to reconcile with our fixed boundary conditions on fixed circles (see equations 4.9–4.11 below and surrounding discussion). The wobbling generates spurious time variability that increases with increasing q; therefore we drop the indirect potential in all `ZEUS90` simulations with $q \geq 1 \times 10^{-3}$ (apparently Lubow et al. 1999 also dropped the indirect potential in their simulations with `ZEUS`; see also Zhu et al. 2011 who found that their planet-disk simulations were not sensitive to the indirect term). Running in

double precision on 128 cores in parallel, ZEUS90 takes 26 seconds to run per planetary orbit for $(q, \alpha, h/r) = (10^{-3}, 10^{-2}, 0.05)$.

4.2.2 Numerical Setup

For our parameter study we vary:

- q from 10^{-4} to 10^{-2},
- α from 10^{-3} to 10^{-1}, and
- h/r from 0.04 to 0.1.

Other properties of our simulations are as follows.

Initial and boundary conditions

Our simulation domain spans 0 to 2π in azimuth, and from $r_{\rm in} = 0.4$ to $r_{\rm out} = 2.5$ in radius (in units where the planet-star separation $r_{\rm p} = 1$). Initial conditions correspond to a steady-state accretion disk having constant α, constant h/r, and a rotation curve modified by the radial pressure gradient:

$$\Sigma = \Sigma_0 \, (r/r_{\rm p})^{-1/2} \,, \tag{4.9}$$

$$v_r = -\frac{3}{2}\,\alpha \left(\frac{h}{r}\right)^2 \sqrt{\frac{G(M_* + M_{\rm p})}{r}} \,, \tag{4.10}$$

$$\Omega = \sqrt{1 - \frac{1}{2}\left(\frac{h}{r}\right)^2} \sqrt{\frac{G(M_* + M_{\rm p})}{r^3}} \,, \tag{4.11}$$

with $\Sigma_0 = 1$ (we could have chosen any value for Σ_0 because we do not calculate the gravity of the disk—the disk exerts no gravitational backreaction on the planet nor does it self-gravitate). At both inner and outer radial boundaries, we fix Σ, v_r and Ω to their values determined by the above equations. These fixed boundary conditions ensure a steady inflow of mass across the simulation domain — as is appropriate for real accretion disks.[3] Figure 4.1 illustrates how our boundary conditions enable a planet-less disk to relax over a viscous diffusion timescale to the equilibrium profile described by equations (4.9)–(4.11).

Ideally the radial boundaries should be placed far enough away that waves launched from the planet damp before they reach the edges of the domain. Goodman & Rafikov (2001) calculated that nonlinear steepening of waves causes them to dissipate over lengthscales of $\sim 3h$ (for $q = 10^{-3}$ and $h/r = 0.05$; the damping length scales as $q^{-0.4}$ and $(h/r)^{2.2}$). Our outer radial boundary of $r_{\rm out} = 2.5$ ($\sim$15–40h away from the planet, depending on our choice for h/r) is distant enough that outward-propagating waves largely dissipate within the domain. For our inner radial boundary of $r_{\rm in} = 0.4$ ($\sim$6–15h away from the planet), the situation is more marginal; depending on the simulation, waves are still present at our disk inner edge. However, the main focus of our work is the surface density deep within the gap, and we have verified that this quantity changes by no more than $\sim$10% as we shrink $r_{\rm in}$ from 0.4 to 0.2. Thus we opt for the larger boundary to keep code timesteps longer. For simplicity we eschew wave-killing zones (cf. de Val-Borro et al. 2006). In practice, the Godunov-type scheme used by PEnGUIn is effective at absorbing waves at fixed boundaries, even more so than using wave-killing zones (Zhaohuan Zhu 2013, personal communication).

[3] By contrast, many popular codes for planet-disk simulations (e.g., FARGO) default to a zero-inflow solution; for a compilation of codes from the community, see, e.g., de Val-Borro et al. (2006).

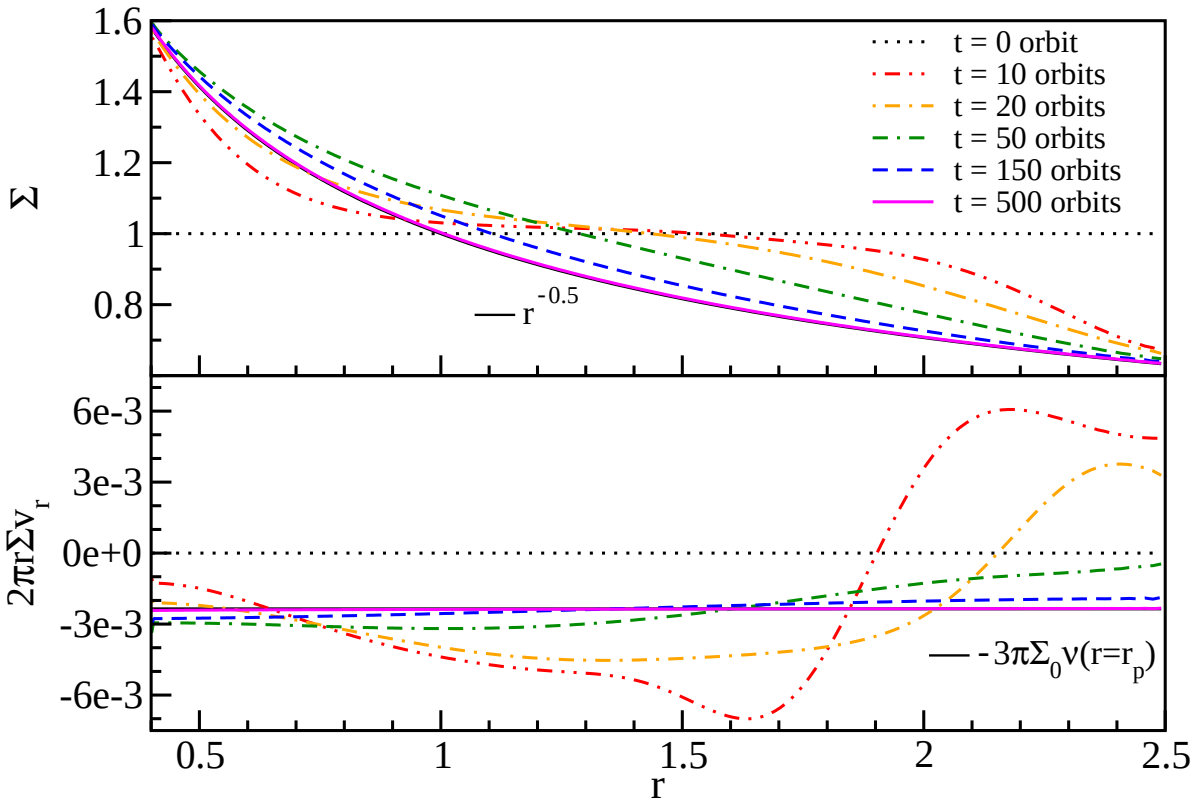

Figure 4.1 Viscous relaxation to steady-state accretion in a disk with $(q, \alpha, h/r) = (0, 0.1, 0.05)$. At $t = 0$, we set $\Sigma = 1$ and $v_r = 0$ except at the boundaries, where conditions are given by equations (4.9)–(4.11). Black solid lines denote the steady-state density profile and accretion rate to which `PEnGUIn` correctly relaxes over a viscous timescale.

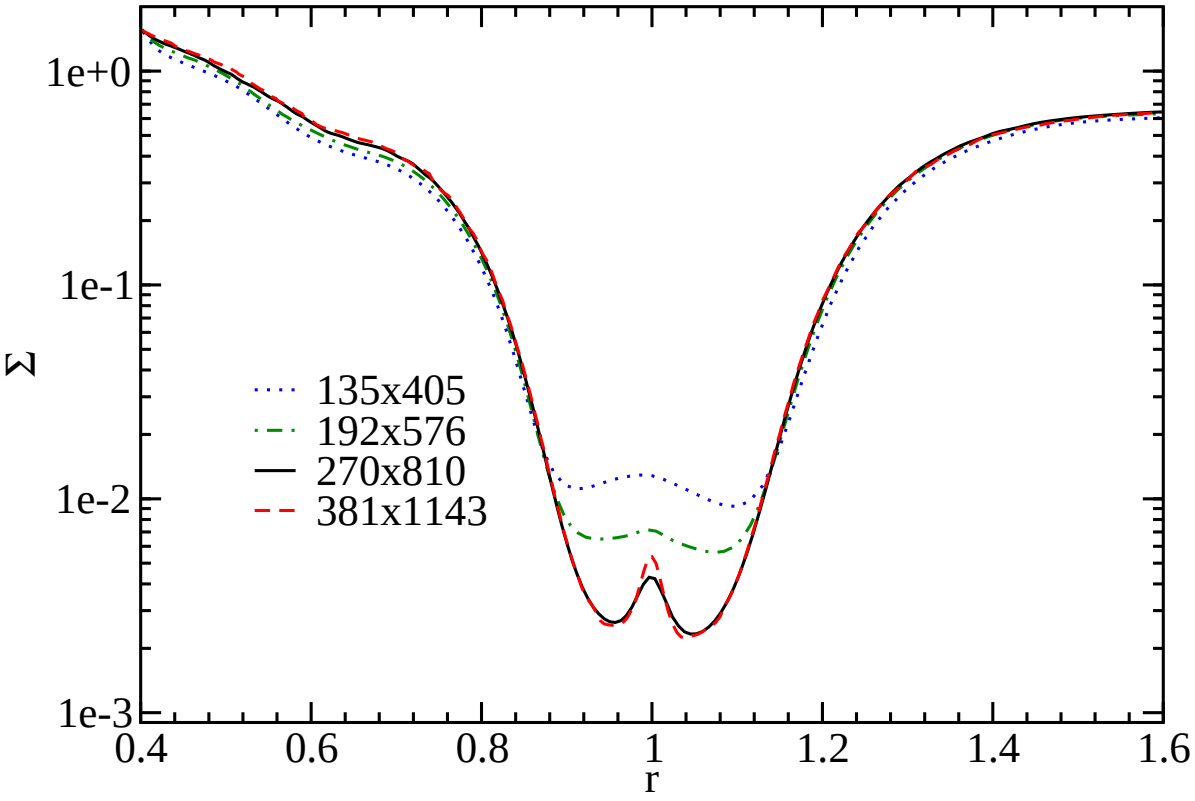

Figure 4.2 Convergence of gap profile with grid resolution for $(q, \alpha, h/r) = (10^{-3}, 10^{-3}, 0.05)$ using PEnGUIn. The dot-dot-dashed curve represents the initial density profile, equal to the density profile in the absence of the planet (equation 4.9). The surface density Σ plotted here is azimuthally averaged. For PEnGUIn science runs, we adopt $270\,(r) \times 810\,(\phi)$ for $h/r = 0.05$, and adjust the cell size to scale with h/r (see Section 4.2.2).

To avoid strong shocks at the beginning of the simulation, the planet mass is ramped from zero to its assigned value over an initial "warm-up" phase. In PEnGUIn, M_p increases according to $M_\mathrm{p}(t)/M_* = q \sin^2[(\Omega_\mathrm{p} t/20)(10^{-3}/q)]$. For $q = 10^{-3}$, this takes 5 orbits. In ZEUS90, the planet mass grows linearly from zero to its desired value in 1 orbit. Both warm-up schemes proved stable.

Grid resolution

Our grid spacings are logarithmic in radius and uniform in azimuth. For $h/r = 0.05$, the resolution is $270\,(r) \times 810\,(\phi)$ for PEnGUIn and 256×864 for ZEUS90 (the latter choice yields square grid cells). Figure 4.2 attests that gap surface densities have largely converged at our standard resolution. We scale our grid cell size with h/r, i.e., with sound speed c_s, so that sound waves of a given frequency are equally well resolved between simulations. Cold disks with small h/r are especially costly, which is why we do not vary h/r below 0.04. Code timesteps scale with grid cell sizes according to the Courant-Friedrichs-Lewy condition, with the Courant number chosen to be 0.5 for PEnGUIn and 0.4 for ZEUS90.

Softening length r_s

For both PEnGUIn and ZEUS90, the planetary potential's softening length is fixed at $r_s = 0.028 r_\mathrm{p}$ or about 4 local grid cell lengths. Equivalently, $r_s = 0.56h$ for $h/r = 0.05$, and $r_s = 0.25$ Hill radii R_H for $q = 10^{-3}$. Any choice for

$r_s \sim h$ or $r_s \sim R_H$ seems reasonable insofar as our 2D treatment of the gas dynamics must break down at distances from the planet less than the vertical thickness of the disk, and because the planet's mass may, in reality, be distributed over a distended envelope or circumplanetary disk. Tests with `PEnGUIn` at $(q, \alpha, h/r) = (10^{-3}, 0.1, 0.05)$ revealed that $r_s \lesssim 0.4h$ caused the surface density to converge substantially more slowly with time. Specifically, for the aforementioned parameters and r_s too small, the gap deepened rapidly, overshot its equilibrium value, and took thousands of orbits to approach a steady state. By contrast, for $r_s = 0.56h$, the surface density equilibrated in a mere ~30 orbits at our standard resolution, with higher grid resolutions yielding similar results.

According to Müller et al. (2012), our choice of softening length yields a 2D gravitational force that matches the vertically averaged, 3D force to within 10% at a distance $\gtrsim 2h$ away from the planet.

Σ_{gap} and convergence with time

Our metric for gap depth is the space- and time-averaged surface density Σ_{gap} in the planet's co-orbital region, normalized to $\Sigma_0 = 1$ (the surface density at $r = r_p = 1$ in the absence of the planet). As judged from snapshots like those shown in Figures 4.3 and 4.4, the annulus spanning $r = r_p - \Delta$ to $r_p + \Delta$ with $\Delta \equiv 2\max(R_H, h)$, excised from $\phi = \phi_p - \Delta/r_p$ to $\phi_p + \Delta/r_p$, is visibly depleted and reasonably uniform. For most simulations, this is the area over which we average Σ to calculate Σ_{gap}.

In a few cases the outer edge of the gap is visibly eccentric (see Section 4.3.2 for more discussion), and the circular annulus we have defined above becomes contaminated with non-gap material and is no longer suitable for measuring Σ_{gap}. In these cases, we keep the circular inner gap edge and the azimuthal excision as defined above, but approximate the outer gap edge with an ellipse having semimajor axis $r_p + \Delta$, and an eccentricity and apsidal orientation estimated by eye from snapshots (for a sampling, jump to Figures 4.10 and 4.11).

Each simulation runs until Σ_{gap} appears to have converged in time; see Figure 4.5 for an example. The time required to reach convergence scales approximately as the viscous timescale r_p^2/ν, shortening somewhat with larger q. Each value of Σ_{gap} that we report in Table 4.1 is averaged over a time interval that starts at time t_{conv} near the end of the simulation, and that lasts for duration Δt. For many models, there is actually no need to time-average because the time variability in Σ_{gap} is fractionally small (less than a few percent). However, some models exhibit greater variability, particularly for q approaching 10^{-2} or $h/r = 0.1$. The fluctuations appear periodic, with periods ranging from 0.5–1 P_p and amplitudes up to order unity. For these more strongly time-variable cases, we also record in Table 4.1 the maximum and minimum values of Σ_{gap} that occur during the averaging interval.

4.3 Results

In Section 4.3.1, we obtain empirical scalings for gap depths at $10^{-4} \leq q \leq 5 \times 10^{-3}$; in Section 4.3.2 we repeat for $5 \times 10^{-3} \leq q \leq 10^{-2}$ and discuss qualitatively the new dynamical phenomena that appear at these highest companion masses; and in Section 4.3.3 we highlight some of the differences between `PEnGUIn` and `ZEUS90`.

4.3.1 Gap Depth Scalings for $10^{-4} \leq q \leq 5 \times 10^{-3}$

Gap depths $\Sigma_{gap}(\Sigma_0)$ are recorded in Table 4.1 and plotted against each of the parameters q, α, and h/r in Figures 4.6, 4.7, and 4.8, respectively. Overall the agreement between the two codes, which utilize completely different algorithms, is remarkably good.

Figure 4.6 attests that for $q \lesssim 5\times10^{-3}$, gap depths scale roughly as q^{-2} — as our analytic scaling (4.3) predicts. For $q \gtrsim 5 \times 10^{-3}$, the curves flatten somewhat (more on the behavior at large q in Section 4.3.2). In Figure 4.7,

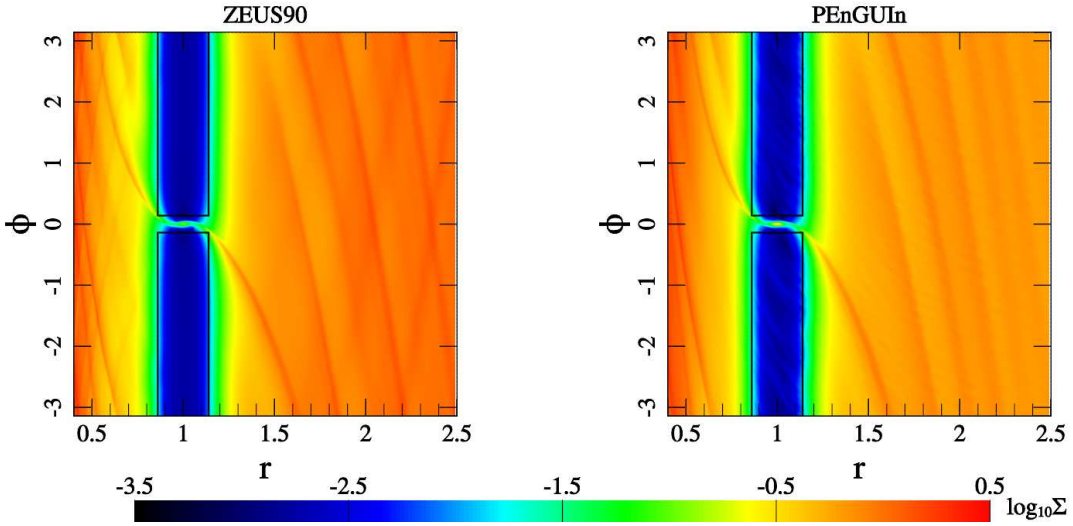

Figure 4.3 Snapshots of simulations with $(q, \alpha, h/r) = (10^{-3}, 10^{-3}, 0.05)$. `PEnGUIn`'s snapshot is taken at $t = 2 \times 10^4 P_{\rm p}$ while `ZEUS90`'s is taken at $t = 1 \times 10^4 P_{\rm p}$. Overall the two codes agree well on the shape and depth of the gap. `ZEUS90` has more trouble converging to the desired outer boundary condition; Σ at $r = 2.5$ deviates from that imposed by equation (4.9) by up to ~50%. Note that `PEnGUIn` does not have the problem in the outer disk that `ZEUS90` does, and moreover succeeds in resolving fine streamers ("filaments") within the gap. The black rectangles indicate the area over which $\Sigma_{\rm gap}$ is averaged.

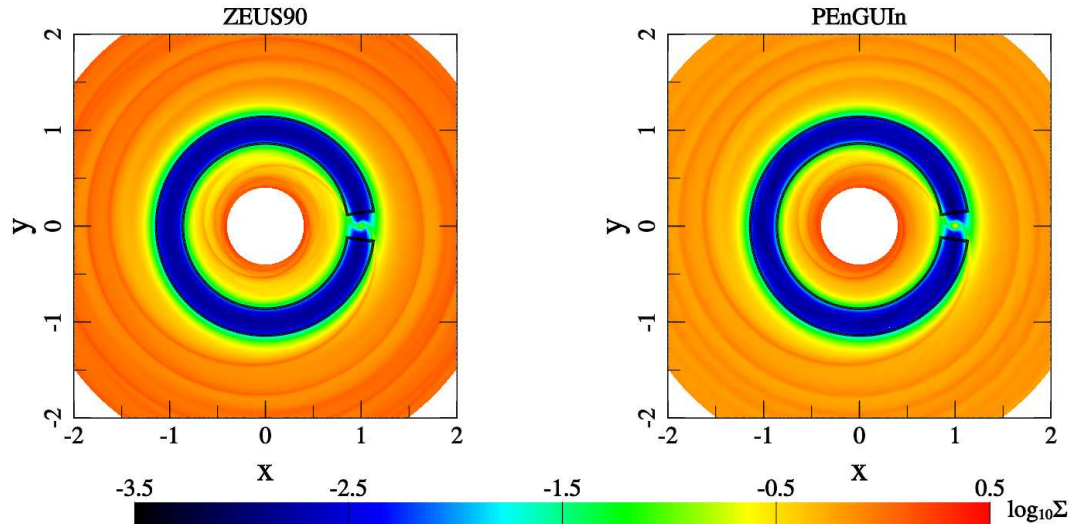

Figure 4.4 Cartesian version of Figure 4.3.

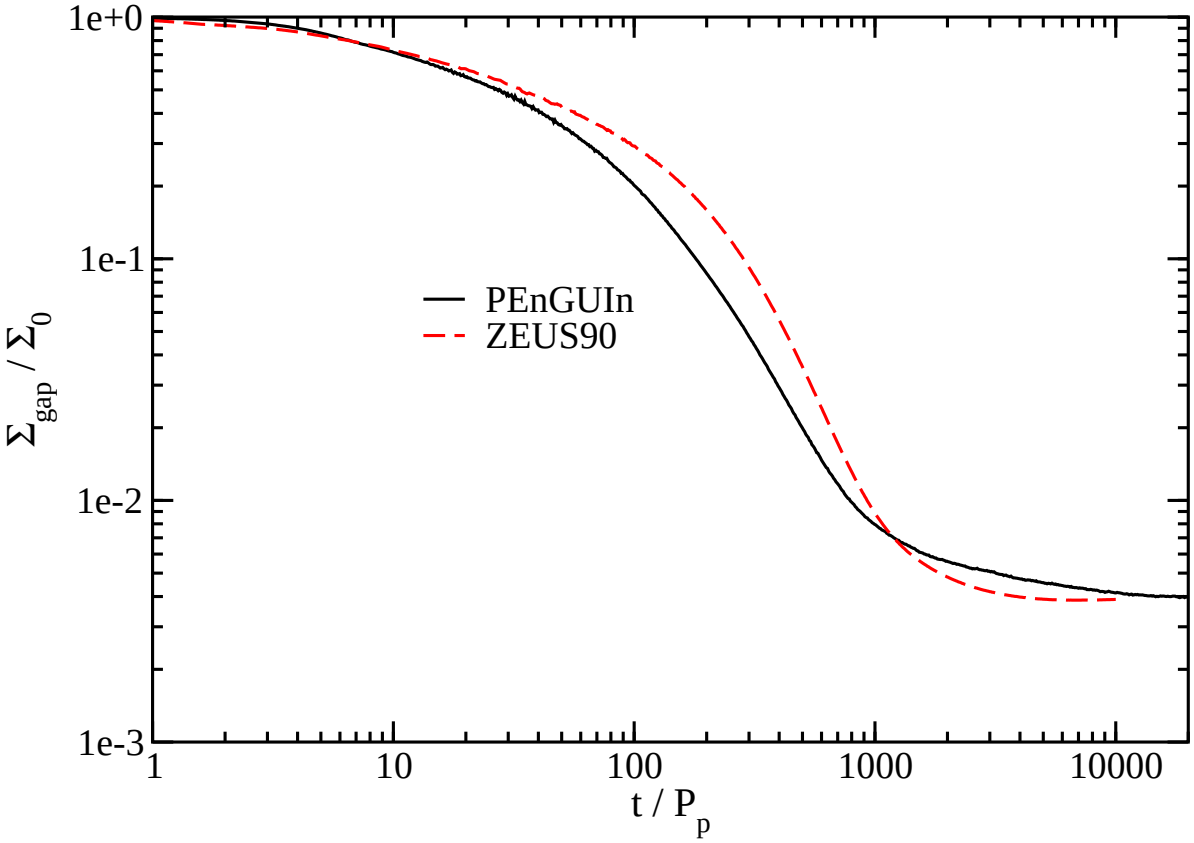

Figure 4.5 Convergence of $\Sigma_{\rm gap}$ with time for $(q, \alpha, h/r) = (10^{-3}, 10^{-3}, 0.05)$. For these parameters, the viscous timescale is formally $r_{\rm p}^2/\nu \sim 6 \times 10^4$ planetary orbits.

Table 4.1. Simulated Gap Depths

			PEnGUIn	ZEUS90	PEnGUIn		ZEUS90		PEnGUIn	ZEUS90
q	α	h/r	$\Sigma_{\rm gap}$[a](Σ_0)		$t_{\rm conv}$[b]($P_{\rm p}$)	Δt[b]($P_{\rm p}$)	$t_{\rm conv}$ ($P_{\rm p}$)	Δt ($P_{\rm p}$)	Comments	
1×10^{-4}	10^{-3}	0.05	4.6×10^{-1}	4.9×10^{-1}	20000	10	10000	10		d
2×10^{-4}	10^{-3}	0.05	1.9×10^{-1}	2.0×10^{-1}	20000	10	10000	10		d
5×10^{-4}	10^{-3}	0.05	3.0×10^{-2}	3.1×10^{-2}	20000	10	10000	10		d
1×10^{-3}	10^{-3}	0.05	4.0×10^{-3}	3.9×10^{-3}	20000	10	10000	10		
2×10^{-3}	10^{-3}	0.05	$8.5^{+0.4}_{-0.3}\times10^{-4}$	$1.2^{+0.1}_{-0.2}\times10^{-3}$	6000	10	6000	10		e
5×10^{-3}	10^{-3}	0.05	$2.1^{+0.2}_{-0.2}\times10^{-4}$	$2.6^{+0.5}_{-0.8}\times10^{-4}$	6000	10	6000	10	f	e
1×10^{-2}	10^{-3}	0.05	$1.7^{+0.4}_{-0.5}\times10^{-4}$	$1.4^{+0.7}_{-0.5}\times10^{-4}$	6000	10	6000	10	f	
1×10^{-3}	10^{-2}	0.05	1.4×10^{-1}	1.5×10^{-1}	2000	10	2000	10		
2×10^{-3}	10^{-2}	0.05	2.7×10^{-2}	2.8×10^{-2}	2000	10	2000	10		
5×10^{-3}	10^{-2}	0.05	$5.5^{+0.9}_{-1.0}\times10^{-3}$	$4.6^{+2.6}_{-1.8}\times10^{-3}$	2000	10	2000	10		
1×10^{-2}	10^{-2}	0.05	$2.0^{+0.6}_{-0.8}\times10^{-3}$	$2.7^{+0.6}_{-0.4}\times10^{-3}$	2000	10	2000	10	e	
1×10^{-3}	10^{-1}	0.05	6.1×10^{-1}	7.2×10^{-1}	300	10	300	10		
2×10^{-3}	10^{-1}	0.05	3.5×10^{-1}	4.9×10^{-1}	300	10	300	10		
5×10^{-3}	10^{-1}	0.05	9.8×10^{-2}	1.7×10^{-1}	300	10	300	10		
1×10^{-2}	10^{-1}	0.05	3.8×10^{-2}	$4.9^{+0.2}_{-0.2}\times10^{-2}$	300	10	300	10		
1×10^{-3}	10^{-3}	0.1	2.2×10^{-1}	$2.4^{+0.1}_{-0.1}\times10^{-1}$	7000	10	7000	10		
5×10^{-3}	10^{-3}	0.1	$1.5^{+0.5}_{-0.1}\times10^{-2}$	$1.6^{+0.1}_{-0.1}\times10^{-2}$	6000	1000	7000	10	c	
1×10^{-2}	10^{-3}	0.1	$8.4^{+0.7}_{-0.6}\times10^{-3}$	$1.0^{+0.1}_{-0.1}\times10^{-2}$	6900	100	7000	10	c	
5×10^{-4}	10^{-3}	0.04	6.0×10^{-3}	6.5×10^{-3}	17000	10	10000	10		d
2×10^{-3}	10^{-3}	0.04	$2.7^{+0.2}_{-0.2}\times10^{-4}$	$2.9^{+0.6}_{-0.6}\times10^{-4}$	10000	10	5000	10		e
2×10^{-3}	10^{-2}	0.04	$7.9^{+1.1}_{-0.9}\times10^{-3}$	$7.0^{+0.4}_{-0.5}\times10^{-3}$	4000	10	4000	10		

[a]Averaged over time and over a partial annulus centered on the planet, as defined in Section 4.2.2 and delineated in Figures 4.3 and 4.4. For visibly eccentric outer disks (see Comments column), the outer edge of the measurement annulus is made eccentric to conform to the gap shape (e.g., Figures 4.10 and 4.11). Maximum and minimum values of $\Sigma_{\rm gap}$ are given for runs for which these values deviate from the time-averaged value by more than a percent. All surface densities are in units where $\Sigma(r = 1) = 1$ in a steadily accreting, planet-less disk.

[b]The time $t_{\rm conv}$ is taken near the end of a simulation, when $\Sigma_{\rm gap}$ appears to have nearly converged to its steady-state value (see Figure 4.5). Formally $t_{\rm conv}$ marks the beginning of the time interval, of duration Δt, over which $\Sigma_{\rm gap}$ is averaged. All times listed are in units of the planetary orbital period, $P_{\rm p}$.

[c]Highly unsteady outer gap edge (e.g., Figure 4.9) and long-term time variability. A longer Δt is chosen to capture the variability.

[d]Indirect potential included (Section 4.2.1).

[e]Eccentric outer disk (e.g., Figures 4.10 and 4.11). Measured eccentricities are ~0.10–0.15 and apsidal precession periods are ~300–600P_p.

[f]An eccentric outer disk is observed at $t \sim 1000P_{\rm p}$, but the eccentricity damps away by $t_{\rm conv}$. The damping is probably artificial (Section 4.3.2).

$\Sigma_{\rm gap}$ appears to scale with α to a power between 1 and 1.5. For comparison, our analytic scaling (4.3) predicts $\Sigma_{\rm gap} \propto \alpha^1$. The empirical dependence on h/r is similarly steeper than the analytic dependence: equation (4.3) predicts that $\Sigma_{\rm gap} \propto (h/r)^5$ whereas Figure 4.8 shows that the power-law indices vary between 5 and 7.

We obtain a "best-fit" power-law relation by minimizing the function $y = \sum[\ln Dq^A\alpha^B(h/r)^C - \ln\Sigma_{\rm gap}]^2/N$ over the parameters (A, B, C, D). The sum is performed over N data points, excluding the discrepant runs at $q = 0.01$ and runs for which $\Sigma_{\rm gap}/\Sigma_0 > 0.2$ (i.e., runs for which gaps hardly open). With these exclusions, there are $N = 13$ data points from `PEnGUIn`, best fitted by

$$10^{-4} \le q \le 5\times10^{-3} \ :$$
$$\Sigma_{\rm gap}/\Sigma_0 = 0.14\left(\frac{q}{10^{-3}}\right)^{-2.16}\left(\frac{\alpha}{10^{-2}}\right)^{1.41}\left(\frac{h/r}{0.05}\right)^{6.61} . \tag{4.12}$$

The $N = 13$ points from `ZEUS90` are best described by a very similar formula:

$$10^{-4} \le q \le 5\times10^{-3} \ :$$
$$\Sigma_{\rm gap}/\Sigma_0 = 0.15\left(\frac{q}{10^{-3}}\right)^{-2.12}\left(\frac{\alpha}{10^{-2}}\right)^{1.42}\left(\frac{h/r}{0.05}\right)^{6.45} . \tag{4.13}$$

Both of these relations fit their respective data to typically better than 20%; the largest deviation in the `PEnGUIn` fit is 40%, corresponding to $(q, \alpha, h/r) = (2\times10^{-3}, 10^{-3}, 0.05)$, and for `ZEUS90` the largest deviation is 50%, corresponding to $(q, \alpha, h/r) = (10^{-3}, 10^{-3}, 0.05)$.

Because our two codes agree so well, and because the fits are good, we are confident the deviations between our empirical scaling (say equation 4.12 from `PEnGUIn`) and our analytic scaling (4.3) are real and reflect physical effects not captured by our analytic scaling. And because our analytic scaling (4.3) matches exactly the scalings found numerically by Duffell & MacFadyen (2013) at low $q \lesssim 10^{-4}$ — whereas our empirical relation (4.12) applies to high $q \gtrsim 10^{-4}$ — these physical effects manifest for giant (Jupiter-like) planets, not lower-mass (Neptune-like) planets.[4] We have not elucidated what this physics is, although there might be some clues from the gap behavior at the very highest values of q we tested, as discussed in Section 4.3.2.

At the same time, we emphasize that the deviations between (4.3) and (4.12), though (probably) real, are not large. If we insist on fitting the `PEnGUIn` data using (4.3) — i.e., if we fix $(A, B, C) = (-2, 1, 5)$ and allow only the coefficient D to float — then the data deviate from (4.3) by typically a factor of 2, and at most a factor of 3. Thus the physical effects not captured by our analytic scaling, whatever they are, do not lead to order-of-magnitude changes in gap depth, at least over the range of parameters tested.

[4] Another difference between our simulations and theirs is that we mandate a steady $\dot{M} \neq 0$ across our entire domain, whereas they adopt (as appears customary for work in this field) wave-killing zones that effectively result in nearly zero-inflow boundary conditions. We have verified, however, that this difference does not matter for $\Sigma_{\rm gap}$; we implemented zero-inflow boundaries in a few runs with `PEnGUIn` and found results for $\Sigma_{\rm gap}$ that matched those with our standard accreting boundaries to better than 1%.

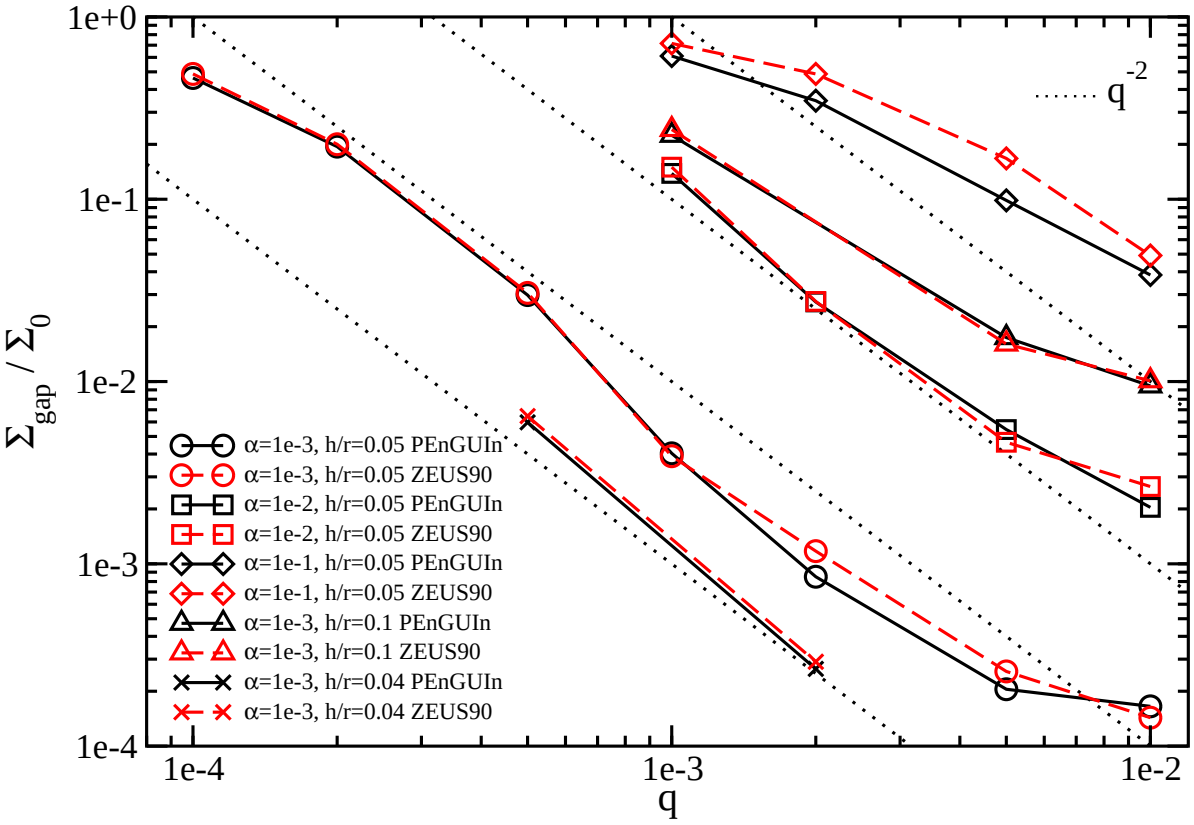

Figure 4.6 $\Sigma_{\rm gap}$ vs. q. Black dotted lines indicate constant power-law slopes of -2, and are shown for reference only. The power-law slopes approximately equal -2 for $q < 5 \times 10^{-3}$, and flatten to -1 for higher q. For formal power-law fits, see the main text.

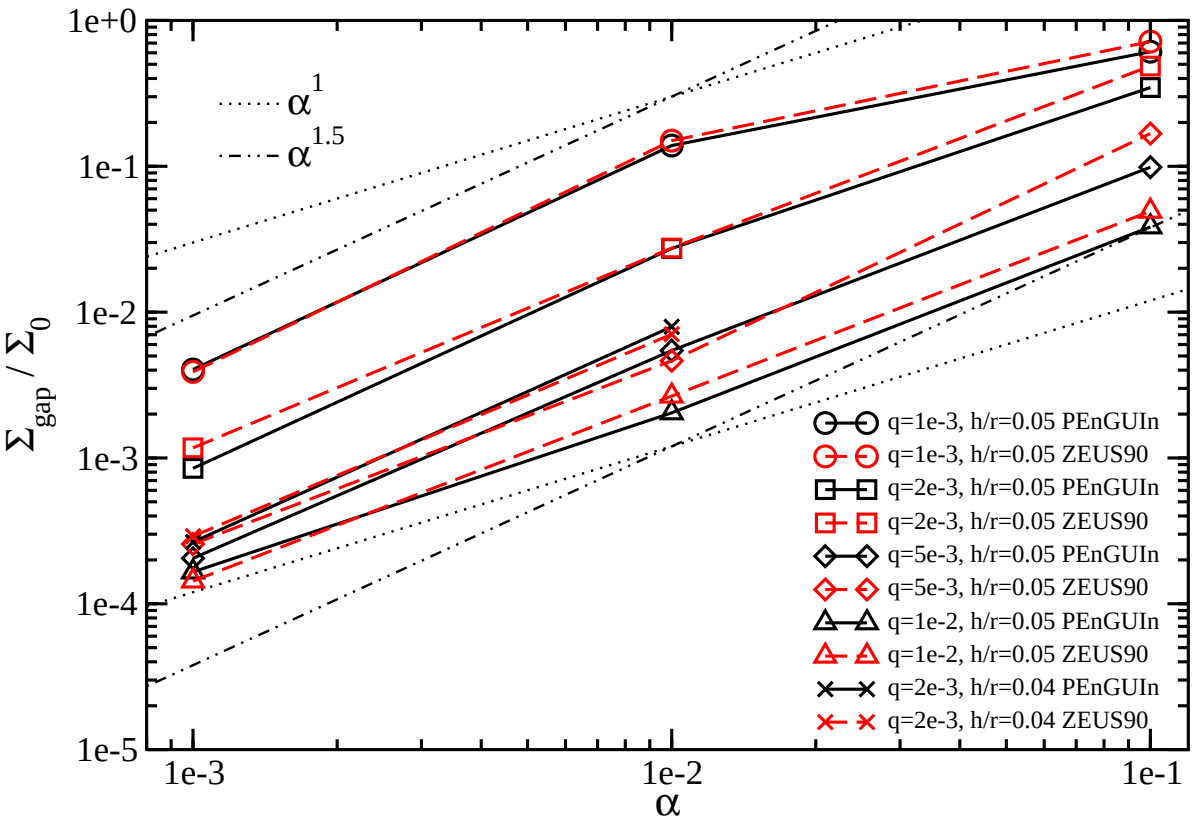

Figure 4.7 Σ_{gap} vs. α. Dotted and dot-dot-dashed lines indicate power-law slopes of 1 and 1.5, bracketing the range exhibited by the data. For formal power-law fits, see main text.

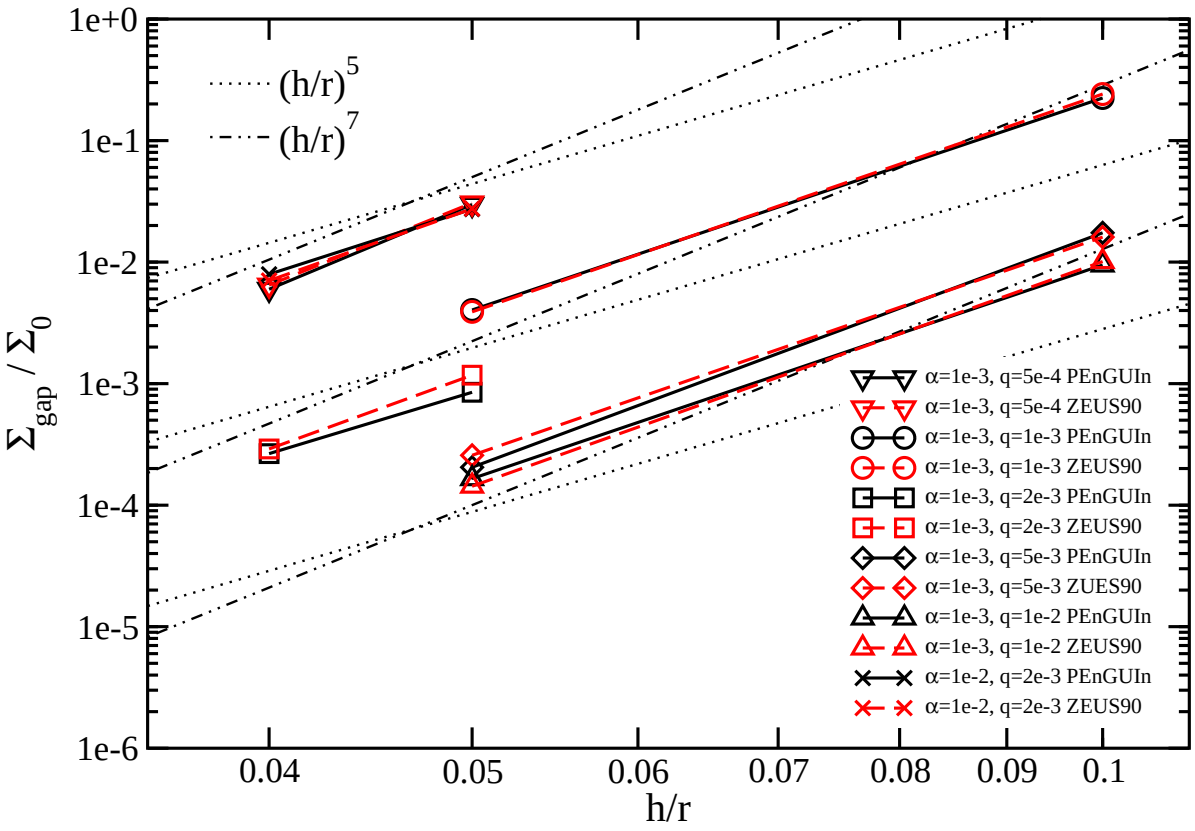

Figure 4.8 Σ_{gap} vs. h/r. Dotted and dot-dot-dashed lines indicate power-law slopes of 5 and 7, bracketing the range exhibited by the data. For formal power-law fits, see main text.

4.3.2 Behavior of Gaps at High $q \gtrsim 5 \times 10^{-3}$

The dependence of $\Sigma_{\rm gap}$ on q flattens at the highest values of q considered (Figure 4.6). Fitting the $N = 8$ points from PEnGUIn for which $q \geq 5 \times 10^{-3}$ and $\Sigma_{\rm gap}/\Sigma_0 < 0.2$ yields:

$$10^{-2} \geq q \geq 5 \times 10^{-3} :$$
$$\Sigma_{\rm gap}/\Sigma_0 = 4.7 \times 10^{-3} \left(\frac{q}{5 \times 10^{-3}}\right)^{-1.00} \left(\frac{\alpha}{10^{-2}}\right)^{1.26} \left(\frac{h/r}{0.05}\right)^{6.12} . \quad (4.14)$$

Similarly for the $N = 8$ points from ZEUS90 we obtain:

$$10^{-2} \geq q \geq 5 \times 10^{-3} :$$
$$\Sigma_{\rm gap}/\Sigma_0 = 5.6 \times 10^{-3} \left(\frac{q}{5 \times 10^{-3}}\right)^{-1.02} \left(\frac{\alpha}{10^{-2}}\right)^{1.34} \left(\frac{h/r}{0.05}\right)^{6.12} . \quad (4.15)$$

Although at first glance one might attribute the flattening of the trend of $\Sigma_{\rm gap}$ with q to the onset of strong shocks, we do not believe this is the correct interpretation. In the strong-shock regime, where disturbances excited by the planet are non-linear at launch, the torque exerted on the disk by the planet scales as $q^1(h/r)^0$ (e.g., Hopkins & Quataert 2011, their section 2.3).[5] Then the same arguments in Section 4.1.1 yield $\Sigma_{\rm gap} \propto q^{-1}(h/r)^2$. This analytic relation reproduces the scaling index for q given by our empirical relations (4.14) and (4.15), but fails to reproduce the empirical scaling index for h/r. Furthermore, the flattening begins at an apparently "universal" q-value of $\sim 5 \times 10^{-3}$ that is independent of h/r, whereas in the strong-shock interpretation, the critical q-value should scale as $(h/r)^3$ (i.e., the expected critical q is given by the so-called thermal mass).

We do not have an explanation for the flatter slope of -1 at high q. We speculate that it might be caused by the most massive companions causing material at the gap edge to "leak" into the gap. The most massive planets disturb the gap edge so strongly that local instabilities send streamers of gas into the gap. These streamers, which de Val-Borro et al. (2006) called "filaments", are prominent in the high-q snapshots in Figure 4.9 (and can be seen even at $q = 10^{-3}$ in the PEnGUIn snapshot in Figure 4.3). The filaments appear to originate from unsteady structures along gap edges; similar structures were seen by, e.g., Kley & Dirksen (2006, their figures 1 and 7).

We also observe evidence for an eccentric outer disk at high q; see Figure 4.10 and the "Comments" column in Table 4.1. The outward transport of angular momentum by waves launched at the 1:3 outer eccentric Lindblad resonance pumps the eccentricity of the outer disk (Lubow, 1991; Papaloizou et al., 2001). The q-value for which disks become eccentric depends on α, h/r, and disk mass (Lubow, 1991; Papaloizou et al., 2001; Kley & Dirksen, 2006; D'Angelo et al., 2006; Dunhill et al., 2013). For a planet held on a fixed circular orbit embedded in a non-gravitating disk for which $h/r = 0.05$ and $\alpha \approx 0.005$, Kley & Dirksen (2006) found that $q \gtrsim 0.003$ led to eccentric disks; for $\alpha \approx 0.01$, the required $q \gtrsim 0.005$. Their findings are in line with ours.

In two runs with PEnGUIn, an eccentricity appears in the outer disk at early times but damps away by the time $\Sigma_{\rm gap}$ converges (see Table 4.1). The eccentricity damping is probably an artifact of our outer circular boundary at $r_{\rm out} = 2.5 r_{\rm p}$, which according to Kley & Dirksen (2006) is too close to the planet to properly simulate eccentric disks. The danger posed by the outer boundary should lessen as the α-viscosity increases and disturbances excited by the planet are more localized; this may be why circularization occurs only for our lowest $\alpha = 10^{-3}$ runs at high q.

We note that the fine-structure filaments threading the gaps are seen in most of our $q \geq 0.001$ simulations,

[5] This scaling can be seen by replacing h with $R_{\rm H}$ in equation (4.1); the torque cut-off distance generally equals $\max(R_{\rm H}, h)$, which in the strong-shock limit equals $R_{\rm H}$.

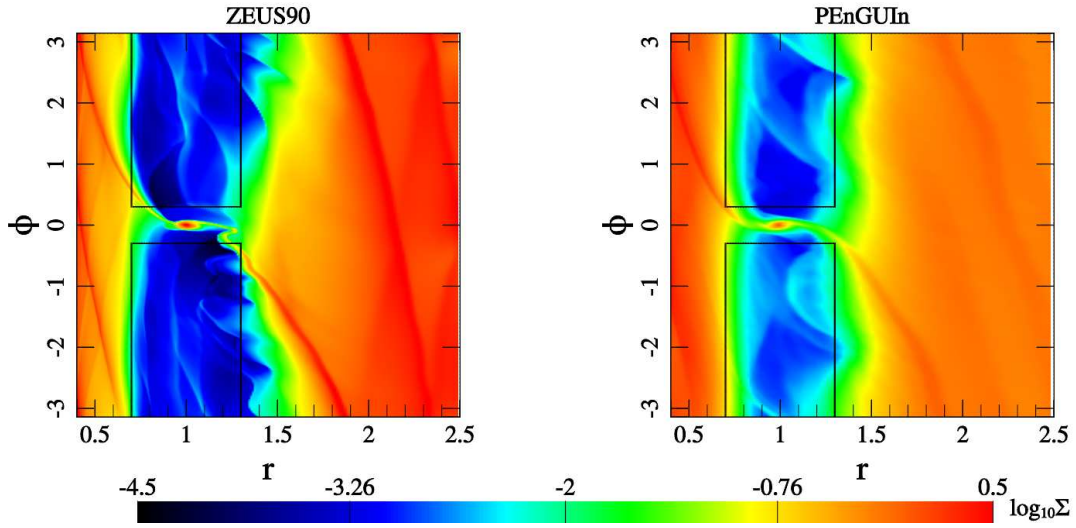

Figure 4.9 Two different examples at high q of unsteady gap edges and streamers filling gaps. The ZEUS90 snapshot is for $(q, \alpha, h/r) = (0.01, 0.01, 0.05)$ and the PEnGUIn snapshot is for $(q, \alpha, h/r) = (0.01, 0.001, 0.1)$.

while only a few of these cases evince eccentric outer disks. Moreover, the filaments observed by de Val-Borro et al. (2006) do not appear associated with disk eccentricity. It is unclear to us whether the filaments and disk eccentricity are directly related.

4.3.3 Code Comparison

Generally PEnGUIn and ZEUS90 agree very well on $\Sigma_{\rm gap}$ (e.g., Figures 4.6–4.8), typically differing by no more than a few tens of percent, and often much better. Because the codes rely on fundamentally different algorithms — one is a shock-capturing Lagrangian-remap code, while the other is an Eulerian code using the upwind method — their agreement lends confidence in the accuracy of our results. Some minor, systematic differences include: (i) $\Sigma_{\rm gap}$ for ZEUS90 is larger than for PEnGUIn; (ii) PEnGUIn usually resolves a higher density peak near the planet; (iii) PEnGUIn typically requires a longer time to converge; and (iv) near the outer boundary where the resolution is coarser, ZEUS90 has difficulty relaxing to the steadily accreting solution described by equations (4.9)–(4.11), deviating from the correct Σ by up to 50%. All of these differences may be attributed to the fact that PEnGUIn uses PPM, which is a fourth-order method for uniform grids (third-order for non-uniform grids), whereas ZEUS90's algorithm is only second-order in space. Thus PEnGUIn tends to be more accurate and less numerically diffusive than ZEUS90, at the cost of taking a longer time to resolve sharp features.

A key innovation of PEnGUIn is its use of GPU technology to accelerate computations. PEnGUIn's speed on a single GTX-Titan graphics card can rival that of a traditional CPU cluster having ~100 cores. For this work we ran PEnGUIn on a desktop computer housing 3 graphics cards. With specialized hardware we can connect up to 8 GPUs to a motherboard. PEnGUIn's scalability with the number of cards approaches linear as the resolution increases; our speed on 3 cards is 2.33 × that of a single card at our standard resolution; if the resolution is doubled (quadrupled), the speed enhancement factor increases to 2.64 (2.92) as PEnGUIn takes more full advantage of GPU's parallelism (e.g. see Figure 2.3 in Section 2.4). Currently PEnGUIn can run on a single node only, and would need to be modified to run on multiple nodes. In a multi-node cluster of GPUs, the scaling of speed with the

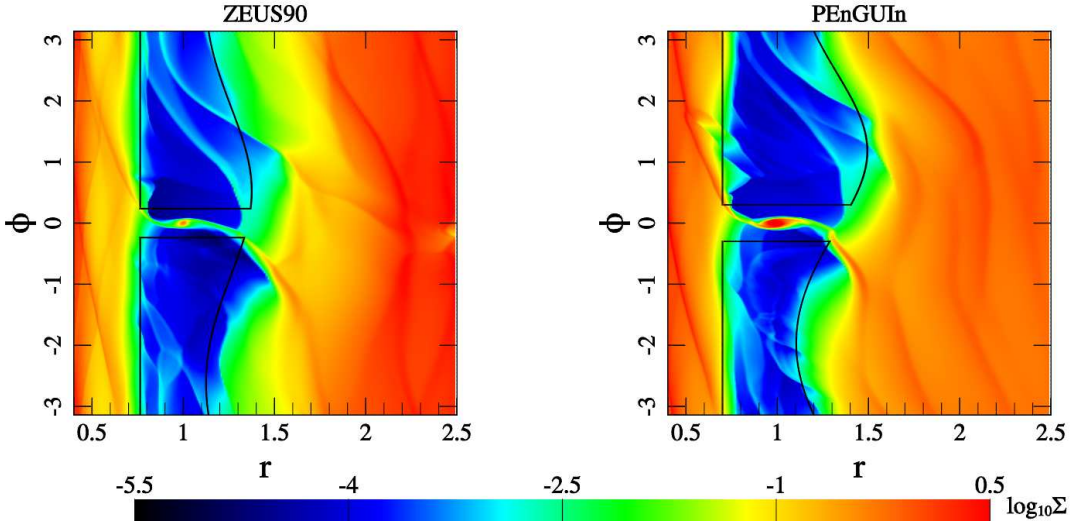

Figure 4.10 Snapshots of eccentric outer disks, one from `ZEUS90` at $(q, \alpha, h/r) = (0.005, 0.001, 0.05)$, and another from `PEnGUIn` at $(q, \alpha, h/r) = (0.01, 0.01, 0.05)$. For the `ZEUS90` run shown, the inner edge of the outer disk (exterior to the planet's orbit) has eccentricity 0.10 and precession period $630P_{\rm p}$. For the `PEnGUIn` run, the eccentricity is 0.15 and the precession period is $380P_{\rm p}$. Black curves enclose the area over which $\Sigma_{\rm gap}$ is computed.

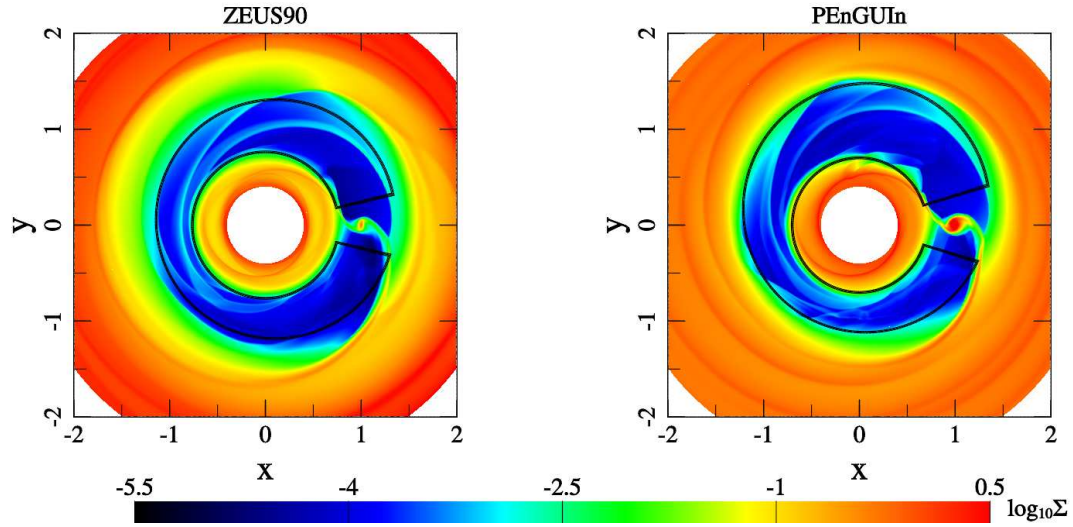

Figure 4.11 Cartesian view of the eccentric disks of Figure 4.10.

number of cards per node is unlikely to be linear because communication between nodes is significantly slower than between cards on a single node.

4.4 Conclusions and Discussions

We established two empirical formulas (4.12 and 4.14) for the surface density contrast, $\Sigma_{\rm gap}/\Sigma_0$, inside and outside the gap carved by a non-accreting giant planet. The first is valid for planet-to-star mass ratios $10^{-4} \le q \le 5\times10^{-3}$, and the second is valid for $5 \times 10^{-3} \le q \le 10^{-2}$. Our formulae are derived from our new, fast, Lagrangian shock-capturing PPM code `PEnGUIn`, and are confirmed by `ZEUS90`. Combining our results with those from the literature, we find that $\Sigma_{\rm gap}$ scales with q, viscosity parameter α, and disk aspect ratio h/r in the following ways:

- At Neptune-like (and perhaps lower) masses, Duffell & MacFadyen (2013) found[6] that $\Sigma_{\rm gap} \propto q^{-2}\alpha^{1}(h/r)^{5}$;
- At Jupiter-like masses, we find that $\Sigma_{\rm gap} \propto q^{-2.2}\alpha^{1.4}(h/r)^{6.6}$ (our equation 4.12);
- At masses near the brown dwarf threshold, we find that $\Sigma_{\rm gap} \propto q^{-1}\alpha^{1.3}(h/r)^{6.1}$ (our equation 4.14).

Our scaling indices for giant planets and quasi-brown dwarfs are supported by two independent codes using different algorithms, and so we are confident in their accuracy. Note that our simulations and those of Duffell & MacFadyen (2013) do share one common set of parameters: $(q, \alpha, h/r) \approx (5\times10^{-4}, 10^{-3}, 0.05)$, for which we find $\Sigma_{\rm gap}/\Sigma_0 = 0.03$ and they find $\Sigma_{\rm gap}/\Sigma_0 = 0.04$ (their Figure 2).[7] We consider this good agreement.

The scaling differences between low mass and high mass, although pointing to real physical effects, do not lead to order-of-magnitude changes in gap depth, at least over the range of parameters surveyed. That is, using the Neptune-like scaling $\Sigma_{\rm gap} \propto q^{-2}\alpha^{1}(h/r)^{5}$ for Jupiter-like planets leads to gap depths that differ (systematically) from those observed in our simulations by factors of only 2–3.

4.4.1 Connecting to Observations of Transition Disks

Are the gaps empty enough to reproduce the low optical depths characterizing the cavities of transitional and pre-transitional disks? Surface density contrasts from models of disks like PDS 70 (Dong et al., 2012) and GM Aur (Calvet et al., 2005) are 10^3 or more. We have found that certain sets of planet-disk parameters can achieve such contrasts. For example, $(q, \alpha, h/r) = (5 \times 10^{-3}, 10^{-3}, 0.05)$ produces contrasts of ~5000 (Table 4.1). Lower mass planets could also be made to work with lower α-viscosities and/or cooler disks with lower h/r; as a further example, $(q, \alpha, h/r) = (2 \times 10^{-3}, 10^{-3}, 0.04)$ generates a contrast of ~3000. The dependence on disk temperature is especially sensitive: $\Sigma_{\rm gap} \propto (h/r)^{6.6} \propto T^{3.3}$.

The surface density contrasts reported in this chapter are all underestimates insofar as we have neglected accretion onto the planet; but arguably the disk accretion rate cannot be reduced by more than a factor of order unity, lest the planet starve the host star and violate observed stellar accretion rates (Zhu et al. 2011; see also Lubow & D'Angelo 2006). We would argue further that our gas surface density contrasts are also underestimates of dust surface density (i.e., optical depth) contrasts, to the extent that mechanisms like dust filtration at outer gap edges (e.g., Zhu et al. 2012) deplete dust relative to gas in gaps.

[6] A caveat is that Duffell & MacFadyen (2013) did not explicitly test the dependence on h/r that they proposed. We used `PEnGUIn` to try to reproduce their low-mass results (data not shown), but encountered the problem that $\Sigma_{\rm gap}$ took too long to converge. What low-mass data we did collect at the end of 20000 orbital periods were consistent with the power-law indices proposed by Duffell & MacFadyen (2013).

[7] Technically, our simulations have spatially constant α and h/r, whereas theirs has spatially constant $\nu = \alpha c_s h$ and $h/r \propto r^{0.25}$. Also, we compute $\Sigma_{\rm gap}$ as an area average, whereas they report the minimum surface density. These differences are probably immaterial.

Given these findings, we feel that when it comes to transition disks, the problem is not so much gap depth, but gap width. A single planet embedded in an accreting disk generates a gap too narrow in radial width to explain the expansive cavities observed in transition disks. To connect to observations would seem to require that we expand our study to include multiple planets or brown dwarfs within a viscous, gravitating disk, as has been done by Zhu et al. (2011). These authors discounted $\alpha \lesssim 0.002$ — and were therefore compelled to invoke additional channels of opacity reduction (e.g., grain growth) to explain transition disks — because multiple planets were found to be dynamically unstable at low α / high Σ (see their page 8). The incompatibility of multiple giant planets with low α is a result that we would like to see confirmed independently and further explored.

4.4.2 A Floor on $\Sigma_{\rm gap}$

All our empirical scalings for $\Sigma_{\rm gap}$ suggest that arbitrarily low values of α generate arbitrarily clean gaps. We expect, however, the scalings to break down for small enough α. Without an intrinsic disk viscosity, a planet may stir the gap edge in such a way as to trigger local instabilities and turbulent diffusion. For example, with $\alpha = 0$, the outer gap edge might be so sharp as to be Rayleigh unstable. There should therefore be a minimum value or "floor" on $\Sigma_{\rm gap}$ caused by a minimum planet-driven viscosity. Just such a floor has been reported by Duffell & MacFadyen (2013); see their Figure 7. Other simulations of planets in inviscid disks, concentrating on orbital migration and not gap depth, have been carried out by Li et al. (2009) and Yu et al. (2010).

Similar arguments suggest that there might also be a floor on $\Sigma_{\rm gap}$ at high q. The streamers/filaments that invade the gap are densest for the highest q-values we tested.

4.4.3 Analytic Derivation

In Section 4.1.1, we presented an analytic derivation of gap depth $\Sigma_{\rm gap}$. We discovered that the power-law scalings in our analytic relation (4.3) match precisely those reported from numerical experiments for low-mass planets by Duffell & MacFadyen (2013). Our analytic relation can even reproduce approximately (deviating systematically by factors of a few) the empirical results we found for gaps carved by Jupiter-like planets.

The success of our breezy analytic derivation for $\Sigma_{\rm gap}$ is surprising. Our derivation is "zero-dimensional" ("0D") because it considers only the total rates of angular momentum transport – a.k.a. the total "angular momentum luminosity" or "angular momentum current" — integrated over all azimuth and radius; it ignores the complicated details of how the torques are actually applied differentially in space. We have already noted in Section 4.1.1 how it is not completely obvious that Σ_0 characterizes the viscous torques in gap edges. Furthermore, our 0D treatment considers only the wave and viscous contributions to the total angular momentum current, and neglects the contribution from advection (i.e., the transport of angular momentum associated with a non-zero radial velocity v_r) and the contribution from azimuthal pressure variations (Crida et al., 2006).

We can address some of these problems and make the leap from 0D to 1D by considering the azimuthally averaged, radially dependent torque balance equation for a steady-state accretion disk perturbed by a planet (see, e.g., equation 3 of Lubow & D'Angelo 2006):

$$\frac{d(3\nu\Sigma\Omega r^2)}{dr} = -\Sigma\Omega r^2 v_r + 2\Sigma r\Lambda(r) \tag{4.16}$$

where the Lindblad torque per unit mass is

$$\Lambda(r) = \mathrm{sgn}\,(r - r_{\rm p})\frac{fGM_* q^2}{2r}\left(\frac{r}{r - r_{\rm p}}\right)^4 \tag{4.17}$$

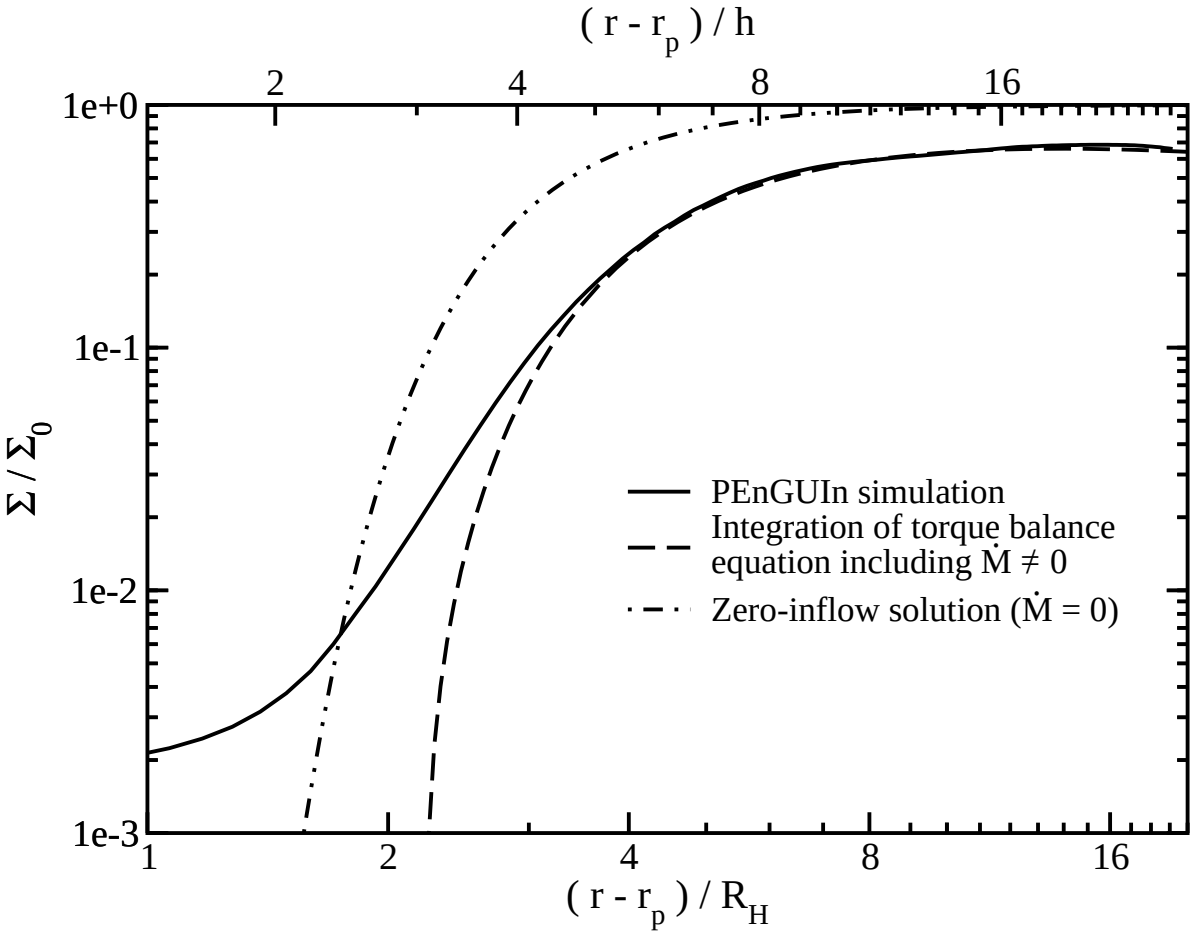

Figure 4.12 Reproducing the simulated gap profile with a 1D analysis. The solid curve is the azimuthally averaged surface density profile outside the planet's orbit for $(q, \alpha, h/r) = (0.001, 0.001, 0.05)$, as calculated from 2D simulations using `PEnGUIn`. Directly integrating the 1D equation (4.18) reproduces well the onset of the gap, if we set $f = 0.2$ (dashed curve). However, the bottom of the gap is not captured at all. Setting $\dot{M} = 0$, as is commonly done in the literature, yields a profile that is similar in shape to the actual profile, but shifted in radius (for the same value of $f = 0.2$; dot-dot-dashed curve).

and f is an order-unity constant. The left-hand side of (4.16) accounts for the viscous torque, while the first term on the right-hand side accounts for angular momentum transport by advection. We define $x \equiv r - r_\mathrm{p}$ and $\dot{M} \equiv 2\pi\Sigma v_r r$ = constant < 0, and approximate $\Omega = \Omega_\mathrm{p}(r/r_\mathrm{p})^{-3/2}$ so that $\nu\Omega r = \nu_\mathrm{p}\Omega_\mathrm{p} r_\mathrm{p}$ = constant (variables subscripted by p take their values at the planet's orbital radius). For the outer disk, equation (4.16) simplifies to

$$\alpha\left(\frac{h}{r}\right)^2 \frac{d(r\Sigma)}{dr} = \frac{|\dot{M}|\Omega_\mathrm{p} r_\mathrm{p}}{6\pi G M_*}\left(\frac{r}{r_\mathrm{p}}\right)^{-1/2} + \frac{f}{3}\, q^2 \left(\frac{r}{x}\right)^4 \Sigma\,. \tag{4.18}$$

From this equation it becomes apparent how the outer accretion disk responds when repelled outward by a planet: for $q > 0$, the surface density gradient $d\Sigma/dr$ steepens, just enough that the viscous torque can exceed the Lindblad torque and maintain a steady flow of mass inward (i.e., carry the $\dot{M}$ imposed at infinity across the planet's orbit).

Setting $\dot{M} = 0$, and working in the WKB limit where $d/dr \gg 1/r$, gives the standard zero-inflow solution: an exponential profile for $\Sigma(r)$, commonly used in the literature (e.g., Lubow et al. 1999; de Val-Borro et al. 2007; Mulders et al. 2013). Keeping $\dot{M} \neq 0$ alters $\Sigma(r)$: it still resembles an exponential but is shifted outward (for

fixed f), as Figure 4.12 demonstrates. We find for the parameters chosen that equation (4.18) describes well the gap profile from our 2D simulations, down to a distance of ~3–4 Hill radii away from the planet. But inside this cut-off distance, the 1D solution fails critically — it falls much too steeply to recover the actual flat-bottomed gap.

The problem of determining the gap depth analytically in 1D appears tantamount to the problem of understanding what happens inside this cut-off distance. Lindblad torques shut off here; the tidal gravitational field of the planet is especially strong; and circulating streamlines give way to horseshoe orbits. One-dimensional analytic treatments may be inadequate to the task of modeling how gas navigates from the outer disk to the inner disk through a series of "horseshoe turns" (Lubow et al., 1999; Kley & Nelson, 2012). As far as analytic treatments go, it may be that to do better than 0D requires at least 2D.

Chapter 5

3D Flow around Earth-Size Planets

A version of this chapter has been accepted for publication in The Astrophysical Journal as "The 3D Flow Field Around an Embedded Planet", Fung, J., Artymowicz, P., and Wu, Y., 2015. Reproduced by permission of the AAS.

5.1 Introduction

In recent years, the field of planetary science has made great strides with the discovery of thousands of planets and planet candidates by the *Kepler* mission (Borucki et al., 2010). The majority of these planets have sizes of about 1 to 4 Earth-radii (super-Earths), located at about 0.1 AU away from their host stars (e.g. Batalha et al., 2013; Petigura et al., 2013). This wealth of data has enabled us to more thoroughly check the accuracy of our theories against observations.

These *Kepler* planets are orders of magnitude less massive than the gap-opening planets we studied in the previous chapter, and their interactions with the disk are more fundamentally linked to the process of planet formation. Despite much effort, however, there remain some discrepancies between theory and observation that urgently need to be bridged. One example is planet migration. Through gravitational interaction with the circumstellar disk, a planet can gain or lose angular momentum and migrate away or towards its host star. Current planet migration theory predicts that Earth-size planets located at about 0.1 AU away from their host stars would migrate inward and fall onto their host stars within a timescale of only about a few thousand years, while the typical lifetime of a protoplanetary disk is about a few million years. The fast inward migration is also known as type I migration (Section 1.3 and references therein; also see Masset & Casoli 2010; Paardekooper et al. 2010, 2011). Because we do observe many planets at these separations from their host stars, this implies our predicted migration rate is orders of magnitude faster than can be tolerated by observational data.

Another example is the accretion of planetary atmospheres. The study of how planetary cores of a few Earth-masses ($M_\oplus$) accrete gas from their protoplanetary disks is essential for understanding how gaseous planets are formed. Current planet accretion theory often uses the Bondi radius r_B to define the extent of an embedded planet's atmosphere (e.g. Pollack et al., 1996; Ikoma et al., 2000; Rafikov, 2006; Lee et al., 2014).

$$r_\mathrm{B} = q\frac{GM_*}{c_\mathrm{s}^2}\,, \tag{5.1}$$

where q is the mass ratio between the planet and the host star, G is Newton's gravitational constant, M_* is the host

star's mass, and c_s is the sound speed of the disk. Treatment of this form neglects the effects of the background Keplerian shear and the disk's vertical structure. On the other hand, recent results in 3D disk-planet interaction by Ormel et al. (2015b) (hereafter OSK15) found that no gas is bound to the planet over a timescale exceeding tens of planetary orbital periods. If their result applies to Earth-size planetary cores, it might prevent their transformation into gaseous giants like Neptune, since the core's envelope may be prevented from cooling, contracting, and accepting more gas from the surrounding flow.

The shortcomings of current theory may in part be due to our lack of understanding about the 3D flow of disk material near the gravitational influence of a planet. Both planet migration and accretion theories rely on understanding the flow topology, yet our current picture of disk-planet interaction is based on a simplified 2D flow, confined to the midplane of the disk, or even a 1D, spherically symmetric flow in the classical Bondi-Hoyle accretion. Figure 5.1 shows the typical 2D corotating streamlines around a planet located at $\mathbf{r}_p = (a, \phi_p)$. In this picture, there are three separate, well defined flow patterns: (i) the disk flow, which can be approximated as deformed, originally circular orbits around the host star, represented by the yellow and green streamlines; (ii) corotational or horseshoe flow, which contains the horseshoe- and tadpole-shaped orbits of the restricted three-body problem, represented by the red and blue streamlines; and (iii) streamlines bound to the planet, represented by the magenta streamlines. We refer the reader to Ormel (2013) for a more thorough analysis of the 2D flow topology. The following established notions in planet formation theory are tied to each of the three flow patterns listed above: (i) density waves are excited in the disk flow; the angular momentum carried by these waves results in the Lindblad torque (also called wave torque) acting on the planet (Section 1.3.1); (ii) the corotational flow exchanges angular momentum with the planet as it transfers from outside to inside the planet's orbit at $r = a$, which results in the corotation torque (also called co-orbital torque or horseshoe drag) (Section 1.3.2); and (iii) the planet is surrounded by an atmosphere separated from the other regions of the disk.

The 2D picture described above is most applicable to large planets, whose Hill radius (or Roche lobe radius):

$$r_H = a\left(\frac{q}{3}\right)^{\frac{1}{3}}, \tag{5.2}$$

where a is the semi-major axis of the planet's orbit, is larger than the local scale height of the disk, h_0, in order to justify the flat disk assumption in the vicinity of the planet. We have seen in the previous chapter that in this limit, the torques exerted by the planet result in gap opening. For super-Earths, however, this condition is not satisfied. For example, if the planet has $1M_\oplus$, or $q \sim 3 \times 10^{-6}$, then it has $r_H \sim 0.01a$, while h_0 is typically between 0.03 to $0.1a$. As a result, one cannot assume that the flow topology can be adequately described by Figure 5.1. What does the horseshoe flow look like above and below the planet's Hill sphere? How does gas flow around the planet and how does it become accreted? These are the questions we aim to answer in this chapter, specifically for super-Earths.

We use high-resolution global simulations to simultaneously resolve the global flow in the horseshoe region and the local flow within the planet's atmosphere. In previous works on 3D disk-planet interaction, reduction of a varying degree of migration torques has been found. Linear analysis by Tanaka et al. (2002) has given a 3D torque weaker by a factor of about two than in a 2D disk, in line with the earlier analytical estimates of Artymowicz (1993a), while fully nonlinear simulations by D'Angelo et al. (2003) have found that for planets with r_H less than 60% of h_0, the torque can be up to an order of magnitude weaker than 2D predictions; and when D'Angelo & Lubow (2010) studied the fully nonlinear case for a small planet ($\sim 1M_\oplus$), this time they have found good agreement with Tanaka et al. (2002). Overall, it appears that 3D effects are more important for the larger than for the smaller planets, which is counter-intuitive. We will demonstrate in this chapter that these results can

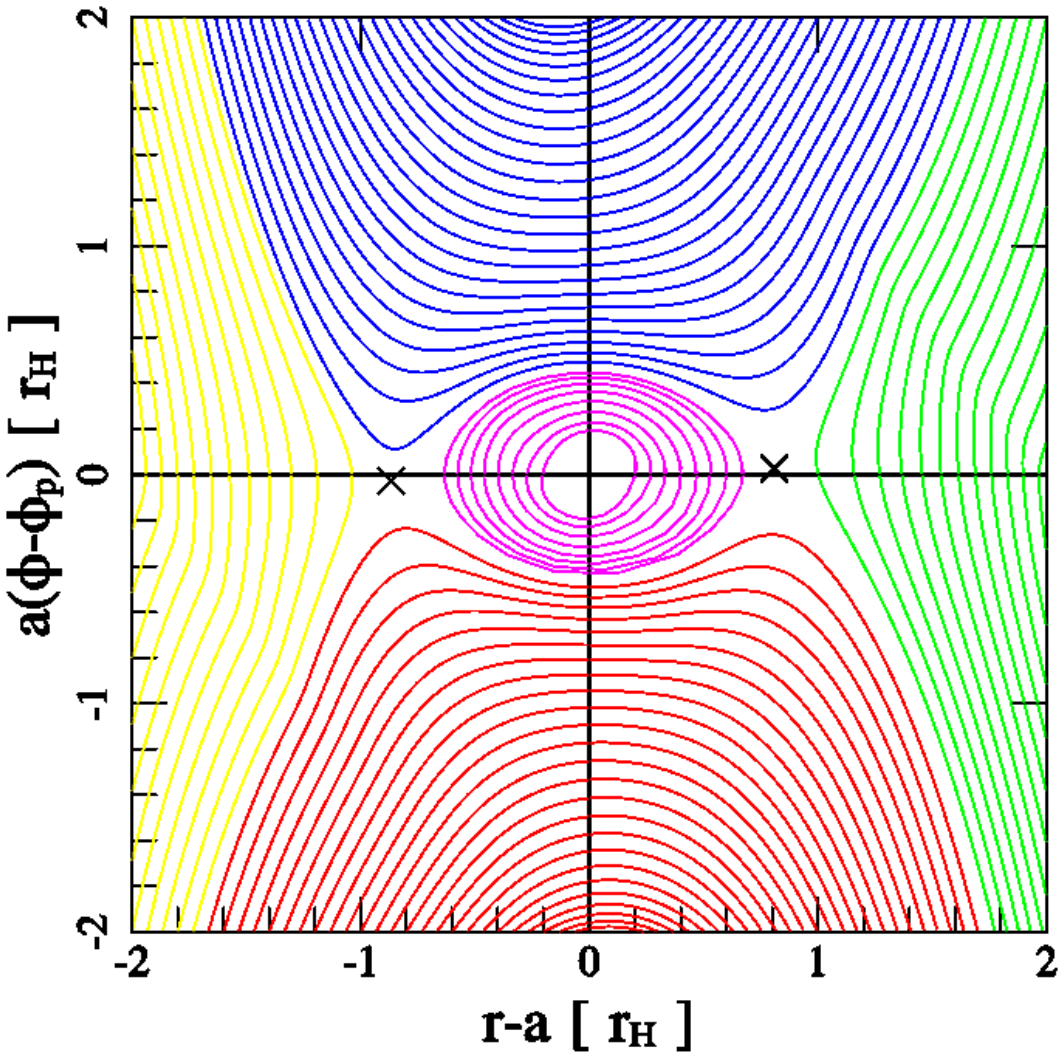

Figure 5.1 Streamlines around a planet in 2D, plotted in the corotating frame of the planet, which is located at the center of this plot. The background Keplerian shear is from bottom to top in the inner disk ($r < a$), and top to bottom in the outer disk ($r > a$). We call the streamlines approaching the planet from the inner disk "inner", and those approaching from the outer disk "outer". The streamlines are color-coded: yellow and green are the inner and outer disk flow; red and blue are the inner and outer horseshoe flow; and magenta is the flow that is bound to the planet. The crosses mark the "stagnation" points, where the velocity is zero. A third stagnation point exists at the location of the planet. This point is irrelevant to our analysis, so we omit to label it. The streamlines here are computed from a 2D simulation using the same setup and resolution as our 3D one (see Section 5.2), but without the vertical dimension, and the planet's potential is not softened.

be tied together and explained, by studying how the flow topology differs from 2D to 3D.

In terms of the accretion of a planet's atmosphere, OSK15's results were obtained from the local 3D simulations of sub-Earth-mass planets' atmospheres, which did not allow them to self-consistently connect their atmospheric flow to the global disk flow. In this work we will not only establish how a planet's atmosphere fits into the co-orbital flow topology, but also probe the interesting regime of planet masses closer to the critical core mass limit.

This chapter is organized as follows. In Section 5.2 we describe our numerical setup. In Section 5.3 we present the flow topology extracted from our simulation. In Section 5.4 we compute the torque on our planet. In Section 5.5 we demonstrate how disk viscosity can affect the flow pattern and consequently the planetary torque. In Section 5.6 we conclude and discuss the implications of our results.

5.2 Setup

To perform global 3D hydrodynamical simulations of disk-planet interaction, we use our GPU-based, Lagrangian, dimensionally-split, shock-capturing hydrodynamics code `PEnGUIn` (Chapter 2). Our simulations are done on a fixed cylindrical grid. It spans the full 2π in the azimuth, $0.7a$ to $1.3a$ in the radial direction, and $0a$ to $0.15a$ in the vertical, where a is the fixed planet-star separation. `PEnGUIn` solves the following Navier-Stokes equations in cylindrical coordinates:

$$\frac{\partial \rho}{\partial t} + (\mathbf{v}\cdot\nabla)\rho = -\rho\,(\nabla\cdot\mathbf{v})\ , \tag{5.3}$$

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{v} = -\frac{1}{\rho}\nabla p - \nabla\Phi + \frac{1}{\rho}\nabla\cdot\mathbb{T}\,, \tag{5.4}$$

where ρ is the density, p the gas pressure, $\mathbf{v}$ the velocity, $\mathbb{T}$ the Newtonian viscous stress tensor which depends on the kinematic viscosity ν, and Φ the gravitational potential of the central star and the planet. Note that unlike previous chapters, we are no longer using vertically averaged variables. In the barycentric frame,

$$\Phi = -\frac{GM_*}{\sqrt{r^2 + r_1^2 + 2rr_1\cos(\phi-\phi_\mathrm{p}) + z^2}} - \frac{qGM_*}{\sqrt{r^2 + r_2^2 - 2rr_2\cos(\phi-\phi_\mathrm{p}) + z^2 + r_\mathrm{s}^2}}\,, \tag{5.5}$$

where $r_1 = qa/(1+q)$ and $r_2 = a/(1+q)$ are the star's and the planet's radial positions, respectively; $\phi_\mathrm{p} - \pi$ and ϕ_p are their angular positions; and r_s is the softening length of the planet's potential. The only difference between this and Equation 4.8 is the inclusion of the vertical dimension z. Recall that M_* is the mass of the star and q is the mass ratio between the planet and the star. We set $GM_*(1+q) = 1$ and $a = 1$ so that the planet's orbital frequency $\Omega_\mathrm{p} = 1$ and period $P_\mathrm{p} = 2\pi$. For convenience, we also denote $v_\mathrm{k} = \sqrt{GM_*(1+q)/r}$ and $\Omega_\mathrm{k} = \sqrt{GM_*(1+q)/r^3}$ as the Keplerian orbital velocity and frequency. We complete our set of equations with an isothermal equation of state: $p = c_\mathrm{s}^2\rho$, where c_s is the constant sound speed of the disk.

5.2.1 Planet and Disk Parameters

We simulate a planet on a fixed circular orbit embedded in a viscous 3D disk. The planet-to-star mass ratio is $q = 1.5\times10^{-5}$, which corresponds to $\sim 5M_\oplus$ for a solar-mass star. We increase the planet mass gradually at the beginning of our simulations, starting from zero to our desired value over $1P_\mathrm{p}$. The relevant length scales in this study are h_0, the scale height of the disk, r_H, the Hill radius within which the gravity of the planet dominates, r_B,

the Bondi radius relevant for accretion, and r_s, the smoothing length. Since c_s is constant, we have $h = c_s/\Omega_k$, which has a radial dependence that goes as $h = h_0(r/a)^{\frac{3}{2}}$, and is equal to $h_0 = 0.03a$ at a. Our planet has $r_H = 0.017a$ (Equation 5.2), and $r_B = qGM_*/c_s^2 = 0.0167a$. Finally, for r_s, unlike in 2D calculations where a non-zero r_s mainly serves as a way to mimic 3D effects, here we include it to avoid singularity. We choose $r_s = 0.1r_H$, or $0.0017a$. Our set of parameters therefore gives us the following hierarchy of length scales: $h > r_H \sim r_B > r_s$.

Another convenient way to quantify the planet's mass is the dimensionless "thermal mass":

$$q_{th} = q\left(\frac{h_0}{a}\right)^{-3}, \tag{5.6}$$

which can also be written as r_B/h_0. The value $q_{th} \approx 1$ marks the division line where a planet becomes sufficiently massive to significantly modify the disk structure, which is sometimes called the "nonlinear" regime. Because we have $q_{th} = 0.56$, which is close to unity, we do expect our planet to have some weak nonlinear effects. The shortest length scale r_s can be interpreted as the physical size of the planet. The Earth-Sun system, for example, would have $r_s \sim 0.04r_H$ if the Earth were placed at 0.1AU. Since r_s is much smaller than both r_H and r_B, we expect it to have little influence on the global flow topology, but has more influence on the density profile of the planet's atmosphere.

The initial density profile of the disk is axisymmetric. It follows a power law in the radial direction, and is set to the hydrostatic solution of an isothermal gas in the vertical direction:

$$\rho(r,z) = \rho_0\left(\frac{r}{a}\right)^{-3} e^{-\frac{z^2}{2h^2}}, \tag{5.7}$$

where we set $\rho_0 = 1$[1]. The surface density, Σ, obtained by integrating Equation 5.7 over z, has the following profile: $\Sigma \propto r^{-\frac{3}{2}}$. This surface density profile is intentionally chosen to test a prediction about corotation torque. It has been shown in 2D that the corotation torque vanishes when there is zero disk vortensity gradient, or $(\nabla\times\mathbf{v})/\Sigma =$ constant (Ward, 1991; Masset & Ogilvie, 2004; Paardekooper et al., 2010), which, for a Keplerian disk, is precisely when $\Sigma \propto r^{-\frac{3}{2}}$. Consequently, if we find a significant corotation torque in our 3D simulation, it will be a new phenomenon not captured by 2D analysis.

The initial velocity field models a steady disk by setting $\partial/\partial t = 0$ in Equation 5.4 and ignoring the potential of the planet: $\mathbf{v} = (v_r, r\Omega, 0)$, where

$$v_r = -3\frac{\nu}{r}\left(\frac{\mathrm{d}\ln\rho}{\mathrm{d}\ln r} + \frac{1}{2}\right), \tag{5.8}$$

$$\Omega = \Omega_k\sqrt{1 + \left(\frac{h}{r}\right)^2\frac{\mathrm{d}\ln\rho}{\mathrm{d}\ln r}}, \tag{5.9}$$

The kinematic viscosity of our disk is set to $\nu = 10^{-6}a^2\Omega_p$, which corresponds to the Shakura-Sunyaev α-viscosity coefficient $\alpha \sim 0.001$ (Shakura & Sunyaev, 1973). This choice determines the viscous diffusion timescale across the horseshoe region. One can estimate this timescale as $t_\nu = w^2/\nu$, where w is the half of the radial width of the horseshoe region. Viscous diffusion can modify the horseshoe flow if t_ν is shorter than the libration timescale of the widest horseshoe orbit, $t_{lib} = (4a/3w)P_p$. If we approximate $w \sim 2r_H$, then our model gives $t_\nu \sim 200P_p$ which is much longer than $t_{lib} \sim 40P_p$. As a result, it is safe to neglect the effects of viscosity in our analysis of the flow topology. In Section 5.3.1 we will give an exact measurement of w, and in Section 5.5 we will consider the scenario where $t_\nu < t_{lib}$.

[1] Since we do not consider the self-gravity of the disk, this normalization has no impact on our results.

5.2.2 Boundary Conditions

Because we cover the full 2π span, our azimuthal boundary condition is periodic. For our radial boundaries, it is important to prevent the reflection of density waves launched by the planet. To achieve this, we impose wave killing zones in $r = \{0.7a, 0.73a\}$ for the inner boundary and $r = \{1.27a, 1.3a\}$ for the outer. Within these zones, we include an artificial damping term:

$$\frac{\partial X}{\partial t} = [X(t = 0) - X(t)] \frac{2c_s|r - r_{\rm kill}|}{(0.03a)^2}, \tag{5.10}$$

where X stands for all disk variables including ρ, p, and $\mathbf{v}$. $r_{\rm kill} = 0.73a$ for the inner boundary, and $1.27a$ for the outer. In the vertical direction, as we only simulate the upper half the of the disk, so we impose a symmetric condition at the disk midplane. At the top we use a reflecting condition to ensure that mass does not leak in or out. In practice, the symmetric and reflecting conditions are equivalent: we copy disk variables across the boundary, and reverse the signs of the z-component velocity and force.

5.2.3 Resolution

Since the focus of this work is to study the co-orbital region, we adopt a nonuniform grid that has enhanced resolution close to the planet. Individual grid size is calculated using the following formula:

$$\Delta x = A(x - x_{\rm p})^2 + \Delta x_{\rm min}, \tag{5.11}$$

where x can be any one of r, ϕ, or z, Δx is the grid size, $x_{\rm p}$ is the location of the planet, and $\Delta x_{\rm min}$ is the desired grid size at the location of the planet. The factor A is found using the following relation:

$$x_{\rm b} = \sqrt{\frac{\Delta x_{\rm min}}{A}} \tan(NA\Delta x_{\rm min}), \tag{5.12}$$

where $x_{\rm b}$ is the distance from the planet's location to the boundary of the grid, which equals to 0.3 in r, π in ϕ, or 0.15 in z . N is the total number of grid cells within $x_{\rm b}$, which equals to 144 in r, 1512 in ϕ, or 72 in z. The entire grid therefore has $288(r) \times 3024(\phi) \times 72(z)$ cells. $\Delta x_{\rm min}$ defines the resolution we desire near the planet, which we choose to be $r_{\rm H}/18 \sim 0.001a$. This also corresponds to about 2 cells per $r_{\rm s}$. Figure 5.2 illustrates how the radial cell size changes across the grid. This prescription has the advantage of creating a high resolution region near the planet that has a roughly uniform cell size, without introducing abrupt changes in resolution which are prone to generating numerical error. Figure 5.3 demonstrates the level of convergence our resolution has achieved with respect to our measurement of the horseshoe half-width, w. Reducing our resolution by a factor of 1.2, or even 1.44, changes our answer by about 1%.

At this resolution, we have $\sim$ 63 million cells in total. Using 3 GTX-Titan graphics cards housed in a single desktop computer, our GPU-accelerated code `PEnGUIn` runs at a speed of 0.84 seconds per timestep, or about 140 minutes per planetary orbit.

5.3 Flow Topology

We run our simulations for $100P_{\rm p}$ to reach a steady state, and then compute the time-averaged density and velocity field from 100 to $101P_{\rm p}$ to obtain our final results. Section 5.3.1 will describe the overall size and shape of the

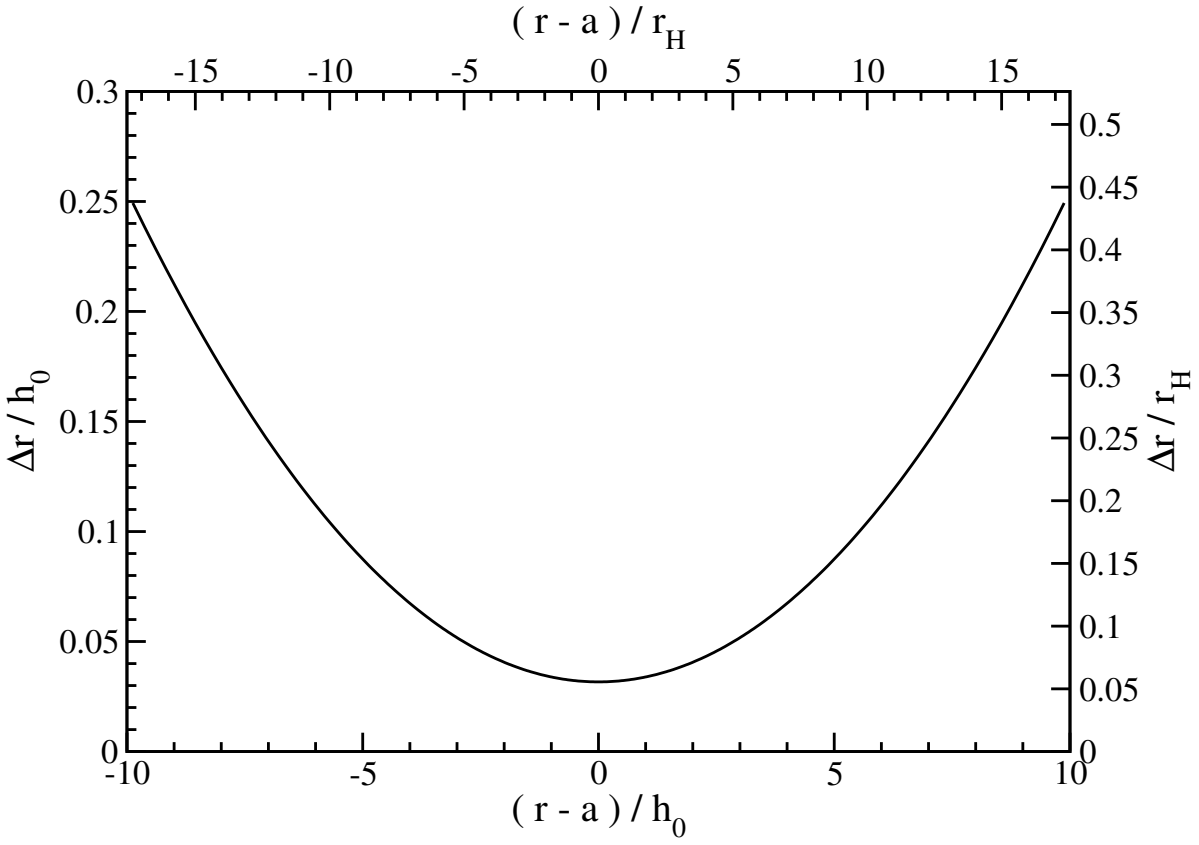

Figure 5.2 Radial resolution of our grid as described by Equations 5.11 and 5.12. Near the planet's location, we have ~ 32 cells per h_0, or ~ 18 cells per r_H.

horseshoe region; Section 5.3.2 will focus on the horseshoe turn, where a close encounter with the planet occurs, and show how this flow interacts with the planet's atmosphere; and finally, Section 5.3.3 will investigate the aftermath of the close encounter.

5.3.1 The Horseshoe Region

Near the midplane, we expect to find horseshoe orbits, like those in Figure 5.1. If we go to a higher altitude, above the planet's Hill sphere, it becomes unclear what kind of trajectory the gas will take to flow around, or across, the Hill sphere. One simple question we can ask is how far up in altitude does the horseshoe region extend.

We use our velocity data to reconstruct the fluid streamlines. At a given z, the largest $|\Delta r| = r - a$ that still performs a horseshoe turn is defined as w, the horseshoe half-width at that z; we will also refer to this streamline as the "widest horseshoe orbit". We measure w at $\phi = \phi_p - 1$ for the inner flow, and $\phi_p + 1$ for the outer one[2]. We find the two differ by about $\sim 1\%$, with the inner orbits being the wider one. We consider this difference insignificant for the scope of this chapter. The w we report is the average of the two.

Figure 5.3 plots w as a function of z. Remarkably, w remains nearly constant in z, even when z reaches $6r_H$ or $\sim 3h_0$. Its average value, weighted by disk density, is $\sim 1.8r_H$, and the corresponding libration time for the widest horseshoe orbit is $t_{lib} \approx 43P_p$. This width is wider than what is expect for planets in the linear regime ($q_{th} \ll 1$), which is $w \sim 1.2\, a\sqrt{aq/h_0}$ (Masset et al., 2006; Paardekooper & Papaloizou, 2009b). In these units, our planet has $w = 1.35\, a\sqrt{aq/h_0}$. This is consistent with the findings of Masset et al. (2006), where they showed that as q_{th} increases to unity, there is an increase in the horseshoe half-width compared to linearly estimated values. We also perform a series of 2D simulations, with varying smoothing lengths, to compare with our result. In Figure 5.4, we see that if one stacks layers of 2D disks, increasing r_s to mimic the effect of increasing altitude, one will not recover the same horseshoe half-width we find in 3D.

To help visualize these trajectories, Figure 5.5 plots the 3D streamlines of the widest horseshoe orbits. Clearly the horseshoe region has a columnar structure. It additionally shows that halfway through the widest horseshoe turns, as the flow crosses the planet's orbit at $r = a$, some of it rapidly accelerates vertically toward the planet, and completes the turns at a significantly lower altitude. In fact, all fluid elements that started at $z \leq 2h_0$ are pulled down to $z \lesssim r_H$ after their horseshoe turns.

On one hand, this flow is columnar; there is almost no vertical variation in the planar velocities. This appears akin to Taylor-Proudman columns, where $\partial \boldsymbol{v}/\partial z \approx 0$ due to a dominating Coriolis force. On the other hand, the vertical flow directly above (and below) the planet implies a rapid v_z variation in z, and a more complex, non-columnar flow structure can be seen immediately after the horseshoe turn (Figure 5.5); these flow patterns do not follow Taylor-Proudman theorem. In the following section, we will show analytically that a columnar flow structure is expected of a Keplerian disk; at the same time, we will account for the non-columnar features we observe.

Columnar Flow in a Keplerian Disk

We rewrite Equation 5.4 in a rotating frame (such as the rotating frame of a planet), where $\boldsymbol{u}$ is the velocity in the rotating frame and $\boldsymbol{\Omega} = (0, 0, \Omega_p)$ is the frame's rotation frequency.

$$\frac{\partial \boldsymbol{u}}{\partial t} + (\boldsymbol{u}\cdot\boldsymbol{\nabla})\boldsymbol{u} + 2\boldsymbol{\Omega}\times\boldsymbol{u} + \boldsymbol{\Omega}\times(\boldsymbol{\Omega}\times\boldsymbol{r}) = -\boldsymbol{\nabla}\Phi - \frac{1}{\rho}\boldsymbol{\nabla}p + \frac{1}{\rho}\boldsymbol{\nabla}\cdot\mathbb{T}, \tag{5.13}$$

[2]It is also possible to measure w post-horseshoe turn, by tracking the streamlines backward in time, but we find the flow there too complex for a clean measurement. See Section 5.3.3 for a discussion on the flow topology there.

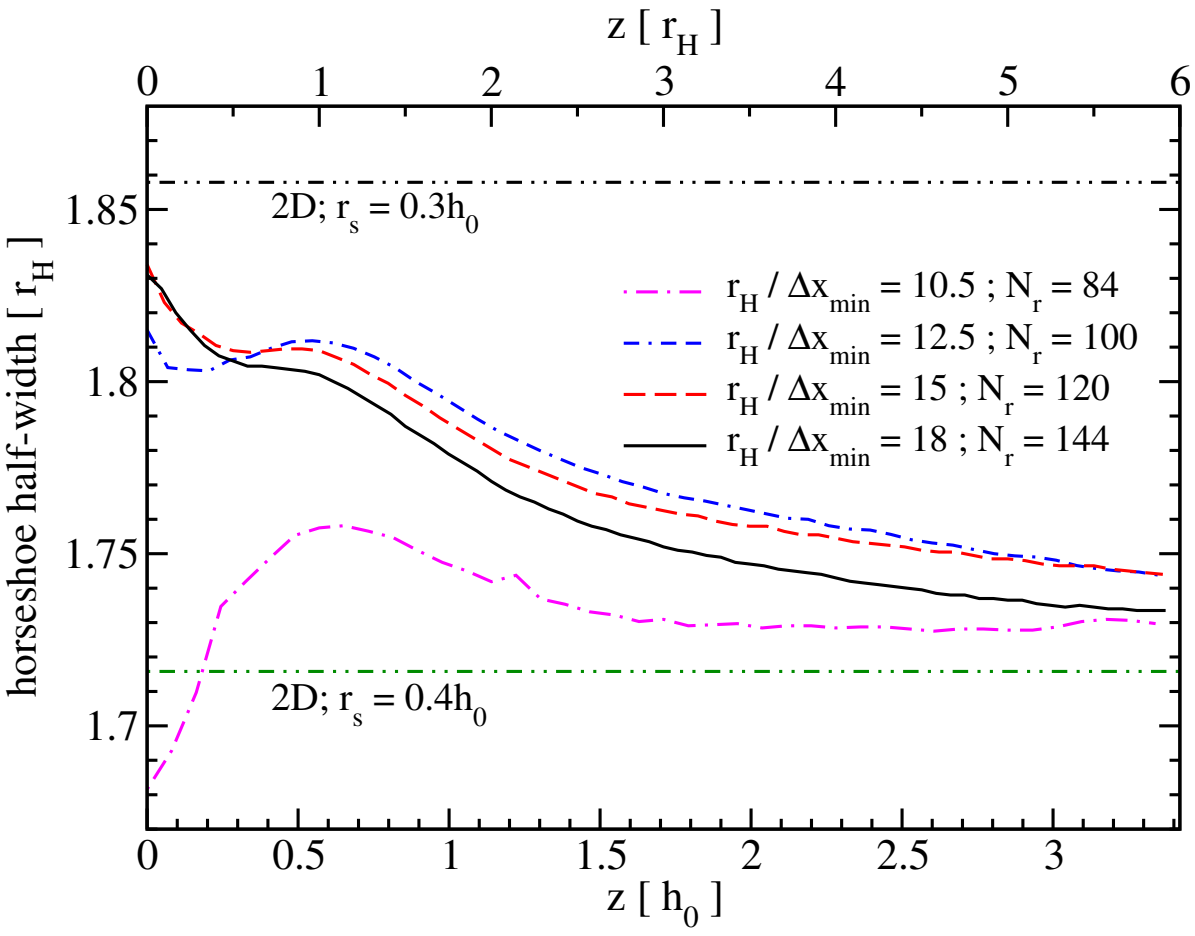

Figure 5.3 Horseshoe half-width as a function of height above the midplane. z refers to the height of the flow before the turn. The magenta dot-dashed curve, blue dot-dash-dashed curve, red dashed curve, and black solid curve are results from different simulations, where the resolution is 20% higher between each curve. The black solid curve is our choice of resolution. This plot shows that our measurement has converged to within 1%. Also shown for comparison are results from 2D simulations with different smoothing lengths, at the same resolution as the black solid curve (also see Figure 5.4).

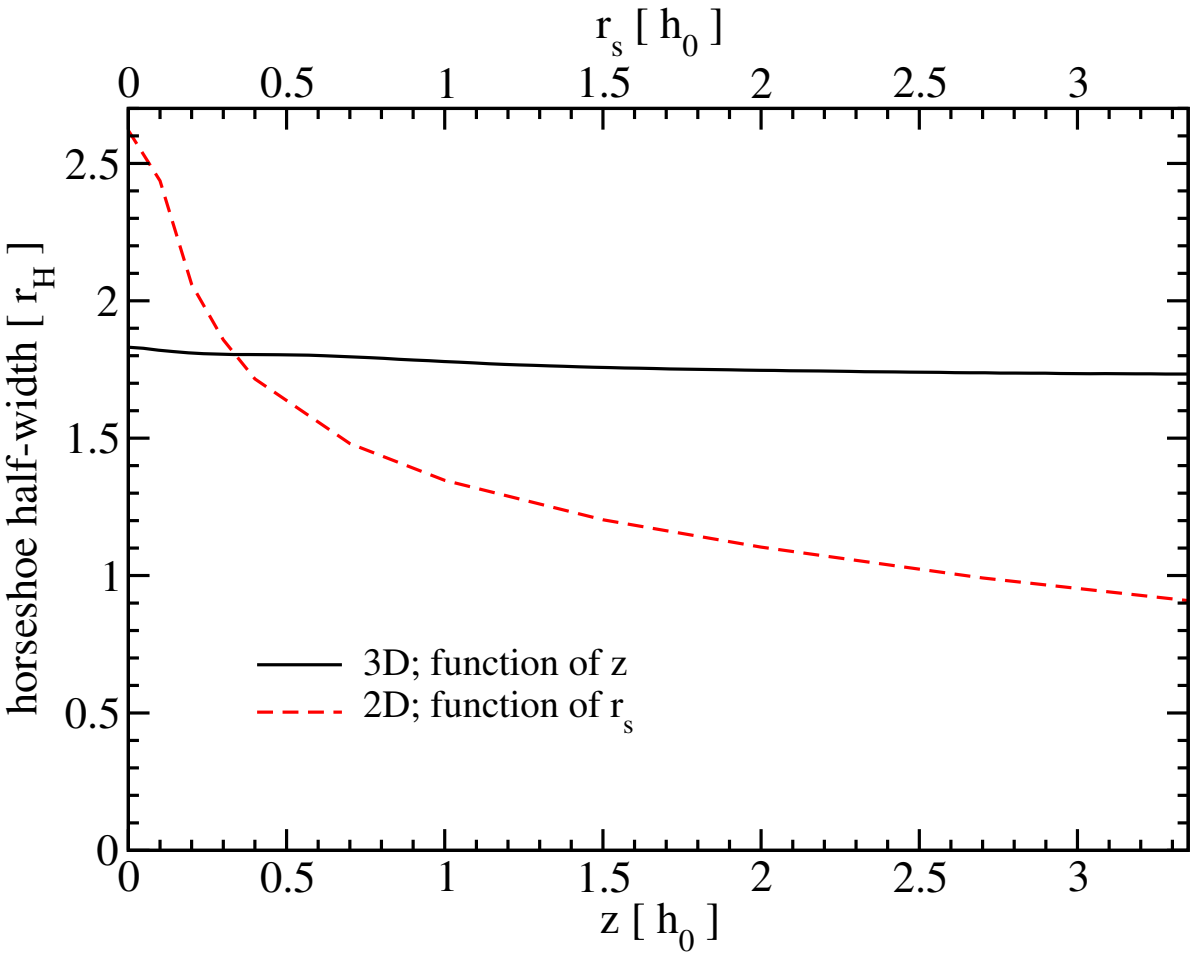

Figure 5.4 Horseshoe half-width as a function of height above the midplane in 3D, and a function of smoothing length in 2D. The black solid curve is the same as the black curve in Figure 5.3. The red dashed curve is from a series of 2D simulations. This plot demonstrates that a 3D disk behaves differently from a combination of 2D layers.

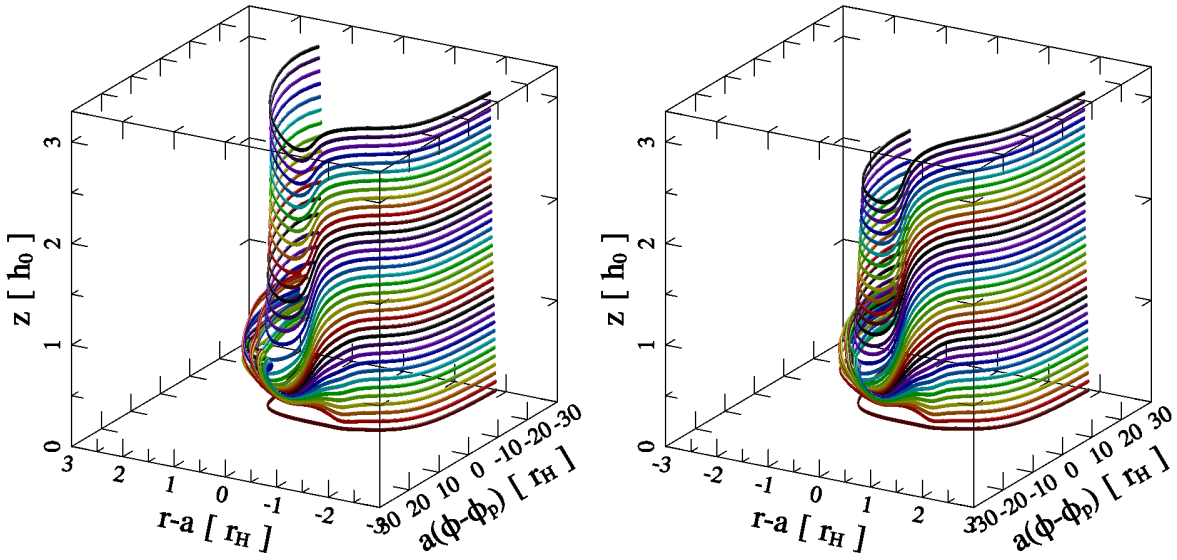

Figure 5.5 Streamlines of the widest horseshoe orbits. The left panel shows the inner flow, and the right shows the outer one. Note that 1) the flow has a columnar structure along the horseshoe turn; 2) most streamlines go through a sharp drop in altitude half-way through their turns, being drawn vertically to the planet; 3) a more complex flow structure is seen near the midplane after the turn; see Figure 5.10 for a close-up picture of the streamlines there.

where the terms on the left-hand side (LHS) are the time-dependent term, the advection term, the Coriolis force term, and the centrifugal force term; on the right-hand side (RHS) are gravity represented by potential Φ, pressure gradient force, and the viscous stress. For a barotropic, steady flow with a large Reynolds number ($Re \approx |\boldsymbol{u}|h_0/\nu \gg 1$), we can transform Equation 5.13 into the vorticity equation by taking the curl of both sides,

$$(\boldsymbol{u}\cdot\boldsymbol{\nabla})\boldsymbol{\omega} = (\boldsymbol{\omega}\cdot\boldsymbol{\nabla})\boldsymbol{u} - \boldsymbol{\omega}(\boldsymbol{\nabla}\cdot\boldsymbol{u})\,, \tag{5.14}$$

where $\boldsymbol{\omega} = \boldsymbol{\nabla}\times\boldsymbol{u} + 2\boldsymbol{\Omega}$ is the total vorticity. The second term on the RHS contains $\boldsymbol{\nabla}\cdot\boldsymbol{u}$, which describes the compressibility of the fluid. We can eliminate this term by combining Equation 5.14 with Equation 5.3 to obtain:

$$(\boldsymbol{u}\cdot\boldsymbol{\nabla})\boldsymbol{\xi} = (\boldsymbol{\xi}\cdot\boldsymbol{\nabla})\,\boldsymbol{u}\,. \tag{5.15}$$

where

$$\boldsymbol{\xi} = \frac{\boldsymbol{\omega}}{\rho} = \frac{\boldsymbol{\nabla}\times\boldsymbol{u} + 2\boldsymbol{\Omega}}{\rho} \tag{5.16}$$

is the vortensity (or potential vorticity) of the fluid. On the LHS of Equation 5.15 we have the advection of $\boldsymbol{\xi}$, and on the right is the vortex tilting term. Consider an incompressible flow (ρ = constant), and the case of a small perturbation in vorticity ($|2\boldsymbol{\Omega}| \gg |\boldsymbol{\nabla}\times\boldsymbol{u}|$). This gives $\boldsymbol{\xi} \approx 2\boldsymbol{\Omega}/\rho$ = constant. Additionally, because $\boldsymbol{\Omega}$ is non-zero only in the z direction, the RHS also simplifies, and Equation 5.15 is reduced to $\partial\boldsymbol{u}/\partial z = 0$, which is the classical Taylor-Proudman theorem stating that there is no vertical variation in the flow.

For a planet's horseshoe region, we can apply a similar analysis, but with relaxed assumptions. First, instead of the incompressibility assumption, we allow the fluid to be compressible, but restrict it to a subsonic flow. Quantitatively this means the shortest length scale over which ρ is allowed to vary is h. The local Keplerian shear

is supersonic far away from the planet, so this assumption also restricts us to a radial range of $r \sim a \pm h$; however, in practice, the Keplerian shear does not usually generate shocks in the disk, so as long as ρ is smooth, our analysis can apply to a larger radial range. Second, we note that a Keplerian disk does not satisfy the assumption $|2\boldsymbol{\Omega}| \gg |\boldsymbol{\nabla} \times \boldsymbol{u}|$, since $|\boldsymbol{\nabla} \times \boldsymbol{u}| = 3\Omega_\mathrm{k}/2$; instead, we assume the vorticity of the disk is mainly in the z direction, such that $\omega_z \gg \omega_r,\ \omega_\phi$. By our assumptions, the LHS of Equation 5.15 has a magnitude less than $\frac{c_\mathrm{s}}{h}\,|\boldsymbol{\xi}|$. Therefore in component form, Equation 5.15 can be estimated as:

$$\frac{c_\mathrm{s}}{h}\,|\omega_r| \gtrsim \left|\omega_r \frac{\partial u_r}{\partial r} + \omega_\phi \frac{1}{r}\frac{\partial u_r}{\partial \phi} + \omega_z \frac{\partial u_r}{\partial z}\right|, \tag{5.17}$$

$$\frac{c_\mathrm{s}}{h}\,\left|\omega_\phi\right| \gtrsim \left|\omega_r \frac{\partial u_\phi}{\partial r} + \omega_\phi \frac{1}{r}\frac{\partial u_\phi}{\partial \phi} + \omega_z \frac{\partial u_\phi}{\partial z}\right|, \tag{5.18}$$

$$\frac{c_\mathrm{s}}{h}\,|\omega_z| \gtrsim \left|\omega_r \frac{\partial u_z}{\partial r} + \omega_\phi \frac{1}{r}\frac{\partial u_z}{\partial \phi} + \omega_z \frac{\partial u_z}{\partial z}\right|. \tag{5.19}$$

Since the flow is subsonic, we can further approximate $\frac{c_\mathrm{s}}{h} \gtrsim \left|\frac{\partial \boldsymbol{u}}{\partial r}\right|, \frac{1}{r}\left|\frac{\partial \boldsymbol{u}}{\partial \phi}\right|$; note that $\left|\frac{\partial (v_\mathrm{k} - r\Omega_\mathrm{p})}{\partial r}\right| \sim \frac{3}{2}\Omega_\mathrm{p} = \frac{3}{2}\frac{c_\mathrm{s}}{h}$, which is within an order of unity to our approximation. Finally, rearranging Equation 5.17 to 5.19, our order-of-magnitude analysis gives:

$$\left|\frac{\partial u_r}{\partial z}\right| \lesssim \frac{c_\mathrm{s}}{h}\left(2\left|\frac{\omega_r}{\omega_z}\right| + \left|\frac{\omega_\phi}{\omega_z}\right|\right), \tag{5.20}$$

$$\left|\frac{\partial u_\phi}{\partial z}\right| \lesssim \frac{c_\mathrm{s}}{h}\left(\left|\frac{\omega_r}{\omega_z}\right| + 2\left|\frac{\omega_\phi}{\omega_z}\right|\right), \tag{5.21}$$

$$\left|\frac{\partial u_z}{\partial z}\right| \lesssim \frac{c_\mathrm{s}}{h}. \tag{5.22}$$

This demonstrates that in the planet's co-orbital region, the variation of v_r and v_ϕ in the z direction is suppressed by a factor of $\max\left(\left|\frac{\omega_r}{\omega_z}\right|, \left|\frac{\omega_\phi}{\omega_z}\right|\right)$; on the other hand, the vertical motion of the gas is allowed to vary over a length scale as short as h.

The physical interpretation of this result can be found in Equation 5.15, which states that a vortex line can only be tilted as much as advection can carry. If there is little planar vorticity to begin with, the advection of vorticity will not be strong enough to tilt the vortex lines to the in the r and ϕ directions, and as a result the flow must remain columnar. v_z, on the other hand, is aligned with the vortex lines, so a vertical acceleration will only lead to the compression or stretching of the vortex lines, which is allowed through the compressibility of the fluid.

We now verify whether the assumption $\omega_z \gg \omega_r,\ \omega_\phi$ holds in the flow around an embedded planet. In Figure 5.6, we plot the ratio f:

$$f \equiv \frac{\int_0^\infty \rho \frac{\sqrt{\omega_r^2 + \omega_\phi^2}}{|\omega_\mathrm{z}|}\,\mathrm{d}z}{\int_0^\infty \rho\,\mathrm{d}z}. \tag{5.23}$$

We find $f \ll 1$ in two regions: one where $r > a$ and $\phi > \phi_\mathrm{p}$, corresponding to the starting half of the outer horseshoe orbits before crossing the planet's orbit; and another one where $r < a$ and $\phi < \phi_\mathrm{p}$, which is the starting half of the inner horseshoe flow. This is consistent with where columnar structure is found (see Figure 5.5). In the regions corresponding to the finishing half of the widest inner and outer horseshoe turns, we find $f \gg 1$, indicating the planar vorticity overtakes vorticity in z. This corresponds to the more complex flow structure after the horseshoe flow crosses $r = a$ near the midplane (see Figure 5.5). In Section 5.3.3 we will further investigate this aspect of the flow.

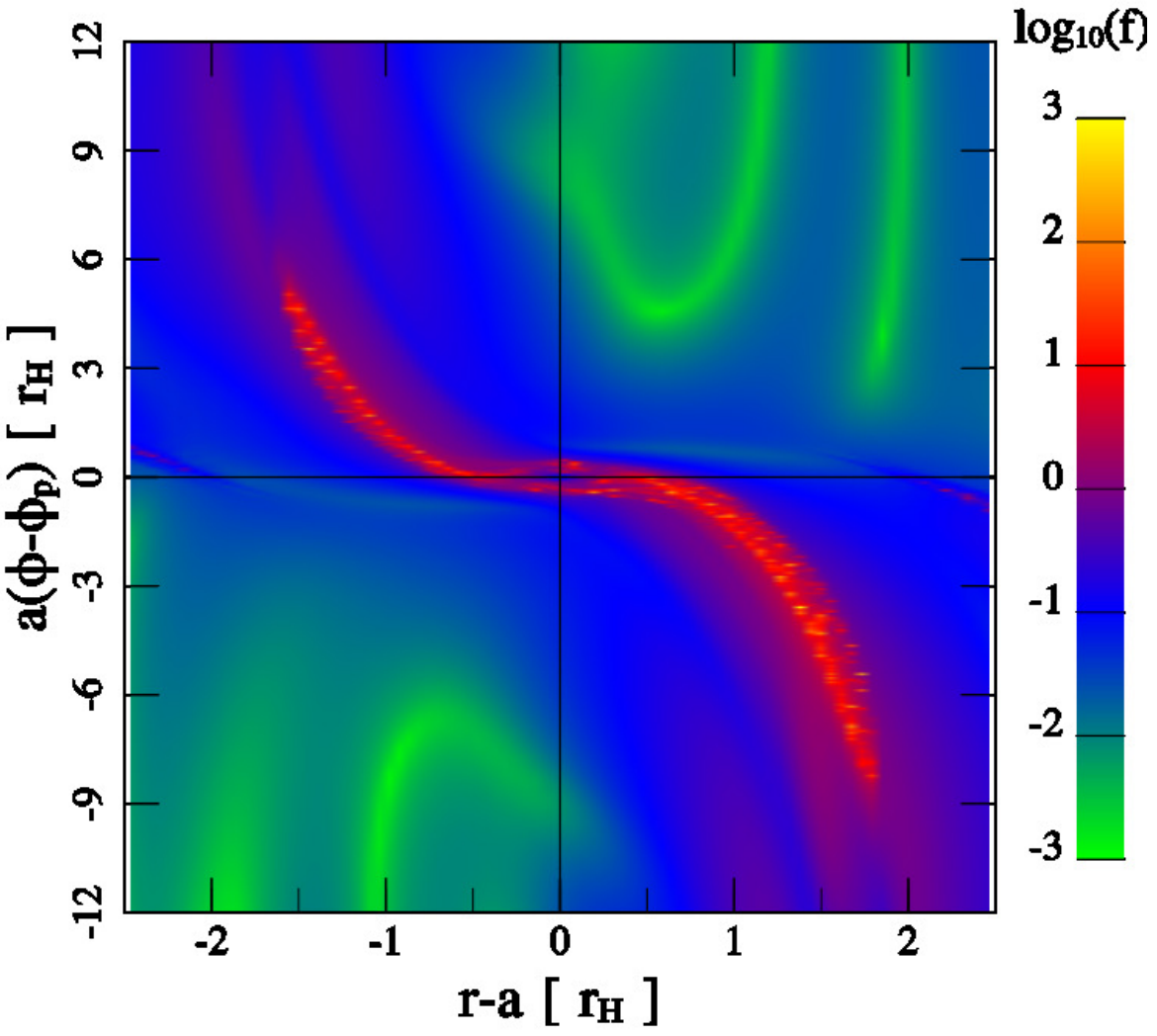

Figure 5.6 Density-weighted vertical average of the planar-to-z vorticity ratio (see Equation 5.23), plotted as a function of r and ϕ. Note that $f \ll 1$ in most regions, except for two streams corresponding to the finishing half of the widest inner and outer horseshoe turns.

5.3.2 Flow in the Planet's Bondi Sphere

In Section 5.3.1 we identify a rapid vertical motion in the horseshoe flow halfway through the horseshoe turn. This is a flow moving toward the planet from directly above (and below) it. Here we investigate how this vertical flow affects the planet's atmosphere. If one assumes the planet's atmosphere has an isothermal hydrostatic structure, it satisfies:

$$\frac{\mathrm{d}\eta}{\mathrm{d}z} = -\frac{\mathrm{d}\Phi}{\mathrm{d}z}\,, \tag{5.24}$$

where η is the enthalpy of the fluid, defined by $\mathrm{d}\eta = \mathrm{d}p/\rho$. This gives a vertical density profile at the planet's location:

$$\rho_{\mathrm{static}}(z) = \rho_0 \exp\left(-\frac{z^2}{2h_0^2} + \frac{r_{\mathrm{B}}}{\sqrt{z^2 + r_{\mathrm{s}}^2}}\right). \tag{5.25}$$

Figure 5.7 plots ρ_{static} together with the density profile we find from simulation. We find that near the planet, there is a large discrepancy between the hydrostatic solution and our simulation result. Within r_{B} of the planet (the Bondi sphere), the density can be an order of magnitude less than ρ_{static}. This is because the gas is not at rest. Figure 5.8 shows the streamlines in the midplane[3] of the disk, with a color code same as Figure 5.1, except for the magenta lines (see caption). Comparing to the 2D streamlines in Figure 5.1, the magenta lines, which represents the flow of material from within the Bondi sphere, have a qualitatively different behavior. In 2D, the atmosphere has closed stream lines bound to the planet; in 3D, there is no static atmosphere. Rather, there is a mass inflow near the planet's poles, and a comparable outflow in the equator (Figure 5.9).

So, similar to the conclusion reached by OSK15 in their study for small planets ($q_{\mathrm{th}} = 0.01$), we find that the flow within the planet's Bondi sphere is not static, but instead circulates with the disk. Moreover, it is clear from Figure 5.5 that the flow in and out of the planet's Bondi sphere is a part of the horseshoe flow, and that the two outflowing streams in Figure 5.8 are simply the continuation of the inner and outer horseshoe turns that has been pulled down from above the midplane. In fact, every magenta line in Figure 5.8 can be traced back to a horseshoe orbit that originated from an altitude of about $0.5 \sim 1h_0$. We call this the transient horseshoe flow, which we will discuss in depth in Section 5.3.3. Then, where is the planet's atmosphere?

There is one region where we do find 3D streamlines that do not leave the vicinity of the planet, which is a small sphere within $1.5r_{\mathrm{s}}$, or $\sim 0.15r_{\mathrm{B}}$, of the planet. Recall that r_{s} is equivalent to the planet's physical size. This means we are finding that the planet's atmosphere is not much larger than the pre-defined planet radius. However, it should be noted that our simulation grid resolves this region by merely 3 cells, so the flow there is not numerically accurate, creating substantial numerical viscosity that lowers the gas density. The explicit viscous force in this region also becomes large as it scales as r_{s}^{-2}, adding to the already substantial numerical viscosity. In reality, viscosity on this length scale should be much weaker than the disk viscosity, since the typical disk eddy size is $\sim h_0$, much larger than r_{s}. With a more accurate and realistic treatment, the amount of bound gas may be larger than what we measure. Other than this region, essentially all of the gas within planet's Bondi sphere are part of a more elaborate horseshoe flow.

Our results share similarities with Tanigawa et al. (2012), who performed local, isothermal, inviscid simulations with a massive planet that has $r_{\mathrm{H}} = h_0$ ($q_{\mathrm{th}} = 3$), and Ayliffe & Bate (2012), who performed SPH simulations with 15-33 $M_\oplus$ planets, taking into account self-gravity and radiation. Figure 5 of Tanigawa et al. (2012) and Figure 12 of Ayliffe & Bate (2012), each showing the mass flux across a sphere of radius $0.3r_{\mathrm{H}}$ around the planet, can be compared to our Figure 5.7. Ayliffe & Bate (2012), in their "post-collapse" regime after mass accretion has slowed down, find that influx is concentrated in the vertical direction, while outflux is only in the midplane.

[3]The midplane is the bottom boundary of our simulation grid, so these streamlines approximate the midplane by following the velocity field of the bottommost cells, which are centered on $z \approx 0.0005a$, with the z-component velocity set to zero.

This is in agreement with our findings. For the more massive planet that Tanigawa et al. (2012) simulated, they find both the influx and outflux of mass are more concentrated along the midplane, even though their influx is still noticeably offset to a higher latitude. This is consistent with our expectation that vertical motion plays a lesser role in the horseshoe flow for planets with $r_H > h_0$. Additionally, Tanigawa et al. (2012) reported that, similar to OSK15's results and our simulation, the gas is unbound at distances larger than $\sim 0.1 r_H$ from the planet. This may be indicating that the unbound nature of the gas in the Hill (for larger planets with $r_B > r_H$) or Bondi (for smaller ones) sphere is irrespective of planet mass.

OSK15 introduced the concept of replenishment timescale, $t_{\rm replenish}$, which measures the total amount of mass within the Bondi sphere, $M_{\rm BS}$, divided by the influx of mass into it, $\dot{M}_{\rm in}$. In our simulation, we have $t_{\rm replenish} \sim \Omega_p^{-1}$. Since we do not include any sink cells to treat planet accretion, we expect the net mass flux across the Bondi sphere to be zero in steady state. We find that $\dot{M}_{\rm in}$ is balanced by $\dot{M}_{\rm out}$, the outflux of mass, to within 1%, with $\dot{M}_{\rm in}$ being the larger one. The Bondi sphere is therefore still slowly accumulating mass after $100 P_p$, but on a timescale much longer than $t_{\rm replenish}$, so it is effectively in a steady state. If we lower resolution by factors of 1.2 to $r_H/\Delta x_{\rm min} = 15$ and 12.5, the flow pattern we see in Figures 5.8 and 5.9 remains qualitatively the same, but $M_{\rm BS}$ decreases by 2% and 5%, while $\dot{M}_{\rm in}$ increases by 10% and 30%, respectively. This indicates that, not surprisingly, pressure support in the Bondi sphere is better established at a higher resolution, and our result has converged to within a level of $\sim 10\%$.

Our measurement of $t_{\rm replenish}$ is much smaller than the $t_{\rm replenish} \sim 100\Omega_p^{-1}$ reported by OSK15 for their $q_{\rm th} = 0.01$ planet. The main difference between our result and theirs, is that the vertical motion near the planet is much faster in our case, which leads to more kinetic support, and lower density near the planet comparing to $\rho_{\rm static}$. Overall this gives a higher $\dot{M}_{\rm in}$, smaller $M_{\rm BS}$, and shorter $t_{\rm replenish}$. For comparison, OSK15 found a density profile that nearly exactly follows the hydrostatic solution (see their Figure 4), which indicates a much slower motion within the Bondi sphere, while D'Angelo et al. (2003) reported supersonic vertical motion above the planet for planet masses $\gtrsim 20 M_\oplus$. This suggests to us that the Bondi sphere is increasingly more kinetic supported as planet mass increases.

5.3.3 Transient Horseshoe Flow and Wake Vortices

By now, it is evident that in 3D, there exists a significant asymmetry in the flow pattern around the planet before and after the horseshoe turn. There is a new flow, which we call the transient horseshoe flow, where fluid at high altitude is pulled down toward the planet, enters its Bondi sphere, and exits radially in the midplane. These are shown by the magenta lines in Figure 5.8 for the midplane.

The word "transient" refers to the fact that this flow only performs the horseshoe turn once, even in steady state. Because of the fall in altitude, gravitational potential energy is converted to kinetic, and the flow gains radial speed as it completes its horseshoe turn. The transient flow is the portion of the horseshoe flow that has gathered enough speed to over-shoot radially and exit the horseshoe region. We analyze the radial outflow in Appendix B and show that, due to the conservation of Bernoulli's constant, the radial outflow at $|r - a| = r_H$ has a speed of $|u_r| \approx 0.6 c_s$, while we measure $0.2 c_s$ to $0.4 c_s$ in our simulation. Appendix B also demonstrates that this outflow speed is independent of planet mass in the limit $q_{\rm th} \gg 1$, but is slower for lower mass planets.

Since material in the horseshoe region is being lost to the transient flow, it must be replenished. This comes from the disk flow lying outside of the horseshoe region: as the transient flow completes its horseshoe turn, it becomes vertically compressed, creating a low pressure region above it, which attracts the high altitude flow right next to the horseshoe region to move in (Figure 5.10). This establishes an exchange of material between the horseshoe region and the disk , as the red and green streamlines in the figure wrap around each other.

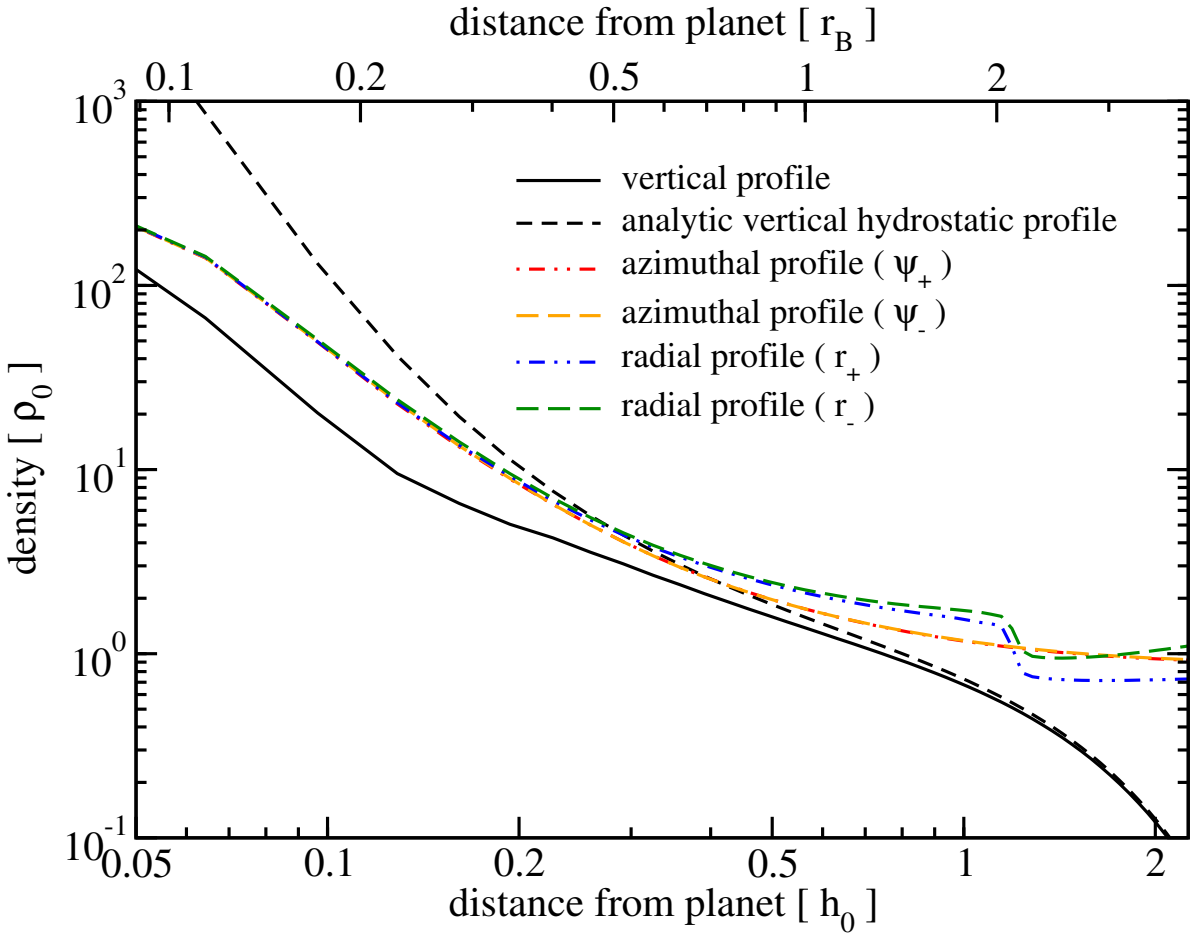

Figure 5.7 Gas density as a function of distance from the planet. Black solid curve plots the vertical density profile; red dash-dot-dotted and orange dashed curves are both azimuthal profiles in the midplane, but in the increasing and decreasing direction of ϕ respectively; similarly, blue dash-dot-dotted and green dashed curves plot the radial profiles in the midplane, and are in the increasing and decreasing direction of r. Black dashed curve is calculated with Equation 5.25. Note the large discrepancy between the black solid and black dashed curves. Comparing the black solid curve to the four profiles in the midplane, we see that the density structure near the planet is flattened by a factor of about 2/3.

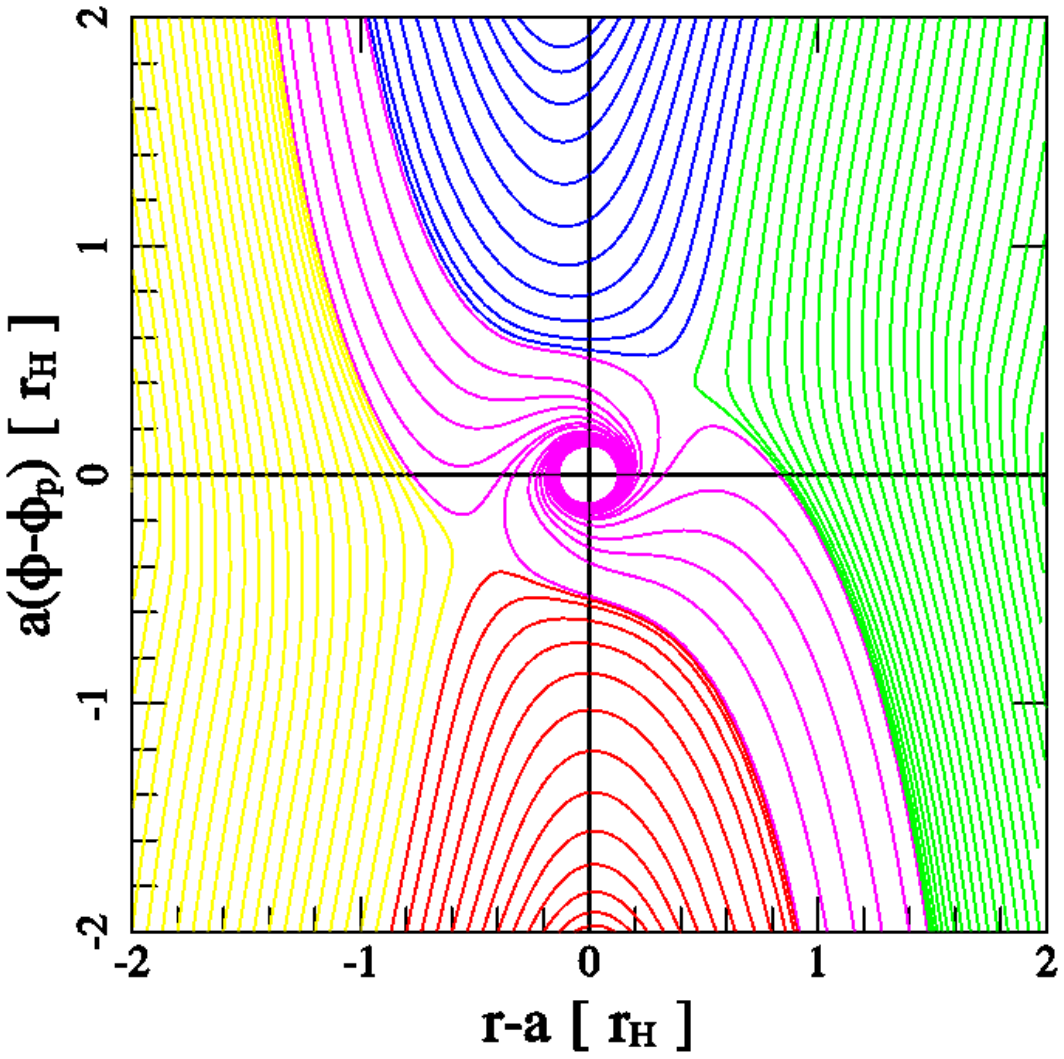

Figure 5.8 Streamlines in the disk midplane. Compare with Figure 5.1 for differences between 2D and 3D flow. Yellow, red, green, and blue streamlines are assigned in the same manner as Figure 5.1. Unlike Figure 5.1, magenta lines are outflows away from the planet, pulled down from initially higher altitudes. They reach as close as $1.5r_s$ from the planet and are unbound.

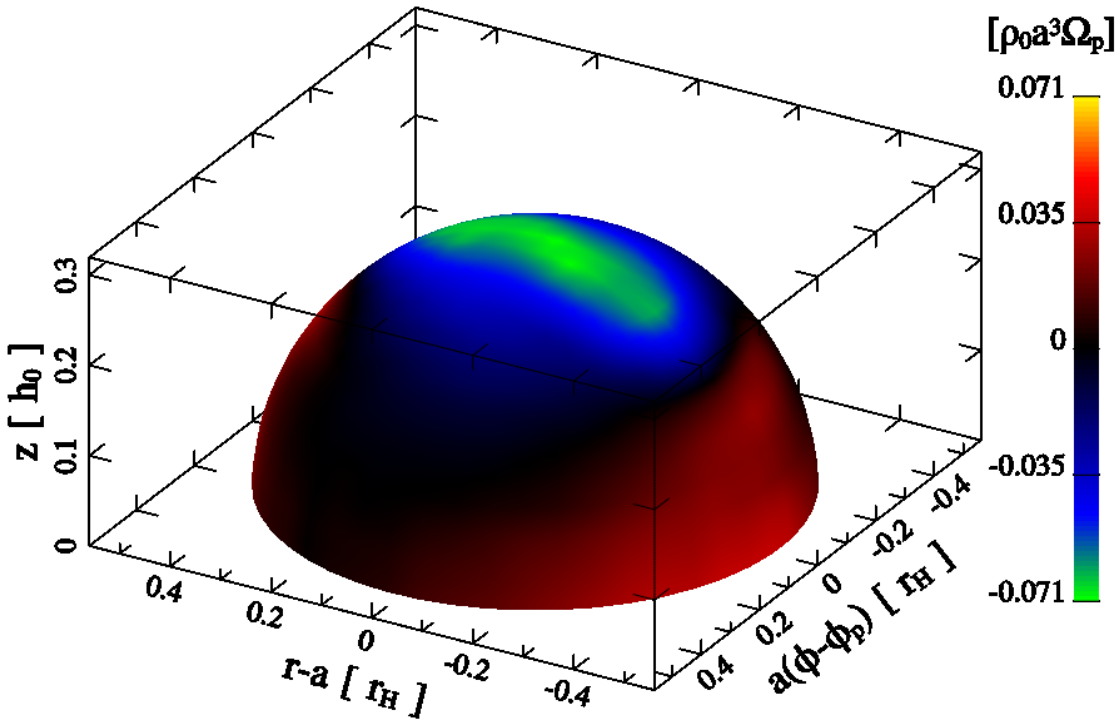

Figure 5.9 Mass flux across the surface of a sphere centered on the planet. The sphere has a radius of $0.5r_{\rm B}$. Blue and green indicate influx; red and yellow are outflux. The speed of the downward flow is about $0.7c_{\rm s}$ in this plot, while the two radial outward flows in the midplane (one not visible from this viewing angle) each has a speed of $\sim 0.2c_{\rm s}$, as is explained in Appendix B. Match this figure with Figure 5.8 for a more complete view of the flow topology near the midplane.

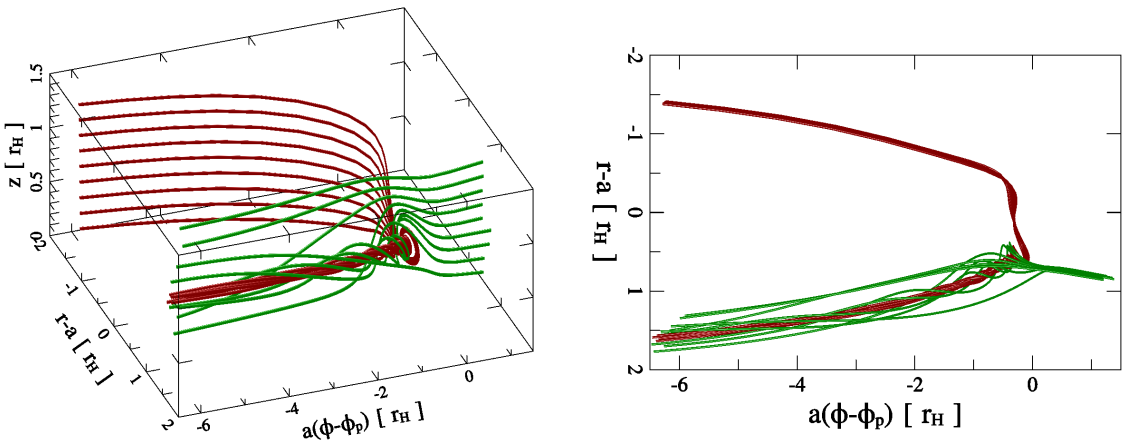

Figure 5.10 Streamlines at the boundary of the horseshoe region. The red lines are inner horseshoe flow; green are outer disk flow. After a close approach to the planet, the red streamlines turn around and descend to the midplane of the disk, sliding underneath the green streamlines. Green lines in higher altitude simply enters the horseshoe region, while lower ones are mixed with the red lines. Similarly, but not shown here, this also happens between the inner horseshoe flow and the outer disk flow.

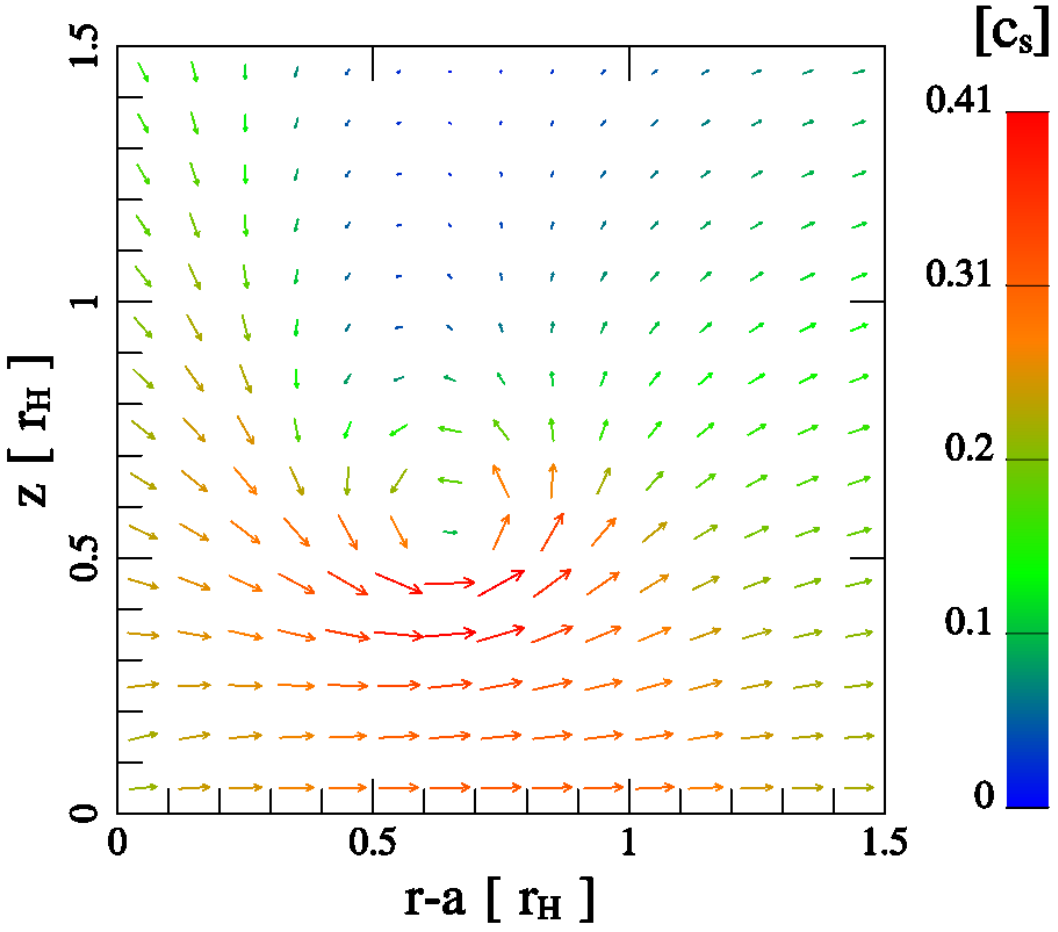

Figure 5.11 Velocity field on a meridional plane at $\phi = \phi_p - 0.5 r_H/a$. The color of the arrows indicates the speed. The fastest radial flow speed is $\sim 0.4 c_s$ in this plot. The vortex roll-up occurs between $0.5 \sim 1 r_H$ away from the planet, and about $0.5 r_H$ above the midplane. The size of the vortex core is about $0.1 \sim 0.2 r_H$.

Finally, the vertically compressed transient flow needs to decompress as it settles into the disk. This is done through a meridional circulation triggered by a "vortex roll-up", illustrated by Figure 5.11. The same meridional flow was also identified by Morbidelli et al. (2014) in their 3D simulation for a Jupiter-mass planet. As the transient flow over-shoots the horseshoe region, it gets deflected upward by the midplane disk material, and rolls over to decompress itself, causing the vortex roll-up. This phenomenon is similar to the behavior of the heads of plumes of fluid intruding into stationary fluid, which tend to roll at the edges to form mushroom-head shapes. However, the closest analogy is the formation of wingtip vortices in aerodynamics of finite-span wings, where higher pressure air underneath most of the wing length begins to move down (in a downwash), but near the wingtip also moves sideways and circulates around the tip to move into the low pressure region above the wing. The downwash and the associated aerodynamical lift are due to the circulation induced around the wing by vortex lines attached to the airfoil spanwise (Joukovsky theorem states the proportionality of lift and circulation). By Kelvin's circulation theorem, the vorticity ω, a solenoidal, divergence-free field, cannot simply end at the ends of the finite-span airfoil. Instead, the vortex line remains continuous and preserves the circulation; it only changes direction and is shed from the wingtips into the wake as two free wingtip vortices separated roughly by one wingspan.

In the protoplanetary disk, we have a close analogy: the gas in the transient horseshoe flow after a passage near the planet is torqued by its gravity and forms two outflow plumes, one directed toward the star and one in the opposite direction. The vertical extent of this flow is $\sim r_{\rm H}$, which plays the role of the aerodynamical wingspan of airfoil; two equivalent airfoils of this length would be positioned vertically and separated radially, also by a distance on the order of a few $r_{\rm H}$. Therefore, the planet's torque on the gas necessarily creates radial plume-like flows of gas, but also a total of four, alternately turning, nearly horizontal vortex lines shed into the disk flow near the interface between the horseshoe and disk regions. Figure 5.10 shows one of these vortices. We call them linear wake vortices, as a reminder that compared with their core diameters, they can be very long, as they are carried by the disk flow.

Furthermore, the equation of vortensity Equation 5.15 supplemented by viscous dissipation term, which we dropped before, can be written as

$$\frac{\mathrm{D}\boldsymbol{\xi}}{\mathrm{D}t} = (\boldsymbol{\xi}\cdot\boldsymbol{\nabla})\,\boldsymbol{u} + \nu\boldsymbol{\nabla}^2\boldsymbol{\xi}\,, \tag{5.26}$$

where $\boldsymbol{\xi}$ is the vortensity given by Equation 5.16. The evolution of vortensity in the core of the vortex can be deduced from this equation, after noticing that vortensity and velocity $\boldsymbol{u}$ are almost parallel and directed along the vortex line.

The first term on the RHS of the above equation is thus the gradient of longitudinal velocity of material along the vortex core multiplied by ξ. As shown in Figure 5.10, the roll-up of the wake vortices happens near the stagnation region, where azimuthal motion with respect to the planet is slow. After the vortex core forms, it is carried at an increasing speed into the disk flow, therefore the velocity gradient is first strongly positive, and further from the planet drops to zero, as the flow becomes an azimuthally non-accelerated disk flow. While the flow accelerates, the full time derivative of vortensity grows, strengthening the wake vortex exponentially (in the direction of vortex line, spatial derivative of velocity equals $\mathrm{D}\ln\boldsymbol{\xi}/\mathrm{D}t$). Only when the vortex starts freely coasting away from the planet, several $r_{\rm H}$ behind its roll-up region, the second term in the equation representing viscous spreading becomes dominant, diffusing the vortex and weakening the vortensity of its core.

Viscous dissipation of the core has an associated timescale equal to the squared diameter d^2 of the core divided by ν, during which fluid elements are carried in azimuth by differential rotation through a distance $\lambda \sim (d^2/\nu)\Omega_{\rm p} r_{\rm H}$. If $d \sim 0.1 r_{\rm H}$ as seen in Figure 5.11, $r_{\rm H} = 0.017a$ appropriate for our $5M_\oplus$ planet, and $\nu = 10^{-6}a^2\Omega_{\rm p}$, then the estimated distance is $\lambda \sim 3r_{\rm H}$. This is an estimate of the distance on which viscosity in our simulation will

double or e-fold the initial diameter by assumption equal to $0.1r_H$. The freely flowing line vortex can spread and effectively disappear after a few length scales λ. This estimate agrees very well with our numerical simulations: vortices are observed to weaken substantially past $10r_H$ behind the planet (Figure 5.6). It is intriguing to ask what will happen to this vortex if the disk is inviscid, which we will investigate in the future.

5.4 Torque on the Planet

As mentioned in Section 5.2.1, our disk has a flat 2D vortensity profile, ω/Σ = constant, which is predicted by 2D analysis to have a vanishing corotation torque. Since vortensity is a conserved quantity along a 2D streamline, a fluid element performing a horseshoe turn will always maintain a density the same as its surrounding if the background vortensity profile is constant; hence there cannot be a density difference along any given r between the inner and outer horseshoe flow, and no corotation torque can be generated. As a result, the only remaining torque is the differential Lindblad torque, which, by the 3D linear calculations by Tanaka et al. (2002), is $-2.19T_0$, where $T_0 = \Sigma_0 a^4 \Omega_p^2 q^2 (h_0/a)^{-2}$, for an isothermal disk with $\Sigma \propto r^{-3/2}$.

The above results may not be applicable in 3D for two main reasons. First, while the 2D vortensity ω/Σ can be a constant in the disk, the 3D vortensity ω/ρ depends on z. For instance, in our model, $\omega/\rho \propto r^{3/2}$ in the midplane even though ω/Σ is constant. This is because $\rho(z = 0) \propto r^{-3}$ (see Equation 5.7). Second, and more importantly, vortensity is a conserved quantity along a streamline in 2D, but not in 3D, as is evident in Equation 5.15. In fact, Figure 5.6 already shows that planar vorticity is generated after the widest horseshoe turns. This section aims to investigate how the corotation torque behaves in 3D.

Before computing the planetary torque in our simulation, we first inspect the density structure near the planet. Figure 5.12 plots the midplane density scaled by the background density and with the axisymmetry density about the planet removed:

$$\Delta\rho = \rho\left(\frac{r}{a}\right)^3 - \frac{1}{2\pi}\int_0^{2\pi} \rho \, \mathrm{d}\phi' \,, \tag{5.27}$$

where ϕ' is the azimuth in a polar coordinate centering on the planet. The axisymmetry part of the density does not contribute a net torque, so it can be safely removed for clarity. One notable problem we can see in Figure 5.12 is the four-armed spiral around the planet. It is a numerical artifact due to the local Cartesian grid geometry around the planet. This problem was also identified by Ormel et al. (2015a). It can be reduced by sufficiently resolving r_s, which unfortunately is not the case for us, since we only resolve r_s by about 2 cells. The four-armed spiral introduces an artificial torque on the planet that needs to be removed. We therefore exclude the torque contribution from within $0.5r_B$ of from the planet, shown in Figure 5.12 as the black circle.

Now we compute the torque. Figure 5.13 plots the torque distribution $\mathrm{d}T/\mathrm{d}r$, which is the amount of torque on the planet by the disk at a given r:

$$\frac{\mathrm{d}T}{\mathrm{d}r} = \int_0^{2\pi}\int_{-\infty}^{\infty} \rho \frac{\partial\Phi}{\partial\phi} \, \mathrm{d}z \, \mathrm{d}\phi \,, \tag{5.28}$$

where T is the net torque on the planet. Figure 5.14 plots the net torque as a function of time, demonstrating that our measurements have converged with time. Figure 5.15 shows how the net torque depends on our choice of the excised region's radius. We find the radius needs to be at least $0.4r_B$, or about 7 grid cells, to fully remove contribution from the non-physical four-armed spiral, and our choice of $0.5r_B$ safely accomplishes that without excessively removing contribution from the disk. We further divide the torque into two components: one from within the planet's Bondi sphere (red curve; contribution from $|\boldsymbol{r} - \boldsymbol{r}_p| < r_B$, represented by the red circle in Figure 5.12), and one from the rest of the disk (blue curve). While the blue curve has the characteristic shape of

the Lindblad torque distribution, the red curve is not a well-known feature. If we integrate each curve, the red curve gives a torque of $+0.50T_0$, while the blue one gives $-1.27T_0$. The net torque is therefore $-0.77T_0$. This is significantly weaker than the result from linear calculation ($-2.19T_0$).

We also perform a 2D simulation to show this result is unique in 3D. The 2D setup is identical to 3D, except we set $r_s = 0.3h_0$, since in Section 5.3.1 we find this smoothing length produces the best matching horseshoe width. Our 2D torque is $-2.86T_0$ (Figure 5.14), comparable to the 3D value from linear calculation. For comparison, D'Angelo et al. (2003) also found 3D torques are about one order of magnitude weaker than 2D when the planet mass is around 10 $M_\oplus$.

Similar to how the blue curve has two bumps near $r - a = \pm h_0$, and red curve also has two separate bumps in the inner and outer disk, but with reversed signs compared to the blue. We will refer to this behavior as "torque reversal". This was also seen by D'Angelo & Lubow (2010) in their Figure 15 where they simulated planets with masses larger than a few $M_\oplus$. This fact, together with Figure 5.12, provides clues to the nature of this torque. In Figure 5.12 we marked two stagnation points, where $|\boldsymbol{u}| = 0$. We see that in the outer disk, the stagnation point is slightly above ϕ_p, while it is below ϕ_p in the inner disk. This is significantly different from the 2D flow pattern (see Figure 5.1), where both stagnation points lie much closer to ϕ_p. Because of this offset, high density regions are created at ϕ larger (less) than ϕ_p in the outer (inner) disk, thus generating a positive (negative) torque near the planet. However, it remains unclear to us why the net contribution from the red curve is positive, i.e., the offset in the outer disk is larger than the inner disk, conveniently reducing the net migration rate.

We also note that even if we ignore the red curve, the blue curve still only contributes a torque of $\sim -1.3T_0$, significantly weaker than $\sim -2.2T_0$ from either linear calculations or 2D simulations. This may have to do with the fact that the blue curve contains both Lindblad and corotation torque. From Figure 5.13 we can see that a large fraction of the torque distribution coincides with the horseshoe region. This prevents us from distinguishing which part belongs to the corotation torque, and which part is the Lindblad torque. However, we can measure the corotation torque with a different method, separate from Figure 5.13.

We follow a fluid element's motion starting at $\phi = \phi_p - 1$ for the inner flow, or $\phi_p + 1$ for the outer, until it completes its horseshoe turn and returns to its starting azimuth. Δl is then the difference in the fluid's specific angular momentum between its start and end points. Combining this with the flow rate in the horseshoe region, the corotation torque, T_{CR}, is:

$$T_{CR}(z) = T_{CR,i}(z) + T_{CR,o}(z) \tag{5.29}$$

$$= \int_{a-w_i}^{a} \int_{-z}^{z} \rho |u_\phi| \Delta l \, \mathrm{d}z' \, \mathrm{d}r \Big|_{\phi=\phi_p-1} \tag{5.30}$$

$$+ \int_{a}^{a+w_o} \int_{-z}^{z} \rho |u_\phi| \Delta l \, \mathrm{d}z' \, \mathrm{d}r \Big|_{\phi=\phi_p+1}, \tag{5.31}$$

where $T_{CR,i}$ is the corotation torque due to the inner horseshoe flow, $T_{CR,o}$ is the outer one; w_i and w_o are the horseshoe half-widths for the inner and outer flow respectively. Furthermore, we can separate the contribution from the transient horseshoe flow by identifying streamlines that settle outside of the horseshoe region. Figure 5.16 plots the differential torque $|\mathrm{d}T_{CR}/\mathrm{d}z|$ on the left panel and cumulative torque $|T_{CR}|$ on the right. Here z refers to starting position of the streamline, not where the torque exchange happens. A caveat with this method is that our velocity field is time-averaged over just $1P_p$, whereas the libration time for these horseshoe orbits are much longer. Nonetheless, because Figure 5.14 shows the 1 P_p time-averaged net torque has little fluctuation over a libration time, this method should be sufficiently accurate for our purpose.

We find that the one-sided torques, $T_{\rm CR,i}$ and $T_{\rm CR,o}$, each has a magnitude of $\sim 50T_0$. The transient horseshoe contributes $\sim 6\%$ of that torque. $T_{\rm CR,o}$ is stronger than $T_{\rm CR,i}$ by $\sim 3\%$, and the net corotation torque is $T_{\rm CR} \sim 1.5T_0$. This sufficiently accounts for the difference between our measured net torque ($-0.77T_0$) and the expected differential Lindblad torque ($-2.19T_0$), which strongly suggests that the corotation torque is responsible for both the torque from the red curve and the reduction of disk torque in the blue curve. Recalling that our disk profile would have zero net corotation torque in 2D, our result accentuates the difference between 3D and 2D torques. The magnitude and sign of the net corotation torque is closely related to the asymmetry in the stagnation point offsets, and will be a topic for the future.

D'Angelo & Lubow (2010) calculated the fully nonlinear 3D torque on a planet with $q \sim 3 \times 10^{-6}$ and a disk profile similar to ours, and found the net torque on the planet to be $T = -2.29T_0$, consistent with previous 2D results, implying the corotation torque is negligible when $\Sigma \propto r^{-3/2}$. Their simulations differ from ours in two ways. First, our planet is 5 times more massive, and over 20 times larger in terms of $q_{\rm th}$. Therefore one can expect our result to deviate somewhat from linear calculations. Second, their simulation is in the regime where $t_\nu < t_{\rm lib}$, while ours has $t_\nu > t_{\rm lib}$ (see Section 5.2.1). We believe the second point is the main cause for the discrepancy between their result and ours. In the next section, we will simulate a more viscous disk, and check whether we can recover their result.

5.5 Dependence on Viscosity

We increase the viscosity in our model to $\nu = 10^{-5}a^2\Omega_{\rm p}$ ($\alpha \sim 0.01$) and investigate how this will affect the flow pattern and the torque on the planet. We will refer to this case as the "viscous" case. Under this setup, the viscous diffusion timescale across the horseshoe region t_ν is $\sim 15P_{\rm p}$, shorter than $t_{\rm lib} \sim 43P_{\rm p}$. As a result, most of the gas in the horseshoe region can only complete one horseshoe turn before being removed due to the background viscous flow. This prevents the flow pattern described in Section 5.3 from fully setting up. We find that in this scenario, the flow has significantly reduced vertical variation, and becomes more similar to the 2D case.

Figure 5.17 plots the midplane streamlines similar to Figure 5.1 and 5.8, but for the viscous case and a 2D simulation with $r_{\rm s} = 0.3h_0$ (unlike Figure 5.1, which has $r_{\rm s} = 0$). In contrast to Figure 5.8, the magenta lines in the viscous case are now inflowing streams entering the Bondi sphere that converges at the stagnation point where the planet is located. They correspond to the gas attempting to stably orbit the planet, but loses angular momentum (with respect to the planet) too rapidly due to both numerical and explicit viscosity. The speed of these magenta lines is very slow close to the planet, less than 1% of $c_{\rm s}$, suggesting much of the flow diverges from the midplane before ever reaching the planet. As a result, most of the influx of mass is still from the vertical direction above the planet. On the other hand, the asymmetry about $r = a$ is much reduced in the viscous case. As we have shown in Section 5.3.3, the asymmetry is caused by the vertical flow near the planet, so, not surprisingly, we find significantly reduced vertical motion. The speed at $0.5r_{\rm B}$ above the planet is $0.1c_{\rm s}$, 7 times slower than our fiducial case (Section 5.3.2).

The more symmetric streamlines also mean the stagnation points now lie much closer to $\phi_{\rm p}$, seen on the left panel of Figure 5.17. This reduces the net torque from the horseshoe region. Figure 5.18 plots the torque distribution of the viscous case as well as that from Figure 5.13 for comparison. It is evident that the torque reversal seen before between the blue and red curve in Figure 5.13 has now been erased by the extra viscosity. The net torque on this planet is $-2.2T_0$, while the 2D case on the right panel of Figure 5.17 has a torque of $-2.86T_0$. These are in agreement with linear calculations of the differential Lindblad torque in 3D (Tanaka et al., 2002) and 2D (Paardekooper & Papaloizou, 2008), respectively. We therefore conclude that, when $t_\nu < t_{\rm lib}$, the

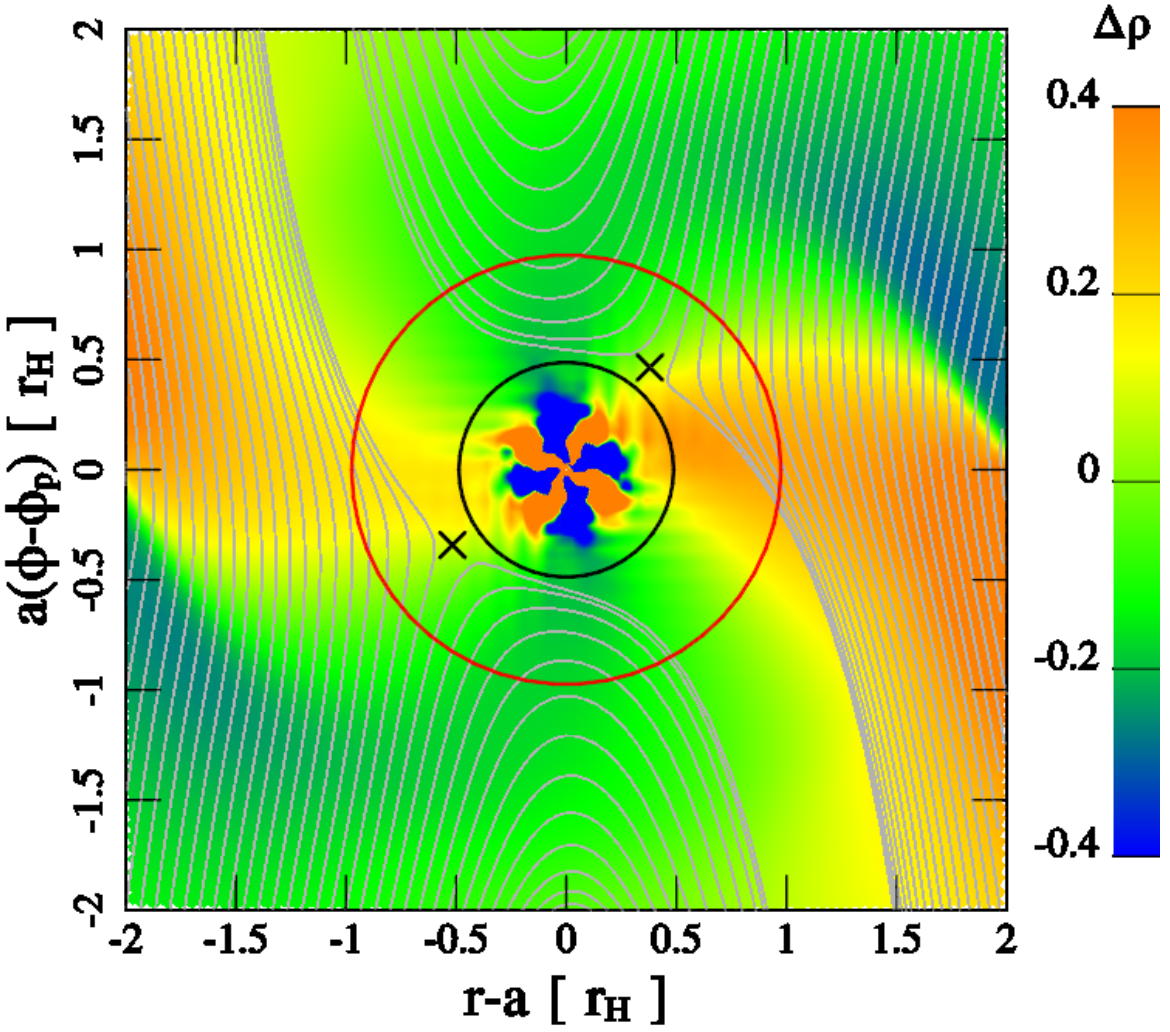

Figure 5.12 The midplane non-axisymmetric density distribution around the planet, scaled by the background density (see Equation 5.27). The gray lines are streamlines in Figure 5.8, except with the magenta lines omitted. The crosses mark the stagnation points. They are located at $\{x, y\} = \{-0.47, -0.36\}$ and $\{0.42, 0.42\}$, in units of $r_{\rm H}$. This is different from the 2D case (Figure 5.1), where the stagnation points lie close to $\phi_{\rm p}$. The black circle has a radius of $0.5r_{\rm B}$. Because of the non-physical four-armed spiral inside the black circle, we exclude this region from our torque calculation. The red circle's radius is $r_{\rm B}$, corresponding to the sphere where the red curve in Figure 5.13 is computed.

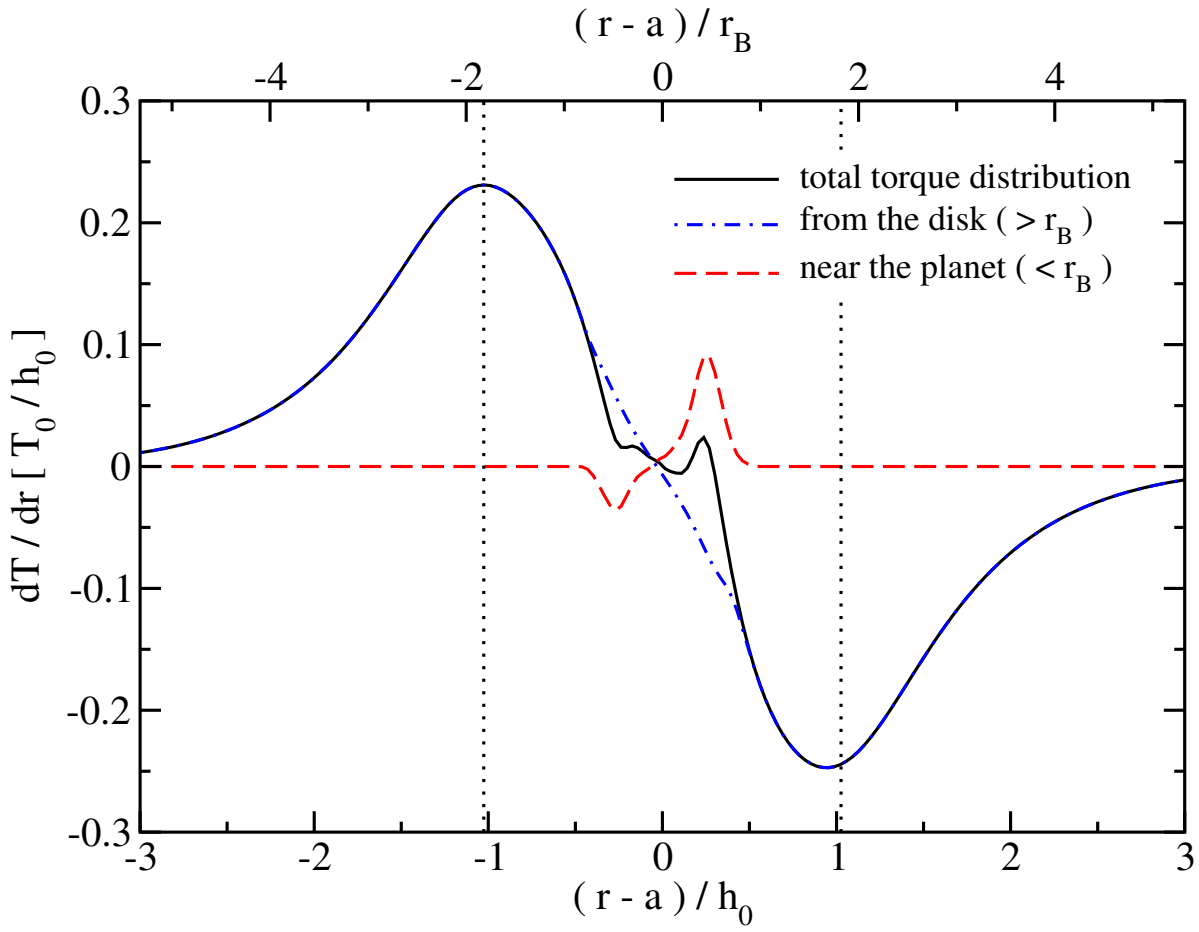

Figure 5.13 Torque distribution as a function r. The black solid curve is the total torque distribution, and is equal to the sum of the red dashed and blue dash-dotted curves. The red curve only includes contribution from within a sphere of 1 r_B around the planet (see Figure 5.12), while the blue curve includes the rest of the disk. The two black dotted lines draw the boundaries of the horseshoe region. The two blue bumps at $\pm h_0$ correspond to the outer and inner Lindblad torques; and the two red bumps near the planet are caused by the stagnation point offsets seen in Figure 5.12.

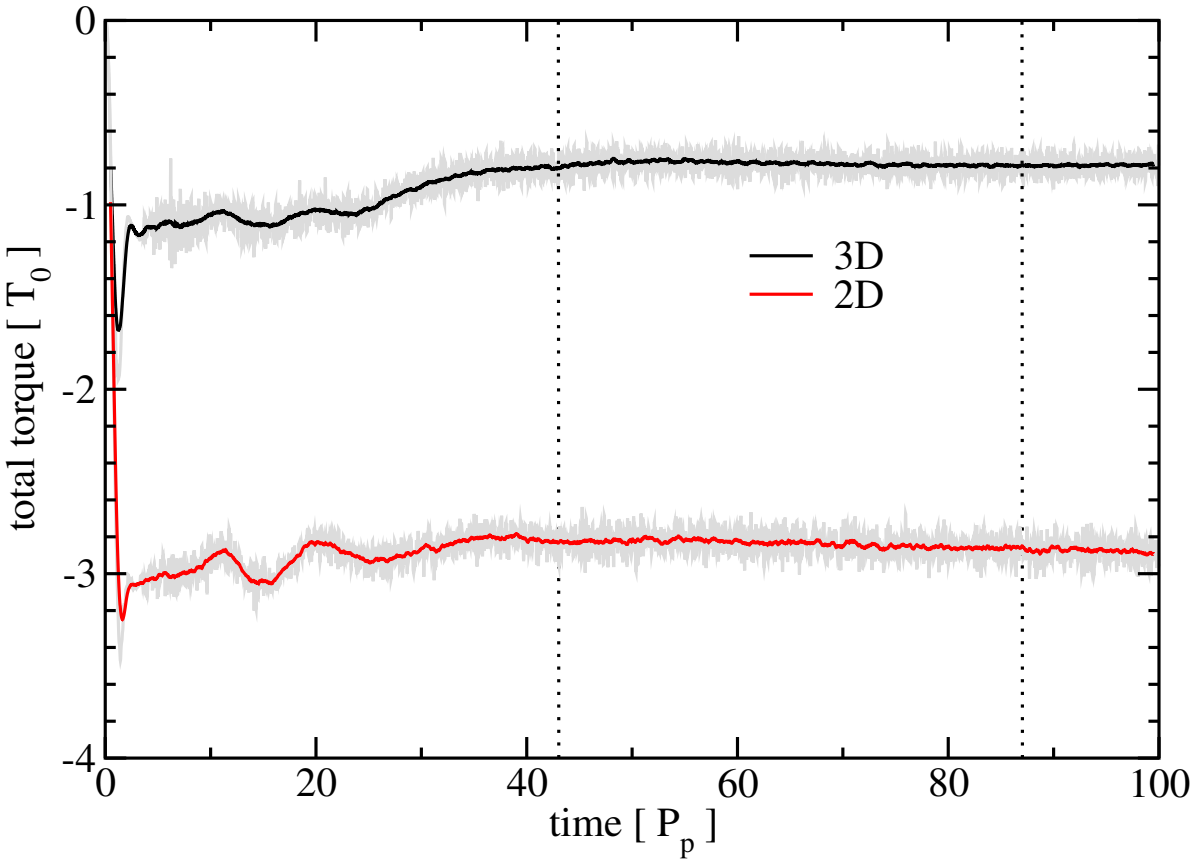

Figure 5.14 Net torque on the planet as a function of time. The black curve is our 3D torque measurement; red is 2D. This 2D case shares the same setup as the 3D one, except for $r_s = 0.3h_0$. Both curves are running-time-averages over $1P_p$. The instantaneous values of the the torques are shown as the gray shades around each curve. The vertical dotted lines mark the libration time of the horseshoe orbit. The first dotted line is at 1 $t_{lib} = 43\ P_p$, and second one is 2 $t_{lib} = 86\ P_p$.

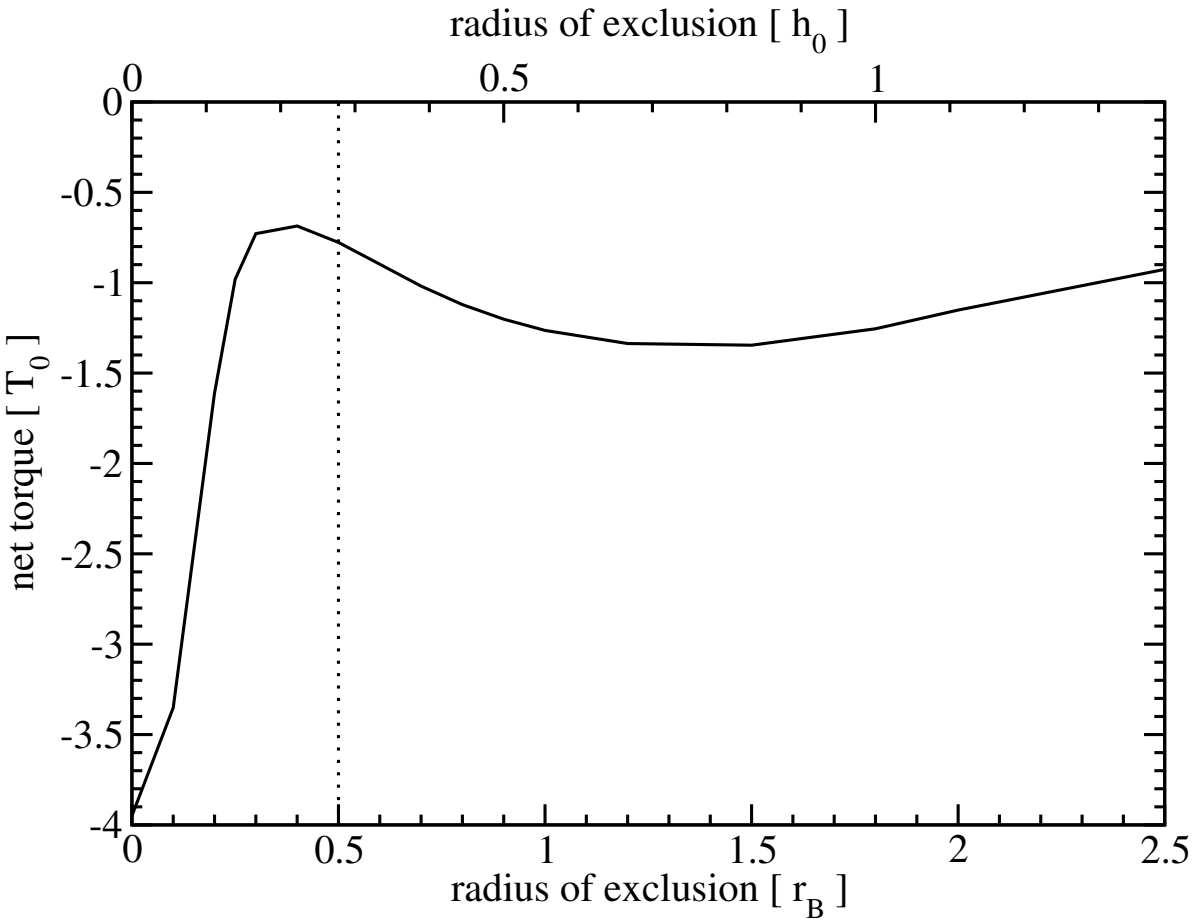

Figure 5.15 Net torque on the planet as a function of the radius of the excluded sphere centered on the planet, shown as the black solid curve. This plot, together with Figure 5.12, shows the non-physical four-armed spiral residing within $\sim 0.4 r_{\rm B}$ from the planet contributes a significant amount of torque that should be excluded from our calculation. The black dotted line labels the radius of exclusion we use, $0.5 r_{\rm B}$, corresponding to the black circle in Figure 5.12.

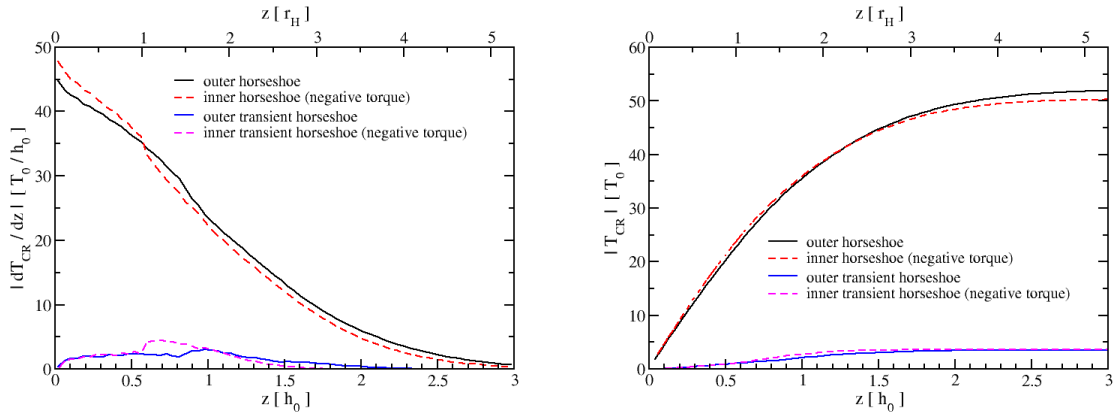

Figure 5.16 Magnitude of the differential corotation torque on the left panel, and magnitude of the cumulative corotation torque on the right, both as functions of height above the midplane. Black solid curve represents contribution from the outer horseshoe flow; red dashed curve corresponds to the inner one. Similarly, blue solid and magenta dashed curves are contributions from the outer and inner flows respectively, but are transient flows that exit the horseshoe region after one turn. Contributions from inner flows are negative in value. On the left panel, one can see that while the regular horseshoe flow provides the strongest torque near the midplane, the transient flow comes from an altitude of $\sim h_0$. On the right panel, it shows that overall the outer horseshoe flow generates a larger torque than inner. The sum of all 4 components is $1.5T_0$.

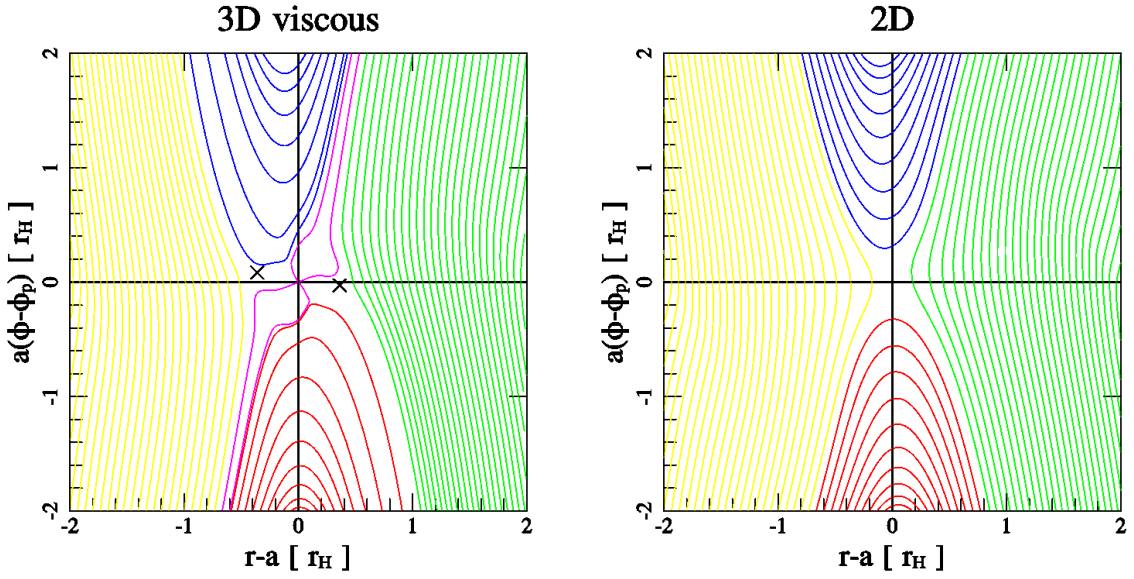

Figure 5.17 Midplane streamlines for the 3D viscous case on the left panel, and a 2D case on the right. This 2D case here is the same as the one in Figure 5.14. Comparing to Figure 5.8, the flow topology in the 3D viscous case is less asymmetric about $r = a$, and therefore is more similar to the 2D case on the right. The stagnation points, labeled as crosses on left, are located at $\{x, y\} = \{-0.36, 0.08\}$ and $\{0.36, 0.03\}$, in units of $r_{\rm H}$. In the 2D case, the large smoothing length ($r_{\rm s} = 0.3h_0$) results in the loss of both stagnation points. Like Figure 5.1, there is a stagnation point at the planet's location on both the left and right panels, which we omit to label.

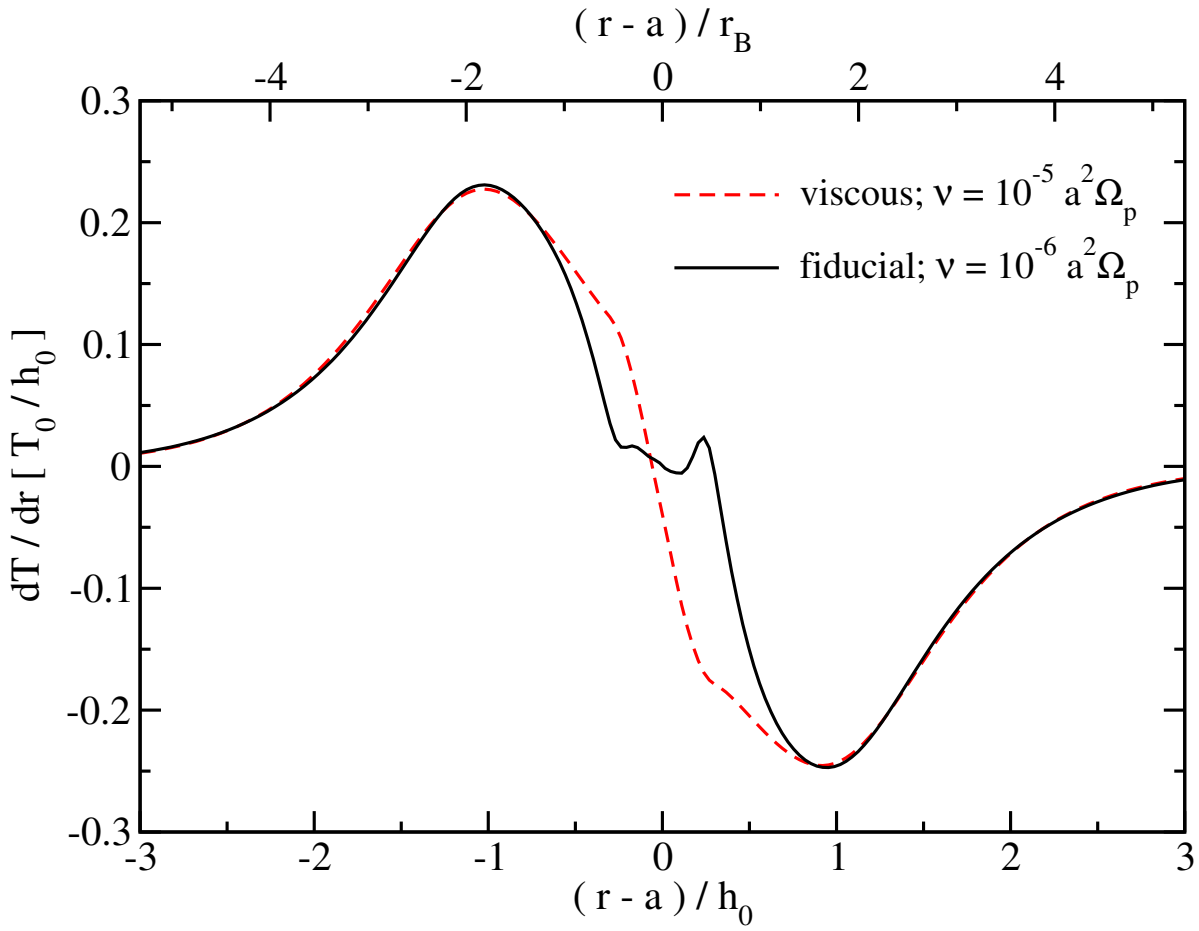

Figure 5.18 Torque distribution as a function r. The black solid curve is identical to the black curve in Figure 5.13. The red dashed curve is the torque distribution of our viscous case. Note the torque reversal near the planet does not exist for the red curve. This is consistent with Figure 5.17, where we see the stagnation points no longer have a large azimuthal offset.

3D flow field will become similar to 2D, and yield a similar torque on the planet as well.

D'Angelo & Bodenheimer (2013) performed 3D simulations of disk-planet interaction including radiative transfer and realistic opacity. They implemented a viscosity of $\nu = 4 \times 10^{-6} a^2 \Omega_p$, and an initial scale height[4] of $h_0 \sim 0.06a$ at the planet's location. If we assume $w \sim 1.2\, a\sqrt{aq/h_0}$, then their model for a $5M_\oplus$ planet is expected to have $t_\nu \sim 14P_p$, much shorter than $t_{lib} \sim 70P_p$. This places their model comparable to our viscous case. Comparing the left panel of our Figure 5.17 to their Figure 10, one can see that the flow patterns are largely similar, and neither of them show significant radial outflow at the midplane from within the Bondi sphere. This suggests that, in comparison to our fiducial case, the vertical inflow does not penetrate as deep into the Bondi sphere, which is consistent with the fact that the inflow speed is also much slower.

Finally, we note that $t_\nu > t_{lib}$ is not only a criterion for disk viscosity, but also sets a lower limit for the planet mass. Namely, it can be rewritten as:

$$q > \left(\frac{8\pi}{3}\right)^{\frac{2}{3}} \left(\frac{h_0}{a}\right) \left(\frac{\nu}{a^2\Omega_p}\right)^{\frac{2}{3}}, \tag{5.32}$$

if we approximate $w \approx a\sqrt{aq/h_0}$. Together with the fact that 3D effects are most relevant for embedded planets: $r_H < h_0$, which can be rewritten as:

$$q < 3\left(\frac{h_0}{a}\right)^3, \tag{5.33}$$

these two criteria bracket the range of planetary mass where we expect the 3D flow field to deviate most from 2D.

5.6 Discussion and Conclusion

We present a detailed picture of the 3D flow topology near an embedded planet on a fixed circular orbit, extracted from 3D hydrodynamical simulations of a $\sim 5M_\oplus$ planet interacting with a circumstellar disk. Our simulations are run with our GPU hydrodynamical code `PEnGUIn`, on a single desktop computer equipped with 3 GTX-Titan graphics cards. We found that the 3D modifications to the horseshoe flow have a significant influence on both the density structure in the planet's Bondi sphere, and the torque exerted on the planet. Below we give a summary of the 3D horseshoe flow:

(1) At the onset of a horseshoe turn, before the close encounter with the planet, the flow is columnar. This results in a nearly constant horseshoe half-width w in z (Figure 5.3).

(2) While a fraction of the horseshoe flow continues in columnar form after the turn, the widest portion is pulled toward the midplane and fall directly on top of the planet (Figure 5.5). This flow plummets deep into the planet's Bondi sphere (Figure 5.9).

(3) The release of potential energy from the fall results in the flow exiting the Bondi sphere near the midplane at a speed of order c_s (Appendix B). Symmetry of the horseshoe streamline about $r = a$ is broken (Figure 5.8). Consequently the widest horseshoe flow will over-shoot the horseshoe region, and exit after just one horseshoe turn. We call this the "transient" horseshoe flow.

(4) As the transient flow pushes into the disk, it is deflected by midplane material, resulting in a vortex roll-up (Figure 5.11). At the same time, the loss of material in the horseshoe region due to the transient flow is

[4]There is some ambiguity in the definition of a scale height in their work, because of the non-trivial temperature profile in their radiation-hydrodynamics treatment. See their Section 4 for details.

replenished by the high altitude flow lying just outside of the region Figure 5.10. This generates a meridional circulation that mixes the flow (Figure 5.10).

(5) The meridional circulation (or the ϕ-direction vortex) is eventually killed by disk viscosity, and the flow resets to the columnar flow in (1) before the next encounter with the planet.

The flow speed inside the Bondi sphere approaches c_s, so the gas density there is much less dense than if it has a hydrostatic structure (Figure 5.7). Nearly all of the gas in the Bondi sphere participates in the horseshoe flow. We found that only gas within a distance of $\sim r_s$ is bound to the planet.

We also found that as a part of the asymmetry in the flow pattern across $r = a$, the stagnation points are now offset from the azimuth of the planet, ϕ_p, where the inner point now lies below ϕ_p and the outer point above. The flow pattern asymmetry corresponds to an imbalance in the inner and outer horseshoe flow, which generates a net corotation torque of $\sim 1.5T_0$. Overall, this results in a net torque of $\sim -0.77T_0$, much reduced from $-2.19T_0$ predicted by linear calculations.

5.6.1 Forming Gaseous Planets

Following OSK15 to measure the flux of mass in and out of the Bondi sphere by $t_{\rm replenish}$, we found $t_{\rm replenish} \sim \Omega_p^{-1}$, which is alarmingly short if the Bondi sphere is the planet's atmosphere. Through streamline analysis, we found that nearly all streamlines within the Bondi sphere are not bound to the planet. It is then more appropriate to classify the Bondi sphere by its flow topology: a part of the transient horseshoe flow. However, this leaves us with little atmosphere. If this is universally true for all planetary cores, then gaseous planets with cores of a few $M_\oplus$, cannot be formed.

There are two major issues with our result in the context of gas accretion. First, it should be reminded that we did find the gas within $1.5r_s$ of the planet to be bound, but it is only resolved by 3 grid cells. Increasing resolution in this region will allow us to be more certain about how much gas is truly bound to the planet. The resolution required to fully resolve the planetary atmosphere is of order the pressure scale height on the surface of the planet: $c_s^2 r_s^2/(qGM_*) \sim 0.1r_s$ for our setup. This kind of resolution is attainable with local simulations around the planet.

Second, and more importantly, our simulation is globally isothermal, which is unrealistic since it does not take into account the heating and cooling of the atmosphere. A planet's atmosphere is expected to be heated through the accretion of gas and planetesimals, and the timescale for it to cool from that heat is much longer than $t_{\rm replenish}$ measured from our simulation; therefore the planet's atmosphere should be more appropriately described as adiabatic rather than isothermal. A more heated and pressurized atmosphere may deflect the transient horseshoe flow and prevent it from intruding into the Bondi sphere, allowing more gas to be bound to the planet. This may have already been observed in the 3D radiation-hydrodynamics simulations by D'Angelo & Bodenheimer (2013), which showed that the planet has a bound atmosphere of the size of its Bondi sphere; however, their relatively large disk viscosity (see Section 5.5) is expected to slow down vertical inflow speed and weaken 3D effects. It remains to be seen how large the bound atmosphere is in an inviscid flow with realistic radiative transfer.

5.6.2 Stopping Type I Migration

We have shown that our 3D disk produces a net corotation torque that is not expected in 2D analysis. As a result, our planet, embedded in a 3D disk, migrates 3 times slower than if it is driven by the differential Lindblad torque alone. This result is subject to a number of uncertainties.

First, a major limitation to the accuracy of our torque measurement is the poor numerical accuracy near the planet (see Figure 5.12). Because of this we have to exclude the $r = 0.5r_B$ sphere around the planet from torque calculation. Ormel et al. (2015a) showed that numerical convergence can be more efficiently achieved if the grid geometry around the planet is polar rather than Cartesian. This is challenging to implement in a global simulation, because the grid should, ideally, also be polar around the star-planet's center of mass. Therefore we did not attempt it. Nevertheless, we believe our torque measurements do capture the essence of 3D effects, because our region of exclusion is sufficiently small that it does not cover the stagnation points, and we used an independent method to measure the corotation torque which gave a consistent result.

Second, we have chosen a disk profile that minimizes the net corotation torque in order to more easily identify differences between 2D and 3D. We have seen in Section 5.4 that the one-sided corotation torque has a magnitude that can overwhelm the Lindblad torque, so an imbalance between the inner and outer horseshoe flow can potentially dominate type I migration rate. This can be accomplished by a modification as simple as changing the disk density or sound speed profile. A future study on how the net 3D corotation torque depends on disk parameters will be valuable to understanding planet migration.

Third, we have not considered thermal physics in our model, which has been shown to be capable of reversing type I migration (e.g. Bitsch & Kley, 2011; Bitsch et al., 2014; Lega et al., 2014; Benítez-Llambay et al., 2015). Before considering the full radiative transfer problem, a possible first step will be to relax our isothermal condition to an adiabatic equation of state. 2D results (Masset & Casoli, 2009; Paardekooper et al., 2010) have shown that an additional corotation torque related to the enthalpy of the fluid is expected for an adiabatic disk. The 3D aspect of this should be studied in more detail.

Fourth, our planet has a fixed circular orbit, despite the fact we are measuring a non-zero net torque on it. From our results , we can speculate that when the planet migrates inward (outward), disk material may be transported from the inner (outer) to the outer (inner) disk via the transient horseshoe flow, which may further unbalance the inner and outer flow and generate a larger net torque on the planet, leading to type III migration (Papaloizou et al., 2007). The 3D dynamical interaction between the planet and the disk has the potential to modify calculations based on fixed planets by a margin as large as the one-sided corotation torque.

Finally, we note that our planet has a relatively large mass comparing to planets that typically fall in the type I regime, which are $\lesssim 1\ M_\oplus$. This raises the question of how well our torque measurement applies to type I migration in general. We believe the horseshoe flow asymmetry that causes the stagnation point offset is applicable to lower mass planets, because the flow pattern we presented shares many similarities with OSK15's (compare our Figure 5.8 to their Figure 3), who simulated the inviscid flow around a planet with $q_{\rm th} = 0.01$. Therefore, lower mass planets should also experience a reduced migration rate like ours. However, it is unclear whether the magnitude of the reduction will be similar. This question will be answered when global inviscid 3D simulations of smaller planets become feasible.

The ability to simulate smaller planets in 3D is very dependent on computational resources. If we attempt to simulate a 1 $M_\oplus$ planet, then in order to have the same resolution across r_B as we do in this chapter, cell sizes would need to decrease by a factor of 5, and the computational time scales as $5^{3+1} = 625$, where the power of 3 comes from the increase in the total number of cells, and 1 from the shortened timestep due to the Courant limit. This amount of computational resources is much beyond what is currently available to us, but in the future, advancements in both hardware, such as access to a large GPU cluster, and software, such as a more sophisticated grid design, may open this pathway for us.

5.6.3 Torque and Viscosity

In Section 5.5 we suggested to use the condition $t_\nu > t_{\rm lib}$ to identify where 3D effects become important to the flow topology and planetary torque. This is also the condition for torque saturation in 2D (Ward, 1991). Since we do measure a non-zero net corotation torque, it is important to ask whether this torque will saturate. Evidently, Figure 5.14 does not show any sign of torque saturation.

The process of torque saturation can be described as follows: because horseshoe orbits are closed in 2D, the entire horseshoe region is completely separated from the rest of the disk, and therefore has only a finite amount of angular momentum to exchange with the planet. Consequently, the net exchange of angular momentum between the horseshoe region and the planet in steady state must be zero. The corotation torque can only be unsaturated if fresh supply of angular momentum enters the horseshoe region, which can be done through viscous diffusion, hence the torque saturation condition. In Section 5.3.3 we showed that there is a constant exchange of material between the horseshoe region and the rest of the disk due to the existence of the transient horseshoe flow. The planet is therefore able to replenish the horseshoe flow without help from disk viscosity. In other words, when $t_\nu > t_{\rm lib}$, 3D effects kick in, and the corotation torque is unsaturated; when $t_\nu < t_{\rm lib}$, disk viscosity dominates, and it is also unsaturated. So corotation torque saturation may be a pure 2D effect. For a future study, it would be interesting to test a range of disk profiles and viscosity in 3D, and find out under what circumstance does torque saturation occur.

Appendix A

Numerical Method for Solving the Linearized IRI Equations

In this Appendix, we document our method for solving Equation 3.16 numerically. To begin, note that Equation 3.16 can only be numerically integrated in the direction of increasing r because of the integral in the fourth term. In principle, it is possible to simply do this integration and find the value of ω that best matches the desired outer boundary condition. This is impractical, however, because any slight error in ω leads to a diverging behavior of η_m at the outer boundary. A better method is to integrate Equation 3.16 simultaneously from the inner boundary outward, and the outer boundary inward, and find the ω that results in a match of the two functions at some intermediate radius $r_{\rm mid}$. To accomplish this, we first define

$$y_m \equiv \int_0^r \frac{\eta_m}{c_s^2}\frac{\mathrm{d}\tau}{\mathrm{d}r'}\mathrm{d}r' \,, \tag{A.1}$$

and then differentiate Equation 3.16 with respect to r:

$$\frac{\partial^3 y_m}{\partial r^3} + a'(r)\frac{\partial^2 y_m}{\partial r^2} + b'(r)\frac{\partial y_m}{\partial r} + c'(r)y_m = 0 \,, \tag{A.2}$$

where

$$\begin{aligned}
a' &\equiv a - 2\frac{\mathrm{d}\ln g}{\mathrm{d}r}, \\
b' &\equiv b - a\frac{\mathrm{d}\ln g}{\mathrm{d}r} + 2\left(\frac{\mathrm{d}\ln g}{\mathrm{d}r}\right)^2 - \frac{1}{g}\frac{\mathrm{d}^2 g}{\mathrm{d}r^2}, \\
c' &\equiv cg, \\
g &\equiv \frac{1}{c_s^2}\frac{\mathrm{d}\tau}{\mathrm{d}r}.
\end{aligned}$$

Thus we can now numerically integrate Equation A.2 in both directions, and recover η_m from y_m. The bound-

ary conditions can be approximated using the WKB method. The WKB form for y_m is

$$y_m = R(r)e^{i\int_0^r k\mathrm{d}r'} , \tag{A.3}$$

$$\frac{\partial y_m}{\partial r} \simeq iky_m , \tag{A.4}$$

where $R(r)$ is a slowly varying function and k is the complex wave number that satisfies $|kr| \gg 1$. Substituting Equation A.3 and Equation A.4 into Equation A.2 we get the following algebraic equation for k:

$$k^3 - ia'k^2 - b'k + ic' = 0 . \tag{A.5}$$

The three solutions of Equation A.5 correspond to the inward traveling ($\mathrm{Re}(k) < 0$), outward traveling ($\mathrm{Re}(k) > 0$), and a third solution that does not exist in the conventional WKB approximation. In fact, it has $|kr| \ll 1$, effectively rendering y_m a constant, which violates the approximation of a tightly winding wave. To accommodate for this solution, we generalize Equation A.3 to allow for a constant offset:

$$y_m = R(r)e^{i\int_0^r k\mathrm{d}r'} + C . \tag{A.6}$$

Substituting this into Equation A.2, we obtain:

$$k^3 - ia'k^2 - b'k + ic'\left(\frac{y_m}{y_m - C}\right) = 0 . \tag{A.7}$$

In the optically thin and thick limits, c' becomes arbitrarily small, and since the $|kr| \ll 1$ solution is already incorporated into the constant offset C, the last term can be dropped, giving back the usual quadratic form:

$$k^2 - ia'k - b' = 0, \tag{A.8}$$

which gives the expected incoming and outgoing solutions for tightly winding waves. We apply the radiative boundary condition, assuming no wave is entering the domain from the boundaries. The other unknowns remaining in Equation A.6 are R and C. For clarity, we will denote variables associated with the solution integrated from the inner boundary with the subscript "in", and the those from the outer boundary with "out".

Recall that y_m is in fact the integral of the perturbation (Equation A.1). At the inner boundary, this quantity is small since inward of the boundary there is only a traveling wave, so we set $C_{\mathrm{in}} = 0$. We choose $R_{\mathrm{in}} = 1$, while R_{out} and C_{out} are determined by the following iterative formulas:

$$R^{i+1}_{\mathrm{out}} = R^i_{\mathrm{out}} \frac{\mathrm{d}^2 y^i_{m,\mathrm{in}}}{\mathrm{d}r^2} \left(\frac{\mathrm{d}^2 y^i_{m,\mathrm{out}}}{\mathrm{d}r^2}\right)^{-1} , \tag{A.9}$$

$$C^{i+1}_{\mathrm{out}} = C^i_{\mathrm{out}} + y^i_{m,\mathrm{in}} - y^i_{m,\mathrm{out}}, \tag{A.10}$$

where i is the current iterative step, the y_m and its derivatives are evaluated at r_{mid}. Convergence typically requires tens or even hundreds of iterations, which is the primary reason for the large amount of computational time required for this method. Lastly, we find the eigenvalue ω by minimizing the following function, evaluated at r_{mid}:

$$f = \left(\frac{\mathrm{Re}\left(\frac{\mathrm{d}y_{m,\mathrm{out}}}{\mathrm{d}r}\right) - \mathrm{Re}\left(\frac{\mathrm{d}y_{m,\mathrm{in}}}{\mathrm{d}r}\right)}{\max\left[\left|\mathrm{Re}\left(\frac{\mathrm{d}y_{m,\mathrm{out}}}{\mathrm{d}r}\right)\right|, \left|\mathrm{Re}\left(\frac{\mathrm{d}y_{m,\mathrm{in}}}{\mathrm{d}r}\right)\right|\right]}\right)^2 + \left(\frac{\mathrm{Im}\left(\frac{\mathrm{d}y_{m,\mathrm{out}}}{\mathrm{d}r}\right) - \mathrm{Im}\left(\frac{\mathrm{d}y_{m,\mathrm{in}}}{\mathrm{d}r}\right)}{\max\left[\left|\mathrm{Im}\left(\frac{\mathrm{d}y_{m,\mathrm{out}}}{\mathrm{d}r}\right)\right|, \left|\mathrm{Im}\left(\frac{\mathrm{d}y_{m,\mathrm{in}}}{\mathrm{d}r}\right)\right|\right]}\right)^2 . \tag{A.11}$$

We use an eighth-order Runge-Kutta method with adaptive step-size control for the numerical integration. We set $r_{\rm mid} = 1$, the inner boundary at $r_{\rm in} = 0.3$, and the outer at $r_{\rm out} = 4$. Minimizing f is also very time consuming because we employ a random sampling method: first we bracket the minimum within a range of likely values for the real and imaginary part of ω, then we randomly select ω within the chosen range, and narrow down the field by preferentially choosing values closer to where f is below a certain threshold. This time consuming method is ultimately superior to methods that involve descending along the gradient of f, because of the numerous local minima that exist.

Appendix B

Radial Flow Speed in the Transient Horseshoe Flow

In this appendix we analytically calculate the radial outflow speed at which the transient horseshoe flow exits the horseshoe region. Bernoulli's constant for a given streamline can be written as:

$$B = -\frac{1}{2}r^2\Omega_{\rm p}^2 + \Phi + \frac{1}{2}|\boldsymbol{u}|^2 + \eta\,. \tag{B.1}$$

We divide η into two components: $\eta = \eta_0 + \eta_{\rm p}$, where η_0 is the background enthalpy profile that balances the star's potential:

$$\frac{{\rm d}\eta_0}{{\rm d}z} = -\frac{\rm d}{{\rm d}z}\frac{GM_*}{\sqrt{r^2+z^2}}\,. \tag{B.2}$$

For $x = r - a$ where $x \ll a$, we can rewrite Equation B.1 as:

$$B = -\frac{3}{2}a^2\Omega_{\rm p}^2 - \frac{3}{2}x^2\Omega_{\rm p}^2 - q\frac{GM_*}{\sqrt{x^2+y^2+z^2}} + \frac{1}{2}|\boldsymbol{u}|^2 + \eta_{\rm p}\,, \tag{B.3}$$

where $\{x, y, z\}$ are the local Cartesian coordinates centering on the planet. In Equation B.3, η_0 has been canceled by the z-dependent part of the star's potential. We will also drop $\eta_{\rm p}$ because it is expected to have a small contribution to Equation B.3 comparing to the planet's gravitational potential, because as shown in Figure 5.7, the density, hence pressure, near the planet is much less than what is needed to balance the planet's gravity.

Consider two points along an inner horseshoe orbit: the first point is just before the fluid is about to perform its horseshoe turn, located at $\boldsymbol{p}_1 = \{x_1, y_1, z_1\}$, where $x_1 < 0$, and its velocity can be approximated as $\boldsymbol{u} = \{u_{\rm r,1,2}, \frac{3}{2}x_1\Omega_{\rm p}, 0\}$, where $\frac{3}{2}x_1\Omega_{\rm p}$ is the local Keplerian shear; the second point is after the turn, having been pulled down to the midplane, at $\boldsymbol{p}_2 = \{x_2, y_2, 0\}$, where $x_2 = -x_1$ and $y_1 = y_2$, and has $\boldsymbol{u} = \{u_{\rm r,2}, \frac{3}{2}x_2\Omega_{\rm p}, 0\}$. For convenience we define $d^2 = x_1^2 + y_1^2 = x_2^2 + y_2^2$ as the distance from the planet on the $x - y$ plane. Equating B at these two points, we have:

$$-q\frac{GM_*}{\sqrt{d^2+z^2}} + \frac{1}{2}u_{\rm r,1}^2 = -q\frac{GM_*}{d} + \frac{1}{2}u_{\rm r,2}^2\,. \tag{B.4}$$

By our results in Section 5.3.1, it is safe to assume all of the widest horseshoe orbits has the same x_1 and y_1

irrespective of their starting z. This allows us to approximate $z^2 \approx h_0^2$ in Equation B.4 since:

$$\frac{\int_0^\infty z^2 \rho(z=0) e^{-\frac{z^2}{2h^2}} \mathrm{d}z}{\int_0^\infty \rho(z=0) e^{-\frac{z^2}{2h^2}} \mathrm{d}z} = h^2 \,, \tag{B.5}$$

Finally, combining Equation B.4 and Equation B.5, we have:

$$|u_{\mathrm{r},2}| = \sqrt{2q\frac{GM}{d}\left(1 - \frac{d}{\sqrt{d^2+h_0^2}}\right) + u_{\mathrm{r},1}^2} \,. \tag{B.6}$$

Since $|u_{\mathrm{r},2}| > |u_{\mathrm{r},1}|$, the horseshoe streamline is asymmetric about $x = 0$. Equation B.6 should be considered an upper limit for two main reasons: first, the streamlines in fact only fall to a fraction of h_0 instead the midplane; and second, an increase in enthalpy as a fluid element moves towards the midplane can compensate for some loss in gravitational potential energy. In fact if η satisfies Equation 5.24, it would leave $|u_{\mathrm{r},1}| = |u_{\mathrm{r},2}|$, resulting in a symmetric horseshoe flow. Equation B.6 is useful however, for allowing us to estimate how this outflow speed scales with planet mass.

d should scale with the horseshoe half-width w, which can have two possible scalings: $w \propto a\sqrt{q(a/h_0)}$ for low mass planets (Masset et al., 2006; Paardekooper & Papaloizou, 2009b), and $w \propto r_{\mathrm{H}}$ for high mass planets (Masset et al., 2006; Peplinski, 2008). In the low mass limit, if we approximate $d \sim a\sqrt{q(a/h_0)}$, and further simplify Equation B.6 using $d \ll h$ (equivalently, $q_{\mathrm{th}} \ll 1$), then we get:

$$|u_{\mathrm{r},2}| \simeq \sqrt{2\sqrt{q_{\mathrm{th}}}c_{\mathrm{s}}^2 + |u_{\mathrm{r},1}|^2} \,. \tag{B.7}$$

In the high mass limit, we approximate $d \sim 2r_{\mathrm{H}}$, and have $r_{\mathrm{H}} \gg h$ (equivalently, $q_{\mathrm{th}} \gg 1$):

$$|u_{\mathrm{r},2}| \simeq \sqrt{\frac{3}{8}c_{\mathrm{s}}^2 + |u_{\mathrm{r},1}|^2} \,. \tag{B.8}$$

These two limits show that the injection of energy due to the change in z is smaller if planet mass decreases, and asymptotes to a constant value for a high mass planet. In our simulation, the radial velocity measured at $|x| = r_{\mathrm{H}}$ ranges from 0.2 to $0.4c_{\mathrm{s}}$, depending on the azimuth at which it is measured (see Figures 5.9 and 5.11). For comparison, Equation B.8 predicts a velocity of $\sim 0.6c_{\mathrm{s}}$ if $|u_{\mathrm{r},1}| \ll c_{\mathrm{s}}$, which is within an order of unity from our numerical result.

Bibliography

Andrews, S. M., & Williams, J. P. 2007, ApJ, 659, 705

Andrews, S. M., Wilner, D. J., Espaillat, C., Hughes, A. M., Dullemond, C. P., McClure, M. K., Qi, C., & Brown, J. M. 2011, ApJ, 732, 42

Andrews, S. M., Wilner, D. J., Hughes, A. M., Qi, C., & Dullemond, C. P. 2009, ApJ, 700, 1502

Artymowicz, P. 1993a, ApJ, 419, 166

—. 1993b, ApJ, 419, 155

Ayliffe, B. A., & Bate, M. R. 2012, MNRAS, 427, 2597

Balbus, S. A., & Hawley, J. F. 1991, ApJ, 376, 214

Batalha, N. M., et al. 2013, ApJS, 204, 24

Benítez-Llambay, P., Masset, F., Koenigsberger, G., & Szulágyi, J. 2015, Nature, 520, 63

Bitsch, B., & Kley, W. 2011, A&A, 530, A41

Bitsch, B., Morbidelli, A., Lega, E., & Crida, A. 2014, A&A, 564, A135

Blondin, J. M., & Lufkin, E. A. 1993, ApJS, 88, 589

Borucki, W. J., et al. 2010, Science, 327, 977

Calvet, N., et al. 2005, ApJL, 630, L185

Casoli, J., & Masset, F. S. 2009, ApJ, 703, 845

Chiang, E., & Murray-Clay, R. 2007, Nature Physics, 3, 604

Chiang, E. I., & Goldreich, P. 1997, ApJ, 490, 368

Colella, P., & Woodward, P. R. 1984, Journal of Computational Physics, 54, 174

Crida, A., Morbidelli, A., & Masset, F. 2006, Icarus, 181, 587

D'Angelo, G., & Bodenheimer, P. 2013, ApJ, 778, 77

D'Angelo, G., Kley, W., & Henning, T. 2003, ApJ, 586, 540

D'Angelo, G., & Lubow, S. H. 2010, ApJ, 724, 730

D'Angelo, G., Lubow, S. H., & Bate, M. R. 2006, ApJ, 652, 1698

de Val-Borro, M., Artymowicz, P., D'Angelo, G., & Peplinski, A. 2007, A&A, 471, 1043

de Val-Borro, M., et al. 2006, MNRAS, 370, 529

Debes, J. H., Jang-Condell, H., Weinberger, A. J., Roberge, A., & Schneider, G. 2013, ApJ, 771, 45

Dodson-Robinson, S. E., & Salyk, C. 2011, ApJ, 738, 131

Dominik, C., & Dullemond, C. P. 2011, A&A, 531, A101

Dong, R., et al. 2012, ApJ, 760, 111

Duffell, P. C., & MacFadyen, A. I. 2013, ApJ, 769, 41

Dunhill, A. C., Alexander, R. D., & Armitage, P. J. 2013, MNRAS, 428, 3072

Espaillat, C., Furlan, E., D'Alessio, P., Sargent, B., Nagel, E., Calvet, N., Watson, D. M., & Muzerolle, J. 2011, ApJ, 728, 49

Espaillat, C., et al. 2007, ApJL, 664, L111

—. 2008, ApJL, 689, L145

Flaherty, K. M., & Muzerolle, J. 2010, ApJ, 719, 1733

Flaherty, K. M., Muzerolle, J., Rieke, G., Gutermuth, R., Balog, Z., Herbst, W., Megeath, S. T., & Kun, M. 2011, ApJ, 732, 83

Flohic, H. M. L. G., & Eracleous, M. 2008, ApJ, 686, 138

Frank, J., King, A., & Raine, D. 2002, Accretion Power in Astrophysics (Cambridge Astrophysics) (Cambridge University Press)

Gammie, C. F. 2001, ApJ, 553, 174

Goldreich, P., Goodman, J., & Narayan, R. 1986, MNRAS, 221, 339

Goldreich, P., & Tremaine, S. 1979, ApJ, 233, 857

—. 1980, ApJ, 241, 425

Goodman, J., & Rafikov, R. R. 2001, ApJ, 552, 793

Haisch, Jr., K. E., Lada, E. A., & Lada, C. J. 2001, ApJL, 553, L153

Hawley, J. F., & Stone, J. M. 1995, Computer Physics Communications, 89, 127

Hayashi, C. 1981, Progress of Theoretical Physics Supplement, 70, 35

Higginbottom, N., Proga, D., Knigge, C., Long, K. S., Matthews, J. H., & Sim, S. A. 2014, ArXiv e-prints

Hopkins, P. F., & Quataert, E. 2011, MNRAS, 415, 1027

Ikoma, M., Nakazawa, K., & Emori, H. 2000, ApJ, 537, 1013

Kirchschlager, F., & Wolf, S. 2013, A&A, 552, A54

Kitamura, Y., Momose, M., Yokogawa, S., Kawabe, R., Tamura, M., & Ida, S. 2002, ApJ, 581, 357

Kley, W. 1998, A&A, 338, L37

Kley, W., & Dirksen, G. 2006, A&A, 447, 369

Kley, W., & Nelson, R. P. 2012, Annual Review of Astronomy and Astrophysics, 50, 211

Kraus, A. L., & Ireland, M. J. 2012, ApJ, 745, 5

Krauss, O., & Wurm, G. 2005, ApJ, 630, 1088

Lee, E. J., Chiang, E., & Ormel, C. W. 2014, ArXiv e-prints

Lega, E., Crida, A., Bitsch, B., & Morbidelli, A. 2014, MNRAS, 440, 683

Li, H., Colgate, S. A., Wendroff, B., & Liska, R. 2001, ApJ, 551, 874

Li, H., Finn, J. M., Lovelace, R. V. E., & Colgate, S. A. 2000, ApJ, 533, 1023

Li, H., Lubow, S. H., Li, S., & Lin, D. N. C. 2009, ApJL, 690, L52

Lin, D. N. C., & Pringle, J. E. 1987, MNRAS, 225, 607

Lin, M.-K. 2013, ApJ, 765, 84

Lovelace, R. V. E., Li, H., Colgate, S. A., & Nelson, A. F. 1999, ApJ, 513, 805

Lubow, S. H. 1991, ApJ, 381, 259

Lubow, S. H., & D'Angelo, G. 2006, ApJ, 641, 526

Lubow, S. H., Seibert, M., & Artymowicz, P. 1999, ApJ, 526, 1001

Masset, F. S., & Casoli, J. 2009, ApJ, 703, 857

—. 2010, ApJ, 723, 1393

Masset, F. S., D'Angelo, G., & Kley, W. 2006, ApJ, 652, 730

Masset, F. S., & Ogilvie, G. I. 2004, ApJ, 615, 1000

Meheut, H., Meliani, Z., Varniere, P., & Benz, W. 2012, A&A, 545, A134

Morbidelli, A., Szulágyi, J., Crida, A., Lega, E., Bitsch, B., Tanigawa, T., & Kanagawa, K. 2014, Icarus, 232, 266

Mulders, G. D., Paardekooper, S.-J., Panić, O., Dominik, C., van Boekel, R., & Ratzka, T. 2013, A&A, 557, A68

Müller, T. W. A., Kley, W., & Meru, F. 2012, A&A, 541, A123

Muzerolle, J., et al. 2009, ApJL, 704, L15

Nelson, R. P., Gressel, O., & Umurhan, O. M. 2013, MNRAS, 435, 2610

Ormel, C. W. 2013, MNRAS, 428, 3526

Ormel, C. W., Kuiper, R., & Shi, J.-M. 2015a, MNRAS, 446, 1026

Ormel, C. W., Shi, J.-M., & Kuiper, R. 2015b, MNRAS, 447, 3512

Owen, J. E., Ercolano, B., Clarke, C. J., & Alexander, R. D. 2010, MNRAS, 401, 1415

Owen, J. E., Hudoba de Badyn, M., Clarke, C. J., & Robins, L. 2013, MNRAS, 436, 1430

Paardekooper, S.-J., Baruteau, C., Crida, A., & Kley, W. 2010, MNRAS, 401, 1950

Paardekooper, S.-J., Baruteau, C., & Kley, W. 2011, MNRAS, 410, 293

Paardekooper, S.-J., & Papaloizou, J. C. B. 2008, A&A, 485, 877

—. 2009a, MNRAS, 394, 2283

—. 2009b, MNRAS, 394, 2297

Papaloizou, J. C. B., Nelson, R. P., Kley, W., Masset, F. S., & Artymowicz, P. 2007, Protostars and Planets V, 655

Papaloizou, J. C. B., Nelson, R. P., & Masset, F. 2001, A&A, 366, 263

Papaloizou, J. C. B., & Pringle, J. E. 1984, MNRAS, 208, 721

—. 1985, MNRAS, 213, 799

—. 1987, MNRAS, 225, 267

Peplinski, A. 2008, PhD thesis, University of Stockholm

Petigura, E. A., Marcy, G. W., & Howard, A. W. 2013, ApJ, 770, 69

Pollack, J. B., Hubickyj, O., Bodenheimer, P., Lissauer, J. J., Podolak, M., & Greenzweig, Y. 1996, Icarus, 124, 62

Quanz, S. P., Avenhaus, H., Buenzli, E., Garufi, A., Schmid, H. M., & Wolf, S. 2013, ApJL, 766, L2

Rafikov, R. R. 2006, ApJ, 648, 666

Semenov, D., Henning, T., Helling, C., Ilgner, M., & Sedlmayr, E. 2003, A&A, 410, 611

Shakura, N. I., & Sunyaev, R. A. 1973, A&A, 24, 337

Shi, J.-M., Krolik, J. H., Lubow, S. H., & Hawley, J. F. 2012, The Astrophysical Journal, 749, 118

Sod, G. A. 1978, Journal of Computational Physics, 27, 1

Sorathia, K. A., Krolik, J. H., & Hawley, J. F. 2013, The Astrophysical Journal, 768, 133

Stone, J. M., & Norman, M. L. 1992, ApJS, 80, 753

Takeuchi, T., & Artymowicz, P. 2001, ApJ, 557, 990

Tanaka, H., Takeuchi, T., & Ward, W. R. 2002, ApJ, 565, 1257

Tanigawa, T., Ohtsuki, K., & Machida, M. N. 2012, ApJ, 747, 47

Thebault, P., Kral, Q., & Augereau, J.-C. 2014, A&A, 561, A16

Toro, E. F. 2009, Riemann solvers and numerical methods for fluid dynamics : a practical introduction (New York: Springer)

Urpin, V., & Brandenburg, A. 1998, MNRAS, 294, 399

Ward, W. R. 1991, in Lunar and Planetary Science Conference, Vol. 22, Lunar and Planetary Science Conference, 1463

Ward, W. R. 1997, ApJL, 482, L211

Yu, C., Li, H., Li, S., Lubow, S. H., & Lin, D. N. C. 2010, ApJ, 712, 198

Zhu, Z., Nelson, R. P., Dong, R., Espaillat, C., & Hartmann, L. 2012, ApJ, 755, 6

Zhu, Z., Nelson, R. P., Hartmann, L., Espaillat, C., & Calvet, N. 2011, ApJ, 729, 47