On some aspects of mathematical and computational models for simulations of granular materials

Gabriel Nóbrega Bufolo

Orientador: Yuri Dumaresq Sobral

Universidade de Brasília

Essa dissertação é submetida para o grau de
Doutor em Matemática

Abril 2023

Universidade de Brasília

Instituto de Ciências Exatas

Departamento de Matemática

On some aspects of mathematical and computational models for simulations of granular materials

por

Gabriel Nóbrega Bufolo*

Tese apresentada ao Departamento de Matemática da Universidade de Brasília, como parte dos requisitos para obtenção do grau de

DOUTOR EM MATEMÁTICA

Brasília, 04 de maio de 2023.

Comissão Examinadora:

Yuri Dumaresq Sobral

Prof. Dr. Yuri Dumaresq Sobral - MAT/UnB (Orientador)

Prof. Dr. Taygoara Felamingo de Oliveira – ENM/UnB (Membro)

Cassio M. Oishi

Prof. Dr. Cassio Machiaveli Oishi – UNESP (Membro)

John Hin

Prof. Dr. Edward John Hinch – University of Cambridge (Membro)

* O autor foi bolsista da CAPES e do CNPq durante a elaboração desta tese.

Ficha catalográfica elaborada automaticamente,
com os dados fornecidos pelo(a) autor(a)

Bufolo, Gabriel Nóbrega
BB929o On some aspects of mathematical and computational models for simulations of granular materials / Gabriel Nóbrega Bufolo; orientador Yuri Dumaresq Sobral. -- Brasília, 2023.
161 p.

Dissertação(Doutorado em Matemática) -- Universidade de Brasília, 2023.

1. Granular Materials. 2. Discrete Element Method. 3. Numerical Analysis. 4. Continuum Mechanics. 5. Machine Learning. I. Dumaresq Sobral, Yuri, orient. II. Título.

Abstract

The discrete element method (DEM) is a numerical technique widely used to simulate granular materials. The temporal evolution of these simulations is often performed using a Verlet-type algorithm, because of its second order and its desirable property of energy conservation. However, when dissipative forces are considered in the model, such as the nonlinear Kuwabara-Kono model, the Verlet method no longer behaves as a second order method, but instead its order decreases to 1.5. This is caused by the singular behavior of the damping force in the Kuwabara-Kono model at the beginning and in the end of particle collisions. In this work, we introduce a simplified problem which reproduces the singularity of the Kuwabara-Kono model and prove that the order of the method decreases from 2 to $1+q$, where $0<q<1$ is the exponent of the nonlinear singular term. Furthermore, we propose a regularized normal force model based on the concept of mollifiers. We show numerically that the Verlet method combined with this regularized force model can integrate collisions with second order accuracy and that the coefficient of restitution of the system tends to increase as a function of the regularization parameter. Furthermore, using the DEM algorithm, we construct a granular Taylor-Couette computer simulation to generate coarse-grained data that will be fed into a SINDy machine learning algorithm in order to infer constitutive laws for granular flows based on the $\mu\left(I\right)$ rheology.

Resumo

O método do elemento discreto (abreviado como DEM, do inglês) é um método numérico amplamente usado para simular materiais granulares. A evolução temporal destas simulações é frequentemente feita usando algoritmos tipo Verlet, por causa de sua segunda ordem e propriedade desejada de conservação de energia. No entento, quando forças dissipativas são incluídas no modelo, como, por exemplo, o modelo não-linear de Kuwabara-Kono, o método de Verlet não mais se comporta como um método de segunda ordem, tendo sua ordem reduzida para 1.5. Isso é causado pelo comportamento singular das forças viscosas no modelo de Kuwabara-Kono no início e fim de colisões de partículas. Neste trabalho, nós introduzimos um problema simplificado que reproduz a singularidade presente no modelo de Kuwabara-Kono e provamos que a ordem do método diminui de 2 para $1 + q$, sendo $0 < q < 1$ o expoente do termo não-linear singular. Além disso, nós propomos um modelo regularizado para forças normais baseado no conceito de mollifiers. Nós mostramos numericamente que o método de Verlet combinado com esse modelo regularizado de forças é capaz de integrar colisões com precisão de segunda ordem e que o coeficiente de restituição do sistema tende a aumentar como uma função do parâmetro regularizador. Além disso, utilizando o algoritmo DEM, nós construimos uma simulação computacional de um escoamento granular de Taylor-Couette para gerar dados *coarse-grained* que serão inseridos no algoritmo de aprendizado SINDy para inferir as equações constitutivas para escoamentos granulares baseado na reologia $\mu(I)$.
Título em português: Sobre alguns aspectos de modelos matemáticos e computacionais para a simulação de materiais granulares.

Acknowledgements

The acknowledgements are likely the most read part of nearly all theses since its non-technical nature allows the layperson to understand it. It may also be the only part of a highly technical thesis on the exact sciences capable of evoking strong emotions of both kinds: eyes tearing up of joy upon reading the heartfelt words of a peer or a sad sigh after finishing and not finding any mentioned of oneself. This, in my opinion, makes it as hard to write as the remaining of the thesis, and almost as important.

However, I trained for four and a half years —the duration of my Ph.D. —to write this thesis, but I did not enjoy such training on how to write proper and eloquent thank you statements. As such, I hope that I will not be too harshly judged for what I written here. Nevertheless, I apologize for any names that I might have missed. You were so many, and the time I had to write this, so little.

To the UnB, UW, CAPES, CNPq, FAPDF and NSF for the funds and services provided to me.

To Prof. Yuri Dumaresq Sobral, my advisor during this Ph.D., for taking me in when I decided to change the direction of my studies to applied mathematics and for going above and beyond in guiding me on academic matters, as well as in life. Language as a tool is not enough to describe how grateful I am to you.

To Prof. Marcelo Fernandes Furtado, my advisor during the first year of my Ph.D., for volunteering to advise me when I naively submitted my Ph.D. application. If you had not taken me in when you did, I would probably not have gotten in at all.

To Prof. John Hinch, the seniorest member on the examining committee of my thesis defense and the former advisor of my advisor, for having imparted in my advisor so many good lessons that he has passed forward to me and for taking the time in your visit to Brazil to attended my thesis defense. I hope I can extract the gold that certainly lies within your suggestions.

To Prof. Cássio Machiaveli Oishi, the external member on the examining committee of my thesis defense and my coworker when I was in Washington state, for accepting the first task when given such short notice and for the support given to me in both those roles. To me, you were a piece of our homeland even when it was 10,000 km away.

To Prof. Taygoara Felamingo de Oliveira, the internal member on the examining committee of my thesis defense. Your insightful suggestions helped me enrich the final version of this thesis.

To Prof. Leandro Martins Cioletti, the alternate member on the examining committee of my thesis defense, for performing such role to the extent of your capabilities even amidst personal issues.

To Prof. Steven L. Brunton, currently the associate director of the AI Institute in Dynamic Systems, for accepting my request for a Ph.D. internship under your supervision. My entire life I had been waiting for the opportunity that you gave me, and the remaining of my life will now be better because of it.

To Prof. J. Nathan Kutz, currently the director of the AI Institute in Dynamic Systems, for corroborating the decision to take me in as an intern in the aforementioned institute. You treated me (and everyone else in the research group) as a friend, and that goes an incredible long way.

To Lauren D. Lederer, currently the managing director of the AI Institute in Dynamic Systems, for all the supply chain issues you dealt with to accommodate me in the institute.

To Dr. Joseph Bakarji, currently a postdoctoral fellow in the University of Washington, who closely assisted my work when I was in Seattle. All your Python and machine learning lessons have been invaluable to the last part of this thesis.

To Prof. Camila de Oliveira Vieira, who was the first Ph.D. graduate from my advisor, for sharing the knowledge that comes with being a pioneer. You never denied help when I asked for it.

To Prof. Igor Lima, Sávio Henrique Chaves Mendes and Saulo Rodrigo Medrado, for all the help in acquiring, setting up and testing the necessary equipment for a hybrid thesis presentation.

To Renato Bufolo, my father, for providing for our family and our home, and for keeping both of my feet on the ground. I would not have had the choice to get here if it weren't for you.

To Roberta de Farias Nóbrega Bufolo, my mother, for caring about my well-being and never giving up on me; for managing our family and our home; for showing me just how powerful motherly love is. Whenever I need, you will cry with and for me and, afterwards, will wipe away my tears.

To Ravena Nóbrega Bufolo, my sister, for, when I found myself in times of trouble, making time amidst your busy schedule to knock on my door to talk to me and cheer me up. You are my standard for work ethic and an inspiration in dealing with adversities.

To Claúdia Maria de Farias Nóbrega, my aunt, for always taking the initiative to offer any help you could. Your active show of support, even when matters were too technical for you to help me with, served as a driving force for me to strive further.

To Laurence Nóbrega, my grandfather, for fostering my curiosity since I was a child, continuously staying open to listen and offer guidance, and always treating my thoughts and concerns with utmost sincerity. "Conosco, ninguém podemos!"

To Lúcia Maria de Farias Nóbrega, my grandmother, for all her prayers. You are my most enthusiastic supporter!

To Víctor Carvalho de Oliveira, who shared a room in the department of mathematics with me, for all the blood, sweat and tears. You were like a brother to me during the time we spent together.

To Anastasia Bizyaeva, Andrei Klishin, Cássio Oishi, Doris Voina, Jonas Kneifl, Joseph Bakarji, Paolo Conti, Prerna Patil, Ryan Raut, and Samuel Otto, my coworkers in the AI Institute in Dynamic Systems, for embracing my presence during my time in Washington state. You made me feel like I belonged and I will forever cherish the memories of the moments we shared.

And to all classmates, staff and other coworkers.

Thank you!

Contents

Introduction

Rice provides more than 20% of the calories consumed worldwide [28]. Between the full maturation of the grain in the fields and the neatly packaged bags of rice in your local supermarket there exists many different processes that have been deliberately performed, among which are the harvest, transport, storage, filtering and packaging. During the aforementioned processes, the most relevant characteristic of rice is not its energy content or its chemical composition, but the fact that it is composed of individual grains. Because of this property, the behavior of large amounts of this substance is not completely akin to a solid or fluid, but something in between... and sometimes, it can even have a gas-like behavior. These materials —the ones formed by macroscopic grains —are called granular materials, and they are incredibly important in the processes that shape the literal and figurative landscapes of our world.

Figure 1: Grains of rice flowing from a combine harvester (a) and an old rice mill (b).

When rain falls, the water that is continuously deposited on top of the irregular soil follows the steepest path downhill. As it flows, it carries along a literal piece —or grain —of soil. Some of those steepest, now lacking some soil that was carried away, become brooks. Given time and circumstance, a few of those brooks can widen to creeks. Over years, those creeks can convert to streams and, over eons, the streams can grow into rivers. These rivers meander and shape the land, carving oxbow lakes, flattening mountains and widening valleys. All of this is possible because soil is a granular medium.

The Amazon rainforest, spanning over five million square kilometers, is so large, that it affects the weather in nearby regions [73]. Being this large, it requires a Brobdingnagian amount of nutrients to sustain itself —among which is phosphorus. The source of this phosphorus can be traced back to more than five thousand kilometers away and across the ocean: the Sahara desert. There, a single grain of sand

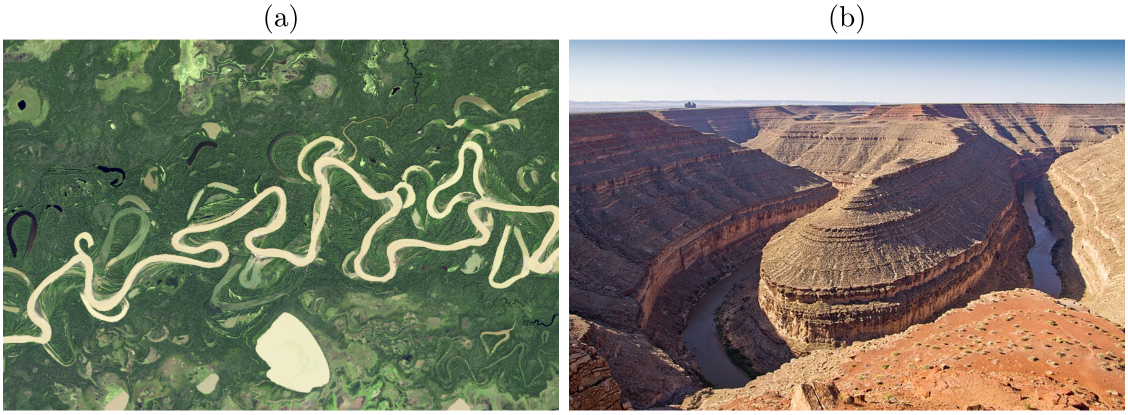


Figure 2: The meandering Amazon (a) and Colorado (b) rivers. Note, in (a), that the many paths it took in the last few decades still remain engraved on the ground. In (b), note the gargantuan trench it has carved into the land —this is the great Canyon!

is picked up by the wind and raised into the atmosphere, where it then rides air currents until it is deposited somewhere in the Amazon basin. Yearly, more than twenty teragrams of dust make this one way trip into the rainforest [88]. However, if the material composing the desert was not made of fine particulate (i.e., if it was not a granular medium), this trip would not be possible. In this case, perhaps our world would be devoid of the Amazon rainforest.

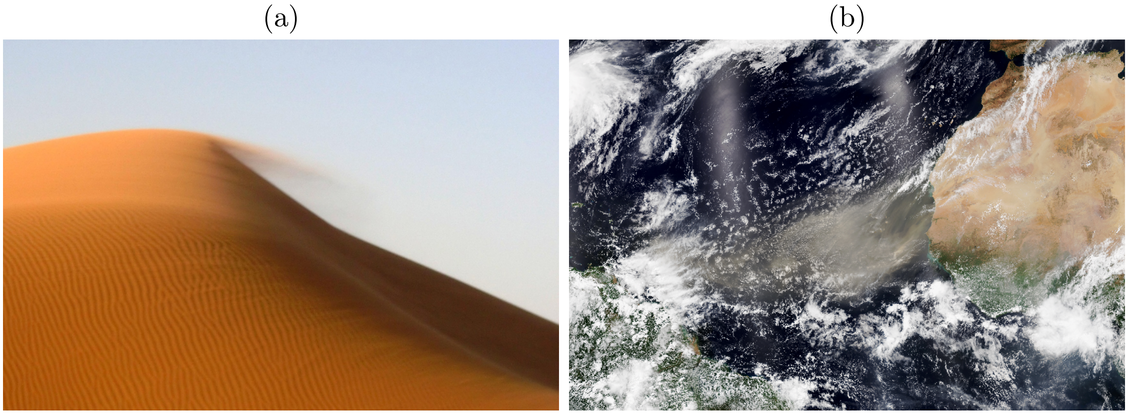


Figure 3: In (b), a satellite image of an immense plume of dust from the Sahara crossing the Atlantic ocean. It may have started as a gust of wind on a dune (a).

The Tycho crater, on the moon, is estimated to be 108 million years old [36]. Its name pays homage to astronomy pioneer Tycho Brahe, who, sadly, died a mere 7 years before the telescope was invented. If he had had access to this invention, even in its infancy, he would have been able to observe the eponymous crater and would likely be puzzled by its ray system. Nowadays, most people have seen this crater and ray system in pictures of the moon —even if they did not notice it. You can see it in fig. 4(a). Ray systems such as the aforementioned one are not exclusive to the moon; they are abundant in Mercury and have also been identified in Mars and even on Earth: the Kamil crater, in the Sahara [80]. In 2018, it was proposed that ray systems form as a result of impacts of asymmetrical meteorites [61]. This proposition was supported by experiments that consist of colliding small, asymmetrical objects with a granular bed —literally colliding them with sand. Even though the surface of the moon is not a granular medium, at least not at the scales of the depth of the Tycho crater, it is still possible to study its morphology by understanding granular phenomena.

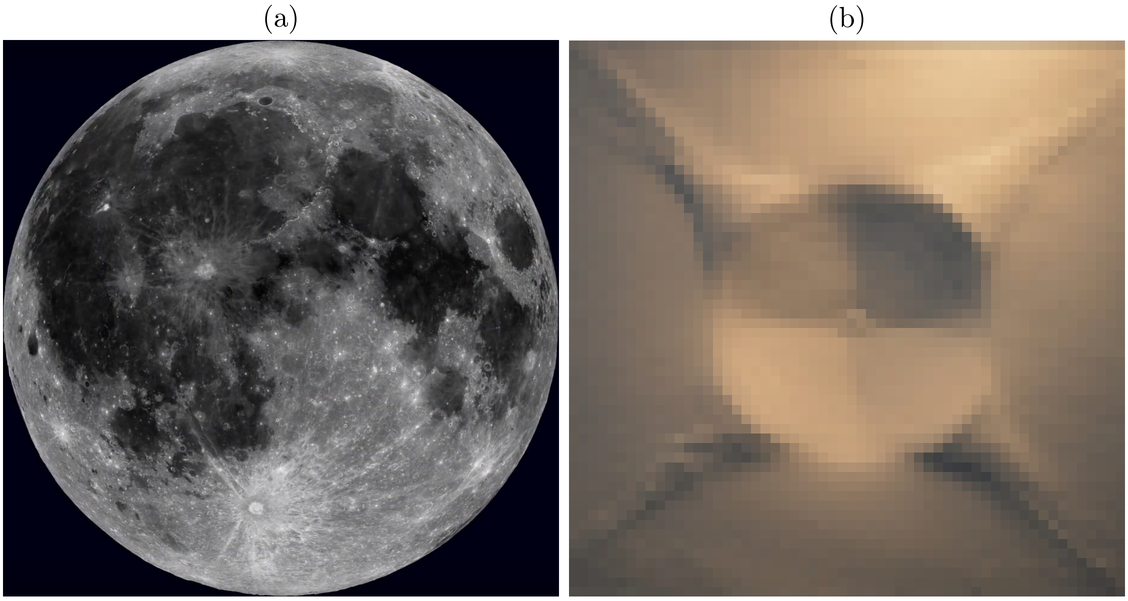


Figure 4: The moon (a). Tycho and its ray system are visible in the bottom quarter of the picture, slightly to the left. In (b), an experimental crater [61] that has its own ray system, due to the non-spherical nature of the impactor. PS: The final version will have a better resolution of (b).

Hopefully, these examples have elicited interest in granular materials to the reader; or, at least, convinced them they are fundamental enough to their lives and the universe that it is a worthwhile endeavor to pursue the advancement of our fundamental knowledge about them. This thesis is my first, humble contribution to that end.

In the first part, we review the foundations of discrete granular simulations. In particular, we describe the standard numerical methods and force schemes used in these simulations. In the second part, our research into discrete granular simulations is shown. First, we study the effect of the Hertz-Kuwabara-Kono force scheme on the order of convergence of the leapfrog integration method. We prove a theorem asserting quantitatively the order penalization of the method and then propose an empirical solution to this penalization. Second, we study the impact of the order of convergence of the chosen integration method into the trajectories of individual particles in the collapse of a granular column. Finally, in the third part, we report on our current research: using machine learning to infer equations describing the velocity field of a granular Taylor-Couette flow.

Part I

Preliminaries

Introduction

A granular medium is broadly defined as a collection of rigid macroscopic particles, whose particle size is larger than $1.00 \times 10^2\,\mu\text{m}$ [3]. This size limitation aims to restrict the types of interaction among particles that dictate the motion of the medium. In particular, interactions such as thermal agitation (Brownian motion) and van der Waals forces can be considered negligible at scales equal or larger than this value.

Despite their medium fitting the same definition given above, flows of granular media can be wildly different, e.g., even the untrained eye can notice the contrast between a wind-driven flow of sand over a dune and the flow of grains on an industrial screw conveyor. Nevertheless, whenever two granular flows share the same macro behavior, they are said to follow the same flow regime. Some examples of flow regimes, as detailed in [57], are the plane shear, annular shear, vertical-chute flow, inclined plane, heap flow and rotating drum. These regimes make the assumption of a dry, monodisperse granular medium. Another example, one which we shall particularly concern ourselves with, is of the collapse of a granular column under the influence of gravity.

Large gravity driven granular flows are abundant in nature and are of particular importance because of how hazardous they can be. Some examples include rock avalanches, debris flow, mudslides, underwater avalanches and sudden landslides. Much research effort has been put into understanding and predicting the triggering and evolution of these types of granular flows, so that the damages caused by them can be somewhat mitigated. For the specific case of dry flows, one model has found much success in the scientific community. This model, described in [5], simplifies the complex topographies generally involved in natural phenomenon to that of a column of grains. The collapse of this column under its own weight provides a surprisingly accurate description of the distances and areas affected by, for instance, a rock avalanche.

An obvious advantage of the model of granular columns is the relative ease at which it can be experimentally tested. However, even with state of the art equipment, it proves to be exceedingly hard to observe certain intrinsic characteristics of the flows performed experimentally. For instance, the force chains that sustain the entire column prior to the collapse or even the individual trajectories of grains, specially of those grains which remain buried deep inside the column during the whole collapse. To circumvent this issue, computer simulations of granular media have been developed and have proved to be a very helpful and reliable tool. Some of these computer simulations provide full knowledge of the state of each particle at each instant in time.

One of the first computational models to simulate granular materials was the one proposed by [17]. It is based on the idea of solving the motion of individual particles by prescribing pairwise contact rules between two particles. This approach is very similar to the molecular dynamics algorithm used in physico-chemical simulation of atoms in molecules [79]. At some point in its development, the model proposed by [17] became known as the "discrete element method" (DEM) and is now one of the most widespread computational tools used to model granular flows in the most varied contexts [32]. In recent years, several variations of DEM algorithm have been proposed [63], and it has also been enhanced by the incorporation of fluid flow using different techniques [42], [85].

The discrete element method will be a major player in the second part of this work. Thus, in this part, we review core concepts involved in the computational simulation of grains with the discrete element method, many of which have been summarized in [71]. Afterwards, we provide validation for our implementation of this method. Finally, some simulations of granular columns are presented and compared to the experimental results of [5].

Chapter 1

Computer simulations of granular materials

1.1 The discrete element method

1.1.1 Particles

To more precisely explain the DEM, a particle is defined as a quadruple $(r, \rho, \vec{x}, \theta)$, where $r \in (0, \infty)$ is its radius, $\rho \in (0, \infty)$ its density, $\vec{x} : [0, \infty) \to \mathbb{R}^2$ its position on the plane as a function of time and $\theta : [0, \infty) \to (-\infty, \infty)$ its accumulated angle of rotation around the axis perpendicular to the coordinate system of the simulation as a function of time. For the sake of simplicity, we chose to have ρ constant for all the particles.

Based on this definition one can visualize a particle as a perfect sphere of radius r made of a material of density ρ. The assumption of the particle as a perfect sphere reduces the computational cost of this model tremendously, at the cost of physical accuracy. The interested reader can find more information on the effects of particle shape in [59].

The angle of rotation must be measured against some arbitrarily chosen reference position, which we chose to be the upwards direction. Also, we follow the convention of positive sign for counterclockwise rotation (see fig. 1.1).

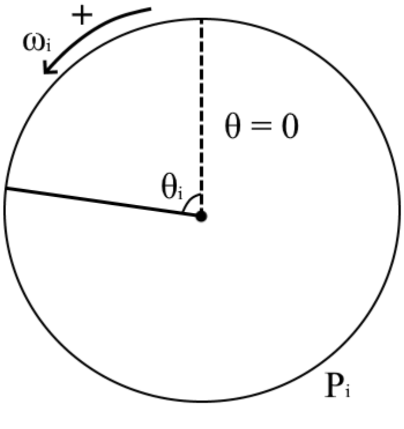


Figure 1.1: The conventions of angular motion. Here, P_i refers to a specific particle while the subindex i in θ_i and ω_i means that these properties refer to P_i.

Some derivatives of $\vec{x}$ and θ have special notations and meaning. For instance $\vec{v} := \vec{x}'$ is the velocity of the particle and $\vec{a} := \vec{v}' = \vec{x}''$ is its acceleration. Similarly, $\omega := \theta'$ is the angular velocity of the particle and $\alpha := \omega' = \theta''$ is its angular acceleration.

Given a particle, we call

$$m := \frac{4}{3}\pi r^3 \rho \tag{1.1}$$

its mass and

$$I := \frac{2}{5} m r^2 \tag{1.2}$$

its moment of inertia. The motivation for such names is clear and their expressions could be derived by assuming each particle as a perfect homogeneous sphere.

1.1.2 Collisions

It is usual to have many different particles in DEM applications. We shall denote them by P_i, with the subindex $i \in \mathbb{N}$ being used to differentiate among them. This shall also apply to all their constituent elements and their derived properties. So, for instance, particle P_i has radius r_i, position $\vec{x}_i\left(t\right)$, and mass m_i. The set of all such indexes will be denoted by $\mathbb{I}$.

For each pair of particles P_i and P_j, we call

$$d_{i,j} := \|\vec{x}_i - \vec{x}_j\| \tag{1.3}$$

their distance function and

$$\xi_{i,j} := \min\left\{0, r_i + r_j - d_{i,j}\right\} \tag{1.4}$$

their overlap function. For each $t \in [0, \infty)$ such that $\xi_{i,j}\left(t\right) > 0$, it is said that the respective particles are “colliding” (or “overlapping” or “in contact”) in the instant t.

Let P_i and P_j be a pair of colliding particles at some instant $t \in [0, \infty)$ and consider the largest (in the inclusion sense) interval that satisfies:

- Contains t,
- Only contains instants in which these two particles are still colliding.

The infimum of such type of interval will be denoted by t_0 and is called the beginning of the contact. Similarly, the supremum of that type of interval is called the end of the collision and will be denoted by t_f. One can also say that “P_i and P_j are colliding during (t_0, t_f)”.

Then, the normal vector $\vec{n}_{i,j}$ of the collision at that instant is defined as

$$\vec{n}_{i,j} := \frac{\vec{x}_j - \vec{x}_i}{\|\vec{x}_j - \vec{x}_i\|}. \tag{1.5}$$

In a similar way, the tangent vector $\vec{t}_{i,j}$ of the collision at such instant can be defined as

$$\vec{t}_{i,j} := \left(n_y, -n_x\right), \tag{1.6}$$

where n_x and n_y are the x-oriented and y-oriented components of $\vec{n}_{i,j}$, respectively. Notice that $\vec{n}_{i,j} = -\vec{n}_{j,i}$ and $\vec{t}_{i,j} = -\vec{t}_{j,i}$. The overlap between two particles and the normal and tangent vectors of a collision are illustrated in fig. 1.2.

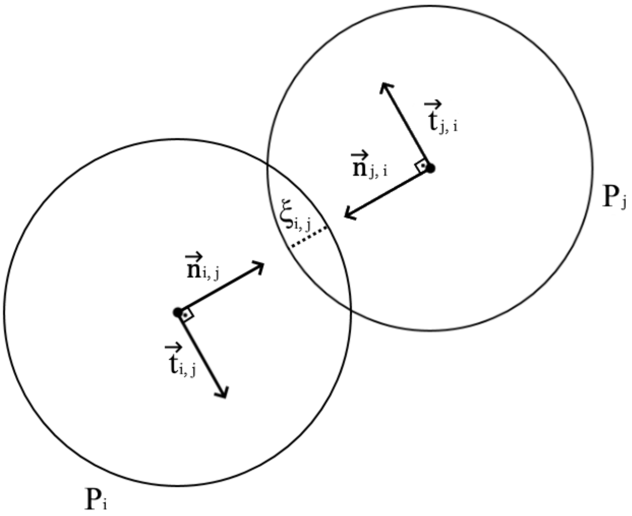


Figure 1.2: Visualization of normal and tangent vectors as well as the overlap between two particles.

The contact point of the collision at the instant t is defined as

$$\vec{c}_{i,j} := \vec{x}_i + \left(r_i - \frac{\xi_{i,j}}{2}\right)\vec{n}_{i,j}, \tag{1.7}$$

which can be visualized in fig. 1.3.

One can also define the relative velocity of the collision at the instant t as

$$\vec{v}_{i,j} := \vec{v}_j - \vec{v}_i. \tag{1.8}$$

Then, the relative normal velocity of the collision in that moment is merely the projection of the relative velocity on the normal vector, that is:

$$v^n_{i,j} := (\vec{v}_j - \vec{v}_i) \cdot \vec{n}_{i,j}, \tag{1.9}$$

where the symbol "·" is used to denote the Euclidean scalar product. The relative shear velocity of the collision in that moment can also be defined, but one must be careful and also take into account the tangential motion at the contact point due to the rotation of the particles. This leads to the following definition:

$$\begin{aligned} v^s_{i,j}(t) &:= (\vec{v}_i(t) - \vec{v}_j(t)) \cdot \vec{t}_{i,j}(t) \\ &- \omega_i(t)\left(r_i - \frac{\xi_{i,j}(t)}{2}\right) + \omega_j(t)\left(r_j - \frac{\xi_{i,j}(t)}{2}\right). \end{aligned} \tag{1.10}$$

The definitions for the above quantities at the instant t can be easily extended to vector functions of time. We will use the same notation for both.

1.1.3 Forces

A force is defined as a triple $\left(\vec{F}, P, \vec{p}\right)$, where $\vec{F} : [0, \infty) \to \mathbb{R}^2$ is the (vectorial) value over time of the force, P is a particle and $\vec{p} : \mathbb{R}^2 \to \mathbb{R}^2$ is the point in which the force acts. We say that $\vec{F}$ acts on P at the point $\vec{p}$.

Before we proceed, an observation must be made. Generally, the function $\vec{F}$ is what is meant when one talks about a force. However, the DEM model we use needs these additional informations, which makes us opt to use the definition that has been presented. Nevertheless, we will use both the symbol $\vec{F}$ when referring either to the triple $\left(\vec{F}, P, \vec{p}\right)$ or to the first element of this triple, the function $\vec{F} : [0, \infty) \to \mathbb{R}^2$. This will also happen with other physical quantities, e.g. torques.

In this work, there are only two types of forces: individual forces and contact forces. The former are characterized by being caused by objects outside the scope of the main model (e.g.: the Earth causing a gravitational force or a magnet causing magnetic forces) while the latter arises from collisions between two particles. These collisions are the focus of the DEM. There is only one type of individual force in this work and, as such, we can write $\vec{F}_i$ to designate the only individual force $\vec{F}$ that acts on P_i. Since there is also only one contact force for each pair of particles, we write $\vec{F}_{i,j}$ to indicate the only contact force $\vec{F}$ that acts on P_i due to its contact with P_j. The set of all forces acting on P_i is denoted by $\mathbb{F}_i$.

Let F_i be an individual force. Then, its point of action is the center of the particle P_i, i.e.,

$$\vec{p} := \vec{x}_i. \tag{1.11}$$

If $\vec{F}_{i,j}$ is a contact force, then its point of action is the contact point instead. Formally:

$$\vec{p} := \vec{c}_{i,j}. \tag{1.12}$$

The vector $\vec{c}_{i,j}$ can be visualized in fig. 1.3. We will discuss more about forces in section 1.3.

1.1.4 Torques

A torque can be defined as a double (τ, P), where $\tau : [0, \infty) \to \mathbb{R}$ is its magnitude and P a particle. It is also said that τ acts on P.

The only torques acting on the particles in this work are due to contact forces. Given a force $\left(\vec{F}, P, \vec{p}\right)$, let $\vec{l}\,(t) := \vec{x}\,(t) - \vec{p}\,(t)$, where $\vec{x}$ is the position of P. The vector $\vec{l}\,(t)$ is called the lever vector and is illustrated in fig. 1.3. Given an instant $t \in [0, \infty)$, the Cartesian coordinates of $\vec{F}$ and $\vec{l}$ are denoted by $\vec{F}\,(t) = (F_x, F_y)$ and $\vec{l}\,(t) = (l_x, l_y)$, respectively. Then, this force produces a torque acting on P with magnitude

$$\tau\,(t) = l_x F_y - l_y F_x. \tag{1.13}$$

1.1.5 Motion

The movement of the particles is determined by Newton's laws of motion. As the reference frame upon which the whole simulation is measured is an inertial one, the first law is a mere consequence of the second. The latter states that

$$\sum_{\vec{F} \in \mathbb{F}_i} \vec{F} = (m_i \vec{v}_i)'. \tag{1.14}$$

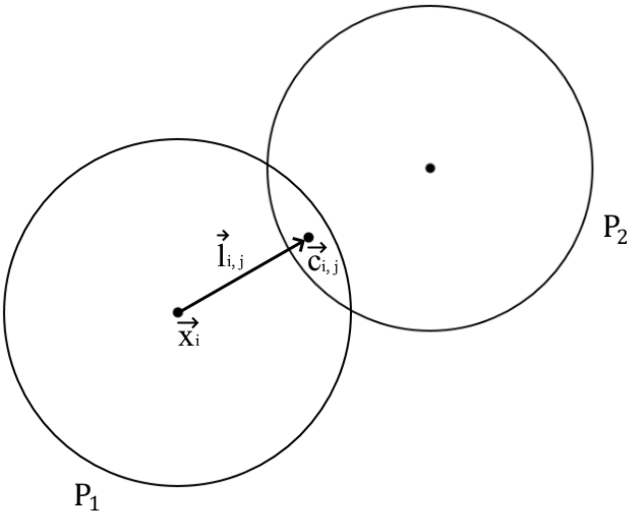


Figure 1.3: The lever vector and the contact point of a collision.

Since each particle has a constant mass, the above equation may be simplified to

$$\sum_{\vec{F}\in\mathbb{F}_i} \vec{F} = m_i a_i. \tag{1.15}$$

The second Newtonian law of motion also describes the rotational dynamics of objects - particles in our case. Already taking into account the constant moment of inertia in this model, it can be formulated as

$$\sum_{\tau\in\mathbb{T}_i} \tau = I_i \alpha_i, \tag{1.16}$$

where $\mathbb{T}_i$ is the set of all torques that act on this same particle.

Newton's third law is incorporated in this model by making all contact forces appear in pairs with opposite direction, each acting on one of the particles that partake in such collision. This effect doesn't happen in individual forces because it is assumed that the source of the force is outside of the scope of this model (e.g.: the Earth for gravitational forces or a magnet for magnetic ones).

Finally, to solve the ODEs in eqs. (1.15) and (1.16), one would have to choose an adequate numerical method. In section 1.4, we introduce our choices of numerical methods for the present work. Then, in section 2.1, we establish some issues associated with each of these numerical methods.

1.2 Spatial hash

As mentioned previously, one of the drawbacks of the DEM is its computational cost. Since the DEM requires that each individual particle in a system is simulated, this means that whenever a contact force $\vec{F}_{i,j}$ is calculated, the program must be able to verify whether P_i and P_j are in contact. A naive implementation would check each possible pair of particles to determine their collision status. This algorithm is $\mathcal{O}(n^2)$ with the number of particles, which would make simulating thousands or tens of thousands of particles impossible.

Fortunately, there are more efficient algorithms, such as the Verlet list algorithm, which was introduced in [81]. The Verlet list algorithm maintains, for each particle, a list of its closest neighbors and periodically updates theses list. In general, this algorithm is $\mathcal{O}(n \log(n))$ with the number of particles.

Another algorithm for neighbor finding which has become mainstream in DEM is the cell linked-list algorithm, which was first introduced in [2]. In this algorithm, the domain of the simulation is partitioned in axis-aligned square regions in such a way that only the particles in adjacent regions need to be checked for collision. In general, this algorithm is $\mathcal{O}(n)$ with the number of particles. A comparison between the Verlet list and cell linked-list algorithms can be found in [21].

A third noteworthy neighbor finding algorithm uses a quadtree to subdivide the domain of the simulation according to the presence and density of particles in each region of the domain of the simulation. It was first introduced in [24] and is particularly well suited for simulations where particle are sparsely distributed in the domain or when there are significant differences in the size of the particles.

In this work, we have implemented an algorithm similar to the cell linked-list algorithm, which is popularly called "spatial hash".

1.2.1 Preliminaries

Let i be the index of a particle in the simulation. The axis-aligned square inside which this particle is inscribed is called its "bounding box". We will call the four points in the corners of the bounding box simply as "corners" or "the corner of the particle i". The bounding box of a particle, as well as the corners of the bounding box, can be visualized in fig. 1.4.

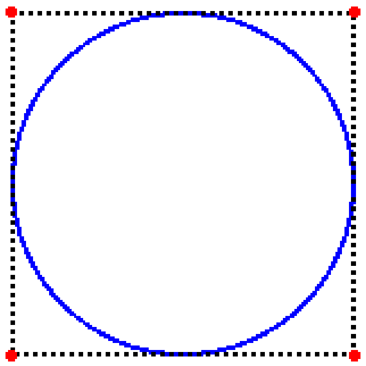


Figure 1.4: A particle (in blue), its bounding box (in black) and its corners (in red).

Let

$$L := 2 \max_{i \in \mathbb{I}} r_i, \tag{1.17}$$

where $\mathbb{I}$ is the set of all indexes of particles. Given $\vec{z} = (z_1, z_2) \in \mathbb{R}^2$ such that $z_1, z_2 > 0$, the vector $\vec{z}_\lfloor \in \mathbb{Z}_+^2$ defined as

$$\vec{z}_\lfloor := \left(\left\lfloor \frac{z_1}{L} \right\rfloor, \left\lfloor \frac{z_2}{L} \right\rfloor \right). \tag{1.18}$$

is called the grid-coordinates of $\vec{z}$. If the positive quadrant of the $\mathbb{R}^2$ plane is partitioned into cells shaped as L-sided axis-aligned squares, then $\vec{z}_\lfloor$ would be the integer coordinates of the cell inside which $\vec{z}$ is. In this way, each different grid-coordinate $\vec{z}_\lfloor$ is corresponded in a one-to-one fashion to the cell whose bottom left corner is at $L\vec{z}_\lfloor$. A cell is said to "contain" $\vec{z}$ if the grid-coordinates of $\vec{z}$ are the same as that of the cell. The relative (to the cell in which is is contained) coordinates of $\vec{z}$ are given by $\vec{z} - L\vec{z}_\lfloor$. Note that, once $\vec{z}_\lfloor$ is known, calculating these relative coordinates are computationally inexpensive, since the expression avoids divisions. The grid, grid-coordinates and relative coordinates are illustrated in fig. 1.5.

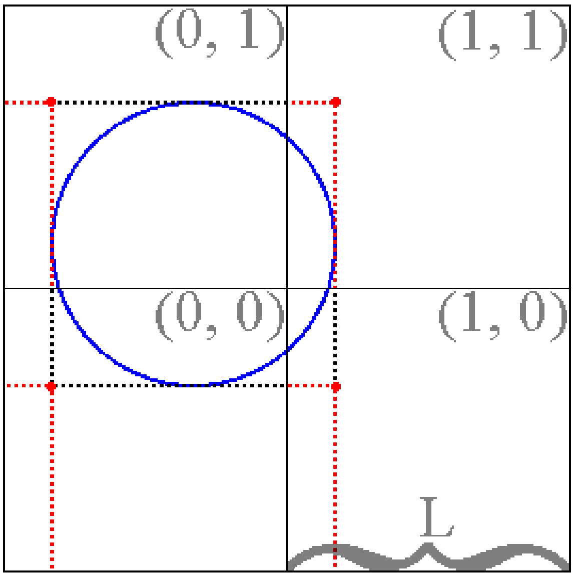


Figure 1.5: A particle in the cell representation of the system. The grid-coordinates of a corner is the ordered pair on the cell where said corner is located. In red, the remainder-coordinates of the corners are displayed as coordinates with respect to the cell inside which the corner is.

1.2.2 The algorithm

Initialization

The goal of the initialization procedure described below is to first populate a data structure with the positions of each corner of each particle. This data structure must be organized in such a way that each cell has its own bucket and that it is possible to retrieve the contents of a bucket with the grid-coordinates of the cell is is associated with in $\mathcal{O}(1)$ time. The initialization procedure is defined as follows:

I.1 For each cell, allocate a contiguous block of memory [1]. Each such block of memory corresponds to a cell in a one-to-one fashion and will be designated as "the block of memory associated of the cell".

I.2 Let i be the index of a particle in the simulation. Convert the coordinates of its corners into grid-coordinates.

[1] This means a block of memory in which memory addresses are sequentially ordered. An example of such contiguously allocated memory is the C++ `std::vector<T>` data structure, which is what was actually used in our code. This sequential order is important because accessing memory that is contiguously allocated is faster, due to cache pre-fetching. General single core optimization techniques and memory architecture are not the main focus of this work though, so we will not go further on justifying this. The interested reader may find more information on [26].

I.3 The grid-coordinates of each cell must be stored in memory and, if the index of the particle is know, its access complexity must be $\mathcal{O}(1)$ with the number of particles.

I.4 Store i in the block of memory of each cell that contains at least one of the corners of the particle.

I.5 Repeat for all particle indexes.

This algorithm simply maps each corner of the bounding box of each particle to the cell inside which it is located.

Retrieval

When a request for the neighbors of a particle (let us assume the index of this particle to be i) is made, the algorithm executes the retrieval procedure described below. In this procedure, the possible candidates are identified by checking the cells inside which the corners of particle i are. These candidates are then grouped in a list, sorted and the redundant copies among them are eliminated from this list, which is the final product of the procedure. The steps of the retrieval are:

R.1 Convert the coordinates of the corners of particle i into grid-coordinates.

R.2 Create a contiguous block of memory (this needs to be done only once for each particle, usually during the start-up of the software).

R.3 Use the grid-coordinates of the corners to access the block of memory of each cell and copy the indexes in that block of memory over to the block of memory created in the last step. Care must be taken to avoid accessing the same cell more than once, in cases where some of the corners of particle i have the same grid-coordinates.

R.4 Sort this block of memory and then remove repeated indexes. The contents of this block are the indexes of the desired neighbors.

This algorithm uses the previous mapping as a quick way to retrieve the neighbors of particles. The sorting and removal of duplicates are the real bottleneck in this implementation of the search for neighbors and this is due to the fact that multiple cells could contain different corners of the same particle, which means that the list obtained from step 3 could have repeated copies of some index. This, if left alone, could cause the same collision to be computed twice, which would not only be redundant and a waste of computational power, but it would also cause the dynamics of the system to misbehave. This misbehavior would be caused by applying the same force multiple times.

Maintenance

During a simulation, particles are constantly moving. Invariably, the corner of a particle will cross the boundary between cells. If these crossings are not accounted for, the data structure, populated in the initialization of the algorithm, quickly becomes obsolete. To avoid this, that data structure needs to be maintained, i.e. the positions of the particles must be updated inside the data structure.

In the Verlet list and cell linked-list algorithms, a similar maintenance is required and it is implement by fully repopulating the respective data structures after a fixed number of steps. One of the main differences of our implementation from those algorithms (together with the usage of corners) is that our implementation never repopulates the entire data structure. Instead, it quickly detects when a particle has crossed between cells and updates only the information regarding that particular particle in the data structure.

The maintenance routine is designed as follows:

M.1 Let i be the index of the particle whose movement has just been calculated. Calculate the coordinates of the corners of particle i relative to the cell each corner is in. It is important that the grid-coordinates used for this computation have not been recalculated since the calculating of the movement of particle i has finished.

M.2 If each entry in this vector is between 0 and L, calculate the movement of other particle and restart the procedure.

M.3 If some entry of that vector is negative or bigger than L, use the grid-coordinates of the corners of the particle i to access the block of memory associated with each cell that contains at least one corner of this particle and remove all of copies of i from each of them.

M.4 Recalculate the grid-coordinates of the corners of particle i.

M.5 Insert i in the block of memory of each cell that contains at least one of its corners.

M.6 Calculate the movement of other particle and restart the procedure. Repeat until the movement of all particles has been processed for this step.

This is the most crucial part of the whole neighbor-finding algorithm. Notice that calculating grid-coordinates is expensive, since it involves floating-point division. To avoid this becoming a performance bottleneck, the grid-coordinates are only calculated when a particle leaves its current cell. The computation performed in step M.1 is, effectively, a cheap way to check whether the particle has left the cell it was in previously, since it only involves subtraction and multiplication. Then, if any of the corners of the particle has changed cell, the rest of the algorithm just removes its index from all cells, recalculates its grid-coordinates and puts back the index of the particle in the correct cells.

1.3 Force Schemes

The choice of contact force schemes influences the dynamic of two-particle collision as well as the overall behavior of a collapse. In this section, all the major force schemes that we considered for this work will be presented. All of these force schemes are summarized in [71].

To keep consistency, we shall denote the component of $\overrightarrow{F}_{i,j}(t)$ in the direction of $\overrightarrow{n}_{i,j}(t)$ as $\overrightarrow{N}_{i,j}(t)$ and the component of $\overrightarrow{F}_{i,j}(t)$ in the direction of $\overrightarrow{t}_{i,j}(t)$ as $\overrightarrow{T}_{i,j}(t)$.

1.3.1 Normal force schemes

In a collision, a force is said to be normal if it is parallel to the normal vector of the collision. Normal force schemes are usually associated with the inelastic deformation of the material that composes the particles involved in the collision.

Because of the inelastic nature of the deformation, normal force commonly have two components: an elastic component and a viscous (or damping) component. The viscous component is characterized by a dependence on the velocities of the particles involved in the collision. Consequently, it is responsible for dissipating the energy of the particles, slowing their movement down. The elastic component does not have an explicit dependence on the velocity and is energy conservative. It is responsible for the bulk of the magnitude of the normal force and the eventual separation of the particles.

The spring-dashpot model

The simplest normal force scheme is the spring-dashpot model, force expression is

$$\vec{N}_{i,j}\left(t\right) = \begin{cases} 0, \text{if } P_i \text{ and } P_j \text{ are not in contact in the instant } t \\ \left(-k_n \xi_{i,j}\left(t\right) - \gamma_n \xi'_{i,j}\left(t\right)\right) \vec{n}_{i,j}\left(t\right), \text{otherwise,} \end{cases} \tag{1.19}$$

where $k_n, \gamma_n \in (0, \infty)$ are parameters of the model. The former is called the elastic constant and, the latter, damping constant.

There are other force schemes that better match experimental data. Nonetheless, a reason one might prefer this scheme is the fact that an analytical solution for a normal collision between two particles interacting via this force scheme can be easily obtained [71] and, in such collision, both the contact duration and coefficient of restitution are independent of both particles velocities.

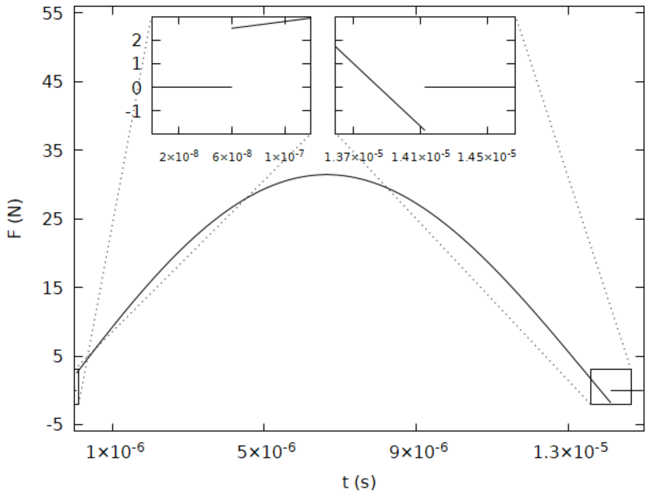


Figure 1.6: Force as a function of time for a normal collision in the spring-dashpot force scheme. The insets zoom on the beginning and end of the collision. Here $m = 1.48 \times 10^{-4}\,\mathrm{kg}$, $k_n = 7.32 \times 10^6\,\mathrm{N/m}$ and $\gamma_n = 2.06\,\mathrm{kg/s}$.

On the other hand, one reason one might want to avoid this model is its behavior at the beginning and end of the collision, as shown in the insets of fig. 1.6. There are discontinuities at both moments, due to the fact that normal relative velocities are not zero at these moments. Also, at the end of the collision, note that the normal force becomes attractive, which has no physical meaning.

Hertz's model with linear damping

A less simplistic approach is given by Hertz's contact mechanics [34]:

$$\vec{N}_{i,j}(t) = \begin{cases} 0, \text{if } P_i \text{ and } P_j \text{ are not in contact in the instant } t \\ \left(-\frac{4}{3}E_{i,j}{r_{i,j}}^{\frac{1}{2}}(\xi_{i,j}(t))^{\frac{3}{2}}\right)\vec{n}_{i,j}(t), \text{otherwise,} \end{cases} \tag{1.20}$$

where

$$r_{i,j} = \frac{1}{\frac{1}{r_i}+\frac{1}{r_j}}. \tag{1.21}$$

The term $E_{i,j} \in (0, \infty)$ on the formula above depends on the Young moduli and Poisson ratios of the materials of both particles. This dependence on the physical properties of the materials of the colliding particles can be considered an advantage of this force scheme, since it allows for a more direct translation of experimental data to the simulations.

As simulations in this work are treated as though the granular media is homogeneous, the sub-indexes can be dropped, i.e. $E_{i,j} = E$. It is usual then to write $\tilde{k}_n = \frac{4}{3}E{r_{i,j}}^{\frac{1}{2}}$. In this work, we decided to make a further simplification and assume $\tilde{k}_n$ be constant, i.e., independent from r_i and r_j. Such hypothesis would be ideal for a particle assembly close to monodispersity.

The above force scheme, however, is not dissipative, which does not reflect the actual physics of most granular collapses. To solve this issue, a linear viscous term is usually added in an ad-hoc fashion. Then, the new expression for the force scheme is

$$\vec{N}_{i,j}(t) = \begin{cases} 0, \text{if } P_i \text{ and } P_j \text{ are not in contact in the instant } t \\ \left(-\tilde{k}_n(\xi_{i,j}(t))^{\frac{3}{2}} - \gamma_n\xi'_{i,j}(t)\right)\vec{n}_{i,j}(t), \text{otherwise.} \end{cases} \tag{1.22}$$

Unfortunately, the inclusion of this linear damping results in the same issues that were present on the spring-dashpot model, i.e. the discontinuities at both ends of the contact and the attractive behavior of the normal force in the last moments of the collision. This can be observed in the insets of fig. 1.7.

Kuwabara-Kono model

More recently, Goro Kuwabara and Kimitoshi Kono [49] proposed to add a non-linear damping term to the elastic Hertzian force. The result is as follows:

$$\vec{N}_{i,j}(t) = \begin{cases} 0, \text{if } P_i \text{ and } P_j \text{ are not in contact in the instant } t \\ \left(-\tilde{k}_n(\xi_{i,j}(t))^{\frac{3}{2}} - \tilde{\gamma}_n\xi'_{i,j}(t)(\xi_{i,j}(t))^{\frac{1}{2}}\right)\vec{n}_{i,j}(t), \text{otherwise,} \end{cases} \tag{1.23}$$

where $\tilde{k}_n$ is the same as in the Hertz model and $\tilde{\gamma}_n$ is a constant. Note that in the original paper, $\tilde{\gamma}_n$ depends on the radii of the particles and the two coefficients of bulk viscosity, but we opted to treat it as a constant for added simplicity.

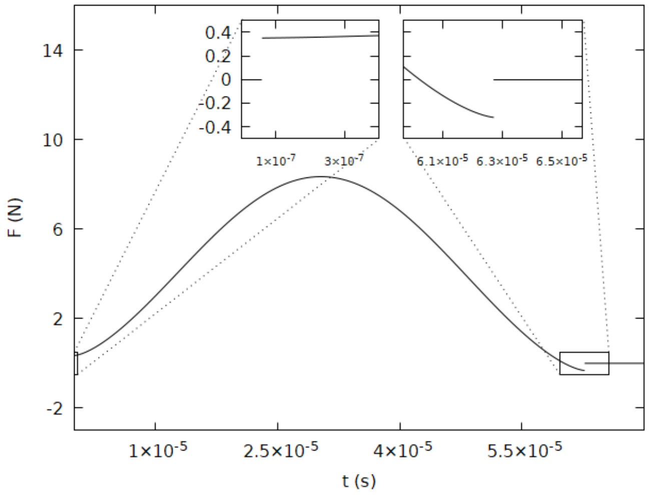


Figure 1.7: Force against time for a normal collision in the Hertz force scheme with linear damping. The insets zoom on the beginning and end of the collision. Here $m = 1.48 \times 10^{-4}\,\mathrm{kg}$, $\tilde{k}_n = 9 \times 10^{7}\,\mathrm{N/m^{1.5}}$ and $\gamma_n = 3.5 \times 10^{-1}\,\mathrm{kg/s}$.

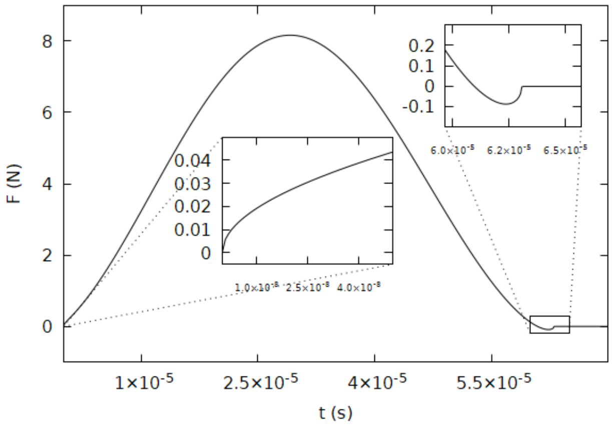


Figure 1.8: Force as a function of time for a normal collision in the Kuwabara-Kono force scheme. The insets zoom on the beginning and end of the collision. Here $m = 1.48 \times 10^{-4}\,\mathrm{kg}$, $\tilde{k}_n = 9 \times 10^{7}\,\mathrm{N/m^{1.5}}$ and $\tilde{\gamma}_n = 1.90 \times 10^{2}\,\mathrm{kg/(m^{0.5}\,s)}$.

Although the original intent of [49] was to extend the original Hertzian model to account for dissipation due to the visco-elastic property of the materials, a useful side effect for our purposes is that it fixes the discontinuities that were present in previously discussed models. However, there still is a residual attractive force at the end of the collision, although with a much lesser magnitude. These two properties can be seen in the insets of fig. 1.8.

Another problem with this model that is not immediately obvious and that will be relevant later in this work is that the derivative of the force with respect to time is unbounded when $\xi_{i,j}(t) \to 0$, which happens at the start and end of each collision. This is due to the term $(\xi_{i,j}(t))^{\frac{1}{2}}$ in the expression of the force, since its derivative with respect to time will have the expression $(\xi_{i,j}(t))^{-\frac{1}{2}}$ in it, which is not smooth when $\xi_{i,j} \to 0$.

Nonetheless, this model is widely used because it reproduces adequately the behavior of normal collision of real particles [49], [71], [75].

1.3.2 Tangential forces

In counterpart to the normal forces, a force in a collision is said to be tangential if it is parallel to the tangential vector of the collision. Tangential force schemes are usually associated with the friction caused by the shearing motion of the surfaces of the contacting particles. Because of this, tangential force schemes are dissipative.

Cundall-Strack model

The Cundall-Strack model [17] is the *de facto* friction model for most DEM simulations. This is due to the fact that it correctly simulates the stability of a pile of grains under its own weight via internal friction, whereas other simpler models fail to reproduce this simple experiment.

If the collision between P_i and P_j starts at some instant $t_0 \in (0, \infty)$, we define

$$\zeta_{i,j}(t) := \int_{t_0}^{t} v_{i,j}^{s}\left(\tilde{t}\right) \mathrm{d}\tilde{t}. \tag{1.24}$$

The quantity $\zeta_{i,j}(t)$ is called the total tangential displacement of this collision at the instant t and its magnitude is the total distance across which the surface of one particle has dragged along the surface of the other from the start of the collision until the instant t. The importance of this value is in determining the compression of an imaginary tangential spring of stiffness $k_s \in (0, \infty)$ (the constant k_s is also referred to as the tangential stiffness of the force scheme). This tangential spring is allowed to compress until the force it exerts is equal to the Coloumb limit of $\mu \|\vec{N}_{i,j}(t)\|$, where $\mu \in [0, 1]$ is called the coefficient of friction. The force exerted by this tangential spring is the tangential force experienced by P_i in the Cundall-Strack force scheme, and it is mathematically described by:

$$\vec{T}_{i,j}(t) = -\min\left\{\left|k_s \zeta_{i,j}(t)\right|, \mu\left\|\vec{N}_{i,j}(t)\right\|\right\} \operatorname{sign}\left(\zeta_{i,j}(t)\right) \vec{t}_{i,j}(t). \tag{1.25}$$

One important observation is that this spring must be allowed to instantly decompress whenever the value of $\|\vec{N}_{i,j}(t)\|$ decreases.

The introduction of this tangential spring plays two major roles in the Cundall-Strack force scheme: it provides a smooth transition of the tangential force from a resting regime to a moving, and it allows particles to "remember" the history of the collisions in which they are taking part of, even when the relative tangential motion of these collisions has already ceased. This last characteristic is what allows the Cundall-Strack force scheme to correctly simulate the stability of a pile of grains.

1.4 Numerical integration methods for the equations of motion

Before we begin, a quick note: in what follows, we will use the term "method" or "numerical method" as synonyms to "numerical integration methods".

For each particle, the ODEs given by eqs. (1.15) and (1.16) must be solved. In this section, numerical methods to find approximations to their solutions will be presented. The description of the methods in the following will focus on eq. (1.15), because eq. (1.16) is analogous.

Before we introduce the aforementioned numerical methods, some simplifications in the notation of eq. (1.15) ought to be made. First, if it is agreed that all the variables in the equation refer to the properties of the same particle, then the sub-indexes (that specify such particle) in eq. (1.15) may be dropped, that is:

$$\sum_{\vec{F}\in\mathbb{F}} \vec{F} = m\vec{a}. \tag{1.26}$$

Then, one can assume that the vectorial nature of the equation is evident and omit the arrows above the vectorial variables:

$$\sum_{F\in\mathbb{F}} F = ma. \tag{1.27}$$

The summation symbol can be dropped and F now refers to the sum of all forces on that particle:

$$F = ma. \tag{1.28}$$

Finally, we write a as x'':

$$F = mx''. \tag{1.29}$$

Now, eq. (1.29) is a second order ODE, whereas the usual numerical methods are obtained for first order ODEs only. However, one can transform eq. (1.29) into a system of two first order ODEs by noting that $v = x'$. Then, the first order form of eq. (1.29) is

$$\begin{cases} x' = v \\ v' = \dfrac{F}{m}. \end{cases} \tag{1.30}$$

Each of these ODEs must also have an initial condition, which is usually the initial position and velocity of the particle at the start of the simulation. Formally:

$$\begin{cases} x' = v \\ v' = \dfrac{F}{m} \\ x\left(0\right) = x_0 \\ v\left(0\right) = v_0, \end{cases} \tag{1.31}$$

where $x_0, v_0 \in \mathbb{R}^2$. Finally, to avoid carrying m around, we consider that it is incorporated in the expression of F:

$$\begin{cases} x' = v \\ v' = F \\ x\left(0\right) = x_0 \\ v\left(0\right) = v_0, \end{cases} \tag{1.32}$$

1.4.1 Explicit Euler method

We start our discussion with the explicit Euler method, because it is the simplest method to integrate ODEs. In this method, the time domain is partitioned into equidistant moments, defined as

$$t_n = n\Delta t, \tag{1.33}$$

where $\Delta t \in (0, \infty)$ is the time step and $n \in \mathbb{N}$. The functions x, v and F are approximated at each time t_n and these approximations are denoted by x_n, v_n and F_n, respectively. Then, the recursive scheme for the explicit Euler method is given by

$$\begin{cases} v_{n+1} = v_n + F_n \Delta t \\ x_{n+1} = x_n + v_n \Delta t. \end{cases} \tag{1.34}$$

To find F_n, one uses the values of x_n and v_n (which are already known by the time F_n needs to be calculated) and the expression of the chosen force scheme (see section 1.3 for more information). The initialization of this method is done via the initial conditions provided by the ODE.

This is a first order method [39] and, additionally, presents a numerical artifact whose manifestation in DEM is as if energy is being injected into the system at every step it is simulating, which leads to a very poor performance in energy conservation [33]. However, this last fact is not a major problem for very dissipative systems, such as collapsing columns of grains. Since these columns are the focus of this work, this alone is not enough reason to discard the explicit Euler method. More important, though, is the fact that this method probably is the most used among the ones we will present, mostly because of its simplicity. Thus, it would be a mistake not to include it in our analyses.

1.4.2 Symplectic Euler and leapfrog methods

The symplectic Euler method is very similar to the explicit Euler method. In fact, the only difference in its recursive scheme is the use of v_{n+1} in the computation of x_{n+1}, instead of v_n. Thus, this is how its recursive scheme looks like:

$$\begin{cases} v_{n+1} = v_n + F_n \Delta t \\ x_{n+1} = x_n + v_{n+1} \Delta t. \end{cases} \tag{1.35}$$

Despite this seemingly small change, this method achieves a much better energy conservation then the explicit Euler method, even if it is still only a first order method [33].

Another method that is closely related to the symplectic Euler method is the leapfrog method. In this method, the position and the velocity of the particles are approximated at staggered moments in time. More precisely, define

$$t_{n+\frac{1}{2}} := \left(n + \frac{1}{2}\right) \Delta t \tag{1.36}$$

and

$$t_{n-\frac{1}{2}} := \left(n - \frac{1}{2}\right) \Delta t, \tag{1.37}$$

where $n \in \mathbb{N}$. Then, the leapfrog method only approximates the velocity v at the moments given by eq. (1.36), and those approximations are denoted by $v_{n+\frac{1}{2}}$ or $v_{n-\frac{1}{2}}$, respectively. Meanwhile, the positions x are still approximated only at the same points as the symplectic Euler method. Thus, the leapfrog method produces a sequence of approximations such as $v_{\frac{1}{2}}, x_1, v_{\frac{3}{2}}, x_2, v_{\frac{5}{2}}, \ldots$

With the notation now out of the way, the recursive scheme for the leapfrog method is:

$$\begin{cases} v_{n+\frac{1}{2}} = v_{n-\frac{1}{2}} + F_n \Delta t \\ x_{n+1} = x_n + v_{n+\frac{1}{2}} \Delta t. \end{cases} \tag{1.38}$$

Note that if one naively sets $v_{-\frac{1}{2}} = v_0$, then this method becomes the symplectic Euler method. As such, to fully harness the benefits of the leapfrog method, one must find v at $t_{-\frac{1}{2}}$ and set $v_{-\frac{1}{2}} = v\left(t_{-\frac{1}{2}}\right) = v\left(-\frac{1}{2}\Delta t\right)$. This, however, may not be trivial and makes the initial conditions depend on the time step.

In addition, caution must be taken in evaluating F_n. The leapfrog method supposes that F does not depend on v. If it does (as is the case for the collapses in this work), one cannot properly evaluate F_n, since the approximated value of v at t_n is not provided by the algorithm. This is a major problem that we will not address in the present work.

Under the hypotheses established above, the leapfrog method is of second order [33]. However, if F depends on v, then using $v_{n-\frac{1}{2}}$ as an approximation for v_n yields a first order method.

1.4.3 Verlet's method for damped systems

The Verlet family of integration methods was first used to predict the motion of planets over long periods of time [33]. The energy conservation properties of these methods play very well with this kind of application. Although they are named after professor Loup Verlet, similar methods were mentioned by Sir Isaac Newton himself in his Principia. Verlet actually rediscovered the method that he had "developed" in many pieces of classical literature, e.g. [76].

The recursive scheme of the most classical method of this family, which is commonly referred to as "the" Verlet method, is

$$x_{n+1} = 2x_n - x_{n-1} + F(x_n)\Delta t^2, \tag{1.39}$$

where the notation abuse $\Delta t^2 = (\Delta t)^2$ is used. Note that:

- The velocity of the particle is not explicitly integrated. Instead, when needed, it can be approximated by a centered finite difference.
- This method utilizes the last two previously calculated positions to determine the following position, while the other methods presented up to now in this work only use the position calculated immediately before.

As the leapfrog method (which is also a member of the Verlet family of methods), the classic Verlet method is of second order. However, both the leapfrog and the classic Verlet methods lose accuracy when the value of F is dependent of the velocity v of the particle. The Verlet method for damped systems solves this issue. It does so by first using the leapfrog method to find an approximation for x_{n+1}, denoted by $\hat{x}_{n+1}$. Then, it estimates v at t_n via a centered finite difference of second order. It can then calculate F_n with enough precision so that performing another leapfrog method with this newly calculated value of F yields a second order overall approximation, despite the dependence of F on v. In a recursive scheme, the method looks like this:

$$\begin{cases} \hat{v}_{n+\frac{1}{2}} & = v_{n-\frac{1}{2}} + F\left(t_n, x_n, v_{n-\frac{1}{2}}\right)\Delta t \\ \hat{x}_{n+1} & = x_n + \hat{v}_{n+\frac{1}{2}}\Delta t \\ v_n & = \dfrac{\hat{x}_{n+1} - x_{n-1}}{2\Delta t} \\ v_{n+\frac{1}{2}} & = v_{n-\frac{1}{2}} + F(t_n, x_n, v_n)\Delta t \\ x_{n+1} & = x_n + v_{n+\frac{1}{2}}\Delta t. \end{cases} \tag{1.40}$$

Even though this is a second order method, the problem associated with the initial condition $v\left(-\dfrac{1}{2}\Delta t\right)$, which was an issue of the original leapfrog method, is still present.

Chapter 2

Validation of the implementation

2.1 Validations for binary collisions

In order to certify that no mistake was made in our computational implementation of the DEM and related concepts, validations were performed. Some of these validations were split in two groups to be displayed in this work: validations for one or two particles and validations for entire collapses (with hundreds of particles). In this section, the validations for scenarios where only one or two particles were used are presented. The validations for column collapses are reported in section 2.2.

Section 2.1.1 introduces the coefficient of normal restitution, the dimensionless initial tangential velocity and the dimensionless final tangential velocity. Then, validations of the behavior of these quantities during collisions in simulations performed with our implementation are given. Some of these collisions were normal and others were tangential. In section 2.1.2, a technique for validating integration methods is presented. Afterwards, the validation of these methods are shown for simulations involving a single particle. Finally, their behavior during binary normal and tangential collision is shown and discussed.

2.1.1 Physical validations

The reference against which we validated the physical behavior of our simulations was [71]. The data includes both theoretical results and numerical results. The simulations performed for this purpose used the Kuwabara-Kono normal force scheme, described in section 1.3.1, with the Cundall-Strack tangential force scheme, described in section 1.3.2. The values for the parameters used in these force schemes, as well as for other parameters for the simulation, can be found in table 2.1.

Table 2.1: Values of the parameters used to perform the simulations used for validation of physical behavior of the binary collisions.

Particles	Normal Forces	Tangential Forces	Simulation
$\rho = 1.3 \times 10^3\,\mathrm{kg/m^3}$	$\tilde{k}_n \approx 9 \times 10^7\,\mathrm{N/m^{1.5}}$	$k_s \approx 9.4 \times 10^{10}\,\mathrm{N/m}$	$\Delta t = 2^{-23}\,\mathrm{s}$
$r = 3\,\mathrm{mm}$	$\gamma = 1.9 \times 10^2\,\mathrm{kg/(m^{0.5}\,s)}$	$\mu = 0.25$	$(\approx 1.19 \times 10^{-7}\,\mathrm{s})$

Normal collision

If P_i and P_j are a pair of particles colliding during (t_0, t_f), one can then define the coefficient of normal restitution of this collision as

$$e_n := -\frac{\left\| \vec{v}^{\,n}_{i,j}\left(t_f\right) \right\|}{\left\| \vec{v}^{\,n}_{i,j}\left(t_0\right) \right\|}. \tag{2.1}$$

The coefficient of normal restitution intrinsically entangles all the properties of the particles and all parameters associated with normal forces. This makes it a good choice to validate the physics of the simulation of a binary normal collision.

To perform this validation, a simulation of two particles is assembled such that they are initially lined up vertically and on the brink of contact (more precisely, the distance between the centers of both particles is exactly twice their radius). This alignment of the particles is necessary to ensure a perfectly normal collision. One of the particles (in our case, the bottommost) is fixed in place for the duration of the simulation, i.e. its position does not change. Then, an initial velocity is prescribed to the particle that was not fixed. The direction of this velocity is towards the fixed particle, i.e. in the normal axis joining their centers. The magnitude of this initial velocity will be denoted simply by v_i. Then, the simulation is allowed to start and is stopped at some moment after the collision has ended. The final velocity of the non-fixed particle is measured and the coefficient of normal restitution is calculated. This process is repeated for a range of values for v_i. The results for the different numerical methods presented previously in section 1.4 can be seen in fig. 2.1.

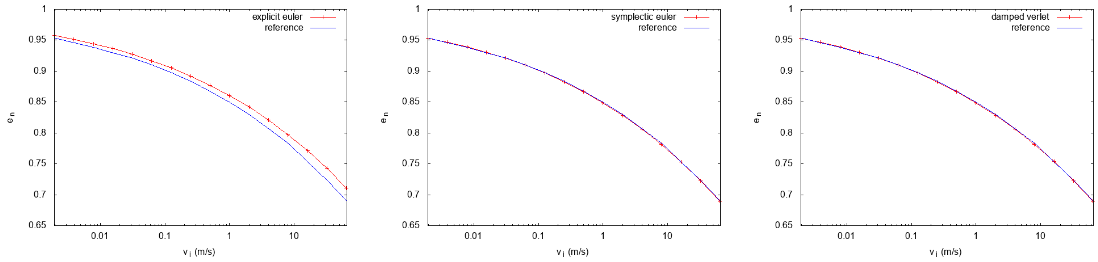


Figure 2.1: Physical validation of normal collisions through the coefficient of normal restitution. The values of the parameters can be found in table 2.1. The reference curve was obtained in [71].

Note that in the leftmost graph of fig. 2.1, the curve of the simulation performed via de explicit Euler method is significantly above the reference curve. This behavior is expected. As mentioned before in section 1.4.1, the explicit Euler method has a tendency to inject energy into systems. Since our choice of parameters leads to a system with relatively little dissipation, the injected energy cannot be dissipated fast enough and accumulates in the system, which then manifests as an increase in the elasticity of the collision.

Tangential collision

Let P_i and P_j be a pair of particles colliding during (t_0, t_f). If there exists a $t \in (t_0, t_f)$ such that $v^s_{i,j}\left(t\right) \neq 0$, then this collision is dubbed a tangential collision. For these collisions, it is useful to define the dimensionless initial and final velocities.

These are

$$\psi_i := \frac{v_{i,j}^s(t_0)}{v_{i,j}^n(t_0)} \tag{2.2}$$

and

$$\psi_f := \frac{v_{i,j}^s(t_f)}{v_{i,j}^n(t_0)}, \tag{2.3}$$

respectively. Then, one technique used in [71] to validate tangential collisions is the comparison of the curve ψ_f *versus* ψ_i of simulations to the same curve obtained from experimental data in [25]. We will validate our simulations by comparing the curve ψ_f *versus* ψ_i from our simulations against the ψ_f *versus* ψ_i curve presented in [71].

To perform this validation, a simulation of two particles is assembled in such a way that they are initially lined up horizontally and on the brink of contact. Then, an initial velocity of magnitude 1 m/s and in the direction of the positive x-axis is prescribed to the leftmost particle, while the rightmost particle is initially at rest. After this, the rightmost particle is shifted vertically by up to the sum of the radii of both particles. In doing so, the particles stop contacting. This if fixed by shifting again the rightmost particle to the left just enough that it contacts the leftmost particle once again. This setup can be seen in fig. 2.2.

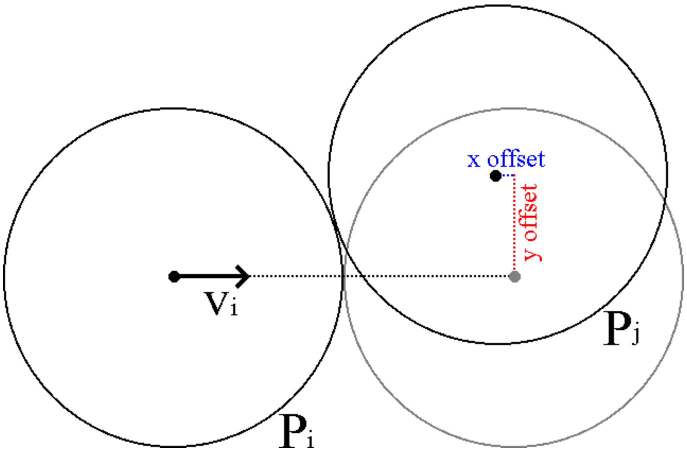


Figure 2.2: Setup for validation of tangential collisions.

The simulation is allowed to start and is stopped at some moment after the collision has ended. During this collision, the positions, velocities and angular velocities of both particles are recorded at each instant. With this information, t_0 and t_f are determined for this collision. Finally, ψ_i and ψ_f can be calculated. This process is repeated several times but the position of the leftmost particle is shifted downwards, so that the data can be collected for collisions with different degrees of obliqueness. The results for each numerical method can be seen in fig. 2.3.

2.1.2 Method validations

One common technique to validate the implementation of a numerical method used to perform the integrations of a given set of equations is to check the order of the implementation against the theoretical order of the method.

A way to use this technique requires one to run several simulations with different time steps but otherwise identical parameters. It is imperative that these simulations

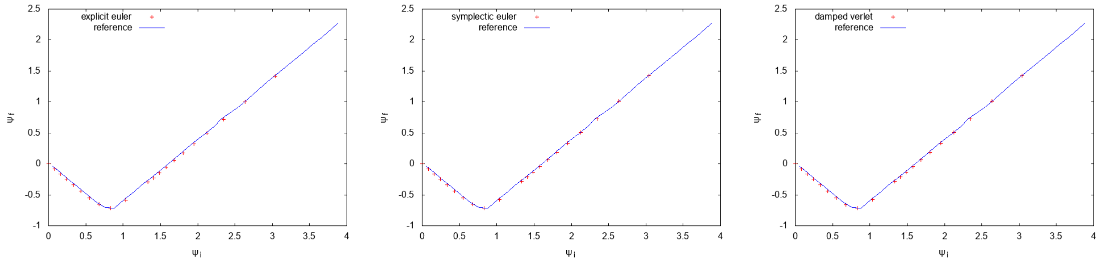


Figure 2.3: Physical validation of a tangential collision. Each red point represents a single collision and their obliqueness increases with ψ_i. The values of the parameters can be found in table 2.1. The reference curve was obtained in [71].

each last for the exact same physical time. Then, for each simulation, the value of Δt and of the last value assumed by some variable which was integrated via the chosen method need to be recorded as a pair. Let us call these values as $z_{\Delta t}$. One must then choose a reference value $\bar{z}$, that is, a value against which the other values of $z_{\Delta t}$ will be compared. Afterwards, calculate $|\bar{z} - z_{\Delta t}|$ (which will be simply called "error" in the plots showing order validation) for each Δt and fit this data to a power law of the form

$$a\Delta t^b, \tag{2.4}$$

via a and b. The value of b is (up to the margin of error) the order of the implementation. If b is sufficiently close to the theoretical order of the method, one can consider the implementation validated. Otherwise, either some necessary hypotheses regarding the chosen method was not observed or the implementation is faulty.

However, some care must be taken to avoid common pitfalls:

(i) The choice of the data interval where the fit is to be performed. For values of Δt that are not sufficiently small, the apparent order of the method can fluctuate. This happens because the grid in which the functions (in this case, forces) are evaluated is still too coarse when Δt is big, so the method may not capture the adequate behavior of the functions that a finer grid will capture.

(ii) On the other hand, one must avoid using values of Δt that are exceedingly small. This is because the reduction in the time step causes the number of steps necessary to fully simulate the same physical time to increase. If the number of steps in a simulation is too big, the accumulation of errors due to floating point arithmetic surpasses the truncation error of the method.

(iii) Ideally, the value for $\bar{z}$ should be chosen from an analytic solution to the problem. However, analytic solutions have not been found for most of the problems we simulate. In this case, one can choose a value among the $z_{\Delta t}$. When doing so, one should choose the value associated with the smallest Δt that has not yet been severely contaminated by accumulated floating point errors. This might not be trivial.

Position of a single particle

We start by validating the order of a single coordinate of the position $\vec{x}$ of a lone particle on which a force acts. More precisely, the force acting on the particle has constant direction and magnitude $F(t) = 1 + \sin(t)$, measured in Newtons. The particle is initially at rest. Other parameters for the simulation can be found in table 2.2.

Table 2.2: Values of the parameters used to perform simulations for order validation of the linear and angular motion of single particles.

Particles	Simulation
$\rho = 2.500 \times 10^3\,\mathrm{kg/m^3}$	$t_f - t_0 = 2^{-13}\,\mathrm{s}$
$r = 5 \times 10^{-1}\,\mathrm{m}$	$\left(\approx 1.22 \times 10^{-4}\,\mathrm{s}\right)$

The reference value for x is the value of the analytic solution to eq. (1.15) at $t = 2^{-13}\,\mathrm{s}$. Any initial position for the particle may be used as initial condition to eq. (1.15), and, since the particle is at rest, $x'(0) = 0$. The plotted results are shown in fig. 2.4.

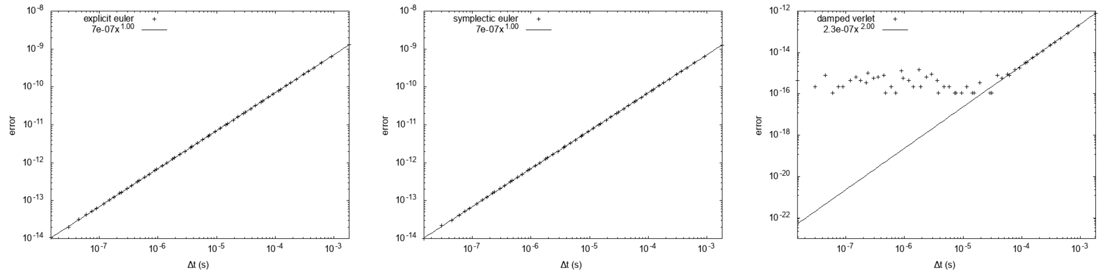


Figure 2.4: Order validation of the position of a single particle. The values of the parameters can be found in table 2.2.

Note that both the explicit and symplectic Euler are $\mathcal{O}(\Delta t)$ methods, whereas the Verlet method for damped systems is $\mathcal{O}(\Delta t^2)$. Also, in the rightmost plot of fig. 2.4, there are points which form a seemingly random pattern as Δt decreases. These points, which appear when the time step is below, approximately, $1 \times 10^{-4}\,\mathrm{s}$, are caused by the accumulation of floating point errors surpassing the truncation error of the method. It only happens in that specific plot because the Verlet method for damped systems is a second order method, which means that its truncation error decreases faster then the other two methods plotted (note the y-axis range of the plots).

Accumulated angle of rotation of a single particle

Even though the law that describes the angular motion of particles is similar to that of linear motion, they both are individually implemented in our software, which means that the order of the method performing them must be validated separately.

In this case, the variable whose order will be validated is the accumulated angle of rotation θ. A simulation with a single particle at rest and with no angular motion is assembled. A prescribed torque (which, as an exception to what was said before, does not come from a force) imposed on this particle. The magnitude of this torque is $\tau(t) = 1 + \sin(t)$, measured in Newtons. Other parameters for the simulation can be found in table 2.2. Keeping the analogy to the linear case, the reference value for θ is the value of the analytic solution to eq. (1.16) at $t = 2^{-13}\,\mathrm{s}$. The plotted results are shown in fig. 2.5.

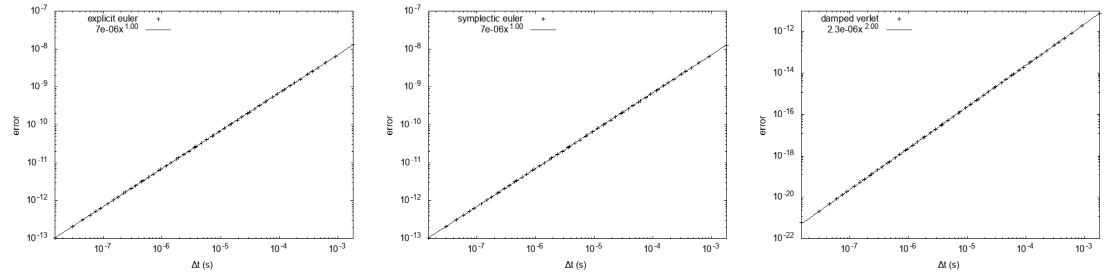


Figure 2.5: Order validation of the angular displacement of a single particle. The values of the parameters can be found in table 2.2.

Note that all the methods behave similarly as in the previous section. Also, the random pattern found in fig. 2.4 was that caused by the accumulation of errors related to the floating point format is not present in fig. 2.5. This is due to the fact that the initial angle of the simulation is $\theta = 0$, while in the validation for the linear order, the initial position was $x > 1$. By design, the floating point format is more precise for values near 0 than it is in other intervals, which justifies this difference.

Position in a binary normal collision

In order to better understand how the Kuwabara-Kono force scheme (see section 1.3.1) interacts with each method presented in section 1.4, an order validation was performed using data from a simulation of a binary collision. The parameters used in the simulations necessary for this validation can be found in table 2.3.

To perform this validation, a simulation was assembled where two particles were lined up vertically. Both particles had no angular motion and the bottommost particle was fixed, i.e. its position did not change. The other particle had an initial velocity with magnitude of 1 m/s and in the direction of the particle in the bottom. This setup guarantees a normal collision. The simulation is then allowed to run for 2^{-16} s ($\approx 1.5 \times 10^{-5}$ s). The variable whose order was checked is (a single coordinate of) the position $\vec{x}$ of the moving particle. The results are presented in fig. 2.6.

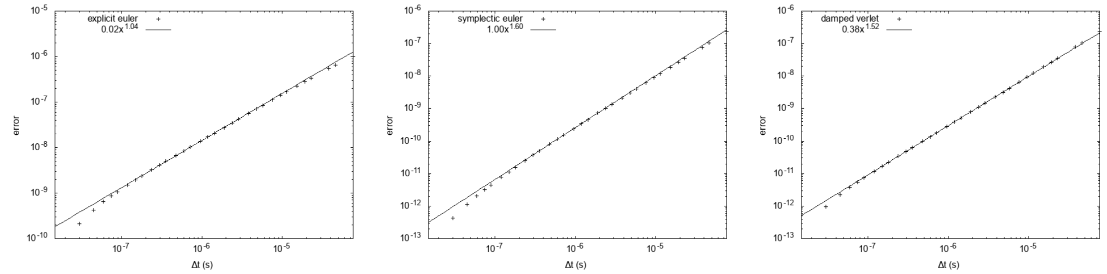


Figure 2.6: Order validation of the position of a particle involved in a binary normal collision. The values of the parameters can be found in table 2.3.

There are two unexpected behaviors in fig. 2.6. The first is the value of b in the center plot of fig. 2.6, i.e. the symplectic Euler method. As mentioned before in section 1.4.2, the symplectic Euler method is of first order. However, the value of b found for our implementation is above this, at $b \approx 1.6$. This can be explained by the fact that the symplectic Euler and the leapfrog method differ only by the fact the former is initialized with $\vec{v}(0)$ and the latter is initialized with $\vec{v}(-0.5\Delta t)$, while their orders are one and two, respectively. Since there are no forces acting on each particle at the start of this simulation, $\vec{v}\,(0) = \vec{v}\,(-0.5\Delta t)$. Then, in the particular case of this validation, the symplectic Euler and leapfrog method are one and the same. This behavior can be seen in fig. 2.7, where the validation performed in section 2.1.2 was repeated for the symplectic Euler method, but the initial velocity of the particle was set to $\vec{v}\,(-0.5\Delta t)$ (which could be easily done because the analytic solution to that problem was known).

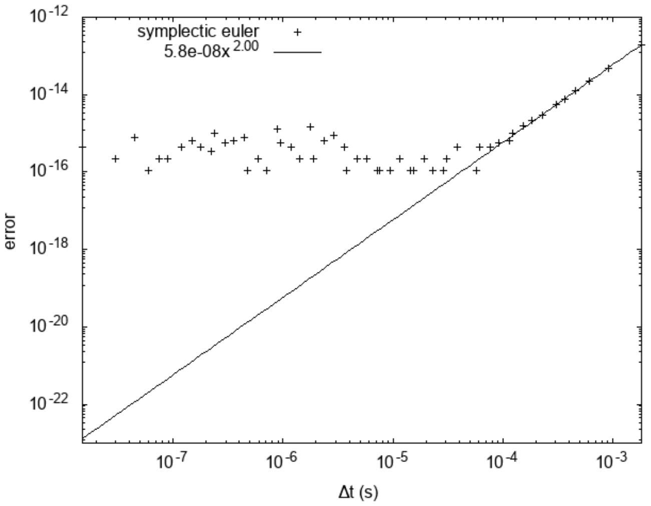


Figure 2.7: The velocity of the particle was set to $v\,(-0.5\Delta t)$ instead of $v\,(0)$ and then simulated with the code responsible for the symplectic Euler method. The result indicates second order accuracy with Δt. This illustrates how the symplectic Euler and leapfrog methods differ only by their initialization process.

This leads us to the second unexpected behavior: the value of b for the Verlet method for damped systems and for the symplectic Euler method are both less than two. For the former, $b \approx 1.52$ and for the latter, $b \approx 1.6$. For the symplectic Euler method, the explanation is quite simple: in the collision that ensues, the forces involved depend on the velocities of the particles, which breaks one of the hypothesis (presented in section 1.4.2) that guarantees second other for the leapfrog method.

The same explanation cannot be used to justify the behavior of the Verlet method for damped systems, since it explicitly addresses forces with velocity dependence. We have strong evidence (see chapter 3) that this behavior is caused by the unbounded derivative of the normal force at $\xi = 0$, which was discussed in section 1.3.1 and can be visualized in fig. 1.8. We have nevertheless fixed this problem with a force scheme which closely matches the Kuwabara-Kono force scheme but still mantains order two when paired with the Verlet method for damped systems, which will be introduced in chapter 5.

Note that the positions of these particles must be such that the distance between their centers is exactly $2r$. This causes the particles to be on the brink of contact, and is necessary. Otherwise, the different time step sizes, which are not integer multiples of each other, would cause the particle to traverse the distance to the other particle in such a way that the first overlap between both particles does not consistently diminishes with the time step. This causes an erratic behavior in the data which makes determining the order of the implementation harder. One example of this is shown in fig. 2.8.

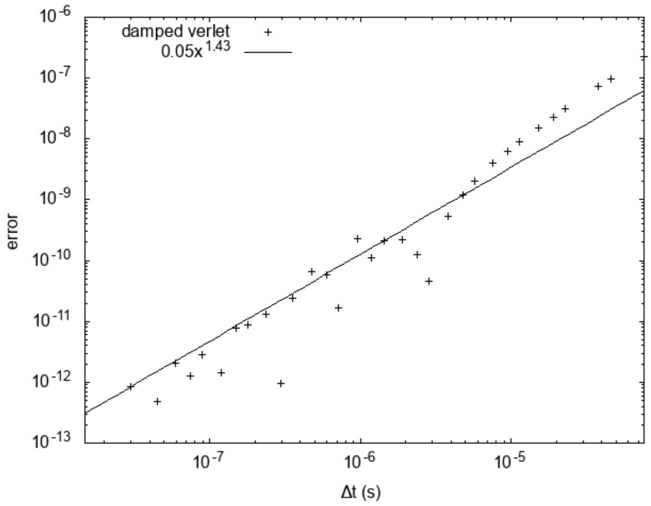


Figure 2.8: An example of an order validation where particles are not initially touching.

Position in a binary tangential collision

The last validation performed aims to understand the interaction between each of the methods presented in section 1.4 and the combination of the Kuwabara-Kono normal force scheme (see section 1.3.1) with the Cundall-Strack tangential force scheme (see section 1.3.2). To do so, a simulation of a tangential collision was assembled. The parameters of this simulation can be found in table 2.3

Table 2.3: Values of the parameters used to perform the simulations used for order validation of binary normal and tangential collisions.

Particles	Normal Forces	Tangential Forces	Simulation
$\rho = 1.9300 \times 10^4\,\mathrm{kg/m^3}$	$\tilde{k}_n \approx 4.4 \times 10^{10}\,\mathrm{N/m^{1.5}}$	$k_s \approx 2.4 \times 10^6\,\mathrm{N/m}$	$t_f - t_0 = 2^{-16}\,\mathrm{s}$
$r = 1\,\mathrm{m}$	$\gamma = 5 \times 10^8\,\mathrm{kg/(m^{0.5}\,s)}$	$\mu = 0.6$	$\left(\approx 1.5 \times 10^{-5}\,\mathrm{s}\right)$

This simulation was constructed as follows:

- A particle at rest with no angular motion was created in the simulation. Let index this particle by i.
- A second particle, indexed by j, was created at the position $\vec{x}_j = \vec{x}_i + \left(\sqrt{3}r, r\right)$. This position puts the center of the particles at a 4.5° angle and makes the particles be on the verge of contact (the reason why this is necessary is explained in section 2.1.2). This particle also has no angular motion.
- An initial velocity of $\vec{v}_j = -\left(1, 0\right)\mathrm{m/s}$ was imposed on particle j.
- The simulation was then allowed to run for $2^{-16}\,\mathrm{s}$ ($\approx 1.5 \times 10^{-5}\,\mathrm{s}$).

Once again, the variable whose order was to be validated was the position. More specifically, the first coordinate of the position of particle j. The results are presented in fig. 2.9. The same behavior described in fig. 2.6 can be observed, and the same explanations given in section 2.1.2 apply.

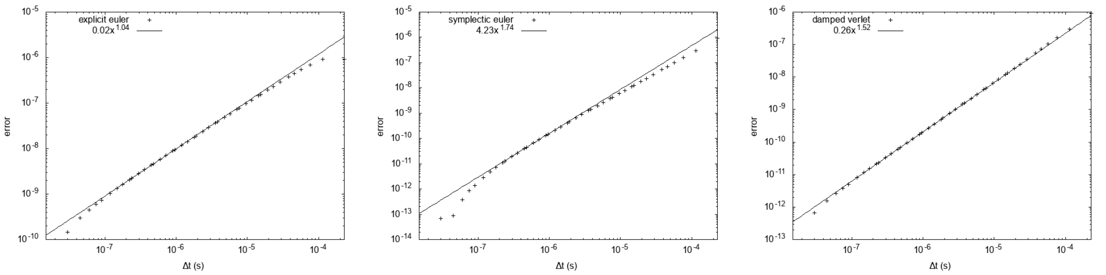


Figure 2.9: Order validation of the position of a particle involved in a binary tangential collision. The values of the parameters can be found in table 2.3.

2.2 Results for granular collapses

In this section, results that validate the granular collapses in our simulations will be presented. Before that can be done, some definitions need to be introduced. We will also take this opportunity to talk about the setup of a granular column.

All the column collapse simulations in this work use the explicit Euler method (see section 1.4.1) and the Hertz normal force scheme with linear damping (see section 1.3.1) together with the Cundall-Strack tangential force scheme (see section 1.3.2). The values of all parameters can be found in table 2.4. Note that in any single simulation, the radii of the particles differ slightly, to prevent the formation of a crystallized lattice. Because of these, the value of r provided in table 2.4 is just the average value of r among all particles of the simulation.

Table 2.4: Values for the parameters in simulations performed for this section.

Particles	Normal Forces	Tangential Forces	Simulation
$\rho = 1.300 \times 10^3\,\mathrm{kg/m^3}$	$\tilde{k}_n \approx 7.3 \times 10^7\,\mathrm{N\,m^{1.5}}$	$\dfrac{\tilde{k}_t}{\tilde{k}_n} = 2/7$	$\Delta t = 5 \times 10^{-7}\,\mathrm{s}$
$r = 3\,\mathrm{mm}$	$\gamma = 3.5 \times 10^1\,\mathrm{kg/s}$	$\mu = 0.5$	

2.2.1 Definitions

The initial bi-dimensional column of particles is roughly an axis-aligned rectangle. We say roughly because the column is composed of particles, whose geometry is circular and, as such, cannot form a perfect rectangle. The column is defined by its height h (i.e., its size in the y-axis) and its basal length L (i.e., its size in the x-axis). With those quantities, one can define the aspect ratio of the initial column as

$$a = \frac{H}{L}. \tag{2.5}$$

The final deposit of a collapse is characterized by two commonly measured properties: the head height and the run out distance. The former can be written as h_∞ and is the maximum vertical distance between all pair of particles that are in contact with the vertical wall. The latter is the maximum horizontal distance between all pair of particles that touch the base of fixed particles and remain part of the main body of the collapse, i.e. are not alone or in isolated small clumps. The run out distance will be denoted by L_∞. These two properties are illustrated in fig. 2.10

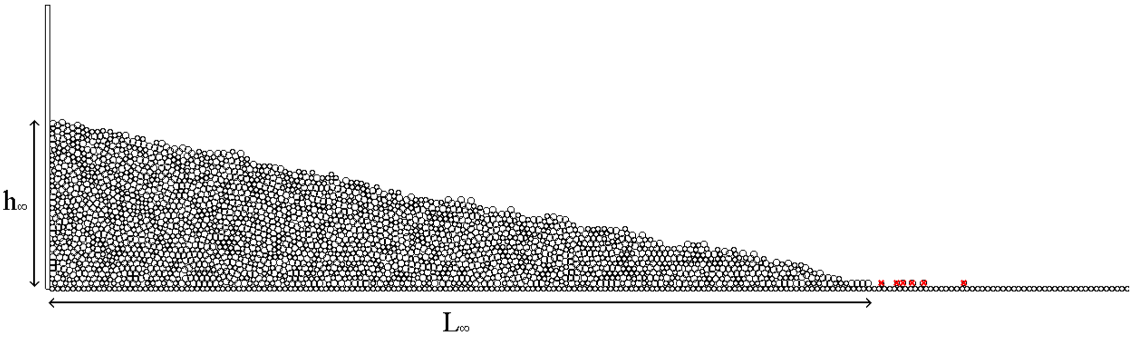


Figure 2.10: Visualization of h_∞ and L_∞. The particles crossed in red are not part of the main body of the collapse and, therefore, are not counted in the determination of the run out distance. This image is the result of the collapse of the column depicted in fig. 2.11.

One can then define the dimensionless counterpart to the head height as

$$\bar{h} = \frac{h}{h_\infty} \tag{2.6}$$

and, similarly,

$$\bar{L} = \frac{L_\infty - L}{L_\infty} \tag{2.7}$$

to the dimensionless run out distance. These dimensionless quantities are often called normalized head height and normalized run out distance.

2.2.2 Setup of initial conditions

A box, open at the top and with walls high enough to avoid spillage, is created. The floor of this box is made of particles which are fixed in place, i.e., their position does not change during the simulation. Their radius is equal to the infimum of the interval from where the radii of the particles that will fill the box are taken. This procedure serves the purpose of providing a more realistic friction between the flowing particles and the surface. [41]

The box was filled by creating an organized grid of mobile particles with positions slightly perturbed to avoid crystallization. The simulation is then allowed to run until the total kinetic energy of each particle was less than 1×10^{-6} J. Afterwards, some particles at the top of the box were trimmed so that the desired aspect ratio (see eq. (2.5)) is achieved. Since this removal of the top particles changes the forces acting on the particles below them, the simulation was run again for 1000 steps so that the decompression can take place before the actual collapse began.

Before initiating the collapse, by the removal of the right wall which held the particles in place, we extended the “floor of fixed particles” far enough to prevent particles from going out of bounds. An example of the final result of this setup is displayed in fig. 2.11.

2.2.3 Results

Ten granular collapses were run, each with approximately 1000 particles and with aspect ratios varying from 0.1 to 8. Additionally, one larger simulation was run with approximately 2500 particles and aspect ratio 1.

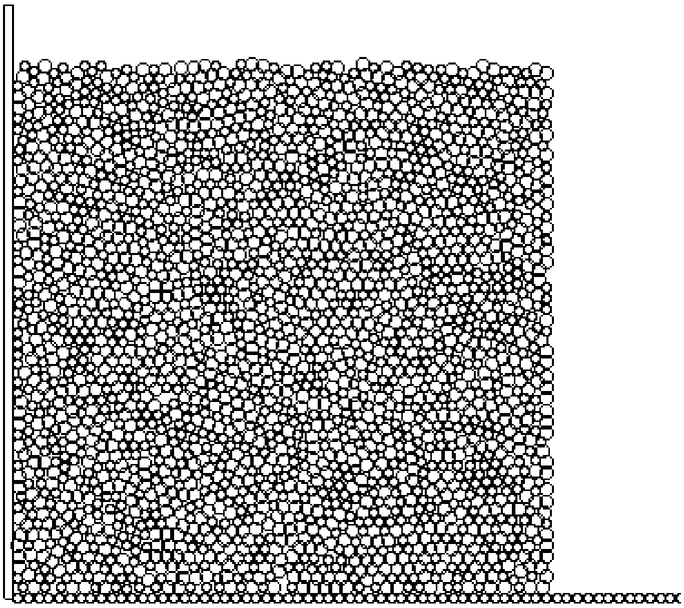


Figure 2.11: The initial condition of a column composed of roughly 1000 particles with $a = 1$. The right wall has already been removed, so the column is ready to collapse. The values of the parameters can be found in table 2.4. This image is the initial condition of the deposit depicted in fig. 2.10.

Deposit profiles

The profile of a deposit consists of the shape formed by its outermost layer of particles. In order to avoid drawing large number of particles when displaying multiple surface profiles together, a bezier-smoothed line generated from the positions of the centers of the particles in the profile is used, as shown in fig. 2.12.

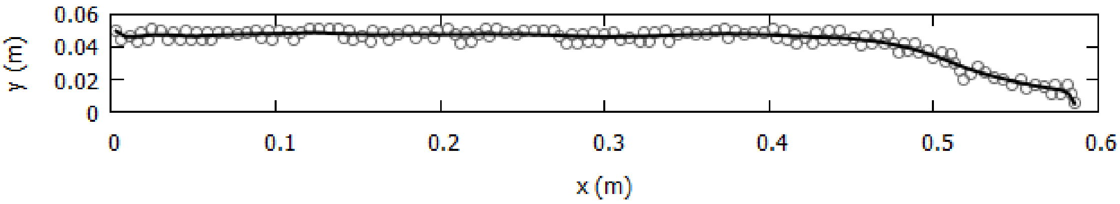


Figure 2.12: Illustration of a deposit surface and its bezier-smoothed surface line. All surface profiles were treated this way.

Depending on the aspect ratio of the initial column, the profile of a deposit can form one of two main shapes: trapezoidal and triangular. The trapezoidal archetype is characterized by virtually no motion of the particles on the surface close to the left wall. This results in a relatively flat top followed by a inclined slope. The aspect ratio of the initial column seems to affect where the flat profile stops and the slope begins. In our simulations, this has happened only to the columns with aspect ratio 0.5 or below, as seen in fig. 2.13.

On the other hand, in the triangular profile all particles in the surface of the column undergo motion, which eliminates the flat top, leaving only a slope. In our simulations, this has happened to all columns with aspect ratio above 0.5. Also, as shown in fig. 2.14, the triangular archetype, when appropriately normalized by

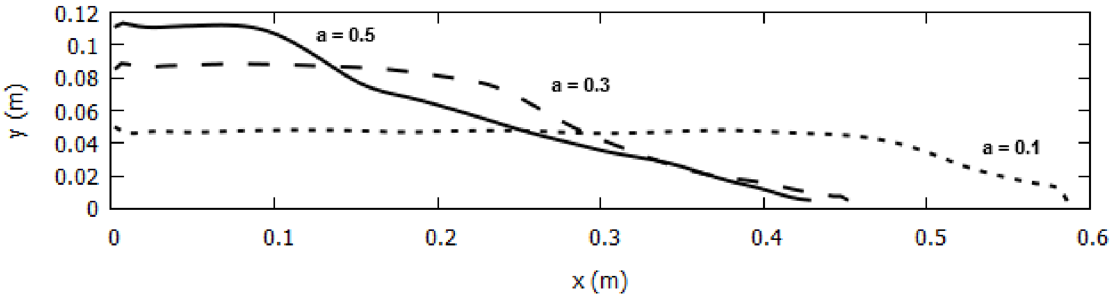


Figure 2.13: The 3 trapezoidal deposit profiles that formed from the 11 simulations that we ran. The values of the parameters use in these simulations can be found in table 2.2.

their respective head height (h_∞) and run out distance (L_∞), seems to converge to a "master" curve.

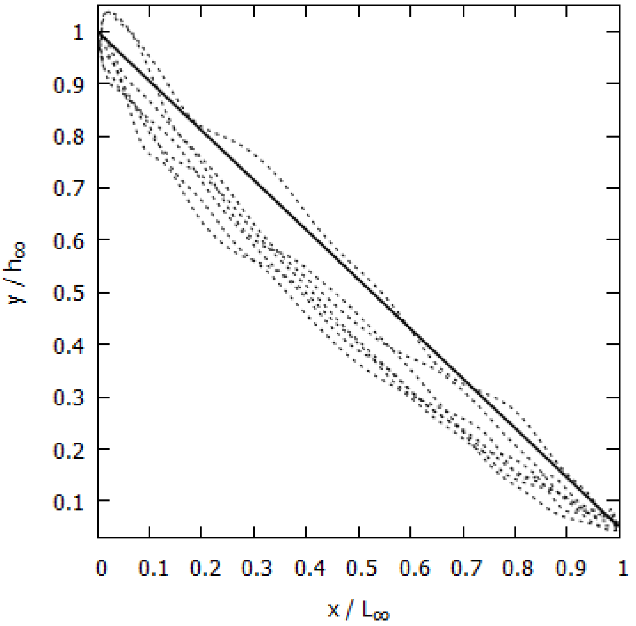


Figure 2.14: The normalized final profile of the collapses that follow the triangular archetype. Note that they seem to converge into a "master" diagonal curve, plotted as a heavy line. The values of the parameters use in these simulations can be found in table 2.2.

Comparison with physical results

We will use the experiments reported in [5] as a reference against which our results will be compared. The results in [5] indicate that both the dimensionless head height and dimensionless run out distance behavior change when the deposit profile changes from trapezoidal to triangular profiles. It is also shown that this transition happens when the aspect ratio is, roughly, 1.5, i.e., when $a < 1.5$ the deposit profile is trapezoidal and when $a > 1.5$ it is triangular.

In our simulations, such finding was not able to be replicated. Instead, the tipping point for the transition between profiles was $a = 0.5$. This discrepancy could be caused by the lack of friction in the vertical wall as well as the relatively low amount of particles in the simulations and remains to be further investigated.

The results in [5] indicate that the head height follows a power law scaling with the aspect ratio, given by

$$\frac{h}{h_\infty} \sim \lambda_1 a^{\alpha_1}, \tag{2.8}$$

for a big enough to cause a triangular profile, where $\alpha_1 \approx 0.6$ and λ_1 is a positive constant that depends on the material, mean radius etc. When fitting our data to a power law, we found $\lambda_1 = 1.469 \pm 0.074$ and $\alpha_1 = 0.606 \pm 0.029$, which indicates that our simulations seem to be working correctly. This result can be seen in the left plot of fig. 2.15.

On the other hand, [5] suggests that, for $a > 2$, the normalized run out distance scales as

$$\frac{(L_\infty - L)}{L_\infty} \sim \lambda_2 a^{\alpha_2}, \tag{2.9}$$

where $\alpha_2 = 0.9 \pm 0.1$ and λ_2 is another constant that depends on the properties of the particles. Our best fit was $\lambda_2 = 1.918 \pm 0.028$ and $\alpha_2 = 0.829 \pm 0.008$, which falls within the error margin. This result result can be seen in the right plot of fig. 2.15.

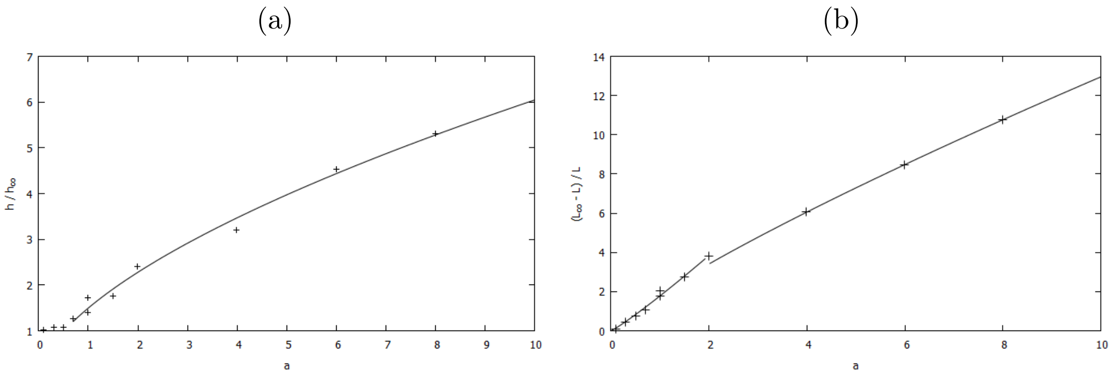


Figure 2.15: Normalized head height (a) and normalized run out distance (b) as a function of aspect ratio. The values of the parameters use in these simulations can be found in table 2.2.

Note that there are two disjoint lines composing the plot in fig. 2.15 (b). This is because the mechanism of collapse of the column changes as the aspect ratio increases. The column fails through shearing when the aspect ratio is small ($a < 2$), with the top corner continuously deforming until it becomes composed of only a few particles. Meanwhile, for large ($a > 2$) aspect ratios, the top part of the column falls mostly down and is then redirected horizontally through collision. As stated before, the results obtained above are for this second case, when $a > 2$.

Part II

Order of timestepping of DEM with a nonlinear force model

Introduction

As was seen in the previous part, the essence of the DEM algorithm lies in the constitutive models of the interaction forces that happen during the contact of particles. In standard dry flows of granular materials, two kinds of forces are normally considered in the description of the contact [71]: elastic forces, that model the restoring effects in the contact, and dissipative forces, that dissipate energy in the contact. The most important dissipative force in granular materials is friction, that is associated to the relative tangential motion of two particles. There are also dissipative forces associated to the normal relative motion of the particles, along the line joining their centers [71]. Friction forces, which are difficult to model [62], are not the the focus of the remainder of this work and will not be discussed any further.

The simplest normal forces that are used in DEM simulations are given by the combination of a Hooke-like elastic force and a linear dissipation force [71]. The appeal of this model, besides its simplicity, is the fact that many analytical solutions related to the properties of the contacts between particles can be derived and they can be used to guide the correct definition of the simulation parameters. Nevertheless, a nonlinear model based on the combination of a Hertzian elastic force [34] with a nonlinear dissipation force [49] matches more closely the data obtained in experimental results with a large range of particle properties [71], [75]. This model, which is known as the Kuwabara-Kono force model, is at the heart of the first chapter of this part.

The equations of motion of the particles, which derive from the application of Newton's laws of motion, in combination with the contact forces and other forces such as gravity, for example, are often integrated in time with symplectic integrators [33], [79], [87]. These methods, which were originally inherited from the Hamiltonian systems found in molecular dynamics simulations, have also proven to be suitable for dissipative systems such as granular materials. Other methods can also be successfully used in DEM, such as Runge-Kutta [55], Euler and even implicit methods [11]. Because of the smallness of the time-step that is needed for the DEM to be physically sound [71], several stability studies were carried out to establish appropriate bounds for the selection of the time step [10], [47]. However, it seems that the order analyses of the integration methods has not attracted the attention of the community. This is precisely the aim of this work.

As noted on section 1.3, all normal force schemes considered in this work have some sort of irregularity in the beginning and the end of a collision. In particular, the Kuwabara-Kono force model (see section 1.3.1) introduces dissipation in the collisions via a term that is proportional to the square root of the overlap between the particles [49]. The local truncation error of the numerical integration algorithm related to this term is, therefore, singular whenever the overlap between the particles is zero, that is, in the beginning and in the end of the collisions. This can be seen in the insets of fig. 1.8. Therefore, it is not expected that the order of the usual integration schemes used on DEM will be preserved when they are applied in conjunction with the Kuwabara-Kono force model. For instance, in section 2.1.2, it was noted that the Verlet method for damped systems (see section 1.4.3) failed to deliver order two when subjected to a normal collision with the Kuwabara-Kono force scheme. This fact seems to have gone unnoticed in the literature until very recently. In fact, it was only when we were writing our findings up that we came across the recent work [40], in which the authors identified that the order of integration was reduced when the Kuwabara-Kono force model was used to simulate problems similar to the Newton's cradle. However, besides its identification, no formal justification for the reduction of the order of the methods was presented in [40], and a numerical, non-singular technique to account for dissipation in the system was proposed in order to allow that the full order of the integration methods could be achieved in their system.

In the first chapter of this part, we study the order reduction phenomenon that is observed when the Verlet [33] algorithm is used in DEM simulations of granular materials using the Kuwabara-Kono force model model. A simplified model, retaining the essential features linked to this phenomenon, is proposed and analysed in detail. We identify that the singular behavior of the first derivative of the non-linear dissipation is the cause of the order reduction phenomenon. We then propose a regularization of the Kuwabara-Kono force model using the concept of mollifiers [23]. This regularized force model, which is non-singular in the beginning and in the end of the collisions, allows the numerical integration to occur with the full theoretical order of the Verlet method.

Inspired by this difference in order, we inquired what the effects of the order of the method were on a macroscopic flow. As per usual, we used the collapse of a granular column under the influence of gravity as our model flow. In the second chapter of this part, we study the effects of different integration methods on the trajectories of individual particles.

Chapter 3

The order of the leapfrog method with the Hertz-Kuwabara-Kono force scheme

3.1 Description of the problem

3.1.1 Normal collision between two particles

Consider two spherical particles, say P_1 and P_2, close to each other in a plane perpendicular to the direction of gravity, in such a way that they evolve to a purely normal collision, that is, a collision in which the relative motion of the particles happens precisely along the axis defined by the centers of the particles. In this case, the collision is free of any tangential forces.

The collision simulation is assembled in such a way that the two particles are lined up along a reference axis. Both particles have no angular motion and one of the particles is fixed, i.e. its position is not evolved in time. The center of the other particle is placed at a distance $r_1 + r_2$ of the center of the fixed particle, with an initial velocity of magnitude $1\,\mathrm{m/s}$ and in the direction of the fixed particle. This setup guarantees a purely normal collision. The simulation is then allowed to run for $105 \times 2^{-13}\,\mathrm{s}$ ($\approx 1.2 \times 10^{-2}\,\mathrm{s}$). The accumulated integration error of the position of the moving particle is calculated and used to determine the order of the Verlet method in eq. (1.40). Two such simulations were run, one using the Kuwabara-Kono force model given in section 1.3.1, with parameters given in in table 3.1, and another one where the purely elastic Hertz force model was used instead, i.e. $\gamma = 0$ in eq. (1.23). The results are presented in fig. 3.1.

Table 3.1: Values of the material and model parameters used to simulate a binary normal collision between two particles.

Particles	Normal Forces	Simulation
$\rho = 19300\,\mathrm{kg/m^3}$	$\tilde{k}_n \approx 4.4 \times 10^{10}\,\mathrm{N/m^{1.5}}$	$t_f - t_0 = 105 \times 2^{-13}\,\mathrm{s}$
$r = 1\,\mathrm{m}$	$\gamma \approx 1.1 \times 10^{9}\,\mathrm{kg}/\left(\mathrm{m^{0.5}s}\right)$	$\left(\approx 1.2 \times 10^{-2}\,\mathrm{s}\right)$

The results of the order analysis depicted in fig. 3.1 reveal an unexpected behavior. When the purely elastic force model is used, the accumulated error in the position decreases as $\mathcal{O}(h^2)$, as it is expected when eq. (1.40) is used to integrate the motion of the particle. However, when the full Kuwabara-Kono force model was used, the curve approximating the accumulated error in the position decreases as $\mathcal{O}(h^{1.5})$, which does not agree with the expected order of the Verlet algorithm in eq. (1.40).

The inclusion of the dissipation term in the model penalized the order of the Verlet algorithm in eq. (1.40). The cause of this order reduction, therefore, has to lie in the $\xi^{1/2}$ term in eq. (1.23): at the beginning and in the end of the collision, when $\xi = 0$, the term $\xi^{1/2}$ no longer satisfies the Lipschitz condition, which is a crucial hypothesis in the Picard-Lindelöf existence and uniqueness theorem [16] and is used in the order analysis of several integration methods [4]. It seems that this issue has not yet been observed nor discussed in the literature. Therefore, in order to further study the origin of this order reduction, we propose to investigate a simplified problem which retains the same problematic feature present in the force model given by eq. (1.23).

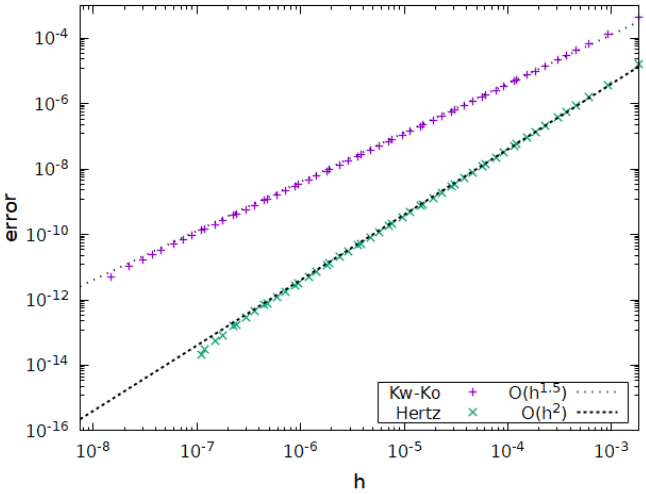


Figure 3.1: Order analysis of the position of the moving particle involved in the binary normal collision. The values of the parameters can be found in table 3.1. The usual Hertz model has no damping term, which means that $\gamma = 0$ in eq. (1.23) in this case.

3.1.2 A simplified problem

In order to formally analyze the order reduction of eq. (1.40) caused by the dissipation term in eq. (1.23), we devised a simplified problem which retais the essential feature we believe is behind the issue identified in section 3.1.1. This simplified model is given by

$$\begin{cases} y''(t) = y(t)^q; \\ y(0) = 0, y'(0) = 1, \end{cases} \tag{3.1}$$

where $q \in (0, 1)$. Note that the right hand side of the differential equation in eq. (3.1) has unbounded derivative at $t = 0$. It is also important to point out that, since the right-hand side of eq. (3.1) has no explicit dependency on y', the Verlet method in eq. (1.40) when applied to eq. (3.1) is equivalent to the standard leapfrog method

[79]. Defining $v = y'$, it gives:

$$\begin{cases} v_{n+\frac{1}{2}} & = v_{n-\frac{1}{2}} + y_n^q \Delta t; \\ y_{n+1} & = y_n + v_{n+\frac{1}{2}} \Delta t. \end{cases} \tag{3.2}$$

To verify that the penalization of the order is still present in eq. (3.1) for $q \in (0, 1)$, the simplified problem was numerically integrated for different values of q using the leapfrog method in eq. (3.2), initialized via the Heun method [79]. The results are presented in fig. 3.2. One can observe that the order of 1.5 obtained in fig. 3.1 also holds in the simplified model eq. (3.1) when $q = 0.5$. In fact, the results presented in fig. 3.2 indicate that the numerical solution of eq. (3.1) via the leapfrog method gives an error decreasing as h^{1+q}.

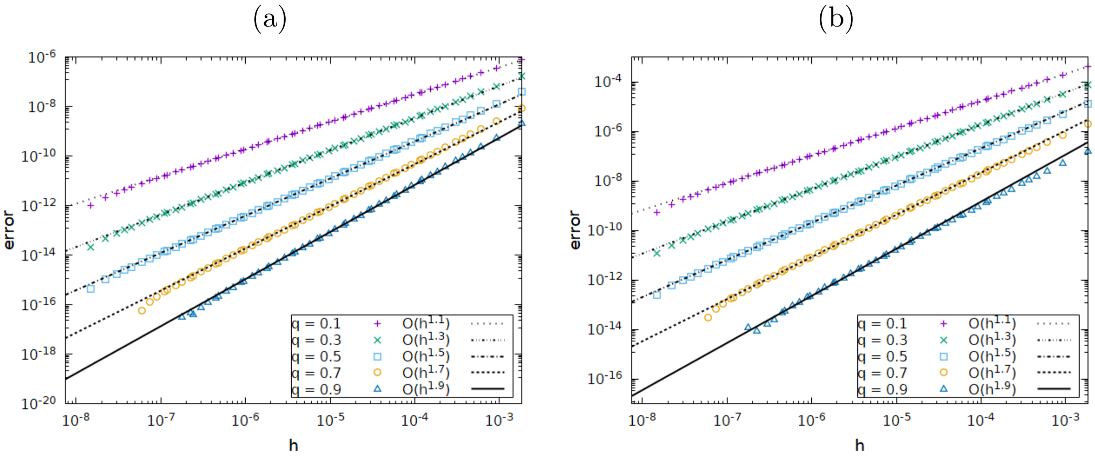


Figure 3.2: Order analyses of the numerical solution of eq. (3.1) via the leapfrog method given in eq. (3.2), for different values of q. In (a), the error for the variable y is presented and, in (b), the error for the v variable is presented.

Therefore, the simplified problem presented in eq. (3.1) retains the same problematic feature observed in the Kuwabara-Kono force model, eq. (1.23), and its enhanced simplicity allows for a more in-depth analytical investigation of the error bounds expected when it is integrated with eq. (3.2). This investigation is carried out in the next section.

3.2 Analysis of the global truncation error

The aim of this section is to prove the following theorem, formulated from the observations in the previous section that the order of the leapfrog method applied the simplified problem given by eq. (3.1) is dependent on the value of q.

Theorem 1. *The order of the leapfrog method given in eq.* (3.2) *applied to the initial value problem (IVP) formulated in eq.* (3.1) *is* $1 + q$, *i.e. the error of the method decreases as* $\mathcal{O}(h^{1+q})$, *where* h *is the size of the time step of the method and* $q \in (0, 1)$ *is the exponent of the nonlinearity of the RHS of eq.* (3.1)*. This result is valid regardless of which numerical method is used to initialize the iterative process of the leapfrog method.*

In the following, we construct the tools that are needed to perform a rigorous order analysis of eq. (3.2) when applied to solve eq. (3.1). The proof we present for of theorem 1 is inspired by the order analysis of a general multi-step method, such as the one presented in [4].

3.2.1 Preliminary results on the solution to the simplified problem

In the following, we consider the simplified problem given in eq. (3.1) as a model problem to explain the penalization of the order of convergence of the numerical solution observed in the convergence results. We start the analysis by showing that eq. (3.1) has a unique solution. To do so, we will need the following result

External Result 1. *Let $D \subset \mathbb{R}^n$ be an open set and $f : D \to \mathbb{R}^n$ a continuous vector field. If $x_0 \in D$ and one of the following two conditions holds:*

(i) f is locally Lipschitz continuous,

(ii) there exists $i_0 \in \{1, 2, \cdots, n\}$ such that $f_{i_0}(x_0) \neq 0$ and f is locally Lipschit continuous when fixing component i_0,

then the problem

$$x' = f(x), x(0) = x_0, \tag{3.3}$$

has a unique local solution.

Proof. See Corollary 3.4 of [15]. □

To apply the previous result to our problem, we consider a slightly modified form of eq. (3.1):

$$\begin{cases} y''(t) = |y(t)|^q; \\ y(0) = 0, \ y'(0) = 1, \end{cases} \tag{3.4}$$

defined on a neighborhood of $t = 0$. Since $q \in (0, 1)$, the absolute value on the RHS is necessary so that the problem is well defined for negative values of $y(t)$.

Proposition 1. *Let $q \in (0, 1)$. Then, there exists $\alpha \in (0, \infty)$ and a unique, twice differentiable function $\mathcal{Y} : (-\alpha, \alpha) \to \mathbb{R}$ which satisfies eq. (3.4) for all $t \in (-\alpha, \alpha)$.*

Proof. Let $F : \mathbb{R}^2 \to \mathbb{R}^2$ be defined as

$$F(x, v) = (v, |x|^q). \tag{3.5}$$

It is easy to see that F is continuous. Moreover, for each $x_0, v_1, v_2 \in \mathbb{R}$, notice that

$$\begin{aligned} \|F(x_0, v_1) - F(x_0, v_2)\|_\infty &= \|(v_1, |x_0|^q) - (v_2, |x_0|^q)\|_\infty \\ &= \|(v_1 - v_2, |x_0|^q - |x_0|^q)\|_\infty \\ &= \|(v_1 - v_2, 0)\|_\infty \\ &= |v_1 - v_2|. \end{aligned} \tag{3.6}$$

Finally,

$$F(0, 1) = (1, 0) \neq 0. \tag{3.7}$$

Thus, the proposition follows from external result 1. □

In order to go back to eq. (3.1) from eq. (3.4), it suffices to remove the absolute value on the RHS of the latter. To do so, we will prove the following result.

Proposition 2. *The function $\mathcal{Y}''$ is continuous.*

Proof. Note that, by proposition 1,

$$\mathcal{Y}''(t) = |\mathcal{Y}(t)|^q, \tag{3.8}$$

for all $t \in (-\alpha, \alpha)$. Since $\mathcal{Y}$ is continuous (because it is differentiable, which itself is again a consequence of proposition 1), the right-hand side of the equation is also continuous. Thus, the left-hand side must also be continuous, which concludes the proof. □

As a second intermediary step in going from eq. (3.4) back to eq. (3.1), we now show that $\mathcal{Y}$ is non-negative over $[0, \alpha)$. The choice of the domain $[0, \alpha)$ is only natural given the physical origin of the problem.

Proposition 3. *For all $t \in [0, \alpha)$, it is true that*

$$\mathcal{Y}(t) \geq 0. \tag{3.9}$$

Moreover, $\mathcal{Y}(t) = 0$ if, and only if, $t = 0$.

Proof. Since $\mathcal{Y}$ is twice differentiable on $[0, \alpha)$ and $\mathcal{Y}''$ is continuous on the same interval (by proposition 2), the Taylor theorem states that, for all $t \in (0, \alpha)$, there exists $c_0 \in (0, t)$ such that

$$\mathcal{Y}(t) = \mathcal{Y}(0) + t\mathcal{Y}'(0) + \frac{t^2}{2}\mathcal{Y}''(c_0). \tag{3.10}$$

Then, by proposition 1, it is true that

$$\begin{aligned}\mathcal{Y}(t) &= \mathcal{Y}(0) + t\mathcal{Y}'(0) + \frac{t^2}{2}\mathcal{Y}''(c_0) \\ &= t + \frac{t^2}{2}|\mathcal{Y}(c_0)|^q.\end{aligned} \tag{3.11}$$

Since, by hypothesis, t is non-negative, and so are $\frac{t^2}{2}$ and $|\mathcal{Y}(c_0)|^q$, we conclude that the left-hand side of the equation must also be non-negative. Furthermore, the only way the right-hand side of the equation can be zero is if $t = 0$, which concludes the proof. □

As a corollary of the last two propositions, we show that $\mathcal{Y}$ is also the unique solution to eq. (3.1) over $[0, \alpha)$

Corollary 1. *For $q \in (0, 1)$, let $\alpha \in (0, \infty)$ and $\mathcal{Y} : [0, \alpha) \to \mathbb{R}$ be the (restriction of the) unique solution to eq. (3.4) given by proposition 1. Then, $\mathcal{Y}$ is the unique solution of eq. (3.1) with domain $[0, \alpha)$.*

Proof. First, we will show that $\mathcal{Y}$ satisfies eq. (3.1). In fact, by proposition 3, it follows that $\mathcal{Y}(t)^q$ is a real number and

$$\mathcal{Y}(t)^q = |\mathcal{Y}(t)|^q, \tag{3.12}$$

for all $t \in [0, \alpha)$. Since $\mathcal{Y}$ satisfies eq. (3.4) and the only difference between it and eq. (3.1) is the absolute value of the RHS, it follows the $\mathcal{Y}$ is a solution to eq. (3.1) over $[0, \alpha)$.

Now, let us go over the uniqueness. Let $\tilde{\mathcal{Y}} : [0, \alpha) \to \mathbb{R}$ be a solution (in the classical sense) of eq. (3.1). Since $\tilde{\mathcal{Y}}$ is a real function, its second derivative $\tilde{\mathcal{Y}}''$ must also be. But since $\tilde{\mathcal{Y}}$ is a solution of eq. (3.1), it must hold that, given $t \in [0, \alpha)$,

$$\tilde{\mathcal{Y}}''(t) = \tilde{\mathcal{Y}}(t)^q \tag{3.13}$$

is true. Thus, $\tilde{\mathcal{Y}}(t)^q$ must be a real number.

Suppose now by contradiction that there exists $t_0 \in [0, \alpha)$ such that $\tilde{\mathcal{Y}}(t_0) < 0$. Then, since $q \in (0, 1)$, $\tilde{\mathcal{Y}}(t_0)^q$ must be non-real, which contradicts the fact that $\tilde{\mathcal{Y}}$ is a real function. Thus, $\tilde{\mathcal{Y}}$ must be a non-negative function.

But this implies that

$$\tilde{\mathcal{Y}}(t)^q = |\tilde{\mathcal{Y}}(t)|^q, \tag{3.14}$$

which, furthermore, implies that $\tilde{\mathcal{Y}}$ is a solution of eq. (3.4). Then, by the uniqueness shown in proposition 1, it must be that $\tilde{\mathcal{Y}} = \mathcal{Y}$, which concludes the proof. □

From now on, the domain of the function $\mathcal{Y}$ is assumed to be $[0, \alpha)$.

Next, we tackle the regularity of $\mathcal{Y}$ via a bootstrap-like argument.

Proposition 4. *The function $\mathcal{Y}$ is smooth on $(0, \alpha)$.*

Proof. The proof will be done by induction on the order of the derivative.

Base Follows directly from proposition 2.

Induction Suppose that there exists $n_0 \in \mathbb{N}$ such that, for all $n \in \mathbb{N}$ with $n \leq n_0$, $\mathcal{Y}^{(n)}$ exists and is continuous on $(0, \alpha)$.

Let $f : [0, \infty) \to \mathbb{R}$ be defined as

$$f(x) = x^q. \tag{3.15}$$

It is easy to see that f is smooth on its domain.

Now, by proposition 1, it is true that

$$\mathcal{Y}^{(n_0)} = \left(\mathcal{Y}^{(2)}\right)^{(n_0-2)} = \left(\mathcal{Y}^q\right)^{(n_0-2)} = \left(f \circ \mathcal{Y}\right)^{(n_0-2)}. \tag{3.16}$$

Note that the right-hand side is the $(n_0 - 2)$-th derivative of the composition of two functions, one of which is smooth and the other which, by the induction hypothesis, is n_0 times differentiable. By the chain rule, this means that the right-hand side is still twice differentiable. Thus, the left-hand side must also be twice differentiable, which implies that it is continuously differentiable.

□

We now shift our focus to the growth of $\mathcal{Y}$ and its first derivatives. We begin with the following result.

Proposition 5. *The functions $\mathcal{Y}, \mathcal{Y}', \mathcal{Y}''$ and $\mathcal{Y}'''$ are all positive on $(0, \alpha)$.*

Proof. This result has already been shown for $\mathcal{Y}$ in proposition 3.

That $\mathcal{Y}''$ is positive follows from the previous statement and the equality

$$\mathcal{Y}''(t) = \mathcal{Y}(t)^q, \tag{3.17}$$

which itself was proved in proposition 1.

The positivity of $\mathcal{Y}'$ can be shown by considering that

$$\mathcal{Y}'(0) = 1, \tag{3.18}$$

which was proved in proposition 1, and the fact that the positivity of $\mathcal{Y}''$ implies that $\mathcal{Y}'$ must be monotonically increasing on $(0, \alpha)$.

Finally, by differentiating both sides of eq. (3.17) with the chain rule, we get

$$\mathcal{Y}'''(t) = q\mathcal{Y}(t)^{q-1}\mathcal{Y}'(t), \tag{3.19}$$

which is positive, because $q \in (0, 1)$ and both $\mathcal{Y}$ and $\mathcal{Y}'$ are also positive. □

As a corollary of the last proposition, we gain further insight into the growth of $\mathcal{Y}$ and its first derivatives.

Corollary 2. *The functions $\mathcal{Y}, \mathcal{Y}', \mathcal{Y}''$ are all monotonically increasing on $(0, \alpha)$.*

Proof. Since a strictly positive derivative implies that the corresponding function is monotonic increasing, this proposition follows immediately from proposition 5. □

Finally, for future reference, we state expressions for $\mathcal{Y}'''$ and $\mathcal{Y}'''$.

$$\begin{aligned} \mathcal{Y}'''(t) &= q\mathcal{Y}(t)^{q-1}\mathcal{Y}'(t), & (3.20)\\ \mathcal{Y}''''(t) &= q\left((q-1)\mathcal{Y}(t)^{q-2}\mathcal{Y}'(t)^2 + \mathcal{Y}(t)^{q-1}\mathcal{Y}(t)^q\right). & (3.21) \end{aligned}$$

3.2.2 The leapfrog method and its local and global truncation errors

We shall now consider the leapfrog method to integrate eq. (3.1) in the interval $[0, \tau] \subset [0, \alpha)$. We define a partition of N subintervals of $[0, \tau]$, which defines a step size

$$h = \frac{\tau}{N}. \tag{3.22}$$

Therefore, the n-th iteration of the leapfrog method defined in eq. (3.2) applied to the IVP in eq. (3.1), for $n \leq N$, is given by:

$$\begin{aligned} v_{n+1} &= v_n + hy_n^q, & v_1 &= \mathcal{Y}'\left(\frac{h}{2}\right);\\ y_{n+1} &= y_n + hv_n + h^2 y_n^q, & y_1 &= \mathcal{Y}(h). \end{aligned} \tag{3.23}$$

Since the order reduction phenomenon is not related to initialization errors, we opt to initialize the leapfrog method, in this theoretical error analysis, using the exact values of $\mathcal{Y}$ and $\mathcal{Y}'$ at the appropriate times required by the leapfrog method. This will not be the case in practical applications, as $\mathcal{Y}$ is unknown. However, the error analysis is not affected by this choice. In section 3.2.6, we discuss the implications to the results when the leapfrog method is initialized with other numerical methods.

The truncation error of the y variable at the n-th step of the leapfrog method applied to eq. (3.1) is defined as follows:

$$T_y\left(n,N,\tau\right) := \mathcal{Y}\left((n+1)\frac{\tau}{N}\right) - \left[\mathcal{Y}\left(n\frac{\tau}{N}\right) + \frac{\tau}{N}\mathcal{Y}'\left(\left(n-\frac{1}{2}\right)\frac{\tau}{N}\right) + \frac{\tau^2}{N^2}\mathcal{Y}\left(n\frac{\tau}{N}\right)^q\right]. \tag{3.24}$$

Similarly, for the v variable, we have that the truncation error is given by:

$$T_v\left(n,N,\tau\right) := \mathcal{Y}'\left(\left(n+\frac{1}{2}\right)\frac{\tau}{N}\right) - \left[\mathcal{Y}'\left(\left(n-\frac{1}{2}\right)\frac{\tau}{N}\right) + \frac{\tau}{N}\mathcal{Y}\left(n\frac{\tau}{N}\right)^q\right]. \tag{3.25}$$

In order to determine the behavior of the truncation errors at the n-th step, the following result is necessary.

Proposition 6. *Let $t \in [0,\alpha)$. Then, it is true that $\mathcal{Y}(t) \geq t$. Also, for all $\tau \in (0,\alpha)$ and $n, N \in \mathbb{N}$ such that $n \leq N$, it is true that $y_n \geq nh$.*

Proof. By proposition 2, Taylor theorem applies to $\mathcal{Y}$ on $[0,\alpha)$ up to order 2. Then. there exists $c_0 \in (0,t)$ such that

$$\mathcal{Y}(t) = \mathcal{Y}(0) + t\mathcal{Y}'(0) + \frac{t^2}{2}\mathcal{Y}''(c_0). \tag{3.26}$$

Substituting the initial conditions given in the IVP (see proposition 1) yields

$$\mathcal{Y}(t) = \mathcal{Y}(0) + t\mathcal{Y}'(0) + \frac{t^2}{2}\mathcal{Y}''(c_0) = 0 + t + \frac{t^2}{2}\mathcal{Y}''(c_0) = t + \frac{t^2}{2}\mathcal{Y}''(c_0). \tag{3.27}$$

Since, by proposition 5, $\mathcal{Y}''$ is a positive function, the first inequality follows.

For the second inequality, note that both (y_n) and (v_n) are monotonically increasing sequences. Then, $v_n \geq v_1 = 1$. In turn, this implies that

$$y_n = y_{n-1} + hv_{n-1} + h^2 y_{n-1}^q = y_{n-1} + h(v_{n-1} + hy_{n-1}^q) = y_{n-1} + hv_n \geq y_{n-1} + h. \tag{3.28}$$

Applying the last equation recursively yields

$$y_n \geq y_1 + (n-1)h = h + (n-1)h = nh. \tag{3.29}$$

□

The following result states that the truncation error at the n-th step can be bounded by functions of n and h.

Lemma 1. *Let $\tau \in (0,\alpha) \cap (0,1)$. Then, there exists $K_1, K_2 \in (0,\infty)$ such that*

$$\left|T_y\left(n,N,\tau\right)\right| \leq K_1 h^{2+q} \quad \text{and} \quad \left|T_v\left(n,N,\tau\right)\right| \leq K_2\left(n-\frac{1}{2}\right)^{2-q} h^{1+q}, \tag{3.30}$$

for all $n, N \in \mathbb{N}$ such that $n < N$.

Proof. By proposition 4, the Taylor theorem applies to both $\mathcal{Y}$ and $\mathcal{Y}'$ on $(0, \alpha)$. For the former we obtain

$$\mathcal{Y}\left((n+1)\,h\right) = \mathcal{Y}(nh) + h\mathcal{Y}'(nh) + \frac{h^2}{2}\mathcal{Y}''(nh) + \frac{h^3}{6}\mathcal{Y}'''(c_0)\,; \tag{3.31}$$

and, for the latter,

$$\mathcal{Y}'\left(\left(n - \frac{1}{2}\right)h\right) = \mathcal{Y}'(nh) - \frac{h}{2}\mathcal{Y}''(nh) + \frac{h^2}{8}\mathcal{Y}'''(c_1)\,, \tag{3.32}$$

where $c_0 \in (nh, (n+1)\,h)$ and $c_1 \in \left(\left(n - \frac{1}{2}\right)h, nh\right)$.

Substituting the previous equations on the definition of $T_y(n, N, t)$ yields

$$\begin{aligned}
T_y(n, N, \tau) &= \mathcal{Y}((n+1)h) - \mathcal{Y}(nh) - h\mathcal{Y}'\left(\left(n - \frac{1}{2}\right)h\right) - h^2\mathcal{Y}(nh)^q \\
&= \mathcal{Y}(nh) + h\mathcal{Y}'(nh) + \frac{h^2}{2}\mathcal{Y}''(nh) + \frac{h^3}{6}\mathcal{Y}'''(c_0) \\
&- \left(\mathcal{Y}(nh) + h\left(\mathcal{Y}'(nh) - \frac{h}{2}\mathcal{Y}''(nh) + \frac{h^2}{8}\mathcal{Y}'''(c_1)\right) + h^2\mathcal{Y}(nh)^q\right) \\
&= \mathcal{Y}(nh) + h\mathcal{Y}'(nh) + \frac{h^2}{2}\mathcal{Y}''(nh) + \frac{h^3}{6}\mathcal{Y}'''(c_0) \\
&- \mathcal{Y}(nh) - h\mathcal{Y}'(nh) + \frac{h^2}{2}\mathcal{Y}''(nh) - \frac{h^3}{8}\mathcal{Y}'''(c_1) + h^2\mathcal{Y}(nh)^q \\
&= h^2\mathcal{Y}''(nh) + \frac{h^3}{6}\mathcal{Y}'''(c_0) - \frac{h^3}{8}\mathcal{Y}'''(c_1) + h^2\mathcal{Y}(nh)^q\,.
\end{aligned} \tag{3.33}$$

By replacing $\mathcal{Y}''(t) = \mathcal{Y}(t)^q$ in the previous equation (which can be done because of proposition 1 and because $nh \in [0, \tau] \subset (-\alpha, \alpha)$), we get

$$\begin{aligned}
T_y(n, N, \tau) &= h^2\mathcal{Y}''(nh) + \frac{h^3}{6}\mathcal{Y}'''(c_0) - \frac{h^3}{8}\mathcal{Y}'''(c_1) + h^2\mathcal{Y}(nh)^q \\
&= h^2\mathcal{Y}(nh)^q + \frac{h^3}{6}\mathcal{Y}'''(c_0) - \frac{h^3}{8}\mathcal{Y}'''(c_1) - h^2\mathcal{Y}(nh)^q \\
&= \frac{h^3}{6}\mathcal{Y}'''(c_0) - \frac{h^3}{8}\mathcal{Y}'''(c_1)\,.
\end{aligned} \tag{3.34}$$

Taking the absolute value of both sides of the equation, using the triangular inequality and the fact that $\mathcal{Y}'''$ is positive (see proposition 5), we arrive at

$$|T_y(n, N, \tau)| = \left|\frac{h^3}{6}\mathcal{Y}'''(c_0) - \frac{h^3}{8}\mathcal{Y}'''(c_1)\right| \leq h^3\left(\mathcal{Y}'''(c_0) + \mathcal{Y}'''(c_1)\right). \tag{3.35}$$

Then, substituting eq. (3.20) in eq. (3.35) yields

$$\begin{aligned}
|T_y(n, N, \tau)| &\leq h^3\left(\mathcal{Y}'''(c_0) + \mathcal{Y}'''(c_1)\right) = h^3\left(q\mathcal{Y}(c_0)^{q-1}\mathcal{Y}'(c_0) + q\mathcal{Y}(c_1)^{q-1}\mathcal{Y}'(c_1)\right) \\
&\leq h^3\left(\mathcal{Y}(c_0)^{q-1}\mathcal{Y}'(c_0) + \mathcal{Y}(c_1)^{q-1}\mathcal{Y}'(c_1)\right),
\end{aligned} \tag{3.36}$$

where the last step is due to $q < 1$. Since $\mathcal{Y}'$ is continuous on the compact interval $[0, \tau]$, there exists $K' \in (0, \infty)$ such that $\mathcal{Y}'(t) \leq K'$, for all $t \in [0, \tau]$. Using this bound, we find that

$$\begin{aligned} |T_y(n, N, \tau)| &\leq h^3 \left(\mathcal{Y}(c_0)^{q-1} \mathcal{Y}'(c_0) + \mathcal{Y}(c_1)^{q-1} \mathcal{Y}'(c_1)\right) \\ &\leq K' h^3 \left(\mathcal{Y}(c_0)^{q-1} + \mathcal{Y}(c_1)^{q-1}\right). \end{aligned} \tag{3.37}$$

Since $q \in (0, 1)$, $q - 1 < 0$. Then, $x \to x^{q-1}$ increases as $x \to 0^+$. Since, by proposition 5, $\mathcal{Y}$ is positive and, by corollary 2, monotonically increasing, this means that $\mathcal{Y}(t)^{q-1}$ increases as $t \to 0^+$. Also, because $c_0 \in (nh, (n+1)h)$, $c_1 \in \left(\left(n - \frac{1}{2}\right) h, nh\right)$ and $n \geq 1$, it follows that $\frac{h}{2} \leq c_0, c_1$. These two facts, when combined, imply that

$$\mathcal{Y}(c_0)^{q-1} \leq \mathcal{Y}\left(\frac{h}{2}\right)^{q-1} \quad \text{and} \quad \mathcal{Y}(c_1)^{q-1} \leq \mathcal{Y}\left(\frac{h}{2}\right)^{q-1}, \tag{3.38}$$

which, when combined with eq. (3.37), yields

$$|T_y(n, N, \tau)| \leq K' h^3 \left(\mathcal{Y}(c_0)^{q-1} + \mathcal{Y}(c_1)^{q-1}\right) \leq K' h^3 \left(\mathcal{Y}\left(\frac{h}{2}\right)^{q-1} + \mathcal{Y}\left(\frac{h}{2}\right)^{q-1}\right). \tag{3.39}$$

Then, by applying proposition 6 and the previously stated fact that $\mathcal{Y}(t)^{q-1}$ increases as $t \to 0^+$, we get

$$\begin{aligned} |T_y(n, N, \tau)| &\leq K' h^3 \left(\mathcal{Y}\left(\frac{h}{2}\right)^{q-1} + \mathcal{Y}\left(\frac{h}{2}\right)^{q-1}\right) \leq K' h^3 \left(\left(\frac{h}{2}\right)^{q-1} + \left(\frac{h}{2}\right)^{q-1}\right) \\ &= K' h^3 \left(\frac{h^{q-1}}{2^{q-1}} + \frac{h^{q-1}}{2^{q-1}}\right) = 2K' h^3 \frac{h^{q-1}}{2^{q-1}} = K' \frac{h^{2+q}}{2^{q-2}} = 2^{2-q} K' h^{2+q} = K_1 h^{2+q}, \end{aligned} \tag{3.40}$$

where $K_1 = 2^{2-q} K'$.

In order to obtain an appropriate bound for $T_v(n, N, \tau)$, we must apply Taylor's theorem to $\mathcal{Y}'\left(\left(n \pm \frac{1}{2}\right) h\right)$ and expand this function up to $\mathcal{O}(h^3)$. This yields

$$\mathcal{Y}'\left(\left(n \pm \frac{1}{2}\right) h\right) = \mathcal{Y}'(nh) \pm \frac{h}{2} \mathcal{Y}''(nh) + \frac{h^2}{8} \mathcal{Y}'''(nh) \pm \frac{h^3}{24} \mathcal{Y}''''(c_\pm), \tag{3.41}$$

where $c_+ \in \left(nh, \left(n + \frac{1}{2}\right) h\right)$ and $c_- \in \left(\left(n - \frac{1}{2}\right) h, nh\right)$. Substituting this equation and eq. (3.32) in the definition of $T_v(n,N,t)$ results in

$$\begin{aligned} T_v(n,N,\tau) &= \mathcal{Y}'\left(\left(n+\frac{1}{2}\right)h\right) - \mathcal{Y}'\left(\left(n-\frac{1}{2}\right)h\right) - h\mathcal{Y}(nh)^q \\ &= \mathcal{Y}'(nh) + \frac{h}{2}\mathcal{Y}''(nh) + \frac{h^2}{8}\mathcal{Y}'''(nh) + \frac{h^3}{24}\mathcal{Y}''''(c_+) \\ &- \left(\mathcal{Y}'(nh) - \frac{h}{2}\mathcal{Y}''(nh) + \frac{h^2}{8}\mathcal{Y}'''(nh) - \frac{h^3}{24}\mathcal{Y}''''(c_-)\right) - h\mathcal{Y}(nh)^q \\ &= \mathcal{Y}'(nh) + \frac{h}{2}\mathcal{Y}''(nh) + \frac{h^2}{8}\mathcal{Y}'''(nh) + \frac{h^3}{24}\mathcal{Y}''''(c_+) \\ &- \mathcal{Y}'(nh) + \frac{h}{2}\mathcal{Y}''(nh) - \frac{h^2}{8}\mathcal{Y}'''(nh) + \frac{h^3}{24}\mathcal{Y}''''(c_-) - h\mathcal{Y}(nh)^q \\ &= h\mathcal{Y}''(nh) + \frac{h^3}{24}\mathcal{Y}''''(c_+) + \frac{h^3}{24}\mathcal{Y}''''(c_-) - h\mathcal{Y}(nh)^q \\ &= h\mathcal{Y}''(nh) + \frac{h^3}{24}\left(\mathcal{Y}''''(c_+) + \mathcal{Y}''''(c_-)\right) - h\mathcal{Y}(nh)^q . \end{aligned} \tag{3.42}$$

By replacing $\mathcal{Y}''(t) = \mathcal{Y}(t)^q$ in the previous equation, just as we did in eq. (3.34), we get

$$\begin{aligned} T_v(n,N,\tau) &= h\mathcal{Y}''(nh) + \frac{h^3}{24}\left(\mathcal{Y}''''(c_+) + \mathcal{Y}''''(c_-)\right) - h\mathcal{Y}(nh)^q \\ &= h\mathcal{Y}^q(nh) + \frac{h^3}{24}\left(\mathcal{Y}''''(c_+) + \mathcal{Y}''''(c_-)\right) - h\mathcal{Y}(nh)^q \\ &= \frac{h^3}{24}\left(\mathcal{Y}''''(c_+) + \mathcal{Y}''''(c_-)\right). \end{aligned} \tag{3.43}$$

Taking the absolute value of both sides of this last equation and using the triangle inequality, we get

$$|T_v(n,N,\tau)| = |\frac{h^3}{24}\left(\mathcal{Y}''''(c_+) + \mathcal{Y}''''(c_-)\right)| \le \frac{h^3}{24}\left(|\mathcal{Y}''''(c_+)| + |\mathcal{Y}''''(c_-)|\right). \tag{3.44}$$

Now, we need to determine a bound for the fourth derivative of $\mathcal{Y}$. We can start doing so by taking the absolute value of both sides of eq. (3.21) and using the triangle inequality, which results in

$$\begin{aligned} |\mathcal{Y}''''(t)| &= |q\left((q-1)\mathcal{Y}(t)^{q-2}\mathcal{Y}'(t)^2 + \mathcal{Y}(t)^{q-1}\mathcal{Y}(t)^q\right)| \\ &\le q\left((1-q)\mathcal{Y}(t)^{q-2}\mathcal{Y}'(t)^2 + \mathcal{Y}(t)^{q-1}\mathcal{Y}(t)^q\right), \end{aligned} \tag{3.45}$$

where we used the fact that that $q-1<0$ and the positivity of $\mathcal{Y}$ and $\mathcal{Y}'$. Now, since $\mathcal{Y}$ is continuous on the compact interval $[0,\tau]$, there exists $K \in (0,\infty)$ such that $\mathcal{Y}(t) \le K$, for all $t \in [0,\tau]$. These arguments lead to the following bound for $\mathcal{Y}''''$:

$$\begin{aligned} |\mathcal{Y}''''(t)| &\le q\left((1-q)\mathcal{Y}(t)^{q-2}\mathcal{Y}'(t)^2 + \mathcal{Y}(t)^{q-1}\mathcal{Y}(t)^q\right) \\ &\le q\left((1-q)K'^2\mathcal{Y}(t)^{q-2} + K^q\mathcal{Y}(t)^{q-1}\right), \end{aligned} \tag{3.46}$$

where K' is as in eq. (3.37). Then, substituting our newly found bound for $\mathcal{Y}''''$ in inequality (3.44) gives

$$\begin{aligned}
&|T_v(n,N,\tau)| \leq \frac{h^3}{24}\left(|\mathcal{Y}''''(c_+)| + |\mathcal{Y}''''(c_-)|\right)\\
&\leq \frac{h^3}{24}\left(q\left((1-q)K'^2\mathcal{Y}(c_+)^{q-2} + K^q\mathcal{Y}(c_+)^{q-1}\right) + q\left((1-q)K'^2\mathcal{Y}(c_-)^{q-2} + K^q\mathcal{Y}(c_-)^{q-1}\right)\right)\\
&\leq \frac{h^3}{24}\left(q\left((1-q)K'^2\mathcal{Y}(c_+)^{q-2} + K^q\mathcal{Y}(c_+)^{q-1} + \ (1-q)K'^2\mathcal{Y}(c_-)^{q-2} + K^q\mathcal{Y}(c_-)^{q-1}\right)\right)\\
&\leq \frac{qh^3}{24}\left((1-q)K'^2\mathcal{Y}(c_+)^{q-2} + K^q\mathcal{Y}(c_+)^{q-1} + (1-q)K'^2\mathcal{Y}(c_-)^{q-2} + K^q\mathcal{Y}(c_-)^{q-1}\right)\\
&\leq \frac{qh^3}{24}\left((1-q)K'^2\mathcal{Y}(c_+)^{q-2} + (1-q)K'^2\mathcal{Y}(c_-)^{q-2} + K^q\mathcal{Y}(c_+)^{q-1} + K^q\mathcal{Y}(c_-)^{q-1}\right)\\
&\leq \frac{qh^3}{24}\left((1-q)K'^2\left(\mathcal{Y}(c_+)^{q-2} + \mathcal{Y}(c_-)^{q-2}\right) + K^q\left(\mathcal{Y}(c_+)^{q-1} + \mathcal{Y}(c_-)^{q-1}\right)\right).
\end{aligned} \tag{3.47}$$

Since $\mathcal{Y}$ is a non-decreasing function and both $q-1<0$ and $q-2<0$, it follows that both $\mathcal{Y}^{q-1}$ and $\mathcal{Y}^{q-2}$ are non-increasing functions. Now, recall that $c_+ \in \left(nh, \left(n+\frac{1}{2}\right)h\right)$ and $c_- \in \left(\left(n-\frac{1}{2}\right)h, nh\right)$. This means that $c_+, c_- \leq \left(n-\frac{1}{2}\right)h$. These two facts, when combined, imply that

$$\begin{aligned}
\mathcal{Y}(c_+)^{q-2} &\leq \mathcal{Y}\left(\left(n-\frac{1}{2}\right)h\right)^{q-2} \quad , \quad \mathcal{Y}(c_-)^{q-2} \leq \mathcal{Y}\left(\left(n-\frac{1}{2}\right)h\right)^{q-2}\\
\mathcal{Y}(c_+)^{q-1} &\leq \mathcal{Y}\left(\left(n-\frac{1}{2}\right)h\right)^{q-1} \text{ and } \ \mathcal{Y}(c_-)^{q-1} \leq \mathcal{Y}\left(\left(n-\frac{1}{2}\right)h\right)^{q-1}.
\end{aligned} \tag{3.48}$$

Replacing these bounds in eq. (3.47) gives

$$\begin{aligned}
|T_v(n,N,\tau)| &\leq \frac{qh^3}{24}\left((1-q)K'^2\left(\mathcal{Y}(c_+)^{q-2} + \mathcal{Y}(c_-)^{q-2}\right) + K^q\left(\mathcal{Y}(c_+)^{q-1} + \mathcal{Y}(c_-)^{q-1}\right)\right)\\
&\leq \frac{qh^3}{24}\left((1-q)K'^2\left(\mathcal{Y}\left(\left(n-\frac{1}{2}\right)h\right)^{q-2} + \mathcal{Y}\left(\left(n-\frac{1}{2}\right)h\right)^{q-2}\right)\right.\\
&\left. + K^q\left(\mathcal{Y}\left(\left(n-\frac{1}{2}\right)h\right)^{q-1} + \mathcal{Y}\left(\left(n-\frac{1}{2}\right)h\right)^{q-1}\right)\right)\\
&= \frac{qh^3}{24}\left(2(1-q)K'^2\mathcal{Y}\left(\left(n-\frac{1}{2}\right)h\right)^{q-2} + 2K^q\mathcal{Y}\left(\left(n-\frac{1}{2}\right)h\right)^{q-1}\right)\\
&\leq \frac{qh^3}{12}\left((1-q)K'^2\mathcal{Y}\left(\left(n-\frac{1}{2}\right)h\right)^{q-2} + K^q\mathcal{Y}\left(\left(n-\frac{1}{2}\right)h\right)^{q-1}\right).
\end{aligned} \tag{3.49}$$

Then, from proposition 6, we obtain

$$\begin{aligned}|T_v(n,N,\tau)| &\leq \frac{qh^3}{12}\left((1-q)\,K'^2\mathcal{Y}\left(\left(n-\frac{1}{2}\right)h\right)^{q-2}+K^q\mathcal{Y}\left(\left(n-\frac{1}{2}\right)h\right)^{q-1}\right)\\ &\leq \frac{qh^3}{12}\left((1-q)\,K'^2\left(\left(n-\frac{1}{2}\right)h\right)^{q-2}+K^q\left(\left(n-\frac{1}{2}\right)h\right)^{q-1}\right).\end{aligned} \tag{3.50}$$

Note that since $\tau < 1$, then this is also the case for $\left(n-\frac{1}{2}\right)h$. This means that

$$\left(\left(n-\frac{1}{2}\right)h\right)^{q-1} \leq \left(\left(n-\frac{1}{2}\right)h\right)^{q-2}, \tag{3.51}$$

which, in turn, means that the truncation error can be written as

$$\begin{aligned}|T_v(n,N,\tau)| &\leq \frac{qh^3}{12}\left((1-q)\,K'^2\left(\left(n-\frac{1}{2}\right)h\right)^{q-2}+K^q\left(\left(n-\frac{1}{2}\right)h\right)^{q-1}\right)\\ &\leq \frac{qh^3}{12}\left((1-q)\,K'^2\left(\left(n-\frac{1}{2}\right)h\right)^{q-2}+K^q\left(\left(n-\frac{1}{2}\right)h\right)^{q-2}\right)\\ &\leq \frac{qh^3}{12}\left(\left((1-q)\,K'^2+K^q\right)\left(\left(n-\frac{1}{2}\right)h\right)^{q-2}\right)\\ &\leq \frac{qh^3}{12}\left((1-q)\,K'^2+K^q\right)\left(\left(n-\frac{1}{2}\right)^{q-2}h^{q-2}\right)\\ &\leq \frac{qh^3}{12}\left(n-\frac{1}{2}\right)^{q-2}h^{q-2}\left((1-q)\,K'^2+K^q\right)\\ &\leq \frac{qh^{3+q-2}}{12}\left(n-\frac{1}{2}\right)^{q-2}\left((1-q)\,K'^2+K^q\right)\\ &\leq \frac{qh^{1+q}}{12}\left(n-\frac{1}{2}\right)^{q-2}\left((1-q)\,K'^2+K^q\right)\\ &\leq \frac{q\left((1-q)\,K'^2+K^q\right)}{12}\left(n-\frac{1}{2}\right)^{q-2}h^{1+q}\\ &= K_2\left(n-\frac{1}{2}\right)^{q-2}h^{1+q},\end{aligned} \tag{3.52}$$

where

$$K_2 = \frac{q\left((1-q)\,K'^2+K^q\right)}{12}. \tag{3.53}$$

□

The first specificity of the simplified problem studied in this work is revealed in lemma 1. When $q \geq 1$, a universal bound for the truncation errors, independent of the variable t, i.e. independent of which precise step n the truncation error is being analyzed, can be used to obtain the second order accuracy of the method [33], [79]. However, the singularity of the first derivative of the RHS of eq. (3.1) at $t = 0$ implies that a universal bound for the truncation error of the v variable does not yield the same expected result when a similar technique is used in this case. This difficulty, which is at the heart of the order penalization behavior that was identified in the previous section, can be overcome by choosing a step dependent bound for the truncation error of the v variable, which now becomes explicitly dependent on the value of n for the n-th step, as observed in eq. (3.52).

Next, we define the total error of the y variable at the n-th step as

$$e_n^y := \mathcal{Y}\left(nh\right) - y_n, \tag{3.54}$$

and the total error of the v variable at the n-th step as

$$e_n^v := \mathcal{Y}'\left(\left(n - \frac{1}{2}\right)h\right) - v_n. \tag{3.55}$$

The following auxiliary result is a quasi-Lipschitz inequality that will be useful when proving bounds for both total errors defined above.

Proposition 7. *For all $\tau \in (0, \alpha)$ and $n, N \in \mathbb{N}$ such that $n \leq N$, it is true that*

$$|\mathcal{Y}\left(nh\right)^q - y_n^q| \leq q\left(nh\right)^{q-1}|e_n^y|. \tag{3.56}$$

Proof. Let $f : (0, \alpha) \to \mathbb{R}$ be defined as $f(x) = x^q$. Since f is smooth in its domain then, by the mean value theorem, given $x_1, x_2 \in (0, \infty)$ such that $x_1 < x_2$, there exists $c_0 \in (x_1, x_2)$ such that

$$|x_1^q - x_2^q| = |f(x_1) - f(x_2)| \leq f'(c_0)|x_1 - x_2| = qc_0^{q-1}|x_1 - x_2|. \tag{3.57}$$

Let

$$x_1 = \min\left\{\mathcal{Y}\left(nh\right), y_n\right\} \quad \text{and} \quad x_2 = \max\left\{\mathcal{Y}\left(nh\right), y_n\right\}. \tag{3.58}$$

Then, since $|x_1 - x_2| = |x_2 - x_1|$, we have

$$|\mathcal{Y}\left(nh\right)^q - y_n^q| = |x_1^q - x_2^q| =\leq qc_0^{q-1}|x_1 - x_2| = qc_0^{q-1}|\mathcal{Y}\left(nh\right) - y_n|. \tag{3.59}$$

By the definition of e_n^y in eq. (3.54), we can rewrite the previous equation as

$$|\mathcal{Y}\left(nh\right)^q - y_n^q| \leq qc_0^{q-1}|\mathcal{Y}\left(nh\right) - y_n| = qc_0^{q-1}|e_n^y|. \tag{3.60}$$

Since $c_0 \in (x_1, x_2)$ and, by proposition 6, we know that $\mathcal{Y}\left(nh\right), y_n \geq nh$, we can infer that $nh \leq c_0$. On the other hand, due to the fact that $q - 1 < 0$, we conclude that $c_0^{q-1} \leq \left(nh\right)^{q-1}$. Replacing this information in eq. (3.60) yields

$$|\mathcal{Y}\left(nh\right)^q - y_n^q| \leq qc_0^{q-1}|e_n^y| \leq q\left(nh\right)^{q-1}|e_n^y|. \tag{3.61}$$

□

For the sake of simplicity, we shall now use a vector notation for both truncation and total errors as follows:

$$\vec{T}_n := \begin{bmatrix} |T_y\left(n, N, \tau\right)| \\ |T_v\left(n, N, \tau\right)| \end{bmatrix} \quad \text{and} \quad \vec{E}_n := \begin{bmatrix} |e_n^y| \\ |e_n^v| \end{bmatrix}. \tag{3.62}$$

In addition, we further simplify the notation by defining

$$w := qh^{1+q}. \tag{3.63}$$

The following result relates the total error at the $(n+1)$-th step to the truncation and total errors at the n-th step by using the following matrix:

$$A_n := \begin{bmatrix} \left(1 + wn^{q-1}\right) & h \\ \frac{w}{h} n^{q-1} & 1 \end{bmatrix}. \tag{3.64}$$

This next result is an estimate for the one-step truncation errors of the numerical approximation, based on the truncation and total errors of the previous step.

Lemma 2. *For all $\tau \in (0, \alpha)$ and $n, N \in \mathbb{N}$ such that $n \leq N$, it is true that*

$$\vec{E}_{n+1} \leq A_n \vec{E}_n + \vec{T}_n, \tag{3.65}$$

where the inequality holds component-wise.

Proof. Rearranging eq. (3.24) we get that

$$\mathcal{Y}\left((n+1)h\right) = \mathcal{Y}(nh) + h\mathcal{Y}'\left(\left(n - \frac{1}{2}\right)h\right) + h^2\mathcal{Y}(nh)^q + T_y(n, N, \tau). \tag{3.66}$$

Next, we subtract the equation for y_n (given in eq. (3.23)) from both sides and manipulate:

$$\begin{aligned}
\mathcal{Y}((n+1)h) - y_{n+1} &= \mathcal{Y}(nh) + h\mathcal{Y}'\left(\left(n - \frac{1}{2}\right)h\right) + h^2\mathcal{Y}(nh)^q + T_y(n, N, \tau) \\
&\quad - \left(y_n + hv_n + h^2 y_n^q\right) \\
&\iff \\
\mathcal{Y}((n+1)h) - y_{n+1} &= \mathcal{Y}(nh) + h\mathcal{Y}'\left(\left(n - \frac{1}{2}\right)h\right) + h^2\mathcal{Y}(nh)^q + T_y(n, N, \tau) \\
&\quad - y_n - hv_n - h^2 y_n^q \\
&\iff \\
\Big[\mathcal{Y}((n+1)h) - y_{n+1}\Big] &= \Big[\mathcal{Y}(nh) - y_n\Big] + \left[h\mathcal{Y}'\left(\left(n - \frac{1}{2}\right)h\right) - hv_n\right] \\
&\quad + \Big[h^2\mathcal{Y}(nh)^q - h^2 y_n^q\Big] + T_y(n, N, \tau) \\
&\iff \\
\Big[\mathcal{Y}((n+1)h) - y_{n+1}\Big] &= \Big[\mathcal{Y}(nh) - y_n\Big] + h\left[\mathcal{Y}'\left(\left(n - \frac{1}{2}\right)h\right) - v_n\right] \\
&\quad + h^2\Big[\mathcal{Y}(nh)^q - y_n^q\Big] + T_y(n, N, \tau)
\end{aligned} \tag{3.67}$$

Then, substituting eqs. (3.54) and (3.55) we get

$$e^y_{n+1} = e^y_n + he^v_n + h^2\left(\mathcal{Y}\left(nh\right)^q - y^q_n\right) + T_y\left(n, N, \tau\right). \tag{3.68}$$

If we now take the absolute value of both sides and use the properties of a norm, we get

$$\begin{aligned} |e^y_{n+1}| &= |e^y_n + he^v_n + h^2\left(\mathcal{Y}\left(nh\right)^q - y^q_n\right) + T_y\left(n, N, \tau\right)| \\ &\leq |e^y_n| + |he^v_n| + |h^2\left(\mathcal{Y}\left(nh\right)^q - y^q_n\right)| + |T_y\left(n, N, \tau\right)| \\ &= |e^y_n| + |h||e^v_n| + |h^2||\mathcal{Y}\left(nh\right)^q - y^q_n| + |T_y\left(n, N, \tau\right)| \\ &= |e^y_n| + h|e^v_n| + h^2|\mathcal{Y}\left(nh\right)^q - y^q_n| + |T_y\left(n, N, \tau\right)|, \end{aligned} \tag{3.69}$$

where we used that $h > 0$ in the last step. By proposition 7,

$$\begin{aligned} |e^y_{n+1}| &\leq |e^y_n| + h|e^v_n| + h^2|\mathcal{Y}\left(nh\right)^q - y^q_n| + |T_y\left(n, N, \tau\right)| \\ &\leq |e^y_n| + h|e^v_n| + h^2 q\left(nh\right)^{q-1}|e^y_n| + |T_y\left(n, N, \tau\right)| \\ &= |e^y_n| + h|e^v_n| + qh^2n^{q-1}h^{q-1}|e^y_n| + |T_y\left(n, N, \tau\right)| \\ &= |e^y_n| + h|e^v_n| + qh^{2+q-1}n^{q-1}|e^y_n| + |T_y\left(n, N, \tau\right)| \\ &= |e^y_n| + h|e^v_n| + qh^{1+q}n^{q-1}|e^y_n| + |T_y\left(n, N, \tau\right)| \\ &= |e^y_n| + qh^{1+q}n^{q-1}|e^y_n| + h|e^v_n| + |T_y\left(n, N, \tau\right)| \\ &= \left(1 + qh^{1+q}n^{q-1}\right)|e^y_n| + h|e^v_n| + |T_y\left(n, N, \tau\right)|. \end{aligned} \tag{3.70}$$

Finally, by the definition of w (eq. (3.63)),

$$\begin{aligned} |e^y_{n+1}| &\leq \left(1 + qh^{1+q}n^{q-1}\right)|e^y_n| + h|e^v_n| + |T_y\left(n, N, \tau\right)| \\ &= \left(1 + wn^{q-1}\right)|e^y_n| + h|e^v_n| + |T_y\left(n, N, \tau\right)|. \end{aligned} \tag{3.71}$$

We now proceed in an analogous manner starting from eq. (3.25). Rearranging eq. (3.25) we get that

$$\mathcal{Y}'\left(\left(n + \frac{1}{2}\right)h\right) = \mathcal{Y}'\left(\left(n - \frac{1}{2}\right)h\right) + h\mathcal{Y}\left(nh\right)^q + T_v\left(n, N, \tau\right). \tag{3.72}$$

Next, we subtract the equation for v_n (given in eq. (3.23)) from both sides and manipulate:

$$\mathcal{Y}'\left(\left(n + \frac{1}{2}\right)h\right) - v_{n+1} = \mathcal{Y}'\left(\left(n - \frac{1}{2}\right)h\right) + h\mathcal{Y}\left(nh\right)^q + T_v\left(n, N, \tau\right) - \left(v_n + hy^q_n\right)$$

$$\Longleftrightarrow$$

$$\mathcal{Y}'\left(\left(n + \frac{1}{2}\right)h\right) - v_{n+1} = \mathcal{Y}'\left(\left(n - \frac{1}{2}\right)h\right) + h\mathcal{Y}\left(nh\right)^q + T_v\left(n, N, \tau\right) - v_n - hy^q_n$$

$$\Longleftrightarrow$$

$$\left[\mathcal{Y}'\left(\left(n + \frac{1}{2}\right)h\right) - v_{n+1}\right] = \left[\mathcal{Y}'\left(\left(n - \frac{1}{2}\right)h\right) - v_n\right] + \left[h\mathcal{Y}\left(nh\right)^q - hy^q_n\right] + T_v\left(n, N, \tau\right)$$

$$\Longleftrightarrow$$

$$\left[\mathcal{Y}'\left(\left(n + \frac{1}{2}\right)h\right) - v_{n+1}\right] = \left[\mathcal{Y}'\left(\left(n - \frac{1}{2}\right)h\right) - v_n\right] + h\left[\mathcal{Y}\left(nh\right)^q - y^q_n\right] + T_v\left(n, N, \tau\right) \tag{3.73}$$

Then, substituting eq. (3.55) we get

$$e^v_{n+1} = e^v_n + h\left(\mathcal{Y}\left(nh\right)^q - y^q_n\right) + T_v\left(n, N, \tau\right). \tag{3.74}$$

If we now take the absolute value of both sides and use the properties of a norm, we get

$$\begin{aligned} |e^v_{n+1}| &= |e^v_n + h\left(\mathcal{Y}\left(nh\right)^q - y^q_n\right) + T_v\left(n, N, \tau\right)| \leq |e^v_n| + |h\left(\mathcal{Y}\left(nh\right)^q - y^q_n\right)| + |T_v\left(n, N, \tau\right)| \\ &= |e^v_n| + |h||\mathcal{Y}\left(nh\right)^q - y^q_n| + |T_v\left(n, N, \tau\right)| = |e^v_n| + h|\mathcal{Y}\left(nh\right)^q - y^q_n| + |T_v\left(n, N, \tau\right)|, \end{aligned} \tag{3.75}$$

where we used that $h > 0$ in the last step. By proposition 7,

$$\begin{aligned} |e^v_{n+1}| &\leq |e^v_n| + h|\mathcal{Y}\left(nh\right)^q - y^q_n| + |T_v\left(n, N, \tau\right)| \leq |e^v_n| + hq\left(nh\right)^{q-1}|e^y_n| + |T_v\left(n, N, \tau\right)| \\ &= |e^v_n| + qhn^{q-1}h^{q-1}|e^y_n| + |T_v\left(n, N, \tau\right)| = |e^v_n| + qh^{1+q-1}n^{q-1}|e^y_n| + |T_v\left(n, N, \tau\right)| \\ &= |e^v_n| + \frac{qh^{1+q}}{h}n^{q-1}|e^y_n| + |T_v\left(n, N, \tau\right)| \end{aligned} \tag{3.76}$$

Finally, by the definition of w (eq. (3.63)),

$$|e^v_{n+1}| \leq |e^v_n| + \frac{qh^{1+q}}{h}n^{q-1}|e^y_n| + |T_v\left(n, N, \tau\right)| = |e^v_n| + \frac{w}{h}n^{q-1}|e^y_n| + |T_v\left(n, N, \tau\right)|. \tag{3.77}$$

Assembling both eqs. (3.71) and (3.77) in matrix form yields the stated result.

□

The result in lemma 2 is used to obtain a general inequality for the total error at the n-th step based only on the truncation errors of all the previous steps. Before we proceed, we state a convention on matrix multiplication.

Convention 1. *Let $k \in \mathbb{N}$ and, for each $n \in \mathbb{N}$, let A_n be a square matrix of dimension k. Then, it is conventioned that*

$$\prod_{i=a}^{b} A_i = \begin{cases} A_a A_{a+1} \cdots A_{b-1} A_b, & a < b; \\ A_a, & a = b; \\ I, & a > b, \end{cases} \tag{3.78}$$

where $a, b \in \mathbb{N}$ and I is the identity matrix of dimension k. Note that the order in which the matrices are multiplied in the case $a < b$ must be respected.

Please be mindful that this convention establishes the order in which the different matrix ought to be multiplied. Since matrix multiplication is not commutative, this is strictly necessary.

This convention fixed, we can proceed to the next result, which is a corollary of lemma 2 and establishes a more general bound for the truncation error, in contrast to the one-step step bound of the previous result.

Corollary 3. *Under the same hypotheses and notation of lemma 2, it holds that*

$$\vec{E}_n \leq \left(\prod_{i=1}^{n-1} A_{n-i}\right)\vec{E}_1 + \sum_{i=1}^{n-1}\left(\left(\prod_{j=1}^{i-1} A_{n-j}\right)\vec{T}_{n-i}\right). \tag{3.79}$$

Proof. The proof will be done by induction on n.

Base For $n = 2$, inequality (3.79) becomes

$$\begin{aligned}
\vec{E}_2 = \vec{E}_n &\le \left(\prod_{i=1}^{n-1} A_{n-i}\right)\vec{E}_1 + \sum_{i=1}^{n-1}\left(\left(\prod_{j=1}^{i-1} A_{n-j}\right)\vec{T}_{n-i}\right)\\
&= \left(\prod_{i=1}^{2-1} A_{2-i}\right)\vec{E}_1 + \sum_{i=1}^{2-1}\left(\left(\prod_{j=1}^{i-1} A_{2-j}\right)\vec{T}_{2-i}\right)\\
&= \left(\prod_{i=1}^{1} A_{2-i}\right)\vec{E}_1 + \sum_{i=1}^{1}\left(\left(\prod_{j=1}^{i-1} A_{2-j}\right)\vec{T}_{2-i}\right)\\
&= A_{2-1}\vec{E}_1 + \left(\prod_{j=1}^{1-1} A_{2-j}\right)\vec{T}_{2-1} = A_1\vec{E}_1 + \left(\prod_{j=1}^{0} A_{2-j}\right)\vec{T}_1\\
&= A_1\vec{E}_1 + I\vec{T}_1 = A_1\vec{E}_1 + \vec{T}_1.
\end{aligned} \tag{3.80}$$

which is what we must prove. If one applies lemma 2 for $n = 1$, it yields

$$\vec{E}_2 = \vec{E}_{1+1} = \vec{E}_{n+1} \le A_n\vec{E}_n + \vec{T}_n = A_1\vec{E}_1 + \vec{T}_1, \tag{3.81}$$

which is the desired expression.

Induction Suppose that there exists $n_0 \in \mathbb{N}$ with $n_0 \le N$ such that inequality (3.79) holds for all $n \le n_0$.

By using lemma 2 with $n = n_0$, we have

$$\vec{E}_{n_0+1} \le A_{n_0}\vec{E}_{n_0} + \vec{T}_{n_0}. \tag{3.82}$$

Then, by the induction hypothesis,

$$\begin{aligned}
\vec{E}_{n_0+1} &\le A_{n_0}\vec{E}_{n_0} + \vec{T}_{n_0}\\
&\le A_{n_0}\left(\left(\prod_{i=1}^{n_0-1} A_{n_0-i}\right)\vec{E}_1 + \sum_{i=1}^{n_0-1}\left(\left(\prod_{j=1}^{i-1} A_{n_0-j}\right)\vec{T}_{n_0-i}\right)\right) + \vec{T}_{n_0}\\
&= A_{n_0}\left(\prod_{i=1}^{n_0-1} A_{n_0-i}\right)\vec{E}_1 + A_{n_0}\sum_{i=1}^{n_0-1}\left(\left(\prod_{j=1}^{i-1} A_{n_0-j}\right)\vec{T}_{n_0-i}\right) + \vec{T}_{n_0}\\
&= \left(\prod_{i=0}^{n_0-1} A_{n_0-i}\right)\vec{E}_1 + \left(\sum_{i=1}^{n_0-1} A_{n_0}\left(\prod_{j=1}^{i-1} A_{n_0-j}\right)\vec{T}_{n_0-i}\right) + \vec{T}_{n_0}\\
&= \left(\prod_{i=0}^{n_0-1} A_{n_0-i}\right)\vec{E}_1 + \sum_{i=1}^{n_0-1}\left(\left(\prod_{j=0}^{i-1} A_{n_0-j}\right)\vec{T}_{n_0-i}\right) + \vec{T}_{n_0}\\
&= \left(\prod_{i=0}^{n_0-1} A_{n_0-i}\right)\vec{E}_1 + \sum_{i=1}^{n_0-1}\left(\left(\prod_{j=0}^{i-1} A_{n_0-j}\right)\vec{T}_{n_0-i}\right) + \vec{T}_{n_0-0}.
\end{aligned} \tag{3.83}$$

Now, by the information presented in convention 1, we get that

$$\begin{aligned}\vec{E}_{n_0+1} &\leq \left(\prod_{i=0}^{n_0-1} A_{n_0-i}\right)\vec{E}_1 + \sum_{i=1}^{n_0-1}\left(\left(\prod_{j=0}^{i-1} A_{n_0-j}\right)\vec{T}_{n_0-i}\right) + \vec{T}_{n_0-0}\\ &= \left(\prod_{i=0}^{n_0-1} A_{n_0-i}\right)\vec{E}_1 + \sum_{i=1}^{n_0-1}\left(\left(\prod_{j=0}^{i-1} A_{n_0-j}\right)\vec{T}_{n_0-i}\right) + I\vec{T}_{n_0-0}\\ &= \left(\prod_{i=0}^{n_0-1} A_{n_0-i}\right)\vec{E}_1 + \sum_{i=1}^{n_0-1}\left(\left(\prod_{j=0}^{i-1} A_{n_0-j}\right)\vec{T}_{n_0-i}\right) + \left(\prod_{j=0}^{-1} A_{n_0-j}\right)\vec{T}_{n_0-0}\\ &= \left(\prod_{i=0}^{n_0-1} A_{n_0-i}\right)\vec{E}_1 + \sum_{i=1}^{n_0-1}\left(\left(\prod_{j=0}^{i-1} A_{n_0-j}\right)\vec{T}_{n_0-i}\right) + \left(\prod_{j=0}^{0-1} A_{n_0-j}\right)\vec{T}_{n_0-0}\\ &= \left(\prod_{i=0}^{n_0-1} A_{n_0-i}\right)\vec{E}_1 + \sum_{i=0}^{n_0-1}\left(\left(\prod_{j=0}^{i-1} A_{n_0-j}\right)\vec{T}_{n_0-i}\right).\end{aligned} \tag{3.84}$$

By making the variable substitution $i' = i + 1$ and $j' = j + 1$, we arrive at

$$\begin{aligned}\vec{E}_{n_0+1} &\leq \left(\prod_{i=0}^{n_0-1} A_{n_0-i}\right)\vec{E}_1 + \sum_{i=0}^{n_0-1}\left(\left(\prod_{j=0}^{i-1} A_{n_0-j}\right)\vec{T}_{n_0-i}\right)\\ &= \left(\prod_{i'=1}^{n_0} A_{n_0-i'+1}\right)\vec{E}_1 + \sum_{i'=1}^{n_0}\left(\left(\prod_{j=0}^{i'-2} A_{n_0-j}\right)\vec{T}_{n_0-i'+1}\right)\\ &= \left(\prod_{i'=1}^{n_0} A_{n_0-i'+1}\right)\vec{E}_1 + \sum_{i'=1}^{n_0}\left(\left(\prod_{j'=1}^{i'-1} A_{n_0-j'+1}\right)\vec{T}_{n_0-i'+1}\right)\\ &= \left(\prod_{i'=1}^{(n_0+1)-1} A_{(n_0+1)-i'}\right)\vec{E}_1 + \sum_{i'=1}^{(n_0+1)-1}\left(\left(\prod_{j'=1}^{i'-1} A_{(n_0+1)-j'}\right)\vec{T}_{(n_0+1)-i'}\right),\end{aligned} \tag{3.85}$$

which is the desired result.

□

Note that the initialization error in eq. (3.23) is zero, i.e. $\vec{E}_1 = 0$. This is because we chose to start the method in eq. (3.23) with exact values at the first step. This was done so that we could omit the term including $\vec{E}_1$ of inequality (3.79) in our future calculations, as to not further complicate this text. Still, in general, this will not be the case and the initial error $\vec{E}_1$ should remain in the RHS of inequality (3.79); e.g. when another numerical method is used to initialize eq. (3.23). In either case, the final result remains unaltered, which gave us confidence in omitting $\vec{E}_1$ from our calculations. This is discussed in more detail in section 3.2.6.

Inequality (3.79) can be further simplified if we realize that for any 2×2 matrix Z with non-negative entries, $Z \leq ZA_n$ for any n. Therefore, the following result holds.

Corollary 4. *Let $\tau \in (0, \alpha)$ and $n, N \in \mathbb{N}$ such that $n < N$. Then, it is true that*

$$\vec{E}_N \leq \sum_{i=1}^{N-1} \left(\prod_{j=1}^{N-1} A_{N-j} \right) \vec{T}_{N-i}, \tag{3.86}$$

where the inequality holds entry-wise.

Proof. Note that

$$\prod_{i=1}^{N} A_{N-i} = \left(\prod_{i=1}^{n} A_{N-i} \right) \prod_{i=n+1}^{N} A_{N-i} \geq \prod_{i=1}^{n} A_{N-i}. \tag{3.87}$$

The fact that, for any $i \in \mathbb{N}$, the entries of A_i are positive implies that any product of A_i also has only positive entries. Thus, in eq. (3.87), we successively use the property stated in the text above for $Z = \prod_{i=1}^{n} A_{N-i}$. Using this inequality in corollary 3 concludes the proof. □

3.2.3 Computing an explicit expression for $\prod_{i=1}^{n-1} A_{n-i}$

Corollary 3 indicates that products of the form $\prod_{j=1}^{n-1} A_{n-j}$ play a central role in understanding the error propagation of the numerical solution of eq. (3.1) via the Leapfrog method in eq. (3.23). On the lemma below, we find an expression for each entry of $\prod_{i=1}^{n-1} A_{n-i}$ in terms of functions a, b, c, d which are defined on the set

$$D := \bigcup_{N=1}^{\infty} \left\{ (n, \theta) \in \{2, \ldots, N\} \times \mathcal{P}(\{1, \ldots, N-1\}) \mid \theta \in \mathcal{P}(\{1, \ldots, n-1\}) \right\}, \tag{3.88}$$

where $\mathcal{P}(\cdot)$ is the power set operator. Intuitively, D is the set of all ordered pairs such that the first element, n, is an integer bigger then 1 and the second element is any set containing only numbers from 1 to $n-1$.

We also obtain coupled, recursive expressions for a, b, c, d and their initial values. In the statement of the following lemma (and, indeed, throughout the remaining of this work), we will use the standard notation $\#\theta$ for the cardinality of the set $\theta \subset \mathbb{N}$.

Lemma 3. *Let $\tau \in (0, \alpha)$ and $n, N \in \mathbb{N}$ such that $2 \leq n \leq N$. Then, it is true that*

$$\prod_{i=1}^{n-1} A_{n-i} = \begin{bmatrix} p_{1,1}(n) & p_{1,2}(n) \\ p_{2,1}(n) & p_{2,2}(n) \end{bmatrix}, \tag{3.89}$$

where

$$p_{1,1}(n) = \sum_{\theta \in \mathcal{P}(\{1,\ldots,n-1\})} a(n, \theta)\, w^{\#\theta} \prod_{i \in \theta} i^{q-1}, \tag{3.90a}$$

$$p_{1,2}(n) = h \sum_{\theta \in \mathcal{P}(\{1,\dots,n-1\})} b(n,\theta)\, w^{\#\theta} \prod_{i \in \theta} i^{q-1}, \tag{3.90b}$$

$$p_{2,1}(n) = \frac{1}{h} \sum_{\theta \in \mathcal{P}(\{1,\dots,n-1\})} c(n,\theta)\, w^{\#\theta} \prod_{i \in \theta} i^{q-1}, \tag{3.90c}$$

$$p_{2,2}(n) = \sum_{\theta \in \mathcal{P}(\{1,\dots,n-1\})} d(n,\theta)\, w^{\#\theta} \prod_{i \in \theta} i^{q-1}, \tag{3.90d}$$

for $a, b, c, d : D \to \mathbb{R}$. Furthermore, these functions satisfy the recursive relations

$$a(n,\theta) = \begin{cases} a(n-1, \theta - \{n-1\}) & \textit{if } (n-1) \in \theta; \\ \big(a(n-1,\theta) + c(n-1,\theta)\big) & \textit{if } (n-1) \notin \theta, \end{cases} \tag{3.91a}$$

$$b(n,\theta) = \begin{cases} b(n-1, \theta - \{n-1\}) & \textit{if } (n-1) \in \theta; \\ \big(b(n-1,\theta) + d(n-1,\theta)\big) & \textit{if } (n-1) \notin \theta, \end{cases} \tag{3.91b}$$

$$c(n,\theta) = \begin{cases} a(n-1, \theta - \{n-1\}) & \textit{if } (n-1) \in \theta; \\ c(n-1,\theta) & \textit{if } (n-1) \notin \theta, \end{cases} \tag{3.91c}$$

$$d(n,\theta) = \begin{cases} b(n-1, \theta - \{n-1\}) & \textit{if } (n-1) \in \theta; \\ d(n-1,\theta) & \textit{if } (n-1) \notin \theta \end{cases} \tag{3.91d}$$

and the initial conditions

$$\begin{aligned} a(2,\emptyset) &= 1, & b(2,\emptyset) &= 1, & c(2,\emptyset) &= 0, & d(2,\emptyset) &= 1, \\ a(2,\{1\}) &= 1, & b(2,\{1\}) &= 0, & c(2,\{1\}) &= 1, & d(2,\{1\}) &= 0. \end{aligned} \tag{3.92}$$

Proof. The proof will be done by induction on n.

Base case For $n = 2$, the left-hand side of eq. (3.89) becomes

$$\prod_{i=1}^{n-1} A_{n-i} = \prod_{i=1}^{2-1} A_{2-i} = \prod_{i=1}^{1} A_{2-i} = A_{2-1} = A_1 = \begin{bmatrix} 1 + w1^{q-1} & h \\ \frac{w}{h} 1^{q-1} & 1 \end{bmatrix} = \begin{bmatrix} 1 + w & h \\ \frac{w}{h} & 1 \end{bmatrix}.$$

Thus, the equality with the right-hand side of eq. (3.89) holds for

$$\begin{aligned} a(2,\emptyset) &= 1, & b(2,\emptyset) &= 1, & c(2,\emptyset) &= 0, & d(2,\emptyset) &= 1, \\ a(2,\{1\}) &= 1, & b(2,\{1\}) &= 0, & c(2,\{1\}) &= 1, & d(2,\{1\}) &= 0. \end{aligned} \tag{3.93}$$

Inductive step Suppose that there exists an $n_0 \in \mathbb{N}$ with $2 < n_0 \leq N$ such that eqs. (3.90a) to (3.90d) hold for all $n \in \mathbb{N}$ with $n \leq n_0$. If $n_0 = N$, the result follows trivially. Otherwise, $n_0 + 1 \leq N$. Therefore, the left-hand side of eq. (3.89) for $n = n_0 + 1$ becomes

$$\prod_{i=1}^{n-1} A_{n-i} = \prod_{i=1}^{(n_0+1)-1} A_{(n_0+1)-i} = \prod_{i=1}^{n_0} A_{n_0+1-i}. \tag{3.94}$$

By making an index substitution $i' = i - 1$, we get

$$\begin{aligned}\prod_{i=1}^{n-1} A_{n-i} &= \prod_{i=1}^{n_0} A_{n_0+1-i} = \prod_{i'=0}^{n_0-1} A_{n_0+1-(i'+1)} \\ &= \prod_{i'=0}^{n_0-1} A_{n_0+1-i'-1} = \prod_{i'=0}^{n_0-1} A_{n_0-i'} \\ &= A_{n_0-0} \prod_{i'=1}^{n_0-1} A_{n_0-i'} = A_{n_0} \prod_{i'=1}^{n_0-1} A_{n_0-i'}.\end{aligned} \tag{3.95}$$

Then, from the definition of A_n in eq. (3.64) and from the induction hypothesis, we have

$$\begin{aligned}\prod_{i=1}^{n-1} A_{n-i} &= A_{n_0} \prod_{i'=1}^{n_0-1} A_{n_0-i'} \\ &= \begin{bmatrix} \left(1 + w n_0^{q-1}\right) & h \\ \frac{w}{h} n_0^{q-1} & 1 \end{bmatrix} \begin{bmatrix} p_{1,1}(n_0) & p_{1,2}(n_0) \\ p_{2,1}(n_0) & p_{2,2}(n_0) \end{bmatrix} \\ &= \begin{bmatrix} \left(1 + w n_0^{q-1}\right) p_{1,1}(n_0) + h p_{2,1}(n_0) & \left(1 + w n_0^{q-1}\right) p_{1,2}(n_0) + h p_{2,2}(n_0) \\ \frac{w}{h} n_0^{q-1} p_{1,1}(n_0) + p_{2,1}(n_0) & \frac{w}{h} n_0^{q-1} p_{1,2}(n_0) + p_{2,2}(n_0) \end{bmatrix}.\end{aligned} \tag{3.96}$$

We now evaluate each entry of the resulting matrix in eq. (3.96), one at a time.

First row, first column The entry in the first row and first column of the resulting matrix in eq. (3.96) is

$$\left(1 + w n_0^{q-1}\right) p_{1,1}(n_0) + h p_{2,1}(n_0). \tag{3.97}$$

Expanding $\left(1 + w n_0^{q-1}\right) p_{1,1}(n_0)$ yields

$$\left(1 + w n_0^{q-1}\right) p_{1,1}(n_0) = p_{1,1}(n_0) + w n_0^{q-1} p_{1,1}(n_0). \tag{3.98}$$

If we substitute $p_{1,1}(n_0)$ by the right-hand side of eq. (3.90a) in the right-hand size of eq. (3.98), we obtain:

$$\begin{aligned} w n_0^{q-1} p_{1,1}(n_0) &= w n_0^{q-1} \sum_{\theta \in \mathcal{P}(\{1,\ldots,n_0-1\})} a(n_0,\theta)\, w^{\#\theta} \prod_{i\in\theta} i^{q-1} \\ &= n_0^{q-1} \sum_{\theta \in \mathcal{P}(\{1,\ldots,n_0-1\})} a(n_0,\theta)\, w^{\#\theta+1} \prod_{i\in\theta} i^{q-1} \\ &= n_0^{q-1} \sum_{\theta \in \mathcal{P}(\{1,\ldots,n_0-1\})} a(n_0,\theta)\, w^{\#(\theta\cup\{n_0\})} \prod_{i\in\theta} i^{q-1} \\ &= \sum_{\theta \in \mathcal{P}(\{1,\ldots,n_0-1\})} a(n_0,\theta)\, w^{\#(\theta\cup\{n_0\})} n_0^{q-1} \prod_{i\in\theta} i^{q-1} \\ &= \sum_{\theta \in \mathcal{P}(\{1,\ldots,n_0-1\})} a(n_0,\theta)\, w^{\#(\theta\cup\{n_0\})} \prod_{i\in(\theta\cup\{n_0\})} i^{q-1}. \end{aligned} \tag{3.99}$$

Now, let

$$P \coloneqq \left\{ (\{n_0\} \cup \theta) \in \mathcal{P}(\{1, \ldots, n_0\}) \mid \theta \in \mathcal{P}(\{1, \ldots, n_0 - 1\}) \right\}. \tag{3.100}$$

We can then rewrite eq. (3.99) as

$$\begin{aligned} w n_0^{q-1} p_{1,1}(n_0) &= \sum_{\theta \in \mathcal{P}(\{1,\ldots,n_0-1\})} a(n_0, \theta)\, w^{\#(\theta \cup \{n_0\})} \prod_{i \in (\theta \cup \{n_0\})} i^{q-1} \\ &= \sum_{\theta \in P} a(n_0, \theta - \{n_0\})\, w^{\#\theta} \prod_{i \in \theta} i^{q-1}. \end{aligned} \tag{3.101}$$

On the other hand, if we substitute $p_{1,2}(n_0)$ by the right-hand side of eq. (3.90c) in the expression $hp_{1,2}(n_0)$, we get

$$\begin{aligned} hp_{1,2}(n_0) &= h\frac{1}{h} \sum_{\theta \in \mathcal{P}(\{1,\ldots,n_0-1\})} c(n_0, \theta)\, w^{\#\theta} \prod_{i \in \theta} i^{q-1} \\ &= \sum_{\theta \in \mathcal{P}(\{1,\ldots,n_0-1\})} c(n_0, \theta)\, w^{\#\theta} \prod_{i \in \theta} i^{q-1}. \end{aligned} \tag{3.102}$$

Combining all the previous expressions, we obtain

$$\begin{aligned} \left(1 + w n_0^{q-1}\right) p_{1,1}(n_0) + hp_{2,1}(n_0) = &\sum_{\theta \in \mathcal{P}(\{1,\ldots,n_0-1\})} a(n_0, \theta)\, w^{\#\theta} \prod_{i \in \theta} i^{q-1} \\ &+ \sum_{\theta \in P} a(n_0, \theta - \{n_0\})\, w^{\#\theta} \prod_{i \in \theta} i^{q-1} \\ &+ \sum_{\theta \in \mathcal{P}(\{1,\ldots,n_0-1\})} c(n_0, \theta)\, w^{\#\theta} \prod_{i \in \theta} i^{q-1} \\ = &\sum_{\theta \in \mathcal{P}(\{1,\ldots,n_0-1\})} a(n_0, \theta)\, w^{\#\theta} \prod_{i \in \theta} i^{q-1} \\ &+ \sum_{\theta \in \mathcal{P}(\{1,\ldots,n_0-1\})} c(n_0, \theta)\, w^{\#\theta} \prod_{i \in \theta} i^{q-1} \\ &+ \sum_{\theta \in P} a(n_0, \theta - \{n_0\})\, w^{\#\theta} \prod_{i \in \theta} i^{q-1} \\ = &\sum_{\theta \in \mathcal{P}(\{1,\ldots,n_0-1\})} \big(a(n_0, \theta) + c(n_0, \theta)\big) w^{\#\theta} \prod_{i \in \theta} i^{q-1} \\ &+ \sum_{\theta \in P} a(n_0, \theta - \{n_0\})\, w^{\#\theta} \prod_{i \in \theta} i^{q-1}. \end{aligned} \tag{3.103}$$

Recalling that

$$\begin{aligned} \mathcal{P}(\{1, \ldots, n_0\}) &= \{A \in \mathcal{P}(\{1, \ldots, n_0\}) \mid n_0 \in A\} \cup \{A \in \mathcal{P}(\{1, \ldots, n_0\}) \mid n_0 \notin A\} \\ &= P \cup \mathcal{P}(\{1, \ldots, n_0 - 1\}), \end{aligned} \tag{3.104}$$

we may rewrite eq. (3.103) as

$$\begin{aligned}\left(1+wn_0^{q-1}\right)p_{1,1}\left(n_0\right)+hp_{2,1}\left(n_0\right) &= \sum_{\theta\in\mathcal{P}(\{1,\ldots,n_0-1\})}\big(a\left(n_0,\theta\right)+c\left(n_0,\theta\right)\big)w^{\#\theta}\prod_{i\in\theta}i^{q-1}\\ &\quad+\sum_{\theta\in P}a\left(n_0,\theta-\{n_0\}\right)w^{\#\theta}\prod_{i\in\theta}i^{q-1}\\ &= \sum_{\theta\in\mathcal{P}(\{1,\ldots,n_0\})}a\left(n_0+1,\theta\right)w^{\#\theta}\prod_{i\in\theta}i^{q-1},\end{aligned} \tag{3.105}$$

where

$$a\left(n_0+1,\theta\right)=\begin{cases}a\left(n_0,\theta-\{n_0\}\right) & \text{if } n_0\in\theta;\\ \big(a\left(n_0,\theta\right)+c\left(n_0,\theta\right)\big) & \text{otherwise.}\end{cases} \tag{3.106}$$

First row, second column This case is analogous to the previous one.

The entry in the first row and second column of the resulting matrix in eq. (3.96) is

$$\left(1+wn_0^{q-1}\right)p_{1,2}\left(n_0\right)+hp_{2,2}\left(n_0\right). \tag{3.107}$$

Expanding $\left(1+wn_0^{q-1}\right)p_{1,2}\left(n_0\right)$ yields

$$\left(1+wn_0^{q-1}\right)p_{1,2}\left(n_0\right)=p_{1,2}\left(n_0\right)+wn_0^{q-1}p_{1,2}\left(n_0\right). \tag{3.108}$$

If we substitute $p_{1,2}\left(n_0\right)$ by the right-hand side of eq. (3.90b) in the right-hand size of eq. (3.108), we obtain:

$$\begin{aligned}wn_0^{q-1}p_{1,2}\left(n_0\right) &= wn_0^{q-1}h\sum_{\theta\in\mathcal{P}(\{1,\ldots,n_0-1\})}b\left(n_0,\theta\right)w^{\#\theta}\prod_{i\in\theta}i^{q-1}\\ &= n_0^{q-1}h\sum_{\theta\in\mathcal{P}(\{1,\ldots,n_0-1\})}b\left(n_0,\theta\right)w^{\#\theta+1}\prod_{i\in\theta}i^{q-1}\\ &= n_0^{q-1}h\sum_{\theta\in\mathcal{P}(\{1,\ldots,n_0-1\})}b\left(n_0,\theta\right)w^{\#(\theta\cup\{n_0\})}\prod_{i\in\theta}i^{q-1}\\ &= h\sum_{\theta\in\mathcal{P}(\{1,\ldots,n_0-1\})}b\left(n_0,\theta\right)w^{\#(\theta\cup\{n_0\})}n_0^{q-1}\prod_{i\in\theta}i^{q-1}\\ &= h\sum_{\theta\in\mathcal{P}(\{1,\ldots,n_0-1\})}b\left(n_0,\theta\right)w^{\#(\theta\cup\{n_0\})}\prod_{i\in(\theta\cup\{n_0\})}i^{q-1}.\end{aligned} \tag{3.109}$$

Now, let

$$P\coloneqq\{(\{n_0\}\cup\theta)\in\mathcal{P}(\{1,\ldots,n_0\})\mid\theta\in\mathcal{P}(\{1,\ldots,n_0-1\})\}. \tag{3.110}$$

We can then rewrite eq. (3.109) as

$$\begin{aligned}wn_0^{q-1}p_{1,2}\left(n_0\right) &= h\sum_{\theta\in\mathcal{P}(\{1,\ldots,n_0-1\})}b\left(n_0,\theta\right)w^{\#(\theta\cup\{n_0\})}\prod_{i\in(\theta\cup\{n_0\})}i^{q-1}\\ &= h\sum_{\theta\in P}b\left(n_0,\theta-\{n_0\}\right)w^{\#\theta}\prod_{i\in\theta}i^{q-1}.\end{aligned} \tag{3.111}$$

On the other hand, if we substitute $p_{2,2}(n_0)$ by the right-hand side of eq. (3.90d) in the expression $hp_{2,2}(n_0)$, we get

$$hp_{2,2}(n_0) = h \sum_{\theta\in\mathcal{P}(\{1,\dots,n_0-1\})} d(n_0,\theta)\, w^{\#\theta} \prod_{i\in\theta} i^{q-1}. \tag{3.112}$$

Combining all the previous expressions, we obtain

$$\begin{aligned}\left(1+wn_0^{q-1}\right) p_{1,2}(n_0) + hp_{2,2}(n_0) &= h \sum_{\theta\in\mathcal{P}(\{1,\dots,n_0-1\})} b(n_0,\theta)\, w^{\#\theta} \prod_{i\in\theta} i^{q-1}\\ &\quad + h \sum_{\theta\in P} b(n_0,\theta-\{n_0\})\, w^{\#\theta} \prod_{i\in\theta} i^{q-1}\\ &\quad + h \sum_{\theta\in\mathcal{P}(\{1,\dots,n_0-1\})} d(n_0,\theta)\, w^{\#\theta} \prod_{i\in\theta} i^{q-1}\\ &= h \sum_{\theta\in\mathcal{P}(\{1,\dots,n_0-1\})} b(n_0,\theta)\, w^{\#\theta} \prod_{i\in\theta} i^{q-1}\\ &\quad + h \sum_{\theta\in\mathcal{P}(\{1,\dots,n_0-1\})} d(n_0,\theta)\, w^{\#\theta} \prod_{i\in\theta} i^{q-1}\\ &\quad + h \sum_{\theta\in P} b(n_0,\theta-\{n_0\})\, w^{\#\theta} \prod_{i\in\theta} i^{q-1}\\ &= h \sum_{\theta\in\mathcal{P}(\{1,\dots,n_0-1\})} \big(b(n_0,\theta)+d(n_0,\theta)\big) w^{\#\theta} \prod_{i\in\theta} i^{q-1}\\ &\quad + h \sum_{\theta\in P} b(n_0,\theta-\{n_0\})\, w^{\#\theta} \prod_{i\in\theta} i^{q-1}.\end{aligned} \tag{3.113}$$

Recalling that

$$\begin{aligned}\mathcal{P}(\{1,\dots,n_0\}) &= \{A\in\mathcal{P}(\{1,\dots,n_0\}) \mid n_0\in A\} \cup \{A\in\mathcal{P}(\{1,\dots,n_0\}) \mid n_0\notin A\}\\ &= P\cup\mathcal{P}(\{1,\dots,n_0-1\}),\end{aligned} \tag{3.114}$$

we may rewrite expression (3.117) as

$$\begin{aligned}\left(1+wn_0^{q-1}\right) p_{1,2}(n_0) + hp_{2,2}(n_0) &= h \sum_{\theta\in\mathcal{P}(\{1,\dots,n_0-1\})} \big(b(n_0,\theta)+d(n_0,\theta)\big) w^{\#\theta} \prod_{i\in\theta} i^{q-1}\\ &\quad + h \sum_{\theta\in P} b(n_0,\theta-\{n_0\})\, w^{\#\theta} \prod_{i\in\theta} i^{q-1}\\ &= h\Bigg(\sum_{\theta\in\mathcal{P}(\{1,\dots,n_0-1\})} \big(b(n_0,\theta)+d(n_0,\theta)\big) w^{\#\theta} \prod_{i\in\theta} i^{q-1}\\ &\quad + \sum_{\theta\in P} b(n_0,\theta-\{n_0\})\, w^{\#\theta} \prod_{i\in\theta} i^{q-1}\Bigg)\\ &= h \sum_{\theta\in\mathcal{P}(\{1,\dots,n_0\})} b(n_0+1,\theta)\, w^{\#\theta} \prod_{i\in\theta} i^{q-1},\end{aligned} \tag{3.115}$$

where

$$b\left(n_0+1,\theta\right)=\begin{cases} b\left(n_0,\theta-\{n_0\}\right) & \text{if } n_0\in\theta;\\ \left(b\left(n_0,\theta\right)+d\left(n_0,\theta\right)\right) & \text{otherwise.}\end{cases} \tag{3.116}$$

Second row, first column The entry in the second row and first column of the resulting matrix is

$$\frac{w}{h}n_0^{q-1}p_{1,1}\left(n_0\right)+hp_{2,1}\left(n_0\right) \tag{3.117}$$

If we substitute $p_{1,1}\left(n_0\right)$ by the right-hand side of eq. (3.90a) in the expression $\frac{w}{h}n_0^{q-1}p_{1,1}\left(n_0\right)$, we obtain

$$\begin{aligned}\frac{w}{h}n_0^{q-1}p_{1,1}\left(n_0\right)&=\frac{w}{h}n_0^{q-1}\sum_{\theta\in\mathcal{P}(\{1,\dots,n_0-1\})}a\left(n_0,\theta\right)w^{\#\theta}\prod_{i\in\theta}i^{q-1}\\&=\frac{1}{h}wn_0^{q-1}\sum_{\theta\in\mathcal{P}(\{1,\dots,n_0-1\})}a\left(n_0,\theta\right)w^{\#\theta}\prod_{i\in\theta}i^{q-1}\\&=\frac{1}{h}n_0^{q-1}\sum_{\theta\in\mathcal{P}(\{1,\dots,n_0-1\})}a\left(n_0,\theta\right)w^{\#\theta+1}\prod_{i\in\theta}i^{q-1}\\&=\frac{1}{h}n_0^{q-1}\sum_{\theta\in\mathcal{P}(\{1,\dots,n_0-1\})}a\left(n_0,\theta\right)w^{\#(\theta\cup\{n_0\})}\prod_{i\in\theta}i^{q-1}\\&=\frac{1}{h}\sum_{\theta\in\mathcal{P}(\{1,\dots,n_0-1\})}a\left(n_0,\theta\right)w^{\#(\theta\cup\{n_0\})}n_0^{q-1}\prod_{i\in\theta}i^{q-1}\\&=\frac{1}{h}\sum_{\theta\in\mathcal{P}(\{1,\dots,n_0-1\})}a\left(n_0,\theta\right)w^{\#(\theta\cup\{n_0\})}\prod_{i\in(\theta\cup\{n_0\})}i^{q-1}.\end{aligned} \tag{3.118}$$

Using the previous equation and substituting $p_{2,1}\left(n_0\right)$ by the right-hand side of eq. (3.90c) in expression (3.117), we obtain

$$\begin{aligned}\frac{w}{h}n_0^{q-1}p_{1,1}\left(n_0\right)+hp_{2,1}\left(n_0\right)&=\frac{1}{h}\sum_{\theta\in P}a\left(n_0,\theta-\{n_0\}\right)w^{\#\theta}\prod_{i\in\theta}i^{q-1}\\&+\frac{1}{h}\sum_{\theta\in\mathcal{P}(\{1,\dots,n_0-1\})}c\left(n_0,\theta\right)w^{\#\theta}\prod_{i\in\theta}i^{q-1}\\&=\frac{1}{h}\Bigg(\sum_{\theta\in P}a\left(n_0,\theta-\{n_0\}\right)w^{\#\theta}\prod_{i\in\theta}i^{q-1}\\&+\sum_{\theta\in\mathcal{P}(\{1,\dots,n_0-1\})}c\left(n_0,\theta\right)w^{\#\theta}\prod_{i\in\theta}i^{q-1}\Bigg)\\&=\frac{1}{h}\sum_{\theta\in\mathcal{P}(\{1,\dots,n_0\})}c\left(n_0+1,\theta\right)w^{\#\theta}\prod_{i\in\theta}i^{q-1},\end{aligned} \tag{3.119}$$

where

$$c\left(n_0+1,\theta\right)=\begin{cases} a\left(n_0,\theta-\{n_0\}\right) & \text{if } n_0\in\theta;\\ c\left(n_0,\theta\right) & \text{otherwise.}\end{cases} \tag{3.120}$$

Second row, second column This case is analogous to the previous one.

The entry in the second row and second column of the resulting matrix is

$$\frac{w}{h} n_0^{q-1} p_{1,2}\left(n_0\right) + h p_{2,2}\left(n_0\right) \tag{3.121}$$

If we substitute $p_{1,2}\left(n_0\right)$ by the right-hand side of eq. (3.90b) in the expression $\frac{w}{h} n_0^{q-1} p_{1,2}\left(n_0\right)$, we obtain

$$\begin{aligned} \frac{w}{h} n_0^{q-1} p_{1,2}\left(n_0\right) &= \frac{w}{h} n_0^{q-1} h \sum_{\theta \in \mathcal{P}(\{1,\ldots,n_0-1\})} b\left(n_0, \theta\right) w^{\#\theta} \prod_{i \in \theta} i^{q-1} \\ &= w n_0^{q-1} \sum_{\theta \in \mathcal{P}(\{1,\ldots,n_0-1\})} b\left(n_0, \theta\right) w^{\#\theta} \prod_{i \in \theta} i^{q-1} \\ &= n_0^{q-1} \sum_{\theta \in \mathcal{P}(\{1,\ldots,n_0-1\})} b\left(n_0, \theta\right) w^{\#\theta+1} \prod_{i \in \theta} i^{q-1} \\ &= n_0^{q-1} \sum_{\theta \in \mathcal{P}(\{1,\ldots,n_0-1\})} v\left(n_0, \theta\right) w^{\#(\theta \cup \{n_0\})} \prod_{i \in \theta} i^{q-1} \\ &= \sum_{\theta \in \mathcal{P}(\{1,\ldots,n_0-1\})} b\left(n_0, \theta\right) w^{\#(\theta \cup \{n_0\})} n_0^{q-1} \prod_{i \in \theta} i^{q-1} \\ &= \sum_{\theta \in \mathcal{P}(\{1,\ldots,n_0-1\})} b\left(n_0, \theta\right) w^{\#(\theta \cup \{n_0\})} \prod_{i \in (\theta \cup \{n_0\})} i^{q-1}. \end{aligned} \tag{3.122}$$

Using the previous equation and substituting $p_{2,2}\left(n_0\right)$ by the right-hand side of eq. (3.90d) in expression (3.121), we obtain

$$\begin{aligned} \frac{w}{h} n_0^{q-1} p_{1,2}\left(n_0\right) + h p_{2,2}\left(n_0\right) &= \sum_{\theta \in P} b\left(n_0, \theta - \{n_0\}\right) w^{\#\theta} \prod_{i \in \theta} i^{q-1} \\ &\quad + h \sum_{\theta \in \mathcal{P}(\{1,\ldots,n_0-1\})} d\left(n_0, \theta\right) w^{\#\theta} \prod_{i \in \theta} i^{q-1} \\ &= \left(\sum_{\theta \in P} b\left(n_0, \theta - \{n_0\}\right) w^{\#\theta} \prod_{i \in \theta} i^{q-1} \right. \\ &\quad \left. + \sum_{\theta \in \mathcal{P}(\{1,\ldots,n_0-1\})} d\left(n_0, \theta\right) w^{\#\theta} \prod_{i \in \theta} i^{q-1} \right) \\ &= \frac{1}{h} \sum_{\theta \in \mathcal{P}(\{1,\ldots,n_0\})} d\left(n_0+1, \theta\right) w^{\#\theta} \prod_{i \in \theta} i^{q-1}, \end{aligned} \tag{3.123}$$

where

$$d\left(n_0+1, \theta\right) = \begin{cases} b\left(n_0, \theta - \{n_0\}\right) & \text{if } n_0 \in \theta; \\ d\left(n_0, \theta\right) & \text{otherwise.} \end{cases} \tag{3.124}$$

□

The result in lemma 3 provides expressions that couple functions a and c, and functions b and d. We can therefore write c in terms of a and d in terms of b as follows.

Proposition 8. *Let $(n, \theta) \in D$ and, if θ is non-empty, let $n_1, \ldots, n_{\#\theta} \in \theta$ be all its $\#\theta$ different elements, indexed in such a way that $n_1 < \ldots < n_{\#\theta}$. Then, the following equalities hold:*

$$c\left(n, \theta\right) = \begin{cases} 0 & \text{if } \theta = \emptyset; \\ 1 & \text{if } \theta = \left\{1\right\}; \\ a\left(n_{\#\theta}, \theta - \left\{n_{\#\theta}\right\}\right) & \text{otherwise} \end{cases} \tag{3.125}$$

and

$$d\left(n, \theta\right) = \begin{cases} 1 & \text{if } \theta = \emptyset; \\ 0 & \text{if } \theta = \left\{1\right\}; \\ b\left(n_{\#\theta}, \theta - \left\{n_{\#\theta}\right\}\right) & \text{otherwise.} \end{cases} \tag{3.126}$$

Proof. The proof will be done by induction on n.

Base If $n = 2$, then, by the definition of D, $\theta \in \mathcal{P}(\{1\})$. This implies that either $\theta = \emptyset$ or $\theta = \{1\}$. By equalities (3.92), we have, in the former case,

$$c\left(2, \theta\right) = c\left(2, \emptyset\right) = 0 \tag{3.127}$$

and

$$d\left(2, \theta\right) = d\left(2, \emptyset\right) = 1. \tag{3.128}$$

On the other hand, in the latter case, we have

$$c\left(2, \theta\right) = c\left(2, \left\{1\right\}\right) = 1 \tag{3.129}$$

and

$$d\left(2, \theta\right) = d\left(2, \left\{1\right\}\right) = 0, \tag{3.130}$$

also by equalities (3.92).

Induction Suppose that there exists $n_0 \in \mathbb{N}$ with $n_0 > 2$ such that the proposition holds for all $(n, \theta) \in D$ with $n \leq n_0$. We shall then prove that the proposition still holds for $n_0 + 1$.

Let $\theta \in \mathcal{P}(\{1, \ldots, n_0\}) = \mathcal{P}(\{1, \ldots, (n_0 + 1) - 1\})$ such that $(n_0, \theta) \in D$. If $n_0 \in \theta$, then $\theta \neq \emptyset$ and $\theta \neq \{1\}$, since $n_0 > 2$. Thus, we want to show that

$$c\left(n_0 + 1, \theta\right) = a\left(n_{\#\theta}, \theta - \left\{n_{\#\theta}\right\}\right) \tag{3.131}$$

and

$$d\left(n_0 + 1, \theta\right) = b\left(n_{\#\theta}, \theta - \left\{n_{\#\theta}\right\}\right). \tag{3.132}$$

Since $n_0 \in \theta$, we can use eq. (3.91c) to get that

$$c\left(n_0 + 1, \theta\right) = a\left(\left(n_0 + 1\right) - 1, \theta - \left\{\left(n_0 + 1\right) - 1\right\}\right) = a\left(n_0, \theta - \left\{n_0\right\}\right) \tag{3.133}$$

and using eq. (3.91d) gives us that

$$d\left(n_0 + 1, \theta\right) = b\left(\left(n_0 + 1\right) - 1, \theta - \left\{\left(n_0 + 1\right) - 1\right\}\right) = b\left(n_0, \theta - \left\{n_0\right\}\right). \tag{3.134}$$

By its definition, $n_{\#\theta}$ is larger then all other elements of θ. However, since $\theta \in \mathcal{P}(\{1, \ldots, n_0\})$ and $n_0 \in \theta$, it follows that n_0 is also larger then all other elements of θ. Thus

$$n_{\#\theta} = \max \theta = n_0, \tag{3.135}$$

which means that

$$c\left(n_0 + 1, \theta\right) = a\left(n_0, \theta - \{n_0\}\right) = a\left(n_{\#\theta}, \theta - \{n_{\#\theta}\}\right) \tag{3.136}$$

and

$$d\left(n_0 + 1, \theta\right) = b\left(n_0, \theta - \{n_0\}\right) = b\left(n_{\#\theta}, \theta - \{n_{\#\theta}\}\right), \tag{3.137}$$

as desired.

On the other hand, if $n_0 \notin \theta$, again from eqs. (3.91c) and (3.91d), we still have that

$$c\left(n_0 + 1, \theta\right) = c\left(n_0, \theta\right) \tag{3.138}$$

and

$$d\left(n_0 + 1, \theta\right) = d\left(n_0, \theta\right). \tag{3.139}$$

Then, by the induction hypothesis, we know that

$$c\left(n_0 + 1, \theta\right) = c\left(n_0, \theta\right) = \begin{cases} 0 & \text{if } \theta = \emptyset; \\ 1 & \text{if } \theta = \{1\}; \\ a\left(n_{\#\theta}, \theta - \{n_{\#\theta}\}\right) & \text{otherwise} \end{cases} \tag{3.140}$$

and

$$d\left(n_0 + 1, \theta\right) = d\left(n_0, \theta\right) = \begin{cases} 1 & \text{if } \theta = \emptyset; \\ 0 & \text{if } \theta = \{1\}; \\ b\left(n_{\#\theta}, \theta - \{n_{\#\theta}\}\right), & \text{otherwise,} \end{cases} \tag{3.141}$$

which concludes the proof.

□

Proposition 8 implies that c and d depend on θ only. Thus, it is justified to replace the previous notation by $c(\theta)$ and $d(\theta)$, whenever suitable. Now, from proposition 8 and eqs. (3.91a) and (3.91b) of lemma 3, we find explicit expressions for a and b in terms of n and θ only, which also result in explicit expressions for c and d. These results are presented in the following.

Lemma 4. *Let $\tau \in (0, \alpha)$ and $(n, \theta) \in D$. If θ is non-empty, let $n_1, \ldots, n_{\#\theta} \in \theta$ be all its $\#\theta$ different elements, indexed in such a way that $n_1 < \ldots < n_{\#\theta}$. Then*

$$a\left(n, \theta\right) = \begin{cases} 1 & \textit{if } \theta = \emptyset; \\ \left(n - n_{\#\theta}\right) \displaystyle\prod_{i=2}^{\#\theta} \left(n_i - n_{i-1}\right) & \textit{otherwise} \end{cases} \tag{3.142}$$

and

$$b\left(n, \theta\right) = \begin{cases} n - 1 & \textit{if } \theta = \emptyset; \\ \left(n - n_{\#\theta}\right)\left(n_1 - 1\right) \displaystyle\prod_{i=2}^{\#\theta} \left(n_i - n_{i-1}\right) & \textit{otherwise.} \end{cases} \tag{3.143}$$

Proof. The proof will be done by induction on n.

Base If $n = 2$, then θ is either $\emptyset$ or $\{1\}$, because $(n, \theta) \in D$. If $\theta = \emptyset$, then eqs. (3.142) and (3.143) agree with equalities (3.92) which says that

$$a\,(n, \theta) = a\,(2, \emptyset) = 1 \tag{3.144}$$

and

$$b\,(n, \theta) = b\,(2, \emptyset) = 2 - 1 = 1. \tag{3.145}$$

On the other hand, if $\theta = \{1\}$, then $\#\theta = 1$ and $n_1 = n_{\#\theta} = 1$. This means that

$$(n - n_{\#\theta}) \prod_{i=2}^{\#\theta} (n_i - n_{i-1}) = (n - n_1) \prod_{i=2}^{1} (n_i - n_{i-1}) = (2 - 1) = 1 \tag{3.146}$$

and

$$\begin{aligned}(n - n_{\#\theta})\,(n_1 - 1) \prod_{i=2}^{\#\theta} (n_i - n_{i-1}) &= (n - n_1)\,(n1 - 1) \prod_{i=2}^{1} (n_i - n_{i-1}) \\ &= (n - n_1)\,(n_1 - 1) = (2 - 1)\,(1 - 1) = 0,\end{aligned} \tag{3.147}$$

which, again, agrees with equalities (3.92).

Induction Suppose that there exists $n_0 \in \mathbb{N}$ with $n_0 > 2$ such that the proposition holds for all $(n, \theta) \in D$ with $n \leq n_0$ and let $(n_0 + 1, \theta) \in D$. We will split this step into six mutually disjoint and all encompassing cases regarding the constituents of θ. These are: (1) $\theta = \emptyset$; (2) $\theta = \{1\}$; (3) $n_0 \notin \theta$, $\#\theta = 1$ and $\theta \neq \{1\}$; (4) $n_0 \notin \theta$ and $\#\theta > 1$; (5) $n_0 \in \theta$ and $\#\theta = 1$; (6) $n_0 \in \theta$ and $\#\theta > 1$.

We shall now analyze each case individually:

(1) If $\theta = \emptyset$, then $n_0 \notin \theta$, which, together with eqs. (3.91a) and (3.91b), implies that

$$\begin{aligned}a\,(n_0 + 1, \theta) &= a\,((n_0 + 1) - 1, \theta) + c\,((n_0 + 1) - 1, \theta) \\ &= a\,(n_0, \theta) + c\,(n_0, \theta)\end{aligned} \tag{3.148}$$

and

$$\begin{aligned}b\,(n_0 + 1, \theta) &= b\,((n_0 + 1) - 1, \theta) + d\,((n_0 + 1) - 1, \theta) \\ &= b\,(n_0, \theta) + d\,(n_0, \theta)\,,\end{aligned} \tag{3.149}$$

respectively.

Then, by equalities (3.92), we have that

$$\begin{aligned}a\,(n_0 + 1, \theta) &= a\,(n_0, \theta) + c\,(n_0, \theta) = a\,(n_0, \emptyset) + c\,(n_0, \emptyset) \\ &= a\,(n_0, \emptyset) + c\,(\emptyset) = a\,(n_0, \emptyset) + 0 = a\,(n_0, \emptyset)\,.\end{aligned} \tag{3.150}$$

and

$$\begin{aligned} b\left(n_0+1,\theta\right) &= b\left(n_0,\theta\right)+d\left(n_0,\theta\right)=b\left(n_0,\emptyset\right)+d\left(n_0,\emptyset\right) \\ &= b\left(n_0,\emptyset\right)+d\left(\emptyset\right)=b\left(n_0,\emptyset\right)+1. \end{aligned} \tag{3.151}$$

Thus, from the induction hypothesis, we conclude that

$$a\left(n_0+1,\theta\right)=a\left(n_0,\emptyset\right)=1 \tag{3.152}$$

and

$$b\left(n_0+1,\theta\right)=b\left(n_0,\emptyset\right)+1=\left(n_0-1\right)+1=\left(n_0+1\right)-1, \tag{3.153}$$

as desired.

(2) If $\theta=\{1\}$, then $n_0\notin\theta$, which implies that eqs. (3.148) and (3.149) are still valid. Then using equalities (3.92) in eqs. (3.148) and (3.149), we get

$$\begin{aligned} a\left(n_0+1,\theta\right) &= a\left(n_0,\theta\right)+c\left(n_0,\theta\right)=a\left(n_0,\{1\}\right)+c\left(n_0,\{1\}\right) \\ &= a\left(n_0,\{1\}\right)+c\left(\{1\}\right)=a\left(n_0,\{1\}\right)+1 \end{aligned} \tag{3.154}$$

and

$$\begin{aligned} b\left(n_0+1,\theta\right) &= b\left(n_0,\theta\right)+d\left(n_0,\theta\right)=b\left(n_0,\{1\}\right)+d\left(n_0,\{1\}\right) \\ &= b\left(n_0,\{1\}\right)+d\left(\{1\}\right)=b\left(n_0,\{1\}\right)+0=b\left(n_0,\{1\}\right). \end{aligned} \tag{3.155}$$

Then, using the induction hypothesis,

$$\begin{aligned} a\left(n_0+1,\theta\right) &= a\left(n_0,\{1\}\right)+1 \\ &= \left(n_0-n_{\#\theta}\right)\prod_{i=2}^{\#\theta}\left(n_i-n_{i-1}\right)+1 \\ &= \left(n_0-n_1\right)\prod_{i=2}^{1}\left(n_i-n_{i-1}\right)+1 \\ &= \left(n_0-n_1\right)+1=\left(n_0-1\right)+1=n_0-1+1=n_0 \end{aligned} \tag{3.156}$$

and

$$\begin{aligned} b\left(n_0+1,\theta\right) &= b\left(n_0,\{1\}\right) \\ &= \left(n_0-n_{\#\theta}\right)\left(n_1-1\right)\prod_{i=2}^{\#\theta}\left(n_i-n_{i-1}\right) \\ &= \left(n_0-n_1\right)\left(n_1-1\right)\prod_{i=2}^{1}\left(n_i-n_{i-1}\right) \\ &= \left(n_0-n_1\right)\left(n_1-1\right)=\left(n_0-1\right)\left(1-1\right)=0. \end{aligned} \tag{3.157}$$

On the other hand, substituting $n=n_0+1$ and $\theta=\{1\}$ in eqs. (3.142) and (3.143) yields

$$\begin{aligned} \left(n-n_{\#\theta}\right)\prod_{i=2}^{\#\theta}\left(n_i-n_{i-1}\right) &= \left(\left(n_0+1\right)-n_1\right)\prod_{i=2}^{1}\left(n_i-n_{i-1}\right)=\left(\left(n_0+1\right)-n_1\right) \\ &= \left(\left(n_0+1\right)-1\right)=n_0+1-1=n_0 \end{aligned} \tag{3.158}$$

and

$$\begin{aligned}(n - n_{\#\theta})(n_1 - 1)\prod_{i=2}^{\#\theta}(n_i - n_{i-1}) &= ((n_0+1) - n_1)(n_1 - 1)\prod_{i=2}^{1}(n_i - n_{i-1})\\ &= ((n_0+1) - n_1)(n_1 - 1)\\ &= ((n_0+1) - 1)(1 - 1) = 0,\end{aligned} \tag{3.159}$$

respectively, which agree with eqs. (3.156) and (3.157), also respectively.

(3) Suppose that $n_0 \notin \theta$, $\#\theta = 1$ and $\theta \neq \{1\}$. Since $n_0 \notin \theta$, then eqs. (3.148) and (3.149) still hold. Now, θ cannot be the empty set, since $\#\theta = 1$. Also, by hypothesis, $\theta \neq \{1\}$. These two conditions and eqs. (3.148) and (3.149) allow us to deduce, using proposition 8, that

$$a(n_0+1,\theta) = a(n_0,\theta) + c(n_0,\theta) = a(n_0,\theta) + a(n_{\#\theta}, \theta - \{n_{\#\theta}\}) \tag{3.160}$$

and

$$b(n_0+1,\theta) = b(n_0,\theta) + d(n_0,\theta) = b(n_0,\theta) + d(n_{\#\theta}, \theta - \{n_{\#\theta}\}), \tag{3.161}$$

respectively. Since $\#\theta = 1$, there exists $m \in \mathbb{N}$ with $2 \leq m \leq N$ and $m \neq n_0$ such that $\theta = \{m\}$, which also means that $n_1 = m = n_{\#\theta}$. This implies that $\theta - \{n_{\#\theta}\} = \emptyset$. Thus, eq. (3.160) becomes

$$a(n_0+1,\theta) = a(n_0,\theta) + a(n_{\#\theta}, \theta - \{n_{\#\theta}\}) = a(n_0,\{m\}) + a(m,\emptyset) \tag{3.162}$$

and eq. (3.161) becomes

$$b(n_0+1,\theta) = b(n_0,\theta) + b(n_{\#\theta}, \theta - \{n_{\#\theta}\}) = b(n_0,\{m\}) + b(m,\emptyset). \tag{3.163}$$

Then, the induction hypothesis then gets us to

$$\begin{aligned}a(n_0+1,\theta) &= a(n_0,\{m\}) + a(m,\emptyset)\\ &= (n_0 - n_{\#\theta})\prod_{i=2}^{\#\theta}(n_i - n_{i-1}) + a(m,\emptyset)\\ &= (n_0 - n_{\#\theta})\prod_{i=2}^{\#\theta}(n_i - n_{i-1}) + 1 \qquad (3.164)\\ &= (n_0 - n_1)\prod_{i=2}^{1}(n_i - n_{i-1}) + 1\\ &= (n_0 - n_1) + 1 = (n_0 - m) + 1 = (n_0 + 1) - m\end{aligned}$$

and

$$\begin{aligned}
b\left(n_0+1,\theta\right) &= b\left(n_0,\{m\}\right)+b\left(m,\emptyset\right)\\
&= \left(n_0-n_{\#\theta}\right)\left(n_1-1\right)\prod_{i=2}^{\#\theta}\left(n_i-n_{i-1}\right)+b\left(m,\emptyset\right)\\
&= \left(n_0-n_{\#\theta}\right)\left(n_1-1\right)\prod_{i=2}^{\#\theta}\left(n_i-n_{i-1}\right)+\left(m-1\right)\\
&= \left(n_0-n_1\right)\left(n_1-1\right)\prod_{i=2}^{1}\left(n_i-n_{i-1}\right)+\left(m-1\right)\\
&= \left(n_0-n_1\right)\left(n_1-1\right)+\left(m-1\right)\\
&= \left(n_0-m\right)\left(m-1\right)+\left(m-1\right)\\
&= \left(\left(n_0-m\right)+1\right)\left(m-1\right)\\
&= \left(\left(n_0+1\right)-m\right)\left(m-1\right).
\end{aligned} \tag{3.165}$$

On the other hand, if we substitute $n = n_0 + 1$ and $\theta = \{m\}$ in the right-hand side of eqs. (3.142) and (3.143), we have

$$\begin{aligned}
\left(n-n_{\#\theta}\right)\prod_{i=2}^{\#\theta}\left(n_i-n_{i-1}\right) &= \left(\left(n_0+1\right)-n_1\right)\prod_{i=2}^{1}\left(n_i-n_{i-1}\right)\\
&= \left(\left(n_0+1\right)-n_1\right)=\left(\left(n_0+1\right)-m\right)
\end{aligned} \tag{3.166}$$

and

$$\begin{aligned}
\left(n-n_{\#\theta}\right)\left(n_1-1\right)\prod_{i=2}^{\#\theta}\left(n_i-n_{i-1}\right) &= \left(\left(n_0+1\right)-n_1\right)\left(n_1-1\right)\prod_{i=2}^{1}\left(n_i-n_{i-1}\right)\\
&= \left(\left(n_0+1\right)-n_1\right)\left(n_1-1\right)\\
&= \left(\left(n_0+1\right)-m\right)\left(m-1\right),
\end{aligned} \tag{3.167}$$

which both agree with the equations obtained using the induction hypothesis.

(4) Suppose that $n_0 \notin \theta$, $\#\theta > 1$. By hypothesis, $n_0 \notin \theta$. Also, since $\#\theta > 1$, it follows that $\theta \neq \emptyset, \{1\}$. This means that eqs. (3.160) and (3.161) still hold. Since $\#\theta > 1$, then $\theta-\{n_{\#\theta}\} \neq \emptyset$, the largest element of $\theta-\{n_{\#\theta}\}$ is $n_{\#\theta-1}$ and $\#\left(\theta-\{n_{\#\theta}\}\right) = \#\theta-1$. Thus, using the induction hypothesis

on eqs. (3.160) and (3.161) yields

$$
\begin{aligned}
a\left(n_0+1,\theta\right) &= a\left(n_0,\theta\right) + a\left(n_{\#\theta},\theta-\left\{n_{\#\theta}\right\}\right) \\
&= \left(n_0-n_{\#\theta}\right)\prod_{i=2}^{\#\theta}\left(n_i-n_{i-1}\right) \\
&+ \left(n_{\#\theta}-n_{\#\left(\theta-\left\{n_{\#\theta}\right\}\right)}\right)\prod_{i=2}^{\#\left(\theta-\left\{n_{\#\theta}\right\}\right)}\left(n_i-n_{i-1}\right) \\
&= \left(n_0-n_{\#\theta}\right)\prod_{i=2}^{\#\theta}\left(n_i-n_{i-1}\right) + \left(n_{\#\theta}-n_{\#\theta-1}\right)\prod_{i=2}^{\#\theta-1}\left(n_i-n_{i-1}\right) \\
&= \left(n_0-n_{\#\theta}\right)\prod_{i=2}^{\#\theta}\left(n_i-n_{i-1}\right) + \prod_{i=2}^{\#\theta}\left(n_i-n_{i-1}\right) \\
&= \left(\left(n_0-n_{\#\theta}\right)+1\right)\prod_{i=2}^{\#\theta}\left(n_i-n_{i-1}\right) \\
&= \left(\left(n_0+1\right)-n_{\#\theta}\right)\prod_{i=2}^{\#\theta}\left(n_i-n_{i-1}\right)
\end{aligned}
\tag{3.168}
$$

and

$$\begin{aligned}
b\left(n_0+1,\theta\right) &= b\left(n_0,\theta\right)+b\left(n_{\#\theta},\theta-\left\{n_{\#\theta}\right\}\right)\\
&= \left(n_0-n_{\#\theta}\right)\left(n_1-1\right)\prod_{i=2}^{\#\theta}\left(n_i-n_{i-1}\right)\\
&\quad+\left(n_{\#\theta}-n_{\#\left(\theta-\left\{n_{\#\theta}\right\}\right)}\right)\left(n_1-1\right)\prod_{i=2}^{\#\left(\theta-\left\{n_{\#\theta}\right\}\right)}\left(n_i-n_{i-1}\right)\\
&= \left(n_1-1\right)\Bigg(\left(n_0-n_{\#\theta}\right)\prod_{i=2}^{\#\theta}\left(n_i-n_{i-1}\right)\\
&\quad+\left(n_{\#\theta}-n_{\#\left(\theta-\left\{n_{\#\theta}\right\}\right)}\right)\left(n_1-1\right)\prod_{i=2}^{\#\left(\theta-\left\{n_{\#\theta}\right\}\right)}\left(n_i-n_{i-1}\right)\Bigg)\\
&= \left(n_1-1\right)\Bigg(\left(n_0-n_{\#\theta}\right)\prod_{i=2}^{\#\theta}\left(n_i-n_{i-1}\right)\\
&\quad+\left(n_{\#\theta}-n_{\#\theta-1}\right)\prod_{i=2}^{\#\theta-1}\left(n_i-n_{i-1}\right)\Bigg)\\
&= \left(n_1-1\right)\left(\left(n_0-n_{\#\theta}\right)\prod_{i=2}^{\#\theta}\left(n_i-n_{i-1}\right)+\prod_{i=2}^{\#\theta}\left(n_i-n_{i-1}\right)\right)\\
&= \left(n_1-1\right)\left(\left(n_0-n_{\#\theta}\right)+1\right)\prod_{i=2}^{\#\theta}\left(n_i-n_{i-1}\right)\\
&= \left(n_1-1\right)\left(\left(n_0+1\right)-n_{\#\theta}\right)\prod_{i=2}^{\#\theta}\left(n_i-n_{i-1}\right)\\
&= \left(\left(n_0+1\right)-n_{\#\theta}\right)\left(n_1-1\right)\prod_{i=2}^{\#\theta}\left(n_i-n_{i-1}\right),
\end{aligned} \tag{3.169}$$

respectively, as wished.

(5) Suppose that $n_0 \in \theta$ and $\#\theta = 1$, i.e. $\theta = \{n_0\}$. The fact that $n_0 \in \theta$ suffices allows us to use eqs. (3.91a) and (3.91b) to write

$$a\left(n_0+1,\theta\right)=a\left(\left(n_0+1\right)-1,\theta-\left\{\left(n_0+1\right)-1\right\}\right)=a\left(n_0,\theta-\left\{n_0\right\}\right) \tag{3.170}$$

and

$$b\left(n_0+1,\theta\right)=b\left(\left(n_0+1\right)-1,\theta-\left\{\left(n_0+1\right)-1\right\}\right)=b\left(n_0,\theta-\left\{n_0\right\}\right). \tag{3.171}$$

Since $\theta = \{n_0\}$, we have

$$a\left(n_0+1,\theta\right)=a\left(n_0,\theta-\left\{n_0\right\}\right)=a\left(n_0,\left\{n_0\right\}-\left\{n_0\right\}\right)=a\left(n_0,\emptyset\right) \tag{3.172}$$

and

$$b\left(n_0+1,\theta\right)=b\left(n_0,\theta-\left\{n_0\right\}\right)=b\left(n_0,\left\{n_0\right\}-\left\{n_0\right\}\right)=b\left(n_0,\emptyset\right). \tag{3.173}$$

Then, by the induction hypothesis, we conclude that

$$a\left(n_0+1,\theta\right)=a\left(n_0,\emptyset\right)=1 \tag{3.174}$$

and

$$b\left(n_0+1,\theta\right)=b\left(n_0,\emptyset\right)=n_0-1. \tag{3.175}$$

Note now that $n_0=n_1=n_{\#\theta}$. Thus, making $n=n_0+1$ and $\theta=\{n_0\}$ in the right-hand side of eqs. (3.142) and (3.143) gives

$$\begin{aligned}\left(n-n_{\#\theta}\right)\prod_{i=2}^{\#\theta}\left(n_i-n_{i-1}\right)&=\left(\left(n_0+1\right)-n_1\right)\prod_{i=2}^{1}\left(n_i-n_{i-1}\right)\\&=\left(n_0+1\right)-n_1=\left(n_0+1\right)-n_0=1\end{aligned} \tag{3.176}$$

and

$$\begin{aligned}\left(n-n_{\#\theta}\right)\left(n_1-1\right)\prod_{i=2}^{\#\theta}\left(n_i-n_{i-1}\right)&=\left(\left(n_0+1\right)-n_1\right)\left(n_1-1\right)\prod_{i=2}^{1}\left(n_i-n_{i-1}\right)\\&=\left(\left(n_0+1\right)-n_1\right)\left(n_1-1\right)\\&=\left(\left(n_0+1\right)-n_0\right)\left(n_0-1\right)\\&=n_0-1,\end{aligned} \tag{3.177}$$

which agrees with eqs. (3.174) and (3.175).

(6) Suppose that $n_0\in\theta$ and $\#\theta>1$. Since $n_0\in\theta$, eqs. (3.170) and (3.171) are still valid. Then, the induction hypothesis gives us

$$\begin{aligned}a\left(n_0+1,\theta\right)=a\left(n_0,\theta-\{n_0\}\right)&=\left(n_0-n_{\#\left(\theta-\left\{n_{\#\theta}\right\}\right)}\right)\prod_{i=1}^{\#\left(\theta-\left\{n_{\#\theta}\right\}\right)}\left(n_i-n_{i-1}\right)\\&=\left(n_0-n_{\#\theta-1}\right)\prod_{i=1}^{\#\theta-1}\left(n_i-n_{i-1}\right)\end{aligned} \tag{3.178}$$

and

$$\begin{aligned}b\left(n_0+1,\theta\right)&=b\left(n_0,\theta-\{n_0\}\right)\\&=\left(n_0-n_{\#\left(\theta-\left\{n_{\#\theta}\right\}\right)}\right)\left(n_1-1\right)\prod_{i=1}^{\#\left(\theta-\left\{n_{\#\theta}\right\}\right)}\left(n_i-n_{i-1}\right)\\&=\left(n_0-n_{\#\theta-1}\right)\left(n_1-1\right)\prod_{i=1}^{\#\theta-1}\left(n_i-n_{i-1}\right)\end{aligned} \tag{3.179}$$

Now, the definition of D together with the facts that $\left(n_0+1,\theta\right)\in D$ and that $n_0\in\theta$ imply that $n_{\#\theta}=n_0$. Thus making $n=n_0+1$ in the right-hand side of eqs. (3.142) and (3.143) gives us

$$\begin{aligned}(n - n_{\#\theta}) \prod_{i=1}^{\#\theta} (n_i - n_{i-1}) &= ((n_0 + 1) - n_{\#\theta}) \prod_{i=1}^{\#\theta} (n_i - n_{i-1}) \\ &= ((n_0 + 1) - n_{\#\theta}) (n_{\#\theta} - n_{\#\theta-1}) \prod_{i=1}^{\#\theta-1} (n_i - n_{i-1}) \\ &= ((n_0 + 1) - n_0) (n_0 - n_{\#\theta-1}) \prod_{i=1}^{\#\theta-1} (n_i - n_{i-1}) \\ &= (n_0 - n_{\#\theta-1}) \prod_{i=1}^{\#\theta-1} (n_i - n_{i-1})\end{aligned} \tag{3.180}$$

and

$$\begin{aligned}(n - n_{\#\theta}) (n_1 - 1) \prod_{i=1}^{\#\theta} (n_i - n_{i-1}) &= ((n_0 + 1) - n_{\#\theta}) (n_1 - 1) \prod_{i=1}^{\#\theta} (n_i - n_{i-1}) \\ &= ((n_0 + 1) - n_{\#\theta}) (n_1 - 1) (n_{\#\theta} - n_{\#\theta-1}) \prod_{i=1}^{\#\theta-1} (n_i - n_{i-1}) \\ &= ((n_0 + 1) - n_0) (n_0 - n_{\#\theta-1}) (n_1 - 1) \prod_{i=1}^{\#\theta-1} (n_i - n_{i-1}) \\ &= (n_0 - n_{\#\theta-1}) (n_1 - 1) \prod_{i=1}^{\#\theta-1} (n_i - n_{i-1}),\end{aligned} \tag{3.181}$$

and these two equations agree with what was obtained via the induction hypothesis.

□

The next corollary combines the expressions of c and d in terms of a and b found in proposition 8 with the explicit expressions of a and b of the previous proposition to derive an explicit expression for c and d.

Corollary 5. *Let $(n, \theta) \in D$. If θ is non-empty, let $n_1, \ldots, n_{\#\theta} \in \theta$ be all its $\#\theta$ different elements, indexed in such a way that $n_1 < \ldots < n_{\#\theta}$. Then*

$$c(n, \theta) = \begin{cases} 0 & \textit{if } \theta = \emptyset; \\ \prod_{i=2}^{\#\theta} (n_i - n_{i-1}) & \textit{otherwise} \end{cases} \tag{3.182}$$

and

$$d(n, \theta) = \begin{cases} 1 & \textit{if } \theta = \emptyset; \\ (n_1 - 1) \prod_{i=2}^{\#\theta} (n_i - n_{i-1}) & \textit{otherwise.} \end{cases} \tag{3.183}$$

Proof. First, consider $\theta = \emptyset$. Then, by proposition 8, we have

$$c\,(n,\theta) = c\,(n,\emptyset) = 0 \tag{3.184}$$

and

$$d\,(n,\theta) = d\,(n,\emptyset) = 1, \tag{3.185}$$

which agree, respectively, with eqs. (3.182) and (3.183).

If $\theta = \{1\}$ then, by applying convention 1 to eqs. (3.182) and (3.183), we have that

$$c\,(n,\theta) = \prod_{i=2}^{\#\theta} (n_i - n_{i-1}) = \prod_{i=2}^{1} (n_i - n_{i-1}) = 1. \tag{3.186}$$

and

$$d\,(n,\theta) = (n_1 - 1) \prod_{i=2}^{\#\theta} (n_i - n_{i-1}) = (n_1 - 1) \prod_{i=2}^{1} (n_i - n_{i-1}) = n_1 - 1 = 1 - 1 = 0, \tag{3.187}$$

where we used that $n_1 = 0$, because 1 is the only element of θ. These still agree with proposition 8.

If $\#\theta = 1$ but $\theta \neq \{1\}$, then the only element of θ is $n_1 = n_{\#\theta}$ and, by proposition 8 and convention 1, we have that

$$c\,(n,\theta) = a\,(n_{\#\theta}, \theta - \{n_{\#\theta}\}) = a\,(n_{\#\theta}, \emptyset) = 1 = \prod_{n=2}^{1} (n_i - n_{i-1}) = \prod_{n=2}^{\#\theta} (n_i - n_{i-1}) \tag{3.188}$$

and

$$\begin{aligned} d\,(n,\theta) &= b\,(n_{\#\theta}, \theta - \{n_{\#\theta}\}) = b\,(n_{\#\theta}, \emptyset) = n - 1 = n_1 - 1 \\ &= (n_1 - 1) \prod_{n=2}^{1} (n_i - n_{i-1}) = (n_1 - 1) \prod_{n=2}^{\#\theta} (n_i - n_{i-1}), \end{aligned} \tag{3.189}$$

which are the RHS of eqs. (3.182) and (3.183), respectively.

Finally, if θ does not match any of the preceding cases, then θ contains at least two elements. Note then that the largest element of $\theta - \{n_{\#\theta}\}$ is $n_{\#\theta-1}$. This, along with proposition 8, implies that

$$c\,(n,\theta) = a\,(n_{\#\theta}, \theta - \{n_{\#\theta}\}) = (n_{\#\theta} - n_{\#\theta-1}) \prod_{n=2}^{\#\theta-1} (n_i - n_{i-1}) = \prod_{n=2}^{\#\theta} (n_i - n_{i-1}) \tag{3.190}$$

and

$$\begin{aligned} d\,(n,\theta) &= b\,(n_{\#\theta}, \theta - \{n_{\#\theta}\}) = (n_{\#\theta} - n_{\#\theta-1})\,(n_1 - 1) \prod_{n=2}^{\#\theta-1} (n_i - n_{i-1}) \\ &= (n_1 - 1)\,(n_{\#\theta} - n_{\#\theta-1}) \prod_{n=2}^{\#\theta-1} (n_i - n_{i-1}) = (n_1 - 1) \prod_{n=2}^{\#\theta} (n_i - n_{i-1}), \end{aligned} \tag{3.191}$$

which concludes the proof. □

In possession of an explicit expression for $\prod_{i=1}^{n-1} A_{n-i}$, we can now establish bounds for the total error given in corollary 4. This will be discussed in the following subsection.

3.2.4 Bounds for the entries of $\prod_{i=1}^{n-1} A_{n-i}$

In this section, we derive auxiliary bounds that will be used in proving theorem 1. The derivation of such bounds require the establishment of some properties of the sets

$$\mathcal{A}_{n,N} = \{\theta \in \mathcal{P}(\{1,\dots,N\}) \mid \#\theta = n\} \tag{3.192}$$

and

$$\mathcal{B}_{n,N} = \{(\theta \cup \{N\}) \in \mathcal{P}(\{1,\dots,N\}) \mid \theta \in \mathcal{A}_{n-1,N-1}\}\,, \tag{3.193}$$

where, in the definition of $\mathcal{A}_{n,N}$, n and N can be zero.

Since the construction of the sets $\mathcal{B}_{n,N}$ involve the use of the sets $\mathcal{A}_{n-1,N-1}$ and the sets $\mathcal{A}_{n,N}$ only exist for non-negative values of n and N, it is not immediate that the sets $\mathcal{B}_{n,N}$ are well defined. This is our next proposition.

Proposition 9. *The sets $\mathcal{B}_{n,N}$ are well defined for all $n, N \in \mathbb{N}$ with $n \leq N$.*

Proof. Since $n, N \in \mathbb{N}$, it follows that $n-1, N-1 \in \mathbb{N} \cup \{0\}$. And since $n \leq N$, it follows that $n-1 \leq N-1$. Thus, $A_{n-1,N-1}$ exists, which is the only point of contention on the definition of $\mathcal{B}_{n,M}$. □

An important property of the sets $\mathcal{A}_{n,N}$ is that, fixed some $n \in \mathbb{N}$, the set sequence $(\mathcal{A}_{n,N})_{N \geq n}$ is monotonically increasing in the sense of subsets, i.e. $\mathcal{A}_{n_0,N} \subset \mathcal{A}_{n_0,M}$ if $N \leq M$. This property shall be referred to hereon as "$\mathcal{A}$ sets are monotonically increasing" and is the subject of the next proposition.

Proposition 10. *Let $n, N, M \in \mathbb{N} \cup \{0\}$ with $n \leq N \leq M$. Then, it is true that $\mathcal{A}_{n,N} \subset \mathcal{A}_{n,M}$.*

Proof. Let $\theta \in \mathcal{A}_{n,N}$. Then, by the definition of $\mathcal{A}_{n,N}$, $\theta \in \mathcal{P}(\{1,\dots,N\})$ and $\#\theta = n$. Since $N \leq M$, this means that $\theta \in \mathcal{P}(\{1,\dots,M\})$ which, together with the fact that $\#\theta = n$, means that $\theta \in \mathcal{A}_{n,M}$. □

The following result states that $\mathcal{A}_{n,N}$ can be decomposed in a disjoint union of sets of the type $\mathcal{B}_{n,N}$.

Proposition 11. *Let $n, N \in \mathbb{N}$ such that $n \leq N$. Then $\mathcal{A}_{n,N} = \bigsqcup_{i=n}^{N} \mathcal{B}_{n,i}$, where $\bigsqcup$ denotes a disjoint union of sets.*

Proof. First, let $M, M' \in \mathbb{N}$ be such that $M' \neq M$. Then, without loss of generality, $M' > M$. Thus $\mathcal{B}_{n,M} \cap \mathcal{B}_{n,M'} = \emptyset$, since all the elements of $\mathcal{B}_{n,M'}$ are sets which contain M' while the elements of $\mathcal{B}_{n,M}$ are sets in $\mathcal{P}(\{1,\cdots,M\})$, which cannot contain M'. This justifies the use of $\bigsqcup$.

Suppose now that $\theta \in \mathcal{A}_{n,N}$. Let $k = \max \theta$ and $\theta' = \theta - \{k\}$. It is clear that $\theta' \in \mathcal{P}(\{1, \dots, k-1\})$ and that

$$\#\theta' = \#\theta - 1 = n - 1. \tag{3.194}$$

Thus, by definition, $\theta' \in \mathcal{A}_{n-1,k-1}$, which, again by definition, means that $\theta = \theta' \cup \{k\} \in \mathcal{B}_{n,k}$. Since $k = \max \theta$ and $\#\theta = n$, then $k \geq n$. However, since $\theta \in \mathcal{P}(\{1, \dots, N\})$, then $k \leq N$. Thus, $\mathcal{B}_{n,k} \subset \bigsqcup_{i=n}^{N} \mathcal{B}_{n,i}$.

On the other hand, suppose that $\theta \in \bigsqcup_{i=n}^{N} \mathcal{B}_{n,i}$. Thus, there exists $k \in \{n, n+1, \dots, N-1, N\}$ such that $\theta \in \mathcal{B}_{n,k}$. By definition, this means that there exists $\theta' \in \mathcal{A}_{n-1,k-1}$ such that

$$\theta = \theta' \cup \{k\}. \tag{3.195}$$

Since $\theta' \in \mathcal{A}_{n-1,k-1}$, then $\theta' \in \mathcal{P}(\{1, \dots, k-1\})$. Thus, $\theta \in \mathcal{P}(\{1, \dots, k\})$. Furthermore, since $k \notin \theta'$ (again, because $\theta' \in \mathcal{P}(\{1, \dots, k-1\})$), we have that

$$\#\theta = \#\left(\theta' \cup \{k\}\right) = \#\theta' + \#\{k\} = \#\theta' + 1. \tag{3.196}$$

However, since $\theta' \in \mathcal{A}_{n-1,k-1}$, then, by definition, $\#\theta' = n - 1$. Therefore,

$$\#\theta = \#\theta' + 1 = (n-1) + 1 = n. \tag{3.197}$$

This means, by definition, that $\theta \in \mathcal{A}_{n,k}$. Thus, since $k \leq N$ and $\mathcal{A}$ sets are monotonically increasing, we have that $\mathcal{A}_{n,k} \subset \mathcal{A}_{n,N}$, which allows us to conclude that $\theta \in \mathcal{A}_{n,N}$. □

For $n, N \in \mathbb{N}$ with $n \leq N$, consider the function $\phi_{n,N} : \mathcal{A}_{n-1,N-1} \to \mathcal{B}_{n,N}$ defined as

$$\phi_{n,N}(\theta) = \theta \cup \{N\}. \tag{3.198}$$

We will now concern ourselves with the question of well definedness of the function $\phi_{n,N}$.

Proposition 12. *Let $n, N \in \mathbb{N}$ with $n \leq N$. Then $\phi_{n,N}$ is well defined.*

Proof. First, since $n, N \in \mathbb{N}$, it follows that $n-1, N-1 \in \mathbb{N} \cup \{0\}$. And since $n \leq N$, it follows that $n - 1 \leq N - 1$. Thus, $\mathcal{A}_{n-1,N-1}$ exists.

Next, let $\theta \in \mathcal{A}_{n-1,N-1}$. Then, by definition, $\theta \subset \{1, \dots, N-1\}$ and $\#\theta = n-1$. Thus, $\phi_{n,N}(\theta) = \theta \cup \{N\}$. Then, by the definition of $\mathcal{B}_{n,N}$ and since $\theta \in \mathcal{A}_{n-1,N-1}$, it follows that $\phi_{n,N}(\theta) = \theta \cup \{N\} \in \mathcal{B}_{n,N}$ □

It will be necessary to use the fact that the function $\phi_{n,N}$ is bijective. Thus, this is our next proposition.

Proposition 13. *Let $n, N \in \mathbb{N}$ with $n \leq N$. Then, $\phi_{n,N}$ is a bijective function.*

Proof. To see that $\phi_{n,M}$ is injective, let $\theta, \theta' \in \mathcal{A}_{n-1,N-1}$ such that

$$\phi_{n,N}(\theta) = \phi_{n,N}(\theta'). \tag{3.199}$$

This means that

$$\theta \cup \{N\} = \theta' \cup \{N\}. \tag{3.200}$$

Since, by the definition of $\mathcal{A}_{n-1,N-1}$, M is neither in θ nor in θ', this means that

$$\theta = \theta \cup \{N\} - \{N\} = \theta' \cup \{N\} - \{N\} = \theta'. \tag{3.201}$$

On the other hand, let $\theta' \in \mathcal{B}_{n,N}$. Then, by the definition of $\mathcal{B}_{n,N}$, there exists $\theta \in \mathcal{A}_{n-1,N-1}$ such that $\theta' = \theta \cup \{N\}$. Then, by the definition of $\phi_{n,N}$,

$$\phi_{n,N}(\theta) = \theta \cup \{N\} = \theta', \tag{3.202}$$

i.e. $\phi_{n,N}$ is surjective. □

With the help of the functions $\phi_{n,N}$, a general inequality regarding certain sums over $\mathcal{A}_{n+1,N-1}$ is obtained in proposition 14.

Proposition 14. *Let* $n, N \in \mathbb{N} \cup \{0\}$ *with* $n+1 \le N-1$ *and* $\gamma : \mathcal{P}(\{1, \ldots, N-1\}) \to [0, \infty)$ *be a function. Then, it is true that*

$$\sum_{\theta \in \mathcal{A}_{n+1,N-1}} \gamma(\theta) \prod_{i \in \theta} i^{q-1} \le \sum_{m=n+1}^{N-1} m^{q-1} \sum_{\theta \in \mathcal{A}_{n,m-1}} \gamma(\theta \cup \{m\}) \prod_{i \in \theta} i^{q-1}. \tag{3.203}$$

Proof. Since $n \in \mathbb{N} \cup \{0\}$, $n+1 \in \mathbb{N}$ and since $N \in \mathbb{N} \cup \{0\}$ and $n+1 \le N-1$, it follows that $N-1 \in \mathbb{N}$. Thus, we can use proposition 11 for $n+1$ and $N-1$. Doing so yields

$$\mathcal{A}_{n+1,N-1} = \bigsqcup_{m=n+1}^{N-1} \mathcal{B}_{n+1,m}, \tag{3.204}$$

which implies that

$$\sum_{\theta \in \mathcal{A}_{n+1,N-1}} \gamma(\theta) \prod_{i \in \theta} i^{q-1} = \sum_{m=n+1}^{N-1} \sum_{\theta \in \mathcal{B}_{n+1,m}} \gamma(\theta) \prod_{i \in \theta} i^{q-1}. \tag{3.205}$$

Then, by proposition 13, we know that the map $\phi_{n+1,m}$ is a bijection from $\mathcal{A}_{n,m-1}$ to $\mathcal{B}_{n+1,m}$, for each $m \in \{n+1, \ldots, N-1\}$. Thus,

$$\begin{aligned}
\sum_{m=n+1}^{N-1} \sum_{\theta \in \mathcal{B}_{n+1,m}} \gamma(\theta) \prod_{i \in \theta} i^{q-1} &= \sum_{m=n+1}^{N-1} \sum_{\theta \in \mathcal{A}_{n,m-1}} \gamma(\theta \cup \{m\}) \prod_{i \in (\theta \cup \{m\})} i^{q-1} \\
&= \sum_{m=n+1}^{N-1} \sum_{\theta \in \mathcal{A}_{n,m-1}} \gamma(\theta \cup \{m\}) m^{q-1} \prod_{i \in \theta} i^{q-1} \qquad (3.206) \\
&= \sum_{m=n+1}^{N-1} m^{q-1} \sum_{\theta \in \mathcal{A}_{n,m-1}} \gamma(\theta \cup \{m\}) \prod_{i \in \theta} i^{q-1}.
\end{aligned}$$

□

Before we proceed, we shall need another external result.

External Result 2. *Let $n \in \mathbb{N}$ and $p \in \mathbb{R}$ such that $p > 0$ but $p \neq 1$. Then*

$$\sum_{k=1}^{n} \frac{1}{n^p} < 1 + \frac{n^{1-p} - 1}{1 - p}. \tag{3.207}$$

Proof. See eq. (25) of [14]. □

With the general bound obtained in proposition 14 and the external result 2 stated, inequalities for certain sums involving the functions a, b, c, d in terms of N and the total integration time τ are obtained in the following lemma. These inequalities are crucial for obtaining estimates for the entries of the matrix $\prod_{i=1}^{n-1} \mathcal{A}_{n-i}$.

Lemma 5. *Let $\tau \in (0, \alpha)$. Then, for all $n, N \in \mathbb{N}$ with $n \leq N - 1$, it is true that*

$$\sum_{\theta \in \mathcal{A}_{n,N-1}} a\left(N, \theta\right) w^{\#\theta} \prod_{i \in \theta} i^{q-1} \leq 2\tau^{n(1+q)}, \tag{3.208a}$$

$$\sum_{\theta \in \mathcal{A}_{n,N-1}} b\left(N, \theta\right) w^{\#\theta} \prod_{i \in \theta} i^{q-1} \leq 2N\tau^{n(1+q)}, \tag{3.208b}$$

$$\sum_{\theta \in \mathcal{A}_{n,N-1}} c\left(N, \theta\right) w^{\#\theta} \prod_{i \in \theta} i^{q-1} \leq \frac{2\tau^{n(1+q)}}{N}, \tag{3.208c}$$

$$\sum_{\theta \in \mathcal{A}_{n,N-1}} d\left(N, \theta\right) w^{\#\theta} \prod_{i \in \theta} i^{q-1} \leq 2\tau^{n(1+q)}. \tag{3.208d}$$

Proof. Since n must be less than or equal to $N - 1$ and n must be in $\mathbb{N}$, it follows that $N \geq 2$. Thus, let $N \in \mathbb{N}$ be such that $N \geq 2$. We will use induction on n to, initially, prove inequality (3.208c).

Base We must prove the proposition for $n = 1$. By making $n = 0$ and $\gamma\left(\theta\right) = c\left(N, \theta\right) w^{\#\theta}$ in proposition 14, we get

$$\begin{aligned}\sum_{\theta \in \mathcal{A}_{1,N-1}} c\left(N, \theta\right) w^{\#\theta} \prod_{i \in \theta} i^{q-1} &= \sum_{\theta \in \mathcal{A}_{n+1,N-1}} c\left(N, \theta\right) w^{\#\theta} \prod_{i \in \theta} i^{q-1} \\ &\leq \sum_{m=1}^{N-1} m^{q-1} \sum_{\theta \in \mathcal{A}_{0,m-1}} c\left(N, \theta \cup \{m\}\right) w^{\#(\theta \cup \{m\})} \prod_{i \in \theta} i^{q-1} \\ &= \sum_{m=1}^{N-1} m^{q-1} \sum_{\theta \in \mathcal{A}_{0,m-1}} c\left(N, \theta \cup \{m\}\right) w^{1+\#\theta} \prod_{i \in \theta} i^{q-1} \\ &= w \sum_{m=1}^{N-1} m^{q-1} \sum_{\theta \in \mathcal{A}_{0,m-1}} c\left(N, \theta \cup \{m\}\right) w^{\#\theta} \prod_{i \in \theta} i^{q-1}.\end{aligned} \tag{3.209}$$

Note that the left-hand side of the equation above is the left-hand side of inequality (3.208c) for $n = 1$. Now, since $\mathcal{A}_{0,m} = \{\emptyset\}$, for all $m \in \mathbb{N}$, we can simplify the above equation to

$$\begin{aligned}\sum_{\theta\in\mathcal{A}_{1,N-1}} c\left(N,\theta\right) w^{\#\theta}\prod_{i\in\theta} i^{q-1} &= w\sum_{m=1}^{N-1} m^{q-1}\sum_{\theta\in\mathcal{A}_{0,m-1}} c\left(N,\theta\cup\{m\}\right) w^{\#\theta}\prod_{i\in\theta} i^{q-1}\\ &= w\sum_{m=1}^{N-1} m^{q-1} c\left(N,\emptyset\cup\{m\}\right) w^{\#\emptyset}\prod_{i\in\emptyset} i^{q-1}\\ &= w\sum_{m=1}^{N-1} m^{q-1} c\left(N,\{m\}\right).\end{aligned}\tag{3.210}$$

Then, we can expand c using corollary 5, which gives

$$\begin{aligned}\sum_{\theta\in\mathcal{A}_{1,N-1}} c\left(N,\theta\right) w^{\#\theta}\prod_{i\in\theta} i^{q-1} &= w\sum_{m=1}^{N-1} m^{q-1} c\left(N,\{m\}\right) = w\sum_{m=1}^{N-1} m^{q-1}\prod_{i=2}^{\#\{m\}}\left(n_i-n_{i-1}\right)\\ &= w\sum_{m=1}^{N-1} m^{q-1}\prod_{i=2}^{1}\left(n_i-n_{i-1}\right) = w\sum_{m=1}^{N-1} m^{q-1}\\ &\leq w\sum_{m=1}^{N} m^{q-1} = w\sum_{m=1}^{N}\frac{1}{m^{1-q}}.\end{aligned}\tag{3.211}$$

We now have to evaluate the sum of the p-series that appears on the last term of eq. (3.211). Then, by external result 2, we get that

$$\sum_{m=1}^{N}\frac{1}{m^{1-q}} < 1+\frac{N^{1-(1-q)}-1}{1-(1-q)},\tag{3.212}$$

valid for $1-q\neq 1$, which is the case in this work. Then, using eq. (3.212), we obtain

$$\begin{aligned}\sum_{\theta\in\mathcal{A}_{1,N-1}} c\left(N,\theta\right) w^{\#\theta}\prod_{i\in\theta} i^{q-1} &\leq w\sum_{m=1}^{N}\frac{1}{m^{1-q}} < w\left(1+\frac{N^{1-(1-q)}-1}{1-(1-q)}\right)\\ &= w\left(1+\frac{N^q-1}{q}\right)\leq w\left(1+\frac{N^q}{q}\right)\\ &\leq w\left(\frac{N^q}{q}+\frac{N^q}{q}\right) = w\frac{2N^q}{q} = \frac{2wN^q}{q}.\end{aligned}\tag{3.213}$$

Then, by using these definitions of w, in eq. (3.63), and of h, in eq. (3.22), we finally get that

$$\begin{aligned}\sum_{\theta\in\mathcal{A}_{1,N-1}} c\left(N,\theta\right) w^{\#\theta}\prod_{i\in\theta} i^{q-1} &\leq \frac{2wN^q}{q} = \frac{2qh^{1+q}N^q}{q} = 2h^{1+q}N^q = 2h\left(Nh\right)^q\\ &= 2\frac{\tau}{N}\left(N\frac{\tau}{N}\right)^q = 2\frac{\tau}{N}\left(\tau\right)^q = \frac{2\tau^{1+q}}{N}.\end{aligned}\tag{3.214}$$

Induction Suppose that there exists $n_0 \in \mathbb{N}$ with $1 < n_0 \leq N-1$ such that the proposition holds for all $n \in \mathbb{N}$ with $1 < n \leq n_0$. If $n_0 = N-1$, the result follows trivially. Otherwise, $n_0 + 1 \leq N-1$.

By taking $\gamma(\theta) = c(N, \theta) w^{\#\theta}$ in proposition 14, we have that

$$\begin{aligned}\sum_{\theta \in \mathcal{A}_{n_0+1,N-1}} c(N,\theta) w^{\#\theta} \prod_{i\in\theta} i^{q-1} &\leq \sum_{m=n_0+1}^{N-1} m^{q-1} \sum_{\theta\in\mathcal{A}_{n_0,m-1}} c(N, \theta\cup\{m\}) w^{\#(\theta\cup\{m\})} \prod_{i\in\theta} i^{q-1} \\ &= \sum_{m=n_0+1}^{N-1} m^{q-1} \sum_{\theta\in\mathcal{A}_{n_0,m-1}} c(N, \theta\cup\{m\}) w^{1+\#\theta} \prod_{i\in\theta} i^{q-1} \\ &= w \sum_{m=n_0+1}^{N-1} m^{q-1} \sum_{\theta\in\mathcal{A}_{n_0,m-1}} c(N, \theta\cup\{m\}) w^{\#\theta} \prod_{i\in\theta} i^{q-1}.\end{aligned} \tag{3.215}$$

Now, observe that if $\theta' = \theta \cup \{m\}$, where $\theta \in \mathcal{A}_{n_0,m-1}$, then we have that, using the indexation of lemma 4, $n_{\#\theta'} = n_{\#\theta+1} = m$. This fact, together with corollary 5, leads to

$$\begin{aligned}\sum_{\theta \in \mathcal{A}_{n_0+1,N-1}} c(N,\theta) w^{\#\theta} \prod_{i\in\theta} i^{q-1} &= w\sum_{m=n_0+1}^{N-1} m^{q-1} \sum_{\theta\in\mathcal{A}_{n_0,m-1}} c(N, \theta\cup\{m\}) w^{\#\theta} \prod_{i\in\theta} i^{q-1} \\ &= w\sum_{m=n_0+1}^{N-1} m^{q-1} \sum_{\theta\in\mathcal{A}_{n_0,m-1}} \left(\prod_{i=2}^{\#(\theta\cup\{m\})} (n_i - n_{i-1})\right) w^{\#\theta} \prod_{i\in\theta} i^{q-1} \\ &= w\sum_{m=n_0+1}^{N-1} m^{q-1} \sum_{\theta\in\mathcal{A}_{n_0,m-1}} (m - n_{\#\theta}) \left(\prod_{i=2}^{\#\theta} (n_i - n_{i-1})\right) w^{\#\theta} \prod_{i\in\theta} i^{q-1}.\end{aligned} \tag{3.216}$$

Using that $m \leq N$ and $n_{\#\theta} > 0$, we have that

$$\begin{aligned}\sum_{\theta \in \mathcal{A}_{n_0+1,N-1}} c(N,\theta) w^{\#\theta} \prod_{i\in\theta} i^{q-1} &= w\sum_{m=n_0+1}^{N-1} m^{q-1} \sum_{\theta\in\mathcal{A}_{n_0,m-1}} (m - n_{\#\theta}) \left(\prod_{i=2}^{\#\theta} (n_i - n_{i-1}) w^{\#\theta}\right) \prod_{i\in\theta} i^{q-1} \\ &\leq w\sum_{m=n_0+1}^{N-1} m^{q-1} \sum_{\theta\in\mathcal{A}_{n_0,m-1}} (N - n_{\#\theta}) \left(\prod_{i=2}^{\#\theta} (n_i - n_{i-1})\right) w^{\#\theta} \prod_{i\in\theta} i^{q-1} \\ &\leq w\sum_{m=n_0+1}^{N-1} m^{q-1} \sum_{\theta\in\mathcal{A}_{n_0,m-1}} (N - 0) \left(\prod_{i=2}^{\#\theta} (n_i - n_{i-1})\right) w^{\#\theta} \prod_{i\in\theta} i^{q-1} \\ &= w\sum_{m=n_0+1}^{N-1} m^{q-1} \sum_{\theta\in\mathcal{A}_{n_0,m-1}} N \left(\prod_{i=2}^{\#\theta} (n_i - n_{i-1})\right) w^{\#\theta} \prod_{i\in\theta} i^{q-1} \\ &= wN\sum_{m=n_0+1}^{N-1} m^{q-1} \sum_{\theta\in\mathcal{A}_{n_0,m-1}} \left(\prod_{i=2}^{\#\theta} (n_i - n_{i-1})\right) w^{\#\theta} \prod_{i\in\theta} i^{q-1}.\end{aligned} \tag{3.217}$$

Then, we can substitute corollary 5 in the last line of eq. (3.217) to obtain that

$$\begin{aligned}\sum_{\theta\in\mathcal{A}_{n_0+1,N-1}} c\left(N,\theta\right) w^{\#\theta}\prod_{i\in\theta} i^{q-1} &\leq wN\sum_{m=n_0+1}^{N-1} m^{q-1}\sum_{\theta\in\mathcal{A}_{n_0,m-1}}\left(\prod_{i=2}^{\#\theta}\left(n_i-n_{i-1}\right)\right) w^{\#\theta}\prod_{i\in\theta} i^{q-1}\\ &\leq wN\sum_{m=n_0+1}^{N-1} m^{q-1}\sum_{\theta\in\mathcal{A}_{n_0,m-1}} c\left(N,\theta\right) w^{\#\theta}\prod_{i\in\theta} i^{q-1}.\end{aligned} \tag{3.218}$$

Now, $c\left(N,\theta\right) w^{\#\theta}\prod_{i\in\theta} i^{q-1}$ is always positive. Furthermore, since $m-1\leq N-1$ and by proposition 10, it is true that $\mathcal{A}_{n_0,m-1}\subset\mathcal{A}_{n_0,N-1}$. Then, we can conclude that

$$\begin{aligned}\sum_{\theta\in\mathcal{A}_{n_0+1,N-1}} c\left(N,\theta\right) w^{\#\theta}\prod_{i\in\theta} i^{q-1} &\leq wN\sum_{m=n_0+1}^{N-1} m^{q-1}\sum_{\theta\in\mathcal{A}_{n_0,m-1}} c\left(N,\theta\right) w^{\#\theta}\prod_{i\in\theta} i^{q-1}\\ &\leq wN\sum_{m=n_0+1}^{N-1} m^{q-1}\sum_{\theta\in\mathcal{A}_{n_0,N-1}} c\left(N,\theta\right) w^{\#\theta}\prod_{i\in\theta} i^{q-1}.\end{aligned} \tag{3.219}$$

Then, by the induction hypothesis, we have that

$$\begin{aligned}\sum_{\theta\in\mathcal{A}_{n_0+1,N-1}} c\left(N,\theta\right) w^{\#\theta}\prod_{i\in\theta} i^{q-1} &\leq wN\sum_{m=n_0+1}^{N-1} m^{q-1}\sum_{\theta\in\mathcal{A}_{n_0,N-1}} c\left(N,\theta\right) w^{\#\theta}\prod_{i\in\theta} i^{q-1}\\ &\leq wN\sum_{m=n_0+1}^{N-1} m^{q-1}\frac{2\tau^{n_0(1+q)}}{N}\\ &= 2w\tau^{n_0(1+q)}\sum_{m=n_0+1}^{N-1} m^{q-1}\\ &\leq 2w\tau^{n_0(1+q)}\sum_{m=n_0+1}^{N} m^{q-1}\\ &\leq 2w\tau^{n_0(1+q)}\sum_{m=2}^{N} m^{q-1}\\ &= 2w\tau^{n_0(1+q)}\sum_{m=2}^{N}\frac{1}{m^{1-q}}\\ &= 2w\tau^{n_0(1+q)}\left(\left(\sum_{m=1}^{N}\frac{1}{m^{1-q}}\right)-1\right).\end{aligned} \tag{3.220}$$

Using eq. (3.212), eq. (3.220) now reads:

$$\begin{aligned}\sum_{\theta\in\mathcal{A}_{n_0+1,N-1}} c\left(N,\theta\right) w^{\#\theta} \prod_{i\in\theta} i^{q-1} &\leq 2w\tau^{n_0(1+q)}\left(\left(\sum_{m=1}^{N}\frac{1}{m^{1-q}}\right)-1\right)\\ &\leq 2w\tau^{n_0(1+q)}\left(\left(1+\frac{N^{1-(1-q)}}{1-(1-q)}\right)-1\right)\\ &= 2w\tau^{n_0(1+q)}\frac{N^q}{q}\\ &= \frac{2wN^q\tau^{n_0(1+q)}}{q}.\end{aligned} \tag{3.221}$$

Finally, by the definitions of w, in eq. (3.63), and of h, in eq. (3.22),

$$\begin{aligned}\sum_{\theta\in\mathcal{A}_{n_0+1,N-1}} c\left(N,\theta\right) w^{\#\theta} \prod_{i\in\theta} i^{q-1} &\leq \frac{2wN^q\tau^{n_0(1+q)}}{q} = \frac{2qh^{1+q}N^q\tau^{n_0(1+q)}}{q}\\ &= 2h\left(Nh\right)^q\tau^{n_0(1+q)} = 2\frac{\tau}{N}\left(N\frac{\tau}{N}\right)^q\tau^{n_0(1+q)}\\ &= 2\frac{\tau}{N}\tau^q\tau^{n_0(1+q)} = 2\frac{\tau^{1+q}}{N}\tau^{n_0(1+q)}\\ &= \frac{2\tau^{(n_0+1)(1+q)}}{N}.\end{aligned} \tag{3.222}$$

This finishes the proof by induction of inequality (3.208c).

Now, we can use inequality (3.208c) to prove inequality (3.208a),

$$\begin{aligned}\sum_{\theta\in\mathcal{A}_{n,N-1}} a\left(N,\theta\right) w^{\#\theta} \prod_{i\in\theta} i^{q-1} &= \sum_{\theta\in\mathcal{A}_{n,N-1}}\left(\left(N-n_{\#\theta}\right)\prod_{i=2}^{\#\theta}\left(n_1-n_{i-1}\right)\right) w^{\#\theta}\prod_{i\in\theta} i^{q-1}\\ &\leq \sum_{\theta\in\mathcal{A}_{n,N-1}} N\prod_{i=2}^{\#\theta}\left(n_1-n_{i-1}\right) w^{\#\theta}\prod_{i\in\theta} i^{q-1}\\ &= N\sum_{\theta\in\mathcal{A}_{n,N-1}}\prod_{i=2}^{\#\theta}\left(n_1-n_{i-1}\right) w^{\#\theta}\prod_{i\in\theta} i^{q-1}\\ &\leq N\frac{2\tau^{n(1+q)}}{N} = 2\tau^{n(1+q)}.\end{aligned} \tag{3.223}$$

Similarly, for inequality (3.208b) we have:

$$\begin{aligned}\sum_{\theta\in\mathcal{A}_{n,N-1}} b\left(N,\theta\right) w^{\#\theta}\prod_{i\in\theta} i^{q-1} &= \sum_{\theta\in\mathcal{A}_{n,N-1}}\left(\left(N-n_{\#\theta}\right)\left(n_1-1\right)\prod_{i=2}^{\#\theta}\left(n_1-n_{i-1}\right)\right) w^{\#\theta}\prod_{i\in\theta} i^{q-1}\\ &\leq \sum_{\theta\in\mathcal{A}_{n,N-1}} N^2\prod_{i=2}^{\#\theta}\left(n_1-n_{i-1}\right) w^{\#\theta}\prod_{i\in\theta} i^{q-1}\\ &= N^2\sum_{\theta\in\mathcal{A}_{n,N-1}}\prod_{i=2}^{\#\theta}\left(n_1-n_{i-1}\right) w^{\#\theta}\prod_{i\in\theta} i^{q-1}\\ &\leq N^2\frac{2\tau^{n(1+q)}}{N} = 2N\tau^{n(1+q)}.\end{aligned} \tag{3.224}$$

Finally, inequality (3.208d) can be proven as:

$$\begin{aligned}\sum_{\theta\in\mathcal{A}_{n,N-1}} d\left(N,\theta\right) w^{\#\theta}\prod_{i\in\theta} i^{q-1} &= \sum_{\theta\in\mathcal{A}_{n,N-1}}\left(\left(n_1-1\right)\prod_{i=2}^{\#\theta}\left(n_1-n_{i-1}\right)\right) w^{\#\theta}\prod_{i\in\theta} i^{q-1}\\ &\leq \sum_{\theta\in\mathcal{A}_{n,N-1}} N\prod_{i=2}^{\#\theta}\left(n_1-n_{i-1}\right) w^{\#\theta}\prod_{i\in\theta} i^{q-1}\\ &= N\sum_{\theta\in\mathcal{A}_{n,N-1}}\prod_{i=2}^{\#\theta}\left(n_1-n_{i-1}\right) w^{\#\theta}\prod_{i\in\theta} i^{q-1}\\ &\leq N\frac{2\tau^{n(1+q)}}{N} = 2\tau^{n(1+q)}.\end{aligned} \tag{3.225}$$

□

Corollary 6. *Let* $\tau\in(0,\alpha)\cap(0,1)$ *and* $N\in\mathbb{N}$. *Then, it is true that*

$$p_{1,1}\left(N\right)\leq\frac{2}{1-\tau^{1+q}}, \tag{3.226a}$$

$$p_{1,2}\left(N\right)\leq\frac{2\tau}{1-\tau^{1+q}}, \tag{3.226b}$$

$$p_{2,1}\left(N\right)\leq\frac{2\tau^{-1}}{1-\tau^{1+q}}, \tag{3.226c}$$

$$p_{2,2}\left(N\right)\leq\frac{2}{1-\tau^{1+q}}. \tag{3.226d}$$

Proof. Since

$$\mathcal{P}(\{1,\ldots,N-1\})=\bigsqcup_{n=0}^{N-1}\mathcal{A}_{n,N-1}, \tag{3.227}$$

one can rewrite eqs. (3.90a) to (3.90d) as the following equations:

$$p_{1,1}\left(N\right)=\sum_{\theta\in\mathcal{P}(\{1,\ldots,N-1\})} a\left(N,\theta\right) w^{\#\theta}\prod_{i\in\theta} i^{q-1}=\sum_{n=0}^{N-1}\sum_{\theta\in\mathcal{A}_{n,N-1}} a\left(N,\theta\right) w^{\#\theta}\prod_{i\in\theta} i^{q-1}; \tag{3.228a}$$

$$p_{1,2}\left(N\right)=h\sum_{\theta\in\mathcal{P}(\{1,\ldots,N-1\})}b\left(N,\theta\right)w^{\#\theta}\prod_{i\in\theta}i^{q-1}=h\sum_{n=0}^{N-1}\sum_{\theta\in\mathcal{A}_{n,N-1}}b\left(N,\theta\right)w^{\#\theta}\prod_{i\in\theta}i^{q-1};\quad(3.228b)$$

$$p_{2,1}\left(N\right)=\frac{1}{h}\sum_{\theta\in\mathcal{P}(\{1,\ldots,N-1\})}c\left(N,\theta\right)w^{\#\theta}\prod_{i\in\theta}i^{q-1}=\frac{1}{h}\sum_{n=0}^{N-1}\sum_{\theta\in\mathcal{A}_{n,N-1}}c\left(N,\theta\right)w^{\#\theta}\prod_{i\in\theta}i^{q-1};\quad(3.228c)$$

$$p_{2,2}\left(N\right)=\sum_{\theta\in\mathcal{P}(\{1,\ldots,N-1\})}d\left(N,\theta\right)w^{\#\theta}\prod_{i\in\theta}i^{q-1}=\sum_{n=0}^{N-1}\sum_{\theta\in\mathcal{A}_{n,N-1}}d\left(N,\theta\right)w^{\#\theta}\prod_{i\in\theta}i^{q-1}.\quad(3.228d)$$

If we now separate the first term of the external sum and write it explicitly, we can transform the equations above as follows:

$$p_{1,1}\left(N\right)=\sum_{n=0}^{N-1}\sum_{\theta\in\mathcal{A}_{n,N-1}}a\left(N,\theta\right)w^{\#\theta}\prod_{i\in\theta}i^{q-1}=a\left(N,\emptyset\right)w^{\#\emptyset}\prod_{i\in\emptyset}i^{q-1}+\sum_{n=1}^{N-1}\sum_{\theta\in\mathcal{A}_{n,N-1}}a\left(N,\theta\right)w^{\#\theta}\prod_{i\in\theta}i^{q-1};\quad(3.229a)$$

$$p_{1,2}\left(N\right)=h\sum_{n=0}^{N-1}\sum_{\theta\in\mathcal{A}_{n,N-1}}b\left(N,\theta\right)w^{\#\theta}\prod_{i\in\theta}i^{q-1}=h\left(b\left(N,\emptyset\right)w^{\#\emptyset}\prod_{i\in\emptyset}i^{q-1}+\sum_{n=1}^{N-1}\sum_{\theta\in\mathcal{A}_{n,N-1}}b\left(N,\theta\right)w^{\#\theta}\prod_{i\in\theta}i^{q-1}\right);\quad(3.229b)$$

$$p_{2,1}\left(N\right)=\frac{1}{h}\sum_{n=0}^{N-1}\sum_{\theta\in\mathcal{A}_{n,N-1}}c\left(N,\theta\right)w^{\#\theta}\prod_{i\in\theta}i^{q-1}=\frac{1}{h}\left(c\left(N,\emptyset\right)w^{\#\emptyset}\prod_{i\in\emptyset}i^{q-1}+\sum_{n=1}^{N-1}\sum_{\theta\in\mathcal{A}_{n,N-1}}c\left(N,\theta\right)w^{\#\theta}\prod_{i\in\theta}i^{q-1}\right);\quad(3.229c)$$

$$p_{2,2}\left(N\right)=\sum_{n=0}^{N-1}\sum_{\theta\in\mathcal{A}_{n,N-1}}d\left(N,\theta\right)w^{\#\theta}\prod_{i\in\theta}i^{q-1}=d\left(N,\emptyset\right)w^{\#\emptyset}\prod_{i\in\emptyset}i^{q-1}+\sum_{n=1}^{N-1}\sum_{\theta\in\mathcal{A}_{n,N-1}}d\left(N,\theta\right)w^{\#\theta}\prod_{i\in\theta}i^{q-1}.\quad(3.229d)$$

Applying lemma 4 and corollary 5, gives us the equations:

$$\begin{aligned}p_{1,1}\left(N\right)&=a\left(N,\emptyset\right)w^{\#\emptyset}\prod_{i\in\emptyset}i^{q-1}+\sum_{n=1}^{N-1}\sum_{\theta\in\mathcal{A}_{n,N-1}}a\left(N,\theta\right)w^{\#\theta}\prod_{i\in\theta}i^{q-1}&(3.230a)\\&=1+\sum_{n=1}^{N-1}\sum_{\theta\in\mathcal{A}_{n,N-1}}a\left(N,\theta\right)w^{\#\theta}\prod_{i\in\theta}i^{q-1};\\p_{1,2}\left(N\right)&=h\left(b\left(N,\emptyset\right)w^{\#\emptyset}\prod_{i\in\emptyset}i^{q-1}+\sum_{n=1}^{N-1}\sum_{\theta\in\mathcal{A}_{n,N-1}}b\left(N,\theta\right)w^{\#\theta}\prod_{i\in\theta}i^{q-1}\right)&(3.230b)\\&=h\left(\left(N-1\right)+\sum_{n=1}^{N-1}\sum_{\theta\in\mathcal{A}_{n,N-1}}b\left(N,\theta\right)w^{\#\theta}\prod_{i\in\theta}i^{q-1}\right);\\p_{2,1}\left(N\right)&=\frac{1}{h}\left(c\left(N,\emptyset\right)w^{\#\emptyset}\prod_{i\in\emptyset}i^{q-1}+\sum_{n=1}^{N-1}\sum_{\theta\in\mathcal{A}_{n,N-1}}c\left(N,\theta\right)w^{\#\theta}\prod_{i\in\theta}i^{q-1}\right)&(3.230c)\\&=\frac{1}{h}\sum_{n=1}^{N-1}\sum_{\theta\in\mathcal{A}_{n,N-1}}c\left(N,\theta\right)w^{\#\theta}\prod_{i\in\theta}i^{q-1};\end{aligned}$$

$$\begin{aligned} p_{2,2}(N) &= d(N,\emptyset)\, w^{\#\emptyset} \prod_{i\in\emptyset} i^{q-1} + \sum_{n=1}^{N-1} \sum_{\theta\in\mathcal{A}_{n,N-1}} d(N,\theta)\, w^{\#\theta} \prod_{i\in\theta} i^{q-1} \\ &= 1 + \sum_{n=1}^{N-1} \sum_{\theta\in\mathcal{A}_{n,N-1}} c(N,\theta)\, w^{\#\theta} \prod_{i\in\theta} i^{q-1}. \end{aligned} \tag{3.230d}$$

Then, by applying lemma 5, the equations become:

$$\begin{aligned} p_{1,1}(N) &= 1 + \sum_{n=1}^{N-1} \sum_{\theta\in\mathcal{A}_{n,N-1}} a(N,\theta)\, w^{\#\theta} \prod_{i\in\theta} i^{q-1} \\ &\leq 1 + \sum_{n=1}^{N-1} 2\tau^{n(1+q)} = 1 + 2\sum_{n=1}^{N-1} \tau^{n(1+q)} \\ &\leq 2 + 2\sum_{n=1}^{N-1} \tau^{n(1+q)} = 2\sum_{n=0}^{N-1} \tau^{n(1+q)}; \end{aligned} \tag{3.231a}$$

$$\begin{aligned} p_{1,2}(N) &= h\left((N-1) + \sum_{n=1}^{N-1} \sum_{\theta\in\mathcal{A}_{n,N-1}} b(N,\theta)\, w^{\#\theta} \prod_{i\in\theta} i^{q-1}\right) \\ &\leq h\left((N-1) + \sum_{n=1}^{N-1} 2N\tau^{n(1+q)}\right) \leq h\left((N-1) + 2N\sum_{n=1}^{N-1} \tau^{n(1+q)}\right); \\ &\leq h\left(2N + 2N\sum_{n=1}^{N-1} \tau^{n(1+q)}\right) = 2hN\sum_{n=0}^{N-1} \tau^{n(1+q)}; \end{aligned} \tag{3.231b}$$

$$\begin{aligned} p_{2,1}(N) &= \frac{1}{h}\sum_{n=1}^{N-1} \sum_{\theta\in\mathcal{A}_{n,N-1}} c(N,\theta)\, w^{\#\theta} \prod_{i\in\theta} i^{q-1} \\ &\leq \frac{1}{h}\sum_{n=1}^{N-1} \frac{2\tau^{n(1+q)}}{N} = \frac{2}{Nh}\sum_{n=1}^{N-1} \tau^{n(1+q)} \\ &= \frac{1}{Nh}\left(2\sum_{n=1}^{N-1} \tau^{n(1+q)}\right) \leq \frac{1}{Nh}\left(2 + 2\sum_{n=1}^{N-1} \tau^{n(1+q)}\right) \\ &= \frac{1}{Nh}\left(2\sum_{n=0}^{N-1} \tau^{n(1+q)}\right) = \frac{2}{Nh}\sum_{n=0}^{N-1} \tau^{n(1+q)}; \end{aligned} \tag{3.231c}$$

$$\begin{aligned} p_{2,2}(N) &= 1 + \sum_{n=1}^{N-1} \sum_{\theta\in\mathcal{A}_{n,N-1}} d(N,\theta)\, w^{\#\theta} \prod_{i\in\theta} i^{q-1} \\ &\leq 1 + \sum_{n=1}^{N-1} 2\tau^{n(1+q)} = 1 + 2\sum_{n=1}^{N-1} \tau^{n(1+q)} \\ &\leq 2 + 2\sum_{n=1}^{N-1} \tau^{n(1+q)} = 2\sum_{n=0}^{N-1} \tau^{n(1+q)}. \end{aligned} \tag{3.231d}$$

Since $\tau \in (0,1)$, the partial sum $\sum_{n=0}^{N-1} \tau^{n(1+q)}$ is bounded above by the sum of the corresponding geometric series, i.e.

$$\sum_{n=0}^{N-1} \tau^{n(1+q)} = \sum_{n=0}^{N-1} \left(\tau^{1+q}\right)^n \leq \sum_{n=0}^{\infty} \left(\tau^{1+q}\right)^n = \frac{2}{1-\tau^{1+q}}. \tag{3.232}$$

Applying this bound, eqs. (3.231a) to (3.231d) become:

$$p_{1,1}(N) \leq 2\sum_{n=0}^{N-1} \tau^{n(1+q)} \leq 2\frac{1}{1-\tau^{1+q}}; \tag{3.233a}$$

$$p_{1,2}(N) \leq 2Nh\sum_{n=0}^{N-1} \tau^{n(1+q)} \leq 2Nh\frac{1}{1-\tau^{1+q}}; \tag{3.233b}$$

$$p_{2,1}(N) \leq \frac{2}{Nh}\sum_{n=0}^{N-1} \tau^{n(1+q)} \leq \frac{2}{Nh}\frac{1}{1-\tau^{1+q}}; \tag{3.233c}$$

$$p_{2,2}(N) \leq 2\sum_{n=0}^{N-1} \tau^{n(1+q)} \leq 2\frac{1}{1-\tau^{1+q}}. \tag{3.233d}$$

Noting that, from eq. (3.22), $Nh = \tau$ concludes the proof. □

3.2.5 Error bounds and order of the method

At this point, we have all the results that are needed to prove theorem 1, which now can be reformulated in a more accurate version, in terms of the total errors defined in eq. (3.54) and eq. (3.55):

Theorem 2. *Let $\tau \in (0, \alpha) \cap (0, 1)$. Then, there exists $\mathcal{K} \in (0, \infty)$ such that*

$$e_N^y \leq \mathcal{K}h^{1+q} \quad \textit{and} \quad e_N^v \leq \mathcal{K}h^{1+q}, \tag{3.234}$$

for every $N \in \mathbb{N}$ with $N \geq 2$.

Proof. Since the inequality in corollary 4 holds component-wise, it can be rewritten as the following two inequalities:

$$|e_N^y| \leq \sum_{i=1}^{N-1} \Bigg(p_{1,1}(N)\,|T_y(N-i, N, \tau)| + p_{1,2}(N)\,|T_v(N-i, N, \tau)| \Bigg); \tag{3.235}$$

$$|e_N^v| \leq \sum_{i=1}^{N-1} \Bigg(p_{2,1}(N)\,|T_y(N-i, N, \tau)| + p_{2,2}(N)\,|T_v(N-i, N, \tau)| \Bigg) \tag{3.236}$$

Using lemma 1 and inequalities (3.226a) and (3.226b) in eq. (3.235) gives us that

$$\begin{aligned}
|e_N^y| &\leq \sum_{i=1}^{N-1} \Bigg(p_{1,1}(N) \left|T_y(N-i,N,\tau)\right| + p_{1,2}(N)\left|T_v(N-i,N,\tau)\right| \Bigg) \\
&\leq \sum_{i=1}^{N-1} \left(p_{1,1}(N) K_1 h^{2+q} + p_{1,2}(N) K_2 \left(N-i-\frac{1}{2}\right)^{q-2} h^{1+q} \right) \\
&\leq \sum_{i=1}^{N-1} \left(\frac{2}{1-\tau^{1+q}} K_1 h^{2+q} + \frac{2\tau}{1-\tau^{1+q}} \left(N-i-\frac{1}{2}\right)^{q-2} K_2 h^{1+q} \right) \\
&= \sum_{i=1}^{N-1} \left(\frac{2K_1}{1-\tau^{1+q}} h^{2+q} + \frac{2\tau K_2}{1-\tau^{1+q}} \left(N-i-\frac{1}{2}\right)^{q-2} h^{1+q} \right) \\
&= \left(\sum_{i=1}^{N-1} \frac{2K_1}{1-\tau^{1+q}} h^{2+q} \right) + \left(\sum_{i=1}^{N-1} \frac{2\tau K_2}{1-\tau^{1+q}} \left(N-i-\frac{1}{2}\right)^{q-2} h^{1+q} \right) \\
&\leq \frac{2NK_1}{1-\tau^{1+q}} h^{2+q} + \frac{2\tau K_2}{1-\tau^{1+q}} h^{1+q} \sum_{i=1}^{N-1} \left(N-i-\frac{1}{2}\right)^{q-2}.
\end{aligned} \tag{3.237}$$

Recalling that $Nh = \tau$ and that $\tau < 1$, we deduce that

$$\begin{aligned}
|e_N^y| &\leq \frac{2NK_1}{1-\tau^{1+q}} h^{2+q} + \frac{2\tau K_2}{1-\tau^{1+q}} h^{1+q} \sum_{i=1}^{N-1} \left(N-i-\frac{1}{2}\right)^{q-2} \\
&= \frac{2NhK_1}{1-\tau^{1+q}} h^{1+q} + \frac{2\tau K_2}{1-\tau^{1+q}} h^{1+q} \sum_{i=1}^{N-1} \left(N-i-\frac{1}{2}\right)^{q-2} \\
&= \frac{2\tau K_1}{1-\tau^{1+q}} h^{1+q} + \frac{2\tau K_2}{1-\tau^{1+q}} h^{1+q} \sum_{i=1}^{N-1} \left(N-i-\frac{1}{2}\right)^{q-2} \\
&= \left(\frac{2K_1}{1-\tau^{1+q}} + \frac{2K_2}{1-\tau^{1+q}} \sum_{i=1}^{N-1} \left(N-i-\frac{1}{2}\right)^{q-2} \right) \tau h^{1+q} \\
&\leq \left(\frac{2K_1}{1-\tau^{1+q}} + \frac{2K_2}{1-\tau^{1+q}} \sum_{i=1}^{N-1} \left(N-i-\frac{1}{2}\right)^{q-2} \right) h^{1+q}.
\end{aligned} \tag{3.238}$$

By symmetry, the summation in the last line of the previous inequality can be rewritten as $\sum_{i=1}^{N-1} \left(i-\frac{1}{2}\right)^{q-2}$, which allows us to write

$$\begin{aligned}
\sum_{i=1}^{N-1} \left(N-i-\frac{1}{2}\right)^{q-2} &= \sum_{i=1}^{N-1} \left(i-\frac{1}{2}\right)^{q-2} = \sum_{i=1}^{N-1} \left(\frac{2i}{2}-\frac{1}{2}\right)^{q-2} = \sum_{i=1}^{N-1} \left(\frac{1}{2}(2i-1)\right)^{q-2} \\
&= \sum_{i=1}^{N-1} \left(\frac{1}{2}\right)^{q-2} (2i-1)^{q-2} = \sum_{i=1}^{N-1} 2^{2-q} (2i-1)^{q-2} \\
&= 2^{2-q} \sum_{i=1}^{N-1} (2i-1)^{q-2}.
\end{aligned} \tag{3.239}$$

Now, note that $\sum_{i=1}^{N-1}(2i-1)^{q-2}$ is a sum of all odd numbers from 1 to $2N-3$, each to the power of $q-2$. Thus, it is less than (or equal to) $\sum_{i=1}^{2N} i^{q-2}$, because all the terms are positive and all the odd terms of the previous sum are included in this one. That is to say that

$$\sum_{i=1}^{N-1}\left(N-i-\frac{1}{2}\right)^{q-2} = 2^{2-q}\sum_{i=1}^{N-1}(2i-1)^{q-2} \leq 2^{2-q}\sum_{i=1}^{2N} i^{q-2}. \tag{3.240}$$

Then, using eq. (3.212), we get the bound

$$\begin{aligned}\sum_{i=1}^{N-1}\left(N-i-\frac{1}{2}\right)^{q-2} &\leq 2^{2-q}\sum_{i=1}^{2N} i^{q-2} \leq 2^{2-q}\left(1+\frac{(2N)^{1-(2-q)}-1}{1-(2-q)}\right)\\ &= 2^{2-q}\left(1+\frac{(2N)^{q-1}-1}{q-1}\right) = 2^{2-q}\left(1+\frac{1-(2N)^{q-1}}{1-q}\right)\\ &\leq 2^{2-q}\left(1+\frac{1}{1-q}\right) = 2^{2-q}\left(\frac{1-q}{1-q}+\frac{1}{1-q}\right)\\ &= 2^{2-q}\left(\frac{1-q+1}{1-q}\right) = \frac{2^{2-q}\,(2-q)}{1-q}.\end{aligned} \tag{3.241}$$

Let

$$\mathcal{K} := 2\max\left\{\frac{2K_1}{1-\tau^{1+q}}, \frac{2^{3-q}\,(2-q)\,K_2}{(1-q)\,(1-\tau^{1+q})}\right\}. \tag{3.242}$$

Then, by inequality (3.241),

$$\begin{aligned}|e_N^y| &\leq \left(\frac{2K_1}{1-\tau^{1+q}}+\frac{2K_2}{1-\tau^{1+q}}\sum_{i=1}^{N-1}\left(N-i-\frac{1}{2}\right)^{q-2}\right)h^{1+q}\\ &\leq \left(\frac{2K_1}{1-\tau^{1+q}}+\frac{2K_2}{(1-\tau^{1+q})}\frac{2^{2-q}\,(2-q)}{1-q}\right)h^{1+q}\\ &= \left(\frac{2K_1}{1-\tau^{1+q}}+\frac{2^{3-q}\,(2-q)\,K_2}{(1-\tau^{1+q})\,(1-q)}\right)h^{1+q} \leq \mathcal{K}h^{1+q}.\end{aligned} \tag{3.243}$$

Similarly, if we start from eq. (3.236) and use lemma 1 and inequalities (3.226c) and (3.226d), we get

$$\begin{aligned}
\left|e_N^v\right| &\leq \sum_{i=1}^{N-1}\left(p_{2,1}(N)\left|T_y(N-i,N,\tau)\right| + p_{2,2}(N)\left|T_v(N-i,N,\tau)\right|\right)\\
&\leq \sum_{i=1}^{N-1}\left(p_{2,1}(N)K_1h^{2+q} + p_{2,2}(N)K_2\left(N-i-\frac{1}{2}\right)^{q-2}h^{1+q}\right)\\
&\leq \sum_{i=1}^{N-1}\left(\frac{2\tau^{-1}}{1-\tau^{1+q}}K_1h^{2+q} + \frac{2}{1-\tau^{1+q}}\left(N-i-\frac{1}{2}\right)^{q-2}K_2h^{1+q}\right)\\
&= \sum_{i=1}^{N-1}\left(\frac{2\tau^{-1}K_1}{1-\tau^{1+q}}h^{2+q} + \frac{2K_2}{1-\tau^{1+q}}\left(N-i-\frac{1}{2}\right)^{q-2}h^{1+q}\right)\\
&= \left(\sum_{i=1}^{N-1}\frac{2\tau^{-1}K_1}{1-\tau^{1+q}}h^{2+q}\right) + \left(\sum_{i=1}^{N-1}\frac{2K_2}{1-\tau^{1+q}}\left(N-i-\frac{1}{2}\right)^{q-2}h^{1+q}\right)\\
&\leq \frac{2\tau^{-1}NK_1}{1-\tau^{1+q}}h^{2+q} + \frac{2K_2}{1-\tau^{1+q}}h^{1+q}\sum_{i=1}^{N-1}\left(N-i-\frac{1}{2}\right)^{q-2}.
\end{aligned} \tag{3.244}$$

Because $Nh=\tau$, we have that $\tau^{-1}h=\dfrac{1}{N}$ and $\tau^{-1}N=h^{-1}$. Thus, we can simplify the previous equation as

$$\begin{aligned}
\left|e_N^v\right| &\leq \frac{2\tau^{-1}NK_1}{1-\tau^{1+q}}h^{2+q} + \frac{2K_2}{1-\tau^{1+q}}h^{1+q}\sum_{i=1}^{N-1}\left(N-i-\frac{1}{2}\right)^{q-2}\\
&= \frac{2h^{-1}K_1}{1-\tau^{1+q}}h^{2+q} + \frac{2K_2}{1-\tau^{1+q}}h^{1+q}\sum_{i=1}^{N-1}\left(N-i-\frac{1}{2}\right)^{q-2}\\
&= \frac{2K_1}{1-\tau^{1+q}}h^{1+q} + \frac{2K_2}{1-\tau^{1+q}}h^{1+q}\sum_{i=1}^{N-1}\left(N-i-\frac{1}{2}\right)^{q-2}\\
&= \left(\frac{2K_1}{1-\tau^{1+q}} + \frac{2K_2}{1-\tau^{1+q}}\sum_{i=1}^{N-1}\left(N-i-\frac{1}{2}\right)^{q-2}\right)h^{1+q}.
\end{aligned} \tag{3.245}$$

Then, by inequality (3.241),

$$\begin{aligned}
\left|e_N^v\right| &\leq \left(\frac{2K_1}{1-\tau^{1+q}} + \frac{2K_2}{1-\tau^{1+q}}\sum_{i=1}^{N-1}\left(N-i-\frac{1}{2}\right)^{q-2}\right)h^{1+q}\\
&\leq \left(\frac{2K_1}{1-\tau^{1+q}} + \frac{2K_2}{\left(1-\tau^{1+q}\right)}\frac{2^{2-q}(2-q)}{1-q}\right)h^{1+q}\\
&= \left(\frac{2K_1}{1-\tau^{1+q}} + \frac{2^{3-q}(2-q)K_2}{\left(1-\tau^{1+q}\right)(1-q)}\right)h^{1+q} \leq \mathcal{K}h^{1+q}.
\end{aligned} \tag{3.246}$$

□

3.2.6 Discussion

We have shown that the order of the leapfrog method, given by eq. (3.2), applied to the simplified problem, given in eq. (3.1), can be, at most, $1+q$, in agreement to the order that we observe in the numerical experiments presented in fig. 3.2. This is, therefore, an indirect confirmation that the presence of the term $\xi^{1/2}$ in the Kuwabara-Kono force model is to be blamed for the order penalization of the Verlet method in DEM simulations.

The argument used to prove this result required the total integration time τ to satisfy $\tau < 1$. Numerical experiments performed for $\tau \geq 1$ indicated that the same order penalization observed in fig. 3.2 and proved in eq. (3.234) is still present in this case. Therefore, it seems that the constraint in τ should not be a requirement for this behavior to be present, but rather a limitation of the approach used in this work to obtain the bounds that were necessary to derive 3.243. There should be another approach to this problem that could provide the adequate bounds without the need to impose $\tau < 1$. In any case, we must note that, in DEM simulations of granular materials, collisions between particles tend to be very short-lived and, in physical units, do not last longer than a few $10^{-6}s$, or $10^{-5}s$ in very rare situations [3], [71] and, therefore, the limitation $\tau < 1$ is definitely not an issue as far as physical systems are concerned.

We have chosen to initialize the leapfrog method in eq. (3.23) with the exact values of the solution at the appropriate time-steps. This choice was made in order to simplify the presentation of the argument. In fact, the initialization of the method has no relevance on the unexpected order penalization of the leapfrog method and the proof of eq. (3.234) can be readily adapted to take into account the initial error introduced by the method used in the initialization of eq. (3.23). In this case, the result presented in inequality (3.79) of corollary 3 would become

$$\vec{E}_n \leq \sum_{i=1}^{n-1} \left(\left(\prod_{j=1}^{i-1} A_{n-j} \right) \vec{T}_{n-i} \right) + \left(\prod_{j=1}^{i-1} A_{n-j} \right) \vec{E}_1, \tag{3.247}$$

where $\vec{E}_1$ is the error of the initialization method. Note that the term in which $\vec{E}_1$ appears has a very similar structure to the first term in the RHS, involving the products of the matrix A_n. Obtaining a version of corollary 4 accounting for $\vec{E}_1$ follows in a similar way as the proof presented in this work. Then, both eq. (3.235) and eq. (3.236) can be rewritten taking into account $\vec{E}_1$ and the final result would be that a similar bound to the one obtained in eq. (3.243) can be constructed, but with potentially different values of $\mathcal{K}$.

Chapter 4

A regularized force model

4.1 Introduction

In section 3.2, we showed that, since the first derivative of y^q, for $q \in (0,1)$, is unbounded near zero, the order of the leapfrog method, given in eq. (3.2) is penalized and cannot be larger than $1+q$. This indicates that the same reason is behind the fact that the Verlet method, given in eq. (1.40), cannot be of order 2 when used to integrate the motion of particles that interact via the Kuwabara-Kono force model, described in eq. (1.23).

In this section, we propose a regularization of the Kuwabara-Kono model based on the concept of mollifiers. The proposed regularization removes the unboundedness of the first derivative of $\xi^{1/2}$ near $t=0$ and allows for an order 2 convergence of the Verlet method. In section 4.2.1, we give a brief introduction to a smoothing technique called “mollification”, which has its origins in the field of mathematical analysis [23], and describe the main technical issues in the application of the mollification process. Then, in section 4.3, a new force scheme closely related to the Kuwabara-Kono force scheme is introduced and results indicating the that, with this newly proposed model, the integration with the Verlet method for damped systems is indeed second order accurate are shown. Finally, still in section 4.3, the physical consequences of the new modifications in force scheme are commented.

4.2 Prerequisites

4.2.1 Mollifiers

Let $\phi : \mathbb{R} \to \mathbb{R}$ be defined as

$$\phi\left(x\right) := \begin{cases} \dfrac{1}{C}\exp\left(\dfrac{1}{x^2-1}\right) & \text{if} -1 < x < 1; \\ 0 & \text{otherwise,} \end{cases} \tag{4.1}$$

where

$$C := \int_{-1}^{1} \exp\left(\frac{1}{x^2-1}\right) \mathrm{d}x \approx 0.444 \tag{4.2}$$

is chosen such that the integral of $\phi(x)$ equals 1. The function defined by eq. (4.1) is called the “standard mollifier” and is depicted in fig. 4.1. The interested reader can find more details as well as several properties of mollifiers in [23].

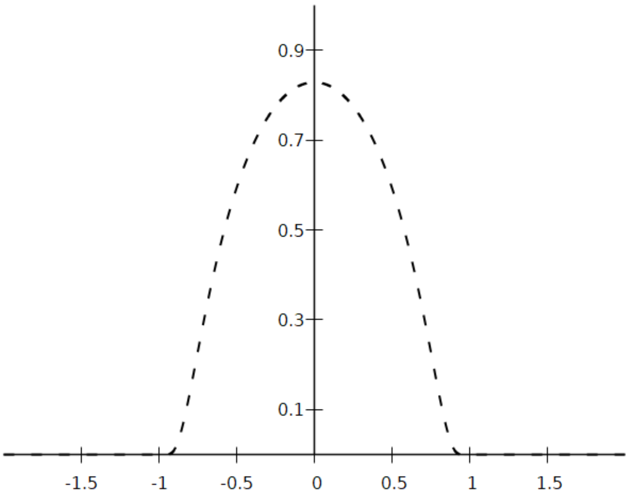


Figure 4.1: The standard mollifier.

For any $\epsilon \in (0, \infty)$, one can then define the real function

$$\phi_\epsilon(x) := \frac{1}{\epsilon}\phi\left(\frac{x}{\epsilon}\right). \tag{4.3}$$

Given $f : \mathbb{R} \to \mathbb{R}$ a locally integrable function, one can define its ϵ-mollification as the convolution of ϕ_ϵ and f. This convolution produces a new real function denoted by $\phi_\epsilon * f$ and which is given by:

$$(\phi_\epsilon * f)(x) = \int_{-\epsilon}^{\epsilon} \phi_\epsilon(z) f(x - z)\,\mathrm{d}z. \tag{4.4}$$

The new function originated by eq. (4.4) has the desired property of being infinitely differentiable in $\mathbb{R}$, while being almost the same function as the original function f, as stated by the result below.

Theorem 3. *Let $f : \mathbb{R} \to \mathbb{R}$ a locally integrable function. Then, the following statements are true:*

(1) $\phi_\epsilon * f$ is infinitely differentiable;

(2) $\phi_\epsilon * f \to f$ almost everywhere as $\epsilon \to 0$;

(3) If f is continuous, then $\phi_\epsilon * f \to f$ uniformly on compact subsets of $\mathbb{R}$.

Proof. See [23]. □

In other words, $\phi_\epsilon * f$ is an infinitely differentiable approximation to f. How well it approximates f depends on how small ϵ is taken. An example will be discussed in section 4.3. In this chapter, all mollifications were computed by using midpoint rule with a partition size of 1000.

4.2.2 Extended square root and ϵ-shift

In order to properly calculate the integral in eq. (4.4), the function f must be defined on $[-\epsilon, \infty)$. In the case of the square root function, which is the one appearing in the dissipation term of eq. (1.23), this is not true. For this reason, we continuously extend the ordinary square root function to negative numbers by making $\sqrt{x} = 0$ if $x < 0$. This extension is denoted by $\sqrt{\cdot} : \mathbb{R} \to \mathbb{R}$ and will substitute the traditional square root function from now on.

If we use the extented square root function to calculate eq. (4.3) we observe that the ϵ-mollification of $\sqrt{\cdot}$ is not zero when $x = 0$, as illustrated in fig. 4.2(a). In fact, notice that the integrand of eq. (4.4) becomes $\phi_\epsilon(z)\sqrt{-z}$ for $x = 0$. Then, if $z \in (-\epsilon, 0)$, the ϵ-mollifier $\phi_\epsilon(z)$ and $\sqrt{-z}$ are non-zero. Since both functions are strictly positive when they are not zero, this results in a positive value for the overall convolution. This could pose a problem for the force model, since it would mean that a non-zero normal force would exist between two particles which are not in contact. A way to prevent this is to right-shift the function $\sqrt{\cdot}$ by ϵ. That is, for each $\epsilon \in (0, \infty)$, define the right-shift function $\tau_\epsilon : \mathbb{R} \to \mathbb{R}$ as

$$\tau_\epsilon(x) = x - \epsilon. \tag{4.5}$$

Then, the ϵ-mollification of $\sqrt{\cdot} \circ \tau_\epsilon$, which is depicted in fig. 4.2(b), is always zero when $x = 0$, i.e. when there is no contact between the particles. For convenience of notation, from now on we define

$${}_\epsilon\sqrt{\cdot} := \phi_\epsilon * \left(\sqrt{\cdot} \circ \tau_\epsilon\right). \tag{4.6}$$

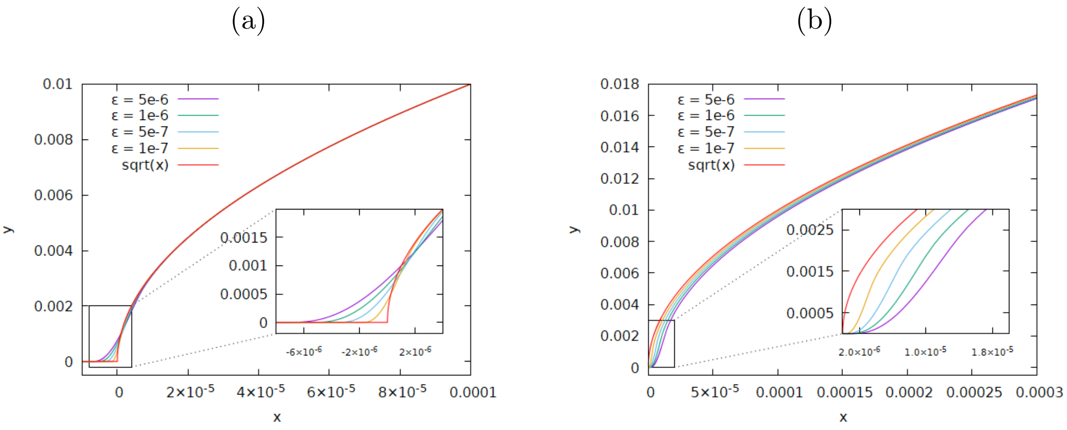


Figure 4.2: (a) Comparison between $\sqrt{\cdot}$ and the ϵ-mollification of $\sqrt{\cdot}$ for different values of ϵ. (b) Comparison between $\sqrt{\cdot}$ and the ϵ-mollification of $\left(\sqrt{\cdot} \circ \tau_\epsilon\right)$, ${}_\epsilon\sqrt{\cdot}$, for different values of ϵ. In both plots, the red curve is $\sqrt{x}$, for comparison. Notice how all curves in (b) go through $(0, 0)$. In both figures, mollifications were computed using a composite midpoint rule with 1000 sub-intervals.

4.2.3 Efficiently computing ${}_\epsilon\sqrt{\cdot}$

Since the expression for the value of an ϵ-mollification is given by an integral on a limited domain (see eq. (4.4)), one can approximate its value by any of many different integration methods available in the literature [4]. However, naively applying any of these methods to eq. (4.4) yields approximations which become very inaccurate near 0, as shown in fig. 4.3(a). These inaccuracies present themselves in two forms: a "delay" before the approximation becomes non-zero and a non-smooth behaviour in some points. These inaccuracies are caused by the finite number of samples each integration method uses to approximate eq. (4.4). This can be better understood by the schematic plots presented in fig. 4.4(a).

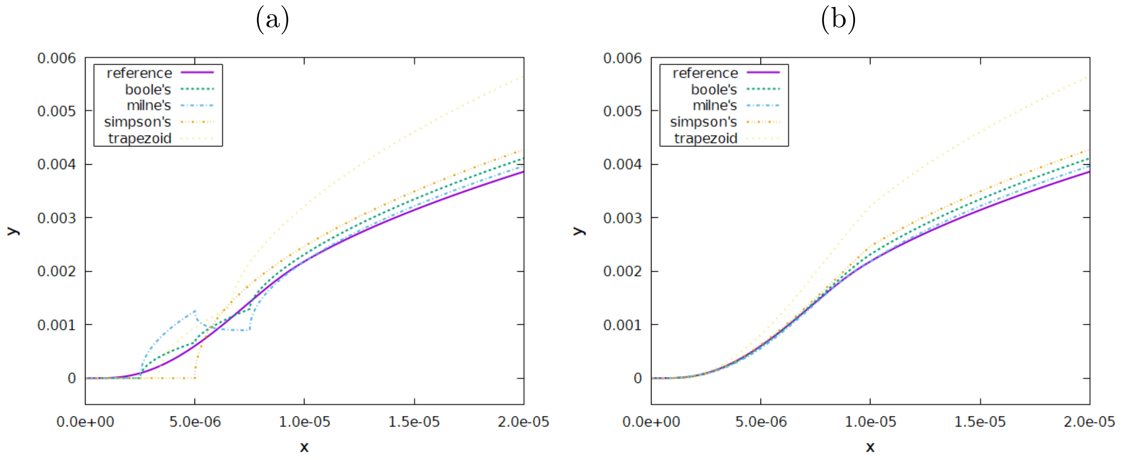


Figure 4.3: (a) Comparison of the approximations of ${}_{\epsilon}\sqrt{\cdot}$, for $\epsilon = 5 \times 10^{-6}$, yielded by different integration methods. (b) Same comparisons, but the integration domain was changed at each x so to match the support of the integrand. In either plot, the same reference curve was used and it was computed using a composite midpoint rule with 1000 sub-intervals.

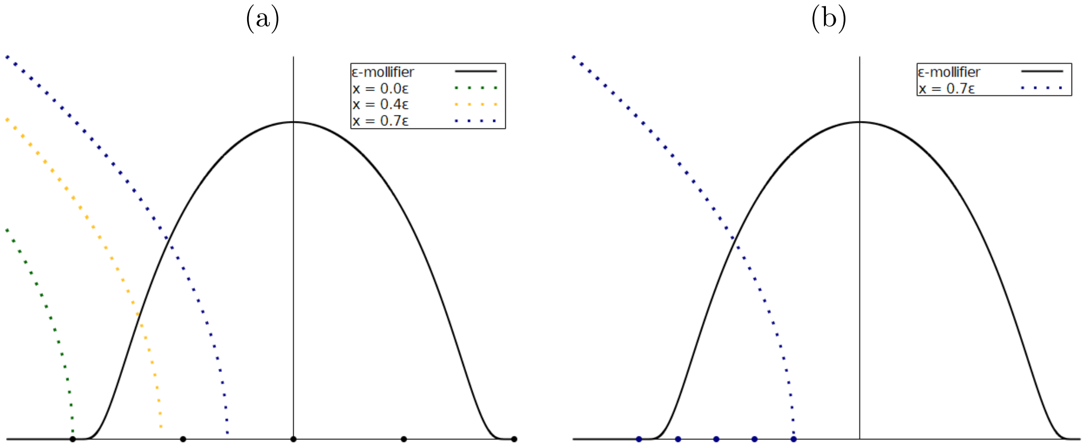


Figure 4.4: Visualization of the factors in the integrand of $({}_{\epsilon}\sqrt{\cdot})(x)$ and the sampling points of the numerical integration method. The straight black lines represent the standard axes. Notice how the support of the product of both functions changes as the point where the convolution is calculated changes. (a) The black dots are the sampled points of the numerical integration method applied to the interval $[-\epsilon, \epsilon]$. (b), the blue dots are the sampled points of the numerical integration method applied only to the support of the product of the integrands, which in this case is $[-\epsilon, -0.3\epsilon]$.

Given $x_0 \in [0, \infty)$, the factors of the integrands of the convolution in ${}_{\epsilon}\sqrt{x_0}$ take the form of $\phi_\epsilon(z)$ and $\sqrt{(x_0 - z - \epsilon)}$. This latter factor is just the square root function right-shifted by $(\epsilon - x_0)$ and then mirrored about the vertical axis. This is illustrated in the colored curves of fig. 4.4(a). When computing ${}_\epsilon\sqrt{x_0}$ with a numerical integration method, the aforementioned inaccuracies arise as the value x_0 crosses a sampling point of the integration method. This causes the approximation to be calculated with more samples, which changes its behavior. For instance, in fig. 4.4(a), the yellow curve indicates that the integral is calculated with only one sampling of the function while for the blue curve the integral is calculated with two points.

One can also observe from fig. 4.4(a) that the support of the product of the integrands of ${}_\epsilon\sqrt{x}$ starts as the empty set when $x = 0$ and increase with x until it becomes the whole interval $[-\epsilon, \epsilon]$ when $x = 2\epsilon$. This suggests that one can improve the approximations of ${}_\epsilon\sqrt{x}$ near $x = 0$ by applying the integration method on the support of the integrand, instead of the entire interval $[-\epsilon, \epsilon]$, as shown in fig. 4.4(b). It can be readily seen that this support is $[-\epsilon, \min\{x - \epsilon, \epsilon\}]$. The results of using this approach to calculate ${}_\epsilon\sqrt{\cdot}$ are displayed in fig. 4.3(b).

4.3 Regularized normal force model

4.3.1 Description and computational validation

Based on the discussion in the subsections above, we propose a regularized model for the normal contact force in which the $\sqrt{\cdot}$ term in eq. (1.23) is substituted by ${}_\epsilon\sqrt{\cdot}$, as defined in eq. (4.6). Therefore, we obtain

$$|\mathbf{F}_N^i(t)| = k\xi(t)^{3/2} + \gamma\xi'(t)\,{}_\epsilon\sqrt{\xi(t)} \tag{4.7}$$

With the removal of the singularity of the first derivative of ${}_\epsilon\sqrt{\xi(t)}$ at $t = 0$, the error of the Verlet method given in eq. (1.40) should decrease as $\mathcal{O}(h^2)$.

We performed simulations for different values of ϵ, in order to evaluate the effect that the mollification parameter has on the overall error behavior. The results are displayed in fig. 4.5 and indicate that, for all values of expected order of the Verlet method in eq. (1.40) was obtained, i.e. the error decreases as $\mathcal{O}(h^2)$. In each of the plots in fig. 4.5, one observes the existence of three distinct regions. The boundary between these regions is highlighted in fig. 4.5(a). For larger h, we observe a region where the error, although monotonically decreasing, does not have with a clear order. Then, as h decreases, there is an intermediate region where the relative error oscillates, the length of which depends on ϵ, and finally, for smaller h, a region where the error becomes monotonically decreasing again, but now decreasing as $\mathcal{O}(h^2)$.

The existence of these regions where the error fluctuates can be explained based on the discussion of section 4.2.3 and, more specifically, by the observations made in the inset of fig. 4.2(b), where it is shown that ${}_\epsilon\sqrt{x}$ and $\sqrt{x}$ are most different when x is close to zero. For values of h larger than about 2ϵ, the behavior of ${}_\epsilon\sqrt{x}$ near 0 is never relevant since, in the first step of the integration, the particle will already have crossed the region where ${}_\epsilon\sqrt{x}$ deviates the most from $\sqrt{x}$ and the duration of the simulation for the order analysis is too short for the particle to have time to bounce back. Thus, ${}_\epsilon\sqrt{x}$ is effectively very close to $\sqrt{x}$ and the model in eq. (4.7) behaves almost as the original Kuwabara-Kono model in eq. (1.23). This is behind the rightmost regions of the plots in fig. 4.5. However, as h becomes significantly

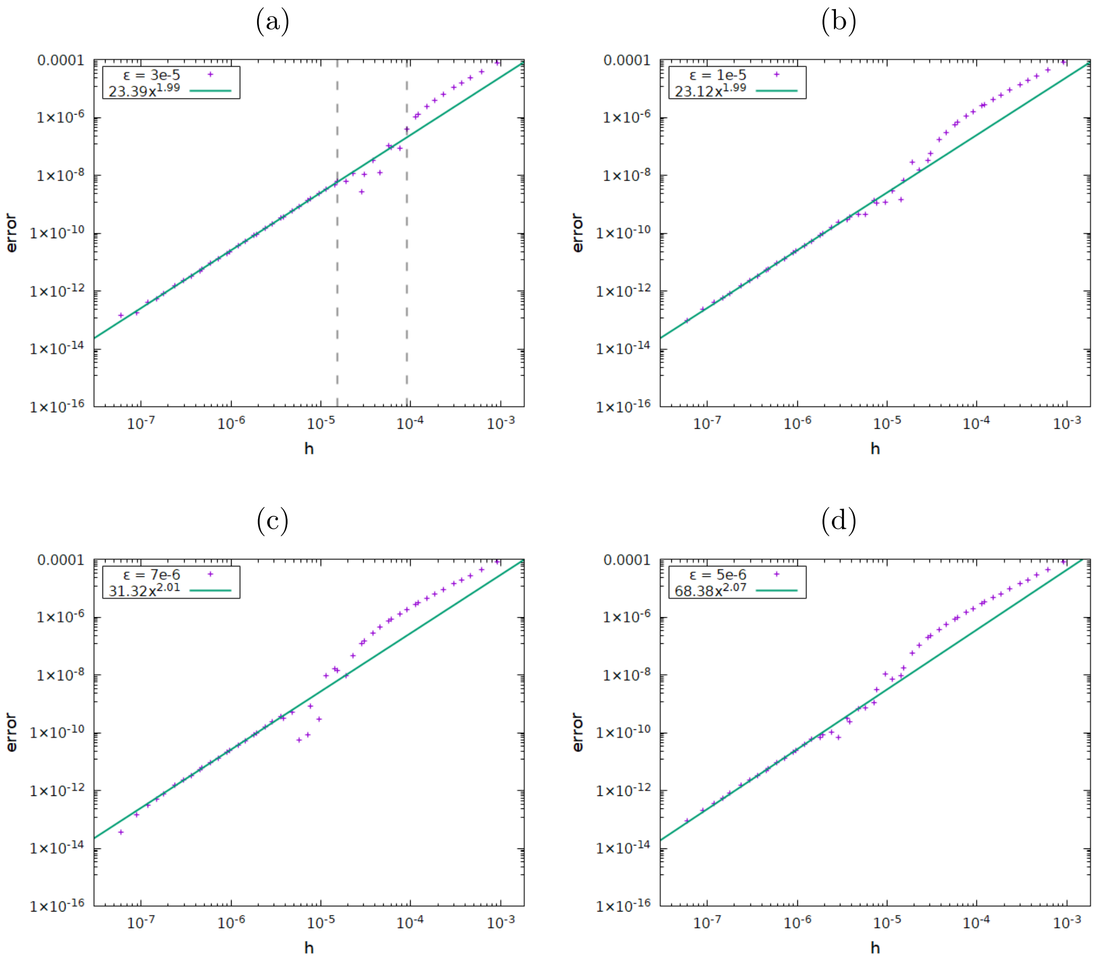


Figure 4.5: Order analyses of the damped Verlet method associated with the Kuwabara-Kono force model where $\sqrt{\cdot}$ is substituted by ${}_{\epsilon}\sqrt{\cdot}$. The physical system being simulated is a binary normal collision and the position of one of these particles is the variable whose order is being analyzed. The values of the parameters used in the simulations are presented in table 3.1.

smaller than 2ϵ, i.e. the leftmost regions in fig. 4.5, the collision of the particles is very well resolved, that is, there will be many time steps in the region of ${}_{\epsilon}\sqrt{x}$ near 0, which means that the integration of the forces near 0 will be well resolved. In these left-most regions, the $\mathcal{O}(h^2)$ convergence as $h \to 0$ is observed. Finally, for intermediary values of h, the integration of the forces will not sample ${}_{\epsilon}\sqrt{x}$ near 0 enough times, which causes the erratic behavior observed in the middle region of fig. 4.5(a).

Therefore, in order to effectively use the regularized model proposed in eq. (4.7), it is necessary that the value of h belongs to the leftmost region of the plots in fig. 4.5, for the chosen value of ϵ. Thus, an adequate combination of ϵ and h must be selected. To quantitatively understand how this selection must be made, we performed order analyses for values of ϵ ranging from 5×10^{-7} to 9×10^{-5}. The values of h used in these order analyses were of the form $m \times 2^{-k}$, where $m \in \{1, 3, 5, 7, 15, 21, 35, 105\}$ and $k \in \mathbb{N} \cap [13, 27]$. We choose these values of m so that h is exactly represented as a double precision floating point number, while the total integration time, which must divide all of the possible values of h, is kept relatively low. For each of these analyses, we selected the biggest value of h, called h_ϵ, such that the error decreases as $\mathcal{O}(h^2)$ for all $h < h_\epsilon$. The results are presented in fig. 4.6. Any pair (ϵ, h_ϵ) below the solid line is a valid choice for which the regularized model in 4.7 integrated with eq. (1.40) will produce an (expected) $\mathcal{O}(h^2)$ decay of the error.

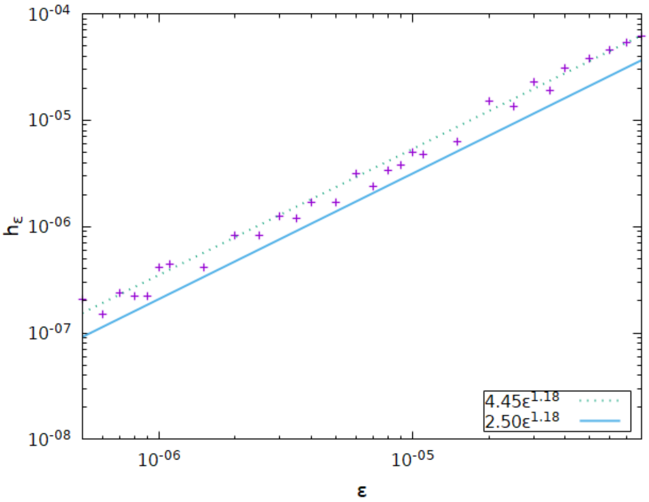


Figure 4.6: Computed value of h_ϵ as a function of ϵ. The dashed line is the best fit of the data to a power law, while the solid line is “safe” choice for h based on all values of ϵ tested. The values of the parameters used in the simulations are presented in table 3.1.

4.3.2 Physical behavior of the regularized model

One can expect that regularizing the force scheme will affect the physical behavior of the contact. In order to evaluate the effect of the regularized model proposed in eq. (4.7) on the physics of the problem, we performed some simulations to determine the influence of the regularization of the normal force, in particular the effects of the choice of the ϵ parameter of eq. (4.7). We also performed qualitative comparisons of the coefficient of restitution of binary collision using eq. (4.7) with the numerical results obtained with eq. (1.23) presented in [71]. The parameters for the simulations performed in this sections are presented in table 4.1, which are identical to the ones used in [71].

Table 4.1: Values of the parameters used to perform the simulations used for validation of physical behavior of the binary collisions.

Particles	Normal Forces	Simulation
$\rho = 1300\,\mathrm{kg/m^3}$	$\tilde{k}_n \approx 9 \times 10^7\,\mathrm{N/m^{1.5}}$	$\Delta t = 2^{-23}\,\mathrm{s}$
$r = 3\,\mathrm{mm}$	$\gamma = 190\,\mathrm{kg/m^{0.5}s}$	$(\approx 1.19 \times 10^{-7}\,\mathrm{s})$

The trajectory of the moving particle in a collision, in an identical setup to the one described in section 3.1.1, was calculated for different values of ϵ using eq. (4.7). The results are shown in fig. 4.7. We observe that the trajectories of the particle obtained from eq. (4.7) are very similar to those obtained with the Kuwabara-Kono force model, eq. (1.23). There are very minor changes in the duration of the collision and the most noticeable difference is that the the maximum overlap of the particles which, in this setup, corresponds to the x coordinate, increases with increasing ϵ.

Therefore, the first conclusion is that the regularized force model slightly softens the materials, allowing for slightly larger maximum overlaps during the collisions.

The Kuwabara-Kono force model, given in eq. (1.23), produces coefficients of normal restitution, defined as the ratio between the velocities before and after the collision,

$$e_n = \frac{|\mathbf{v}_{after}|}{|\mathbf{v}_{before}|}, \tag{4.8}$$

that depend on the normal impact velocity [71]. This dependence cannot be easily inferred from fig. 4.7. Thus, simulations were performed in which the coefficient of normal restitution was measured for different impact velocities and for different values of ϵ. The results can be seen in fig. 4.8.

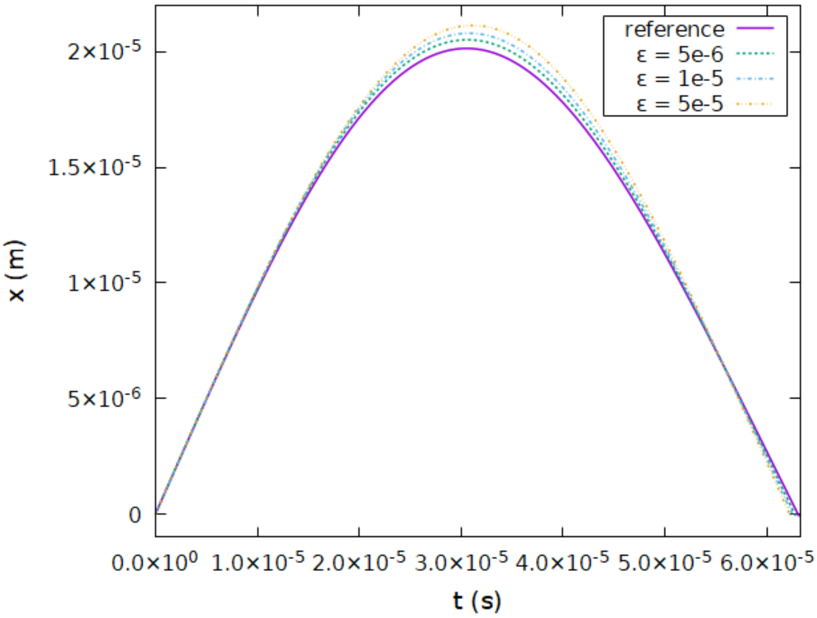


Figure 4.7: Time evolution of the position of the mobile particle in a normal binary collision~~,~~ using the regularized model in eq. (4.7), for different values of ϵ. The reference curve uses the Kuwabara-Kono force model, given in eq. (1.23), instead. The impact velocity used is 1 m/s and the values of the remaining parameters are presented in table 4.1.

As shown in fig. 4.8, collisions become more elastic (e_n closer to 1) as the value of ϵ increases, especially for very low normal impact velocities. The maximum overestimation of the restitution coefficient with respect to the reference value is of about 17% for normal impact velocities just under 1 meter per second, when $\epsilon = 10^{-5}$. This effect, however, becomes significantly less pronounced as the normal impact velocities increase, especially as smaller values of ϵ can be chosen. In conclusion, from fig. 4.8 we observe that the regularized model in eq. (4.7) produces slightly more elastic responses and hinders the efficiency of the normal dissipative terms during binary collisions.

Nevertheless, the results above indicate that the regularized model in eq. (4.7) can be used as a viable alternative to model normal forces in DEM simulations. The differences observed with respect to the Kuwabara-Kono model, in eq. (1.23), are not too relevant if one considers the large variety and spread in the experimental data obtained for normal impacts of particles [75]: although the Kuwabara-Kono model

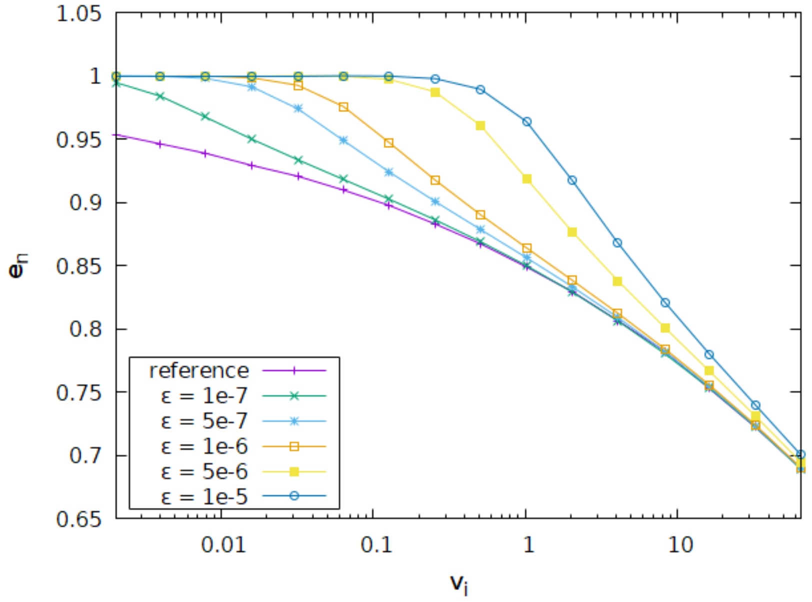


Figure 4.8: Visualization of the dependence of the coefficient of normal constitution on the impact velocity, for different values of ϵ. The values of the parameters can be found in table 4.1. Reference curves obtained from [71].

is the most widely used model for normal forces in DEM, it does not reproduce precisely the trends for all kinds of particles and materials.

In addition, the fact that the coefficient of restitution is slightly closer to 1 in the regularized model should not be of major concern for the majority of applications: since most of the energy dissipated at low speed contacts in granular flows is due to frictional (tangential) processes, and not due to normal collisions [3], the overall nature of the granular flow will not be changed. Normal collisions only dominate energy dissipation in very fast flows. In any case, these conclusions still need further investigation and are the subject of our current studies. Finally, on a surprising note, preliminary results indicate that when eq. (4.6) was also applied to the square root factor in the elastic term $\xi^{3/2}$ in eq. (4.7), the results for the coefficient of restitution were surprisingly closer to the ones obtained by the non-regularized Kuwaba-Kono model. This results, which most likely comes from a non-linear interaction between the regularized dissipation term and the mollified (regular) elastic term, also needs further investigation.

4.3.3 Discussion

In this work, we have identified that, contrary to the expected, the order of the Verlet method, widely used in DEM simulations of granular materials, is not 2 when the model for the normal force used is the Kuwabara-Kono force model, but it is lower lower instead. This is due to the fact that, in this model, there is a square-root factor linked to the dissipative term that has a singular derivative in the beginning and in the end of particle collisions. A detailed theoretical analysis was carried out in order to identify this problem and to show that, in fact, the penalization of the order of the method is linked to this singularity. In a simplified problem, which contains a nonlinear term y^q, for $q \in (0, 1)$, we have identified the same issue and we have proved that the order of the Verlet method or, in this simplified scenario, the lepfrog method, is $1 + q$.

We have proposed a regularized force model, based on an extension of the Kuwabara-Kono model, in which the square-root function appearing in the dissipation term is replaced by a mollified square-root function. This mollified function, which is infinitelly differentiable, allows for an actual order 2 integration of the equations of motion in DEM with the Verlet method. The regularized force model, however, mimics slightly softer materials and slightly less dissipative collisions with respect to a similar simulation carried out for the same parameters using the Kuwabara-Kono model.

Fully characterising the regularized model, and its implications on the motion of large granular assemblies, is one of our current interest of research and preliminary results in this direction are reported on chapter 5. We also want to understand the influence of using the mollified square-root function on the elastic term of the regularized force model that was proposed in this work. Finally, we are also interested in expanding the analyses carried out in this work for higher order methods, such as the higher order symplectic methods presented in [87] or Runge-Kutta methods [55], and also some implicit methods [11].

Chapter 5

On the relation between trajectories of individual particles and the order of numerical integration

5.1 Description of the problem

In granular collapses, one is often interested in the dynamics of the system as a whole. This entails answering questions such as: what are the run out distance and head height? How long does the collapse last? How much of the material remains unmoved?

Little attention is paid to the behavior of individual particles. However, there are some situations where finding out the trajectory of individual particles is the goal or where one can better classify granular flows by studying classes of particle trajectories [18], [19].

Other reason such studies might have been avoided is just the sheer complexity of tracking single grains among millions in 3d granular flows experiments. Even if theoretically possible, the requirement of measurement and data acquisition equipment, and post processing makes it a challenging endeavor.

This issue can be worked around by tracking particles in computer simulations of granular flows. In particular, DEM simulations offer the perfect environment, since, once the simulation is properly set up, tracking individual particles in DEM simulations is trivial. However, a completely different issue presents itself once one starts to study the movement of grains through computer simulations: the stability of the numerical method in face of the inherently chaotic behavior of granular flows. This will be the focus of this article.

In order to exemplify such problem, a DEM simulation of a granular collapse following a somewhat naive approach was set up. The approach is considered naive because it uses a first order method, i.e. the symplectic Euler method. In this simulation, particles were arranged in a roughly rectangular column. Initially, this column of particles is contained in both sides by vertical walls, which are composed of similar such particles that have their positions fixed in time, i.e. they do not move. At $t = 0\,\mathrm{s}$, the right-most wall is removed and the material is allowed to flow. The particles collide following the Hertz normal contact model [34] with Kuwabara-

Kono damping [49] and the tangential contact is handled through the Cundall-Strack friction model [17]. The numerical parameters of the simulation are given in table 5.1.

Table 5.1: Simulation parameters

Name	Symbol	Value
Number of particles	N	1.416×10^3
Aspect Ratio	a	$\approx 5 \times 10^{-1}$
Physical time of simulation	T	$3.0 \times 10^1\,\mathrm{s}$
Average radius	$\bar{r}$	$1\,\mathrm{m}$
Density	ρ	$1.9300 \times 10^4\,\mathrm{kg/m^3}$
Normal elastic constant (Hertz)	$\tilde{k}_n$	$\approx 4.5 \times 10^{11}\,\mathrm{N/m^{1.5}}$
Normal damping constant (Kuwabara-Kono)	γ	$5 \times 10^8\,\mathrm{kg/(m^{0.5}\,s)}$
Tangential elastic constant (Cundall-Strack)	k_s	$^2/_7\tilde{k}_n$
Friction coefficient	μ	6×10^{-1}

In order to test the numerical stability of the trajectories of the particles, the same simulation (i.e. same parameters and same initial conditions) was ran with time steps varying from $\Delta t = 2^{-13} \approx 1 \times 10^{-4}$ to $\Delta t = 2^{-19} \approx 2 \times 10^{-6}$ seconds.

Note that the choice of parameter was made such that the contact duration during a typical collision would be (relatively) long lasting, in order to allow several different time steps value without compromising the the integrity of collisions [71]

For some hand-picked particles, the comparison of their trajectories for each choice of time step is presented in fig. 5.1.

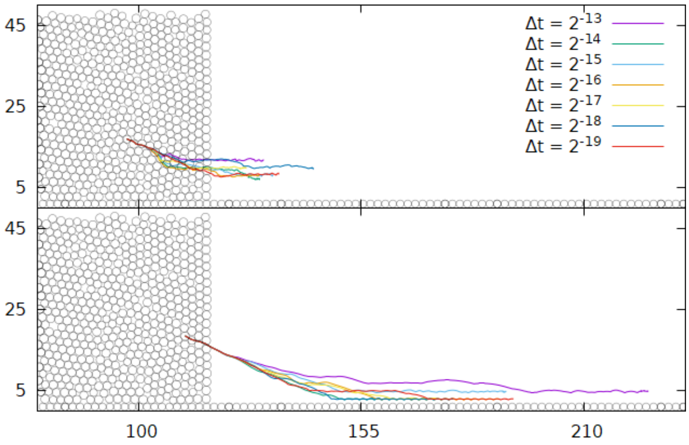


Figure 5.1: Trajectories of two particles from simulations of equal parameters (see table 5.1) and same initial conditions, but varying time steps.

Note that in the early stages of the motion of the particles the trajectories coincide. However, after some time they diverge and then become wildly different.

By using a low-order method, accuracy in the trajectory of individual particles is severely limited, which may cause trajectories to diverge sooner. Of course, as of now this is just speculation. We will dive further into the effects of numerical integration on trajectory convergence in section 5.3.

5.2 Prerequisites

5.2.1 Mollifiers

For the remainder of this chapter, unless specifically stated otherwise, all simulations employ the force scheme derived in chapter 4.

5.2.2 Higher precision storage formats

All computations performed in a computer are subject to one crucial constraint: storage capacity. A certain amount of memory must be attributed to each relevant quantity in a calculation. In modern computer architectures, this memory amount is fixed by the manufacturer (instead of the user) and is usually 32 or 64 bits. The number is then stored in a floating point format and computations are performed via floating point arithmetic. This storage capacity is then reflected in both the precision of the stored number and in rounding errors originating from arithmetic operations.

This is relevant for this work for two reasons. First, because the rounding errors from floating point operations might accelerate the divergence of the trajectories of particles, specially since the smaller the time step size, the more operations one need to perform to reach the same moment in time, which implies in a higher amount of cumulative rounding errors. The second reason is that, once we start studying higher order methods, we shall need higher precision to track our results, since the largest value of time step possible for a well performed simulation is dictated by the material and is usually in the order of $1 \times 10^{-5}\,\mathrm{s}$ to $1 \times 10^{-7}\,\mathrm{s}$.

If one wishes to mitigate the aforementioned issues of precision and rounding errors, one might be tempted to just increase the amount of bits of memory used to store the terms of the calculation. Unfortunately, not only are those amounts fixed by the manufacturer, but the entire architecture of the computer is built around them. Along with that, the architecture of the processor is usually built to natively support floating point arithmetic on data of that specifically size. This means that, on most computers, the only way to perform floating point arithmetic with precision higher then 64 bits is to emulate it on software, which is slow. In performance intensive programs, such as DEM simulations, this alternative becomes unfeasible.

Another alternative is to use double (or quad) words [35], [43]. In this storage format, numbers are represented as an unevaluated sum of two (resp. four) words. In order to take advantage of the architecture of modern computers, the chosen word type is usually a double precision (64 bit) floating point number. In this way, double words (which are called “double doubles”, because they consist of two double precision floating point numbers) store the “bulk” value of a number in the first word and use the second word to store a number several orders of magnitude smaller then in the first word, which takes advantage of the higher precision of floating point formats near zero to make this storage format much more accurate then just a standard double precision floating point number alone.

For this work, a double double basic arithmetic library was implemented in C++ using the algorithms provided in [43]. For more elaborated functions, the techniques in [35] were employed.

5.2.3 A metric for simulations

So far, we have only studied the convergence of the trajectory of a single individual particle. In order to obtain a more complete description of the convergence of particle trajectories in the entire simulation, we need to introduce a new metric which, given two simulations, measures how far apart, on average, are the trajectories of corresponding particles.

More formally, suppose we run two simulations (i.e. numeric approximations of systems of coupled ODEs) and denote each by H_i with $i \in \{1, 2\}$. Each of these simulations has the same parameters and initial condition, with the only difference between them being the size of the time step. Let Δt_i be the time steps size associated with the simulation k and assume that there exists $C \in \mathbb{N}$ such that $\Delta t_1 = C\Delta t_2$. If this condition is not met, then the only step in which the simulations would be representing approximations to the same moment in time would be at $t = 0$. From here on out, C represents the smallest such positive integer.

Let $T \in \mathbb{R}$ be the desired physical time of duration of the simulation and define

$$\mathbb{T} \coloneqq \{n\Delta t_1 \in \mathbb{R} \mid n \in \mathbb{N} \text{ and } n\Delta t_1 \leq T\}. \tag{5.1}$$

Note that $\mathbb{T}$ is precisely the set of all moments in time that are “reachable” by both simulations, i.e. there exists $n, m \in \mathbb{N}$ such that $n\Delta t_1 = m\Delta t_2 \in \mathbb{T}$.

Now, let $N \in \mathbb{N}$ be the number of particles in these simulations and $x_{i,j} : \mathbb{T} \to \mathbb{R}^2$ be the function that ascribes a 2-dimensional position to the particle $j \in \{1, 2, \dots, M\}$ in the simulation i at each moment in time $t \in \mathbb{T}$. Then, we define, for $t \in \mathbb{T}$,

$$\mathrm{d}(H_1, H_2)(t) \coloneqq \frac{1}{N}\sum_{j=1}^{N}|x_{1,j}(t) - x_{2,j}(t)|. \tag{5.2}$$

This metric can also be nondimensionalized via the formula

$$\bar{\mathrm{d}}(H_1, H_2)(\bar{t}) \coloneqq \frac{1}{r}\,\mathrm{d}(H_1, H_2)(\bar{t}T), \tag{5.3}$$

where $r \in \mathbb{R}$ is the average grain radius, $\bar{t} \in [0, 1]$ is the dimensionless time and T is as above. Figure 5.2 illustrates the definition of d.

5.3 Influence of order of numerical integration method

5.3.1 Comparison

We will start with Euler's method, since it is the simplest and, perhaps, the most common integration method. The Hertz-Kuwabra-Kono force scheme was used, since, as shown in fig. 3.1, Euler's method does not suffer order penalization under this force scheme. The parameters used can be viewed in table 5.1. The time step sizes varied between $\Delta t = 2^{-13}\mathrm{s} \approx 1 \times 10^{-4}\,\mathrm{s}$ and $\Delta t = 2^{-19}\mathrm{s} \approx 2 \times 10^{-6}\,\mathrm{s}$. If we denote by H_k the simulation whose time step size is $\Delta t = 2^{-k}\mathrm{s}$, then fig. 5.3(a) shows $\bar{\mathrm{d}}(H_k, H_{19})(\bar{t})$ for the remaining values of k.

The resulting plot (fig. 5.3(a)) has a roughly sigmoid shape, which we can split into three regions for the purposes of analysis. The boundary between these regions is not sharp.

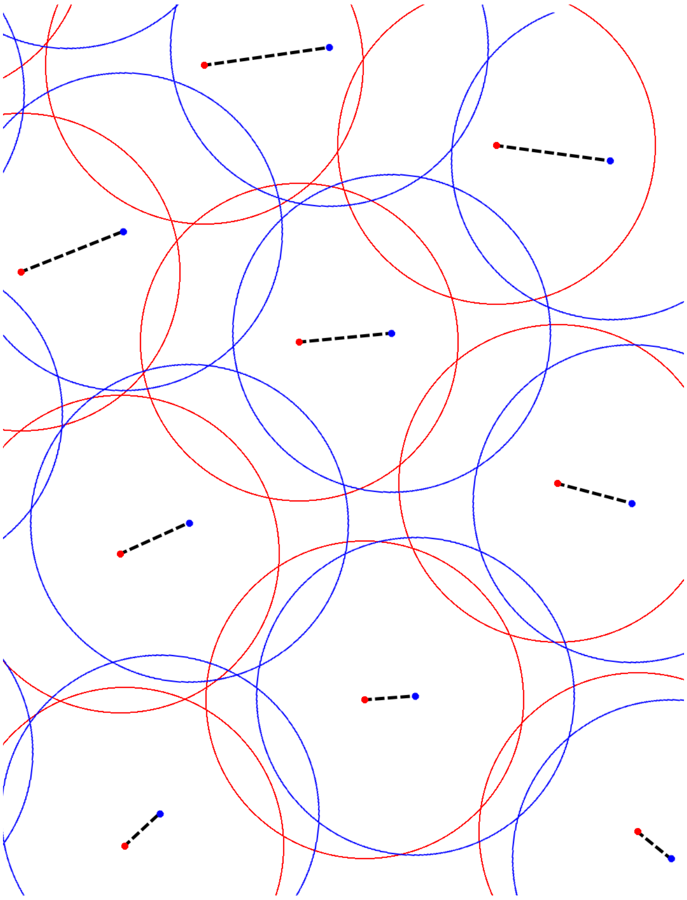


Figure 5.2: Given a moment $t \in \mathbb{T}$, let the red and blue particles represent two distinct simulations with the same parameters and initial condition. Then the d metric for this pair of simulations would be the average length of the dotted black lines.

In the first region, which roughly corresponds with the interval $\bar{t}\,[0, 0.05]$, d grows sub-linearly and its value is negligible (with respect to the average particle radius, by which it is normalized). In this region, the numerical error is still small enough that chaotic behavior has not started to significantly take place yet.

The second region can be characterized by a (perturbed) linear growth of d. It is also in this region that d becomes considerable in relation to r. This region starts approximately at the end of the last region $\bar{t} = 0.05$ and extends almost to $\bar{t} = 0.5$. The errors accumulated in the previous region are sufficient enough to cause particles in this region to start to behave chaotically. It is because of this that the distance to the reference steadily increases in this region.

Finally, the third region encompasses the remaining of the cart, i.e. from $\bar{t} = 0.5$ to $\bar{t} = 1$. In it, the rate of change of d slows down until it completely stops. Note that after $\bar{t} = 0.7$, there is very little, if any, change. This simply means that both the reference simulation and the compared simulation have reached equilibrium.

The second method to be analyzed, also because of its common usage in the literature, is the Predictor-Corrector Verlet method, an adaptation of the Verlet method for non-conservative systems. Since the Predictor-Corrector Verlet method converges as $\mathcal{O}(\Delta t^2)$, it becomes necessary to use the mollified Hertz-Kuwabara-Kono force scheme described in section 5.2.1, with $\epsilon = 5 \times 10^{-6}$. To avoid biases in the comparison, we kept the same parameters from the simulations performed with Euler's method, i.e. those of table 5.1. The time step sizes varied between

$\Delta t = 2^{-15}\text{s} \approx 3 \times 10^{-5}\,\text{s}$ and $\Delta t = 2^{-22}\text{s} \approx 2 \times 10^{-7}\,\text{s}$ and the reference simulation had time step size $\Delta t = 2^{-23}\text{s} \approx 1 \times 10^{-7}\,\text{s}$.

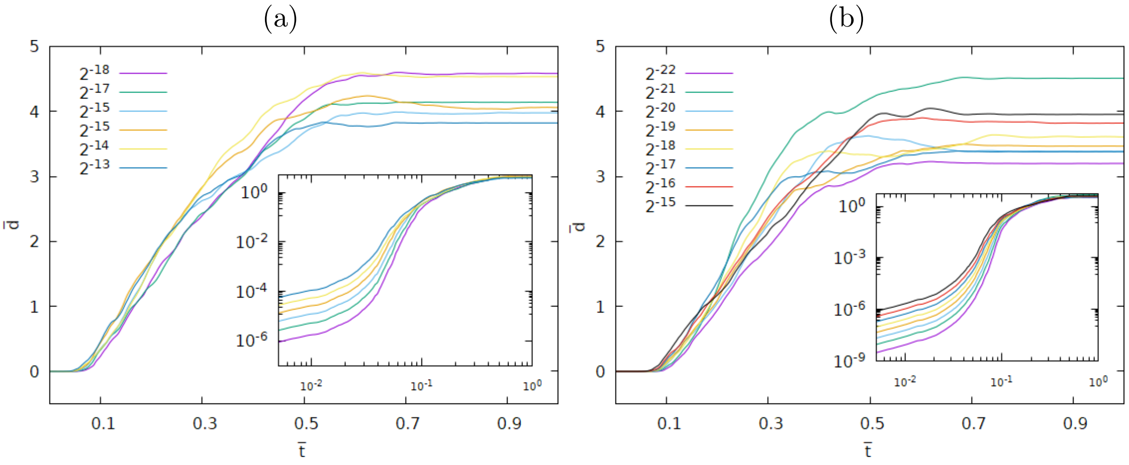


Figure 5.3: All simulations related to these plots used the parameters of table 5.1 and the same initial condition. All graphs here show the value of $\bar{\mathrm{d}}(H, R)(\bar{t})$ over dimensionless time $\bar{t}$, where H is a simulation whose time step size is denoted in the key and R is a reference simulation, which is different between figs. 5.3(a) and 5.3(b). Simulations used in the charts of fig. 5.3(a) where solved via Euler's method, using the Hertz-Kuwabara-Kono force scheme and the time step size used for R is $\Delta t = 2^{-19}\text{s} \approx 2{\times}10^{-6}\,\text{s}$. Simulations used in the charts of fig. 5.3(b) where solved via the Predictor-Corrector Verlet method, using the mollified Hertz-Kuwabara-Kono force scheme in chapter 4 and the time step size used for R is $\Delta t = 2^{-23}\text{s} \approx 1 \times 10^{-7}\,\text{s}$. Both insets show the respective data on a log-log scale.

One can see the same three regions that were previously identified for the simulations performed with Euler's method in fig. 5.3(b). Notably, though, the length of the first region in fig. 5.3(b) is significantly bigger. This fact can be better appreciated by looking at the different scales of the log-log insets in both figures.

5.3.2 Future work

Currently, we are trying to implement a method of order $\mathcal{O}(h^3)$ to see if the trend continues and, if it does, to quantitatively relate the duration of trajectory coincidence with the order of the method.

Unfortunately, many numerical integration methods are not suitable for our applications. Implicit methods are too slow for DEM simulations, the 4th order Runge-Kutta would require the code base to be restructured as would all predictor-corrector methods that the author is aware of. Multistep methods seem to be good candidates, but initializing them with the desired order has proven to be challenging.

Part III

Machine learning and continuum models of granular materials

Foreword

As of the submission of this work, the field of machine learning has very recently took the media's attention by storm worldwide with the release of the chatbot Chat-GPT [60] and some text-to-image converters such as Stable Diffusion [68]. Although the world became shocked to see how far machine learning has come, the field itself is not new and has been changing the world around us for at least a decade —image recognition being a prime example of this.

Because of its prominence, I decided to look at opportunities to involve myself and my field of research with the area. This culminated in a 2-month long visit to the AI Institute in Dynamical Systems, in the University of Washington (UW) —Seattle. However, before this visit began, it was necessary to find a research problem in granular materials in which using machine learning would be suitable.

After learning that I would be able to carry out this visit, my contact with the folks at the UW intensified. During this contact, I learned about the Sparse Identification of Non-linear Dynamics [9] algorithm (SINDy for short), which is a machine learning algorithm that infers differential equations from data. This algorithm is due to Nathan Kutz and Steven Brunton, two professors at UW and director/associate director of the AI institute in Dynamical Systems, respectively. Thus, we decided on finding a problem in granular media in which the SINDy algorithm could be used.

Let us set aside granular materials for a moment to talk about the adjacent field of fluid dynamics. The main object of study of fluid dynamics is the flow; i.e., the movement of fluid. It does so mainly through an Eulerian specification of the flow field; i.e., by modeling it as a continuous vector field of instantaneous velocities at each point of the domain and at each moment in time. This stands in contrast to the Lagrangian specification of granular flows we have used in the first part of this thesis; i.e., modeling the positions, velocities and interactions of individual grains in order to retrieve the macro behavior of the flow.

There are many different reasons to choose one specification over the other. Some examples are: difficulty in modeling, limitation in predictions or computational cost. This last one, computational cost, is an unfortunate struggle of the Lagrangian specification, and the reason for this is quite evident: a pinch of sand contains in the order of 10^4 individual grains, which is the same order of magnitude of the largest simulations done for this work. These simulations took a couple of days. The largest simulations known to this author are in the order of 10^6. This is still way less than the number of grains involved in some large scale natural phenomena, e.g. an avalanche, the propagation of a dune or a landslide.

On the other hand, because of its Eulerian approach, fluid dynamics models can simulate flows of colossal scales, such as ocean currents or, indeed, the entirety of Earth's atmosphere!

Given the information above, the natural question that arises is: is it possible to model granular flows through an Eulerian framework? It appears that the answer to this questions cannot be a simple “yes” or “no”, for there are flows in which the arrangement of the particles is relevant for the macro behavior of the flow (e.g., the flow of sufficiently large particles through a funnel can become jammed depending on the micro structure near the exit) and flows which have already been succesfully modeled through an Eulerian framework; e.g., the flow down an inclined plane. However, a better answer is still a subject of much research.

One of the most promising attempts so far is the $\mu(I)$ rheology [57]. This approach treats granular flows as a continuum medium and uses the classical equations of continuum mechanics —i.e., the continuity and Cauchy momentum equations —fitted with a frictional stress tensor based on the non-dimensional local parameter I. This parameter —called the inertial number —can be understood as a ratio of two time scales: a macroscopic timescale defined by the mean time it takes for a granular layer to slide a distance of one average granular diameter in relation to an adjacent granular layer and a microscopic timescale given by the time it takes a particle under some amount of pressure to fill a grain-size hole. The effect of this definition of I is that if its value at a point in the domain and a moment in time is large (i.e., $I \gg 1$), then the behavior of the flow locally is fluid-like. On the other hand, when the value of I approaches zero, the granular flow locally comes to a halt and becomes more solid-like.

The key aspect of the $\mu(I)$ rheology then is that it takes the inertial number as a parameter to produce a friction coefficient —the eponymous $\mu(I)$ —which acts as a constant of proportionality between the magnitude of the stress tensor and the local pressure. This is the reason why this rheology is deemed frictional: it is highly dependent on the pressure.

As stated before, the $\mu(I)$ rheology has been the most successful approach to modeling granular media as fluids, having achieved successes in some types of flows and showing good progress in others [27], [51], [74], [86]. Thus, we decided that a good research problem for my stay in the UW would be to run Lagrangian simulations of a simple flow and use the SINDy algorithm to confirm or correct the $\mu(I)$ rheology for this flow. This, of course, poses another question: which flow would we choose?

Our first instinct was to study a Couette flow, because of its simplicity. However, in order to perform a computer simulation of this flow, it is paramount to have periodic boundary conditions. Unfortunately, the code base we have developed for the other projects listed in this thesis did not support periodic boundaries and it was deemed that it would take too long to implement it, specially in the context of the time frame of my visit.

Thus, we settled on a similar flow which does not necessitate periodic boundaries: a 2d Taylor-Couette flow. This flow consists of grains confined between concentric circles, which spin in place. This spin forces the grains to shear, resulting in a flow. In other words, instead of periodic boundary condition, we “twisted” the entire domain into a circle so that the two extremes would be connected.

There is still a piece of the puzzle missing, namely how to process Lagrangian data in order to obtain Eulerian equations. The trick is to first transform the discrete data into a continuous field. This can be done via technique called “coarse-graining” [30].

In the literature, one can find examples of coarse-graining being used to do exactly what we need. This technique consists of “spreading out” the value of point-like objects over an area in order to derive a continuum field from it.

In the following, we will explore all of these subjects. In the first section, we will describe Taylor-Couette flows and discuss the difficulties in how to simulate it. In the second section, we will briefly introduce the coarse-graining technique and show its application to a Taylor-Couette flow. In the third section, we will give a brief overview of the SINDy algorithm. Finally, on the fourth section we will show our preliminary results and discuss future work.

Introduction

Machine learning has achieved tremendous success lately, with many of its breakthrough innovations already starting to bleed into the mainstream. For instance, human-like chatbots that can, in many standardized tests, achieve scores within 10% of the best humans are already a reality [60].

However, within academia, many fields are still figuring out how to best employ machine learning within themselves. In the field of fluid dynamics, some early applications of machine learning include using neural networks learn to the solutions of ordinary and partial differential equations [20], [31], [50], and the progress in this front is still ongoing [12], [64]. A more recent approach is to use the technique of dynamic mode decomposition to extract spatial and temporal coherent structures from time series data of fluid flows, resulting in a low-dimensional linear model for the evolution of these dominant coherent structures [22], [48]. More advances of machine learning in fluid dynamics can be found in the review [8].

The field of granular materials is taking a longer time to incorporate machine learning techniques. Recently, there have been efforts to use neural networks for creating predictive models for stress-strain relations of granular materials under compression [13], [77], [84]. Furthermore, some research has gone into speeding up DEM simulations by training a convolutional neural network to replace the direct calculation of particle-particle and particle-boundary collisions [54] as well as using support vector machine and random forest algorithms to predict the outcome of DEM simulations regarding fragment formation [67]. Still, there does seem to be a lot of open opportunities to employ machine learning in the context of granular materials.

The SINDy (acronym for Sparse Identification of Non-linear Dynamics) [9] is a relatively recent development in the field of machine learning. It uses a regression type algorithm to infer differential equations from data and has been already applied to a plethora of varied and complex problems, such as reduced-order models of fluid dynamics [53] and plasma dynamics [46], turbulence closures [6], mesoscale ocean closures [89], nonlinear optics [40], computational chemistry [72], and numerical integration schemes [78].

In this part, we disclose the extent of our research in applying the SINDy method to extract a constitutive law for dense granular flows.

Chapter 6

Prerequisites

6.1 Granular Taylor-Couette flows

In the world of fluid dynamics, the flow consisting of a viscous fluid confined in the gap between two rotating concentric cylinders is named the Taylor–Couette flow. For Newtonian fluids and low angular velocities of the cylinders, an analytic solution to this problem is known and holds up quite well to real data.

A device equipped to perform and measure exclusively slow Taylor-Couette flows is known as a viscometer or rheometer. This is because the known analytic solutions ties the velocity of and torque exerted by the cylinders with the viscosity of the fluid. The first two properties are easy to measure, which allows one to determine the viscosity of Newtonian fluids.

6.1.1 Viscous Taylor-Couette flows

In order to better understand this viscosity measuring characteristic of the Taylor-Couette flow, let us build a 2-dimensional mathematical model for it. Consider two concentric circles of radii $R_1, R_2 \in (0,\infty)$ with $R_1 < R_2$. Furthermore, consider that the inner circle rotates with angular velocity $\Omega \in (0,\infty)$ and the outer circle is static. Finally, suppose that the gap between the circles is filled with a Newtonian fluid of density $\rho \in (0,\infty)$ and dynamic viscosity $\mu \in (0,\infty)$. We are also going to make the following hypotheses: the flow has reached a steady state which is axisymmetric, azimuthal and satisfies the no-slip boundary condition.

The equation that governs this type of flow is the Navier-Stokes incompressible equation, which is

$$\rho\left(\frac{\partial\vec{\boldsymbol{u}}}{\partial t}+\vec{\boldsymbol{u}}\cdot\nabla\vec{\boldsymbol{u}}\right)=-\nabla P+\mu\Delta\vec{\boldsymbol{u}}, \tag{6.1}$$

where $\vec{\boldsymbol{u}} : \mathbb{R}^2\times\mathbb{R}\to\mathbb{R}^2, \vec{\boldsymbol{u}}\left(r,\theta,t\right)=\left(u_r\left(r,\theta,t\right),u_\theta\left(r,\theta,t\right)\right)$ is the 2-dimensional velocity field of the flow in polar coordinates about the center of the circles and $P:\mathbb{R}^2\times\mathbb{R}\to\mathbb{R}$ is the scalar pressure field in the same coordinate system.

Since we are considering the resulting flow to be steady, the time derivative of $\vec{\boldsymbol{u}}$ must be zero. Thus, we are left with

$$\rho\vec{\boldsymbol{u}}\cdot\nabla\vec{\boldsymbol{u}}=-\nabla P+\mu\Delta\vec{\boldsymbol{u}}. \tag{6.2}$$

Writing out the explicit expression for the equation above in polar coordinates gives

$$\begin{cases} r : \rho \left(u_r \dfrac{\partial u_r}{\partial r} + \dfrac{u_\theta}{r} \dfrac{\partial u_r}{\partial \theta} - \dfrac{u_\theta^2}{r} \right) = -\dfrac{\partial P}{\partial r} + \mu \left(\dfrac{1}{r} \dfrac{\partial}{\partial r} \left(r \dfrac{\partial u_r}{\partial r} \right) + \dfrac{1}{r^2} \dfrac{\partial^2 u_r}{\partial \theta^2} - \dfrac{u_r}{r^2} - \dfrac{2}{r^2} \dfrac{\partial u_\theta}{\partial \theta} \right); \\ \theta : \rho \left(u_r \dfrac{\partial u_\theta}{\partial r} + \dfrac{u_\theta}{r} \dfrac{\partial u_\theta}{\partial \theta} + \dfrac{u_r u_\theta}{r} \right) = -\dfrac{1}{r} \dfrac{\partial P}{\partial \theta} + \mu \left(\dfrac{1}{r} \dfrac{\partial}{\partial r} \left(r \dfrac{\partial u_\theta}{\partial r} \right) + \dfrac{1}{r^2} \dfrac{\partial^2 u_\theta}{\partial \theta^2} + \dfrac{2}{r^2} \dfrac{\partial u_r}{\partial \theta} - \dfrac{u_\theta}{r^2} \right). \end{cases} \tag{6.3}$$

Since the flow is axisymmetric, there is no change of velocity with respect to θ, which means that $\dfrac{\partial}{\partial \theta} \equiv 0$. Thus

$$\begin{cases} r : \rho \left(u_r \dfrac{\partial u_r}{\partial r} - \dfrac{u_\theta^2}{r} \right) = -\dfrac{\partial P}{\partial r} + \mu \left(\dfrac{1}{r} \dfrac{\partial}{\partial r} \left(r \dfrac{\partial u_r}{\partial r} \right) - \dfrac{u_r}{r^2} \right); \\ \theta : \rho \left(u_r \dfrac{\partial u_\theta}{\partial r} + \dfrac{u_r u_\theta}{r} \right) = \mu \left(\dfrac{1}{r} \dfrac{\partial}{\partial r} \left(r \dfrac{\partial u_\theta}{\partial r} \right) - \dfrac{u_\theta}{r^2} \right). \end{cases} \tag{6.4}$$

From the hypothesis of no movement in the radial direction, we get that $u_r = 0$, therefore simplifying the equations to

$$\begin{cases} r : -\rho \dfrac{u_\theta^2}{r} = -\dfrac{\partial P}{\partial r}; \\ \theta : 0 = \mu \left(\dfrac{1}{r} \dfrac{\partial}{\partial r} \left(r \dfrac{\partial u_\theta}{\partial r} \right) - \dfrac{u_\theta}{r^2} \right). \end{cases} \tag{6.5}$$

By doing a few more simplifying steps and switching the sides of the equations, we are left with

$$\begin{cases} r : \dfrac{\partial P}{\partial r} = \rho \dfrac{u_\theta^2}{r}; \\ \theta : \dfrac{\partial^2 u_\theta}{\partial r^2} + \dfrac{1}{r} \dfrac{\partial u_\theta}{\partial r} - \dfrac{u_\theta}{r^2} = 0. \end{cases} \tag{6.6}$$

The solution for the equation in the θ coordinate is

$$u_\theta (r, \theta, t) = ar + \frac{b}{r}, \tag{6.7}$$

where a and b are constants that depend upon the initial conditions. The equation in r gives the scalar pressure field, once u_θ is known.

By imposing the no-slip condition, namely that

$$\begin{cases} u_\theta (R_1, \theta, t) = \Omega R_1; \\ u_\theta (R_2, \theta, t) = 0 \end{cases} \tag{6.8}$$

hold, we get the final solution

$$u_\theta (r, \theta, t) = \frac{\Omega R_1^2}{R_2^2 - R_1^2} \left(-r + \frac{R_2^2}{r} \right). \tag{6.9}$$

Now, in order to find an expression for the magnitude of the torque in the cylinder, one must first find the magnitude of the force that drives it. However, the magnitude of this force must be equal to the frictional forces of the fluid that act on the walls of the cylinder, since they are its reaction. Since these forces act on the plane whose normal points radially outward (in the $\vec{r}$ direction) and the force itself points in the direction tangent to the cylinders (the $\overrightarrow{\boldsymbol{\theta}}$ direction), their strength must be given by the $\sigma_{r\theta}$ component of the stress tensor at $r = R_1$. This component, in cylindrical coordinates, is given by

$$\sigma_{r\theta}\left(r,\theta,t\right) = \mu\left(\frac{\partial u_r}{\partial\theta} + \frac{\partial u_\theta}{\partial r} - \frac{u_\theta}{r}\right). \tag{6.10}$$

Applying the flow hypotheses and the solution given in eq. (6.9), one gets

$$\sigma_{r\theta}\left(r,\theta,t\right) = -2\mu\frac{\Omega R_1^2 R_2^2}{R_2^2 - R_1^2}\frac{1}{r^2}. \tag{6.11}$$

Therefore,

$$\sigma_{r\theta}\left(R_1,\theta,t\right) = \frac{-2\mu\Omega R_2^2}{R_2^2 - R_1^2}. \tag{6.12}$$

The pointwise torque along the surface of the cylinder can be found by multiplying the negative of eq. (6.12) by R_1 (the negative is because the force due to the fluid is the reaction to the force we want to find) and the total torque per unit height of the cylinder is found by multiplying the previous result by $2\pi R_1$. Thus, if τ is the total torque, it is given by

$$\tau = \frac{4\pi\mu\Omega R_2^2 R_1^2}{R_2^2 - R_1^2}. \tag{6.13}$$

If the length of the gap between the cylinders, $R_2 - R_1$, is much smaller then their radii, as is usually the case in rheometers, one concludes that

$$\tau = \frac{4\pi\mu\Omega R_2^2 R_1^2}{R_2^2 - R_1} = \frac{4\pi\mu\Omega R_2^2 R_1^2}{\left(R_2 + R_1\right)\left(R_2 - R_1\right)} \approx \frac{4\pi\mu\Omega R^4}{2R\left(R_2 - R_1\right)} = \frac{2\pi\mu\Omega R^3}{\left(R_2 - R_1\right)}, \tag{6.14}$$

which can be rewritten as

$$\mu \approx \frac{\delta\tau}{VS}, \tag{6.15}$$

where $\delta \coloneqq R_2 - R_1$ is the length of the gap between the cylinders, $S \coloneqq 2\pi R$ and $V = \Omega R$ is the velocity of the rotating cylinder. Note that, since $R_2 - R_1$ is considered much smaller than the radii of the cylinders, we also assumed that $R_1 \approx R_2$ and substituted the individual radii by R.

Because of this relevance as a viscosity measuring flow, as well as it being one of the simplest shear-driven flows that would not require the implementation of periodic boundaries in the existing code base of our work, we decided on using the Taylor-Couette flow as a source of data to learn equations on.

6.1.2 The $\mu\left(I\right)$ rheology

Exchanging the fluid in the gap of the cylinders for some granular medium gives us the granular Taylor-Couette flow. This flow has been experimentally studied previously by other authors [58], [82]. Some key findings include a phase-transition behavior depending on the packing fraction of the assembly [37] and the existence of non-local behavior in the flow [45]. These results show that even simple granular flows can have complex behavior.

The aim of our research is to find, via machine learning, a continuum constitutive equation to describe this flow. As will be seen later, knowing what type of terms could appear in this equation is necessary in order to use the SINDy algorithm. Thus, it is useful to have a starting point; i.e., an equation from which to extrapolate. This is where the $\mu\left(I\right)$ rheology comes in [57].

The $\mu\left(I\right)$ rheology is currently accepted in the literature as the closest to correct rheology for granular flows. It was introduced in [57] and is based on the dimensionless parameter I, which is called the inertial number and is defined as

$$I = \frac{d\|\overleftrightarrow{\boldsymbol{D}}\|}{\sqrt{P/\rho_p}}, \tag{6.16}$$

where d is the average diameter of a particle, $\|\overleftrightarrow{\boldsymbol{D}}\|$ is the shear rate defined in terms of the second invariant of the rate of strain stress tensor, P is the pressure and ρ_p is the grain density (not the density of the grains!), which takes into account the packing fraction of the grains.

One interpretation of the parameter I is as a "phase" meter; i.e., its value dictates if the granular medium behaves as a solid, liquid or gas. For values of I near zero, the medium behaves as a solid, with very little motion, mostly limited to individual particles; i.e., no "bulk" motion. For values of I away from zero but less than one, the medium flows densely, as a liquid. Finally, when $I \geq 1$, there is a dilute flow of the particles and a high average distance between them —a granular gas.

Another way to interpret I is as a ratio of time scales [3]: the micro time scale, defined as

$$t_{\text{micro}} = \frac{d}{\sqrt{P/\rho_p}}, \tag{6.17}$$

which is the typical time it takes for a particle-sized hole in a lattice to be filled; and the macro scale, which is

$$t_{\text{macro}} = \frac{1}{\|\overleftrightarrow{\boldsymbol{D}}\|}, \tag{6.18}$$

which represents the mean time it takes for grain in a "shearing layer" to cross another grain in an adjacent "shearing layer" [3]. Thus, I can be understood as

$$I = \frac{t_{\text{micro}}}{t_{\text{macro}}}. \tag{6.19}$$

This is, of course, aligned with the above mentioned interpretation: when $t_{\text{micro}} \ll t_{\text{macro}}$, lattice-filling happens much faster than shearing and thus the medium behaves as a solid, whereas when $t_{\text{macro}} \ll t_{\text{micro}}$, then shearing is faster and fluid-like behavior dominates the flow.

The $\mu\left(I\right)$ rheology then establishes that the stress tensor $\overleftrightarrow{\boldsymbol{\sigma}}$ is given by

$$\overleftrightarrow{\boldsymbol{\sigma}} = -P\overleftrightarrow{\boldsymbol{I}} + \mu\left(I\right)P\frac{\overleftrightarrow{\boldsymbol{D}}}{\|\overleftrightarrow{\boldsymbol{D}}\|}, \tag{6.20}$$

where $\mu\left(I\right)$ is a parameter that depends only on I, $\overleftrightarrow{\boldsymbol{D}}$ is defined as

$$\overleftrightarrow{\boldsymbol{D}} = \frac{1}{2}\left(\nabla\overrightarrow{\boldsymbol{u}} + \nabla\overrightarrow{\boldsymbol{u}}^{\intercal}\right) \tag{6.21}$$

and $\|\overleftrightarrow{\boldsymbol{D}}\|$ is its second invariant. It also states that the packing fraction ϕ is a function only of I; i.e.

$$\phi = \phi\left(I\right). \tag{6.22}$$

Finally, it states that granular flows are similar (albeit not equal) to Bingham fluids: there exists a threshold below which no flow happens. The key differences from a Bingham fluid is that the effective viscosity depends on the local pressure and the yield stress is not a material constant, but also has a dependence on the local pressure.

We will now proceed to finding equations, via the $\mu\left(I\right)$ rheology, for the granular Taylor-Couette flow. The purpose of this derivation is twofold: first, to inspect which kind of terms appear in the equation (the importance of this will become clearer in section 6.3) and second, to illustrate the difference from the Newtonian fluid case and show that the granular case is much more intractable.

Let us start by recalling some tensorial calculus in 2-dimensional polar coordinates. Thus, for the remainder of this subsection, $\overrightarrow{\boldsymbol{u}}$ is a 2-dimensional vector field in polar coordinates with radial component u_r and angular component u_θ, $\overleftrightarrow{\boldsymbol{\sigma}}$ is a second order 2-dimensional tensor in polar coordinates, with components given by

$$\overleftrightarrow{\boldsymbol{\sigma}} = \begin{bmatrix} \sigma_{rr} & \sigma_{r\theta} \\ \sigma_{\theta r} & \sigma_{\theta\theta}, \end{bmatrix} \tag{6.23}$$

and all differential operators are represented in 2-dimensional polar coordinates. The gradient of a vector field is given by

$$\nabla\overrightarrow{\boldsymbol{u}} = \begin{bmatrix} \dfrac{\partial u_r}{\partial r} & \dfrac{1}{r}\left(\dfrac{\partial u_r}{\partial\theta} - u_\theta\right) \\ \dfrac{\partial u_\theta}{\partial r} & \dfrac{1}{r}\left(\dfrac{\partial u_\theta}{\partial\theta} + u_r\right) \end{bmatrix}, \tag{6.24}$$

the divergence of a tensor by

$$\nabla\cdot\overleftrightarrow{\boldsymbol{\sigma}} = \frac{1}{r}\left(r\frac{\partial\sigma_{rr}}{\partial r} + \frac{\partial\sigma_{r\theta}}{\partial\theta} + \sigma_{rr} - \sigma_{\theta\theta}\right)\hat{\boldsymbol{r}} + \frac{1}{r}\left(r\frac{\partial\sigma_{\theta r}}{\partial r} + \frac{\partial\sigma_{\theta\theta}}{\partial\theta} + \sigma_{r\theta} + \sigma_{\theta r}\right)\hat{\boldsymbol{\theta}}. \tag{6.25}$$

and its second invariant is

$$\|\overleftrightarrow{\boldsymbol{\sigma}}\| = \sqrt{\sigma_{r\theta}\sigma_{\theta r} - \sigma_{rr}\sigma_{\theta\theta}}. \tag{6.26}$$

By making the (experimentally backed) hypotheses of axisymmetry ($\dfrac{\partial}{\partial\theta} \equiv 0$) and no net movement in the radial direction ($u_r \equiv 0$), we get

$$\nabla\overrightarrow{\boldsymbol{u}} = \begin{bmatrix} 0 & \dfrac{-u_\theta}{r} \\ \dfrac{\partial u_\theta}{\partial r} & 0 \end{bmatrix} \tag{6.27}$$

and

$$\nabla\cdot\overleftrightarrow{\boldsymbol{\sigma}} = \left(\frac{\partial\sigma_{rr}}{\partial r} + \frac{\sigma_{rr} - \sigma_{\theta\theta}}{r}\right)\hat{\boldsymbol{r}} + \left(\frac{\partial\sigma_{\theta r}}{\partial r} + \frac{\sigma_{r\theta} + \sigma_{\theta r}}{r}\right)\hat{\boldsymbol{\theta}}. \tag{6.28}$$

Now, by eqs. (6.20) and (6.21), we have

$$\overleftrightarrow{\boldsymbol{D}} = \begin{bmatrix} 0 & \frac{1}{2}\left(\frac{\partial u_\theta}{\partial r} - \frac{u_\theta}{r}\right) \\ \frac{1}{2}\left(\frac{\partial u_\theta}{\partial r} - \frac{u_\theta}{r}\right) & 0 \end{bmatrix}, \tag{6.29}$$

$$\|\overleftrightarrow{\boldsymbol{D}}\| = \frac{1}{2}\left|\frac{\partial u_\theta}{\partial r} - \frac{u_\theta}{r}\right| \tag{6.30}$$

and

$$\overleftrightarrow{\boldsymbol{\sigma}} = \begin{bmatrix} -P & \mathfrak{s}\mu\left(I\right)P \\ \mathfrak{s}\mu\left(I\right)P & -P \end{bmatrix}, \tag{6.31}$$

where

$$\mathfrak{s} = \frac{\left(\frac{\partial u_\theta}{\partial r} - \frac{u_\theta}{r}\right)}{\left|\frac{\partial u_\theta}{\partial r} - \frac{u_\theta}{r}\right|} = \operatorname{sign}\left(\frac{\partial u_\theta}{\partial r} - \frac{u_\theta}{r}\right). \tag{6.32}$$

Using again that $\frac{\partial}{\partial \theta} \equiv 0$ and $u_r \equiv 0$, it is true that

$$\frac{\partial \sigma_{\theta r}}{\partial r} = \mathfrak{s}\left(\mu'\left(I\right)I'P + \mu\left(I\right)P'\right), \tag{6.33}$$

where $\cdot'$ denotes differentiation with respect to r and assuming that $\mathfrak{s}$ remains constant throughout the radial direction. Thus,

$$\nabla \cdot \overleftrightarrow{\boldsymbol{\sigma}} = \left(-\frac{\partial P}{\partial r}\right)\hat{\boldsymbol{r}} + \mathfrak{s}\left(\mu'\left(I\right)I'P + \mu\left(I\right)P' + \frac{2\mu\left(I\right)P}{r}\right)\hat{\boldsymbol{\theta}}. \tag{6.34}$$

Recall that the Cauchy Momentum Equation states that

$$\frac{\mathrm{D}\vec{\boldsymbol{u}}}{\mathrm{Dt}} = \frac{1}{\rho}\nabla \cdot \overleftrightarrow{\boldsymbol{\sigma}}, \tag{6.35}$$

where ρ is the density of the fluid. The LHS is just the material derivative of $\vec{\boldsymbol{u}}$. In polar coordinates, it reads

$$\frac{\mathrm{D}\vec{\boldsymbol{u}}}{\mathrm{Dt}} = \left(\frac{\partial u_r}{\partial t} + u_r\frac{\partial u_r}{\partial r} + \frac{u_\theta}{r}\frac{\partial u_r}{\partial \theta} - \frac{u_\theta^2}{r}\right)\hat{\boldsymbol{r}} + \left(\frac{\partial u_\theta}{\partial t} + u_r\frac{\partial u_\theta}{\partial r} + \frac{u_\theta}{r}\frac{\partial u_\theta}{\partial \theta} + \frac{u_r u_\theta}{r}\right)\hat{\boldsymbol{\theta}}. \tag{6.36}$$

Under the previously stated hypotheses, this simplifies to

$$\frac{\mathrm{D}\vec{\boldsymbol{u}}}{\mathrm{Dt}} = -\frac{u^2}{r}\hat{\boldsymbol{r}} + \left(\frac{\partial u}{\partial t}\right)\hat{\boldsymbol{\theta}}. \tag{6.37}$$

If we also assume steady state (again, it is a reasonable hypothesis given the experimental results), we obtain that

$$\frac{\mathrm{D}\vec{\boldsymbol{u}}}{\mathrm{Dt}} = -\frac{u^2}{r}\hat{\boldsymbol{r}}. \tag{6.38}$$

This finally allows us to write the continuum equations for the Taylor-Couette granular flow, as implied by the $\mu\left(I\right)$ rheology. Those would be

$$\begin{cases} r : P' = \rho\dfrac{u^2}{r} \\ \theta : \left(\mu'\left(I\right)I'P + \mu\left(I\right)P' + \dfrac{2\mu\left(I\right)P}{r}\right) = 0. \end{cases} \tag{6.39}$$

One can readily observe the complexity of the equation. In particular, there exists a dependence of $\mu(I)$ and its derivative with respect to r, even though the exact expression μ is never given —it is flow dependent and usually found via experiments or simulations.

6.2 Coarse-graining

Coarse-graining is a popular technique to convert data generated from a Lagrangian framework to an Eulerian framework. In other words, from a DEM simulation, one can use coarse-graining to obtain continuous fields (e.g., flow velocity or density of the medium), over the entirety of the domain.

This technique was first created in the field of chemistry, where there was an effort to connect molecular scale dynamics with macroscopic continuum dynamics. One could argue that these efforts started with the classical studies by Boltzmann and were later taken on by [38], [7] and others [65]. Later (much later), the field of granular matter adopted the coarse-graining technique, starting with [83], as it suited the goal of obtaining a continuum description of granular media.

Although the present paper does not focus on granular gases, the presented results are valid for them as well, but, in models where the collisions are taken to be instantaneous temporal as well as spatial coarse-graining need to be invoked [29] (a minor modification of the presented formulation).

There are two types of coarse-graining (which may be combined), insofar as the dimensions being coarse-grained are concerned: spatial and temporal coarse-graining. In the former, the properties being coarse-grained are averaged throughout space while in the latter, throughout time. For our purpose, only spatial coarse-graining appears to be necessary. This is not the case for all granular phenomena; e.g., when the grains are highly agitated, but only a small part of their movement correlates with the bulk movement of the flow (i.e., their behavior is gas-like), it is necessary to use temporal coarse-graining as well as spatial coarse-graining [29].

6.2.1 Coarse-grained fields

The first step when one wishes to coarse-grain a granular flow is to choose an appropriate coarse-graining function. A coarse-grainig function, for a n-dimensional system, is a function $\phi : \mathbb{R}^n \to \mathbb{R}$ which is positive, symmetric around zero and integrates to unity; i.e., $\displaystyle\int_{\mathbb{R}^n} \phi\left(\vec{\boldsymbol{x}}\right) \mathrm{d}\vec{\boldsymbol{x}} = 1$. It is also desirable, although not strictly necessary, that it is compactly supported. Some examples of coarse-graining functions found in the literature include the normalized Gaussian (which is not compactly supported) and $\phi\left(\vec{\boldsymbol{x}}\right) = H\left(w - \|\vec{\boldsymbol{x}}\|\right)/\Omega_{n,w}$, where H is the Heaviside function, $w \in (0, \infty)$ is some width and $\Omega_{n,w} \in (0, \infty)$ is the n-dimensional volume of a hypersphere of radius w. In this work, we shall only work with compactly supported coarse-graining functions.

For spatial only coarse-graining with compactly supported coarse-graining functions —which, again, is what was needed for this work —the radius of the support of the coarse-graining function defines a coarse-graining scale (which is also called a spatial “resolution” or simply “coarse-graining width”) [30]. If this resolution is chosen too small, the continuum hypothesis is not adequately satisfied and the measured fields become too erratic (see fig. 7.3). A sanity check to avoid too small scales is to plot the coarse-grained density as a function of the coarse-graining width: for small values this graphic is erratic, yet it plateaus when the coarse-graining width becomes large enough. On the other hand, choose a scale that is too large for the problem and one fails to capture local behavior of the flow. Indeed, sub-resolution scale information is not included in the coarse-grained fields. In this sense, coarse-graining is a lossy compression scheme.

Assuming that a coarse-graining scale $w \in (0, \infty)$ has been chosen and the coarse-graining function ϕ_w has been decided on, we shall now define the relevant coarse-grained fields for this work. All the definitions presented below are originally from [30] or in references therein. We shall start by one which has already been mentioned: the coarse-grained volume density. It is defined as

$$\overline{\rho}\left(\overrightarrow{\boldsymbol{x}}, t\right) := \sum_{i=1}^{N} V_i \phi_w \left(\overrightarrow{\boldsymbol{x}} - \overrightarrow{\boldsymbol{x}}_i\left(t\right)\right). \tag{6.40}$$

Note that this definition is akin to a discrete convolution between the desired property (a.k.a. the particle volume V_i) and the coarse-graining function. In this sense, the coarse graining function acts as the kernel of this convolution, which, essentially, “spreads” throughout the domain the value of the property one is coarse-graining, according to the distribution prescribed by the coarse-graining function. Thus, most other elementary fields are defined in much the same vein, that is, the coarse-grained density field is

$$\rho\left(\overrightarrow{\boldsymbol{x}}, t\right) := \sum_{i=1}^{N} m_i \phi_w \left(\overrightarrow{\boldsymbol{x}} - \overrightarrow{\boldsymbol{x}}_i\left(t\right)\right), \tag{6.41}$$

and the coarse-grained momentum density field is

$$\overrightarrow{\boldsymbol{p}}\left(\overrightarrow{\boldsymbol{x}}, t\right) := \sum_{i=1}^{N} m_i \overrightarrow{\boldsymbol{v}}_i\left(t\right) \phi_w \left(\overrightarrow{\boldsymbol{x}} - \overrightarrow{\boldsymbol{x}}_i\left(t\right)\right). \tag{6.42}$$

The coarse-grained velocity field is defined differently, for reasons that will become clear once we discuss the issue of boundaries. Its definition is the momentum density field divided by the density field; i.e.,

$$\overrightarrow{\boldsymbol{V}}\left(\overrightarrow{\boldsymbol{x}}, t\right) := \frac{\overrightarrow{\boldsymbol{p}}\left(\overrightarrow{\boldsymbol{x}}, t\right)}{\rho\left(\overrightarrow{\boldsymbol{x}}, t\right)}. \tag{6.43}$$

For convenience, we shall use the following two shorthands:

$$\begin{aligned}
\overrightarrow{\boldsymbol{x}}_{i,j}\left(t\right) &:= \overrightarrow{\boldsymbol{x}}_i\left(t\right) - \overrightarrow{\boldsymbol{x}}_j\left(t\right), \\
\overrightarrow{\boldsymbol{v}}'_i\left(\overrightarrow{\boldsymbol{x}}, t\right) &:= \overrightarrow{\boldsymbol{v}}_i\left(t\right) - \overrightarrow{\boldsymbol{V}}\left(\overrightarrow{\boldsymbol{x}}, t\right).
\end{aligned} \tag{6.44}$$

In order to define the coarse-grained stress tensor, it is useful to first define the collisional stress and the kinematic stress. The first is the stress coming from the contacts between particles; it is defined as

$$\sigma^c_{\alpha\beta}\left(t\right) := -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \overrightarrow{\boldsymbol{f}}_{i,j,\alpha}\left(t\right) \overrightarrow{\boldsymbol{x}}_{i,j,\beta}\left(t\right) \int_0^1 \phi_w \left(\overrightarrow{\boldsymbol{x}}\left(t\right) - \overrightarrow{\boldsymbol{x}}_i\left(t\right) + s\overrightarrow{\boldsymbol{x}}_{i,j}\left(t\right)\right) \mathrm{d}s. \tag{6.45}$$

In the definition above, the purpose of the integral term is to average the position of the center of the two particles. This can be more readily understood if one rewrites

$$\vec{\boldsymbol{x}}\,(t) - \vec{\boldsymbol{x}}_i\,(t) + s\,\vec{\boldsymbol{x}}_{i,j}\,(t) = \vec{\boldsymbol{x}}\,(t) - \left(s\,\vec{\boldsymbol{x}}_j + (1-s)\,\vec{\boldsymbol{x}}_i\right) \tag{6.46}$$

and notes that $s\,\vec{\boldsymbol{x}}_j + (1-s)\,\vec{\boldsymbol{x}}_i$ is just a parametrization of the line joining $\vec{\boldsymbol{x}}_i$ to $\vec{\boldsymbol{x}}_j$. The later —that is, the kinematic stress —is the stress derived from the movement of the grains as a flow; it is defined as

$$\sigma^k_{\alpha\beta}\,(t) := -\sum_{i=1}^{N} m_i \vec{\boldsymbol{v}}'_{i,\alpha}\,(t)\,\vec{\boldsymbol{v}}'_{i,\beta}\,(t)\,\phi_w\left(\vec{\boldsymbol{x}} - \vec{\boldsymbol{x}}_i\,(t)\right). \tag{6.47}$$

The coarse-grained stress tensor is just the addition of the collisional and kinematic stresses; i.e.,

$$\overleftrightarrow{\boldsymbol{\sigma}} := \overleftrightarrow{\boldsymbol{\sigma}}^c + \overleftrightarrow{\boldsymbol{\sigma}}^k. \tag{6.48}$$

As long as one stays sufficiently away from the boundaries of the domain —a distance of w, to be precise —the main equations of continuum mechanics will hold for the coarse-grained fields, namely the continuity equation,

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \rho \vec{\boldsymbol{u}} = 0 \tag{6.49}$$

and the Cauchy momentum equation

$$\rho \frac{\mathrm{D}\vec{\boldsymbol{u}}}{\mathrm{Dt}} = \nabla \cdot \overleftrightarrow{\boldsymbol{\sigma}} \tag{6.50}$$

This is shown in [30].

6.2.2 Coarse-graining near boundaries

Now we address the challenges in the coarse-graining procedure near boundaries. One of such challenges is that when the distance between the point where the field is calculated and the boundary of the domain is less then the support of the coarse graining function, the value of the field becomes artificially smaller, and the less this distance becomes, the smaller the field will seem to be. This is caused by the lack of particles outside the domain, which causes the effective support of the function to become smaller (i.e., it becomes the intersection of the original support and the domain) and the coarse-graining function to have an integral of less than unity. As explained in [66], one of the ways to solve this issue is to dynamically normalize the coarse-graining function by the n-dimensional volume of its effective support. This can be troublesome for irregular boundaries and non-trivial functions. This process can be better understood by looking at section 6.2.2.

Here we also go back to the definition of the coarse-grained velocity field: since it is defined as a ratio between two other coarse-grained fields, it is already normalized. This, together with the computational speed are the reasons to define coarse-graining in this way, instead as a discrete convolution.

Unfortunately, this technique has a drawback that we have sidestepped, which will be presented in section 7.1. Another unfortunate fact —one that we have yet to bypass —is that the coarse-grained continuity (eq. (6.49) and Cauchy momentum (eq. (6.50)) equations do not hold near walls. This will prove to be a challenge in the future.

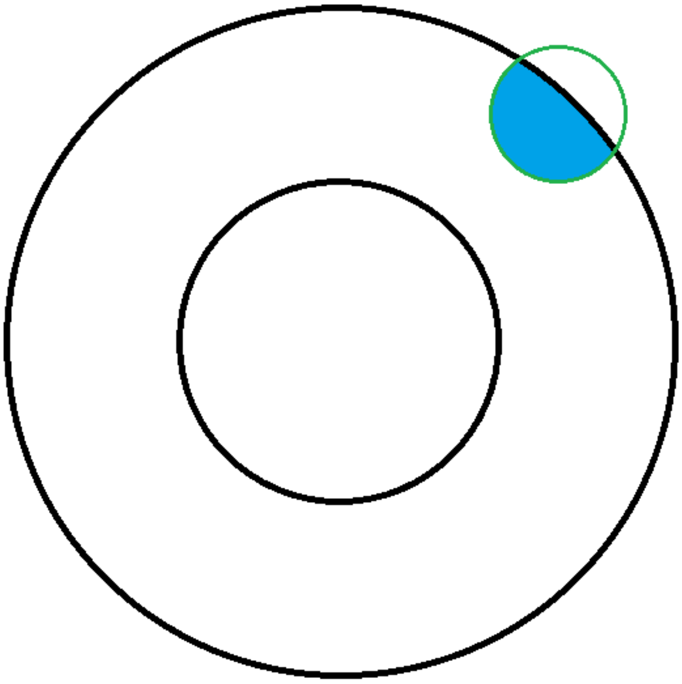


Figure 6.1: If the green circle represents the support of a point in the domain whose coarse-grained fields are desired, the normalization factor would be the integral of the coarse-graining function over the blue region.

6.3 The SINDy algorithm

The recent advances in machine learning have allowed a variety of new tools to be developed. From miraculous chatbots to art emulating software, the progress has been nothing but astounding. Nevertheless, even though these new tools can solve a variety of problems, they often do not offer additional understanding about these problems —they act as black boxes. Indeed, most machine learning models are able to interpolate results from collected data, but they fail to extrapolate the data to regions where no data was collected to begin with. For example, a neural network trained to recognize weather patterns in a region may fail to predict the weather in a completely unrelated region, while a model based on differential equations does not have to concern itself with this issue. Thus, the question that arises is how to extract these differential equations from data in the first place. This is where the SINDy (acronym for Sparse Identification of Non-linear Dynamics) [9] method comes into play.

Consider a n-dimensional dynamical system of the form

$$\frac{d}{dt}\overrightarrow{\boldsymbol{x}}(t)=\overrightarrow{\boldsymbol{f}}\left(\overrightarrow{\boldsymbol{x}}(t)\right), \tag{6.51}$$

where $\overrightarrow{\boldsymbol{x}}:\mathbb{R}\to\mathbb{R}^n$; $\overrightarrow{\boldsymbol{x}}(t)=(x_1(t),\cdots,x_n(t))$ is the system state as a function of time and the function $\overrightarrow{\boldsymbol{f}}:\mathbb{R}^n\to\mathbb{R}^n$ represents the dynamic constraints that define the equations of the system. One key observation is that for many systems of interest, the function $\overrightarrow{\boldsymbol{f}}$ is composed of few terms. We call this property "sparsity" and say that "$\overrightarrow{\boldsymbol{f}}$ is sparse". In order to appreciate this sparsity, just think of how many PDEs you have heard about with more than 10 terms.

Now, let $t_1,\cdots,t_M\in\mathbb{R}$ be the times which constitute the M samples of $\overrightarrow{\boldsymbol{x}}$ and consider the $M \times n$ matrices

$$\begin{aligned} \boldsymbol{X} &= \begin{bmatrix} \overrightarrow{\boldsymbol{x}}^{\mathsf{T}}(t_1) \\ \vdots \\ \overrightarrow{\boldsymbol{x}}^{\mathsf{T}}(t_M) \end{bmatrix} = \begin{bmatrix} x_1(t_1) & \cdots & x_n(t_1) \\ \vdots & \ddots & \vdots \\ x_1(t_M) & \cdots & x_n(t_M) \end{bmatrix}, \\ \boldsymbol{X}' &= \begin{bmatrix} \overrightarrow{\boldsymbol{x}}'^{\mathsf{T}}(t_1) \\ \vdots \\ \overrightarrow{\boldsymbol{x}}'^{\mathsf{T}}(t_M) \end{bmatrix} = \begin{bmatrix} x_1'(t_1) & \cdots & x_n'(t_1) \\ \vdots & \ddots & \vdots \\ x_1'(t_M) & \cdots & x_n'(t_M) \end{bmatrix}, \end{aligned} \tag{6.52}$$

where the values in $\boldsymbol{X}'$ can be either measured or numerically approximated from the samples of $\overrightarrow{\boldsymbol{x}}$.

The next step is to define a library of candidate terms for the expression of $\overrightarrow{\boldsymbol{f}}$. Let $g_1, \cdots, g_N : \mathbb{R}^n \to \mathbb{R}$ be the N candidate functions. Then, the library $\Theta(\boldsymbol{X})$ is the matrix whose j-th column is the vector $\left(g_j\left(\overrightarrow{\boldsymbol{x}}(t_1)\right), \cdots, g_j\left(\overrightarrow{\boldsymbol{x}}(t_M)\right)\right)$. For example, if $n = 2$ and the candidate functions are as in table 6.1, then the library $\Theta(\boldsymbol{X})$ will be given by

$$\Theta(\boldsymbol{X}) = \begin{bmatrix} 1 & x_1(t_1) & x_2(t_1) & x_1(t_1)x_2(t_1) & x_1(t_1)^2 & x_2(t_1)^2 & \sin(x_1(t_1)) & \sin(x_2(t_1)) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_1(t_M) & x_2(t_M) & x_1(t_M)x_2(t_M) & x_1(t_M)^2 & x_2(t_M)^2 & \sin(x_1(t_M)) & \sin(x_2(t_M)) \end{bmatrix}. \tag{6.53}$$

The choice of these candidate terms is crucial and should ideally be made based on the terms that appear in equations of related problems. If the correct terms are not present in the library, the correct equation will not be found.

Table 6.1: Example of candidate functions for the Θ library in dimension 2.

1	$g_1(x_1, x_2) = 1$	
$\overrightarrow{\boldsymbol{x}}$	$g_2(x_1, x_2) = x_1$	$g_3(x_1, x_2) = x_2$
$\overrightarrow{\boldsymbol{x}}^2$	$g_4(x_1, x_2) = x_1x_2$ $g_5(x_1, x_2) = x_1^2$	 $g_6(x_1, x_2) = x_2^2$
$\sin\left(\overrightarrow{\boldsymbol{x}}\right)$	$g_7(x_1, x_2) = \sin(x_1)$	$g_8(x_1, x_2) = \sin(x_2)$

Finally, for $j \in \{1, \cdots, M\}$, let $\overrightarrow{\boldsymbol{\xi}}_j = (\xi_{1,j}, \cdots, \xi_{N,j}) \in \mathbb{R}^N$ be some randomly initialized vectors and define the $N \times M$ matrix

$$\Xi = \begin{bmatrix} \overrightarrow{\boldsymbol{\xi}}_1 & \cdots & \overrightarrow{\boldsymbol{\xi}}_N \end{bmatrix} = \begin{bmatrix} \xi_{1,1} & \cdots & \xi_{1,M} \\ \vdots & \ddots & \vdots \\ \xi_{N,1} & \cdots & \xi_{N,M} \end{bmatrix}. \tag{6.54}$$

This matrix (Ξ) is called the matrix of the coefficients and the goal of the SINDy algorithm is to tune the values $\xi_{i,j}$ such that

$$\boldsymbol{X}' \approx \boldsymbol{\Theta}(\boldsymbol{X})\,\Xi \tag{6.55}$$

is as close as possible to eq. (6.51). Note that, under eq. (6.55), the vector $\overrightarrow{\boldsymbol{\xi}}_j$ is a bundle of all the coefficients of the j-th coordinate of the system in eq. (6.51). Going back to our example, eq. (6.55) would represent the following set of equations:

$$\begin{aligned} x'_j\left(t_i\right) = {} & \xi_{1,j} + \xi_{2,j} x_1\left(t_i\right) + \xi_{3,j} x_2\left(t_i\right) + \xi_{4,j} x_1\left(t_i\right) x_2\left(t_i\right) \\ & + \xi_{5,j} x_1\left(t_i\right)^2 + \xi_{6,j} x_2\left(t_i\right)^2 + \xi_{7,j} \sin\left(x_1\left(t_i\right)\right) + \xi_{8,j} \sin\left(x_2\left(t_i\right)\right). \end{aligned} \tag{6.56}$$

More precisely, the goal of the SINDy algorithm is to minimize some loss function L; e.g., the LASSO (least absolute shrinkage and selection operator) function,

$$L\left(\overrightarrow{\boldsymbol{\xi}}_1, \cdots, \overrightarrow{\boldsymbol{\xi}}_N\right) = \sum_{k=1}^{N} \|\boldsymbol{X}'_k - \boldsymbol{\Theta}\left(\boldsymbol{X}\right) \overrightarrow{\boldsymbol{\xi}}_k\|_2 + \lambda \|\overrightarrow{\boldsymbol{\xi}}_k\|_1, \tag{6.57}$$

where $\boldsymbol{X}'_k$ is the k-th column of the matrix $\boldsymbol{X}'$, $\lambda \in [0, \infty)$ is a hyperparameter, $\|\cdot\|_2$ is the L_2 norm and $\|\cdot\|_1$ is the L_1 norm. The L_2 norm term represents the error while the L_1 norm term promotes sparsity on the $\overrightarrow{\boldsymbol{\xi}}_k$, with the hyperparameter λ controlling how important this "sparsification" is in relation to the minimization of the error.

Formally, then, the vectors $\overrightarrow{\boldsymbol{\xi}}_k$ can be defined as

$$\overrightarrow{\boldsymbol{\xi}}_k = \underset{\overrightarrow{\boldsymbol{\xi}}'}{\arg\min} \|\boldsymbol{X}'_k - \boldsymbol{\Theta}\left(\boldsymbol{X}\right) \overrightarrow{\boldsymbol{\xi}}'\|_2 + \lambda \|\overrightarrow{\boldsymbol{\xi}}'\|_1 \tag{6.58}$$

and the equations found by the SINDy method are

$$\frac{d}{dt} \overrightarrow{\boldsymbol{x}}_k\left(t\right) = \overrightarrow{\boldsymbol{f}}_k\left(\overrightarrow{\boldsymbol{x}}\left(t\right)\right) := \overrightarrow{\boldsymbol{\Theta}}\left(\overrightarrow{\boldsymbol{x}}^{\intercal}\left(t\right)\right) \overrightarrow{\boldsymbol{\xi}}_k, \tag{6.59}$$

where $\overrightarrow{\boldsymbol{\Theta}}\left(\overrightarrow{\boldsymbol{x}}^{\intercal}\left(t\right)\right)$ is the symbolic vector of functions (as opposed to the matrix $\boldsymbol{\Theta}\left(\boldsymbol{X}\right)$, which is a data matrix). Thus,

$$\frac{d}{dt} \overrightarrow{\boldsymbol{x}}\left(t\right) = \overrightarrow{\boldsymbol{f}}\left(\overrightarrow{\boldsymbol{x}}\left(t\right)\right) := \boldsymbol{\Xi}^{\intercal} \left[\overrightarrow{\boldsymbol{\Theta}}\left(\overrightarrow{\boldsymbol{x}}^{\intercal}\left(t\right)\right)\right]^{\intercal}. \tag{6.60}$$

There have been many improvements in the SINDy method since its inception. For instance, the PDE functional identification of non-linear dyanamics [69] (PDEFIND method or algorithm, for short) expands the original functionality of the SINDy method to partial differential equations. The weak SINDy [56] solves the weak formulation of PDEs via the Galerkin method, which provides high robustness to noisy data. There is also SINDyPI [44] (PI stands for parallel, implicit), which tackles implicitly defined differential equations. Since eq. (6.39) is defined implicitly and cannot be explicitized, SINDyPI seems like the best choice for our current problem.

Chapter 7

Preliminary developments

7.1 Preliminary developments

In order to construct the initial setup of the simulation of a granular Taylor-Couette flow, an arbitrary center $c \in \mathbb{R}^2$ was chosen as were two radii $R_i, R_L \in (0, \infty)$, with $R_i < R_L$. The center combined with the radii define two circles, whose perimeters were filled with particles. More precisely, the particles were placed with their centers in the perimeter in such a way that they were tangent to their neighbors. The radii of the particles along the perimeter of the circle defined by R_L is not important and can be chosen arbitrarily. For the particles along the one defined by R_i, the choice of their radii will be described later. These particles do not interact in anyway among themselves and are not affected by any other particles. However, they do affect other particles, imparting on then a force as if they were any normal particle. This is done in order to have well defined bounds for the system.

The next step is to fill the gap between the two circles with particles. In order to avoid crystallization, particles with two different radii were used: a large group of particles with a smaller radius and a smaller number of particles of bigger radius. After the total number of particles was decided, the filling process began: to fill the gap, particles were randomly placed in a square that circumscribes the circle defined by R_L and, wherever a particle would be placed outside the larger circle, within the smaller circle or intercepting any previously placed particle (including the ones alongside the perimeter of both circles), that placement would be skipped and a new random placement tried. In order for this process to finish reasonably quickly (say, a few seconds at most), it is necessary that the value of R_L be large. Otherwise, the process may saturate the gap before the number of desired particles is reached or it may have trouble finding available spaces to place remaining particles.

Now, the aim is to make the granular medium in the gap reach an acceptable packing fraction. By acceptable, it is meant a value large enough that a phase transition of the force chains [37] will not happen when the inner circle starts to rotate, but not so large that the individual grains have trouble slipping by each other. After this value is decided, a simulation on the aforementioned system is ran, imparting in the particles on the outer perimeter a velocity towards the center of the system, such that they will arrive at the center in 5 seconds. The magnitude of their velocity is not important, although it should be slow enough so that the particles in the gap do not develop overlaps bigger than what is reasonably accepted by the DEM model (5% to 10% of their radii). Every few frames, a snapshot of the whole system is saved.

After this simulation has finished, the approximate packing fraction of each snapshot is calculated by dividing the sum of the areas of each particles (note that the interpenetration is not accounted for) by the approximated area of the gap; the latter being estimated as the area between the circles centered on c and of radii $R_i + \bar{r_i}, R_L - \bar{r_L}$. Here, R_L is the radius of the outer circle at the specific snapshot where the area is being measured, $\bar{r_i}$ (respectively, r_L) is the average radius of the particles that constitute the inner (respectively, outer) wall. The addition (respectively, subtraction) is done in order to account that in the final assembly, most particles will not fit between the particles of the walls. After the packing fractions have been calculated, one slightly above the target packing fraction is selected. This is done because in the next steps of the construction of the initial conditions more space will be introduced to allow the particles to adjust a bit. How much above the target was decided on a trial and error basis. The R_L of this simulation shall henceforth be called R_o.

Following the compaction of the granular medium in the gap, the next step is to replace the particles in the outer wall. Before the compaction, the particles of the outer wall need to be tangent to each other, otherwise the particles in the gap may escape to the outside of the outer wall. However, during the compaction process, the radius of the outer wall is reduced, which necessarily causes the particles of the outer wall to bunch together. This result is undesirable, because it would cause multiple particle of the outer wall to contact particles in the gap simultaneously, which itself would result in a force multiple times larger than what is desired. Thus, it is necessary to replace these particles with adequately spaced ones. If these particles have a smaller radii than the ones being replaced, then some area will be gained, which will eventually lower the overall packing fraction.

The new outer wall shall be composed, as before, of particles tangentially adjacent whose centers are fixed on the perimeter of the circle of center c and radius R_o and that do not interact in anyway among themselves and are not affected by any other particles, but that do affect other particles, imparting on then a force as if they were any normal particle. The difference now is that these particles will not have a homogeneous radius. Instead, there will be a row of smaller particles, followed by a bigger one, then another row of smaller particles, and so forth. (Note that the radii of these particles is not the same as of the big nor of the small particles in the media) This inhomogeneous wall is necessary in order to create a “rugosity” of the walls and is crucial in order to promote the correct movement of the granular medium. If this is not done, the particles nearest to the wall will not move much. Apparently, this was necessary even in physical experiments [37]. Also, in the first step of the construction of the initial condition, the inner ring was also constructed in this manner.

If one ponders for a moment, one will observe that it is not trivial to construct such a wall. This is because the number of particles must be an integer and the perimeter of a circle of radius R is given by $2\pi R$, most combinations of parameters will leave a gap in the wall whose size does not fit any other particle. Since we needed to tinker repeatedly with the rugosity of the walls, we came up with an algorithm to devise perfect parameters. We will now explain this algorithm. Throughout the explanation of this algorithm, whenever two large particles are said to be adjacent, this means that they are both in the same wall and that there is a path within this wall such that there is no other large particle lying on it. This path shall be called "adjacency path". First, three parameters are specified:

- A "rugosity size" (r_t). This is the radius of the larger particles in the wall, a.k.a. the "teeth".
- A "suggested rugosity spacing" (ℓ). This is the length of the adjacency path of two adjacent large particles.
- A "suggested small wall disk radius" ($\tilde{r}_w$). This is the radius of the smaller particles.

A few notes on these parameters. First, these parameters are the same for both the inner and outer wall. Second, while the last two parameters have the title of "suggested", the first one lacks it. This choice is arbitrary. As explained before, some parameters need to be adjusted in order to fill the perimeter, and we decided on those.

The algorithm first decides how many large particles there will be in the whole perimeter through the expression

$$N_\mathrm{t} := \left\lfloor \frac{2\pi R}{\ell} \right\rfloor, \tag{7.1}$$

where R is the radius of the inner or outer circle. Then, this number is used to calculate the angular space of adjacent large particles; i.e. the (smallest) angular distance between them. This is simply

$$\theta_t := \frac{2\pi}{N_t}. \tag{7.2}$$

In order to calculate how many small particles fit between two adjacent large particles, the remaining angular space θ_Δ between them is calculated. This is the angular space between their centers minus the angular space taken by their length. This is given by

$$\theta_\Delta := \theta_t - 2 \arcsin\left(\frac{r_t}{2R}\right). \tag{7.3}$$

Then, the estimated angular space taken by a small particle θ_t is calculated from $\tilde{r}_w$ via

$$\theta_t := 2 \arcsin\left(\frac{\tilde{r}_w}{2R}\right), \tag{7.4}$$

and the number of small particles between two adjacent large particles N_w is set to

$$N_w := \left\lfloor \frac{\theta_\Delta}{2\theta_t} \right\rfloor. \tag{7.5}$$

Finally, the actual radius of the small particles r_w is set to

$$r_w = R \tan\left(\frac{\theta_\Delta}{2N_w}\right), \tag{7.6}$$

in order to actually fit N_w small particles between adjacent large particles.

Once the outer wall has been properly replaced, the assembly is simulated for one second. This is done in order to allow the grains inside to dissipate residual kinetic energy from the compaction process and to accommodate within the new boundary. All the parameters used in the construction of the initial conditions are provided in table 7.1. Their values were chosen from parameters already present in the literature [37], [52], [70], [82], in order to have validated examples as reference. The finalized, empty Taylor-Couette cell and a detail of the wall composition can be visualized in fig. 7.1.

Table 7.1: Initial condition parameters

Name	Symbol	Value
Total Disks	N	2911
Big Disks	N_b	400
Small Disks	N_s	2511
Big Disk Radius	r_b	$4.5 \times 10^{-3}\,\mathrm{m}$
Small Disk Radius	r_s	$3.7 \times 10^{-3}\,\mathrm{m}$
Inner Ring Radius	R_i	$1.042 \times 10^{-1}\,\mathrm{m}$
Outer Ring Radius	R_o	$2.492 \times 10^{-1}\,\mathrm{m}$
Target Packing Fraction	φ	0.83203
Rugosity Size	r_t	$1.0 \times 10^{-3}\,\mathrm{m}$
Suggested Rugosity Spacing	ℓ	$5.0 \times 10^{-3}\,\mathrm{m}$
Suggested Small Wall Disk Radius	$\tilde{r}_w$	$1.3 \times 10^{-4}\,\mathrm{m}$

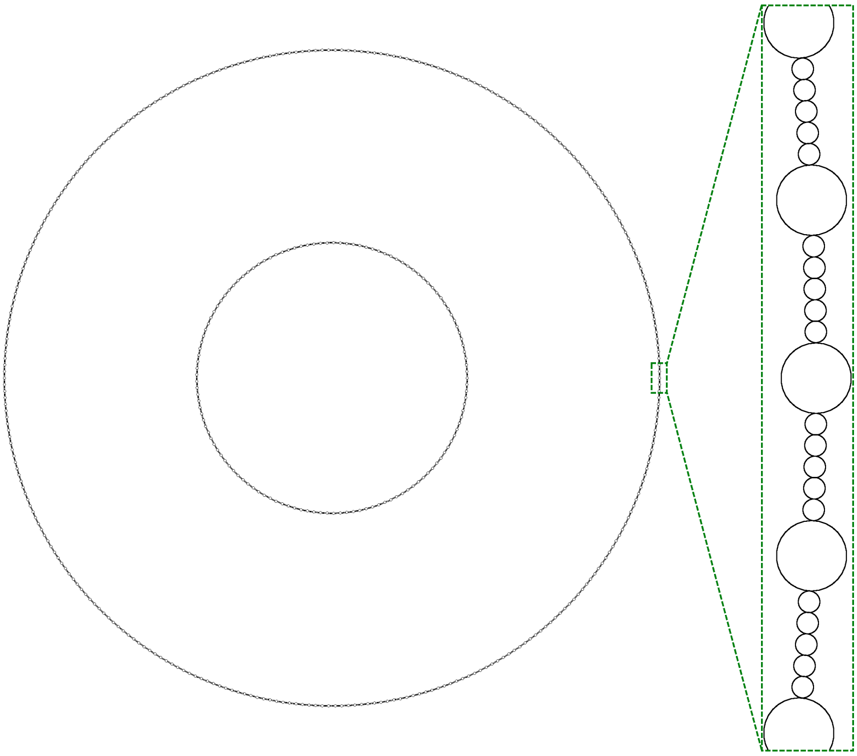


Figure 7.1: Illustration of the empty Taylor-Couette cell with a zoom on a region of the outer wall. Note that the wall is made out of two types of particle and their arrangement is alternated, in order to provide a certain amount of rugosity. The inner wall is also composed in this manner.

The outer wall is now programmed to rotate with some angular speed ω and the simulation is executed. The integration is performed via the symplectic Euler method with some time step size Δt. The normal collision forces follow the Hooke law with linear damping scheme (see section 1.3.1) and the tangential collision follows the Cundall-Strack (see section 1.3.2) friction scheme. A snapshot of the assembly was taken every 1/24th of a second. After some physical time T, the simulation concludes. All the parameters can be seen in better detail in table 7.2 and the simulation process is depicted in fig. 7.2

Table 7.2: Simulation parameters

Name	Symbol	Value
Time Step Size	Δt	$\approx 2.548\,63 \times 10^{-5}\,\mathrm{s}$
Simulation Duration	T	$1.80 \times 10^{2}\,\mathrm{s}$
Disk Height	h	$6 \times 10^{-4}\,\mathrm{m}$
Inner Wall Angular Speed	ω	$5 \times 10^{-2}\,\mathrm{rad/s}$
Density	ρ	$1.060 \times 10^{3}\,\mathrm{kg/m^3}$
Elastic Constant	k_n	$3.52 \times 10^{2}\,\mathrm{N/m}$
Tangential Stiffness Ratio	k_s/k_n	0.15/0.19
Damping Constant	γ	$1.9 \times 10^{-1}\,\mathrm{kg/s}$
Friction Constant	μ	0.44

From the same initial conditions and parameters, another simulation is ran. The only differences are that the duration is shortened to $T = 1 \times 10^{-2}\,\mathrm{s}$ and snapshots are taken every $1 \times 10^{-4}\,\mathrm{s}$. This second simulation is useful for validating properties which involve temporal derivatives.

In order to perform the coarse-graining procedure described in section 6.2, it is necessary first to choose a coarse-graining function. Through trial and error, we settled on the ϵ-mollifier defined previously in eq. (4.3). Note that the normalization constant needs to be changed because of the 2-dimensional nature of this problem, in contrast to the 1-dimensional nature of the previous problem.

To calculate the normalization factor, we use angular coordinates centered at c and the Fubini theorem to exchange an area integral by two linear integrals. Finally, we deploy Boole's rule [1] twice.

Choosing the value of ϵ is equivalent to choosing the coarse-graining scale w, because ϵ dictates the size of the support of the ϵ-mollifier. This decision is made in accordance to [52]. First, a snapshot of the system is chosen; this snapshot can be chosen arbitrarily. Then, for different values of w, the value of the resulting coarse-grained volume density is averaged among equally spaced points along a ring in the middle of the gap between the two walls and plotted against the respective w. The plot should plateau at some point and the smallest value from this plateau is chosen as the de facto w. This plot is presented in fig. 7.3.

The details of the coarse-graining procedure given until here are the same for both the long-running ($T = 1.80 \times 10^{2}\,\mathrm{s}$) and short-running ($T = 1 \times 10^{-2}\,\mathrm{s}$) simulations. However, from now on, their treatment shall be different. This is because we wish to validate different features from them, and these features have different dependences on the time-scale.

We start with the long-running simulation. From its results, we wish to validate the velocity and volume density profiles of our granular assembly. For this, N_θ

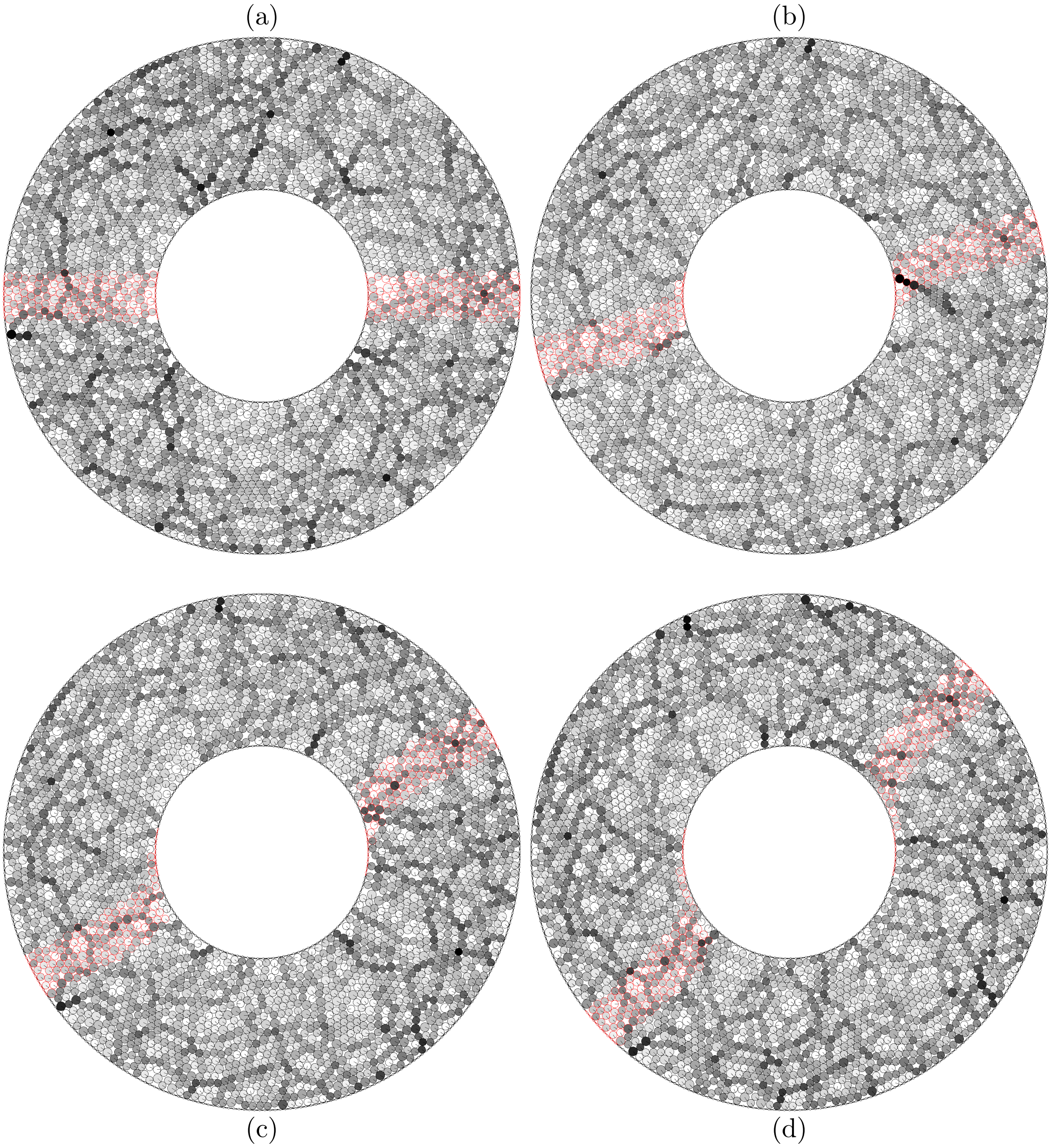


Figure 7.2: Different snapshots of a simulation of a granular Taylor-Couette cell. At $t = 0\,\mathrm{s}$, a rectangular region passing through the cente of the system is chosen and the particles lying therein are marked in red so that one can observe the shearing near the inner wall as the simulation evolvs. The shades of gray that fill the inside of a particle shows how much stress that particle is under (the darker, the higher the stress). The physical time of the snapshots are $t = 0\,\mathrm{s}$ (a), $t = 5\,\mathrm{s}$ (b), $t = 1.0 \times 10^1\,\mathrm{s}$ (c) and $t = 1.5 \times 10^1\,\mathrm{s}$ (d). The parameters used are displayed in tables 7.1 and 7.2.

equally spaced points over the angular interval $[0, 2\pi)$ were selected as well as N_r points in the radial range $[R_i - w, R_o + w]$. The radial range is extended by w in both directions because otherwise the coarse-graining function would not be able to capture the velocity/density of the region immediately adjacent to both of the walls. Afterwards, this domain shall be linearly compressed into $[R_i, R_o]$.

The coarse-graining procedure described in section 6.2 was then performed in each of the last 600 hundred snapshots of the long-running simulation; i.e., the last $2.5 \times 10^1\,\mathrm{s}$ of simulation. Afterwards, the values obtained for each of the $N_r \times N_\theta$

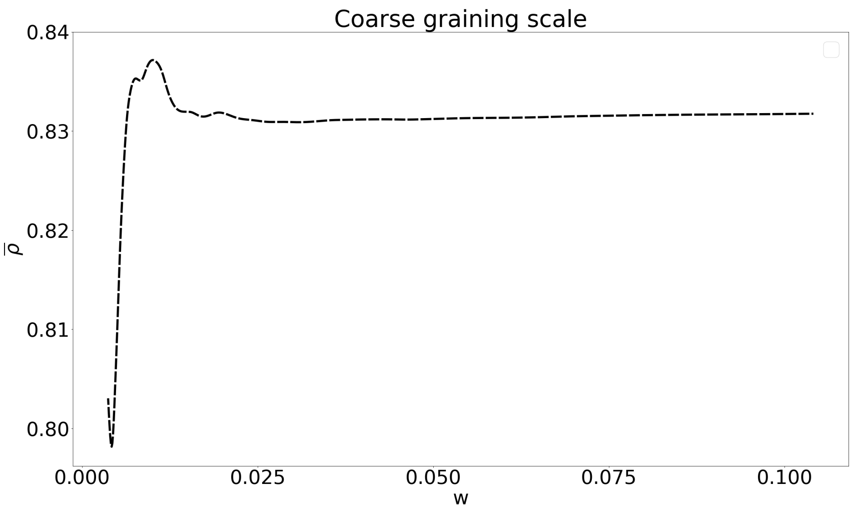
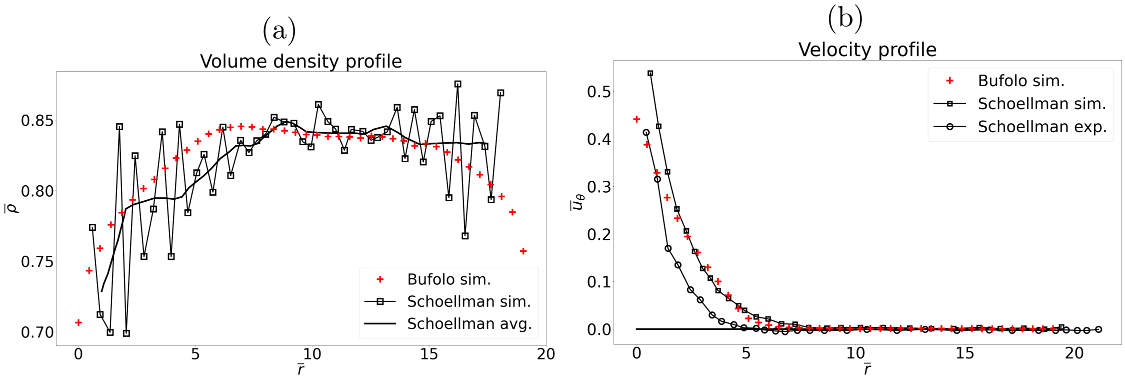


Figure 7.3: Plot of the coarse-grained volume density field $\bar{\rho}$ as a function of the radius w of the support of the coarse-graining function. Note that as w increases, $\bar{\rho}$ plateaus. The coarse-grained volume density field was averaged among equally spaced points along a ring in the middle of the gap between the two walls.

points was averaged over the the snapshots. Finally, for each of the radial points, the N_θ points associated with that radial point where averaged, leaving us with N_r points averaged both in time and angle. The radial spaced was then linearly compressed back into the $[R_i, R_o]$ range and plotted. This plot and the reference data taken from [70] can be seen in fig. 7.4. The parameters used for this coarse-graining are displayed in table 7.3

Table 7.3: Parameters for the coarse-graining procedure of the long-running simulation.

Name	Symbol	Value
Coarse-graining Scale	w	3.5×10^{-2} m
Number of Radial Points	N_r	48
Number of Angular Points	N_θ	16


(a)

(b)

Figure 7.4: Volume density (a) and velocity (b) profiles of the long-running simulation. Here, $\bar{\rho}$ is the volume density and $\bar{u_\theta}$ is the tangential component of the velocity, averaged over time and angular space. The variable $\bar{r}$ the radial distance from the inner wall, normalized by the average grain radius. (Recall that this final range is a mapping from a larger range)

We proceed to the short-running simulation. The goal now is to validate the continuity equation, namely

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \rho \overrightarrow{\boldsymbol{u}} = 0, \tag{7.7}$$

and the Cauchy momentum equation; i.e., eq. (6.35). For reference, we rewrite the Cauchy momentum equation here in the form LHS = 0:

$$\frac{\mathrm{D}}{\mathrm{Dt}} \left(\rho \overrightarrow{\boldsymbol{u}}\right) - \nabla \cdot \overleftrightarrow{\boldsymbol{\sigma}} = 0. \tag{7.8}$$

In order to compute the derivatives involved, we have used a sixth order finite difference scheme. This order was chosen by trial and error. In order to correctly estimate the time derivatives present in both the equations, the time difference between two consecutive snapshots must be sufficiently small. Through trial and error, we have found the temporal spacing of $1 \times 10^{-4}\,\mathrm{s}$ to be small enough. For the same reason, the spatial derivatives require the spacing between nearby points to be smaller than of those of the long-running simulation. Again through trial and error, we found that a spatial scale of $5 \times 10^{-4}\,\mathrm{m}$ was adequate.

Because of these small spatial and temporal scales, it became unfeasible to perform this analysis through the entire domain or in the long-running simulation. Therefore, we limited ourselves to a thin, radially aligned, strip that ran from the inner wall to the outer wall. This strip was $0.0005 \times N_x$ (or about 1.4 small particles) wide and $0.0005 \times N_y$ (or the size of the entire assembly) long (i.e., in the radial direction). Because of how thin it is, we treated it using Cartesian coordinates. The values for N_x and N_y can be found in table 7.4.

Table 7.4: Parameters for the coarse-graining procedure of the short-running simulation.

Name	Symbol	Value
Coarse-graining Scale	w	$3.5 \times 10^{-2}\,\mathrm{m}$
Number of Points in the x-axis	N_x	21
Number of Points in the y-axis	N_y	500

Finally, we ran the coarse-graining procedure, calculated the necessary derivatives via finite differences, calculated the value of the LHS of eq. (7.7) (and similarly for eq. (6.35)) and plotted their average over time and over the x coordinate. The results are shown in figs. 7.5 and 7.6.

7.2 Discussion

It is visible in fig. 7.4 that the data from our simulations closely matches the data in [70]. In fig. 7.4(a), there are a few mismatches, specially towards the outer rim of the Taylor-Couette cell. However, in the data from [70] there are also some large fluctuations, which could help explain this discrepancy. On the other hand, the data in fig. 7.4(b) appears to match our results much closer, with the only caveat being that our data does not match exactly with neither the experimental nor the simulation data from [70], but it is solidly between both, becoming closer to the experimental data for regions away from the inner rim.

For figs. 7.5 and 7.6, our goal is to get the plots as close to zero in as much as the plot as possible, as that means that the respective equations are holding

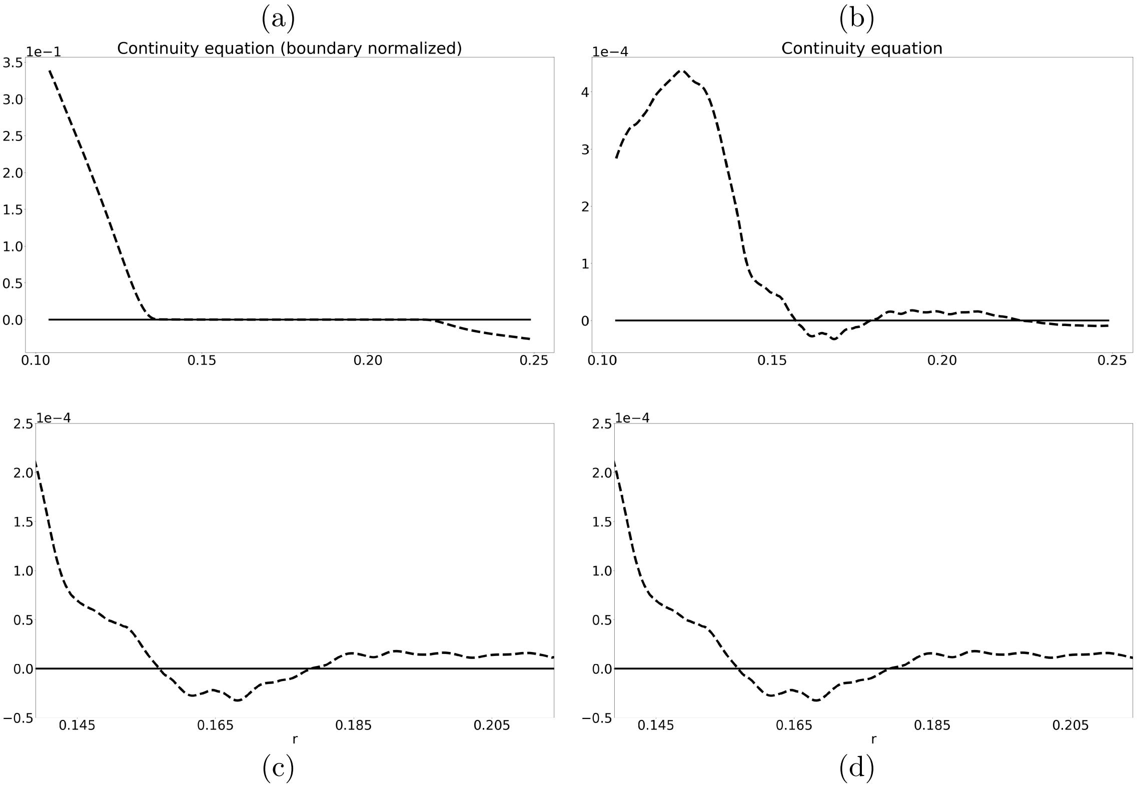


Figure 7.5: Plots of the value of the LHS of the continuity equation (eq. (7.7)). In (a) and (c), the normalization near the boundary was perfomed, while in (b) and (d) it was not. Figures (c) and (d) exclude the regions within w of either boundary. Take note of the scales in the top left corner of each plot.

well throughout the domain. In fig. 7.5, we observe that this is mostly the case, even though one can note that in the case where boundary normalization is being performed, the validity of the equation near the boundary (see fig. 7.5(a)) is slightly jeopardized. This is expected though, because the boundary normalization affects the terms of eq. (7.7) in an uneven manner. In all other cases, the continuity equation is valid up to 1×10^{-4}.

As for the Cauchy momentum equation (eq. (7.8)), fig. 7.6 shows that its validity is much more tenuous. It absolutely does not hold near the boundary region (see figs. 7.6(a) and 7.6(b)). Again, this is expected; Since all the particle composing the walls do not feel the forces of the particles in the gap and the movement (or lack thereof) of the particles in the walls is not caused by a force, but by a prescribed evolution, the stresses near them cannot be correctly calculated. On the other hand, in the region outside of the influence of the boundary, the Cauchy momentum equation holds up to 1×10^{-1}.

7.3 Future work

There is still much to be done in this project. First and foremost, a more in-depth analysis needs to be performed on the validity of the Cauchy momentum equation in the regions away from the boundary: even though 1×10^{-1} is significant enough to observe a trend, it is not good enough to affirm with certainty that the coarse-grained process is without fault, as the coarse-graining procedure must result

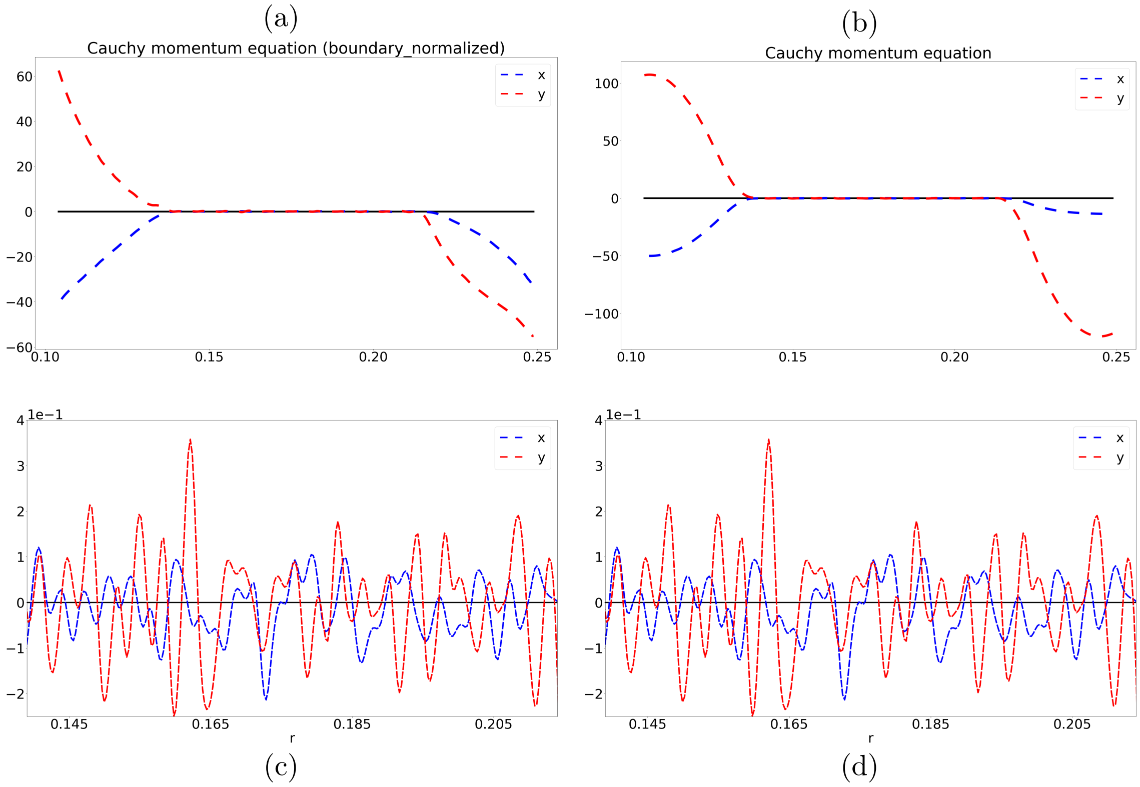


Figure 7.6: Plots of the value of the x (in blue) and y (in red) coordinates of the LHS of the Cauchy momentum equation (eq. (7.8)). In (a) and (c), the normalization near the boundary was perfomed, while in (b) and (d) it was not. Figures (c) and (d) exclude the regions within w of either boundary and have a scale factor in the top left corner of each plot.

in fields which satisfy the Cauchy momentum equation [30]. However, it may be the case that what we observe in fig. 7.6(c) and fig. 7.6(d) are just noise caused by the discretization process.

Following this, it is still necessary to use the SINDy procedure to discover the governing differential equations of this process. As discussed before, suitable guesses of the type of terms that are expected in the equation have to be made. It has been already stated that we plan to use the $\mu(I)$ rheology for this purpose, so an intermediate step is to verify its validity in our data set. One possible test is to calculate the value of I across the domain and then isolate $\mu(I)$ in eq. (6.39) and calculate it from the data. Finally, checking if $I \rightarrow \mu(I)$ is in fact a function may provide a sanity check on the validity of that rheology in our data.

It may also be the case that our data has some sort of non-local behavior. If this situation arises, there is an extensive review [45] of non-local models for granular materials.

Bibliography

[1] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. United States Department of Commerce, National Bureau of Standards (NBS), 1964, ISBN: 0-486-61272-4.

[2] M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids*. USA: Clarendon Press, 1989, ISBN: 0198556454.

[3] B. Andreotti, Y. Forterre, and O. Pouliquen, *Granular Media: Between Fluid and Solid*. Cambridge University Press, 2013. DOI: `10.1017/CBO9781139541008`.

[4] K. Atkinson, *An Introduction to Numerical Analysis*. Wiley, 1978, ISBN: 9780471029854. [Online]. Available: `https://books.google.com.br/books?id=ByPpQ1nt3esC`.

[5] N. Balmforth and R. Kerswell, "Granular collapse in two dimensions", *Journal of Fluid Mechanics*, vol. 538, pp. 399–428, Sep. 2005.

[6] S. Beetham, R. Fox, and J. Capecelatro, "Sparse identification of multiphase turbulence closures for coupled fluid–particle flows", *Journal of Fluid Mechanics*, vol. 914, May 2021. DOI: `10.1017/jfm.2021.53`.

[7] M. Born and K. Huang, *Dynamical Theory of Crystal Lattices*. Oxford University Press, 1954, ISBN: 0-19-850369-5.

[8] S. L. Brunton, B. R. Noack, and P. Koumoutsakos, "Machine learning for fluid mechanics", *Annual Review of Fluid Mechanics*, vol. 52, no. 1, pp. 477–508, 2020. DOI: `10.1146/annurev-fluid-010719-060214`. eprint: `https://doi.org/10.1146/annurev-fluid-010719-060214`. [Online]. Available: `https://doi.org/10.1146/annurev-fluid-010719-060214`.

[9] S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems", *Proceedings of the National Academy of Sciences*, vol. 113, no. 15, pp. 3932–3937, Mar. 2016. DOI: `10.1073/pnas.1517384113`. [Online]. Available: `https://doi.org/10.1073%2Fpnas.1517384113`.

[10] S. J. Burns and K. J. Hanley, "Establishing stable time-steps for dem simulations of non-collinear planar collisions with linear contact laws", *International Journal for Numerical Methods in Engineering*, vol. 110, no. 2, pp. 186–200, 2017. DOI: `https://doi.org/10.1002/nme.5361`. eprint: `https://onlinelibrary.wiley.com/doi/pdf/10.1002/nme.5361`. [Online]. Available: `https://onlinelibrary.wiley.com/doi/abs/10.1002/nme.5361`.

[11] E. M. Campello, "A computational model for the simulation of dry granular materials", *International Journal of Non-Linear Mechanics*, vol. 106, pp. 89–107, 2018, ISSN: 0020-7462. DOI: https://doi.org/10.1016/j.ijnonlinmec.2018.08.010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020746218301409.

[12] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud, *Neural ordinary differential equations*, 2019. arXiv: 1806.07366 [cs.LG].

[13] Z. Cheng and J. Wang, "Estimation of contact forces of granular materials under uniaxial compression based on a machine learning model", *Granular Matter*, vol. 24, 2021, ISSN: 1674-7755. DOI: 10.1007/s10035-021-01160-z. [Online]. Available: https://doi.org/10.1007/s10035-021-01160-z.

[14] E. Chlebus, "An approximate formula for a partial sum of the divergent p-series", *Applied Mathematics Letters*, vol. 22, no. 5, pp. 732–737, 2009, ISSN: 0893-9659. DOI: https://doi.org/10.1016/j.aml.2008.07.007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893965908002747.

[15] J. Cid, "On uniqueness criteria for systems of ordinary differential equations", *Journal of Mathematical Analysis and Applications*, vol. 281, no. 1, pp. 264–275, 2003, ISSN: 0022-247X. DOI: https://doi.org/10.1016/S0022-247X(03)00096-9. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022247X03000969.

[16] A. Coddington, *An Introduction to Ordinary Differential Equations* (Dover Books on Mathematics). Dover Publications, 1961, ISBN: 9780486659428. [Online]. Available: https://books.google.com.br/books?id=wGxmfLq4b%5C_4C.

[17] P. Cundall and O. Strack, "A discrete numerical model for granular assemblies", English (US), *Geotechnique*, vol. 29, no. 1, pp. 47–65, Mar. 1979, ISSN: 0016-8505. DOI: 10.1680/geot.1979.29.1.47.

[18] F. Cunha, R. Gontijo, and Y. Sobral, "Symmetry breaking of particle trajectories due to magnetic interactions in a dilute suspension", *Journal of Magnetism and Magnetic Materials*, vol. 326, pp. 240–250, 2013, ISSN: 0304-8853. DOI: https://doi.org/10.1016/j.jmmm.2012.08.032. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0304885312007202.

[19] F. R. Da Cunha and E. J. Hinch, "Shear-induced dispersion in a dilute suspension of rough spheres", *Journal of Fluid Mechanics*, vol. 309, pp. 211–223, 1996. DOI: 10.1017/S0022112096001619.

[20] M. W. M. G. Dissanayake and N. Phan-Thien, "Neural-network-based approximations for solving partial differential equations", *Communications in Numerical Methods in Engineering*, vol. 10, no. 3, pp. 195–201, 1994. DOI: https://doi.org/10.1002/cnm.1640100303. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cnm.1640100303. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/cnm.1640100303.

[21] J. Domínguez, "Dualsphysics: Towards high performance computing using sph technique", Ph.D. dissertation, Nov. 2014.

[22] "Dynamic mode decomposition of numerical and experimental data", *Journal of Fluid Mechanics*, vol. 656, pp. 5–28, 2010. DOI: `10.1017/S0022112010001217`.

[23] L. C. Evans, *Partial differential equations*. Providence, R.I.: American Mathematical Society, 2010.

[24] R. Finkel and J. Bentley, "Quad trees a data structure for retrieval on composite keys", *Acta Informatica*, vol. 4, pp. 1–9, 2004.

[25] S. Foerster, M. Louge, H. Chang, and K. Allia, "Measurements of the collision properties of small spheres", *Physics of Fluids - PHYS FLUIDS*, vol. 6, pp. 1108–1115, Mar. 1994. DOI: `10.1063/1.868282`.

[26] A. Fog. "Optimizing software in c++, An optimization guide for windows, linux, and mac plataforms". (Jan. 31, 2021), [Online]. Available: `https://www.agner.org/optimize/optimizing_cpp.pdf`.

[27] A. Franci and M. Cremonesi, "3d regularized $\mu(I)$-rheology for granular flows simulation", *Journal of Computational Physics*, vol. 378, pp. 257–277, 2019, ISSN: 0021-9991. DOI: `https://doi.org/10.1016/j.jcp.2018.11.011`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S0021999118307290`.

[28] N. K. Fukagawa and L. H. Ziska, "Rice: Importance for global nutrition.", *Journal of nutritional science and vitaminology*, vol. 65, S2–S3, supplement 2019. DOI: `https://doi.org/10.3177/jnsv.65.S2`.

[29] B. J. Glasser and I. Goldhirsch, "Scale dependence, correlations, and fluctuations of stresses in rapid granular flows", *Physics of Fluids*, vol. 13, no. 2, pp. 407–420, 2001. DOI: `10.1063/1.1338543`. eprint: `https://doi.org/10.1063/1.1338543`. [Online]. Available: `https://doi.org/10.1063/1.1338543`.

[30] I. Goldhirsch, "Stress, stress asymmetry and couple stress: From discrete particles to continuous fields", *Granular Matter*, vol. 12, pp. 239–252, May 2010. DOI: `10.1007/s10035-010-0181-z`.

[31] R. González-García, R. Rico-Martínez, and I. Kevrekidis, "Identification of distributed parameter systems: A neural net based approach", *Computers & Chemical Engineering*, vol. 22, S965–S968, 1998, European Symposium on Computer Aided Process Engineering-8, ISSN: 0098-1354. DOI: `https://doi.org/10.1016/S0098-1354(98)00191-4`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S0098135498001914`.

[32] Y. Guo and J. S. Curtis, "Discrete element method simulations for complex granular flows", *Annual Review of Fluid Mechanics*, vol. 47, no. 1, pp. 21–46, 2015. DOI: `10.1146/annurev-fluid-010814-014644`. eprint: `https://doi.org/10.1146/annurev-fluid-010814-014644`. [Online]. Available: `https://doi.org/10.1146/annurev-fluid-010814-014644`.

[33] E. Hairer, C. Lubich, and G. Wanner, "Geometric numerical integration illustrated by the störmer–verlet method", *Acta Numerica*, vol. 12, pp. 399–450, 2003. DOI: `10.1017/S0962492902000144`.

[34] H. Hertz, "Ueber die berührung fester elastischer körper.", vol. 1882, no. 92, pp. 156–171, 1882. DOI: `doi:10.1515/crll.1882.92.156`. [Online]. Available: `https://doi.org/10.1515/crll.1882.92.156`.

[35] Y. Hida, S. Li, and D. Bailey, “Quad-double arithmetic: Algorithms, implementation, and application”, Jul. 2001.

[36] H. Hiesinger, C. H. van der Bogert, J. H. Pasckert, *et al.*, “How old are young lunar craters?”, *Journal of Geophysical Research: Planets*, vol. 117, no. E12, 2012. DOI: `https://doi.org/10.1029/2011JE003935`. eprint: `https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2011JE003935`. [Online]. Available: `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2011JE003935`.

[37] D. Howell, R. P. Behringer, and C. Veje, “Stress fluctuations in a 2d granular couette experiment: A continuous transition”, *Phys. Rev. Lett.*, vol. 82, pp. 5241–5244, 26 Jun. 1999. DOI: `10.1103/PhysRevLett.82.5241`. [Online]. Available: `https://link.aps.org/doi/10.1103/PhysRevLett.82.5241`.

[38] J. H. Irving and J. G. Kirkwood, “The statistical mechanical theory of transport processes. iv. the equations of hydrodynamics”, *The Journal of Chemical Physics*, vol. 18, no. 6, pp. 817–829, 1950. DOI: `10.1063/1.1747782`. eprint: `https://doi.org/10.1063/1.1747782`. [Online]. Available: `https://doi.org/10.1063/1.1747782`.

[39] A. Iserles, *A First Course in the Numerical Analysis of Differential Equations* (Cambridge Texts in Applied Mathematics), 2nd ed. Cambridge University Press, 2008. DOI: `10.1017/CBO9780511995569`.

[40] G. James, K. Vorotnikov, and B. Brogliato, “Kuwabara-Kono numerical dissipation: a new method to simulate granular matter”, *IMA Journal of Applied Mathematics*, vol. 85, no. 1, pp. 27–66, Feb. 2020, ISSN: 0272-4960. DOI: `10.1093/imamat/hxz034`. eprint: `https://academic.oup.com/imamat/article-pdf/85/1/27/32894870/hxz034.pdf`. [Online]. Available: `https://doi.org/10.1093/imamat/hxz034`.

[41] L. Jing, C. Y. Kwok, Y. F. Leung, and Y. D. Sobral, “Characterization of base roughness for granular chute flows”, *Phys. Rev. E*, vol. 94, p. 052 901, 5 Nov. 2016. DOI: `10.1103/PhysRevE.94.052901`. [Online]. Available: `https://link.aps.org/doi/10.1103/PhysRevE.94.052901`.

[42] L. Jing, C. Y. Kwok, Y. F. Leung, and Y. D. Sobral, “Extended cfd–dem for free-surface flow with multi-size granules”, *International Journal for Numerical and Analytical Methods in Geomechanics*, vol. 40, no. 1, pp. 62–79, 2016. DOI: `https://doi.org/10.1002/nag.2387`. eprint: `https://onlinelibrary.wiley.com/doi/pdf/10.1002/nag.2387`. [Online]. Available: `https://onlinelibrary.wiley.com/doi/abs/10.1002/nag.2387`.

[43] M. Joldes, J.-M. Muller, and V. Popescu, “Tight and rigorous error bounds for basic building blocks of double-word arithmetic”, *ACM Transactions on Mathematical Software*, vol. 44, pp. 1–27, Oct. 2017. DOI: `10.1145/3121432`.

[44] K. Kaheman, J. N. Kutz, and S. L. Brunton, “Sindy-pi: A robust algorithm for parallel implicit sparse identification of nonlinear dynamics”, *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 476, no. 2242, p. 20 200 279, 2020. DOI: `10.1098/rspa.2020.0279`. eprint: `https://royalsocietypublishing.org/doi/pdf/10.1098/rspa.2020.0279`. [Online]. Available: `https://royalsocietypublishing.org/doi/abs/10.1098/rspa.2020.0279`.

[45] K. Kamrin, “Non-locality in granular flow: Phenomenology and modeling approaches”, *Front. Phys.*, vol. 7, 2019. DOI: `10.3389/fphy.2019.00116`. [Online]. Available: `https://www.frontiersin.org/articles/10.3389/fphy.2019.00116/full`.

[46] A. A. Kaptanoglu, K. D. Morgan, C. J. Hansen, and S. L. Brunton, “Physics-constrained, low-dimensional models for magnetohydrodynamics: First-principles and data-driven approaches”, *Phys. Rev. E*, vol. 104, p. 015 206, 1 Jul. 2021. DOI: `10.1103/PhysRevE.104.015206`. [Online]. Available: `https://link.aps.org/doi/10.1103/PhysRevE.104.015206`.

[47] H. Kruggel-Emden, M. Sturm, S. Wirtz, and V. Scherer, “Selection of an appropriate time integration scheme for the discrete element method (dem)”, *Computers & Chemical Engineering*, vol. 32, no. 10, pp. 2263–2279, 2008, ISSN: 0098-1354. DOI: `https://doi.org/10.1016/j.compchemeng.2007.11.002`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S0098135407002864`.

[48] J. N. Kutz, S. L. Brunton, B. W. Brunton, and J. L. Proctor, *Dynamic Mode Decomposition*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2016. DOI: `10.1137/1.9781611974508`. eprint: `https://epubs.siam.org/doi/pdf/10.1137/1.9781611974508`. [Online]. Available: `https://epubs.siam.org/doi/abs/10.1137/1.9781611974508`.

[49] G. Kuwabara and K. Kono, “Restitution coefficient in a collision between two spheres”, *Japanese Journal of Applied Physics*, vol. 26, no. Part 1, No. 8, pp. 1230–1233, Aug. 1987. DOI: `10.1143/jjap.26.1230`. [Online]. Available: `https://doi.org/10.1143/jjap.26.1230`.

[50] I. Lagaris, A. Likas, and D. Fotiadis, “Artificial neural networks for solving ordinary and partial differential equations”, *IEEE Transactions on Neural Networks*, vol. 9, no. 5, pp. 987–1000, 1998. DOI: `10.1109/72.712178`. [Online]. Available: `https://doi.org/10.1109%2F72.712178`.

[51] P.-Y. Lagrée, L. Staron, and S. Popinet, “The granular column collapse as a continuum: Validity of a two-dimensional navier–stokes model with a $\mu(I)$-rheology”, *Journal of Fluid Mechanics*, vol. 686, pp. 378–408, 2011. DOI: `10.1017/jfm.2011.335`.

[52] M. Lätzel, S. Luding, and H. J. Herrmann, “Macroscopic material properties from quasi-static, microscopic simulations of a two-dimensional shear-cell”, *Granular Matter*, vol. 2, pp. 123–135, 3 Jun. 2000. DOI: `10.1007/s100350000048`. [Online]. Available: `https://doi.org/10.1007/s100350000048`.

[53] J.-C. Loiseau and S. L. Brunton, “Constrained sparse galerkin regression”, *Journal of Fluid Mechanics*, vol. 838, pp. 42–67, 2018. DOI: `10.1017/jfm.2017.823`.

[54] L. Lu, X. Gao, J.-F. Dietiker, M. Shahnam, and W. Rogers, “Machine learning accelerated discrete element modeling of granular flows”, *Chemical Engineering Science*, vol. 245, p. 116 832, Jun. 2021. DOI: `10.1016/j.ces.2021.116832`.

[55] G. H. B. Martins, W. A. M. Morgado, S. M. D. Queirós, and A. P. F. Atman, "Large-deviation quantification of boundary conditions on the brazil nut effect", *Phys. Rev. E*, vol. 103, p. 062 901, 6 Jun. 2021. DOI: 10.1103/PhysRevE.103.062901. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevE.103.062901.

[56] D. A. Messenger and D. M. Bortz, "Weak sindy for partial differential equations", *Journal of Computational Physics*, vol. 443, p. 110 525, 2021, ISSN: 0021-9991. DOI: https://doi.org/10.1016/j.jcp.2021.110525. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0021999121004204.

[57] G. Midi, "On dense granular flows", *European Physical Journal E*, vol. 14, pp. 341–365, Aug. 2004. DOI: 10.1140/epje/i2003-10153-0.

[58] D. M. Mueth, "Measurements of particle dynamics in slow, dense granular couette flow", *Phys. Rev. E*, vol. 67, p. 011 304, 2 Jan. 2003. DOI: 10.1103/PhysRevE.67.011304. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevE.67.011304.

[59] C. Nouguier-Lehon, B. Cambou, and E. Vincens, "Influence of particle shape and angularity on the behaviour of granular materials: A numerical analysis", *International Journal for Numerical and Analytical Methods in Geomechanics*, vol. 27, no. 14, pp. 1207–1226, 2003. DOI: https://doi.org/10.1002/nag.314. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/nag.314. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/nag.314.

[60] OpenAI, *Gpt-4 technical report*, 2023. arXiv: 2303.08774 [cs.CL].

[61] F. Pacheco-Vázquez, "Ray systems and craters generated by the impact of nonspherical projectiles", *Phys. Rev. Lett.*, vol. 122, p. 164 501, 16 Apr. 2019. DOI: 10.1103/PhysRevLett.122.164501. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.122.164501.

[62] E. Pennestrì, V. Rossi, P. Salvini, and P. P. Valentini, "Review and comparison of dry friction force models", *Nonlinear Dynamics*, vol. 83, no. 4, pp. 1785–1801, Mar. 2016, ISSN: 1573-269X. DOI: 10.1007/s11071-015-2485-3. [Online]. Available: https://doi.org/10.1007/s11071-015-2485-3.

[63] F. Radjai and F. Dubois, *Discrete-element modeling of granular materials.* Wiley-Iste, 2011, 425 p. [Online]. Available: https://hal.archives-ouvertes.fr/hal-00691805.

[64] M. Raissi and G. E. Karniadakis, "Hidden physics models: Machine learning of nonlinear partial differential equations", *Journal of Computational Physics*, vol. 357, pp. 125–141, Mar. 2018. DOI: 10.1016/j.jcp.2017.11.039. [Online]. Available: https://doi.org/10.1016%2Fj.jcp.2017.11.039.

[65] L. E. Reichl, *A Modern Course in Statistical Physics.* Wiley and numerous references therein, 1998.

[66] A. Ries, L. Brendel, and D. Wolf, "Coarse graining strategies at walls", *Computational Particle Mechanics*, vol. 1, pp. 177–190, Jun. 2014. DOI: 10.1007/s40571-014-0023-6.

[67] D. Rim, E. N. MillÃ¡n, B. Planes, E. M. Bringa, and L. G. Moyano, "Cluster analysis for granular mechanics simulations using Machine Learning Algorithms", en, *Entre Ciencia e IngenierÃa*, vol. 14, pp. 82–87, Dec. 2020, ISSN: 1909-8367. [Online]. Available: `http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S1909-83672020000200082&nrm=iso`.

[68] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, *High-resolution image synthesis with latent diffusion models*, 2022. arXiv: `2112.10752 [cs.CV]`.

[69] S. H. Rudy, S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Data-driven discovery of partial differential equations", *Science Advances*, vol. 3, no. 4, e1602614, 2017. DOI: `10.1126/sciadv.1602614`. eprint: `https://www.science.org/doi/pdf/10.1126/sciadv.1602614`. [Online]. Available: `https://www.science.org/doi/abs/10.1126/sciadv.1602614`.

[70] S. Schöllmann, "Simulation of a two-dimensional shear cell", *Phys. Rev. E*, vol. 59, pp. 889–899, 1 Jan. 1999. DOI: `10.1103/PhysRevE.59.889`. [Online]. Available: `https://link.aps.org/doi/10.1103/PhysRevE.59.889`.

[71] J. Shäfer, S. Dippel, and D. Wolf, "Force Schemes in Simulations of Granular Materials", *Journal de Physique I*, vol. 6, no. 1, pp. 5–20, 1996. DOI: `10.1051/jp1:1996129`. [Online]. Available: `https://hal.archives-ouvertes.fr/jpa-00247176`.

[72] M. Sorokina, S. Sygletos, and S. Turitsyn, "Sparse identification for nonlinear optical communication systems: Sino method", *Optics Express*, vol. 24, p. 30 433, Dec. 2016. DOI: `10.1364/OE.24.030433`.

[73] A. Staal, O. Tuinenburg, J. Bosmans, *et al.*, "Forest-rainfall cascades buffer against drought across the amazon", *Nature Climate Change*, vol. 8, Jun. 2018. DOI: `10.1038/s41558-018-0177-y`.

[74] L. Staron, P. .-. Lagrée, and S. Popinet, "Continuum simulation of the discharge of the granular silo", *The European Physical Journal E*, vol. 37, 1 Jan. 2014. DOI: `10.1140/epje/i2014-14005-6`. [Online]. Available: `https://doi.org/10.1140/epje/i2014-14005-6`.

[75] A. Stevens and C. Hrenya, "Comparison of soft-sphere models to measurements of collision properties during normal impacts", *Powder Technology*, vol. 154, no. 2, pp. 99–109, 2005, ISSN: 0032-5910. DOI: `https://doi.org/10.1016/j.powtec.2005.04.033`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S0032591005001658`.

[76] C. Störmer, "Méthode d'intégration numérique des équations différentielles ordinaires", presented at the C.R. Congress Internat. 1920, Strassbourg, 1921, pp. 243–257.

[77] I. G. Tejada and P. Antolin, *Use of machine learning for unraveling hidden correlations between particle size distributions and the mechanical behavior of granular materials*, 2020. arXiv: `2006.05711 [cond-mat.dis-nn]`.

[78] S. Thaler, L. Paehler, and N. A. Adams, "Sparse identification of truncation errors", *Journal of Computational Physics*, vol. 397, p. 108 851, 2019, ISSN: 0021-9991. DOI: `https://doi.org/10.1016/j.jcp.2019.07.049`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S0021999119305352`.

[79] J. Thijssen, *Computational Physics*, 2nd ed. Cambridge University Press, 2007. DOI: 10.1017/CBO9781139171397.

[80] S. Urbini, I. Nicolosi, A. Zeoli, *et al.*, “Geological and geophysical investigation of kamil crater, egypt”, *Meteoritics & Planetary Science*, vol. 47, no. 11, pp. 1842–1868, 2012. DOI: https://doi.org/10.1111/maps.12023. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/maps.12023. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/maps.12023.

[81] L. Verlet, “Computer "experiments" on classical fluids. i. thermodynamical properties of lennard-jones molecules”, *Phys. Rev.*, vol. 159, pp. 98–103, 1 Jul. 1967. DOI: 10.1103/PhysRev.159.98. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRev.159.98.

[82] W. Wang, X. Liu, and k. Liu, “Experimental research on force transmission of dense granular assembly under shearing in taylor–couette geometry”, *Tribology Letters*, vol. 48, 2 Nov. 2012. DOI: 10.1007/s11249-012-0009-6. [Online]. Available: https://doi.org/10.1007/s11249-012-0009-6.

[83] J. Weber, “Recherches concernant les contraintes intergranulaires dans les milieux pulvérulents application aux lois de similitude dans les études sur modèles réduits de problèmes de mécanique des sols pulvérulents”, 1966.

[84] M. Wu, Z. Xia, and J. Wang, “Constitutive modelling of idealised granular materials using machine learning method”, *Journal of Rock Mechanics and Geotechnical Engineering*, vol. 15, no. 4, pp. 1038–1051, 2023, ISSN: 1674-7755. DOI: https://doi.org/10.1016/j.jrmge.2022.08.002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1674775522001688.

[85] G. Yang, L. Jing, C. Kwok, and Y. Sobral, “A comprehensive parametric study of lbm-dem for immersed granular flows”, *Computers and Geotechnics*, vol. 114, p. 103 100, 2019, ISSN: 0266-352X. DOI: https://doi.org/10.1016/j.compgeo.2019.103100. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0266352X19301569.

[86] G. Yang, S. Yang, L. Jing, C. Kwok, and Y. Sobral, “Efficient lattice boltzmann simulation of free-surface granular flows with $\mu(I)$-rheology”, *Journal of Computational Physics*, vol. 479, p. 111 956, 2023, ISSN: 0021-9991. DOI: https://doi.org/10.1016/j.jcp.2023.111956. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0021999123000517.

[87] H. Yoshida, “Construction of higher order symplectic integrators”, *Physics Letters A*, vol. 150, no. 5, pp. 262–268, 1990, ISSN: 0375-9601. DOI: https://doi.org/10.1016/0375-9601(90)90092-3. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0375960190900923.

[88] H. Yu, M. Chin, T. Yuan, *et al.*, “The fertilizing role of african dust in the amazon rainforest: A first multiyear assessment based on data from cloud-aerosol lidar and infrared pathfinder satellite observations”, *Geophysical Research Letters*, vol. 42, no. 6, pp. 1984–1991, 2015. DOI: https://doi.org/10.1002/2015GL063040. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2015GL063040. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015GL063040.

[89] L. Zanna and T. Bolton, “Data-driven equation discovery of ocean mesoscale closures”, *Geophysical Research Letters*, 2020. DOI: `https://doi.org/10.1029/2020GL088376`.