Journal of Computational Physics 231 (2012) 759–794

Contents lists available at ScienceDirect

Journal of Computational Physics

journal homepage: www.elsevier.com/locate/jcp

Smoothed particle hydrodynamics and magnetohydrodynamics ☆

Daniel J. Price

Centre for Stellar and Planetary Astrophysics & School of Mathematical Sciences, Monash University, Melbourne Vic. 3800, Australia

ARTICLE INFO

Article history:
Available online 15 December 2010

Keywords:
Particle methods
Hydrodynamics
Smoothed particle hydrodynamics
Magnetohydrodynamics (MHD)
Astrophysics

ABSTRACT

This paper presents an overview and introduction to smoothed particle hydrodynamics and magnetohydrodynamics in theory and in practice. Firstly, we give a basic grounding in the fundamentals of SPH, showing how the equations of motion and energy can be self-consistently derived from the density estimate. We then show how to interpret these equations using the basic SPH interpolation formulae and highlight the subtle difference in approach between SPH and other particle methods. In doing so, we also critique several 'urban myths' regarding SPH, in particular the idea that one can simply increase the 'neighbour number' more slowly than the total number of particles in order to obtain convergence. We also discuss the origin of numerical instabilities such as the pairing and tensile instabilities. Finally, we give practical advice on how to resolve three of the main issues with SPMHD: removing the tensile instability, formulating dissipative terms for MHD shocks and enforcing the divergence constraint on the particles, and we give the current status of developments in this area. Accompanying the paper is the first public release of the NDSPMHD SPH code, a 1, 2 and 3 dimensional code designed as a testbed for SPH/SPMHD algorithms that can be used to test many of the ideas and used to run all of the numerical examples contained in the paper.

1. Introduction

Smoothed particle hydrodynamics (SPH), originally formulated by Lucy [53] and Gingold and Monaghan [37], is by now very widely used for many diverse applications in astrophysics, geophysics, engineering and in the film and computer games industry. Whilst numerous excellent reviews already exist (e.g. [59,64,78,92]), there remain – particularly in the astrophysical domain – some widespread misconceptions about its use, and more importantly, its fundamental basis.

Our aim in this – mainly pedagogical – review is therefore not to provide a comprehensive survey of SPH applications, nor the implementation of particular physical models, but to re-address the fundamentals about why and how the method works, and to give practical guidance on how to formulate general SPH algorithms and avoid some of the common pitfalls in using SPH. Since such an understanding is critical to the development of robust and accurate methods for magnetohydrodynamics (MHD) in SPH (hereafter referred to as "smoothed particle magnetohydrodynamics", SPMHD), this will lead us naturally on to review the background and current status in this area – particularly relevant given the importance of MHD in most, if not all, astrophysical problems. Whilst the paper is written with an astrophysical flavour in mind (given the topical issue of JCP for which it is written), the principles are general and thus are applicable in any of the areas in which SPH is applied.

☆ This review and associated material germinated as lectures and tutorials given as part of the ASTROSIM summer school on computational astrophysics held during July 2010 in Toruń, Poland. A video of the original lectures can be viewed online at http://supercomputing.astri.umk.pl/. The paper also presents an updated version of much of the otherwise unpublished material in my PhD thesis [78].

E-mail address: daniel.price@monash.edu
URL: http://users.monash.edu.au/~dprice/SPH/

0021-9991/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jcp.2010.12.011

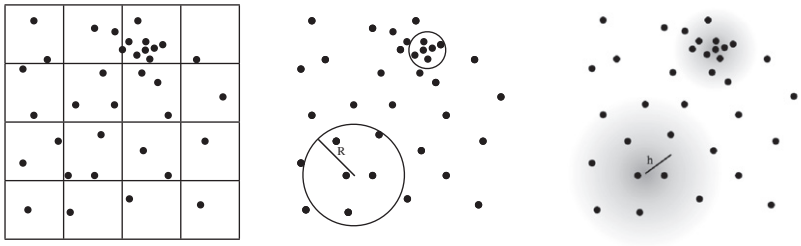

Fig. 1. Computing a continuous density field from a collection of point mass particles. (a) In particle-mesh methods (left panel) the density is computed by interpolating the mass to a grid (or simply dividing the mass by the volume). However, this tends to over/under resolve clustered/sparse regions. (b) An alternative not requiring a mesh is to construct a local volume around the sampling point, solving the clustering problem by scaling the sample volume according to the local number density of particles. (c) This panel shows the approach adopted in SPH, where the density is computed via a weighted sum over neighbouring particles, with the weight decreasing with distance from the sample point according to a scale factor h.

Finally, alongside this article I have released a public version of my NDSPMHD SPH/SPMHD code, along with a set of easy-to-follow numerical exercises – consisting of setup and input files for the code and step-by-step instructions for each problem in 1, 2 and 3 dimensions – the problems themselves having been chosen to illustrate many of the theoretical points in this paper. Indeed, NDSPMHD has been used to compute all of the test problems and examples shown. The hope is that this will become a useful resource.[1] not only for advanced researchers but also for students embarking on an SPH-based research topic.

2. The foundations of SPH: calculating density

The usual introductory lines on SPH refer to it as a "Lagrangian particle method for solving the equations of hydrodynamics". However, SPH starts with a basis much more fundamental than that, as the answer to the following question:

How does one compute the density from an arbitrary distribution of point mass particles?

This problem arises in many areas other than hydrodynamics, for example in obtaining the solution to Poisson's equation for the gravitational field $\nabla^2\Phi = 4\pi G\rho(\mathbf{r})$ when a (continuous) density field is represented by a collection of point masses.

2.1. Approaches to computing the density

Three common approaches are summarised in Fig. 1. Perhaps the most straightforward (Fig. 1(a)) is to construct a mesh of some sort and divide the mass in each cell by the volume. This basic approach forms the basis of hybrid particle-mesh methods such as Marker-In-Cell e.g. [39] and Particle-In-Cell [42] schemes, where one can further improve the density estimate using any of the standard particle-cell interpolation methods, such as Cloud-In-Cell (CIC), Triangular-Shaped-Cloud (TSC) etc. However there are clear limitations – firstly that a fixed mesh will inevitably over/under-sample dense/sparse regions (respectively) when the mass distribution is highly clustered[2]; and secondly a loss of accuracy, speed and consistency because of the need to interpolate both to/from the particles, for example to compute forces.

The second approach (Fig. 1(b)) is to remove the mesh entirely and instead calculate the density based on a local sampling of the mass distribution, for example in a sphere centred on the location of the sampling point (which may or may not be the location of a particle itself). The most basic scheme would be to divide the total mass by the sampling volume, i.e.,

$$\rho(\mathbf{r}) = \frac{\sum_{b=1}^{N_{neigh}} m_b}{\frac{4}{3}\pi R^3}. \tag{1}$$

The problem of resolving clustered/sparse regions can be easily addressed in this method by adjusting the size of the sampling volume according to the local number density of sampling points, for example by computing with a fixed "number of neighbours" for each particle – as shown in Fig. 1. However, this leads to a very noisy estimate, since the density estimate will be very sensitive to whether a distant particle on the edge of the volume is "in" or "out" of the estimate (with $\delta\rho \propto 1/N_{neigh}$ for equal mass particles). This leads naturally to the idea that one should progressively down-weight the contributions from neighbouring particles as their relative distance increases, in order that changes in distant particles have a progressively smaller influence on the local estimate (that is, the density estimate is *smoothed*).

[1] NDSPMHD is available from http://users.monash.edu.au/dprice/SPH/. Note that we do not advocate the use of NDSPMHD as a "performance" code in 3D, since it is not designed for this purpose and excellent parallel 3D codes already exist (such as the GADGET code by Springel [98]). Rather it is meant as a testbed for algorithmic experimentation and understanding.

[2] More recently, this problem has been addressed by the use of adaptively refined meshes to calculate the density field e.g. [19].

2.2. The SPH density estimator

This third approach forms the basis of SPH and is shown in Fig. 1(c): Here the density is computed using a weighted summation over nearby particles, given by

$$\rho(\mathbf{r}) = \sum_{b=1}^{N_{neigh}} m_b W(\mathbf{r} - \mathbf{r}_b, h), \tag{2}$$

where W is an (as yet unspecified) weight function with dimensions of inverse volume and h is a scale parameter determining the rate of fall-off of W as a function of the particle spacing (also yet to be determined). Conservation of total mass $\int \rho \mathrm{dV} = \sum_{b=1}^{N_{part}} m_b$ implies a normalisation condition on W given by

$$\int_V W(\mathbf{r}' - \mathbf{r}_b, h)\mathrm{d}V' = 1. \tag{3}$$

The accuracy of the density estimate then rests on the choice of a sufficiently good weight function (hereafter referred to as the *smoothing kernel*). Elementary considerations suggest that a good density kernel should have at least the following properties:

1. A weighting that is positive, decreases monotonically with relative distance and has smooth derivatives;
2. Symmetry with respect to $(\mathbf{r} - \mathbf{r}')$ – i.e., $W(\mathbf{r}' - \mathbf{r}, h) \equiv W(|\mathbf{r}' - \mathbf{r}|, h)$; and
3. A flat central portion so the density estimate is not strongly affected by a small change in position of a near neighbour.

A natural choice that satisfies all of the above properties is the Gaussian:

$$W(\mathbf{r} - \mathbf{r}', h) = \frac{\sigma}{h^d} \exp\left[-\frac{(\mathbf{r} - \mathbf{r}')^2}{h^2}\right], \tag{4}$$

where d refers to the number of spatial dimensions and σ is a normalisation factor given by $\sigma = \left[1/\sqrt{\pi},\ 1/\pi,\ 1/(\pi\sqrt{\pi})\right]$ in [1,2,3] dimensions. The Gaussian satisfies condition 1 particularly well since it is infinitely smooth (differentiable) – and gives in practice an excellent density estimate. However it has the practical disadvantage of requiring interaction with all of the particles in the domain [with computational cost of $\mathcal{O}(N^2)$ if computing the density at the particle locations], despite the fact that the relative contribution from neighbouring particles quickly becomes negligible with increasing distance. Thus in practice it is better to use a kernel that is Gaussian-like in shape but truncated at a finite radius (e.g. a few times the scale length, h). Using kernels with such "compact support" means a much more efficient density evaluation, since the cost scales like $\mathcal{O}(N_{neigh}N)$, but inevitably leads to a more noisy density estimate since one is more sensitive to small changes in the local distribution.

2.3. Kernel functions with compact support

There are many kernel functions which fit this bill. The most well-used (for SPH at least) are the Schoenberg [96] B-spline functions [58,64,66], generated as the Fourier transform

$$M_n(x, h) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \left[\frac{\sin(kh/2)}{kh/2}\right]^n \cos(kx)\mathrm{d}k. \tag{5}$$

These give progressively better approximations to the Gaussian at higher n, both by increasing the radius of compact support and by increasing smoothness, since each function M_n is continuous up to the $\{n-2\}$th derivatives. Since we minimally require continuity in at least the first and second derivatives, the lowest order B-spline useful for SPH is the M_4 (cubic) spline truncated at $2h$:

$$w(q) = \sigma \begin{cases} \frac{1}{4}(2-q)^3 - (1-q)^3, & 0 \leqslant q < 1; \\ \frac{1}{4}(2-q)^3, & 1 \leqslant q < 2; \\ 0. & q \geqslant 2, \end{cases} \tag{6}$$

where for convenience we use $W(|\mathbf{r} - \mathbf{r}'|, h) \equiv \frac{1}{h^d} w(q)$, where $q = |\mathbf{r} - \mathbf{r}'|/h$ and σ is a normalisation constant given by $\sigma = [2/3, 10/(7\pi), 1/\pi]$ in [1,2,3] dimensions. Next are the M_5 quartic, truncated at $2.5h$:

$$w(q) = \sigma \begin{cases} \left(\frac{5}{2}-q\right)^4 - 5\left(\frac{3}{2}-q\right)^4 + 10\left(\frac{1}{2}-q\right)^4, & 0 \leqslant q < \frac{1}{2}; \\ \left(\frac{5}{2}-q\right)^4 - 5\left(\frac{3}{2}-q\right)^4, & \frac{1}{2} \leqslant q < \frac{3}{2}; \\ \left(\frac{5}{2}-q\right)^4, & \frac{3}{2} \leqslant q < \frac{5}{2}; \\ 0. & q \geqslant \frac{5}{2}, \end{cases} \tag{7}$$

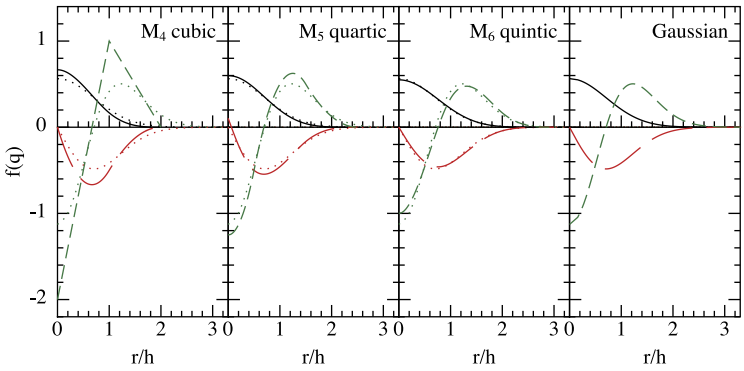

Fig. 2. The M_4 (cubic, truncated at $2h$), M_5 (quartic, truncated at $2.5h$) and M_6 (quintic, truncated at $3h$) Schoenberg [96] B-spline kernel functions (solid lines) and their first (long-dashed) and second (short-dashed) derivatives, compared to the Gaussian (right panel and dotted lines in other panels). Notice that although the "number of neighbours" increases in the M_5 and M_6 functions compared to the cubic spline, the smoothing scale h retains the same meaning with respect to the Gaussian. Thus, using the higher order B-splines is a way to increase the smoothness of the kernel summations without altering the resolution length, and is very different to simply increasing the number of neighbours under the cubic spline.

with normalisation $\sigma = [1/24, 96/(1199\pi), 1/(20\pi)]$, and the M_6 quintic, truncated at $3h$:

$$w(q) = \sigma \begin{cases} (3-q)^5 - 6(2-q)^5 + 15(1-q)^5, & 0 \leqslant q < 1; \\ (3-q)^5 - 6(2-q)^5, & 1 \leqslant q < 2; \\ (3-q)^5, & 2 \leqslant q < 3; \\ 0. & q \geqslant 3, \end{cases} \tag{8}$$

with normalisation $\sigma = [1/120, 7/(478\pi), 1/(120\pi)]$ (e.g. [69]). These kernel functions and their first and second derivatives are shown for comparison with the Gaussian in Fig. 2.

One important aspect to draw from our discussion so far is the clear meaning attached to the smoothing length h as specifying the fall-off of the kernel weighting with respect to the particle separation. In particular, it is clear that referring to the "number of neighbours" does not have any meaning *per se* for Gaussian and Gaussian-like kernels: For the Gaussian the number of neighbours is in principle infinite, but there nevertheless exists a well-defined smoothing scale, h. The higher order B-splines (Fig. 2) also demonstrate that it is possible to change the "neighbour number" – by progressing to higher n in the series – *without* changing the smoothing length. It is a widely-propagated myth that one can achieve formal convergence in SPH by "increasing the number of neighbours" (e.g. more slowly than the total number of particles). However, there are very important differences between simply "stretching" the cubic spline to accommodate a larger neighbour number – which amounts to changing the ratio of h to particle spacing – and using a kernel that has a larger radius of compact support but retains the same h. That is, in no sense is the SPH density estimate (our "approach 3") the same as approach 2 shown in Fig. 1(b). We will return to this point later.

There are obviously other kernels, and other families of kernels, that satisfy the above properties e.g. [24]. However, more detailed investigations into kernels e.g. [35] tend to merely confirm the points made above – namely that Bell-shaped, symmetric, monotonic kernels provide the best density estimates. We will examine the formal errors in the kernel density estimate shortly, but first we turn to the issue of setting the smoothing length, h.

2.4. Setting the smoothing length

Early SPH simulations e.g. [37] simply employed a spatially constant resolution length h, though one which was allowed to change as a function of time according to the densest part of a calculation.[3] However, as is evident from Fig. 1, it is clearly desirable to resolve both clustered and sparse regions evenly – that is, with a roughly constant ratio of h to the mean local particle separation. Thus, a natural choice for setting the smoothing length is to relate to the local number density of particles, i.e.,

$$h(\mathbf{r}) \propto n(\mathbf{r})^{-1/d}; \quad n(\mathbf{r}) = \sum_b W[\mathbf{r} - \mathbf{r}_b, h(\mathbf{r})]. \tag{9}$$

For equal mass particles, this is equivalent to making h proportional to the density itself (since $1/n \equiv m/\rho$). Since in turn density is itself a function of smoothing length, this leads to the idea of an iterative summation to simultaneously obtain the (mutually dependent) $\rho(\mathbf{r})$ and $h(\mathbf{r})$ [62,89,99]. Computed at the location of particle a, we have a set of two simultaneous equations

[3] Similar to the spatially fixed but time-evolved gravitational softening lengths still employed in many cosmological simulations.

$$\rho(\mathbf{r}_a) = \sum_b m_b W(\mathbf{r}_a - \mathbf{r}_b, h_a); \quad h(\mathbf{r}_a) = \eta\left(\frac{m_a}{\rho_a}\right)^{1/d}, \tag{10}$$

where η is a parameter specifying the smoothing length in units of the mean particle spacing $(m/\rho)^{1/d}$. These two equations can be solved simultaneously using standard root-finding methods such as Newton–Raphson or Bisection and most "modern" SPH codes employ such a procedure (for reasons that will become clear). Note that enforcing the relation given in (10) is approximately equivalent to keeping the "mass inside the smoothing sphere" constant [99], since for example in three dimensions

$$M^a_{tot} = \int_{V_a} \rho dV \approx \frac{4}{3}\pi R^3_{kern}\rho_a, \tag{11}$$

where R_{kern} is the kernel radius ($2h$ for the cubic spline), so $M_{tot} = const$ implies $h^3\rho = const$. Since for equal mass particles $M_{tot} = m\ N_{neigh}$, this also means that the number of neighbours should be approximately constant if the relationship in (10) (or for unequal mass particles, Eq. 9) is enforced. Indeed a "number of neighbours" parameter can be used in place of the parameter η, using

$$N_{neigh,1D} = 2\zeta\eta; \quad N_{neigh,2D} = \pi(\zeta\eta)^2; \quad N_{neigh,3D} = \frac{4}{3}\pi(\zeta\eta)^3, \tag{12}$$

where ζ is the compact support radius in units of h (i.e., $\zeta = 2$ for the cubic spline). However, this is problematic for several reasons. Firstly it gives the dangerous impression that N_{neigh} is a free parameter unrelated to h, whereas changing N_{neigh} explicitly changes h – more specifically, the ratio of h to particle spacing (i.e., η) – and corresponds to "stretching" the cubic spline as discussed above. Secondly, whereas η carries the same meaning in 1, 2 and 3 dimensions, the N_{neigh} parameter changes, making it difficult to relate the results of one and two dimensional test problems to three dimensional simulations. Thirdly, N_{neigh} is often used as an integer parameter, whereas it is clear from (10) that the $h - \rho$ (or $h - n$) iterations can be performed to arbitrary accuracy (that is, to fractional neighbour numbers) – which is also necessary if one is to assume that the relationship is differentiable. Finally, N_{neigh} is only related to the *true* number of neighbours so long as the (number) density of particles within the smoothing sphere is approximately constant (that is, so far as the integral in Eq. (11) can be approximated by $\frac{4}{3}\pi R^3\rho$). So at best a N_{neigh} parameter only characterises the *mean* neighbour number – and there can be strong fluctuations about this mean, for example in strong density gradients.

Earlier adaptive SPH implementations employed density estimates involving either an average smoothing length $\bar{h} = \frac{1}{2}(h_a + h_b)$ e.g. [5] or an average of the smoothing kernels $\overline{W}_{ab} = \frac{1}{2}[W_{ab}(h_a) + W_{ab}(h_b)]$ e.g. [40]. However, this inevitably leads to heuristic methods for setting the smoothing length itself – for example by evolving the time derivative of Eq. (10) with "corrections" to try to keep the neighbour number approximately constant e.g. [5] or simply enforcing a constant neighbour number either approximately or exactly e.g. [40]. It also considerably complicated attempts to incorporate derivatives of the smoothing length – necessary for exact energy and entropy conservation – into the equations of motion [73]. By contrast, the mathematical meaning of Eq. (10) is clear and it is straightforward to take derivatives involving the smoothing length.

Finally, the density estimate computed via (10) is time-independent, depending only on particle positions and masses and thus explicitly answering our original question of how to compute a density field from point mass particles. This also means it has wide applicability to many other problems beyond SPH – for example Price and Federrath [85] use it to construct a density field from Lagrangian tracer particles in grid-based simulations of supersonic turbulence; and it forms the basis of the adaptive gravitational force softening method introduced by Price and Monaghan [89].

2.5. Errors in the density estimate

The formal errors in the density estimate may be determined by writing the density summation as an integral – that is, assuming $m \equiv \rho \mathrm{d}V$ and that the summation is well sampled, giving

$$\langle\rho(\mathbf{r})\rangle = \int \rho(\mathbf{r}')W(\mathbf{r} - \mathbf{r}', h)\mathrm{d}V'. \tag{13}$$

where $\langle\cdot\cdot\rangle$ refers to a smoothed estimate. Expanding $\rho(\mathbf{r}')$ in a Taylor series about $\mathbf{r}$, we have

$$\langle\rho(\mathbf{r})\rangle = \rho(\mathbf{r})\int W(\mathbf{r} - \mathbf{r}', h)dV' + \nabla\rho(\mathbf{r})\cdot\int(\mathbf{r}' - \mathbf{r})W(\mathbf{r} - \mathbf{r}', h)dV' + \nabla^\alpha\nabla^\beta\rho(\mathbf{r})\int\delta\mathbf{r}^\alpha\delta\mathbf{r}^\beta W(\mathbf{r} - \mathbf{r}', h)dV' + \mathcal{O}(h^3) \tag{14}$$

so that if the normalisation condition (3) is satisfied and a symmetric kernel $W(\mathbf{r} - \mathbf{r}', h) = W(\mathbf{r}' - \mathbf{r}, h)$ is used such that the odd error terms vanish, the error in the density interpolant is $\mathcal{O}(h^2)$. In principle it is also possible to construct kernels such that the second moment is also zero, resulting in errors of $\mathcal{O}(h^4)$ see [58]. The disadvantage of such kernels is that the kernel function becomes negative in some part of the domain, resulting in a potentially negative density evaluation. Achieving such higher order in practice also requires that the kernel is extremely well sampled, leading to substantial additional cost requirements. One possibility given the iterations necessary to solve (10) would be to automatically switch between high order and low order kernels during the iterations (e.g. if a negative density occurs), thus leading to high order interpolation

in smooth regions but a low order interpolation where the density changes rapidly. The errors in the discrete version are discussed further in Section 4.3.

2.6. Alternatives to the SPH density estimate

Finally, it should be noted that the three general approaches described in Section 2.2 are not the only methods that can be employed for estimating the density. A fourth alternative that has received recent attention involves the use of Delaunay or Voronoi tessellation – the former proposed by Pelupessy et al. [75] and the latter developed into a full hydrodynamics scheme by Serrano et al. [97] and Heß and Springel [41]. These are promising approaches that in principle can offer all the same advantages as SPH in terms of exact conservation – since it can be derived similarly from a Hamiltonian formulation – but with an improved density estimate and an exact partition of unity.

3. From density to equations of motion

The reader at this point may wonder why we have spent so long discussing nothing else except the density estimate. The reason is that this is the only real freedom one has if one wishes to obtain a fully conservative SPH algorithm, at least in the absence of dissipative terms. This is because the rest of the SPH algorithm can be derived entirely *from* the density estimate.

3.1. The discrete Lagrangian

The derivation starts with the discrete Lagrangian. As is usual, the Lagrangian is simply given by

$$L = T - V, \tag{15}$$

where T and V are the kinetic and potential (in this case, thermal) energies, respectively. For a system of point masses with velocity $\mathbf{v} \equiv \mathrm{d}\mathbf{r}/\mathrm{d}t$ and internal energy per unit mass u, we have

$$L = \sum_b m_b \left[\frac{1}{2} v_b^2 - u_b(\rho_b, s_b)\right], \tag{16}$$

where in general the thermal energy u can be specified as a function of the thermodynamic variables ρ and s (the density and entropy, respectively). Although (16) can be considered as a discrete version of the continuum Lagrangian for hydrodynamics e.g. [30,71,95]

$$L = \int \left[\rho v^2 - \rho u(\rho, s)\right] \mathrm{d}V, \tag{17}$$

one is free to consider the discrete Hamiltonian system, it's associated symmetries and equations of motion directly – that is, without explicit reference to the continuum system. In other words the Hamiltonian properties are directly present in the discrete system and the motions will be constrained to obey the symmetries and conservation properties of the discrete Lagrangian.

3.2. Least action principle and the Euler–Lagrange equations

The equations of motion for such a system can be derived from the principle of least action, that is minimising the action

$$S = \int L \mathrm{d}t, \tag{18}$$

such that $\delta S = \int \delta L \mathrm{d}t = 0$, where δ is a variation with respect to a small change in the particle coordinates $\delta\mathbf{r}$. Assuming that the Lagrangian can be written as a differentiable function of the particle positions $\mathbf{r}$ and velocities $\mathbf{v}$, we have

$$\delta S = \int \left(\frac{\partial L}{\partial \mathbf{v}} \cdot \delta\mathbf{v} + \frac{\partial L}{\partial \mathbf{r}} \cdot \delta\mathbf{r}\right) \mathrm{d}t = 0. \tag{19}$$

Integrating by parts, using the fact that $\delta\mathbf{v} = \mathrm{d}(\delta\mathbf{r})/\mathrm{d}t$, where $\mathrm{d}/\mathrm{d}t \equiv \partial/\partial t + \mathbf{v} \cdot \nabla$ gives

$$\delta S = \int \left\{\left[-\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \mathbf{v}}\right) + \frac{\partial L}{\partial \mathbf{r}}\right] \cdot \delta\mathbf{r}\right\} \mathrm{d}t + \left[\frac{\partial L}{\partial \mathbf{v}} \cdot \delta\mathbf{r}\right]_{t_0}^{t} = 0. \tag{20}$$

So if we assume that the variation vanishes at the start and end times, then since the variation $\delta\mathbf{r}$ is arbitrary, the equations of motion are given by the Euler–Lagrange equations, here taken with respect to particle a:

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \mathbf{v}_a}\right) - \frac{\partial L}{\partial \mathbf{r}_a} = 0. \tag{21}$$

We have somewhat laboured the point here because it is important to understand the assumptions we have made by employing the Euler–Lagrange equations to derive the equations of motion. The first is that in using Eq. (21) we are not explicitly considering the discreteness of the time integral. So when we refer below to "exact conservation" (e.g. of energy and momentum) we mean "solely governed by errors in the time integration scheme".[4]

The other, more critical, assumption we have made in employing (21) is that the Lagrangian is differentiable. This means that we have explicitly excluded the possibility of discontinuous solutions to the equations of motion. What this means in practice is that any discontinuities present in the system (for example in the initial conditions) require careful treatment – for example by adding dissipative terms that smooth discontinuities to a resolvable scale (i.e., a few h) such that they can be treated as no longer discontinuous. We give some practical examples of this in Section 6.3. A better way would be to account for the neglected surface integral terms directly in the Lagrangian e.g. [48], though it is not clear how one would go about doing so in an SPH context.

3.3. Equations of motion

All that remains in order to derive the equations of motion is to compute the derivatives in (21) by writing the terms in the Lagrangian as a function of the particle coordinates and velocities. From (16) we have

$$\frac{\partial L}{\partial \mathbf{v}_a} = m_a \mathbf{v}_a; \quad \frac{\partial L}{\partial \mathbf{r}_a} = -\sum_b m_b \left.\frac{\partial u_b}{\partial \rho_b}\right|_s \frac{\partial \rho_b}{\partial \mathbf{r}_a}, \tag{22}$$

the latter since u is a function of ρ and s, and we assume that the entropy s is constant (i.e., no dissipation). Note that the former gives the canonical momenta of the system ($\mathbf{p} \equiv \partial L/\partial \mathbf{v}$). This step is straightforward for hydrodynamics, but it can also be used to derive the conservative momentum variable in the case of more complicated physics – for example in relativistic SPH e.g. [67,92].

3.3.1. Thermodynamics of the fluid

From the first law of thermodynamics we have

$$\mathrm{d}U = T\mathrm{d}S - P\mathrm{d}V, \tag{23}$$

where $\delta Q \equiv T\mathrm{d}S$ is the heat added to the system (per unit volume) and $\delta W \equiv P\mathrm{d}V$ is the work done by expansion and compression of the fluid. We do not compute the volume directly in SPH, but instead we can use the volume estimate given by $V = m/\rho$ and thus the change in volume given by $\mathrm{d}V = -m/\rho^2 \mathrm{d}\rho$. Using quantities per unit mass instead of per unit volume (i.e., $\mathrm{d}u$ instead of $\mathrm{d}U$), we have

$$\mathrm{d}u = T\mathrm{d}s + \frac{P}{\rho^2}\mathrm{d}\rho, \tag{24}$$

such that at constant entropy, the change in thermal energy is given by

$$\left.\frac{\partial u_b}{\partial \rho_b}\right|_s = \frac{P}{\rho^2}. \tag{25}$$

3.3.2. The density gradient

So far we have not made any explicit reference to SPH or kernel interpolation. This arises because of the spatial derivative of the density (Eq. 22), that we obtain by differentiating our density estimate (Eq. 10). Noting that we are taking the gradient of the density estimate at particle b with respect to the coordinates of particle a, this is given by

$$\frac{\partial \rho_b}{\partial \mathbf{r}_a} = \frac{1}{\Omega_b}\sum_c m_c \frac{\partial W_{bc}(h_b)}{\partial \mathbf{r}_a}(\delta_{ba} - \delta_{ca}), \tag{26}$$

where $W_{bc}(h_b) \equiv W(\mathbf{r}_b - \mathbf{r}_c, h_b)$, δ_{ba} is a Dirac delta function referring to the particle indices and we have assumed that the smoothing length is itself a function of density [i.e., $h = h(\rho)$], giving a term accounting for the gradient of the smoothing length given by

$$\Omega_a \equiv \left[1 - \frac{\partial h_a}{\partial \rho_a}\sum_b m_b \frac{\partial W_{ab}(h_a)}{\partial h_a}\right], \tag{27}$$

where for the standard $h - \rho$ relationship (Eq. 10) the derivative is given by

$$\frac{\partial h}{\partial \rho} = -\frac{h}{\rho \mathrm{d}}, \tag{28}$$

where d is the number of spatial dimensions.

[4] It should be noted that it is quite possible to also derive the time integration scheme from a Lagrangian – for example, Monaghan [64] gives the appropriate Lagrangian for the symplectic and time-reversible leapfrog scheme.

3.3.3. Equations of motion

Using (26) and (25) in (22) we have

$$\frac{\partial L}{\partial \mathbf{r}_a} = -\sum_b m_b \frac{P_b}{\Omega_b \rho_b^2} \sum_c m_c \frac{\partial W_{bc}(h_b)}{\partial \mathbf{r}_a}(\delta_{ba} - \delta_{ca}), \tag{29}$$

which, upon simplification, gives the equations of motion from the Euler–Lagrange equations in the form

$$\frac{\mathrm{d}\mathbf{v}_a}{dt} = -\sum_b m_b \left[\frac{P_a}{\Omega_a \rho_a^2}\frac{\partial W_{ab}(h_a)}{\partial \mathbf{r}_a} + \frac{P_b}{\Omega_b \rho_b^2}\frac{\partial W_{ab}(h_b)}{\partial \mathbf{r}_a}\right]. \tag{30}$$

For a constant smoothing length these equations simplify to the standard SPH expression [59]

$$\frac{\mathrm{d}\mathbf{v}_a}{\mathrm{d}t} = -\sum_b m_b \left(\frac{P_a}{\rho_a^2} + \frac{P_b}{\rho_b^2}\right)\nabla_a W_{ab}. \tag{31}$$

3.3.4. Conservation properties

Although we cannot yet provide interpretation to the equations of motion so derived, we can note a number of interesting properties. The first is that the total linear momentum is conserved exactly, since

$$\frac{\mathrm{d}}{\mathrm{d}t}\sum_a m_a \mathbf{v}_a = \sum_a m_a \frac{\mathrm{d}\mathbf{v}_a}{\mathrm{d}t} = -\sum_a \sum_b m_a m_b \left(\frac{P_a}{\rho_a^2} + \frac{P_b}{\rho_b^2}\right)\nabla_a W_{ab} = 0, \tag{32}$$

where the double summation is zero because of the antisymmetry in the kernel gradient[5] (the reader may verify that this is also true for Eq. (30)). Secondly, the total angular momentum is also exactly conserved, since

$$\frac{\mathrm{d}}{\mathrm{d}t}\sum_a \mathbf{r}_a \times m_a \mathbf{v}_a = \sum_a m_a \left(\mathbf{r}_a \times \frac{\mathrm{d}\mathbf{v}_a}{\mathrm{d}t}\right) = -\sum_a \sum_b m_a m_b \left(\frac{P_a}{\rho_a^2} + \frac{P_b}{\rho_b^2}\right)\mathbf{r}_a \times (\mathbf{r}_a - \mathbf{r}_b)\widetilde{F}_{ab} = 0, \tag{33}$$

where for convenience we have written the gradient of the kernel in the form $\nabla_a W_{ab} = \mathbf{r}_{ab}\widetilde{F}_{ab}$, and the last term is zero again because of the antisymmetry in the double summation [since $(\mathbf{r}_a \times \mathbf{r}_b) = -(\mathbf{r}_b \times \mathbf{r}_a)$].

The above conservation properties follow directly from the symmetries in the original Lagrangian and (by extension) the SPH density estimate (Eq. 10) – linear momentum conservation because the Lagrangian and density estimate are invariant to translations, and angular momentum conservation because they are invariant to rotations of the particle coordinates. This is important in thinking about possible modifications to the SPH scheme (for example using non-spherical kernels would result in the non-conservation of angular momentum because the density estimate would no longer be invariant to rotations).

Finally, we note that although the equations of motion depend only on the relative positions of the particles, they *do* depend on the absolute value of the pressure. That is, Eqs. (30) and (31) contain a force between the particles that is non-zero even when the pressure is constant. We discuss the importance of – and problems associated with – this 'spurious' force in Section 5.

3.4. Energy equation

The remaining part of the (dissipationless) SPH algorithm – the energy equation – can also be derived from the Hamiltonian dynamics. Here we have the choice of evolving either the thermal energy u, the total specific energy $e = \frac{1}{2}v^2 + u$ or an entropy variable $K = P/\rho^\gamma$. Equations for each of these can be derived, as given below. It is important to note, however, that – provided the equations are derived from the Lagrangian formulation – there is no difference in SPH between evolving any of these variables, *except* due to the timestepping algorithm. This is rather different to the situation in an Eulerian code where the finite differencing of the advection terms mean that writing the equations in conservative form (i.e., using e) differs more substantially from evolving u or K.

3.4.1. Internal energy

The evolution equation for the internal energy (in the absence of dissipation), from Eq. (25), is given by

$$\frac{\mathrm{d}u_a}{\mathrm{d}t} = \frac{P_a}{\rho_a^2}\frac{\mathrm{d}\rho_a}{\mathrm{d}t}. \tag{34}$$

Taking the time derivative of the density sum (Eq. 10), we obtain an evolution equation for u of the form

$$\frac{\mathrm{d}u_a}{\mathrm{d}t} = \frac{P_a}{\Omega_a \rho_a^2}\sum_b m_b(\mathbf{v}_a - \mathbf{v}_b)\cdot\nabla_a W_{ab}(h_a). \tag{35}$$

[5] This can be easily seen by swapping the summation indices a and b in the double sum and adding half of the original term to half of the rearranged term, giving zero.

3.4.2. Total energy

The conserved (total) energy is found from the Lagrangian via the Hamiltonian

$$H = \sum_a \mathbf{v}_a \cdot \frac{\partial L}{\partial \mathbf{v}_a} - L = \sum_a m_a \left(\frac{1}{2} v_a^2 + u_a \right), \tag{36}$$

which is simply the total energy of the SPH particles, E, since the Lagrangian does not explicitly depend on the time. Taking the (Lagrangian) time derivative of (36), we have

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \sum_a m_a \left(\mathbf{v}_a \cdot \frac{\mathrm{d}\mathbf{v}_a}{\mathrm{d}t} + \frac{\mathrm{d}u_a}{\mathrm{d}t} \right). \tag{37}$$

Substituting (30) and (35) and rearranging we find

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \sum_a m_a \frac{\mathrm{d}e_a}{\mathrm{d}t} = -\sum_a \sum_b m_a m_b \left[\frac{P_a}{\Omega_a \rho_a^2} \mathbf{v}_b \cdot \nabla_a W_{ab}(h_a) + \frac{P_b}{\Omega_b \rho_b^2} \mathbf{v}_a \cdot \nabla_a W_{ab}(h_b) \right] = 0. \tag{38}$$

This equation shows that the total energy is also exactly conserved by the SPH scheme (where the double sum is zero again because of the antisymmetry with respect to the particle index, similar to the conservation of linear momentum discussed above). The conservation of total energy is a consequence of the symmetry of the Lagrangian (16) with respect to time as well as invariance under time translations. Eq. (38) also shows that the dissipationless evolution equation for the specific energy e is given by

$$\frac{\mathrm{d}e_a}{\mathrm{d}t} = -\sum_b m_b \left[\frac{P_a}{\Omega_a \rho_a^2} \mathbf{v}_b \cdot \nabla_a W_{ab}(h_a) + \frac{P_b}{\Omega_b \rho_b^2} \mathbf{v}_a \cdot \nabla_a W_{ab}(h_b) \right]. \tag{39}$$

3.4.3. Entropy

For the specific case of an ideal gas equation of state, where

$$P = K(s)\rho^\gamma, \tag{40}$$

it is possible to use the function $K(s)$ as the evolved variable [99], where the evolution of K is given by

$$\frac{\mathrm{d}K}{\mathrm{d}t} = \frac{\gamma - 1}{\rho^{\gamma-1}} \left(\frac{\mathrm{d}u}{\mathrm{d}t} - \frac{P}{\rho^2} \frac{\mathrm{d}\rho}{\mathrm{d}t} \right) = \frac{\gamma - 1}{\rho^{\gamma-1}} \left(\frac{\mathrm{d}u}{\mathrm{d}t} \right)_{diss}. \tag{41}$$

The thermal energy is then evaluated using

$$u = \frac{K}{\gamma - 1} \rho^{\gamma-1}. \tag{42}$$

Since $\mathrm{d}K/\mathrm{d}t = 0$ in the absence of dissipation, using K has the advantage that the evolution is independent of the time-integration algorithm. The disadvantage is that it is more difficult to apply to non-ideal equations of state. This is sometimes referred to as the 'entropy-conserving' form of SPH [after 99] – which is somewhat misleading since the entropy per particle is also exactly conserved if (35) or (39) are used provided the smoothing length gradient terms are correctly accounted for (i.e., $\mathrm{d}u/\mathrm{d}t - P/\rho^2 \mathrm{d}\rho/\mathrm{d}t = 0$), apart from minor differences arising from the timestepping scheme. So the term 'entropy-conserving' more correctly refers to the correct accounting of smoothing length gradient terms and a consistent formulation of the energy equation than whether or not an entropy variable is evolved.

3.5. Summary

In summary, our full system of equations for ρ, $\mathbf{v}$ and u is given by

$$\rho_a = \sum_b m_b W(\mathbf{r}_a - \mathbf{r}_b, h_a); \quad h = h(\rho), \tag{43}$$

$$\frac{\mathrm{d}\mathbf{v}_a}{\mathrm{d}t} = -\sum_b m_b \left[\frac{P_a}{\Omega_a \rho_a^2} \nabla_a W_{ab}(h_a) + \frac{P_b}{\Omega_b \rho_b^2} \nabla_a W_{ab}(h_b) \right], \tag{44}$$

$$\frac{\mathrm{d}u_a}{\mathrm{d}t} = \frac{P_a}{\Omega_a \rho_a^2} \sum_b m_b (\mathbf{v}_a - \mathbf{v}_b) \cdot \nabla_a W_{ab}(h_a), \tag{45}$$

where in place of (45) we could equivalently use either (39) or (41). The reader will note that so far we have not even mentioned the continuum equations of hydrodynamics – we have merely specified the physics that goes into the Lagrangian (Eq. 16), the thermodynamics of the fluid (Eq. 24) and the manner in which the density is calculated (Eq. 10) and with these have directly derived the discrete equations of the Hamiltonian system. In order to interpret them we need to know how to translate our SPH Eqs. (43)–(45) into their continuum equivalents.

Before doing so, it is worth recapping briefly the assumptions we have made in arriving at (43)–(45) from the Lagrangian (Eq. 16). These are

(i) That the time integration and thus time derivatives, $\mathrm{d}/\mathrm{d}t$, are computed exactly (though this assumption can in principle be relaxed);
(ii) That the Lagrangian, and by implication the density and thermal energies are differentiable;
(iii) That there is no change in entropy, such that the first law of thermodynamics $\mathrm{d}u = P\mathrm{d}V$ is satisfied, and that the change in particle volume is given by $\mathrm{d}V = -m/\rho^2\mathrm{d}\rho$.

The second and third assumptions in particular come into play in dealing with shocks and other kinds of discontinuities, which will we discuss further in Section 6.3.

3.6. Alternative formulations

Within the constraints of a Hamiltonian SPH formulation, it is clear that there is only a very limited freedom to change the algorithm without breaking some of the conservation properties of the scheme. So there are only two basic ways to change the (dissipationless) algorithm that are consistent with the Hamiltonian approach (other than changing the first law of thermodynamics): (i) change the way the density is calculated or (ii) introduce additional physical terms, and associated constraints, into the Lagrangian. Examples of the former are the consistent formulation of variable smoothing length terms – requiring the iterative solution of h and ρ in the density sum (Eq. 10) – by Springel and Hernquist [99] and Monaghan [62], and an incorporation of boundary correction forces [51]. Examples of the latter include consistent derivations of relativistic SPH [67,92], adaptive gravitational force softening [89], sub-resolution turbulence models [62,63] and MHD [81,87].

4. Kernel interpolation theory and SPH derivatives

The usual way of introducing SPH is to start with a formal discussion of kernel interpolation theory. We have taken a rather different approach in this review and will introduce this theory primarily only to *interpret* the equations that we have derived from the discrete Lagrangian, and also as a way of discussing how to go about introducing additional physics. However, we will *not* use the linear error properties of the interpolation scheme to define the method – apart from our construction of the density estimate discussed above. The reason is that focussing on linear errors over the Hamiltonian properties of SPH misses some of the subtle but important non-linear behaviour that makes SPH work in practice, which we will come to discuss. However, first let us proceed:

4.1. Kernel interpolation: The basics

Kernel interpolation theory starts with the identity

$$A(\mathbf{r}) = \int A(\mathbf{r}')\delta(\mathbf{r}-\mathbf{r}')\mathrm{d}\mathbf{r}', \tag{46}$$

where A is an arbitrary scalar variable and δ refers to the Dirac delta function. This integral is then approximated by replacing the delta function with a smoothing kernel W with finite width h, i.e.,

$$A(\mathbf{r}) = \int A(\mathbf{r}')W(\mathbf{r}-\mathbf{r}',h)\mathrm{d}\mathbf{r}' + O(h^2). \tag{47}$$

where W has the property

$$\lim_{h\to 0} W(\mathbf{r}-\mathbf{r}',h) = \delta(\mathbf{r}-\mathbf{r}'), \tag{48}$$

and normalisation $\int_V W\mathrm{d}V' = 1$, as we have already discussed in Section 2.2. Finally the integral interpolant (Eq. 47) is discretised onto a finite set of interpolation points (the particles) by replacing the integral by a summation and the mass element ρdV with the particle mass m, i.e.,

$$\begin{aligned}\langle A(\mathbf{r})\rangle &= \int \frac{A(\mathbf{r}')}{\rho(\mathbf{r}')}W(\mathbf{r}-\mathbf{r}',h)\rho(\mathbf{r}')\mathrm{d}\mathbf{r}',\\ &\approx \sum_{b=1}^{N_{neigh}} m_b\frac{A_b}{\rho_b}W(\mathbf{r}-\mathbf{r}_b,h),\end{aligned} \tag{49}$$

This 'summation interpolant' is the basis of all SPH formalisms. The reader will note that choosing $A = \rho$ results in the SPH density estimate (Eq. 10). In this paper we have argued that the density estimate (10) is in some sense more fundamental than the summation interpolant (49), since the equations of motion can be derived without reference to (49). On the other hand, the summation interpolant gives a general way of a interpolating a quantity at any point in space $A(\mathbf{r})$ from quantities defined solely on the particles themselves (i.e., A_b, ρ_b,m_b),[6] and in turn to a general way of formulating SPH equations. In particular, gradient terms may be straightforwardly calculated by taking the derivative of (49), giving

$$\nabla A(\mathbf{r}) = \frac{\partial}{\partial \mathbf{r}} \int \frac{A(\mathbf{r}')}{\rho(\mathbf{r}')} W(\mathbf{r}-\mathbf{r}', h)\rho(\mathbf{r}')d\mathbf{r}' + O(h^2), \tag{50}$$

$$\approx \sum_b m_b \frac{A_b}{\rho_b} \nabla W(\mathbf{r}-\mathbf{r}_b, h). \tag{51}$$

For vector quantities the expressions are similar, simply replacing A with $\mathbf{A}$ in (49), giving

$$\mathbf{A}(\mathbf{r}) \approx \sum_b m_b \frac{\mathbf{A}_b}{\rho_b} W(\mathbf{r}-\mathbf{r}_b, h), \tag{52}$$

$$\nabla \cdot \mathbf{A}(\mathbf{r}) \approx \sum_b m_b \frac{\mathbf{A}_b}{\rho_b} \cdot \nabla W(\mathbf{r}-\mathbf{r}_b, h), \tag{53}$$

$$\nabla \times \mathbf{A}(\mathbf{r}) \approx -\sum_b m_b \frac{\mathbf{A}_b}{\rho_b} \times \nabla W(\mathbf{r}-\mathbf{r}_b, h), \tag{54}$$

$$\nabla^j A^i(\mathbf{r}) \approx \sum_b m_b \frac{A^i_b}{\rho_b} \nabla^j W(\mathbf{r}-\mathbf{r}_b, h). \tag{55}$$

The problem is that using these expressions 'as is' in general leads to quite poor gradient estimates, and we can do better by considering the errors in the above approximations. However, the basic interpolants given above give us a general way of interpreting SPH expressions such as those derived in Section 3.

4.2. Interpretation of the Hamiltonian SPH equations

We are now able to provide interpretation to our Hamiltonian-SPH Eqs. (43)–(45) derived in Section 3 using the basic identities (51)–(55). We begin with the density summation (43). Taking the time derivative, we have

$$\frac{\mathrm{d}\rho_a}{\mathrm{d}t} = \frac{1}{\Omega_a} \sum_b m_b (\mathbf{v}_a - \mathbf{v}_b) \cdot \nabla_a W_{ab}(h_a), \tag{56}$$

which for a constant smoothing length simplifies to

$$\frac{\mathrm{d}\rho_a}{\mathrm{d}t} = \sum_b m_b (\mathbf{v}_a - \mathbf{v}_b) \cdot \nabla_a W_{ab}(h). \tag{57}$$

Using (53) we can translate each of the terms according to

$$\frac{\mathrm{d}\rho_a}{\mathrm{d}t} = \mathbf{v}_a \cdot \sum_b \frac{m_b}{\rho_b} \rho_b \nabla_a W_{ab} - \sum_b \frac{m_b}{\rho_b} (\rho_b \mathbf{v}_b) \cdot \nabla_a W_{ab} \approx \mathbf{v}_a \cdot \nabla\rho - \nabla \cdot (\rho \mathbf{v}) \approx -\rho_a (\nabla \cdot \mathbf{v})_a. \tag{58}$$

So remarkably (57) (and by inference, 56) – which we obtained simply by taking the time derivative of the density sum – represents a (particular) SPH discretisation of the continuity equation. Indeed, the density summation is therefore an exact, time-independent, solution to the SPH continuity equation. This should not be particularly surprising, since the continuity equation derives from the conservation of mass, which is self-evidently enforced on the particles since m is fixed.

Our force term (44), assuming a constant h (Eq. 31) can be translated according to

$$-\sum_b m_b \left(\frac{P_a}{\rho_a^2} + \frac{P_b}{\rho_b^2}\right) \nabla_a W_{ab} \approx -\frac{P}{\rho^2} \nabla\rho - \nabla\left(\frac{P}{\rho}\right) \approx -\frac{\nabla P}{\rho}, \tag{59}$$

where we have used the basic identity (51) to translate the two terms into their continuum equivalents.

Finally, the thermal energy Eq. (45) can be translated, from (34) and (58), giving

$$\frac{\mathrm{d}u}{\mathrm{d}t} = -\frac{P}{\rho} \nabla \cdot \mathbf{v}. \tag{60}$$

[6] Eq. 49 naturally also forms the basis for visualisation of SPH simulations, where one wishes to reconstruct the field in all of the spatial volume given quantities defined on particles: This is the approach implemented in SPLASH [79].

In other words, it is evident that the Lagrangian has indeed given us valid discretisations for the equations of hydrodynamics, and therefore that the equations of our Hamiltonian system (43)–(45) do indeed solve these equations in discrete form – a remarkable achievement given the relatively few assumptions that were made. Yet the discretisations we derived are clearly not the basic ones arising from kernel interpolation theory (51)–(54). In order to examine the errors in these discretisations we need to understand the errors arising from the basic gradient operators, and how more general derivative operators can be constructed.

4.3. Errors

The errors introduced by the approximation (47) are similar to those in the density estimate (Section 2.5). That is, if we expand $A(\mathbf{r}')$ in a Taylor series about $\mathbf{r}$ [5,59], we find

$$\langle A(\mathbf{r})\rangle = \int \left[A(\mathbf{r}) + (\mathbf{r}'-\mathbf{r})^{\alpha}\frac{\partial A}{\partial \mathbf{r}^{\alpha}} + \frac{1}{2}(\mathbf{r}'-\mathbf{r})^{\alpha}(\mathbf{r}'-\mathbf{r})^{\beta}\frac{\partial^2 A}{\partial \mathbf{r}^{\alpha}\partial \mathbf{r}^{\beta}} + \mathcal{O}((\mathbf{r}-\mathbf{r}')^3)\right] W(\mathbf{r}-\mathbf{r}', h)\mathrm{d}\mathbf{r}', \tag{61}$$

such that for symmetric $W \equiv W(|\mathbf{r}-\mathbf{r}'|, h)$ and normalised ($\int W \mathrm{d}V' = 1$) kernels we have

$$\langle A(\mathbf{r})\rangle = A(\mathbf{r}) + \frac{1}{2}\frac{\partial^2 A}{\partial \mathbf{r}^{\alpha}\partial \mathbf{r}^{\beta}}\int (\mathbf{r}'-\mathbf{r})^{\alpha}(\mathbf{r}'-\mathbf{r})^{\beta} W(|\mathbf{r}-\mathbf{r}'|, h)\mathrm{d}\mathbf{r}' + \mathcal{O}[(\mathbf{r}'-\mathbf{r})^4], \tag{62}$$

giving an interpolation that is second order accurate [$\mathcal{O}(h^2)$] unless higher order kernels are used (see Section 2.5). However, the errors in the discrete version (Eq. 49) are not identical, since they depend on the degree to which the discrete summations approximate the integrals – specifically on the degree to which the discrete normalisation conditions are satisfied. Following Price [78] we can perform a similar analysis on the summation interpolant (49, here assumed to be computed on particle a) by expanding A_b in a Taylor series around $\mathbf{r}_a$, giving

$$\langle A_a\rangle = \sum_b m_b \frac{A_b}{\rho_b} W_{ab} = A_a \sum_b \frac{m_b}{\rho_b} W_{ab} + \nabla A_a \cdot \sum_b \frac{m_b}{\rho_b}(\mathbf{r}_b - \mathbf{r}_a) W_{ab} + \mathcal{O}(h^2). \tag{63}$$

This shows that in practice, the interpolation will only truly be second order accurate if the conditions

$$\sum_b \frac{m_b}{\rho_b} W_{ab} \approx 1; \quad \text{and} \quad \sum_b \frac{m_b}{\rho_b}(\mathbf{r}_b - \mathbf{r}_a) W_{ab} \approx 0, \tag{64}$$

hold. The degree to which this is true depends strongly on the particle distribution within the kernel radius, and the properties of the kernel when a finite number of neighbours are employed – in particular, the ratio of smoothing length to particle spacing Δp. Fig. 3 shows the first condition computed for the B-spline kernels and the Gaussian as a function of $h/\Delta p$ in 1D (solid/black lines), showing that in general the above conditions are maintained very well, provided the particles are regular. Thus, maintaining a regular particle arrangement, together with an appropriate choice of $h/\Delta p$, can be very important in obtaining accurate results with SPH. Used another way, the conditions (64) are essentially the criteria for a 'good density kernel' – that is, a good density kernel is one which satisfies these conditions well given the typical particle distributions encountered in SPH. Note that the second of these is much easier to satisfy than the first, since it requires only a reasonably

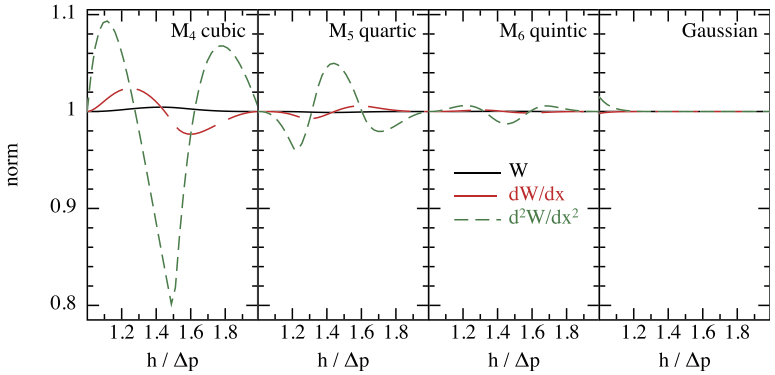

Fig. 3. Accuracy with which the normalisation conditions on the kernel, the kernel gradient and the kernel second derivative are computed with fixed h on a one-dimensional line of particles with the M_4–M_6 kernels (from left) compared to the Gaussian (right), as a function of the ratio of smoothing length to particle spacing $h/\Delta p$. Bell-shaped kernels with compact support lead to excellent density estimates (solid/black lines), reasonable gradient estimates (long dashed/red lines) but poor second derivative estimates (short dashed/green lines) when a finite number of neighbours are employed. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

symmetric particle distribution to be satisfied. In general reasonable fulfilment of the first condition amounts to the conditions (1)–(3) on the density kernel described in Section 2.2.

The errors resulting from the gradient interpolation (50) may be estimated in a similar manner by again expanding $A(\mathbf{r}')$ in a Taylor series about $\mathbf{r}$, giving

$$\begin{aligned}\nabla A(\mathbf{r}) &= \int \left[A(\mathbf{r}) + (\mathbf{r}'-\mathbf{r})^\alpha \frac{\partial A}{\partial \mathbf{r}^\alpha} + \frac{1}{2}(\mathbf{r}'-\mathbf{r})^\beta(\mathbf{r}'-\mathbf{r})^\gamma \frac{\partial^2 A}{\partial \mathbf{r}^\beta \partial \mathbf{r}^\gamma} + \mathcal{O}[(\mathbf{r}-\mathbf{r}')^3]\right]\nabla W(|\mathbf{r}-\mathbf{r}'|,h)\mathrm{d}\mathbf{r}', \\ &= A(\mathbf{r})\int \nabla W \mathrm{d}\mathbf{r}' + \frac{\partial A}{\partial \mathbf{r}^\alpha}\int (\mathbf{r}'-\mathbf{r})^\alpha \nabla W \mathrm{d}\mathbf{r}' + \frac{1}{2}\frac{\partial^2 A}{\partial \mathbf{r}^\beta \partial \mathbf{r}^\gamma}\int (\mathbf{r}'-\mathbf{r})^\beta(\mathbf{r}'-\mathbf{r})^\gamma \nabla W \mathrm{d}\mathbf{r}' + \mathcal{O}[(\mathbf{r}'-\mathbf{r})^3], \\ &= \nabla A(\mathbf{r}) + \frac{1}{2}\frac{\partial^2 A}{\partial \mathbf{r}^\beta \partial \mathbf{r}^\gamma}\int (\mathbf{r}'-\mathbf{r})^\beta(\mathbf{r}'-\mathbf{r})^\gamma \nabla W(r)\mathrm{d}\mathbf{r}' + \mathcal{O}[(\mathbf{r}'-\mathbf{r})^3],\end{aligned} \tag{65}$$

where we have used the fact that $\int \nabla W \mathrm{d}\mathbf{r}' = 0$ for even kernels, whilst the second term integrates to unity for even kernels satisfying the normalisation condition (3). The resulting errors in the integral interpolant for the gradient are therefore also of $\mathcal{O}(h^2)$. As previously, the errors in the discrete version (51) can be found by expanding A_b in a Taylor series around $\mathbf{r}_a$, giving

$$\begin{aligned}\langle \nabla A_a\rangle &= \sum_b m_b \frac{A_b}{\rho_b}\nabla_a W_{ab}, \\ &= A_a \sum_b \frac{m_b}{\rho_b}\nabla_a W_{ab} + \frac{\partial A_a}{\partial \mathbf{r}^\alpha}\sum_b \frac{m_b}{\rho_b}(\mathbf{r}_b-\mathbf{r}_a)^\alpha \nabla_a W_{ab} + \mathcal{O}(h^2),\end{aligned} \tag{66}$$

where the summations represent SPH approximations to the integrals in the second line of (65). So the gradient errors resulting from (51) would in principle be similarly governed by the extent to which these discrete summations approximate the integrals, i.e., how well the conditions

$$\sum_b \frac{m_b}{\rho_b}\nabla_a W_{ab} \approx \mathbf{0}; \quad \text{and} \quad \sum_b \frac{m_b}{\rho_b}(\mathbf{r}_b-\mathbf{r}_a)^\alpha \nabla_a^\beta W_{ab} \approx \delta^{\alpha\beta}; \tag{67}$$

hold. The latter term is shown for the B-spline kernels by the long-dashed/red lines in Fig. 3. The difference is that for gradients we can explicitly use the error terms to construct more accurate gradient operators, and we do this below.

4.4. First derivatives

From (66) we immediately see that a straightforward improvement to the gradient estimate (51) can be obtained by a simple subtraction of the first error term (i.e., the term in (66) that is present even in the case of a constant function), giving e.g. [59]

$$\langle \nabla A_a\rangle = \sum_b m_b \frac{(A_b - A_a)}{\rho_b}\nabla_a W_{ab}, \tag{68}$$

which, interpreting each term according to (51), is an SPH estimate of

$$\nabla A \approx \langle \nabla A\rangle - A\langle \nabla 1\rangle. \tag{69}$$

Since the first error term in (66) is removed, the interpolation is exact for constant functions and indeed this is obvious from the form of (68). The interpolation can be made exact for linear functions via a matrix inversion of the second error term in (66), i.e., solving

$$\chi^{\alpha\beta}\frac{\partial A_a}{\partial \mathbf{r}^\alpha} = \sum_b \frac{m_b}{\rho_b}(A_b - A_a)\nabla^\beta W_{ab}, \quad \chi^{\alpha\beta} \equiv \sum_b \frac{m_b}{\rho_b}(\mathbf{r}_b-\mathbf{r}_a)^\alpha \nabla^\beta W_{ab}. \tag{70}$$

where $\nabla^\beta \equiv \partial/\partial \mathbf{r}^\beta$. This normalisation is somewhat cumbersome in practice, since χ is a matrix quantity, requiring considerable extra storage (in three dimensions this means storing $3\times 3 = 9$ extra quantities for each particle) and also since calculation of this term requires prior knowledge of the density.

A similar interpolant for the gradient follows by using

$$\nabla A \approx \frac{1}{\rho}[\langle \nabla(\rho A)\rangle - A\langle \nabla \rho\rangle] = \frac{1}{\rho_a}\sum_b m_b(A_b - A_a)\nabla_a W_{ab}, \tag{71}$$

which again is exact for a constant A. Expanding A_b in a Taylor series, we see that in this case the interpolation of a linear function can be made exact by solving

$$\chi^{\alpha\beta}\frac{\partial A_a}{\partial \mathbf{r}^\alpha} = \sum_b m_b(A_b - A_a)\nabla^\beta W_{ab}, \quad \chi^{\alpha\beta} \equiv \sum_b m_b(\mathbf{r}_b-\mathbf{r}_a)^\alpha \nabla^\beta W_{ab}. \tag{72}$$

which has some advantages over (70) in that it can be computed without prior knowledge of the density.

However, the gradient operator we derived in the equations of motion (31) does not correspond to any of the above possibilities. This operator is given by

$$\nabla A \approx \rho\left[\frac{A}{\rho^2}\langle\nabla\rho\rangle + \left\langle\nabla\left(\frac{A}{\rho}\right)\right\rangle\right] = \rho_a\sum_b m_b\left(\frac{A_a}{\rho_a^2}+\frac{A_b}{\rho_b^2}\right)\nabla_a W_{ab}. \tag{73}$$

Expanding A_b in a Taylor series about $\mathbf{r}_a$, we have

$$\rho_a A_a\sum_b m_b\left(\frac{1}{\rho_a^2}+\frac{1}{\rho_b^2}\right)\nabla_a W_{ab} + \frac{\partial A_a}{\partial \mathbf{r}^\alpha}\rho_a\sum_b\frac{m_b}{\rho_b^2}(\mathbf{r}_b-\mathbf{r}_a)^\alpha\nabla_a W_{ab} + \mathcal{O}(h^2), \tag{74}$$

from which we see that for a constant function the error in (73) is governed by the extent to which

$$\sum_b m_b\left(\frac{1}{\rho_a^2}+\frac{1}{\rho_b^2}\right)\nabla_a W_{ab} \approx \mathbf{0}. \tag{75}$$

Although a simple subtraction of the first term in (74) from the original expression (73) eliminates this error, this would not give the form we derived in Section 3.3. Indeed, retaining the exact conservation of momentum requires that such error terms are not eliminated, the consequences of which we will discuss in Section 5.

4.5. Generalised first derivative operators

Finally, an infinite variety of gradient operators – of two basic types – can be constructed by noting that

$$\nabla A = \frac{1}{\phi}[\nabla(\phi A) - A\nabla\phi] \approx \sum_b\frac{m_b}{\rho_b}\frac{\phi_b}{\phi_a}(A_b-A_a)\nabla_a W_{ab}, \tag{76}$$

and

$$\nabla A = \phi\left[\frac{A}{\phi^2}\nabla\phi + \nabla\left(\frac{A}{\phi}\right)\right] \approx \sum_b\frac{m_b}{\rho_b}\left(\frac{\phi_b}{\phi_a}A_a+\frac{\phi_a}{\phi_b}A_b\right)\nabla_a W_{ab}, \tag{77}$$

where ϕ is *any* arbitrary, differentiable scalar quantity defined on the particles. Indeed, for a given ϕ, the pair of operators defined by (76) and (77) can be shown to form a conjugate pair[7] – and choosing one (e.g. for the density/thermal energy evolution) tends to lead to the other (e.g. in the equations of motion). For example, (68) and (71) correspond to using $\phi = 1$ and $\phi = \rho$, respectively in (76), whilst (73) correponds to using $\phi = \rho$ in (77) and arises in the equations of motion because it is the conjugate operator to (71) that arises in the density gradient (57).

Various 'alternative' formulations of the SPH equations have been proposed that correspond to a particular choice of ϕ e.g. [40,55,91]. For example Hernquist and Katz [40] suggested using an acceleration equation of the form

$$\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} = -\sum_b m_b\left(2\frac{\sqrt{P_a P_b}}{\rho_a\rho_b}\right)\nabla_a W_{ab}, \tag{78}$$

corresponding to the symmetric operator (Eq. 77) with $\phi = \sqrt{P}/\rho$. It can be readily shown that ensuring exact conservation of energy with such a formulation would require using the conjugate operator (i.e., Eq. 76 with $\phi = \sqrt{P}/\rho$) in the thermal energy equation (although it should be noted that simultaneous exact conservation of energy and entropy is not possible with *any* alternative formulation).

4.6. First derivatives of vector quantities

All of the above discussion applies also to vector derivatives, simply using (53) or (54) in place of (51) and likewise resulting in two basic operators for each type of derivative, given by

$$\langle\nabla\cdot\mathbf{A}\rangle_a \approx \sum_b\frac{m_b}{\rho_b}\frac{\phi_b}{\phi_a}(\mathbf{A}_b-\mathbf{A}_a)\cdot\nabla_a W_{ab}; \quad \langle\nabla\times\mathbf{A}\rangle_a \approx -\sum_b\frac{m_b}{\rho_b}\frac{\phi_b}{\phi_a}(\mathbf{A}_b-\mathbf{A}_a)\times\nabla_a W_{ab} \tag{79}$$

and

$$\langle\nabla\cdot\mathbf{A}\rangle_a \approx \sum_b\frac{m_b}{\rho_b}\left(\frac{\phi_b}{\phi_a}\mathbf{A}_a+\frac{\phi_a}{\phi_b}\mathbf{A}_b\right)\cdot\nabla_a W_{ab}; \quad \langle\nabla\times\mathbf{A}\rangle_a \approx -\sum_b\frac{m_b}{\rho_b}\left(\frac{\phi_b}{\phi_a}\mathbf{A}_a+\frac{\phi_a}{\phi_b}\mathbf{A}_b\right)\times\nabla_a W_{ab}, \tag{80}$$

[7] The conjugate nature of the symmetric and antisymmetric SPH gradient operators was first noted by Cummins and Rudman[21] in the context of projection schemes for SPH.

where as previously ϕ is an arbitrary (differentiable) scalar quantity. For general vector derivatives written in tensor notation the corresponding expressions are given by

$$\langle \nabla^j A^i \rangle_a \approx \sum_b \frac{m_b}{\rho_b}\frac{\phi_b}{\phi_a}\left(A^i_b - A^i_a\right)\nabla^j_a W_{ab}; \quad \text{or } \langle \nabla^j A^i \rangle_a \approx \sum_b \frac{m_b}{\rho_b}\left(\frac{\phi_b}{\phi_a}A^i_a + \frac{\phi_a}{\phi_b}A^i_b\right)\nabla^j_a W_{ab}. \tag{81}$$

These operators form the basic building blocks for formulating quite general SPH equations. Higher order operators can also be constructed for vector derivatives using matrix inversions, similar to (70) and (72).

4.7. Particle methods "the wrong way": an SPH formulation based on linear errors and exact derivatives

Based on the discussion given in Sections 4.3, 4.4, 4.5, 4.6 a straightforward approach would be to simply go ahead and discretise the continuum equations of hydrodynamics (or magnetohydrodynamics) using the most accurate gradient estimates possible. For example employing (68) the hydrodynamic equations of motion $\mathrm{d}\mathbf{v}/\mathrm{d}t = -\nabla P/\rho$ could be written in the form

$$\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} = \sum_b m_b\left(\frac{P_a - P_b}{\rho_a\rho_b}\right)\nabla_a W_{ab}, \tag{82}$$

or any of the alternatives offered by choosing ϕ appropriately in (76). Indeed such a formulation was originally examined (and discarded) by Morris [68,69] but has been recently (re-) proposed by Abel [1] (the latter author employing $\phi = 1/\rho$). We could even proceed to more accurate derivatives using Eq. (70) or Eq. (72) that are exact to linear order. Yet, these operators are clearly different from the equation of motion we derived from the Lagrangian (30), though on the basis of the linear error properties alone (Section 4.4) would seem to be a much better choice. On the other hand, it is clear that (82) does not exactly conserve linear (or angular) momentum, nor total energy.

The key difference in approach is that the above analysis tells us about the *linear* errors, whereas the Hamiltonian formulation tells us about the *non-linear* properties of the system (i.e., symmetries, conservation and constraints on the behaviour of the global system). These turn out to be crucial for long term stability and accuracy, and we will look at what this means in practice in Section 5. That said, "linear errors do not lie", so it will always be true that – provided the particle distribution is regular – formulations such as (82) or similar will be more accurate for linear or weakly non-linear problems (i.e., those run for a short time and/or not involving strong shocks) and with sufficient resolution can always be made to give accurate results.

Ideally of course, one would want *both* exact derivatives *and* exact conservation. In SPH at least, it seems one cannot have both – and to my knowledge this is yet to be convincingly demonstrated by any particle method, though it is perhaps possible using tesselation schemes.

5. Why a bad derivative leads to good derivatives: The importance of local conservation

The paradox we face is that, whilst the Lagrangian derivation gave us valid discretisations of the equations of hydrodynamics, these are not the most accurate discretisations that are possible based on a linear error analysis. Indeed, on the basis of the linear error properties, the gradient estimate derived for the acceleration equation would be a very *poor* choice of gradient operator, yet it is the only operator which respects all of the symmetry and conservation properties in the Lagrangian.

To understand what these errors mean in practice we can perform a simple thought experiment (which we will also run as a numerical experiment): Consider a distribution of particles in a closed (e.g. periodic) box at constant pressure. In particular, we will consider a uniform random particle distribution. This is the worst-case error scenario, since errors in the interpolation will essentially be Monte-Carlo ($\propto 1/\sqrt{N}$). Now consider what will happen if use one of the more accurate gradient estimators, for example (82). Clearly, since the pressure is constant, there will be no force and hence no particle motion. That is, the force vanishes when the pressure is constant *regardless of the particle distribution*. If instead we use the Hamiltonian formulation, then the pairwise force between particles is given by

$$\underbrace{-m_a m_b}_{-ve}\underbrace{\left[\frac{P}{\rho_a^2} + \frac{P}{\rho_b^2}\right]}_{+ve}\underbrace{F_{ab}}_{-ve}(\hat{\mathbf{r}}_a - \hat{\mathbf{r}}_b). \tag{83}$$

where we have written the kernel gradient using $\nabla_a W_{ab} \equiv \hat{\mathbf{r}}_{ab} F_{ab}$. Since for positive definite kernels (Fig. 2) the kernel derivative function is always negative, the overall result – assuming positive pressure – is a *positive* (that is, repulsive) force between the particles directed along their line of sight. This force will only vanish when the error term (75) becomes zero – that is, when the particles are *regular*. What this means is that *the particles care about their (bad) arrangement* and will rearrange themselves until the condition (75) holds, corresponding to a particle distribution that is locally isotropic and regular. This state corresponds to the minimum energy state of the Hamiltonian system – i.e., that which minimises the action (18), and in this state the symmetric gradient estimate (73) is computed with good accuracy (since by definition the motions act to minimise the errors in this estimate). Monaghan [64] gives some specific examples of this.

In other words, the Hamiltonian SPH formulation – or more generally any formulation which respects local conservation of momentum between particle pairs – contains an *intrinsic* "re-meshing" procedure and the particles are constrained to remain locally ordered at all times. With an 'accurate' but non-conservative pressure estimate the particles are insensitive to their arrangement and can become arbitrarily randomised in the course of a simulation (according to the streamlines of the flow). Thus, methods based on an 'exact derivatives' approach e.g. [26,54] inevitably require the addition of explicit (and ad hoc) re-meshing procedures, as the interpolation errors on a randomised particle distribution will be terrible regardless of the order of the interpolation scheme. So in practice the 'accurate' gradient estimate will give much *less* accurate results. This is the paradox of SPH: One deliberately chooses a bad gradient estimate (in the linear errors) in order to obtain a good gradient estimate (because the particles stay regular).

One of the implications of this "intrinsic remeshing" is that not all initial conditions represent stable configurations for the particles. In particular, the cubic lattice – though simple – does not represent a minimum energy state and is only quasi-stable for certain ratios of the smoothing length to the particle separation [7,69]. In general, given a small perturbation or sufficient time, the cubic lattice will transition to the more isotropic hexagonal-close-packed lattice arrangement and will do so almost immediately if $h/\Delta p$ is small (e.g. $\eta \lesssim 1.1$).

5.1. Example 1: Settling of a random particle distribution

Our first numerical example (using NDSPMHD) demonstrates this 'settling' in practice. The setup is a two dimensional domain $x, y \in [0,1]$ with periodic boundary conditions. The particles are given an initially uniform thermal energy, density is calculated according to the sum and the pressure is determined using $P = (\gamma - 1)\rho u$ with $\gamma = 5/3$. The thermal energy is set to give a sound speed $c_s = 1$, giving $u = 0.9$. The end result should therefore be a uniform pressure equilibrium. The particle distributions after 0.5 (top row) and 10 (bottom row) sound crossing times are shown in Fig. 4 using the Hamiltonian SPH formulation (43)–(45) (left and centre columns, without and with artificial viscosity, respectively) and the equations of motion computed using a 'relative pressure' formulation (specifically, Eq. (76) with $\phi = 1/\rho$ as employed by Abel [1]) (right column). With a locally conservative formulation the random initial configuration settles rapidly into a regular particle distribution (left and centre columns), leading in turn to good gradient estimates. This settling occurs even in the absence of an artificial viscosity term (left panels), though adding viscosity does help speed the settling process (centre panels). However, using a "more accurate" but non-conservative gradient estimate there is no regularisation of the particle distribution, leading to

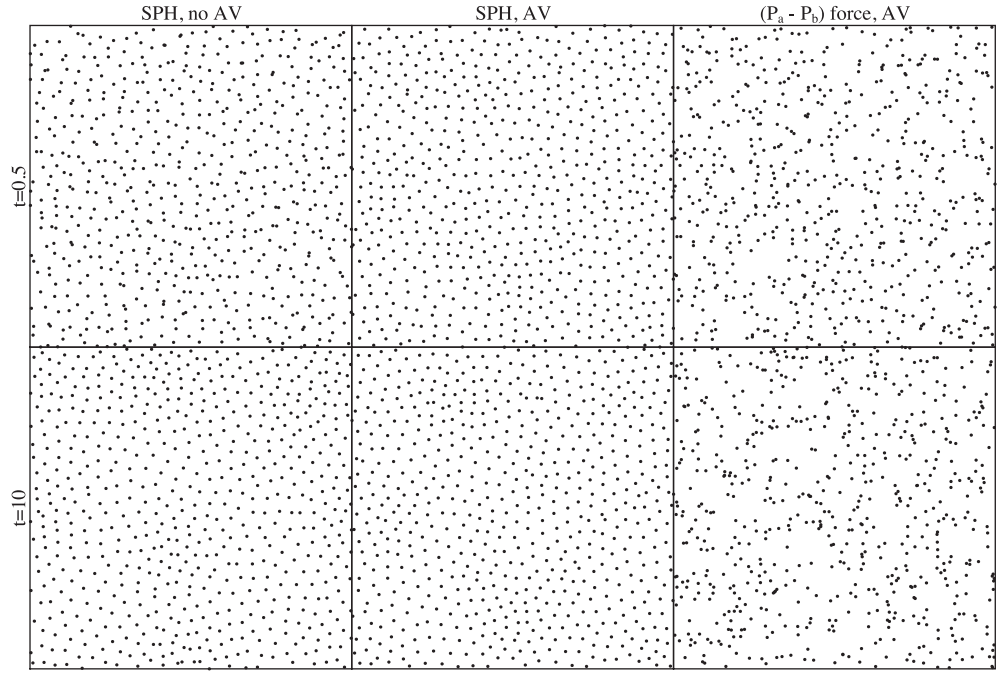

Fig. 4. Settling of an initially random particle distribution due to pairwise conservation of momentum in the pressure gradient. The left and centre columns show the results with the standard (Hamiltonian) SPH method, both without (left) and with (centre) artificial viscosity, after 0.5 (top) and 10 (bottom) sound crossing times. The right column shows the results when a "relative pressure" formulation is adopted. With a momentum-conserving force the particles are sensitive to their arrangement and will regularise accordingly (left and centre columns), whereas "more accurate" but non-conserving formulations (right) compute gradients that are insensitive to the particle arrangement and thus require explicit re-meshing procedures. Note that although the application of artificial viscosity helps the settling to proceed faster (centre), it is primarily a pressure-driven effect and occurs even if no viscosity is applied (left).

poor gradient estimates due to the random nature of the particle distribution. The total energy is also conserved exactly by the SPH formulations, whilst in the relative pressure formulation the total energy grows exponentially.

5.2. Example 2: A 2D shock tube

The other "classic" example of particle settling is the behaviour of SPH particles in a multidimensional shock tube problem, where there is a 1D compression of the particle distribution (e.g. along the x axis). Since the shock induces a highly anisotropic compression – and thus a highly non-preferred particle arrangement – the mutual repulsion of SPH particles will eventually produce a post-shock "remeshing" of the particle distribution, involving transverse motions of the particles. An example is given in Fig. 5, showing the particle distribution at $t = 0.1$ in a two dimensional Sod shock tube problem (described further in Section 6.3.4) in which the particles were initially placed on two hexagonal close packed lattices upstream and downstream of the shock (initially placed at the origin). The particles can be seen to "break" (at $x \approx 0.14$) from the highly anisotropic compression-induced 'lines' at $0.14 \lesssim x \lesssim 0.19$, leading to a more isotropic particle distribution further downstream ($x \lesssim 0.12$). This example also illustrates the fact that one inevitably has some motions at the resolution scale that are not related to the physical problem, but related to the implicit "regularisation" of the particle distribution present in locally conservative SPH formulations.

5.3. Corollary: Negative pressures and the tensile instability

The corollary of the above is that the particles require a *positive* pressure in order to remain ordered. If the net pressure (or stress) becomes negative, the net force between a particle pair will become *attractive*, causing a catastrophic numerical instability. For example, with a pressure gradient of the form

$$\frac{\mathrm{d}\mathbf{v}_a}{\mathrm{d}t} = -\sum_b m_b \left(\frac{P_a - P_0}{\rho_a^2} + \frac{P_b - P_0}{\rho_b^2}\right) \nabla_a W_{ab}, \tag{84}$$

the pairwise force will become negative when $P_0 > P$, and in this situation the particles clump together unphysically. This is known as the 'tensile instability' [61] and occurs in SPH when a stress tensor is employed that can result in (physically) negative stresses. In particular, this is the case for MHD [76] and in elastic dynamics [38]. The occurrence of the tensile instability was one of the main initial difficulties with the development of MHD in SPH and is discussed in detail in Section 8.

5.4. The pairing instability: Why one cannot simply use 'more neighbours'.

Another, more benign, instability in the particle distribution occurs with the cubic spline and other bell-shaped kernels depending on the ratio of smoothing length to particle spacing. This is due to the shape of the kernel gradient term for these kernels (see Fig. 2), and is a consequence of the fact that these kernels are designed to give good density estimates (Section 2.2), rather than necessarily being the best choice for calculating gradients. In particular, the kernel gradient in these kernels contains a maximum (negative) value at $r/h \sim 2/3$ and tends to zero at the origin (Fig. 2). This characteristic is desirable for a good density estimate – as it means one is insensitive to a small change in the position of a near neighbour – but means that the mutual repulsive force tends to zero for neighbouring particles placed "within the hump" of the kernel gradient. The net result is that two particles spaced closer than the location of the "hump" in the gradient form a "pair", eventually falling on top of each other.

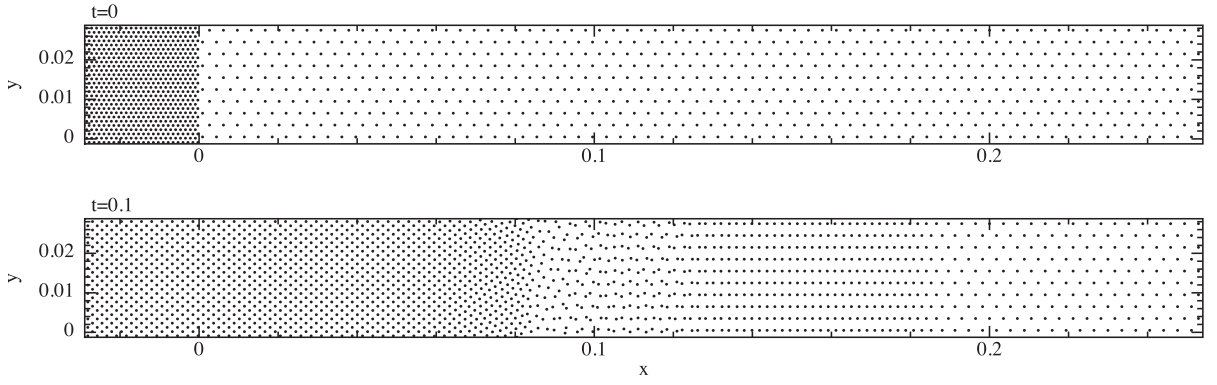

Fig. 5. Particle settling in a two dimensional shock tube problem. The particles are initially arranged on hexagonal close packed lattices either side of the shock (top panel). As the shock propagates (bottom panel, showing $t = 0.1$) it induces a one dimensional compression in the particles and thus a highly anisotropic particle arrangement, which "remeshes" to a more isotropic arrangement downstream from the shock, involving small motions of particles in the y-direction.

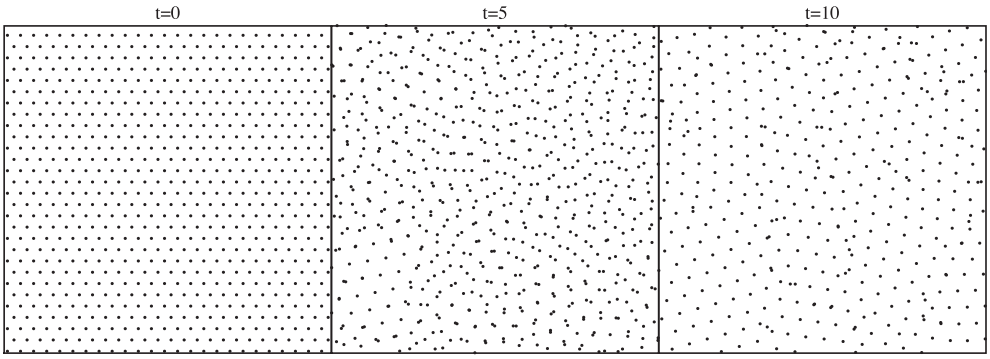

Fig. 6. The pairing instability in action: The (2D) setup is similar to that shown in Fig. 4 except that the particles are initially placed on a close packed lattice and we use the M_4 cubic spline kernel with a large ratio of smoothing length to particle spacing (here $\eta = 1.5$, corresponding to $\sim$28 and 100 neighbours in 2 and 3D, respectively). After a few sound crossing times ($t = 5$, centre panel) particles form 'pairs' which proceed to merge into a locally hexagonal "glass-like" lattice arrangement with almost exactly half the resolution of the initial conditions (right panel, shown after 10 sound crossing times). Although fairly benign – and easily avoided by a sensible choice of η – the pairing instability is the main reason one cannot simply "stretch" the cubic spline to large neighbour numbers to achieve convergence. Instead, one should use a kernel with a larger radius of compact support but the same ratio of smoothing length to particle spacing, such as the M_5 or M_6 splines.

For the cubic and other B-spline kernels complete merging occurs when $h \gtrsim 1.5\Delta p$ (i.e., $\eta \gtrsim 1.5$ or $\gtrsim 100$ Neighbours in 3D for the cubic spline), corresponding to the placement of the first neighbour "inside the hump". There is also an intermediate regime $1.225 \lesssim \eta \lesssim 1.5$ (62–100 Neighbours in 3D) where a close-packed or cubic lattice is unstable to pair formation, but where the pairs do not completely merge. These empirical regimes are confirmed by detailed stability analysis of the SPH equations in 2D [7,69] that explicitly show that instability occurs – though with small energies – for large $h/\Delta p$.

Fig. 6 (and our example 3) shows the pairing instability in action: The setup is as for example 1 but with $\eta = 1.5$ in the cubic spline kernel instead of $\eta = 1.2$ and with particles placed initially on a hexagonal close-packed lattice (an otherwise very stable configuration: left panel). After a few sound crossing times (centre panel) particles begin to form pairs, with these pairs eventually merging completely (right panel) to give a locally hexagonal "glass-like" configuration, but with exactly half the resolution of the initial conditions!

Though fixes have been proposed,[8] none are entirely satisfactory. However, the pairing instability, unlike the tensile instability, is quite benign. For example the density change associated with the transition in Fig. 6 is of order 1% for the cubic spline and 0.1% for the M_6 quintic – but entails a factor-of-two loss in spatial resolution and is therefore a waste of computational resources. Furthermore, it can be easily avoided by a sensible choice of η (we recommend $\eta = 1.2$ for the B-spline kernels, corresponding to $N_{neigh} = 57.9$ for the cubic spline in 3D). The pairing instability is the main reason one cannot simply "stretch" the cubic spline to large neighbour numbers in order to obtain convergence and demonstrates at least one good reason why η (or N_{neigh}) should not be regarded as a free parameter in SPH simulations.

6. Second derivatives and dissipation terms in SPH and SPMHD

6.1. The SPH Laplacian

Our remaining "basic" issue regarding both SPH and SPMHD regards the formulation of second derivative terms. As for first derivative terms we can start with the basic summation interpolant (49) and simply take derivatives analytically, e.g.

[8] Thomas and Couchman [106] suggested modifying the gradient of the cubic spline kernel, using

$$w'(q) = -\sigma \begin{cases} -1, & 0 \leqslant q < 2/3; \\ -3q + \frac{9}{4}q^2, & 2/3 \leqslant q < 1; \\ \frac{3}{4}(2-q)^2, & 1 \leqslant q < 2; \\ 0, & q \geqslant 2. \end{cases} \tag{85}$$

with W itself unchanged and σ equal to the usual normalisation factor for the cubic spline (i.e., $1/\pi$ in 3D). That is, the "hump" is removed by simply making the kernel gradient constant within $r/h < 2/3$. Whilst it cures the pairing instability, one should be careful about employing such a gradient in practice since the kernel gradient (85) is no longer correctly normalised (i.e., Eq. 67b no longer holds, even in the continuum limit) meaning that as the region within $r/h < 2/3$ is increasingly well sampled the numerical sound speed and other quantities will be systematically wrong. Though one could attempt to re-normalise the new gradient kernel, this results in a low weighting in the outer regions that in turn leads to poor gradient estimates.
Similarly, whilst perhaps a satisfactory 'gradient kernel' could be derived without a pairing instability, in the derivation from a Lagrangian there is no freedom over the kernel gradient since it derives directly from the gradient of density – that is, if one separates the gradient kernel from the density kernel then either the total energy (from Eq. 37) or the entropy will no longer exactly be conserved (the latter if $\mathrm{d}u/\mathrm{d}t \neq P/\rho^2 d\rho/\mathrm{d}t$ and thus $\mathrm{d}K/\mathrm{d}t \neq 0$ in Eq. 41).

$$\langle\nabla^2 A\rangle_a \approx \sum_b m_b \frac{A_b}{\rho_b}\nabla_a^2 W_{ab}. \tag{86}$$

Expanding A_b in a Taylor series about $\mathbf{r}_a$, we have

$$\sum_b m_b \frac{A_b}{\rho_b}\nabla_a^2 W_{ab} = A_a \sum_b \frac{m_b}{\rho_b}\nabla_a^2 W_{ab} + \nabla^\alpha A_a \sum_b \frac{m_b}{\rho_b}\delta\mathbf{r}^\alpha \nabla_a^2 W_{ab} + \frac{1}{2}\nabla^\alpha\nabla^\beta A_a \sum_b \frac{m_b}{\rho_b}\delta\mathbf{r}^\alpha\delta\mathbf{r}^\beta\nabla_a^2 W_{ab} + \cdots \tag{87}$$

As previously, we can immediately improve the estimate by subtracting the first error term from both sides, giving an interpolant of the form

$$\sum_b \frac{m_b}{\rho_b}(A_b - A_a)\nabla_a^2 W_{ab} = \nabla^\alpha A_a \sum_b \frac{m_b}{\rho_b}\delta\mathbf{r}^\alpha \nabla_a^2 W_{ab} + \frac{1}{2}\nabla^\alpha\nabla^\beta A_a \sum_b \frac{m_b}{\rho_b}\delta\mathbf{r}^\alpha\delta\mathbf{r}^\beta\nabla_a^2 W_{ab} + \cdots, \tag{88}$$

which vanishes when A is constant. However, the accuracy of our second derivative estimate will depend on the remaining error terms in (88), corresponding to the normalisation conditions:

$$\sum_b \frac{m_b}{\rho_b}\delta\mathbf{r}\nabla^2 W_{ab} = \mathbf{0}; \quad \text{and} \quad \frac{1}{2}\sum_b \frac{m_b}{\rho_b}\delta\mathbf{r}^\alpha\delta\mathbf{r}^\beta\nabla^2 W_{ab} = \delta^{\alpha\beta}, \tag{89}$$

such that

$$\nabla^2 A_a \approx \sum_b \frac{m_b}{\rho_b}(A_b - A_a)\nabla_a^2 W_{ab}. \tag{90}$$

The problem is that the conditions (89) are very poorly satisfied using the second derivative of a compact bell-shaped kernel function (short-dashed/green lines in Fig. 2, showing the second term in (89)), since the second derivative changes sign inside the kernel domain (for the cubic spline it is also discontinuous) and must therefore be extremely well sampled in each direction to give accurate results. Thus in practice we require a better functional form – a "second derivative kernel".

The actual formulation commonly employed derives from early work by Monaghan and first published by Brookshaw [11], where the SPH Laplacian is written in the form

$$\nabla^2 A_a \approx 2\sum_b \frac{m_b}{\rho_b}(A_a - A_b)\frac{F_{ab}}{|r_{ab}|}, \tag{91}$$

where we use the definition $\nabla_a W_{ab} \equiv \hat{\mathbf{r}}_{ab} F_{ab}$ such that F_{ab} is the scalar part of the kernel gradient term. Thus in effect we use the *first derivative* kernel function divided by the particle spacing to give a second derivative. Whilst the interpretation of (91) from the integral representation is in general rather complicated see, e.g. [31,47,64,78], it can be more easily understood as if we had simply used (90) with a new kernel Y_{ab} instead of W_{ab}, defined according to

$$\nabla^2 Y_{ab} \equiv -\frac{2F_{ab}}{|r_{ab}|}. \tag{92}$$

The left figure in Fig. 7 shows Y' constructed in the above manner from the M_4 and M_6 kernel gradients, and may be compared with the standard kernel second derivatives shown in Fig. 2. The right figure shows the normalisation condition (89) for these 'kernels'. Comparison with Fig. 3 shows that, indeed, the second derivative is much better estimated with kernel second derivative functions that are monotonically decreasing and positive, such as those constructed via (92) from the bell-shaped kernels.

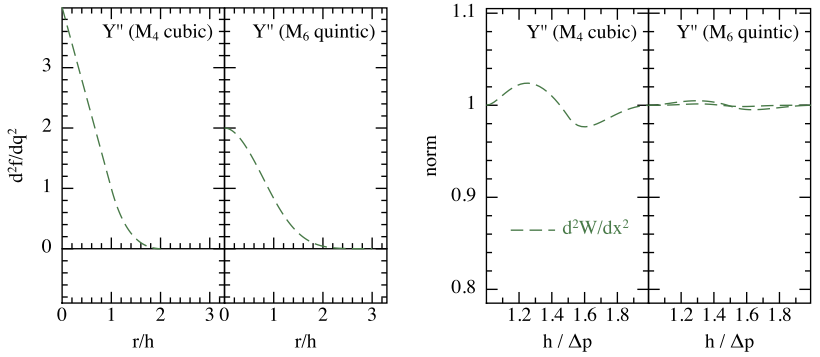

Fig. 7. Second derivative kernel functions $Y''(q) \equiv -2w'(q)/q$ constructed from the first derivatives of the M_4 cubic and M_6 quintic kernel functions (left figure, compare to the standard second derivatives shown in Fig. 2), together with the accuracy with which the second derivative normalisation conditions (89) are satisfied for a fixed ratio of $h/\Delta p$ (compare with the standard second derivative functions shown in Fig. 3).

Understanding the Brookshaw [11] Laplacian as equivalent to (90) with an alternative kernel also helps to interpret more complicated expressions. For example, it is quite straightforward to show that

$$\sum_b \frac{m_b}{\rho_b}(\kappa_a+\kappa_b)(A_a-A_b)\frac{F_{ab}}{|r_{ab}|}\approx \nabla\cdot(\kappa\nabla A), \tag{93}$$

by writing $-2F_{ab}/|r_{ab}|\equiv\nabla^2 Y_{ab}$ and interpreting each term via (86). Cleary and Monaghan [18] proposed an alternative average of the κ terms given by

$$\sum_b \frac{m_b}{\rho_b}\frac{4\kappa_a\kappa_b}{(\kappa_a+\kappa_b)}(A_a-A_b)\frac{F_{ab}}{|r_{ab}|}\approx \nabla\cdot(\kappa\nabla A), \tag{94}$$

which was formulated to give smooth derivatives when κ is discontinuous. The above expression forms the basis of formulations of thermal conductivity in SPH [18], though has been similarly used to model a wide range of dissipative terms including viscosity e.g. [17,44], salt diffusion [65] and in an astrophysical context for the treatment of radiation in the flux-limited diffusion approximation [109,110].

6.2. Vector second derivatives

Second derivatives of vector quantities do not quite follow the same analogy as the Laplacian, because in general $\nabla^i\nabla^j W_{ab}$ involves a mix of the first and second derivatives of the dimensionless kernel function (see Appendix A in Price [81]). Thus, the proof is more involved (see [31,64] for details), but the basic expressions for vector second derivatives are given by

$$\langle\nabla^2\mathbf{A}\rangle = -2\sum_b \frac{m_b}{\rho_b}(\mathbf{A}_a-\mathbf{A}_b)\frac{F_{ab}}{|r_{ab}|}, \tag{95}$$

$$\langle\nabla(\nabla\cdot\mathbf{A})\rangle = -\sum_b \frac{m_b}{\rho_b}\left[\left(\delta^k_k+2\right)(\mathbf{A}_{ab}\cdot\hat{\mathbf{r}}_{ab})\hat{\mathbf{r}}_{ab}-\mathbf{A}_{ab}\right]\frac{F_{ab}}{|r_{ab}|}, \tag{96}$$

where $\delta^k_k\equiv d$ i.e., the number of spatial dimensions. A corollary of the above is that a second derivative computed purely along the line of sight between the particles (e.g. constructed so as to conserve angular momentum) corresponds to

$$-\sum_b \frac{m_b}{\rho_b}(\mathbf{A}_{ab}\cdot\hat{\mathbf{r}}_{ab})\frac{F_{ab}}{|r_{ab}|} = \frac{1}{d+2}\nabla(\nabla\cdot\mathbf{A})+\frac{1}{2(d+2)}\nabla^2\mathbf{A}. \tag{97}$$

An alternative approach is to calculate vector second derivatives by taking two first derivatives. Although more expensive, this has been the approach adopted in several formulations of physical viscosity, e.g. for SPH modelling of accretion discs [34,108].

6.3. Artificial dissipation terms in SPH and SPMHD

6.3.1. Interpretation of SPH artificial viscosity terms

This brings us to the formulation and interpretation of artificial dissipation terms in SPH. The 'standard' formulation of artificial viscosity is given by [59]

$$\left(\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t}\right)_{diss} = -\sum_b m_b\frac{-\alpha\bar{c}_{s,ab}\mu_{ab}+\beta\mu^2_{ab}}{\bar{\rho}_{ab}}\hat{\mathbf{r}}_{ab}F_{ab};\quad \mu_{ab}=\frac{h\mathbf{v}_{ab}\cdot\mathbf{r}_{ab}}{\mathbf{r}_{ab}+\epsilon\bar{h}^2_{ab}}, \tag{98}$$

where barred quantities correspond to an average, i.e., $\bar{\rho}_{ab}=(\rho_a+\rho_b)/2$, c_s is the sound speed, α and β are dimensionless parameters (typically $\alpha=1$ and $\beta=2$) and $\epsilon\sim0.01$ is a small parameter to prevent divergences. This form was chosen because it is Galilean invariant, vanishes for rigid body rotation and conserves total linear and angular momentum [59]. If we neglect the non-linear (β) term, set $\epsilon=0$ and assume that c_s, h and ρ are approximately constant over the kernel radius, then using (97) we can directly translate this expression into the continuum form

$$\frac{1}{d+2}\alpha c_s h\nabla(\nabla\cdot\mathbf{v})+\frac{1}{2(d+2)}\alpha c_s h\nabla^2\mathbf{v}. \tag{99}$$

Comparison with the compressible Navier–Stokes equations shows that the artificial viscosity is therefore equivalent to a physical viscosity with shear and bulk coefficients proportional to the resolution length, e.g. [3,52,64,72]

$$\nu\approx\frac{1}{2(d+2)}\alpha c_s h;\quad \zeta=\frac{5}{3}\nu\approx\frac{5}{6(d+2)}\alpha c_s h. \tag{100}$$

Since for shock-capturing the viscosity is only applied when particles are approaching ($\mathbf{v}_{ab}\cdot\mathbf{r}_{ab}<0$), in a uniform shear flow – or accretion disc – the viscosity coefficients will be approximately half of these values.

6.3.2. General formulation of dissipative terms in SPH and SPMHD

A more general formulation of dissipative terms was proposed by Monaghan [60], based on an analogy with Riemann solvers and the need to formulate dissipative terms for ultra-relativistic shocks [16]. The general principle is that dissipative terms in the conservative variables involves jumps in those variables (the "left" and "right" states of the Riemann problem), multiplied by eigenvalues that can be interpreted as signal velocities.

For the equations of hydrodynamics the conservative variables are the density, specific momentum and energy (ρ, $\mathbf{v}$ and $e = \frac{1}{2}v^2 + u$, respectively), and the terms take the form[9] [60,80]

$$\left(\frac{\mathrm{d}\mathbf{v}_a}{\mathrm{d}t}\right)_{diss} = \sum_b m_b \frac{\alpha v_{sig}(\mathbf{v}_a - \mathbf{v}_b)\cdot\hat{\mathbf{r}}_{ab}}{\bar{\rho}_{ab}}\hat{\mathbf{r}}_{ab}\overline{F_{ab}}, \tag{101}$$

$$\left(\frac{\mathrm{d}e_a}{\mathrm{d}t}\right)_{diss} = \sum_b m_b \frac{(e^*_a - e^*_b)}{\bar{\rho}_{ab}}\overline{F_{ab}}, \tag{102}$$

where $e^*_a = \frac{1}{2}\alpha v_{sig}(\mathbf{v}_a\cdot\hat{\mathbf{r}}_{ab})^2 + \alpha_u v^u_{sig}u_a$ refers to an energy including only components along the line of sight joining the particles, $\overline{F_{ab}} \equiv [F_{ab}(h_a) + F_{ab}(h_b)]/2$ (equivalent to writing $\overline{F}_{ab} = \hat{\mathbf{r}}_{ab}\cdot\overline{\nabla_a W_{ab}}$) and α and α_u are the dimensionless artificial viscosity and thermal conductivity parameters, respectively. The signal speed v_{sig} refers to the maximum (averaged) signal speed between a particle pair. For hydrodynamics we use

$$v_{sig} = \begin{cases} \frac{1}{2}[c_{s,a} + c_{s,b} - \beta\mathbf{v}_{ab}\cdot\hat{\mathbf{r}}_{ab}]; & \mathbf{v}_{ab}\cdot\hat{\mathbf{r}}_{ab} \leqslant 0; \\ 0; & \mathbf{v}_{ab}\cdot\hat{\mathbf{r}}_{ab} > 0. \end{cases} \tag{103}$$

Eq. (101), with the above v_{sig}, provides a standard artificial viscosity term similar to (98) (careful expansion shows they differ only by a factor of $h/|r_{ab}|$ and the no-longer-necessary ϵ term). The dissipation term in the total energy also contains a term involving $(u_a - u_b)$ which acts to smooth jumps in the thermal energy (e.g. at a contact discontinuity). The interpretation of this term is clearer from the contribution to the thermal energy evolution, given by

$$\left(\frac{\mathrm{d}u}{\mathrm{d}t}\right)_{diss} = -\sum_b \frac{m_b}{\bar{\rho}_{ab}}\left[\frac{1}{2}\alpha v_{sig}(\mathbf{v}_{ab}\cdot\hat{\mathbf{r}}_{ab})^2 + \alpha_u v^u_{sig}(u_a - u_b)\right]\overline{F}_{ab}, \tag{104}$$

where the second term, from (91), can be directly interpreted as a $\nabla^2 u$ term. Note that the signal velocity v^u_{sig} used in the artificial conductivity does not have to be the same as that used for the viscosity. In particular, Price [80] proposed using $v^u_{sig} = \sqrt{|P_a - P_b|/\bar{\rho}_{ab}}$ to equalise the pressure across contact discontinuities, whilst Wadsley et al. [107] proposed a conductivity term equivalent to using $v^u_{sig} = |\mathbf{v}_{ab}\cdot\hat{\mathbf{r}}_{ab}|$ (we adopt the former for the tests shown in this paper).

6.3.3. Switches for viscosity terms

One of the key issues in practice is to ensure that sufficient dissipation is applied to discontinuities, but that such dissipation is effectively turned off in smooth parts of the flow by designing appropriate switches. Morris and Monaghan [70] suggested allowing the parameter α to be individual to each particle, with an evolution equation of the form

$$\frac{d\alpha}{dt} = \mathcal{S} + \frac{\alpha - \alpha_{min}}{\tau}; \quad \tau = \frac{h}{\sigma c_s}, \tag{105}$$

where $\mathcal{S}$ is a source term that grows large at the discontinuity [e.g. $\mathcal{S} = \max(0, -\nabla\cdot\mathbf{v})$ for shocks], τ is the decay time, set such that α decays to α_{min} over several smoothing lengths (typically $\sigma = 0.1$ and $\alpha_{min} = 0.1$). A similar switch can be employed for the thermal conductivity parameter α_u, with Price and Monaghan [88] adopting a source term given by $\mathcal{S}_u = 0.1h\nabla^2 u$. More sophisticated switches for shock detection are also possible, with a promising recent alternative suggested by Cullen and Dehnen [20]. Directly employing the Riemann solution is another possibility e.g. [15,45].

6.3.4. Examples 4 and 5: One and two dimensional shock tubes, and Kelvin–Helmholtz instabilities

Two specific examples of the dissipative terms in practice are shown in Fig. 8 (applying only viscosity) and Fig. 9 (applying artificial viscosity *and* conductivity), showing the results of a one dimensional Sod shock tube problem (left subfigure) and a two dimensional Kelvin–Helmholtz (K–H) instability problem (right subfigure). The 1D shock tube is setup using a total of 450 particles in $-0.5 < x < 0.5$ with conditions to the left of the origin given by $[\rho_L, P_L, v_L] = [1, 1, 0]$ and to the right by $[\rho_R, P_R, v_R] = [0.125, 0.1, 0]$ with $\gamma = 5/3$. Importantly, purely discontinuous initial conditions are employed so that the contact discontinuity is not already smoothed. The K–H instability problem is setup with a 2:1 density ratio and equal mass particles, identical to the setup described in Price [80] and using 512×512 particles in the low density fluid. Whilst shocks are smoothed by the artificial viscosity term (Figs. 8 and 9), with only viscosity the jump in thermal energy at the contact discontinuity is not treated, resulting in a 'blip' in the pressure profile (Fig. 8). This manifests in the 2D K–H problem as

[9] Note that in SPH no dissipation is required in the density equation provided the density is computed from a sum. This is because (10) represents an integral form of the continuity equation and can be shown to differ from (56) by a surface integral term that is non-zero only at boundaries or equivalently, discontinuities in the flow. See Price [80] for more details.

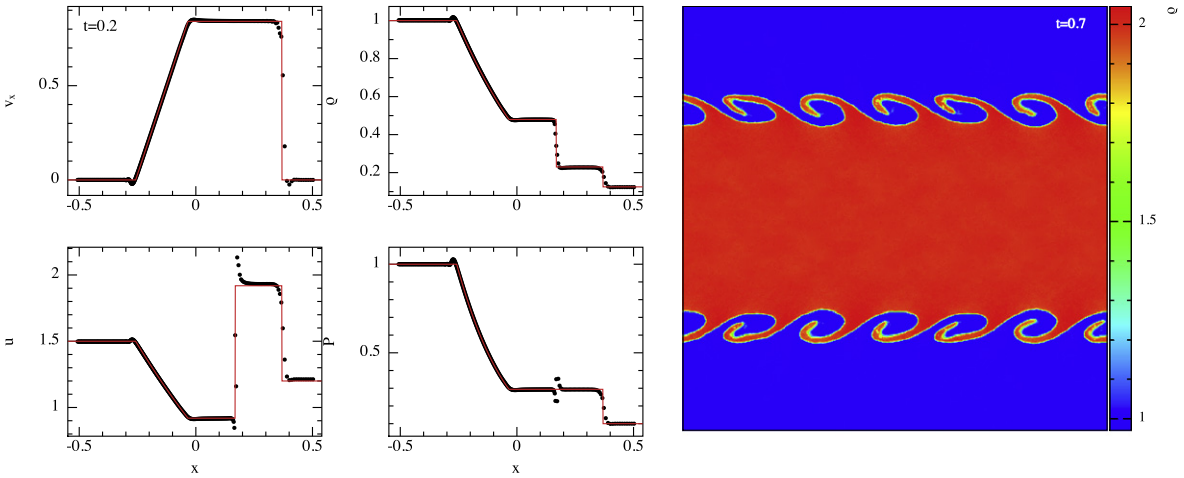

Fig. 8. Treatment of discontinuities in SPH: A 1D Sod shock tube problem (left) and 2D Kelvin–Helmholtz instability (right), applying only artificial viscosity terms and with unsmoothed initial conditions. Whilst in the 1D shock tube problem (left panel) the shock is smoothed over several particle spacings by the viscosity term, there are problems at the contact discontinuity causing a 'blip' in the pressure. This leads to a suppression of mixing across contact discontinuities in 2D (right panel).

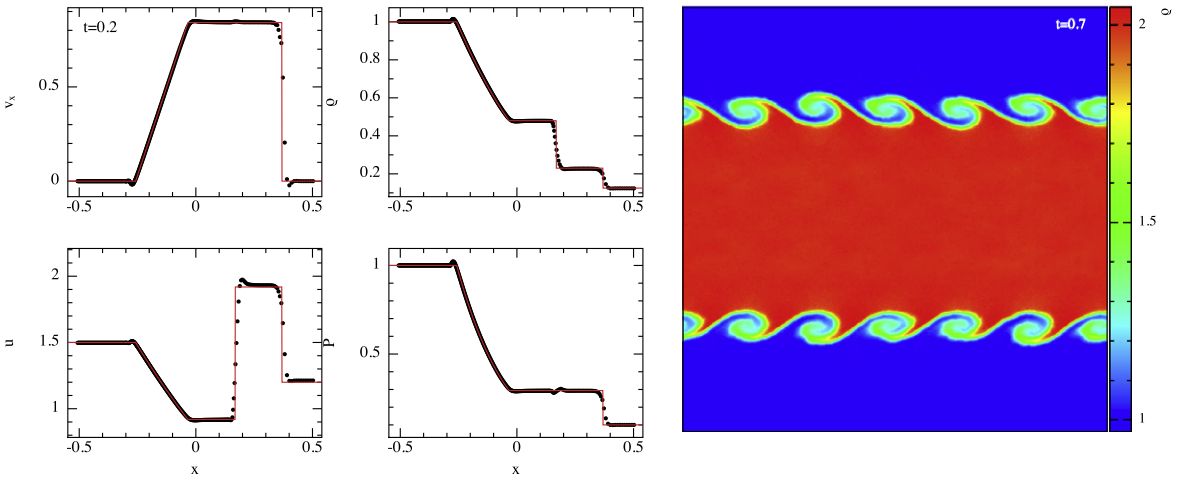

Fig. 9. As in Fig. 8 but with artificial conductivity applied as well as artificial viscosity. This smoothes the contact discontinuity, removing the pressure 'blip' (left panel) and restoring the mixing in the 2D K–H problem (right panel).

an 'artificial surface tension' effect, caused by the very same kind of pressure blip across the boundary (contact discontinuity) between the dense and the light fluids, suppressing mixing of the two. With conductivity applied (Fig. 9) the pressure is smooth across the contact discontinuity in both problems, which for the K–H problem means that the two fluids mix correctly. The lack of treatment of contact discontinuities in standard SPH codes (i.e., with no artificial conductivity term) explains the discrepancy between grid and SPH results somewhat infamously highlighted by Agertz et al. [2].

Fig. 10 shows the same shock tube in 2D (as already discussed briefly in Fig. 5). The main difference to 1D is that the "noise" due to the particle resettling behind the shock front is visible (left subfigure). It is often asserted that SPH performs poorly on 2D shocks for this reason, however the noise can be very effectively minimised (at some additional cost) by employing the M_6 quintic kernel instead of the cubic spline (right subfigure), giving results comparable to the 1D version and illustrating in practice how the higher M_n kernels can be used to obtain convergence in SPH.

7. Smoothed particle magnetohydrodynamics from a Lagrangian

We can follow the same general approach to constructing an SPMHD algorithm as for hydrodynamics (Section 3): Write down the Lagrangian, use appropriate physical constraints and use this to consistently derive the resultant equations of motion.

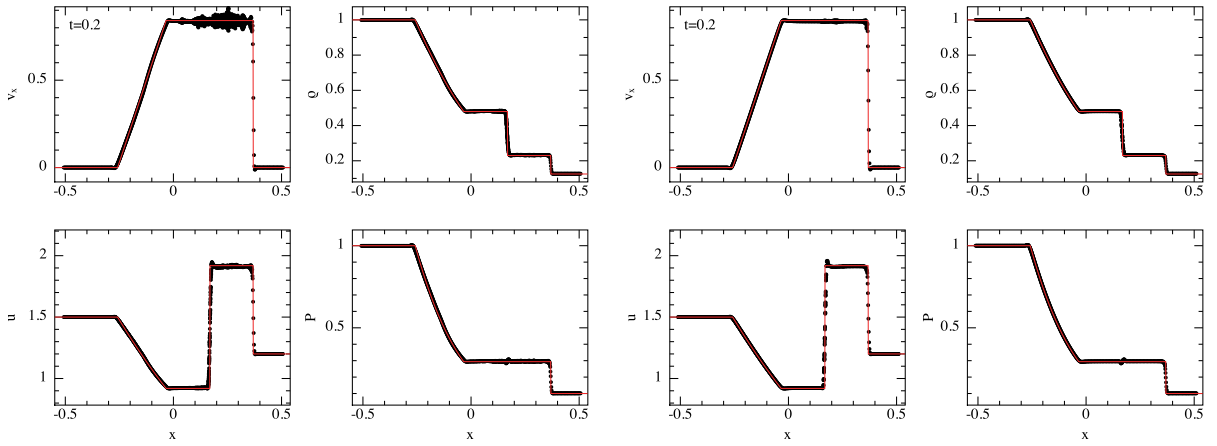

Fig. 10. The Sod shock tube in 2D. Using the cubic spline kernel (left panel), there is additional 'noise' in the 2D velocity field (compared to 1D, Fig. 9) due to the transverse "remeshing" motions of particles behind the shock front in multidimensions (see Fig. 5). However, this can be quite effectively minimised by using a smoother kernel (right panel, using the M_6 quintic).

7.1. MHD Lagrangian

For MHD, the Lagrangian is given by [87]

$$L_{\rm MHD} = \sum_b m_b \left[\frac{1}{2}v_b^2 - u_b(\rho_b, s_b) - \frac{1}{2\mu_0}\frac{B_b^2}{\rho_b}\right], \tag{106}$$

corresponding simply to the subtraction of a magnetic energy term from the hydrodynamic version (Eq. 16). In the continuum limit this corresponds to the standard MHD Lagrangian used by many authors e.g. [33,74]

$$L_{\rm MHD} = \int \left(\frac{1}{2}\rho v^2 - \rho u - \frac{1}{2\mu_0}B^2\right){\rm d}V. \tag{107}$$

The difference to the hydrodynamic case is that, unlike the thermal energy term, neither the magnetic field $\mathbf{B}$ nor the change in the magnetic field can be written directly as a function of the particle coordinates, so we cannot straightforwardly employ the Euler–Lagrange Eq. (21). Instead, we can use the more general form of the variational principle given by $\delta S = \int \delta L {\rm d}t = 0$ (Section 3.2), where from (106) we have

$$\delta L = m_a \mathbf{v}_a \cdot \delta \mathbf{v}_a - \sum_b m_b \left[\left.\frac{\partial u_b}{\partial \rho_b}\right|_s \delta\rho_b + \frac{1}{2\mu_0}\left(\frac{B_b}{\rho_b}\right)^2 \delta\rho_b + \frac{1}{\mu_0}\mathbf{B}_b \cdot \delta\left(\frac{\mathbf{B}_b}{\rho_b}\right)\right] \tag{108}$$

and the perturbation is with respect to a small change in the particle coordinates $\delta\mathbf{r}$. So we can derive the equations of motion provided that we are able to express the change in the magnetic field $\delta\mathbf{B}$ as a function of the *change* in particle coordinates – equivalent to being able to write down an expression for the Lagrangian time derivative ${\rm d}\mathbf{B}/{\rm d}t$ [or equivalently, ${\rm d}(\mathbf{B}/\rho)/{\rm d}t$] since ${\rm d}/{\rm d}t \equiv \delta/\delta t$. In other words, in order to derive the SPMHD equations of motion it is necessary to specify not only the density estimate but also the manner in which the magnetic field is evolved.

7.2. SPMHD formulation of the induction equation

Given the induction equation for (ideal) MHD, written in the Lagrangian form

$$\frac{\rm d}{{\rm d}t}\left(\frac{\mathbf{B}}{\rho}\right) = \left(\frac{\mathbf{B}}{\rho}\cdot\nabla\right)\mathbf{v} \tag{109}$$

and using our general formulations of SPH derivatives given in Section 4.6, it is straightforward to write down an SPH version of the form

$$\frac{\rm d}{{\rm d}t}\left(\frac{\mathbf{B}_a}{\rho_a}\right) = -\sum_b m_b(\mathbf{v}_a - \mathbf{v}_b)\frac{\mathbf{B}_a}{\Omega_a\rho_a^2}\cdot\nabla W_{ab}(h_a), \tag{110}$$

which is equivalent to using the antisymmetric derivative (81(a)) with $\phi = \rho$.

7.3. Equations of motion

The perturbations required in (108), from (56) and (110) are therefore given by

$$\delta\rho_b = \frac{1}{\Omega_b}\sum_c m_c(\delta\mathbf{r}_b - \delta\mathbf{r}_c)\cdot\nabla_b W_{bc}(h_b), \tag{111}$$

$$\delta\left(\frac{\mathbf{B}_b}{\rho_b}\right) = -\sum_c m_c(\delta\mathbf{r}_b - \delta\mathbf{r}_c)\frac{\mathbf{B}_b}{\Omega_b\rho_b^2}\cdot\nabla_b W_{bc}(h_b), \tag{112}$$

giving, from (108) and using (25)

$$\delta S = \int \delta L\mathrm{d}t = \int\left\{m_a\mathbf{v}_a\cdot\delta\mathbf{v}_a - \sum_b m_b\left[\left(\frac{P_b + \frac{1}{2}B_b^2/\mu_0}{\Omega_b\rho_b^2}\right)\sum_c m_c\nabla_b W_{bc}(h_b)(\delta_{ba}-\delta_{ca})\right]\cdot\delta\mathbf{r}_a \right.$$
$$\left. + \sum_b m_b\left[\frac{\mathbf{B}_b}{\mu_0}\sum_c m_c\frac{\mathbf{B}_b}{\Omega_b\rho_b^2}\cdot\nabla_b W_{bc}(h_b)(\delta_{ba}-\delta_{ca})\right]\cdot\delta\mathbf{r}_a\right\}\mathrm{d}t = 0, \tag{113}$$

where $\delta_{ba}\equiv\delta\mathbf{r}_b/\delta\mathbf{r}_a$. Integrating the velocity term by parts, simplifying the double summations using the Kronecker deltas and the antisymmetry of the kernel gradient, and assuming the perturbations $\delta\mathbf{r}_a$ are arbitrary, we find that the SPMHD equations of motion are given by

$$\frac{\mathrm{d}\mathbf{v}_a}{\mathrm{d}t} = -\sum_b m_b\left[\frac{P_a + \frac{1}{2\mu_0}B_a^2}{\Omega_a\rho_a^2}\nabla_a W_{ab}(h_a) + \frac{P_b + \frac{1}{2\mu_0}B_b^2}{\Omega_b\rho_b^2}\nabla_a W_{ab}(h_b)\right] + \frac{1}{\mu_0}\sum_b m_b\left[\frac{\mathbf{B}_a(\mathbf{B}_a\cdot\nabla_a W_{ab}(h_a))}{\Omega_a\rho_a^2} + \frac{\mathbf{B}_b(\mathbf{B}_b\cdot\nabla_a W_{ab}(h_b))}{\Omega_b\rho_b^2}\right]. \tag{114}$$

In tensor notation these can be written more compactly in the form

$$\frac{\mathrm{d}v_a^i}{\mathrm{d}t} = \sum_b m_b\left[\frac{S_a^{ij}}{\Omega_a\rho_a^2}\nabla_a^j W_{ab}(h_a) + \frac{S_b^{ij}}{\Omega_b\rho_b^2}\nabla_a^j W_{ab}(h_b)\right], \tag{115}$$

where S^{ij} is the MHD stress tensor, defined according to

$$S^{ij} \equiv -\left(P + \frac{1}{2\mu_0}B^2\right)\delta^{ij} + \frac{1}{\mu_0}B^iB^j. \tag{116}$$

As for the hydrodynamic case (Section 3) it is readily seen that the equations of motion conserve linear momentum exactly, due to the pairwise symmetry in the force. However, the MHD equations, unlike their hydrodynamic counterparts, do not exactly conserve angular momentum, since the perturbation to the magnetic field, (112) – and hence the anisotropic force term derived from it – is not invariant to rotations. It is interesting to note that the anisotropic magnetic force term in (114) derives entirely from the numerical representation of the induction Eq. (110), whilst the isotropic term derives purely from the magnetic energy term in the Lagrangian and the density perturbation (111).

7.4. Energy equation

The evolution equations for thermal energy and entropy – in the absence of dissipation – are identical to their hydrodynamic counterparts. The total energy evolution can be deduced from the Hamiltonian as in Section 3.4.2. The corresponding expression for MHD is given by

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \sum_a m_a\left[\mathbf{v}_a\cdot\frac{\mathrm{d}\mathbf{v}_a}{\mathrm{d}t} + \frac{\mathrm{d}u_a}{\mathrm{d}t} + \frac{1}{2}\frac{B_a^2}{\rho_a^2}\frac{d\rho_a}{\mathrm{d}t} + \mathbf{B}_a\cdot\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\mathbf{B}_a}{\rho_a}\right)\right]. \tag{117}$$

Using (115), (35), (56) and (110), it can be shown that the total energy evolves according to

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \sum_a m_a\sum_b m_b\left[\left(\frac{S^{ij}}{\Omega\rho^2}\right)_a v_b^i\nabla_a^j W_{ab}(h_a) + \left(\frac{S^{ij}}{\Omega\rho^2}\right)_b v_a^i\nabla_a^j W_{ab}(h_b)\right] = 0 \tag{118}$$

and is thus also conserved exactly. This further implies an evolution equation for the specific energy $e \equiv \frac{1}{2}v^2 + u + \frac{1}{2\mu_0}B^2/\rho$ of the form

$$\frac{\mathrm{d}e_a}{\mathrm{d}t} = \sum_b m_b\left[\left(\frac{S^{ij}}{\Omega\rho^2}\right)_a v_b^i\nabla_a^j W_{ab}(h_a) + \left(\frac{S^{ij}}{\Omega\rho^2}\right)_b v_a^i\nabla_a^j W_{ab}(h_b)\right]. \tag{119}$$

7.5. Interpretation of the Hamiltonian SPMHD equations

Having used the SPH forms of the continuity and induction equations in the form

$$\frac{\mathrm{d}\rho}{\mathrm{d}t} = -\rho\frac{\partial v^i}{\partial x^i}, \tag{120}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{B^i}{\rho}\right) = -\frac{B^j}{\rho}\frac{\partial v^i}{\partial x^j}, \tag{121}$$

it is straightforward to show (using 55) that our expressions (115) and (119) derived above are SPH representations of the MHD acceleration and energy equations in the form

$$\frac{\mathrm{d}v^i}{\mathrm{d}t} = \frac{1}{\rho}\frac{\partial S^{ij}}{\partial x^j}, \tag{122}$$

$$\frac{\mathrm{d}e}{\mathrm{d}t} = \frac{\partial(v^i S^{ij})}{\partial x^j}, \tag{123}$$

where S^{ij} is the MHD stress tensor given by (116). That we have explicitly *used* the induction equation to derive the equations of motion and energy is useful to our later discussion regarding what is a consistent formulation of monopole terms in the MHD equations (Section 10.1).

7.5.1. MHD Example 1: Advection of a current loop

Our first MHD example demonstrates that some problems that prove very difficult for Eulerian schemes are almost trivial in a Lagrangian scheme such as SPMHD. The problem involves the advection of a loop of current across the computational domain, was introduced by Gardiner and Stone [36] to test their ATHENA MHD code and presents a challenging problem for grid-based MHD schemes. The setup used here is identical to that in Rosswog and Price [93] and we refer the reader to that paper (or the NDSPMHD setup file) for full details. The current loop itself is given by the vector potential $A_z = A_0(R - \sqrt{x^2+y^2})$, with A_0 set to give a weak field (plasma β of 2×10^6) (here we compute and evolve the magnetic field, **B**). All particles in the domain ($-1 < x < 1$, $-0.5 < y < 0.5$) are given a constant initial velocity along the box diagonal, with the magnitude set such that $t = 1$ represents one crossing of the domain. The results are shown in Fig. 11, at $t = 0$ (left panel) and after 1000 (this is not a misprint!) crossings of the computational domain (right panel). In the absence of the explicit addition of resistivity terms, there is *no* change in the current or the magnetic energy and the advection is computed exactly.

8. The tensile instability in MHD

The rather large caveat to using the equations of motion derived in Section 7.3 is that in MHD the total stress can become negative, meaning that the force between particles can become attractive rather than repulsive and the equations will be unstable to the tensile instability discussed in Section 5.3. The regime of instability is evident if we consider the force in just one spatial dimension (and at constant h), given by

$$\frac{\mathrm{d}v_a^x}{\mathrm{d}t} = -\sum_b m_b\left[\frac{P_a - \frac{1}{2\mu_0}B_x^2}{\rho_a^2} + \frac{P_b - \frac{1}{2\mu_0}B_x^2}{\rho_b^2}\right]\frac{\mathrm{d}W_{ab}}{\mathrm{d}x}, \tag{124}$$

so the force will become attractive (along the field lines) when $\frac{1}{2}B^2/\mu_0 > P$, i.e., when magnetic pressure exceeds gas pressure. A more detailed stability analysis e.g. [7,68,69,76] confirms that this is also the criterion for instability in more than one spatial dimension. Thus, whilst for particular applications that remain in the regime where magnetic pressure is smaller than gas pressure it is possible to use the conservative formulation e.g. [28], in most cases stabilising the tensile instability is the first and most basic requirement for a stable SPMHD algorithm.

The physical reason for the MHD instability is that the momentum-conserving force corresponds to

$$\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} = \frac{\nabla\cdot\mathbf{S}}{\rho} \equiv -\frac{\nabla P}{\rho} + \frac{(\nabla\times\mathbf{B})\times\mathbf{B}}{\mu_0\rho} + \frac{\mathbf{B}\nabla\cdot\mathbf{B}}{\mu_0\rho}, \tag{125}$$

which differs from equations of motion written in terms of the Lorentz force (where $\mathbf{J} = \nabla\times\mathbf{B}/\mu_0$),

$$\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} = -\frac{\nabla P}{\rho} + \frac{\mathbf{J}\times\mathbf{B}}{\rho}, \tag{126}$$

by the "monopole term" in (125). This term gives a force directed along the magnetic field and proportional to the (numerically non-zero) divergence of the magnetic field, and is the source of the instability in the conservative formulation if it cannot be counteracted by pressure. As a result, Eq. 126 was used as the basis of a number of early SPMHD formulations in

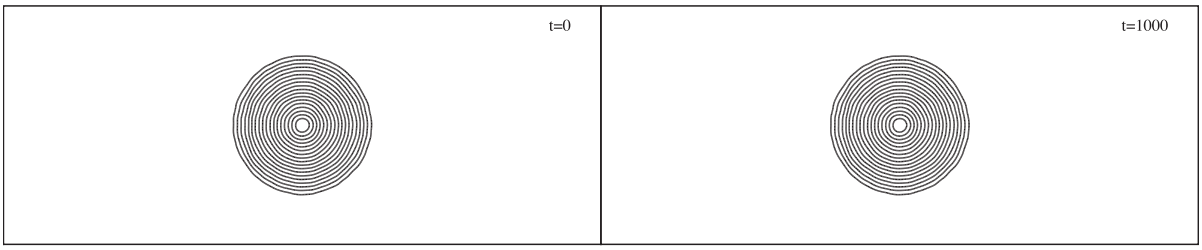

Fig. 11. Advection of a current loop across a 2D domain. Magnetic field lines are shown at $t = 0$ (left panel) and after 1000 crossings of the computational domain (right panel). In the absence of the explicit addition of resistivity terms, there is *no* change in the magnetic field and the advection is computed exactly in SPMHD, irrespective of the direction of propagation, numerical resolution or advection velocity.

which the force was simply computed using standard curl operators such as (54) or (79) e.g. [13,14,43,56]. However, the poor conservation properties of such formulations means that MHD shocks are not well captured.

Dealing with such 'source terms' is not an issue unique to SPMHD and requires careful consideration in all numerical MHD formulations (see Section 10.1). Of course, $\nabla \cdot \mathbf{B} = 0$ should be zero physically due to the non-existence of magnetic monopoles in the Universe, so both (126) and (125) are valid in the continuum limit. The problem is that the divergence term is not exactly zero numerically – so one is forced to make a choice. As noted by Toth [104], it is also not a simple matter of enforcing the $\nabla \cdot \mathbf{B} = 0$ constraint in some way, since what is necessary to achieve a force that is both conservative *and* exactly perpendicular to $\mathbf{B}$ is to constrain $\nabla \cdot \mathbf{B}$ to be zero *in the discretisation that the term appears in the force equation*[10]. In SPMHD this is equivalent to requiring both exact derivatives and exact conservation which, as discussed in Section 4.7, does not appear to be possible.

8.1. Fix 1: subtract a constant from the stress

The original paper by Phillips and Monaghan [76] proposed a simple fix involving a prior sweep over the particles to find the maximum (negative) stress, which would then simply be subtracted (as a constant) from the stress in the equations of motion, giving

$$\frac{\mathrm{d}v_a^i}{\mathrm{d}t} = \sum_b m_b \left[\frac{S_a^{ij} - S_{max}^{ij}}{\Omega_a \rho_a^2} \nabla_a^j W_{ab}(h_a) + \frac{S_b^{ij} - S_{max}^{ij}}{\Omega_b \rho_b^2} \nabla_a^j W_{ab}(h_b) \right], \tag{127}$$

which conserves momentum but not total energy. The caveats are that there is a computational cost involved to compute S_{max}^{ij} and if this term is large it can lead to unphysical effects in the simulation. On the other hand, this is a simple technique that removes the instability and has relatively few side effects provided the correction is small. It is particularly useful if, for example, the simulation is dominated by large (constant) external stresses, whereby explicitly subtracting the external component of the stress (e.g. due to an externally imposed magnetic field) can serve to stabilise the formulation.

8.2. Fix 2: use a more accurate but non-conservative gradient estimate in the anisotropic force

Morris [69] proposed a compromise approach whereby one retains conservative form in the isotropic part of the force, but adopts a more accurate but non-conservative derivative estimate (that vanishes when the stress is constant) in the anisotropic term. Adapted for variable smoothing lengths, the resultant force is given by Price and Monaghan [88], Price [81].

$$\frac{\mathrm{d}v_a^i}{\mathrm{d}t} = -\sum_b m_b \left[\frac{P_a + \frac{1}{2\mu_0} B_a^2}{\Omega_a \rho_a^2} \nabla_a^i W_{ab}(h_a) + \frac{P_b + \frac{1}{2\mu_0} B_b^2}{\Omega_b \rho_b^2} \nabla_a^i W_{ab}(h_b) \right] + \frac{1}{\mu_0} \sum_b m_b \frac{(B_i B_j)_b - (B_i B_j)_a}{\rho_a \rho_b} \overline{\nabla^j W}_{ab}, \tag{128}$$

where $\overline{\nabla W}_{ab} \equiv [\nabla W_{ab}(h_a) + \nabla W_{ab}(h_b)]/2$. The 'Morris approach' does not exactly conserve momentum or total energy, but the errors due to non-conservation are in practice quite small even on strong shock tube problems see [78]. Note, however, that retaining local conservation in the *isotropic* term is important for the reasons discussed in Section 5 and also in order to obtain the correct jump conditions on MHD shocks. The caveat to this approach is that the remaining non-conservation of momentum can become problematic, especially in simulations run for a long time period, where secular effects can accumulate and cause a loss of symmetry, and that it cannot be turned "off" in the otherwise stable weak-field regime.

[10] This was initially thought impossible by Toth [104], though Toth [105] later showed such a discretisation could be achieved for grid codes. However, there does not appear to be an equivalent formulation in SPMHD, since the tensile instability occurs even in one dimension where the divergence constraint can be trivially enforced using $B_x = const$, but $\partial B_x/\partial x \neq 0$ in the force equation (c.f. Eq. 124).

8.3. Fix 3: subtract the unphysical source term

Borve et al. [6] suggested an approach whereby conservation form is retained, but the monopole source term is explicitly subtracted. Extending their expression to incorporate variable smoothing length terms, the corrected equations of motion are given by

$$\frac{\mathrm{d}v_a^i}{\mathrm{d}t} = \sum_b m_b \left[\frac{S_a^{ij}}{\Omega_a \rho_a^2} \nabla_a^j W_{ab}(h_a) + \frac{S_b^{ij}}{\Omega_b \rho_b^2} \nabla_a^j W_{ab}(h_b) \right] - \widehat{B}_a^i \sum_b m_b \left[\frac{B_a^j}{\Omega_a \rho_a^2} \nabla_a^j W_{ab}(h_a) + \frac{B_b^j}{\Omega_b \rho_b^2} \nabla_a^j W_{ab}(h_b) \right]. \quad (129)$$

The corrective term with $\widehat{B}_a^i = B_a^i$ is equivalent to the source term added to the momentum equation in the eight-wave Riemann solver [77] – with the key point from an SPH standpoint being that the discretisation used for the correction term is identical to that in which the divergence term occurs in the force equation. Borve et al. [7] also showed that it is not necessarily required to always choose $\widehat{B}_a^i = B_a^i$ in order to achieve stability, showing in particular that $\widehat{B}_a^i = \frac{1}{2} B_a^i$ is sufficient. Borve et al. [8] give a more general method for setting $\widehat{B}^i$ such that – amongst other things – no correction is applied when the gas pressure exceeds the magnetic pressure (we refer the reader to their paper for more details). Whilst adding the correction term violates exact conservation of momentum and energy, it does so only insofar as the divergence term in (129) is non-zero. Like the Morris approach, (129) is a good general method for removing the tensile instability in SPMHD – but has the advantage of being able to be switched *off* in the regime where the conservative formulation is stable.

8.4. Other methods

Finally, other methods have been proposed for dealing with the tensile instability in other physical problems, such as elastic dynamics. In particular Monaghan [61] suggested explicitly adding a short range repulsive force between particles to counteract the unphysical attractive force, similar to what would occur in atoms. Price and Monaghan [86] initially applied this approach to the MHD equations and it was shown by Price [78] to be equivalent to a modification of the kernel gradient in the anisotropic part of the MHD force (the second term in Eq. (114)). However it proved problematic in highly compressible calculations where the smoothing lengths could vary dramatically (leading to a difficulty in defining "short range") and at large negative stress was found to result in large errors in the numerical sound speed see [78]. It was therefore abandoned by Price and Monaghan [88] in favour of either the Morris or Borve et al. [6] approaches.

8.4.1. MHD Example 2: Circularly polarized Alfvén wave in 2.5D

An example of the tensile instability in practice is shown in Fig. 12, which shows the particle distribution (left panels in each figure) and x-component of the magnetic field (right panels in each figure) in the 2.5D circularly polarized Alfvén wave test from Toth [104]. The initial conditions [88,104] are $\rho = 1$, $P = 0.1$, $v_\parallel = 0$, $B_\parallel = 1$, $v_\perp = B_\perp = 0.1\sin(2\pi r_\parallel)$ and $v_z = B_z = 0.1\cos(2\pi r_\parallel)$ with $\gamma = 5/3$ (where $r_\parallel = x\cos\theta + y\sin\theta$), with the wave vector directed at an angle of $\theta = 30^\circ$ with respect to the x-axis in the computational domain $0 < x < 1/\cos\theta$, $0 < y < 1/\sin\theta$ with periodic boundary conditions.

Using a conservative SPMHD formulation (left figure, shown at $t = 1$) whole lines of particles can be seen to have attracted each other in the direction of $B_\parallel$, which proceeds to destroy the calculation at later times. With either the Morris or Borve

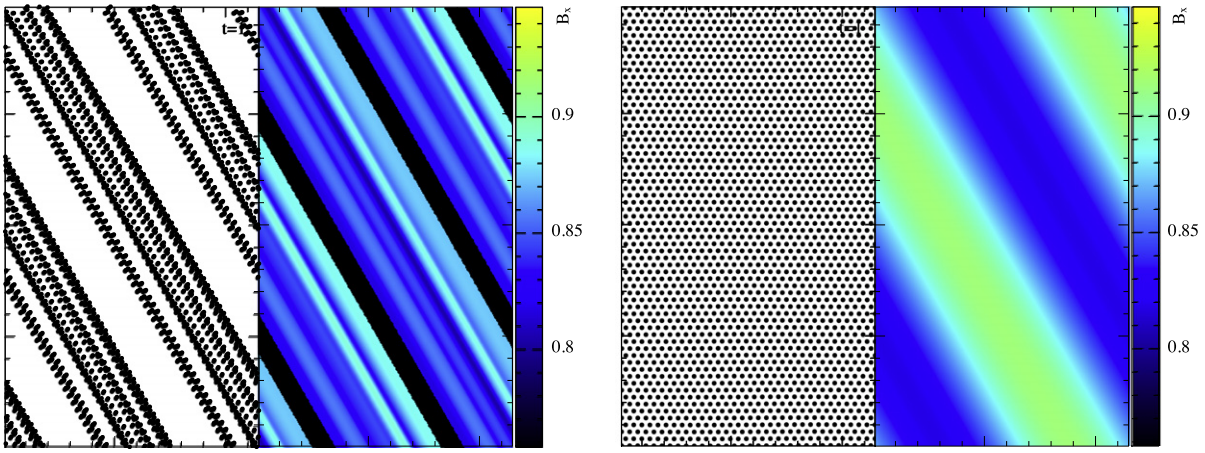

Fig. 12. The tensile instability in spmhd: the figures show the particle distribution (left panel in each figure) and x-component of the magnetic field (right panel in each figure) in the 2.5D circularly polarized Alfvén wave test using the (unstable) conservative SPMHD force (left figure) and with a stable formulation (right figure), shown after 1 wave crossing time. Untreated, the MHD tensile instability results in a catastrophic clumping of particles along the magnetic field lines (left figure) which proceeds to destroy the calculation. Physically, it can be attributed to non-zero divergence terms in the conservative form of the MHD force, and can be stabilised by explicitly subtracting the unphysical 'source term' (right figure).

et al. [6] approach (right figure, also shown at $t = 1$) the particle arrangement – and hence the wave – is stable and propagates correctly.

9. Dissipation terms and shocks in SPMHD

The second important issue for SPMHD is the formulation of dissipation terms appropriate for MHD shocks. We can follow the same general principle as in Section 6.3.2, starting with (102) as the basic equation, except that in MHD the energy variable is given by

$$e_a^* = \frac{1}{2}\alpha v_{sig}(\mathbf{v}_a \cdot \hat{\mathbf{r}}_{ab})^2 + \alpha_u v_{sig}^u u_a + \alpha_B v_{sig}^B \frac{B_a^2}{2\mu_0 \bar{\rho}_{ab}}, \tag{130}$$

and the signal velocity in the kinetic energy term is generalised to

$$v_{sig} = \begin{cases} \frac{1}{2}[v_a + v_b - \beta \mathbf{v}_{ab} \cdot \hat{\mathbf{r}}_{ab}]; & \mathbf{v}_{ab} \cdot \hat{\mathbf{r}}_{ab} \leqslant 0; \\ 0; & \mathbf{v}_{ab} \cdot \hat{\mathbf{r}}_{ab} > 0, \end{cases} \tag{131}$$

where v is the fastest wave speed for MHD (i.e., the fast MHD mode), given by

$$v_a = \frac{1}{\sqrt{2}}\left[\left(c_{s,a}^2 + \frac{B_a^2}{\mu_0\rho_a}\right) + \sqrt{\left(c_{s,a}^2 + \frac{B_a^2}{\mu_0\rho_a}\right)^2 - 4\frac{c_{s,a}^2(\mathbf{B}_a \cdot \hat{\mathbf{r}}_{ab})^2}{\mu_0\rho_a}}\right]^{1/2}. \tag{132}$$

The key constraint in deriving dissipative terms in the other equations is that the contribution to the entropy from the dissipative terms in the total energy equation must be positive definite in order to satisfy the second law of thermodynamics. For the kinetic energy term this is satisfied provided a term is added to the acceleration equation (i.e., Eq. 101), giving a positive definite contribution to the thermal energy (Eq. 104). It may also be shown that the thermal energy term results in a positive definite contribution to the entropy (Appendix B of Price and Monaghan [86]). For the magnetic energy term, positivity means that the overall dissipative contribution to the thermal energy equation is given by

$$\left(\frac{\mathrm{d}u}{\mathrm{d}t}\right)_{diss} = -\sum_b \frac{m_b}{\bar{\rho}_{ab}}\left[\frac{1}{2}\alpha v_{sig}(\mathbf{v}_a - \mathbf{v}_b)^2 + \alpha_u v_{sig}^u(u_a - u_b) + v_{sig}^B \frac{\alpha_B}{2\mu_0\bar{\rho}_{ab}}(\mathbf{B}_a - \mathbf{B}_b)^2\right]\overline{F_{ab}}, \tag{133}$$

which is positive definite because F_{ab} is negative definite (Fig. 2) and the terms inside the brackets are positive. Using (133), we can deduce the term which must arise in the induction equation using

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{B^2}{2\mu_0\rho}\right)_{diss} = \left(\frac{\mathrm{d}e}{\mathrm{d}t}\right)_{diss} - \mathbf{v}\cdot\left(\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t}\right)_{diss} - \left(\frac{\mathrm{d}u}{\mathrm{d}t}\right)_{diss}, \tag{134}$$

giving a term in the induction equation of the form

$$\left(\frac{\mathrm{d}\mathbf{B}_a}{\mathrm{d}t}\right)_{diss} = \rho_a \sum_b m_b \frac{\alpha_B v_{sig}^B}{\bar{\rho}_{ab}^2}(\mathbf{B}_a - \mathbf{B}_b)\overline{F}_{ab}, \tag{135}$$

or, evolving $\mathbf{B}/\rho$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\mathbf{B}_a}{\rho_a}\right)_{diss} = \sum_b m_b \frac{\alpha_B v_{sig}^B}{\bar{\rho}_{ab}^2}(\mathbf{B}_a - \mathbf{B}_b)\overline{F}_{ab}. \tag{136}$$

The interpretation of this term is straightforward using (95): We have derived an artificial resistivity term

$$\left(\frac{\mathrm{d}\mathbf{B}_a}{\mathrm{d}t}\right)_{diss} = \eta_M \nabla^2 \mathbf{B}_a; \quad \eta_M \approx \frac{1}{2}\alpha_B v_{sig}^B |r_{ab}|. \tag{137}$$

For the artificial resistivity it is also better to use a signal velocity that differs from the one used in the viscosity. A simple choice (used in NDSPMHD) is to use an averaged Alfvén speed, e.g.

$$v_{sig}^B = \frac{1}{2}\sqrt{v_{A,a}^2 + v_{A,b}^2}; \quad v_A^2 = \frac{B^2}{\mu_0\rho}. \tag{138}$$

Price and Monaghan [86,88] also show how an artificial resistivity can be derived that uses only components of the magnetic field perpendicular to the line of sight. However, Price and Monaghan [88] found that the version using the jump in total energy (Eqs. 130, 135) performed better in practice. The parameter α_B can also be evolved using a switch similar to (105), where Price and Monaghan [88] suggest a source term of the form

$$\mathcal{S} = \max\left(\frac{|\nabla \times \mathbf{B}|}{\sqrt{\mu_0 \rho}}, \frac{|\nabla \cdot \mathbf{B}|}{\sqrt{\mu_0 \rho}}\right). \tag{139}$$

Note that artificial thermal conductivity and resistivity should in general be applied regardless of whether or not the particle pair is approaching or receding.

9.1. MHD Example 3: shock tube problems in MHD

The application of specific dissipative terms for shock-capturing is illustrated in Fig. 13, which shows a "1.75D" MHD shock tube problem (i.e., 1D but with 3D magnetic and velocity fields) from Ryu and Jones [94], using 800 particles and the standard M_4 cubic spline kernel. The setup is $(\rho, P, v_x, v_y, v_z, B_y, B_z) = [1.08, 0.95, 1.2, 0.01, 0.5, 3.6/(4\pi)^{1/2}, 2/(4\pi)^{1/2}]$ for $x < 0$, whilst for $x \geqslant 0$ we have $(\rho, P, v_x, v_y, v_z, B_y, B_z) = [1, 1, 0, 0, 0, 4/(4\pi)^{1/2}, 2/(4\pi)^{1/2}]$, with $B_x = 2/(4\pi)^{1/2}$ and $\gamma = 5/3$. The main point is that this shows all 7 possible discontinuities in MHD: Three waves (slow, Alfvén and fast modes) propagating in each direction, plus the contact discontinuity. Since each requires treatment, we require artificial viscosity, thermal conductivity and resistivity to treat the jumps in velocity, thermal energy and magnetic field, respectively.

Fig. 14 shows a 2D version of the Brio and Wu [10] shock tube test, considered a standard test for MHD codes e.g. [4,22,94,103], using 800×24 particles and the M_6 quintic kernel. Initially $(\rho, P, v_x, v_y, B_y) = [1, 1, 0, 0, 1]$ for $x < 0$, with $(\rho, P, v_x, v_y, B_y) = [0.125, 0.1, 0, 0, -1]$ for $x \geqslant 0$ with $B_x = 0.75$ everywhere and $\gamma = 2.0$. Profiles of all SPMHD particles are plotted at $t = 0.1$ whilst the numerical solution from Balsara [4] is given by the solid (red) line. The solution is comparable to that achievable in 1D [93, e.g.] – at this resolution the only noticeable deviation from the Balsara solution is the slight mismatch in thermal energy at $x \approx 0.15$–0.3, a result of the low spatial resolution ($10\times$ lower than in the $x < 0$ region) in this low density region.

10. The divergence constraint in SPMHD

The third major issue, common to all numerical MHD codes, is how to enforce the $\nabla \cdot \mathbf{B} = 0$ constraint. The problem arises because the constraint occurs only as an *initial condition* in the MHD equations, since from the induction equation we have

$$\frac{\partial \mathbf{B}}{\partial t} = -\nabla \times (\mathbf{v} \times \mathbf{B}); \quad \frac{\partial}{\partial t}(\nabla \cdot \mathbf{B}) = 0, \tag{140}$$

The problem is that the truncation error in the numerical solution means that such a condition cannot be maintained indefinitely. What to do about it falls into three general approaches: "ignore", "clean" or "prevent" the numerical growth

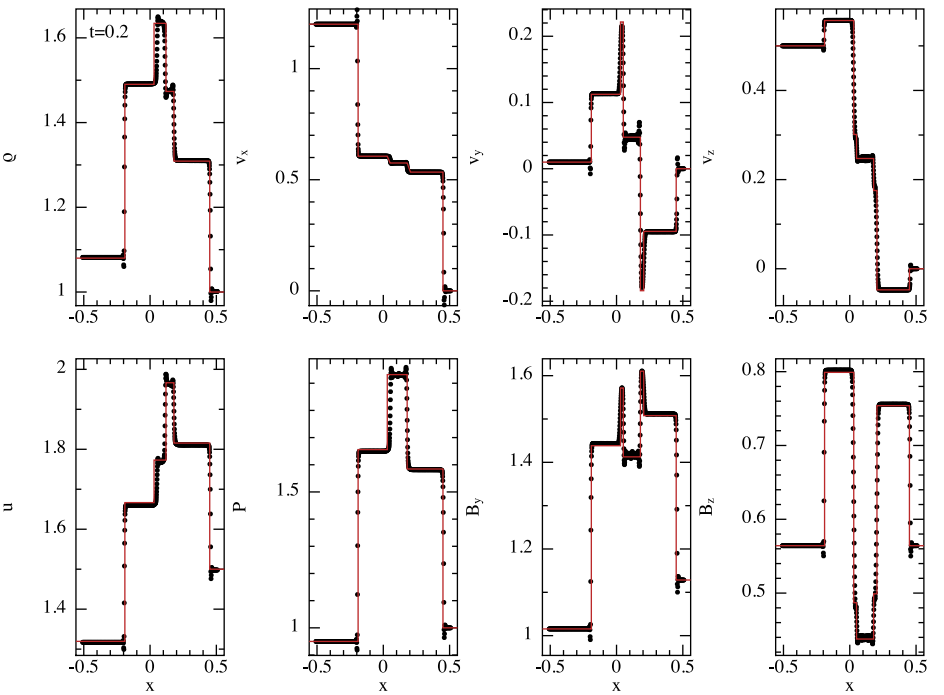

Fig. 13. A "1.75D" MHD shock tube problem, showing the formation of 7 discontinuities: a slow, Alfvén and fast wave propagating in each direction, plus the contact discontinuity. Capture of all of these discontinuities requires the application of artificial viscosity, thermal conductivity and resistivity to treat jumps in v, u and B, respectively. The M_4 cubic spline kernel has been used.

of monopole terms. We will discuss each of these in the context of SPMHD, but first turn to the issue of formulating consistent equations given that $\nabla \cdot \mathbf{B} \neq 0$.

10.1. $\nabla \cdot \boldsymbol{B} \neq 0$: Monopole terms in the evolution equations

We have already touched on the 'source terms' issue in the context of the tensile instability in SPMHD, which arises from the fact that the MHD force written as the divergence of a stress tensor differs from the exactly-perpendicular-to-B Lorentz force by a term proportional to $\nabla \cdot \mathbf{B}$ – and that this term does not vanish even in the trivial case of 1D where $B_x = const$. A similar issue arises in the induction Eq. (140), which in "conservative form" is given by

$$\frac{\partial \mathbf{B}}{\partial t} = \nabla \cdot (\mathbf{vB} - \mathbf{Bv}) = -(\mathbf{v} \cdot \nabla)\mathbf{B} + (\mathbf{B} \cdot \nabla)\mathbf{v} - \mathbf{B}(\nabla \cdot \mathbf{v}) + \mathbf{v}(\nabla \cdot \mathbf{B}), \tag{141}$$

identical to (140), and conserving the *volume* integral of the flux $\int \mathbf{B} \mathrm{d}V$. However, if the monopole term $\mathbf{v}(\nabla \cdot \mathbf{B})$ is neglected, then we have (using the Lagrangian time derivative)

$$\frac{\mathrm{d}\mathbf{B}}{\mathrm{d}t} = (\mathbf{B} \cdot \nabla)\mathbf{v} - \mathbf{B}(\nabla \cdot \mathbf{v}), \tag{142}$$

taking the divergence of which shows that the monopoles evolve according to an equation similar to the continuity equation for density

$$\frac{\partial}{\partial t}(\nabla \cdot \mathbf{B}) + \nabla \cdot (\mathbf{v} \nabla \cdot \mathbf{B}), \tag{143}$$

meaning that the volume integral $\int_V (\nabla \cdot \mathbf{B}) \mathrm{d}V$, and by Green's theorem the *surface* integral of the flux, $\oint_S \mathbf{B} \cdot \mathrm{d}\mathbf{S}$ is conserved. So adopting (142) represents a 'monopole conserving' form of the induction equation, such that the surface flux integral will be conserved even if the divergence errors are non-zero. On the basis of this idea Powell et al. [77] suggested the 8-wave formulation, with the induction equation given by (142) and the momentum and energy equations given by

$$\frac{\mathrm{d}v^i}{\mathrm{d}t} = \frac{1}{\rho}\frac{\partial S^{ij}}{\partial x^j} - \frac{B^i}{\rho}\frac{\partial B^j}{\partial x^j}, \tag{144}$$

$$\frac{\mathrm{d}e}{\mathrm{d}t} = \frac{1}{\rho}\frac{\partial (v_i S^{ij})}{\partial x^j} - \frac{v_i B^i}{\rho}\frac{\partial B^j}{\partial x^j}. \tag{145}$$

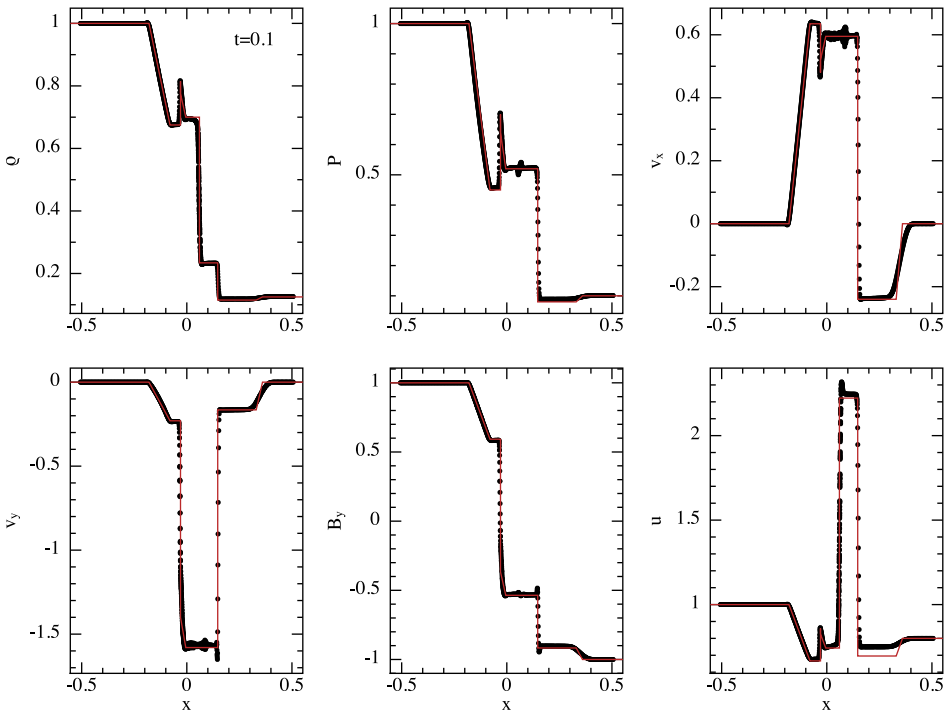

Fig. 14. The Brio–Wu MHD shock tube problem in 2D, computed using 800 × 24 particles and the M_6 quintic kernel. With the quintic, results comparable to the 1D solution can be obtained. The numerical solution from Balsara [4] is given by the solid (red) line. All particles are plotted. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

However, Janhunen [46] and Dellar [25] have strongly argued that a consistent formulation in the presence of monopoles, whilst adopting (142) instead of (140), should nevertheless still conserve linear (but not angular) momentum and energy. For SPMHD we have already proved that this is the case because we derived the momentum-conserving force (115) from the surface-flux-and-monopole conserving induction Eq. (109). This is a somewhat moot point in practice though, since we require the subtraction of the source term in the momentum equation anyway in order to stabilise the tensile instability (Section 8.3), giving essentially the 8-wave formulation.

10.2. $\nabla \cdot \mathbf{B} \approx 0$: The "ignore" approach

The simplest approach to the divergence constraint is to do nothing at all. That is, simply monitor the divergence error and "trust" the simulations if it remains small. In SPMHD it is usual to monitor the divergence error using the dimensionless quantity

$$\frac{h\nabla \cdot \mathbf{B}}{|B|}, \tag{146}$$

typically hoping that it remains smaller than a few percent. For a number of problems, including many of the test problems shown in this paper, this approach works surprisingly well. There is also the question of what one means by the "divergence error", since (146) can be computed using a variety of operators (Section 4.4) and the 'stencil' in which one measures divergence changes in SPH as soon as one changes the particle positions. However there are clearly many problems where this approach is not sufficient.

10.3. $\nabla \cdot \mathbf{B} \to 0$: The "clean" approach

Divergence cleaning in SPMHD also faces the problem of defining what $\nabla \cdot \mathbf{B} = 0$ actually means. Clearly from a stability point of view it would be desirable for the divergence to be zero in the discretisation in which it enters the force equation, but we have already seen that this is not possible even in 1D because the conservative SPH derivative does not vanish for constant functions. However, it may be hoped that cleaning in a particular discretisation will at least produce solenoidal fields to the order of the truncation error [i.e., $\mathcal{O}(h^2)$]. Approaches to divergence cleaning fall into three general categories based on the solution to an elliptic, parabolic or hyperbolic partial differential equation.

10.3.1. Elliptic cleaning

Elliptic cleaning corresponds to using projection methods: Solving either for a correction term,

$$\mathbf{B} = \mathbf{B}^* - \nabla\phi; \quad \nabla^2\phi = \nabla \cdot \mathbf{B}^*, \tag{147}$$

or alternatively for the physical (solenoidal) component of the field,

$$\mathbf{B} = \nabla \times \mathbf{A}; \quad \nabla^2\mathbf{A} = -\nabla \times \mathbf{B}^*. \tag{148}$$

SPH methods for both of the above have been discussed by Price and Monaghan [88]. The main issue is that, if one simply uses the Green's function solution – analogous to the computation of gravitational forces – with the source function computed with a chosen div or curl operator, the projection is only approximate (that is, the Poisson equation is not discretely satisfied[11]). For 3D simulations with particles on individual timesteps, it is also impractical and highly inefficient to compute a global projection every n timesteps.

10.3.2. Parabolic cleaning

A parabolic approach corresponds to adding a $\nabla \cdot \mathbf{B}$ diffusion term to the induction equation, i.e.,

$$\left(\frac{\mathrm{d}\mathbf{B}}{\mathrm{d}t}\right)_{clean} = \nabla \cdot (\eta_c \nabla \cdot \mathbf{B}). \tag{149}$$

Having formulated our artificial dissipative terms using the jump in total magnetic energy (130) means that the artificial resistivity already contains a term of this form (see 137), so applying artificial resistivity (with a switch that responds to the divergence, Eq. 139) corresponds to a form of parabolic divergence cleaning. This is partly why the "ignore" approach is not as bad as it sounds, and has successfully used for a number of problems e.g. [49,12,29].

10.3.3. Hyperbolic/parabolic cleaning

Dedner et al. [23] suggested a hyperbolic cleaning method for the MHD equations, given by

$$\frac{\mathrm{d}\mathbf{B}}{\mathrm{d}t} = -\mathbf{B}(\nabla \cdot \mathbf{v}) + (\mathbf{B} \cdot \nabla)\mathbf{v} - \nabla\psi, \tag{150}$$

[11] The approximate nature of the projection is related to the question of how one should "soften" the Green's function solution, similar to the case for gravity. What is required is a generalisation of the softening formulation given by Price and Monaghan [89] which shows how the gravitational Poisson equation can be discretely satisfied if the SPH density is used as the source term, so this would be an approach worth pursuing.

where ψ is an additional variable that evolves according to [88]

$$\frac{\mathrm{d}\psi}{\mathrm{d}t} = -c_h^2(\nabla\cdot\mathbf{B}) - \frac{\psi}{\tau}, \tag{151}$$

where for SPH it is natural to use $c_h \equiv v_{sig}$ and $\tau \equiv \alpha_h h/v_{sig}$ where $\alpha_h \in [0,1]$ is a dimensionless parameter. The first term in (151) corresponds to hyperbolic propagation of the "divergence wave" (at speed c_h), whilst ψ/τ is a parabolic decay term (with decay time τ). This can be shown combining the equations (assuming $\mathbf{v} = const$) to yield a wave equation,

$$\frac{1}{c_h^2}\frac{\partial^2\psi}{\partial t^2} - \nabla^2\psi + \tau\frac{\partial\psi}{\partial t} = 0. \tag{152}$$

As reported by Price and Monaghan [88] and others e.g. [57], the original Dedner et al. [23] formulation contained a parameter that is *not* dimensionless, but that is equivalent to the wavelength of critical damping λ_c in the parabolic decay term (evident by writing $\tau \equiv \lambda_c/c_h$ in the above). Indeed, Price and Monaghan [88] found that whilst using (150), (151) could be effective at cleaning well-resolved divergence errors ($\lambda_c \gg h$), it was found to have little or no effect when used in "real" problems (where $\lambda_c \sim h$), giving at most a factor of $\sim$2 reduction in $\nabla\cdot\mathbf{B}_{\max}$. It was also found that a poor choice of parameters could *increase* the divergence error substantially via constructive interference of divergence waves.

More recently Stasyszyn and Dolag [29] report better success using this method, by two adaptations: (i) accounting for the magnetic energy dissipation in the energy equation, and (ii) implementation of a limiter to control the amount by which the magnetic field can change in a single timestep (Dolag 2010, priv. commun.) Nevertheless, it does not appear that the method is being used on real problems e.g. [49], suggesting that it remains ineffective.

10.4. $\nabla\cdot\mathbf{B} = 0$: The "prevent" approach

The third approach is to try to 'prevent' divergence errors altogether by formulating the MHD equations such that $\nabla\cdot\mathbf{B} = 0$ is satisfied by construction. Unfortunately, the approach most successfully adopted in grid based schemes – the constrained transport method [32,36] – cannot be directly applied in an SPH context because it requires the computation of surface rather than volume integrals.

10.4.1. Euler potentials

Perhaps the closest to a 'constrained transport' approach in SPMHD is the formulation in terms of Euler potentials [101,102] (also referred to as "Clebsch variables", e.g. by Phillips and Monaghan [76]), α_E and β_E, where the magnetic field is expressed as

$$\mathbf{B} = \nabla\alpha_E \times \nabla\beta_E. \tag{153}$$

Taking the divergence shows that $\nabla\cdot\mathbf{B} = 0$ is satisfied by construction. A further advantage is that for ideal MHD, the induction Eq. (109) becomes simply

$$\frac{\mathrm{d}\alpha_E}{\mathrm{d}t} = 0, \quad \frac{\mathrm{d}\beta_E}{\mathrm{d}t} = 0, \tag{154}$$

which corresponds physically to the advection of magnetic field lines by Lagrangian particles [100]. That is, the Euler potentials reflect the physical fact that the field lines are 'frozen' to the fluid in ideal MHD and are therefore a natural description for a Lagrangian scheme. For SPH it is straightforward to write down expressions for (153) using any of our standard first derivative operators (Section 4.4) and to simply use the resultant $\mathbf{B}$ in the SPMHD equations of motion – though this will give only approximate conservation of energy, see Price [81]. However, in order to capture shocks, some dissipation in the magnetic field is required. For this reason Price and Bate [82] and Rosswog and Price [93] introduced dissipation terms of the form

$$\left(\frac{\mathrm{d}\alpha_E^a}{\mathrm{d}t}\right)_{diss} = \sum_b \frac{m_b}{\bar{\rho}_{ab}}\alpha_B v_{sig}^B(\alpha_E^a - \alpha_E^b)\overline{F_{ab}}; \quad \left(\frac{\mathrm{d}\beta_E^a}{\mathrm{d}t}\right)_{diss} = \sum_b \frac{m_b}{\bar{\rho}_{ab}}\alpha_B v_{sig}^B\left(\beta_E^a - \beta_E^b\right)\overline{F_{ab}}, \tag{155}$$

corresponding to the continuum equations (assuming $\eta_{\mathrm{M}} \approx const$)

$$\frac{\mathrm{d}\alpha_E}{\mathrm{d}t} \approx \eta_{\mathrm{M}}\nabla^2\alpha; \quad \frac{\mathrm{d}\beta_E}{\mathrm{d}t} = \eta_{\mathrm{M}}\nabla^2\beta, \tag{156}$$

where η_{M} is given by (137). However it is clear that this is *not* a consistent formulation of (physical) resistivity for the Euler potentials, since we have neglected mixed terms that are of comparable magnitude to those in (156) e.g. [9]. It is also difficult to ensure that the contribution to the entropy is positive definite,[12] although Price [81] shows how the latter requirement can be achieved for the vector potential, and similar formulations should be used in place of (155) for the Euler potentials.

Nevertheless, the control of the divergence errors means that the Euler potentials have found successful application to a range of problems e.g. [82,27,50], most notably in star formation e.g. [82–84] since one is able to form stars in the presence of

[12] Rosswog and Price [93] simply used Eq. (133) with $\mathbf{B}$ calculated from (153) such that the contribution is guaranteed to be positive. However, the total energy is only approximately conserved by this procedure. See Price [81].

magnetic fields without blow-up of the numerical solution due to divergence errors.[13] The caveat is that the Euler potentials do not capture the full physics of MHD, since the Helicity $\mathbf{A}\cdot\mathbf{B}$ is identically zero, meaning that topologically non-trivial fields cannot be represented (or by corollary, created during a simulation) using (153), (154). For example, the Euler potentials cannot be used to follow multiple complete windings of a magnetic field in a rotating fluid, a corollary of the fact that the potentials are simply advected by the particles (Eq. 154), so reconstruction of the field via (153) requires a one-to-one mapping between the initial and final particle positions. These restrictions are easily demonstrated by simple test problems [9] and mean in practice that one is limited to studying problems with initially simple field geometries, and that important field winding processes (i.e., any dynamo action) are missed. On the other hand, it may be possible to formulate a more general Euler potentials description without such limitations.

10.4.2. The vector potential

Another possibility for a divergence-free formulation without the restrictions of the Euler potentials is to simply employ the vector potential $\mathbf{B}=\nabla\times\mathbf{A}$ directly. Indeed in 2D they are exactly equivalent, since for a 2D field $\alpha_E\equiv A_z(x,y)$ and $\beta_E\equiv z$. The main disadvantage of the vector potential in 3D is that the evolution equation is complicated and requires an appropriate gauge choice. For example, setting the scalar potential $\phi=\mathbf{v}\cdot\mathbf{A}$ one can obtain a Galilean-covariant evolution equation for $\mathbf{A}$ of the form [81]

$$\frac{\mathrm{d}\mathbf{A}}{\mathrm{d}t}=-\mathbf{A}\times(\nabla\times\mathbf{v})-(\mathbf{A}\cdot\nabla)\mathbf{v}+\mathbf{v}\times\mathbf{B}_{ext}=-A^j\nabla v^j+\mathbf{v}\times\mathbf{B}_{ext}. \tag{157}$$

Using this, one can proceed to derive an SPMHD algorithm for the vector potential from the Lagrangian (106). In particular, Price [81] shows that by writing the curl according to

$$\mathbf{B}_a=(\nabla\times\mathbf{A})_a+\mathbf{B}_{ext}=\frac{1}{\rho_a}\sum_b m_b(\mathbf{A}_a-\mathbf{A}_b)\times\nabla_a W_{ab}+\mathbf{B}_{ext}, \tag{158}$$

the effective evolution equation for $\mathbf{B}$ is given by

$$\frac{\mathrm{d}\mathbf{B}_a}{\mathrm{d}t}=\frac{1}{\rho_a}\sum_b m_b(\mathbf{A}_a-\mathbf{A}_b)\times[(\mathbf{v}_a-\mathbf{v}_b)\cdot\nabla]\nabla_a W_{ab}+\frac{1}{\rho_a}\sum_b m_b\left(\frac{\mathrm{d}\mathbf{A}_a}{\mathrm{d}t}-\frac{\mathrm{d}\mathbf{A}_b}{\mathrm{d}t}\right)\times\nabla_a W_{ab}-\frac{\mathbf{B}_{int}}{\rho_a}\frac{\mathrm{d}\rho_a}{\mathrm{d}t}. \tag{159}$$

Thus, writing the evolution equation for $\mathbf{A}$, (157), in the form

$$\frac{\mathrm{d}\mathbf{A}_a}{\mathrm{d}t}=\frac{A^a_j}{\rho_a}\sum_b m_b(v^j_a-v^j_b)\nabla_a W_{ab}+(\mathbf{v}\times\mathbf{B}_{ext})_a, \tag{160}$$

one can self-consistently derive equations of motion for the vector potential from (108) by writing the perturbation to $\delta\mathbf{B}$ in terms of $\delta\mathbf{A}$ using (159) and $\delta\mathbf{A}$ in terms of $\delta\mathbf{r}$ using (160), giving

$$\begin{aligned}\frac{\mathrm{d}\mathbf{v}_a}{\mathrm{d}t}=&-\sum_b m_b\left(\frac{P_a-\frac{3}{2\mu_0}B_a^2}{\rho_a^2}+\frac{P_b-\frac{3}{2\mu_0}B_b^2}{\rho_b^2}\right)\nabla_a W_{ab}-\frac{1}{\mu_0}\sum_b m_b\left\{\left(\frac{\mathbf{B}_a}{\rho_a^2}+\frac{\mathbf{B}_b}{\rho_b^2}\right)\cdot[(\mathbf{A}_a-\mathbf{A}_b)\times\nabla]\right\}\nabla_a W_{ab}\\&-\frac{2}{\mu_0}\sum_b m_b\left(\frac{\mathbf{B}_a}{\rho_a^2}+\frac{\mathbf{B}_b}{\rho_b^2}\right)\cdot\mathbf{B}_{ext}\nabla_a W_{ab}+\frac{1}{\mu_0}\sum_b m_b\left(\frac{\mathbf{B}_a}{\rho_a^2}+\frac{\mathbf{B}_b}{\rho_b^2}\right)\mathbf{B}_{ext}\cdot\nabla_a W_{ab}\\&-\sum_b m_b\left[\frac{\mathbf{A}_a}{\rho_a^2}\mathbf{J}_a\cdot\nabla_a W_{ab}+\frac{\mathbf{A}_b}{\rho_b^2}\mathbf{J}_b\cdot\nabla_a W_{ab}\right],\end{aligned} \tag{161}$$

where $\mathbf{J}$ the magnetic current is computed using the symmetric curl operator (the conjugate to 158),

$$\mathbf{J}_a\equiv\frac{(\nabla\times\mathbf{B})_a}{\mu_0}\equiv-\frac{\rho_a}{\mu_0}\sum_b m_b\left[\frac{\mathbf{B}_a}{\rho_a^2}+\frac{\mathbf{B}_b}{\rho_b^2}\right]\times\nabla_a W_{ab}. \tag{162}$$

The proof that the above equations do indeed translate into an continuum expression of the conservative MHD equations, as well as a more general version taking full account of smoothing length gradient terms is given by Price [81]. However, the major flaw, from simple inspection of (161), is that the isotropic (first) term will give a negative pressure – and thus the tensile instability – whenever $3B^2/(2\mu_0)>P$, a much more restrictive regime than for the usual SPMHD equations (see Section 8). This is verified in test problems, but unlike the standard SPMHD instability, proves difficult to stabilise without significantly affecting the solution. A second instability was also found to arise purely due to the evolution of $\mathbf{A}$ according to (157), resulting in unconstrained growth of vector potential components, most likely related to the need to explicitly enforce a gauge condition alongside the evolution of the vector potential. Price [81] therefore concluded that using the vector potential was not a viable approach for SPMHD.

[13] Price and Rosswog [90] also employed the Euler potentials to simulate magnetic field growth during Neutron star mergers. Whilst similar results were found using a $\mathbf{B}$-based formulation, the simulations with Euler potentials were later found to be incorrect, showing exaggerated field growth because the boundary conditions on the potentials for the stars were not correctly accounted for.

11. Summary

In summary, we have given an overview of SPH methodology, starting with the density estimate as the basis of all SPH formulations (Section 2) and showing how the equations of motion and energy can be self-consistently derived from the density estimate using a variational principle (Section 3). Kernel interpolation theory has been introduced mainly as a way of interpreting the SPH equations, and we have discussed how linear error analysis can be used to construct more accurate and very general derivative operators (Section 4). In Section 5 the importance of local conservation was highlighted with respect to maintaining a regular particle distribution and thus accurate derivatives, giving us an understanding of how the tensile instability arises in SPMHD and why one should be careful in setting the ratio of smoothing length to particle spacing. Second derivatives in SPH were discussed in Section 6, mainly as a way of formulating dissipative terms necessary to capture shocks and other kinds of discontinuities. In the second half of the paper, we have shown how SPMHD, like SPH, can also be formulated from a variational principle (Section 7) and have addressed the three main issues with regards to the accuracy of SPMHD: the tensile instability (Section 8), the formulation of shock-capturing dissipation terms (Section 9) and the enforcement of the divergence-free condition on the magnetic field (Section 10). Finally, this paper marks the public release of the NDSPMHD SPH/SPMHD code that can be used to test and verify all of the ideas and methods that have been discussed.

Acknowledgments

My knowledge and understanding of SPH would be nothing without the endless insight derived from my long-time mentor and colleague (and godfather-of-SPH), Joe Monaghan. In gathering my thoughts for this paper The author acknowledges particularly useful discussions with Klaus Dolag, Evghenii Gaburov, Volker Springel and Guillaume Laibe as well as many of the students at the ASTROSIM summer school – thanks in particular go to Michał Hanasz for the invitation to present this material and the excellent organisation of the school. Figures were produced using SPLASH [79], a freely available visualisation tool for SPH that is downloadable from http://users.monash.edu.au/∼dprice/splash/. Thanks to James Wetter for the excellent work on the new giza backend. Finally, thanks to the (anonymous) referees for helpful comments on the manuscript.

References

[1] T. Abel, rpSPH: a much improved smoothed particle hydrodynamics algorithm, March 2010, arxiv:1003.0937.
[2] O. Agertz, B. Moore, J. Stadel, D. Potter, F. Miniati, J. Read, L. Mayer, A. Gawryszczak, A. Kravtsov, Å. Nordlund, F. Pearce, V. Quilis, D. Rudd, V. Springel, J. Stone, E. Tasker, R. Teyssier, J. Wadsley, R. Walder, Fundamental differences between SPH and grid methods, MNRAS 380 (September) (2007) 963–978.
[3] P. Artymowicz, S.H. Lubow, Dynamics of binary-disk interaction. 1: resonances and disk gap sizes, ApJ 421 (February) (1994) 651–667.
[4] D.S. Balsara, Total variation diminishing scheme for adiabatic and isothermal magnetohydrodynamics, ApJS 116 (May) (1998) 133.
[5] W. Benz, Smoothed particle hydrodynamics – a review, in: J.R. Buchler (Ed.), The Numerical Modelling of Nonlinear Stellar Pulsations, Kluwer, 1990, pp. 269–288.
[6] S. Børve, M. Omang, J. Trulsen, Regularized smoothed particle hydrodynamics: a new approach to simulating magnetohydrodynamic shocks, ApJ 561 (November) (2001) 82–93.
[7] S. Børve, M. Omang, J. Trulsen, Two-dimensional MHD smoothed particle hydrodynamics stability analysis, ApJS 153 (August) (2004) 447–462.
[8] S. Børve, M. Omang, J. Trulsen, Multidimensional MHD shock tests of regularized smoothed particle hydrodynamics, ApJ 652 (December) (2006) 1306–1317.
[9] A. Brandenburg, Magnetic field evolution in simulations with Euler potentials, MNRAS 401 (January) (2010) 347–354.
[10] M. Brio, C.C. Wu, An upwind differencing scheme for the equations of ideal magnetohydrodynamics, J. Comput. Phys. 75 (April) (1988) 400–422.
[11] L. Brookshaw, A method of calculating radiative heat diffusion in particle simulations, Proc. Astron. Soc. Aust. 6 (1985) 207–210.
[12] F. Bürzle, P.C. Clark, F. Stasyszyn, T. Greif, K. Dolag, R.S. Klessen, P. Nielaba, Protostellar collapse and fragmentation using an MHD GADGET, August 2010, arxiv:1008.3790.
[13] S.E. Byleveld, H. Pongracic, The influence of magnetic fields on star formation, PASA 13 (January) (1996) 71–74.
[14] A.H. Cerqueira, E.M. de Gouveia Dal Pino, Three-dimensional magnetohydrodynamic simulations of radiatively cooling, pulsed jets, ApJ 560 (October) (2001) 779–791.
[15] S.-H. Cha, A.P. Whitworth, Implementations and tests of Godunov-type particle hydrodynamics, MNRAS 340 (March) (2003) 73–90.
[16] E. Chow, J.J. Monaghan, Ultrarelativistic SPH, J. Comput. Phys. 134 (1997) 296–305.
[17] P. Cleary, J. Ha, V. Alguine, T. Nguyen, Flow modelling in casting processes, Appl. Math. Model. 26 (2) (2002) 171–190.
[18] P.W. Cleary, J.J. Monaghan, Conduction modelling using smoothed particle hydrodynamics, J. Comput. Phys. 148 (January) (1999) 227–264.
[19] H.M.P. Couchman, Mesh-refined P3M – a fast adaptive N-body algorithm, ApJL 368 (February) (1991) L23–L26.
[20] L. Cullen, W. Dehnen, Inviscid smoothed particle hydrodynamics, MNRAS (July) (2010) 1126.
[21] S.J. Cummins, M. Rudman, An SPH projection method, J. Comput. Phys. 152 (July) (1999) 584–607.
[22] W. Dai, P.R. Woodward, Extension of the piecewise parabolic method to multidimensional ideal magnetohydrodynamics, J. Comput. Phys. 115 (December) (1994) 485–514.
[23] A. Dedner, F. Kemm, D. Kröner, C.-D. Munz, T. Schnitzer, M. Wesenberg, Hyperbolic divergence cleaning for the MHD equations, J. Comput. Phys. 175 (January) (2002) 645–673.
[24] W. Dehnen, Towards optimal softening in three-dimensional N-body codes – I: minimizing the force error, MNRAS 324 (June) (2001) 273–291.
[25] P.J. Dellar, A note on magnetic monopoles and the one-dimensional MHD Riemann problem, J. Comput. Phys. 172 (September) (2001) 392–398.
[26] G. Dilts, Moving-least-squares-particle hydrodynamics – I: consistency and stability, Int. J. Numer. Meth. Eng. 44 (8) (1999) 1115–1155.
[27] C.L. Dobbs, D.J. Price, Magnetic fields and the dynamics of spiral galaxies, MNRAS 383 (January) (2008) 497–512.
[28] K. Dolag, M. Bartelmann, H. Lesch, SPH simulations of magnetic fields in galaxy clusters, A&A 348 (August) (1999) 351–363.
[29] K. Dolag, F. Stasyszyn, An MHD GADGET for cosmological simulations, MNRAS 398 (October) (2009) 1678–1697.
[30] C. Eckart, Variation principles of hydrodynamics, Phys. Fluids 3 (May) (1960) 421–427.
[31] P. Español, M. Revenga, Smoothed dissipative particle dynamics, Phys. Rev. E 67 (2) (2003) 026705.
[32] C.R. Evans, J.F. Hawley, Simulation of magnetohydrodynamic flows – A constrained transport method, ApJ 332 (September) (1988) 659–677.
[33] G. Field, Magnetic helicity in astrophysics, in: R. Epstein, W. Feldman (Eds.), AIP Conf. Proc. 144: Magnetospheric Phenomena in Astrophysics, 1986, pp. 324–341.

[34] O. Flebbe, S. Muenzel, H. Herold, H. Riffert, H. Ruder, Smoothed particle hydrodynamics: physical viscosity and the simulation of accretion disks, ApJ 431 (August) (1994) 754–760.
[35] D.A. Fulk, D.W. Quinn, An analysis of 1-D smoothed particle hydrodynamics kernels, J. Comput. Phys. 126 (June) (1996) 165–180.
[36] T.A. Gardiner, J.M. Stone, An unsplit Godunov method for ideal MHD via constrained transport, J. Comput. Phys. 205 (May) (2005) 509–539.
[37] R.A. Gingold, J.J. Monaghan, Smoothed particle hydrodynamics – Theory and application to non-spherical stars, MNRAS 181 (November) (1977) 375–389.
[38] J. Gray, J.J. Monaghan, R.P. Swift, SPH elastic dynamics, Comput. Meth. Appl. Mech. Eng. 190 (October) (2001) 6641–6662.
[39] F.H. Harlow, J.E. Welch, Numerical calculation of time-dependent viscous incompressible flow of fluid with free surface, Phys. Fluids 8 (December) (1965) 2182–2189.
[40] L. Hernquist, N. Katz, TREESPH – A unification of SPH with the hierarchical tree method, ApJS 70 (June) (1989) 419–446.
[41] S. Heß, V. Springel, Particle hydrodynamics with tessellation techniques, MNRAS 406 (August) (2010) 2289–2311.
[42] R.W. Hockney, J.W. Eastwood, Computer Simulation Using Particles, McGraw-Hill, New York, 1981.
[43] J.G. Hosking, A.P. Whitworth, Modelling ambipolar diffusion with two-fluid smoothed particle hydrodynamics, MNRAS 347 (January) (2004) 994–1000.
[44] X.Y. Hu, N.A. Adams, A multi-phase SPH method for macroscopic and mesoscopic flows, J. Comput. Phys. 213 (April) (2006) 844–861.
[45] S. Inutsuka, Reformulation of smoothed particle hydrodynamics with Riemann solver, J. Comput. Phys. 179 (June) (2002) 238–267.
[46] P. Janhunen, A positive conservative method for magnetohydrodynamics based on HLL and Roe methods, J. Comput. Phys. 160 (May) (2000) 649–661.
[47] M. Jubelgas, V. Springel, K. Dolag, Thermal conduction in cosmological SPH simulations, MNRAS preprint, April 2004, 000.
[48] A.V. Kats, Variational principle and canonical variables in hydrodynamics with discontinuities, Physica D (May) (2001) 459–474.
[49] H. Kotarba, S.J. Karl, T. Naab, P.H. Johansson, K. Dolag, H. Lesch, F.A. Stasyszyn, Simulating magnetic fields in the antennae galaxies, ApJ 716 (June) (2010) 1438–1452.
[50] H. Kotarba, H. Lesch, K. Dolag, T. Naab, P.H. Johansson, F.A. Stasyszyn, Magnetic field structure due to the global velocity field in spiral galaxies, MNRAS 397 (August) (2009) 733–747.
[51] S. Kulasegaram, J. Bonet, R.W. Lewis, M. Profit, A variational formulation based contact algorithm for rigid boundaries in two-dimensional SPH applications, Comput. Mech. 33 (2004) 316–325.
[52] G. Lodato, D.J. Price, On the diffusive propagation of warps in thin accretion discs, MNRAS 405 (June) (2010) 1212–1226.
[53] L.B. Lucy, A numerical approach to the testing of the fission hypothesis, Astron. J. 82 (December) (1977) 1013–1024.
[54] J.L. Maron, G.G. Howes, Gradient particle magnetohydrodynamics: a lagrangian particle code for astrophysical magnetohydrodynamics, ApJ 595 (September) (2003) 564–572.
[55] S. Marri, S.D.M. White, Smoothed particle hydrodynamics for galaxy-formation simulations: improved treatments of multiphase gas, of star formation and of supernovae feedback, MNRAS 345 (October) (2003) 561–574.
[56] Z. Meglicki, Analysis and applications of smoothed particle magnetohydrodynamics. PhD thesis, Australian National University, 1995
[57] A. Mignone, P. Tzeferacos, A second-order unsplit Godunov scheme for cell-centered MHD: the CTU-GLM scheme, J. Comput. Phys. 229 (March) (2010) 2117–2138.
[58] J.J. Monaghan, Extrapolating B. Splines for Interpolation, J. Comput. Phys. 60 (September) (1985) 253.
[59] J.J. Monaghan, Smoothed particle hydrodynamics, Ann. Rev. Astron. Astrophys. 30 (1992) 543–574.
[60] J.J. Monaghan, SPH and Riemann solvers, J. Comput. Phys. 136 (September) (1997) 298–307.
[61] J.J. Monaghan, SPH without a tensile instability, J. Comput. Phys. 159 (April) (2000) 290–311.
[62] J.J. Monaghan, SPH compressible turbulence, MNRAS 335 (September) (2002) 843–852.
[63] J.J. Monaghan, Energy transfer in a particle α model, J. Turb. 5 (March) (2004) 12.
[64] J.J. Monaghan, Smoothed particle hydrodynamics, Rep. Prog. Phys. 68 (8) (2005) 1703–1759.
[65] J.J. Monaghan, H.E. Huppert, M.G. Worster, Solidification using smoothed particle hydrodynamics, J. Comput. Phys. 206 (July) (2005) 684–705.
[66] J.J. Monaghan, J.C. Lattanzio, A refined particle method for astrophysical problems, A&A 149 (August) (1985) 135–143.
[67] J.J. Monaghan, D.J. Price, Variational principles for relativistic smoothed particle hydrodynamics, MNRAS 328 (December) (2001) 381–392.
[68] J.P. Morris, A study of the stability properties of smooth particle hydrodynamics, PASA 13 (January) (1996) 97–102.
[69] J.P. Morris, Analysis of smoothed particle hydrodynamics with applications. PhD thesis, Monash University, Melbourne, Australia, 1996b
[70] J.P. Morris, J.J. Monaghan, A switch to reduce SPH viscosity, J. Comput. Phys. 136 (1997) 41–50.
[71] P.J. Morrison, Hamiltonian description of the ideal fluid, Rev. Mod. Phys. 70 (April) (1998) 467–521.
[72] J.R. Murray, SPH simulations of tidally unstable accretion discs in cataclysmic variables, MNRAS 279 (March) (1996) 402–414.
[73] R.P. Nelson, J.C.B. Papaloizou, Variable smoothing lengths and energy conservation in smoothed particle hydrodynamics, MNRAS 270 (September) (1994) 1.
[74] W.A. Newcomb, Lagrangian and Hamiltonian methods in magnetohydrodynamics, Nucl. Fusion Suppl. 2 (1962) 451–463.
[75] F.I. Pelupessy, W.E. Schaap, R. van de Weygaert, Density estimators in particle hydrodynamics. DTFE versus regular SPH, A&A 403 (May) (2003) 389–398.
[76] G.J. Phillips, J.J. Monaghan, A numerical method for three-dimensional simulations of collapsing, isothermal, magnetic gas clouds, MNRAS 216 (October) (1985) 883–895.
[77] K.G. Powell, P.L. Roe, T.J. Linde, T.I. Gombosi, D.L. de Zeeuw, A solution-adaptive upwind scheme for ideal magnetohydrodynamics, J. Comput. Phys. 154 (September) (1999) 284–309.
[78] D.J. Price, Magnetic fields in astrophysics. PhD thesis, University of Cambridge, Cambridge, UK, 2004, astro-ph/0507472.
[79] D.J. Price, SPLASH: an interactive visualisation tool for smoothed particle hydrodynamics simulations, Publ. Astron. Soc. Aust. 24 (October) (2007) 159–173.
[80] D.J. Price, Modelling discontinuities and Kelvin-Helmholtz instabilities in SPH, J. Comput. Phys. 227 (2008) 10040–10057.
[81] D.J. Price, Smoothed particle magnetohydrodynamics – IV. Using the vector potential, MNRAS 401 (November) (2010) 1475–1499.
[82] D.J. Price, M.R. Bate, The impact of magnetic fields on single and binary star formation, MNRAS 377 (May) (2007) 77–90.
[83] D.J. Price, M.R. Bate, The effect of magnetic fields on star cluster formation, MNRAS 385 (April) (2008) 1820–1834.
[84] D.J. Price, M.R. Bate, Inefficient star formation: the combined effects of magnetic fields and radiative feedback, MNRAS 398 (2009) 33–46.
[85] D.J. Price, C. Federrath, A comparison between grid and particle methods on the statistics of driven, supersonic, isothermal turbulence, MNRAS (Jun.e) (2010) 960.
[86] D.J. Price, J.J. Monaghan, Smoothed particle magnetohydrodynamics – I. Algorithm and tests in one dimension, MNRAS 348 (Febuary) (2004) 123–138.
[87] D.J. Price, J.J. Monaghan, Smoothed particle magnetohydrodynamics – II. Variational principles and variable smoothing-length terms, MNRAS 348 (Febuary) (2004) 139–152.
[88] D.J. Price, J.J. Monaghan, Smoothed particle magnetohydrodynamics – III. Multidimensional tests and the $\nabla \cdot B = 0$ constraint, MNRAS 364 (December) (2005) 384–406.
[89] D.J. Price, J.J. Monaghan, An energy-conserving formalism for adaptive gravitational force softening in smoothed particle hydrodynamics and N-body codes, MNRAS 374 (Febuary) (2007) 1347–1358.
[90] D.J. Price, S. Rosswog, Producing ultrastrong magnetic fields in neutron star mergers, Science 312 (May) (2006) 719–722.
[91] B.W. Ritchie, P.A. Thomas, Multiphase smoothed-particle hydrodynamics, MNRAS 323 (May) (2001) 743–756.
[92] S. Rosswog, Astrophysical smooth particle hydrodynamics, New Astron. Rev. 53 (April) (2009) 78–104.
[93] S. Rosswog, D. Price, MAGMA: a three-dimensional, Lagrangian magnetohydrodynamics code for merger applications, MNRAS 379 (August) (2007) 915–931.

[94] D. Ryu, T.W. Jones, Numerical magetohydrodynamics in astrophysics: algorithm and tests for one-dimensional flow, ApJ 442 (March) (1995) 228–258.
[95] R. Salmon, Hamiltonian fluid mechanics, Ann. Rev. Fluid Mech. 20 (1988) 225–256.
[96] I.J. Schoenberg, Contributions to the problem of approximation of equidistant data by analytic functions. A: on the problem of smoothing or graduation – a 1st class of analytic approximation formulae, Q. Appl. Math. 4 (1946) 45.
[97] M. Serrano, P. Español, I. Zúñiga, Voronoi fluid particle model for euler equations, J. Stat. Phys. 121 (October) (2005) 133–147.
[98] V. Springel, The cosmological simulation code GADGET-2, MNRAS 364 (December) (2005) 1105–1134.
[99] V. Springel, L. Hernquist, Cosmological smoothed particle hydrodynamics simulations: the entropy equation, MNRAS 333 (July) (2002) 649–664.
[100] D.P. Stern, The motion of magnetic field lines, Space Sci. Rev. 6 (1966) 147.
[101] D.P. Stern, Euler potentials, Am. J. Phys. 38 (April) (1970) 494–501.
[102] D.P. Stern, Representation of magnetic fields in space, Rev. Geophys. Space Phys. 14 (May) (1976) 199–214.
[103] J.M. Stone, J.F. Hawley, C.R. Evans, M.L. Norman, A test suite for magnetohydrodynamical simulations, ApJ 388 (April) (1992) 415–437.
[104] G. Tóth, The $\nabla \cdot B = 0$ constraint in shock-capturing magnetohydrodynamics codes, J. Comput. Phys. 161 (July) (2000) 605–652.
[105] G. Tóth, Conservative and orthogonal discretization for the Lorentz force, J. Comput. Phys. 182 (October) (2002) 346–354.
[106] P.A. Thomas, H.M.P. Couchman, Simulating the formation of a cluster of galaxies, MNRAS 257 (July) (1992) 11–31.
[107] J.W. Wadsley, G. Veeravalli, H.M.P. Couchman, On the treatment of entropy mixing in numerical cosmology, MNRAS 387 (June) (2008) 427–438.
[108] S.J. Watkins, A.S. Bhattal, N. Francis, J.A. Turner, A.P. Whitworth, A new prescription for viscosity in smoothed particle hydrodynamics, A&AS 119 (October) (1996) 177–187.
[109] S.C. Whitehouse, M.R. Bate, Smoothed particle hydrodynamics with radiative transfer in the flux-limited diffusion approximation, MNRAS 353 (October) (2004) 1078–1094.
[110] S.C. Whitehouse, M.R. Bate, J.J. Monaghan, A faster algorithm for smoothed particle hydrodynamics with radiative transfer in the flux-limited diffusion approximation, MNRAS 364 (December) (2005) 1367–1377.