Carnegie Mellon University

MELLON COLLEGE OF SCIENCE

THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF Doctor of Philosophy

TITLE Nonlocal Aggregation Dynamics with Near-Neighbor Weighting, and the Associated Geometry of Measures

PRESENTED BY Jeff N. Eisenbeis

ACCEPTED BY THE DEPARTMENT OF Mathematical Sciences

Robert Pego
Dejan Slepcev — MAJOR PROFESSOR — January 8, 2018 — DATE

Thomas Bohman — DEPARTMENT HEAD — January 8, 2018 — DATE

APPROVED BY THE COLLEGE COUNCIL

Rebecca W. Doerge — DEAN — January 8, 2018 — DATE

Nonlocal aggregation dynamics with near-neighbor weighting, and the associated geometry of measures

Jeff N. Eisenbeis

Dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

at

Carnegie Mellon University, Department of Mathematical Sciences
Pittsburgh, Pennsylvania

Advised by Professors Robert Pego and Dejan Slepčev

Dedicated to my beloved wife and daughter, forever

ACKNOWLEDGMENTS

Thank you to my advisors Dejan and Bob, who are supermen. Thank you for being who you are, and doing what you do for the world (and for me).

Thank you to my committee members Jack Schaeffer, Jian-Guo Liu, and Adrian Tudorascu who go back years in their helping of me.

Thank you to the Department of Math Sciences, including its first-class above-and-beyond staff like Stella Andreoletti and Nancy Watson.

Thank you to the department, my advisors, and the National Science Foundation for all financial support.

Thank you to all my friends and peers over the years here.

Thank you to my family over the years, *not* here. I tried to get out of here, really.

Contents

CHAPTER 1

Introduction

In this thesis we introduce a new weighted-averaging variant of the familiar "nonlocal biological aggregation equation" in Euclidean space, with weights dependent on the nearness of neighbors, which are added for more realism and flexible modeling. We discover how the gradient flow structure of the original equation is realized again via the introduction of a new metric tensor, one that penalizes movement in crowded configurations (nonlocally). We interpret this metric tensor and its global metric, examine the formal differential geometry structure, understand its boundedness when infinite spreading can occur, and finally establish the topology for a version of the metric defined in a bounded set. Numerical simulations follow to illustrate the behavior of the aggregation dynamics and the metric's geodesics.

1.1. Collective dynamics

We begin with the general concept of collective behavior dynamics. *Collective dynamics*, as used in the scientific literature, always seems to refer to groupings of individuals of some kind who follow rules on an individual basis, usually identical, that lead to overall group behavior *without* centralized or external governance. An immediate tangible example might be the clustering of fish into schools, whose schooling evolutions of shape and density and size demonstrate remarkable response to their environments, seemingly intelligently as a collective, despite that no central intelligence seems to be present and that no communication between fish beyond near neighbors seems to occur.

The emergence of such group behavior, group patterns, group disposition, and group anything from a limited set of specific individual rules is often called *emergent behavior*. Emergent behavior is a fascinating mechanism which may explain many large-scale phenomena in the natural world, with deep philosophical meaning and much to study.

This is a broad general principle and a broad topic of discourse, no doubt, within all the sciences including math. Keeping the theme in mind, let us narrow it to the kinds of dynamic models under much study in the math literature in recent years, in particular limited to *spatial* models tracking the positions or distribution of individuals/mass of a group under simple interaction rules that encourage their grouping.

For example, let us consider models which might be applicable to track the "point" locations of schooling fish, or flocking birds, or herding penguins, or clustering bacteria, or unusual particles

in physics, or traveling formations of search-and-rescue robots. The first note about these models is that they are aspirational; as far as this author can tell, limited success has been met so far in applying the mathematical models discussed in this chapter to *actual* application-benefitting data-fitting scientific usage. That is not to say they are not useful; indeed, many interesting *phenomena* are observed which align with and possibly explain the striking emergent phenomena observed within experimental scientific disciplines.

This raises the second note about these models. A common theme surrounding them is *simplicity*. As far as this author can tell, simplistic models—which is a nice way of saying overly simplistic, from a practical point of view—is a hallmark of applied mathematical modeling for two legitimate reasons. First is the issue of mathematical tractability. This is clearly a benefit for useful analysis to proceed. However, it does not quite justify the large focus and effort given to models which are without a doubt very, very simplistic, to the point that nobody would believe the application can be governed by such a thing. The second more subtle reason is about simplicity of explanation: if a phenomenon observed in the world is also observed in a very simple model, a model which perhaps captures the important characteristics of the system, this can be very appealing as an explanation of the world phenomenon, despite that the world is of course more complicated. This is some kind of scientific principle of simplest explanation, perhaps simply what is known as Ockham's razor.

We now take a look at the distinguishing characteristics of spatial tracking models studied in recent decades, as mentioned above, which set the backdrop for the new model of this thesis.

1.2. The variety of spatial collective dynamics models

Again, we are interested in collective dynamics models that track the positions or distribution of individuals/mass of a group. The physical space is presumed to be Euclidean space $\mathbb{R}^d$, with practical applicability for $d = 2$ or 3.

Looking through the literature, we notice the following distinguishing characteristics between models, which may help to classify them and the resulting investigations:

• particles vs. density or measure distribution

ODE models are used to track individuals as particles, whereas PDE or nonlocal PDE (integro-differential) are used to track the distribution of group members from a higher-level mean-field perspective. Examples of the former are found in [**78, 33, 32**] and the latter in [**37, 12, 36**].

• position matching ("aggregation") vs. velocity matching ("flocking") vs. combination of the two

Position matching represents the attempt of individuals to be near others (but not too near, sometimes), without taking into account where those neighbors are headed; whereas velocity matching represents an attempt to move under (any) formation with others, without regard to separation.

When both are employed by a model, usually zones of activation are used, with attractive position matching used in an outer distance zone, and velocity matching used within an intermediate zone, and repulsive position matching within a nearest zone. Examples of position matching are found in [**55, 43**], velocity matching in [**78, 33**], and their combination in [**67, 32**].

• *when position matching is included*: attractive vs. repulsive vs. combination of the two

Position matching may in general involve only attraction, or only repulsion, or both. For example, attractive-only and repulsive-only phenomena are considered in [**11**].

• *when repulsion is included, in the case of density or measure distribution*: nonlocal repulsion (integro-spatial) vs. local repulsion (differential operator)

For PDEs, repulsion may be obtained by either a local PDE operator or by nonlocal integration over nearby mass. For example, the former in [**73**] and the latter in [**61**].

• radially symmetric interaction vs. asymmetric

Interactions may be taken symmetrically around the individual, which is the most common case considered in models, or asymmetrically, as in [**53, 60**].

• homogeneous individuals vs. heterogeneous

Individuals may interact with all other individuals identically, which is the most common case considered in models, or different "species" may be present.

• equal averaging of pairwise interactions vs. weighted averaging

Usually, interactions are defined first pairwise, and then the whole of pairwise interactions are resolved through some averaging. The most common case considered is equal averaging of these interactions. Alternatively, weights may be calculated, as in [**63, 64**].

• First-order time derivative (non-inertial) vs. second-order time derivative (inertial)

Interactions may directly impact an individual's velocity, as in first-order models, or may impact the individual's acceleration, as in second-order. Examples of the former are found in [**78, 81**] and the latter in [**23**].

• velocity preference ("self-propelled") vs. without

A preferred velocity may be built into the model, to represent typical flock motion, which is most common in second-order models such as in [**30**].

• other modeling ingredients: diffusion, external potential, density constraints, speed constraints, spatial boundary, ...

Other modeling ingredients are sometimes combined with the above, such as diffusion in [**76**], external potential in [**6**], limits on density "packing", limits on attainable speed, and spatial boundaries as in [**82**].

• spatial dimension d

Specific dimensions are sometimes studied for phenomena particular to them, such as 2D patterns and double milling in [**23, 30**].

1.3. Nonlocal aggregation

Let us now turn to the traditional nonlocal aggregation model, studied for example in [**55, 21, 12**] in its density/measure version. We call this model "pure" aggregation to distinguish from the new "weighted" aggregation introduced in this thesis.

1.3.1. Particle model.

Pure aggregation is well studied, with particle model

$$\dot{x_i} = -\frac{1}{N}\sum_{j\neq i} \nabla W\left(x_i - x_j\right) \tag{1.3.1}$$

which is an ODE in the configuration space $X = \mathbb{R}^{dN}$, for N particles positioned at locations $x_1, ... x_N$ in physical space $\mathbb{R}^d$ as a function of time.

In the classification scheme above of section 1.2, this model is particle, position matching, attractive or repulsive or combination of the two, radially symmetric or asymmetric, homogeneous in individuals, equal averaging, first order, without velocity preference, without other modeling ingredients, and in arbitrary dimension.

W is always taken in this thesis to be radially symmetric in $C^1\left(\mathbb{R}^d \setminus \{0\}\right)$, with possible singularity at the origin. In this thesis the radial profile of any radially symmetric function is notated with a raised circle, so that $W^\bullet$ is the radial profile of W, i.e. for all x

$$W\left(x\right) = W^\bullet(|x|)\,.$$

The potential function W describes the "pairwise interaction" between particles, the whole of which are resolved through equal averaging. As a "potential", it acts modulo addition by a constant, as seen in the model by the appearance of its derivative only. The choice to model interaction forces by the gradient of a potential function is motivated to allow energy considerations, so that W appears without derivative in an appropriate energy functional as seen in chapter 3. W is often chosen to be attractive at long range and repulsive at short range, i.e. $W^\bullet$ decreasing near the origin and

later increasing, modeling individuals' desires to aggregate in clusters while retaining some personal space. In other cases W may be purely attractive or repulsive, to model a collapsing aggregate or an expanding one.

This is a first-order model in time. That is, there is no acceleration or inertia involved. This modeling treats the interaction magnitude from ∇W directly as an implied velocity for the particle, which makes sense for applications in which individuals may rapidly accelerate to their typical velocities. In a slight abuse of terminology, the magnitude of ∇W experienced by a particle is still called "force".

The optional $\frac{1}{N}$ scaling prevents *additivity* of the forces. This is desirable if forces are presumed to be encoded in ∇W with reasonable magnitudes for the application, in which case they should not be multiplied by a large number in the presence of a large number of neighbors. In particular, this allows varying N, such as studying $N \to \infty$. The $\frac{1}{N}$ scaling may also be interpreted as assigning $\frac{1}{N}$ mass to each particle (in the sense of magnitude of influence), which corresponds to using probability measures in a measure-valued model, or equivalently as simply making the summation in the model an *average*.

This brings us to the density/measure model of pure aggregation.

1.3.2. Density and measure model.

The measure-valued counterpart of (1.3.1) may be obtained by representing the particles as Dirac measures, discarding particle labeling, which is given by the probability measure

$$\mu_t = \frac{1}{N} \sum_i \delta_{x_i(t)}.$$

Then the right-hand side of (1.3.1) may be written as a convolution with this measure,

$$-\left(\nabla W * \mu_t\right)\left(x_i\left(t\right)\right).$$

This presumes for now that W is radially symmetric with $\nabla W\left(0\right) = 0$. The resulting continuity equation

$$\partial_t \mu_t + \operatorname{div}\left(\mu_t v\right) = 0 \tag{1.3.2}$$
$$v = -\nabla W * \mu_t$$

then has the benefit of also allowing more general measures. In particular, it allows the modeling of a continuum *density* μ (with respect to Lebesgue measure), and indeed more generally any suitable finite measure μ in $\mathbb{R}^d$, with the dependent velocity field v advecting μ through time according to the spatial convolution.

The measure model has the advantage over the original ODE in representing individuals at the large scale, on average, in the way that PDEs and continuum models frequently are used (such as for fluid dynamics instead of tracking individual molecules). But both models have their utility.

Because the measure model's attractive and repulsive forces are both modeled nonlocally by integration, the label "nonlocal aggregation" is well motivated.

1.4. Gradient flow and optimal transport

Pure aggregation (1.3.2) is an example of a *gradient flow*. This means there exists a functional on the configuration space of measures which is not only a Lyapunov function but which also dissipates *maximally*, by the direction of evolution in the configuration space. That is, the evolution is a steepest descent, as identified in chapter 3. To define such maximal dissipation one also needs a metric on the configuration space, which here is the 2-Wasserstein metric, also known as the quadratic Monge-Kantorovich distance, elaborated below.

Note pure aggregation's ODE (1.3.1) also is a gradient flow, under an analogous energy and metric on its configuration space.

Many classical PDEs of probability densities are in fact gradient flows with respect to the 2-Wasserstein metric, such as the heat equation (diffusion) and the Fokker-Planck equation, investigated by Jordan, Kinderlehrer, Otto [**49**]. A quite general theory for gradient flows in metric spaces, and particularly in spaces of probability measures, has been elucidated by Ambrosio, Gigli, and Savaré [**3**].

The basis of the 2-Wasserstein metric lies within the other general theory to be mentioned here, that of *optimal transport*. The rudimentary basics of optimal transport are presumed background for this thesis, as can be found in the introductory chapter of a book like Villani's [**79**]. The 2-Wasserstein distance between two measures is none other than the square root of their optimal transport cost, using cost function equal to the square of Euclidean distance.

Optimal transport traces back to Monge's original formulation [**62**], which was picked up over a century later by Kantorovitch [**50, 51, 52**] at first unaware of the former. Kantorovitch's formulation relaxed Monge's into the modern one considered today, with Monge's problem a constrained version which is often more difficult. Two decades later Sudakov [**72**] advanced the theory with an almost-correct proof of existence of a minimizer to Monge's problem.

Optimal transport theory has undergone rapid advancement in recent decades, with noteworthy advancements: by Brenier [**16**] with his polar factorization of L^2 vector-valued maps to yield rearrangements to gradients of convex maps; by Caffarelli [**18**] on regularity conditions for these; by Gangbo and McCann [**44, 45**] for cost functions strictly convex or a strictly concave function of

Euclidean distance; by McCann [**59**] with his displacement interpolation and displacement convexity; by Evans and Gangbo [**42**] on formulating optimal transport through differential equations; by Cordero-Erausquin, McCann, and Schmuckenschläger [**31**] on characterization of the optimal transport maps and displacement interpolation on manifolds; and by Benamou and Brenier [**5**] for their connection of the 2-Wasserstein optimal transport to a "least action" Eulerian ("fluid mechanics") displacement formulation.

Summary works on optimal transport include that by Rachev and Rüschendorf [**66**]; those by Ambrosio and later with Gigli [**1, 2**]; those by Villani [**79, 80**]; and that by Santambrogio [**68**].

Another novel direction was taken by Otto [**65**], with the point of view that Eulerian displacement interpolations may be taken formally as curves in Riemannian manifolds. The formal tangent vector at a configuration in the space of measures is taken as an equivalence class of velocity fields: those that via the continuity equation yield the instantaneous measure evolution. Textbooks, such as [**38**], contain the differential geometry theory alluded to.

Lastly, as an example of the possible variety of gradient flows using atypical metrics, we mention Erbar's recent work [**40**] which introduces another nonlocal modification of the 2-Wasserstein metric, as this thesis does, to identify a gradient flow where a gradient flow was otherwise not obviously present. He considers a nonlocal continuity equation, in contrast to the nonlocal mobility considered here. The "jump" continuity equation therein indeed exemplifies the possible variety of gradient flows yet to be noticed.

1.5. The contributions and content of this thesis

The principal contributions of this thesis are regarded as follows.

• The case is made for weighted averaging as an important improvement to the traditional pure aggregation equation.

• It is discovered that such weighted aggregation is a gradient flow under a new metric tensor for measures.

• Such new metric tensor and its global metric are introduced, which penalize motion in "crowded" configurations as a generalization of the 2-Wasserstein metric.

• Mass spreading is studied in this metric space of measures, and it is found that conditions for finite-cost infinite spreading imply boundedness of the metric.

• Topological equivalence to weak (-*) convergence of measures is established for a version of the metric on a bounded set that is first defined for regular measures and then extended by completion.

The remainder of the thesis is organized as follows.

Chapter 2 concerns the motivational development of weighted aggregation from pure aggregation with its weaknesses. In chapter 3 it is discovered that weighted aggregation is a gradient flow with the introduction of a new metric tensor. Chapter 4 develops this new geometry of measures with its global metric and geodesics. In chapter 5 the possibility of finite-cost infinite spreading is noted, which leads to an examination of the conditions for such and the implications on boundedness of the metric. Chapter 6 introduces an alternative development of the desired metric using regular measures and extending by completion, with the development limited to measures on a given bounded set, and establishes topological and uniform equivalence to the 2-Wasserstein metric there. Chapter 7 briefly lists some open directions of research following the developments in this thesis. Finally, chapter 8 contains the visual results of numerical experimentation performed on the geodesics of the particle version of the metric as well as on the weighted aggregation ODE. Of particular note is an interesting behavior found experimentally we call "polar milling", which involves recirculation of individuals inside clusters during their migration to merge with other distant clusters.

CHAPTER 2

Development of the aggregation model

2.1. Motivation for a change

From a modeling point of view, the pure aggregation equation (1.3.1) suffers from a weakness similar to one already identified in the Cucker-Smale model for biological flocking [**63**]. Here however, at issue is position attraction-repulsion ("aggregation") instead of velocity matching ("flocking"). The weakness can be seen as follows.

Consider a rather typical W with a repulsive regime near the origin and an attractive regime on the rest of its domain, and suppose W flattens at large distance, meaning the attraction diminishes to near zero. This commonly models a diminishing benefit or interest in aggregating far away. The catch is that the nearly-zero forces from distant particles have a diluting effect on close-proximity forces where W is more active. (The model is averaging, not additive.) This unduly slows the motion of a particle aggregating with nearby neighbors, *relative* to the motion of other particles not much diluted.

For instance, given a local cluster not yet in equilibrium, joined by a more massive companion cluster far away, there exists sufficiently large separation such that the small cluster's dynamics will be dominated by itself, and thus it will aggregate, yet it will act in slow motion compared to the motion occurring in the larger cluster. Indeed, slower than it would act if the larger cluster were not present. This seems undesirable.

So, one may wonder: Why are we treating the near-zero forces at large distance, which model diminishing benefit, *the same as* the near-zero forces at near distance which model desirable spacing?

Both tip the average toward "do nothing", but one seems to be important to the agents to "do nothing", whereas the other seems to be about *unimportance*. Perhaps it should be paid less attention.

A modest adaptation to the model, then... Why not weighted averaging?

2.2. Adapting the model: weighting[1]

Let us replace the $\frac{1}{N}$ in equation (1.3.1) with a weight on each term in the summation, to form a more general convex combination.

The only information at our disposal is particle positions, so let us assign weight according to proximity: Let nonnegative $a^\bullet \in C([0,\infty))$ prescribe relative "attention" given at distance, to be used for the weights of neighbors. For example, attention function $a^\bullet(r) := \frac{1}{1+r}$ would represent a diminishing attention with distance, favoring near neighbors.

As hinted at by the notation, $a^\bullet$ is then revolved about the origin to produce radially symmetric nonnegative $a \in C\left(\mathbb{R}^d\right)$ with $a^\bullet$ as its radial profile.

The sought convex combination is then written

$$\dot{x_i} = \sum_{j \neq i} \bar{\theta}_{ij}(\vec{x}) \nabla W(x_j - x_i) \tag{2.2.1}$$

$$\bar{\theta}_{ij}(\vec{x}) := \frac{a(x_i - x_j)}{\sum_{k \neq i} a(x_i - x_k)}$$

with $\vec{x}$ denoting the vector of $x_1, \ldots x_N$.

From the model it is seen that attention values $a(\cdot)$ are *relative* (to each other), meaning the model is invariant to scaling of a, and a therefore may be normalized. For example, whenever a is bounded, we normalize to $\sup a\left(\mathbb{R}^d\right) = 1$. Note we do not write $\|a\|_\infty$ for reasons of generalization later when a may be discontinuous, and an almost-everywhere description of a is insufficient.

We keep W in its role prescribing *pairwise interaction*, which in particular is an agent's behavior when seeing only one other agent. The whole of pairwise interactions are resolved through weighted average.

The reader will note that weighted averaging is precisely the remedy applied by Motsch and Tadmor to the Cucker-Smale model of flocking under velocity matching [**63**], to improve its weaknesses similar to those of pure aggregation cited here. The same authors also recently applied this to a "consensus" model in opinion dynamics [**64**], which is more like the position aggregation of this thesis. The methods and objectives in those works were slightly different than here. Both rely on an attractive-only interaction potential W, with the consensus model, a special case of the model here, specifically focused on $W^\bullet(r) := r^2$ as the radial profile of W.

[1]Note the adapted model of this section becomes modified once further in the next section.

We must also note that the landmark flocking model of Vicsek et al [**78**] itself can be seen to have used weighting, albeit of a discrete kind: neighbors are included or excluded depending on proximity. This is none other than the attention function $a = \chi_{\mathrm{B}(0,R)}$, the characteristic function of a ball. No doubt many modern simulations also share this practicality. The model here can approximate this attention function by continuous functions a.

2.3. Self attention

The model (2.2.1) is fine for functions a that are strictly positive. However, we may also wish to consider nonnegative functions a allowed to take value zero, as motivated for example in the previous paragraph. This would allow in general the modeling of finite "attention horizons", meaning functions a with bounded support, which may be a reasonable desire in applications.

We can see that if a vanishes somewhere, then the right hand side of (2.2.1) is undefined in many configurations. Indeed, conceptually, we have not imagined what a particle should do when it can "see" no others, meaning when all neighbors are assigned attention zero for this particle. If we prescribe a rule for this situation, the only justifiable rule seems to be "do nothing". Any other rule would involve a direction chosen arbitrarily. Of course, the rule "do nothing" is also pure aggregation's description for an isolated particle.

How do we revise (2.2.1) for this?

If we define the form $0/0$ as zero, or equivalently if we provide cases on the right hand side of (2.2.1) to yield zero for this case, then the ODE becomes generally not well-posed. For example, in a 2-particle world initialized by separation of distance one, with attractive W and with $a^\bullet$ positive on $[0, 1)$ and zero at one, there is no uniqueness of solutions on future time. And of course the equation with exceptions becomes harder to study.

Another natural way may be to give some attention to self. Self interaction is then of course defined to be zero. One may call it the "laziness" tendency, a bit of zero inertia that must continually be overcome by neighbors to excite the individual. It can be chosen arbitrarily small.

Wishfully thinking, perhaps experimentalists may even find benefit from such a parameter which is *not* arbitrarily small.

Thus we add regularizing constant $a_0 \geq 0$ to the denominator of $\bar{\theta}_{ij}$ and remove the bar for notation:

$$\theta_{ij}\left(\vec{x}\right) := \frac{a\left(x_i - x_j\right)}{a_0 + \sum_{k \neq i} a\left(x_i - x_k\right)}.$$

The new ODE taken with the new θ_{ij},

$$\dot{x_i} = \sum_{j \neq i} \theta_{ij}\left(\vec{x}\right) \nabla W\left(x_j - x_i\right), \tag{2.3.1}$$

then remains a convex combination by thinking of the summation over $j \neq i$ as including an implicit "$j = i$" term of zero.

Hereafter we assume one of a, a_0 is strictly positive, for the denominator of $\theta_{ij}\left(\vec{x}\right)$. This expression arises often, so we give it notation

$$\alpha_i\left(\vec{x}\right) := a_0 + \sum_{k \neq i} a\left(x_i - x_k\right) > 0$$

"the total attention of (or mass seen by) particle i in configuration $\vec{x}$".

Note the case $a_0 = 0$ is allowed, so no restriction has been made by incorporating a_0 into the model when a is positive.

One way to interpret the term a_0 when positive is as the simplest way to interpolate between the new "do nothing" instruction and the expected behavior as neighbors increase. More complicated schemes may be desirable in another study, for example scaling a_0 with N.

Note a "$\frac{1}{N}$"-type term scaling each $\dot{x_i}$ is *not* necessary in this model and is absent, as had been seen in the pure aggregation model (1.3.1) for purposes of scaling N. Namely, as more particles are added to a system, and N therefore increases, the particles do *not* feel greater and greater forces as they would in the pure aggregation model without its $\frac{1}{N}$ term. Pure aggregation's $\frac{1}{N}$ makes its summation an average, which the new weighted aggregation model already has with its weighted sum.

Finally, in support of the new self-attention term a_0, let us observe that pure aggregation itself as written in equation (1.3.1) already essentially contains self attention. Indeed, the care that would be necessary to avoid it, namely by scaling the model with $\frac{1}{N-1}$ instead of with $\frac{1}{N}$, is never normally taken.

Attention function a will frequently be chosen to be radially decreasing, though not required at this point. Radially decreasing a represents the natural case of individual particles paying more attention to nearer neighbors, and is the basis for the idea of modifying the original aggregation equation to resolve its motivating weaknesses.

The ODE (2.3.1) may be given classical ODE existence and uniqueness through suitable conditions on a and W, such as locally Lipschitz a and ∇W.

2.4. Features of the new model

With modeling complete for the *particle* version of weighted aggregation, let us collect a few immediate observations in comparison with pure aggregation.

- Weighted aggregation generalizes pure aggregation: A constant function a in the model (2.3.1) yields the model (1.3.1) with possibly different scaling from N, which is asymptotically the same for large N. In particular, if $a_0 = a(0)$ then the $\frac{1}{N}$ is obtained, and $a_0 = 0$ yields scaling $\frac{1}{N-1}$. Constant a has the interpretation of equal attention to all other particles, i.e. pure aggregation.
- As with pure aggregation written with the the $\frac{1}{N}$ factor, weighted aggregation always produces a net force of reasonable magnitude for an individual, as encoded in the magnitudes of ∇W; in this case within the convex hull of the pairwise forces.
- Unlike in pure aggregation, N is no longer explicit in the formulation, aside from summation lengths.
- θ_{ij} is not generally symmetric: The weight of particle i influencing particle j is not the same as the weight of particle j influencing particle i.
- Unlike in pure aggregation, center of mass is not generally conserved.

Let us also elaborate the possible behavioral improvements brought on by weighted averaging. Revisiting the weakness described in section 2.1, it should be apparent that the issue has been addressed, if function a is shaped appropriately: namely, that a radially decays with sufficient rates over the length scales to attenuate unwanted distant forces while preserving wanted local ones. Besides this motivating reason, the weighting may introduce other modeling benefits:

- Paying more attention to nearby individuals may be a modeling advantage in a variety of circumstances. For example, the pairwise interaction force need not decay with distance to benefit from this.
- Paying more attention to nearby individuals may allow more spread-out equilibria, which may be desirable. Consider, for example, an attractive-repulsive interaction potential W containing a "sweet spot" distance of zero force that represents desirable aggregation spacing. If an attention function is chosen that attenuates beyond this distance, it may better allow such consistent lattice-like spacing in aggregate packs without as much incentive to pack tighter due to distant neighbors attracting each other. This is especially promising if the number of individuals is caused to increase.
- Attention shaping may be flexible for many purposes, not only those focusing larger attention on nearer neighbors.
- Allowing asymmetric a has the further potential to incorporate a "directed attention", such as animals' fields of vision or one-sided attention in traffic modeling.

Finally, a feature of significance from the point of view of analysis is that the new model remains a gradient flow, as is pure aggregation, under an appropriate new metric. This is described in the next chapter after first in this chapter attending to the measure-valued side of everything just discussed.

2.5. The model on measures

Let us catch up the modeling as it applies to a measure-valued version for weighted aggregation, which modifies that for pure aggregation in equation (1.3.2).

Again using the attention function a to prescribe relative "attention" given at each distance, we write the corresponding measure-valued model

$$\begin{aligned} &\partial_t \mu_t + \operatorname{div}\left(\mu_t v\right) = 0 \\ &v = -\frac{(a\nabla W) * \mu_t}{a * \mu_t} \end{aligned} \tag{2.5.1}$$

where convolutions are in space, the finite measures $(\mu_t)_t$ vary in time, and velocity field v varies in space and time.

If $(\mu_t)_t$ through all time is a Lebesgue density, meaning absolutely continuous with respect to Lebesgue measure, the continuity equation reads as a PDE with μ_t the density function. Existence may be taken in the sense of L^p solutions, as in [**12**].

Alternatively, if $(\mu_t)_t$ through time is a finite Borel measure in $\mathbb{R}^d$, and if a and W are smooth enough, such as a and ∇W locally Lipschitz, the continuity equation may be interpreted as a measure evolution in the Lagrangian sense by taking $\mu_t = \Phi(t,\cdot)_{\#}\mu_0$, the push-forward of the initial measure μ_0, where $\Phi(t,x) \in \mathbb{R}^d$ denotes the time t solution of the IVP

$$\dot{\xi}(s) = v(s, \xi(s)), \;\; \xi(0) = x \in \mathbb{R}^d.$$

Or, in more generality, equation (2.5.1) may be taken to have existence in the sense of distributions, as the basis in [**21, 3**].

In any case, v must satisfy the prescribed dependence on μ in the second equation of (2.5.1).

Some conditions are also needed on a and W to define the expression for v. Nonnegative radially symmetric $a \in C\left(\mathbb{R}^d\right)$ which arose in the particle model is here taken also to be bounded with $a(0) > 0$; the latter to ensure by continuity at the origin that $a * \mu_t$ is nonzero on the support of μ_t, i.e. where v must be defined; and the former to provide that the convolution is finite. W is also taken Lipschitz for the convolution in the numerator to be finite.

We must also comment about W at the origin. We presume $W \in C^1\left(\mathbb{R}^d \setminus \{0\}\right)$ as mentioned in the background describing the pure aggregation model. However, it should be noted that in both pure aggregation as well as the new weighted aggregation, singularities at the origin are not as easily feasible in the measure-valued model as in the particle model. Namely, in the event that $(\mu_t)_t$ begins with or develops a point mass. (This of course *does* happen when the model is used to represent particles.) For such measure flows it is convenient to have $\nabla W(0) = 0$. However, this may be relaxed, taking the gradient as the minimal element of the subdifferential to allow pointy W at the origin, as developed in [**21**]. This approach allows continuation of the solution after point masses develop.

Note we have little choice regarding so-called "self attention" in the measure-valued model. If a measure μ contains a point mass, necessarily it has a "built-in" self attention "$a_0 = a(0)$". This corresponds to each infinitesimal unit of that mass paying attention to the remainder of that mass according to function a taken at distance zero. On the other hand, if μ contains components that are absolutely continuous with respect to Lebesgue density, these have *no* self attention, we might say.

Similarly to what occurs with the pure aggregation models, the new weighted aggregation models for the particle case and the measure case match for point masses: If we assume $\nabla W(0) = 0$, and if μ_t is the Dirac sum $\mu_t = \frac{1}{N}\sum_i \delta_{x_i(t)}$, or indeed $\mu_t = \sum_i \delta_{x_i(t)}$ without the "$\frac{1}{N}$", then the measure model (2.5.1) reduces to the particle model (2.3.1) with $a_0 = a(0)$.

CHAPTER 3

Gradient flow structure

In this chapter we examine the gradient-flow structure of the weighted aggregation models developed in chapter 2. We first illustrate the intuition as applied to the particle model, and proceed to develop the same in the general measure-valued model.

3.1. Energy dissipation

The dissipative free energy (Lyapunov function) for pure aggregation equation (1.3.1) has the structure of an "interaction energy" [**79**], which generally takes the form

$$\mathcal{E}_W(\vec{x}) := \frac{1}{2}\sum_{ij} W(x_i - x_j)$$

whenever W represents a pairwise potential energy on particles. This is exactly the case for pure aggregation (1.3.1) with its potential W.

Does the weighted aggregation model (2.3.1) have a dissipative energy? We might expect any such to also be an interaction energy, based on the interactive nature of the equation, as opposed to an internal energy, external potential energy, or combination thereof [**79**]. The answer is affirmative, although a naive guess like

$$\mathcal{E}_{guess}(\vec{x}) := \sum_{ij} \theta_{ij}(\vec{x})\, W(x_i - x_j)$$

would not work. Recall θ_{ij} is not even symmetric.

An energy is quickly noticed by writing equation (2.3.1) as

$$\dot{x_i} = -\frac{1}{\alpha_i(\vec{x})}\sum_{j\neq i} a(x_i - x_j)\,\nabla W(x_i - x_j)$$

where due to the radial symmetry of a and W, one observes that

$$a\nabla W = \nabla\widetilde{W}$$

for some radially symmetric $\widetilde{W} \in C^1\left(\mathbb{R}^d \setminus \{0\}\right)$, namely with radial profile

$$\widetilde{W}^\bullet(r) = \int_0^r a^\bullet W^{\bullet\,\prime}$$

using normalization $\widetilde{W}(0) = 0$.

Thus

$$\dot{x_i} = -\frac{1}{\alpha_i(\vec{x})} \sum_{j \neq i} \nabla \widetilde{W}(x_i - x_j). \tag{3.1.1}$$

Apparently $\widetilde{W}$ then is the potential for an interaction energy

$$\mathcal{E}(\vec{x}) := \frac{1}{2} \sum_{ij} \widetilde{W}(x_i - x_j) \tag{3.1.2}$$

which we briefly verify.

Let $\vec{x}, \vec{u} \in \mathbb{R}^{dN}$ be an arbitrary configuration and tangent vector to the configuration, respectively. The differential of $\mathcal{E}$ in direction $\vec{u}$ is

$$d\mathcal{E}(\vec{x})(\vec{u}) = \sum_i \nabla_{x_i} \mathcal{E}(\vec{x}) \cdot u_i = \sum_i \left[\sum_{j \neq i} \nabla \widetilde{W}(x_i - x_j) \right] \cdot u_i$$

by symmetry, and

$$= -\sum_i \alpha_i(\vec{x}) v_i(\vec{x}) \cdot u_i$$

written in terms of the evolution vector field

$$v_i(\vec{x}) = -\frac{1}{\alpha_i(\vec{x})} \sum_{j \neq i} \nabla \widetilde{W}(x_i - x_j),$$

the right-hand-side of (3.1.1).

In the direction of evolution $\vec{u} = \vec{v}(\vec{x})$, the differential

$$d\mathcal{E}(\vec{x})(\vec{v}(\vec{x})) = -\sum_i \alpha_i(\vec{x}) |v_i(\vec{x})|^2 < 0$$

at non-equilibria $\vec{x}$. Thus $\mathcal{E}$ is dissipative, i.e. a Lyapunov function.

3.2. Gradient energy dissipation

However, there is further structure to the differential. Consider momentarily the differential for pure aggregation:

$$d\mathcal{E}_W(\vec{x})(\vec{u}) = -\vec{v}(\vec{x}) \cdot \vec{u}$$

in which case

$$\vec{v}(\vec{x}) = -\nabla_{\vec{x}} \mathcal{E}_W(\vec{x}).$$

That is, the evolution is a *gradient flow* of the functional $\mathcal{E}_W$.

More generally, if the differential can be written as *any* inner product with $\vec{v}(\vec{x})$, possibly configuration-dependent and varying continuously, which is to say a Riemannian metric, then a gradient-flow structure remains. Such inner product more precisely is the *tensor* of the metric, which leads the configuration space $\mathbb{R}^{dN}$ to be taken as a Riemannian manifold.

Indeed, our differential from the previous section is

$$d\mathcal{E}(\vec{x})(\vec{u}) = -g_{\vec{x}}(\vec{v}(\vec{x}), \vec{u})$$

where

$$g_{\vec{x}}(\vec{v}, \vec{u}) := \sum_i \alpha_i(\vec{x}) v_i \cdot u_i$$

which is a Riemannian metric tensor, recalling that $\alpha_i(\vec{x}) > 0$.

Thus our evolution vector field

$$\vec{v}(\vec{x}) = -\mathrm{grad}_g \mathcal{E}(\vec{x})$$

is a gradient flow.

3.3. The model on measures

The dissipative energy for the measure-valued model is

$$\mathcal{E}(\mu) := \frac{1}{2}\int \left(\widetilde{W} * \mu\right) d\mu$$

at measure (configuration) μ.

For this model, the "tangent vectors" of an evolution are taken formally as the velocity field $v : \mathbb{R}^d \to \mathbb{R}^d$ in the continuity equation,

$$\partial_t \mu_t + \operatorname{div}(\mu_t v) = 0,$$

following the submersion formalism introduced by Otto [**65**]. However, it is clear that many distinct velocity fields in general may satisfy this equation for a given measure μ, meaning they contribute the same instantaneous change to the measure. This necessitates an equivalence relation between vector fields which yield the same divergence. That is, at a configuration μ, vector fields $v, \hat{v} : \mathbb{R}^d \to \mathbb{R}^d$ are considered equivalent if

$$\operatorname{div}(\mu v) = \operatorname{div}(\mu \hat{v}).$$

We formally show the energy $\mathcal{E}$ provides a gradient flow as follows.

Given arbitrary configuration μ and tangent vector $u : \mathbb{R}^d \to \mathbb{R}^d$, we wish to calculate that

$$d\mathcal{E}(\mu)(u) = -g_\mu(v(\mu), u)$$

for some "Riemannian" metric tensor g, where

$$v(\mu) := -\frac{(a\nabla W) * \mu}{a * \mu}.$$

Let $\mu_t := \Phi(t, \cdot)_{\#}\mu$ where $\Phi(t, x) \in \mathbb{R}^d$ denotes the time t solution of the IVP

$$\dot{\xi}(s) = u(\xi(s)), \xi(0) = x \in \mathbb{R}^d.$$

Note at time zero this evolution has configuration μ with tangent vector u, though it is simpler than the aggregation evolution.

Its energy is

$$\begin{aligned}\mathcal{E}(\mu_t) &= \frac{1}{2}\iint \widetilde{W}(x-y)\,d\mu_t(x)\,d\mu_t(y)\\ &= \frac{1}{2}\iint \widetilde{W}(\Phi(t,x)-\Phi(t,y))\,d\mu(x)\,d\mu(y)\end{aligned}$$

as the integral of a push-forward measure, or two nested, and

$$\begin{aligned}d\mathcal{E}(\mu)(u) &= \delta_t\mathcal{E}(\mu_t)\,{}_{|t=0}\\ &= \frac{1}{2}\iint \nabla\widetilde{W}(\Phi(t,x)-\Phi(t,y))\cdot[\delta_t\Phi(t,x)-\delta_t\Phi(t,y)]\,d\mu(x)\,d\mu(y)\,{}_{|t=0}\\ &= \frac{1}{2}\iint \nabla\widetilde{W}(\Phi(t,x)-\Phi(t,y))\cdot[u(\Phi(t,x))-u(\Phi(t,y))]\,d\mu(x)\,d\mu(y)\,{}_{|t=0}\\ &= \frac{1}{2}\iint \nabla\widetilde{W}(x-y)\cdot[u(x)-u(y)]\,d\mu(x)\,d\mu(y)\\ &= \iint \nabla\widetilde{W}(x-y)\cdot[u(x)]\,d\mu(x)\,d\mu(y)\\ &= \int\left(\nabla\widetilde{W}*\mu\right)\cdot u\,d\mu\\ &= -\int(a*\mu)\,v(\mu)\cdot u\,d\mu.\end{aligned}$$

This motivates a candidate metric tensor of the form

$$g_\mu(u,v) := \int (a*\mu)\,u\cdot v\,d\mu$$

for the Lagrangian description of the dynamics, to provide the anticipated gradient flow. For an Eulerian formulation, which is needed to develop the PDE theory, we need to introduce into this metric tensor form the equivalence relation mentioned above, which we pick up in the next chapter.

CHAPTER 4

The Riemannian geometry

In this chapter we consider the metric tensor that led to gradient flow in the previous chapter as well as the resulting global metric from the tensor. We also attempt to interpret these and their properties. In doing so we again proceed first with the particle version, i.e. the "particle metric", and subsequently the "measure metric" suitable for densities and more general measures. The measure metric is seen to generalize the well-known 2-Wasserstein metric on $\mathbb{R}^d$, with specialization to it when a is constant. This, again, is also the condition for weighted aggregation to specialize to pure aggregation.

4.1. Measure metric over equivalence classes

First we should be more precise about the measure metric. Above it is noted that at a "configuration" measure μ, vector fields $v, \hat{v} : \mathbb{R}^d \to \mathbb{R}^d$ taken as formal tangent vectors must be considered equivalent if

$$\text{div } (\mu v) = \text{div } (\mu \hat{v}) \,,$$

meaning they instantaneously evolve μ in the same way. In light of this we must define metric tensor g on equivalence classes of vector fields, not on the vector fields. So, which representative should be used to define g in the formula for g of the previous chapter? The answer is the minimal one, in the sense of yielding the minimum norm on the equivalence class. The interpretation is that no fruitless advecting should be regarded. Thus g should be defined

$$g_\mu (v, v) := \inf \left\{ \int (a * \mu) \, |\hat{v}|^2 \, d\mu \;\middle|\; \hat{v} : \mathbb{R}^d \to \mathbb{R}^d \text{ with div } (\mu \hat{v}) = \text{div } (\mu v) \right\}.$$

This suffices to characterize g fully, since in general an inner product is characterized fully by its norm via the polarization identity.

4.2. Riemannian manifold structure

4.2.1. Basic theory.

Let us call X the configuration space, whether we are working with the particle case or the measure-valued case: either $\mathbb{R}^{dN}$ in the particle case, or some space of Borel probability measures in $\mathbb{R}^d$ in the measure-valued case. In the measure-valued case we suppose only to have a formal metric tensor, leading to a formal Riemannian manifold.

In the particle case, note X is a differentiable manifold with single chart: identity.

In either case, a metric tensor g was identified in the previous chapter, making X a Riemannian manifold, as follows.

Following the classical theory of differential geometry, a smooth curve $\gamma : [0,1] \to X$ has length defined

$$L(\gamma) := \int_0^1 \|\dot{\gamma}(t)\|_{\gamma(t)} \, dt$$

where $\dot{\gamma}(t)$ is the tangent vector of $\gamma(t)$ and

$$\|\cdot\|_{\gamma(t)} := \sqrt{g_{\gamma(t)}(\cdot,\cdot)}.$$

X is then made a metric space by defining distance

$$d(\vec{y},\vec{z}) := \inf\{L(\gamma) \mid \text{smooth curve } \gamma \text{ is such that } \gamma(0) = \vec{y}, \gamma(1) = \vec{z}\}$$

where $\vec{y},\vec{z}$ are replaced by μ,ν in the measure-valued case. We often speak of a *source configuration* $\vec{y}$ or μ, and a *target configuration* $\vec{z}$ or ν. This implies that the curves connecting them are under study.

Geodesics, then, are defined as the curves that locally connect their points with minimum length. At any point on a geodesic, there is a ball within which all other points on the geodesic lie at distance equal to the length of the geodesic's segment connecting them. That is, the geodesic is "global" within the ball.

4.2.2. Global metric.

From this abstract theory we now state the resulting global particle metric and global measure metric that arise from the metric tensors identified in chapter 3. The latter metric is first defined on a space of measures too large for it generally to be finite, and then restricted as a way of characterizing the appropriate space. This in fact allows dependence on function a for the metric's natural space, which is not as straightforward as $\mathrm{P}_2\left(\mathbb{R}^d\right)$ that suits the 2-Wasserstein metric.

For $a_0 \geq 0$ and nonnegative radially symmetric $a \in C\left(\mathbb{R}^d\right)$ with one of a, a_0 strictly positive, the metric given to $\mathbb{R}^{dN}$ is

$$d_e\left(\vec{y},\vec{z}\right) := \inf\left\{ \int_0^1 \sqrt{\sum_i \alpha_i\left(\vec{x}\left(t\right)\right)\left|\dot{x}_i\left(t\right)\right|^2}dt \quad \middle| \quad \vec{x} \in C^1\left(\left[0,1\right];\mathbb{R}^{dN}\right) \text{ with } \vec{x}\left(0\right)=\vec{y}, \vec{x}\left(1\right)=\vec{z} \right\}$$

where $\alpha_i\left(\vec{x}\right) := a_0 + \sum_{k\neq i} a\left(x_i - x_k\right) > 0$.

This metric may be read intuitively as "the infimum time-integral of the attention-weighted-ℓ^2 norm on particle speeds."

Let $\mathcal{A} = \left\{ a : \mathbb{R}^d \to [0,\infty) \;\middle|\; a \text{ is Borel measurable, radially symmetric, bounded, continuous at the origin with } a\left(0\right) > 0, \text{ and normalized to } \sup a\left(\mathbb{R}^d\right) = 1 \right\}$.

For $a \in \mathcal{A}$ the metric on measures is defined by first establishing it as a function on $\mathrm{P}\left(\mathbb{R}^d\right) \times \mathrm{P}\left(\mathbb{R}^d\right)$, possibly taking value ∞ when the infimum is empty,

$$\hat{d}_{\mathrm{E}}\left(\mu,\nu\right) := \inf\left\{ \int_0^1 \sqrt{\int_{\mathbb{R}^d} \left(a * \mu_t\right)\left(x\right)\left|v\left(t,x\right)\right|^2 \; d\mu_t\left(x\right)}dt \quad \middle| \quad \left(\left(\mu_t\right)_{t\in[0,1]}, v\right) \in V_{\mu,\nu} \right\}$$

where $V_{\mu,\nu}$ is defined as the set of $\left(\left(\mu_t\right)_{t\in[0,1]}, v\right)$ such that the curve $\left(\mu_t\right)_{t\in[0,1]}$ in $\mathrm{P}\left(\mathbb{R}^d\right)$ satisfies $\mu_0 = \mu$, $\mu_1 = \nu$, and such that $v \in \mathrm{L}^1\left(\left[0,1\right];\mathrm{L}^2\left(\left(a*\mu_t\right)d\mu_t\right)\right)$ gives $\partial_t \mu_t + \operatorname{div}\left(\mu_t\, v\right) = 0$ in the sense of distributions.

Let

$$\mathrm{P}_{\mathrm{E}}\left(\mathbb{R}^d\right) = \left\{ \mu \in \mathrm{P}\left(\mathbb{R}^d\right) \;\middle|\; \hat{d}_{\mathrm{E}}\left(\mu,\delta_0\right) < \infty \right\}$$

where δ_0 is the Dirac measure at the origin, and on $\mathrm{P}_{\mathrm{E}}\left(\mathbb{R}^d\right)$ define $d_{\mathrm{E}} = \hat{d}_{\mathrm{E}}$.

The metric may be read intuitively as "the infimum time-integral of the attention-weighted-$\mathrm{L}^2\left(d\mu_t\right)$ norm on the speed functional."

Immediately we must observe the Eulerian versus Lagrangian viewpoint here. The definition as written is Eulerian, in that a velocity field describes the measure evolution locally via the continuity equation. Or one might say that the point paths in $\mathbb{R}^d$ that are transporting the measure are tracked according to their velocity distribution over $\mathbb{R}^d$. The alternative Lagrangian viewpoint is to track these point paths according to *their* dynamic positions in $\mathbb{R}^d$ (and therefore also their velocities) in the same way that characteristics are used in classical PDE theory. Either viewpoint leads essentially to the same objects, at least for the Wasserstein metrics as shown in [**3**], which

develops more technically than we get into here. Ideally a comparable theory might be developed for the more complicated metric here, which we do not attempt.

When $V_{\mu,\nu}$ is defined in chapter 6 for a class of smooth measures, the Eulerian velocity field is tied directly to the Lagrangian paths by the obvious ODE.

Similarly, when $\hat{d}_{\mathrm{E}}$ above is compared to the 2-Wasserstein metric's displacement interpolation form with its corresponding Lagrangian interpretation as described in [**79**] and [**3**], it is clear $\hat{d}_{\mathrm{E}}$ yields the 2-Wasserstein metric when $a \equiv 1$. Thus because $a \leq 1$ we have $\hat{d}_{\mathrm{E}} \leq d_{\mathrm{W}_2} < \infty$ on $\mathrm{P}_2\left(\mathbb{R}^d\right)$, so $\mathrm{P}_2\left(\mathbb{R}^d\right) \subset \mathrm{P}_{\mathrm{E}}\left(\mathbb{R}^d\right)$ and $d_{\mathrm{E}} \leq d_{\mathrm{W}_2}$ on $\mathrm{P}_2\left(\mathbb{R}^d\right)$.

Incidentally, absolute continuity of curves in $\left(\mathrm{P}_2\left(\mathbb{R}^d\right), d_{\mathrm{W}_2}\right)$ is inherited by d_{E}, following the definition in [**3**] since $d_{\mathrm{E}} \leq d_{\mathrm{W}_2}$.

On another note, observe the role continuity of a has played up to now. In the aggregation modeling, for convenience a was assumed continuous, with the effect that the modified interaction potential $\widetilde{W}$ then is continuously differentiable as we expected of an interaction potential. Subsequently, in the development of the gradient flow structure, continuous a turned out to be necessary in the particle case to make a continuously varying metric tensor, thereby resulting in a true Riemannian manifold. However, on the measure metric side, this seems not so necessary to lead to our formal Riemannian manifold, as is especially intuitive for curves in the measure space consisting entirely of Lebesgue densities. To examine this formal measure geometry in sufficient generality, continuity of a has no longer been assumed, with the exception of continuity at the origin. This allows such attention function favorites as $a = \chi_{\mathrm{B}(0,1)}$. Of course, smoother attention functions are expected for calculating Euler-Lagrange geodesics, below.

4.2.3. Cost and cost rate.

Finally, as a convenience of terminology in both the particle and measure cases, we use the term *cost* to refer to the time integral above for a specific curve through the configuration space, and a specific corresponding velocity field (in the case of d_{E}), not necessarily the infimum of such. That is, a particular curve $\vec{x} \in C^1\left([0,1]; \mathbb{R}^{dN}\right)$ has cost

$$\int_0^1 \sqrt{\sum_i \alpha_i\left(\vec{x}\left(t\right)\right) \left|\dot{x}_i\left(t\right)\right|^2} dt$$

and a particular curve $\left(\mu_t\right)_{t\in[0,1]}$ with a corresponding velocity field v has cost

$$\int_0^1 \sqrt{\int_{\mathbb{R}^d} \left(a * \mu_t\right)(x) \left|v\left(t,x\right)\right|^2 \, d\mu_t\left(x\right)} dt.$$

Similarly, but in an abuse of terminology, the square of the integrand in the time integral will be referred to as the *cost rate*. Thus the cost is actually the time integral of the square root of the cost rate.

The cost terminology allows convenient discussion of specific configuration curves, including the bounds they imply on the above global metrics, and is preferred in this thesis over the differential geometry terminology *length* or *arclength* of the curve, helping to distinguish from the terminology of length and distance in $\mathbb{R}^d$.

4.3. Interpretations

4.3.1. Comparison to 2-Wasserstein.

First let us compare the new metric to the 2-Wasserstein.

The measure metric generalizes the 2-Wasserstein metric on $\mathbb{R}^d$, as seen by the displacement interpolation formulation of the 2-Wasserstein established by Benamou and Brenier [**5**] which has metric tensor

$$\bar{g}_\mu(v,v) := \int |v|^2 \, d\mu.$$

The new metric tensor

$$g_\mu(v,v) := \int (a * \mu) |v|^2 \, d\mu$$

reduces to it (up to a factor) precisely when a is a constant function. This is consistent with the gradient flow story above, since in this case the modified potential $\widetilde{W}$ also reduces to W, reducing the energy $\mathcal{E}$ to the standard interaction energy, which is traditionally paired with the 2-Wasserstein to yield gradient flow of pure aggregation. This also matches intuition; constant attention function a corresponds to the ability of mass to "see" all other mass, at all distances, thereby eliminating the reward in the metric given to movement that is away from other mass. This also suggests a reverse way of thinking about the Wasserstein metric, that mass is penalized for moving within the same Euclidean space as all other mass—when really, the metric is intended to penalize movement for movement's sake.

Note the new metric has *only* a displacement interpolation formulation, unlike the 2-Wasserstein metric which also has an optimal transportation formulation. The simplicity of the 2-Wasserstein displacement interpolation formulation explains why it admits an optimal transportation formulation with the displacement interpolation removed, "statically". In the new metric the complexity of attention interaction along the curve of displacement generally precludes this.

4.3.2. Interpretation of the function a.

The effect of the new term in the integrand is to make the metric tensor "doubly" nonlocal, with an inner integration determining the penalty size at each location of the outer integral. The so-called attention function a has changed roles from "attention" dependent on distance, in aggregation, to *penalty* dependent on distance. Mass in motion is charged a cost rate for its speed squared times a penalty for proximity to nearby mass.

The metric thus has interesting interpretations, and potential use in other applications besides the classic aggregation equation. Recall a is frequently taken in this thesis as radially decreasing, meaning greater penalty for motion when near large mass. Perhaps in some applications it is dangerous for mass to move rapidly near other mass. Or in other applications, perhaps mass simply faces resistance moving near other mass, such as with animals moving in congested areas or crowded particles in physics. So, "crowd motion" seems to be a relevant concept for the metric.

Observe also the role the metric plays in weighted aggregation. One goal of the introduction of weighting was to alleviate pure aggregation's tendency of diluting: that distant large clusters can cause a slow-motion effect on a small cluster. However, the effect of the new tensor (with suitable radially decreasing a) is to *slow* the members of the large cluster. This is relative to the speeds found in the members of the small cluster, which gives the *effect* of speeding up the small cluster. So slowing dense crowds is again the purpose of the metric.

4.3.3. Relation between the particle metric and the measure metric.

Finally, let us examine again the relationship between the new measure metric and the new particle metric for some N, when $a_0 = a(0)$. The value $a(0)$ is effectively self-attention under the measure metric, under which the mass inside a point mass sees all the rest of the point mass. If a point mass happens to remain intact, under the measure metric, the cost of its travel is exactly the same as with the particle metric.

In some sense, the particle metric represents a kind of submanifold in the space of measures under the measure metric, a submanifold which is a *true* Riemannian manifold inside the larger formal Riemannian manifold that is the latter. The submanifold would be taken as all measures that are a $\frac{1}{N}$ normalization of an N-sum of Dirac measures, $\mu = \frac{1}{N}\sum_i \delta_{x_i}$. That is, the point masses share the same fraction of mass. The particle metric's cost infimum is a *constrained* version of the measure metric's: interpolating curves must maintain the same form, i.e. the point masses cannot split. In this viewpoint the submanifold is "curved" within the larger space: to connect a source measure and a target measure in the submanifold, the measure metric may in general find shorter curves within the larger space by leaving the submanifold; that is, by spreading (smearing) point masses into densities during the transit that the point masses otherwise must traverse. Indeed, with appropriate attention function a, *always* such spreading will occur.

But the wrinkle is in particle labeling. Every particle configuration under the particle metric corresponds to precisely one measure under the measure metric; but a given measure so described corresponds to *many* particle configurations, due to particle ordering. Indeed, swapping the order of particles represents nontrivial distance under the particle metric, but zero distance to the measure metric. This is the limited sense in which the larger formal Riemannian manifold has this Riemannian manifold as a submanifold.

The particle metric, though simple, may be seen to have value in two aspects: Firstly, it provides insight into the more interesting measure metric and suggests dynamics and structure and characteristics for the measure metric to satisfy. Secondly, the particle metric is itself a description of the geometry underlying the *ODE* model of weighted aggregation developed in chapter 2, a model which may well be of value to scientific modeling, despite itself being less interesting mathematically than the nonlocal PDE model of the same.

4.4. Geodesic equations

The qualitative nature of this metric's geodesics should now be apparent. As opposed to the geodesics of 2-Wasserstein, which have mass follow straight lines, the mass is now encouraged to *spread radially* as it travels, and in particular for radially decreasing function a, the mass is encouraged to *spread out*. In this case much of the mass is likely to invest in spatial paths a bit longer, at a cost for that of course, in order to travel much of that distance in less crowdedness, making up the cost in penalty savings.

We can especially imagine this whenever a final configuration in $\mathbb{R}^{dN}$ is merely some translation of the source configuration: the likely geodesic would be expected to spread out a bit initially as it begins to traverse the needed translation, and then to complete most of that translation at some spread arrangement, before finally settling back into tighter formation as it reaches its destination configuration.

Now, calculations.

4.4.1. The particle metric.

The particle metric is a classical Riemannian metric on $X = \mathbb{R}^{dN}$, so by classical differential geometry, the curve energy defined for a smooth curve $\gamma : [0, 1] \to X$,

$$\int_0^1 \|\dot{\gamma}(t)\|^2_{\gamma(t)}\, dt,$$

can be minimized to obtain length-minimizing geodesics parametrized proportionally to arclength, or in our terminology, cost-minimizing.

For geodesics calculations we further take $a \in C^1\left(\mathbb{R}^d\right)$, or possibly only $a \in C^1\left(\mathbb{R}^d \setminus \{0\}\right)$ whenever particles along a geodesic curve never coincide.

The corresponding Euler-Lagrange equations are as follows. For all $j \in \{1 \dots N\}$

$$\begin{aligned} 0 &= \left(\nabla_{\gamma_j} - \partial_t \nabla_{\dot{\gamma}_j}\right) \|\dot{\gamma}\|_\gamma^2 \\ &= \left(\nabla_{\gamma_j} - \partial_t \nabla_{\dot{\gamma}_j}\right) \sum_i \alpha_i(\gamma) |\dot{\gamma}_i|^2 \\ &= |\dot{\gamma}_j|^2 \sum_{k \neq j} \nabla a(\gamma_j - \gamma_k) + \sum_{i \neq j} |\dot{\gamma}_i|^2 \nabla a(\gamma_j - \gamma_i) - 2\partial_t \left[\alpha_j(\gamma) \dot{\gamma}_j\right] \end{aligned}$$

with

$$\partial_t \left[\alpha_j(\gamma) \dot{\gamma}_j\right] = \ddot{\gamma}_j \alpha_j(\gamma) + \dot{\gamma}_j \sum_{k \neq j} \nabla a(\gamma_j - \gamma_k) \cdot (\dot{\gamma}_j - \dot{\gamma}_k).$$

Changing indices and grouping, we may write the ODE succinctly as

$$\forall i \quad \alpha_i(\gamma) \ddot{\gamma}_i = \sum_{j \neq i} \frac{|\dot{\gamma}_i|^2 + |\dot{\gamma}_j|^2}{2} \nabla a(\gamma_i - \gamma_j) + \left[\sum_{j \neq i} \nabla a(\gamma_i - \gamma_j) \cdot (\dot{\gamma}_j - \dot{\gamma}_i)\right] \dot{\gamma}_i. \tag{4.4.1}$$

This is the fully simplified formulation of the geodesics. It is, however, also worth keeping in mind some intermediary formulations of the above, namely with a "∂_t" kept unexpanded:

$$\forall i \quad \partial_t \left[\alpha_i(\gamma) \dot{\gamma}_i\right] = \sum_{j \neq i} \frac{|\dot{\gamma}_i|^2 + |\dot{\gamma}_j|^2}{2} \nabla a(\gamma_i - \gamma_j) \tag{4.4.2}$$

as well as, more expanded,

$$\forall i \quad \alpha_i(\gamma) \ddot{\gamma}_i = \sum_{j \neq i} \frac{|\dot{\gamma}_i|^2 + |\dot{\gamma}_j|^2}{2} \nabla a(\gamma_i - \gamma_j) - \partial_t \left[\alpha_i(\gamma)\right] \dot{\gamma}_i. \tag{4.4.3}$$

Each of these three formulations of varying expandedness provides some insight.

Equation (4.4.1) is the expanded expression showing how individual particle acceleration is determined by the whole of particle positions, velocities, the gradient of the penalty/attention function, and total mass/attention seen by the particle in question.

Equation (4.4.2) is most compact, with the Euler-Lagrange time derivative unexpanded, showing its relation to particle speeds and changing attention.

Equation (4.4.3) keeps the large bracket from equation (4.4.1) unexpanded, as the time derivative of "total attention", providing perhaps the most insightful view on what is happening. Consider the expressions of this equation. Each particle experiences acceleration in precisely two ways: Firstly, as seen in the summation term, *toward* every other particle, supposing attention function a is radially decreasing. This matches the intuition above that expects particles to often first spread apart and later collapse together. Indeed, velocity away from a neighbor particle, followed by slowing and then reversing to become velocity toward the neighbor, is precisely acceleration toward the neighbor throughout. Secondly, as seen in the "∂_t" term, each particle experiences acceleration *in its current direction* based on changing crowdedness. This is positive acceleration when leaving crowdedness and deceleration when entering. Indeed, $\alpha_i(\gamma)$ represents "total attention of (or mass seen by) particle i in configuration γ", so its derivative is the changing crowdedness as seen by the particle. Interestingly, the magnitude of this acceleration is such that *relative* rate-of-uncrowding yields *relative* acceleration (due to the "$\alpha_i(\gamma)$" term on the left side of the equation, and due to $\dot{\gamma}_i$ containing its own magnitude).

4.4.2. The measure metric.

We next write the equations we expect to be geodesics for the measure metric, obtained directly as a formal limit of equation (4.4.2). We write it for a density ρ varying in time, in relation to its formal tangent vector, the velocity field v varying in time. The time derivative ∂_t in equation (4.4.2) becomes the material derivative $\partial_t + v \cdot \nabla$:

$$\forall x \quad (\partial_t + v(x) \cdot \nabla)\left[(a * \rho)(x)\, v(x)\right] = \int \frac{|v(x)|^2 + |v(y)|^2}{2} \nabla a(x-y)\, \rho(y)\ dy$$

which is rewritten

$$0 = (\partial_t + v(x) \cdot \nabla)\left[(a * \rho)(x)\, v(x)\right] - \frac{|v(x)|^2}{2} \nabla (a * \rho)(x) - \frac{1}{2} \nabla \left[a * \left(\rho |v|^2\right)\right](x) \tag{4.4.4}$$

This equation couples with the continuity equation

$$\partial_t \rho + \operatorname{div}(\rho v) = 0$$

to formally yield the expected geodesics.

Next observe that any such v corresponding to a geodesic has a bit of structure, due to a Helmholtz-like decomposition of $\mathrm{L}^2((a * \rho)\, \rho\, dx)$: Formally, we can take

$$v = \frac{\nabla p}{a * \rho}$$

for some "pressure" p varying in space and time. To see this, consider alternatively that a geodesic exists having ρ, v with $v = \frac{\nabla p}{a*\rho} + w$ for some p and for nonzero w that is orthogonal in space, with respect to $\mathrm{L}^2\left((a*\rho)\,\rho\,dx\right)$, to the set of vector fields of the form $\frac{\nabla q}{a*\rho}$. Recall that the geodesic locally minimizes the cost

$$\begin{aligned} & \int_0^1 \int (a*\rho)\,|v|^2 \rho\,dx\,dt \\ = & \int_0^1 \int (a*\rho)\,|\frac{\nabla p}{a*\rho} + w|^2 \rho\,dx\,dt \\ = & \int_0^1 \int (a*\rho)\left[|\frac{\nabla p}{a*\rho}|^2 + |w|^2\right]\rho\,dx\,dt. \end{aligned}$$

Also by orthogonality, for all $q \in \mathrm{H}^1\left(\mathbb{R}^d\right)$

$$\begin{aligned} 0 &= \int (a*\rho)\,\frac{\nabla q}{a*\rho} \cdot w\rho\,dx \\ &= -\int q\,\mathrm{div}\,(w\rho)\,\,dx. \end{aligned}$$

Thus

$$\mathrm{div}\,(w\rho) = 0$$

which means that $\tilde{v} = \frac{\nabla p}{a*\rho}$ shares the same equivalence class as v, and $\tilde{v}$ yields better cost, thus $\tilde{v}$ could be taken instead of v.

Setting $v = \frac{\nabla p}{a*\rho}$ in the Euler-Lagrange equation (4.4.4) yields formally

$$\begin{aligned} 0 &= \partial_t \nabla p + \frac{1}{a*\rho}\nabla p \cdot \nabla\nabla p - \frac{|\nabla p|^2}{2\,(a*\rho)^2}\nabla\,(a*\rho) - \frac{1}{2}\nabla\left[a * \frac{\rho|\nabla p|^2}{(a*\rho)^2}\right] \\ &= \partial_t \nabla p + \frac{1}{2}\frac{(a*\rho)\,\nabla|\nabla p|^2 - |\nabla p|^2\nabla\,(a*\rho)}{(a*\rho)^2} - \frac{1}{2}\nabla\left[a * \frac{\rho|\nabla p|^2}{(a*\rho)^2}\right] \\ &= \partial_t \nabla p + \frac{1}{2}\nabla\frac{|\nabla p|^2}{a*\rho} - \frac{1}{2}\nabla\left[a * \frac{\rho|\nabla p|^2}{(a*\rho)^2}\right] \\ &= \nabla\left\{\partial_t p + \frac{1}{2}\frac{|\nabla p|^2}{a*\rho} - \frac{1}{2}\left[a * \frac{\rho|\nabla p|^2}{(a*\rho)^2}\right]\right\}. \end{aligned}$$

Thus the expression

$$\partial_t p + \frac{1}{2}\frac{|\nabla p|^2}{a * \rho} - \frac{1}{2}\left[a * \frac{\rho|\nabla p|^2}{(a*\rho)^2}\right]$$

is a function of time uniform in space, which may be chosen to be zero (because any other time function would merely give a different p with the same spatial derivative at each time).

Thus another description of our formal geodesics, using pressure p, is the coupled system

$$\begin{aligned} \partial_t \rho + \operatorname{div}\left(\rho \frac{\nabla p}{a * \rho}\right) &= 0 \\ \partial_t p + \frac{1}{2}\frac{|\nabla p|^2}{a * \rho} - \frac{1}{2}\left[a * \frac{\rho|\nabla p|^2}{(a*\rho)^2}\right] &= 0. \end{aligned}$$

CHAPTER 5

Finite-cost infinite spreading

In our consideration of the geometry of the particle metric and measure metric, we now turn to a narrower study: that of "spreading".

Mass spreading has been motivated in the prior sections as a primary phenomenon of this geometry, as observed by the geodesic equation behavior and by the intuitive interpretation of the effects of the penalty function a, namely in the case of radially decreasing a. The "spread out and translate" strategy is clearly evident for curves between a source configuration and a target configuration when the target configuration is merely a translation of the source. This strategy also is a plausible description of curves between other source and target configurations, whenever various translations must be pursued by components of the source configuration. (Note under the 2-Wasserstein metric in displacement interpolation form, components of the source configuration must in general pursue translations toward target positions as determined by the optimal coupling.)

The penalty function a may be regarded as a *lessening* of the cost of the 2-Wasserstein, since $a \equiv 1$ corresponds to the 2-Wasserstein and for convenience we have normalized bounded a to $\sup a\left(\mathbb{R}^d\right) = 1$. The introduction of penalty a then raises the possibility that mass motion may become dramatically cheaper as the mass spreads, corresponding to decaying a, possibly to the point that a bound exists on the cost of a spreading plan regardless of the amount of that spreading.

Said another way, the possibility has been introduced that mass may be able to spread "infinitely" at finite cost.

5.1. Conventions

We identify here a few key conventions used in this chapter.

$\mathrm{P}_c\left(\mathbb{R}^d\right)$ denotes the space of compactly supported Borel probability measures in $\mathbb{R}^d$.

e_1 denotes $(1, 0, 0, \ldots 0) \in \mathbb{R}^d$, used frequently in the inputs of radially symmetric functions.

$\mathcal{L}^d$ denotes Lebesgue measure in $\mathbb{R}^d$.

For $x \in \mathbb{R}^d$, δ_x denotes the Dirac measure at x: for all Borel $A \subset \mathbb{R}^d$, $\delta_x(A) = \chi_A(x)$.

For Borel probability measures μ, ν in $\mathbb{R}^d$, $\Pi(\mu, \nu)$ denotes the set of couplings of μ and ν, i.e. the set of Borel measures in $\mathbb{R}^d \times \mathbb{R}^d$ having marginals μ and ν: for all $\pi \in \Pi(\mu, \nu)$ and Borel $A \subset \mathbb{R}^d$, $\pi\left(A \times \mathbb{R}^d\right) = \mu(A)$ and $\pi\left(\mathbb{R}^d \times A\right) = \nu(A)$.

$C > 0$ is used for "adaptive constant C" notation in this chapter, so that it may change value from one expression to the next when present. Our constants C are always independent of time t found in the expressions, since we frequently estimate an expression of t before later time integrating it, and any other independences may be seen from context.

5.2. Examples

5.2.1. Particle example.

For a first example of finite-cost infinite spreading, consider the simplest: two particles in $\mathbb{R}^d$ begin coincident, and then head in opposite directions. Self attention is set as zero (because otherwise finite-cost infinite spreading cannot occur). Parametrizing time by separation distance, we find the cost at time T to be

$$\int_0^T \sqrt{2a^\bullet(t)\left(\frac{1}{2}\right)^2}\,dt$$

where for convenience we do not restrict the time interval to be $[0, 1]$. A bound exists on this for all T precisely when

$$\int_0^\infty \sqrt{a^\bullet(t)}dt < \infty.$$

That is, we have finite-cost infinite spreading if and only if $\sqrt{a^\bullet}$ is integrable and $a > 0$, the latter required because $a_0 = 0$.

5.2.2. Measure example.

Likewise, examples are not difficult to find for the measure metric of this thesis. The measure metric has the disadvantage that "self attention" is inherently built-in and cannot be set to zero. Here we say "disadvantage" meaning from the point of view of the curve attempting to cut cost. On the other hand, the measure metric has the advantage that it can spread all its mass infinitesimally, in a smearing sort of way, not being stuck with clumps of mass as with the particle metric.

For the simplest example consider again a penalty function a with bounded support, and in fact the simplest such,

$$a = \chi_{\mathrm{B}(0,1)}$$

the characteristic function of the unit ball at the origin. Consider a measure curve that begins as the Dirac measure at the origin and then spreads mass in all directions uniformly. This is described by an expanding sphere which on its surface holds uniform distribution of the probability measure.

For this curve and often throughout this chapter, we find a Lagrangian description of the curve useful. Notice here for our expanding sphere the normal practice of initializing Lagrangian paths at the start of the curve evolution does not work. This is due to the singularity we have specified at that time: a point mass which must instantaneously disintegrate into many (all) directions, which Lagrangian paths starting there cannot achieve. Instead, any other time may be chosen to specify an initial measure which is then pushed forward and backward in time.

Let us prescribe at $t = 1$ initial measure the unit sphere S^{d-1} bearing uniform distribution of mass one. Let us evolve it forward and backward in time by Lagrangian paths originating on the sphere, directed radially outward forward in time with speed one. The result is the desired measure curve $(\mu_t)_{t\in[0,\infty)}$ in $\mathrm{P}_c\left(\mathbb{R}^d\right)$ which at time t is uniformly concentrated on the sphere of radius t, for all $t \in (0, \infty)$, and which at time zero is the Dirac measure at the origin due to finite-time collapse of paths.

For $t > 0$ the resulting cost rate

$$\int_{\mathbb{R}^d} (a * \mu_t)(x) \left|v(t,x)\right|^2 \, d\mu_t(x),$$

with $v \equiv 1$, simplifies to

$$(a * \mu_t)(te_1)$$

by symmetry of the integrand on the support of μ_t; which for our function a is then

$$\left(\chi_{\mathrm{B}(0,1)} * \mu_t\right)(te_1)$$

which equals the fraction of the surface of $t\,S^{d-1}$ that is within distance one of its pole at te_1.

For $t > 1$ this has both upper and lower bound of the form, under different constants C independent of t,

$$C\frac{1}{t^{d-1}}.$$

Thus the total cost

$$\int_0^\infty \sqrt{\int_{\mathbb{R}^d} (a * \mu_t)(x) \left|v(t,x)\right|^2 \, d\mu_t(x)}\, dt,$$

when split as an integral from 0 to 1 and from 1 to ∞, has the latter bounded above and below by, under different constants C,

$$C \int_1^\infty t^{-\frac{d-1}{2}}\, dt$$

which is finite if and only if $d \geq 4$ because $d \in \mathbb{N}$.

Note we have integrated to ∞ in the cost integral, as a convenience, rather than to 1 as in the definition. The meaning can be interpreted as follows. For any time $T > 0$ specifying a truncation of the (unending) measure curve $(\mu_t)_{t\in[0,\infty)}$, the resulting truncated curve $(\mu_t)_{t\in[0,T]}$ may then be time reparametrized to the interval $[0, 1]$ to meet the definition, without changing its cost. If the cost integral taken from zero to ∞ of the unending curve is finite, such "finite-cost infinite spreading" merely represents a bound on the costs of *all* truncated curves of this spreading, independent of spreading size.

5.3. Canonical spreading: the unit explosion

The previous example showed that for cutoff penalty $a = \chi_{\mathrm{B}(0,1)}$, four spatial dimensions provide sufficient "directions" for the mass to spread infinitely at finite cost, and any fewer number of dimensions does not—for spherical expansion.

This leads to the question of whether there are better geometric spreading strategies. The unsurprising improvement to spherical spreading and the likely overall best candidate seems to be spreading out as a ball, namely a uniformly dense one. We give this a name.

Definition. The *unit explosion* in $\mathbb{R}^d$ is the curve $(\mu_t)_{t\in[0,\infty)}$ in $\mathrm{P}_c\left(\mathbb{R}^d\right)$ defined by $\mu_0 = \delta_0$, the Dirac measure at the origin, and for $t \in (0, \infty)$

$$d\mu_t = \frac{\chi_{\mathrm{B}(0,t)}}{\mathcal{L}^d\left(\mathrm{B}\left(0,t\right)\right)} d\mathcal{L}^d.$$

We say the *explosion at* $x \in \mathbb{R}^d$ to mean the definition modified by centering at x instead of the origin.

We say the *explosion to size* $R > 0$ to mean the truncation of the curve to $(\mu_t)_{t\in[0,R]}$.

We say the *mass* $m \geq 0$ *explosion* to mean a scaling of all the measures by m, giving a curve in the space of finite measures but not necessarily probability measures.

The Lagrangian description of the unit explosion is as follows. At $t = 1$ let the initial measure be taken from above,

$$d\mu_1 = \frac{\chi_{\mathrm{B}(0,1)}}{\mathcal{L}^d\left(\mathrm{B}\left(0,1\right)\right)} d\mathcal{L}^d$$

i.e. the unit ball with uniform density. Let the Lagrangian paths originate from $\mathrm{B}(0,1)$ and for each such $x_0 \in \mathrm{B}(0,1)$ let $x_0 t$ be its location at time $t \in [0,\infty)$, forward and backward in time. Note the preservation of uniform density in the push-forward measure by scaling, and the finite-time collapse at $t = 0$ giving $\mu_0 = \delta_0$.

Let us see whether this spreading strategy improves upon spherical spreading for the cutoff penalty $a = \chi_{\mathrm{B}(0,1)}$.

First, calculating for arbitrary penalty a, the unit explosion $(\mu_t)_{t\in[0,\infty)}$ has cost rate for $t > 0$

$$\int_{\mathbb{R}^d} (a * \mu_t)(x) \, |v(t,x)|^2 \; d\mu_t(x)$$

with $v(t,x) := \frac{x}{t}$; which equals

$$\begin{aligned} & C\frac{1}{t^d} \int_{\mathrm{B}(0,t)} (a * \mu_t)(x) \left|\frac{x}{t}\right|^2 \; dx \\ = \; & C\frac{1}{t^d} \int_0^t (a * \mu_t)(re_1) \left(\frac{r}{t}\right)^2 r^{d-1} \; dr \end{aligned}$$

(by radial symmetry)

$$\begin{aligned} = \; & C\frac{1}{t^d} \int_0^1 (a * \mu_t)(pte_1) \, p^2 (pt)^{d-1} \, t \; dp \\ = \; & C \int_0^1 p^{d+1} \int_{\mathbb{R}^d} a(y - pte_1) \; d\mu_t(y) \; dp \\ = \; & C \int_0^1 p^{d+1} \frac{1}{t^d} \int_{\mathrm{B}(0,t)} a(y - pte_1) \; dy \, dp \\ = \; & C\frac{1}{t^d} \int_0^1 p^{d+1} \int_{\mathrm{B}(pte_1,t)} a(y) \; dy \, dp. \end{aligned} \tag{5.3.1}$$

Inserting $a = \chi_{\mathrm{B}(0,1)}$ and bounding from above, the cost rate expression (5.3.1) is less than or equal to

$$\begin{aligned} & C\frac{1}{t^d} \int_0^1 p^{d+1} \int_{\mathbb{R}^d} \chi_{\mathrm{B}(0,1)}(y) \; dy \, dp \\ & = C\frac{1}{t^d}. \end{aligned}$$

Whereas instead bounding from below, for $t > 2$ the cost rate expression (5.3.1) is greater than or equal to

$$C\frac{1}{t^d} \int_0^{\frac{1}{2}} p^{d+1} \int_{\mathrm{B}(pte_1,t)} \chi_{\mathrm{B}(0,1)}(y) \; dy \, dp$$

$$= C\frac{1}{t^d}$$

noting for all $p \in \left(0, \frac{1}{2}\right)$, $\mathrm{B}\left(0,1\right) \subset \mathrm{B}\left(pte_1, t\right)$.

Thus the total cost

$$\int_0^\infty \sqrt{\int_{\mathbb{R}^d} \left(a * \mu_t\right)\left(x\right) \left|v\left(t,x\right)\right|^2 \, d\mu_t\left(x\right)} \, dt,$$

when split as an integral from 0 to 2 and from 2 to ∞, has the latter bounded above and below by, under different constants C,

$$C\int_2^\infty t^{-\frac{d}{2}} \, dt$$

which is finite if and only if $d \geq 3$ because $d \in \mathbb{N}$.

So we find that expansion as a ball allows one fewer dimension than spherical expansion, for sufficient "directions" for the mass to spread infinitely at finite cost, for cutoff penalty a.

5.4. Explodability

Definition. Let $a \in \mathcal{A}$, defined in section 4.2.2. We say a is *explodable* if

$$\int_0^\infty \sqrt{\int_{\mathbb{R}^d} \left(a * \mu_t\right)\left(x\right) \left|\frac{x}{t}\right|^2 \, d\mu_t\left(x\right)} \, dt \; < \infty$$

for unit explosion $\left(\mu_t\right)_{t\in[0,\infty)}$.

Let us note some of the situations in which a is explodable or not.

5.4.1. Simple conditions.

The calculation of the previous section showed that $\chi_{\mathrm{B}(0,1)}$ is explodable if and only if the spatial dimension is three or higher. An immediate consequence is that *no* $a \in \mathcal{A}$ is explodable in dimensions one or two. This is noted by recalling $a \in \mathcal{A}$ is continuous at the origin with $a(0) > 0$, so that $a \geq \varepsilon \chi_{\mathrm{B}(0,\delta)}$ for some $\varepsilon, \delta > 0$, while noting $\varepsilon \chi_{\mathrm{B}(0,\delta)}$ shares the same cost calculation as $\chi_{\mathrm{B}(0,1)}$.

A second immediate consequence is that any $a \in \mathcal{A}$ with bounded support is explodable if and only if the spatial dimension is three or higher. This follows from the previous observation and from the boundedness requirement of $a \in \mathcal{A}$, so that $a \leq \chi_{\mathrm{B}(0,\delta)}$ for some $\delta > 0$.

So then, we consider $a \in \mathcal{A}$ with unbounded support in $\mathbb{R}^d$ with $d \geq 3$. A simple sufficient condition for explodability is that a is integrable. This is evident from expression (5.3.1), so that the total cost

$$\begin{aligned} & \int_0^\infty \sqrt{\int_{\mathbb{R}^d} (a * \mu_t)(x) \, |v(t,x)|^2 \; d\mu_t(x)} \, dt \\ = \; & C \int_0^\infty \sqrt{\frac{1}{t^d} \int_0^1 p^{d+1} \int_{\mathrm{B}(pte_1,t)} a(y) \; dy \, dp} \, dt, \end{aligned}$$

which when split as an integral from 0 to 1 and from 1 to ∞, has the latter bounded above by

$$\begin{aligned} & C \int_1^\infty \sqrt{\frac{1}{t^d} \int_0^1 p^{d+1} \int_{\mathbb{R}^d} a(y) \; dy \, dp} \, dt \\ = \; & C \int_1^\infty t^{-\frac{d}{2}} \, dt \; < \infty. \end{aligned}$$

However, this asks a rather conservative decay of penalty a, especially when the dimension d is large, as seen by a consideration of power laws.

5.4.2. Power law decay.

Again assume $d \geq 3$. Consider penalty function $a \in \mathcal{A}$

$$a(x) := (|x| + 1)^{-q}$$

for some $q \geq 0$.

Cost rate expression (5.3.1) is written

$$\begin{aligned} & C\frac{1}{t^d}\int_0^1 p^{d+1}\int_{\mathrm{B}(pte_1,t)} a(y)\ dy\,dp \\ = \ & C\int_0^1 p^{d+1}\int_{\mathrm{B}(pe_1,1)} a(tz)\ dz\,dp \\ = \ & C\int_0^1 p^{d+1}\int_{\mathrm{B}(pe_1,1)} (t|z|+1)^{-q}\ dz\,dp. \end{aligned} \tag{5.4.1}$$

Suppose $q > 2$. This expression is bounded above by

$$C\int_0^1 p^{d+1}\int_{\mathrm{B}(pe_1,1)} (t|z|+1)^{-\hat{q}}\ dz\,dp$$

for $\hat{q} = \min\{2.5, q\}$, which then is bounded above by

$$\begin{aligned} & C\int_0^1 p^{d+1}\int_{\mathrm{B}(pe_1,1)} (t|z|)^{-\hat{q}}\ dz\,dp \\ = \ & Ct^{-\hat{q}}\int_0^1 p^{d+1}\int_{\mathrm{B}(pe_1,1)} |z|^{-\hat{q}}\ dz\,dp \\ = \ & Ct^{-\hat{q}} \end{aligned}$$

noting $|z|^{-\hat{q}}$ is integrable at the origin because $\hat{q} < d$.

So the total cost

$$\int_0^\infty \sqrt{\int_{\mathbb{R}^d} (a * \mu_t)(x)\,|v(t,x)|^2\ d\mu_t(x)}\ dt,$$

when split as an integral from 0 to 1 and from 1 to ∞, has the latter bounded above by

$$C\int_1^\infty t^{-\frac{\hat{q}}{2}}\ dt\ < \infty$$

because $\hat{q} > 2$.

Conversely, suppose $0 \le q \le 2$. For $t > 1$ the cost rate expression (5.4.1) is bounded below by

$$\begin{aligned} & C\int_0^1 p^{d+1}\int_{\mathrm{B}(pe_1,1)} (t\,|z|+t)^{-q}\; dz\,dp \\ = \;& Ct^{-q}\int_0^1 p^{d+1}\int_{\mathrm{B}(pe_1,1)} (|z|+1)^{-q}\; dz\,dp \\ = \;& Ct^{-q}. \end{aligned}$$

So the total cost

$$\begin{aligned} & \int_0^\infty \sqrt{\int_{\mathbb{R}^d} (a*\mu_t)(x)\,|v(t,x)|^2\; d\mu_t(x)}\; dt \\ \geq \;& C\int_1^\infty t^{-\frac{q}{2}}\; dt \;=\infty. \end{aligned}$$

Thus the power law is explodable if and only if $q > 2$ and $d \geq 3$. This gives a rough characterization of explodability in $\mathcal{A}$.

5.5. Cheap translation

Up to now in this chapter we have examined the cost for which mass concentrated at a point may spread infinitely. We should ask now what benefits may be reaped from this spreading when it can occur, from the curve's point of view as far as connecting source and target configurations with short curves, i.e. low cost.

It is natural to wonder whether arbitrarily large spreading may in fact cause the cost of subsequent translation to be arbitrarily small. Specifically, whether an explosion to an arbitrarily large size might control the cost of subsequent translation of that large ball to be arbitrarily small per distance translated.

This together with explodability would imply that a Dirac measure can be transported over *arbitrary* distance in $\mathbb{R}^d$ to a target Dirac measure, at barely above twice the cost of the unit explosion: the Dirac would merely explode to appropriate size, then translate exactly the displacement between the source and target Dirac measures, at little cost, and finally implode back to a Dirac measure. That is, we would have a bound on the pairwise metric distance between all Dirac measures.

Of course, the explosion size might well dwarf the size of the translation, engulfing the target location during spreading. But no matter. The center still translates the displacement.

It turns out the answer is affirmative for every explodable $a \in \mathcal{A}$. Note this does not require a to be radially decreasing nor its radial profile to have a limit at infinity (of zero, as it would).

The answer should be qualified, however: arbitrarily cheap translation is not obtained from *sufficiently large* explosion size, but rather from *some* large sizes, with always such a size available beyond any current or given size.

To see this, consider the concluding measure after explosion at the origin to size $R > 0$,

$$d\mu_R = \frac{\chi_{\mathrm{B}(0,R)}}{\mathcal{L}^d\left(\mathrm{B}\left(0,R\right)\right)} d\mathcal{L}^d.$$

Consider a curve in $\mathrm{P}_c\left(\mathbb{R}^d\right)$ that is a constant-speed translation of this measure in some direction in $\mathbb{R}^d$ for some duration of time. Such curve has cost rate independent of time

$$\begin{aligned}
& \int_{\mathbb{R}^d} \left(a * \mu_R\right)(x) \left|v_0\right|^2 \, d\mu_R\left(x\right) \\
= \; & C \frac{1}{R^d} \int_{\mathrm{B}(0,R)} \left(a * \mu_R\right)(x) \, dx \\
= \; & C \frac{1}{R^d} \int_0^R \left(a * \mu_R\right)(re_1) \, r^{d-1} \, dr \\
= \; & C \frac{1}{R^d} \int_0^1 \left(a * \mu_R\right)(pRe_1) \, (pR)^{d-1} \, R \, dp \\
= \; & C \frac{1}{R^d} \int_0^1 p^{d-1} \int_{\mathrm{B}(pRe_1,R)} a\left(y\right) \, dy \, dp \\
\leq \; & C \frac{1}{R^d} \int_0^1 p^{d-1} \int_{\mathrm{B}(0,2R)} a\left(y\right) \, dy \, dp \\
= \; & C \frac{1}{R^d} \int_{\mathrm{B}(0,2R)} a\left(y\right) \, dy. \qquad (5.5.1)
\end{aligned}$$

If a is explodable, from expression (5.3.1) we have for unit explosion $(\mu_t)_{t\in[0,\infty)}$

$$\begin{aligned}
\infty &> \int_0^\infty \sqrt{\int_{\mathbb{R}^d} (a * \mu_t)(x) \left|\frac{x}{t}\right|^2 \, d\mu_t(x)} \, dt \\
&= C \int_0^\infty \sqrt{\frac{1}{t^d} \int_0^1 p^{d+1} \int_{\mathrm{B}(pte_1,t)} a(y) \, dy \, dp} \, dt \\
&\geq C \int_0^\infty \sqrt{\frac{1}{t^d} \int_0^{\frac{1}{2}} p^{d+1} \int_{\mathrm{B}\left(0,\frac{t}{2}\right)} a(y) \, dy \, dp} \, dt \\
&= C \int_0^\infty \sqrt{\frac{1}{t^d} \int_{\mathrm{B}\left(0,\frac{t}{2}\right)} a(y) \, dy} \, dt \\
&= C \int_0^\infty \sqrt{\frac{1}{(4s)^d} \int_{\mathrm{B}(0,2s)} a(y) \, dy} \, ds \\
&= C \int_0^\infty \sqrt{\frac{1}{s^d} \int_{\mathrm{B}(0,2s)} a(y) \, dy} \, ds.
\end{aligned}$$

Thus

$$\liminf_{s\to\infty} \frac{1}{s^d} \int_{\mathrm{B}(0,2s)} a(y) \, dy = 0$$

which provides the claimed control of bound (5.5.1).

5.6. Metric boundedness

In the previous section we found that for all explodable $a \in \mathcal{A}$ the set of Dirac measures in $\mathbb{R}^d$ is bounded with respect to the measure metric of this thesis d_{E}.

We should ask, then, how this applies to other measures in $\mathrm{P}_{\mathrm{E}}\left(\mathbb{R}^d\right)$. It is natural to wonder whether *any* given measure can find geometric means to spread "infinitely" in some sense, at finite cost. The intuition certainly is there: mass concentrated at a point, i.e. a Dirac measure, seems to be about the worst case for mass attempting to escape from itself, at least in the case of radially decreasing penalty function a. But in what directions would the mass components of a more arbitrary measure spread, and in what "infinitely spreading" geometric formation?

It is also natural to wonder, if such finite-cost infinite spreading is available to more arbitrary measures, whether they may also enjoy arbitrarily cheap translation after spreading as do the Dirac measures; and then whether this results in a bounded metric space on the whole.

It turns out the answers are affirmative for the subspace $\mathrm{P}_2\left(\mathbb{R}^d\right) \subset \mathrm{P}_{\mathrm{E}}\left(\mathbb{R}^d\right)$, for radially decreasing explodable $a \in \mathcal{A}$. A bound on the diameter of $\mathrm{P}_2\left(\mathbb{R}^d\right)$ is given by a factor of the cost of the unit

explosion, with the factor dependent on dimension d. The "infinitely spreading formation" mentioned above is seen shortly within the argument.

5.6.1. Radial monotonicity lemma.

To name the bound on diameter we need first to collect an observation of monotonicity when integrating radially symmetric, radially decreasing functions on balls. Fix ball size $\hat{r} > 0$, and consider the convolution function of x

$$a * \chi_{\mathrm{B}(0,\hat{r})}(x) = \int_{\mathrm{B}(x,\hat{r})} a.$$

The convolution is radially symmetric by radial symmetry of a. Further, the convolution is radially decreasing by radially decreasing, radially symmetric a, as a result of rearrangement theory.

Next observe $\overline{\mathrm{B}(0,\hat{r})}$ can be covered by a finite number of copies of itself where each is translated to have its center on the boundary of $\overline{\mathrm{B}(0,\hat{r})}$. This is evident by noting that any such translated copy contains a nontrivial sector of $\overline{\mathrm{B}(0,\hat{r})}$: namely, the intersection of $\overline{\mathrm{B}(0,\hat{r})}$ with the convex cone generated by a nonempty nonsingleton spherical cap, which is the boundary of $\overline{\mathrm{B}(0,\hat{r})}$ intersected with the copy.

The number of copies required for the cover depends only on dimension d; since d is constant in this section let Ξ denote the minimum number.

The closure of $\mathrm{B}(0,\hat{r})$ was used for convenience to cover the origin and its cones, but of course does not matter when we now integrate over it and its copies.

Thus for all $\hat{r} > 0, x \in \mathbb{R}^d, y \in \mathrm{B}(0,\hat{r})$

$$\int_{\mathrm{B}(x,\hat{r})} a \quad \leq \quad \int_{\mathrm{B}(0,\hat{r})} a \quad \leq \quad \Xi \int_{\mathrm{B}(\hat{r}e_1,\hat{r})} a \quad \leq \quad \Xi \int_{\mathrm{B}(y,\hat{r})} a. \tag{5.6.1}$$

5.6.2. Theorem statement and beginning of the proof.

Theorem. Let radially decreasing, explodable $a \in \mathcal{A}$. Let Ξ be defined as above. Let $\overline{\mathrm{P}}_2\left(\mathbb{R}^d\right)$ denote the closure of $\mathrm{P}_2\left(\mathbb{R}^d\right)$ in $\left(\mathrm{P}_{\mathrm{E}}\left(\mathbb{R}^d\right), d_{\mathrm{E}}\right)$ defined in section 4.2.2. Then $\overline{\mathrm{P}}_2\left(\mathbb{R}^d\right)$ is a bounded metric space with diameter less than or equal to $2\sqrt{\Xi}$ times the cost of the unit explosion.

Proof.

Let $\mu, \nu \in \mathrm{P}_2\left(\mathbb{R}^d\right)$, $\varepsilon > 0$, and let $\mathcal{C} \in (0,\infty)$ be the cost of the unit explosion.

We show $d_{\mathrm{E}}(\mu,\nu) < 2\sqrt{\Xi}\,\mathcal{C} + 24\varepsilon$ by identifying an intermediate measure $\mu_* \in \mathrm{P}_c\left(\mathbb{R}^d\right)$ such that $d_{\mathrm{E}}(\mu,\mu_*) < \sqrt{\Xi}\,\mathcal{C} + 12\varepsilon$ and $d_{\mathrm{E}}(\nu,\mu_*) < \sqrt{\Xi}\,\mathcal{C} + 12\varepsilon$. The bound on the distance between μ and μ_* is

shown by identifying 5 intermediate measures in sequence between *them*. By identical construction using ν in place of μ, the distance ν to μ_* may likewise be bounded.

First we approximate μ and ν each by a finite sum of partial Dirac measures, using the denseness of such in $\left(\mathrm{P}_2\left(\mathbb{R}^d\right), d_{\mathrm{W}_2}\right)$. Let

$$\mu_0 = \sum_{i=1}^{M} m_i \delta_{x_i}$$

where $m_i > 0$, $\sum m_i = 1$, $\forall i \neq j \;\; x_i \neq x_j$, and $d_{\mathrm{W}_2}(\mu, \mu_0) < \varepsilon$. We have

$$d_{\mathrm{E}}(\mu, \mu_0) < \varepsilon$$

because $d_{\mathrm{E}} \leq d_{\mathrm{W}_2}$.

Likewise let ν_0 approximate ν with its own sum, which we do not label.

Let $D > 0$ such that the supports of μ_0 and ν_0 lie within $\mathrm{B}(0, D)$.

Let $R > 0$ such that an explosion at the origin to size R concludes as a measure that may translate at speed D with cost rate less than $\varepsilon^2 \Xi^{-1}$, as provided in section 5.5; and let $\mu_* \in \mathrm{P}_c\left(\mathbb{R}^d\right)$ be this concluding measure.

Let curve $(\mu_t)_{t\in[0,R]}$ in $\mathrm{P}_c\left(\mathbb{R}^d\right)$ be defined as the sum over i of the mass m_i explosion at x_i to size R. Note μ_0 so defined matches the μ_0 we already have. For all $t \in (0, R]$,

$$d\mu_t = \sum_i m_i \; \rho_{i,t} \; d\mathcal{L}^d$$

where $\rho_{i,t}(x) := \bar{\rho}_t(x - x_i)$ with $\bar{\rho}_t$ the Lebesgue density function of the unit explosion at time t.

Some discussion now to explain what comes next. Intuitively, this "simultaneous explosion" $(\mu_t)_{t\in[0,R]}$ is precisely the curve we want in order to progress from μ_0 to a measure spread out in all directions. This is the "infinitely spreading formation" mentioned in the introduction of this section. Its concluding measure μ_R, like μ_0, is one of our 5 intermediate measures from μ to μ_*. However, the velocity fields of the individual explosion components begin to overlap at some time, possibly before time R. These conflicting velocity fields are resolved by approximating the explosion components of $(\mu_t)_{t\in[0,R]}$ with "particle explosion" components which encounter no conflict with each other. First, the original explosion components expand from their point masses to some small size for which none have yet overlapped; then such concluding measures are approximated by finite sums of partial Dirac measures, and these massed particles continue to expand precisely along the point paths that had

been transporting the continuum explosion components, to time R; and finally those spread-out massed particles are noted to approximate μ_R.

We continue the proof in the next subsection.

5.6.3. The explosion phase.

Let $T \in (0, R)$ such that $\forall i \neq j,\ T < \frac{1}{3}|x_i - x_j|$. T is the time (and size) at which the continuum explosion components are stopped and approximated.

Let $\delta > 0$ such that $\delta R < \varepsilon$, $2\delta < \frac{\varepsilon^2}{R^2}$, and such that for all $q \in [-2\delta R, 2\delta R]$,

$$\left| \int_{\mathbb{R}^d} \left[a^\bullet - a_q^\bullet \right] (|x|)\ dx \right| \ < \ \frac{\varepsilon^2}{R^2} \mathcal{L}^d \left(\mathrm{B}\left(0, T\right) \right)$$

where for $q \in \mathbb{R}$ and $r \geq 0,\ a_q^\bullet(r) := a^\bullet\left((r-q)_+\right)$ with $s_+ := \max\{s, 0\}$.

The last condition is justified because

$$\lim_{q \to 0} \int_{\mathbb{R}^d} \left[a^\bullet - a_q^\bullet \right] (|x|)\ dx = 0$$

from the theory of such shift functions.

For $t \in (0, R]$ let $B_t := \bigcup_i \mathrm{B}\left(x_i, t\right)$, and for $x \in B_T$ let $x_\square$ denote the nearest member of $\{x_1, \ldots x_M\}$ to x.

Let $\Phi : [0, R] \times B_T \to \mathbb{R}^d$ by

$$\Phi_t(x) := \frac{t}{T}\left(x - x_\square\right) + x_\square.$$

Note $\Phi_T(\cdot)$ is the identity map on B_T, and for all $t \in [0, R]$,

$$\mu_t = \Phi_t(\cdot)_\# \mu_T.$$

Let finite collection of Borel sets $(A_{ij})_{i,j}$ in $\mathbb{R}^d$ be such that for each i, $(A_{ij})_j$ is a partition of $\mathrm{B}\left(x_i, T\right)$, and for all i, j, $\mathrm{diam}\left(A_{ij}\right) < \delta T$ and $\mu_T\left(A_{ij}\right) > 0$. For each i, j let $y_{ij} \in A_{ij}$, and let

$$\sigma_T = \sum_{ij} \mu_T\left(A_{ij}\right) \delta_{y_{ij}}.$$

Let $\pi_T \in \Pi\left(\mu_T, \sigma_T\right)$ be the coupling resulting from the map for which for all i, j each member of A_{ij} maps to y_{ij}.

For $t \in [T, R]$ let

$$\sigma_t := \Phi_t(\cdot)_\# \sigma_T$$

and

$$\pi_t := (\Phi_t(\cdot) \times \Phi_t(\cdot))_{\#} \pi_T.$$

Our 3 intermediate measures between μ_0 and μ_R are now identified: $\mu_T, \sigma_T, \sigma_R$.

Note for all $t \in [T, R]$ and (x, y) in the support of π_t,

$$|x - y| < \delta t.$$

Also note for all $t \in [T, R]$,

$$d_{\mathrm{E}}(\mu_t, \sigma_t) \leq d_{W_2}(\mu_t, \sigma_t) \leq \frac{t}{T} \mathrm{diam}(A_{ij}) < \delta t \leq \delta R < \varepsilon.$$

Thus $d_{\mathrm{E}}(\mu_T, \sigma_T) < \varepsilon$ and $d_{\mathrm{E}}(\sigma_R, \mu_R) < \varepsilon$.

We have left to bound $d_{\mathrm{E}}(\mu_0, \mu_T)$, $d_{\mathrm{E}}(\sigma_T, \sigma_R)$, and $d_{\mathrm{E}}(\mu_R, \mu_*)$. The third is done in section 5.6.6, whereas the former two we take up now and through the next two sections, which are achieved by bounding the costs of the curves $(\mu_t)_{t\in[0,T]}$ and $(\sigma_t)_{t\in[T,R]}$ in $\mathrm{P}_c(\mathbb{R}^d)$.

Note both curves admit a global velocity field from the definition of d_{E} given directly by the Lagrangian paths of Φ, although with a wrinkle in the case of the latter curve. The Lagrangian paths of Φ are not guaranteed to be free of intersections after time T. However, note no pair of paths may intersect twice, due to their linear paths. Thus the *finitely many* paths characterizing $(\sigma_t)_{t\in[T,R]}$ experience intersections at *finitely many* times. The velocity field from the definition of d_{E} belongs to $\mathrm{L}^1([0,1]; \mathrm{L}^2((a * \sigma_t)\, d\sigma_t))$; thus the velocity field here needn't be defined at any times of intersections, and our cost estimates are not disturbed by these times.

The curve $(\mu_t)_{t\in[0,T]}$ admits a velocity field v satisfying for all $t \in (0, T)$ and $x \in B_T$,

$$v(t, \Phi_t(x)) = \partial_t \Phi_t(x) = \frac{x - x_{\square}}{T}.$$

Whereas the curve $(\sigma_t)_{t\in[T,R]}$ admits a velocity field v satisfying the very same condition for a.e. $t \in (T, R)$ and all x in $(y_{ij})_{i,j}$.

We use the same variable v for both because the time intervals $(0, T)$ and (T, R) are disjoint.

For all $t \in (0, T)$, the cost rate of the curve $(\mu_t)_{t\in[0,T]}$ is

$$\begin{aligned} & \int_{\mathbb{R}^d} (a * \mu_t)(x) \, |v(t, x)|^2 \, d\mu_t(x) \\ = & \int_{B_T} (a * \mu_t)(\Phi_t(\tilde{x})) \, |\partial_t \Phi_t(\tilde{x})|^2 \, d\mu_T(\tilde{x}) \end{aligned}$$

and for a.e. $t \in (T,R)$, the cost rate of the curve $(\sigma_t)_{t\in[T,R]}$ is

$$\begin{aligned} & \int_{\mathbb{R}^d} (a*\sigma_t)(y)\,|v(t,y)|^2\ d\sigma_t(y) \\ = & \int_{B_T} (a*\sigma_t)(\Phi_t(\tilde{y}))\,|\partial_t\Phi_t(\tilde{y})|^2\ d\sigma_T(\tilde{y})\,. \end{aligned}$$

We now compare the latter cost rate expression from each of these on the interval (T,R), despite that (μ_t) is not employed on $[T,R]$. After finding the two to be within $\frac{2\varepsilon^2}{R^2}$, we subsequently bound the one from (μ_t) on the whole interval $(0,R)$ as a means of bounding the sum of the costs of the two curves.

5.6.4. Cost rate comparison during the approximated explosion.

For a.e. $t \in (T,R)$,

$$\begin{aligned} & \left| \int_{B_T} (a*\mu_t)(\Phi_t(\tilde{x}))\,|\partial_t\Phi_t(\tilde{x})|^2\ d\mu_T(\tilde{x}) - \int_{B_T} (a*\sigma_t)(\Phi_t(\tilde{y}))\,|\partial_t\Phi_t(\tilde{y})|^2\ d\sigma_T(\tilde{y}) \right| \\ = & \left| \iint_{B_T^2} \left[(a*\mu_t)(\Phi_t(\tilde{x}))\,|\partial_t\Phi_t(\tilde{x})|^2 - (a*\sigma_t)(\Phi_t(\tilde{y}))\,|\partial_t\Phi_t(\tilde{y})|^2 \right] d\pi_T(\tilde{x},\tilde{y}) \right| \end{aligned}$$

which, following the form "$bc-de = b[c-e]+e[b-d]$", equals

$$\left| \iint_{B_T^2} \left[(a*\mu_t)(\Phi_t(\tilde{x})) \left[|\partial_t\Phi_t(\tilde{x})|^2 - |\partial_t\Phi_t(\tilde{y})|^2 \right] + |\partial_t\Phi_t(\tilde{y})|^2 \left[(a*\mu_t)(\Phi_t(\tilde{x})) - (a*\sigma_t)(\Phi_t(\tilde{y})) \right] \right] d\pi_T(\tilde{x},\tilde{y}) \right|$$

which is less than or equal to

$$\iint_{B_T^2} \left[(a*\mu_t)(\Phi_t(\tilde{x})) \left| |\partial_t\Phi_t(\tilde{x})|^2 - |\partial_t\Phi_t(\tilde{y})|^2 \right| + |\partial_t\Phi_t(\tilde{y})|^2 \left| (a*\mu_t)(\Phi_t(\tilde{x})) - (a*\sigma_t)(\Phi_t(\tilde{y})) \right| \right] d\pi_T(\tilde{x},\tilde{y})\,.$$

For a.e. $t \in (T,R)$ and all $(\tilde{x},\tilde{y})$ in the support of π_T,

$$\begin{aligned} & (a*\mu_t)(\Phi_t(\tilde{x})) \left| |\partial_t\Phi_t(\tilde{x})|^2 - |\partial_t\Phi_t(\tilde{y})|^2 \right| \\ \leq & \left| |\partial_t\Phi_t(\tilde{x})|^2 - |\partial_t\Phi_t(\tilde{y})|^2 \right| \\ = & \frac{1}{T^2} \left| |\tilde{x}-\tilde{x}_\square|^2 - |\tilde{y}-\tilde{y}_\square|^2 \right| \\ = & \frac{1}{T^2} \left| |\tilde{x}-\tilde{x}_\square|^2 - |\tilde{y}-\tilde{x}_\square|^2 \right| \end{aligned}$$

because $\tilde{x},\tilde{y}$ belong to the same $A_{ij} \subset \mathrm{B}(x_i,T)$ for some i,j. Following the form

$$\left| |u|^2 - |w|^2 \right| = \left| |u|+|w| \right| \ \left| |u|-|w| \right| \leq (|u|+|w|)\,|u-w|$$

for $u, w \in \mathbb{R}^d$, the previous expression is then less than or equal to

$$\frac{1}{T^2}\left(\left|\tilde{x}-\tilde{x}_{\square}\right|+\left|\tilde{y}-\tilde{x}_{\square}\right|\right)\left|\left(\tilde{x}-\tilde{x}_{\square}\right)-\left(\tilde{y}-\tilde{x}_{\square}\right)\right|$$

$$<\frac{1}{T^2}\ (T+T)\ \delta T<\frac{\varepsilon^2}{R^2}$$

because $|\tilde{x}-\tilde{y}|<\delta T$.

Also for a.e. $t \in (T, R)$ and all $(\tilde{x}, \tilde{y})$ in the support of π_T,

$$\begin{aligned}
& \left|\partial_t \Phi_t(\tilde{y})\right|^2\left|\left(a * \mu_t\right)\left(\Phi_t(\tilde{x})\right)-\left(a * \sigma_t\right)\left(\Phi_t(\tilde{y})\right)\right| \\
= & \frac{1}{T^2}\left|\tilde{y}-\tilde{y}_{\square}\right|^2\left|\iint_{B_t^2}\left[a\left(x-\Phi_t(\tilde{x})\right)-a\left(y-\Phi_t(\tilde{y})\right)\right] d\pi_t(x, y)\right| \\
< & \left|\iint_{B_t^2}\left[a\left(x-\Phi_t(\tilde{x})\right)-a\left(y-\Phi_t(\tilde{y})\right)\right] d\pi_t(x, y)\right| .
\end{aligned}$$

For a.e. $t \in (T, R)$ and all $(\tilde{x}, \tilde{y})$ in the support of π_T and (x, y) in the support of π_t,

$$\begin{aligned}
& a\left(y-\Phi_t(\tilde{y})\right) \\
= & a^{\bullet}\left(\left|\left(x-\Phi_t(\tilde{x})\right)+\left(\Phi_t(\tilde{x})-\Phi_t(\tilde{y})\right)+(y-x)\right|\right) \\
\geq & a^{\bullet}\left(\left|x-\Phi_t(\tilde{x})\right|+\delta R+\delta R\right)
\end{aligned}$$

because radial profile $a^{\bullet}$ of a is decreasing, and $\left(\Phi_t(\tilde{x}), \Phi_t(\tilde{y})\right)$ belongs to the support of π_t.

This equals, using the shift notation defined above,

$$a_{-2\delta R}^{\bullet}\left(\left|x-\Phi_t(\tilde{x})\right|\right) .$$

Thus for a.e. $t \in (T,R)$ and all $(\tilde{x},\tilde{y})$ in the support of π_T,

$$\begin{aligned}
& \iint_{B_t^2} \left[a\left(x-\Phi_t\left(\tilde{x}\right)\right)-a\left(y-\Phi_t\left(\tilde{y}\right)\right)\right] d\pi_t\left(x,y\right) \\
\leq & \iint_{B_t^2} \left[a^\bullet\left(\left|x-\Phi_t\left(\tilde{x}\right)\right|\right)-a_{-2\delta R}^\bullet\left(\left|x-\Phi_t\left(\tilde{x}\right)\right|\right)\right] d\pi_t\left(x,y\right) \\
= & \int_{\mathbb{R}^d} \left[a^\bullet-a_{-2\delta R}^\bullet\right]\left(\left|x-\Phi_t\left(\tilde{x}\right)\right|\right)\ d\mu_t\left(x\right) \\
= & \int_{\mathbb{R}^d} \left[a^\bullet-a_{-2\delta R}^\bullet\right]\left(\left|\hat{x}\right|\right)\ d\mu_t\left(\hat{x}\right) \\
= & \sum_i m_i \int_{\mathbb{R}^d} \left[a^\bullet-a_{-2\delta R}^\bullet\right]\left(\left|\hat{x}\right|\right)\ \rho_{i,t}\left(\hat{x}\right)\ d\hat{x} \\
\leq & \frac{1}{\mathcal{L}^d\left(\mathrm{B}\left(0,T\right)\right)} \int_{\mathbb{R}^d} \left[a^\bullet-a_{-2\delta R}^\bullet\right]\left(\left|\hat{x}\right|\right)\ d\hat{x} \\
< & \frac{\varepsilon^2}{R^2}
\end{aligned}$$

noting $a^\bullet - a_{-2\delta R}^\bullet \geq 0$.

Similarly, for a.e. $t \in (T,R)$ and all $(\tilde{x},\tilde{y})$ in the support of π_T and (x,y) in the support of π_t,

$$\begin{aligned}
& a\left(y-\Phi_t\left(\tilde{y}\right)\right) \\
= & a^\bullet\left(\left|\left(x-\Phi_t\left(\tilde{x}\right)\right)+\left(\Phi_t\left(\tilde{x}\right)-\Phi_t\left(\tilde{y}\right)\right)+\left(y-x\right)\right|\right) \\
\leq & a^\bullet\left(\left[\left|x-\Phi_t\left(\tilde{x}\right)\right|-\left|\left(\Phi_t\left(\tilde{x}\right)-\Phi_t\left(\tilde{y}\right)\right)+\left(y-x\right)\right|\right]_+\right) \\
\leq & a^\bullet\left(\left[\left|x-\Phi_t\left(\tilde{x}\right)\right|-\delta R-\delta R\right]_+\right) \\
= & a_{2\delta R}^\bullet\left(\left|x-\Phi_t\left(\tilde{x}\right)\right|\right)
\end{aligned}$$

and so

$$\begin{aligned}
& \iint_{B_t^2} \left[a\left(x-\Phi_t\left(\tilde{x}\right)\right)-a\left(y-\Phi_t\left(\tilde{y}\right)\right)\right] d\pi_t\left(x,y\right) \\
\geq & \iint_{B_t^2} \left[a^\bullet\left(\left|x-\Phi_t\left(\tilde{x}\right)\right|\right)-a_{2\delta R}^\bullet\left(\left|x-\Phi_t\left(\tilde{x}\right)\right|\right)\right] d\pi_t\left(x,y\right) \\
= & \sum_i m_i \int_{\mathbb{R}^d} \left[a^\bullet-a_{2\delta R}^\bullet\right]\left(\left|\hat{x}\right|\right)\ \rho_{i,t}\left(\hat{x}\right)\ d\hat{x} \\
\geq & \frac{1}{\mathcal{L}^d\left(\mathrm{B}\left(0,T\right)\right)} \int_{\mathbb{R}^d} \left[a^\bullet-a_{2\delta R}^\bullet\right]\left(\left|\hat{x}\right|\right)\ d\hat{x} \\
> & -\frac{\varepsilon^2}{R^2}
\end{aligned}$$

noting $a^\bullet - a_{2\delta R}^\bullet \leq 0$.

Thus

$$\left| \iint_{B_t^2} \left[a\left(x - \Phi_t\left(\tilde{x}\right)\right) - a\left(y - \Phi_t\left(\tilde{y}\right)\right) \right] d\pi_t\left(x, y\right) \right| < \frac{\varepsilon^2}{R^2}$$

and so we obtain a bound on the original expression of this section: For a.e. $t \in (T, R)$,

$$\begin{aligned} & \left| \int_{B_T} \left(a * \mu_t\right)\left(\Phi_t\left(\tilde{x}\right)\right) \left|\partial_t \Phi_t\left(\tilde{x}\right)\right|^2 \, d\mu_T\left(\tilde{x}\right) - \int_{B_T} \left(a * \sigma_t\right)\left(\Phi_t\left(\tilde{y}\right)\right) \left|\partial_t \Phi_t\left(\tilde{y}\right)\right|^2 \, d\sigma_T\left(\tilde{y}\right) \right| \\ < & \iint_{B_T^2} \left[\frac{\varepsilon^2}{R^2} + \frac{\varepsilon^2}{R^2} \right] d\pi_T\left(\tilde{x}, \tilde{y}\right) = \frac{2\varepsilon^2}{R^2}. \end{aligned}$$

5.6.5. Final bound on the explosions.

Now, time integrating the square roots of these cost rates,

$$\begin{aligned} & \left| \int_T^R \sqrt{\int_{B_T} \left(a * \mu_t\right)\left(\Phi_t\left(\tilde{x}\right)\right) \left|\partial_t \Phi_t\left(\tilde{x}\right)\right|^2 \, d\mu_T\left(\tilde{x}\right)} dt - \int_T^R \sqrt{\int_{B_T} \left(a * \sigma_t\right)\left(\Phi_t\left(\tilde{y}\right)\right) \left|\partial_t \Phi_t\left(\tilde{y}\right)\right|^2 \, d\sigma_T\left(\tilde{y}\right)} dt \right| \\ \leq & \int_T^R \left| \sqrt{\int_{B_T} \left(a * \mu_t\right)\left(\Phi_t\left(\tilde{x}\right)\right) \left|\partial_t \Phi_t\left(\tilde{x}\right)\right|^2 \, d\mu_T\left(\tilde{x}\right)} - \sqrt{\int_{B_T} \left(a * \sigma_t\right)\left(\Phi_t\left(\tilde{y}\right)\right) \left|\partial_t \Phi_t\left(\tilde{y}\right)\right|^2 \, d\sigma_T\left(\tilde{y}\right)} \right| dt \\ \leq & \int_T^R \left[\frac{\varepsilon}{R} + \frac{R}{\varepsilon} \left| \int_{B_T} \left(a * \mu_t\right)\left(\Phi_t\left(\tilde{x}\right)\right) \left|\partial_t \Phi_t\left(\tilde{x}\right)\right|^2 \, d\mu_T\left(\tilde{x}\right) - \int_{B_T} \left(a * \sigma_t\right)\left(\Phi_t\left(\tilde{y}\right)\right) \left|\partial_t \Phi_t\left(\tilde{y}\right)\right|^2 \, d\sigma_T\left(\tilde{y}\right) \right| \right] dt \end{aligned}$$

following the form "$\left|\sqrt{b} - \sqrt{c}\right| \leq d + \frac{1}{d}\left|b - c\right|$". This then is less than

$$\int_T^R \left[\frac{\varepsilon}{R} + \frac{R}{\varepsilon} \frac{2\varepsilon^2}{R^2} \right] \, dt < 3\varepsilon.$$

Thus

$$d_{\mathrm{E}}\left(\mu_0, \mu_T\right) + d_{\mathrm{E}}\left(\sigma_T, \sigma_R\right) < 3\varepsilon + \int_0^R \sqrt{\int_{B_T} \left(a * \mu_t\right)\left(\Phi_t\left(\tilde{x}\right)\right) \left|\partial_t \Phi_t\left(\tilde{x}\right)\right|^2 \, d\mu_T\left(\tilde{x}\right)} \, dt.$$

For all $t \in (0, R)$,

$$\begin{aligned}
&\int_{B_T} (a * \mu_t)(\Phi_t(\tilde{x}))\, |\partial_t \Phi_t(\tilde{x})|^2 \; d\mu_T(\tilde{x}) \\
= &\int_{B_T} (a * \mu_t)(\Phi_t(\tilde{x})) \left| \frac{\tilde{x} - \tilde{x}_\square}{T} \right|^2 \; d\mu_T(\tilde{x}) \\
= &\int_{\mathbb{R}^d} \left(a * \sum_i m_i\, \rho_{i,t} \right)(x) \sum_j \left| \frac{x - x_j}{t} \right|^2 m_j\, \rho_{j,t}(x)\; dx \\
= &\sum_i m_i \sum_j m_j \int_{\mathbb{R}^d} (a * \rho_{i,t})(x) \left| \frac{x - x_j}{t} \right|^2 \rho_{j,t}(x)\; dx \\
= &\sum_i m_i \sum_j m_j \int_{\mathbb{R}^d} (a * \bar{\rho}_t)(x - x_i) \left| \frac{x - x_j}{t} \right|^2 \bar{\rho}_t(x - x_j)\; dx \\
= &\sum_i m_i \sum_j m_j \int_{\mathbb{R}^d} (a * \bar{\rho}_t)(y + x_j - x_i) \left| \frac{y}{t} \right|^2 \bar{\rho}_t(y)\; dy \\
= &\sum_i m_i \sum_j m_j \int_{\mathrm{B}(0,t)} (a * \bar{\rho}_t)(y + x_j - x_i) \left| \frac{y}{t} \right|^2 \bar{\rho}_t(y)\; dy \\
\leq &\sum_i m_i \sum_j m_j \int_{\mathrm{B}(0,t)} \Xi\, (a * \bar{\rho}_t)(y) \left| \frac{y}{t} \right|^2 \bar{\rho}_t(y)\; dy
\end{aligned}$$

by inequality (5.6.1) applied to $a * \bar{\rho}_t$. Because $\sum_i m_i = 1$, this equals

$$\Xi \int_{\mathrm{B}(0,t)} (a * \bar{\rho}_t)(y) \left| \frac{y}{t} \right|^2 \bar{\rho}_t(y)\; dy$$

which is Ξ times the cost rate for the unit explosion.

Thus

$$d_{\mathrm{E}}(\mu_0, \mu_T) + d_{\mathrm{E}}(\sigma_T, \sigma_R) < 3\varepsilon + \sqrt{\Xi}\, \mathcal{C}.$$

5.6.6. The translation phase.

Finally, we bound $d_{\mathrm{E}}(\mu_R, \mu_*)$.

Let curve $(\mu_t)_{t \in [R, R+1]}$ in $\mathrm{P}_c(\mathbb{R}^d)$ be obtained by translating each density component $m_i\, \rho_{i,R}$ to $m_i\, \bar{\rho}_R$ simultaneously during one unit of time. That is, for all $t \in [R, R+1]$,

$$d\mu_t = \sum_i m_i\, \sigma_{i,t}\, d\mathcal{L}^d$$

where $\sigma_{i,t}(x) := \bar{\rho}_R(x - (R + 1 - t)\, x_i)$ with $\bar{\rho}_t$ still defined as above, i.e. the Lebesgue density function of the unit explosion at time t. Note $\sigma_{i,t}$ is unrelated to σ_t despite the similar label.

Observe that μ_R so defined matches the μ_R already defined above, because $\sigma_{i,R} = \rho_{i,R}$.

Moreover,

$$\begin{aligned} d\mu_{R+1} &= \sum_i m_i \, \sigma_{i,R+1} \, d\mathcal{L}^d \\ &= \sum_i m_i \, \bar{\rho}_R(x) \; d\mathcal{L}^d \\ &= \bar{\rho}_R(x) \; d\mathcal{L}^d \\ &= d\mu_*. \end{aligned}$$

The situation is now analogous to that for the simultaneous explosion above. The curve $(\mu_t)_{t\in[R,R+1]}$ at some times experiences overlap of its density components $m_i \, \sigma_{i,t}$, with differing velocity fields for those components. The identical approximation procedure is again applied as before, approximating the continuum translations curve $(\mu_t)_{t\in[R,R+1]}$ by a massed-particle translations curve which we do not here label. Note in this case there is no need for an initial continuum evolution before the massed-particle approximation begins.

Omitting the repeated details, the distance between μ_R and its massed-particle approximation is again bounded by ε, as is the distance between the translated massed particles and μ_{R+1}. The difference of the cost expressions along the way is again bounded by 3ε.

The cost rate expression for $(\mu_t)_{t\in[R,R+1]}$ is then bounded as follows. Noting that for each i, $\sigma_{i,t}$ through time is a translation with constant velocity $-x_i$, we write the cost rate expression directly as

$$\begin{aligned} & \int_{\mathbb{R}^d} \left(a * \sum_i m_i \, \sigma_{i,t} \right)(x) \sum_j |x_j|^2 \, m_j \, \sigma_{j,t}(x) \; dx \\ = \; & \sum_i m_i \sum_j m_j \int_{\mathbb{R}^d} (a * \bar{\rho}_R)(x - (R+1-t)\, x_i) \, |x_j|^2 \, \bar{\rho}_R(x - (R+1-t)\, x_j) \; dx \\ = \; & \sum_i m_i \sum_j m_j \int_{\mathrm{B}(0,R)} (a * \bar{\rho}_R)(y + (R+1-t)(x_j - x_i)) \, |x_j|^2 \, \bar{\rho}_R(y) \; dy \\ < \; & \sum_i m_i \sum_j m_j \int_{\mathrm{B}(0,R)} \Xi \, (a * \bar{\rho}_R)(y) \, D^2 \, \bar{\rho}_R(y) \; dy \\ = \; & \Xi \int_{\mathrm{B}(0,R)} (a * \bar{\rho}_R)(y) \, D^2 \, \bar{\rho}_R(y) \; dy \\ < \; & \Xi \, \varepsilon^2 \, \Xi^{-1} = \varepsilon^2 \end{aligned}$$

by the definition of R above.

Thus

$$\int_R^{R+1} \sqrt{\int_{\mathbb{R}^d} \left(a * \sum_i m_i \, \sigma_{i,t} \right)(x) \sum_j |x_j|^2 \, m_j \, \sigma_{j,t}(x) \, dx} \, dt < \varepsilon$$

and

$$d_{\mathrm{E}}(\mu_R, \mu_*) < 2\varepsilon + 3\varepsilon + \varepsilon.$$

This concludes the final bound on the distances between the intermediate measures. We have shown $d_{\mathrm{E}}(\mu, \mu_*) < \sqrt{\Xi \mathcal{C}} + 12\varepsilon$. Since $d_{\mathrm{E}}(\nu, \mu_*)$ may be bounded likewise, this completes the proof. □

CHAPTER 6

The measure metric obtained by extension, and its topology

In this chapter we would like to obtain a version of the measure metric of this thesis in a different way, by first defining it for a moderate class of measures where we can understand it well and its topology, and then by extending this via metric completion to work on more general measures. This allows the topological understanding of the simple metric to be inherited by the more general metric. The more general measures we limit also, for simplicity, to those taking support inside a large ball in $\mathbb{R}^d$.

6.1. Conventions

We identify here a few key conventions used in this chapter.

In this thesis and particularly used this chapter, the symbol "$\subset$" means subset, not necessarily strict.

Subscripts on sets, when not used for indices, are used to denote set neighborhood in $\mathbb{R}^d$, as in

$$A_\varepsilon := \left\{ x \in \mathbb{R}^d \;\middle|\; \exists y \in A \text{ with } |y - x| < \varepsilon \right\}.$$

The terminology of *topologically stronger* and *uniformly stronger* metrics is followed as in [**70**]. *Topologically stronger* simply means "has a finer topology", i.e. the identity map from the stronger metric to the weaker metric is continuous. *Uniformly stronger* means more: such identity map is uniformly continuous.

6.2. Development

Fix a closed ball $K \subset \mathbb{R}^d$ and let $\mathrm{P}(K)$ denote the set of Borel probability measures with support in K.

Let d_{W_2} denote the 2-Wasserstein metric on $\mathrm{P}(K)$ from optimal transport theory,

$$d_{\mathrm{W}_2}(\mu, \nu) := \inf \left\{ \sqrt{\int_{K \times K} |x - y|^2 \, d\pi(x, y)} \;\middle|\; \pi \in \Pi(\mu, \nu) \right\}$$

where $\Pi(\mu,\nu)$ is the set of couplings of μ and ν.

Let d_{LP} denote the Lévy-Prokhorov metric on $\mathrm{P}(K)$,

$$d_{\mathrm{LP}}(\mu,\nu) := \inf\{\varepsilon > 0 \ \mid \forall \text{Borel } A \subset K,\ \mu(A) \le \nu(A_\varepsilon) + \varepsilon \text{ and } \nu(A) \le \mu(A_\varepsilon) + \varepsilon\}.$$

For $r > 0$ define a modification of the Lévy-Prokhorov metric (which will be shown to be a metric) on $\mathrm{P}(K)$ as

$$d_{\mathrm{LP}_r}(\mu,\nu) := \inf\{\varepsilon > 0 \,|\, \forall \text{Borel } A \subset K \text{ with } \operatorname{diam}(A) \le r,\ \mu(A) \le \nu(A_\varepsilon) + \varepsilon \text{ and } \nu(A) \le \mu(A_\varepsilon) + \varepsilon\}.$$

Of course for all r, $d_{\mathrm{LP}_r} \le d_{\mathrm{LP}}$, and for $r \ge \operatorname{diam}(K)$, $d_{\mathrm{LP}_r} = d_{\mathrm{LP}}$.

First we make a needed comparison of the above.

Lemma. For all r, d_{LP_r} is topologically stronger than d_{W_2} on $\mathrm{P}(K)$.

Proof.

Let $r > 0$, $\mu \in \mathrm{P}(K)$, and $\varepsilon \in (0, r)$.

Let disjoint Borel cover $(A_i)_{i=1}^m$ of K such that $\forall i$ $\operatorname{diam}(A_i) \le \varepsilon$ and $\mu(\partial A_i) = 0$ where ∂ denotes boundary.

(For example, all boundaries may be established by a collection of hyperplanes, each of which may translate to all but a countable number of shifts while avoiding positive measure, its shifts partitioning $\mathbb{R}^d$.)

Let $\overline{\delta} = \frac{\varepsilon}{m} > 0$ and let $\delta \in \left(0, \overline{\delta}\right)$ s.t. $\forall i$ $\mu((\partial A_i)_\delta) < \overline{\delta}$.

(monotone convergence of μ on intersections of sets of the form $(\partial A_i)_{\frac{1}{n}}$, uniform over finite i)

Let $\nu \in \mathrm{P}(K)$ s.t. $d_{\mathrm{LP}_r}(\mu,\nu) < \delta$. $\forall i$ $\nu(A_i) \le \mu((A_i)_\delta) + \delta \le \mu(A_i) + \mu((\partial A_i)_\delta) + \delta < \mu(A_i) + 2\overline{\delta}$ so $\left(\nu(A_i) - 2\overline{\delta}\right) \vee 0 \in [0,\, \mu(A_i) \wedge \nu(A_i)]$.

Let $\pi \in \Pi(\mu,\nu)$ such that for all i at least $\left(\nu(A_i) - 2\overline{\delta}\right) \vee 0$ mass is coupled within A_i, the remainder coupled arbitrarily.

For all i at most $2\overline{\delta}$ of ν's mass in A_i is coupled outside A_i. Thus at most $2\overline{\delta}m$ of ν's total mass is coupled at distance greater than ε.

$$d_{\mathrm{W}_2}(\mu,\nu) \leq \sqrt{\int_{K\times K} |x-y|^2 \, d\pi(x,y)} \leq \sqrt{2\overline{\delta}m \cdot \mathrm{diam}(K)^2 + 1\cdot\varepsilon^2} = \sqrt{2\mathrm{diam}(K)^2\varepsilon + \varepsilon^2}. \qquad \square$$

d_{LP_r} is a metric (shown nondegenerate by the lemma, i.e. distinct measures have nonzero distance), with $d_{\mathrm{LP}_r} \leq d_{\mathrm{LP}}$, so $(\mathrm{P}(K), d_{\mathrm{LP}_r})$ is compact since $(\mathrm{P}(K), d_{\mathrm{LP}})$ is. Thus d_{LP_r} is *uniformly* stronger than d_{W_2} on $\mathrm{P}(K)$ by the lemma that it is topologically stronger.

Next we define the measure metric, first in limited scope on "nice" measures.

Let $\widetilde{\mathrm{P}}(K)$ denote the following subset of $\mathrm{P}(K)$: each measure absolutely continuous with respect to Lebesgue measure with C^∞ density bounded above & below (by positive constant) on K.

Fix $a \in \mathcal{A}$, defined in section 4.2.2, and from it define $\widetilde{d}_{\mathrm{E}}$ (which will be shown to be a metric) on $\widetilde{\mathrm{P}}(K)$ as

$$\widetilde{d}_{\mathrm{E}}(\mu,\nu) := \inf\left\{ \int_0^1 \sqrt{\int_{\mathbb{R}^d} (a * \mu_t)(x) \, |v(t,x)|^2 \, d\mu_t(x)} dt \;\middle|\; \left((\mu_t)_{t\in[0,1]}, v\right) \in \widetilde{V}_{\mu,\nu} \right\}$$

where $\widetilde{V}_{\mu,\nu}$ is defined as the set of $\left((\mu_t)_{t\in[0,1]}, v\right)$ such that $v \in C^1\left([0,1]\times\mathbb{R}^d; \mathbb{R}^d\right)$, $\mu_1 = \nu$, and $\forall t \;\; \mu_t = \Phi(t,\cdot)_{\#}\mu$ where $\Phi(t,x) \in \mathbb{R}^d$ denotes the time $t \in [0,1]$ solution of the IVP

$$\dot{\xi}(s) = v(s, \xi(s)), \;\; \xi(0) = x \in \mathbb{R}^d$$

given by classical ODE theory.

We now compare this to the modified Lévy-Prokhorov metric on $\widetilde{\mathrm{P}}(K)$.

Lemma. For sufficiently small r, $\widetilde{d}_{\mathrm{E}}$ is uniformly stronger than d_{LP_r} on $\widetilde{\mathrm{P}}(K)$.

Proof.

Let $r, \beta > 0$ such that on $[0, 3r]$, $a^\bullet > \beta$ where $a^\bullet$ is the radial profile of a.

Let $\varepsilon \in (0, r)$, and suppose there exist $\mu, \nu \in \widetilde{\mathrm{P}}(K)$ s.t. $d_{\mathrm{LP}_r}(\mu,\nu) > \varepsilon$.

Let Borel $A \subset K$ such that $\mathrm{diam}(A) \leq r$ and $\mu(A) > \nu(A_\varepsilon) + \varepsilon$, swapping the labels of μ, ν if necessary.

Suppose there exists $\left((\mu_t)_{t\in[0,1]}, v\right) \in \widetilde{V}_{\mu,\nu}$, and let the corresponding map Φ be as defined above.

Let Φ^{-1} denote the inverse of Φ in its second argument given by the classical theory, i.e. $\forall t \in [0,1], x \in \mathbb{R}^d \quad \Phi^{-1}(t, \Phi(t,x)) = x$ and $\Phi\left(t, \Phi^{-1}(t,x)\right) = x$.

Let

$$\overline{T} = \min\left\{ t \in [0,1] \;\middle|\; \mu\left(A \cap \Phi^{-1}\left(t, {A_\varepsilon}^c\right)\right) \geq \frac{\varepsilon}{2} \right\} \in (0,1].$$

(The bracketed set is nonempty because $\mu\left(A \cap \Phi^{-1}\left(1, {A_\varepsilon}^c\right)\right) = \mu(A) - \mu\left(A \cap \Phi^{-1}(1, A_\varepsilon)\right) = \mu(A) - \nu(\Phi(1,A) \cap A_\varepsilon) \geq \mu(A) - \nu(A_\varepsilon) > \varepsilon$, and is closed by closedness of $\left[\frac{\varepsilon}{2}, 1\right]$ and smooth $\mu_t^A := \Phi(t,\cdot)_{\#}\mu_{|A}$.)

Let

$$B = A \cap \Phi^{-1}\left(\overline{T}, {A_\varepsilon}^c\right)$$

and for $x \in B$ let

$$T(x) := \min\left\{ t \in [0,1] \;\middle|\; \Phi(t,x) \in {A_\varepsilon}^c \right\} \in \left(0, \overline{T}\right].$$

(The bracketed set is nonempty by $\overline{T}$ membership and closed by continuity of Φ.)

$$\begin{aligned}
& \int_0^1 \sqrt{\int_{\mathbb{R}^d} (a * \mu_t)(x) \left|v(t,x)\right|^2 \, d\mu_t(x)}\, dt \\
\geq \; & \int_0^1 \int_{\mathbb{R}^d} \sqrt{(a * \mu_t)(x)}\, \left|v(t,x)\right| \, d\mu_t(x)\, dt
\end{aligned}$$

(Jensen)

$$\begin{aligned}
= \; & \int_0^1 \int_{\mathbb{R}^d} \sqrt{\int_{\mathbb{R}^d} a(y - \Phi(t,x))\, d\mu_t(y)}\, \left|v(t,\Phi(t,x))\right| \, d\mu(x)\, dt \\
= \; & \int_{\mathbb{R}^d} \int_0^1 \sqrt{\int_{\mathbb{R}^d} a(y - \Phi(t,x))\, d\mu_t(y)}\, \left|v(t,\Phi(t,x))\right| \, dt\, d\mu(x)
\end{aligned}$$

(Tonelli)

$$\begin{aligned}
\geq \; & \int_B \int_0^{T(x)} \sqrt{\int_{A_\varepsilon} a(y - \Phi(t,x))\, d\mu_t(y)}\, \left|v(t,\Phi(t,x))\right| \, dt\, d\mu(x) \\
> \; & \int_B \int_0^{T(x)} \sqrt{\beta \mu_t(A_\varepsilon)}\, \left|v(t,\Phi(t,x))\right| \, dt\, d\mu(x)
\end{aligned}$$

(a difference of members of A_ε having been given to a, with $\operatorname{diam}(A_\varepsilon) < 3r$)

$$> \sqrt{\beta \frac{\varepsilon}{2}} \int_B \int_0^{T(x)} \left|v(t,\Phi(t,x))\right| \, dt\, d\mu(x)$$

(noting $\mu_t(A_\varepsilon) \geq \mu_t(\Phi(t,A) \cap A_\varepsilon) = \mu\left(A \cap \Phi^{-1}(t, A_\varepsilon)\right) = \mu(A) - \mu\left(A \cap \Phi^{-1}(t, {A_\varepsilon}^c)\right) > \varepsilon - \frac{\varepsilon}{2}$ for all $t < T(x) \leq \overline{T}$)

$$\geq \sqrt{\beta \frac{\varepsilon}{2}} \int_B \varepsilon \, d\mu(x)$$

(arclength of path $\Phi(\cdot, x)$ bounded below by the distance between its endpoints, the first of which lies in $B \subset A$ and the second in ${A_\varepsilon}^c$)

$$\geq \sqrt{\beta \frac{1}{2}} \frac{1}{2} \varepsilon^{2.5}.$$

Thus $\widetilde{d}_{\mathrm{E}}(\mu, \nu)$ is greater than or equal to the previous line. □

The following elementary observation of metric completion is deduced since not readily found in a reference.

Lemma. Uniformly equivalent metrics have uniformly equivalent completions, and in particular, topologically equivalent completions.

Proof.

Let uniformly equivalent metrics d_1, d_2 on a set X and let continuous moduli of continuity σ_1, σ_2 for the identity maps $(X, d_1) \to (X, d_2)$ and $(X, d_2) \to (X, d_1)$. From $d_2 \leq \sigma_1 \circ d_1$ and $d_1 \leq \sigma_2 \circ d_2$ we see the d_1-Cauchy sequences are precisely the d_2-Cauchy sequences, and for two such Cauchy sequences $(x_n), (y_n)$,

$$d_2^*((x_n), (y_n)) = \lim_n d_2(x_n, y_n) \leq \lim_n \sigma_1(d_1(x_n, y_n)) = \sigma_1\left(\lim_n d_1(x_n, y_n)\right) = \sigma_1(d_1^*((x_n), (y_n)))$$

where star denotes the metric completed, and likewise $d_1^* \leq \sigma_2 \circ d_2^*$. □

6.3. Completion of the metric

Theorem. Let $\widetilde{\mathrm{P}}(K)$ and $\widetilde{d}_{\mathrm{E}}$ be defined as above for $a \in \mathcal{A}$. Then the completion of $\left(\widetilde{\mathrm{P}}(K), \widetilde{d}_{\mathrm{E}}\right)$ is the space $\mathrm{P}(K)$ with metric that is uniformly equivalent to $(\mathrm{P}(K), d_{\mathrm{W}_2})$, and so metrizes the topology of weak (-*) convergence of measures.

Proof.

Claim: for $\mu, \nu \in \widetilde{\mathrm{P}}(K)$,

$$d_{\mathrm{W}_2}(\mu,\nu) = \inf\left\{ \int_0^1 \sqrt{\int_{\mathbb{R}^d} |v(t,x)|^2 \ d\mu_t(x)}dt \quad \middle| \quad \left((\mu_t)_{t\in[0,1]}, v\right) \in \widetilde{V}_{\mu,\nu} \right\}.$$

That is, the optimal transport formulation of the 2-Wasserstein metric matches its displacement interpolation form on $\widetilde{\mathrm{P}}(K)$ for $\widetilde{V}_{\mu,\nu}$ as defined above.

To see this, note for the 2-Wasserstein metric there exists an optimal coupling of μ,ν given by a map from K to K that is a C^∞ diffeomorphism due to the smoothness of μ and ν, as in De Philippis and Figalli's Thm 3.3 [**35**] and earlier developed by Caffarelli [**18, 19**]. This results in linear displacement-interpolation Lagrangian paths between μ,ν that are smooth and invertible in space, as noted in Santambrogio's Lemma 5.29 [**68**]. The resulting induced Eulerian velocity field v, as formed in Santambrogio's Proposition 5.30 [**68**], inherits this smoothness and achieves the above infimum as a minimum, with associated curve $(\mu_t)_{t\in[0,1]}$ defined by the push-forward of μ along those Lagrangian paths.

Thus

$$\begin{aligned} d_{\mathrm{W}_2}(\mu,\nu) = \inf\left\{ \int_0^1 \sqrt{\int_{\mathbb{R}^d} (\mathbf{1} * \mu_t)(x)\, |v(t,x)|^2 \ d\mu_t(x)}dt \quad \middle| \quad \left((\mu_t)_{t\in[0,1]}, v\right) \in \widetilde{V}_{\mu,\nu} \right\} \\ \geq \widetilde{d}_{\mathrm{E}}(\mu,\nu) \end{aligned}$$

because $a \leq 1$.

Thus $\widetilde{d}_{\mathrm{E}}$ is a metric on $\widetilde{\mathrm{P}}(K)$ under the metric axioms as follows. $\widetilde{d}_{\mathrm{E}} < \infty$ by the previous inequality. The distance between any measure and itself is evidently zero from the definition of $\widetilde{d}_{\mathrm{E}}$ using $v \equiv 0$. $\widetilde{d}_{\mathrm{E}}$ is nondegenerate, i.e. distinct measures have nonzero distance, by the lemma that it is uniformly stronger than d_{LP_r}. Metric symmetry and the triangle inequality are evident from the definition of $\widetilde{d}_{\mathrm{E}}$ as a displacement interpolation; the former by the invertibility of the ODE, and the latter by any intermediate measure serving as a constraint on the original curve, with the two legs of the curve optimized independently and using time reparametrization.

By the first two lemmas and the fact that $\widetilde{d}_{\mathrm{E}} \leq d_{\mathrm{W}_2}$ on $\widetilde{\mathrm{P}}(K)$, on $\widetilde{\mathrm{P}}(K)$ we have that d_{W_2} is uniformly stronger than $\widetilde{d}_{\mathrm{E}}$ which is uniformly stronger than d_{LP_r} for some r, which is uniformly stronger than d_{W_2}. By the completion lemma, $\left(\widetilde{\mathrm{P}}(K), \widetilde{d}_{\mathrm{E}}\right)$ and $\left(\widetilde{\mathrm{P}}(K), d_{\mathrm{W}_2}\right)$ have uniformly equivalent completions, and therefore topologically equivalent completions. The resulting topology is therefore the topology of d_{W_2} on $\mathrm{P}(K)$, by denseness of $\widetilde{\mathrm{P}}(K)$ in $\mathrm{P}(K)$ with respect to d_{W_2}. □

CHAPTER 7

Open directions for further research

To close out the discussion before the numerical results of chapter 8, we collect here a short list of research directions open for the taking, as motivated by the developments in this thesis.

• Development of a well-posedness theory for weighted aggregation (2.5.1). For example, weak solutions set in Sobolev spaces as developed in [**12**] for pure aggregation.

• An understanding of the effects that attention functions have on the H-stability of weighted aggregation (2.5.1) . (See H-stability in for example [**30**].)

• Investigation of the curvature of the formal Riemannian manifold $\left(\mathrm{P}_{\mathrm{E}}\left(\mathbb{R}^d\right), d_{\mathrm{E}}\right)$.

• A study of displacement convexity for internal, potential, and interaction energies along the geodesics of d_{E}. (Note when semi-convex, the gradient flow theory of [**3**] may provide well-posedness.)

• On $\mathrm{P}(K)$ does the completion of metric $\widetilde{d}_{\mathrm{E}}$ in chapter 6 match the metric d_{E} defined in chapter 4?

• Metric space completion: When a is explodable, the unit explosion yields a Cauchy sequence in $\left(\mathrm{P}_{\mathrm{E}}\left(\mathbb{R}^d\right), d_{\mathrm{E}}\right)$ lacking a limit. Plausibly this space's completion may be characterized in a natural way by accounting for portions of mass that have spread out to have negligible concentration everywhere, leaving sub-probability measures. A candidate space might be

$$\mathrm{P}'_{\mathrm{E}}\left(\mathbb{R}^d\right) := \left\{ m\mu \;\middle|\; m \in [0,1], \mu \in \mathrm{P}_{\mathrm{E}}\left(\mathbb{R}^d\right) \right\}$$

together with a metric like

$$D_E\left(m\mu, \tilde{m}\nu\right) := \lim_{n\to\infty} d_{\mathrm{E}}\left(\mu_n, \nu_n\right)$$

where μ_n and ν_n are Cauchy sequences in $\left(\mathrm{P}_{\mathrm{E}}\left(\mathbb{R}^d\right), d_{\mathrm{E}}\right)$ weakly converging in measure to $m\mu$ and $\tilde{m}\nu$, respectively.

• Metric tensor localization: Consider a sequence of attention functions that converge as mollifiers to the Dirac delta distribution. These belong to $\mathcal{A}$ except for the normalization $\sup a\left(\mathbb{R}^d\right) = 1$ which must be dropped. The limit of the corresponding metric tensors is

$$g_\rho(v,v) := \int |v|^2 \rho^2 \, d\mathcal{L}^d$$

which is a local (no longer nonlocal) metric tensor with simple yet interesting metric, indeed admitting characterization as the homogeneous H^{-1} distance (and also corresponding to constant "mobility function" in [**27**].)

CHAPTER 8

Simulation and visualization

In this final chapter, the graphical results of various numerical experiments are provided to evidence and illustrate the theory of this thesis. Additionally, a concept application "hierarchical clustering" is tested out, and additionally after that, an interesting behavior is discovered for aggregation which we call "polar milling".

Experiments are shown for the geodesics of the particle metric that was defined in section 4.2.2, as well as for the weighted aggregation ODE (2.3.1). Both are kept to 2D simulations for purposes of displaying the resulting graphics in this 2D thesis, although arbitrary dimension may be run numerically.

All numerics and plotting were performed in MATLAB licensed by Carnegie Mellon University, and coded by this author, using mostly built-in commands and solvers as named below.

8.1. Geodesics visualization

For geodesics, the derived Euler-Lagrange equation (4.4.1) of section 4.4 is numerically solved as a boundary value problem.

MATLAB's boundary value solver "bvp5c" was used, which is described as a finite difference code that implements the four-stage Lobatto IIIa formula. Relative tolerance and absolute tolerance of 10^{-3} and 10^{-6} respectively were used, with number of mesh points allowed to 10^9.

Attention a for this section is chosen to have exponential decay with $a(0) = 1$, pictured in figure 8.1.1. Self attention is set at $a_0 = a(0)$.

For the first experiment, see figure 8.1.2. The convention of ordering subfigures in this chapter is: From the upper-left subfigure, subfigures are chronological left-to-right, and then the next row down left-to-right, etc.

A source configuration of five particles has been chosen, shown in the upper-left subfigure; and a target configuration has also been specified, which is that same formation *translated* from its start, shown in the lower-right subfigure. The sequence of subfigures shows snapshots of the resulting geodesic curve in the configuration space. Observe how the pentagon-like shape spreads out as it

begins to translate in formation, then does most of its translating in that spread-out formation, and finally collapses back to the tight pentagon-like shape.

Figure 8.1.3 shows the paths that were taken by each of the five particles.

Next see figure 8.1.4 for the second experiment. Here two particles are required to swap positions. Observe that they do so while keeping some separation rather than passing through each other. Figure 8.1.5 shows their paths. Naturally, by symmetry, another geodesic may be obtained by reflection through the line joining the two points.

Finally for this section, a more "arbitrary" geodesic is shown. See figure 8.1.6. A loose pile of particles is the source configuration, and two of them are required to end at the right side-by-side, whereas the other three are required to end at the left in a triangular formation, as shown in the final subfigure. Figure 8.1.7 of their paths reveals the characteristic spreading, in this case spreading within two separate "components" of the source configuration which must travel in generally the same direction.

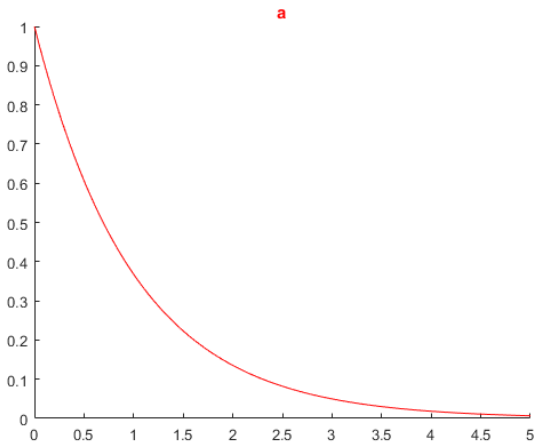

FIGURE 8.1.1. Attention profile $a^\bullet$

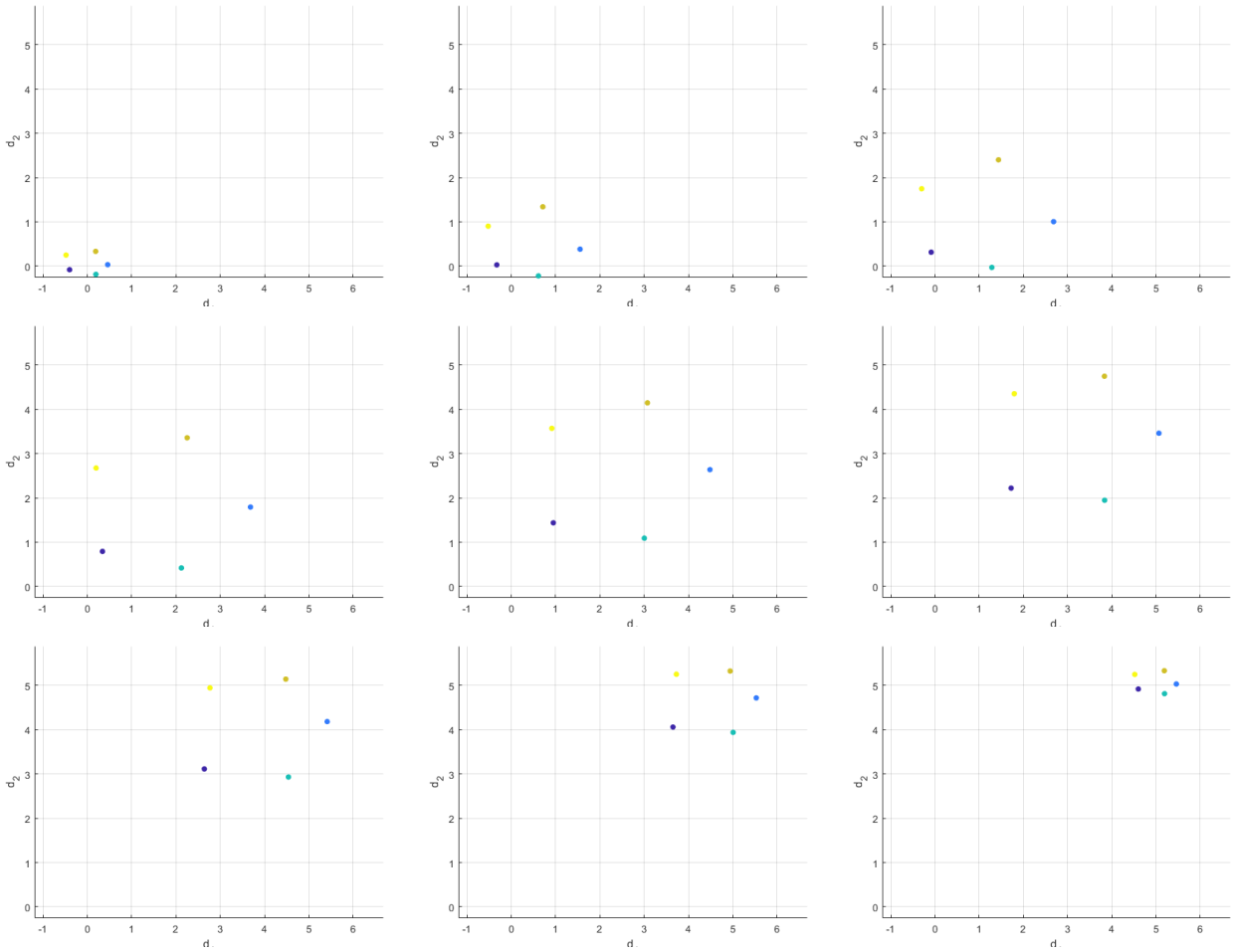

FIGURE 8.1.2. When the target configuration is merely a translation of the source configuration.

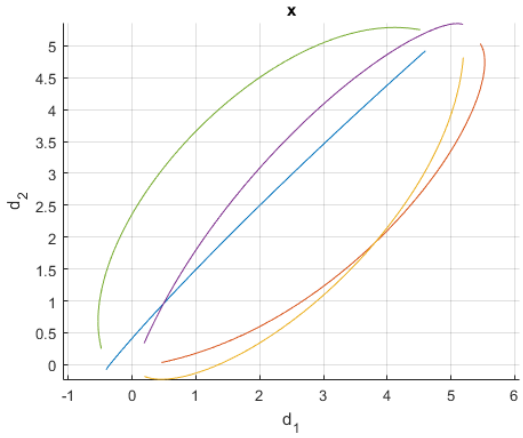

FIGURE 8.1.3. The same curve through the configuration space, drawn with tracer lines.

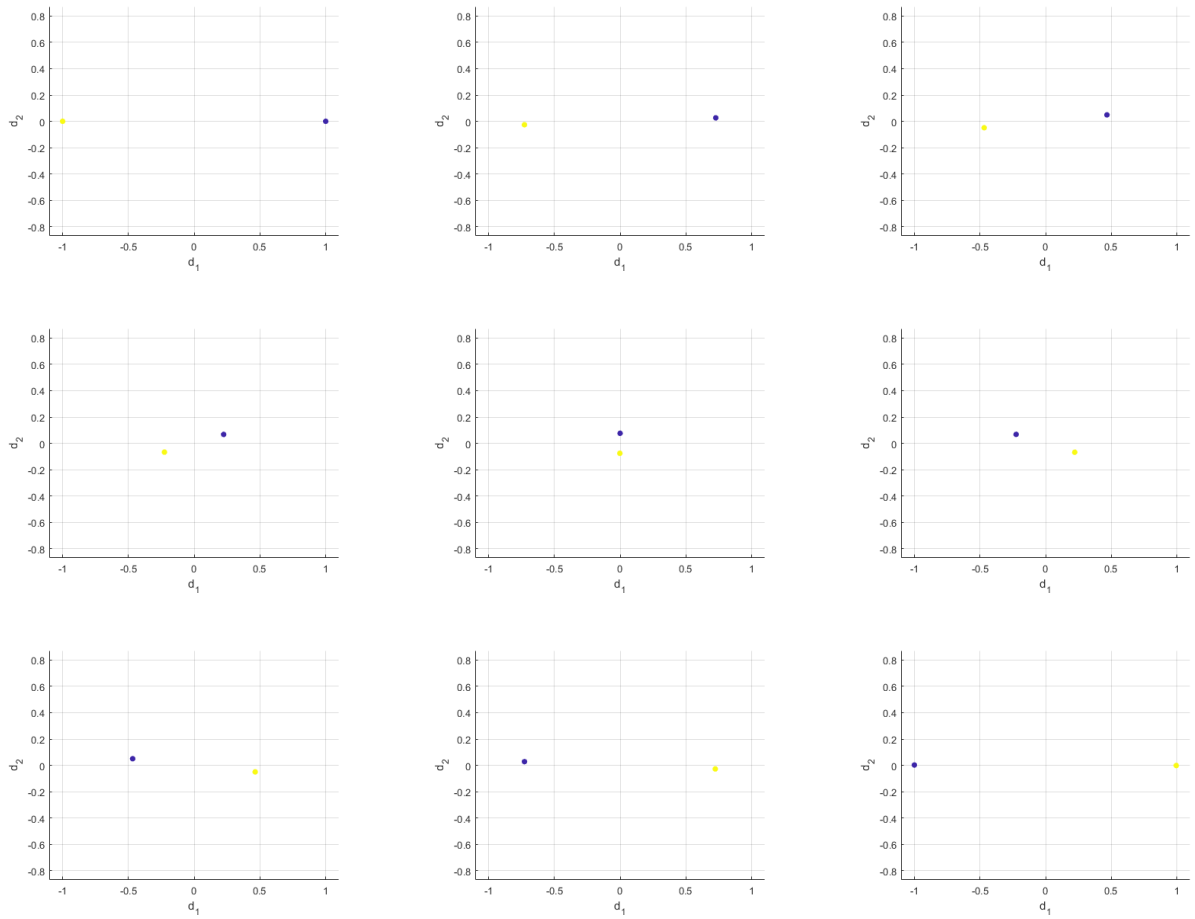

FIGURE 8.1.4. Two particles must swap places.

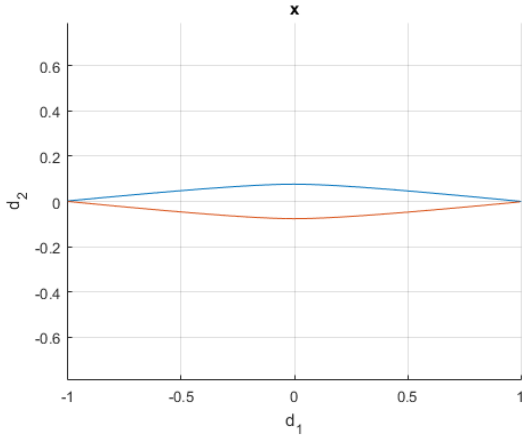

FIGURE 8.1.5. The same curve through the configuration space, drawn with tracer lines.

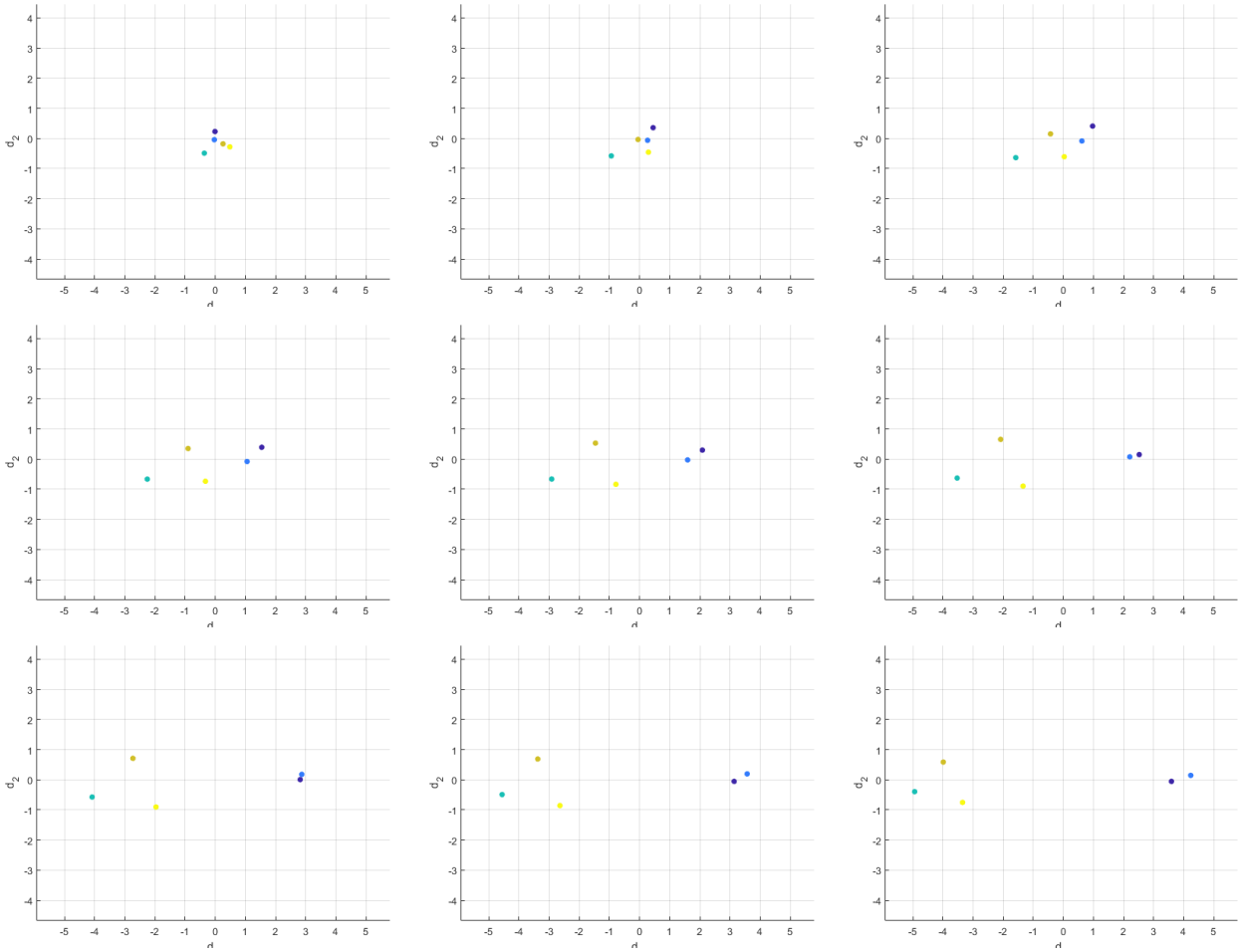

FIGURE 8.1.6. More arbitrary source and target configurations. Spreading still occurs between components headed in the same direction.

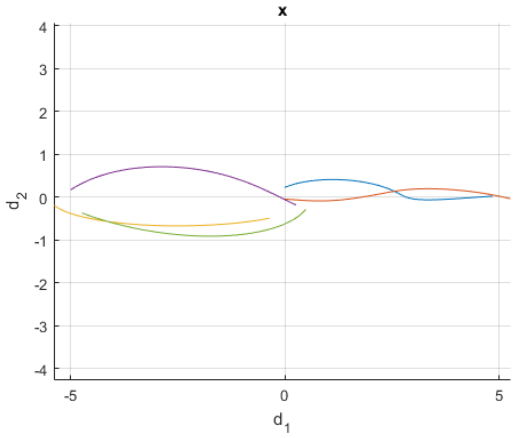

FIGURE 8.1.7. The same curve through the configuration space, drawn with tracer lines.

8.2. Weighted aggregation visualization

For weighted aggregation, the ODE (2.3.1) of section 2.3 is numerically solved as an initial value problem.

MATLAB's initial value solver "ode113" was used, which is described as a variable-step, variable-order (VSVO) Adams-Bashforth-Moulton PECE solver of orders 1 to 13. Relative tolerance and absolute tolerance of 10^{-5} and 10^{-6} respectively were used.

Attention a for this section is chosen to have bounded support, with a 3rd degree polynomial sloping downward from $a^\bullet(0) = 1$ through $a^\bullet(2) = \frac{1}{2}$ to $a^\bullet(4) = 0$. W for this section is chosen as a Morse potential. Both are pictured in figure 8.2.1. Self attention is set at $a_0 = a(0)$.

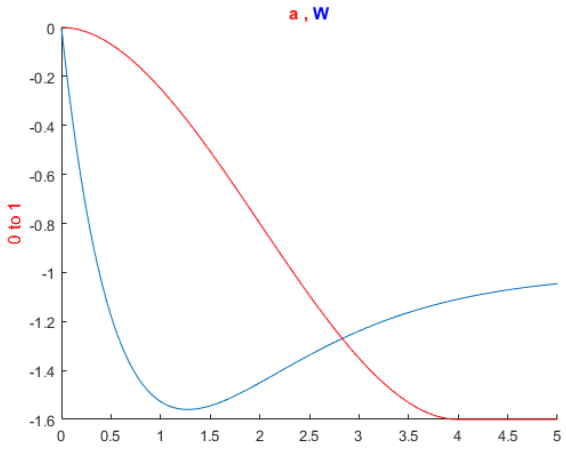

FIGURE 8.2.1. Attention profile $a^\bullet$ (red) and interaction-potential profile $W^\bullet$ (blue). The vertical axis is used for $W^\bullet$, whereas $a^\bullet$ varies from 0 to 1.

First see figure 8.2.2. The initial configuration is on the average a bit too tightly packed for comfort, according to the preferred spacing established by W. This causes the resulting aggregation to be a generally spreading-out evolution, settling into a comfortable lattice-like ball by the time of the final subfigure. Observe that this lattice spacing roughly matches the pairwise comfort separation established by W, which is facilitated by the attention function which does not allow much attention beyond roughly this separation distance, as seen in figure 8.2.1.

Next, in figure 8.2.3 an initial configuration is on the average *too* spread out for comfort, according to the preferred spacing established by W. This causes the resulting aggregation to be a generally collapsing evolution, which does not yet reach a comfortable configuration by the time of the final subfigure. Observe the temporary "filament-like" formations, as particles *first* attempt to reach comfort spacing with primarily their nearest neighbors, before seeking more distant aggregations,

thanks again to the attention function which does not allow much attention beyond roughly the pairwise comfort separation.

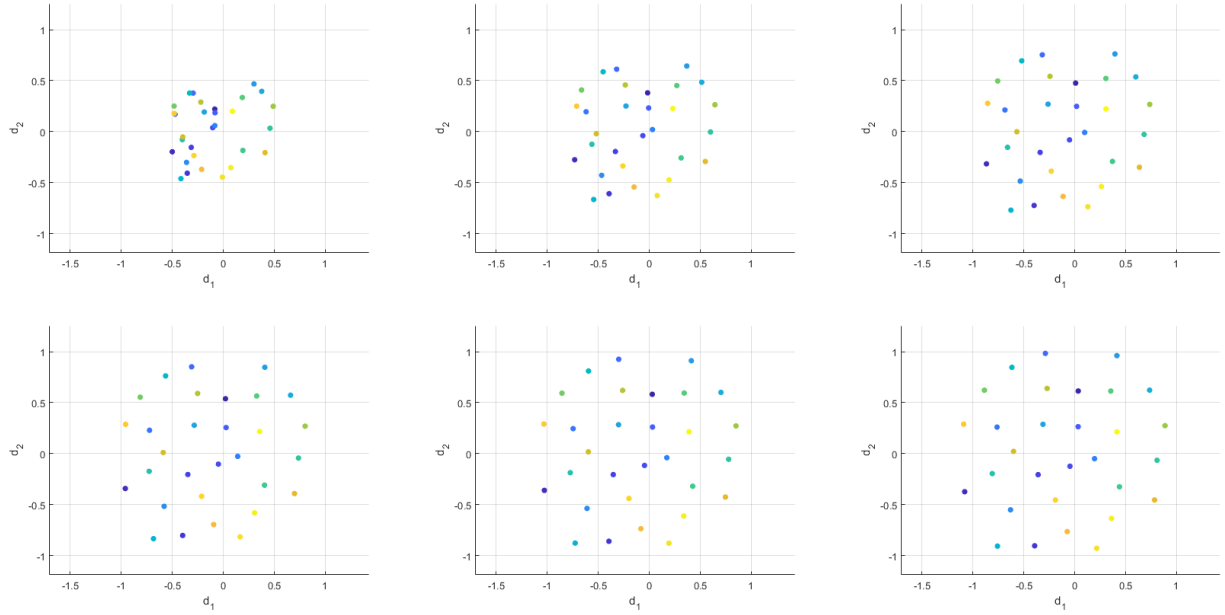

FIGURE 8.2.2. Evolution from an initial configuration having average group spacing smaller than that preferred by W.

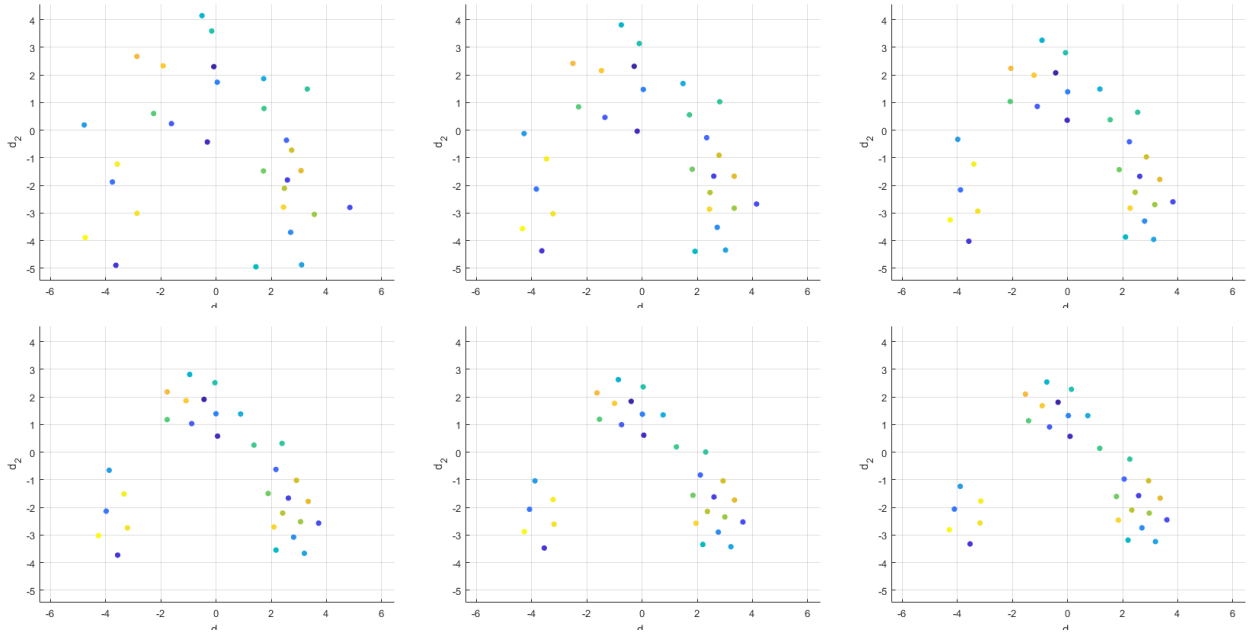

FIGURE 8.2.3. Evolution from an initial configuration having average group spacing larger than that preferred by W.

The next two figures, 8.2.4 and 8.2.5, compare pure aggregation (in the former) to weighted aggregation (in the latter). Both begin with the same initial configuration: Three remote particles separated a fair distance from a much larger crowd.

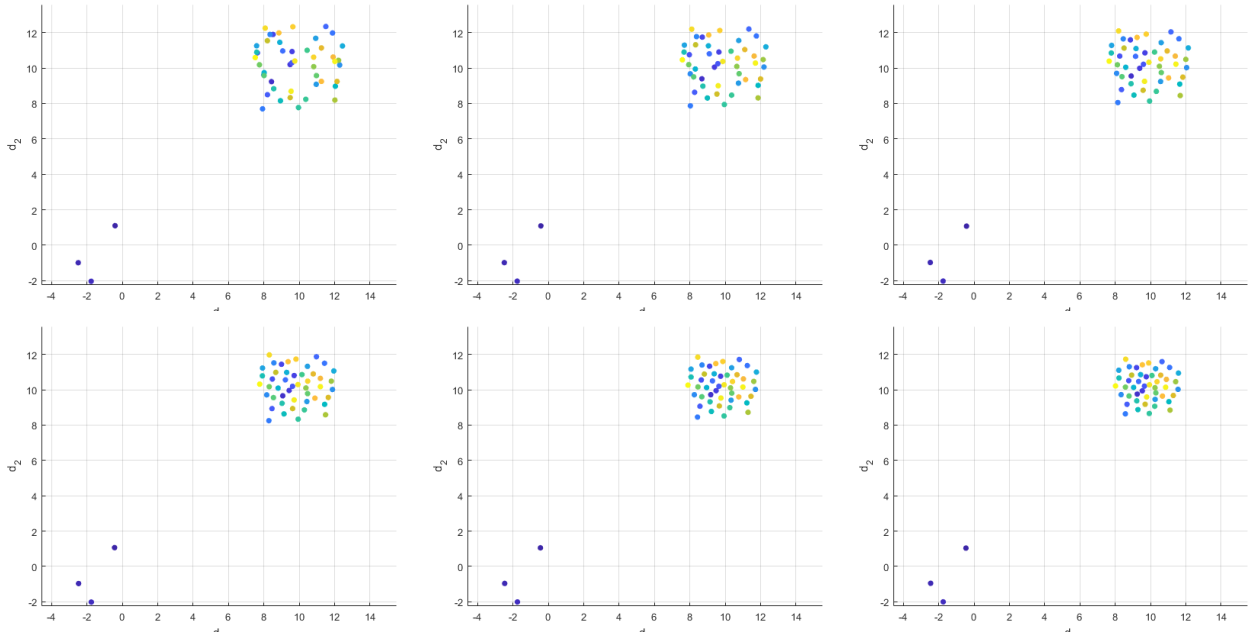

FIGURE 8.2.4. A weakness of pure aggregation evidenced: a distant large group diluting the interaction of a small group.

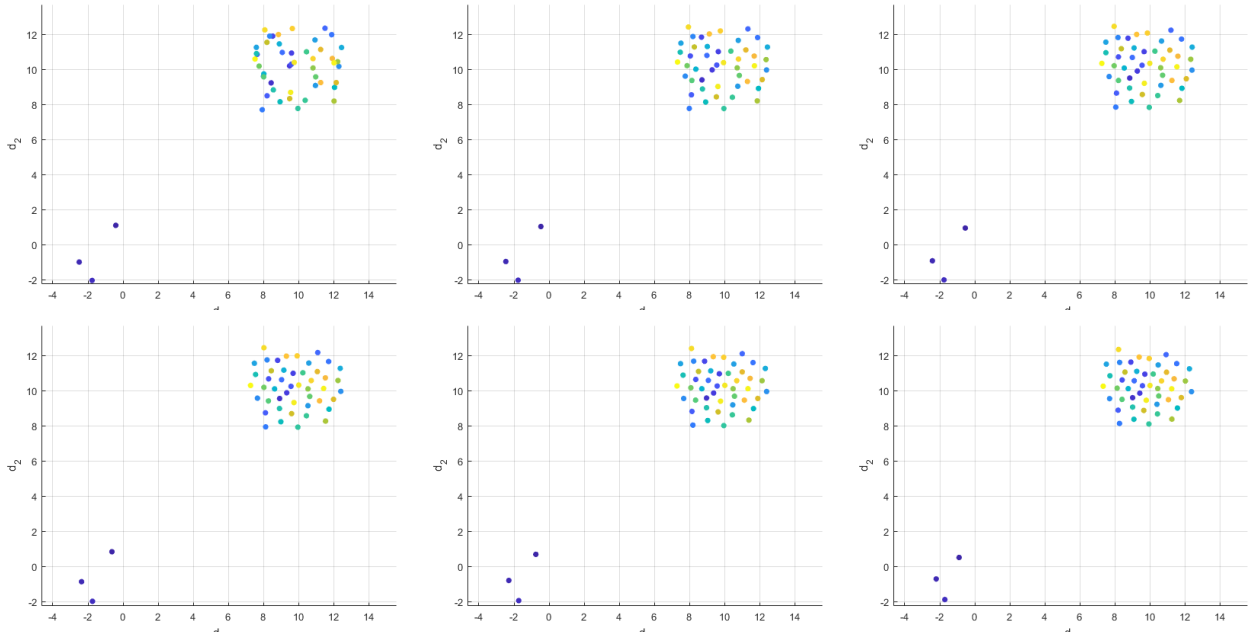

FIGURE 8.2.5. The introduction of weighting improves the aforementioned weakness.

Observe two effects. First, the aggregation that occurs within the large crowd, which does not have much to do with the three remote particles, packs much tighter in the case of pure aggregation than in this weighted aggregation, as seen by comparing the final subfigure from each figure. This evidences the effect of attention on maintaining a lattice-like spacing that is more similar to the pairwise comfort distance established by W. Theoretically this is expected because without attention limited to mostly one's next neighbors, particles experience that most of the lattice is too far from them.

Second, observe the slower response of the three remote particles in the case of pure aggregation. In both cases they begin to approach a comfort spacing with each other, though this evolution is "unrealistically" slowed in the case of pure aggregation. The final subfigure helps to see the different amounts of progress, though it is slightly subtle. Figure 8.2.6 helps to see this by indicating the particle velocities at the start of evolution. Of course, this experiment setup does not involve a very large distant cluster, at very large distance, so the "unrealistic" effects can be worse in more dramatic situations.

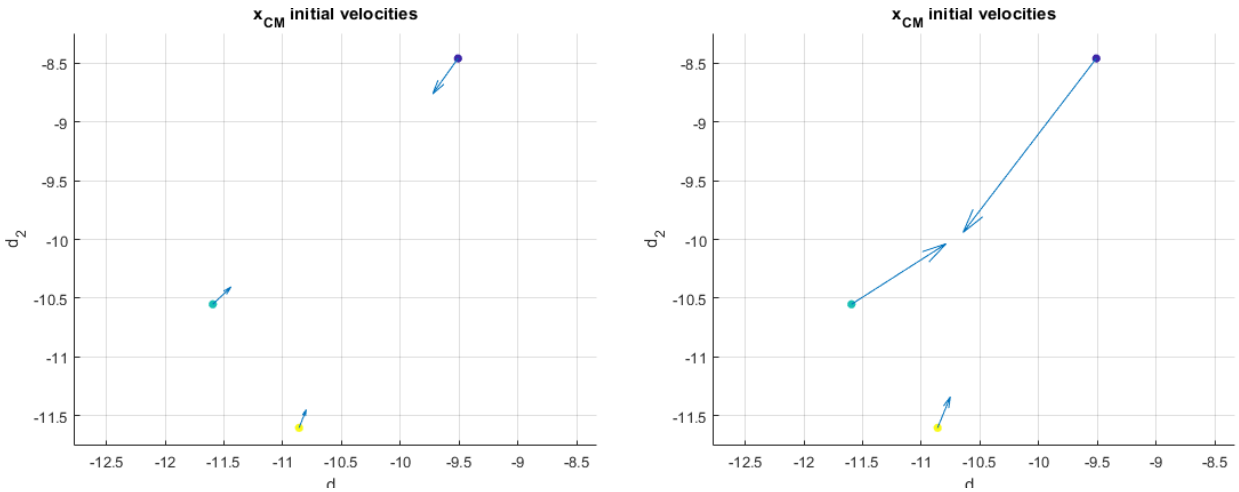

FIGURE 8.2.6. Close-up of the smaller group with initial velocities shown. On the left, pure aggregation, and on the right, weighted aggregation.

8.3. Application: hierarchical clustering

This section again simulates weighted aggregation as was done in the previous section. The section should not be taken too seriously, but instead is just a first trial of an interesting idea.

The idea derives from the modern computer science discipline of machine learning. One task there which needs solution methods is the problem of "clustering" for automated classification. In its simplest form, given data points in $\mathbb{R}^d$ one wishes to group them as a means of classifying which belong together in association.

Thus a toy application (for now) for aggregation, and especially weighted aggregation, might be the conducting of this information-theoretic clustering via *actual* clustering in the dynamical sense of evolution equations. An attractive-only interaction potential W would be chosen. The attention function provides means for particles to first focus strongly on joining their nearest neighbors, and then look around in bigger and bigger length scales. This in fact provides information-theoretic clustering known as *hierarchical* clustering, because a hierarchy of grouping may be determined from which particles join which others first.

See references such as [**48, 84, 83, 69**] for a proper understanding of this problem and its recent state of the art.

Attention a for this section is chosen to be zero near the origin and then jump to one before exponentially decaying. The narrow zone of zero allows for particle merging within that tolerance. W is chosen for this section to be the 2-norm function, i.e. $W^\bullet$ is the identity function, providing simple finite-time collapse. (Although, collapse is precluded due to particle attention dropping to zero.) Both are pictured in figure 8.3.1. Self attention is set at $a_0 = 10^{-6}$ for regularity.

See figure 8.3.2 for the result of the experiment, although it is difficult to visually track in a few snapshots. The particle paths figure 8.3.3 is more revealing, showing the actual hierarchy of grouping that occurred.

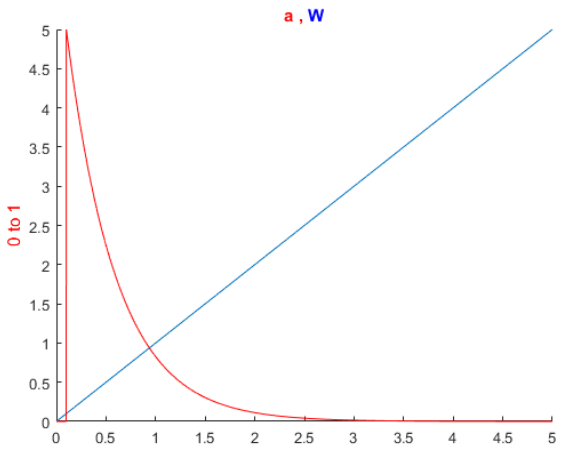

FIGURE 8.3.1. Attention profile $a^\bullet$ (red) and interaction-potential profile $W^\bullet$ (blue). The vertical axis is used for $W^\bullet$, whereas $a^\bullet$ varies from 0 to 1.

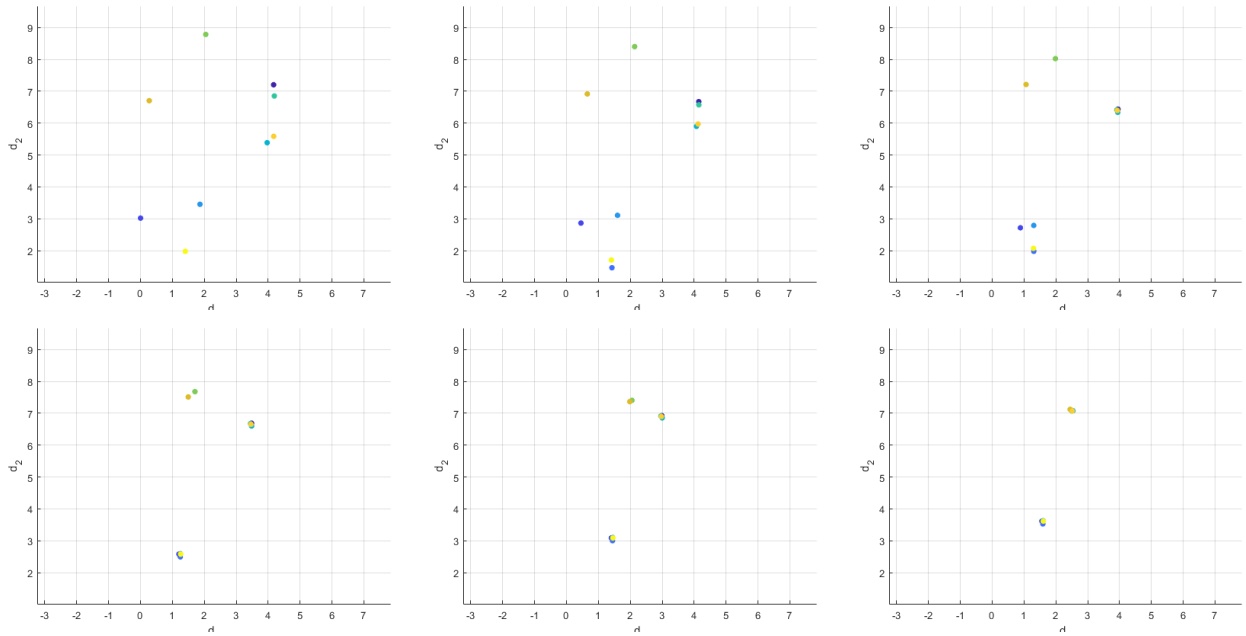

FIGURE 8.3.2. Interesting aspirational application: agglomerative hierarchical clustering (unsupervised machine learning).

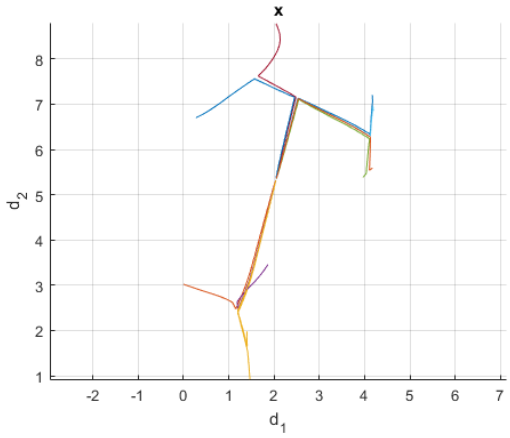

FIGURE 8.3.3. The previous graphic shown with tracer lines.

8.4. Polar milling of distant traveling aggregates

Finally, this section again simulates weighted aggregation as was done in the previous two sections. Here we discover an interesting behavior we call "polar milling".

The same numeric solver and settings were used in this experiment as in the previous two sections, with the exception that relative tolerance was tightened to 10^{-6} to increase confidence in the phenomenon.

Attention a is chosen for this section to have exponential decay with $a(0) = 1$. W is chosen as a Morse potential. Both are pictured in figure 8.4.1. Self attention is set at $a_0 = a(0)$.

Consider an aggregate of particles that has settled into a lattice-like ball corresponding to its own equilibrium, but which lies in the same Euclidean space as a very distant aggregate of other particles. Imagine this distance is so great that during the duration of our experiment, the separating distance won't change significantly, despite that our ball will be moving toward the distant aggregate (and it toward our ball, presumably). Our ball *will* move toward the distant aggregate whenever our ball is roughly in self-equilibrium.

The idealized approximate model is found by substituting all displacements between distant particles with one large unchanging displacement. Doing so in equation (3.1.1) from section 3.1 yields the following.

For all $i \in \mathcal{B}$ where $\mathcal{B}$ is the set of indices making up our ball,

$$\begin{aligned}
\dot{x_i} &= \frac{1}{\alpha_i(\vec{x})} \sum_{j \neq i} \nabla \widetilde{W}(x_j - x_i) \\
&= \frac{1}{\alpha_i(\vec{x})} \left[\sum_{j \in \mathcal{B} \setminus i} \nabla \widetilde{W}(x_j - x_i) + \sum_{j \in \mathcal{B}^C} \nabla \widetilde{W}(x_j - x_i) \right] \\
&\approx \frac{1}{\alpha_i(\vec{x})} \left[\sum_{j \in \mathcal{B} \setminus i} \nabla \widetilde{W}(x_j - x_i) + \sum_{j \in \mathcal{B}^C} \nabla \widetilde{W}(y) \right] \\
&= \frac{1}{\alpha_i(\vec{x})} \left[\sum_{j \in \mathcal{B} \setminus i} \nabla \widetilde{W}(x_j - x_i) + z \right]
\end{aligned}$$

for some $y, z \in \mathbb{R}^d$.

Noteworthy is that this ODE *too* is a gradient flow with respect to the metric of this thesis. The energy is a sum of the interaction energy we have used previously and a new external potential energy,

$$\mathcal{E}^U(\vec{x}) := \frac{1}{2} \sum_{ij} \widetilde{W}(x_i - x_j) + \sum_i U(x_i)$$

where external potential $U : \mathbb{R}^d \to \mathbb{R}$ is simply $U(x) := -z \cdot x$.

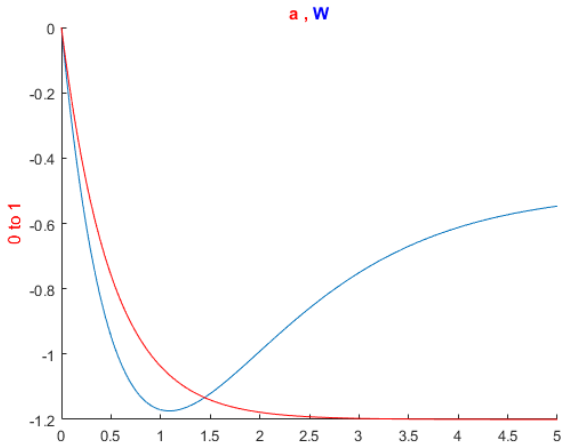

FIGURE 8.4.1. Attention profile $a^\bullet$ (red) and interaction-potential profile $W^\bullet$ (blue). The vertical axis is used for $W^\bullet$, whereas $a^\bullet$ varies from 0 to 1.

This makes sense: U with its simple downward slope in the direction of z causes particles to drift in that direction, subject to the metric tensor on the configuration space which provides the $\frac{1}{\alpha_i(\vec{x})}$ term in the above ODE as before.

This ODE was numerically simulated using $z = e_1$, with evolution shown in figure 8.4.2. It is difficult to see, but something is happening dynamically within the ball as it translates. Figure 8.4.3 helps see slightly more, but still this is difficult to visually track without continuum-time animation. Thus the author *reports* observing milling in which particles near the leading (rightmost) pole move into the ball interior, and subsequently exit the interior at the trailing (leftmost) pole, and then migrate around the outside to the rightmost pole again.

Figure 8.4.4 helps identify this visually.

The 3D case was also simulated, with the resulting graphics capable of rotation during the evolution, which allowed observation of the 3D milling from varying angles: particles entered the ball interior at the leading pole, and exited at the trailing pole, and migrated around the 2D exterior to repeat. The 3D case especially motivates the term "polar mill" as opposed to the 2D case which may appear to fall under the preexisting term "double mill".

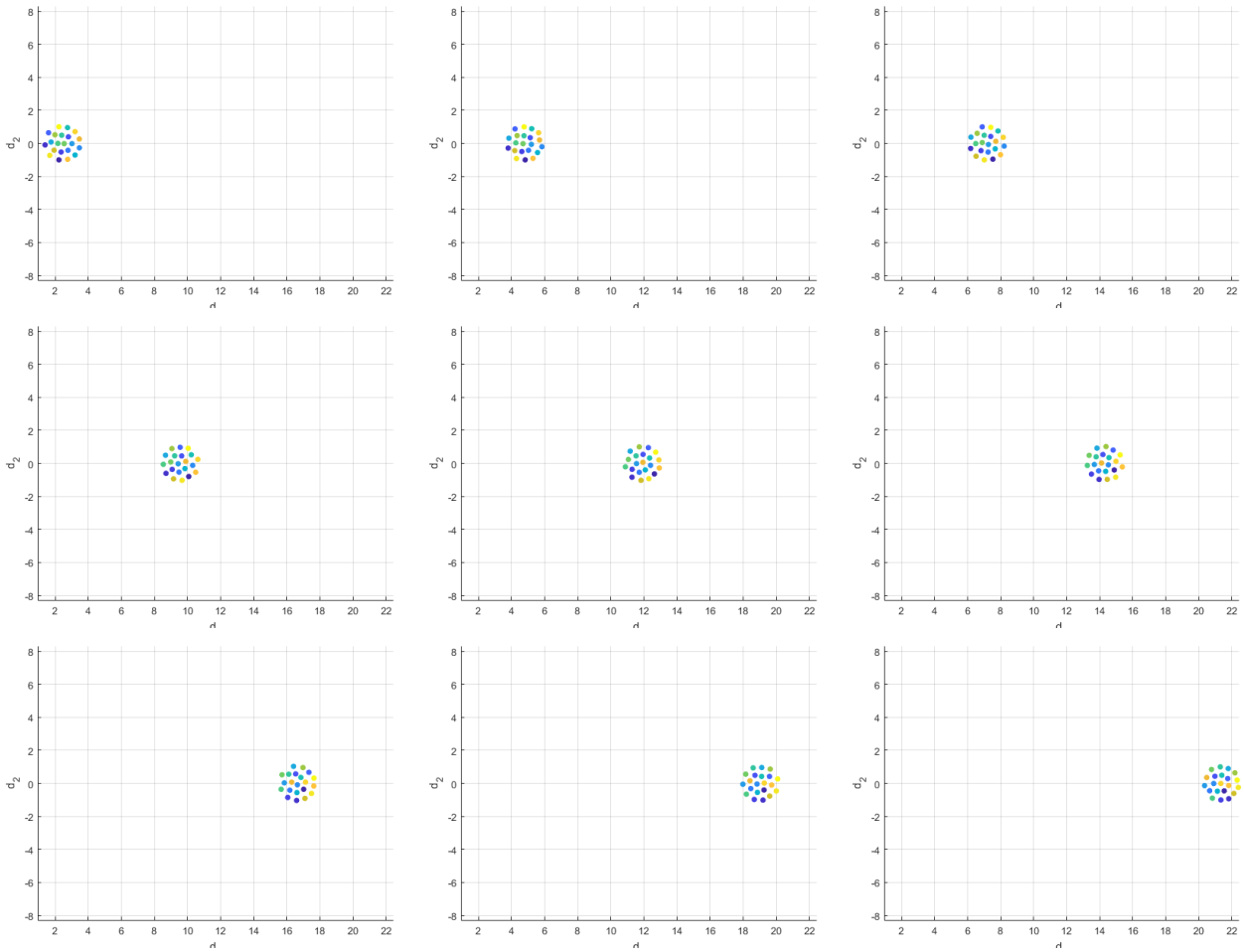

FIGURE 8.4.2. Polar milling of an aggregate as it translates, although difficult to see the milling without animation.

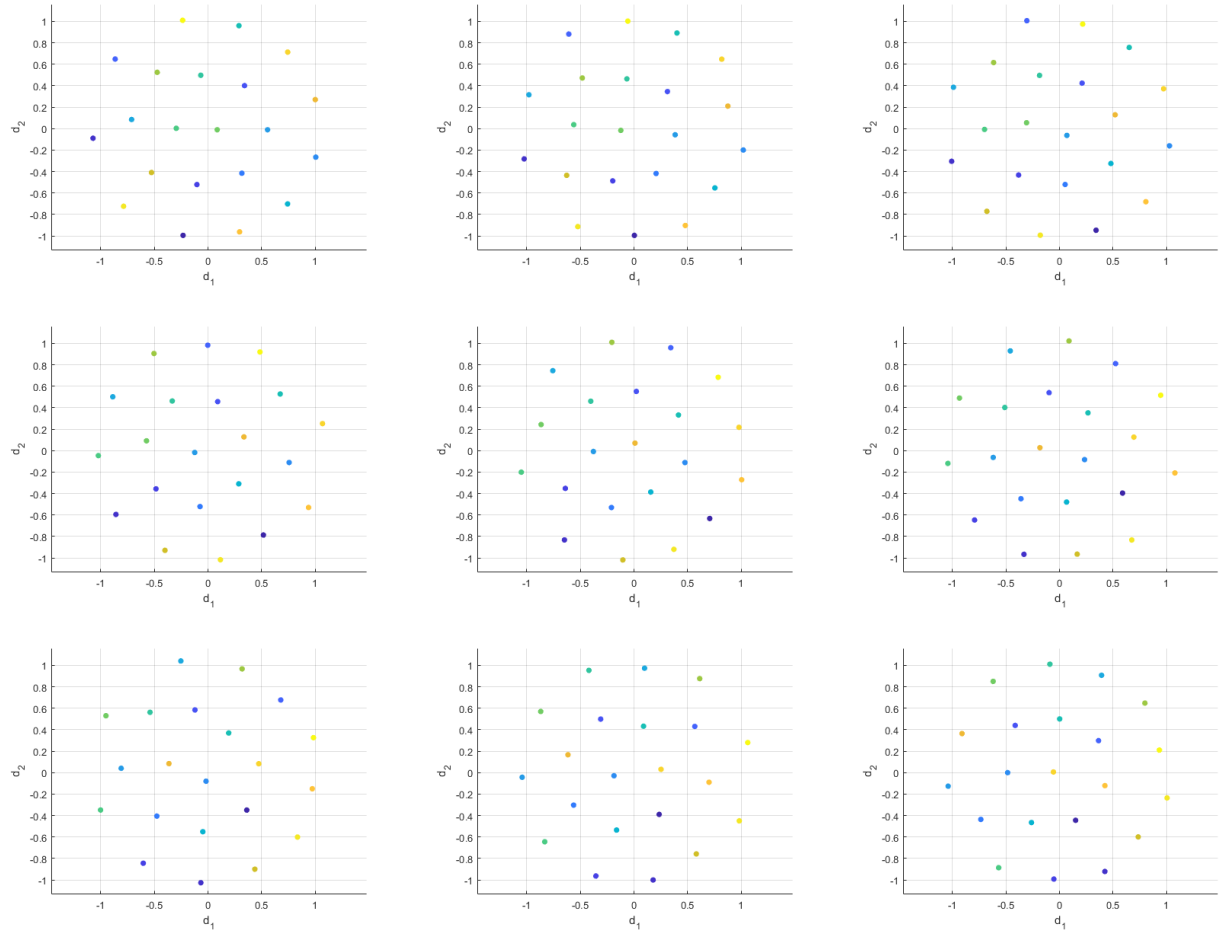

FIGURE 8.4.3. "Zoom in" on the polar milling during translation, by plotting coordinates with respect to center of mass. Still somewhat difficult to see the milling without animation.

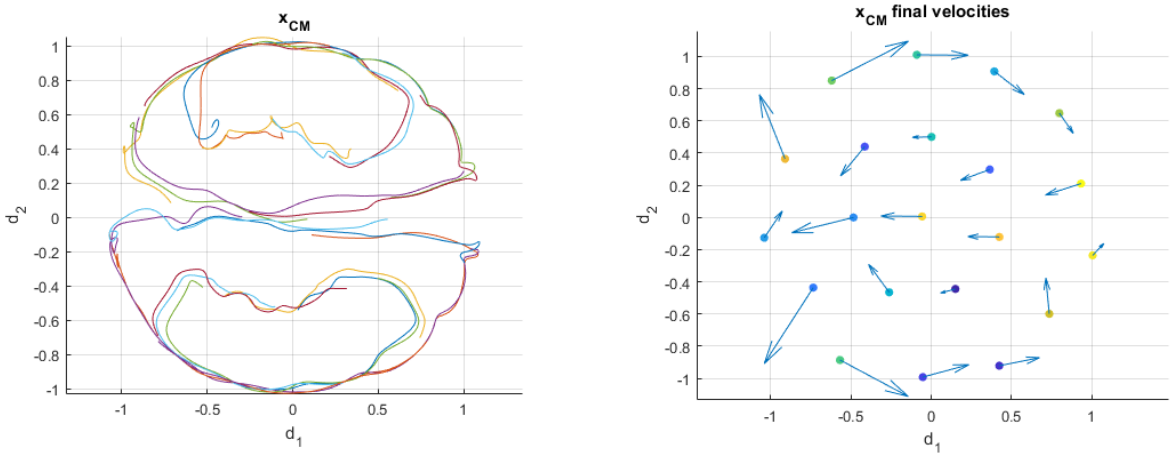

FIGURE 8.4.4. On the left, tracer lines show the particle paths of polar milling. On the right, velocities are shown at the end of the polar milling evolution of the previous figure. Both are in coordinates with respect to center of mass.

Bibliography

[1] L. Ambrosio. Lecture notes on optimal transport problems. In *Mathematical aspects of evolving interfaces (Funchal, 2000)*, volume 1812 of *Lecture Notes in Math.*, pages 1–52. Springer, Berlin, 2003.

[2] L. Ambrosio and N. Gigli. A user's guide to optimal transport. In *Modelling and optimisation of flows on networks*, volume 2062 of *Lecture Notes in Math.*, pages 1–155. Springer, Heidelberg, 2013.

[3] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition, 2008.

[4] D. Balagué, J. A. Carrillo, T. Laurent, and G. Raoul. Nonlocal interactions by repulsive-attractive potentials: radial ins/stability. *Phys. D*, 260:5–25, 2013.

[5] J.-D. Benamou and Y. Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numer. Math.*, 84(3):375–393, 2000.

[6] A. J. Bernoff and C. M. Topaz. A primer of swarm equilibria. *SIAM J. Appl. Dyn. Syst.*, 10(1):212–250, 2011.

[7] A. L. Bertozzi, J. A. Carrillo, and T. Laurent. Blow-up in multidimensional aggregation equations with mildly singular interaction kernels. *Nonlinearity*, 22(3):683–710, 2009.

[8] A. L. Bertozzi, J. B. Garnett, and T. Laurent. Characterization of radially symmetric finite time blowup in multidimensional aggregation equations. *SIAM J. Math. Anal.*, 44(2):651–681, 2012.

[9] A. L. Bertozzi, T. Kolokolnikov, H. Sun, D. Uminsky, and J. von Brecht. Ring patterns and their bifurcations in a nonlocal model of biological swarms. *Commun. Math. Sci.*, 13(4):955–985, 2015.

[10] A. L. Bertozzi and T. Laurent. Finite-time blow-up of solutions of an aggregation equation in $\mathbf{R}^n$. *Comm. Math. Phys.*, 274(3):717–735, 2007.

[11] A. L. Bertozzi, T. Laurent, and F. Léger. Aggregation and spreading via the Newtonian potential: the dynamics of patch solutions. *Math. Models Methods Appl. Sci.*, 22(suppl. 1):1140005, 39, 2012.

[12] A. L. Bertozzi, T. Laurent, and J. Rosado. L^p theory for the multidimensional aggregation equation. *Comm. Pure Appl. Math.*, 64(1):45–83, 2011.

[13] I. Bica, T. Hillen, and K. J. Painter. Aggregation of biological particles under radial directional guidance. *J. Theoret. Biol.*, 427:77–89, 2017.

[14] S. Boi, V. Capasso, and D. Morale. Modeling the aggregative behavior of ants of the species *polyergus rufescens*. *Nonlinear Anal. Real World Appl.*, 1(1):163–176, 2000. Spatial heterogeneity in ecological models (Alcalá de Henares, 1998).

[15] G. A. Bonaschi, J. A. Carrillo, M. Di Francesco, and M. A. Peletier. Equivalence of gradient flows and entropy solutions for singular nonlocal interaction equations in 1D. *ESAIM Control Optim. Calc. Var.*, 21(2):414–441, 2015.

[16] Y. Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Comm. Pure Appl. Math.*, 44(4):375–417, 1991.

[17] M. Burger, V. Capasso, and D. Morale. On an aggregation model with long and short range interactions. *Nonlinear Anal. Real World Appl.*, 8(3):939–958, 2007.

[18] L. A. Caffarelli. The regularity of mappings with a convex potential. *J. Amer. Math. Soc.*, 5(1):99–104, 1992.

[19] L. A. Caffarelli. Boundary regularity of maps with convex potentials. II. *Ann. of Math. (2)*, 144(3):453–496, 1996.

[20] J. A. Cañizo, J. A. Carrillo, and F. S. Patacchini. Existence of compactly supported global minimisers for the interaction energy. *Archive for Rational Mechanics and Analysis*, 217(3):1197–1217, Sep 2015.

[21] J. A. Carrillo, M. Di Francesco, A. Figalli, T. Laurent, and D. Slepčev. Global-in-time weak measure solutions and finite-time aggregation for nonlocal interaction equations. *Duke Math. J.*, 156(2):229–271, 2011.

[22] J. A. Carrillo, M. Di Francesco, A. Figalli, T. Laurent, and D. Slepčev. Confinement in nonlocal interaction equations. *Nonlinear Anal.*, 75(2):550–558, 2012.

[23] J. A. Carrillo, M. R. D'Orsogna, and V. Panferov. Double milling in self-propelled swarms from kinetic theory. *Kinet. Relat. Models*, 2(2):363–378, 2009.

[24] J. A. Carrillo, M. Fornasier, J. Rosado, and G. Toscani. Asymptotic flocking dynamics for the kinetic Cucker-Smale model. *SIAM J. Math. Anal.*, 42(1):218–236, 2010.

[25] J. A. Carrillo, M. Fornasier, G. Toscani, and F. Vecil. Particle, kinetic, and hydrodynamic models of swarming. In *Mathematical modeling of collective behavior in socio-economic and life sciences*, Model. Simul. Sci. Eng. Technol., pages 297–336. Birkhäuser Boston, Inc., Boston, MA, 2010.

[26] J. A. Carrillo, F. James, F. Lagoutière, and N. Vauchelet. The Filippov characteristic flow for the aggregation equation with mildly singular potentials. *J. Differential Equations*, 260(1):304–338, 2016.

[27] J. A. Carrillo, S. Lisini, G. Savaré, and D. Slepčev. Nonlinear mobility continuity equations and generalized displacement convexity. *J. Funct. Anal.*, 258(4):1273–1309, 2010.

[28] J. A. Carrillo, S. Martin, and V. Panferov. A new interaction potential for swarming models. *Phys. D*, 260:112–126, 2013.

[29] J. A. Carrillo, R. J. McCann, and C. Villani. Contractions in the 2-Wasserstein length space and thermalization of granular media. *Arch. Ration. Mech. Anal.*, 179(2):217–263, 2006.

[30] Y.-l. Chuang, M. R. D'Orsogna, D. Marthaler, A. L. Bertozzi, and L. S. Chayes. State transitions and the continuum limit for a 2D interacting, self-propelled particle system. *Phys. D*, 232(1):33–47, 2007.

[31] D. Cordero-Erausquin, R. J. McCann, and M. Schmuckenschläger. A Riemannian interpolation inequality à la Borell, Brascamp and Lieb. *Invent. Math.*, 146(2):219–257, 2001.

[32] I. D. Couzin, J. Krause, R. James, G. D. Ruxton, and N. R. Franks. Collective memory and spatial sorting in animal groups. *J. Theoret. Biol.*, 218(1):1–11, 2002.

[33] F. Cucker and S. Smale. Emergent behavior in flocks. *IEEE Trans. Automat. Control*, 52(5):852–862, 2007.

[34] F. Cucker and S. Smale. On the mathematics of emergence. *Jpn. J. Math.*, 2(1):197–227, 2007.

[35] G. De Philippis and A. Figalli. The Monge-Ampère equation and its link to optimal transportation. *Bull. Amer. Math. Soc. (N.S.)*, 51(4):527–580, 2014.

[36] P. Degond, J.-G. Liu, S. Motsch, and V. Panferov. Hydrodynamic models of self-organized dynamics: derivation and existence theory. *Methods Appl. Anal.*, 20(2):89–114, 2013.

[37] P. Degond and S. Motsch. Continuum limit of self-driven particles with orientation interaction. *Math. Models Methods Appl. Sci.*, 18(suppl.):1193–1215, 2008.

[38] M. P. a. do Carmo. *Riemannian geometry*. Mathematics: Theory & Applications. Birkhäuser Boston, Inc., Boston, MA, 1992. Translated from the second Portuguese edition by Francis Flaherty.

[39] R. Eftimie, G. de Vries, and M. A. Lewis. Complex spatial group patterns result from different animal communication mechanisms. *Proc. Natl. Acad. Sci. USA*, 104(17):6974–6979, 2007.

[40] M. Erbar. Gradient flows of the entropy for jump processes. *Ann. Inst. Henri Poincaré Probab. Stat.*, 50(3):920–945, 2014.

[41] L. C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 1998.

[42] L. C. Evans and W. Gangbo. *Differential equations methods for the Monge-Kantorovich mass transfer problem*. American Mathematical Soc., 1999.

[43] R. C. Fetecau, Y. Huang, and T. Kolokolnikov. Swarm dynamics and equilibria for a nonlocal aggregation model. *Nonlinearity*, 24(10):2681–2716, 2011.

[44] W. Gangbo and R. J. McCann. Optimal maps in Monge's mass transport problem. *C. R. Acad. Sci. Paris Sér. I Math.*, 321(12):1653–1658, 1995.

[45] W. Gangbo and R. J. McCann. The geometry of optimal transportation. *Acta Math.*, 177(2):113–161, 1996.

[46] S.-Y. Ha and J.-G. Liu. A simple proof of the Cucker-Smale flocking dynamics and mean-field limit. *Commun. Math. Sci.*, 7(2):297–325, 2009.

[47] Y. Huang and A. Bertozzi. Asymptotics of blowup solutions for the aggregation equation. *Discrete Contin. Dyn. Syst. Ser. B*, 17(4):1309–1331, 2012.

[48] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Comput. Surv.*, 31(3):264–323, Sept. 1999.

[49] R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the Fokker-Planck equation. *SIAM J. Math. Anal.*, 29(1):1–17, 1998.

[50] L. V. Kantorovich. On translation of mass (in Russian), C R. Doklady. *Acad. Sci. USSR*, 37:199–201, 1942.

[51] L. V. Kantorovich. A problem of Monge. *Uspekhi Mat. Nauk*, 3(24):225–226, 1948.

[52] L. V. Kantorovitch. On the translocation of masses. *Management Science*, 5(1):1–4, 1958.

[53] Y. Katz, K. Tunstrøm, C. C. Ioannou, C. Huepe, and I. D. Couzin. Inferring the structure and dynamics of interactions in schooling fish. *Proceedings of the National Academy of Sciences*, 108(46):18720–18725, 2011.

[54] T. Kolokolnikov, J. A. Carrillo, A. Bertozzi, R. Fetecau, and M. Lewis. Emergent behaviour in multi-particle systems with non-local interactions [Editorial]. *Phys. D*, 260:1–4, 2013.

[55] T. Laurent. Local and global existence for an aggregation equation. *Comm. Partial Differential Equations*, 32(10-12):1941–1964, 2007.

[56] I. Lebar Bajec, N. Zimic, and M. Mraz. The computational beauty of flocking: boids revisited. *Math. Comput. Model. Dyn. Syst.*, 13(4):331–347, 2007.

[57] A. J. Leverentz, C. M. Topaz, and A. J. Bernoff. Asymptotic dynamics of attractive-repulsive swarms. *SIAM J. Appl. Dyn. Syst.*, 8(3):880–908, 2009.

[58] A. J. Majda and A. L. Bertozzi. *Vorticity and incompressible flow*, volume 27 of *Cambridge Texts in Applied Mathematics*. Cambridge University Press, Cambridge, 2002.

[59] R. J. McCann. A convexity principle for interacting gases. *Adv. Math.*, 128(1):153–179, 1997.

[60] P. A. Milewski and X. Yang. A simple model for biological aggregation with asymmetric sensing. *Commun. Math. Sci.*, 6(2):397–416, 2008.

[61] A. Mogilner, L. Edelstein-Keshet, L. Bent, and A. Spiros. Mutual interactions, potentials, and individual distance in a social aggregation. *J. Math. Biol.*, 47(4):353–389, 2003.

[62] G. Monge. *Mémoire sur la théorie des déblais et des remblais*. De l'Imprimerie Royale, 1781.

[63] S. Motsch and E. Tadmor. A new model for self-organized dynamics and its flocking behavior. *J. Stat. Phys.*, 144(5):923–947, 2011.

[64] S. Motsch and E. Tadmor. Heterophilious dynamics enhances consensus. *SIAM Rev.*, 56(4):577–621, 2014.

[65] F. Otto. The geometry of dissipative evolution equations: the porous medium equation. *Comm. Partial Differential Equations*, 26(1-2):101–174, 2001.

[66] S. T. Rachev and L. Rüschendorf. *Mass transportation problems. Vol. I.* Probability and its Applications (New York). Springer-Verlag, New York, 1998. Theory.

[67] C. W. Reynolds. Flocks, herds and schools: A distributed behavioral model. In *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '87, pages 25–34, New York, NY, USA, 1987. ACM.

[68] F. Santambrogio. *Optimal transport for applied mathematicians*, volume 87 of *Progress in Nonlinear Differential Equations and their Applications*. Birkhäuser/Springer, Cham, 2015. Calculus of variations, PDEs, and modeling.

[69] A. Saxena, M. Prasad, A. Gupta, N. Bharill, O. P. Patel, A. Tiwari, M. J. Er, W. Ding, and C.-T. Lin. A review of clustering techniques and developments. *Neurocomputing*, 267(Supplement C):664 – 681, 2017.

[70] E. Schechter. *Handbook of analysis and its foundations*. Academic Press, Inc., San Diego, CA, 1997.

[71] R. Simione, D. Slepčev, and I. Topaloglu. Existence of ground states of nonlocal-interaction energies. *Journal of Statistical Physics*, 159(4):972–986, May 2015.

[72] V. Sudakov. *Geometric Problems in the Theory of Infinite-dimensional Probability Distributions*. Number no. 141 in Geometric Problems in the Theory of Infinite-dimensional Probability Distributions. American Mathematical Society, 1979.

[73] H. Sun, D. Uminsky, and A. L. Bertozzi. Stability and clustering of self-similar solutions of aggregation equations. *J. Math. Phys.*, 53(11):115610, 18, 2012.

[74] E. Tadmor and C. Tan. Critical thresholds in flocking hydrodynamics with non-local alignment. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 372(2028):20130401, 22, 2014.

[75] C. M. Topaz and A. L. Bertozzi. Swarming patterns in a two-dimensional kinematic model for biological groups. *SIAM J. Appl. Math.*, 65(1):152–174, 2004.

[76] C. M. Topaz, A. L. Bertozzi, and M. A. Lewis. A nonlocal continuum model for biological aggregation. *Bull. Math. Biol.*, 68(7):1601–1623, 2006.

[77] C. M. Topaz, M. R. D'Orsogna, L. Edelstein-Keshet, and A. J. Bernoff. Locust dynamics: behavioral phase change and swarming. *PLoS Comput. Biol.*, 8(8):e1002642, 11, 2012.

[78] T. Vicsek, A. Czirók, E. Ben-Jacob, I. Cohen, and O. Shochet. Novel type of phase transition in a system of self-driven particles. *Phys. Rev. Lett.*, 75(6):1226–1229, 1995.

[79] C. Villani. *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003.

[80] C. Villani. *Optimal transport*, volume 338 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2009. Old and new.

[81] J. H. von Brecht, D. Uminsky, T. Kolokolnikov, and A. L. Bertozzi. Predicting pattern formation in particle interactions. *Math. Models Methods Appl. Sci.*, 22(suppl. 1):1140002, 31, 2012.

[82] L. Wu and D. Slepčev. Nonlocal interaction equations in environments with heterogeneities and boundaries. *Comm. Partial Differential Equations*, 40(7):1241–1281, 2015.

[83] D. Xu and Y. Tian. A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2):165–193, Jun 2015.

[84] R. Xu, J. Xu, and D. C. Wunsch. A comparison study of validity indices on swarm-intelligence-based clustering. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(4):1243–1256, 2012.