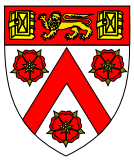
Asymptotic Behaviour and Derivation of Mean Field Models

Thomas James Holding



Trinity College

This dissertation is submitted for the degree of Doctor of Philosophy.

August, 2016.

Abstract

This thesis studies various problems related to the asymptotic behaviour and derivation of mean field models from systems of many particles.

Chapter 1 introduces mean field models and their derivation, and then summarises the following chapters of this thesis.

Chapters 2, 3 and 4 directly study systems composed of many particles.

In Chapter 2 we prove quantitative propagation of chaos for systems of interacting SDEs with interaction kernels that are merely Hölder continuous (the usual assumption being Lipschitz). On the way we prove the existence of differentiable stochastic flows for a class of degenerate SDEs with rough coefficients and a uniform law of large numbers for SDEs.

Chapters 3 and 4 study the asymptotic behaviour of the Arrow-Hurwicz-Uzawa gradient method, which is a dynamical system for locating saddle points of concave-convex functions. This method is widely used in distributed optimisation over networks, for example in power systems and in rate control in communication networks. Chapter 3 gives an exact characterisation of the limiting solutions of the gradient method on the full space for arbitrary concave-convex functions. In Chapter 4 we extend this result to the subgradient method where the dynamics of the gradient method are restricted to an arbitrary convex set.

Chapters 5, 6 and 7 study the stability of mean field models. Chapters 5 and 6 prove an instability criterion for non-monotone equilibria of the Vlasov-Maxwell system. In Chapter 5 we study a related problem in approximation of the spectra of families of unbounded self adjoint operators. In Chapter 6 we show how the instability problem for Vlasov-Maxwell can be reduced to this spectral problem.

In Chapter 7 we give a proof of well-posedness of a class of solutions to the Vlasov-Poisson system with unbounded spatial density.

Chapters 8 and 9 change track and study the dynamics of a solute in a fluid background. In Chapter 8 we study a simple model for this phenomena, the kinetic Fokker-Planck equation, and show contraction of its semi-group in the Wasserstein distance when the spatial variable lies on the torus. Chapter 9 studies a more complex model of passive transport of a solute under a large and highly oscillatory fluid field. We prove a homogenisation result showing convergence to an effective diffusion equation for the transported solute profile.

Declaration

This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except where specified in the text. This dissertation is not substantially the same as any that I have submitted for a degree or diploma or other qualification at any other university. This dissertation does not exceed the prescribed limit of 60 000 words.

Thomas James Holding
August, 2016

Acknowledgements

Most of all, I would like to thank my supervisors Clément Mouhot and José A. Carrillo for their support, enthusiasm, broad and varied mathematical insights and whose work ethic is an inspiration I cannot hope to match. It has been a privilege to be their student. Previously to their supervision, I would also like to mention Clément's kinetic theory course that was inspirational to me while I was taking Part III and it is not an exaggeration to say that it has shaped my mathematical interests ever since.

I would like to thank my collaborators, Jonathan Ben-Artzi, Helge Dietert, Jo Evans, Harsha Hutridurga, Ioannis Lestas, Evelyne Miot, and Jeffrey Rauch, all of whom it has been a pleasure to work with and without whom I would not have been able to complete this thesis. In particular, I am grateful to Jonathan and Ioannis for helping me improve my technical writing during the first years of my time in the CCA, a thankless task for which the others can be grateful, to Jeff for always being willing to take the time aside to state and prove (from memory) interesting results for my benefit, to my academic siblings Jo and Helge for tolerating my incoherence and over enthusiasm (which I know was grating at the time), and to Harsha, with whom I spent the most time working face to face and spent many a fondly remembered hour staring at seemingly impossible problems and then talking about other things.

My examiners can be thankful to Harsha as his decision to give up smoking may improve his health, but decreased our productivity just enough to cost this thesis an additional chapter, but not to Evelyne, whose remarkable dedication during the final month allowed the addition of a different chapter to replace it.

Speaking of remarkable dedication at the last minute, special thanks must go to

Helge Dietert and Vittoria Silvestri for selflessly proof reading the introductory chapter within a single day of receiving it a ridiculous two days before my submission, and to Tom Begley and Karen Habermann for assistance with printing and binding on the day of submission.

More specific acknowledgements related to the work in this thesis will be given at the start of each chapter.

I would like to thank those who are still or were part of the kinetic group during my time in Cambridge: Marc Briant, Ludovic Cesbron, Helge Dietert, Amit Einav, Jo Evans, Megan Griffin-Pickering, Franca Hoffmann, Harsha Hutridurga, Mikaela Iacobelli, Sara Merino-Acetuno, Davide Piazzoli[1], Ariane Trescases and of course my supervisor Clément Mouhot. The Cambridge 'kinetic coffee', now apparently world renowned (or so I am told), was always a pleasure to attend, whether we discussed mathematics, food, how to cross the road in different countries or the correct way to keep hens. In particular, I thank my academic older siblings Marc and Sara for going though it all two years before me and being proof that there is light at the end of the tunnel, Amit for always knowing the best places to eat in any locality, Franca, my fermionic academic sister, with whom I share two supervisors but seemingly never a home city (and see most of at conferences), and Jo for always being there in times of stress to distract me with her mathematical problems and her conversation[2].

I would also like to mention José's other students who reside at Imperial College London, Francesco Patacchini, Markus Schmidtchen and of course the already mentioned Franca Hoffmann, all of whom it has been a pleasure to spend time with when I visit Imperial or attend conferences.

I would like to thank the probability group for putting up with me attending their seminars and eating their cake, in spite of the obvious fact that I have never passed a written exam in the field, and likely never will, and the control group in the engineering department for graciously accepting me when I crashed their seminars and criticised their finance department approved cake, and for

[1]With whom I share the much coveted 'Smoothest handover between talks award' from the MASDOC CCA Conference 2015, which holds pride of place on the office wall.

[2]Her (accidentally) last words to me in Cambridge were that I'm a bit odd, but well within the usual range for mathematicians, which I suppose I'll have to take as a compliment...

making my first conference abroad in Florence an experience that I remember fondly.

The CCA third cohort will always hold a special place in my heart, the first year being a baptism of fire that none of us will forget, which has welded us together in a way that has held ever since. Our lunch time conversations are the highlight of the workday, and their camaraderie in the otherwise lonely reality of working towards a PhD will be one of the things I will miss the most after leaving. Among many CCA students in my own and other years, some deserve special mention: my office mates Henry Jackson and Kim Moore for putting up with me for these four years, Kim for somehow making it appear that the office cats were not in any way due to me, and Henry for tolerating the many simulacra of him that were constructed during times of idleness, Vittoria Silvestri for her continual insistence that I am both hard working and competent, despite me providing ample evidence to the contrary of each, Rob Hocking in whose life there are more entertaining and ridiculous stories than would seem to sensibly fit, and whom it was a pleasure to work with (on a project that will be ready for his thesis, but not my own), Karen Habermann - honorary third cohort member - who has been volunteered into looking after the cats once we leave. Lastly I would like to give special thanks to Henry, Jo, Kim, and Vittoria for their moral support in our final year and for their enduring confidence in me, which far surpasses my own.

Finally, I am thankful for the continual support of my family, especially my parents, without which things would have been very different, and last but not least, I would like to thank Imre Leader for his support at key points during my time in Cambridge.

Contents

Chapter 1

Introduction

We begin this thesis with a look through the menagerie of mean field equations and their formal derivation. During this expository adventure we will briefly hint at some of the results presented in this thesis. After that we will give a more detailed description of the results contained in this manuscript, chapter by chapter.

1.1 Layout of this thesis

This thesis is composed of nine chapters. Aside from this introductory chapter, these consist of research papers, namely [87, 92, 93, 18, 19, 95, 46, 88]. These chapters are self contained and can be read individually in any order, although it should be noted that Chapter 6 applies the results of Chapter 5, and that Chapter 4 builds on the results of Chapter 3.

1.2 Mean field models

Mean field models describe the evolution of a density of a very large number of interacting particles where the force on each particle is approximated by an averaged *mean field* force over all the other particles. They are of central importance

in the mathematical study of many phenomena in the physical, biological and social sciences, ranging from the dynamics of plasmas and galaxies to the swarming and flocking of fish and birds. Examples of mean field models include the Vlasov-Poisson and Vlasov-Maxwell equations for galaxies and plasmas, the vorticity formulation of the two dimensional Euler equation for incompressible fluids, the Hartree equation in quantum mechanics and the aggregation and aggregation-diffusion equations. We refer the reader to [77, 113, 182] for viewpoints from both mathematics and physics.

1.2.1 The Boltzmann equation.

Although it is not strictly a mean field model, any history of such models must begin with the Boltzmann equation (see e.g. [36]) as it was the first instance of an equation derived from the interaction of many particles, in this case the colliding molecules that compose a rarefied gas.

The most basic description of such a gas would be the complete list of the positions and velocities of each of the $N \gg 1$ particles that compose it, which is a vector in $\mathbb{R}^{6N}$. The phase-space distribution function of this system is a (non-negative) function $f^N(z_1, z_2, \ldots, z_N)$ that maps $\mathbb{R}^{6N}$ to $\mathbb{R}_+$ and is symmetric under the relabelling of particles. (Here and later we we use $z_i = (x_i, v_i) \in \mathbb{R}^{3+3}$ to denote a coordinate in phase-space.) However, from a practical perspective, the only relevant physical quantity is the first marginal

$$f^{1,N}(x_1, v_1) = \int f^N(z_1, z_2, \ldots, z_N)\, dz_2 \cdots dz_N \tag{1.2.1}$$

also known as the one-particle distribution function, which describes the statistics of a single particle and allows the computation of macroscopic quantities such as temperature and local density. Due to symmetry under relabelling of particles, the special choice of z_1 in the above formula is arbitrary and irrelevant. Integrating out any other choice of $N-1$ phase-space variables would give the same function. More generally we can define the k-th marginal (k-particle distribution function) $f^{k,N}(z_1, \ldots, z_k)$ by the formula

$$f^{k,N}(z_1, \ldots, z_k) = \int f^N(z_1, \ldots, z_k, z_{k+1}, \ldots, z_N)\, dz_{k+1} \cdots dz_N.$$

Although not as observationally physically relevant as the first marginal $f^{1,N}$, the higher marginals are important as they describe correlations between the particles. They are also important mathematically, as they arise when one tries to close an evolution equation on the one-particle distribution function $f^{1,N}$. Indeed, the full N-particle distribution evolves according to the Liouville equation

$$\partial_t f^N(t, z_1, \dots, z_N) + \sum_{i=1}^{N} v_i \cdot \nabla_{x_i} f^N(t, z_1, \dots, z_N) = \{\text{changes in } f^N \text{ from collisions}\} \tag{1.2.2}$$

where we have kept the collisional term schematic. To obtain the time derivative of the one-particle distribution function $f^{1,N}$, we integrate out the variables $z_2, \dots z_N$ in the Liouville equation (1.2.2). However, the effect of collisions upon the *one*-particle distribution function $f^{1,N}$ cannot be computed from $f^{1,N}$ alone, because, as the collisions are pairwise, they depend upon the *two*-particle distribution function $f^{2,N}$. Thus we have obtained

$$\partial_t f^{1,N}(t, x_1, v_1) + v_1 \cdot \nabla_{x_1} f^{1,N}(t, x_1, v_1) = \{\text{collisional term involving } f^{2,N}\}.$$

Similarly, when integrating (1.2.2) to obtain an equation for $f^{2,N}$, the collisional term involves $f^{3,N}$, and repeating for $f^{3,N}, f^{4,N}, \dots$ we obtain a hierarchy of equations: the *Bogoliubov-Born-Green-Kirkwood-Yvon (BBGKY) hierarchy* (see e.g. the lecture notes [77]), in which each equation involves the function described by the next equation.

Building upon the earlier work of Maxwell [141], Boltzmann theorised that the velocities of two particles just prior to performing a collision are approximately independent, i.e. that

$$f^{2,N}(t, x_1, v_1, x_2, v_2) \approx f^{1,N}(t, x_1, v_1) f^{1,N}(t, x_2, v_2). \tag{1.2.3}$$

This assumption, now commonly known as 'molecular chaos' and referred to by Boltzmann at the time as the 'Stoßzahlansatz', allowed him to close an equation on the time evolution of the one-particle distribution function $f(t, x, v) = f^{1,N}(t, x, v)$:

$$\partial_t f(t, x, v) + v \cdot \nabla_x f(t, x, v) = Q(f, f)$$

where Q is the collision operator, which is a bilinear integral operator acting only

in the v variable.

The Boltzmann equation is not of mean field type. This is because, in its derivation as a limit $N \to \infty$ of Newtonian dynamics, to ensure that the number of collisions a typical particle experiences per unit time remains constant, it is necessary to scale the *particle size* with N, in what is called the 'Boltzmann-Grad limit' (see e.g. the recent work [70]). This scaling is a different nature to mean field models, such as the Vlasov equation described below, in which the *inter-particle forces* are scaled instead.

We close the discussion of the Boltzmann equation by remarking that the rigorous derivation of the Boltzmann equation from Newtonian dynamics remains one of the major open problems in kinetic theory and mathematical physics, which we do not touch upon in this thesis. The best result in this direction remains Lanford's 1975 proof [120] that the Boltzmann-Grad limit holds for a very short time, which is of the order of the time between two successive collisions of a typical particle. Recently, Lanford's result has been revisited and refined by Gallagher, Saint-Raymond and Texier [70], which we recommend to the interested reader as a starting point.

1.2.2 The Vlasov equation.

The other important equation in the history of mean field models is the Vlasov equation [77, 182]. Consider N particles with positions and velocities $(Z^{i,N})_{i=1}^N = (X^{i,N}, V^{i,N})_{i=1}^N \in \mathbb{R}^{6N}$. We assume that these evolve under Newton's laws with pairwise interaction forces given by a kernel K, i.e.

$$\begin{cases} \dot{X}_t^{i,N} = V_t^{i,N} \\ \dot{V}_t^{i,N} = \dfrac{1}{N} \displaystyle\sum_{j=1}^{N} K(X_t^{i,N}, X_t^{j,N}) \end{cases} \tag{1.2.4}$$

Here the factor N^{-1} in front of the interaction term ensures that the force on a single particle stays of order one independently of N, and we define $K(x,x) = 0$ for notational simplicity.

While the Vlasov equation can be derived from the BBGKY hierarchy, a more

convenient tool in this case turns out to be the *empirical measure* μ^N defined by

$$\mu_t^N = \frac{1}{N}\sum_{i=1}^{N} \delta_{Z_t^{i,N}} \tag{1.2.5}$$

which is the average of Dirac masses at each particle's phase-space position. Formally, the empirical measure μ^N is a weak solution to the following equation

$$\begin{cases} \partial_t \mu_t^N + v \cdot \nabla_x \mu_t^N + F_t \cdot \nabla_v \mu_t^N = 0 \\ F_t(x) = \int K(x,y)\, d\mu_t^N(y) \end{cases} \tag{1.2.6}$$

where we note that the integral in the definition of F is equal to $\frac{1}{N}\sum_{j=1}^{N} K(x, X_t^{j,N})$ by the definition of the empirical measure. The equation (1.2.6) is a general form of the Vlasov equation. At a formal level, we expect that in taking $N \to \infty$ so that μ^N approaches a smooth density, we would obtain a classical solution or at the very least a solution that is a function and not a measure.

For Coulombic interactions, which in the attractive case model stars in a galaxy and in the repulsive case model ions in a plasma [114], (1.2.6) is known as the *Vlasov-Poisson equation* and may be written as

$$\begin{cases} \partial_t f_t(x,v) + v \cdot \nabla_x f_t(x,v) - \nabla_x \phi_t(x) \cdot \nabla_v f_t(x,v) = 0, \\ -\Delta \phi_t(x) = \pm \rho_t(x), \quad \rho_t(x) = \int f_t(x,v)\, dv, \end{cases} \tag{1.2.7}$$

where the repulsive and attractive cases are the $+$ and $-$ cases respectively. The Vlasov-Poisson system is widely used in the modelling of both plasmas and galaxies as it captures kinetic effects such as Landau damping, the relaxation towards equilibrium of small perturbations in a plasma predicted mathematically by Landau in 1946 [119], which, although observed experimentally, is not predicted by fluid mechanical models such as *Magneto-Hydro-Dynamics (MHD)*. The nonlinear theory of Landau damping was recently made mathematically rigorous by Mouhot and Villani [151, 150], and this has lead to the development of a general theory of such damping in this kind of kinetic and fluid equations, from which we mention Dietert [45] on the Kuramoto model, and Bedrossian and Masmoudi [15] on the two dimensional Euler equation, among many others.

For plasmas, the Vlasov-Poisson equation assumes that the interactions are given

by an electrostatic force, which corresponds to assuming infinite speed of propagation of electromagnetic radiation. To remove this assumption, we must replace the electrostatic interactions with a full coupling with Maxwell's equations for a dynamic electromagnetic field. By doing so we recover the *Vlasov-Maxwell system*:

$$\begin{cases} \partial_t f^{\pm}(t,x,v) + \hat{v} \cdot \nabla_x f^{\pm}(t,x,v) \pm (E(t,x) + \hat{v} \times B(t,x)) \cdot \nabla_v f^{\pm}(t,x,v) = 0, \\ \nabla_x \cdot B = 0, \qquad\qquad \partial_t B = -\nabla_x \times E, \\ \nabla_x \cdot E = \displaystyle\int_{\mathbb{R}^3} (f^+ - f^-)\, dv, \quad \partial_t E = \nabla_x \times B - \int_{\mathbb{R}^3} \hat{v}(f^+ - f^-)\, dv. \end{cases} \tag{1.2.8}$$

Here $f^{\pm}(t,x,v)$ are the phase space densities of positively and negatively charged particles, and $\hat{v} = v/\sqrt{1+|v|^2}$ is the relativistic velocity.

The Vlasov-Maxwell equation (1.2.8) has the structure of a wave equation and has finite speed of propagation. In contrast, Vlasov-Poisson (1.2.7) has infinite speed of propagation of information, even if the maximum speed of the particles is bounded. The additional difficulty of the dynamic electromagnetic field makes analysis of the Vlasov-Maxwell system (1.2.8) more challenging than the simpler Vlasov-Poisson equation (1.2.7). For example, global existence and uniqueness of smooth solutions is known in the full three dimensional setting for Vlasov-Poisson (1.2.7), the original result due to Pfaffelmoser [168], later simplified by Schaeffer [177], with an alternative method given by Lions and Perthame [133]. For the Vlasov-Maxwell system, global existence and uniqueness of smooth solutions in three dimensions is known only under some assumption of symmetry, a result due to Glassey and Schaeffer [74] refining their earlier results in dimension [73, 75]. The full three dimensional Cauchy problem remains an open problem.

In this thesis we present two new results on the Vlasov-Poisson and Vlasov-Maxwell systems. In Chapters 5 and 6 we present new linear instability results for the Vlasov-Maxwell system (1.2.8). In Chapter 7 we present a new well-posedness result for the Vlasov-Poisson system (1.2.7) for a class of solutions with unbounded density.

1.2.3 First order systems.

Another class of mean field models arises from first order dynamics, where the second order Newtonian dynamics in (1.2.4) are replaced by a first order system where only the particle positions are considered. Indeed, consider N particles in $\mathbb{R}^d$ with positions $(X^{i,N})_{i=1}^N \in \mathbb{R}^{Nd}$, that evolve as

$$\dot{X}_t^{i,N} = \frac{1}{N}\sum_{j=1}^{N} K(X_t^{i,N}, X_t^{j,N}) \tag{1.2.9}$$

for an interaction kernel $K(x,y)$. The forces are scaled in the same way as in (1.2.4), ensuring that they remain of order one independent of N.

The corresponding empirical measure is

$$\mu_t^N = \frac{1}{N}\sum_{i=1}^{N} \delta_{X_t^{i,N}} \tag{1.2.10}$$

which is a probability measure on $\mathbb{R}^d$. Formally, the empirical measure (1.2.10) is a weak solution to the following non-linear first order equation

$$\begin{cases} \partial_t \mu_t^N + \nabla \cdot (F_t \mu_t^N) = 0, \\ F_t(x) = \int K(x,y) d\mu_t^N(y). \end{cases} \tag{1.2.11}$$

Although, strictly speaking, (1.2.11) (respectively (1.2.9)) includes the second order system (1.2.4) (respectively (1.2.6)) as a special case, the second order structure of (1.2.4), (1.2.6) has distinctive properties that make the analysis different. In particular, second order systems are slightly more stable, in the sense that changes in the forces have to filter through the velocities before affecting the positions, a property exploited in Chapter 7 to show improved well-posedness results for the Vlasov-Poisson system (1.2.7). Furthermore, in the case where noise is added (discussed in Section 1.2.4), it is often more natural to consider noise in the velocity variable only and not in the spatial dynamics. This distinction changes the form of the results in Chapter 2 (summarised below in Section 1.3.5) on propagation of chaos for noisy systems.

A first example of a specific first order mean field model is the vorticity formu-

lation of the two dimensional incompressible Euler system. Although the two dimensional Euler equation is not derived as a mean field limit of a particle system. In its vorticity formulation, it has an interpretation, originally due to Helmholtz [84], as the $N \to \infty$ limit of the interaction of point vortices, where the total vortex strength normalised to a constant.

$$\begin{cases} \partial_t \omega_t + u_t \cdot \nabla \omega_t = 0 \\ u_t = \dfrac{x^\perp}{|x|^2} * \omega_t \end{cases} \tag{1.2.12}$$

In many ways the two dimensional vorticity equation (1.2.12) behaves like a first order counterpart to the Vlasov-Poisson system in two spatial dimensions (four phase space dimensions). They share the same strength of interaction kernel, the only difference between them being that the Biot-Savart law in the vorticity equation (1.2.12) is directed perpendicular to the Coulomb law in the Vlasov-Poisson system (1.2.7). Unlike the Vlasov-Maxwell (1.2.8) and Vlasov-Poisson (1.2.7) equations, we do not present any results on the vortex equation in this thesis.

1.2.4 Systems with noise.

The vorticity equation (1.2.12) is derived from the two dimensional incompressible Euler equation. The two dimensional incompressible Navier-Stokes equation instead produces the following equation in vorticity formulation, which we will refer to as the *viscous vorticity* equation.

$$\begin{cases} \partial_t \omega_t + u_t \cdot \nabla \omega_t - \Delta \omega_t = 0, \\ u_t = \dfrac{x^\perp}{|x|^2} * \omega_t, \end{cases} \tag{1.2.13}$$

where $x^\perp = (-x_2, x_1)$ for $x = (x_1, x_2) \in \mathbb{R}^2$. This is the formal limit of a particle system in the same way as the inviscid case (1.2.12). However, here the particle system is a system of stochastic differential equations (SDEs). Instead of writing the specific system for the viscous vorticity equation we consider the more general

system of first order SDEs for N particles $(X^{i,N})_{i=1}^N \in \mathbb{R}^{Nd}$:

$$dX_t^{i,N} = \frac{1}{N}\sum_{j=1}^{N} K(X_t^{i,N}, X_t^{j,N})\,dt + dB_t^{i,N} \tag{1.2.14}$$

where $(B^{i,N})_{i=1}^N$ are standard independent Brownian motions. Formally, the mean field limit of this system of SDEs is the non-linear convection diffusion equation (which could also be called a non-linear Fokker-Planck equation, or the PDE for the law of a McKean-Vlasov equation)

$$\begin{cases} \partial_t f_t + \nabla \cdot (F_t f_t) - \dfrac{1}{2}\Delta f_t = 0, \\ F_t(x) = \displaystyle\int K(x,y) f_t(y)\,dy \end{cases} \tag{1.2.15}$$

The empirical measure μ^N corresponding to the particle system $(X^{i,N})_{i=1}^N$ is defined by the same formula (1.2.10) as in the deterministic case. However, there is an important distinction: for noisy systems the empirical measure is *random*, so is a random probability measure on $\mathbb{R}^d$.

For second order systems governed by Newton's laws, models often consider noise in the velocities only. Indeed, let $(X^{i,N}, V^{i,N})_{i=1}^N \in \mathbb{R}^{2Nd}$ be the positions and velocities of N particles, then they evolve as

$$\begin{cases} dX_t^{i,N} = V_t^{i,N} dt \\ dV_t^{i,N} = \dfrac{1}{N}\displaystyle\sum_{j=1}^{N} K(X_t^{i,N}, X_t^{j,N}) + dB_t^{i,N} \end{cases} \tag{1.2.16}$$

where again $(B^{i,N})_{i=1}^N$ are standard independent Brownian motions. The formal mean field limit of this particle system is the non-linear kinetic Fokker-Planck equation

$$\begin{cases} \partial_t f_t + v \cdot \nabla_x f_t + \nabla_v \cdot (F_t f_t) - \dfrac{1}{2}\Delta_v f_t = 0, \\ F_t(x) = \displaystyle\int K(x,y) f_t(y,v)\,dvdy, \end{cases} \tag{1.2.17}$$

where the unknown is the phase space density $f_t(x,v)$.

1.2.5 Propagation of chaos

The notion of molecular chaos (1.2.3) was key in the formal derivation of the Boltzmann equation above. One might naively hope that this property would be made exact, that the N-particle distribution function would exactly tensorise, i.e.

$$f^N(t, z_1, z_2, \dots, z_N) = f^{\otimes N}(t, z_1, z_2, \dots, z_N) = f(t, z_1) f(t, z_2) \cdots f(t, z_N). \tag{1.2.18}$$

We could certainly impose that the initial condition $f^N(0, z_1, \dots, z_N)$ has this property, but we cannot impose this at later times as the solution is then given to us by the particle dynamics. We then have to check: are we lucky enough that this tensorisation is propagated? The answer, of course, is no. The collisions cause correlations between the particles and the solution for any positive time is no longer tensorised. As a result we cannot ask for perfect tensorisation.

It was the work of Kac [106] that gave the correct formulation: the tensorisation holds only in the limit $N \to \infty$ when looking at the kth marginal for k fixed in the limit.

Definition 1.2.1 (Chaotic family of measures)**.** *We say that a family $(f^N)_{N=1}^{\infty}$ of laws on $\mathbb{R}^{dN}$ which are symmetric under particle permutations is chaotic if there is a law f on $\mathbb{R}^d$ such that*

$$\lim_{N \to \infty} f^{k,N}(z_1, z_2, \dots, z_k) = f^{\otimes k}(z_1, z_2, \dots, z_k) = f(z_1) f(z_2) \cdots f(z_k)$$

holds for each $k \in \mathbb{N}$ where the limit holds in the weak topology on the space of probability measures.*

Kac proposed this definition for the purpose of the derivation of the homogeneous Boltzmann equation from a random walk on the energy sphere, a stochastic system now known as Kac's process. He showed, by a beautiful combinatorial argument, that, although the tensorisation property (1.2.18) is not propagated by the stochastic process, the property of chaoticity is propagated. This phenomenon is known as the *propagation of chaos* and is the central problem in rigorously deriving mean field models from systems of interacting particles.

Later Sznitman [185] gave an alternate definition based upon the empirical measure, which will be of more use in this thesis. Recall that in general the empirical measure is a random variable, and in particular can be obtained from the distribution function $f^N(z_1, \ldots, z_N)$ by constructing a probability space and random variables $(Z^{i,N})_{i=1}^N$ with law f^N and then defining $\mu^N = \frac{1}{N}\sum_{i=1}^N \delta_{Z^{i,N}}$.

Definition 1.2.2 (Chaotic family of measures (2))**.** *We say that a family $(f^N)_{N=1}^\infty$ of laws on $\mathbb{R}^{dN}$ which are symmetric under particle permutations is chaotic if there is a probability measure f on $\mathbb{R}^d$ such that*

$$\lim_{N\to\infty} \mathbb{E}d(\mu^N, f) = 0$$

where d metrises weak convergence of probability measures.

This definition turns out to be equivalent to the previous definition due to Kac, and this equivalence is *quantitative* in the speed of convergence. This convergence is measured in terms of particular metrics on the space of probability measures. In particular, we define the Wasserstein-p distance:

Definition 1.2.3 (Wasserstein distance)**.** *Let $p \in [1, \infty)$ and μ, ν be probability measures on $\mathbb{R}^d$ with p-th moment finite. Then the Wasserstein-p distance between μ and ν is given by*

$$\mathcal{W}_p^p(\mu, \nu) = \inf\left\{\int |x-y|^p \, d\pi(x,y) : \pi \text{ is a coupling of } \mu \text{ and } \nu\right\}$$

where by a coupling we mean a measure on the product space $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ and ν.

The Wasserstein-p distance metrises weak convergence of probability measures for those measures which have finite p-th moment. In the special case of $p = 1$ it is also known as the Monge-Kantorovich-Wasserstein-(Rubinstein) (MKW) metric and can be defined equivalently by duality as:

Definition 1.2.4 (Monge-Kantorovich-Wasserstein-(Rubinstein) metric)**.** *Let μ and ν be probability measures on $\mathbb{R}^d$ with first moment finite. Then the MKW distance is given by*

$$d_{\mathrm{MKW}}(\mu, \nu) = \sup\left\{\int h \, d\mu - \int h \, d\nu : h \text{ is 1-Lipschitz}\right\}.$$

The Wasserstein distances are the basis for the rich theory of *optimal transportation* (see e.g. [198] for an overview). The Wasserstein-2 distance in particular has a deep geometric interpretation related both to *gradient flows* (see e.g. [9]) and to curvature of metric spaces [198]. These distances are used throughout this thesis.

1.3 Summary of Chapter 2

It is the derivation of (1.2.15) and (1.2.17) in the limit $N \to \infty$ of (1.2.14) and (1.2.16) respectively that is the topic of the second chapter (Chapter 2) of this thesis. In this summary we shall only give a brief overview of the existing literature, preferring to concentrate on the philosophical differences in the various approaches. A more extensive picture is given at the start of Chapter 2. The importance of this problem in kinetic theory and mathematical physics was discussed in the preceding Section 1.2.

1.3.1 Compactness versus quantitative.

It is important to distinguish between *compactness* methods and *quantitative* methods. The former gives convergence without a rate, but has the advantage of being easier to establish. For most convergence results in the field a compactness proof is given first, and only later and with different techniques is a quantitative estimate on the rate established. In many models only a compactness proof with no rate is known, for example in the viscous vorticity model (1.2.13) (see the recent study of Fournier, Hauray and Mischler [66] and the references therein.) As our results in this thesis are quantitative, we shall not spend much time summarising the full extent of compactness methods.

1.3.2 Stochastic versus deterministic.

A second important distinction is between the deterministic dynamics of (1.2.9) and (1.2.4) and the stochastic (noisy) dynamics of (1.2.14) and (1.2.16). De-

terministic systems have two advantages over their noisy counterparts. Firstly, very fine control over particle positions and velocities can be established, which allows conditions on inter-particle distances, for example, to be propagated in time, an approach followed by Hauray and Jabin [81] for Vlasov with singular forces. In noisy systems, close encounters between particles are uncommon overall, but due to noise are also extremely unlikely to never occur, making such propagation estimates impossible. Secondly, the empirical measure is a weak solution of the corresponding limit equation, which fails in the noisy case. This allows the application of weak-strong stability estimates on the limit equation. These make quantitative proofs of propagation of chaos easier than in the noisy case. However, for compactness methods the situation is reversed, with the noise regularising the system and making undesirable events also uncommon, giving compactness.

1.3.3 Quantitative methods for the deterministic case.

The standard baseline quantitative result on deterministic systems is due to Dobrushin [49] and also Braun and Hepp [29], both on the Vlasov equation (1.2.6) with a kernel K that is globally *Lipschitz* in both variables. Their proofs rely on the observation that, when K is Lipschitz, the Vlasov equation (1.2.6) is well-posed in the space of measures equipped with the weak* topology. This allows an explicit computation, now commonly known as the *Dobrushin estimate*, on the dependence of the solution to the Vlasov equation upon its initial datum in the Wasserstein distance. The same estimate applies to the first order system (1.2.11) again assuming that K is Lipschitz. For more singular kernels, results are more limited, a notable result being due to Hauray and Jabin [81].

1.3.4 Quantitative methods for the stochastic case.

For noisy systems (1.2.15) and (1.2.17), the empirical measure is no longer a weak solution to the limit equation (even in the formal sense), and the Dobrushin estimate does not directly apply. Instead, for Lipschitz kernels K one applies a coupling technique popularised by Sznitman in his lecture notes [185], where

the particle system is coupled to a system of N *mutually independent* particles with the vector field of the limit equation. This method has the advantage that the driving Brownian motions effectively 'disappear' from the mathematical estimates, and a Dobrushin style proof is then possible.

For non-Lipschitz kernels very few quantitative results exist in the literature. We mention the work of Fournier and Hauray [65] on a particle model for the Landau equation where the kernel $K(x, y)$ has a singularity of strength $|x - y|^\alpha$ with $\alpha \in (0, 1)$ on the diagonal, and is Lipschitz elsewhere, and also the recent results of Jabin and Wang [100, 101] for kernels $K(x, y) = W(x - y)$ which are, respectively, L^∞ or, (roughly speaking) $W^{-1,\infty}$. However, these results require specially prepared initial conditions that are uniformly bounded above and below by exponentials or Gaussians.

1.3.5 The results of Chapter 2

The main results of Chapter 2 are quantitative propagation of chaos for the noisy systems (1.2.14) and (1.2.16) for interaction kernels that are merely *Hölder continuous*. In particular, the results apply to kernels that are *nowhere Lipschitz* and to general initial data. We give an informal statement of these results below, leaving the exact formulation and extensions to the chapter itself.

Theorem 1.3.1. *Let the initial condition f_0 be in L^2 with some decay at infinity. Let $[0, T]$ be an arbitrary finite time interval. Consider the particle systems (1.2.14) and (1.2.16) with initial positions i.i.d. with law f_0. Let f_t be the corresponding solution to the respective limit equations (1.2.15) and (1.2.17). Assume:*

- First order system: *Let $K \in C^{0,\alpha}(\mathbb{R}^d \times \mathbb{R}^d; \mathbb{R}^d)$ with $\alpha \in (0, 1)$.*
- Second order system: *Let $K \in C^{0,\alpha}(\mathbb{R}^d \times \mathbb{R}^d; \mathbb{R}^d)$ with $\alpha \in (1/3, 1)$.*

Then

$$\mathbb{E} \sup_{t\in[0,T]} d_{MKW}(\mu_t^N, f_t) \leq CN^{-\gamma} \tag{1.3.1}$$

for an explicit constant $\gamma > 0$ depending only on the dimension d and upon α.

The proof brings together ideas from *empirical process theory* (see e.g. [193]) and

stochastic flows for SDEs with rough coefficients ([116] describes the smooth case in detail).

The main hindrance to proving propagation of chaos in systems such as (1.2.4), (1.2.9), (1.2.14) and (1.2.16) is the non-independence of the particles as this obstructs an application of the law of large numbers. Although the particles will have the same individual laws they interact through the force field that depends upon all of the particles. Every proof of propagation of chaos must deal with this problem at some key point. For example, the coupling method of Sznitman [185] uses the Lipschitz assumption on K to view the particle system as a perturbation of N independent particles to which the law of large numbers directly applies.

To prove Theorem 1.3.1 we take a different and novel approach based upon proving a *uniform law of large numbers*. We now describe this in the first order case. We define the *empirical process* $((X^{b,i,N})_{i=1}^N)_{b\in\mathcal{C}}$ as the solutions to the SDE

$$\begin{cases} dX_t^{b,i,N} = b_t(X_t^{b,i,N})dt + dB_t^{i,N} \\ X_0^{i,N} \text{ i.i.d. with law } f_0 \end{cases}$$

where b ranges over a set $\mathcal{C}$ of all Hölder continuous vector fields with norm bounded by a constant, i.e.

$$\mathcal{C} = \{b \in C^{0,\alpha}([0,T]\times\mathbb{R}^d;\mathbb{R}^d) : \|b\|_{C^{0,\alpha}} \leq C\}. \tag{1.3.2}$$

Note that for each $b \in \mathcal{C}$ fixed, the N particles $(X^{i,N})_{i=1}^N$ are independent and identically distributed. Let $\mu_t^{b,N}$ be the empirical measure corresponding to $(X^{b,i,N})_{i=1}^N$. Let f_t^b be the law of $X_t^{b,1,N}$ for each b fixed. The key observation is that, if we choose

$$b_t^N(x) = \frac{1}{N}\sum_{i=1}^N K(x, X_t^{i,N}) \tag{1.3.3}$$

with $X_t^{i,N}$ given by the original particle system (1.2.14), then, at least formally,

$$(X_t^{b^N,i,N})_{i=1}^N = (X_t^{i,N})_{i=1}^N. \tag{1.3.4}$$

We can avoid dealing with the original dependent particle system (1.2.14) by the

inequality

$$\begin{aligned} d_{\mathrm{MKW}}(\mu_t^N, f_t) &\leq d_{\mathrm{MKW}}(\mu_t^{b^N,N}, f_t^{b^N}) + d_{\mathrm{MKW}}(f_t^{b^N}, f_t) \\ &\leq \sup_{b\in\mathcal{C}} d_{\mathrm{MKW}}(\mu_t^{b,N}, f_t^b) + d_{\mathrm{MKW}}(f_t^{b^N}, f_t). \end{aligned}$$

The key being that the only appearance of b^N on the final line is as the vector field in a *PDE* rather than an *SDE*, which is a much smoother object. The remaining problem is to estimate the supremum:

Theorem 1.3.2. *Let the assumptions of Theorem 1.3.1 hold. Then*

$$\mathbb{E}\sup_{b\in\mathcal{C}} d_{\mathrm{MKW}}(\mu_t^{b,N}, f_t^b) \leq CN^{-\gamma'}$$

for an explicit $\gamma' > 0$ depending only upon the dimension d and upon α.

Remark 1.3.1. *In stating the above theorem we have swept under the rug a major technical issue of defining the empirical process $(\mu^{b,N})_{b\in\mathcal{C}}$ in such a way that the supremum makes sense. In particular, the statement above holds only in a formal sense, assuming that everything can be defined. The resolution of this issue is left to Chapter 2 itself.*

1.3.6 Significance of the results

The more application minded reader may ask, given that most models in the physical and biological sciences have a kernel of the form $K(x, y) = W(x - y)$ with W smooth apart from a singularity at the origin, whether it is of interest to consider kernels that are nowhere smooth. This question has a simple response: Theorem 1.3.1 should not be considered a result about *singular* kernels as such, rather it should be considered an improvement in what should be thought of as a singular kernel. The baseline regularity below which a kernel is considered singular has previously been taken to be Lipschitz; Theorem 1.3.1 shows that this can be reduced to Hölder continuous for noisy systems.

1.4 Arrow-Hurwicz-Uzawa Gradient method

In the previous Section 1.3 we discussed the modelling and study of systems comprised of many individual interacting components via mean field equations and described their rigorous derivation. In this section we will present another example of a system comprised of many interacting individual elements. However, in this case mean field approximation will not be appropriate and we will instead work with the high dimensional dynamics directly. This will lead to the study of the Arrow-Hurwicz-Uzawa *gradient method*, which was introduced by the eponymous authors [10] as a numerical method of finding saddle points of concave-convex functions.

Given a function $\varphi(x, y) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ which is concave-convex (concave in x for fixed y, convex in y for fixed x), the *gradient method* is the system of ordinary differential equations for $\mathbf{z}(t) = (x(t), y(t)) \in \mathbb{R}^{n+m}$ given by

$$\begin{cases} \dfrac{dx_i}{dt} = \dfrac{\partial \varphi}{\partial x_i}, & i = 1, 2, \dots, n, \\ \dfrac{dy_j}{dt} = -\dfrac{\partial \varphi}{\partial y_j}, & j = 1, 2, \dots, m. \end{cases} \tag{1.4.1}$$

The problem of finding saddle points of a concave-convex function φ is a central problem in numerical analysis and optimisation (see e.g. [27, 61] among many others). In particular the dual problem of a convex optimisation problem is to find a saddle point of the associated concave-convex Lagrangian (see e.g. [27]). This link between the gradient method and optimisation problems has been exploited [110, 61] in the control of many distributed systems, such as wireless networks, multi-path routing [183, 199, 124], power systems, microeconomics [194] and in a general framework for distributed optimisation over networks.

In Section 1.5 and Section 1.6 we shall describe the results on the asymptotic behaviour of this system and the related *subgradient method*, where the dynamics (1.4.1) are restricted to a convex domain, that are proved in this thesis. First, however, we will give more details of the network optimisation problems for which the gradient method is applied, and show how these can be interpreted as many particle systems.

1.4.1 Motivation: Network Utility Maximisation (NUM)

Consider the simple concave optimisation problem (see e.g. [27])

$$\max_{x\in\mathbb{R}^n} U(x) \tag{1.4.2}$$

for a concave function $U : \mathbb{R}^n \to \mathbb{R}$.

Network Utility Maximisation (NUM) is concerned with the specific case where the utility function U is distributed as a sum over the nodes of a network. We present here a simple example of such a problem and show how the gradient dynamics (1.4.1) can be applied to yield a distributed algorithm for its solution. For more general examples and a more detailed exposition of the theory we encourage the reader to consult [28, 164, 183, 110]. Another more specific example, the multi-path routing problem, is discussed in Chapter 4 as an application of the results therein.

For our purposes a network is just an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with $\mathcal{V} = \{1, 2, \ldots, n\}$. Each node (vertex) $i \in \mathcal{V}$ is associated with a variable $x_i \in \mathbb{R}$. We assume that the optimisation problem (1.4.2) splits into a sum over the graph as follows:

$$\max_{x\in\mathbb{R}^n} \sum_{i\in\mathcal{V}} U_i(x_i, x_{n_{i,1}}, \ldots, x_{n_{i,d(i)}}). \tag{1.4.3}$$

Here, for a vertex $i \in \mathcal{V}$, $d(i)$ is its *degree* (number of edges), and $(n_{i,j})_{j=1}^{d(i)}$ are its *neighbours* (vertices which share an edge with i). Each U_i is a concave real-valued function that only depends upon the variable x_i associated with the vertex i, and the variables associated with its *neighbouring vertices*.

If each U_i only depended upon x_i and not its neighbours then the problem (1.4.3) would tensorise completely over the graph and could be solved fully locally, (in a *distributed* manner), by computing, for each $i \in \mathcal{V}$, the solution to the optimisation problem

$$\max_{x_i\in\mathbb{R}} U_i(x_i).$$

One could imagine that each node is capable of this computation on its own. However in the general case this is not possible and in any algorithm there must be communication between the nodes.

To transform the optimisation problem (1.4.3) into a form more amenable to distributed computation, we apply the technique of *dual decomposition* (see e.g. [28]). For each vertex i, we add the additional variables $x_{i,1}, \ldots, x_{i,d(i)}$ which will hold local (to i) copies of the variables $x_{n_{i,1}}, \ldots, x_{n_{i,d(i)}}$, respectively, of the neighbouring vertices. Then the *relaxed* problem

$$\max_{x_i, x_{i,j}} \sum_{i \in \mathcal{V}} U_i(x_i, x_{i,1}, \ldots, x_{i,d(i)})$$

is fully distributed and can be solved locally at each vertex. Of course, this is a relaxation too far and this problem is not equivalent to (1.4.3). To restore this missing correspondence we must add the constraint that, for each vertex $i \in \mathcal{V}$, all the local copies of the variable x_i are equal. This yields the constrained optimisation problem

$$\max_{\substack{x_i, x_{j,i} \\ x_i = x_{j,i} \text{ if } (i,j) \in \mathcal{E}}} \sum_{i \in \mathcal{V}} U_i(x_i, x_{i,1}, \ldots, x_{i,d(i)}) \tag{1.4.4}$$

which is equivalent to (1.4.3).

A simple approach to constructing a distributed algorithm for solving (1.4.4) is to relax the constraints with Lagrange multipliers, and solve the corresponding *saddle point problem* for the resulting Lagrangian:

$$\varphi(x, y) = \sum_{i \in \mathcal{V}} U_i(x_i, x_{i,1}, \ldots, x_{i,d(i)}) + \sum_{e_k = (i,j) \in \mathcal{E}} y_k (x_i - x_{j,i}) \tag{1.4.5}$$

where we have abused notation and let x denote the full vector of the x_i's and the $x'_{i,j}s$, the second sum is over *directed edges* and $y \in \mathbb{R}^{2|\mathcal{E}|}$ are Lagrange multipliers. Finding a saddle point of (1.4.5) is equivalent[1] to solving (1.4.4) and hence (1.4.3).

A key observation of the celebrated work [110] of Kelly, Maulloo and Tan is that many of the then existing distributed methods for solving the NUM problem can be interpreted as applications of the gradient method to the associated Lagrangian. Indeed, for the example (1.4.5), when we write the gradient dynamics

[1] In general this would require a feasibility condition on the problem (1.4.3) (see e.g. [27]). However, in this case the problem is so simple that these conditions are automatically satisfied.

(1.4.1) we obtain:

$$\begin{cases} \dfrac{dx_i}{dt} = \dfrac{\partial U_i}{\partial x_i}(x_i, x_{i,1}, \dots, x_{i,d(i)}) + \displaystyle\sum_{e_k=(i,j)\in\mathcal{E}} y_k, & i \in \mathcal{V}, \\ \dfrac{dx_{i,j}}{dt} = \dfrac{\partial U_i}{\partial x_{i,j}}(x_i, x_{i,1}, \dots, x_{i,d(i)}) - \displaystyle\sum_{e_k=(i',j)\in\mathcal{E}} y_k, & i \in \mathcal{V}, (i,j) \in \mathcal{E}, \\ \dfrac{dy_k}{dt} = -x_i + x_{j,i}, & k = 1, \dots, 2|\mathcal{E}|, \end{cases} \tag{1.4.6}$$

where the sums are again over directed edges and $e_k = (i, j)$ in the equation for y_k. These dynamics are localised in the sense that, for each vertex $i \in \mathcal{V}$, to update the local variables $x_i, x_{i,1}, \dots, x_{i,d(i)}$ of the vertex, we only need know these local variables themselves, and the Lagrange multipliers corresponding to edges touching i.

The system (1.4.6) is an example of a many particle system. However, unlike the systems of interacting particles described in Section 1.2, the 'particles' (nodes in the graph) are not interchangeable because of the geometry of the graph $\mathcal{G}$. Moreover, the structure of the graphs arising in applications vary considerably and it is not (in general) reasonable to model them as *random graphs*, which is one way in which exchangability could be recovered. For these reasons, among others, the analysis of the system (1.4.6) is done directly without passing to a scaling limit[2].

1.5 Summary of Chapter 3

Chapter 3 studies the asymptotic behaviour and limiting solutions of the gradient method (1.4.1) applied to arbitrary concave-convex function in C^2.

1.5.1 Previous results

When either the concavity or the convexity of φ is *strict*, convergence of the gradient method (1.4.1) to a saddle point was proved by Arrow, Hurwicz and Usawa in

[2]It should be noted, however, that the many particle dynamics may have been derived *from* a scaling limit, for example from the *fluid limit* of a Markov chain.

[10]. Without the strictness assumption, they observed the presence of oscillatory behaviour and non-convergence. However, in many important applications [61], strictness is lacking, but still convergence is observed and proved. This includes the large class of augmented Lagrangian methods, such as Penalty functions and the alternating direction method of multipliers [68, 27].

The majority of such convergence proofs follow the same formula. First it is observed that, for any saddle point $\bar{\mathbf{z}} = (\bar{x}, \bar{y})$, the Euclidean distance

$$W(t) = \frac{1}{2}|\mathbf{z}(t) - \bar{\mathbf{z}}|^2$$

is non-increasing along trajectories of the gradient method. As a consequence of an application of LaSalle's theorem[3] we obtain that any solution converges to the set of solutions that lie at a constant distance from any saddle point. One then studies these solutions and shows that the only possible such solution is a saddle point itself, thus proving convergence.

Despite the simple form and wide applicability of the gradient dynamics (1.4.1), a complete classification of the limiting behaviour is absent from the control literature. It is this gap that Chapter 3 aims to fill. The main result is that the limiting solutions of the gradient method solve an explicit linear ODE, which we state a shortened version of below. We state the result for when $\mathbf{0}$ is a saddle point, the general result being then obtained by a translation of coordinates.

Theorem 1.5.1. *Let $\varphi \in C^2$ be an arbitrary concave-convex and $\mathbf{0}$ be a saddle point of φ. Then the limiting solutions of the gradient method are exactly the solutions to the linear ODE*

$$\dot{\mathbf{z}}(t) = \begin{bmatrix} 0 & \varphi_{xy}(\mathbf{0}) \\ -\varphi_{yx}(\mathbf{0}) & \end{bmatrix} \mathbf{z}(t) \tag{1.5.1}$$

for an explicit set of initial conditions.

Here $\varphi_{xy}, \varphi_{yx}$ are the matrices of partial derivatives of φ.

In addition to this result we also state and prove several corollaries, which, how-

[3]The application of LaSalle's theorem is justified here as we study the gradient method whose trajectories are smooth. For the subgradient method this is no longer the case and a more careful analysis is needed.

ever, we will leave to the chapter itself.

The proofs of Theorem 1.5.1 are geometric in nature and embarrassingly(!) elementary. The key observation is that not only is the Euclidean distance from any saddle point non-increasing along the dynamics of the gradient method, but also the Euclidean distance between any two solutions is itself non-increasing, i.e.

$$\frac{d}{dt}|\mathbf{z}(t) - \mathbf{z}'(t)|^2 \leq 0$$

for any two solutions $\mathbf{z}(t)$ and $\mathbf{z}'(t)$ of the gradient method (1.4.1). A property which we refer to in the chapter (and the following chapter) as *pathwise stability*.[4] With this observation in hand, one can then deduce geometric properties of the set of limiting solutions culminating in showing that they solve the explicit linear ODE given above.

1.6 The subgradient method

In many applications it is required that the dynamics of the gradient method (1.4.1) are restricted to lie in a closed convex set $K \subseteq \mathbb{R}^{n+m}$. For example, in applications to optimisation, this is needed to ensure that the Lagrangian multipliers remain non-negative in the relaxation of inequality constraints. This is achieved by projecting the dynamics onto K and the resulting system is known as the *subgradient method*.[5]

A major issue in the study of the subgradient method is that it is a non-smooth differential equation, i.e. its trajectories are not differentiable. This means that the classical LaSalle type theorems (e.g. [112]) do not apply, as noted in [38] where tools from non-smooth analysis are used. We do however note that, if K is an affine subspace, then the resulting subgradient dynamics are smooth.

[4] We use the term pathwise stable, rather than *incrementally stable*, *contractive* or *monotone* as these later terms have different (conflicting!) definitions in the different mathematics and engineering communities.

[5] For brevity we do not write the subgradient method explicitly in this introduction, as doing so would require too many definitions. In the special case of positive orthant constraints the subgradient method is given by (1.7.1) below.

1.7 Summary of Chapter 4

In Chapter 4 we extend the results of Chapter 3 to the subgradient method on a convex set K. The main result is that the limiting solutions of the non-smooth subgradient method solve an explicit smooth system of differential equations. We state a simplified version of this result below.

Theorem 1.7.1. *Let $K \subseteq \mathbb{R}^{n+m}$ be non-empty closed and convex. Let $\varphi : K \to \mathbb{R}$ be C^2 and concave-convex. Then the ω-limit set of the subgradient method on K applied to φ is contained in the unique minimal face F of K that contains all the saddle points of φ. Furthermore, the limiting solutions are solutions to the (smooth) subgradient method on V where V is the affine span of the face F.*

As a consequence of this result, the study of the limiting solutions of the non-smooth subgradient method reduces to the study of the limiting solutions of a smooth system, the subgradient method on an affine subspace. In particular, the results in Chapter 3 apply to this system and give an exact classification of limiting behaviour.

To illustrate Theorem 1.7.1 consider the special case of positive orthant constraints. For concreteness the positive orthant is given by

$$K = \mathbb{R}^n_+ \times \mathbb{R}^m_+ = \{(x, y) \in \mathbb{R}^{n+m} : x_i \geq 0, i = 1, \ldots, n, \text{ and } y_j \geq 0, j = 1, \ldots, m\}.$$

In this case the subgradient method has the simple form

$$\begin{cases} \dfrac{dx_i}{dt} = \left[\dfrac{\partial \varphi}{\partial x_i} \right]^+_{x_i} & i = 1, 2, \ldots, n, \\ \dfrac{dy_j}{dt} = \left[-\dfrac{\partial \varphi}{\partial y_j} \right]^+_{y_j} & j = 1, 2, \ldots, m, \end{cases} \tag{1.7.1}$$

where we have used the notation

$$[a]^+_b = \begin{cases} a & \text{if } b > 0 \text{ or } a > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Here the projection operators $[\cdot]^+$ act to ensure that the dynamics preserve the

coordinatewise non-negativity of the solution $\mathbf{z}(t) = (x(t), y(t))$. Such constraints arise naturally in many applications, for example in ensuring the non-negativity of Lagrange multipliers or transmission rates.

In this case, the faces correspond to the fixing some subset of coordinates x_i, y_j to zero, and keeping the rest non-negative, i.e. each face is of the form

$$F = \{(x, y) \in \mathbb{R}^n_+ \times \mathbb{R}^m_+ : x_i = 0 \text{ for } i \in I, y_j = 0 \text{ for } j \in J\}$$

where $I \subseteq \{1, \dots, n\}$ and $J \subseteq \{1, \dots, m\}$ are arbitrary subsets of coordinates.

The above Theorem 1.7.1 then implies that the subgradient method is convergent to a saddle point, if the gradient method obtained from the function $\varphi(x, y)$ by fixing an arbitrary subset of the coordinates to zero is convergent.

The proof of Theorem 1.7.1 on the subgradient method is based upon the same observation as the proof of Theorem 1.5.1 for the gradient method, that the Euclidean distance between any two solutions is non-increasing in time. However, due to the non-smoothness of the dynamics, we take a more abstract view, using tools from topological dynamics. This results in a proof that is no longer as elementary as that for Theorem 1.5.1.

1.8 Instabilities of the Vlasov-Maxwell system

Chapters 5 and 6 are devoted to the study of instabilities of the relativistic Vlasov-Maxwell system (1.2.8). We summarise them below.

1.8.1 The (in)stability problem

The *stability* or *instability* of stationary solutions to (1.2.8) is a classical problem in mathematical physics (see e.g. [114]). An early result on this problem is the linear stability criterion of Penrose [167] for spatially homogeneous equilibria. Later results include [96, 139] among many others. In [130, 132], Lin and Strauss established a sharp linear stability for *monotone* (defined in the next paragraph)

equilibria, and obtained non-linear stability results in [131]. Their result associates to each monotone equilibrium of (1.2.8) a Schrödinger operator $\mathcal{L}^0$ acting in the spatial variable only, and states that the equilibrium is linearly stable if $\mathcal{L}^0 \geq 0$ (i.e. has no negative eigenvalues), and is linearly unstable if $\mathcal{L}^0$ possesses a negative eigenvalue.

1.8.2 Monotonicity of equilibria

Typically one assumes (justified by the so-called 'Jean's theorem'[6]) that the equilibrium can be written in the form $f^{0,\pm}(x,v) = \mu^{\pm}(e^{\pm}, p^{\pm})$, i.e. as a function of the microscopic energy $e^{\pm}$ and momentum $p^{\pm}$ which are conserved along solutions to (1.2.8).[7] Such an equilibrium is called *monotone* if

$$\frac{\partial \mu^{\pm}}{\partial e^{\pm}} < 0 \quad \text{whenever } \mu^{\pm} > 0. \tag{1.8.1}$$

Equilibria that do not satisfy this condition are called *non-monotone*. The assumption of monotonicity is very often made in the study of (in)stability of the Vlasov Poisson (1.2.7) and Vlasov Maxwell (1.2.8) systems, as it is believed to make equilibria more stable. However, there are many interesting examples of non-monotone equilibria, both stable and unstable, and the mathematical theory in these cases is far more sparse. A notable exception being the work of Penrose [167] for homogeneous equilibria.

1.9 Summary of Chapters 5 and 6

In Chapter 6 we extend the linear *instability* result of Lin and Strauss [132] to non-monotone equilibria. We give a simplified statement below.

Theorem 1.9.1. *Let $f^{0,\pm}$ be an equilibrium of the Vlasov-Maxwell system in either 1.5D symmetry or 3D with cylindrical symmetry. Assume that $f^{0,\pm}(x,v) =$*

[6] Which is not strictly a 'theorem' as it is not always true (see Remark 6.1.1). It is, however, always convenient.

[7] The exact form of the conserved quantities $e^{\pm}$ and $p^{\pm}$ depends upon the symmetries of the considered equilibria, so we will not record them at this point.

$\mu^{\pm}(e^{\pm}, p^{\pm})$ is C^1 with compact support in the x variable. Then there exists a Schrödinger operator[8] *$\mathcal{L}^0$ acting only in the x variable, such that if $\mathcal{L}^0$ has a negative eigenvalue then there exists a growing mode solution*

$$(e^{\lambda t} f^{\pm}(x, v), e^{\lambda t} E(x), e^{\lambda t} B(x)), \quad \lambda > 0$$

to the RVM system (1.2.8) linearised around the $f^{0,\pm}$.

This builds on the work of Ben-Artzi [16, 17] who studied the stability question for non-monotone equilibria in the simpler 1.5D periodic setting. By taking the Laplace transform and inverting the linearised Vlasov equation, the problem of linear instability is converted in the problem of finding a growth rate $\lambda > 0$ for which a self-adjoint operator $\mathcal{M}^{\lambda}$ has a non-trivial kernel. We wish to *count* the number of negative eigenvalues of $\mathcal{M}^{\lambda}$ for λ large and small. If they are different, then the continuity in λ of the spectrum ensures that at some intermediate λ an eigenvalue of $\mathcal{M}^{\lambda}$ must cross zero and give a non-trivial kernel. However, the operator $\mathcal{M}^{\lambda}$ comes from Maxwell's equations and has the schematic form

$$\mathcal{M}^{\lambda} = \mathcal{A} + \mathcal{K}^{\lambda} = \begin{bmatrix} -\Delta + 1 & 0 \\ 0 & \Delta - 1 \end{bmatrix} + \mathcal{K}^{\lambda}, \tag{1.9.1}$$

where $\mathcal{K}^{\lambda}$ is a symmetric bounded arising from computing the current and charge densities $j(t, x)$ and $\rho(t, x)$ from linearised Vlasov equation. Due to the opposite signs of the Laplacians and the unbounded domain, the operator $\mathcal{M}^{\lambda}$ has essential spectrum extending both to positive and negative infinity. This means that eigenvalues may be absorbed or emitted from the essential spectrum, and the counting argument cannot work. To get around this problem we use the results of Chapter 5 (described in the next paragraph), to show it is sufficient to apply the counting argument to finite dimensional approximations to $\mathcal{M}^{\lambda}$.

Summary of Chapter 5 In Chapter 5 we consider the problem of approximating the discrete spectra of families of self-adjoint operators that are merely strongly continuous. We look at families $\left\{\mathcal{M}^{\lambda}\right\}_{\lambda \in [0,1]}$ of the form (1.9.1) where

[8] In Chapter 6 the theorem is stated in a different way without explicit reference to the operator $\mathcal{L}^0$. We differ here as this version is slightly shorter (if less precise).

$\left\{\mathcal{K}^\lambda\right\}_{\lambda\in[0,1]}$ is a bounded, symmetric and strongly continuous family. Such families have, for every $\lambda \in [0,1]$, discrete spectrum inside $(-1,1)$ and it is these eigenvalues we wish to approximate. We show that, under uniform relative compactness assumptions on the strongly continuous bounded perturbation $\mathcal{K}^\lambda$, this is possible and give an explicit construction of such finite dimensional approximations, in such a way that the spectrum of the approximations converges in compact subsets of $(-1,1)$ to the spectrum of $\mathcal{M}^\lambda$, uniformly in the parameter λ.

1.10 Uniqueness for the Vlasov-Poisson system

The Cauchy problem for the Vlasov-Poisson system (1.2.7) has received considerable attention in the literature, with classical solutions constructed in [192, 11, 168, 178] and weak solutions in [133, 71] (see also the discussion in Pages 23 to 24).

In [143] Miot established the uniqueness of weak solutions to Vlasov-Poisson under the assumption that the L^p norms of the spatial density ρ grow at most linearly in p, i.e. that

$$\sup_{t\in[0,T]} \sup_{p\geq 1} \frac{1}{p} \|\rho(t)\|_{L^p(\mathbb{R}^d)} \leq C < \infty. \tag{1.10.1}$$

This generalises the uniqueness established by Loeper in [134] where ρ is assumed to obey the stronger bound

$$\sup_{t\in[0,T]} \|\rho(t)\|_{L^\infty(\mathbb{R}^d)} \leq C < \infty. \tag{1.10.2}$$

The argument in [134] uses a log-Lipschitz Grönwall estimate, which, although this is not done in the paper[9], would give a stability estimate of the form

$$\mathcal{W}_2(f_2(t), f_2(t)) \leq C\mathcal{W}_2(f_1(0), f_2(0))^{\exp(-ct)} \tag{1.10.3}$$

in the Wasserstein-2 distance (see Definition 1.2.3) for any two solutions f_1, f_2

[9] Such estimates are often not done explicitly in the literature as they are usually easily deduced from uniqueness proof. Note however, that this is not the case in [143].

of the Vlasov-Poisson system (1.2.7) each obeying the regularity bound (1.10.2). This estimate is valid up to the time when the right hand side reaches some macroscopic size (1/9 for example). This bound grows (for large t) like an exponential tower $e^{e^{ct}}$. The argument of Miot in [143], however, does not establish continuous dependence upon initial data.

1.11 Summary of Chapter 7

In Chapter 7 we clarify the results of Loeper [134] and Miot [143] by interpolating the function spaces (1.10.2) and (1.10.1) in the framework of exponential Orlicz spaces (see e.g. [170]). We establish continuous dependence estimates (similar to (1.10.3)) which hold when the spatial density is assumed to belong to these spaces.

In particular, we improve the result of [143] by establishing continuous dependence with a bound that grows like the exponential tower $e^{e^{ct}}$, and we improve the growth bound deduced from [134] down to a stretched exponential e^{ct^2}.

The proofs have two ingredients. Firstly, we establish an estimate on the Newton kernel (K in (1.2.7)) when convolved with functions in an exponential Orlicz space. Secondly, we exploit that the Vlasov-Poisson system originates from second order Newtonian mechanics (also exploited in [143]). Schematically, the Vlasov-Poisson system acts like a second order ODE

$$\ddot{X} = F(X)$$

as the forces depend only on the particle positions and not their velocities. The key observation is that second order differential inequalities

$$\ddot{X} \leq \varphi(X), \qquad X \geq 0, \quad \dot{X} \geq 0$$

can be closed when φ is only $\log^2$-Lipschitz, (first order inequalities can only[10] be closed for φ log-Lipschitz), and give better growth bounds than first order differential inequalities.

[10] This can be extended slightly to functions like $x|\log(x)|\log|\log(x)|$, etc.

1.12 Asymptotic behaviour of solutes in a fluid background

The mathematical modelling of particles suspended in a fluid background is a topic of great importance in both fluid and statistical mechanics. On a microscopic level, the rapid oscillatory motion of particles suspended in water, reported in 1828 by Brown [30] and eponymously named Brownian motion, later led to the development of the first stochastic models of particle motion successively by Einstein, Smoluchowski and Langevin [55, 200, 121]. On the macroscopic level, understanding the spreading of a passively transported solute in a fluid is important both in applications, for example in understanding efficient methods for mixing fluids, but also experimentally, where observations of fluids are taken by observing passively transported dyes or tracer particles (see e.g. [190]).

1.12.1 The Langevin equation

Langevin, building upon the prior work of Einstein [55] and Smoluchowski [200], developed the first dynamical theory of Brownian motion [121]. He theorised that a suspended particle with position X_t and velocity V_t satisfied Newton's Laws of motion with a force F_t due to collisions with surrounding smaller fluid particles. Thus,

$$\begin{cases} \dfrac{dX_t}{dt} = V_t, \\ \dfrac{dV_t}{dt} = F_t. \end{cases} \tag{1.12.1}$$

The force F is split into two parts. The first kind of collision is due to the modelled particle, having velocity V_t, displacing fluid particles due to its motion. This is a frictional force $-\lambda V_t$. The second part of the force ξ_t is due to random density fluctuations in the fluid background. Hence,

$$\frac{dV_t}{dt} = -\lambda V_t + \xi_t.$$

Due to the separation of time scales between the slowly moving tracked particle and the more rapidly moving fluid background, the fluctuation force ξ has un-

correlated time marginals, i.e. ξ_t and ξ_s are independent for $t \neq s$. Furthermore, being due to fluctuations they are on average zero. In modern language $d\xi_t$ is a brownian motion (meaning Wiener process, rather than the Brownian motion due to Brown), and the pair X, V solve the SDE

$$\begin{cases} dX_t = V_t, \\ dV_t = -\lambda V_t + dB_t, \end{cases} \tag{1.12.2}$$

where B is a standard brownian motion. The forward Kolmogorov (Fokker-Planck) equation for the evolution of the probability law of the SDE (1.12.2) is the kinetic Fokker-Planck equation

$$\partial_t f(t,x,v) + v \cdot \nabla_x f(t,x,v) = \nabla_v \cdot (\lambda v f(t,x,v) + \frac{1}{2}\nabla_v f(t,x,v)) \tag{1.12.3}$$

where the unknown is a probability density function f.

1.13 Summary of Chapter 8

In Chapter 8 we study the mixing properties of the dynamics (1.12.2) in a periodic spatial domain. In particular, we ask the question:

Question 1.13.1. *Are the dynamics (1.12.2) contractive in the Wasserstein-2 distance? More explicitly, do there exist constants $c, \gamma > 0$ such that*

$$\mathcal{W}_2(\mu_t, \nu_t) \leq ce^{-\gamma t}\mathcal{W}_2(\nu_0, \mu_0) \tag{1.13.1}$$

where μ_t, ν_t are the laws of any two solutions to the SDE (1.12.2) for $(X, V) \in \mathbb{T} \times \mathbb{R}$?

Such questions of convergence to equilibrium and contraction (in various metrics) of the kinetic Fokker-Planck equation have been a central object of study in statistical mechanics, and are approached using both analytic and probabilistic methods.

From an analytic perspective, one typically works in a L^2 space weighted by the inverse of the equilibrium measure. In the spatially homogeneous setting, where

only velocities are considered, contractivity is established by showing that the generator is *coercive* in this L^2 space, which implies contractivity of the corresponding semi-group. In the inhomogeneous setting the generator is no longer coercive, leading Villani to develop the theory of *hypocoercivity* [197], where 'skew' norms are constructed for which the generator *is* coercive. Despite this success, analytic methods have trouble accessing the probabilistic definition of the Wasserstein-2 distance, with a notable exception of the case where the dynamics are a $\mathcal{W}_2$ gradient flow, which holds in the spatially homogeneous Fokker-Planck equation [105], but does not hold in the inhomogeneous setting considered here.

From a probabilistic point of view a common approach is to construct a *coupling* between two stochastic processes that realises the desired bound in the metric between the laws. In the spatially homogeneous setting, and in the case of a spatial confining potential that is sufficiently well behaved, this is successful and has been used to establish contraction in the Wasserstein-2 distance [25, 24].

Having a *periodic spatial domain* simplifies the analytic theory. However, for coupling techniques, which work in the case of a 'nice' spatially confining potential, no such coupling result in $\mathcal{W}_2$ has ever been established, which, considering the simplicity of the problem is somewhat of a puzzle. In Chapter 8 we present a result that explains this failure:

Theorem 1.13.1. *There exists no 'nice' coupling of stochastic processes that verifies the contraction property* (1.13.1)*.*

Here 'nice' roughly means adapted to the same filtration. However, we do construct an explicit coupling that achieves convergence:

Theorem 1.13.2. *For any initial data μ_0, ν_0, there exists a 'nice' coupling that verifies the following bound*

$$\mathcal{W}_2(\mu_t, \nu_t) \leq ce^{-\gamma t}(\sqrt{\mathcal{W}_2(\mu_0, \nu_0)} + \mathcal{W}_2(\mu_0, \nu_0)). \tag{1.13.2}$$

In fact this dependence on the square root (which is much larger when the distance is small) is shown to be optimal for such couplings. It is however *not* optimal in the sense that the semi-group *is* a contraction.

Theorem 1.13.3. *Question 1.13.1 is answered in the affirmative, in that (1.13.1) holds.*

But, due to the previous results, any such coupling that attains this bound cannot be a 'nice' coupling of two stochastic processes.

1.14 Macroscopic transport of tracers

A common approximation to the Langevin dynamics (1.12.2) is to consider the overdamped limit where the time scale of velocity changes are taken to be much faster than the time scale of position changes. In such a regime, (1.12.2) becomes instead the equation

$$dX_t = cdB_t.$$

However, this is missing a large part of the picture, as we have neglected the macroscopic motion of the fluid itself and spatial variation of the molecular diffusion coefficient. When one takes these effects into account and passes to the corresponding PDE for the density one obtains

$$\partial_t u(t,x) + b(x) \cdot \nabla u(t,x) + \nabla \cdot (D(x) \nabla u(t,x)) = 0 \tag{1.14.1}$$

where $u : [0,\infty) \times \mathbb{R}^d \to \mathbb{R}$ is the concentration of the solute, b is the velocity field of the fluid (assumed to be incompressible) and D is the molecular diffusion matrix.

1.15 Summary of Chapter 9

In Chapter 9 we study the dispersion of solutes in the presence of *strong convection* and *rapidly oscillating* coefficients. This corresponds to the parabolic problem, a version of (1.14.1),

$$\partial_t u^\varepsilon(t,x) + \frac{1}{\varepsilon} b\left(x, \frac{x}{\varepsilon}\right) \cdot \nabla u^\varepsilon(t,x) - \nabla \cdot \left(D\left(x, \frac{x}{\varepsilon}\right) \nabla u^\varepsilon(t,x) \right) = 0 \tag{1.15.1}$$

for a small parameter $\epsilon \ll 1$. Such problems arise from a rescaling $t, x \to t\epsilon, x\epsilon$ of (1.14.1), which corresponds to looking at long time and large space behaviour, and also in models of turbulence (see e.g. [137]).

We assume for simplicity that the functions $b(x, y)$ and $D(x, y)$ are 1-periodic in the second variable. We are interested in obtaining an effective equation (without the presence of ε) that is valid in the $\varepsilon \to 0$ limit.

Problems of this form fall under the framework of *homogenisation* theory (see e.g. [1]). This problem in particular has been addressed in [142] under the assumption that the fluid flow is purely periodic, i.e. $b = b(x/\varepsilon)$ and has zero average, i.e. $\int_{[0,1]^d} b(y)\, dy = 0$.

The first step towards removing these assumptions was the development of the *two-scale expansions with drift* method [140, 2]. This is a generalisation of the classical two-scale convergence method [1], modified to account for the strong convection. Heuristically, the two-scale expansion with drift posits the following formal asymptotic expansion for the solution of (1.15.1):

$$u^\varepsilon(t, x) = \sum_{i=0}^{\infty} \varepsilon^i u_i \left(t, x - \frac{b^* t}{\varepsilon}, \frac{x}{\varepsilon} \right), \tag{1.15.2}$$

where $u_i(t, x, y)$ are assumed to be 1-periodic in the y variable, and b^* is a constant drift vector. This formal expansion is made rigorous by a notion of weak convergence against oscillating test functions of the same form as the terms in (1.15.2), which allows the identification of an equation for u_0 (which is proved to not depend upon y). This equation is in Lagrangian coordinates for the presumed effective macroscopic flow.

The limitation of this method is that the velocity field b must be assumed to take the form $b = b^* + \tilde{b}(x/\varepsilon)$ where $\tilde{b}(y)$ has zero average.

In Chapter 9 we develop a method of *convergence along mean flows* which generalises the two-scale expansions with drift method to more general vector fields. In analogy with (1.15.2) we posit the formal asymptotic expansion:

$$u^\varepsilon(t, x) = \sum_{i=0}^{\infty} \varepsilon^i u_i \left(t, \Phi_{-t/\varepsilon}(x), \frac{t}{\varepsilon}, \frac{x}{\varepsilon} \right), \tag{1.15.3}$$

where $u_i(t, x, \tau, y)$ are 1-periodic in y and now depend upon a *fast time* variable τ. Here $\Phi_t(x)$ is the flow of the ODE generated by the mean flow

$$\dot{X} = \bar{b}(X) := \int_{[0,1]^d} b(X, y)\, dy.$$

We develop a notion of weak convergence along flows, and a corresponding weak compactness result, to make this expansion rigorous and obtain an equation for u_0 (which is proved to not depend on the fast variables y and τ).

An aside: Two-scale convergence. The method of *two-scale convergence* was developed by Nguetseng and Allaire [1] for periodic homogenisation and in its simplest form consists in identifying weak limits $u^\varepsilon(x) \rightharpoonup u^0(x, y)$ of the form

$$\lim_{\varepsilon \to 0} \int u^\varepsilon(x)\varphi(x, x/\varepsilon)\, dx = \int \int_{[0,1]^d} u^0(x, y)\varphi(x, y)\, dydx$$

for test functions $\varphi(x, y)$ and limit u^0 which are periodic in their second variable. Two-scale convergence has the advantage of being simpler and more straightforward to apply than the earlier *method of oscillating test function* of Tartar (see e.g. his extensive monograph [188]), but is less general.

In the classical theory of two-scale convergence one has corrector results which state that if u_ε two-scale converges to u_0, and ∇u_ε two-scale converges, then it two scale converges to $\nabla_x u_0 + \nabla_y u_1$ for some function $u_1 \in H^1_y$. In the case of the theory we develop of convergence along mean flows, we have that, instead, ∇u_ε converges to $J(\tau, x)\nabla u_0 + \nabla_y u_1$ where J is the Jacobian of the flow Φ. It is this dependence of the limits upon a fast time variable τ that necessitates the appearance of τ in (1.15.3) (it is not present in (1.15.2)).

The appearance of the Jacobian in the limit means that to obtain an effective equation for u_0 we must take the *fast time average* of the limit equation, i.e. take limits of the form

$$Mf := \lim_{l \to \infty} \frac{1}{2l} \int_{-l}^{l} f(x, \tau)\, d\tau \tag{1.15.4}$$

for the coefficients of the equation. To ensure that such limits exist we draw upon the theory of *Banach Algebras with mean value* (w.m.v.). This theory was constructed to deal with the problem of non-periodic homogenisation (see e.g.

[35, 155, 156, 8, 176, 157]), and uses the Gelfand transform (see e.g. [122]) to represent functions on a non-compact domain (e.g. $\mathbb{R}$) as functions on a compact space. We make the assumption that the Jacobian J lies in such an algebra. In particular it must be uniformly bounded in the fast time variable, an assumption that is necessary for the formal validity of the asymptotic expansion (1.15.3).

The assumption that the Jacobian J is uniformly bounded is needed to obtain a bound on the enhancement of diffusion due to *Lagrangian stretching* where the convection enlarges spatial gradients. In such cases the solution u^ε can decay to zero in a very short $t \ll 1$ (as $\varepsilon \to 0$) time scale, and the limit u^0 can be identically zero for any positive time. For general fluid fields b the behaviour of the Jacobian can be very complicated and it is difficult to obtain a complete picture of this phenomenon. However, in the case of *completely integrable Hamiltonian flows*, which in particular includes all *two dimensional incompressible flows* and all *shear flows*, the Jacobian behaves in a far simpler manner. In a forthcoming work [89] (which unfortunately was not complete in time to be part of this thesis) we shall address the behaviour in this case, obtaining rigorous asymptotic expansions both for a *time boundary layer* where diffusion is driven by the large convection, and for times of order one, where the large convection has no effect as the solution is invariant along streamlines.

Chapter 2

Propagation of chaos for Hölder continuous interaction kernels

We develop a new technique for establishing quantitative propagation of chaos for systems of interacting particles. Using this technique we prove propagation of chaos for diffusing particles whose interaction kernel is merely Hölder continuous, even at long ranges. Moreover, we do not require specially prepared initial data. On the way, we establish a law of large numbers for SDEs that holds over a class of vector fields simultaneously. The proofs bring together ideas from empirical process theory and stochastic flows.

Acknowledgements

We would like to thank Vittoria Silvestri for a helpful discussion regarding the continuity of the stochastic process defined in Section 2.2.2.1, Zhenfu Wang for discussion of [100] and for comments on the presentation of an earlier draft of this chapter and José A. Carrillo for mathematical comments during the preparation of this work. We also wish to thank Franco Flandoli for a discussion that lead to the discovery of an error in a previous version of this chapter.

2.1 Introduction

We consider the following system of N particles diffusing in $\mathbb{R}^d$:

$$\begin{cases} dX_t^{i,N} = b_t^N(X_t^{i,N})dt + dB_t^{i,N}, \quad i = 1, \dots, N, \\ b_t^N(x) = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} K(x, X_t^{i,N}), \\ (X_0^{i,N})_{i=1}^N \text{ i.i.d. with law } f_0. \end{cases} \tag{2.1.1}$$

where $B^{i,N}$ are i.i.d. standard d-dimensional Brownian motions and $K(x, y) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ is the interaction kernel. We are interested in the derivation of a *mean-field* model in the $N \gg 1$ regime for the density of particles $f_t(x)$. One expects that f should solve the non-linear convection-diffusion equation:

$$\begin{cases} \partial_t f_t + \nabla \cdot (b_t^\infty f_t) - \frac{1}{2}\Delta f_t = 0, \qquad (t, x) \in (0, T) \times \mathbb{R}^d, \\ b_t^\infty(x) = \displaystyle\int f_t(y) K(x, y)\, dy, \\ f_0(x) \text{ initial condition.} \end{cases} \tag{2.1.2}$$

To rigorously derive this limit one has to show that the *empirical measure*

$$\mu_t^N = \frac{1}{N}\sum_{i=1}^{N} \delta_{X_t^{i,N}} \tag{2.1.3}$$

converges weakly to the solution f_t to (2.1.2) as $N \to \infty$. This convergence of the empirical measure to a deterministic limit is known as chaoticity of the particle system [185]. At the initial time $t = 0$ this property is given, as the particles are i.i.d.. At any later time the particles are no longer independent. Establishing that, nevertheless, they are chaotic and the empirical measure converges as $N \to \infty$ is the problem of showing *propagation of chaos*. Of particular interest is making these notions *quantitative* - obtaining explicit polynomial (in N) bounds on some probability distance $d(\mu_t^N, f_t)$.

Establishing propagation of chaos is a central part of the rigorous mathematical derivation of macroscopic or mesoscopic continuum models from microscopic laws governing the motion of particles [76, 98]. The notion dates back to Boltzmann's idea of *molecular chaos* used for the derivation of the Boltzmann equation from

Newtonian collisions of gas particles. More generally, the notion of propagation of chaos is used in the derivation of the Vlasov-Poisson and Vlasov-Maxwell equations for galaxies and plasmas, for models of swarming [33, 34], in the Vortex dynamics interpretation of the Euler equation [66], the particles method for numerical integration of PDEs, the theory of particle filters in statistics [147], the derivation of the spatially homogeneous Boltzmann equation from Kac's model [144], among many others.

The main general technique for establishing quantitative propagation of chaos in this regime is the coupling method of Sznitman [185], which requires K to be *Lipschitz continuous*. Following this, many authors have obtained results assuming that $K(x, y)$ is Lipschitz except for a singularity at $x = y$, and in the presence of specially prepared initial conditions [65, 33, 81, 98]. For these techniques, the presence of the noise is a hindrance as it makes it harder to control the distances between particles.

Main result: In this work we develop a new method for establishing quantitative propagation of chaos, and apply it to give quantitative estimates of propagation of chaos of the system (2.1.1) under the assumption that K *is merely Hölder continuous*. In particular, this covers cases where K is *nowhere Lipschitz* and the result does not require specially prepared initial conditions. However, the result relies completely on the presence of noise, and fails at the first hurdle in its absence.

Second order systems. We also consider second order Langevin systems of the form:

$$\begin{cases} dX_t^{i,N} = V_t^{i,N} dt, \\ dV_t^{i,N} = b_t^N(X_t^{i,N})dt - \kappa V_t^{i,N} dt + dB_t^{i,N}, \\ b_t^N(x) = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} K(x, X_t^{i,N}), \\ (X_0^{i,N}, V_0^{i,N})_{i=1}^N \text{ i.i.d. with law } f_0(x, v). \end{cases} \qquad i = 1, \dots, N, \tag{2.1.4}$$

where again $B_t^{i,N}$ are i.i.d. standard d-dimensional Brownian motions, and κ is a constant (possibly zero). $X^{i,N} \in \mathbb{R}^d$ is the spatial position of the ith particle and $V^{i,N} \in \mathbb{R}^d$ is its velocity. The associated *mean-field* model is the non-linear

kinetic Fokker-Planck equation:

$$\begin{cases} \partial_t f_t + v\cdot\nabla_x f_t - \kappa\nabla_v\cdot(vf_t) + b_t^\infty\cdot\nabla_v f_t - \frac{1}{2}\Delta_v f_t = 0, \ (t,x,v)\in(0,T)\times\mathbb{R}^d\times\mathbb{R}^d, \\ b_t^\infty(x) = \int\left(\int f_t(y,v)\,dv\right)K(x,y)\,dy, \\ f_0(x,v) \text{ initial condition.} \end{cases} \tag{2.1.5}$$

Such systems model Newtonian particles under pairwise interaction forces that depend only on the spatial positions of the particles, and whose velocities are driven by independent white noises.

The empirical measure is given by

$$\mu_t^N = \frac{1}{N}\sum_{i=1}^N \delta_{(X_t^{i,N},V_t^{i,N})}. \tag{2.1.6}$$

Main result: We give quantitative estimates of propagation of chaos of the system (2.1.4) under the assumption that K *is Hölder continuous with Hölder exponent greater than* $2/3$. Again this applies to interaction kernels that are *nowhere Lipschitz* and the result does not require specially prepared initial conditions. The restriction of the Hölder exponent to be at least $2/3$ is due to the degeneracy of the generator of the diffusion process in the spatial variable. The generator is only *hypoelliptic* rather than *elliptic* and this reduces the regularising effect on the dynamics.

On the way to proving the propagation of chaos, we also establish a *Glivenko-Cantelli* theorem [193] for SDEs over all bounded Hölder continuous vector fields, (with a similar result in the second order case). This will be discussed in more detail below in Section 2.2.2 with a more precise statement, but we provide an informal statement below.

Glivenko-Cantelli theorem for SDEs (informal statement) Let X_t^b solve $dX_t = b_t(X_t)\,dt + dB_t$ for a vector field b, and $(X_t^{b,i,N})_{i=1}^N$ be N i.i.d. copies of X^b. Then

$$\mathbb{E}\sup_b \sup_{t\in[0,T]} d(\mu_t^{b,N}, f_t^b) \le CN^{-\gamma}$$

where $\mu^{b,N}$ is the empirical measure associated with $(X^{b,i,N})_{i=1}^N$, f^b is the law of

X^b, d is some metric on the space of probability measures and the supremum is over all smooth vector fields b with α-Hölder norm bounded by a uniform constant.

The proof uses recent results on stochastic flows for rough drifts (see e.g. [63, 12, 146, 59, 201, 60] among others), and methods from empirical process theory.

2.1.1 Layout of the chapter

The chapter is laid out as follows. In Section 2.1.2 we give preliminary definitions. Section 2.2 presents the main results of the chapter. In Section 2.3 we discuss prior work, compare our method to existing techniques and discuss open questions. The proofs of the results begin in Section 2.4 where we give the proof of Theorem 2.2.4. Then in Section 2.5 we apply the results proved in Section 2.4 to prove Theorem 2.2.1. Section 2.6 provides the proof of Proposition 2.2.2. Finally Section 2.A presents some properties of metric entropy which are used in the earlier sections of the chapter.

2.1.2 Preliminaries

Before we state the main results of this work we require some preliminary definitions.

We will always work with a single probability space $(\Omega, \mathcal{F}, \mathbb{P})$ that contains N i.i.d. Brownian motions $(B^{i,N})_{i=1}^N$ defined for times $[0,T]$ for a fixed final time T. We emphasise that throughout this work N is a fixed number and we will never take a limit $N \to \infty$. We denote the L^p norm on the probability space as $\|\!|\cdot|\!\|_p$. *Deterministic* norms are denoted with a double bar, e.g. $\|\cdot\|_{\mathrm{L}^\infty(\mathbb{R}^d)}$. The space of Borel probability measures on $\mathbb{R}^d$ is denoted $\mathrm{P}(\mathbb{R}^d)$, those with finite pth moment are denoted $\mathrm{P}_p(\mathbb{R}^d)$. We also make use of the following norm.

Definition 2.1.1 (Sub-Gaussian norm)**.** *For a random variable X we define the*

sub-Gaussian*[193, 196] norm*[1] $|||X|||$ *by*

$$|||X||| = \sup_{p \geq 1} \frac{1}{\sqrt{p}} |||X|||_p .$$

When $|||X|||$ *is finite we say that the random variable* X *is* sub-Gaussian. *If the random variable with law* $\mu \in \mathrm{P}(\mathbb{R}^d)$ *is* sub-Gaussian *then we say that* μ *is* sub-Gaussian.

The space of random variables with finite sub-Gaussian norm coincides with an exponential Orlicz space [193]. It is easy to see that if X is a random variable with $|||X||| = c$, then X obeys the tail bound

$$\mathbb{P}(|X| > u) \leq C \exp\left(-\frac{Cu^2}{c^2}\right) \tag{2.1.7}$$

for any $u > 0$, and absolute constants C. We refer the reader to [193, 196] for more details.

We define Lip1 to be the set of functions $h : \mathbb{R}^d \to \mathbb{R}$ that are 1-Lipschitz and vanish at zero.

To metrise weak convergence in $\mathrm{P}(\mathbb{R}^d)$, we define the following metrics:

Definition 2.1.2 (Bounded Lipschitz metric)**.** *For* $\mu, \nu \in \mathrm{P}(\mathbb{R}^d)$ *we define the* bounded-Lipschitz (BL) metric d_{BL} *by*

$$d_{\mathrm{BL}}(\mu, \nu) = \sup\left\{\int h \, d\mu - \int h \, d\nu : h \in \mathrm{Lip1}, \|h\|_{\mathrm{L}^\infty(\mathbb{R}^d)} \leq 1\right\}.$$

Definition 2.1.3 (Monge-Kantorovich-Wasserstein-(Rubinstein) metric)**.** *Given* μ *and* ν *in* $\mathrm{P}_1(\mathbb{R}^d)$ *we define the* Monge-Kantorovich-Wasserstein-(Rubinstein) (MKW) metric d_{MKW} *by*

$$d_{\mathrm{MKW}}(\mu, \nu) = \sup\left\{\int h \, d\mu - \int h \, d\nu : h \in \mathrm{Lip1}\right\}.$$

[1] The sub-Gaussian norm is sometimes called the ψ_2 norm and denoted $\|\cdot\|_{\psi_2}$. This notation is used, for example, in [193].

The MKW metric can also be defined using optimal transportation as

$$d_{\mathrm{MKW}}(\mu,\nu) = \inf\left\{\int |x-y|\, d\pi(x,y) : \pi \in \mathrm{P}(\mathbb{R}^d \times \mathbb{R}^d) \text{ has marginals } \mu \text{ and } \nu\right\}.$$

We will mostly use the first definition as it more amenable to the tools of empirical process theory.

Integral to the proofs is the concept of metric entropy[193].

Definition 2.1.4 (Metric entropy)**.** *Let $(\mathcal{X}, d)$ be a totally bounded metric space. Then the* metric entropy *of $\mathcal{X}$ is defined for $\varepsilon > 0$ by*

$$H(\varepsilon, \mathcal{X}, d) = \inf\{\log(m) : (x_i)_{i=1}^m \text{ is an } \varepsilon\text{-net of } \mathcal{X}\},$$

where an ε-net is a set of points $(x_i)_{i=1}^m \subseteq \mathcal{X}$ for which every $x \in \mathcal{X}$ has $d(x, x_i) \leq \varepsilon$ for some i.

We summarise the various properties and estimates of metric entropy that are used throughout this work in Section 2.A for the convenience of the reader.

The usual Lebesgue spaces will be denoted $\mathrm{L}^p(\mathbb{R}^d)$ for $p \in [1,\infty]$. We denote the usual Hölder spaces on $\mathbb{R}^d$ for k a positive integer and $\alpha \in (0,1)$ as $\mathrm{C}^{k,\alpha}(\mathbb{R}^d)$, and the (fractional) Sobolev space of differentiability $s \geq 0$ and integrability $p \in [1,\infty]$ as $\mathrm{W}^{s,p}(\mathbb{R}^d)$. When we wish to emphasis which variable a norm is respect to we will denote it with a subscript, e.g. $\mathrm{L}^q_x(\mathbb{R}^d)$.

To allow sets of functions that do not decay at infinity to have finite metric entropy we make use of weighted spaces. For $x \in \mathbb{R}^d$ we define $\langle x \rangle = \sqrt{1+|x|^2}$.

Definition 2.1.5 (Weighted L^p spaces)**.** *For $r \in \mathbb{R}$ and $q \in [1,\infty]$ we define $\mathrm{L}^{r,q}(\mathbb{R}^d)$ as the space of functions h such that $h\langle x\rangle^r \in \mathrm{L}^q(\mathbb{R}^d)$ with the norm*

$$\|h\|_{\mathrm{L}^{r,q}(\mathbb{R}^d)} = \|h\langle x\rangle^r\|_{\mathrm{L}^q(\mathbb{R}^d)} \left(= \left(\int |h|^q \langle x\rangle^{rq}\, dx\right)^{1/q} \text{ when } q \neq \infty\right).$$

In particular, $\mathrm{L}^{0,q}(\mathbb{R}^d) = \mathrm{L}^q(\mathbb{R}^d)$.

Next we define of 'abstract Hölder spaces'. We will use these in the proofs rather than the usual Sobolev spaces because they behave more naturally under com-

position with Hölder continuous functions.

Definition 2.1.6. *Let* $(\mathrm{V}, \|\cdot\|_{\mathrm{V}})$ *be a Banach space of functions* $U \to \mathbb{R}$ *for* $U = \mathbb{R}^d$ *or* $U = [0,T] \times \mathbb{R}^d$. *Let* $\alpha \in (0,1]$ *and* k *be a non-negative integer, then we define the space* $\Lambda^{k,\alpha}(\mathrm{V})$ *as those functions* $h \in \mathrm{V}$ *for which the following norm is finite*

$$\|h\|_{\Lambda^{k,\alpha}(\mathrm{V})} = \sup_{|\beta| \le k} \left\|\partial^\beta h\right\|_{\mathrm{V}} + \sup_{|\beta|=k} \sup_{y,z \in, y \neq z} \left\| \frac{(\partial^\beta h)(\cdot + z) - (\partial^\beta h)(\cdot + y)}{|y-z|^\alpha} 1_{\cdot + z \in U, \cdot + y \in U} \right\|_{\mathrm{V}}.$$

where β *ranges over multi-indices.*

Note that for $\alpha \in (0,1)$, $\Lambda^{0,\alpha}(\mathrm{L}^\infty(\mathbb{R}^d))$ is the Hölder space $\mathrm{C}^{0,\alpha}(\mathbb{R}^d)$. More generally, if $q \in [1,\infty]$, $s = k + \alpha$ is not an integer then $\Lambda^{k,\alpha}(\mathrm{L}^q(\mathbb{R}^d))$ is the Besov space $B^s_{q,\infty}$ (see [189] for the definitions and properties of the Besov spaces). In particular the fractional Sobolev space $\mathrm{W}^{s,q}(\mathbb{R}^d)$ embeds continuously in $\Lambda^{k,\alpha}(\mathrm{L}^q(\mathbb{R}^d))$.

Due to the driving Brownian motions, the natural regularity for the vector field b^N in (2.1.1) is a parabolic space.

Definition 2.1.7 (Parabolic space)**.** *Let* $(\mathrm{V}, \|\cdot\|_{\mathrm{V}})$ *be a Banach space of functions* $[0,T] \times \mathbb{R}^d \to \mathbb{R}$. *Let* $\alpha \in (0,1]$ *then we define the parabolic space* $\Lambda^{0,\alpha}_{para}(\mathrm{V})$ *is the space of functions* $h \in \mathrm{V}$ *with the following norm finite*

$$\|h\|_{\Lambda^{0,\alpha}_{para}(\mathrm{W})} = \|h\|_{\mathrm{W}} + \sup_{\substack{(s,y),(t,z) \in \mathbb{R}^{1+d}, \\ (s,y) \neq (t,z)}} \left\| \frac{h(\cdot + t, \cdot + z) - h(\cdot + s, \cdot + y)}{|y-z|^\alpha + |s-t|^{\alpha/2}} 1_{\cdot + (t,z) \in U, \cdot + (s,y) \in U} \right\|_{\mathrm{W}}.$$

We also define the particular case of the parabolic Hölder spaces.

Definition 2.1.8 (Parabolic Hölder space)**.** *For* $\alpha \in (0,1)$ *the* parabolic Hölder space $\mathrm{C}^{0,\alpha}_{para}$ *is the space of continuous functions* $\varphi : [0,T] \times \mathbb{R}^d \to \mathbb{R}$ *with the norm*

$$\|\varphi\|_{\mathrm{C}^{0,\alpha}_{para}([0,T]\times\mathbb{R}^d)} = \sup_{(t,x) \in [0,T] \times \mathbb{R}^d} |\varphi(t,x)| + \sup_{\substack{(s,y),(t,x) \in ([0,T]\times\mathbb{R}^d)^2, \\ (s,y) \neq (t,x)}} \frac{|\varphi(t,x) - \varphi(s,y)|}{|t-s|^{\alpha/2} + |x-y|^\alpha}.$$

Just as $\Lambda^{0,\alpha}(\mathrm{L}^\infty(\mathbb{R}^d))$ is the Hölder space $\mathrm{C}^{0,\alpha}(\mathbb{R}^d)$ (for $\alpha \in (0,1)$) the parabolic Hölder space $\mathrm{C}^{0,\alpha}_{para}(\mathbb{R}^d)$ is equal to $\Lambda^{0,\alpha}_{para}(\mathrm{L}^\infty(\mathbb{R}^d))$.

Stochastic flows are the analogue of the flow map of an ODE in the stochastic setting [116].

Definition 2.1.9 ($\mathrm{C}^{k,\beta}$ stochastic flow)**.** *We say that a random map $\phi_{s,t} : \mathbb{R}^n \to \mathbb{R}^n$ is a $\mathrm{C}^{k,\beta}$ ($k \geq 1, \beta \in (0,1)$) stochastic flow of diffeomorphisms if it satisfies the following*

1. *$\phi_{t,t}$ is the identity map almost surely.*
2. *$\phi_{u,t} \circ \phi_{s,u} = \phi_{s,t}$ holds as maps, almost surely for all $s < u < t$.*
3. *$\phi_{s,t}(x)$ is k-times differentiable with respect to x and all the derivatives are continuous, with the k-th derivative β-Hölder continuous. Furthermore, the map $\phi_{s,t} : \mathbb{R}^n \to \mathbb{R}^n$ is a diffeomorphism of $\mathbb{R}^n$ almost surely.*

Note that a $\mathrm{C}^{k,\beta}$ stochastic flow need not be globally $\mathrm{C}^{k,\beta}$ as both it and its derivatives may grow without bound as $|x| \to \infty$.

We say that a stochastic differential equation (for $X \in \mathbb{R}^n$) generates if $\mathrm{C}^{k,\beta}$ stochastic flow of diffeomorphisms if the solution map

$$\begin{aligned} \phi_{s,t} : \mathbb{R}^n &\to \mathbb{R}^n \\ X_s &\mapsto X_t \end{aligned}$$

has a version that is a $\mathrm{C}^{k,\beta}$ stochastic flow of diffeomorphisms.

Definition 2.1.10 (Glivenko-Cantelli class)**.** *Let $\mathbb{Q}$ be a probability measure on a measurable space $(\mathcal{X}, \mathcal{A})$ and $\mathcal{F}$ a class of measurable functions $\mathcal{X} \to \mathbb{R}$. We say that $\mathcal{F}$ is a* Glivenko-Cantelli *class (with respect to $\mathbb{Q}$) if*

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{N} \sum_{i=1}^{N} f(X^{i,N}) - \int f \, d\mathbb{Q} \right| \to 0$$

in probability or almost surely, where $(X^{i,N})_{i=1}^\infty$ are i.i.d. with law $\mathbb{Q}$.

Remark 2.1.1. *Strictly speaking, the convergence in the definition above should be in* outer *probability or* outer *almost surely (see [193, §1] for the definition and*

properties of the outer integral), as the supremum may fail to be measurable in general. In this work, however, all considered suprema will be measurable and we will have no need of the more technical definition of the outer integral.

In this work we consider both the first order many particle system (2.1.1) and the second order many particle system (2.1.4) along with their respective limit equations. We hope that the reader will admit us the abuse of notation of using the same symbols X, f, μ^N for each, as which is considered will be clear from the context.

2.2 Main results

In this section we present the main results of this chapter.

2.2.1 Propagation of chaos

The first such results are on propagation of chaos.

2.2.1.1 First order systems.

For first order systems we have the full regularising effect of the driving noise and we only require that the interaction kernel K is Hölder continuous for some positive exponent. Under this assumption we achieve sub-Gaussian concentration of the Wasserstein distance over compact time intervals around the initial distance. Recall that the norm $|||\cdot|||$ is defined by Definition 2.1.1 and is equivalent to a sub-Gaussian tail bound (2.1.7). Note that we *cannot* expect $|||d_{\mathrm{MKW}}(\mu_t^N, f_t)|||$ to be small in general, as even at the initial time we have that

$$|||d_{\mathrm{MKW}}(\mu_0^N, f_0)||| = \infty$$

unless f_0 has sub-Gaussian tails. For this reason we bound instead a compensated quantity $d_{\mathrm{MKW}}(\mu_t^N, f_t) - d_{\mathrm{MKW}}(\mu_0^N, f_0)$. Bounds upon $d_{\mathrm{MKW}}(\mu_0^N, f_0)$ are

a separate question to *propagation* of chaos and are considered elsewhere (e.g. [48, 64]).

Theorem 2.2.1 (Propagation of chaos for first order systems)**.** *Let $(X^{i,N})_{i=1}^N$ be a solution of the first order many particle system (2.1.1), μ^N be the associated empirical measure given by (2.1.3) and $f_t(x)$ be the solution to the limit equation (2.1.2). Then the following hold:*

1. ***Hölder interactions:*** *Let $K \in \mathrm{C}^{0,\alpha}(\mathbb{R}^d \times \mathbb{R}^d; \mathbb{R}^d)$ and $f_0 \in \mathrm{P}_p(\mathbb{R}^d) \cap \mathrm{L}^{r,2}(\mathbb{R}^d)$ for some $\alpha \in (0,1]$, $2 \neq p > 1$ and $r > 1 + (d/2)$. Then it holds that*

$$\left\| \sup_{t\in[0,T]} d_{\mathrm{MKW}}(\mu_t^N, f_t) - d_{\mathrm{MKW}}(\mu_0^N, f_0) \right\| \leq CN^{-\gamma}, \quad \text{for any } \gamma < \frac{1}{2 + \max(\frac{d+2}{\alpha^2}, \frac{d}{p-1})}. \tag{2.2.1}$$

2. ***Sobolev interactions:*** *Let $K(x,y) = W(x-y)$ for $W \in \mathrm{W}^{s,q}(\mathbb{R}^d; \mathbb{R}^d)$ with $1 \geq s > (2+d)/q$ and $q \in (2,\infty]$. Assume that $f_0 \in \mathrm{L}^{r,q'}(\mathbb{R}^d) \cap \mathrm{P}_p(\mathbb{R}^d)$ for some $2 \neq p > d/q$, $r > (d/q) + (d/2) + 1$ and where $(1/q) + (1/q') = 1/2$. Then it holds that*

$$\left\| \sup_{t\in[0,T]} d_{\mathrm{MKW}}(\mu_t^N, f_t) - d_{\mathrm{MKW}}(\mu_0^N, f_0) \right\| \leq CN^{-\gamma}, \quad \text{for any } \gamma < \frac{1}{2 + \max(\frac{d+2}{s^2}, \frac{d}{p-1})}. \tag{2.2.2}$$

By combining the above theorem and the results on the convergence of $d_{\mathrm{MKW}}(\mu_0^N, f_0)$ in [64] we can obtain the following simple corollary.

Corollary 2.2.1. *Under the assumptions of Theorem 2.2.1 we have*

$$\mathbb{E} \sup_{t\in[0,T]} d_{\mathrm{MKW}}(\mu_t^N, f_t) \leq \mathbb{E} d_{\mathrm{MKW}}(\mu_0^N, f_0) + CN^{-\gamma}$$

with γ as given in the respective cases (1), (2) of Theorem 2.2.1. Furthermore, if p is large enough (depending only on d) then it holds that

$$\mathbb{E} \sup_{t\in[0,T]} d_{\mathrm{MKW}}(\mu_t^N, f_t) \leq CN^{-\gamma}.$$

We now give some remarks on Theorem 2.2.1.

Remark 2.2.1. *We do not require any special preparation for the initial condition f_0, and the initial particle positions are chosen i.i.d. according to f_0. In particular, any compactly supported uniformly bounded density will satisfy all the assumptions of the theorem. The restriction to i.i.d. initial particle positions is made throughout this work, both to simplify the proofs and as it is a natural initial condition. However, the author believes that there is no fundamental reason why this could not be dropped with additional work.*

Remark 2.2.2. *In both cases in the above theorem the $d/(p-1)$ term in γ comes from the metric entropy of the space* Lip1 *in a weighted L^∞ norm (see Proposition 2.A.1). The requirement that $p \neq 2$ is merely to avoid complicating the theorem statements with logarithmic correction terms for this critical weight. Results for $p = 2$ can be obtained by using the inclusion $\mathrm{P}_2 \subset \mathrm{P}_p$ for any $p < 2$. We maintain the avoidance of $p = 2$ throughout the other results of the chapter.*

Remark 2.2.3. *If we consider instead the bounded Lipschitz metric d_{BL}, then the same method of proof allows us to estimate*

$$\left\| \sup_{t\in[0,T]} d_{\mathrm{BL}}(\mu_t^N, f_t) \right\| \leq CN^{-\gamma}$$

under the same assumptions as Theorem 2.2.1 and with the same corresponding values of γ. This is because the bounded Lipschitz metric is almost surely bounded by 2 and there is no need to compensate for the initial error. Additionally, the exponent γ can be improved to anything less than $1/(2+\max((d+2)/\alpha^2))$ in (1) and $1/(2+\max((d+2)/s^2))$ in (2), as the metric entropy estimate for the bounded Lipschitz functions is smaller than that for Lip1 *(see also Remark 2.2.2). Analogous results in the bounded Lipschitz metric can be formulated for the Glivenko-Cantelli theorems for SDEs (Theorems 2.2.4 and 2.2.6) presented later in this work, and for the second order propagation of chaos result Theorem 2.2.2 below.*

Remark 2.2.4. *The assumption on K in (2) may be weakened to*

$$K \in \Lambda^{0,s}(\mathrm{L}_y^\infty(\mathbb{R}^d; \mathrm{L}_x^q(\mathbb{R}^d; \mathbb{R}^d)))$$

for the same s and q. (This space is defined in Definition 2.1.6.) Note that this

condition is implied by the assumption in (2). We provide the proof under this weaker assumption. Note that with this weakened assumption the case $q = \infty$ collapses into (1) with $\alpha = s$ as $\Lambda^{0,s}(\mathrm{L}^\infty(\mathbb{R}^d \times \mathbb{R}^d; \mathbb{R}^d))$ is exactly the Hölder space $\mathrm{C}^{0,s}(\mathbb{R}^d \times \mathbb{R}^d; \mathbb{R}^d)$.

Remark 2.2.5. *In case (2) in the above theorem the assumptions on K imply that $K \in \mathrm{C}^{0,\alpha}(\mathbb{R}^d \times \mathbb{R}^d; \mathbb{R}^d)$ for $\alpha = s - (d/q) > 0$ by Sobolev embedding. So in this sense case (2) is weaker than case (1). The advantage of (2) is that γ is obtained from the Sobolev exponent s instead of the (smaller) Hölder exponent α. In particular, note that singularities of the form*

$$K(x, y) = W(x - y) = |x - y|^\alpha, \qquad \alpha \in (0, 1)$$

are (locally) only α-Hölder, while they are (locally) in the Sobolev space $\mathrm{W}^{1,q}$ for any $q < d/(1 - \alpha)$. For this type of singularity the exponent γ obtained by (2) is substantially better than that from (1).

Remark 2.2.6. *The integrability assumptions on f_0 in the above theorem are not optimal, and the author has made little attempt to optimise them. In particular, the proof does not use the regularising effect (gain of integrability and smoothness) of the parabolic limit equation in the PDE estimates, instead relying on more elementary L^2 energy estimates. A more careful analysis is beyond the scope of this thesis, in which we are primarily concerned with the assumptions on K and upon the exponent γ obtained, and not on optimal integrability of the initial data.*

2.2.1.2 Second order systems.

In the second order case we have a similar result, but with the restriction that $\alpha > 2/3$ as we have merely a gain of $2/3$ derivatives from the hypoelliptic regularising effect of the noise. However, the particle spatial trajectories have better time regularity properties than in the first order case (at least C^1 rather than $\mathrm{C}^{0,(1/2)-\varepsilon}$) and as a result we obtain a better exponent γ.

Theorem 2.2.2 (Propagation of chaos for second order systems)**.**
Let $(X^{i,N}, V^{i,N})_{i=1}^N$ be a solution of the second order many particle system (2.1.4), μ^N be the associated empirical measure given by (2.1.6) and $f_t(x)$ be the solution

to the limit equation (2.1.5). Then the following hold:

1. ***Hölder interactions:*** *Let $K \in \mathrm{C}^{0,\alpha}(\mathbb{R}^d \times \mathbb{R}^d; \mathbb{R}^d)$ and $f_0 \in \mathrm{P}_p(\mathbb{R}^d \times \mathbb{R}^d) \cap \mathrm{L}^{r,2}(\mathbb{R}^d \times \mathbb{R}^d)$ for $\alpha \in (2/3, 1]$, $2 \neq p > 1$ and $r > 1 + d$. Then there are finite constants c, C such that the following holds*

$$\left\| \left[\sup_{t\in[0,T]} d_{\mathrm{MKW}}(\mu_t^N, f_t) - c d_{\mathrm{MKW}}(\mu_0^N, f_0) \right]_+ \right\| \leq CN^{-\gamma}, \tag{2.2.3}$$

where here and throughout $[a]_+$ is the positive part of a, and γ is given by

$$\gamma = \frac{1}{2 + \max(\frac{d+1}{\alpha^2}, \frac{d}{p-1})}.$$

2. ***Sobolev interactions:*** *Let $K(x, y) = W(x-y)$ for some $W \in \mathrm{W}^{s,q}(\mathbb{R}^d; \mathbb{R}^d)$ with $q \in (2, \infty]$, $q > (d+1)/s$ and $3/2 \geq s > (2/3) + (d/q)$. Let $f_0 \in \mathrm{P}_p(\mathbb{R}^d \times \mathbb{R}^d) \cap \mathrm{L}^{r,q'}(\mathbb{R}^d \times \mathbb{R}^d)$ for some $p > d/q$ with also $p > 2$, $r > (d/q) + d + 1$ and where $(1/q) + (1/q') = 1/2$. Then there are finite constants c, C such that the following holds*

$$\left\| \left[\sup_{t\in[0,T]} d_{\mathrm{MKW}}(\mu_t^N, f_t) - c d_{\mathrm{MKW}}(\mu_0^N, f_0) \right]_+ \right\| \leq CN^{-\gamma},$$

where

$$\gamma = \begin{cases} \dfrac{1}{2 + \frac{d+1}{s^2}}, & \text{if } s \leq 1, \\ \dfrac{1}{2 + \max(\frac{d+1}{s}, d)}, & \text{otherwise.} \end{cases}$$

Using the elementary inequality $x \leq [x - y]_+ + y$ for $x, y \geq 0$, we can use Theorem 2.2.2 to obtain bounds on the expectation:

Corollary 2.2.2. *Under the assumptions of Theorem 2.2.2 we have*

$$\mathbb{E} \sup_{t\in[0,T]} d_{\mathrm{MKW}}(\mu_t^N, f_t) \leq C\mathbb{E} d_{\mathrm{MKW}}(\mu_0^N, f_0) + CN^{-\gamma}$$

with γ as given in the respective cases (1), (2) of Theorem 2.2.2. Furthermore,

if p is large enough (depending only on d) then it holds that

$$\mathbb{E} \sup_{t\in[0,T]} d_{\mathrm{MKW}}(\mu_t^N, f_t) \leq CN^{-\gamma}.$$

We make some remarks on Theorem 2.2.2 (see also the remarks after Theorem 2.2.1 which are applicable here as well).

Remark 2.2.7. *In the first order case (Theorem 2.2.1) we admitted as evident the well-posedness of both the limit PDE (2.1.2) and the particle system (2.1.1). In the second order case neither is a priori obvious due to the degeneracy of the noise and the roughness of the coefficients. We note here that the well-posedness of the particle system (2.1.4) is a consequence of the existence of a differentiable stochastic flow (see [201]). That the limit PDE (2.1.5) is also well-posed may be obtained from this by standard methods. However, we point out to the reader that for most of the chapter the proofs are done on mollified (C_b^1) vector fields, and so existence and uniqueness is not a concern.*

Remark 2.2.8. *We need to subtract $cd_{\mathrm{MKW}}(\mu_0^N, f_0)$ (for $c > 1$) and take the positive part, while in Theorem 2.2.1 we could choose $c = 1$ and the non-negativity of the expression was automatic. This is because the initial error may be amplified by the dynamics as the vector fields involved are not bounded (see the proof of Lemma 2.4.4 for details).*

Remark 2.2.9. *The assumption on K in (1) can be weakened to*

$$K \in \Lambda^{0,s}(\mathrm{L}_y^\infty(\mathbb{R}^d; \mathrm{L}_x^q(\mathbb{R}^d; \mathbb{R}^d)))$$

for $s \in (2/3, 1]$, $q > (d+1)/s$ and $p > d/q$. (This space is defined in Definition 2.1.6). This includes the result stated in the theorem as the case $q = \infty$.

The assumption on K in (2) can be weakened when $s > 1$ to

$$W \in \Lambda^{1,(s-1)}(\mathrm{L}^q(\mathbb{R}^d; \mathbb{R}^d))$$

under the additional assumption that $p \geq 4$. This implies the assumption on K in the theorem statement. Note that the $s \leq 1$ case of this weakened assumption was included in the relaxation of (1) directly above.

In each case the proof is given for these weakened assumptions.

2.2.2 Empirical process & Glivenko-Cantelli theorems for SDEs

In this subsection we define the empirical process hinted at in the informal statement Section 2.1. We present first the definitions and results for first order systems.

2.2.2.1 First order systems.

Let $\tilde{\mathcal{C}}^\alpha$ be the set of vector fields given by

$$\tilde{\mathcal{C}}^\alpha = \{b : \|b\|_{\mathrm{C}([0,T];\mathrm{C}^{0,\alpha}(\mathbb{R}^d;\mathbb{R}^d))} \leq C\} \tag{2.2.4}$$

for some $\alpha \in (0,1)$ and $C < \infty$ fixed. For any $b \in \tilde{\mathcal{C}}$ and any $N \in \mathbb{N}$ denote $(X^{b,i,N})_{i=1}^N$ as the solution to

$$\begin{cases} dX_t^{b,i,N} = b_t(X_t^{b,i,N})dt + dB_t^{i,N}, \\ X_0^{b,i,N} = X_0^{i,N} \end{cases} \quad i = 1, \dots, N \tag{2.2.5}$$

where $X_0^{i,N}$ and $B^{i,N}$ are the same as in (2.1.1). Note that $(X^{b,i,N})_{i=1}^N$ are i.i.d. by construction. The *empirical process* $(\mu_t^{b,N})_{t\in[0,T],b\in\tilde{\mathcal{C}}^\alpha}$ is defined by

$$\mu_t^{b,N} = \frac{1}{N}\sum_{i=1}^N \delta_{X_t^{b,i,N}}. \tag{2.2.6}$$

Note that for any $b \in \tilde{\mathcal{C}}^\alpha$ and any $i \in \{1, \dots, N\}$, the law of $X_t^{b,i,N}$ is given by f_t^b, the solution to the following parabolic PDE:

$$\begin{cases} \partial_t f_t^b + \nabla \cdot (b_t f_t^b) - \frac{1}{2}\Delta f_t^b = 0, & (t,x) \in (0,T) \times \mathbb{R}^d, \\ f_0^b(x) = f_0(x) \text{ initial condition.} \end{cases} \tag{2.2.7}$$

We would like to be able to consider $\mu_t^{b,N}$ for *random* $b \in \tilde{\mathcal{C}}^\alpha$. To do so we need that the stochastic process $(X^{b,i,N})_{i=1}^N$ indexed by $t \in [0,T]$ and $b \in \tilde{\mathcal{C}}^\alpha$ be (almost

surely) continuous. Let us be precise about this for the benefit of readers less familiar with such notions. We wish to construct a (random) map (in other words a stochastic process) φ defined by

$$\begin{aligned}\varphi : ([0,T], |\cdot|) \times (\tilde{\mathcal{C}}^\alpha, \mathrm{L}^\infty([0,T]; \mathrm{L}^{-r,\infty}(\mathbb{R}^d;\mathbb{R}^d))) &\to (\mathbb{R}^d)^N \\ (t,b) &\mapsto (X_t^{b,i,N})_{i=1}^N,\end{aligned}$$

for some $r > 0$. The statement that φ is continuous at a point $(t,b) \in [0,T] \times \tilde{\mathcal{C}}^\alpha$ means that for any sequence $t_n, b_n \in [0,T] \times \tilde{\mathcal{C}}^\alpha$ converging to t, b as $n \to \infty$ in the topologies in the above display, we have that $(X_{t_n}^{b_n,i,N})_{i=1}^N \to (X_t^{b,i,N})_{i=1}^N$ as $n \to \infty$ in $\mathbb{R}^{d\cdot N}$.

We ask that φ be almost surely continuous, i.e.

$$\mathbb{P}(\,\forall (t,b) \in [0,T] \times \tilde{\mathcal{C}}^\alpha, \varphi \text{ is continuous at } (t,b)) = 1. \tag{2.2.8}$$

This is a *much stronger* requirement than with the quantifiers switched, i.e.

$$\forall (t,b) \in [0,T] \times \tilde{\mathcal{C}}^\alpha, \mathbb{P}(\varphi \text{ is continuous at } (t,b)) = 1. \tag{2.2.9}$$

The former implies the latter but not vice versa.

The size of the index set $\tilde{\mathcal{C}}^\alpha$ causes a technical issue that although (2.2.9) can be shown, it is impossible to show (2.2.8):

Proposition 2.2.1. *Let $f_0 \in \mathrm{P}_p(\mathbb{R}^d)$ for some $p > 1$. Then the process $(X_t^{b,i,N})_{i=1}^N$ indexed by $b \in \tilde{\mathcal{C}}^\alpha$ and $t \in [0,T]$ cannot be modified to give an almost surely continuous process with the $\mathrm{L}^\infty([0,T]; \mathrm{L}^{-r,\infty}(\mathbb{R}^d;\mathbb{R}^d))$ (any $r > 0$) topology on $\tilde{\mathcal{C}}^\alpha$.*

This is because, roughly speaking, constructing this process would give uniqueness for SDEs with *random* α-Hölder coefficients, and there are simple counterexamples. We refer the reader to the proof of Proposition 2.2.1 for further details.

For this reason we define $\mathcal{C}^\alpha$ as the set of *smooth* (C_b^1 in x) vector fields in $\tilde{\mathcal{C}}^\alpha$, i.e.

$$\mathcal{C}^\alpha = \mathrm{C}([0,T]; \mathrm{C}_b^1(\mathbb{R}^d;\mathbb{R}^d)) \cap \tilde{\mathcal{C}}^\alpha. \tag{2.2.10}$$

Note that $\mathcal{C}^\alpha$ contains vector fields with C_b^1 norm arbitrarily large. Also $\mathcal{C}^\alpha$ is dense in $\tilde{\mathcal{C}}^\alpha$ in the $\mathrm{L}^\infty([0,T];\mathrm{L}^{-r,\infty}(\mathbb{R}^d;\mathbb{R}^d))$ topology for any $r>0$.

For this set of vector fields we *can* construct a continuous stochastic process.

Theorem 2.2.3. *Let $f_0 \in \mathrm{P}_p(\mathbb{R}^d)$ for some $p>1$. Then the process $(X_t^{b,i,N})_{i=1}^N$ defined by (2.2.5) and indexed by $t\in[0,T], b\in\mathcal{C}^\alpha$ has a modification that is continuous, where $\mathcal{C}^\alpha$ is equipped with the $\mathrm{L}^\infty([0,T];\mathrm{L}^{-r,\infty}(\mathbb{R}^d;\mathbb{R}^d))$ topology (any $r\in(0,p)$). As a consequence, the same holds for the empirical process $(\mu_t^{b,N})_{t\in[0,T],b\in\mathcal{C}^\alpha}$ given by (2.2.6) above, mapping into the space of probability measures equipped with the weak topology.*

In the style of language of (2.2.8), this theorem states that we can construct the process $\mu_t^{b,N}$ in such a way that

$$\mathbb{P}\left(\begin{array}{l}\text{For any sequence } (t_n,b_n)\to(t,b) \text{ as } n\to\infty \text{ in } [0,T]\times\mathcal{C}^\alpha,\\ \text{we have } \mu_{t_n}^{b_n,N}\to\mu_t^{b,N} \text{ as } n\to\infty \text{ weakly in } \mathrm{P}(\mathbb{R}^d).\end{array}\right)=1$$

where the convergence of b_n is in $\mathrm{L}^\infty([0,T];\mathrm{L}^{-r,\infty}(\mathbb{R}^d;\mathbb{R}^d))$.

The inability to construct the process on the full set of α-Hölder continuous vector fields $\tilde{\mathcal{C}}^\alpha$ means that the following results are a priori, in the sense that they must be applied to smoothed (C_b^1) vector fields, but are uniform in the degree of smoothness.

Our main result on this empirical process is that the Wasserstein distance between the empirical measure $\mu_t^{b,N}$ and the law f_t^b has (polynomial in N) sub-Gaussian concentration about the initial distance $d_{\mathrm{MKW}}(\mu_0^N,f_0)$.

Theorem 2.2.4 (Glivenko-Cantelli theorem for SDEs)**.** *Let $f_0\in\mathrm{P}_p(\mathbb{R}^d)$ for some $2\neq p>1$. Assume that $\mathcal{C}\subset\mathcal{C}^\alpha$ obeys the metric entropy bound*

$$H(\varepsilon,\mathcal{C},\|\cdot\|_{\mathrm{L}^\infty([0,T];\mathrm{L}^{-r,\infty}(\mathbb{R}^d;\mathbb{R}^d))})\leq C\varepsilon^{-k} \tag{2.2.11}$$

for some $r\in(1,p)$. Then it holds that

$$\left\|\sup_{t\in[0,T],b\in\mathcal{C}} d_{\mathrm{MKW}}(\mu_t^{b,N},f_t^b)-d_{\mathrm{MKW}}(\mu_0^N,f_0)\right\|\leq CN^{-\gamma}, \tag{2.2.12}$$

with

$$\gamma = \frac{1}{2 + \max(d, d/(p-1), k)}.$$

As discussed below the statement of Theorem 2.2.1, we can easily use this bound to obtain estimates on the expectation of the Wasserstein distance.

Corollary 2.2.3. *Under the assumptions of Theorem 2.2.4 it holds that*

$$\mathbb{E} \sup_{t\in[0,T], b\in\mathcal{C}} d_{\mathrm{MKW}}(\mu_t^{b,N}, f_t^b) \leq \mathbb{E} d_{\mathrm{MKW}}(\mu_0^N, f_0) + CN^{-\gamma},$$

where

$$\gamma = \frac{1}{2 + \max(d, d/(p-1), k)}.$$

Moreover, if p is large enough depending only on d, k then it holds that

$$\mathbb{E} \sup_{t\in[0,T], b\in\mathcal{C}} d_{\mathrm{MKW}}(\mu_t^{b,N}, f_t^b) \leq CN^{-\gamma}, \quad \gamma = \frac{1}{2 + \max(d, k)}. \tag{2.2.13}$$

Remark 2.2.10. *Similar results with weaker non-polynomial rates may be easily obtained with minor modification of the proof for the case that different types of estimates on the metric entropy hold. In particular convergence to zero of (2.2.12) as $N \to \infty$ will hold for any set $\mathcal{C} \subset \mathcal{C}^\alpha$ that is totally bounded in the norm used in (2.2.11).*

Remark 2.2.11. *Despite the density of $\mathcal{C}^\alpha$ in $\tilde{\mathcal{C}}^\alpha$ in the $\mathrm{L}^\infty([0,T]; \mathrm{L}^{-r,\infty}(\mathbb{R}^d; \mathbb{R}^d))$ norm, we cannot replace the subset $\mathcal{C}$ with its closure in $\mathcal{C}^\alpha$ in this norm in the above theorem. This is due to the difficulty in defining the process considered over such a large index set (see Proposition 2.2.1).*

Remark 2.2.12. *We call this result a Glivenko-Cantelli theorem as it implies that*

$$\mathcal{F} = \{\omega \mapsto h(X_t^b(\omega)) : h \text{ is 1-Lipschitz}, t \in [0,T], b \in \mathcal{C}\},$$

is a Glivenko-Cantelli class with respect to the Weiner measure (see Definition 2.1.10).

The proof of Theorem 2.2.4 also provides better estimates of weaker measures of distance, see Proposition 2.4.2 in Section 2.4 below.

Applications of Theorem 2.2.4 combined with well known metric entropy of function spaces [54, 189] gives explicit convergence rates. In particular, for the parabolic Hölder scale of spaces we obtain:

Corollary 2.2.4. *Let $f_0 \in \mathrm{P}_p(\mathbb{R}^d)$ for some $2 \neq p > 1$, and let $\mathcal{C}^\alpha_{para}$ be given by*

$$\mathcal{C}^\alpha_{para} = \mathrm{C}([0,T]; \mathrm{C}^1_b(\mathbb{R}^d;\mathbb{R}^d)) \cap \{b : \|b\|_{\mathrm{C}^{0,\alpha}_{para}([0,T]\times\mathbb{R}^d;\mathbb{R}^d)} \leq C\}$$

for some constants $C \in (0,\infty)$ and $\alpha \in (0,1)$. Then it holds that

$$\left\| \sup_{t\in[0,T], b\in\mathcal{C}} d_{\mathrm{MKW}}(\mu^{b,N}_t, f^b_t) - d_{\mathrm{MKW}}(\mu^N_0, f_0) \right\| \leq CN^{-\gamma}, \tag{2.2.14}$$

where

$$\gamma = \frac{1}{2 + \max(\frac{d+2}{\alpha}, \frac{d}{p-1})}.$$

As before, this estimate can be combined with estimates of the initial distance $d_{\mathrm{MKW}}(\mu^N_0, f_0)$ to obtain results corresponding to Corollary 2.2.3. Similar results can be easily obtained for the non-parabolic spaces. While we do not claim that the γ in (2.2.14) is optimal, the result is optimal for the Hölder scale in the sense that no such estimate is possible for $\alpha = 0$. In fact:

Proposition 2.2.2. *Let $f_0 \in \mathrm{P}_p(\mathbb{R}^d)$ for some $p > 1$ and let $\mathcal{C}^0$ be given by*

$$\mathcal{C}^0 = \mathrm{C}([0,T]; \mathrm{C}^1_b(\mathbb{R}^d;\mathbb{R}^d)) \cap \{b : \|b\|_{\mathrm{C}([0,T]\times\mathbb{R}^d;\mathbb{R}^d)} \leq C\}$$

for some constant $C \in (0,\infty)$. Then it holds that

$$\inf_{N\geq 1} \mathbb{E} \sup_{t\in[0,T], b\in\mathcal{C}} d_{\mathrm{BL}}(\mu^{b,N}_t, f^b_t) \geq c > 0. \tag{2.2.15}$$

Note that the use of d_{BL} in (2.2.15) is a stronger statement than if d_{MKW} were used as the Wasserstein metric generates a stronger topology. The proof of Proposition 2.2.2 is provided in Section 2.6 where a stochastic control problem is introduced, which is solvable only if no such uniform law of large numbers can hold.

2.2.2.2 Second order systems.

The definitions in the second order case are analogous to those in the first order case, but with give them in full for completeness. Let $\tilde{\mathcal{C}}^\alpha$ be the set of vector fields given by

$$\tilde{\mathcal{C}}^\alpha = \{b : \|b\|_{\mathrm{C}([0,T];\mathrm{C}^{0,\alpha}(\mathbb{R}^d;\mathbb{R}^d))} \leq C\} \tag{2.2.16}$$

for some constants C, α with $\alpha \in (2/3, 1)$. Define $(X^{b,i,N}, V^{b,i,N})_{i=1}^N$ for $b \in \tilde{\mathcal{C}}^\alpha$ and $N \in \mathbb{N}$ as the solution to

$$\begin{cases} dX_t^{b,i,N} = V_t^{b,i,N} dt, \\ dV_t^{b,i,N} = b_t(X_t^{b,i,N})dt - \kappa V_t^{b,i,N} dt + dB_t^{i,N}, \quad i = 1, \dots, N, \\ (X_0^{b,i,N}, V_0^{b,i,N}) = (X_0^{i,N}, V_0^{i,N}), \end{cases} \tag{2.2.17}$$

where $X_0^{i,N}, V_0^{i,N}$ and $B^{i,N}$ are the same as in (2.1.4).

The *empirical process* $(\mu_t^{b,N})_{t\in[0,T],b\in\tilde{\mathcal{C}}^\alpha}$ is defined by

$$\mu_t^{b,N} = \frac{1}{N}\sum_{i=1}^N \delta_{(X_t^{b,i,N}, V_t^{b,i,N})}. \tag{2.2.18}$$

For any $b \in \tilde{\mathcal{C}}^\alpha$ and $i \in \{1, \dots, N\}$, the law of $(X_t^{i,N}, V_t^{i,N})$ is f_t^b which solves the following degenerate parabolic PDE:

$$\begin{cases} \partial_t f_t^b + v\cdot\nabla_x f_t^b - \kappa\nabla_v\cdot(v f_t^b) + b_t\cdot\nabla_v f_t^b - \frac{1}{2}\Delta_v f_t^b = 0, \ (t,x) \in (0,T)\times\mathbb{R}^d\times\mathbb{R}^d, \\ f_0^b(x,v) = f_0(x,v) \text{ initial condition.} \end{cases} \tag{2.2.19}$$

As before we must work with a dense smooth subset

$$\mathcal{C}^\alpha = \mathrm{C}([0,T];\mathrm{C}_b^1(\mathbb{R}^d;\mathbb{R}^d)) \cap \tilde{\mathcal{C}}^\alpha. \tag{2.2.20}$$

The stochastic process indexed by $\mathcal{C}^\alpha$ has a continuous modification.

Theorem 2.2.5. *Let $f_0 \in \mathrm{P}_p(\mathbb{R}^{2d})$ for some $p > 1$. Then the process $(X_t^{b,i,N}, V_t)_{i=1}^N$ indexed by $t \in [0,T], b \in \mathcal{C}^\alpha$ has a modification that is continuous, where $\mathcal{C}^\alpha$ is equipped with the $\mathrm{L}^\infty([0,T];\mathrm{L}^{-r,\infty}(\mathbb{R}^d;\mathbb{R}^d))$ topology (any $r > 0$). As a consequence, the same holds for the empirical process $(\mu_t^{b,N})_{t\in[0,T],b\in\mathcal{C}}$ in the weak topology on $\mathrm{P}(\mathbb{R}^d)$.*

For the second order system the main result of this subsection is the following.

Theorem 2.2.6 (Glivenko-Cantelli theorem for SDEs (second order case))**.** *Let $f_0 \in \mathrm{P}_p(\mathbb{R}^{2d})$ for some $2 \neq p > 1$ and assume $\alpha \in (2/3, 1)$. Assume that $\mathcal{C} \subset \mathcal{C}^\alpha$ obeys the metric entropy bound*

$$H(\varepsilon, \mathcal{C}, \|\cdot\|_{\mathrm{L}^\infty([0,T];\mathrm{L}^{-r,\infty}(\mathbb{R}^d;\mathbb{R}^d))}) \leq C\varepsilon^{-k} \tag{2.2.21}$$

for some $r \in (1, p)$. Let $\mu^{b,N}$ and f^b be defined by (2.2.18) and (2.2.19) respectively. Then it holds that

$$\left\| \left[\sup_{t\in[0,T], b\in\mathcal{C}} d_{\mathrm{MKW}}(\mu_t^{b,N}, f_t^b) - c d_{\mathrm{MKW}}(\mu_0^N, f_0) \right]_+ \right\| \leq CN^{-\gamma}, \tag{2.2.22}$$

where γ is given by

$$\gamma = \frac{1}{2 + \max(d, d/(p-1), k)}. \tag{2.2.23}$$

As before we can use this result to obtain estimates like the following.

Corollary 2.2.5. *Under the assumptions of Theorem 2.2.6, the following holds*

$$\mathbb{E} \sup_{t\in[0,T], b\in\mathcal{C}} d_{\mathrm{MKW}}(\mu_t^{b,N}, f_t^b) \leq C\mathbb{E} d_{\mathrm{MKW}}(\mu_0^N, f_0) + CN^{-\gamma}, \tag{2.2.24}$$

with γ given by (2.2.23). Moreover, if p is large enough depending only on d, k then it holds that

$$\mathbb{E} \sup_{t\in[0,T], b\in\mathcal{C}} d_{\mathrm{MKW}}(\mu_t^{b,N}, f_t^b) \leq CN^{-\gamma}, \quad \gamma = \frac{1}{2 + \max(d, k)}. \tag{2.2.25}$$

As in the first order case we can obtain results in the Hölder scale of spaces. In this case, however, it makes more sense to consider the usual non-parabolic spaces.

Corollary 2.2.6. *Let $f_0 \in \mathrm{P}_p(\mathbb{R}^{2d})$ for some $2 \neq p > 1$. Let $\alpha \in (2/3, 1)$ and consider the class of α-Hölder functions defined by*

$$\mathcal{C} = \mathrm{C}([0,T]; \mathrm{C}_b^1(\mathbb{R}^d; \mathbb{R}^d)) \cap \{b : \|b\|_{\mathrm{C}^{0,\alpha}([0,T]\times\mathbb{R}^d;\mathbb{R}^d)} \leq C\}.$$

Then it holds that

$$\left\| \left[\sup_{t\in[0,T],b\in\mathcal{C}} d_{\mathrm{MKW}}(\mu_t^{b,N}, f_t^b) - c d_{\mathrm{MKW}}(\mu_0^N, f_0) \right]_+ \right\| \leq CN^{-\gamma}, \tag{2.2.26}$$

where

$$\gamma = \frac{1}{2 + \max(\frac{d+1}{\alpha}, \frac{d}{p-1})}.$$

As before this can be used to bound the expectation of the supremum.

Remark 2.2.13. *In the second order case the exponent γ is bounded below on the range of α considered, (for $p > 2$), i.e.*

$$\gamma = \frac{1}{2 + \frac{d+1}{\alpha}} > \frac{2}{7 + 3d} > 0. \tag{2.2.27}$$

Remark 2.2.14. *The lower bound of $2/3$ on α seems unlikely to be optimal in the sense that compactness methods would likely yield a Glivenko-Cantelli theorem for $\alpha \in (0, 1)$ but without an explicit convergence rate. We do not pursue such results here.*

2.3 Prior work and discussion

There has been much prior work on propagation of chaos of the particle system (2.1.1). This has been split between the *noisy case* considered in this manuscript, and the *noiseless case* where the driving Brownian motions are absent.

2.3.1 Lipschitz interactions

The first quantitative results in propagation of chaos are due to Dobrushin [49] in the noiseless case, and then later Sznitman in the case with noise considered in this work. Both these results rely on the interaction kernel K being Lipschitz continuous. Dobrushin observed that the empirical measure μ^N is a weak solution to the limit equation and then established that, under the assumption that K is Lipschitz, the limit equation is well-posed in the space of measures using the

MKW distance, from which propagation of chaos then follows from the convergence of initial data. In the case with noise, the empirical measure is no longer a weak solution to the limit equation. To get around this problem, Sznitman [185] developed a coupling method to prove propagation of chaos. This will be described in detail below in Section 2.3.4.

2.3.2 Singularity only at the origin

A subsequent line of enquiry was into interaction kernels which are Lipschitz apart from a single singularity where K or its derivative blows up in a specified manner. The case $K(x,y) = W(x-y)$ with W Lipschitz away from the origin[2] has received much attention as it models, for example, gravitational attraction.

2.3.2.1 Noiseless case

In [81] and later papers by various authors, propagation of chaos is established in the case without noise for interaction kernels satisfying the bounds $|W(x)| \leq C|x|^{-\alpha}$ and $|\nabla W(x)| \leq C|x|^{-\alpha-1}$ for some $\alpha < 1$. As in the proof of Dobrushin, these works rely on weak-strong stability estimates on the limit equation. However, to avoid the singularity at the origin, control must be obtained over the minimum distance between particles, and this requires specially prepared initial particle positions to control these distances at the initial time. A comprehensive review is given in [98].

2.3.2.2 Noisy case

Subsequently to proving the results of this work, the author was surprised to find that the noisy case has been considered *harder* than the noiseless case. This is in stark contrast with the comparison of existence and uniqueness theory for ODEs and SDEs where noise allows for less regular vector fields. When, however, one considers that to handle a singularity at the origin one must control the

[2] In these cases one must disallow self-interaction, so that the force term $b_t^N(X_t^{i,N})$ on the ith particle in (2.1.1) is replaced with $\frac{1}{N-1}\sum_{j=1,j\neq i}^N K(X_t^{i,N}, X_t^{j,N})$. The results of this chapter also apply to this case, see Section 2.3.6.1 below.

distances between particles and avoid near collisions, this makes more sense. Among recent work along these lines is [65] where propagation of chaos is obtained for a system similar to (2.1.1) with $W(z) = z|z|^{\alpha-1}$ for some $\alpha \in (0,1)$, (so W is α-Hölder continuous). Another recent work is [82] where the 1-dimensional Vlasov-Poisson-Fokker-Planck equation is considered, and the interaction kernel is the sign function, i.e. constant except for a jump at the origin. As in the works mentioned in the previous paragraph, the proof in [65] uses control over particle distances. A review is given in [102].

2.3.3 Bounded interactions or bounded potentials

In a recent work [100] an intriguing combinatorial argument is made to prove propagation of chaos for systems with *bounded* interaction kernels $W(z) \in \mathrm{L}^\infty$, later extended to bounded potentials [101] (roughly speaking $W(z) \in \mathrm{W}^{-1,\infty}$) in both the noisy and noiseless cases under the condition that $\operatorname{div} W = 0$. These works rely on controlling the *relative-entropy* between the solution to the N-particle Liouville equation and the limit solution. An advantage of these works over the results in this chapter in the $K(x,y) = W(x-y)$ case is that the assumptions on the interaction kernel are weaker in the sense that L^∞ (even $\mathrm{W}^{-1,\infty}$) rather than Hölder regularity is asked. However, this comes at the cost of assuming that W is divergence free, and rather surprisingly, rather non-generic assumptions on the initial datum f_0, which cannot be taken to be smooth with compact support, for example.

2.3.4 The coupling method of Sznitman

To prove propagation of chaos for Lipschitz interactions K in the noisy case Sznitman introduced a coupling method, where the particles (2.1.1) are coupled to an auxiliary particle system with the vector field b^N replaced by the vector field of the limit equation b^∞. In the notations of (2.2.5), the auxiliary particle system is $(X^{b^\infty,i,N})_{i=1}^N$. We give a heuristic description of the proof below.

2.3.4.1 Heuristic description

By the triangle inequality we observe that

$$\mathbb{E}d(\mu_t^N, f_t) \leq \mathbb{E}d(\mu_t^N, \mu_t^{b^\infty,N}) + \mathbb{E}d(\mu_t^{b^\infty,N}, f_t). \tag{2.3.1}$$

The second term is the expected difference between the empirical measure of N i.i.d. samples from their law f_t, and so tends to zero as $N \to \infty$ by the law of large numbers. Using Lipschitz continuity of the vector field b^∞ we obtain the bound

$$|X_t^{i,N} - X_t^{b^\infty,i,N}| \leq \|\nabla b^\infty\|_{\mathrm{L}^\infty} \int_0^t |X_s^{i,N} - X_s^{b^\infty,i,N}|\, ds + \int_0^t |b_s^N(X_s^{i,N}) - b^\infty(X_s^{i,N})|\, ds,$$

and one concludes via the Grönwall inequality that

$$d(\mu_t^N, \mu_t^{b^\infty,N}) \leq e^{t\|\nabla b^\infty\|_{\mathrm{L}^\infty}} \frac{1}{N} \sum_{i=1}^N \int_0^t |b_s^N(X_s^{i,N}) - b^\infty(X_s^{i,N})|\, ds.$$

That is, the particle system depends smoothly on the vector field. Then one uses Lipschitz continuity of K to obtain that the vector field depends smoothly on the particle positions, and closes the argument.

2.3.4.2 Limitations

This coupling method relies heavily on stability estimates on the particle system, and uses no stability estimates on the limit PDE. This can be seen in (2.3.1) where for the first term, one uses stability estimates, and on the second term, the law of large numbers is used. By coupling in this way, we are philosophically viewing the limit PDE as a perturbation of the particle system; viewing a smoother system as a perturbation of a rougher system.

2.3.5 A new coupling method

To get around the limitation of the coupling method above, we reverse the roles of the particle system and the limit PDE in (2.3.1). We wish to apply stability

estimates on the limit equation to the second term in (2.3.1) and the law of large numbers to the first term. This way we make better use of the two powerful tools at our disposal: *estimates on parabolic PDEs* and *the law of large numbers*.

To apply the law of large numbers to the first term, we must necessarily couple with a continuum object and not a discrete particle system. The only choice is to couple with $f_t^{b^N}$ defined by (2.2.7), that is, with the limit PDE with the vector field b^∞ replaced with vector field of the particle system b^N. (Note that $f_t^{b^N}$ is a random variable, even though it is a continuum object). Again the discussion below is heuristic. We refer the reader to the proof of Theorem 2.2.1 in Section 2.5 below for a more complete and rigorous presentation.

2.3.5.1 Heuristic description

By the triangle inequality we have

$$\mathbb{E}d(\mu_t^N, f_t) \leq \mathbb{E}d(\mu_t^N, f_t^{b^N}) + \mathbb{E}d(f_t^{b^N}, f_t). \tag{2.3.2}$$

This can be rewritten as

$$\mathbb{E}d(\mu_t^N, f_t) \leq \mathbb{E}d(\mu_t^{b^N,N}, f_t^{b^N}) + \mathbb{E}d(f_t^{b^N}, f_t^{b^\infty}).$$

Using stability estimates on the limit equation one readily obtains that it is sufficient to bound the first term on the right hand side by something that tends to zero as $N \to \infty$.

For the first term we wish to apply the law of large numbers. We have, however, a problem. The particle system is identically distributed, but not independent, so we cannot apply the law of large numbers directly. Moreover, we know very little about b^N. Indeed, our lack of knowledge of how to estimate b^N was our motivation for constructing this method. Because of this, we give up all hope of understanding b^N, and instead use the trivial bound

$$d(\mu_t^{b^N,N}, f_t^{b^N}) \leq \sup_{b \in \mathcal{C}} d(\mu_t^{b,N}, f_t^b)$$

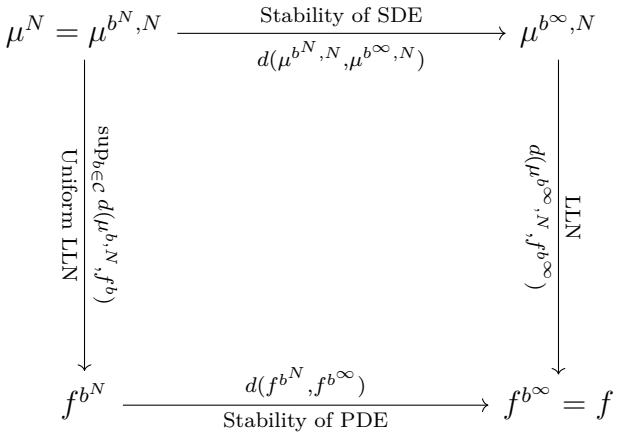


Figure 2.1: Illustration of the coupling method of Sznitman and the coupling method proposed in this chapter. To compare μ^N and f one can either go right then down, which is the coupling method of Sznitman, or down then right, which is the coupling method proposed here.

where $b \in \mathcal{C}$ ranges over all possible vector fields.[3] This coupling and the coupling of Sznitman are illustrated in Fig. 2.1.

One is then left with the problem of bounding $\mathbb{E}\sup_{b\in\mathcal{C}} d(\mu_t^{b,N}, f_t^b)$. If the supremum where outside the expectation this would be easy, as it is the empirical measure of N i.i.d. samples compared to their law f_t^b and the usual law of large numbers applies. In this way, we have exchanged the non-independence of the particles with taking a supremum over a very large set. This technique is commonly used in proving consistency of estimators in theoretical statistics [193], but to the authors knowledge has not been applied to the problem of propagation of chaos in this way before.

That this supremum can be bounded (Theorem 2.2.4) is perhaps surprising. The proof crucially relies on the existence of a differentiable stochastic flow associated with the SDE (2.2.5) for a single particle.

[3] Of course, in practice $\mathcal{C}$ will not be all possible vector fields, but merely those in some norm bounded set. Thus there will be some asymptotically (as $N \to \infty$) small chance that $b \notin \mathcal{C}$, which must be handled separately. We omit this here for brevity.

2.3.6 Discussion

2.3.6.1 Simple extensions

No self-interaction The propagation of chaos result Theorem 2.2.1 can be easily extended to the case of no self-interaction, where $b_t^N(X_t^{i,N})$ on the ith particle in (2.1.1) is replaced with $\frac{1}{N-1}\sum_{j=1,j\neq i}^N K(X_t^{i,N}, X_t^{j,N})$. This may be done by considering instead the interaction kernel $\tilde{K}$ given by

$$\tilde{K}(x,y) = K(x,y) - K(x,x).$$

Note that as K is always at least bounded and continuous, there is no problem defining $K(x,x)$ and it will be uniformly bounded.

Multi-particle interactions Theorem 2.2.1 also easily extends to the case where b^N is instead given by

$$b_t^N(x) = \frac{1}{N^m}\sum_{i_1,\dots,i_m=1}^N K(x, X_t^{i_1,N}, \dots, X_t^{i_m,N})$$

for $K \in \mathrm{C}^{0,\alpha}$. All the additional work happens at the PDE level and is straightforward.

2.3.7 Open questions

2.3.7.1 The curse of dimensionality

The existence of differentiable stochastic flows for SDEs with Hölder continuous vector fields (see e.g. [63, 12, 146, 59]) and their use in this chapter leads to an obvious question:

Can one use these stochastic flow results on the particle system (2.1.1) *and adapt the coupling method of Sznitman directly?*

In this work we have avoided this by using only the existence of a differentiable stochastic flow for a *single particle*, a system of fixed dimension d. The main ob-

stacle in applying stochastic flow results to the whole system is that the dimension of the particle system (2.1.1) is Nd which blows up as $N \to \infty$. Answering the above question would require a more careful analysis that kept track of how the constants in the proofs of these papers depend upon the dimension. It is conceivable that the special structure of (2.1.1) could be exploited to obtain bounds on these constants uniformly in N. The author plans to consider this in future.

We do remark, however, that the approach taken in this chapter has conceptual advantages and brings new insight into the problem of establishing propagation of chaos.

2.3.7.2 The $\mathrm{C}^{0,0+}$ barrier

The results of this chapter show that, in the presence of noise, the regularity barrier of Lipschitz continuity of K for quantitative propagation of chaos can be reduced to K being Hölder continuous. However, stochastic flows are known to exist for vector fields in L^p, $p > d$ (see [12]). This leads to the following question:

Can the barrier in the noisy case be reduced to $K \in \mathrm{W}^{s,p}$, $p > d$, $s > 0$*?*

The method used in this work fails in this case, as it requires the vector fields to be uniformly bounded for the key estimate Corollary 2.4.1.

2.4 Empirical process & Glivenko-Cantelli

In this section we prove the Glivenko-Cantelli results (Theorems 2.2.4 and 2.2.6) and Proposition 2.4.2 which will be used in the proof of the propagation of chaos result in Section 2.5. Before we move on to these results we will begin by establishing that the stochastic process $X^{b,i,N}$ is almost surely continuous.

2.4.1 The stochastic process

In this subsection we will prove that the stochastic process $(X_t^{b,i,N})_{i=1}^N$ indexed by $b \in \mathcal{C}$ has a continuous modification (Theorems 2.2.3 and 2.2.5) and show that it is impossible to construct a continuous modification of the same process indexed by the larger set $\tilde{\mathcal{C}}$ (Proposition 2.2.1). As the proof in the second order case is no harder, we leave it to the reader. Furthermore, we may without loss of generality consider the single particle ($N = 1$) case due to independence of the particles. For this reason we drop the i, N indices in this subsection.

Key to understanding both continuity for $\mathcal{C}$ and discontinuity for $\tilde{\mathcal{C}}$ is the observation that if $Y_t = X_t - B_t$ and X_t solves (2.4.3), then Y_t solves

$$dY_t = b_t(Y_t + B_t)dt, \quad Y_0 = X_0, \tag{2.4.1}$$

which is an ODE with drift vector field $\tilde{b}_t(x, \omega) = b_t(x + B_t(\omega))$.

Proof of Proposition 2.2.1. We argue by contradiction. Suppose that an almost surely continuous version exists, and without loss of generality that $d = 1$. To start with assume that $f_0 = \delta_0$. Define the random vector field $\tilde{b}_t(x, \omega) = \min(|x - B_t(\omega)|^\alpha, 1)$. Then by the computation above we deduce that $Y_t^{\tilde{b}} = X_t^{\tilde{b}} - B_t$ almost surely solves the ODE

$$dY_t = \min(|Y_t|^\alpha, 1)dt, \quad Y_0 = 0. \tag{2.4.2}$$

Note that this does not uniquely determine Y as the above ODE does not have unique solutions. However, as the process is continuous we can identify $Y_t^{\tilde{b}}$ as the unique limit of any sequence of Y^{b^n} with (random) $b^n \in \mathrm{C}_b^1$ and $b^n \to \tilde{b}$ in $\mathrm{L}^{-r,\infty}(\mathbb{R}^d; \mathbb{R}^d)$ almost surely. As these approximating vector fields are in C_b^1 the path $X_t^{b^n}$ is unique for each n. But we can easily construct a sequence b^n such that Y^{b^n} to converge to either the zero solution to (2.4.2) or a non-zero solution, contradicting uniqueness of this limit.

Extending this proof to more general initial conditions than $f_0 = \delta_0$ may be done by replacing the function $\min(|Y_t|^\alpha, 1)$ with a vector field in $\mathrm{C}^{0,\alpha}$ that exhibits non-uniqueness for the corresponding ODE at every point in $\mathbb{R}$. We leave this to

the reader. □

Proof of Theorem 2.2.3. Away from a fixed null set, $\mathcal{N}$ say, B_t is a continuous path on $[0,T]$. By the computation around Eq. (2.4.1), we have that for any $b \in \mathcal{C}$, $Y_t^b = X_t^b - B_t$ is the solution to a random ODE, and the random vector field is continuous and has spatial Lipschitz constant bounded by $L_b := \|b\|_{\mathrm{C}([0,T];\mathrm{C}_b^1(\mathbb{R}^d;\mathbb{R}^d))}$ away from the null set $\mathcal{N}$. Hence we can solve this ODE to construct Y_t^b (and thus also X_t^b) uniquely. Moreover, by standard Grönwall estimates on the solution we deduce that for any $\tilde{b} \in \mathcal{C}$ it holds that

$$\sup_{t\in[0,T]} |X_t^b - X_t^{\tilde{b}}| \leq T\exp(L_b T)\left\|b-\tilde{b}\right\|_{\mathrm{L}^\infty([0,T];\mathrm{L}^\infty(\mathbb{R}^d;\mathbb{R}^d))}$$

away from $\mathcal{N}$. To complete the proof it now suffices to replace this L^∞ estimate with an $\mathrm{L}^{-r,\infty}$ estimate. This may be done by a simple localisation argument. We leave this to the reader. □

Remark 2.4.1. *Although as evidenced by Proposition 2.2.1 the process indexed by $\tilde{\mathcal{C}}$ cannot have an almost surely continuous version, and in the proof of continuity (Theorem 2.2.3) of the process indexed by $\mathcal{C}$ we used the C_b^1 bound, we nevertheless have uniform 'Lipschitz continuity at each point' in the sense that, due to Corollary 2.4.1 (presented below), we have*

$$\sup_{b\in\tilde{\mathcal{C}}} \mathbb{E} \sup_{t\in[0,T],\tilde{b}\in\mathcal{C}} \frac{|X_t^b - X_t^{\tilde{b}}|}{\left\|b-\tilde{b}\right\|} < \infty$$

where the norm on $b-\tilde{b}$ is $\mathrm{L}^{-r,\infty}([0,T]\times\mathbb{R}^d;\mathbb{R}^d)$. It is this estimate that will be key to the later analysis, and we will never use the C_b^1 norm of the vector fields considered.

2.4.2 Estimates on the SDEs

Before we begin the proof of Theorem 2.2.4 proper, we will obtain some preliminary estimates on the SDEs:

$$dX_t = b_t(X_t)dt + dB_t \tag{2.4.3}$$

and

$$\begin{cases} dX_t = V_t dt, \\ dV_t = -b(X_t)dt - \kappa V_t dt + dB_t. \end{cases} \tag{2.4.4}$$

2.4.2.1 Growth bounds

We first obtain some simple a priori growth estimates for (2.4.3) and (2.4.4) which will be used throughout the sequel.

Lemma 2.4.1. *Let X solve (2.4.3) then*

$$\sup_{t\in[0,T]} |X_t| \leq |X_0| + C \|b\|_{\mathrm{L}^\infty([0,T]\times\mathbb{R}^d;\mathbb{R}^d)} + \sup_{t\in[0,T]} |B_t|.$$

Let (X, V) solve (2.4.4) then

$$\sup_{t\in[0,T]} (|X_t| + |V_t|) \leq C(|X_0| + |V_0| + \|b\|_{\mathrm{L}^\infty([0,T]\times\mathbb{R}^d;\mathbb{R}^d)} + \sup_{t\in[0,T]} |B_t|).$$

As a consequence, $\sup_{t\in[0,T]} |X_t|$ (respectively $\sup_{t\in[0,T]}(|X_t| + |V_t|)$) possesses as many moments as $|X_0|$ (respectively $|X_0| + |V_0|$).

Proof. The first claim on (2.4.3) is immediate from the definition of solution in integral form. For the first claim on (2.4.4) we first estimate

$$\begin{aligned} |X_t| &\leq |X_0| + \int_0^t |V_s|\, ds \\ |V_t| &\leq |V_0| + t \|b\|_{\mathrm{L}^\infty([0,T]\times\mathbb{R}^d;\mathbb{R}^d)} + \sup_{s\in[0,t]} |B_s| + |\kappa| \int_0^t |V_s|\, ds \end{aligned}$$

and then conclude with the Grönwall inequality on $|X_t| + |V_t|$. The remaining claims now follow from the triangle inequality and well known results on Brownian motion (see e.g. [148]). We omit the details. □

2.4.2.2 Reference processes

Next we define a 'reference process' in each of the first and second order cases. This process will have the property that the difference between it and our actual

process will be sub-Gaussian. For the first order case we define $(\widetilde{X}_t^{i,N})_{i=1}^N$ as

$$\widetilde{X}_t^{i,N} = X_0^{i,N}, \qquad i = 1, \dots, N, \qquad t \in [0,T]. \tag{2.4.5}$$

While in the second order case we instead define $(\widetilde{X}^{i,N}, \widetilde{V}_t^{i,N})_{i=1}^N$ as the solution to the following ODE with random initial condition:

$$\begin{aligned} d\widetilde{X}_t^{i,N} &= \widetilde{V}_t^{i,N} dt, \\ d\widetilde{V}_t^{i,N} &= -\kappa \widetilde{V}_t^{i,N} dt, \qquad i = 1, \dots, N \\ (\widetilde{X}_0^{i,N}, \widetilde{V}_0^{i,N}) &= (X_0^{i,N}, V_0^{i,N}). \end{aligned} \tag{2.4.6}$$

In both first and second order cases the reference process is nothing other than the solution to the corresponding SDE with $b = 0$ and the driving noise removed.

Being a linear ODE, the equation (2.4.6) can be explicitly solved to give

$$(\widetilde{X}_t^{i,N}, \widetilde{V}_t^{i,N}) = \left(X_0^{i,N} + V_0^{i,N} \frac{1 - e^{-\kappa t}}{\kappa}, V_0^{i,N} e^{-\kappa t} \right), \qquad i = 1, \dots, N. \tag{2.4.7}$$

We further define (in each case) the empirical measure corresponding to the reference process as $\widetilde{\mu}_t^N$ and the common law of each reference particle as $\widetilde{f}_t$. (Although $\widetilde{f}_t$ is the solution to a transport equation, we do not have explicit need of this fact.)

As discussed above, the reason we consider these reference processes is that the increment between the stochastic process we care about and the reference process is sub-Gaussian, even though each individual process may not be sub-Gaussian (which will be the case if the initial measure f_0 is not sub-Gaussian).

Lemma 2.4.2. *Let $Z_t^{b,i,N} = X_t^{b,i,N} - \widetilde{X}_t^{i,N}$ for the first order case (resp. $Z_t^{b,i,N} = (X_t^{b,i,N} - \widetilde{X}_t^{i,N}, V_t^{b,i,N} - \widetilde{V}_t^{i,N})$ for the second order case), where $X_t^{b,i,N}$ solves (2.2.5) for $b \in \mathcal{C}$ and $\widetilde{X}_t^{i,N}$ is the reference process defined by (2.4.5), (resp. $(X_t^{b,i,N}, V^{b,i,N})$ solves (2.2.17) for $b \in \mathcal{C}$ and $(\widetilde{X}_t^{i,N}, \widetilde{V}_t^{i,N})$ is the reference process defined by (2.4.6).) Then for each i, we have the following almost sure bound:*

$$\sup_{t \in [0,T]} |Z_t^{b,i,N}| \leq C \left(1 + \sup_{t \in [0,T]} |B_t^{i,N}| \right),$$

and the following time increment bound for any $t \in [0,T], \varepsilon > 0$,

$$\mathbb{E} \sup_{s\in[0,T],|s-t|^{1/3}\le\varepsilon} |Z_t^{b,i,N} - Z_s^{b,i,N}| \le C\varepsilon.$$

Proof. We first prove the almost sure bounds. The first inequality in the first order case follows directly from the integral form of the SDE (2.4.3), noting that $\widetilde{X}_t^{i,N}$ is nothing other than the initial condition $X_0^{i,N}$. For the second order case we first observe that

$$d(e^{\kappa t}V_t) = e^{\kappa t}b_t(X_t)dt + e^{\kappa t}dB_t, \tag{2.4.8}$$

(where we have omitted the b, i, N indices for brevity), so that

$$\begin{aligned}(e^{\kappa t}V_t - V_0) &= \int_0^t e^{\kappa s}b_s(X_s)\,ds + \int_0^t e^{\kappa s}dB_s \\ &= \int_0^t e^{\kappa s}b_s(X_s)\,ds + e^{\kappa t}B_t - \kappa\int_0^t e^{\kappa s}B_s\,dt\end{aligned} \tag{2.4.9}$$

by stochastic integration by parts. From this the bound on $|V_t - \widetilde{V}_t| = |V_t - e^{-\kappa t}V_0|$ is easily deduced as $b \in \mathcal{C}$ is uniformly bounded. The bound on $|X_t - \widetilde{X}_t|$ is now deduced by integrating this bound on $[0,t]$.

Now we prove the time increment estimates. For the first order system we have, from the integral form of the ODE,

$$\begin{aligned}\sup_{s\in[0,T],|s-t|^{1/3}\le\varepsilon} |(X_t - \widetilde{X}_t) - (X_s - \widetilde{X}_s)| \le \varepsilon^2 \sup_{b\in\mathcal{C}} \|b\|_{\mathrm{L}^\infty([0,T]\times\mathbb{R}^d;\mathbb{R}^d)} \\ + \sup_{s\in[0,T],|t-s|^{1/3}\le\varepsilon} |B_s - B_t|\end{aligned}$$

The supremum over $\mathcal{C}$ is bounded by a constant. That the remaining part has expectation bounded by a constant times ε can either be seen as a consequence of the law of the iterated logarithm (see e.g. [148]) or that the 1/3-Hölder norm of Brownian motion has finite expectation (see e.g. [195]).

The corresponding estimate for the second order case is similar using instead the integral form of (2.4.9). We leave it to the reader. □

Using the reference processes we can obtain an estimate on the Wasserstein dis-

tance $d_{\mathrm{MKW}}(\mu_t^{b,N}, f^b)$ using the following inequality. Note that, as is evident from its proof, establishing (2.4.10) requires no properties of the particle system other than the relationship of the empirical measures to the laws.

Lemma 2.4.3. *Let $X_t^{b,i,N}$ solve (2.2.5) for $b \in \mathcal{C}$ and $\widetilde{X}_t^{i,N}$ be the reference process defined by (2.4.5). Then the following holds:*

$$\begin{aligned} d_{\mathrm{MKW}}(\mu_t^{b,N}, f_t^b) &\leq d_{\mathrm{MKW}}(\widetilde{\mu}_t^N, \widetilde{f}_t) + \\ &+ \sup_{h \in \mathrm{Lip1}} \left(\frac{1}{N} \sum_{i=1}^{N} (h(X_t^{b,i,N}) - h(\widetilde{X}_t^{i,N})) - \mathbb{E}(h(X_t^{b,i,N}) - h(\widetilde{X}_t^{i,N})) \right). \end{aligned} \tag{2.4.10}$$

In the second order case where $(X_t^{b,i,N}, V^{b,i,N})$ solve (2.2.17) for $b \in \mathcal{C}$ and $(\widetilde{X}_t^{i,N}, \widetilde{V}_t^{i,N})$ is the reference process defined by (2.4.6) the corresponding inequality to (2.4.10) holds, i.e. with $X^{b,i,N}$ replaced with $(X^{b,i,N}, V^{b,i,N})$ and so on. We omit writing this inequality for brevity.

Proof. We only give the proof in the first order case for brevity. The second order case is analogous and we leave it to the reader. We have

$$\begin{aligned} d_{\mathrm{MKW}}(\mu_t^N, f_t^b) &= \sup_{h \in \mathrm{Lip1}} \left(\frac{1}{N} \sum_{i=1}^{N} h(X_t^{b,i,N}) - \mathbb{E}h(X_t^{b,i,N}) \right) \\ &\leq \sup_{h \in \mathrm{Lip1}} \left(\frac{1}{N} \sum_{i=1}^{N} (h(X_t^{b,i,N}) - h(\widetilde{X}_t^{i,N})) - \mathbb{E}(h(X_t^{b,i,N}) - h(\widetilde{X}_t^{i,N})) \right) \\ &\quad + \sup_{h \in \mathrm{Lip1}} \left(\frac{1}{N} \sum_{i=1}^{N} h(\widetilde{X}_t^{i,N}) - \mathbb{E}h(\widetilde{X}_t^{i,N}) \right). \end{aligned}$$

The final supremum is nothing other that $d_{\mathrm{MKW}}(\widetilde{\mu}_t^N, \widetilde{f}_t)$. The proof is complete. □

As the reference processes are simple (by choice) and their evolution is deterministic, we have the following control over the distance of the reference empirical measure to the law.

Lemma 2.4.4. *In the first and second order cases the following holds:*

$$\sup_{t \in [0,T]} d_{\mathrm{MKW}}(\widetilde{\mu}_t^N, \widetilde{f}_t) \leq c d_{\mathrm{MKW}}(\mu_0^N, f_0). \tag{2.4.11}$$

Moreover, in the first order case c may be taken to be equal to 1.

Proof. For the first order case the claim (with $c = 1$) is obvious from the definitions. For the second order case we argue directly by evolving an initial coupling between f_0 and μ_0^N along the trajectories of the ODE flow given by (2.4.6). Indeed, let $\pi_0 \in \mathrm{P}(\mathbb{R}^{2d} \times \mathbb{R}^{2d})$ be any coupling between μ_0^N and f_0, and define π_t as the pushforward of π_0 by the (autonomous, deterministic, smooth) flow ϕ_t of the ODE (2.4.6). Then $\pi_t \in \mathrm{P}(\mathbb{R}^{2d} \times \mathbb{R}^{2d})$ is a coupling of $\widetilde{\mu}_t^N$ and $\widetilde{f}_t$ and we have the bound

$$\begin{aligned}\int |(x^1, v^1) - (x^2, v^2)| \, d\pi_t(x^1, v^1, x^2, v^2) \\ &= \int |\phi_t(x^1, v^1) - \phi_t(x^2, v^2)| \, d\pi_0(x^1, v^1, x^2, v^2) \\ &\leq \|\nabla \phi_t\|_{\mathrm{L}^\infty(\mathbb{R}^{2d}; \mathbb{R}^{2d})} \int |(x^1, v^1) - (x^2, v^2)| \, d\pi_0(x^1, v^1, x^2, v^2).\end{aligned}$$

The ODE flow ϕ_t is uniformly Lipschitz in (x, v) over times $t \in [0, T]$, so the L^∞ norm above is bounded by a constant. Thus $d_{\mathrm{MKW}}(\widetilde{f}_t, \widetilde{\mu}_t^N)$ is bounded by a constant times the final integral in the above display. The claim of the lemma now follows by taking the infimum over all couplings π_0 between f_0 and μ_0^N. □

2.4.2.3 Local Lipschitz dependence upon the field

We now recall that in both the first and second order cases the SDE generates a differentiable stochastic flow. The first order case is established in [63] (see also [12, 146, 59] for results along the same lines). For the second order case see [201] (see also [60]). We have need of a simple corollary that provides global bounds in space with a weight.

Theorem 2.4.1. *Let $b \in \mathrm{L}^\infty([0, T]; \mathrm{C}^{0,\alpha}(\mathbb{R}^d; \mathbb{R}^d))$, with $\alpha \in (0, 1)$ (resp. $\alpha \in (2/3, 1)$) then the SDE (2.4.3) (resp. (2.4.4)) generates a $\mathrm{C}^{1,\alpha'}$ stochastic flow (see Definition 2.1.9) $\phi_{s,t} : \mathbb{R}^d \to \mathbb{R}^d$ (resp. $\mathbb{R}^{2d} \to \mathbb{R}^{2d}$), for α' depending only upon α. For any $p \in [1, \infty)$ and $r > 0$ there is a constant $C_{p,r}$ depending only on p, r and $\|b\|_{\mathrm{L}^\infty([0,T]; \mathrm{C}^{0,\alpha}(\mathbb{R}^d; \mathbb{R}^d))}$ but not b itself, such that*

$$\left\| \sup_{0 \leq s \leq t \leq T} \|\nabla \phi_{s,t}\|_{\mathrm{L}^{-r,\infty}(\mathbb{R}^d; \mathbb{R}^{d \times d})} \right\|_p \leq C_{p,r} < \infty \tag{2.4.12}$$

(resp. the same claim with $\mathbb{R}^d$ replaced with $\mathbb{R}^{2d}$).

As discussed above, the existence of the stochastic flow is shown elsewhere, so it suffices to prove (2.4.12).

Proof of (2.4.12). We give the proof in the first order case, the second order proof being analogous. Local estimates of the form

$$\left\| \sup_{0\le s\le t\le T} \|\nabla\phi_{s,t}\|_{\mathrm{L}^\infty(Q;\mathbb{R}^{d\times d})} \right\|_p \le C$$

for $p \in [1,\infty)$ and Q a unit cube in $\mathbb{R}^d$ follow from [63], and as the bounds upon b are global, these estimates are uniform over unit cubes Q. Now let A_n for $n = 1, 2, 3, \dots$ be the annulus $\{2^n \le |x| \le 2^{n+1}\}$ and $A_0 = \{|x| \le 2\}$ be subsets of $\mathbb{R}^d$. Then it holds that

$$\sup_{0\le s\le t\le T} \|\nabla\phi_{s,t}\|_{\mathrm{L}^{-r,\infty}(\mathbb{R}^d;\mathbb{R}^{d\times d})} \le \sum_{n=0}^{\infty} 2^{-nr} \sup_{0\le s\le t\le T} \|\nabla\phi_{s,t}\|_{\mathrm{L}^\infty(A_n;\mathbb{R}^{d\times d})}. \tag{2.4.13}$$

Each A_n can be covered by m_n unit cubes $Q_{n,i}$ and m_n can be chosen to be at most $C2^{nd}$. Hence, it holds that

$$\left\| \|\nabla\phi_{s,t}\|_{\mathrm{L}^\infty(A_n;\mathbb{R}^{d\times d})} \right\|_p \le \left\| \max_{i=1}^{m_n} \|\nabla\phi_{s,t}\|_{\mathrm{L}^\infty(Q_{n,i};\mathbb{R}^{d\times d})} \right\|_p.$$

We apply the elementary inequality $\left\|\max_{i=1}^m |X_i|\right\|_p \le C_p m^{1/p} \max_{i=1}^m \|X_i\|$ which may be obtained by bounding the maximum of the X_i with the sum of the $|X_i|$. This yields

$$\left\| \|\nabla\phi_{s,t}\|_{\mathrm{L}^\infty(A_n;\mathbb{R}^{d\times d})} \right\|_p \le C_p 2^{nd/p}. \tag{2.4.14}$$

Therefore, combining (2.4.14) and (2.4.13) we obtain

$$\left\| \sup_{0\le s\le t\le T} \|\nabla\phi_{s,t}\|_{\mathrm{L}^{-r,\infty}(\mathbb{R}^d;\mathbb{R}^{d\times d})} \right\|_p \le C_p \sum_{n=0}^{\infty} 2^{nd/p-nr}$$

and this sum is convergent for all p sufficiently large. The estimate for smaller p then follows by bounding with the estimate for larger p. □

We obtain the following corollary of Theorem 2.4.1, which holds for both the first

and second order systems.

Corollary 2.4.1. *Let $b \in \mathcal{C}$, $f_0 \in \mathrm{P}_p(\mathbb{R}^d)$ (respectively $\mathrm{P}_p(\mathbb{R}^{2d})$) for some $p > r > 1$. Then there exists a random variable L with finite expectation uniform over $b \in \mathcal{C}$, such that for any $\tilde{b} \in \mathcal{C}$ and $t \in [0,T]$ it holds that*

$$|X_t^b - X_t^{\tilde{b}}| \leq L \int_0^t \left\| b_s - \tilde{b}_s \right\|_{\mathrm{L}^{-r,\infty}(\mathbb{R}^d;\mathbb{R}^d)} ds$$

in the first order case, and

$$|X_t^b - X_t^{\tilde{b}}| + |V_t^b - V_t^{\tilde{b}}| \leq L \int_0^t \left\| b_s - \tilde{b}_s \right\|_{\mathrm{L}^{-r,\infty}(\mathbb{R}^d;\mathbb{R}^d)} ds$$

respectively in the second order case (with a different L).

Proof. We give the proof in the first order case. The second order case is analogous and no harder. Let $\phi, \tilde{\phi}$ be the associated stochastic flows given by Theorem 2.4.1 and let

$$J = \sup_{0 \leq s \leq t \leq T} \sup_{|x| \leq \sup_{t \in [0,T]} |X_t^{\tilde{b}}|} |\nabla \phi_{s,t}(x)|.$$

Define the function $\psi(u) = (\phi_{t,u} \circ \tilde{\phi}_{0,u})(X_0)$. Then $\psi(0) = X_t^b$ and $\psi(t) = X_t^{\tilde{b}}$. We wish to estimate $|\psi(t) - \psi(0)| \leq \int_0^t \left| \frac{d\psi}{ds} \right| ds$, but it is not immediately clear how to evaluate the derivative due to the presence of the non-differentiable Brownian motions. Instead we prove the moral equivalent using Riemann sums. Let $0 = t_0 < t_1 < \cdots < t_n = t$ be a partition of $[0,t]$ of maximum width h. Now consider

$$\begin{aligned}
|\psi(t_{k+1}) - \psi(t_k)| &= |(\phi_{t_{k+1},t} \circ \tilde{\phi}_{0,t_{k+1}})(X_0) - (\phi_{t_k,t} \circ \tilde{\phi}_{0,t_k})(X_0)| \\
&= |(\phi_{t_{k+1},t} \circ \tilde{\phi}_{0,t_{k+1}})(X_0) - (\phi_{t_{k+1},t} \circ \phi_{t_k,t_{k+1}} \circ \tilde{\phi}_{0,t_k})(X_0)| \\
&\leq J |\tilde{\phi}_{0,t_{k+1}}(X_0) - (\phi_{t_k,t_{k+1}} \circ \tilde{\phi}_{0,t_k})(X_0)| \\
&\leq J |(\tilde{\phi}_{t_k,t_{k+1}} \circ \tilde{\phi}_{0,t_k})(X_0) - (\phi_{t_k,t_{k+1}} \circ \tilde{\phi}_{0,t_k})(X_0))| \\
&\leq J \left| \int_{t_k}^{t_{k+1}} \tilde{b}_s(\tilde{\phi}_{t_k,s}(X_{t_k}^{\tilde{b}})) - b_s(\phi_{t_k,s}(X_{t_k}^{\tilde{b}}))\, ds \right| \\
&\leq J \int_{t_k}^{t_{k+1}} |\tilde{b}_s(X_{t_k}^{\tilde{b}}) - b_s(X_{t_k}^{\tilde{b}})|\, ds \\
&\quad + J \int_{t_k}^{t_{k+1}} |b_s(\tilde{\phi}_{t_k,s}(X_{t_k}^{\tilde{b}})) - b_s(\phi_{t_k,s}(X_{t_k}^{\tilde{b}}))|\, ds,
\end{aligned} \tag{2.4.15}$$

Note that we have the simple estimate

$$\sup_x |\tilde{\phi}_{u,s}(x) - \phi_{u,s}(x)| \leq \int_u^s \|b_\tau\|_{\mathrm{L}^\infty(\mathbb{R}^d;\mathbb{R}^d)} + \left\|\tilde{b}_\tau\right\|_{\mathrm{L}^\infty(\mathbb{R}^d;\mathbb{R}^d)} \, d\tau \leq C|s-u|.$$

Therefore, as b_s is α-Hölder continuous, the final integral in (2.4.15) is bounded by $CJ|t_{k+1} - t_k|^{1+\alpha}$ for some constant C, and when we take the partition width h to zero in the sum in the display below, this term contributes nothing. Hence,

$$\begin{aligned} |X_t^b - X_t^{\tilde{b}}| &= |\psi(t) - \psi(0)| \\ &\leq \lim_{h\to 0} \sum_{k=0}^{n-1} |\psi(t_{k+1}) - \psi(t_k)| \\ &= J \int_0^t |b_s(X_s^{\tilde{b}}) - \tilde{b}_s(X_s^{\tilde{b}})| \, ds. \end{aligned}$$

Next we note that

$$J \int_0^t |b_s(X_s^{\tilde{b}}) - \tilde{b}_s(X_s^{\tilde{b}})| \, ds \leq J \sup_{0\leq s\leq T} \left\langle X_s^{\tilde{b},i,N} \right\rangle^r \int_0^t \left\|b_s - \tilde{b}_s\right\|_{\mathrm{L}^{-r,\infty}(\mathbb{R}^d;\mathbb{R}^d)} \, ds.$$

Now define

$$\begin{aligned} L :&= J \sup_{0\leq s\leq T} \left\langle X_s^{\tilde{b},i,N} \right\rangle^r \\ &= \left(\sup_{0\leq s\leq T} \left\langle X_s^{\tilde{b},i,N} \right\rangle^r \right) \left(\sup_{0\leq s\leq t\leq T} \sup_{|x|\leq \sup_{t\in[0,T]} |X_t^{\tilde{b}}|} |\nabla \phi_{s,t}(x)| \right). \end{aligned}$$

Note that for any $\varepsilon > 0$,

$$L \leq \left(\sup_{0\leq t\leq T} \left\langle X_t^{\tilde{b}} \right\rangle^{r+\varepsilon} \right) \sup_{0\leq s\leq t\leq T} \|\nabla \phi_{s,t}\|_{\mathrm{L}^{-\varepsilon,\infty}}$$

so that by Hölder's inequality

$$\mathbb{E}L \leq \left\| \sup_{0\leq t\leq T} \left\langle X_t^{\tilde{b}} \right\rangle \right\|_p^\varepsilon \left\| \sup_{0\leq s\leq t\leq T} \|\nabla \phi_{s,t}\|_{\mathrm{L}^{-\varepsilon,\infty}} \right\|_q$$

with $(1/p') + (1/q) = 1$ with $p'(r+\varepsilon) = p$. (We ensure $p' > 1$ by taking ε sufficiently small and using $r < p$.) These are both finite by Lemma 2.4.1 and Theorem 2.4.1 respectively. The proof is complete. □

2.4.3 The empirical process theory argument

To control the supremum in (2.2.12) we will use the following key proposition.

Recall that a semi-metric space is a metric space without the triangle inequality. The definition of metric entropy extends without modification to semi-metric spaces. For a random variable X valued in a Banach space V we say that X is centred if $\mathbb{E}g(X) = 0$ for all g in the dual of V.

Proposition 2.4.1. *Let $(\mathcal{X}, d)$ be a totally bounded semi-metric space and $(\mathrm{V}, \|\cdot\|_{\mathrm{V}})$ be a separable Banach space. Let $\varphi : \mathcal{X} \to \mathrm{V}$ be a centred random map, Y a non-negative sub-Gaussian random variable and $(G(x,\varepsilon))_{x\in\mathcal{X},\varepsilon\in(0,1]}$ be a family of random variables. Let $(\varphi^{1,N}, Y^{i,N}, G^{i,N}), \ldots, (\varphi^{N,N}, Y^{N,N}, G^{N,N})$ be i.i.d. copies of (φ, Y, G) and assume the following:*

(i) ***Metric entropy bounds:*** *The semi-metric space $\mathcal{X}$ obeys the following bound*

$$H(\varepsilon, \mathcal{X}, d) \leq C_h \varepsilon^{-k}$$

for constants $C_h, h > 0$.

(ii) ***'Pointwise Lipschitz' condition:*** *For every $x \in \mathcal{X}, \varepsilon \in (0,1]$ we have*

$$\sup_{d(x,\tilde{x})\leq\varepsilon} \|\varphi(x) - \varphi(\tilde{x})\|_{\mathrm{V}} \leq G(x,\varepsilon) \tag{2.4.16}$$

and there is a constant C_G such that for all $\varepsilon \in (0,1]$,

$$\sup_{x\in\mathcal{X}} \mathbb{E}G(x,\varepsilon) \leq C_G\varepsilon. \tag{2.4.17}$$

(iii) ***Dominating sub-Gaussians (envelope function):*** *We have*

$$\sup_{x\in\mathcal{X}} \|\varphi(x)\|_{\mathrm{V}} \leq Y$$

with $|||Y||| \leq C_Y$, and if $V \neq \mathbb{R}$ then also $Y \leq C_Y$ almost surely.

(iv) ***Pointwise law of large numbers:*** *We have*

$$\sup_{x\in\mathcal{X}} \mathbb{E}\left\| \frac{1}{N}\sum_{i=1}^{N} \varphi^{i,N}(x)\right\|_{\mathrm{V}} \leq C_{\mathrm{V}} N^{-1/2}.$$

Then it holds that

$$\left\| \sup_{x\in\mathcal{X}} \left\| \frac{1}{N}\sum_{i=1}^{N} \varphi^{i,N}(x)\right\|_{\mathrm{V}} \right\| \leq C(C_G + (C_Y + C_{\mathrm{V}})\sqrt{C_h})N^{-\gamma}, \quad \gamma = \frac{1}{2+k}.$$

Note that in the above proposition the usual case is that $V = \mathbb{R}$. In which case assumption (iv) follows from assumption (iii) and the usual law of large numbers under second moment conditions.

In order to prove this proposition and also for later proofs, we will need a couple of standard results on the sub-Gaussian norm.

Lemma 2.4.5 (Law of large numbers)**.** *Let $X_1, \ldots, X_N$ be i.i.d. centred sub-Gaussian random variables. Then*

$$\left\| \frac{1}{N}\sum_{i=1}^{N} X_i \right\| \leq C \left\| X_1 \right\| N^{-1/2}$$

for an absolute constant C.

The proof is a simple corollary of [196, Lemma 5.9.].

Lemma 2.4.6 (Orlicz maximal inequality)**.** *Let $X_1, \ldots, X_m$ be real sub-Gaussian random variables, not necessarily independent. Then*

$$\left\| \max_{i=1}^{m} |X_i| \right\| \leq C \max_{i=1}^{m} \left\| X_i \right\| \sqrt{\log(1+m)}$$

for an absolute constant C.

We refer the reader to [193, §2.2.] for details of the proof.

We will also need a variant of Talagrand's inequality for empirical processes [186].

Theorem 2.4.2 (Talagrand's inequality for Banach spaces)**.** *Let $(\mathrm{V}, \|\cdot\|_{\mathrm{V}})$ be*

a separable Banach space and $X_1, \dots, X_N$ be i.i.d. centered V*-valued random variables with $\|X_1\|_{\mathrm{V}} \leq 1$ almost surely. Then,*

$$\left\| \left\| \frac{1}{N} \sum_{i=1}^{N} X_i \right\|_{\mathrm{V}} - \mathbb{E} \left\| \frac{1}{N} \sum_{i=1}^{N} X_i \right\|_{\mathrm{V}} \right\| \leq C N^{-1/2}$$

For an absolute constant C.

Proof of Proposition 2.4.1. Let $\varepsilon > 0$ to be chosen and let $(x^m)_{m=1}^M$ be an ε-net of $\mathcal{X}$. By assumption (i) M may be taken to be at most $\exp(C_h \varepsilon^{-k})$. Let $m \in \{1, \dots, M\}$ be arbitrary, and $x \in \mathcal{X}$ be in the ε-ball centred at x^m. Then we have the bound

$$\left\| \frac{1}{N} \sum_{i=1}^{N} \varphi^{i,N}(x) \right\|_{\mathrm{V}} \leq \left\| \frac{1}{N} \sum_{i=1}^{N} (\varphi^{i,N}(x) - \varphi^{i,N}(x^m)) \right\|_{\mathrm{V}} + \left\| \frac{1}{N} \sum_{i=1}^{N} \varphi^{i,N}(x^m) \right\|_{\mathrm{V}}. \tag{2.4.18}$$

Consider the summands in the first term on the right hand side. By assumptions (ii) and (iii), we have

$$\begin{aligned} \left\| \varphi^{i,N}(x) - \varphi^{i,N}(x^m) \right\|_{\mathrm{V}} &\leq \min(2Y^{i,N}, G^{i,N}(x^m, \varepsilon)) \\ &=: \mathbb{E} \min(2Y^{i,N}, G^{i,N}(x^m, \varepsilon)) + A^{i,N,m} \end{aligned} \tag{2.4.19}$$

where $(A^{i,N,m})_{i=1}^N$ (defined by the last equality) are i.i.d. uniformly sub-Gaussian centered random variables. To control the last term in (2.4.18) we split into two cases. Firstly, if $\mathrm{V} \neq \mathbb{R}$ then we write the last term in (2.4.18) as

$$\begin{aligned} \left\| \frac{1}{N} \sum_{i=1}^{N} \varphi^{i,N}(x^m) \right\|_{\mathrm{V}} &= \mathbb{E} \left\| \frac{1}{N} \sum_{i=1}^{N} \varphi^{i,N}(x^m) \right\|_{\mathrm{V}} \\ &\quad + \left(\left\| \frac{1}{N} \sum_{i=1}^{N} \varphi^{i,N}(x^m) \right\|_{\mathrm{V}} - \mathbb{E} \left\| \frac{1}{N} \sum_{i=1}^{N} \varphi^{i,N}(x^m) \right\|_{\mathrm{V}} \right). \end{aligned}$$

Hence, by assumption (iv) and Talagrand's inequality (Theorem 2.4.2), we have the bound

$$\left\| \left\| \frac{1}{N} \sum_{i=1}^{N} \varphi^{i,N}(x^m) \right\|_{\mathrm{V}} \right\| \leq C(C_{\mathrm{V}} + C_Y) N^{-1/2}. \tag{2.4.20}$$

Secondly, if $\mathrm{V} = \mathbb{R}$ then we can directly apply the law of large numbers for

sub-Gaussian random variables (Lemma 2.4.5) to obtain that

$$\left\| \left\| \frac{1}{N} \sum_{i=1}^{N} \varphi^{i,N}(x^m) \right\|_{\mathrm{V}} \right\| \leq C C_Y N^{-1/2}. \tag{2.4.21}$$

Hence, by assumption (ii), (2.4.19) and whichever of (2.4.20) or (2.4.21) applies we have

$$\begin{aligned}
\sup_{x \in \mathcal{X}, d(x,x^m) \leq \varepsilon} \left\| \frac{1}{N} \sum_{i=1}^{N} \varphi^{i,N}(x) \right\|_{\mathrm{V}} &\leq \mathbb{E} \min(2Y^{i,N}, G^{i,N}(x^m, \varepsilon)) \\
&\quad + \underbrace{\left| \frac{1}{N} \sum_{i=1}^{N} A^{i,N,m} \right| + \left\| \frac{1}{N} \sum_{i=1}^{N} \varphi^{i,N}(x^m) \right\|_{\mathrm{V}}}_{=:B^m} \\
&\leq C_G \varepsilon + B^m.
\end{aligned}$$

By the law of large numbers for sub-Gaussian random variables (Lemma 2.4.5), the uniform sub-Gaussian bounds on $A^{i,N,m}$ and the above bounds on the average of $\varphi^{i,N}(x^m)$, we have

$$\left\| B^m \right\| \leq C \left\| A^{1,N,m} \right\| N^{-1/2} + C(C_{\mathrm{V}} + C_Y) N^{-1/2} \leq 4C(C_Y + C_{\mathrm{V}}) N^{-1/2}.$$

Putting the estimates over the ε-net together, and using the Orlicz maximal inequality (Lemma 2.4.6), we obtain

$$\begin{aligned}
\left\| \sup_{x \in \mathcal{X}} \left\| \frac{1}{N} \sum_{i=1}^{N} \varphi^{i,N}(x) \right\|_{\mathrm{V}} \right\| &\leq \left\| \max_{m=1,\dots,M} \sup_{x \in \mathcal{X}, d(x,x^m) \leq \varepsilon} \left\| \frac{1}{N} \sum_{i=1}^{N} \varphi^{i,N}(x) \right\|_{\mathrm{V}} \right\| \\
&\leq \left\| \max_{m=1,\dots,M} C_G \varepsilon + B^m \right\| \\
&\leq C_G \varepsilon + \left\| \max_{m=1,\dots,M} B^m \right\| \\
&\leq C_G \varepsilon + C \sqrt{\log(1+M)} \max_{i=1,\dots,M} \left\| B^m \right\| \\
&\leq C_G \varepsilon + C(C_Y + C_{\mathrm{V}}) \sqrt{C_h} \varepsilon^{-k/2} N^{-1/2}.
\end{aligned}$$

By choosing $\varepsilon = N^{-1/(2+k)}$ we obtain the claimed result. □

In addition to Theorems 2.2.4 and 2.2.6 we will also prove a proposition that will be used in the proofs of Theorems 2.2.1 and 2.2.2.

Proposition 2.4.2. *Let $f^b, \mu^{b,N}$ be as in (2.2.7),(2.2.6) in the first order case, and respectively (2.2.19),(2.2.18) in the second order case. Let $f_0 \in \mathrm{P}_p(\mathbb{R}^d)$ (respectively $f_0 \in \mathrm{P}_p(\mathbb{R}^{2d})$) for some $p > 1$. Let $\mathcal{C}$ be a bounded subset of $\mathcal{C}^\alpha$ for some $\alpha \in (0,1)$ (respectively $\alpha \in (2/3, 1)$) which satisfies*

$$H(\varepsilon, \mathcal{C}, \|\cdot\|_{\mathrm{L}^\infty([0,T];\mathrm{L}^{-r,\infty}(\mathbb{R}^d;\mathbb{R}^d))}) \leq C\varepsilon^{-k} \tag{2.4.22}$$

for some $r \in (1,p)$. Let $h : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a bounded function satisfying

$$\sup_{y,\delta\in\mathbb{R}^d,\delta\neq 0} \frac{\|h(\cdot, y+\delta) - h(\cdot, y)\|_{\mathrm{L}^{-r',q}(\mathbb{R}^d)}}{|\delta|^\beta} < \infty,$$

where $\beta \in (0,1]$, $q \in [1,\infty]$ and $r' > d/q$. Then we have

$$\left\| \sup_{b\in\mathcal{C}} \sup_{t\in[0,T]} \left\| \frac{1}{N}\sum_{i=1}^N h(\cdot, X_t^{b,i,N}) - \mathbb{E}h(\cdot, X_t^{b,i,N}) \right\|_{\mathrm{L}^{-r',q}(\mathbb{R}^d)} \right\| \leq CN^{-\gamma}, \quad \gamma = \frac{1}{2+\frac{k}{\beta}}.$$

Proof of Theorems 2.2.4 and 2.2.6 and Proposition 2.4.2. We begin by presenting the proofs of Theorems 2.2.4 and 2.2.6. We first note that by Lemmas 2.4.3 and 2.4.4 we have

$$\begin{aligned} &\left[\sup_{t\in[0,T],b\in\mathcal{C}} d_{\mathrm{MKW}}(\mu_t^{b,N}, f_t^b) - cd_{\mathrm{MKW}}(\mu_0^N, f_0) \right]_+ \leq \\ &\sup_{h\in\mathrm{Lip}1,t\in[0,T],b\in\mathcal{C}} \left(\frac{1}{N}\sum_{i=1}^N (h(X_t^{b,i,N}) - h(\widetilde{X}_t^{i,N})) - \mathbb{E}(h(X_t^{b,i,N}) - h(\widetilde{X}_t^{i,N})) \right), \end{aligned} \tag{2.4.23}$$

in the first order case (with $c = 1$) and the corresponding inequality in the second order case (with $c > 1$).

From here on we give the proof for the first order system (Theorem 2.2.4). The proof for the second order system (Theorem 2.2.6) is analogous (using instead the second order versions of the above lemmas) and we leave it to the reader. The proof follows from the application of Proposition 2.4.1 with a carefully chosen map φ and semi-metric space $(\mathcal{X}, d)$.

We set $(\mathcal{X}, d)$ to be

$$([0,T], |\cdot|^{1/3}) \times (\mathcal{C}, \|\cdot\|_{\mathrm{L}^\infty([0,T];\mathrm{L}^{-r,\infty}(\mathbb{R}^d;\mathbb{R}^d))}) \times (\mathrm{Lip}1, \|\cdot\|_{\mathrm{L}^{-p,\infty}(\mathbb{R}^d;\mathbb{R}^d)})$$

with the product metric, where $|\cdot|$ is the standard Euclidean norm. By assumption the metric entropy of the second space in the above display is bounded by $C\varepsilon^{-k}$. By the results in [54, 161] the metric entropy of the third space in the above display is bounded by $C\varepsilon^{-\min(d,d/(p-1))}$ (see Proposition 2.A.1 for details.) As the metric entropy of $([0,T],|\cdot|^{1/3})$ is logarithmic in ε, the metric entropy of (X,d) is controlled, using Lemma 2.A.1 by

$$H(\varepsilon, X, d) \leq C\varepsilon^{-d/(p-1)} + C\varepsilon^{-k} + C\log(1/\varepsilon) \leq C\varepsilon^{-\max(d,d/(p-1),k)}. \tag{2.4.24}$$

We define $\varphi^{i,N} : X \to \mathbb{R}$ for $i = 1, \dots, N$, by

$$\varphi^{i,N}(t,b,h) = (h(X_t^{b,i,N}) - h(\widetilde{X}_t^{i,N})) - \mathbb{E}(h(X_t^{b,i,N}) - h(\widetilde{X}_t^{i,N})). \tag{2.4.25}$$

With this choice of φ and X the supremum on the right hand side of (2.4.23) will be equal to

$$\sup_{x\in X}\left|\frac{1}{N}\sum_{i=1}^{N}\varphi^{i,N}(x)\right|.$$

Using that h is 1-Lipschitz, it follows from Lemma 2.4.2 that $\sup_{x\in X}|\varphi^{i,N}(x)|$ is bounded by a sub-Gaussian random variable $Y^{i,N}$. Thus assumptions (i) and (iii) of Proposition 2.4.1 are satisfied, and as the target space V is simply $\mathbb{R}$, assumption (iv) is also satisfied. This just leaves the verification of assumption (ii). We will compute this for each variable (t,b,h) in turn. For brevity we compute this for $\varphi^{i,N}(t,b,h) = h(X_t^{b,i,N}) - h(\widetilde{X}_t^{i,N})$, the estimate for the centered version (2.4.25) follows easily.

Let $L^{i,N}$ be as in Corollary 2.4.1 applied respectively to $X^{b,i,N}$, so that $L^{i,N}$ are i.i.d. with finite expectation. Then as $h \in \mathrm{Lip1}$ we have

$$|\varphi^{i,N}(t,b,h) - \varphi^{i,N}(t,\tilde{b},h)| \leq L^{i,N}T\left\|b-\tilde{b}\right\|_{\mathrm{L}^\infty([0,T];\mathrm{L}^{-r,\infty}(\mathbb{R}^d;\mathbb{R}^d))}.$$

Next we consider

$$\begin{aligned}|\varphi^{i,N}(t,b,h) - \varphi^{i,N}(t,b,\tilde{h})| &= |(h(X_t^{b,i,N}) - h(\widetilde{X}_t^{i,N})) - (\tilde{h}(X_t^{b,i,N}) - \tilde{h}(\widetilde{X}_t^{i,N}))| \\ &\leq \left(\sup_{s\in[0,T]}\left\langle X_s^{b,i,N}\right\rangle^p + \left\langle \widetilde{X}_s^{i,N}\right\rangle^p\right)\left\|h-\tilde{h}\right\|_{\mathrm{L}^{-p,\infty}(\mathbb{R}^d)}\end{aligned}$$

and we can control the expectation of this supremum by Lemma 2.4.1 for $X^{b,i,N}$ (the estimate for $\tilde{X}^{i,N}$ being easier) and the pth moment of X_0.

Lastly we control the dependence on time. Let $t \in [0,T]$, then we have

$$\sup_{s\in[0,T],|t-s|^{1/3}\leq\varepsilon} |\varphi^{i,N}(t,b,h) - \varphi^{i,N}(s,b,h)| \leq$$
$$\sup_{s\in[0,T],|t-s|^{1/3}\leq\varepsilon} \|h\|_{\mathrm{Lip}} |(X_t^{b,i,N} - \widetilde{X}_t^{i,N}) - (X_s^{b,i,N} - \widetilde{X}_s^{i,N})|$$

and the expectation of the supremum on the right hand side is controlled by Lemma 2.4.2.

This completes the proof of Theorem 2.2.4.

We now prove Proposition 2.4.2. The proof again relies on Proposition 2.4.1 and a carefully chosen map φ and semi-metric space $(\mathcal{X}, d)$. We split the proof into two cases. Firstly we handle $q \in [1,\infty)$. Let $\mathrm{V} = \mathrm{L}^{-r',q}(\mathbb{R}^d)$ and

$$\mathcal{X} = ([0,T] \times |\cdot|^{\alpha\beta/3}) \times (\mathcal{C}, \|\cdot\|^{\beta}_{\mathrm{L}^\infty([0,T];\mathrm{L}^{-r,\infty}(\mathbb{R}^d;\mathbb{R}^d))})$$

with the product metric where $r \in (1,p)$. As in the above proof, we estimate the metric entropy of X,

$$H(\varepsilon, \mathcal{X}, d) \leq C\varepsilon^{-k/\beta}$$

where the dominant term comes from the second space in the definition of X (using Lemma 2.A.3). We define $\varphi^{i,N} : \mathcal{X} \to \mathrm{V}$ by

$$(\varphi^{i,N}(t,b))(x) = h(x, X_t^{b,i,N}) - \mathbb{E}h(x, X_t^{b,i,N}).$$

This is uniformly bounded in $\mathrm{V} = \mathrm{L}^{-r',q}(\mathbb{R}^d)$ by $2\sup_{y\in\mathbb{R}^d} \|h(\cdot,y)\|_{\mathrm{L}^{-r',q}(\mathbb{R}^d)}$ which is finite as h is uniformly bounded and $\langle x\rangle^{-r'q}$ is integrable by assumption. To verify the 'Pointwise Lipschitz' condition we argue in the same way as in the proof of Theorem 2.2.4. We have, again working with the uncentered version for brevity,

$$\left\|\varphi^{i,N}(t,b) - \varphi^{i,N}(s,\tilde{b})\right\|_{\mathrm{V}} \leq |X_t^{i,N,b} - X_s^{i,N,\tilde{b}}|^{\beta} \sup_{\delta,y\in\mathbb{R}^d,\delta\neq 0} \frac{\|h(\cdot,y+\delta) - h(\cdot,y)\|_{\mathrm{V}}}{|\delta|^{\beta}}$$

This supremum is finite by assumption, and the expectation of $|X_t^{i,N,b} - X^{i,N,\tilde{b}}|^{\beta}$

can be controlled as in the proof of Theorem 2.2.4 using also the β appearing in the metric of $\mathcal{X}$.

It remains to check the pointwise law of large numbers. We compute

$$\begin{aligned}
&\sup_{(t,b)\in X} \mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\varphi^{i,N}(t,b)\right\|_{\mathrm{V}} \\
&= \sup_{(t,b)\in X} \mathbb{E}\left(\int_{\mathbb{R}^d}\left|\frac{1}{N}\sum_{i=1}^{N}h(x,X_t^{i,b,N})-\mathbb{E}h(x,X_t^{i,b,N})\right|^q \langle x\rangle^{-qr'}\,dx\right)^{1/q} \\
&\leq C\sup_{(t,b)\in X}\left(\int_{\mathbb{R}^d}\mathbb{E}\left|\frac{1}{N}\sum_{i=1}^{N}h(x,X_t^{i,b,N})-\mathbb{E}h(x,X_t^{i,b,N})\right|^q \langle x\rangle^{-qr'}\,dx\right)^{1/q} \\
&\leq C\sup_{(t,b)\in X}\left(\int_{\mathbb{R}^d}\left(\sup_{x'\in\mathbb{R}^d}\mathbb{E}\left|\frac{1}{N}\sum_{i=1}^{N}h(x',X_t^{i,b,N})-\mathbb{E}h(x',X_t^{i,b,N})\right|^q\right)\langle x\rangle^{-qr'}\,dx\right)^{1/q} \\
&\leq C\sup_{x\in\mathbb{R}^d,(t,b)\in X}\left\|\frac{1}{N}\sum_{i=1}^{N}h(x,X_t^{i,b,N})-\mathbb{E}h(x,X_t^{i,b,N})\right\|_q,
\end{aligned}$$

where we have used that $\langle x\rangle^{-qr'}$ is integrable on $\mathbb{R}^d$ twice, first on the third line to apply Jensen's inequality and then on the fourth line after bringing the expectation out of the integral. As h is a uniformly bounded function, we can apply Lemma 2.4.5 (or just the usual law of large numbers) to obtain that

$$\sup_{(t,b)\in\mathcal{X}}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\varphi^{i,N}(t,b)\right\|_{\mathrm{V}} \leq CN^{-1/2}$$

as required. This completes the proof of Proposition 2.4.2 for $p<\infty$.

Now suppose that $q=\infty$. In this case we set $\mathrm{V}=\mathbb{R}$ and

$$\mathcal{X}=([0,T]\times|\cdot|^{\alpha\beta/3})\times(\mathcal{C},\|\cdot\|^{\beta}_{\mathrm{L}^\infty([0,T];\mathrm{L}^{-r,\infty}(\mathbb{R}^d;\mathbb{R}^d))})\times(\mathbb{R}^d,\rho_{r',\beta})$$

with the product metric, where $r\in(1,p)$. Here $\rho_{r',\beta}$ is a semi-metric on $\mathbb{R}^d$ defined by

$$\rho_{r',\beta}(x,y)=\frac{\min(|x-y|^\beta,1)}{(1+\min(|x|,|y|))^{\min(r',1)}}.$$

It is easy to check that the metric entropy of $(\mathbb{R}^d,\rho_{r',\beta})$ is logarithmic for any $r'>0$ and $\beta\in(0,1]$. Hence, as in the $p<\infty$ case, the metric entropy of $\mathcal{X}$ has

the bound

$$H(\varepsilon, \mathcal{X}, d) \leq C_\gamma \varepsilon^{-k/\beta}.$$

We define $\varphi^{i,N} : \mathcal{X} \to \mathbb{R}$ by

$$\varphi^{i,N}(t, b, x) = h(x, X_t^{b,i,N}) - \mathbb{E}h(x, X_t^{b,i,N}).$$

This is uniformly bounded by 2. The 'Pointwise Lipschitz' conditions for $[0, T]$ and $\mathcal{C}$ are the same as in the proof of Theorem 2.2.4, except that h is only β-Hölder continuous instead of Lipschitz, which is taken into account in the choice of metrics on $[0, T]$ and $\mathcal{C}$. This leaves the estimate for $(\mathbb{R}^d, \rho_{r',\beta})$. Again consider the uncentred case to ease notation. Let $x, y \in \mathbb{R}^d$ with $|x-y| \leq 1$, then we have, where without loss of generality $|x| \geq |y|$,

$$\begin{aligned}
|\varphi^{i,N}(t, b, x) - \varphi^{i,N}(t, b, y)| &= |h(x, X_t^{b,i,N}) - h(y, X_t^{b,i,N})| \\
&\leq |\langle x\rangle^{r'} h(x, X_t^{b,i,N}) - \langle y\rangle^{r'} h(y, X_t^{b,i,N})| \langle x\rangle^{-r'} \\
&\quad + \langle y\rangle^{r'} |h(y, X_t^{b,i,N})| |\langle x\rangle^{-r'} - \langle y\rangle^{-r'}| \\
&\leq C|x-y|^\beta \langle x\rangle^{-r'} + C \langle y\rangle^{r'} \langle y\rangle^{-r'-1} |x-y| \\
&\leq C\rho_{r',\beta}(x, y).
\end{aligned}$$

where we have again considered the uncentered case to ease notation, and we have assumed without loss of generality that $|x| \geq |y|$. This completes the proof of Proposition 2.4.2. □

2.5 Propagation of chaos

In this section we prove the propagation of chaos results Theorems 2.2.1 and 2.2.2. This section is organised as follows. In Section 2.5.1 we prove Theorem 2.2.1 after stating a pair of preliminary lemmas without proof. In Section 2.5.2 we present the proof of Theorem 2.2.2 again after stating without proof a pair of lemmas. Note that the proof of Theorem 2.2.2 is very similar to that of Theorem 2.2.1 so we give only the differences. Finally in Sections 2.5.3 and 2.5.4 we provide the postponed proofs of the lemmas.

2.5.1 The first order case

As a first step, we must obtain a prior estimates on the time regularity of the vector field b^N and show that the contribution of b^N not being sufficiently regular to (2.2.1) is of lower order. As the proofs are technical we present them at the end of this section.

In the first order case we expect that $b^N \in \Lambda^{0,\alpha'}_{para}(\mathrm{L}^q([0,T]\times\mathbb{R}^d;\mathbb{R}^d))$ for $\alpha' < \alpha$ as $X^{i,N}$ will be merely (almost 1/2)-Hölder continuous in time due to the driving noise.

Lemma 2.5.1. *[Time regularity (first order case)] Let the interaction kernel $K(x,y)$ lie in $\Lambda^{0,\alpha}(\mathrm{L}^\infty_y(\mathbb{R}^d;\mathrm{L}^q_x(\mathbb{R}^d)))$ with $\alpha \in (0,1]$ and $f_0 \in \mathrm{P}_1(\mathbb{R}^d)$. Define the event E_A by*

$$E_A = \{\left\|b^N\right\|_{\Lambda^{0,\alpha'}_{para}(\mathrm{L}^q([0,T]\times\mathbb{R}^d))} > A\}$$

for any $\alpha' \in (0,\alpha)$. Then there exists $A > 0$ such that we have the bound

$$\left\|\left\|1_{E_A}\left(\sup_{t\in[0,T]} d_{\mathrm{MKW}}(\mu^N_t, f_t) - d_{\mathrm{MKW}}(\mu^N_0, f_0)\right)\right\|\right\| \leq CN^{-1/2}.$$

where C and A depend only on α' and the norm of K.

Next we bound the dependence of the laws f^b_t using simple energy estimates. As we will work throughout the proof with mollified kernels, we prove the results for smooth vector fields (with constants independent of the degree of smoothness) which avoids any issues with existence or uniqueness. Again, we delay the proof until the end of this section. In the first order case we have:

Lemma 2.5.2 (Weighted energy estimate (first order case))**.** *Let $b, \tilde{b} \in \mathrm{L}^\infty([0,T]\times\mathbb{R}^d;\mathbb{R}^d)$ be continuous in t and C^1_b in x, and $f_0 \in \mathrm{L}^{p+r,q}(\mathbb{R}^d)$ for some $r,p > 0$ and $q \in [2,\infty)$, then*

$$\left\|f^b_t - f^{\tilde{b}}_t\right\|_{\mathrm{L}^{p,2}(\mathbb{R}^d)} \leq C\int_0^t \left\|b_s - \tilde{b}_s\right\|_{\mathrm{L}^{-r,q'}(\mathbb{R}^d;\mathbb{R}^d)} dt, \qquad \frac{1}{q'} + \frac{1}{q} = \frac{1}{2},$$

where C depends only on f_0, $\|b\|_{\mathrm{L}^\infty([0,T]\times\mathbb{R}^d)}$ and $\left\|\tilde{b}\right\|_{\mathrm{L}^\infty([0,T]\times\mathbb{R}^d)}$.

With these lemmas we are ready to prove the main propagation of chaos re-

sult.

Proof of Theorem 2.2.1. We divide the proof into 5 steps.

Step 1. Mollification of the interaction kernel. Note first that Sobolev embedding implies that $K(x,y)$ is in $\mathrm{C}^{0,\alpha}(\mathbb{R}^d;\mathbb{R}^d)$ for some $\alpha > 0$ in all cases of the theorem. Now let K_n be a sequence of smooth interaction kernels obeying the same bounds as K. Then using Corollary 2.4.1 on the entire $N \cdot d$ dimensional system we deduce that the solutions $X_t^{n,i,N}$ to the SDE system (2.1.1) with K replaced by K_n converge almost surely to the solution $X_t^{i,N}$ of the original system (2.1.1). Using this, it is sufficient to prove propagation of chaos for a smooth kernel K with constants depending only on the bounds assumed in the theorem. Thus from here on in the proof K shall be assumed to be in C_b^1. As a consequence all the considered vector fields b will also lie in $\mathrm{C}([0,T];\mathrm{C}_b^1(\mathbb{R}^d;\mathbb{R}^d))$, and we can freely apply Theorem 2.2.4 to such fields.

Step 2. Choice of functional space and exponents. We have assumed that $K(x,y) \in \Lambda^{0,s}(\mathrm{L}_y^\infty(\mathbb{R}^d;\mathrm{L}_x^q(\mathbb{R}^d)))$ (note that, as explained in Remark 2.2.4, case (1) of Theorem 2.2.1 is included in case (2) as $q = \infty$). By the assumption on f_0 we may choose r, r' such that

$$f_0 \in \mathrm{L}^{r+r',q'}(\mathbb{R}^d), \quad r > d/q, \quad r' > (d/2) + 1.$$

Note that with these choices we have the continuous inclusions:

$$\mathrm{L}^{r',2}(\mathbb{R}^d) \hookrightarrow \mathrm{L}^{1,1}(\mathbb{R}^d) \hookrightarrow (\mathrm{P}_1(\mathbb{R}^d), d_{\mathrm{MKW}}), \tag{2.5.1}$$

which will be how we control the Wasserstein distance between functions (as opposed to measures).

Step 3. Regularity of the interaction field. Let $s' < s$, then by Lemma 2.5.1 we may assume $b^N \in \mathcal{C}$ where

$$\mathcal{C} = \mathrm{C}([0,T];\mathrm{C}_b^1(\mathbb{R}^d;\mathbb{R}^d)) \cap \{b : \|b\|_{\Lambda_{para}^{0,s'}(\mathrm{L}^q([0,T]\times\mathbb{R}^d;\mathbb{R}^d))} \leq A\}, \tag{2.5.2}$$

for some $A < \infty$. Note that by Proposition 2.A.1(4) we have the metric entropy

bound

$$H(\varepsilon, \mathcal{C}, \mathrm{L}^\infty([0,T]; \mathrm{L}^{-p',\infty}(\mathbb{R}^d; \mathbb{R}^d))) \leq C\varepsilon^{-(d+2)/s'} \tag{2.5.3}$$

for any $p' \in (d/q, p)$. [Recall that $p > d/q$ by assumption.] Here we have used that $q > (d+2)/s$ so that s' can be chosen large enough for $q > (d+2)/s'$ to hold. Note further that, due to Sobolev embedding, all elements $b \in \mathcal{C}$ have a uniform bound in $\mathrm{C}^{0,\alpha}_{para}$, i.e. $\|b\|_{\mathrm{C}^{0,\alpha}_{para}([0,T]\times\mathbb{R}^d;\mathbb{R}^d))} \leq CA$ for $\alpha = s - (d+2)/q > 0$ and an absolute constant C.

Step 4. Consistency: The uniform law of large numbers on the particles. Note that the particle system $(X_t^{i,N})_{i=1}^N$ is equal to $(X_t^{b^N,i,N})_{i=1}^N$, and the limit process f_t is equal to $f_t^{b^\infty}$.

By the triangle inequality,

$$\begin{aligned}
&\sup_{t\in[0,T]} d_{\mathrm{MKW}}(\mu_t^N, f_t) - d_{\mathrm{MKW}}(\mu_0^N, f_0) \\
&\quad \leq \left(\sup_{t\in[0,T]} d_{\mathrm{MKW}}(\mu_t^{b^N,N}, f_t^{b^N}) - d_{\mathrm{MKW}}(\mu_0^N, f_0) \right) + \sup_{t\in[0,T]} d_{\mathrm{MKW}}(f_t^{b^N}, f_t^{b^\infty}).
\end{aligned}$$

Using Theorem 2.2.4 with $\mathcal{C}$ given by (2.5.2) and using (2.5.3), the first term can be bounded as

$$\begin{aligned}
&\left\| \sup_{t\in[0,T]} d_{\mathrm{MKW}}(\mu_t^{b^N,N}, f_t^{b^N}) - d_{\mathrm{MKW}}(\mu_0^N, f_0) \right\| \\
&\quad \leq \left\| \sup_{b\in\mathcal{C}} \sup_{t\in[0,T]} d_{\mathrm{MKW}}(\mu_t^{b,N}, f_t^b) - d_{\mathrm{MKW}}(\mu_0^N, f_0) \right\| \leq CN^{-\gamma_1},
\end{aligned}$$

where

$$\gamma_1 = \frac{1}{2 + \max(\frac{d+2}{s'}, \frac{d}{p-1})}.$$

Step 5. Stability: Estimates on the limit equation. This leaves the other distance $d_{\mathrm{MKW}}(f_t^{b^N}, f_t^{b^\infty})$, for which we will use estimates on the limit equation.

Step 5.1. Dependence of f upon the field. By applying the energy estimate

(Lemma 2.5.2) we obtain

$$\begin{aligned}\left\|f_t^{b^N}-f_t^{b^\infty}\right\|_{\mathrm{L}^{1,1}(\mathbb{R}^d)} &\le C\left\|f_t^{b^N}-f_t^{b^\infty}\right\|_{\mathrm{L}^{r',2}(\mathbb{R}^d)}\\ &\le C\int_0^t\left\|b_s^N-b_s^\infty\right\|_{\mathrm{L}^{-r,q}(\mathbb{R}^d;\mathbb{R}^d)}\,ds\end{aligned}\tag{2.5.4}$$

where the first continuous inclusion in (2.5.1) is used for the first line and $f_0\in \mathrm{L}^{r+r',q'}(\mathbb{R}^d)$ is needed to apply the energy estimate for the second.

Step 5.2. Dependence of the field upon f. For $b\in\mathcal{C}$, define $b_t^{b,N}$ by

$$b_t^{b,N}(x)=\frac{1}{N}\sum_{i=1}^N K(x,X_t^{b,i,N}),$$

so that $b_t^N=b_t^{b^N,N}$. Next define $b_t^{b,\infty}$ by

$$b_t^{b,\infty}(x)=\int K(x,y)f_t^b(y)\,dy,$$

so that $b_t^\infty=b_t^{b^\infty,\infty}$. Then we have

$$\begin{aligned}\left\|b_t^N-b_t^\infty\right\|_{\mathrm{L}^{-r,q}(\mathbb{R}^d;\mathbb{R}^d)} &\le \left\|b_t^N-b_t^{b^N,\infty}\right\|_{\mathrm{L}^{-r,q}(\mathbb{R}^d;\mathbb{R}^d)}+\left\|b_t^{b^N,\infty}-b_t^{b^\infty,\infty}\right\|_{\mathrm{L}^{-r,q}(\mathbb{R}^d;\mathbb{R}^d)}\\ &\le \sup_{b\in\mathcal{C}}\left\|b_t^{b,N}-b_t^{b,\infty}\right\|_{\mathrm{L}^{-r,q}(\mathbb{R}^d;\mathbb{R}^d)}+\left\|b_t^{b^N,\infty}-b_t^{b^\infty,\infty}\right\|_{\mathrm{L}^{-r,q}(\mathbb{R}^d;\mathbb{R}^d)}\end{aligned}\tag{2.5.5}$$

By applying Proposition 2.4.2 to K we can control the first of these by

$$\left\|\sup_{b\in\mathcal{C}}\sup_{t\in[0,T]}\left\|b_t^{b,N}-b_t^{b,\infty}\right\|_{\mathrm{L}^{-r,q}(\mathbb{R}^d;\mathbb{R}^d)}\right\|\le CN^{-\gamma_2},\qquad \gamma_2=\frac{1}{2+\frac{d+2}{s'^2}}.\tag{2.5.6}$$

Here we have used that $b_t^{b,\infty}=\mathbb{E}b_t^{b,N}$ for b deterministic, that $r>d/q$, that K is bounded and that $K\in\Lambda^{0,s}(\mathrm{L}^\infty_y(\mathbb{R}^d;\mathrm{L}^q_x(\mathbb{R}^d;\mathbb{R}^d)))$ implies

$$\sup_{0\neq(\delta_1,\delta_2)\in\mathbb{R}^d\times\mathbb{R}^d}\left\|\frac{K(x+\delta_1,y+\delta_2)-K(x,y)}{|\delta_1|^2+|\delta_2|^2}\right\|_{\mathrm{L}^\infty_y(\mathbb{R}^d;\mathrm{L}^q_x(\mathbb{R}^d;\mathbb{R}^d))}<\infty$$

and by taking $\delta_1=0$ and using that L^q embeds continuously into $\mathrm{L}^{-r,q}$, we recover the assumption of Proposition 2.4.2.

The second of the terms on the right of Eq. (2.5.5) can be controlled by

$$\begin{aligned}\left\|b_t^{b^N,\infty} - b_t^{b^\infty,\infty}\right\|_{\mathrm{L}^{-r,q}(\mathbb{R}^d;\mathbb{R}^d)} &\leq C \sup_{x\in\mathbb{R}^d} \int |K(x,y)||f_t^{b^N}(y) - f_t^{b^\infty}(y)|\, dy \\ &\leq C \left\|f_t^{b^N} - f_t^{b^\infty}\right\|_{\mathrm{L}^1(\mathbb{R}^d)} \leq C \left\|f_t^{b^N} - f_t^{b^\infty}\right\|_{\mathrm{L}^{1,1}(\mathbb{R}^d)}\end{aligned} \tag{2.5.7}$$

where for $q = \infty$ the first inequality is clear, and for $q < \infty$ we have used that $\langle x\rangle^{-rq}$ is integrable on $\mathbb{R}^d$ to obtain it.

Step 5.3. Grönwall estimate. Combining (2.5.7) with the previous estimates (2.5.4),(2.5.5),(2.5.6) yields

$$\left\|f_t^{b^N} - f_t^{b^\infty}\right\|_{\mathrm{L}^{1,1}(\mathbb{R}^d)} \leq Y + C\int_0^t \left\|f_s^{b^N} - f_s^{b^\infty}\right\|_{\mathrm{L}^{1,1}(\mathbb{R}^d)} ds$$

where Y is a non-negative sub-Gaussian random variable with norm bound $|||Y||| \leq CN^{-\gamma_2}$. Therefore, applying the Grönwall inequality we have

$$\sup_{t\in[0,T]} \left\|f_t^{b^N} - f_t^{b^\infty}\right\|_{\mathrm{L}^{1,1}(\mathbb{R}^d)} \leq CY,$$

and as the $\mathrm{L}^{1,1}$ distance controls the Wasserstein distance (this is the second continuous inclusion in (2.5.1)), we have proved the theorem. □

2.5.2 The second order case

We now move onto the second order case. We begin, as before, with estimates on the time regularity of the interaction field.

In the second order case we expect higher regularity for b^N as K is evaluated at the spatial positions $X^{i,N}$ which are time differentiable. However, we need additional moments to control the velocities.

Lemma 2.5.3. *[Time regularity (second order case)] Let c be the constant in the claim of Lemma 2.4.4. Then the following hold:*

1. *Let $K(x,y) \in \Lambda^{0,\alpha}(\mathrm{L}^\infty_y(\mathbb{R}^d;\mathrm{L}^q_x(\mathbb{R}^d)))$ for some $\alpha \in (0,1]$ and let $f_0 \in \mathrm{P}_2(\mathbb{R}^d\times$*

$\mathbb{R}^d$). *Let* E_A *be the event*

$$E_A = \{\left\|b^N\right\|_{\Lambda^{0,\alpha}(\mathrm{L}^q([0,T]\times\mathbb{R}^d))} > A\}.$$

Then there exists $A > 0$ *such that we have the bound*

$$\left\|1_{E_A}\left[\sup_{t\in[0,T]} d_{\mathrm{MKW}}(\mu_t^N, f_t) - c d_{\mathrm{MKW}}(\mu_0^N, f_0)\right]_+\right\| \leq CN^{-1/2},$$

where C *and* A *depend only on the norm of* K.

2. *Let* $K(x,y) = W(x-y)$ *for* $W \in \Lambda^{1,\alpha}(\mathrm{L}^q(\mathbb{R}^d;\mathbb{R}^d))$ *for some* $\alpha \in (0,1/2)$ *and let* $f_0 \in \mathrm{P}_4(\mathbb{R}^{2d})$. *Let* E_A *be the event*

$$E_A = \{\left\|b^N\right\|_{\Lambda^{1,\alpha}(\mathrm{L}^q([0,T]\times\mathbb{R}^d))} > A\}.$$

Then there exists $A > 0$ *such that we have the bound*

$$\left\|1_{E_A}\left[\sup_{t\in[0,T]} d_{\mathrm{MKW}}(\mu_t^N, f_t) - c d_{\mathrm{MKW}}(\mu_0^N, f_0)\right]_+\right\| \leq CN^{-1/2},$$

where C *and* A *depend only on the norm of* K.

The second order energy estimate is:

Lemma 2.5.4 (Weighted energy estimate (second order case))**.** *Let the vector fields* $b, \tilde{b}$ *lie in* $\mathrm{L}^\infty([0,T]\times\mathbb{R}^d;\mathbb{R}^d)$ *and be continuous in* t *and* C_b^1 *in* x, *and* $f_0 \in \mathrm{L}^{p+r,q}(\mathbb{R}^d\times\mathbb{R}^d)$ *for some* $r,p>0$ *and* $q\in[2,\infty)$, *then*

$$\left\|f_t^b - f_t^{\tilde{b}}\right\|_{\mathrm{L}^{p,2}(\mathbb{R}^{2d})} \leq C\int_0^t \left\|b_s - \tilde{b}_s\right\|_{\mathrm{L}^{-r,q'}(\mathbb{R}^d;\mathbb{R}^d)} dt, \qquad \frac{1}{q'} + \frac{1}{q} = \frac{1}{2},$$

where C *depends only on* f_0, $\|b\|_{\mathrm{L}^\infty([0,T]\times\mathbb{R}^d;\mathbb{R}^d)}$ *and* $\left\|\tilde{b}\right\|_{\mathrm{L}^\infty([0,T]\times\mathbb{R}^d;\mathbb{R}^d)}$.

Proof of Theorem 2.2.2. We model the proof on that of Theorem 2.2.1, and thus split it into 5 steps. Much of the proof is analogous to that of Theorem 2.2.1. Therefore we only explain the differences.

Step 1. Mollification of the interaction kernel. This is identical to the corresponding step in the proof of Theorem 2.2.1. We thus omit it.

Step 2. Choice of functional space and exponents. By the assumptions on f_0 we may choose r, r' such that the following holds:

$$f_0 \in \mathrm{L}^{r+r',q'}(\mathbb{R}^d \times \mathbb{R}^d), \quad r > d/q, \quad r' > d+1.$$

As in the proof of Theorem 2.2.1 we have the continuous inclusions:

$$\mathrm{L}^{r',2}(\mathbb{R}^d \times \mathbb{R}^d) \hookrightarrow \mathrm{L}^{1,1}(\mathbb{R}^d \times \mathbb{R}^d) \hookrightarrow (\mathrm{P}_1(\mathbb{R}^d \times \mathbb{R}^d), d_{\mathrm{MKW}}).$$

Step 3. Regularity of the interaction field. The choice of $\mathcal{C}$ depends upon which of assumptions (1) and (2) is made. More precisely which of the relaxed assumptions in Remark 2.2.9 is made. In each case:

1. By Lemma 2.5.3(1) we may assume that $b^N \in \mathcal{C}$ where

   $$\mathcal{C} = \mathrm{C}([0,T]; \mathrm{C}^1_b(\mathbb{R}^d; \mathbb{R}^d)) \cap \{b : \|b\|_{\Lambda^{0,s}(\mathrm{L}^q([0,T]\times\mathbb{R}^d;\mathbb{R}^d))} \leq A\},$$

   for some $A < \infty$. Note that the metric entropy of $\mathcal{C}$ is bounded using Proposition 2.A.1 as

   $$H(\varepsilon, \mathcal{C}, \|\cdot\|_{\mathrm{L}^\infty([0,T];\mathrm{L}^{-p',\infty}(\mathbb{R}^d;\mathbb{R}^d))}) \leq C\varepsilon^{-(d+1)/s} \tag{2.5.8}$$

   where we have used that $q > (d+1)/s$ and used $p > d/q$ to take $p' \in (d/q, p)$.

2. By Lemma 2.5.3(2) we may assume that $b^N \in \mathcal{C}$ where

   $$\mathcal{C} = \mathrm{C}([0,T]; \mathrm{C}^1_b(\mathbb{R}^d; \mathbb{R}^d)) \cap \{b : \|b\|_{\Lambda^{1,(s-1)}(\mathrm{L}^q([0,T]\times\mathbb{R}^d;\mathbb{R}^d))} \leq A\},$$

   where we have used that $p \geq 4$. The corresponding metric entropy estimate provided by Proposition 2.A.1 is given again by (2.5.8).(5) for the same choice of p' and use of assumptions, but note here $s > 1$.

Note that Sobolev embedding implies the bound $\|b\|_{\mathrm{C}^{0,\beta}([0,T]\times\mathbb{R}^d;\mathbb{R}^d)} \leq CA$ for $\beta = \min(1, s-(d+1)/q)$ for any $b \in \mathcal{C}$. Moreover, Sobolev embedding also implies the bound $\|b\|_{\mathrm{C}([0,T];\mathrm{C}^{0,\alpha}(\mathbb{R}^d;\mathbb{R}^d))} \leq CA$ for $b \in \mathcal{C}$ and $\alpha = s - d/q > 2/3$ by assumption. Therefore, we have sufficient regularity to apply Theorem 2.2.6 in the next step.

Step 4. Consistency: The uniform law of large numbers on the particles. This is identical to the corresponding step in the proof of Theorem 2.2.1, except we instead apply Theorem 2.2.6 and here

$$\gamma_1 = \frac{1}{2 + \max(\frac{d+1}{s}, d)}$$

(noting that $p > 2$ by assumption).

Step 5. Stability: Estimates on the limit equation.

Step 5.1. Dependence of f upon the field. This is analogous to step 5 in the proof of Theorem 2.2.1 using the energy estimate Lemma 2.5.4 and we leave it to the reader.

Step 5.2. Dependence upon the field upon f. The only differences between this step and the corresponding step in the proof of Theorem 2.2.1 are that here $b_t^{b,\infty}$ is defined by

$$b_t^{b,\infty}(x) = \int K(x,y) \left(\int f^b(y,v)\, dv \right) dy$$

and (2.5.7) is replaced by

$$\begin{aligned} &\left\| b_t^{b^N,\infty} - b_t^{b^\infty,\infty} \right\|_{\mathrm{L}^{-r,q}(\mathbb{R}^d;\mathbb{R}^d)} \\ &\quad \le C \sup_{x \in \mathbb{R}^d} \int |K(x,y)| \left| \int f_t^{b^N}(y,v)\, dv - \int f_t^{b^\infty}(y,v)\, dv \right| dy \\ &\quad \le C \left\| f_t^{b^N} - f_t^{b^\infty} \right\|_{\mathrm{L}^1(\mathbb{R}^d \times \mathbb{R}^d)} \le C \left\| f_t^{b^N} - f_t^{b^\infty} \right\|_{\mathrm{L}^{1,1}(\mathbb{R}^d \times \mathbb{R}^d)}. \end{aligned}$$

Lastly, γ_2 is here instead given by

$$\gamma_2 = \begin{cases} \frac{1}{2+\frac{d+1}{s^2}}, & \text{if } s \le 1 \\ \frac{1}{2+\frac{d+1}{s}}, & \text{otherwise.} \end{cases}$$

Step 5.3. Grönwall estimate. This is identical to the proof of Theorem 2.2.1 and we omit it. □

2.5.3 Proof of the time regularity lemmas

For the proof of Lemmas 2.5.1 and 2.5.3 we require the following simple estimate.

Lemma 2.5.5. *Let E be an event and K be bounded. Then*

$$\left\| 1_E \left[\sup_{t\in[0,T]} d_{\mathrm{MKW}}(\mu_t^N, f_t) - c d_{\mathrm{MKW}}(\mu_0^N, f_0) \right]_+ \right\| \leq C\mathbb{P}(E) + CN^{-1/2}, \tag{2.5.9}$$

where c is chosen as in Lemma 2.4.4.

Proof. We present only the first order case for brevity, the second order case being analogous. From an identical computation to that used in the proof of Lemma 2.4.3 and then using Lemma 2.4.4 we deduce that

$$\begin{aligned}
&\left[\sup_{t\in[0,T]} d_{\mathrm{MKW}}(\mu_t^N, f_t) - c d_{\mathrm{MKW}}(\mu_0^N, f_0) \right]_+ \\
&\quad \leq \sup_{t\in[0,T]} \sup_{h\in \mathrm{Lip}1} \left(\frac{1}{N}\sum_{i=1}^N (h(X_t^{i,N}) - h(\widetilde{X}_t^{i,N})) - \mathbb{E}(h(X_t^{i,N}) - h(\widetilde{X}_t^{i,N})) \right) \\
&\quad \leq \frac{1}{N}\sum_{i=1}^N \underbrace{C\left(1 + \sup_{t\in[0,T]} |B_t^{i,N}|\right)}_{=:A^{i,N}}
\end{aligned}$$

where we have used Lemma 2.4.2 to obtain the final line. Hence the left hand side of (2.5.9) is bounded by

$$\begin{aligned}
\left\| 1_E \frac{1}{N}\sum_{i=1}^N A^{i,N} \right\| &\leq \left\| 1_E \mathbb{E}A^{i,N} \right\| + \left\| 1_E \frac{1}{N}\sum_{i=1}^N (A^{i,N} - \mathbb{E}A^{i,N}) \right\| \\
&\leq (\mathbb{E}A^{i,N})\mathbb{P}(E) + \left\| \frac{1}{N}\sum_{i=1}^N (A^{i,N} - \mathbb{E}A^{i,N}) \right\| \\
&\leq C\mathbb{P}(E) + C\left\| A^{1,N} \right\| N^{-1/2}
\end{aligned}$$

where we have used the law of large numbers for sub-Gaussian random variables (Lemma 2.4.5) on the last line. As $A^{1,N}$ is a sub-Gaussian random variable, the proof is complete. □

We now continue on to the proofs of the time regularity lemmas.

Proof of Lemma 2.5.1. By using Lemma 2.5.5 it suffices to find A such that

$$\mathbb{P}(\left\|b^N\right\|_{\Lambda^{0,\alpha'}_{para}(\mathrm{L}^q([0,T]\times\mathbb{R}^d;\mathbb{R}^d))} > A) \leq CN^{-1/2}.$$

By the definition of b^N we have the estimate

$$\begin{aligned}\left\|b^N\right\|_{\Lambda^{0,\alpha'}_{para}(\mathrm{L}^q([0,T]\times\mathbb{R}^d;\mathbb{R}^d))} &= \left\|\frac{1}{N}\sum_{i=1}^N K(x, X_t^{i,N})\right\|_{\Lambda^{0,\alpha'}_{para}(\mathrm{L}^q([0,T]\times\mathbb{R}^d;\mathbb{R}^d))} \\ &\leq \frac{1}{N}\sum_{i=1}^N \left\|K(x, X_t^{i,N})\right\|_{\Lambda^{0,\alpha'}_{para}(\mathrm{L}^q([0,T]\times\mathbb{R}^d;\mathbb{R}^d))},\end{aligned}$$

and this is bounded by

$$\begin{aligned}&\|K\|_{\Lambda^{0,\alpha}(\mathrm{L}^\infty_y(\mathbb{R}^d;\mathrm{L}^q_x(\mathbb{R}^d;\mathbb{R}^d)))}\left(1+\frac{1}{N}\sum_{i=1}^N \left\|X_t^{i,N}\right\|_{\mathrm{C}^{0,\alpha'/(2\alpha)}([0,T];\mathbb{R}^d)}\right) \\ &\leq \frac{1}{N}\sum_{i=1}^N \underbrace{C\left(1+\left\|B_t^{i,N}\right\|_{\mathrm{C}^{0,\alpha'/(2\alpha)}([0,T];\mathbb{R}^d)}\right)}_{=:A^{i,N}}\end{aligned}$$

where $(A^{i,N})_{i=1}^N$ are i.i.d. random variables with finite second moments (sub-Gaussian even, see [195]). Set $A = 2\mathbb{E}A^{1,N}$, then from Chebyshev's inequality we have

$$\begin{aligned}\mathbb{P}(\left\|b^N\right\|_{\Lambda^{0,\alpha'}_{para}(\mathrm{L}^q([0,T]\times\mathbb{R}^d;\mathbb{R}^d))} > A) &\leq \mathbb{P}\left(\frac{1}{N}\sum_{i=1}^N (A^{i,N} - \mathbb{E}A^{i,N}) > 2\mathbb{E}A^{1,N}\right) \\ &\leq \frac{\mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^N A^{i,N}\right)}{|\mathbb{E}A^{1,N}|^2} \leq CN^{-1},\end{aligned}$$

which completes the proof of the lemma. □

To prove the second claim of Lemma 2.5.3 we shall need a simple lemma.

Lemma 2.5.6. *Let $W \in \mathrm{W}^{1,1}_{loc}$ and $g \in \mathrm{C}^1([0,T];\mathbb{R}^d)$. Then $W(x-g(t))$ has weak time derivative given by*

$$\partial_t[W(x-g(t))] = -g'(t)\cdot(\nabla W)(x-g(t)).$$

Proof. Let $\varphi \in \mathcal{D}([0,T]\times\mathbb{R}^d)$ be a test function, and let the pairing of a distri-

bution in $\mathcal{D}'([0,T]\times\mathbb{R}^d)$ and a test function in $\mathcal{D}([0,T]\times\mathbb{R}^d)$ be $\langle\cdot,\cdot\rangle$. Then we have

$$\begin{aligned}
\langle \partial_t[W(x-g(t))], \varphi(t,x)\rangle &= -\langle W(x-g(t)), (\partial_t\varphi)(t,x)\rangle \\
&= -\langle W(x), (\partial_t\varphi)(t,x+g(t))\rangle \\
&= -\langle W(x), \partial_t[\varphi(t,x+g(t))] - g'(t)\cdot(\nabla\varphi)(t,x+g(t))\rangle \\
&= 0 + \langle W(x), g'(t)\cdot(\nabla[\varphi(t,x+g(t))]\rangle \\
&= -\langle g'(t)\cdot\nabla W(x), \varphi(t,x+g(t))\rangle \\
&= -\langle g'(t)\cdot(\nabla W)(x-g(t)), \varphi(t,x)\rangle
\end{aligned}$$

which is the claim of the lemma. □

Proof of Lemma 2.5.3. We shall prove each claim in turn. For both claims, by Lemma 2.5.5 it is sufficient to bound the probability of the bad event.

1. As in the proof of Lemma 2.5.1 we compute

$$\begin{aligned}
&\left\|b^N\right\|_{\Lambda^{0,\alpha}(\mathrm{L}^q([0,T]\times\mathbb{R}^d;\mathbb{R}^d))} \\
&\quad= \left\|\frac{1}{N}\sum_{i=1}^N K(\cdot, X_t^{i,N})\right\|_{\Lambda^{0,\alpha}(\mathrm{L}^q([0,T]\times\mathbb{R}^d;\mathbb{R}^d))} \\
&\quad\le \frac{1}{N}\sum_{i=1}^N \left\|K(\cdot, X_t^{i,N})\right\|_{\Lambda^{0,\alpha}(\mathrm{L}^q([0,T]\times\mathbb{R}^d;\mathbb{R}^d))} \\
&\quad\le \|K\|_{\Lambda^{0,\alpha}(\mathrm{L}^\infty_y(\mathbb{R}^d;\mathrm{L}^q_x(\mathbb{R}^d;\mathbb{R}^d)))}\left(1+\frac{1}{N}\sum_{i=1}^N \left\|X_t^{i,N}\right\|_{\mathrm{C}^{0,1}([0,T];\mathbb{R}^d;\mathbb{R}^d)}\right) \\
&\quad\le \frac{1}{N}\sum_{i=1}^N \underbrace{C(1+\sup_{t\in[0,T]}|V_t^{i,N}|)}_{=:A^{i,N}}.
\end{aligned}$$

Define $A = 2\mathbb{E}A^{1,N}$ and $E = \{\frac{1}{N}\sum_{i=1}^N A^{i,N} \ge A\}$. Then,

$$\mathbb{P}(E) \le \frac{\mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^N A^{i,N}\right)}{|\mathbb{E}A^{1,N}|^2} \le CN^{-1}$$

by Chebyshev's inequality using that $A^{i,N}$ are i.i.d. with finite second moment by Lemma 2.4.1.

2. As $X^{i,N}$ is continuously time differentiable, we can apply Lemma 2.5.6 to obtain

$$\begin{aligned}\partial_t b_t^N(x) &= \frac{1}{N}\sum_{i=1}^N \partial_t[K(x, X_t^{i,N})] = \frac{1}{N}\sum_{i=1}^N \partial_t[W(x - X_t^{i,N})] \\ &= \frac{1}{N}\sum_{i=1}^N V_t^{i,N} \cdot (\nabla W)(x - X_t^{i,N}).\end{aligned}$$

Furthermore, the x derivatives satisfy

$$\nabla b_t^N(x) = \frac{1}{N}\sum_{i=1}^N (\nabla W)(x - X_t^{i,N}),$$

which is always easier to bound than $\partial_t b^N$, so we omit these bounds.

Taking the L^q norm we have

$$\left\|\partial_t b_t^N\right\|_{\mathrm{L}^q([0,T]\times\mathbb{R}^d;\mathbb{R}^d)} \le C\left\|\nabla W\right\|_{\mathrm{L}^q(\mathbb{R}^d;\mathbb{R}^{d\times d})} \frac{1}{N}\sum_{i=1}^N \sup_{s\in[0,T]} |V_s^{i,N}|,$$

and similarly,

$$\begin{aligned}\sup_{x,y\in\mathbb{R}^d, x\neq y} \frac{1}{|x-y|^\alpha} &\left\|\partial_t b^N(x+\cdot) - \partial_t b^N(y+\cdot)\right\|_{\mathrm{L}^q([0,T]\times\mathbb{R}^d;\mathbb{R}^d)} \\ &\le C\left\|\nabla W\right\|_{\Lambda^{0,\alpha}(\mathrm{L}^q(\mathbb{R}^d;\mathbb{R}^d))} \frac{1}{N}\sum_{i=1}^N \sup_{s\in[0,T]} |V_s^{i,N}|,\end{aligned}$$

and we can estimate the $V^{i,N}$ terms in the same way as part (1). In the same way

$$\begin{aligned}\sup_{s,t\in[0,T], s\neq t} \frac{1}{|t-s|^\alpha} &\left\|\partial_t b_t^N - \partial_t b_s^N\right\|_{\mathrm{L}^q([0,T]\times\mathbb{R}^d;\mathbb{R}^d)} \\ &\le C\left\|\nabla W\right\|_{\Lambda^{0,\alpha}(\mathrm{L}^q(\mathbb{R}^d;\mathbb{R}^d))} \left(\frac{1}{N}\sum_{i=1}^N \sup_{s\in[0,T]} |V_s^{i,N}|\right)^2 \\ &\quad + C\left\|\nabla W\right\|_{\mathrm{L}^q(\mathbb{R}^d;\mathbb{R}^{d\times d})} \frac{1}{N}\sum_{i=1}^N \left\|V^{i,N}\right\|_{\mathrm{C}^{0,\alpha}([0,T];\mathbb{R}^d)}.\end{aligned}$$

All of these terms may be controlled using the methods in part (1) and the proof of Lemma 2.5.1 as $\alpha < 1/2$. We omit the details. □

2.5.4 Proof of the energy estimates

We now provide the proofs of the two energy estimates.

Proof of Lemma 2.5.2. For brevity, let $f_t = f_t^b$ and $\tilde{f}_t = f_t^{\tilde{b}}$. [We abuse notation in this proof and use $\tilde{f}$ to refer to the definition in the previous sentence rather than the law of the reference process.] Let $g = f - \tilde{f}$, then g_t solves

$$\begin{cases} \partial_t g_t + \nabla \cdot (b_t g_t) - \frac{1}{2}\Delta g_t = -\nabla \cdot (\tilde{f}_t(b_t - \tilde{b}_t)), \quad (t,x) \in [0,T] \times \mathbb{R}^d, \\ g_0 = 0. \end{cases}$$

We multiply this equation by $g_t \langle x \rangle^{2p}$ and integrate by parts. This yields

$$\begin{aligned} \frac{d}{dt} \|g_t \langle x \rangle^p\|^2_{\mathrm{L}^2(\mathbb{R}^d)} - \int g_t b_t \cdot \nabla(g_t \langle x \rangle^{2p})\, dx + \int \nabla g_t \cdot \nabla(g_t \langle x \rangle^{2p})\, dx \\ = \int \tilde{f}_t(b_t - \tilde{b}_t) \cdot \nabla(g_t \langle x \rangle^{2p})\, dx. \end{aligned}$$

We bound the right hand side using Hölder's inequality by

$$|RHS| \leq \left\|b_t - \tilde{b}_t\right\|_{\mathrm{L}^{-r,q'}(\mathbb{R}^d;\mathbb{R}^d)} \left\|\tilde{f}_t\right\|_{\mathrm{L}^{p+r,q}(\mathbb{R}^d)} \left(\|g_t\|_{\mathrm{L}^{p,2}(\mathbb{R}^d)} + \|\nabla g_t\|_{\mathrm{L}^{p,2}(\mathbb{R}^d;\mathbb{R}^d)}\right)$$

and similarly the second term on the left hand side using instead that the norm $\|b\|_{\mathrm{L}^\infty([0,T]\times\mathbb{R}^d;\mathbb{R}^d)}$ is bounded by a constant. Using Young's inequality, and that ∇ hitting $\langle x \rangle^{2p}$ produces terms of lower order, we obtain

$$\frac{d}{dt} \|g_t\|^2_{\mathrm{L}^{p,2}(\mathbb{R}^d)} \leq C \|g_t\|^2_{\mathrm{L}^{p,2}(\mathbb{R}^d)} + C \left\|\tilde{f}_t\right\|^2_{\mathrm{L}^{p+r,q}(\mathbb{R}^d)} \left\|b_t - \tilde{b}_t\right\|^2_{\mathrm{L}^{-r,q'}(\mathbb{R}^d;\mathbb{R}^d)}.$$

Hence, by Grönwall's inequality, we have

$$\|g_t\|^2_{\mathrm{L}^{p,2}(\mathbb{R}^d)} \leq C \int_0^t \left\|\tilde{f}_s\right\|^2_{\mathrm{L}^{p+r,q}(\mathbb{R}^d)} \left\|b_s - \tilde{b}_s\right\|^2_{\mathrm{L}^{-r,q'}(\mathbb{R}^d;\mathbb{R}^d)} ds,$$

which implies that

$$\|g_t\|_{\mathrm{L}^{p,2}(\mathbb{R}^d)} \leq C \int_0^t \left\|\tilde{f}_s\right\|_{\mathrm{L}^{p+r,q}(\mathbb{R}^d)} \left\|b_s - \tilde{b}_s\right\|_{\mathrm{L}^{-r,q'}(\mathbb{R}^d;\mathbb{R}^d)} ds.$$

Thus it suffices to obtain a bound, independent of $\tilde{b}$,

$$\left\|\tilde{f}\right\|_{\mathrm{L}^\infty([0,T];\mathrm{L}^{p+r,q}(\mathbb{R}^d))} \leq C.$$

This may be done using the equation for $\tilde{f}_t$, the assumed L^q moment bound on f_0 and the same technique as above multiplying by $|\tilde{f}_t|^{q-1}\langle x\rangle^{q(p+r)}$ instead of $g_t\langle x\rangle^{2p}$. We omit the details. □

The proof of the weighted energy estimate in the second order case is slightly different.

Proof of Lemma 2.5.4. As in the proof of Lemma 2.5.2, let $f_t = f_t^b$, $\tilde{f}_t = f_t^{\tilde{b}}$ and $g = f - \tilde{f}$. Then g solves

$$\begin{cases} \partial_t g_t + v\cdot\nabla_x g_t - \kappa\nabla_v\cdot(vg_t) + b_t\cdot\nabla_v g_t - \dfrac{1}{2}\Delta_v g_t = -(b_t-\tilde{b}_t)\cdot\nabla_v\tilde{f}_t, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{for } (t,x,v)\in[0,T]\times\mathbb{R}^d\times\mathbb{R}^d, \\ g_0 = 0. \end{cases}$$

By multiplying this equation by $g_t\langle(x,v)\rangle^{2p}$ and then integrating by parts, we obtain the following weighted energy estimate

$$\begin{aligned} &\frac{d}{dt}\|g_t\|^2_{\mathrm{L}^{p,2}(\mathbb{R}^{2d})} + \int v\cdot\nabla_x(g_t\langle(x,v)\rangle^{2p})g_t\,dxdv + \kappa\int g_t v\cdot\nabla_v(g_t\langle(x,v)\rangle^{2p})\,dxdv \\ &- \int g_t b_t\cdot\nabla_v(g_t\langle(x,v)\rangle^{2p})\,dxdv + \frac{1}{2}\int\nabla_v g_t\cdot\nabla_v(g_t\langle(x,v)\rangle^{2p})\,dxdv \\ &= \int\tilde{f}_t(b_t-\tilde{b}_t)\cdot\nabla_v(g_t\langle(x,v)\rangle^{2p})\,dxdv. \end{aligned}$$

In the similar way as in the proof of Lemma 2.5.2, as ∇_x, ∇_v hitting $\langle(x,v)\rangle^{2p}$ give terms of lower order and as $b_t, \tilde{b}_t$ are independent of v, we have

$$\begin{aligned} &\frac{d}{dt}\|g_t\|^2_{\mathrm{L}^{p,2}(\mathbb{R}^{2d})} + \|\nabla_v g_t\|^2_{\mathrm{L}^{p,2}(\mathbb{R}^{2d};\mathbb{R}^d)} \leq C\|g_t\|^2_{\mathrm{L}^{p,2}(\mathbb{R}^{2d})} + \\ &\quad + C\left\|b_t-\tilde{b}_t\right\|_{\mathrm{L}^{-r,q'}(\mathbb{R}^d;\mathbb{R}^d)}\left\|\tilde{f}_t\right\|_{\mathrm{L}^{p+r,q}(\mathbb{R}^{2d})}\left(\|g_t\|_{\mathrm{L}^{p,2}(\mathbb{R}^{2d})} + \|\nabla_v g_t\|_{\mathrm{L}^{p,2}(\mathbb{R}^{2d};\mathbb{R}^d)}\right) \end{aligned}$$

By using Young's inequality and then the Grönwall inequality we obtain, as in

the proof of Lemma 2.5.2,

$$\|g_t\|_{\mathrm{L}^{p,2}(\mathbb{R}^{2d})} \leq C \int_0^t \left\|\tilde{f}_s\right\|_{\mathrm{L}^{p+r,q}(\mathbb{R}^{2d})} \left\|b_s - \tilde{b}_s\right\|_{\mathrm{L}^{-r,q'}(\mathbb{R}^d;\mathbb{R}^d)} ds.$$

The claim of the lemma then follows from a bound on $\sup_{t\in[0,T]} \left\|\tilde{f}_t\right\|_{\mathrm{L}^{p+r,q}(\mathbb{R}^{2d})}$ which may be obtained using the assumption that $f_0 \in \mathrm{L}^{p+r,q}(\mathbb{R}^{2d})$ and similar energy estimates to the above. We leave this to the reader. □

2.6 Counterexample

In this section we will prove Proposition 2.2.2. We begin by introducing a sorting problem, which if the uniform law of large numbers holds over a class $\mathcal{C}$, is unsolvable.

Problem 2.6.1. *Given a class of vector fields $\mathcal{C}$ and even N, consider N particles evolving as (2.2.5) with initial law f_0. Tag the first $N/2$ particles* red *and the rest* blue. *Can we choose a (random) $b^N \in \mathcal{C}$ (depending on N) so that the red and blue particles are sorted to the right and left respectively, uniformly in N, i.e.*

$$\inf_{N\to\infty} \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} h(X_t^{b^N,i,N}, \mathrm{colour}(X^{b^N,i,N})) - \mathbb{E}h(X_t^{b^N,i,N}, \mathrm{colour}(X^{b^N,i,N}))\right] > 0 \tag{2.6.1}$$

where

$$h(x_1,\dots,x_d,c) = g(x_1)\begin{cases} 1 & \text{if } c = \text{red} \\ -1 & \text{if } c = \text{blue} \end{cases} \tag{2.6.2}$$

and $g(x)$ is a smoothed version of the sign function.

A simple argument by contradiction implies the following lemma.

Lemma 2.6.1. *If Problem 2.6.1 is solvable for a class of vector fields $\mathcal{C}$, then the uniform law of large numbers for SDEs cannot hold over this class. In particular, if Problem 2.6.1 is solvable for $\mathcal{C}^0$ given by (2.2.2) then Proposition 2.2.2 is true.*

We will now exhibit an explicit vector field $b \in \mathcal{C}^0$ that solves Problem 2.6.1, thus proving Proposition 2.2.2. This turns out to be quite simple.

Firstly, we note that, whatever $b \in \mathcal{C}^0$ is chosen, the set of times at which any two particles are in the same position is of measure zero almost surely. This is due to the absolute continuity with respect to Brownian motion due to Girsanov's theorem. Define the function $\psi_\varepsilon(t)$ as

$$\psi_\varepsilon(t) = \prod_{i=1}^{N/2} \prod_{j=N/2+1}^{N} \psi\left(\frac{X_t^{b^N,i,N} - X_t^{b^N,j,N}}{\varepsilon}\right) \tag{2.6.3}$$

for $\varepsilon > 0$ arbitrary, and $\psi(x)$ a function that is zero for $|x| \leq 1/2$ and 1 for $|x| \geq 1$. Then ψ_ε is adapted, almost surely continuous and zero whenever a red and blue particle are within $\varepsilon/2$ of each other, and 1 when no such particles are within ε of each other. Furthermore, it is a simple computation to show that

$$\lim_{\varepsilon \to 0} \mathbb{E} \int_0^T 1_{\psi_\varepsilon(t) \neq 1} \, dt = T,$$

no matter what $b^N \in \mathcal{C}^0$ is chosen, although (of course) this limit will not be uniform in N. Now let $\varepsilon > 0$ be chosen so that the expectation of the above integral is at least $3T/4$.

Let $\eta_\varepsilon(x) = \eta(x/\varepsilon)$ where $\eta(x)$ is a smooth bump function with $\eta(0) = 1$ and $\eta(x) = 0$ for $|x| \geq 1/2$. Now define b^N as

$$b^N(x) = \psi_\varepsilon(t) \left(\sum_{i=1}^{N/2} \eta_\varepsilon(x - X_t^{b^N,i,N}) - \sum_{i=N/2+1}^{N} \eta_\varepsilon(x - X_t^{b^N,i,N}) \right),$$

then b^N is adapted, uniformly bounded by 1, smooth in x and continuous in t almost surely. Moreover, when $\psi_\varepsilon(t) = 1$, b^N is equal to 1 on every red particle and -1 on every blue particle. Therefore, every red particle is pushed by b^N at least the distance $\int_0^T 1_{\psi_\varepsilon = 1} - 1_{\psi_\varepsilon \neq 1} \, dt$ to the right and similarly every blue particle to the left. By choice of ε the expectation of this is at least $T/2$. Hence, the expectation (2.6.1) is bounded away from zero by a fixed constant independent of N, completing the proof.

2.A Metric entropy

In this section we summarise the properties and estimates of metric entropy that are used in the rest of the chapter. The results henceforth are either well known or simple corollaries of well known results. The reader is encouraged to consult [54, 189] for an exposition of metric entropy in the context of functional analysis and [193] for a more statistical viewpoint (cf. [161]).

Lemma 2.A.1 (Metric entropy of product spaces)**.** *Let (X, d_X), (Y, d_Y) be totally bounded metric spaces. Define the product metric $d_{X\times Y}$ on $X \times Y$, by*

$$d_{X\times Y}((x, y), (x', y')) = \max(d_X(x, x'), d_Y(y, y')).$$

Then it holds that

$$H(\varepsilon, X \times Y, d_{X\times Y}) \leq H(\varepsilon, X, d_X) + H(\varepsilon, Y, d_Y).$$

In particular, if $H(\varepsilon, X, d_X) \leq C\varepsilon^{-k_X}$ and $H(\varepsilon, Y, d_Y) \leq C\varepsilon^{-k_Y}$ then we have the bound $H(\varepsilon, X \times Y, d) \leq C\varepsilon^{-\max(k_X,k_Y)}$ for any metric d equivalent to $d_{X\times Y}$.

Proof. Let $x_1, \dots x_n$ be an ε-net of (X, d_X) and $y_1, \dots, y_m$ be an ε-net of (Y, d_Y). Then $\{(x_i, y_j) : 1 \leq i \leq n, 1 \leq j \leq m\}$ is an ε-net of $(X \times Y, d_{X\times Y})$. The claims follow. □

Lemma 2.A.2 (Metric entropy of finite dimensional spaces)**.** *Let K be a compact set in $\mathbb{R}^d$, and $|\cdot|$ be the Euclidean norm*

$$H(\varepsilon, K, |\cdot|) \leq C \log(1/\varepsilon).$$

Proof. It suffices to consider $K = [0, 1]^d$ and by Lemma 2.A.1 we need only consider $K = [0, 1]$. Then an explicit ε-net is given by $\{k\varepsilon : k \in \mathbb{N}, k \leq 1/\varepsilon\}$. □

Lemma 2.A.3 (Change of metric)**.** *Let $\mathcal{C}$ be a totally bounded subset of a metric space (X, d), then it holds that*

$$H(\varepsilon, \mathcal{C}, d^\alpha) \leq CH(\epsilon^{1/\alpha}, \mathcal{C}, d) \tag{2.A.1}$$

where $d^\alpha(x,x') = |d(x,x')|^\alpha$ *for* $\alpha \in (0,1]$*. In particular, if* $H(\varepsilon, \mathcal{C}, d^\alpha) \leq C\varepsilon^{-k}$ *then* $H(\varepsilon, \mathcal{C}, d^\alpha) \leq C\varepsilon^{-k/\alpha}$*.*

Proof. Let $(x_n)_{n=1}^m$ be a ε-net with respect to d of $\mathcal{C}$, then $(x_n)_{n=1}^m$ is also an ε^α net of $\mathcal{C}$ with respect to d^α. The particular claim follows easily. □

The main estimates of metric entropy required for the rest of the chapter are given in the proposition below.

Proposition 2.A.1 (Metric entropy of smooth functions)**.** *Let* $p > 1$*, then the following hold:*

1. ***Lipschitz functions:*** *Let* $p \neq 2$*, then the Lipschitz functions on* $\mathbb{R}^d$ *obey:*

$$H(\varepsilon, \mathrm{Lip}1, \|\cdot\|_{\mathrm{L}^{-p,\infty}(\mathbb{R}^d)}) \leq C\varepsilon^{-\min(d, d/(p-1))}.$$

2. ***Hölder functions:*** *For* $\alpha \in (0,1]$*, the Hölder functions* $\mathcal{C}^\alpha = \{f \in \mathrm{C}^{0,\alpha}(\mathbb{R}^d) : \|f\|_{\mathrm{C}^{0,\alpha}(\mathbb{R}^d)} \leq C\}$*, obey the bound:*

$$H(\varepsilon, \mathcal{C}^\alpha, \|\cdot\|_{\mathrm{L}^{-p,\infty}(\mathbb{R}^d)}) \leq C\varepsilon^{-d/\alpha}.$$

3. ***Parabolic Hölder functions:*** *For* $\alpha \in (0,1]$*, the parabolic Hölder functions*

$$\mathcal{C}^\alpha_{para} = \{f \in \mathrm{C}^{0,\alpha}_{para}([0,T] \times \mathbb{R}^d) : \|f\|_{\mathrm{C}^{0,\alpha}_{para}([0,T]\times\mathbb{R}^d)} \leq C\},$$

obey the bound:

$$H(\varepsilon, \mathcal{C}^\alpha_{para}, \|\cdot\|_{\mathrm{L}^\infty([0,T];\mathrm{L}^{-p,\infty}(\mathbb{R}^d))}) \leq C\varepsilon^{-(d+2)/\alpha}.$$

4. ***Parabolic*** L^q ***Hölder functions:*** *Let* $k \in \{0,1,\dots\}$*,* $\alpha \in (0,1]$ *and* $q \in [1,\infty]$ *with* $k + \alpha > (d+2)/q$*. The set of parabolic* L^q *Hölder functions* $\mathcal{C}^{k,\alpha}_{q,para} = \{f \in \Lambda^{k,\alpha}_{para}(\mathrm{L}^q([0,T]\times\mathbb{R}^d)) : \|f\|_{\Lambda^{k,\alpha}_{para}(\mathrm{L}^q([0,T]\times\mathbb{R}^d))} \leq C\}$ *obeys the bound:*

$$H(\varepsilon, \mathcal{C}^{k,\alpha}_{q,para}, \|\cdot\|_{\mathrm{L}^\infty([0,T];\mathrm{L}^{-p,\infty}(\mathbb{R}^d))}) \leq C\varepsilon^{-(d+2)/(k+\alpha)}.$$

for any $p > d/q$*.*

5. L^q ***Hölder functions:*** *Let $k \in \{0, 1, \dots\}$, $\alpha \in (0, 1]$ and $q \in [1, \infty]$ with $k + \alpha > (d+1)/q$. The set of L^q Hölder functions*

$$\mathcal{C}_q^{k,\alpha} = \{f \in \Lambda^{k,\alpha}(\mathrm{L}^q([0,T] \times \mathbb{R}^d)) : \|f\|_{\Lambda^{k,\alpha}(\mathrm{L}^q([0,T]\times\mathbb{R}^d))} \leq C\},$$

obeys the bound:

$$H(\varepsilon, \mathcal{C}_q^{k,\alpha}, \|\cdot\|_{\mathrm{L}^\infty([0,T];\mathrm{L}^{-p,\infty}(\mathbb{R}^d))}) \leq C\varepsilon^{-(d+1)/(k+\alpha)},$$

for any $p > (k+\alpha) - (d+1)/q$.

In all cases the estimates for vector valued (i.e. in $\mathbb{R}^n$) functions are the same up to change in constants due to Lemma 2.A.1.

Proof. We prove each in turn.

1. Let $h \in \mathrm{Lip1}$ be arbitrary, then we have $h(0) = 0$ and therefore $|h(x)| \leq |x|$. As a result we have $x \mapsto h(x)\langle x \rangle \in \mathrm{C}_b^1(\mathbb{R}^d)$ with a bound on the C_b^1 norm independent of $h \in \mathrm{Lip1}$. Let B be the unit ball in C_b^1. We deduce that

   $$H(\varepsilon, \mathrm{Lip1}, \|\cdot\|_{\mathrm{L}^{-p,\infty}(\mathbb{R}^d)}) \leq H(\varepsilon, B, \|\cdot\|_{\mathrm{L}^{-(p-1),\infty}(\mathbb{R}^d)}).$$

   The bound then follows from the estimates on entropy numbers in [54] (cf. [161] for an exposition in terms of metric entropy).

2. This result follows directly from [161, Corollary 3.1].

3. This follows from Lemma 2.A.4 below and the identification (2.A.2),(2.A.2) with $d_1 = 1$, $d_2 = d$, $p = \infty$, and the parabolic anisotropy $\mathfrak{p}$ defined below.

4. This follows from Lemma 2.A.4 in the same way as (3).

5. This follows from [161, Theorem 1.1]. □

We now provide a simple estimate on the metric entropy of weighted spaces of anisotropic regularity, which was we needed for Proposition 2.A.1(3)-(4). We make no claim of optimality or originality in this result, which the author has included because of the inability to find a reference.

Definition 2.A.1 (Anisotropy)**.** *An* anisotropy *of* $\mathbb{R}^d$ *is tuple* $\mathfrak{a} = (\mathfrak{a}_1, \dots, \mathfrak{a}_d) \in \mathbb{R}^d$ *such that*

$$\mathfrak{a}_i > 0 \text{ for each } i = 1, \dots, d, \text{ and } \sum_{i=1}^{d} \mathfrak{a}_i = d.$$

An anisotropy $\mathfrak{a}$ corresponds to the anisotropic distance $|\cdot|_{\mathfrak{a}}$ on $\mathbb{R}^d$ given by

$$|x|_{\mathfrak{a}} = |(x_1, \dots, x_d)|_{\mathfrak{a}} = \sum_{i=1}^{d} |x_i|^{\mathfrak{a}_i}.$$

Note that for $\mathfrak{a} = (1, \dots, 1)$ the distance $|\cdot|_{\mathfrak{a}}$ is equivalent to the usual Euclidean distance on $\mathbb{R}^d$.

We record in particular that

$$\mathfrak{p} := \frac{d+1}{d+(1/2)}(1/2, \underbrace{1, 1, \dots, 1}_{d \text{ times}}) \in \mathbb{R}^{1+d}$$

is the *parabolic* anisotropy. Here the prefactor is to ensure that the sum of the indices is $d + 1$.

Given an anisotropy $\mathfrak{a}$ and a subset $U \subseteq \mathbb{R}^d$ it is possible to define the Besov space of anisotropic regularity $B^{s,\mathfrak{a}}_{p,q}(U)$ for $p, q \in (0, \infty]$ and $s \in \mathbb{R}$. As we do not require the full (somewhat lengthy) definition of these spaces we do not provide them. Instead we refer the reader to [189, §5] for their full definition and for more details about anisotropies and spaces of anisotropic regularity.

For our purposes it is sufficient to note that

$$\Lambda^{k,\alpha}_{para}(\mathrm{L}^q([0,T] \times \mathbb{R}^d)) = B^{\mathfrak{p},s}_{q,q}([0,T] \times \mathbb{R}^d) \tag{2.A.2}$$

for any $q \in [1, \infty]$, non-negative integer k, $\alpha \in (0, 1]$ and

$$s = \frac{d+1}{d+2}(k + \alpha). \tag{2.A.3}$$

Lemma 2.A.4. *Fix* $d_1, d_2 \in \mathbb{N}$, $s > 0, \in [1, \infty], r > 0$ *and an anisotropy* $\mathfrak{a}$ *such that* $s > d/q$ *and* $r > d_2/q$ *where* $d = d_1 + d_2$*. Define the set* $\mathcal{C}$ *as those functions*

$h \in B^{\mathfrak{a},s}_{q,q,loc}([0,1]^{d_1} \times \mathbb{R}^{d_2})$ *satisfying the estimate*

$$\sup_{Q} \|h\|_{B^{s,\mathfrak{a}}_{q,q}([0,1]^{d_1} \times Q)} \leq C'$$

for a fixed constant C'*, where the supremum is over unit cubes* $Q \subseteq \mathbb{R}^{d_2}$*. Then,*

$$H(\varepsilon, \mathcal{C}, \|\cdot\|_{\mathrm{L}^{-r,\infty}([0,1]^{d_1} \times \mathbb{R}^{d_2})}) \leq C\varepsilon^{-d/s}.$$

Proof of Lemma 2.A.4. Fix $\varepsilon > 0$ and let $R = R(\varepsilon)$ to be chosen be a constant. (All constants C will be uniform in ε and R). For $n \in \mathbb{Z}^{d_2}$, let Q_n be the cube $[0,1]^{d_1} \times (n + [-1/2, 1/2]^{d_2})$ and let $\mathcal{Q} = \{Q_n : |n| \leq R\}$. Define the set E by

$$E = \prod_{Q_n \in \mathcal{Q}} E_n$$

where E_n is a $\varepsilon|n|^r$-net of $\mathcal{C}|_{Q_n}$ (the restriction of functions in $\mathcal{C}$ to Q_n). Note that $H(\varepsilon|n|^r, \mathcal{C}|_{Q_n}, \|\cdot\|_{\mathrm{L}^\infty(Q_n)}) \leq C|n|^{-rd/s}\varepsilon^{-d/s}$ by [189, theorem 5.30.]. Therefore, E_n can be chosen so that $\log|E_n| \leq C|n|^{-rd/s}\varepsilon^{-d/s}$ and hence

$$\log|E| \leq C\varepsilon^{-k} \sum_{n \in \mathbb{Z}^{d_2}, |n| < R} |n|^{-rd/s} \leq C\varepsilon^{-d/s} \sum_{n \in \mathbb{Z}^{d_2}} |n|^{-rd/s} \leq C\varepsilon^{-d/s}$$

as $rd/s > d_2$ by assumption, so the sum is finite.

Define the set of functions F, by collecting for each $(e_n)_{n \in \mathbb{Z}^{d_2}, |n| \leq R} \in E$ a function $h \in \mathcal{C}$ with the property that for each n, $\|h - e_n\|_{\mathrm{L}^\infty(Q_n)} \leq \varepsilon|n|^r$ if such a function exists. Then $\log|F| \leq \log|E| \leq C\varepsilon^{-d/s}$ and we claim that F is a $C\varepsilon$-net of $\mathcal{C}$ in the $\mathrm{L}^{-r,\infty}([0,1]^{d_1} \times \mathbb{R}^{d_2})$. Indeed, suppose that $h \in \mathcal{C}$, then for each cube Q_n ($|n| < R$) we have a function $e_n \in E_n$ with $\|h - e_n\|_{\mathrm{L}^\infty(Q_n)} \leq \varepsilon|n|^r$ by construction. Then by construction of F we have $g \in F$ with $\|g - e_n\|_{\mathrm{L}^\infty(Q_n)} \leq \varepsilon|n|^r\varepsilon$, (such a function must exist as we could have chosen h in the construction of F). Therefore,

$$\|g - h\|_{\mathrm{L}^{-r,\infty}(\bigcup_{Q_n \in \mathcal{Q}} Q_n)} \leq \|g - e\|_{\mathrm{L}^{-r,\infty}(\bigcup_{Q_n \in \mathcal{Q}} Q_n)} + \|g - e\|_{\mathrm{L}^{-r,\infty}(\bigcup_{Q_n \in \mathcal{Q}} Q_n)} \leq C\varepsilon$$

where $e(x) = e_n(x)$ for $x \in Q_n$. While on $\{x : |x| > R\}$ we have that $|h - g| \langle x \rangle^{-r} \leq 2CR^{-r}$ as functions in $\mathcal{C}$ are uniformly bounded by Sobolev embedding. By choosing R sufficiently large we may ensure that this is less than ε. □

Chapter 3

Stability and instability in gradient dynamics - Part I

It is known that for a strictly concave-convex function, the gradient method introduced by Arrow, Hurwicz and Uzawa [10], has guaranteed global convergence to its saddle point. Nevertheless, there are classes of problems where the function considered is not strictly concave-convex, in which case convergence to a saddle point is not guaranteed. In the chapter we provide a characterization of the asymptotic behaviour of the gradient method, in the general case where this is applied to a general concave-convex function. We prove that for any initial conditions the gradient method is guaranteed to converge to a trajectory described by an explicit linear ODE. In the following Chapter 4 this study is extended to the subgradient method, where the dynamics are constrained in a prescribed convex set.

Acknowledgements

The work in this chapter was done in collaboration with Ioannis Lestas and forms the first part of a two part paper [92] in preparation.

3.1 Introduction

Finding the saddle point of a concave-convex function is a problem that is relevant in many applications in engineering and economics and has been addressed by various communities. It includes, for example, optimization problems that are reduced to finding the saddle point of a Lagrangian. The gradient method, first introduced by Arrow, Hurwicz and Uzawa [10] has been widely used in this context as it leads to decentralized update rules for network optimization problems. It has therefore been extensively used in areas such as resource allocation in communication and economic networks (e.g. [97], [110], [183], [61]).

Nevertheless, in broad classes of problems there are features that render the analysis of the asymptotic behaviour of gradient dynamics nontrivial. In particular, although for a strictly concave-convex function convergence to a saddle point via gradient dynamics is ensured, when this strictness is lacking, convergence is not guaranteed and oscillatory solutions can occur. The existence of such oscillations has been reported in specific applications [10], [61], [94], [165], however, an exact characterization of those for a general concave-convex function has not been studied in the literature and is one of the aims of Part I of this work.

Furthermore, when subgradient methods are used to restrict the dynamics in a convex domain (needed, e.g., in optimization problems), the dynamics become non-smooth in continuous-time. This increases significantly the complexity in the analysis as classical Lyapunov and LaSalle type techniques (e.g. [112]) cannot be applied. This is also reflected in the alternative approach taken for the convergence proof in [10] for subgradient dynamics applied to a strictly concave-convex Lagrangian with positivity constraints. Furthermore, an interesting recent study [38] pointed out that the invariance principle for hybrid automata in [135] cannot be applied in this context, and gave an alternative proof, by means of Caratheodory's invariance principle, to the convergence result in [10] mentioned above. In general, rigorously proving convergence for the subgradient method, even in what would naively appear to be simple cases, is a non-trivial problem, and requires much machinery from non-smooth analysis.

Our aim in this and the following chapter is to carry out a detailed study of the asymptotic behaviour of continuous-time gradient dynamics in a general setting,

where the function with respect to which these dynamics are applied is not necessarily *strictly* concave-convex. Furthermore, we provide a framework of results that allow one to study the asymptotic behaviour of the subgradient method (3.3.2) with *smooth analysis* as opposed to *non-smooth analysis*.

Our main contributions can be summarized as follows:

- In this chapter, we consider the gradient method applied on a general concave-convex function in an unconstrained domain, and provide an exact characterization to the limiting solutions, which can in general be oscillatory. In particular, we show that despite the nonlinearity of the dynamics the trajectories converge to solutions that satisfy a linear ODE that is explicitly characterized. Furthermore, we show that when such oscillations occur, the dynamic behaviour can be problematic, in the sense that arbitrarily small stochastic perturbations can lead to an unbounded variance.

- In the following Chapter 4, we consider the subgradient method applied to a general concave-convex function with the trajectories restricted in a general convex domain. We show that despite the non-smooth character of these dynamics, their limiting behaviour is given by the solutions of one of an explicit family of *smooth* differential equations. This therefore allows to remove the complications associated with non-smooth analysis, and prove convergence in broad classes of problems.

It should be noted that there is a direct link between the results in Part I and Part II as the smooth dynamics, that are proved to be associated with the asymptotic behaviour of the subgradient method, are a class of dynamics that can be analysed with the framework introduced in Part I. Applications of the results in Part I will therefore be discussed in Part II, as in many cases (e.g. optimization problems with inequality constraints) a restricted domain for the concave-convex function needs to be considered.

Finally, we would also like to comment that the methodology used for the derivations in the two chapters is of independent technical interest. In Part I the analysis is based on various geometric properties established for the saddle points of a concave-convex function. In Part II the non-smooth analysis is carried out by means of some more abstract results on dynamical systems that are applicable

in this context, while also making use of the notion of a face of a convex set to characterize the asymptotic behaviour of the dynamics.

This chapter is structured as follows. In Section 3.2 we introduce various definitions and preliminaries that will be used throughout the chapter. In Section 3.3 the problem formulation is given and the main results are presented in Section 3.4, i.e. characterization of the limiting behaviour of gradient dynamics. This section also includes an extension to a class of subgradient dynamics that restrict the trajectories on affine spaces. This is a technical result that will be used in Part II to characterize the limiting behaviour of general subgradient dynamics. The proofs of the results are finally given in Section 3.6.

3.2 Preliminaries

3.2.1 Notation

Real numbers are denoted by $\mathbb{R}$ and non-negative real numbers as $\mathbb{R}_+$. For vectors $x, y \in \mathbb{R}^n$ the inequality $x < y$ holds if the corresponding inequality holds for each pair of components, $d(x, y)$ is the Euclidean metric and $|x|$ denotes the Euclidean norm.

The space of $k-$times continuously differentiable functions is denoted by C^k. For a sufficiently differentiable function $f(x, y) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ we denote the vector of partial derivatives of f with respect to x as f_x, respectively f_y. The Hessian matrices with respect to x and y are denoted f_{xx} and f_{yy} with f_{xy} and f_{yx} denoting the matrices of mixed partial derivatives in the appropriate arrangement. For a vector valued function $g : \mathbb{R}^n \to \mathbb{R}^m$ we let g_x denote the matrix formed by partial derivatives of the elements of g.

For a matrix $A \in \mathbb{R}^{n \times m}$ we denote its kernel and transpose by $\ker(A)$ and A^T respectively. If A is in addition symmetric, we write $A < 0$ if A is negative definite.

3.2.1.1 Geometry

For subspaces $E \subseteq \mathbb{R}^n$ we denote the orthogonal complement as $E^\perp$, and for a set of vectors $E \subseteq \mathbb{R}^n$ we denote their span as $\text{span}(E)$, their affine span as $\text{aff}(E)$ and their convex hull as $\text{Conv}(E)$. The addition of a vector $v \in \mathbb{R}^n$ and a set $E \subseteq \mathbb{R}^n$ is defined as $v + E = \{v + u : u \in E\}$.

For a set $K \subset \mathbb{R}^n$, we denote the interior, relative interior, boundary and closure of K as $\operatorname{int} K$, $\operatorname{relint} K$, ∂K and $\overline{K}$ respectively, and we say that K and M are orthogonal and write $K \perp M$ if for any two pairs of points $\mathbf{k}, \mathbf{k}' \in K$ and $\mathbf{m}, \mathbf{m}' \in M$, we have $(\mathbf{k}' - \mathbf{k})^T(\mathbf{m} - \mathbf{m}') = 0$.

Given a set $E \subseteq \mathbb{R}^n$ and a function $\phi : E \to E$ we say that ϕ is an isometry of (E, d) or simply an isometry, if for all $x, y \in E$ we have $d(\phi(x), \phi(y)) = d(x, y)$.

For $x \in \mathbb{R}, y \in \mathbb{R}_+$ we define $[x]_y^+ = x$ if $y > 0$ and $\max(0, x)$ if $y = 0$.

3.2.1.2 Convex geometry

When we consider a concave-convex function $\varphi(x, y) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ (see Definition 3.2.1) we shall denote the pair $\mathbf{z} = (x, y) \in \mathbb{R}^{n+m}$ in bold, and write $\varphi(\mathbf{z}) = \varphi(x, y)$. The full Hessian matrix will then be denoted $\varphi_{\mathbf{zz}}$. Vectors in $\mathbb{R}^{n+m}$ and matrices acting on them will be denoted in bold font (e.g. $\mathbf{A}$). Saddle points (see Definition 3.2.2) of φ will be denoted $\bar{\mathbf{z}} = (\bar{x}, \bar{y}) \in \mathbb{R}^{n+m}$.

For a closed convex set $K \subseteq \mathbb{R}^n$ and $\mathbf{z} \in \mathbb{R}^n$, we define the maximal orthogonal linear manifold to K through $\mathbf{z}$ as

$$M_K(\mathbf{z}) = \mathbf{z} + \text{span}(\{\mathbf{u} - \mathbf{u}' : \mathbf{u}, \mathbf{u}' \in K\})^\perp \tag{3.2.1}$$

and the normal cone to K through $\mathbf{z}$ as

$$N_K(\mathbf{z}) = \{\mathbf{w} \in \mathbb{R}^n : \mathbf{w}^T(\mathbf{z}' - \mathbf{z}) \leq 0 \text{ for all } \mathbf{z}' \in K\}. \tag{3.2.2}$$

When K is an affine space $N_K(\mathbf{z})$ is independent of $\mathbf{z} \in K$ and is denoted N_K. If K is in addition non-empty, then we define the projection of $\mathbf{z}$ onto K as

$\mathbf{P}_K(\mathbf{z}) = \text{argmin}_{\mathbf{w}\in K}\, d(\mathbf{z}, \mathbf{w})$.

3.2.2 Convex analysis

3.2.2.1 Concave-convex functions and saddle points

Definition 3.2.1 (Concave-convex function)**.** *Let $K \subseteq \mathbb{R}^{n+m}$ be non-empty closed and convex. We say that a function $\varphi(x, y) : K \to \mathbb{R}$ is concave-convex on K if for any $(x', y') \in K$, $\varphi(x, y')$ is a concave function of x and $\varphi(x', y)$ is a convex function of y. If either the concavity or convexity is always strict, we say that φ is strictly concave-convex on K.*

Definition 3.2.2 (Saddle point)**.** *For a concave-convex function $\varphi : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ we say that $(\bar{x}, \bar{y}) \in \mathbb{R}^{n+m}$ is a saddle point of φ if for all $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$ we have the inequality $\varphi(x, \bar{y}) \leq \varphi(\bar{x}, \bar{y}) \leq \varphi(\bar{x}, y)$.*

If φ is in addition C^1 then $(\bar{x}, \bar{y})$ is a saddle point if and only if $\varphi_x(\bar{x}, \bar{y}) = 0$ and $\varphi_y(\bar{x}, \bar{y}) = 0$.

3.2.2.2 Dynamical systems

Definition 3.2.3 (Flows and semi-flows)**.** *A triple (ϕ, X, ρ) is a flow (resp. semi-flow) if (X, ρ) is a metric space, ϕ is a continuous map from $\mathbb{R}\times X$ (resp. $\mathbb{R}_+ \times X$) to X which satisfies the two properties*

(i) For all $x \in X$, $\phi(0, x) = x$.

(ii) For all $x \in X$, $t, s \in \mathbb{R}$ (resp. $\mathbb{R}_+$),

$$\phi(t + s, x) = \phi(t, \phi(s, x)). \tag{3.2.3}$$

When there is no confusion over which (semi)-flow is meant, we shall denote $\phi(t, x(0))$ as $x(t)$. For sets $A \subseteq \mathbb{R}$ (resp. $\mathbb{R}_+$) and $B \subseteq X$ we define $\phi(A, B) = \{\phi(t, x) : t \in A, x \in B\}$.

Definition 3.2.4 (Global convergence)**.** *We say that a (semi)-flow (ϕ, X, ρ) is globally convergent, if for all initial conditions $x \in X$, the trajectory $\phi(t, x)$ converges to the set of equilibrium points of (ϕ, X, ρ) as $t \to \infty$, i.e.*

$$\inf\{d(\phi(t, x), y) : y \text{ an equilibrium point}\} \to 0 \text{ as } t \to \infty.$$

A specific form of incremental stability, which we will refer to as pathwise stability, will be needed in the analysis that follows.

Definition 3.2.5 (Pathwise stability)**.** *We say that a semi-flow (ϕ, X, ρ) is pathwise stable*[1] *if for any two trajectories $x(t), x'(t)$ the distance $\rho(x(t), x'(t))$ is non-increasing in time.*

As the subgradient method has a discontinuous vector field we need the notion of Carathéodory solutions of differential equations.

Definition 3.2.6 (Carathéodory solution)**.** *We say that a trajectory $\mathbf{z}(t)$ is a Carathéodory solution to a differential equation $\dot{\mathbf{z}} = \mathbf{f}(\mathbf{z})$, if $\mathbf{z}$ is an absolutely continuous function of t, and for almost all times t, the derivative $\dot{\mathbf{z}}(t)$ exists and is equal to $\mathbf{f}(\mathbf{z}(t))$.*

3.3 Problem formulation

The main object of study in this work is the *gradient method* on an arbitrary concave-convex function in C^2.

Definition 3.3.1 (Gradient method)**.** *Given φ a C^2 concave-convex function on $\mathbb{R}^{n+m}$, we define the gradient method as the flow on $(\mathbb{R}^{n+m}, d)$ generated by the differential equation*

$$\begin{aligned} \dot{x} &= \varphi_x, \\ \dot{y} &= -\varphi_y. \end{aligned} \tag{3.3.1}$$

[1] Notions similar to this have been studied before by many different mathematical communities under different names, such as *monotone*, *contractive* or *dissipative* systems. As these terms are already used in the control community for different concepts, we shall not use them to avoid confusion.

It is clear that the saddle points of φ are exactly the equilibrium points of (3.3.1). In Section 3.A.1 we consider the addition of constant gains to the gradient method.

In the following Chapter 4 we study instead the *subgradient method* where the gradient method (Definition 3.3.1) is restricted to a convex set K by the addition of a projection term to the differential equation (3.3.1).

Definition 3.3.2 (Subgradient method)**.** *Given a non-empty closed convex set $K \subseteq \mathbb{R}^{n+m}$ and a C^2 function φ that is concave-convex on K, we define the subgradient method on K as a semi-flow on (K, d) consisting of Carathéodory solutions of*

$$\begin{aligned} \dot{\mathbf{z}} &= \mathbf{f}(\mathbf{z}) - \mathbf{P}_{N_K(\mathbf{z})}(\mathbf{f}(\mathbf{z})), \\ \mathbf{f}(\mathbf{z}) &= \begin{bmatrix} \varphi_x & -\varphi_y \end{bmatrix}^T. \end{aligned} \tag{3.3.2}$$

In Section 3.A.1 we consider the addition of constant gains to the subgradient method. The gradient method is then the subgradient method on $\mathbb{R}^{n+m}$. The equilibrium points of the subgradient method on K are exactly the K-restricted saddle points.

We briefly summarise the contributions of this work in the bullet points below.

- We provide an exact classification of the limiting solutions of the gradient method (3.3.1) applied to arbitrary concave-convex functions which is not assumed to be strictly concave-convex. Despite the non-linearity of the gradient dynamics, we show that these limiting solutions solve an explicit linear ODE given by derivatives of the concave-convex function at a saddle point.

- We apply this classification to give exact characterisations of limiting behaviour in special cases, for example in Lagrangians originating from optimisation problems.

- We show how the lack of convergence of the gradient method can lead to instability in the presence of noise.

- In the following Chapter 4 we show that the limiting behaviour of the

subgradient method on arbitrary convex domains is reduced to the limiting behaviour on affine subspaces. To assist in the analysis of these dynamics, we extend the exact classification of limiting solutions described in the first bullet point to the subgradient method on affine subspaces.

The simple form of the gradient method (Definition 3.3.1) and the subgradient method (Definition 3.3.2) makes them attractive methods for finding saddle points, and respectively restricted saddle points. This use dates back to Arrow, Hurwicz and Uzawa [10] who introduced the method.

More recently the localised structure of the system (3.3.1) when applied to network optimization problems has led to a renewed interest [61], [110], [183], [126], where the (sub)gradient method has been applied to various network resource allocation problems.

The study of the convergence of the (sub)gradient method was originated by Arrow, Hurwicz and Uzawa [10], who proved convergence of the gradient method, under the assumption of *strict* concave-convexity. However, in the absence of strict concave-convexity, the dynamics are more complex, and non-convergent oscillatory behaviour has been observed in some cases (see e.g. [61]).

The study of the subgradient method is much complicated by the discontinuity of the vector field, which prevents the application of the classical Lyapunov or LaSalle theorems or other tools of smooth analysis. This problem is the subject of the following Chapter 4 and we refer the reader to the discussion therein.

3.4 Main Results

This section presents the main results of the chapter. Before stating these we give some preliminary results.

It was proved in [90] that the gradient method is pathwise stable, which is stated in the proposition below. For the readers convenience we include a proof in an appendix.

Proposition 3.4.1. *Let φ be C^2 and concave-convex on $\mathbb{R}^{n+m}$, then the gradient*

method (3.3.1) *is pathwise stable.*

Because saddle points are equilibrium points of the gradient method we obtain the well known result below.

Corollary 3.4.1. *Let φ be C^2 and concave-convex on $\mathbb{R}^{n+m}$, then the distance of a solution of* (3.3.1) *to any saddle point is non-increasing in time.*

By an application of LaSalle's theorem we obtain:

Corollary 3.4.2. *Let φ be C^2 and concave-convex on $\mathbb{R}^{n+m}$, then the gradient method* (3.3.1) *converges to a solution of* (3.3.1) *which has constant distance from any saddle point.*

Thus classifying the limiting behaviour of the gradient method reduces to the problem of finding all solutions which lie a constant distance from *any* saddle point. In order to facilitate the presentation of the results, for a given concave-convex function φ we define the following sets:

- $\bar{\mathcal{S}}$ will denote the set of saddle points of φ.
- $\mathcal{S}$ will denote the set of solutions to (3.3.1) that are a constant distance from any saddle point of φ.

Note that if $\bar{\mathcal{S}} = \mathcal{S} \neq \emptyset$ then Corollary 3.4.2 gives the convergence of the gradient method to a saddle point.

Our first main result is that solutions of the gradient method converge to solutions that satisfy an explicit linear ODE.

To present our results we define the following matrices of partial derivatives of φ

$$\mathbf{A}(\mathbf{z}) = \begin{bmatrix} 0 & \varphi_{xy}(\mathbf{z}) \\ -\varphi_{yx}(\mathbf{z}) & 0 \end{bmatrix}, \qquad \mathbf{B}(\mathbf{z}) = \begin{bmatrix} \varphi_{xx}(\mathbf{z}) & 0 \\ 0 & -\varphi_{yy}(\mathbf{z}) \end{bmatrix}. \tag{3.4.1}$$

For simplicity of notation we shall state the result for $\mathbf{0} \in \bar{\mathcal{S}}$; the general case may be obtained by a translation of coordinates. It is common in applications to consider the gradient method with constant gains (3.A.1), as discussed in Section 3.A.1, the results stated here may be adapted to this case by a coordinate transformation.

Theorem 3.4.1. *Let φ be C^2 and concave-convex on $\mathbb{R}^{n+m}$. Let $\mathbf{0} \in \bar{\mathcal{S}}$ then solutions in $\mathcal{S}$ solve the linear ODE:*

$$\dot{\mathbf{z}}(t) = \mathbf{A}(\mathbf{0})\mathbf{z}(t). \tag{3.4.2}$$

Furthermore, a solution $\mathbf{z}(t)$ to (3.4.2) is in $\mathcal{S}$ if and only if for all $t \in \mathbb{R}$ and $r \in [0, 1]$,

$$\mathbf{z}(t) \in \ker(\mathbf{B}(r\mathbf{z}(t))) \cap \ker(\mathbf{A}(r\mathbf{z}(t)) - \mathbf{A}(\mathbf{0})) \tag{3.4.3}$$

where $\mathbf{A}(\mathbf{z})$ and $\mathbf{B}(\mathbf{z})$ are defined by (3.4.1).

To explain the significance of this result we make the following remarks.

Remark 3.4.1. *Despite the* non-linearity *of the gradient dynamics (3.3.1), the limiting solutions solve an* linear *ODE with explicit coefficients depending only on the derivatives of φ at the saddle point.*

Remark 3.4.2. *The strength of the result is that proving that there are no non-trivial limiting solutions implies global convergence of the gradient dynamics. In this way, the problem of showing global convergence reduces to checking for the existence of these limiting solutions.*

Remark 3.4.3. *The condition (3.4.3) appears to be very hard to check, as it requires knowledge of the trajectory for all times $t \in \mathbb{R}$. However, in applications to proving convergence this make the result* stronger*, as it makes it easier to prove that trajectories do* not *satisfy the condition.*

Remark 3.4.4 (Localisation)**.** *This result uses only local information about the concave-convex function φ, in the sense that if φ is only concave-convex on a convex subset $K \subseteq \mathbb{R}^{n+m}$ which contains $\mathbf{0}$, then any trajectory $\mathbf{z}(t)$ of the gradient method (3.3.1) that lies a constant distance from any saddle point in K and* does not leave K at any time t *will obey the conditions of the theorem.*

As a simple illustration of the use of this result we show how to recover the well known result that the gradient method is globally convergent under the assumption that φ is strictly concave-convex.

Example 3.4.1. *Suppose φ is strictly concave (the strictly convex case is similar), then φ_{xx} is of full rank except at isolated points, and the condition (3.4.3)*

can only hold if $x(t) = 0$. *Then the ODE* (3.4.2) *implies that* $y(t)$ *is constant, and hence* $(x(t), y(t))$ *is a saddle point. Thus the only limiting solution of the gradient method are the saddle points, which establishes global convergence.*

From Theorem 3.4.1 we deduce some further results that give a more easily understandable classification of the limiting solutions of the gradient method for simpler forms of φ.

In particular, the 'linear' case occurs when φ is a quadratic function, as then the gradient method (3.3.1) is a linear system of ODEs. In this case $\mathcal{S}$ has a simple explicit form in terms of the Hessian matrix of φ at $\mathbf{0} \in \bar{\mathcal{S}}$, and in general this provides an inclusion as described below, which can be used to prove global convergence of the gradient method using only local analysis at a saddle point.

Theorem 3.4.2. *Let* φ *be* C^2, *concave-convex on* $\mathbb{R}^{n+m}$ *and* $\mathbf{0} \in \bar{\mathcal{S}}$. *Then define*

$$\mathcal{S}_{linear} = \operatorname{span}\{\mathbf{v} \in \ker(\mathbf{B}) : \mathbf{v} \text{ is an eigenvector of } \mathbf{A}\} \tag{3.4.4}$$

where $\mathbf{A} = \mathbf{A}(\mathbf{0})$ *and* $\mathbf{B} = \mathbf{B}(\mathbf{0})$ *in* (3.4.1). *Then* $\mathcal{S} \subseteq \mathcal{S}_{linear}$ *with equality if* φ *is a quadratic function.*

Here we draw an analogy with the recent study [13] on the discrete time gradient method in the quadratic case. There the gradient method is proved to be semi-convergent if and only if $\ker(\mathbf{B}) = \ker(\mathbf{A} + \mathbf{B})$, i.e. if $\mathcal{S}_{\text{linear}} \subseteq \bar{\mathcal{S}}$. Theorem 3.4.2 includes a continuous time version of this statement.

Next we give an illustration of how the presence of oscillatory solutions to the gradient method can lead to instabilities. Consider the addition of a driving white noise to the dynamics (3.3.1). This gives the following stochastic differential equations

$$\begin{aligned} dx(t) &= \varphi_x dt + \Sigma^x dB^x(t) \\ dy(t) &= -\varphi_y dt + \Sigma^y dB^y(t) \end{aligned} \tag{3.4.5}$$

where $B^x(t), B^y(t)$ are independent standard Brownian motions in $\mathbb{R}^n, \mathbb{R}^m$ respectively, and Σ^x, Σ^y are positive definite symmetric matrices in $\mathbb{R}^{n\times n}, \mathbb{R}^{m\times m}$ respectively.

Theorem 3.4.3. *Let* $\varphi \in C^2$ *be concave-convex on* $\mathbb{R}^{n+m}$. *Let* $\mathbf{0} \in \bar{\mathcal{S}}$ *and* $\mathcal{S}$

contain a bi-infinite line. Consider the noisy dynamics (3.4.5). Then, for any initial condition, the variance of the solution tends to infinity as $t \to \infty$, in that

$$\mathbb{E}|\mathbf{z}(t)|^2 \to \infty \text{ as } t \to \infty. \tag{3.4.6}$$

where $\mathbb{E}$ denotes the expectation operator.

The condition that $\mathcal{S}$ contains a bi-infinite line is satisfied, for example, as soon as $\mathcal{S}$ is more than a single point if φ is a quadratic function, and commonly occurs in applications, e.g. in the multi-path routing example given in the following Chapter 4.

One of the main applications of the gradient method is to the dual formulation of concave optimization problems where some of the constraints are relaxed by Lagrange multipliers. When all the relaxed constraints are linear, φ has the form

$$\varphi(x, y) = U(x) + y^T(Dx + e) \tag{3.4.7}$$

where D is a constant matrix and e a constant vector. Under the assumption that U is analytic we obtain a simple exact characterisation of $\mathcal{S}$. One specific case of this was studied by the authors previously in [90], but without the analyticity condition.

Theorem 3.4.4. *Let φ be defined by (3.4.7) with U analytic and $D \in \mathbb{R}^{m\times n}$, $e \in \mathbb{R}^m$ constant. Assume that $(\bar{x}, \bar{y}) = \bar{\mathbf{z}}$ is a saddle point of φ. Then $\mathcal{S}$ is given by*

$$\mathcal{S} = \bar{\mathbf{z}} + \operatorname{span}\{(x, y) \in \mathcal{W} \times \mathbb{R}^m : (x, y) \text{ is an eigenvector of } \begin{bmatrix} 0 & D^T \\ -D & 0 \end{bmatrix}\}$$

$$\mathcal{W} = \{x \in \mathbb{R}^n : s \mapsto U(sx + \bar{x}) \text{ is linear for } s \in \mathbb{R}\}.$$

Furthermore $\mathcal{W}$ is an affine subspace.

3.4.1 The subgradient method on affine subspaces

We now extend the exact classification (Theorem 3.4.1) to the subgradient method on affine subspaces. As discussed above, in part II of this work we show that all

limiting behaviour of the subgradient method on any convex domain is described by the subgradient on affine subspaces. Let V be an affine subspace of $\mathbb{R}^{n+m}$ and let $\mathbf{\Pi} \in \mathbb{R}^{(n+m)^2}$ be the orthogonal projection matrix onto the orthogonal complement of the normal cone N_V. Then the subgradient method (3.3.2) on V is given by

$$\dot{\mathbf{z}} = \mathbf{\Pi}\mathbf{f}(\mathbf{z}) \tag{3.4.8}$$

where $\mathbf{f}(\mathbf{z}) = \begin{bmatrix}\varphi_x & -\varphi_y\end{bmatrix}^T$. We generalise Theorem 3.4.1 for this projected form of the gradient method. As with the statement of Theorem 3.4.1, we state the result for $\mathbf{0}$ being an equilibrium point; the general case may be obtained by a translation of coordinates.

Theorem 3.4.5. *Let $\mathbf{\Pi} \in \mathbb{R}^{(n+m)^2}$ be an orthogonal projection matrix, φ be C^2 and concave-convex on $\mathbb{R}^{n+m}$, and $\mathbf{0}$ be an equilibrium point of (3.4.8). Then the trajectories $\mathbf{z}(t)$ of (3.4.8) that lie a constant distance from any equilibrium point of (3.4.8) are exactly the solutions to the linear ODE:*

$$\dot{\mathbf{z}}(t) = \mathbf{\Pi}\mathbf{A}(\mathbf{0})\mathbf{\Pi}\mathbf{z}(t) \tag{3.4.9}$$

that satisfy, for all $t \in \mathbb{R}$ and $r \in [0,1]$, the condition

$$\mathbf{z}(t) \in \ker(\mathbf{\Pi}\mathbf{B}(r\mathbf{z}(t))\mathbf{\Pi}) \cap \ker(\mathbf{\Pi}(\mathbf{A}(r\mathbf{z}(t)) - \mathbf{A}(\mathbf{0}))\mathbf{\Pi}) \tag{3.4.10}$$

where $\mathbf{A}(\mathbf{z})$ and $\mathbf{B}(\mathbf{z})$ are defined by (3.4.1).

Remark 3.4.5. *As with Theorem 3.4.1, this result can be localised for when φ is not concave-convex on the whole of $\mathbb{R}^{n+m}$, (see Remark 3.4.4).*

3.5 Modification method

The main applications of the (sub)gradient method are to solving constrained optimisation problems. In these cases the gradient method is not appropriate, and instead the subgradient method must be employed to ensure the Lagrange multipliers remain non-negative. For this reason the majority of the examples will be given in the following Chapter 4 where the subgradient method is studied. However, as an example for illustration we describe a method of modifying the

concave-convex function φ so that the gradient method converges to a saddle point.

Such methods are used in network optimisation (see e.g. [10], [61]), where it is important to preserve the localised structure of the dynamics, which makes the use of higher order information difficult. Here we consider a method were auxiliary variables are used to give convergence. This was used in [90] to achieve convergence without any additional information transfer. Here we present a natural generalisation to any concave-convex function φ. In Section 3.5.1 we give an example of how this method can be applied in distributed optimisation problem to yield guaranteed convergence without requiring additional information transfer.

In the subsequent Chapter 4 we extend this method to the subgradient method and apply it to the explicit example of multi-path routing over a communication network.

We define the modified concave-convex function φ' as,

$$\begin{aligned} \varphi'(x', x, y) &= \varphi(x, y) + \psi(Mx - x') \\ \psi : \mathbb{R}^{n'} \to \mathbb{R}, M \in \mathbb{R}^{n' \times n} &\text{ is a constant matrix} \\ \psi \in C^2 &\text{ is strictly concave with } \max \psi(0) = 0 \end{aligned} \tag{3.5.1}$$

where x' is a set of n' auxiliary variables. It is easy to see that under this condition φ' is concave-convex in $((x', x), y)$. We will denote the versions of the sets defined below Corollary 3.4.2 for φ' with a prime, e.g. $\bar{\mathcal{S}}'$ is the set of saddles points of φ'.

There is an equivalence between saddle points of φ and φ'. If $(x, y) \in \bar{\mathcal{S}}$ then a simple computation shows that $(Mx, x, y) \in \bar{\mathcal{S}}'$. Conversely if $(x', x, y) \in \bar{\mathcal{S}}'$ then we must have $Mx = x'$ and $(x, y) \in \bar{\mathcal{S}}$. In this way searching for a saddle point of φ may be done by searching for a saddle point (x', x, y) of φ' and discarding the extra x' variables.

The condition we require on the matrix M is given in the statement of the result below, and the reason for this assumption is evident in the proof. We do remark however, that taking $n' = n$ and M as the $n \times n$ identity matrix will always satisfy

the given condition, and can also preserve the locality of the gradient method as described in Section 3.5.1 below.

We remark that by the duality of the gradient method between the parameters x and y we could have instead (or as well as) added auxiliary variables y' and obtained the same results.

We establish that under conditions on the matrix M, this modification method gives global convergence of the subgradient method (3.3.2) on any affine subspace, and thus also the gradient method (3.3.1). The proof of this proposition is provided in Section 3.7 in a slightly more general case.

Proposition 3.5.1. *Let φ be C^2 and concave-convex on $\mathbb{R}^{n+m}$. Let φ' satisfy (3.5.1) and $M \in \mathbb{R}^{n' \times n}$ be such that $\ker(M) \cap \ker(\varphi_{xx}(\bar{\mathbf{z}})) = \{0\}$ for some equilibrium point $\bar{\mathbf{z}}$ of the subgradient method on V. Then the gradient method (3.3.2) applied to φ' is globally convergent.*

3.5.1 Distributed optimisation problem

Consider the optimisation problem

$$\max_{x \in \mathbb{R}^n, Ax = Y} \sum_{l=1}^{L} U_l(x) \tag{3.5.2}$$

where U_l are (in general non-strictly) concave functions that each depend only upon some subset $I_l \subseteq \{1, \ldots, n\}$ of the components of $x \in \mathbb{R}^n$. Functions that are given by a sum of this kind arise in various kinds of distributed optimisation problems, for example in communication networks. The optimisation problem (3.5.2) has the associated Lagrangian

$$\varphi(x, y) = \sum_{i=l}^{n} U_l(x) + y^T(Y - Ax) \tag{3.5.3}$$

where $y \in \mathbb{R}^m$ are Lagrange multipliers. The resulting gradient dynamics (3.3.1) are given by

$$\begin{aligned} \dot{x}_i &= \sum_{\substack{l \in \{1,\dots,L\}, \\ i \in I_l}} \frac{\partial U_l}{\partial x_j} - \sum_{j=1}^{m} A_{ji} y_j, \\ \dot{y} &= Ax - Y. \end{aligned} \tag{3.5.4}$$

These dynamics are localised in the sense that to update x_i the only components of x that are needed are those x_j with $i, j \in I_l$ for some l. Similarly, the components of y may also be localised depending upon the structure of the matrix A.

Because the Lagrangian φ is not strictly concave-convex, the gradient dynamics (3.5.4) are not guaranteed to converge to a saddle point. For this reason we consider the modified function φ' defined by (3.5.1). This results in the dynamics

$$\begin{aligned} \dot{x}_i &= \sum_{\substack{l \in \{1,\dots,L\}, \\ i \in I_l}} \frac{\partial U_l}{\partial x_j} - \sum_{j=1}^{m} A_{ji} y_j + \sum_{k=1}^{n'} M_{ki} \frac{\partial \psi}{\partial u_k}(Mx - x'), \\ \dot{x}'_k &= \frac{\partial \psi}{\partial u_k}(Mx - x'), \\ \dot{y} &= Ax - Y. \end{aligned}$$

If the function ψ and matrix M are chosen appropriately these dynamics are still localised. For example, if we take $\psi(u) = -|u|^2$ and M the n by n identity matrix, then each component x'_k is associated with the corresponding component x_k and the pair (x_k, x'_k) require no additional information to be updated compared to the original dynamics (3.5.4).

3.6 Proofs of the main results

In this section we prove the main results of the chapter which are stated in Section 3.4.

3.6.1 Outline of the proofs

We first give a brief outline of the derivations of the results to improve the readability. Before we give this summary we need to define some additional notation.

Given $\bar{\mathbf{z}} \in \bar{\mathcal{S}}$, we denote the set of solutions to the gradient method (3.3.1) that are a constant distance from $\bar{\mathbf{z}}$, (but not necessarily other saddle points), as $\mathcal{S}_{\bar{\mathbf{z}}}$. It is later proved that $\mathcal{S}_{\bar{\mathbf{z}}} = \mathcal{S}$ but until then the distinction is important.

3.6.1.1 Gradient method

Subsections 3.6.2 and 3.6.3 provide the proofs of Theorems 3.4.1-3.4.4 and Theorem 3.4.5.

First in Section 3.6.2 we use the pathwise stability of the gradient method (Proposition 3.4.1) and geometric arguments to establish convexity properties of $\mathcal{S}$. Lemma 3.6.1 and Lemma 3.6.2 tell us that $\bar{\mathcal{S}}$ is convex and can only contain bi-infinite lines in degenerate cases. Lemma 3.6.3 gives an orthogonality condition between $\mathcal{S}$ and $\bar{\mathcal{S}}$ which roughly says that the larger $\bar{\mathcal{S}}$ is, the smaller $\mathcal{S}$ is. These allow us to prove the key result of the section, Lemma 3.6.5, which states that any convex combination of $\bar{\mathbf{z}} \in \bar{\mathcal{S}}$ and $\mathbf{z}(t) \in \mathcal{S}_{\bar{\mathbf{z}}}$ lies in $\mathcal{S}_{\bar{\mathbf{z}}}$.

In Section 3.6.3 we use the geometric results of Section 3.6.2 to prove Theorems 3.4.1-3.4.4. We split the Hessian matrix of φ into symmetric and skew-symmetric parts, which allows us to express the gradients φ_x, φ_y in terms of line integrals from an (arbitrary) saddle point $\mathbf{0}$. This line integral formulation together with Lemma 3.6.5 allow us to prove Theorem 3.4.1, from which Theorem 3.4.2 is then deduced.

To prove Theorem 3.4.3 we first prove a lemma Lemma 3.6.7 (analogous to Lemma 3.6.2) that tells us that $\mathcal{S}$ containing a bi-infinite line implies the presence of a quantity conserved by all solutions of the gradient dynamics (3.3.1). In the presence of noise, the variance of this quantity converges to infinity and allows us to prove Theorem 3.4.3.

To prove Theorem 3.4.4 we construct a quantity $V(\mathbf{z})$ that is conserved by solu-

tions in $\mathcal{S}$. In the case considered this has a natural interpretation in terms of the utility function $U(x)$ and the constraints $g(x)$.

Finally Theorem 3.4.5 is proved by modifying the above proof to take into account the addition of the projection matrix.

3.6.2 Geometry of $\bar{\mathcal{S}}$ and $\mathcal{S}$

In this section we will use the gradient method to derive geometric properties of convex-concave functions. We will start with some simple results which are then used as a basis to derive Lemma 3.6.5 the main result of this section. On the way we illustrate how the gradient method can be used to prove results (Lemma 3.6.1 and Lemma 3.6.2) on the geometry of concave-convex functions.

Lemma 3.6.1. *Let $\varphi \in C^2$ be concave-convex on $\mathbb{R}^{n+m}$, then $\bar{\mathcal{S}}$, the set of saddle point of φ, is closed and convex.*

Proof. Closure follows from continuity of the derivatives of φ. For convexity let $\bar{\mathbf{a}}, \bar{\mathbf{b}} \in \bar{\mathcal{S}}$ and $\mathbf{c}$ lie on the line between them. Consider the two closed balls about $\bar{\mathbf{a}}$ and $\bar{\mathbf{b}}$ that meet at the single point $\mathbf{c}$, as in Fig. 3.1. By Proposition 3.4.1, $\mathbf{c}$ is an equilibrium point as the motion of the gradient method starting from $\mathbf{c}$ is constrained to stay within both balls. It is hence a saddle point. □

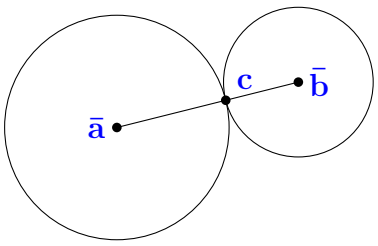


Figure 3.1: $\bar{\mathbf{a}}$ and $\bar{\mathbf{b}}$ are two saddle points of φ which is C^2 and concave-convex on $\mathbb{R}^{n+m}$. By Proposition 3.4.1 any solution of (3.3.1) starting from $\mathbf{c}$ is constrained for all positive times to lie in each of the balls about $\bar{\mathbf{a}}$ and $\bar{\mathbf{b}}$.

Lemma 3.6.2. *Let φ be C^2 and concave-convex on $\mathbb{R}^{n+m}$. Let the set of saddle points of φ contain the infinite line $L = \{\mathbf{a} + s\mathbf{b} : s \in \mathbb{R}\}$ for some $\mathbf{a}, \mathbf{b} \in \mathbb{R}^{n+m}$. Then φ is translation invariant in the direction of L, i.e. $\varphi(\mathbf{z}) = \varphi(\mathbf{z} + s\mathbf{b})$ for any $s \in \mathbb{R}$.*

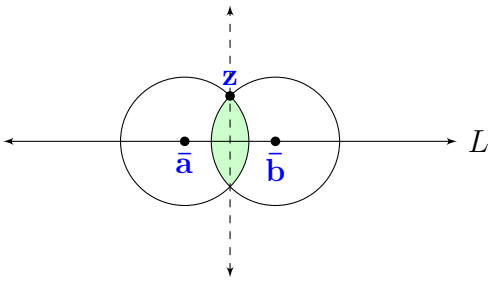


Figure 3.2: $\bar{\mathbf{a}}$ and $\bar{\mathbf{b}}$ are two saddle points of φ which is C^2 and concave-convex on $\mathbb{R}^{n+m}$. Solutions of (3.3.1) are constrained to lie in the shaded region for all positive time by Proposition 3.4.1.

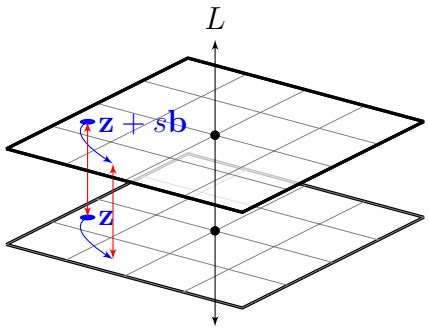


Figure 3.3: L is a line of saddle points of φ which is C^2 and concave-convex on $\mathbb{R}^{n+m}$. Solutions of (3.3.1) starting on hyperplanes normal to L are constrained to lie on these planes for all time. $\mathbf{z}$ lies on one normal hyperplane, and $\mathbf{z}+s\mathbf{b}$ lies on another. Considering the solutions of (3.3.1) starting from each we see that by Proposition 3.4.1 the distance between these two solutions must be constant and equal to $|s\mathbf{b}|$.

Proof. We do this in two steps. First we will prove that the motion of the gradient method is restricted to linear manifolds normal to L. Let $\mathbf{z}$ be a point and consider the motion of the gradient method starting from $\mathbf{z}$. As illustrated in Fig. 3.2 we pick two saddle points $\bar{\mathbf{a}}, \bar{\mathbf{b}}$ on L, then by Proposition 3.4.1 the motion starting from $\mathbf{z}$ is constrained to lie in the (shaded) region, which is the intersection of the two closed balls about $\bar{\mathbf{a}}$ and $\bar{\mathbf{b}}$ which have $\mathbf{z}$ on their boundaries. The intersection of the regions generated by taking a sequence of pairs of saddle points off to infinity is contained in the linear manifold normal to L.

Next we claim that for $s \in \mathbb{R}$ the motion starting from $\mathbf{z}+s\mathbf{b}$ is exactly the motion starting from $\mathbf{z}$ shifted by $s\mathbf{b}$. As illustrated in Fig. 3.3, by Proposition 3.4.1 the motion from $\mathbf{z} + s\mathbf{b}$ must stay a constant distance $s|\mathbf{b}|$ from the motion from $\mathbf{z}$.

This uniquely identifies the motion from $\mathbf{z} + s\mathbf{b}$ and proves the claim. Finally we deduce the full result by noting that the second claim implies that φ is defined up to an additive constant on each linear manifold as the motion of the gradient method contains all the information about the derivatives of φ. As φ is constant on L, the proof is complete. □

We now use these techniques to prove orthogonality results about solutions in $\mathcal{S}$.

Lemma 3.6.3. *Let $\varphi \in C^2$ be concave-convex on $\mathbb{R}^{n+m}$, and $\mathbf{z}$ be a trajectory in $\mathcal{S}$, then $\mathbf{z}(t) \in M_{\bar{\mathcal{S}}}(\mathbf{z}(0))$ for all $t \in \mathbb{R}$.*

Proof. If $\bar{\mathcal{S}} = \{\bar{\mathbf{z}}\}$ or $\emptyset$ the claim is trivial. Otherwise we let $\bar{\mathbf{a}} \neq \bar{\mathbf{b}} \in \bar{\mathcal{S}}$ be arbitrary, and consider the spheres about $\bar{\mathbf{a}}$ and $\bar{\mathbf{b}}$ that touch $\mathbf{z}(t)$. By Proposition 3.4.1, $\mathbf{z}(t)$ is constrained to lie on the intersection of these two spheres which lies inside $M_L(\mathbf{z}(0))$ where L is the line segment between $\bar{\mathbf{a}}$ and $\bar{\mathbf{b}}$. As $\bar{\mathbf{a}}$ and $\bar{\mathbf{b}}$ were arbitrary this proves the lemma. □

Lemma 3.6.4. *Let φ be C^2 and concave-convex on $\mathbb{R}^{n+m}$, $\bar{\mathbf{z}} \in \bar{\mathcal{S}}$ and $\mathbf{z}(t) \in \mathcal{S}_{\bar{\mathbf{z}}}$ lie in $M_{\bar{\mathcal{S}}}(\mathbf{z}(0))$ for all t. Then $\mathbf{z}(t) \in \mathcal{S}$.*

Proof. If $\bar{\mathcal{S}} = \{\bar{\mathbf{z}}\}$ the claim is trivial. Let $\bar{\mathbf{a}} \in \bar{\mathcal{S}} \setminus \{\bar{\mathbf{z}}\}$ be arbitrary. Then by Lemma 3.6.1 the line segment L between $\bar{\mathbf{a}}$ and $\bar{\mathbf{z}}$ lies in $\bar{\mathcal{S}}$. Let $\mathbf{b}$ be the intersection of the extension of L to infinity in both directions and $M_{\bar{\mathcal{S}}}(\mathbf{z}(0))$. Then the definition of $M_{\bar{\mathcal{S}}}(\mathbf{z}(0))$ tells us that the extension of L meets $M_{\bar{\mathcal{S}}}(\mathbf{z}(0))$ at a right angle. $d(\mathbf{b}, \bar{\mathbf{z}})$ is constant and $d(\mathbf{z}(t), \bar{\mathbf{z}})$ as $\mathbf{z}(t) \in \mathcal{S}$, which implies that $d(\mathbf{z}(t), \bar{\mathbf{a}})$ is also constant (as illustrated in Fig. 3.4). Indeed, we have

$$\begin{aligned} d(\mathbf{z}(t), \bar{\mathbf{a}})^2 &= d(\mathbf{z}(t), \mathbf{b})^2 + d(\mathbf{b}, \bar{\mathbf{a}})^2 \\ &= d(\mathbf{z}(t), \bar{\mathbf{z}})^2 - d(\mathbf{b}, \bar{\mathbf{z}})^2 + d(\mathbf{b}, \bar{\mathbf{a}})^2 \end{aligned} \tag{3.6.1}$$

and all the terms on the right hand side are constant. □

Using these orthogonality results we prove the key result of the section, a convexity result between $\mathcal{S}_{\bar{\mathbf{z}}}$ and $\bar{\mathbf{z}}$.

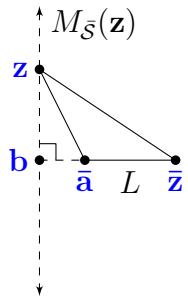


Figure 3.4: $\bar{\mathbf{a}}$ and $\bar{\mathbf{z}}$ are saddle points of φ which is C^2 and concave-convex on $\mathbb{R}^{n+m}$, and L is the line segment between them. $\mathbf{z}$ is a point on a solution in $\mathcal{S}_{\bar{\mathbf{z}}}$ which lies on $M_{\bar{\mathcal{S}}}(\mathbf{z})$ which is orthogonal to L by definition. $\mathbf{b}$ is the point of intersection between $M_{\bar{\mathcal{S}}}(\mathbf{z})$ and the extension of L.

Lemma 3.6.5. *Let φ be C^2 and concave-convex on $\mathbb{R}^{n+m}$, $\bar{\mathbf{z}} \in \bar{\mathcal{S}}$ and $\mathbf{z}(t) \in \mathcal{S}_{\bar{\mathbf{z}}}$. Then for any $s \in [0,1]$, the convex combination $\mathbf{z}'(t) = (1-s)\bar{\mathbf{z}} + s\mathbf{z}(t)$ lies in $\mathcal{S}_{\bar{\mathbf{z}}}$. If in addition $\mathbf{z} \in \mathcal{S}$, then $\mathbf{z}'(t) \in \mathcal{S}$.*

Proof. Clearly $\mathbf{z}'$ is a constant distance from $\bar{\mathbf{z}}$. We must show that $\mathbf{z}'(t)$ is also a solution to (3.3.1). We argue in a similar way to Fig. 3.3 but with spheres instead of planes. Let the solution to (3.3.1) starting at $\mathbf{z}'(0)$ be denoted $\mathbf{z}''(t)$. We must show this is equal to $\mathbf{z}'(t)$. As $\mathbf{z}(t) \in \mathcal{S}$ it lies on a sphere about $\bar{\mathbf{z}}$, say of radius r, and by construction $\mathbf{z}'(0)$ lies on a smaller sphere about $\bar{\mathbf{z}}$ of radius rs. By Proposition 3.4.1, $d(\mathbf{z}(t), \mathbf{z}''(t))$ and $d(\mathbf{z}''(t), \bar{\mathbf{z}})$ are non-increasing, so that $\mathbf{z}''(t)$ must be within rs of $\bar{\mathbf{z}}$ and within $r(1-s)$ of $\mathbf{z}(t)$. The only such point is $\mathbf{z}'(t) = (1-s)\bar{\mathbf{z}} + s\mathbf{z}(t)$ which proves the claim. For the additional statement, we consider another saddle point $\bar{\mathbf{a}} \in \bar{\mathcal{S}}$ and let L be the line segment connecting $\bar{\mathbf{a}}$ and $\bar{\mathbf{z}}$. By Lemma 3.6.3, $\mathbf{z}(t)$ lies in $M_{\bar{\mathcal{S}}}(\mathbf{z}(0))$, so by construction, $\mathbf{z}'(t) \in M_{\bar{\mathcal{S}}}(\mathbf{z}'(0))$, (as illustrated by Fig. 3.5). Hence, by Lemma 3.6.4, $\mathbf{z}'(t) \in \mathcal{S}$. □

Proposition 3.6.1 is a further convexity result that can be proved by means of similar methods, using Lemma 3.6.6 which is proved in the next section using analytic techniques.

Lemma 3.6.6. *Let φ be C^2 and concave-convex on $\mathbb{R}^{n+m}$. Let $\mathbf{z}(t), \mathbf{z}'(t) \in \mathcal{S}$. Then $d(\mathbf{z}(t), \mathbf{z}'(t))$ is constant.*

Proposition 3.6.1. *Let φ be C^2 and concave-convex on $\mathbb{R}^{n+m}$, then $\mathcal{S}$ is convex.*

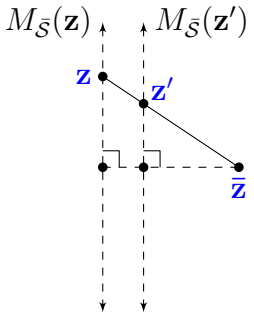


Figure 3.5: $\bar{\mathbf{z}}$ is a saddle point of φ which is C^2 and concave-convex on $\mathbb{R}^{n+m}$. $\mathbf{z}$ is a point on a solution in $\mathcal{S}$ and $\mathbf{z}'$ is a convex combination of $\mathbf{z}$ and $\bar{\mathbf{z}}$. $M_{\bar{\mathcal{S}}}(\mathbf{z})$ and $M_{\bar{\mathcal{S}}}(\mathbf{z}')$ are parallel to each other by definition.

Proof. The proof is very similar to that of Lemma 3.6.5. Let $\mathbf{z}(t), \mathbf{z}'(t) \in \mathcal{S}$, and $s \in (0,1)$. Set $\mathbf{w}(t) = s\mathbf{z}(t) + (1-s)\mathbf{z}'(t)$. By Lemma 3.6.6 we know that $d = d(\mathbf{z}(t), \mathbf{z}'(t))$ is constant. Denote the solution of the gradient method starting from $\mathbf{w}(0)$ as $\mathbf{w}'(t)$. We must prove that $\mathbf{w}'(t) = \mathbf{w}(t)$ and that $\mathbf{w}(t) \in \mathcal{S}$. First we imagine two closed balls centered on $\mathbf{z}(t)$ and $\mathbf{z}'(t)$ and of radii sd and $(1-s)d$ respectively. By Proposition 3.4.1, $\mathbf{w}'(t)$ is constrained to lie within both of these balls. For each t there is only one such point and it is exactly $\mathbf{w}(t)$. Next we let $\bar{\mathbf{a}} \in \bar{\mathcal{S}}$ be arbitrary, then $d(\bar{\mathbf{a}}, \mathbf{w}(t))$ is determined by $d(\mathbf{z}(t), \mathbf{z}'(t)), d(\bar{\mathbf{a}}, \mathbf{z})$ and $d(\bar{\mathbf{a}}, \mathbf{z}'(t))$, (as illustrated by Fig. 3.6). Indeed, we may assume by translation that $\bar{\mathbf{a}} = \mathbf{0}$, and then

$$\begin{aligned} d(\bar{\mathbf{a}}, \mathbf{w}(t))^2 &= d(\mathbf{0}, \mathbf{z}(t) + (1-s)\mathbf{z}'(t))^2 \\ &= s^2 d(\mathbf{0}, \mathbf{z}(t))^2 + (1-s)^2 d(\mathbf{0}, \mathbf{z}'(t))^2 - 2s(1-s)\mathbf{z}^T(t)\mathbf{z}'(t) \end{aligned} \tag{3.6.2}$$

The first two terms in (3.6.2) are constant by Lemma 3.6.6 and the third can be computed as

$$2\mathbf{z}^T(t)\mathbf{z}'(t) = d(\mathbf{z}(t), \mathbf{z}'(t))^2 - d(\mathbf{0}, \mathbf{z}(t))^2 - d(\mathbf{0}, \mathbf{z}'(t))^2 \tag{3.6.3}$$

which is constant for the same reason. □

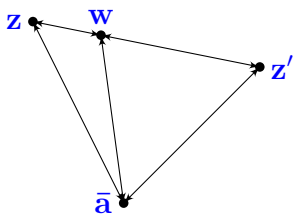


Figure 3.6: $\mathbf{z}$ and $\mathbf{z}'$ are two elements of $\mathcal{S}$ and $\mathbf{w}$ is a convex combination of them. $\bar{\mathbf{a}}$ is a saddle point in $\bar{\mathcal{S}}$. We know all the distances are constant except possibly $d(\mathbf{w}, \bar{\mathbf{a}})$, but this is uniquely determined by the other four distances.

3.6.3 Classification of $\mathcal{S}$

We will now proceed with a full classification of $\mathcal{S}$ and prove Theorems 3.4.1-3.4.4. For notational convenience we will make the assumption (without loss of generality) that $\mathbf{0} \in \bar{\mathcal{S}}$. Then we compute $\varphi_x(\mathbf{z}), \varphi_y(\mathbf{z})$ from line integrals from $\mathbf{0}$ to $\mathbf{z}$. Indeed, letting $\hat{\mathbf{z}}$ be a unit vector parallel to $\mathbf{z}$, we have

$$\begin{bmatrix} \varphi_x(\mathbf{z}) \\ -\varphi_y(\mathbf{z}) \end{bmatrix} = \left(\int_0^{|\mathbf{z}|} \begin{bmatrix} \varphi_{xx}(s\hat{\mathbf{z}}) & \varphi_{xy}(s\hat{\mathbf{z}}) \\ -\varphi_{yx}(s\hat{\mathbf{z}}) & -\varphi_{yy}(s\hat{\mathbf{z}}) \end{bmatrix} ds \right) \hat{\mathbf{z}}. \tag{3.6.4}$$

Together with the definition of the matrices $\mathbf{A}(\mathbf{z})$ and $\mathbf{B}(\mathbf{z})$ given by (3.4.1) we obtain

$$\begin{bmatrix} \varphi_x(\mathbf{z}) \\ -\varphi_y(\mathbf{z}) \end{bmatrix} = \int_0^{|\mathbf{z}|} (\mathbf{A}(s\hat{\mathbf{z}}) + \mathbf{B}(s\hat{\mathbf{z}}))\hat{\mathbf{z}}\, ds. \tag{3.6.5}$$

We are now ready to prove the first main result.

Proof of Theorem 3.4.1. Define the set $\mathcal{X}$ as solutions of the ODE (3.4.2) which obey the condition (3.4.3) for all $t \in \mathbb{R}$ and $r \in [0, 1]$. Then Theorem 3.4.1 is the statement that $\mathcal{X} = \mathcal{S}$. For brevity we define the matrix $\mathbf{B}'(\mathbf{z})$ by

$$\mathbf{B}'(\mathbf{z}) = \mathbf{B}(\mathbf{z}) + (\mathbf{A}(\mathbf{z}) - \mathbf{A}(\mathbf{0})). \tag{3.6.6}$$

As $\mathbf{A}(\mathbf{z})$ is skew symmetric and $\mathbf{B}(\mathbf{z})$ is symmetric we have

$$\ker(\mathbf{B}'(\mathbf{z})) = \ker(\mathbf{B}(\mathbf{z})) \cap \ker(\mathbf{A}(\mathbf{z}) - \mathbf{A}(\mathbf{0})),$$

so that condition (3.4.3) is equivalent to

$$\mathbf{z}(t) \in \ker(\mathbf{B}'(r\mathbf{z}(t))) \text{ for all } t \in \mathbb{R}, r \in [0,1]. \tag{3.6.7}$$

We will prove that $\mathcal{X} \subseteq \mathcal{S}_\mathbf{0}$, $\mathcal{X} \subseteq \mathcal{S}$ and $\mathcal{S}_\mathbf{0} \subseteq \mathcal{X}$. As the other inclusion $\mathcal{S} \subseteq \mathcal{S}_\mathbf{0}$ is clear this will prove the theorem.

Step 1: $\mathcal{X} \subseteq \mathcal{S}_\mathbf{0}$. For any non-zero point $\mathbf{z}$ we can compute the partial derivatives of φ at $\mathbf{z}$ using the line integral formula (3.6.5) and (3.6.6),

$$\begin{bmatrix} \varphi_x(\mathbf{z}) \\ -\varphi_y(\mathbf{z}) \end{bmatrix} = \mathbf{A}(\mathbf{0})\mathbf{z} + \int_0^{|\mathbf{z}|} \mathbf{B}'(s\hat{\mathbf{z}})\hat{\mathbf{z}} ds \tag{3.6.8}$$

where $\mathbf{z} = |\mathbf{z}|\hat{\mathbf{z}}$. If $\mathbf{z}(t) \in \mathcal{X}$, then $\dot{\mathbf{z}}(t) = \mathbf{A}(\mathbf{0})\mathbf{z}(t)$, and by skew-symmetry of $\mathbf{A}(\mathbf{0})$, $|\mathbf{z}(t)|$ is constant, which means that $\mathbf{z}(t)$ is a constant distance from $\mathbf{0}$. Furthermore, the assumption that $\mathbf{z}(t) \in \ker(\mathbf{B}'(r\mathbf{z}(t)))$ for $r \in [0,1]$ implies that the integrand in (3.6.8) vanishes, and $\mathbf{z}(t)$ is a solution of the gradient method.

Step 2: $\mathcal{X} \subseteq \mathcal{S}$. Let $\bar{\mathbf{z}}$ be arbitrary. Consider the function $t \mapsto d(\mathbf{z}(t), \bar{\mathbf{z}})^2$. By expanding in the orthonormal basis of eigenvectors of $\mathbf{A}(\mathbf{0})$ we observe that this function is a linear combination of continuous periodic functions. As, by Proposition 3.4.1, this function is also non-increasing, it must be constant.

Step 3: $\mathcal{S}_\mathbf{0} \subseteq \mathcal{X}$. Let $\mathbf{z}(t) \in \mathcal{S}_\mathbf{0}$ and $R = |\mathbf{z}(t)|$ which is constant. For $r \in [0, R]$, define $\mathbf{z}(t; r) = (r/R)\mathbf{z}(t)$, so that $\mathbf{z}(t; 0) = \mathbf{0}$ and $\mathbf{z}(t; R) = \mathbf{z}(t)$. Note that the corresponding unit vector $\hat{\mathbf{z}}(t; r) = \hat{\mathbf{z}}(t)$ does not depend on r. The convexity result Lemma 3.6.5 implies that $\mathbf{z}(t; r) \in \mathcal{S}_\mathbf{0}$, and is a solution of the gradient method. We shall compute the time derivative of this in two ways. First, we use (3.3.1) and (3.6.8) to obtain,

$$\dot{\mathbf{z}}(t; r) = \mathbf{A}(\mathbf{0})\mathbf{z}(t; r) + \int_0^r \mathbf{B}'(s\hat{\mathbf{z}}(t))\hat{\mathbf{z}}(t)\, ds. \tag{3.6.9}$$

Second, we use the explicit definition of $\mathbf{z}(t; r)$ in terms of $\mathbf{z}(t)$ to obtain,

$$\dot{\mathbf{z}}(t; r) = \frac{r}{R}\mathbf{A}(\mathbf{0})\mathbf{z}(t) + \frac{r}{R}\int_0^R \mathbf{B}'(s\hat{\mathbf{z}}(t))\hat{\mathbf{z}}(t)\, ds. \tag{3.6.10}$$

Equating (3.6.9) and (3.6.10) we deduce that

$$\int_0^r \mathbf{B}'(s\hat{\mathbf{z}}(t))\hat{\mathbf{z}}(t)\,ds = \frac{r}{R}\int_0^R \mathbf{B}'(s\hat{\mathbf{z}}(t))\hat{\mathbf{z}}(t)\,ds. \tag{3.6.11}$$

Differentiating with respect to r we have,

$$\mathbf{B}'(r\hat{\mathbf{z}}(t))\hat{\mathbf{z}}(t) = \frac{1}{R}\int_0^R \mathbf{B}'(s\hat{\mathbf{z}}(t))\hat{\mathbf{z}}(t)\,ds. \tag{3.6.12}$$

The right hand side of this is independent of r, which implies that the left hand side is also independent of r, and is thus equal to its value at $r = 0$, so that

$$\mathbf{B}'(r\hat{\mathbf{z}}(t))\hat{\mathbf{z}}(t) = \mathbf{B}'(\mathbf{0})\hat{\mathbf{z}}(t) = \mathbf{B}(\mathbf{0})\hat{\mathbf{z}}(t). \tag{3.6.13}$$

Putting this back into our expression for $\dot{\mathbf{z}}$ we find that

$$\dot{\mathbf{z}}(t) = \mathbf{A}(\mathbf{0})\mathbf{z}(t) + \mathbf{B}(\mathbf{0})\mathbf{z}(t), \tag{3.6.14}$$

but as $|\mathbf{z}(t)|$ is constant, $\mathbf{A}(\mathbf{0})$ skew symmetric, and $\mathbf{B}(\mathbf{0})$ symmetric, $\mathbf{B}(\mathbf{0})\mathbf{z}(t)$ must vanish, which, together with (3.6.13) shows that $\mathbf{z}(t) \in \mathcal{X}$. □

Corollary 3.6.1. *Let φ be C^2 and concave-convex on $\mathbb{R}^{n+m}$ and there be a saddle point $\bar{\mathbf{z}}$ which is locally asymptotically stable. Then $\mathcal{S} = \bar{\mathcal{S}} = \{\bar{\mathbf{z}}\}$.*

Proof. By local asymptotic stability of $\bar{\mathbf{z}}$, $\mathcal{S} \cap \mathcal{B} = \{\bar{\mathbf{z}}\}$ for some open ball $\mathcal{B}$ about $\bar{\mathbf{z}}$. Then by Proposition 3.6.1, $\mathcal{S}$ is convex, and we deduce that $\mathcal{S} = \{\bar{\mathbf{z}}\}$. □

The proof of Lemma 3.6.6 is now very simple.

Proof of Lemma 3.6.6. Using Theorem 3.4.1 we have that

$$\mathbf{z}(t) - \mathbf{z}'(t) = e^{t\mathbf{A}(\mathbf{0})}(\mathbf{z}(0) - \mathbf{z}'(0))$$

which has constant magnitude as $\mathbf{A}(\mathbf{0})$ is skew symmetric. □

To prove Theorem 3.4.3 we require the following lemma which shows the existence of a conserved quantity of the gradient dynamics.

Lemma 3.6.7. *Let φ be C^2 and concave-convex on $\mathbb{R}^{n+m}$. Suppose that $\mathcal{S}$ contains a bi-infinite line $L = \{\mathbf{a} + s\mathbf{v} : s \in \mathbb{R}\}$. Assume that $\mathbf{0} \in \bar{\mathcal{S}}$. Then $W(t;\mathbf{z}) = |(e^{t\mathbf{A}(\mathbf{0})}\mathbf{v})^T\mathbf{z}|^2$ is a conserved quantity for any solution $\mathbf{z}$ of (3.3.1).*

Proof. As $\mathcal{S}$ is closed and convex (Proposition 3.6.1) we may assume that the line passes though the origin and take $\mathbf{a} = \mathbf{0}$. Let $\mathbf{v}(t) = e^{t\mathbf{A}(\mathbf{0})}\mathbf{v}$ and note that $\lambda\mathbf{v}(t)$ is a solution to the gradient method (3.3.1) by Theorem 3.4.1 for any $\lambda \in \mathbb{R}$. We follow the strategy of the first part of the proof of Lemma 3.6.2 with $-\lambda\mathbf{v}(t), \lambda\mathbf{v}(t)$ replacing the saddle points $\bar{\mathbf{a}}$,$\bar{\mathbf{b}}$. Indeed, let $\mathbf{z}(t)$ be any solution to (3.3.1) and let $\lambda' = \mathbf{v}^T\mathbf{z}(0)$. Then for any $t \geq 0$, Proposition 3.4.1 implies that $\mathbf{z}(t)$ must satisfy

$$d(\pm\lambda\mathbf{v}(t), \mathbf{z}(t)) \leq d(\pm\lambda\mathbf{v}(0), \mathbf{z}(0)), \tag{3.6.15}$$

where by $\pm$ we mean that the equation holds for each of $+$ and $-$. In the same way as in the proof of Lemma 3.6.2, taking the intersection of these balls for a sequence $\lambda \to \infty$ we deduce that $\mathbf{z}(t)$ is contained in the linear manifold normal to the line through the origin and $\mathbf{v}(t)$, and passing through $\lambda'\mathbf{v}(t)$. Indeed, by squaring (3.6.15) and expanding we obtain

$$|\mathbf{z}(t)|^2 \mp 2\lambda\mathbf{v}(t)^T\mathbf{z}(t) \leq |\mathbf{z}(0)|^2 \mp 2\lambda\mathbf{v}(0)^T\mathbf{z}(0).$$

By dividing through by λ and taking the limit $\lambda \to \infty$ we deduce that $\mathbf{v}(t)^T\mathbf{z}(t)$ is equal to $\mathbf{v}(0)^T\mathbf{z}(0)$ which implies that $W(t;\mathbf{z})$ is conserved. □

Proof of Theorem 3.4.3. Consider the conserved quantity $W(t;\mathbf{z})$ that was given by Lemma 3.6.7. Applying Itō's lemma and taking expectations, we have

$$\frac{d}{dt}\mathbb{E}W(t;\mathbf{z}(t)) = \mathbb{E}\dot{W}(t;\mathbf{z}(t)) + \tfrac{1}{2}\mathbb{E}\,\mathrm{Tr}(\mathbf{\Sigma}^T W_{\mathbf{zz}}\mathbf{\Sigma})$$

where $\mathbf{\Sigma} = \mathrm{diag}(\Sigma^x, \Sigma^y)$, $\dot{W}$ is the total derivative along the deterministic flow (3.3.1) and Tr is the trace operator. As W is conserved along the deterministic flow, $\dot{W} = 0$ and a simple computation shows that the second term is independent of $\mathbf{z}$ and bounded below by a strictly positive constant. Therefore $\mathbb{E}W(t;\mathbf{z}(t))$ grows at least linearly in time. It remains to note that $W(t;\mathbf{z}) \leq |e^{t\mathbf{A}(\mathbf{0})}\mathbf{v}|^2|\mathbf{z}|^2 \leq |\mathbf{v}|^2|\mathbf{z}|^2$, so that $|\mathbf{z}(t)| \geq cW(t;\mathbf{z}(t))$ for a constant $c > 0$. This implies that also $\mathbb{E}|\mathbf{z}(t)|^2 \to \infty$ and completes the proof of the proposition. □

The convexity of $\mathcal{S}$ allow us to deduce that the average position of any limiting trajectory is a saddle point.

Corollary 3.6.2. *Let φ be C^2 and concave-convex on $\mathbb{R}^{n+m}$ and $\mathbf{z}(t) \in \mathcal{S}$, then the average position of $\mathbf{z}(t)$ defined by*

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \mathbf{z}(t)\, dt \tag{3.6.16}$$

exists and lies inside $\bar{\mathcal{S}}$.

Proof. That the limit exists follows from expanding $\mathbf{z}(t) = e^{t\mathbf{A}(\mathbf{0})}\mathbf{z}(0)$ into eigenmodes and noting that, as $\mathbf{A}(\mathbf{0})$ is skew symmetric, each individual limit exists.

To prove that the limit is in $\bar{\mathcal{S}}$ we consider, for $T > 0$, the function $\mathbf{z}(t;T) = \frac{1}{T}\int_0^T \mathbf{z}(t+s)\, ds$. This is a linear combination of the functions $t \mapsto \mathbf{z}(t+s)$ which are in $\mathcal{S}$, so by the convexity result Lemma 3.6.6, it also lies in $\mathcal{S}$. As $T \to \infty$ this tends to a constant independent of t, which by closure of $\mathcal{S}$ also lies in $\mathcal{S}$. But it is a constant, so it is also in $\bar{\mathcal{S}}$. □

To prove Theorem 3.4.2 and Theorem 3.4.4 we require the lemma below (this can be proved in the same way as a similar result in [90]).

Lemma 3.6.8. *Let X be a linear subspace of $\mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$ a normal matrix. Let*

$$Y = \operatorname{span}\{v \in X : v \text{ is an eigenvector of } A\}. \tag{3.6.17}$$

Then Y is the largest subset of X that is invariant under A.

We note that invariance of a subspace under A is equivalent to invariance of the subspace under the group e^{tA}.

Proof of Theorem 3.4.2. **Step 1: $\mathcal{S}_{\text{linear}} \subseteq \mathcal{S}$ when φ is a quadratic function.** We will use the characterisation of $\mathcal{S}$ given by Theorem 3.4.1. By Lemma 3.6.8, $\mathcal{S}_{\text{linear}}$ is invariant under $e^{t\mathbf{A}(\mathbf{0})}$, so that $\mathbf{z}(0) \in \mathcal{S}_{\text{linear}} \implies \mathbf{z}(t) = e^{t\mathbf{A}(\mathbf{0})}\mathbf{z}(0) \in \mathcal{S}_{\text{linear}}$. Hence if $\mathbf{z}(0) \in \mathcal{S}_{\text{linear}}$ then $\mathbf{z}(t) \in \ker(\mathbf{B}'(\mathbf{0}))$ for all time t, and as φ is a quadratic function, $\mathbf{B}'(\mathbf{z})$ is constant, so this is enough to show $\mathcal{S}_{\text{linear}} \subseteq \mathcal{S}$.

Step 2: $\mathcal{S} \subseteq \mathcal{S}_{\text{linear}}$. Let $\mathbf{z}(t) \in \mathcal{S}$, then by Theorem 3.4.1 taking $r = 0$ we have

$\mathbf{z}(t) = e^{t\mathbf{A}(\mathbf{0})} \in \ker(\mathbf{B}'(\mathbf{0}))$ for all $t \in \mathbb{R}$. Thus $\mathcal{S}$ lies inside the largest subset of $\ker(\mathbf{B}'(\mathbf{0}))$ that is invariant under the action of the group $e^{t\mathbf{A}(\mathbf{0})}$, which by Lemma 3.6.8 is exactly $\mathcal{S}_{\text{linear}}$. □

In order to prove Theorem 3.4.4 we give a different interpretation of the condition in Theorem 3.4.1. The condition $\mathbf{z} \in \ker(\mathbf{B}(s\mathbf{z}))$ for all $s \in [0,1]$ looks like a line integral condition. Indeed, if we define a function $V(\mathbf{z})$ by

$$V(\mathbf{z}) = \mathbf{z}^T \left(\int_0^1 \int_0^1 \mathbf{B}(ss'\mathbf{z}) s \, ds' \, ds \right) \mathbf{z} \tag{3.6.18}$$

then as $\mathbf{B}(\mathbf{z})$ is symmetric negative semi-definite we have that $V(\mathbf{z}) = 0$ if and only if $\mathbf{z} \in \ker(\mathbf{B}(s\mathbf{z}))$ for every $s \in [0,1]$. This still leaves the condition $\mathbf{z} \in \ker(\mathbf{A}(s\mathbf{z}) - \mathbf{A}(\mathbf{0}))$ for all $s \in [0,1]$, and the function V has no natural interpretation in general. However in the specific case where φ is the Lagrangian of a concave optimization problem where the relaxed constraints are linear, we do have an interpretation. In this case the assumption that $\mathbf{0}$ is a saddle point is no longer generic and we must translate coordinates explicitly. Let the Lagrangian of the optimization problem be given by

$$\begin{gathered} \varphi(x', y') = U'(x') + {y'}^T g'(x') \\ U' \in C^2 \text{ and concave}, \quad g' \text{ linear with } g'_x = D. \end{gathered} \tag{3.6.19}$$

We pick a saddle point $(\bar{x}', \bar{y}')$, and shift to new coordinates $(x, y) = (x' - \bar{x}', y' - \bar{y}')$ so that $(0,0)$ is a saddle point in the new coordinates. After expanding we obtain

$$\varphi(x, y) = (U'(x + \bar{x}') + \bar{y}'^T g'(x + \bar{x}')) + y^T g'(x + \bar{x}') \tag{3.6.20}$$

which is a Lagrangian originating from the utility function

$$U(x) = U'(x + \bar{x}') + \bar{y}'^T g'(x + \bar{x}') \tag{3.6.21}$$

and constraints $g(x) = g'(x + \bar{x}')$. Without loss of generality we assume that $U(0) = 0$. As $g(x)$ is a linear function we have

$$\mathbf{B}(\mathbf{z}) = \begin{bmatrix} U_{xx}(x) & 0 \\ 0 & 0 \end{bmatrix} \tag{3.6.22}$$

so that $V(\mathbf{z})$ is independent of y, and in fact by direct computation we have $V(\mathbf{z}) = U(x)$. This leads us to the following lemma.

Lemma 3.6.9. *Let (3.6.19) hold. Then $\mathcal{S}$ is the largest subset of $U^{-1}(\{0\})\times\mathbb{R}^m = \{(x,y) \in \mathbb{R}^{n+m} : U(x) = 0\}$ that is invariant under evolution by the group $e^{t\mathbf{A}(\mathbf{0})}$, where U is given by (3.6.21).*

Proof. Denote the set defined in the lemma as $\mathcal{Y}$.

Step 1: $\mathcal{S} \subseteq \mathcal{Y}$. By the computation above we know that $\mathbf{z} \in U^{-1}(\{0\}) \times \mathbb{R}^{n+m}$ if and only if $\mathbf{z} \in \ker(\mathbf{B}(s\mathbf{z}))$ for all $s \in [0,1]$. Thus by Theorem 3.4.1, we have $\mathcal{S} \subseteq U^{-1}(\{0\}) \times \mathbb{R}^m$ and as $\mathcal{S}$ is invariant under the action of $e^{t\mathbf{A}(\mathbf{0})}$.

Step 2: $\mathcal{Y} \subseteq \mathcal{S}$. If $\mathbf{z}(0)$ is in the largest subset of $U^{-1}(\{0\}) \times \mathbb{R}^m$ invariant under the action of $e^{t\mathbf{A}(\mathbf{0})}$, then $\mathbf{z}(t)$ is in this set for all $t \in \mathbb{R}$. Defining $\mathbf{z}(t) = e^{t\mathbf{A}(\mathbf{0})}\mathbf{z}(0)$, we have $\mathbf{z}(t) \in \ker(\mathbf{B}(s\mathbf{z}(t)))$ for all $s \in [0,1]$, so $\mathbf{z}(t) \in \mathcal{S}$ by Theorem 3.4.1. □

To obtain a more exact expression for $\mathcal{S}$, we make use of the assumption that U is analytic.

Lemma 3.6.10. *Let (3.6.19) hold and in addition U given by (3.6.21) be analytic. Then the following hold:*

(i) $U^{-1}(\{0\}) = \operatorname{span}(U^{-1}(\{0\}))$.

(ii) $\mathcal{S} = \{e^{t\mathbf{A}(\mathbf{0})}\mathbf{z}(0) : \mathbf{z}(0) \in \mathcal{Q}\}$ *where*

$$\mathcal{Q} = \operatorname{span}\left\{(x,y) \in U^{-1}(\{0\})\times\mathbb{R}^m : (x,y) \text{ is an eigenvector of } \begin{bmatrix} 0 & D^T \\ -D & 0 \end{bmatrix}\right\}$$

Proof. We begin with (i). Recall we have assumed without loss of generality that $U(0) = 0$. As $U^{-1}(\{0\})$ is the set of maxima of a concave function, it is convex. If $U^{-1}(\{0\})$ is the single point 0, then (i) is trivial. Otherwise let L be a line segment (of strictly positive length) in $U^{-1}(\{0\})$, and let $\hat{L}$ be the bi-infinite extension of L. Let f be a linear bijection from $\mathbb{R}$ to $\hat{L}$, and let $I \subset \mathbb{R}$ be the interval in $\mathbb{R}$ given by $f^{-1}(L)$. Then $U(f(t)) : \mathbb{R} \to \mathbb{R}$ is an analytic function whose restriction to I vanishes. Hence $U(f(t))$ vanishes everywhere on $\mathbb{R}$, which is equivalent to U vanishing on $\hat{L}$. By varying the choice of L, we deduce that $U^{-1}(\{0\})$ contains

infinite lines in every direction in $\operatorname{span}(U^{-1}(\{0\}))$ and by convexity is equal to $\operatorname{span}(U^{-1}(\{0\}))$.

(ii) is a consequence of Lemma 3.6.9 and Lemma 3.6.8. □

Lastly, we translate back into the original coordinates.

Lemma 3.6.11. *Let (3.6.19) hold and U' be analytic, then*

$$U^{-1}(\{0\}) = \{x \in \mathbb{R}^n : \mathbb{R} \ni s \mapsto U'(sx + \bar{x}') \text{ is linear}\}$$

where U is given by (3.6.21).

Proof. Suppose that $x \in U^{-1}(\{0\})$ then by Lemma 3.6.10 $U(sx) = 0$ for all $s \in \mathbb{R}$. Recall that $U - U'$ is a linear function. Hence $U'(sx + \bar{x}')$ is linear as a function of $s \in \mathbb{R}$. Now suppose that $U'(sx + \bar{x}')$ is linear as a function of $s \in \mathbb{R}$ for some $x \in \mathbb{R}^n$, then $U(sx)$ is also linear. But $U(0) = 0$ and $U_x(0) = 0$, as $\mathbf{0}$ is a saddle point of φ, so by linearity $U(sx) = 0$ for all $s \in \mathbb{R}$. □

Proof of Theorem 3.4.4. This is just a simple combination of Lemma 3.6.11 and Lemma 3.6.10. □

We now consider the case of the projected gradient method.

Proof of Theorem 3.4.5. We show how to adapt the proof of the results on the gradient method. We denote the set of equilibrium points of the projected gradient method as $\bar{\mathcal{S}}^{\mathbf{\Pi}}$ and similarly $\mathcal{S}^{\mathbf{\Pi}}, \mathcal{S}^{\mathbf{\Pi}}_{\bar{\mathbf{z}}}$, in analogy with $\mathcal{S}, \mathcal{S}_{\bar{\mathbf{z}}}$.

We first note that the projected gradient method is pathwise stable which can be verified directly. Together with the assumption that $\mathbf{0} \in \bar{\mathcal{S}}^{\mathbf{\Pi}}$, this means that the reasoning in Section 3.6.2 applies, and in particular a version of Lemma 3.6.5 holds, i.e.

Lemma 3.6.12. *Let φ be C^2 and concave-convex on $\mathbb{R}^{n+m}$, $\mathbf{\Pi} \in \mathbb{R}^{(n+m)^2}$ be an orthogonal projection matrix, $\bar{\mathbf{z}} \in \bar{\mathcal{S}}^{\mathbf{\Pi}}$ and $\mathbf{z}(t) \in \mathcal{S}^{\mathbf{\Pi}}_{\bar{\mathbf{z}}}$. Then for any $s \in [0, 1]$, the convex combination $\mathbf{z}'(t) = (1 - s)\bar{\mathbf{z}} + s\mathbf{z}(t)$ lies in $\mathcal{S}^{\mathbf{\Pi}}_{\bar{\mathbf{z}}}$. If in addition $\mathbf{z} \in \mathcal{S}^{\mathbf{\Pi}}$, then $\mathbf{z}'(t) \in \mathcal{S}^{\mathbf{\Pi}}$.*

Equation (3.6.4) becomes

$$\mathbf{\Pi}\begin{bmatrix}\varphi_x(\mathbf{z})\\-\varphi_y(\mathbf{z})\end{bmatrix}\mathbf{\Pi} = \left(\int_0^{|\mathbf{z}|}\mathbf{\Pi}\begin{bmatrix}\varphi_{xx}(s\hat{\mathbf{z}}) & \varphi_{xy}(s\hat{\mathbf{z}})\\-\varphi_{yx}(s\hat{\mathbf{z}}) & -\varphi_{yy}(s\hat{\mathbf{z}})\end{bmatrix}\mathbf{\Pi}ds\right)\hat{\mathbf{z}}$$

and we replace (3.4.1) with

$$\widetilde{\mathbf{A}}(\mathbf{z}) = \mathbf{\Pi}\begin{bmatrix}0 & \varphi_{xy}(\mathbf{z})\\-\varphi_{yx}(\mathbf{z}) & 0\end{bmatrix}\mathbf{\Pi}, \qquad \widetilde{\mathbf{B}}(\mathbf{z}) = \mathbf{\Pi}\begin{bmatrix}\varphi_{xx}(\mathbf{z}) & 0\\0 & -\varphi_{yy}(\mathbf{z})\end{bmatrix}\mathbf{\Pi}$$

The remainder of the proofs carry through unaltered. □

3.7 Proof of convergence for the modification method

Instead of proving Proposition 3.5.1 directly we state and prove the following slightly more general result which contains Proposition 3.5.1 as the special case that $V = \mathbb{R}^{n+m}$.

Proposition 3.7.1. *Let $V \subseteq \mathbb{R}^{n+m}$ be an affine subspace, and φ be C^2 and concave-convex on V. Let φ' satisfy (3.5.1) and $M \in \mathbb{R}^{n'\times n}$ be such that $\ker(M)\cap\ker(\varphi_{xx}(\bar{\mathbf{z}})) = \{0\}$ for some equilibrium point $\bar{\mathbf{z}}$ of the subgradient method on V. Then the subgradient method (3.3.2) on $\mathbb{R}^{n'}\times V$ is globally convergent.*

Proof. By translation of coordinates we may assume that $\bar{\mathbf{z}}' = (M\bar{x}, \bar{x}, \bar{y}) = \mathbf{0} \in V$ is an equilibrium point. Let $\mathbf{\Pi}'$ be the orthogonal projection matrix onto the subspace $\mathbb{R}^{n'}\times V$. We decompose $\mathbf{\Pi}'$ on $\mathbb{R}^{n'}\times\mathbb{R}^{n+m}$ as

$$\mathbf{\Pi}' = \begin{bmatrix}I & \mathbf{0}\\\mathbf{0} & \mathbf{\Pi}\end{bmatrix}. \tag{3.7.1}$$

Now let $\mathbf{z}(t) = (x'(t), x(t), y(t))$ be a limiting solution of the subgradient method on K and let $(\tilde{x}(t), \tilde{y}(t)) = \mathbf{\Pi}(x(t), y(t))$.

Step 1: $x'(t)$ is constant. By applying Theorem 3.4.5 (noting Remark 3.4.5) we have that $\mathbf{z}$ solves (3.4.9). By the form of $\mathbf{A}(\mathbf{0})$ we deduce that $\dot{x}'(t) = 0$.

Step 2: $\tilde{x}(t)$ and $\tilde{y}(t)$ are constant. From the condition (3.4.10) we have that $\mathbf{B}(s\mathbf{z})\mathbf{\Pi}\mathbf{z} = 0$ for $r \in [0,1]$, from which it follows that

$$0 = \mathbf{z}^T\mathbf{\Pi}'\mathbf{B}(r\mathbf{z})\mathbf{\Pi}'\mathbf{z} = u^T\psi_{uu}u + \tilde{x}^T\varphi_{xx}\tilde{x} - \tilde{y}^T\varphi_{yy}\tilde{y} \tag{3.7.2}$$

where ψ_{uu} is the Hessian matrix of ψ evaluated at $u = M\tilde{x} - x'$. As each term is non-positive and ψ is strictly concave we deduce that $M\tilde{x} - x' = 0$ and $\tilde{x} \in \ker(\varphi_{xx}(\mathbf{0}))$. Thus $M\tilde{x}(t)$ is constant. By the condition that $\ker(M) \cap \ker(\varphi_{xx}) = \{0\}$ we deduce that $\tilde{x}(t)$ is constant. Then the form of $\mathbf{A}(\mathbf{0})$ allows us to deduce that $\tilde{y}(t)$ is also constant.

Step 3: $x(t)$ and $y(t)$ are constant. The vector field in (3.4.9) is orthogonal to $\ker(\mathbf{\Pi})$, so that $(\tilde{x}(t), \tilde{y}(t))$ being constant implies that $(x(t), y(t))$ are constant.

This proves that any limiting solution to the subgradient method on V is an equilibrium point, and therefore that the subgradient method on V is globally convergent. □

3.A Appendix

3.A.1 The addition of constant gains

It is common in applications to consider the gradient method with constant gains, i.e.

$$\begin{aligned} \dot{x}_i &= \gamma_i^x \varphi_{x_i} \quad \text{for } i = 1, \dots, n, \\ \dot{y}_j &= -\gamma_j^y \varphi_{y_j} \quad \text{for } j = 1, \dots, m. \end{aligned} \tag{3.A.1}$$

for $\varphi \in C^2$ a concave-convex function on $\mathbb{R}^{n+m}$ and γ_i^x, γ_j^y positive constants. However, in the setting of an arbitrary concave-convex, this is not a generalisation, and it is sufficient to study the gradient method (3.3.1) without gains, by a coordinate transformation that we now describe.

Let Λ be a diagonal matrix defined from the gains by

$$\mathbf{\Lambda} = \operatorname{diag}(\sqrt{\gamma_1^x}, \dots, \sqrt{\gamma_n^x}, \sqrt{\gamma_1^y}, \dots, \sqrt{\gamma_m^y}). \tag{3.A.2}$$

Given a concave-convex function φ we define a new concave-convex function φ' by

$$\varphi'(\mathbf{z}') = \varphi(\mathbf{\Lambda}\mathbf{z}'). \tag{3.A.3}$$

Let $\mathbf{z}'(t)$ be a solution to the gradient method (3.3.1) without gains applied to φ', then $\mathbf{z}(t) := \mathbf{\Lambda}\mathbf{z}'(t)$ is a solution to the gradient method (3.A.1) applied to φ with gains. Indeed, we have

$$\dot{\mathbf{z}}(t) = \mathbf{\Lambda}\dot{\mathbf{z}}'(t) = \mathbf{\Lambda}^2 \begin{bmatrix} \varphi_x(\mathbf{\Lambda}\mathbf{z}'(t)) \\ -\varphi_y(\mathbf{\Lambda}\mathbf{z}'(t)) \end{bmatrix} = \mathbf{\Lambda}^2 \begin{bmatrix} \varphi_x(\mathbf{z}(t)) \\ -\varphi_y(\mathbf{z}(t)) \end{bmatrix}$$

and the $\mathbf{\Lambda}^2$ term gives the gains.

Thus any properties of the gradient method with gains can be obtained from the gradient method without gains applied to a suitably modified function.

However, applying this transformation to the subgradient method has the effect of altering the metric in the convex projection.

Definition 3.A.1 (Subgradient method with gains)**.** *Given a non-empty closed convex set $K \subseteq \mathbb{R}^{n+m}$, $\varphi \in C^2$ a concave-convex function on K and a set of positive gains γ_i^x, γ_j^y as in (3.A.1), we define the subgradient method on K with gains as a semi-flow on (K, d) consisting of Carathéodory solutions of*

$$\dot{\mathbf{z}} = \mathbf{f}(\mathbf{z}) - \mathbf{P}_{N_K(\mathbf{z}), d_{\mathbf{\Lambda}^{-1}}}(\mathbf{f}(\mathbf{z})) \tag{3.A.4}$$

where $\mathbf{f}(\mathbf{z})$ is the vector field of the gradient method with gains (3.A.1) and $\mathbf{P}_{M, d_{\mathbf{\Lambda}^{-1}}}$ is a weighted convex projection given by

$$\mathbf{P}_{M, d_{\mathbf{\Lambda}^{-1}}}(\mathbf{z}) = \operatorname{argmin}_{\mathbf{w} \in M} d(\mathbf{\Lambda}^{-1}\mathbf{w}, \mathbf{\Lambda}^{-1}\mathbf{z}) \tag{3.A.5}$$

where Λ is defined in terms of the gains by (3.A.2).

This change arises from the stretching of the domain K in the applied coordinate transformation.

Remark 3.A.1. *The form of non-negativity constraints is not affected by this change to the metric in the convex projection. For example, if the y coordinates are restricted to be non-negative and the x coordinates unconstrained, then the*

subgradient method with gains (3.A.4) *is given by*

$$\begin{aligned} \dot{x}_i &= \gamma_i^x \varphi_{x_i} \qquad \text{for } i = 1, \dots, n, \\ \dot{y}_j &= [-\gamma_j^y \varphi_{y_j}]_{y_j}^+ \quad \text{for } j = 1, \dots, m. \end{aligned} \tag{3.A.6}$$

This holds more generally for any convex set K with boundaries aligned to the coordinate axes.

3.A.2 Proof of Proposition 3.4.1

Proof of Proposition 3.4.1. Let $(x'(t), y'(t))$ and $(x(t), y(t))$ be two solutions of (3.3.1) and define $2W(t) = |x'(t) - x(t)|^2 + |y'(t) - y(t)|^2$. Then we have

$$\begin{aligned} \dot{W} &= (x' - x)^T(\dot{x}' - \dot{x}) + (y' - y)^T(\dot{y}' - \dot{y}) \\ &= (x' - x)^T(\varphi_x' - \varphi_x) - (y' - y)^T(\varphi_y' - \varphi_y) \\ &= \int_0^1 \frac{d}{ds} \left\{ (x' - x)^T \varphi_x \circ \gamma(s) - (y' - y)^T \varphi_y \circ \gamma(s) \right\} ds \end{aligned}$$

where φ_x' denotes φ_x at (x', y'), and $\gamma(s) = ((x' - x)s + x, (y' - y)s + y)$ traverses the line from x to x' linearly. Note that only the partial derivatives of φ depend on s. Continuing, letting $\hat{x} = x' - x$ and $\hat{y} = y' - y$,

$$\begin{aligned} \dot{W} &= \int_0^1 \hat{x}^T \varphi_{xx} \circ \gamma(s) \hat{x} \, ds + \int_0^1 \hat{x}^T \varphi_{xy} \circ \gamma(s) \hat{y} \, ds + \\ &\quad - \int_0^1 \hat{y}^T \varphi_{yy} \circ \gamma(s) \hat{y} \, ds - \int_0^1 \hat{y}^T \varphi_{yx} \circ \gamma(s) \hat{x} \, ds \\ &= \int_0^1 \hat{x}^T \varphi_{xx} \circ \gamma(s) \hat{x} \, ds - \int_0^1 \hat{y}^T \varphi_{yy} \circ \gamma(s) \hat{y} \, ds \end{aligned}$$

By concavity/convexity we have that $\varphi_{xx}, \varphi_{yy}$ are negative/positive semi-definite which shows that $\dot{W} \leq 0$. □

Chapter 4

Stability and instability in gradient dynamics - Part II: The subgradient method

In this chapter we extend the results of Chapter 3 to the subgradient method where the dynamics are constrained to lie in a prescribed convex set. Having a discontinuous vector field, the convergence of the subgradient method is non-trivial to study as the common tools of *smooth* analysis such as the classical LaSalle and Lyapunov theorems do not apply. We provide a general framework of results that reduce the study of the asymptotic behaviour of the subgradient method to the study of a explicit family of *smooth* systems, to which more standard techniques can be applied.

Acknowledgements

The work in this chapter was done in collaboration with Ioannis Lestas and forms the second part of a two part paper in preparation [93].

4.1 Introduction

In Chapter 3 we studied the asymptotic behaviour of the gradient method when this is applied on a general concave-convex function in an unconstrained domain, and provided an exact characterization to its limiting solutions. Nevertheless, in many applications, such as primal/dual algorithms in optimization problems, it becomes necessary to constrain the system states in a prescribed convex set, e.g. positivity constraints on Lagrange multipliers or constraints on physical quantities like data flow, and prices/commodities in economics [97], [110], [183], [61]. The subgradient method is used in such cases, which is a version of the gradient method with a projection term in the vector field additionally included, so as to ensure that the trajectories do not leave the desired set.

In discrete time, there is an extensive literature on the subgradient method, via its application in optimization problems (see e.g. [153]). However, in many applications, for example power networks [32], [51], [108] and classes of data network problems [110], [183] continuous time models are considered. It is thus important to have a good understanding of the subgradient dynamics in a continuous time setting, which could also facilitate analysis and design by establishing links with other more abstract results in dynamical systems theory.

A main complication in the study of the subgradient method arises from the fact the this is a *non-smooth* nonlinear ODE with a discontinuous vector field due to the projections involved. This prohibits the direct application of classical Lyapunov or LaSalle theorems (e.g. [112]), which is reflected in the direct approach used by Arrow, Hurwicz and Uzawa in [10] that avoids the use of such tools. More recently, the work of Feijer and Paganini [61] unified the previously ad-hoc and application focused analysis of primal dual gradient dynamics in network optimisation, and proposed that the switching in the dynamics be interpreted in the framework of hybrid automata, where a LaSalle Invariance principle was recently obtained in [135]. However, as recently pointed out in [38], there are cases where the assumptions required in [135] do not hold. In [38], the LaSalle invariance principle for discontinuous Carathéodory systems is applied to prove convergence of the subgradient method under positivity constraints and the assumption of strict concavity. In [173] the subgradient method is used to solve

linear programs with inequality constraints. In general, proving convergence for the subgradient method even in simple cases, is a non-trivial problem that requires the non-smooth character of the system to be explicitly addressed.

Our aim in this chapter is to provide a framework of results that allow one to study the asymptotic behaviour of the subgradient method (4.3.2) with *smooth analysis* as opposed to *non-smooth analysis*. One of our main results is that the limiting behaviour of the subgradient method constrained to an arbitrary convex domain is given by the solutions of one of an explicit family of *smooth* differential equations. With this result, proving convergence of the subgradient method may be done with standard Lyapunov and LaSalle type stability tools, reducing the barrier to obtaining rigorous convergence proofs in applications.

We illustrate our results by means of various examples and also apply those to modification schemes in network optimization, that provide convergence guarantees while maintaining the decentralized structure of the dynamics. We also discuss an application to the problem of multi-path congestion control (e.g. [199, 110, 126, 129]), where we prove convergence of a modification scheme, that achieves optimality without requiring any additional information transfer.

The chapter is structured as follows. Various preliminaries from convex analysis and dynamical systems are given in Section 4.2. The problem formulation is given in Section 4.3 and the main results are presented in Section 4.4, where various examples that illustrate those are also discussed. Applications to modification methods in network optimization, and to the problem of multipath routing are given in Section 4.5. The proofs of the result are finally given in Section 4.6.

Addendum: Subsequent to the preparation of this thesis, the author became aware of the work [204], which contains a related result to Theorem 4.4.1 on stability of projected dynamical systems and minimal faces. We believe the application to the subgradient method in the considered applications remains novel.

4.2 Preliminaries

We use the same notation and definitions as the previous Chapter 3 and we refer the reader to the preliminaries section therein. In addition we also have need of the following.

4.2.1 Convex analysis

4.2.1.1 Concave-convex functions and saddle points

For a function φ that is concave-convex on $\mathbb{R}^{n+m}$ the (standard) notion of a saddle point was given in Definition 3.2.2 in the previous chapter. We now consider φ restricted to a non-empty closed convex set $K \subseteq \mathbb{R}^{n+m}$, in which case the notion of saddle point needs to be modified to incorporate the constraints.

Definition 4.2.1 (Restricted saddle point)**.** *Let $K \subseteq \mathbb{R}^{n+m}$ be non-empty closed and convex. For a concave-convex function $\varphi : K \to \mathbb{R}$, we say that $(\bar{x}, \bar{y}) \in K$ is a K-restricted saddle point of φ if for all $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$ with $(x, \bar{y}), (\bar{x}, y) \in K$ we have the inequality $\varphi(x, \bar{y}) \leq \varphi(\bar{x}, \bar{y}) \leq \varphi(\bar{x}, y)$.*

If in addition $\varphi \in C^1$ then $\bar{\mathbf{z}} = (\bar{x}, \bar{y}) \in K$ is a K-restricted saddle point if and only if the vector of partial derivatives $\begin{bmatrix} \varphi_x(\bar{\mathbf{z}}) & -\varphi_y(\bar{\mathbf{z}}) \end{bmatrix}^T$ lies in the normal cone $N_K(\bar{\mathbf{z}})$.

Any K-restricted saddle point in the interior of K is also a saddle point. If $C \subseteq K$ is closed and convex and $\bar{\mathbf{z}} \in C$ is a K-restricted saddle point, then $\bar{\mathbf{z}}$ is also a C-restricted saddle point.

However, it in general does not hold that if $\varphi : \mathbb{R}^{n+m} \to \mathbb{R}$ has a saddle point, and K is closed convex and non-empty, then φ has a K-restricted saddle point. (An explicit example contradicting this is given later in Example 4.4.3(ii)). In this chapter we will only consider cases where at least one K-restricted saddle point exists, leaving the problem of showing existence to the specific application.

4.2.1.2 Concave programming

Concave programming (see e.g. [27]) is concerned with the study of optimization problems of the form

$$\max_{x \in C, g(x) \geq 0} U(x) \tag{4.2.1}$$

where $U : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}^m$ are concave functions and $C \subseteq \mathbb{R}^n$ is non-empty closed and convex. This is associated with the Lagrangian

$$\varphi(x, y) = U(x) + y^T g(x) \tag{4.2.2}$$

where $y \in \mathbb{R}^m_+$ are the Lagrange multipliers.

Theorem 4.2.1. *Let g be concave and Slater's condition hold, i.e.*

$$\exists x' \in \operatorname{relint} C \text{ with } g(x') > 0. \tag{4.2.3}$$

Then $\bar{x}$ is an optimum of (4.2.1) if and only if $\exists \bar{y}$ with $(\bar{x}, \bar{y})$ a $C \times \mathbb{R}^m_+$-restricted saddle point of (4.2.2).

The min-max optimization problem associated finding a K-restricted saddle point of (4.2.2) is the dual problem of (4.2.1).

4.2.1.3 Faces of convex sets

Some of the main results of this chapter refer to faces of a convex set. We refer the reader to [78, Chap. 1.8.] for further discussion of such topics.

Definition 4.2.2 (Face of a convex set)**.** *Given a non-empty closed convex set K, a face F of K is a subset of K that has both the following properties:*

(i) F is convex.

(ii) For any line segment $L \subseteq K$, if $(\operatorname{relint} L) \cap F \neq \emptyset$ then $L \subseteq F$.

For the readers convenience we recall some standard properties of faces:

(a) The intersection of two faces of K is a face of K.

(b) The empty set and K itself are both faces of K. If a face F is neither $\emptyset$ or K it is called a proper face.

(c) If F is a face of K and F' is a face of F, then F' is a face of K.

(d) For a face F of K, the normal cone $N_K(\mathbf{z})$ is independent of the choice of $\mathbf{z} \in \text{relint}(F)$. In these cases we drop the $\mathbf{z}$ dependence and write it as N_F.

(e) K may be written as the disjoint union:

$$K = \bigcup \{\text{relint } F : F \text{ is a face of } K\}. \tag{4.2.4}$$

Property (a) above leads to the following definition.

Definition 4.2.3 (Minimal face containing a set)**.** *For a convex set K and a subset $A \subseteq K$ we define the* minimal face containing A *as*

$$\bigcap \{F : F \text{ is a face of } K \text{ and } A \subseteq F\}$$

which is a face by property (a) above.

4.2.2 Dynamical systems

Definition 4.2.4 (Flows and semi-flows)**.** *A triple (ϕ, X, ρ) is a flow (resp. semi-flow) if (X, ρ) is a metric space, ϕ is a continuous map from $\mathbb{R} \times X$ (resp. $\mathbb{R}_+ \times X$) to X which satisfies the two properties*

(i) For all $x \in X$, $\phi(0, x) = x$.

(ii) For all $x \in X$, $t, s \in \mathbb{R}$ (resp. $\mathbb{R}_+$),

$$\phi(t + s, x) = \phi(t, \phi(s, x)). \tag{4.2.5}$$

When there is no confusion over which (semi)-flow is meant, we shall denote $\phi(t, x(0))$ as $x(t)$. For sets $A \subseteq \mathbb{R}$ (resp. $\mathbb{R}_+$) and $B \subseteq X$ we define $\phi(A, B) = \{\phi(t, x) : t \in A, x \in B\}$.

Definition 4.2.5 (ω-limit set)**.** *Given a semi-flow (ϕ, X, ρ) we denote the set of ω-limit points of trajectories as*

$$\Omega(\phi, X, \rho) = \bigcup_{x \in X} \bigcap_{t \geq 0} \overline{\phi([t, \infty), x)}. \tag{4.2.6}$$

where $\overline{A}$ denotes the closure of $A \subseteq X$ in (X, ρ).

Definition 4.2.6 (Invariant sets)**.** *For a semi-flow (ϕ, X, ρ) we say that a set $A \subseteq X$ is positively invariant if $\phi(\mathbb{R}_+, A) \subseteq A$. If ϕ is also a flow we say that A is negatively invariant if $\phi((-\infty, 0], A) \subseteq A$. If $\phi(t, A) = A$ for all $t \in \mathbb{R}$ then we say A is invariant.*

Definition 4.2.7 (Sub-(semi)-flow)**.** *For a flow (resp. semi-flow) (ϕ, X, ρ) and an invariant (resp. positively invariant) set $A \subseteq X$ we obtain the sub-flow (resp. sub-semi-flow) by restricting $\phi(t, x)$ to act on $x \in A$ and denote it as (ϕ, A, ρ).*

Definition 4.2.8 (Global convergence)**.** *We say that a (semi)-flow (ϕ, X, ρ) is globally convergent, if for all initial conditions $x \in X$, the trajectory $\phi(t, x)$ converges to the set of equilibrium points of (ϕ, X, ρ) as $t \to \infty$, i.e.*

$$\inf\{d(\phi(t, x), y) : y \text{ an equilibrium point}\} \to 0 \text{ as } t \to \infty.$$

In Chapter 3 much of the analysis relied on a specific form of incremental stability which we reproduce below for the convenience of the reader.

Definition 4.2.9 (Pathwise stability)**.** *We say that a semi-flow (ϕ, X, ρ) is pathwise stable*[1] *if for any two trajectories $x(t), x'(t)$ the distance $\rho(x(t), x'(t))$ is non-increasing in time.*

This is linked with the following special class of (semi)-flows.

Definition 4.2.10 ((Semi)-Flow of isometries)**.** *We say that a (semi)-flow (ϕ, X, ρ) is a (semi)-flow of isometries if for every $t \in \mathbb{R}$ (resp. $\mathbb{R}_+$), the function $\phi(t, \cdot) : X \to X$ is an isometry, i.e. for all $x, y \in X$ it holds that $\rho(\phi(t, x), \phi(t, y)) = \rho(x, y)$.*

[1] As was noted on Page 131, we use the term 'pathwise stability' to avoid the confusion around existing similar notions with conflicting definitions in different places.

Finally we have need of the notion of Carathéodory solutions of differential equations.

Definition 4.2.11 (Carathéodory solution)**.** *We say that a trajectory* $\mathbf{z}(t)$ *is a Carathéodory solution to a differential equation* $\dot{\mathbf{z}} = \mathbf{f}(\mathbf{z})$*, if* $\mathbf{z}$ *is an absolutely continuous function of* t*, and for almost all times* t*, the derivative* $\dot{\mathbf{z}}(t)$ *exists and is equal to* $\mathbf{f}(\mathbf{z}(t))$*.*

Note that we do not require that $\mathbf{f}$ satisfies the assumptions of the Carathéodory existence theorem.

4.3 Problem formulation

The main object of study in this work is the *subgradient method* on an arbitrary concave-convex function in C^2 and an arbitrary convex domain K. We first recall the definition of the *gradient method*, which is studied in Chapter 3.

Definition 4.3.1 (Gradient method)**.** *Given* φ *a* C^2 *concave-convex function on* $\mathbb{R}^{n+m}$*, we define the gradient method as the flow on* $(\mathbb{R}^{n+m}, d)$ *generated by the differential equation*

$$\begin{aligned} \dot{x} &= \varphi_x \\ \dot{y} &= -\varphi_y. \end{aligned} \tag{4.3.1}$$

The *subgradient method* is obtained by restricting the gradient method to a convex set K by the addition of a projection term to the differential equation (4.3.1).

Definition 4.3.2 (Subgradient method)**.** *Given a non-empty closed convex set* $K \subseteq \mathbb{R}^{n+m}$ *and a* C^2 *function* φ *that is concave-convex on* K*, we define the subgradient method on* K *as a semi-flow on* (K, d) *consisting of Carathéodory solutions of*

$$\begin{aligned} \dot{\mathbf{z}} &= \mathbf{f}(\mathbf{z}) - \mathbf{P}_{N_K(\mathbf{z})}(\mathbf{f}(\mathbf{z})) \\ \mathbf{f}(\mathbf{z}) &= \begin{bmatrix} \varphi_x & -\varphi_y \end{bmatrix}^T. \end{aligned} \tag{4.3.2}$$

where the notion of Carathéodory solution to a differential equation is defined in Definition 4.2.11.

As explained in Section 3.A.1, all the results of this chapter can be translated to the subgradient method with gains by a transformation of coordinates.

Remark 4.3.1. *For (non-affine) convex sets K the subgradient method* (4.3.2) *is a* non-smooth *system. The vector field is discontinuous due to the convex projection term, independently of the regularity of the function φ or of the boundary of K. This is in contrast to the gradient method* (4.3.1)*, which is a* smooth *system, as it inherits the regularity of the function φ.*

The equilibrium points of the subgradient method on K are exactly the K-restricted saddle points.

We briefly summarise the contributions of this work in the bullet points below.

- We give conditions through which convergence can be deduced for the subgradient method (which is a *non-smooth* system due to the presence of switching) via the study of solutions to explicit *smooth* ODEs derived from the form of the concave-convex function and the convex domain.
- These smooth ODEs are given by the subgradient method on *affine subspaces*, this links with the previous Chapter 3, where the convergence properties of these smooth systems are studied.
- We give a proof of the convergence of the subgradient method applied to any strictly concave-convex function for arbitrary convex domains. Furthermore, we provide example applications of our results on the subgradient method to various methods of modifying a concave-convex function to give convergence. In particular, we give an application to the problem of congestion control in multi-path routing.

The study of the (sub)gradient method was originated by Arrow, Hurwicz and Uzawa [10], who took a direct approach and established convergence of the subgradient method with positivity constraints under the assumption of strict concave-convexity [10]. More recently, Feijer and Paganini [61] attempted to use the invariance principle for hybrid automata [135] to modernise and unify the ad-hoc approaches that had dominated until then. They provided a new proof of the convergence result of Arrow, Hurwicz and Uzawa, and also proved convergence of

a number of modification methods where strict concave-convexity is absent. However, recently it has been pointed out by Cherukuri, Mallada and Cortés [38] that there are cases where, when interpreted as a hybrid automata, the subgradient method does not satisfy all the assumptions of the invariance principle in [135]. In [38] an invariance principle for Carathéodory solutions is applied to prove convergence with positivity constraints for strictly concave-convex functions which are linear in the second variable (i.e. of the form (4.2.2)).

4.4 Main Results

This section states the main results of the chapter.

The aim of this work is to study the convergence properties of the subgradient method (Definition 4.3.2) applied to general concave-convex functions which lack strict concavity and on an arbitrary convex domain.

We divide the results into three parts, which are outlined below for the convenience of the reader.

- In Section 4.4.1 we describe the essential problems that arise in analysis of the non-smooth dynamics of the subgradient method, and then develop tools to deal with this problem. In Proposition 4.4.1 we give a invariance principle for pathwise stable semi-flows, which applies without any smoothness assumption on the dynamics. We then study semi-flows generated by projecting a pathwise stable ODE onto a closed convex set, and obtain the key result, Theorem 4.4.1, that says that the dynamics on the ω-limit set are smooth.

- In Section 4.4.2 we apply these tools to the subgradient method (4.3.2). In Theorem 4.4.2 we show that the limiting solutions of the (non-smooth) subgradient method on a convex set are given by the dynamics of the (smooth) subgradient method on an affine subspace, and describe exactly the set of limiting solutions when there is an internal saddle point. This allows us to obtain Corollary 4.4.1, a criterion for global asymptotic stability of the subgradient method.

- In Section 4.4.3 we combine Theorem 4.4.2 with the results of Chapter 3 (for convenience of the reader reproduced in Section 4.A) to obtain a general convergence criterion (Theorem 4.4.3) for the subgradient method.

We illustrate these results with examples throughout.

4.4.1 Pathwise stability and convex projections

If one wishes to extend the results of Chapter 3 to the subgradient method on a non-empty closed convex set $K \subseteq \mathbb{R}^{n+m}$, then one runs into two problems, both coming from the discontinuity of the vector field in (4.3.2). The first is that the previously simple application of LaSalle's theorem would become much more technical - needing tools from non-smooth analysis. The second, more fundamental, problem is that LaSalle's theorem only gives convergence to a set of trajectories, and it remains to characterise this set. The trajectories in this set still satisfy an ODE with a discontinuous vector field, and we do not have uniqueness of the solution backwards in time - we still only have a semi-flow.

To solve these issues we reinterpret the prior results in terms of a simple property which is still present in the subgradient method.

The main tool used to prove the results in Chapter 3 was pathwise stability, (Definition 4.2.9), which says that the Euclidean distance between any two solutions is non-increasing with time. (We will later prove such a result for the subgradient method). Intuitively, one would think that the distance between any two of the limiting solutions would be constant, and indeed, one can verify that this is the case directly from (4.4.22) (below) and the skew-symmetry of $\mathbf{A}(\mathbf{0})$ given by (4.A.1). A more abstract way of saying this is that the sub-flow obtained by considering the gradient method acting on the set of limiting solutions is a *flow of isometries*. In fact, this can be proved more directly for any pathwise stable semi-flow.

Proposition 4.4.1. *Let (ϕ, X, d) be a pathwise stable semi-flow (see Definition 4.2.9) with $X \subseteq \mathbb{R}^{n+m}$ which has an equilibrium point $\bar{\mathbf{z}}$. Let Ω be its ω-limit set. Then the sub-semi-flow (ϕ, Ω, d) (see Definition 4.2.7) defines a flow of isometries (see Definition 4.2.10). Moreover, Ω is a convex set.*

Note here that (ϕ, Ω, d) is a *flow* rather than a *semi-flow*. This comes from the simple observation that an isometry is always invertible, so we can define, for $t \geq 0$, $\phi(-t, \cdot) : \Omega \to \Omega$ as $\phi(t, \cdot)^{-1}$.

Remark 4.4.1. *Care should be taken in interpreting the backwards flow given by Proposition 4.4.1. There could be multiple trajectories in X that meet at a point in $y \in \Omega$ at time $t = 0$, but exactly one of these trajectories will lie in Ω for all times $t \in \mathbb{R}$.*

We would like to note that we are not the first to make this observation. Indeed, we deduce this result from a more general result in [44] which was published in 1970.

We consider pathwise stable differential equations which are projected onto a convex set, and make the following set of assumptions.[2]

$$\begin{aligned} &(\phi, K, d) \text{ is the semi-flow of Carathéodory solutions of} \\ &\qquad \dot{\mathbf{z}} = \mathbf{f}(\mathbf{z}) - \mathbf{P}_{N_K(\mathbf{z})}(\mathbf{f}(\mathbf{z})) \text{ where,} \\ &\qquad K \subseteq \mathbb{R}^n \text{ is non-empty, closed and convex} \\ &C^1 \ni \mathbf{f} : K \to \mathbb{R}^n \text{ satisfies, for all } \mathbf{z}, \mathbf{w} \in K, \\ &\qquad (\mathbf{f}(\mathbf{z}) - \mathbf{f}(\mathbf{w}))^T(\mathbf{z} - \mathbf{w}) \leq 0. \end{aligned} \tag{4.4.1}$$

A simple first result is that the projected dynamics are still pathwise stable.

Lemma 4.4.1. *Let (4.4.1) hold. Then (ϕ, K, d) is pathwise stable.*

Our main result on such projected differential equations is that, even though the projection term gives a discontinuous vector field, when we restrict our attention to the ω-limit set, the vector field is C^1. This allows us to replace *non-smooth analysis* with *smooth analysis* when studying the asymptotic behaviour of such systems.

Theorem 4.4.1. *Let (4.4.1) hold and assume that the semi-flow (ϕ, K, d) has an equilibrium point. Let Ω be its ω-limit set. Then (ϕ, Ω, d) defines a flow of isometries given by solutions to the following differential equation, which has a*

[2] That the final assumed inequality in (4.4.1) holds for the subgradient method is evident from the proof of the pathwise stability of the gradient method (Proposition 3.4.1.)

C^1 vector field,

$$\dot{\mathbf{z}} = \mathbf{f}(\mathbf{z}) - \mathbf{P}_{N_V}(\mathbf{f}(\mathbf{z})). \tag{4.4.2}$$

Here V is the affine span of the (unique) minimal face (see Definition 4.2.3) of K that contains the set of equilibrium points of the semi-flow. If $\mathbf{f}$ is defined on all of V then Ω is contained in the ω-limit set of the flow generated by (4.4.2) on V.

Remark 4.4.2. *The existence of a minimal face of K that contains the set of equilibrium points is a simple consequence of the definition of a face (see Definition 4.2.2 and the discussion that follows). The important part of Theorem 4.4.1 is that the dynamics on Ω are given by (4.4.2), i.e. the projection operator $\mathbf{P}_{N_K(\mathbf{z})}$ in (4.3.2) becomes $\mathbf{P}_{N_V}$ which does not depend on the position $\mathbf{z}$.*

Remark 4.4.3. *In the statement of Theorem 4.4.1, it is not necessarily the case that $\mathbf{P}_{N_V}(\mathbf{f}(\mathbf{z})) = \mathbf{P}_{N_K(\mathbf{z})}(\mathbf{f}(\mathbf{z}))$ for all $\mathbf{z}$ in the minimal face provided by the Theorem, this is only guaranteed for $\mathbf{z} \in \Omega$. Neither is it the case that $N_K(\mathbf{z})$ must be constant and equal to N_V on Ω, only that the projection of $\mathbf{f}(\mathbf{z})$ onto N_V and $N_K(\mathbf{z})$ must be equal.*

4.4.2 Subgradient method

We now apply theses results to the subgradient method. Our first result reduces the study of the convergence on general convex domains, where the subgradient method is non-smooth, to the study of convergence of the subgradient method on affine spaces, on which the subgradient method is smooth. We also give an exact classification of the asymptotic behaviour in the case of an internal saddle point.

As in Chapter 3, given a concave-convex function φ we define $\bar{\mathcal{S}}$ and $\mathcal{S}$ respectively as the set of saddle points and the set of solutions to the gradient method (4.3.1) that lie a constant distance from any saddle point.

Theorem 4.4.2. *Let $K \subseteq \mathbb{R}^{n+m}$ be non-empty, closed and convex. Let φ be C^2, concave-convex on K and have a K-restricted saddle point. Let (ϕ, K, d) denote the subgradient method (4.3.2) on K and Ω be its ω-limit set. Then Ω is convex, and (ϕ, Ω, d) defines a flow of isometries. Exactly one of the following holds:*

(i) There is an internal saddle point $\bar{\mathbf{z}} \in \bar{\mathcal{S}} \cap \operatorname{int} K$ *and*

$$\Omega = \{\mathbf{z}(t) \in \mathcal{S} : \mathbf{z}(\mathbb{R}) \subseteq K\}. \tag{4.4.3}$$

(ii) $\bar{\mathcal{S}} \cap \operatorname{int} K = \emptyset$*. Let* F *be the minimal face containing all* K*-restricted saddle points and* V *be the affine span of* F*, which is proper. Then trajectories* $\mathbf{z}(t)$ *of* (ϕ, Ω, d) *solve the ODE:*

$$\dot{\mathbf{z}} = \mathbf{f}(\mathbf{z}) - \mathbf{P}_{N_V}(\mathbf{f}(\mathbf{z})), \tag{4.4.4}$$

where $\mathbf{f}(\mathbf{z}) = \begin{bmatrix} \varphi_x & -\varphi_y \end{bmatrix}^T$*. Furthermore, if* φ *is also concave-convex on* V *then* Ω *is contained in the* ω*-limit set of the subgradient method* (4.3.2) *on* V*.*

Remark 4.4.4. *The ODE* (4.4.4) *is the subgradient method on the affine subspace* V*. The limiting solutions of this smooth system were classified in Chapter 3. Later, in Section 4.4.3 we use the results of Chapter 3 together with Theorem 4.4.2 to obtain a convergence criterion for the subgradient method. This is used subsequently give proofs for the applications (Section 4.5).*

The corresponding result for the subgradient method with constant gains can be obtained by a coordinate transformation as described in Section 3.A.1.

Note that Theorem 4.4.2(i) is a special case of Theorem 4.4.2(ii) when the (non-proper) face is the whole set K, and the affine span of K is $\mathbb{R}^{n+m}$.

Theorem 4.4.2(i) and the results in Chapter 3 give a full characterisation of the limiting solutions of the subgradient method when there is an internal saddle point, however, in applications it is common for this to not hold, e.g. in the case of a Lagrangian originating from an optimisation problem where at least one of the inequality constraints is binding. In such cases Theorem 4.4.2(ii) applies and gives a smooth ODE that the limiting solutions must solve.

We now present several examples to illustrate the application of Theorem 4.4.2 in some simple cases.

Example 4.4.1. *Consider the case where* $K \subseteq \mathbb{R}^{n+m}$ *is strictly convex. In this case the proper faces of* K *are given by* $\{\mathbf{w}\}$ *for each* $\mathbf{w} \in \partial K$*, i.e. each consist*

of a single point of the boundary of K. The subgradient method on a single point is trivially globally convergent.

By applying the two cases of Theorem 4.4.2, we obtain that:

(i) If there is a saddle point in the interior of K, then then the subgradient method (4.3.2) on K is globally convergent if and only if the (unconstrained) gradient method (4.3.1) is globally convergent.

(ii) If there is no saddle point in the interior of K, but there is a K-restricted saddle point, then the subgradient method is globally convergent.

There are cases where the unconstrained gradient method (4.3.1) is globally convergent, but the subgradient method is not, as the next example illustrates.

Example 4.4.2. *Define the concave-convex function*

$$\varphi(x_1, x_2, y) = -\frac{1}{2}|x_1|^2 + (x_1 + x_2)y. \tag{4.4.5}$$

This has a single saddle point at $(0, 0, 0)$, and corresponds to the optimisation problem

$$\max_{x_1+x_2=0} -\frac{1}{2}|x_1|^2 \tag{4.4.6}$$

where the constraint is relaxed with the Lagrange multiplier y. On this function the gradient method is the linear system

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{y} \end{bmatrix} = \begin{bmatrix} -1 & 0 & 1 \\ 0 & 0 & 1 \\ -1 & -1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ y \end{bmatrix}. \tag{4.4.7}$$

It is easily verified that all the eigenvalues of this matrix lie in the left half plane, so that the gradient method is globally convergent. Now consider the family of convex sets defined by

$$K_a = \{(x_1, x_2, y) \in \mathbb{R}^3 : x_1 \geq a\} \tag{4.4.8}$$

for $a \in \mathbb{R}$. The subgradient method on K_a is given by the system

$$\begin{aligned} \dot{x}_1 &= [-x_1 + y]^+_{x_1 - a} \\ \dot{x}_2 &= y \\ \dot{y} &= -x_1 - x_2. \end{aligned} \tag{4.4.9}$$

The convergence of the subgradient method on K_a depends crucially on the value of a. There are three cases:

(i) $a < 0$: In this case the saddle point $(0,0,0)$ lies in the interior of K_a so that Theorem 4.4.2(i) applies, and as the unconstrained gradient method is globally convergent, so is the subgradient method on K_a.

(ii) $a > 0$: Here the unconstrained saddle point $(0,0,0)$ lies outside K_a. A simple computation shows that the point $(a,a,0)$ is the only K_a-restricted saddle point. Thus, Theorem 4.4.2(ii) applies. The only proper face of K_a is the set

$$F_a = \{(a, x_2, y) : x_2, y \in \mathbb{R}\}. \tag{4.4.10}$$

The subgradient method on F_a is the system

$$\begin{bmatrix} \dot{x}_2 \\ \dot{y} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} x_2 \\ y \end{bmatrix} + \begin{bmatrix} 0 \\ -a \end{bmatrix} \tag{4.4.11}$$

together with the equality $x_1 = a$. This matrix has imaginary eigenvalues $\pm i$, showing that the subgradient method on F_a is not globally convergent. This does yet imply that the subgradient method on K_a is not globally convergent, as we have not verified that some subset of these oscillatory solutions to the subgradient method on F_a are also solutions of the subgradient method on K_a. However, it is easy to verify that this is indeed the case, so that the subgradient method on K_a is not globally convergent when $a > 0$.

(iii) $a = 0$: In this case the saddle point $(0,0,0)$ lies on the boundary of K_0, so that Theorem 4.4.2(ii) applies, and the analysis of the subgradient method on F_0 is the same as in case (ii) above. However, when we check whether any oscillatory solutions of the subgradient method on F_0 are also solutions of the subgradient method on K_0, we find that there are no such solutions. Indeed, to be a solution to the subgradient method on F_0 and the subgradient

method on K_0 we must have both $x_1 = a = 0$ and $-x_1 + y \leq 0$ by (4.4.9). Then (4.4.9) implies that $y = 0$ and then that $x_1 = 0$. So the only such solution is the saddle point. Therefore the subgradient method is K_0 on globally convergent.

This shows that the subgradient method on K_a undergoes a bifurcation at $a = 0$.

The following example illustrates that the subgradient method can be globally convergent when the gradient method is not.

Example 4.4.3. *Define the concave-convex function*

$$\varphi(x_1, x_2, y) = -\frac{1}{2}|x_2|^2 + x_1 y. \tag{4.4.12}$$

This has a single saddle point at $(0, 0, 0)$ and corresponds to the optimisation problem

$$\max_{x_1=0} -\frac{1}{2}|x_2|^2 \tag{4.4.13}$$

where the constraint is relaxed via the Lagrange multiplier y. The gradient method applied to φ is the linear system

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{y} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & -1 & 0 \\ -1 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ y \end{bmatrix} \tag{4.4.14}$$

whose matrix has eigenvalues $-1, \pm i$ so the gradient method is not globally convergent. We again consider the subgradient method on the closed convex set K_a defined by (4.4.8) for $a \in \mathbb{R}$ splitting into three cases:

(i) $a < 0$: As in Example 4.4.2(i) the saddle point $(0, 0, 0)$ lies in the interior of K_a. As the unconstrained gradient method is not globally convergent, Theorem 4.4.2(i) implies that the subgradient method on K_a is also not globally convergent.

(ii) $a > 0$*: The subgradient method on* K_a *is given by*

$$\begin{aligned} \dot{x}_1 &= [y]^+_{x_1 - a} \\ \dot{x}_2 &= -x_2 \\ \dot{y} &= -x_1 \end{aligned} \tag{4.4.15}$$

The saddle point $(0,0,0)$ *lies outside* K_a*. For* $(\bar{x}_1, \bar{x}_2, \bar{y})$ *to be a* K_a*-restricted saddle point,* (4.4.15) *implies that* $\bar{x_1} = \bar{x}_2 = 0$*, but this is impossible in* K_a*, so there are no* K_a*-restricted saddle points. This can also be understood in terms of the optimisation problem* (4.4.13) *which has empty feasible set if we impose the further condition that* $x_1 \geq a > 0$*. This means that none of our results apply, but a direct analysis of* (4.4.15) *shows that* $\dot{y} \leq -a < 0$ *so that* $y(t) \to -\infty$ *as* $t \to \infty$*, and the system is not globally convergent.*

(iii) $a = 0$*: Solving* (4.4.15) *for the* K_0*-restricted saddle points yields the continuum* $\{(0,0,y) : y \leq 0\}$*. None of these lie in the interior of* K_0 *so Theorem 4.4.2(ii) applies. The only proper face of* K_0 *is* F_0 *defined by* (4.4.10)*. On* F_0*, the subgradient method is the system*

$$\begin{bmatrix} \dot{x}_2 \\ \dot{y} \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_2 \\ y \end{bmatrix} \tag{4.4.16}$$

together with the equality $x_1 = 0$*, which is clearly globally convergent, noting that the set of* F_0*-restricted saddle points is* $\{(0,0,y) : y \in \mathbb{R}\}$*. Therefore the subgradient method on* K_0 *is also globally convergent.*

So in this case the subgradient method on K_a *starts non-convergent for* $a < 0$*, becomes globally convergent for* $a = 0$ *and finally looses all its equilibrium points when* $a > 0$*.*

Although the minimal face F in Theorem 4.4.2(ii) is given as the intersection of all faces that contain K-restricted saddle points, it can be useful to obtain convergence criteria that do not depend upon knowledge of all K-restricted saddle points. We note that if the subgradient method is globally convergent on any affine span of a face of K, then global convergence is implied.

Corollary 4.4.1. *Let* $K \subseteq \mathbb{R}^{n+m}$ *be non-empty, closed and convex. Let* φ *be* C^2

and concave-convex on $\mathbb{R}^{n+m}$. Let φ have a K-restricted saddle point. Assume that, for any face F of K that contains a K-restricted saddle point, the subgradient method on $\mathrm{aff}(F)$ is globally convergent. Then the subgradient method on K is globally convergent.

Example 4.4.4. *To illustrate this result, let us consider the case of positivity constraints, where (x, y) are restricted to $K = \mathbb{R}^n_+ \times \mathbb{R}^m_+$. Here the faces of K are given by sets of the form*

$$\{(x, y) \in \mathbb{R}^n_+ \times \mathbb{R}^m_+ : x_i = 0, y_j = 0 \text{ for } i \notin I, j \notin J\}$$

where $I \subseteq \{1, \ldots, n\}$ and $J \subseteq \{1, \ldots, m\}$ are sets of indices. The affine span of such a face is then given by

$$\{(x, y) \in \mathbb{R}^{n+m} : x_i = 0, y_j = 0 \text{ for } i \notin I, j \notin J\}. \tag{4.4.17}$$

Thus, by Corollary 4.4.1, checking convergence of the subgradient method in this case may be done by checking convergence of the gradient method with any arbitrary set of coordinates fixed as zero[3].

In some cases the faces of the constraint set K have an interpretation in terms of the specific problem.

Example 4.4.5. *Consider the optimisation problem*

$$\max_{g_j(x) \geq 0, j \in \{1,\ldots,m\}} U(x) \tag{4.4.18}$$

where $U, g_j : \mathbb{R}^n \to \mathbb{R}$ are concave functions in C^2. This is associated with the Lagrangian

$$\varphi(x, y) = U(x) + \sum_{j \in \{1,\ldots,m\}} y_j g_j(x) \tag{4.4.19}$$

where $y \in \mathbb{R}^m$ is a vector of Lagrange multipliers[4]*. To ensure that the Lagrange multipliers are non-negative we define the constraint set $K = \mathbb{R}^n \times \mathbb{R}^m_+$. As in Example 4.4.4 the affine spans of the faces of K are given by (4.4.17) for*

[3] This result was presented previously by the authors in [91].

[4] For simplicity of presentation we shall assume throughout the example that there is no duality gap in the problems considered.

$I = \{1, \dots, m\}$ and J any subset of $\{1, \dots, m\}$. The subgradient method applied on such a face corresponds to the gradient method on the modified Lagrangian

$$\varphi'(x, y) = U(x) + \sum_{j \in J} y_j g_j(x) \tag{4.4.20}$$

which is associated with the modified optimisation problem

$$\max_{g_j(x)=0, j \in J} U(x) \tag{4.4.21}$$

where, compared to (4.4.18), the inequality constraints are replaced by equality constraints, and some subset of the constraints are removed.

If φ is concave-convex on $\mathbb{R}^{n+m}$, which happens if, for example, the constraints $g_j(x) \geq 0$ are linear, then Corollary 4.4.1 applies. We obtain that the subgradient method on K applied to φ is globally convergent, if, for any $J \subseteq \{1, \dots, m\}$, the gradient method applied to the Lagrangian φ' corresponding to the modified optimisation problem (4.4.21) is globally convergent.

In general, φ is only concave-convex on K, so Corollary 4.4.1 does not apply. However, by applying Theorem 4.4.2 we deduce that the subgradient method is globally convergent, if, for any $J \subseteq \{1, \dots, m\}$, every solution of the gradient method applied to the Lagrangian φ' corresponding to the modified optimisation problem (4.4.21), that additionally lies in K for all times t, converges to the set of saddle points.

We note that this result is not sharp, i.e. there are optimisation problems for which the subgradient method is globally convergent, but the gradient method on one of the modified problem defined above fails to converge.

4.4.3 A general convergence criterion

By combining Theorem 4.4.2 with the results on the limiting solutions of the (smooth) subgradient method on affine subspaces given in Chapter 3 we obtain the following convergence criterion for the subgradient method on arbitrary convex sets and arbitrary concave-convex functions. This states that the subgradient method is globally convergent, if it has no trajectory satisfying an explicit linear

ODE.

The matrices $\mathbf{A}(\mathbf{z})$ and $\mathbf{B}(\mathbf{z})$ that will be used in the statement of the theorem were defined in Chapter 3. For the readers convenience we reproduce both these definitions and the statement of the result needed to prove Theorem 4.4.3 in the Section 4.A. The theorem is stated under the assumption that $\mathbf{0} \in K$ is a K-restricted saddle point. The general case is obtained by a translation of coordinates.

Theorem 4.4.3. *Let K be non-empty, closed and convex in $\mathbb{R}^{n+m}$ with $\mathbf{0} \in K$. Let $\varphi \in C^2$ be concave-convex on K and have $\mathbf{0}$ as a K-restricted saddle point. Let F be the minimal face of K that contains all K-restricted saddle points and let $\mathbf{\Pi}$ be the orthogonal projection matrix onto the orthogonal complement of N_F. Then if the subgradient method on K applied to φ has no non-constant trajectory $\mathbf{z}(t)$ that satisfies the linear ODE*

$$\dot{\mathbf{z}}(t) = \mathbf{\Pi}\mathbf{A}(\mathbf{0})\mathbf{\Pi}\mathbf{z}(t) \tag{4.4.22}$$

and the condition, for all $r \in [0,1]$ and $t\mathbb{R}$,

$$\mathbf{z}(t) \in \ker(\mathbf{\Pi}\mathbf{B}(r\mathbf{z}(t))\mathbf{\Pi}) \cap \ker(\mathbf{\Pi}(\mathbf{A}(r\mathbf{z}(t)) - \mathbf{A}(\mathbf{0}))\mathbf{\Pi}), \tag{4.4.23}$$

then the subgradient method is globally convergent.

Remark 4.4.5. *Although the condition (4.4.23) appears difficult to verify, it is only necessary to show that the condition* does not *hold. This turns out to be easy in many cases, for example in the proofs of the convergence of the modification methods (Theorem 4.5.2).*

4.5 Applications

In this section we apply the results of Section 4.4 to obtain global convergence in a number cases. First we consider the subgradient method under strict concave-convexity on arbitrary convex domains. Second we look at some examples of methods of modifying a concave-convex function to obtain convergence. Lastly we apply one such modification method to the problem of congestion control in

multi-path routing.

The proofs for this section are provided in Section 4.7.

4.5.1 Convergence under strict concave-convexity on arbitrary convex domains

The convergence of the subgradient method when applied to functions $\varphi \in C^2$ which are strictly concave-convex, (i.e. at least one of the concavity or convexity is strict), was proved by Arrow, Hurwicz and Uzawa [10] under positivity constraints. More recently, [61] and [38] revisited this result, giving more modern proofs in the case where the concave-convex function φ has the form (4.2.2) with U and g strictly concave. The more general case of restriction of a general concave-convex function to an arbitrary convex set K appears to be unknown in the literature. (The theory for discrete time subgradient methods is more complete, see e.g. [153]). Here we prove that for non-empty closed convex set K the subgradient method on K applied to a strictly concave-convex function is globally convergent.

Theorem 4.5.1. *Let $K \subseteq \mathbb{R}^{n+m}$ be non-empty, closed and convex. Let φ be C^2 and strictly concave-convex on K, and have a K-restricted saddle point. Then the subgradient method (4.3.2) on K is globally convergent.*

4.5.2 Modification methods for convergence

We will consider methods for modifying φ so that the (sub)gradient method converges to a saddle point. Such methods are used in network optimisation (see e.g. [10], [61]), where it is important to preserve the localised structure of the dynamics, which makes the use of higher order information difficult. One such method was described in the previous Chapter 3 (Section 3.5) for the gradient method. We will extend this method to the subgradient method, describe two more such methods, and then give convergence results.

4.5.2.1 Auxiliary variables method

In Chapter 3 we described a modification method for the gradient method (Section 3.5). We now recall this method and extend it to the subgradient method restricted to an arbitrary convex domain. We refer the reader to Section 3.5 for some additional discussion. In Section 4.5.3 below we give an example application of this method to the problem of multi-path congestion control.

Given a concave-convex function φ defined on a convex domain K, we define the modified concave-convex function $\varphi' : \mathbb{R}^{n'} \times K \to \mathbb{R}$ as

$$\begin{aligned} \varphi'(x', x, y) &= \varphi(x, y) + \psi(Mx - x') \\ \psi : \mathbb{R}^{n'} \to \mathbb{R}, M \in \mathbb{R}^{n' \times n} &\text{ is a constant matrix} \\ \psi \in C^2 &\text{ is strictly concave with } \max \psi(0) = 0 \end{aligned} \tag{4.5.1}$$

where x' is a set of n' auxiliary variables. We define the augmented convex domain as $K' = \mathbb{R}^{n'} \times K$. Note that the additional auxiliary variables are not restricted and are allowed to take values in the whole of $\mathbb{R}^{n'}$. Note that the $n \times n$ identity matrix always satisfies the assumptions upon M above.

Note that there is a correspondence between K-restricted saddle points of φ and K'-restricted saddle points of φ'. If $(\bar{x}, \bar{y})$ is a K-restricted saddle point of φ, then $(M\bar{x}, \bar{x}, \bar{y})$ is a K'-restricted saddle point of φ'. In the reverse direction, if $(\bar{x}', \bar{x}, \bar{y})$ is a K'-restricted saddle point of φ' then $M\bar{x} = \bar{x}'$ and $(\bar{x}, \bar{y})$ is a K-restricted saddle point of φ.

4.5.2.2 Penalty function method

For this and the next method we will assume that the concave-convex functions φ is a Lagrangian originating from a concave optimization problem (see Section 4.2.1.2). We will assume that the Lagrangian φ satisfies

$$\begin{aligned} \varphi(x, y) &= U(x) + y^T g(x) \\ C^2 \ni U : \mathbb{R}^n &\to \mathbb{R} \text{ is concave} \\ C^2 \ni g : \mathbb{R}^n &\to \mathbb{R}^m \text{ is concave.} \end{aligned} \tag{4.5.2}$$

We consider a so called penalty method (see e.g. [68]). This method adds a penalising term to the Lagrangian based directly on the constraint functions. The new Lagrangian φ' is defined by

$$\begin{aligned} \varphi'(x,y) &= \varphi(x,y) + \psi(g(x)) \\ C^2 \ni \psi : \mathbb{R}^m \to \mathbb{R} &\text{ is strictly concave with } \psi_u > 0 \\ \psi(u) = 0 &\iff u \geq 0. \end{aligned} \tag{4.5.3}$$

It is easy to see that the saddle points of φ and φ' are the same. This method, when applied (with proper choice of ψ) to distributed optimization problems, does not destroy the local nature of the gradient method, and implementation is possible with only minimal additional information transfer. In particular the additional transfer is only between neighbouring nodes.

Remark 4.5.1. *This method has been considered previously by many authors, (see [61] and the references therein), either without constraints, or with positivity constraints, i.e.* $K = \mathbb{R}^n_+ \times \mathbb{R}^m_+$. *Theorem 4.5.2 below applies to all non-empty closed convex set* $K \subseteq \mathbb{R}^{n+m}$.

4.5.2.3 Constraint modification method

We next recall a method proposed by Arrow et al.[10] and later studied in [61]. Here we instead modify the constraints to enforce strict concavity. The Lagrangian (4.5.2) is modified to become:

$$\begin{aligned} \varphi'(x,y) &= U(x) + y^T \psi(g(x)) \\ C^2 \ni U : \mathbb{R}^n \to \mathbb{R} &\text{ is concave} \\ C^2 \ni g : \mathbb{R}^n \to \mathbb{R}^m &\text{ is concave} \\ C^2 \ni \psi = [\psi^1, \ldots, \psi^m]^T : \mathbb{R}^m \to \mathbb{R}^m & \\ \psi^j(0) = 0, \psi^j_u \geq 0 &\text{ and } \psi^j_{uu} < 0 \text{ for } j = 1, \ldots m. \end{aligned} \tag{4.5.4}$$

It is clear that the saddle points of the modified and original Lagrangian will be the same. Again, the method preserves the local structure of the gradient method, requiring only minimal additional information transfer.

Remark 4.5.2. *Previous works [10],[61],[38] have proved convergence of this method with positivity constraints, i.e.* $K = \mathbb{R}^n_+ \times \mathbb{R}^m_+$. *Theorem 4.5.2 below applies to any constraint set which is a product set* $K = K_x \times K_y$ *with* $K_x \subseteq \mathbb{R}^n$, $K_y \subseteq \mathbb{R}^m$ *both non-empty closed and convex.*

4.5.2.4 Convergence results

We now give a global convergence results of the for each of the methods described above on convex domains.

Theorem 4.5.2 (Convergence of modification methods)**.** *Assume that* φ, φ' *and* K *satisfy one of the following:*

1. Auxiliary variable method: *Let* $\varphi \in C^2$ *be concave-convex on* $K \subseteq \mathbb{R}^{n+m}$ *a non-empty closed convex set. Let* φ' *and* K' *be defined by* (4.5.1) *and the text directly below it.*

2. Penalty function method: *Let* φ *have the form* (4.5.2), φ' *be defined by* (4.5.3) *and* $K \subseteq \mathbb{R}^{n+m}$ *be an arbitrary non-empty closed convex set.*

3. Constraint modification method: *Let* φ *have the form* (4.5.2), φ' *be given by* (4.5.4) *and* $K = K_x \times K_y$ *with* $K_x \subseteq \mathbb{R}^n$, $K_y \subseteq \mathbb{R}^m$ *both non-empty closed and convex.*

Then, if φ *has a* K*-restricted saddle point, then the subgradient method* (4.3.2) *on* K *applied to* φ' *is globally convergent.*

Remark 4.5.3. *None of the modification methods produce a* strictly *concave-convex function* φ'. *Because of this, these convergence results do not follow from Theorem 4.5.1 and require a detailed proof. The lack of strict concave-convexity is important in many applications where converting the problem into a strictly concave-convex problem is impractical.*

Remark 4.5.4. *Each of the convergence results in Theorem 4.5.2 is proved using Theorem 4.4.3. Despite the complexity of the non-linear and non-smooth systems of ODEs involved, the application of Theorem 4.4.3 makes the convergence easy to verify, which is evident from the simplicity of the proofs themselves.*

4.5.3 Multi-path congestion control

Multipath routing is a problem that has received considerable attention within the communications literature due to the significant advantages it can provide relative to congestion control algorithms that use single paths [124]. Nevertheless their implementation is not directly obvious as the availability of multiple routes can render the network prone to route flapping instabilities [202].

A classical approach to analyse such algorithms is to formulate them as solving a network optimization problem where aggregate user utilities are maximized subject to capacity constraints [110]. In the seminal work in [110] it was noted that when capacity constraints are relaxed with penalty functions and primal algorithms are considered, then convergence can be guaranteed despite the presence of multiple routes. In order, however, to achieve the network capacities, dual or primal/dual algorithms need to be deployed [183]. Nevertheless, when multiple routes per source/destination pair are available the corresponding optimization problem is known to be not strictly convex and the use of classical gradient dynamics can lead to unstable behaviour [199], [107], [10]. In order to address this issue various studies have considered relaxations that lead to a modified optimization problem that is strictly convex [199], [62]. This leads to algorithms with guaranteed convergence, but with the equilibrium solution deviating from that of the solution of the original optimization problem.

Here we consider a multi-path routing problem with a fixed number of routes per source/destination pair, as in [110], [199], [126], [129]. For such schemes we investigate algorithms that allow the corresponding network optimization problem to be solved without requiring any relaxation in its solution or any additional information exchange. In particular, we show that this is feasible by incorporating appropriate higher order dynamics in the local update rules.

4.5.3.1 Problem formulation

We consider a multi-path routing problem where each source/destination pair has a fixed number of routes.

In particular, we consider a network that consists of sources $s_1, \ldots, s_m$, routes

$r_1, \ldots, r_n$, and links $l_1, \ldots, l_l$. Each source s_i is associated with a unique destination for a message which is to be routed. Every route r_j has a unique source s_i, and we write $r_j \sim s_i$ to mean that s_i is the source associated with route r_j. Routes r_j each use a number of links, and we write $r_j \sim l_k$ to mean that the link l_k is used by the route r_j. The desired running capacity of the link l_k is denoted C_k, and $0 \leq C \in \mathbb{R}^l$ is the vector of these capacities. We let A be the connectivity matrix, so that $A_{kj} = 1$ if $l_k \sim r_j$ and 0 otherwise. In the same way we set $H_{ij} = 1$ if $s_i \sim r_j$ and 0 otherwise. x_j denotes the current usage of the route r_j. We associate to each source s_i a strictly concave, increasing utility function U_i.

The problem of maximising total utility over the network is stated as

$$\max_{x \geq 0, Ax \leq C} \sum_{s_i} U_i \left(\sum_{r_j \sim s_i} x_j \right). \tag{4.5.5}$$

Here the first sum is over all sources s_i, and the second over routes r_j with $r_j \sim s_i$. (We shall use such notation throughout this section.) This optimisation problem is associated with the Lagrangian

$$\varphi(x, y) = \sum_{s_i} U_i \left(\sum_{r_j \sim s_i} x_j \right) + y^T (C - Ax). \tag{4.5.6}$$

where $y \in \mathbb{R}^l_+$ are Lagrange multipliers that relax the $Ax \leq C$ constraint. A common approach in the context of congestion control is to consider primal-dual dynamics originating from this Lagrangian so as to deduce decentralized algorithms for solving the network optimisation problem (4.5.5) [110],[183]. This gives rise to the subgradient method

$$\begin{aligned} \dot{x}_j &= \left[U_i' \left(\sum_{s_i \sim r_k} x_k \right) - \sum_{l_k \sim r_j} y_k \right]^+_{x_j} \\ \dot{y}_k &= \left[\sum_{l_k \sim r_j} x_j - C_k \right]^+_{y_k} \end{aligned} \tag{4.5.7}$$

where $s_i \sim x_j$ in the equation for $\dot{x}_j$ and U_i' is the derivative of the utility function U_i. Note that the equilibrium points of (4.5.7) are saddle points of the Lagrangian

(under the positivity constraints on x and y) and hence also solutions of the optimization problem (4.5.5) (Slater's condition is assumed to hold throughout this section).

Remark 4.5.5. *The dynamics* (4.5.7) *are nothing other than the subgradient method* (4.3.2) *on the positive orthant* $\mathbb{R}^{n+l}_+$ *applied to the Lagrangian* (4.5.6).

The dynamics (4.5.7) are also localised in the sense that the update rules for x_j depend only on the current usage, x_k, of routes with the same source and of the congestion signals associated with links on these routes. In the same way the update rules for congestion signals y_k depend only on the usage of routes using the associated link.

4.5.3.2 Instability

The dynamics (4.5.7) inherit the stability properties of the subgradient method discussed in Section 4.4. In particular the distance of $(x(t), y(t))$ from any saddle point $(\bar{x}, \bar{y})$ is non-increasing. However, the lack of strict concavity of the Lagrangian (4.5.6) leads to a lack of global convergence of the dynamics (4.5.7) in some situations as we shall describe below.

To simplify the situation we shall assume that there is a strictly positive saddle point $\bar{\mathbf{z}} > 0$. In this situation Theorem 4.4.2(i) applies, and the convergence properties are the same as the unconstrained gradient method. The structure of the problem suggests an application of Theorem 3.4.2. Here a simple computation yields that $\mathcal{S}_{\text{linear}}$ is equal to $\bar{\mathcal{S}}$ (we use the notation of Chapter 3) unless the following *algebraic* condition on the network topology holds:

$$\exists u \in \ker(H) \setminus \{0\}, \lambda > 0 \text{ such that } A^T A u = \lambda u. \tag{4.5.8}$$

Theorem 3.4.2 tells us that global convergence holds if (4.5.8) does not hold, but in fact more is true.

Proposition 4.5.1. *Let* $\bar{\mathbf{z}} = (\bar{x}, \bar{y}) > 0$ *be a saddle point of* φ *defined by* (4.5.6) *and* $U_i \in C^2$ *be be strictly concave and strictly increasing. Then the dynamics* (4.5.7) *are globally convergent if and only if* (4.5.8) *does not hold.*

The algebraic criterion (4.5.8) on the network topology is satisfied by many networks, for example the network in Fig. 4.1.

We also remark that under the condition (4.5.8), the system is sensitive to noise in the sense that the unconstrained dynamics satisfy the conditions of Theorem 3.4.3.

4.5.3.3 Modified dynamics

Here we present a modification of the dynamics (4.5.7), that, while still fully localised, give guaranteed convergence to an optimal solution of (4.5.5).

We use the auxiliary variables method described in Section 4.5.2.1. We define a modified optimisation problem

$$\max_{\substack{x\geq 0, x'\in\mathbb{R}^n \\ Ax\leq C}} \sum_{s_i} U_i\left(\sum_{r_j\sim s_i} x_j\right) - \frac{1}{2}\sum_{r_k}\kappa_k|x'_k - x_k|^2 \tag{4.5.9}$$

where $x' \in \mathbb{R}^n$ is an additional vector to be optimised over, and $\kappa_k > 0$ are arbitrary constants. It is important to note that this has the same optimal x points as (4.5.5). This gives rise to a modified Lagrangian

$$\begin{aligned}\varphi'(x', x, y) = &\sum_{s_i} U_i\left(\sum_{r_j\sim s_i} x_j\right) + y^T(C - Ax) \\ &- \frac{1}{2}\sum_{r_k}\kappa_k|x'_k - x_k|^2.\end{aligned} \tag{4.5.10}$$

The new dynamics are given by the following subgradient method.

$$\begin{aligned}\dot{x}_j &= \left[U'_i\left(\sum_{s_i\sim r_k} x_k\right) - \sum_{l_k\sim r_j} y_k + \kappa_j(x'_j - x_j)\right]^+_{x_j} \\ \dot{x}'_j &= \kappa_j(x_j - x'_j) \\ \dot{y}_k &= \left[\sum_{l_k\sim r_j} x_j - C_k\right]^+_{y_k}.\end{aligned} \tag{4.5.11}$$

Remark 4.5.6. *The dynamics (4.5.11) are the subgradient method (4.3.2) on*

$\mathbb{R}^n_+ \times \mathbb{R}^n \times \mathbb{R}^l_+$ *applied to the modified Lagrangian* (4.5.10). *The Lagrangian* (4.5.10) *corresponds to* (4.5.1) *with* $\psi(z) = -|z|^2/2$ *and* M *the* $n \times n$ *identity matrix.*

It is apparent (as discussed in Section 4.5.2.1) that the equilibrium points of the modified dynamics (4.5.11) and the original dynamics (4.5.7) are in correspondence. We remark that the new dynamics are analogous to the addition of a low pass filter to the unmodified dynamics (4.5.7).

These dynamics are still localised. Each route r_k is now associated with its usage, x_k, and a new variable x'_k. To update x_k the only additional information required over the unmodified scheme is the value of x'_k, and to update x'_k one only needs x_k. Thus the new variables x'_k are local to the updaters of x_k.

Convergence of the modified dynamics to an optimum of the original problem now follows immediately from Theorem 4.5.2.1).

Theorem 4.5.3. *Let* $U_i \in C^2$ *be strictly concave and strictly increasing. Then solutions of* (4.5.11) *converge as* $t \to \infty$ *to maxima of the original problem* (4.5.5).

Remark 4.5.7. *The use of derivative action to damp oscillatory behaviour has been studied previously in the context of node based multi-path routing in [162] by incorporating derivative action in a price signal that gets communicated (i.e. a form of prediction is needed) and a local stability result was derived. This has also been used in gradient dynamics in game theory in [180]. A control scheme similar to* (4.5.11) *for multi-path routing was proposed in [129] and studied in both continuous and discrete time. In [129] the scheme differs from* (4.5.11) *in that the* x_j *variables are updated instantaneously. In our context this would be*

$$x(t) = \underset{x \geq 0, Ax \leq C}{\operatorname{argmax}} \varphi'(x'(t), x, y(t)). \tag{4.5.12}$$

4.5.3.4 Numerical results

In this section we present numerical simulations to illustrate our analytic results. We consider the two networks in Fig. 4.1 and Fig. 4.4.

In Fig. 4.2 and Fig. 4.3 we use the network in Fig. 4.1 with capacities all set to

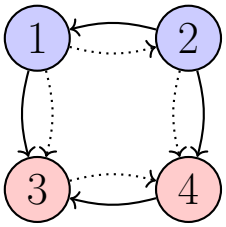


Figure 4.1: A first example network. Sources at 1 and 2 transmit to the destinations 4 and 3 respectively. Each has a choice of two routes. Routes associated with the source at 1 are dotted lines, while those associated with the source at 2 are solid lines.

1. The utility functions were chosen as $\log(1+x)$ and $1-e^{-x}$ for the sources at 1 and 2 respectively. The parameters κ_j were all set to 1. This network satisfies the condition (4.5.8) and this is apparent in the oscillating modes of the unmodified dynamics (4.5.7), shown in Fig. 4.2, that do not decay. However, when we apply the modified dynamics (4.5.11) to this network, we obtain the rapid convergence to the equilibrium shown in Fig. 4.3.

In Fig. 4.5 and Fig. 4.6 we use the network in Fig. 4.4. We take the utility function as $\log(1+x)$, and the capacities all set to 0.5. The parameters κ_j were all set to 1. On this network the original dynamics Eq. (4.5.7) converge to equilibrium, shown in Fig. 4.5, but there is transient oscillatory behaviour. When we instead implement the modified dynamics (4.5.11), we see an improved performance with more rapid convergence and damping of the oscillations.

4.6 Proofs of the main results

In this section we prove the main results of the chapter which are stated in Section 4.4.

4.6.1 Outline of the proofs

We first give a brief outline of the derivations of the results to improve the readability.

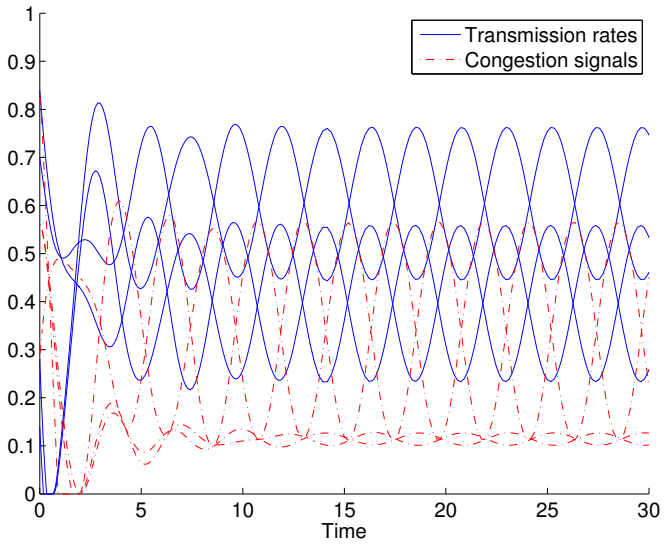


Figure 4.2: The unmodified dynamics (4.5.7) running on the network given in Fig. 4.1 with all link capacities set to 1 and the utility functions are $\log(1+x)$ and $1-e^{-x}$ for the sources at 1 and 2 respectively. In this network the condition (4.5.8) holds, and there is oscillatory behaviour which does not decay.

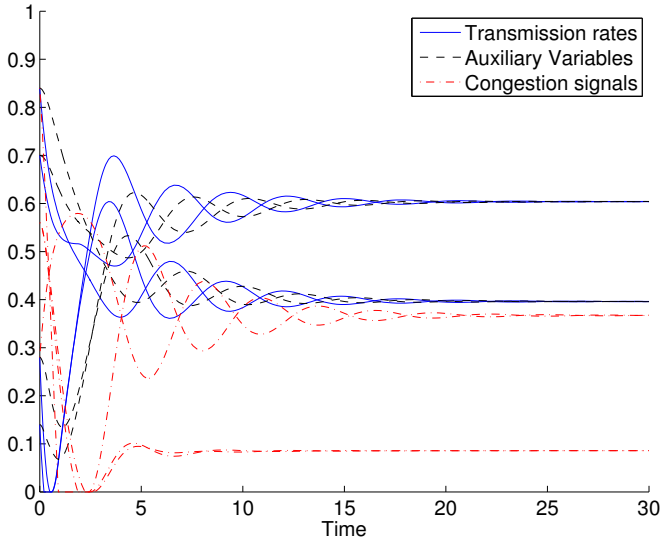


Figure 4.3: The modified dynamics (4.5.11) running on the network given in Fig. 4.1 with all link capacities set to 1, $\kappa_j = 1$ for all j. The utility functions are $\log(1+x)$ and $1-e^{-x}$ for the sources at 1 and 2 respectively. In this network the condition (4.5.8) holds, but the modification of the dynamics causes rapid convergence to equilibrium.

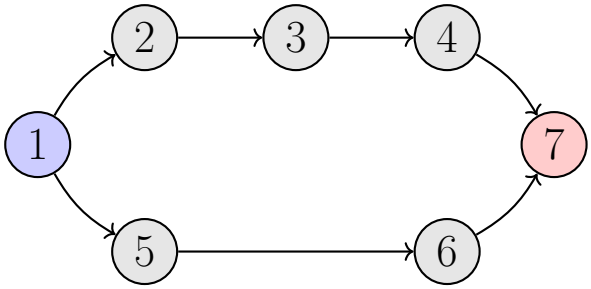


Figure 4.4: A second example network. A single source at 1 transmits to the destination 7. It has a choice of two routes.

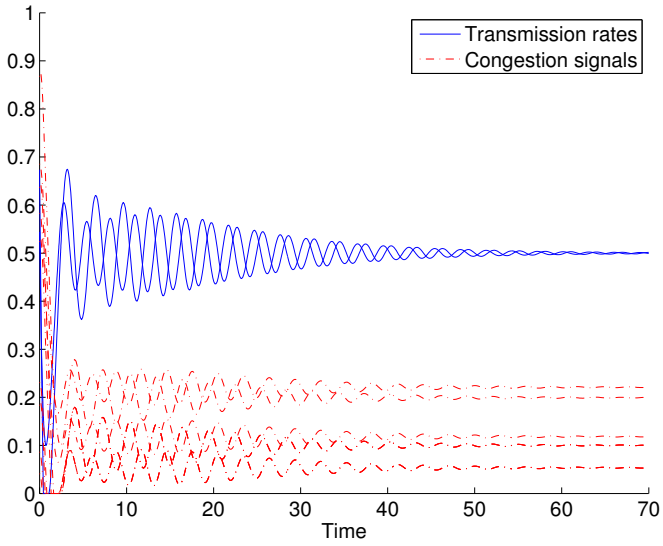


Figure 4.5: The unmodified dynamics (4.5.7) running on the network given in Fig. 4.4 with all link capacities set to 0.5 and the utility function is $\log(1+x)$. The system is asymptotically stable, but displays transient oscillatory behaviour.

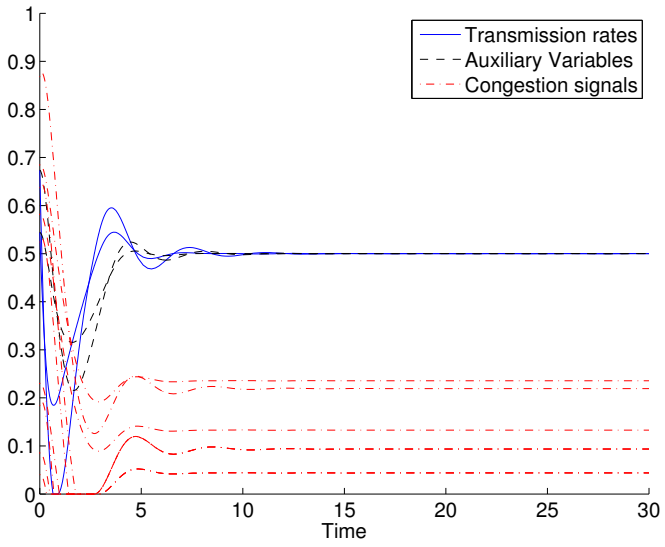


Figure 4.6: The modified dynamics (4.5.11) running on the network given in Fig. 4.4 with all link capacities set to 0.5, $\kappa_j = 1$ for all j and the utility function is $\log(1 + x)$. The oscillatory behaviour of the unmodified dynamics in Fig. 4.5 is damped, and the system rapidly converges to equilibrium.

4.6.1.1 Pathwise stability and convex projections

In Section 4.6.2 we prove the results described in Section 4.4.1.

We revisit some of the literature on topological dynamical systems [44], quoting a more general result Theorem 4.6.1, from which Proposition 4.4.1 is deduced. Then Lemma 4.4.1 is proved using the convexity of the domain K. The combination of these results allow us to prove the main result of the subsection, Theorem 4.4.1, using that the convex projection term cannot break the isometry property of the flow on the ω-limit set.

4.6.1.2 Subgradient method

In Section 4.6.3 the results of Section 4.4.2 are then deduced from those of Section 4.4.1.

4.6.2 Convergence to a flow of isometries

In this section we provide the proofs of Proposition 4.4.1, Lemma 4.4.1 and Theorem 4.4.1.

We begin by revisiting the literature on topological dynamical systems, in which a type of incremental stability is studied, and show how this leads to an invariance principle for pathwise stability.

Definition 4.6.1 (Equicontinuous semi-flow)**.** *We say that a flow (resp. semi-flow)* (ϕ, X, ρ) *is* equicontinuous *if for any* $x(0) \in X$ *and* $\varepsilon > 0$ *there is a* $\delta = \delta(x(0), \varepsilon)$ *such that if* $\rho(x'(0), x(0)) < \delta$ *then*

$$\rho(x(t), x'(t)) \leq \varepsilon \text{ for all } t \in \mathbb{R} \text{ (resp. } \mathbb{R}_+). \tag{4.6.1}$$

Remark 4.6.1. *In the control literature equicontinuity of a semi-flow would correspond to 'semi-global non-asymptotic incremental stability', but we shall keep the term equicontinuity for brevity and consistency with [44].*

Definition 4.6.2 (Uniformly almost periodic flow)**.** *We say that a flow* (ϕ, X, ρ) *is* uniformly almost periodic *if for any* $\varepsilon > 0$ *there is a* syndetic *set* $A \subseteq \mathbb{R}$*, (i.e.* $\mathbb{R} = A + B$ *for some compact set* $B \subseteq \mathbb{R}$*), for which*

$$\rho(\phi(t, x), x) \leq \varepsilon \text{ for all } t \in A, x \in X. \tag{4.6.2}$$

For the readers convenience we reproduce the results, [44, Theorem 8] and [56, Proposition 4.4.], that we will use.

Theorem 4.6.1 (G. Della Riccia [44])**.** *Let* (ϕ, X, ρ) *be an equicontinuous semi-flow and let* X *be either locally compact or complete. Let* Ω *be its* ω*-limit set. Then* (ϕ, Ω, ρ) *is an equicontinuous semi-flow of homeomorphisms of* Ω *onto* Ω*. This generates an equicontinuous flow.*

The backwards flow given by Theorem 4.6.1 is only unique on Ω, (see Remark 4.4.1 which also applies here).

Proposition 4.6.1 (R. Ellis [56])**.** *Let* (ϕ, X, ρ) *be a flow, with* X *compact. Then the following are equivalent:*

(i) The flow is equicontinuous.

(ii) The flow is uniformly almost periodic.

In our case we study pathwise stability which is a particular form of equicontinuity. We prove stronger results in this special case.

Proof of Proposition 4.4.1. By Theorem 4.6.1 (ϕ, Ω, d) is an equicontinuous flow with an equilibrium point $\bar{\mathbf{z}}$. Let $R > 0$ be arbitrary, and define

$$Y_R = \left\{ \mathbf{z}(0) \in \Omega : \sup_{t \in \mathbb{R}} d(\mathbf{z}(t), \bar{\mathbf{z}}) \leq R \right\}. \tag{4.6.3}$$

As the flow is equicontinuous, Y_R is a closed bounded subset of $\mathbb{R}^{n+m}$ and hence compact, and moreover, the union of the sets Y_R over $R \geq 0$ is Ω. By Proposition 4.6.1 the flow (ϕ, Y_R, d) is uniformly almost periodic. By pathwise stability, $d : Y_R \times Y_R \to \mathbb{R}$ is a non-increasing along the direct product flow, and is a continuous function on a compact set. Hence we have the inequality, for any two points $\mathbf{z}(0), \mathbf{z}'(0) \in Y_R$,

$$\begin{aligned} \lim_{t \to -\infty} d(\mathbf{z}(t), \mathbf{z}'(t)) &= \sup_{t \in \mathbb{R}} d(\mathbf{z}(t), \mathbf{z}'(t)) \\ &\geq \inf_{t \in \mathbb{R}} d(\mathbf{z}(t), \mathbf{z}'(t)) = \lim_{t \to \infty} d(\mathbf{z}(t), \mathbf{z}'(t)). \end{aligned} \tag{4.6.4}$$

We claim that the two limits are equal. Indeed, by uniform almost periodicity there are sequences $t_n \to \infty$ and $t'_n \to -\infty$ as $n \to \infty$ for which

$$0 = \lim_{n \to \infty} d(\mathbf{z}(t_n), \mathbf{z}(0)) = \lim_{n \to \infty} d(\mathbf{z}(t'_n), \mathbf{z}(0)) \tag{4.6.5}$$

and the analogous limits hold for $\mathbf{z}'$ for the same sequences t_n, t'_n. Hence, by continuity of d, we have

$$\lim_{t \to -\infty} d(\mathbf{z}(t), \mathbf{z}'(t)) = d(\mathbf{z}(0), \mathbf{z}'(0)) = \lim_{t \to \infty} d(\mathbf{z}(t), \mathbf{z}'(t)). \tag{4.6.6}$$

Hence $d(\mathbf{z}(t), \mathbf{z}'(t))$ is constant. By picking R big enough, this holds for any $\mathbf{z}(0), \mathbf{z}'(0) \in \Omega$, which completes the proof that the sub-semi-flow generates a flow of isometries.

It remains to show that Ω is convex. To this end let $\mathbf{z}(t), \mathbf{z}'(t)$ be two trajectories

of (ϕ, Ω, d). Let that $\lambda \in (0,1)$ and define $\mathbf{z}''(t) = \lambda \mathbf{z}(t) + (1-\lambda)\mathbf{z}'(t)$. By the same argument as used in the proof of Proposition 3.6.1 we deduce that $\mathbf{z}''(t)$ is a trajectory of the original semi-flow, but (as argued above) by uniform almost periodicity of (ϕ, Ω, d) we have a sequence of times $t_n \to \infty$ for which $d(\mathbf{z}(t_n), \mathbf{z}(0)) \to 0$ as $n \to \infty$ and the same limit for $\mathbf{z}'(t)$. Hence $d(\mathbf{z}''(t_n), \mathbf{z}''(0)) \to 0$ also, showing that $\mathbf{z}''(0)$ is in the ω-limit set. □

We now work under the set of assumptions (4.4.1) and consider projected pathwise stable differential equations.

Proof of Lemma 4.4.1. Let $\mathbf{z}(t)$ and $\mathbf{z}'(t)$ be two arbitrary solutions to the projected ODE, and define $W(t) = \frac{1}{2}|\mathbf{z}(t) - \mathbf{z}'(t)|^2$. Then W is absolutely continuous and for almost all times $t \geq 0$ we have,

$$\begin{aligned} \dot{W}(t) &= (\mathbf{z}(t) - \mathbf{z}'(t))^T (\dot{\mathbf{z}}(t) - \dot{\mathbf{z}}'(t)) \\ &= (\mathbf{z}(t) - \mathbf{z}'(t))^T (\mathbf{f}(\mathbf{z}(t)) - \mathbf{f}(\mathbf{z}'(t))) + \\ &\quad - (\mathbf{z}(t) - \mathbf{z}'(t))^T \mathbf{P}_{N_K(\mathbf{z}(t))}(\mathbf{f}(\mathbf{z}(t))) + \\ &\quad + (\mathbf{z}(t) - \mathbf{z}'(t))^T \mathbf{P}_{N_K(\mathbf{z}'(t))}(\mathbf{f}(\mathbf{z}'(t))). \end{aligned} \tag{4.6.7}$$

The first term is non-positive due to the assumption that the original ODE was pathwise stable. The other two terms are non-positive due to the definition of the normal cone. □

We now use the isometry property together with the geometry of the convex projection term to obtain the key result of this section, Theorem 4.4.1, which states that the limiting dynamics of a pathwise stable ODE restricted to a convex set K have C^1 smooth vector field and lie inside one of the faces of K.

To prove the theorem we will make use of a simple lemma on faces of convex sets.

Lemma 4.6.1. *Let $K \subseteq \mathbb{R}^n$ be non-empty closed and convex and $A \subseteq K$. Let F be the minimal face of K containing A, (see Definition 4.2.3), then* relint(F) *intersects* Conv A.

The statement of this lemma and the idea behind its proof are illustrated by Fig. 4.7.

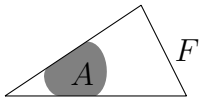


Figure 4.7: This figure illustrates the claim of Lemma 4.6.1. The triangle F is the minimal face containing the convex set A (shaded region). If A intersects two subfaces of F, then, as shown, to be convex it must also intersect the relative interior of F.

Proof. As faces are convex, the minimal face containing A is the same as the minimal face containing $\operatorname{Conv} A$. So we are free to assume without loss of generality that A is convex. Assume for a contradiction that $A \cap \operatorname{relint}(F) = \emptyset$. Define the set $\mathcal{F}$ as

$$\{C : C \text{ is a proper face of } F \text{ and } A \cap (\operatorname{relint} C) \neq \emptyset\}.$$

Note that every point in the relative boundary of F lies in the relative interior of some proper face of F by property (e) below Definition 4.2.2. This implies that $\mathcal{F}$ is not empty. Now, either there is a face C in $\mathcal{F}$ that contains all other faces in $\mathcal{F}$, or there are two faces $F_1, F_2 \in \mathcal{F}$ such that there is no face $F_3 \in \mathcal{F}$ containing both F_1 and F_2. In the first case, C is a face containing A that is strictly contained in F, contradicting minimality of F. In the second case let $x_i \in (\operatorname{relint} F_i) \cap A$ for $i = 1, 2$, (note that $x_1 \neq x_2$ by property (e) of faces), and let x_3 be some point in the open line segment between x_1 and x_2. By convexity of A, $x_3 \in A$. Hence x_3 lies in $\operatorname{relint}(F_3)$ for some face F_3, and $F_3 \in \mathcal{F}$, as otherwise x_3 would lie in $\operatorname{relint}(F)$ contradicting the assumption that $(\operatorname{relint} F) \cap A = \emptyset$. We claim that F_3 contains both F_1 and F_2, a contradiction. Indeed, first we note that $x_1, x_2 \in F_3$ by property (ii) in Definition 4.2.2 as $x_3 \in F_3$. Then, as F_i is convex and $x_i \in \operatorname{relint}(F_i)$, F_i can be written as the union of line segments which have x_i as an interior point (i.e. not an end point). But each of these line segments touches F_3 at x_i, so by Definition 4.2.2(ii) each lies entirely within F_3. □

Proof of Theorem 4.4.1. **Step 1: Identification of the limiting equation.** First, by Lemma 4.4.1 and Proposition 4.4.1 (ϕ, Ω, d) is a flow of isometries. Now let F be the minimal face that contains Ω, i.e. the intersection of all faces that contain Ω, and N_F be its normal cone. (In step 2 of the proof we will identify this face more precisely). We note that the vector field in (4.4.1) must be directed parallel to V, as otherwise trajectories would leave F, contradicting $\Omega \subseteq F$.

It is sufficient to show that if $\mathbf{z} = \mathbf{z}(0) \in \Omega$ with $\mathbf{n}(t) = \mathbf{P}_{N_K(\mathbf{z}(t))}(\mathbf{f}(\mathbf{z}(t)))$ then $\mathbf{n}(t)$ is orthogonal to F. If $\mathbf{z}(t) \in \operatorname{relint} K$ then $N_K(\mathbf{z}(t)) = N_F$ and the orthogonality holds. Otherwise $\mathbf{z}(t)$ lies in the relative boundary of F.

As each solution of the differential equation (4.4.1) holds only for almost all times t and we wish to consider an uncountably infinite family of solutions, we run the risk of taking an uncountable union of sets of measure zero, (which does not necessarily have zero measure). Avoiding this makes the proof technical. To better communicate the idea of the proof, we shall first give the proof that would work if the differential equations held for all times t.

Step 1.1: Heuristic (unrigorous) proof.

Let $C = \operatorname{Conv} \Omega$, then, by the definition of a face, $\Omega \subseteq F$ implies that $C \subseteq F$. From Lemma 4.6.1 and the minimality of F we deduce that C must intersect $\operatorname{relint} F$. Thus there are $\mathbf{x}(0), \mathbf{y}(0) \in \Omega$ and $\lambda \in (0,1)$ with $\mathbf{w} = \lambda \mathbf{x}(0) + (1-\lambda)\mathbf{y}(0) \in \operatorname{relint} F$. Set $W = \frac{1}{2}|\mathbf{x}(t) - \mathbf{z}(t)|^2$. By the isometry property of the flow we know that $\dot{W} = 0$ at t. Following the computation (4.6.7) in the proof of Lemma 4.4.1 we deduce that $(\mathbf{x} - \mathbf{z})^T \mathbf{n} = 0$. Similarly we obtain $(\mathbf{y} - \mathbf{z})^T \mathbf{n} = 0$. Taking a convex combination of these equalities, we obtain

$$(\mathbf{w} - \mathbf{z})^T \mathbf{n} = \lambda (\mathbf{x} - \mathbf{z})^T \mathbf{n} + (1-\lambda)(\mathbf{y} - \mathbf{z})^T \mathbf{n} = 0 + 0 = 0 \tag{4.6.8}$$

and as $\mathbf{w}$ is in the relative interior of F this implies that $\mathbf{n}$ is orthogonal to F.

Step 1.2: Rigorous proof. We now give the fully rigorous proof. We must show that the set of times t when $\mathbf{n}(t)$ is not orthogonal to F is of measure zero. Let Ω' be a countable dense subset of Ω that contains $\mathbf{z}(0)$. By invariance of Ω under the flow ϕ, the set $\phi(t, \Omega') = \{\phi(t, \mathbf{x}) : \mathbf{x} \in \Omega'\}$ is also dense in Ω for any $t \in \mathbb{R}$. Then the set

$$\begin{aligned} A = \{t \in [0, \infty) : \exists \mathbf{x}(0) \in \Omega' \text{ such that} \\ \dot{\mathbf{x}}(t) \neq \mathbf{f}(\mathbf{x}(t)) - \mathbf{P}_{N_K(\mathbf{x}(t))}(\mathbf{f}(\mathbf{x}(t)))\} \end{aligned} \tag{4.6.9}$$

is the countable union of measure zero sets, and is hence of measure zero. From the isometry property and by considering $W(t) = \frac{1}{2}|\mathbf{x}(t) - \mathbf{z}(t)|^2$ with $\mathbf{x}(0) \in \Omega'$, it follows that $(\mathbf{x}(t) - \mathbf{z}(t))^T \mathbf{n}(t) = 0$ for all $\mathbf{x}(0) \in \Omega'$ and $t \in [0, \infty) \setminus A$. Thus,

for $t \in [0,\infty) \setminus A$, $(\mathbf{x} - \mathbf{z}(t))^T \mathbf{n}(t) = 0$ for all $\mathbf{x}$ in a dense subset of Ω, and hence for any $\mathbf{x} \in \Omega$. The proof now follows as step 1.1. above.

Step 2: Identification of the limiting face. Finally we will show that the face F defined above is in fact the minimal face F' containing the equilibrium points of the semi-flow (ϕ, K, d). We argue by contradiction. If $F \neq F'$ then there must be some trajectory $\mathbf{z}(t)$ in Ω and a time t_0 with $\mathbf{z}(t_0) \in F \setminus F'$. For $T > 0$ we define $\mathbf{z}(t;T) = \frac{1}{2T}\int_{-T}^{T} \mathbf{z}(t+s)\,ds$. For any finite T this is a convex combination of trajectories in Ω, and as Ω is convex by Proposition 4.4.1, $t \mapsto \mathbf{z}(t;T)$ is a trajectory in Ω. Next, as the semi-flow is uniformly almost periodic due to Proposition 4.6.1 the trajectory $\mathbf{z}(t)$ is an almost periodic function. Therefore, the limit $T \to \infty$ of $\mathbf{z}(t;T)$ exists (see e.g. [56]), and this limit is clearly a constant ($\mathbf{z}'$ say) independent of t. As Ω is closed, $\mathbf{z}' \in \Omega$ and being a constant, is an equilibrium point of the semi-flow.

To obtain a contradiction we argue that $\mathbf{z}' \notin F'$ which is impossible as F' contains all equilibrium points. Indeed, this follows as the trajectory $\mathbf{z}(t)$, being almost periodic and passing through $\mathbf{z}(t_0) \in F \setminus F'$ spends a positive proportion of its time in $F \setminus F'$. Therefore, there is a $\delta > 0$ such that for any sufficiently large T, the average $\mathbf{z}(t;T)$ satisfies $d(\mathbf{z}(t;T), F) \geq \delta$ and this property carries over to the limit $\mathbf{z}'$. □

4.6.3 Subgradient method

In this section we give the proofs of the results of Section 4.4.2.

Proof of Theorem 4.4.2. We apply Theorem 4.4.1, noting the pathwise stability of the gradient method (incrementally-stable-fullspace). Let F be the minimal face given by Theorem 4.4.1. There are two cases.

Case 1. $\bar{\mathcal{S}} \cap \operatorname{int} K$ is non-empty. Then, as F must contain all K-restricted saddle points, it must contain a point in the interior of K. The only such face is K itself whose affine span is $\mathbb{R}^{n+m}$ (as K has non-empty interior) which has normal cone $\{\mathbf{0}\}$. Thus we are in case (i), (4.4.2) is the gradient method (4.3.1) and (4.4.3) holds.

Case 2. $\bar{\mathcal{S}} \cap \operatorname{int} K$ is empty. We are in case (ii) of the theorem. The claims of (ii) follow directly from Theorem 4.4.1. □

4.6.4 A general convergence criterion

In this section we give the proofs of Section 4.4.3.

Proof of Theorem 4.4.3. By Theorem 4.4.2(i) and (ii) any solution $\mathbf{z}(t)$ in the ω-limit set of the subgradient method on K solves (4.4.4). By using $\mathbf{\Pi}$, the orthogonal projection matrix onto the orthogonal complement of N_V, the ODE (4.4.4) can be written as (4.A.2). Thus, by Theorem 4.A.1 (in the Section 4.A), $\mathbf{z}(t)$ satisfies (4.4.22) and (4.4.23) for all $t \in \mathbb{R}$ and $r \in [0,1]$. Therefore, if there are no non-constant trajectories of the subgradient method on K satisfying these conditions then the ω-limit set consists only of equilibrium points and the subgradient method on K is globally convergent. □

4.7 Proofs of the examples

In this section we provide the proofs of the results presented in Section 4.5.

4.7.1 Convergence under strict concave-convexity on arbitrary convex domains

Proof of Theorem 4.5.1. We adapt the reasoning in Example 3.4.1, using instead Theorem 4.4.3. We consider the strictly concave case. The strictly convex case is the same, but switching the roles of x and y. By translation of coordinates we may assume that $\mathbf{0}$ is a K-restricted saddle point. Let $V, F, \mathbf{\Pi}$ be as in Theorem 4.4.3, and let $\mathbf{z}(t) = (x(t), y(t))$ be a trajectory of the subgradient method on K satisfying (4.4.22) and (4.4.23) for all $t \in \mathbb{R}$ and $r \in [0,1]$.

Step 1: $x(t) = 0$. $\mathbf{z}(t)$ lies in V for all times t, which implies that $\mathbf{\Pi}\mathbf{z}(t) = \mathbf{z}(t)$.

(4.4.23) implies that, for all $r \in [0,1]$ and $t \in \mathbb{R}$,

$$x(t)^T \varphi_{xx}(r\mathbf{z}(t))x(t) - y(t)^T \varphi_{yy}(r\mathbf{z}(t))y(t) = 0.$$

By the concavity and convexity of φ_{xx} and φ_{yy} respectively we deduce that

$$x(t)^T \varphi_{xx}(r\mathbf{z}(t))x(t) = 0.$$

Strict concavity of φ implies that φ_{xx} is of full rank except at isolated points. Thus, by varying $r \in [0,1]$ we deduce that $x(t) = 0$.

Step 2: $y(t)$ is constant.

Let $\mathbf{\Pi}$ be decomposed on $\mathbb{R}^n \times \mathbb{R}^m$ as

$$\mathbf{\Pi} = \begin{bmatrix} \Pi_{11} & \Pi_{12} \\ \Pi_{21} & \Pi_{22} \end{bmatrix}. \tag{4.7.1}$$

Then (4.4.22) (for $x(t) = 0$) is

$$\dot{y} = \Pi_{21} \varphi_{xy}(\mathbf{0}) y.$$

The relation $\mathbf{\Pi}\dot{\mathbf{z}} = \dot{\mathbf{z}}$ then implies that

$$\Pi_{11}\dot{x} + \Pi_{12}\dot{y} = \dot{x}$$

and as $\dot{x} = 0$ we deduce that $\Pi_{12}\Pi_{21}\varphi_{xy}(\mathbf{0})y = 0$ and hence that $\Pi_{21}\varphi_{xy}(\mathbf{0})y = 0$, i.e. $\dot{y} = 0$. Therefore all limiting solutions of the subgradient method on K are equilibrium points and the subgradient method on K is globally convergent. □

4.7.2 Modification methods

We consider each method in turn. Theorem 4.5.2 may then be obtained by combining all the results proved in each subsection below.

4.7.2.1 Auxiliary variables method

The majority of the proof for the auxiliary variables method was completed in Chapter 3. We provide the remainder below.

Proof of Theorem 4.5.2.1). It was proved in Proposition 3.7.1 that for any affine subspace V containing a V-restricted saddle point the subgradient method applied to φ' is globally convergent. By Theorem 4.4.2 this implies global convergence of the subgradient method for any closed convex K. □

4.7.2.2 Penalty function method

The proof below shows that this method gives convergence by adding just enough strict convexity to eliminate the oscillations caused by the matrix $\mathbf{A}(\mathbf{0})$.

Proposition 4.7.1. *Let $K \subseteq \mathbb{R}^{n+m}$ be non-empty closed and convex. Let (4.5.2), (4.5.3) hold, and assume that there exists a K-restricted saddle point. Then the subgradient method (4.3.2) on K is globally convergent.*

Proof. By the change of variables

$$\begin{aligned}\tilde{\varphi}'(x,y) &= \varphi'(x+\bar{x}, y+\bar{y}) \\ &= U(x+\bar{x}) + \bar{y}^T g(x+\bar{x}) + y^T g(x+\bar{x}) + \psi(g(x+\bar{x})) \\ &= \tilde{U}(x) + y^T \tilde{g}(x) + \psi(\tilde{g}(x))\end{aligned}$$

for a K-restricted saddle point $(\bar{x}, \bar{y})$ we may assume that $\mathbf{0}$ is a K-restricted saddle point. We apply Theorem 4.4.3 and let $F, V, \mathbf{\Pi}$ be as in Theorem 4.4.3 and $\mathbf{z}(t) = (x(t), y(t))$ be a trajectory of the subgradient method on K satisfying (4.4.22) and (4.4.23) for all $t \in \mathbb{R}$ and $r \in [0,1]$.. Define $(\tilde{x}(t), \tilde{y}(t)) = \tilde{\mathbf{z}}(t) = \mathbf{\Pi z}(t)$. We compute that

$$\mathbf{A}(\mathbf{0}) = \begin{bmatrix} 0 & g_x(0)^T \\ -g_x(0) & 0 \end{bmatrix}. \tag{4.7.2}$$

Step 1: $g_x(0)\tilde{x}(t) = 0$**.**

The condition (4.4.23) implies that the following expression is zero for all $s \in [0,1]$,

$$\tilde{\mathbf{z}}^T \mathbf{B}(s\mathbf{z})\mathbf{w} = \tilde{x}^T \varphi_{xx} \tilde{x} + [g_x \tilde{x}]^T \psi_{uu} [g_x \tilde{x}] + \psi_u (\tilde{x}^T g_{xx} \tilde{x}) \tag{4.7.3}$$

where φ_{xx} is evaluated at $s\mathbf{z}$, with g_x, g_{xx} at sx, and ψ_{uu}, ψ_{u^k} at $u = g(sx)$, and where $x^T g_{xx} x$ is the vector with ith component $x^T g^i_{xx} x$ where $g = [g^1, \dots, g^m]^T$. All the terms are non-positive by the assumptions on ψ and φ. Strict concavity of ψ and that (4.7.3) vanishes for all $s \in [0,1]$ implies that $g_x(sx)\tilde{x} = 0$ for all $s \in [0,1]$. In particular $g_x(0)\tilde{x}(t) = 0$.

Step 2: $\tilde{x}(t)$ is constant.

Let $\mathbf{\Pi}$ be decomposed as in (4.7.1). Then $\tilde{x}, \tilde{y}$ satisfy

$$\dot{\tilde{x}} = \Pi_{11} g_x(0)^T \tilde{y} \qquad\qquad \dot{\tilde{y}} = -\Pi_{21} g_x(0)^T \tilde{y}.$$

Taking the time derivative of $g_x(0)\tilde{x} = 0$ we obtain $g_x(0)\Pi_{11} g_x(0)^T \tilde{y} = 0$. As Π_{11} is positive semi-definite, $\ker(g_x(0)\Pi_{11} g_x(0)^T) = \ker(\Pi_{11} g_x(0)^T)$, and hence $\dot{\tilde{x}} = \Pi_{11} g_x(0)^T \tilde{y} = 0$ and $\tilde{x}(t)$ is constant.

Step 3: $\tilde{y}(t)$ is constant. The relation $\mathbf{\Pi}\dot{\tilde{\mathbf{z}}} = \dot{\tilde{\mathbf{z}}}$ implies that $\Pi_{11}\dot{\tilde{x}} + P_{12}\dot{\tilde{y}} = \dot{\tilde{x}} = 0$ and $0 = \Pi_{12}\dot{\tilde{y}} = -\Pi_{12}\Pi_{21} g_x(0)^T \tilde{y}$. Therefore, again, as $\Pi_{12}\Pi_{21}$ is positive semi-definite we have $\tilde{y}^T g_x(0)\Pi_{12}\Pi_{21} g_x(0)^T \tilde{y} = 0$ and $\Pi_{21} g_x(0)^T \tilde{y} = 0 = -\dot{\tilde{y}}$, which implies $\tilde{y}$ is constant. □

4.7.2.3 Constraint modification method

We first consider the case without constraints. The proof below shows that the method works by disrupting the linear structure of the oscillating solutions by changing $\mathbf{A}(\mathbf{z})$ to ensure it is not equal to $\mathbf{A}(\mathbf{0})$, (where $\mathbf{0}$ is a saddle).

Proposition 4.7.2. *Let (4.5.4) hold and $\bar{\mathcal{S}} \neq \emptyset$. Then $\mathcal{S} = \bar{\mathcal{S}}$ and the gradient method (4.3.1) is globally convergent.*

Proof. Without loss of generality we may assume that $\mathbf{0}$ is a saddle point of φ. We use the classification of $\mathcal{S}$ given by Theorem 3.4.2 and use the notation therein.

We first compute,

$$\mathbf{A}(\mathbf{z}) = \begin{bmatrix} 0 & (\psi_g g_x)^T \\ -\psi_g g_x & 0 \end{bmatrix}. \tag{4.7.4}$$

Let $\mathbf{z}(t) = (x(t), y(t)) \in \mathcal{S}_{\text{linear}}$ then we have

$$0 = \frac{d}{ds}[(\psi_g^i(g(sx))^T g_x(sx)x]_{s=0} \text{ for } i = 1, \dots, m \tag{4.7.5}$$

Then by applying the chain rule we obtain

$$0 = [g_x(0)x]^T \psi_{gg}^i(0)[g_x(0)x] + \psi_g^i(0)^T (x^T g_{xx}(0)x), \tag{4.7.6}$$

where $x^T g_{xx}(0)x$ is the vector with components $x^T g_{xx}^i x$ where $g = [g^1, \dots g^m]^T$. All the terms are non-positive due to the assumptions on ψ and g. As $\psi_{gg}^i < 0$ we have $g_x(0)x = 0$. Hence $\dot{y} = 0$ and therefore $\dot{x}$ is constant. As $|x|^2 + |y|^2$ is also constant this means that $\dot{x}$ is zero. Therefore the $\mathcal{S}_{\text{linear}} = \bar{\mathcal{S}}$ and the gradient method is globally convergent. □

Now we extend the stability to the subgradient method on sets which have a product structure. Due to Corollary 4.4.1 this is essentially an exercise in algebra.

Corollary 4.7.1. *Let $K = K_x \times K_y$ for $K_x \subseteq \mathbb{R}^n$ and $K_y \subseteq \mathbb{R}^m$ non-empty closed and convex. Let (4.5.4) hold and there be a K-restricted saddle point. Then the subgradient method (4.3.2) on K is globally convergent.*

Proof. By Corollary 4.4.1 it suffices to prove that the subgradient method converges on $\text{aff}(F)$ where F is an arbitrary face of K that contains a K-restricted saddle point $\bar{\mathbf{z}}$. By translation of coordinates we may assume that $\bar{\mathbf{z}} = \mathbf{0}$. By the product structure of K, $V = \text{aff}(F)$ must also decompose into $V = V_x \times V_y$ with $V_x \subseteq \mathbb{R}^n$ and $V_y \subseteq \mathbb{R}^m$ affine subspaces. Let the orthogonal projection matrices onto V_x, V_y, which exist as $(0,0) \in V_x \times V_y$, be P, Q respectively. Then the subgradient method on V, satisfies, for $(x, y) \in V$,

$$\dot{x} = P\varphi_x = \varphi_x^V, \quad \dot{y} = -Q\varphi_y = -\varphi_y^V \tag{4.7.7}$$

where $\varphi^V(x, y) := \varphi(Px, Qy)$. By a rotation of coordinate bases we may assume

that $V_x = \mathbb{R}^{n'} \times \{0\}$ and $V_y = \mathbb{R}^{m'} \times \{0\}$ for some $n' \leq n$ and $m' \leq m$. Then $\varphi^V : \mathbb{R}^{n'} \times \mathbb{R}^{m'} \to \mathbb{R}$ is of the form (4.5.4) and Proposition 4.7.2 gives convergence. □

4.7.3 Multi-path congestion control

Proof of Proposition 4.5.1. The *if* claim follows directly from the discussion above. For the *only if* we explicitly construct a trajectory that does not converge. Let u satisfy (4.5.8), then it can be directly verified that

$$\mathbf{z}(t) = \bar{\mathbf{z}} + ce^{t\mathbf{A}(\bar{\mathbf{z}})} \begin{bmatrix} u \\ -Au \end{bmatrix}$$

is a solution (for any $c > 0$) of the unconstrained gradient method (4.3.1) applied to φ. By taking c small enough using that $\bar{\mathbf{z}} > 0$ (and skew-symmetry of $\mathbf{A}(\bar{\mathbf{z}})$) we can ensure that $\mathbf{z}(t) > 0$ for all $t \in \mathbb{R}$, and hence $\mathbf{z}(t)$ is also a solution of the subgradient dynamics (4.5.7). □

4.A Appendix

We now recall a result proved in the previous Chapter 3 on the limiting solutions of the subgradient method on affine subspaces. To present this result we recall from Chapter 3 the definition of the following matrices of partial derivatives of φ.

$$\mathbf{A}(\mathbf{z}) = \begin{bmatrix} 0 & \varphi_{xy}(\mathbf{z}) \\ -\varphi_{yx}(\mathbf{z}) & 0 \end{bmatrix}, \qquad \mathbf{B}(\mathbf{z}) = \begin{bmatrix} \varphi_{xx}(\mathbf{z}) & 0 \\ 0 & -\varphi_{yy}(\mathbf{z}) \end{bmatrix}. \tag{4.A.1}$$

Consider the ODE (4.4.4) in more detail. Let $\mathbf{\Pi} \in \mathbb{R}^{(n+m)^2}$ be the orthogonal projection matrix onto the orthogonal complement of N_V. Then the ODE (4.4.4) can be written as

$$\dot{\mathbf{z}} = \mathbf{\Pi}\mathbf{f}(\mathbf{z}) \tag{4.A.2}$$

where $\mathbf{f}(\mathbf{z}) = \begin{bmatrix} \varphi_x & -\varphi_y \end{bmatrix}^T$. The result is stated for $\mathbf{0}$ being an equilibrium point; the general case may be obtained by a translation of coordinates.

Theorem 4.A.1. *[Theorem 3.4.5] Let* $\mathbf{\Pi} \in \mathbb{R}^{(n+m)^2}$ *be an orthogonal projection matrix,* φ *be* C^2 *and concave-convex on* $\mathbb{R}^{n+m}$*, and* $\mathbf{0}$ *be an equilibrium point of* (4.A.2)*. Then the trajectories* $\mathbf{z}(t)$ *of* (4.A.2) *that lie a constant distance from any equilibrium point of* (4.A.2) *are exactly the solutions to the linear ODE:*

$$\dot{\mathbf{z}}(t) = \mathbf{\Pi}\mathbf{A}(\mathbf{0})\mathbf{\Pi}\mathbf{z}(t)$$

that satisfy, for all $t \in \mathbb{R}$ *and* $r \in [0, 1]$*, the condition*

$$\mathbf{z}(t) \in \ker(\mathbf{\Pi}\mathbf{B}(r\mathbf{z}(t))\mathbf{\Pi}) \cap \ker(\mathbf{\Pi}(\mathbf{A}(r\mathbf{z}(t)) - \mathbf{A}(\mathbf{0}))\mathbf{\Pi})$$

where $\mathbf{A}(\mathbf{z})$ *and* $\mathbf{B}(\mathbf{z})$ *are defined by* (4.A.1)*.*

Remark 4.A.1. *As discussed in Chapter 3, this result can be localised for when* φ *is not concave-convex on the whole of* $\mathbb{R}^{n+m}$*. In particular trajectories in the* ω*-limit set of the subgradient method given by Theorem 4.4.2(ii) satisfy the conditions given in Theorem 4.A.1.*

Chapter 5

Approximations of strongly continuous families of unbounded self-adjoint operators

The problem of approximating the discrete spectra of families of self-adjoint operators that are merely strongly continuous is addressed. It is well-known that the spectrum need not vary continuously (as a set) under strong perturbations. However, it is shown that under an additional compactness assumption the spectrum does vary continuously, and a family of symmetric finite-dimensional approximations is constructed. An important feature of these approximations is that they are valid for the entire family uniformly. An application of this result to the study of plasma instabilities is illustrated.

Acknowledgements

The work in this chapter was done in collaboration with Jonathan Ben-Artzi and appears in a similar form in [18].

5.1 Introduction

5.1.1 Overview

We present a method for obtaining finite-dimensional approximations of the discrete spectrum of families of self-adjoint operators. We are interested in operators that decompose into a system of two coupled Schrödinger operators with opposite signs (see (5.1.1) below). However our results are applicable to "standard" Schrödinger operators, and in fact we prove our main result, Theorem 5.1.1, for Schrödinger operators first, see Theorem 3′. We are interested in the following problem:

Problem 5.1.1. *Consider the family of self-adjoint unbounded operators*

$$\mathcal{M}^\lambda = \mathcal{A} + \mathcal{K}^\lambda = \begin{bmatrix} -\Delta + 1 & 0 \\ 0 & \Delta - 1 \end{bmatrix} + \begin{bmatrix} \mathcal{K}^\lambda_{++} & \mathcal{K}^\lambda_{+-} \\ \mathcal{K}^\lambda_{-+} & \mathcal{K}^\lambda_{--} \end{bmatrix}, \quad \lambda \in [0,1] \tag{5.1.1}$$

acting in an appropriate subspace of $L^2(\mathbb{R}^d) \oplus L^2(\mathbb{R}^d)$*, where* $\{\mathcal{K}^\lambda\}_{\lambda\in[0,1]}$ *is a bounded, symmetric and strongly continuous family.* Is it possible to construct explicit finite-dimensional self-adjoint approximations of $\mathcal{M}^\lambda$ whose spectrum in compact subsets of $(-1,1)$ converges to that of $\mathcal{M}^\lambda$ uniformly in λ?

This problem is motivated by Maxwell's equations, which in the Lorenz gauge may be written as the following elliptic system for the electromagnetic potentials ϕ and $\mathbf{A}$ (after taking a Laplace transform in time):

$$\begin{cases} (-\boldsymbol{\Delta} + \lambda^2)\mathbf{A} + \mathbf{j} = \mathbf{0} \\ (\Delta - \lambda^2)\phi + \rho = 0 \end{cases} \tag{5.1.2}$$

where ρ and $\mathbf{j}$ are the charge and current densities, respectively. The specific problem we have in mind, treated separately in Chapter 6, is that of instabilities of the relativistic Vlasov-Maxwell system describing the evolution of collisionless plasmas and it is outlined in Section 5.6 below. The Vlasov equation provides the coupling of the two equations in (5.1.2), making the system self-adjoint (see, for instance, the expressions (5.6.5) and (5.6.6)).

5.1.2 The main result

Let us first summarise the notation we use throughout this article. For operators we use upper case calligraphic letters, such as $\mathcal{T}$. The spectrum of $\mathcal{T}$ is denoted $\mathrm{sp}(\mathcal{T})$. For the sesquilinear form associated to an operator we use the same letter in lower case Fraktur font. Hence the operator $\mathcal{T}$ has the associated form $\mathfrak{t}$. The space of bounded linear operators on a Hilbert space $\mathfrak{H}$ is denoted $\mathfrak{B}(\mathfrak{H})$. Domains of operators or forms are denoted by $\mathfrak{D}$. The graph norms of an operator $\mathcal{T}$ and a form $\mathfrak{t}$ are denoted $\|\cdot\|_{\mathcal{T}}$ and $\|\cdot\|_{\mathfrak{t}}$, respectively. Strong, strong resolvent and norm resolvent convergence are denoted by $\xrightarrow{s}$, $\xrightarrow{s.r.}$ and $\xrightarrow{n.r.}$, respectively. For brevity, we denote $\overline{\mathbb{N}} = \mathbb{N} \cup \{\infty\}$. We also recall the definition of a sectorial form:

Definition 5.1.1. *A form $\mathfrak{t}$ is said to be* sectorial *if its numerical range $\Theta(\mathfrak{t})$ (that is, the set $\{\mathfrak{t}[u,u] : \|u\| = 1,\, u \in \mathfrak{D}(\mathfrak{t})\} \subseteq \mathbb{C}$) is a subset of a sector of the form*

$$\{\zeta : |\arg(\zeta - \gamma)| \leq \theta\}, \qquad \theta \in [0, \pi/2), \quad \gamma \in \mathbb{R}.$$

Let $\mathfrak{H} = \mathfrak{H}_+ \oplus \mathfrak{H}_-$ be a (separable) Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\|$ and let

$$\mathcal{A}^\lambda = \begin{bmatrix} \mathcal{A}^\lambda_+ & 0 \\ 0 & -\mathcal{A}^\lambda_- \end{bmatrix} \quad \text{and} \quad \mathcal{K}^\lambda = \begin{bmatrix} \mathcal{K}^\lambda_{++} & \mathcal{K}^\lambda_{+-} \\ \mathcal{K}^\lambda_{-+} & \mathcal{K}^\lambda_{--} \end{bmatrix}, \quad \lambda \in [0,1]$$

be two families of operators on $\mathfrak{H}$ depending upon the parameter $\lambda \in [0,1]$, where the family $\mathcal{A}^\lambda$ is also assumed to be defined for λ in an open neighbourhood D of $[0,1]$ in the complex plane. The two families $\mathcal{A}^\lambda$ and $\mathcal{K}^\lambda$ satisfy:

i) Sectoriality: The families $\{\mathcal{A}^\lambda_\pm\}_{\lambda \in D}$ are *holomorphic of type (B)*[1]. That is, they are families of sectorial operators and the associated sesquilinear forms $\mathfrak{a}^\lambda_\pm$ are *holomorphic of type (a)*: all $\{\mathfrak{a}^\lambda_\pm\}_{\lambda \in D}$ are sectorial and closed, with domains that are independent of λ and dense in $\mathfrak{H}_\pm$,[2] and $D \ni \lambda \mapsto \mathfrak{a}^\lambda_\pm[u,v]$ are holomorphic for any $u, v \in \mathfrak{D}(\mathfrak{a}^\lambda_\pm)$. Furthermore, we assume that $\mathcal{A}^\lambda_\pm$ are self-adjoint for $\lambda \in [0,1]$.

[1] We adopt the terminology of Kato [109].

[2] Hence we shall remove the λ superscript when discussing the domains of $\mathfrak{a}^\lambda$ and $\mathfrak{a}^\lambda_\pm$.

ii) Gap: $\mathcal{A}_{\pm}^{\lambda} > 1$ for every $\lambda \in [0,1]$.

iii) Bounded perturbation: $\{\mathcal{K}^{\lambda}\}_{\lambda\in[0,1]} \subset \mathfrak{B}(\mathfrak{H})$ is a self-adjoint strongly continuous family.

iv) Compactness: There exist self-adjoint operators $\mathcal{P}_{\pm} \in \mathfrak{B}(\mathfrak{H}_{\pm})$ which are relatively compact with respect to $\mathcal{A}_{\pm}^{\lambda}$, satisfying $\mathcal{K}^{\lambda} = \mathcal{K}^{\lambda}\mathcal{P}$ for all $\lambda \in [0,1]$ where

$$\mathcal{P} = \begin{bmatrix} \mathcal{P}_{+} & 0 \\ 0 & \mathcal{P}_{-} \end{bmatrix}.$$

Finally, if the family $\mathcal{A}^{\lambda}$ does not have a compact resolvent we assume:

v) Compactification of the resolvent: There exist holomorphic forms $\{\mathfrak{w}_{\pm}^{\lambda}\}_{\lambda\in D}$ of type (a) and associated operators $\{\mathcal{W}_{\pm}^{\lambda}\}_{\lambda\in D}$ of type (B) such that for $\lambda \in [0,1]$, $\mathcal{W}_{\pm}^{\lambda}$ are self-adjoint and non-negative. Define

$$\mathcal{W}^{\lambda} = \begin{bmatrix} \mathcal{W}_{+}^{\lambda} & 0 \\ 0 & -\mathcal{W}_{-}^{\lambda} \end{bmatrix}, \quad \lambda \in D,$$

and

$$\mathcal{A}_{\varepsilon}^{\lambda} := \mathcal{A}^{\lambda} + \varepsilon\mathcal{W}^{\lambda}, \quad \lambda \in D,\ \varepsilon \geq 0 \tag{5.1.3}$$

with respective associated forms $\mathfrak{w}^{\lambda}$ and $\mathfrak{a}_{\varepsilon}^{\lambda}$. Then we assume that $\mathfrak{D}(\mathfrak{w}^{\lambda}) \cap \mathfrak{D}(\mathfrak{a})$ are dense for all $\lambda \in D$ and the inclusion $(\mathfrak{D}(\mathfrak{w}^{\lambda}) \cap \mathfrak{D}(\mathfrak{a}), \|\cdot\|_{\mathfrak{a}_{\varepsilon}^{\lambda}}) \to (\mathfrak{H}, \|\cdot\|)$ is compact for some $\lambda \in D$ and all $\varepsilon > 0$.

Goal. Define the family of (unbounded) operators $\{\mathcal{M}^{\lambda}\}_{\lambda\in[0,1]}$, acting in $\mathfrak{H}$, as

$$\mathcal{M}^{\lambda} = \mathcal{A}^{\lambda} + \mathcal{K}^{\lambda}, \quad \lambda \in [0,1]. \tag{5.1.4}$$

It is these operators that we wish to approximate.

The Projections. Let $\mathcal{A}_\varepsilon^\lambda$ be as in (5.1.3), and define

$$\mathcal{M}_\varepsilon^\lambda = \mathcal{A}_\varepsilon^\lambda + \mathcal{K}^\lambda, \quad \lambda \in [0,1]. \tag{5.1.5}$$

Let

- $\{e_{\varepsilon,k}^\lambda\}_{k\in\mathbb{N}} \subset \mathfrak{H}$ be a complete orthonormal set of eigenfunctions of $\mathcal{A}_\varepsilon^\lambda$,
- $\mathcal{G}_{\varepsilon,n}^\lambda : \mathfrak{H} \to \mathfrak{H}$ be the orthogonal projection operators onto $\operatorname{span}(e_{\varepsilon,1}^\lambda, \dots, e_{\varepsilon,n}^\lambda)$,
- $\widetilde{\mathcal{M}}_{\varepsilon,n}^\lambda$ be the n-dimensional operator defined as the restriction of $\mathcal{M}_\varepsilon^\lambda$ to $\mathcal{G}_{\varepsilon,n}^\lambda(\mathfrak{H})$.

For $\lambda \in [0,1]$, $\varepsilon \geq 0$ and $n \in \mathbb{N}$ we define the measures (where we *always* take multiplicities into account!)

$$\nu_{\lambda,\varepsilon} = \sum_{x \in \mathrm{sp}_{\mathrm{pp}}(\mathcal{M}_\varepsilon^\lambda) \setminus \mathrm{sp}_{\mathrm{ess}}(\mathcal{M}_\varepsilon^\lambda)} \delta_x$$

and for any $\varepsilon > 0$ the measures

$$\widetilde{\nu}_{\lambda,\varepsilon,n} = \sum_{x \in \mathrm{sp}(\widetilde{\mathcal{M}}_{\varepsilon,n}^\lambda)} \delta_x,$$

where δ_x is the standard Dirac delta function centred at x. Consider a cutoff function ϕ_η satisfying

$$\phi_\eta(x) = \begin{cases} 1 & x \in [-1,1] \\ 0 & x \in \mathbb{R} \setminus (-1-\eta, 1+\eta) \end{cases}, \quad \phi_\eta \in C(\mathbb{R}, [0,1]), \quad \eta \in (0,\alpha). \tag{$*$}$$

Finally, define the measures

$$\mu_{\lambda,\varepsilon}^\eta = \phi_\eta \nu_{\lambda,\varepsilon}$$

and

$$\widetilde{\mu}_{\lambda,\varepsilon,n}^\eta = \phi_\eta \widetilde{\nu}_{\lambda,\varepsilon,n}.$$

Our main result is formulated for the general case where the spectrum of $\mathcal{A}^\lambda$ may

have a continuous part:

Theorem 5.1.1. *The mappings* $[0,1] \times [0,\infty) \ni (\lambda, \varepsilon) \mapsto \mu^{\eta}_{\lambda,\varepsilon}$ *and* $[0,1] \ni \lambda \mapsto \widetilde{\mu}^{\eta}_{\lambda,\varepsilon,n}$ *(here* $\varepsilon > 0$*) are weakly continuous and as* $n \to \infty$, $d_{BL}(\widetilde{\mu}^{\eta}_{\lambda,\varepsilon,n}, \mu^{\eta}_{\lambda,\varepsilon}) \to 0$ *uniformly in* $\lambda \in [0,1]$.

Remark 5.1.1. *Note that the above statement does not depend on the particular choice of cutoff function* ϕ_η*, as long as the requirements in* $(*)$ *are satisfied.*

A Simpler Case: Semi-Bounded Operators. As the notation becomes quite cumbersome due to the decomposition $\mathfrak{H} = \mathfrak{H}_+ \oplus \mathfrak{H}_-$, we shall first treat the simpler case of semi-bounded operators. Let $\mathcal{A}^\lambda$ and $\mathcal{K}^\lambda$, where $\lambda \in [0,1]$, be two families of operators on some Hilbert space $\mathfrak{H}$ (which is not assumed to decompose as before) where the family $\mathcal{A}^\lambda$ is also assumed to be defined for λ in an open neighbourhood D of $[0,1]$ in the complex plane. For the sake of precision, we repeat the assumptions (i)-(v) reformulated for this case.

i) Sectoriality: The family $\mathcal{A}^\lambda$ is sectorial of type (B) in $\lambda \in D$ and self-adjoint for $\lambda \in [0,1]$.

ii) Semi-boundedness: $\mathcal{A}^\lambda > 1$ for every $\lambda \in [0,1]$.

iii) Bounded perturbation: $\{\mathcal{K}^\lambda\}_{\lambda\in[0,1]} \subset \mathfrak{B}(\mathfrak{H})$ is a self-adjoint strongly continuous family.

iv) Compactness: There exists a self-adjoint operator $\mathcal{P} \in \mathfrak{B}(\mathfrak{H})$ which is relatively compact with respect to $\mathcal{A}^\lambda$, satisfying $\mathcal{K}^\lambda = \mathcal{K}^\lambda \mathcal{P}$ for all $\lambda \in [0,1]$.

v) Compactification of the resolvent: There exist holomorphic forms $\{\mathfrak{w}^\lambda\}_{\lambda\in D}$ of type (a) and associated operators $\{\mathcal{W}^\lambda\}_{\lambda\in D}$ of type (B) such that for $\lambda \in [0,1]$, $\mathcal{W}^\lambda$ are self-adjoint and non-negative. Define

$$\mathcal{A}^\lambda_\varepsilon := \mathcal{A}^\lambda + \varepsilon \mathcal{W}^\lambda, \quad \lambda \in D,\ \varepsilon \geq 0$$

with respective associated forms $\mathfrak{w}^\lambda$ and $\mathfrak{a}^\lambda_\varepsilon$. Then we assume that $\mathfrak{D}(\mathfrak{w}^\lambda) \cap \mathfrak{D}(\mathfrak{a})$ are dense for all $\lambda \in D$ and the inclusion $(\mathfrak{D}(\mathfrak{w}^\lambda) \cap \mathfrak{D}(\mathfrak{a}), \|\cdot\|_{\mathfrak{a}^\lambda_\varepsilon}) \to (\mathfrak{H}, \|\cdot\|)$ is compact for some $\lambda \in D$ and all $\varepsilon > 0$.

We define the projections and measures as above and therefore do not repeat the definition again.

Theorem 3′. *In the semi-bounded case the mappings* $[0,1] \times [0,\infty) \ni (\lambda, \varepsilon) \mapsto \mu^{\eta}_{\lambda,\varepsilon}$ *and* $[0,1] \ni \lambda \mapsto \widetilde{\mu}^{\eta}_{\lambda,\varepsilon,n}$ *(here* $\varepsilon > 0$*) are weakly continuous and as* $n \to \infty$, $d_{BL}(\widetilde{\mu}^{\eta}_{\lambda,\varepsilon,n}, \mu^{\eta}_{\lambda,\varepsilon}) \to 0$ *uniformly in* $\lambda \in [0,1]$.

In the subsequent sections we will prove Theorem 3′ before proving Theorem 5.1.1 in Section 5.5.

5.1.3 Discussion

One of the main driving forces behind the study of linear operators in the 20th century was the development of quantum mechanics. Particular attention had been given to the characterisation of the spectra of such operators, as it encodes many important physical properties (such as energy levels, for instance). When operators become too complex, a typical approach is to view them as perturbations of simpler operators whose spectrum is well understood. Two of the classic texts on this topic are those written by Kato [109] and Reed and Simon [171]. Both are still widely cited to this day. We also refer to Simon's review paper [181] and the references therein.

Recently, Hansen [80] presented new techniques for approximating spectra of linear operators (self-adjoint and non-self-adjoint) from a more computational point of view. In [184], Strauss presents a new method for approximating eigenvalues and eigenvectors of self-adjoint operators via an algorithm that is itself self-adjoint, and which does not produce spectral pollution. Both papers provide extensive references to additional literature in the field. We also mention [115], where analysis similar to ours is performed for bounded operators. We note that spectral pollution (the appearance of spurious eigenvalues within gaps in the essential spectrum when approximating) has attracted significant attention [42, 127, 128]. We do not encounter this issue here because of how the problem is set up: the trial spaces are (and therefore commute with) the spectral projectors of the block diagonal parts of the unperturbed operator, see e.g. [128] for more discussion of this topic.

The question that we are motivated by is somewhat different. We are interested in the simultaneous approximation of *families of operators*, rather than approximating a single fixed linear operator. This may be viewed as perturbation theory with two parameters: the continuous parameter λ representing small continuous perturbations generating the family of operators, and the discrete parameter n representing the dimension of the finite-dimensional approximation. One of the important aspects of this theory is that the finite-dimensional approximations approximate the entire family of operators uniformly in λ. Previously, in [17, Proposition 2.5] a much weaker result of this type was obtained, where the resolvent set of Schrödinger operators with a compact resolvent was shown to be stable under similar perturbations. We also mention [43, 41, 104] where the convergence of the so-called *Hill's method* (or Fourier-Floquet-Hill) is studied. This is a numerically-oriented method for studying spectra of periodic differential operators (not necessarily self-adjoint) and involves the truncation of the associated Fourier series. We refer in particular to [104] for an instance where this method is also applied to a *family* of operators.

There are two substantial difficulties in proving our results. If the spectrum of $\mathcal{A}^\lambda$ were discrete for some λ (and therefore for all λ) we would have a natural way to construct approximations by projecting onto increasing subspaces associated to the eigenvalues of $\mathcal{M}^\lambda$. However we do not require the spectrum to be discrete, and, indeed, in the type of problems we have in mind it is not. This necessitates the introduction of yet another perturbation parameter, ε, related to the compactification of the resolvent. The other difficulty is in ensuring that the finite-dimensional approximations approximate the whole family of operators uniformly in λ. To this end, the compactness assumption (iv) plays a crucial role.

We make several remarks on Theorem 5.1.1 and Theorem 3′ and the assumptions (i)-(v):

Remark 5.1.2. *The compactness requirements (iv) on $\mathcal{P}$ are motivated by (5.1.1). If $\mathcal{A}$ has a compact resolvent (e.g. when acting in $L^2(\mathbb{T}^d) \oplus L^2(\mathbb{T}^d)$ where $\mathbb{T}^d$ is the d-dimensional torus) we may take $\mathcal{P}$ to be the identity. Otherwise (e.g. for $L^2(\mathbb{R}^d) \oplus L^2(\mathbb{R}^d)$) if the perturbations $\mathcal{K}^\lambda$ are compactly supported in the sense*

that

$$\bigcup_{\lambda\in[0,1],u\in\mathfrak{H}} \operatorname{supp}(\mathcal{K}^\lambda u) \subset K \tag{5.1.6}$$

where $K = K_+ \times K_- \subset \mathbb{R}^d \times \mathbb{R}^d$ *is compact, then we may take* $\mathcal{P}_\pm$ *as multiplications by the indicator functions of the sets* $K_\pm$*. Indeed, we first note that* (5.1.6) *implies that for all* λ, $\mathcal{K}^\lambda = \mathcal{P}\mathcal{K}^\lambda$*. Then as* $\mathcal{K}^\lambda$ *and* $\mathcal{P}$ *are symmetric, we deduce that* $\mathcal{K}^\lambda = (\mathcal{K}^\lambda)^* = (\mathcal{K}^\lambda)^*\mathcal{P}^* = \mathcal{K}^\lambda\mathcal{P}$ *as required. That* $\mathcal{P}$ *is relatively compact with respect to* $-\Delta$ *follows from Rellich's theorem. We also remark that this choice of* $\mathcal{P}$ *is in fact the natural inclusion map from* L^2 *to* $L^2(K)$*.*

Remark 5.1.3. *Care must be taken regarding the spaces we view operators as acting on. If we view* $\mathcal{M}^\lambda_{\varepsilon,n} = \mathcal{G}^\lambda_{\varepsilon,n}\mathcal{M}^\lambda_\varepsilon\mathcal{G}^\lambda_{\varepsilon,n} : \mathfrak{H} \to \mathfrak{H}$ *then* 0 *will* always *be a spurious eigenvalue with infinite multiplicity. To remove this unwanted eigenvalue we must instead consider* $\widetilde{\mathcal{M}}^\lambda_{\varepsilon,n} : \mathfrak{H}^\lambda_{\varepsilon,n} \to \mathfrak{H}^\lambda_{\varepsilon,n}$ *where* $\mathfrak{H}^\lambda_{\varepsilon,n} = \mathcal{G}^\lambda_{\varepsilon,n}(\mathfrak{H})$ *is the* n*-dimensional space corresponding to the eigenprojection* $\mathcal{G}^\lambda_{\varepsilon,n}$*.*

Remark 5.1.4. *Property (ii) implies that there exists* $\alpha(\lambda) > 0$ *such that* $(-\alpha(\lambda) - 1, 1 + \alpha(\lambda))$ *is in the resolvent set of* $\mathcal{A}^\lambda$*. Since the spectrum is continuous in* $\lambda \in [0,1]$ *this implies that there is a uniform constant* $\alpha > 0$ *such that* $(-\alpha - 1, 1 + \alpha)$ *is in the resolvent set of* $\mathcal{A}^\lambda$ *for all* $\lambda \in [0,1]$*.*

Remark 5.1.5. *We finally remark that the construction of a compactifying operator* $\mathcal{W}$ *in general is not easy. We have in mind an application to a case where this is applied to* $-\Delta$ *and then it is simple: any reasonable unbounded potential will do.*

This chapter is organised as follows. In Section 5.2 we present some results related to general properties (such as self-adjointness, equivalence of norms, etc.) of the various operators. In Section 5.3 we construct the finite-dimensional approximations to our family of operators, which are used in Section 5.4 to prove Theorem 3′. In Section 5.5 these results are extended to families of operators which are not positive, proving Theorem 5.1.1. Finally, in Section 5.6 we give a brief description of an application of these results related to plasma instabilities, which is the subject of Chapter 6 where one can find the full details.

5.2 Preliminary results

We remind the reader that in this section, as well as in Section 5.3 and Section 5.4 we treat the semi-bounded case (Theorem 3′).

Considering the definition (5.1.4) and the subsequent specifications of the properties of the various operators and associated forms, we have the following results.

Lemma 5.2.1. *The forms $\mathfrak{m}^\lambda$ have the same domains as the forms $\mathfrak{a}^\lambda$, and are independent of λ. For any $\lambda \in [0,1]$, $\mathcal{M}^\lambda$ is self-adjoint and has the same essential spectrum and domain as $\mathcal{A}^\lambda$. In particular its spectrum inside $(-\infty, 1]$ is discrete.*

Proof. The equality $\mathfrak{D}(\mathfrak{m}^\lambda) = \mathfrak{D}(\mathfrak{a}^\lambda)$ holds since $\mathcal{K}^\lambda$ is bounded for each λ. The fact that the domains are independent of λ was assumed above in the sectoriality assumption (i). Self-adjointness follows from the Kato-Rellich theorem, due to $\mathcal{A}^\lambda$ being self-adjoint for $\lambda \in [0,1]$ and the symmetry assumption (iii) on $\mathcal{K}^\lambda$. The essential spectrum result follows from Weyl's theorem as $\mathcal{K}^\lambda = \mathcal{K}^\lambda \mathcal{P}$ is relatively compact with respect to $\mathcal{A}^\lambda$ (for any λ) because $\mathcal{P}$ is. □

Next, we turn our attention to the map $\lambda \mapsto \mathcal{M}^\lambda$. Intuitively, one would expect $\mathcal{M}^\lambda$ to have continuity properties similar to those of $\mathcal{K}^\lambda$ and therefore be merely continuous in the strong resolvent sense. In fact, due to the relative compactness assumption on $\mathcal{P}$ we have more:

Proposition 5.2.1. *The family $\{\mathcal{M}^\lambda\}_{\lambda\in[0,1]}$ is norm resolvent continuous.*

Proof. Fix some $\lambda \in [0,1]$ and let $[0,1] \ni \lambda_n \to \lambda$ as $n \to \infty$. It is sufficient to prove

$$\left\|(\mathcal{M}^{\lambda_n} + i)^{-1} - (\mathcal{M}^\lambda + i)^{-1}\right\|_{\mathfrak{B}(\mathfrak{H})} \to 0 \text{ as } n \to \infty.$$

Using the triangle inequality we have

$$\begin{aligned}\left\|(\mathcal{M}^{\lambda_n} + i)^{-1} - (\mathcal{M}^\lambda + i)^{-1}\right\|_{\mathfrak{B}(\mathfrak{H})} \leq{}& \left\|(\mathcal{M}^{\lambda_n} + i)^{-1} - (\mathcal{A}^{\lambda_n} + \mathcal{K}^\lambda + i)^{-1}\right\|_{\mathfrak{B}(\mathfrak{H})} \\ &+ \left\|(\mathcal{A}^{\lambda_n} + \mathcal{K}^\lambda + i)^{-1} - (\mathcal{M}^\lambda + i)^{-1}\right\|_{\mathfrak{B}(\mathfrak{H})}.\end{aligned}$$

By observing that $\{\mathcal{A}^\sigma + \mathcal{K}^\lambda\}_{\sigma\in D}$ is also a holomorphic family of type (B) we deduce that the second term tends to zero as $n \to \infty$. For the first term we follow the method used to deduce the second Neumann series (see [109, II-(1.13)])

$$(\mathcal{A}^{\lambda_n} + \mathcal{K}^{\lambda_n} + i)^{-1} = (\mathcal{A}^{\lambda_n} + \mathcal{K}^{\lambda} + i)^{-1}(1 + (\mathcal{K}^{\lambda_n} - \mathcal{K}^{\lambda})(\mathcal{A}^{\lambda_n} + \mathcal{K}^{\lambda} + i)^{-1})^{-1}$$

which is valid whenever $\left\|(\mathcal{K}^{\lambda_n} - \mathcal{K}^{\lambda})(\mathcal{A}^{\lambda_n} + \mathcal{K}^{\lambda} + i)^{-1}\right\|_{\mathfrak{B}(\mathfrak{H})} < 1$. By the norm resolvent continuity of operator inversion and again using the norm resolvent continuity of the family $\{\mathcal{A}^\sigma + \mathcal{K}^\lambda\}_{\sigma\in[0,1]}$, it is sufficient to show that

$$\left\|(\mathcal{K}^{\lambda_n} - \mathcal{K}^{\lambda})(\mathcal{A}^{\lambda} + \mathcal{K}^{\lambda} + i)^{-1}\right\|_{\mathfrak{B}(\mathfrak{H})} \to 0 \text{ as } n \to \infty. \tag{5.2.1}$$

We observe that $\mathcal{A}^\lambda + \mathcal{K}^\lambda$ is self-adjoint with the same domain as $\mathcal{A}^\lambda$ by Lemma 5.2.1, so $\mathcal{P}$ is also relatively compact with respect to $\mathcal{A}^\lambda + \mathcal{K}^\lambda$. By assumption (iv) we have

$$(\mathcal{K}^{\lambda_n} - \mathcal{K}^{\lambda})(\mathcal{A}^{\lambda} + \mathcal{K}^{\lambda} + i)^{-1} = (\mathcal{K}^{\lambda_n} - \mathcal{K}^{\lambda})\mathcal{P}(\mathcal{A}^{\lambda} + \mathcal{K}^{\lambda} + i)^{-1}.$$

This is a composition of a strongly convergent sequence of operators and the compact operator $\mathcal{P}(\mathcal{A}^\lambda + \mathcal{K}^\lambda + i)^{-1}$. The compactness converts the strong convergence to norm convergence and proves (5.2.1). □

5.3 Constructing approximations

We first treat approximations of operators with discrete spectra, which are naturally defined via a sequence of increasing projection operators. For brevity, we call these approximations *n-approximations* ("n" refers to the dimension of the projection). Then, our strategy when treating operators with a *continuous* spectrum is to first "perturb" them by adding a family of unbounded operators (think of adding an unbounded potential to a Laplacian) depending upon a small parameter ε. For each $\varepsilon > 0$ these perturbations are assumed to eliminate any continuous spectrum, so that then we may apply an n-approximation. We therefore call these (ε, n)-*approximations*. We start with a standard result for which we could not find a good reference and we therefore state and prove it here.

Lemma 5.3.1. *Let $\mathfrak{H}$ be a Hilbert space and let $\mathcal{T}_n \xrightarrow{s.r.} \mathcal{T}$ as $n \to \infty$ with $\mathcal{T}_n, \mathcal{T}$*

self-adjoint operators on $\mathfrak{H}$. *Let* $\mathcal{K}_n \xrightarrow{s} \mathcal{K}$ *as* $n \to \infty$ *with* $\mathcal{K}_n, \mathcal{K}$ *bounded self-adjoint operators on* $\mathfrak{H}$. *Then* $\mathcal{T}_n + \mathcal{K}_n$ *and* $\mathcal{T} + \mathcal{K}$ *are self-adjoint in* $\mathfrak{H}$ *and* $\mathcal{T}_n + \mathcal{K}_n \xrightarrow{s.r.} \mathcal{T} + \mathcal{K}$.

Proof. The self-adjointness follows from the Kato-Rellich theorem. For the convergence it is sufficient to prove that $(\mathcal{T}_n + \mathcal{K}_n + \alpha i)^{-1} \xrightarrow{s} (\mathcal{T} + \mathcal{K} + \alpha i)^{-1}$ for some real $\alpha \neq 0$. As the $\mathcal{K}_n$ are strongly convergent, by the uniform boundedness principle they are uniformly bounded in operator norm by some $M \geq \|\mathcal{K}\|_{\mathfrak{B}(\mathfrak{H})}$. Letting $\alpha = 2M$, and using the second Neumann series,

$$\begin{aligned}(\mathcal{T}_n + \mathcal{K}_n + \alpha i)^{-1} &= (\mathcal{T}_n + \alpha i)^{-1}(1 + \mathcal{K}_n(\mathcal{T}_n + \alpha i)^{-1})^{-1} \\ &= (\mathcal{T}_n + \alpha i)^{-1} \sum_{k=0}^{\infty} (-1)^k (\mathcal{K}_n(\mathcal{T}_n + \alpha i)^{-1})^k\end{aligned}$$

is convergent uniformly in n as $\|\mathcal{K}_n(\mathcal{T}_n + \alpha i)^{-1}\|_{\mathfrak{B}(\mathfrak{H})} \leq M/\alpha = 1/2 < 1$. As $n \to \infty$ each term of the series converges strongly to the corresponding term of the series for $(\mathcal{T} + \mathcal{K} + \alpha i)^{-1}$ and as the series convergences uniformly in n we may may swap the order of summation and take strong limits. □

5.3.1 Operators with discrete spectra

In this paragraph we assume that $\mathcal{A}^\lambda$ has discrete spectrum and compact resolvent for some λ (and, in fact, for all λ, as $\mathcal{A}^\lambda$ is a holomorphic family of type (B)[3]). We exploit a property of self-adjoint holomorphic families [109, VII Theorem 3.9 and VII Remark 4.22]: all eigenvalues of $\mathcal{A}^\lambda$ can be represented by functions which are holomorphic on $[0, 1]$. That is, there exists a sequence of scalar-valued functions $\{\mu_k^\lambda\}_{k\in\mathbb{N}}$ which are all holomorphic functions of $\lambda \in [0, 1]$ that represents all the repeated eigenvalues of $\mathcal{A}^\lambda$. Moreover, there exists a sequence of vector-valued functions $\{e_k^\lambda\}_{k\in\mathbb{N}}$ which are all also holomorphic functions of $\lambda \in [0, 1]$ such that for every $\lambda \in [0, 1]$, $\{e_k^\lambda\}_{k\in\mathbb{N}}$ form a complete orthonormal family of corresponding eigenvectors. An immediate consequence is that the unitary operator defined by

$$\begin{aligned}\mathcal{U}_\sigma^\lambda &: \mathfrak{H} \to \mathfrak{H} \\ e_k^\sigma &\mapsto e_k^\lambda \qquad \text{for any } k \in \mathbb{N}\end{aligned}$$

[3] See property (i) in Section 5.1.2 for a precise definition.

is jointly holomorphic in $\lambda, \sigma \in [0,1]$, i.e. possesses a locally convergent power series in the two variables λ, σ. We now define the n-truncation operator by

$$\mathcal{G}_n^\lambda : \mathfrak{H} \to \mathfrak{H}$$
$$e_k^\lambda \mapsto \begin{cases} e_k^\lambda & \text{if } k \leq n, \\ 0 & \text{if } k > n. \end{cases}$$

Since the eigenfunctions form a complete orthonormal set we have the convergence $\mathcal{G}_n^\lambda \xrightarrow{s} 1$ as $n \to \infty$ for fixed λ. Additionally by expressing $\mathcal{G}_n^\lambda = \mathcal{U}_\sigma^\lambda \mathcal{G}_n^\sigma \mathcal{U}_\lambda^\sigma$ for some fixed $\sigma \in [0,1]$ we see that $\mathcal{G}_n^\lambda \xrightarrow{s} 1$ as $n \to \infty$. Moreover, for any sequence $\lambda_n \to \lambda$ we have $\mathcal{G}_n^{\lambda_n} \xrightarrow{s} 1$ as $n \to \infty$. For notational convenience we define $\mathcal{G}_\infty^\lambda = 1$ for all $\lambda \in [0,1]$.

We now define the finite-dimensional approximations of $\mathcal{A}^\lambda$ and $\mathcal{M}^\lambda$ by

$$\mathcal{A}_n^\lambda = \mathcal{G}_n^\lambda \mathcal{A}^\lambda \mathcal{G}_n^\lambda \quad \text{and} \quad \mathcal{M}_n^\lambda = \mathcal{G}_n^\lambda \mathcal{M}^\lambda \mathcal{G}_n^\lambda, \tag{5.3.1}$$

respectively. It is too much to hope for convergence $\mathcal{M}_n^\lambda \xrightarrow{n.r.} \mathcal{M}^\lambda$ as $n \to \infty$, but we can hope for $\mathcal{M}_n^\lambda \xrightarrow{s.r.} \mathcal{M}^\lambda$. Indeed:

Lemma 5.3.2. *For any sequence $\lambda_n \to \lambda \in [0,1]$ as $n \to \infty$, we have the convergence $\mathcal{M}_n^{\lambda_n} \xrightarrow{s.r.} \mathcal{M}^\lambda$.*

Proof. By the stability of strong resolvent continuity with respect to bounded strongly continuous perturbations (see Lemma 5.3.1), it is sufficient to prove that $\mathcal{A}_n^{\lambda_n} \xrightarrow{s.r.} \mathcal{A}^\lambda$ as $n \to \infty$ and that $\mathcal{G}_n^{\lambda_n} \mathcal{K}^{\lambda_n} \mathcal{G}_n^{\lambda_n} \xrightarrow{s} \mathcal{K}^\lambda$. The latter is true as it is the composition of strong convergences of bounded operators. For the former it is sufficient to show that $(\mathcal{A}_n^{\lambda_n} + i)^{-1} \xrightarrow{s} (\mathcal{A}^\lambda + i)^{-1}$ as $n \to \infty$. Splitting this term as

$$(\mathcal{A}_n^{\lambda_n} + i)^{-1} = \mathcal{G}_n^{\lambda_n} (\mathcal{A}_n^{\lambda_n} + i)^{-1} \mathcal{G}_n^{\lambda_n} + (1 - \mathcal{G}_n^{\lambda_n})(\mathcal{A}_n^{\lambda_n} + i)^{-1}(1 - \mathcal{G}_n^{\lambda_n}),$$

(where we have used the fact that $\mathcal{G}_n^{\lambda_n}$ is a spectral projection which commutes with $(\mathcal{A}_n^{\lambda_n} + i)^{-1}$), we see that the second term converges strongly to zero since $(\mathcal{A}_n^{\lambda_n} + i)^{-1}$ is uniformly bounded and since $\mathcal{G}_n^{\lambda_n} \xrightarrow{s} 1$. For the first term on the right hand side, note that

$$\mathcal{G}_n^{\lambda_n} (\mathcal{A}_n^{\lambda_n} + i)^{-1} \mathcal{G}_n^{\lambda_n} = \mathcal{G}_n^{\lambda_n} (\mathcal{A}^{\lambda_n} + i)^{-1} \mathcal{G}_n^{\lambda_n}$$

which converges strongly to $(\mathcal{A}^\lambda + i)^{-1}$ by the composition of strong convergences. □

5.3.2 Operators with continuous spectra

We are now ready to turn to the general case of families $\{\mathcal{A}^\lambda\}_{\lambda\in[0,1]}$ that may have continuous spectra. Such operators require (ε, n)-approximations. The ε-approximations $\mathcal{A}^\lambda_\varepsilon$ of $\mathcal{A}^\lambda$ were defined in (5.1.3) and the corresponding approximations $\mathcal{M}^\lambda_\varepsilon$ were defined in (5.1.5).

Lemma 5.3.3. *1. For any $\varepsilon > 0$, $\{\mathcal{A}^\lambda_\varepsilon\}_{\lambda\in D}$ is a holomorphic family of type (B) with compact resolvent.*

2. For any $\lambda \in [0,1], \varepsilon \geq 0$, $\mathcal{A}^\lambda_\varepsilon$ is self-adjoint and we have $\mathcal{A}^\lambda_\varepsilon \geq \mathcal{A}^\lambda \geq 1 + \alpha$, where α was defined in Remark 5.1.4.

Proof. The second claim is obvious since $\mathcal{W}^\lambda \geq 0$. For the first we must show that $\mathfrak{a}^\lambda_\varepsilon$ is sectorial and that its domain $\mathfrak{D}(\mathfrak{a}^\lambda_\varepsilon)$ is independent of λ and dense in $\mathfrak{H}$, and that for any fixed $u \in \mathfrak{D}(\mathfrak{a}^\lambda_\varepsilon)$ the function $\mathfrak{a}^\lambda_\varepsilon[u]$ is holomorphic in $\lambda \in D$. For any $\lambda \in D$, $\mathfrak{a}^\lambda_\varepsilon$ is the sum of the sectorial forms $\mathfrak{a}^\lambda$ and $\varepsilon\mathfrak{w}^\lambda$ so by [109, VI§1.6-Theorem 1.33] it is closed and sectorial with domain $\mathfrak{D}(\mathfrak{a}) \cap \mathfrak{D}(\mathfrak{w}^\lambda)$, which is independent of λ since both $\mathcal{A}^\lambda$ and $\mathcal{W}^\lambda$ are holomorphic families of type (B). Furthermore, we assumed that $\mathfrak{D}(\mathfrak{a}) \cap \mathfrak{D}(\mathfrak{w}^\lambda)$ is dense in $\mathfrak{H}$. For any fixed $u \in \mathfrak{D}(\mathfrak{a}^\lambda_\varepsilon)$, $\mathfrak{a}^\lambda_\varepsilon[u] = \mathfrak{a}^\lambda[u] + \varepsilon\mathfrak{w}^\lambda[u]$ is the sum of two holomorphic functions of $\lambda \in D$, so $\mathfrak{a}^\lambda_\varepsilon[u]$ is also holomorphic in D. Finally by the assumption that the inclusion $(\mathfrak{D}(\mathfrak{a}^\lambda_\varepsilon), \|\cdot\|_{\mathfrak{a}^\lambda_\varepsilon}) \hookrightarrow \mathfrak{H}$ is compact we deduce that the resolvent of $\mathcal{A}^\lambda_\varepsilon$ is compact. □

For each $\varepsilon > 0$ the operator $\mathcal{A}^\lambda_\varepsilon$ has a discrete spectrum, and therefore the n-approximations of $\mathcal{A}^\lambda_\varepsilon$ and $\mathcal{M}^\lambda_\varepsilon$ may be defined analogously to (5.3.1) via the projection operators

$$\mathcal{G}^\lambda_{\varepsilon,n} : \mathfrak{H} \to \mathfrak{H}$$
$$e^\lambda_{\varepsilon,k} \mapsto \begin{cases} e^\lambda_{\varepsilon,k} & \text{if } k \leq n, \\ 0 & \text{if } k > n, \end{cases}$$

(where $\{e^\lambda_{\varepsilon,k}\}_{k\in\mathbb{N}}$ are normalised eigenfunctions of $\mathcal{A}^\lambda_\varepsilon$) as

$$\mathcal{A}^\lambda_{\varepsilon,n} = \mathcal{G}^\lambda_{\varepsilon,n}\mathcal{A}^\lambda_\varepsilon\mathcal{G}^\lambda_{\varepsilon,n} \quad \text{and} \quad \mathcal{M}^\lambda_{\varepsilon,n} = \mathcal{G}^\lambda_{\varepsilon,n}\mathcal{M}^\lambda_\varepsilon\mathcal{G}^\lambda_{\varepsilon,n}.$$

We know by Lemma 5.3.2 that the family $\{\mathcal{A}^\lambda_{\varepsilon,n}\}_{\lambda\in[0,1],n\in\overline{\mathbb{N}}}$ is continuous in the strong resolvent sense. In addition, we have:

Lemma 5.3.4. *The family $\{\mathcal{A}^\lambda_\varepsilon\}_{\lambda\in[0,1],\varepsilon\in[0,\infty)}$ is continuous in the strong resolvent sense.*

Proof. By the equivalence of strong and weak convergence of the resolvent for self-adjoint operators [172, VIII, Problem 20(a)] it is sufficient to prove that $(\mathcal{A}^\lambda_\varepsilon+1)^{-1}$ is weakly continuous jointly in λ and ε. Without loss of generality we restrict to $\varepsilon\in[0,1]$ the general case being no harder. Let $U\subseteq D$ be an open set containing the interval $[0,1]$ such that for $\lambda\in U$, $\operatorname{Re}\mathfrak{a}^\lambda \geq 1$ and $\operatorname{Re}\mathfrak{w}^\lambda \geq -1$. Then, for $\lambda\in U$ and $\varepsilon\in[0,1]$ the forms $\mathfrak{a}^\lambda_\varepsilon$ are closed and sectorial, with $\operatorname{Re}\mathfrak{a}^\lambda_\varepsilon \geq 0$. Hence the associated operators have the resolvent bound $\left\|(\mathcal{A}^\lambda_\varepsilon+\zeta)^{-1}\right\|_{\mathfrak{B}(\mathfrak{H})} \leq 1/\operatorname{Re}\zeta$ for $\operatorname{Re}\zeta>0$. In particular,

$$\sup_{\varepsilon\in[0,1],\lambda\in U}\left\|(\mathcal{A}^\lambda_\varepsilon+1)^{-1}\right\|_{\mathfrak{B}(\mathfrak{H})} \leq 1. \tag{5.3.2}$$

Now fix $u,v\in\mathfrak{H}$, let $\varepsilon_n\to\varepsilon_\infty\in[0,\infty)$ and define the sequence of holomorphic functions $f_n: U\to\mathbb{C}$ by

$$f_n(\lambda) = \left\langle (\mathcal{A}^\lambda_{\varepsilon_n}+1)^{-1}u - (\mathcal{A}^\lambda_{\varepsilon_\infty}+1)^{-1}u, v\right\rangle$$

with $f_\infty = 0$. To prove the joint weak continuity of the resolvent it is clearly sufficient to show that $f_n\to 0$ uniformly over $\lambda\in[0,1]$. The case $\varepsilon_\infty>0$ is straightforward so we assume that $\varepsilon_\infty = 0$. Without loss of generality we may assume that $\varepsilon_n\neq 0$ for all n. We will use a simple corollary of Montel's theorem (see e.g. [174, Theorem 14.6]) that states that a sequence of holomorphic functions that is uniformly bounded on an open set $U\subseteq\mathbb{C}$ and converges pointwise in U converges uniformly on any compact set $K\subset U$. The uniform boundedness of f_n follows from (5.3.2) above. Thus it suffices to show that $f_n\to 0$ pointwise. To this end we will establish pointwise convergence of the corresponding forms $\mathfrak{a}^\lambda_{\varepsilon_n}$.

Indeed,

$$\forall \lambda \in D, w \in \mathfrak{D}(\mathfrak{a}^\lambda_{\varepsilon_n}), \quad \mathfrak{a}^\lambda_{\varepsilon_n}[w] - \mathfrak{a}^\lambda[w] = \varepsilon_n \mathfrak{w}^\lambda[w] \to 0 \quad \text{as } n \to \infty.$$

For $n \in \mathbb{N}$ the forms have common form domain $\mathfrak{D}(\mathfrak{a}) \cap \mathfrak{D}(\mathfrak{w})$, which is a form core for $\mathfrak{a}^\lambda$, and the sequence of form differences $\mathfrak{a}^\lambda_{\varepsilon_n} - \mathfrak{a}^\lambda$ is uniformly sectorial. Thus due to [109, VIII.§3.2-Theorem 3.6] $\mathcal{A}^\lambda_{\varepsilon_n} \xrightarrow{s.r.} \mathcal{A}^\lambda$ as $n \to \infty$, which implies the pointwise convergence $f_n \to 0$ and completes the proof. □

Corollary 5.3.1. *The family* $\{\mathcal{M}^\lambda_\varepsilon\}_{\lambda\in[0,1],\varepsilon\in[0,\infty)}$ *is continuous in the strong resolvent sense.*

Proof. This follows from the stability of strong resolvent continuity with respect to bounded strongly continuous perturbations. □

5.4 Proof of Theorem 3′

We split the proof into first proving upper and lower semi-continuity of the spectrum. Informally, we recall that *upper*-semicontinuity of spectra means that the spectrum cannot expand when perturbed, while *lower*-semicontinuity means that the spectrum cannot shrink when perturbed. Then the claims on the measures $\mu^\eta_{\lambda,\varepsilon}$ and $\tilde{\mu}^\eta_{\lambda,\varepsilon,n}$ will be addressed.

Proof of Theorem 3′. We split the proof into three parts, denoted **I, II, III**.

I. Claim: along any sequence $(\lambda_m, \varepsilon_m) \to (\lambda_\infty, \varepsilon_\infty)$ **it holds that** $\mu^\eta_{\lambda_m,\varepsilon_m} \rightharpoonup \mu^\eta_{\lambda_\infty,\varepsilon_\infty}$ **as** $m \to \infty$**.** Indeed, we have to show that for any bounded continuous function f it holds that, as $m \to \infty$,

$$\int f \,\mathrm{d}\mu^\eta_{\lambda_m,\varepsilon_m} = \sum_{y\in\mathrm{sp}(\mathcal{M}^{\lambda_m}_{\varepsilon_m})} \phi_\eta(y) f(y) \to \sum_{y\in\mathrm{sp}(\mathcal{M}^{\lambda_\infty}_{\varepsilon_\infty})} \phi_\eta(y) f(y) = \int f \,\mathrm{d}\mu^\eta_{\lambda_\infty,\varepsilon_\infty} \tag{5.4.1}$$

where (as before) multiplicity is taken into account in the summations. Without loss of generality we assume that $f \geq 0$. We know that the spectrum of $\mathcal{M}^{\lambda_\infty}_{\varepsilon_\infty}$ inside the support of ϕ_η is discrete, consisting of a finite number of eigenvalues,

each of finite multiplicity. Let them be $\sigma_1, \dots, \sigma_M$ of respective multiplicities $N_1, \dots, N_M$. We split the proof of (5.4.1) into two steps.

I1. Claim: $\liminf_m \sum_{y\in \mathrm{sp}(\mathcal{M}^{\lambda_m}_{\varepsilon_m})} \phi_\eta(y) f(y) \geq \sum_{y\in \mathrm{sp}(\mathcal{M}^{\lambda_\infty}_{\varepsilon_\infty})} \phi_\eta(y) f(y)$.

By the strong resolvent convergence of $\mathcal{M}^{\lambda_m}_{\varepsilon_m}$ to $\mathcal{M}^{\lambda_\infty}_{\varepsilon_\infty}$ (Corollary 5.3.1) we know that for any $\delta > 0$ small enough there are only finitely many ms for which $\mathcal{M}^{\lambda_m}_{\varepsilon_m}$ does not have, for each $i = 1, \dots, M$, at least N_i eigenvalues (counting multiplicity!) within δ of σ_i. Thus, by the continuity and non-negativity of $\varphi_\eta f$, for any $\varepsilon' > 0$, we may choose δ small enough so that

$$\sum_{y\in \mathrm{sp}(\mathcal{M}^{\lambda_m}_{\varepsilon_m})} \phi_\eta(y) f(y) \geq \sum_{y\in \mathrm{sp}(\mathcal{M}^{\lambda_\infty}_{\varepsilon_\infty})} \phi_\eta(y) f(y) - \varepsilon'$$

for all but finitely many ms, which completes I1.

I2. Claim: $\limsup_m \sum_{y\in \mathrm{sp}(\mathcal{M}^{\lambda_m}_{\varepsilon_m})} \phi_\eta(y) f(y) \leq \sum_{y\in \mathrm{sp}(\mathcal{M}^{\lambda_\infty}_{\varepsilon_\infty})} \phi_\eta(y) f(y)$.

We first claim that for all but finitely many ms we have

$$\#\left(\mathrm{sp}(\mathcal{M}^{\lambda_m}_{\varepsilon_m}) \cap [-\eta - 1, 1 + \eta]\right) \leq \#\left(\mathrm{sp}(\mathcal{M}^{\lambda_\infty}_{\varepsilon_\infty}) \cap [-\eta - 1, 1 + \eta]\right) =: M',$$

counting multiplicities. Indeed, suppose not. Then there would exist a subsequence (for which we abuse notation and still denote by m) for which $\mathcal{M}^{\lambda_m}_{\varepsilon_m}$ has (at least) $M' + 1$ distinct eigenvalues (counting multiplicity). Say $\sigma_{m,1}, \dots, \sigma_{m,M'+1}$ with normalised eigenfunctions $u_{m,1}, \dots, u_{m,M'+1}$. By compactness of $[-\eta - 1, 1 + \eta]^{M'+1}$ we may pass to a subsequence (again we retain the index m) on which $\sigma_{m,i} \to \sigma_{\infty,i}$ for each $i = 1, \dots, M' + 1$ and some $\sigma_{\infty,i}$s. By Proposition 5.4.1(i) below we may pass to successive subsequences to obtain a final subsequence (still denoted m) for which additionally $u_{m,i} \to u_{\infty,i}$ strongly as $m \to \infty$ for each i where $u_{\infty,i}$ is a normalised eigenfunction of $\mathcal{M}^{\lambda_\infty}_{\varepsilon_\infty}$ with eigenvalue $\sigma_{\infty,i}$. Moreover, as all the operators involved are self-adjoint, for each m the eigenfunctions $\{u_{m,i}\}_{i=1}^{M'+1}$ form an orthonormal system, and as orthonormality is preserved by strong limits, this holds also for $\{u_{\infty,i}\}_{i=1}^{M'+1}$. But this implies that $\mathcal{M}^{\lambda_\infty}_{\varepsilon_\infty}$ has at least $M' + 1$ eigenvalues in $[-\eta - 1, 1 + \eta]$, a contradiction, proving the claim.

We can now complete the proof of I2. Suppose that the claimed bound fails, then

there would exist $\varepsilon' > 0$ and a subsequence (still denoted m) for which

$$\sum_{y \in \mathrm{sp}(\mathcal{M}^{\lambda_m}_{\varepsilon_m})} \phi_\eta(y) f(y) \geq \sum_{y \in \mathrm{sp}(\mathcal{M}^{\lambda_\infty}_{\varepsilon_\infty})} \phi_\eta(y) f(y) + \varepsilon'$$

for each m. Let $M_m = \#\left(\mathrm{sp}(\mathcal{M}^{\lambda_m}_{\varepsilon_m}) \cap [-\eta - 1, 1 + \eta]\right)$. Then by the previous claim we know that for all but finitely many ms we have $M_m \leq M'$. Thus some number $M'' \in \{1, \ldots, M'\}$ is equal to infinitely many of the M_ms. We pass to this subsequence (still denoted m) so that $M_m = M''$ for every m. Let these eigenvalues be $\{\sigma_{m,i}\}_{i=1}^{M''}$. As in the proof of the claim above, after passing to another subsequence we have $\sigma_{m,i} \to \sigma_{\infty,i}$ for each i where $\{\sigma_{\infty,i}\}_{i=1}^{M''}$ are distinct (counting multiplicity) eigenvalues of $\mathcal{M}^{\lambda_\infty}_{\varepsilon_\infty}$. Hence, by continuity and non-negativity of $f\phi_\eta$, we have

$$\begin{aligned}\sum_{y \in \mathrm{sp}(\mathcal{M}^{\lambda_\infty}_{\varepsilon_\infty})} \phi_\eta(y) f(y) &\geq \sum_{i=1}^{M''} \phi_\eta(\sigma_{\infty,i}) f(\sigma_{\infty,i}) \\ &= \lim_{m \to \infty} \sum_{i=1}^{M''} \phi_\eta(\sigma_{m,i}) f(\sigma_{m,i}) \geq \sum_{y \in \mathrm{sp}(\mathcal{M}^{\lambda_\infty}_{\varepsilon_\infty})} \phi_\eta(y) f(y) + \varepsilon'\end{aligned}$$

where the limit is on the subsequence we obtained. This is a contradiction which completes I2, and the weak convergence $\mu^\eta_{\lambda_m, \varepsilon_m} \rightharpoonup \mu^\eta_{\lambda_\infty, \varepsilon_\infty}$ follows.

II. Claim: for any $\varepsilon > 0$ and $n \in \mathbb{N}$ fixed and along any sequence $\lambda_m \to \lambda_\infty$ it holds that $\widetilde{\eta}_{\lambda_m, \varepsilon, n} \rightharpoonup \widetilde{\eta}_{\lambda_\infty, \varepsilon, n}$. This may be shown either by the same proof as in I, or we may simply note that the operators involved are finite dimensional matrices whose coefficients vary continuously in λ.

III. Claim: for any fixed $\varepsilon > 0$ we have $d_{BL}(\widetilde{\mu}^\eta_{\lambda, \varepsilon, n}, \mu^\eta_{\lambda, \varepsilon}) \to 0$ uniformly in $\lambda \in [0, 1]$ as $n \to \infty$. The convergence $\widetilde{\mu}^\eta_{\lambda_n, \varepsilon, n} \rightharpoonup \mu^\eta_{\lambda_\infty, \varepsilon}$ along any sequence $\lambda_n \to \lambda_\infty$ follows from the same proof as in I, replacing the reference to Proposition 5.4.1(i) with Proposition 5.4.1(ii). Uniform convergence follows from the compactness of $[0, 1]$. Indeed, suppose that this uniform convergence does not hold. Then there would exist $\delta > 0$ such that, for infinitely many ns it holds that $d_{BL}(\widetilde{\mu}^\eta_{\lambda_n, \varepsilon, n}, \mu^\eta_{\lambda_n, \varepsilon}) > \delta$ for some $\lambda_n \in [0, 1]$. Extract a subsequence (we abuse

notation and retain the index n) for which $\lambda_n \to \lambda_\infty \in [0,1]$. From I we know that for all but finitely many ns we must have $d_{BL}(\mu^\eta_{\lambda_n,\varepsilon}, \mu^\eta_{\lambda_\infty,\varepsilon}) < \delta/2$. Therefore, by the triangle inequality

$$d_{BL}(\widetilde{\mu}^\eta_{\lambda_n,\varepsilon,n}, \mu^\eta_{\lambda_\infty,\varepsilon}) \geq \left| d_{BL}(\widetilde{\mu}^\eta_{\lambda_n,\varepsilon,n}, \mu^\eta_{\lambda_n,\varepsilon}) - d_{BL}(\mu^\eta_{\lambda_n,\varepsilon}, \mu^\eta_{\lambda_\infty,\varepsilon}) \right| > \delta/2$$

for infinitely many ns, which is a contradiction to the weak convergence $\widetilde{\mu}^\eta_{\lambda_n,\varepsilon,n} \rightharpoonup \mu^\eta_{\lambda_\infty,\varepsilon}$. □

The missing ingredient in the above proof is:

Proposition 5.4.1. *Let $\sigma_n \to \sigma$ as $n \to \infty$ with $\sigma_n, \sigma \in (-\infty, 1]$ and $\lambda_n \to \lambda$ as $n \to \infty$ with $\lambda_n, \lambda \in [0,1]$. Then the following hold.*

1. *Let $\varepsilon_n \to \varepsilon \geq 0$ as $n \to \infty$, and $\{u_n\}_{n=1}^\infty$ be a sequence with $\|u_n\| = 1$, $u_n \in \mathfrak{D}(\mathcal{M}^\lambda_{\varepsilon_n})$ and $\mathcal{M}^{\lambda_n}_{\varepsilon_n} u_n = \sigma_n u_n$. Then $\{u_n\}_{n=1}^\infty$ has a subsequence strongly converging to some $u \neq 0$, which satisfies $\mathcal{M}^\lambda_\varepsilon u = \sigma u$.*

2. *Let $\varepsilon > 0$ be fixed, and $\{u_n\}_{n=1}^\infty$ be a sequence with $\|u_n\| = 1$, $\mathcal{G}^{\lambda_n}_{\varepsilon,n} u_n = u_n$ and $\mathcal{M}^{\lambda_n}_{\varepsilon,n} u_n = \sigma_n u_n$. Then $\{u_n\}_{n=1}^\infty$ has a subsequence strongly converging to some $u \neq 0$, which satisfies $\mathcal{M}^\lambda_\varepsilon u = \sigma u$.*

Proof. As the proof of the first claim is slightly simpler and otherwise the same, we only give the proof for the second claim, leaving the first to the reader. Each u_n solves the equation

$$\mathcal{G}^{\lambda_n}_{\varepsilon,n} \mathcal{A}^{\lambda_n}_\varepsilon \mathcal{G}^{\lambda_n}_{\varepsilon,n} u_n - \sigma_n u_n + \mathcal{G}^{\lambda_n}_{\varepsilon,n} \mathcal{K}^{\lambda_n} \mathcal{G}^{\lambda_n}_{\varepsilon,n} u_n = 0.$$

The requirement that $u_n = \mathcal{G}^{\lambda_n}_{\varepsilon,n} u_n$ and the fact that $\mathcal{G}^{\lambda_n}_{\varepsilon,n}$ commutes with $\mathcal{A}^{\lambda_n}_\varepsilon$ means that this is equivalent to

$$\mathcal{A}^{\lambda_n}_\varepsilon u_n = \sigma_n u_n - \mathcal{G}^{\lambda_n}_{\varepsilon,n} \mathcal{K}^{\lambda_n} u_n. \tag{5.4.2}$$

Taking the inner product with u_n we estimate,

$$\mathfrak{a}^0[u_n] \leq C\mathfrak{a}^{\lambda_n}[u_n] \leq C\mathfrak{a}^{\lambda_n}_{\varepsilon_n}[u_n] \leq C\sigma_n \|u_n\|^2 + C \sup_{\lambda\in[0,1]} \left\|\mathcal{K}^\lambda\right\|_{\mathfrak{B}(\mathfrak{H})} \|u_n\|^2 \leq C' \tag{5.4.3}$$

where C is independent of n and comes from the relative form boundedness of the holomorphic family $\{\mathcal{A}^\lambda\}_{\lambda\in D}$ (see [109, VII-§4.2]) and the supremum is finite by the uniform boundedness principle as $\{\mathcal{K}^\lambda\}_{\lambda\in[0,1]}$ is strongly continuous. Hence for all n we have $\left\||\mathcal{A}^0|^{1/2}u_n\right\|^2 \leq C'$, where $|\mathcal{A}^0|^{1/2}$ is the square root of the positive self-adjoint operator $\mathcal{A}^0$. By assumption, $\mathcal{P}$ is relatively compact with respect to $\mathcal{A}^0$, and hence also to $|\mathcal{A}^0|^{1/2}$. Indeed, the inverse of $|\mathcal{A}^0|^{1/2}$ can be expressed using the functional calculus (see [109, V-§3.11-Equation 3.43]) of the self-adjoint operator $\mathcal{A}^0$ as

$$|\mathcal{A}^0|^{-1/2} = \frac{1}{\pi}\int_0^\infty \zeta^{-1/2}(\mathcal{A}^0+\zeta)^{-1}\,d\zeta$$

where the integral is absolutely convergent in operator norm due to the bound $\|(\mathcal{A}^0+\zeta)^{-1}\|_{\mathfrak{B}(\mathfrak{H})} \leq (1+\zeta)^{-1}$ for $\zeta \geq 0$. By composing both sides of this equation on the left with $\mathcal{P}$ and moving $\mathcal{P}$ inside the integral (which is possible as $\mathcal{P}$ is bounded and the integral converges absolutely in norm) we deduce that $\mathcal{P}|\mathcal{A}^0|^{-1/2}$ is given by an absolutely norm convergent integral of compact operators, and is hence compact.

Thus we may pass to a subsequence (though we retain the subscript n) for which

$$\mathcal{P}u_n \to v \in \mathfrak{H}.$$

Then by rewriting (5.4.2) and using $\mathcal{K}^\lambda = \mathcal{K}^\lambda\mathcal{P}$ for all $\lambda \in [0,1]$ we have

$$u_n = -(\mathcal{A}^{\lambda_n}_\varepsilon - \sigma_n)^{-1}\mathcal{G}^{\lambda_n}_{\varepsilon,n}\mathcal{K}^{\lambda_n}\mathcal{P}u_n \tag{5.4.4}$$

where the resolvent exists by the assumption that $\mathcal{A}^\lambda \geq 1+\alpha$ for all $\lambda \in [0,1]$. As remarked before $\mathcal{G}^\lambda_{\varepsilon,n} \xrightarrow{s} 1$ uniformly in $\lambda \in [0,1]$ so that $\mathcal{G}^{\lambda_n}_{\varepsilon,n} \xrightarrow{s} 1$ as $n \to \infty$. Therefore by the composition of strong convergences

$$u_n \to -(\mathcal{A}^\lambda_\varepsilon - \sigma)^{-1}\mathcal{K}^\lambda v := u$$

as $n \to \infty$. Then as u_n is strongly convergent, necessarily $v = \mathcal{P}u$ and the assertion of the proposition follows. □

5.5 Non-positive operators: proof of Theorem 5.1.1

We define the ε-approximations of $\mathcal{A}^\lambda_\pm$ as before in terms of a pair of holomorphic families $\mathcal{W}^\lambda_\pm$ with the same assumptions. The eigenprojections of $\mathcal{A}^\lambda_\varepsilon$ are then denoted by $\mathcal{G}^\lambda_{\pm,\varepsilon,n}$ and we define

$$\mathcal{G}^\lambda_{\varepsilon,n} = \begin{bmatrix} \mathcal{G}^\lambda_{+,\varepsilon,n} & 0 \\ 0 & \mathcal{G}^\lambda_{-,\varepsilon,n} \end{bmatrix}$$

and

$$\mathcal{A}^\lambda_{\varepsilon,n} = \mathcal{G}^\lambda_{\varepsilon,n}\mathcal{A}^\lambda_\varepsilon\mathcal{G}^\lambda_{\varepsilon,n}$$
$$\mathcal{M}^\lambda_{\varepsilon,n} = \mathcal{G}^\lambda_{\varepsilon,n}\mathcal{M}^\lambda_\varepsilon\mathcal{G}^\lambda_{\varepsilon,n}.$$

All the preceding proofs of continuity can be adapted to this case. Indeed, Proposition 5.2.1 holds without modification, while Lemma 5.3.2 and Lemma 5.3.4 can be extended by using the identity

$$\left(\begin{bmatrix} \mathcal{T}_+ & 0 \\ 0 & \mathcal{T}_- \end{bmatrix} + i\right)^{-1} = \begin{bmatrix} (\mathcal{T}_+ + i)^{-1} & 0 \\ 0 & (\mathcal{T}_- + i)^{-1} \end{bmatrix}$$

and the stability of norm (resp. strong) continuity to symmetric bounded norm (reps. strongly) continuous perturbations. With these continuity results, the proof of lower semi-continuity of Σ and Σ_ε can be easily adapted. The compactness result Proposition 5.4.1 that establishes the upper semi-continuity needs a little more modification. Recall that the discrete region of the spectrum is the gap $(-\alpha-1, 1+\alpha)$ rather than the half-line $(-\infty, 1+\alpha)$. We restate the compactness result below.

Proposition 5.5.1. *Let $\sigma_n \to \sigma$ as $n \to \infty$ with $\sigma_n, \sigma \in [-1,1]$ and $\lambda_n \to \lambda$ as $n \to \infty$ with $\lambda_n, \lambda \in [0,1]$. Then the following hold.*

1. *Let $\varepsilon_n \to \varepsilon \geq 0$ as $n \to \infty$, and $\{u_n\}_{n=1}^\infty$ be a sequence with $\|u_n\| = 1$, $u_n \in \mathfrak{D}(\mathcal{M}^\lambda_{\varepsilon_n})$ and $\mathcal{M}^{\lambda_n}_{\varepsilon_n}u_n = \sigma_n u_n$. Then $\{u_n\}_{n=1}^\infty$ has a subsequence strongly converging to some $u \neq 0$, which satisfies $\mathcal{M}^\lambda_\varepsilon u = \sigma u$.*

2. *Let $\varepsilon > 0$ be fixed, and $\{u_n\}_{n=1}^\infty$ be a sequence with $\|u_n\| = 1$, $\mathcal{G}^{\lambda_n}_\varepsilon u_n = u_n$*

and $\mathcal{M}^{\lambda_n}_{\varepsilon,n} u_n = \sigma_n u_n$. *Then* $\{u_n\}_{n=1}^{\infty}$ *has a subsequence strongly converging to some* $u \neq 0$, *which satisfies* $\mathcal{M}^{\lambda}_{\varepsilon} u = \sigma u$.

Proof (sketched). We need only change (5.4.3) to the two estimates

$$\begin{aligned}\mathfrak{a}^0_{\pm}[u^{\pm}_k] &\leq C_{\pm}\mathfrak{a}^{\lambda_k}_{\pm}[u^{\pm}_k] \leq C_{\pm}\mathfrak{a}^{\lambda_k}_{\pm,\varepsilon_k}[u^{\pm}_k] \\ &\leq C_{\pm}|\sigma_k| \left\|u^{\pm}_k\right\|^2 + C_{\pm} \sup_{\lambda\in[0,1]} \left\|\mathcal{K}^{\lambda}\right\|_{\mathfrak{B}(\mathfrak{H})} \|u_k\|^2 \leq C'\end{aligned}$$

obtained by taking the inner product of (5.4.2) with $u^{\pm}_k$ where $u_k = (u^+_k, u^-_k) \in \mathfrak{H}_+ \times \mathfrak{H}_-$, from which the relative compactness of $\mathcal{P}u_k$ follows as before, and lastly note that $\mathcal{A}^{\lambda}_{\pm} \geq 1+\alpha$ implies that the resolvent $(\mathcal{A}^{\lambda_k}_{\varepsilon_k} - \sigma_k)^{-1}$ exists in (5.4.4). □

This proves Theorem 5.1.1.

5.6 An application: plasma instabilities

The discussion in this section is informal. As stability analysis typically relies on a detailed understanding of the spectrum of the linearised problem, most results in this direction require delicate spectral analysis. However, an outstanding open problem has been stability analysis of plasmas that do not possess special symmetries (such as periodicity or monotonicity[4]) due to the more complicated structure of the spectrum. A significant obstacle has been the existence of an essential spectrum extending to both $\pm\infty$. Let us briefly outline the problem, which is treated in detail in Chapter 6.

Plasmas are typically modelled by the relativistic Vlasov-Maxwell system: Letting $f = f(t,x,v)$ be a probability density function measuring the density of electrons that at time $t \geq 0$ are located at the point $x \in \mathbb{R}^d$, have momentum $v \in \mathbb{R}^d$ and velocity $\hat{v} = v/\sqrt{1+|v|^2}$, the (relativistic) Vlasov equation

$$\frac{\partial f}{\partial t} + \hat{v} \cdot \nabla_x f + \mathbf{F} \cdot \nabla_v f = 0 \tag{5.6.1}$$

[4] *Monotonicity*, roughly speaking, means that there are fewer particles at higher energies. For a precise definition see e.g. [17].

is a transport equation describing their evolution due to the Lorentz force $\mathbf{F} = -\mathbf{E} - \hat{v} \times \mathbf{B}$. Here we have taken the mass of the electrons and the speed of light to be 1 for simplicity. The fields $\mathbf{E} = \mathbf{E}(t,x)$ and $\mathbf{B} = \mathbf{B}(t,x)$ are the (self-consistent) electric and magnetic fields, respectively. They satisfy Maxwell's equations (written here for their respective potentials ϕ and $\mathbf{A}$, satisfying $\mathbf{E} = -\nabla\phi - \partial_t \mathbf{A}$ and $\mathbf{B} = \nabla \times \mathbf{A}$ in the Lorenz gauge $\partial_t \phi + \nabla \cdot \mathbf{A} = 0$):

$$\begin{cases} (-\boldsymbol{\Delta} + \partial_t^2)\mathbf{A} - \mathbf{j} = \mathbf{0}, \\ (\Delta - \partial_t^2)\phi + \rho = 0, \end{cases} \tag{5.6.2}$$

where $\rho = \rho(t,x) = -\int f \, dv$ is the charge density and $\mathbf{j} = \mathbf{j}(t,x) = -\int \hat{v} f \, dv$ is the current density (negative signs are due to the electrons charge). Linearising (5.6.1) we obtain

$$\frac{\partial f}{\partial t} + \hat{v} \cdot \nabla_x f + \mathbf{F^0} \cdot \nabla_v f = -\mathbf{F} \cdot \nabla_v f^0, \tag{5.6.3}$$

where f^0 and $\mathbf{F^0}$ are the equilibrium density and force field, respectively, and f and $\mathbf{F}$ are their first order perturbations. Maxwell's equations do not require linearisation as they are already linear. We seek solutions to (5.6.2)-(5.6.3) that grow exponentially in time. Therefore, substituting into (5.6.3) the ansatz that all time-dependent quantities behave like $e^{\lambda t}$ with $\lambda > 0$, we get

$$\lambda f + \hat{v} \cdot \nabla_x f + \mathbf{F^0} \cdot \nabla_v f = -\mathbf{F} \cdot \nabla_v f^0.$$

An inversion of this equation leaves us with the integral expression

$$f = -(\lambda + (\hat{v}, \mathbf{F^0}) \cdot \nabla_{x,v})^{-1} (\mathbf{F} \cdot \nabla_v f^0) \tag{5.6.4}$$

which depends upon λ as a parameter. By substituting the expression (5.6.4) into Maxwell's equations (5.6.2), f is eliminated as an unknown, and the only unknowns left are ϕ and $\mathbf{A}$. Note that an immediate benefit is that the problem now only involves the spatial variable x, and not the full phase-space variables x, v.

We are therefore left with the task of showing that Maxwell's equations are sat-

isfied with the parameter $\lambda > 0$. Gauss' equation, for instance, becomes

$$(\Delta - \lambda^2)\phi = -\rho = \int f \, dv = -\int (\lambda + (\hat{v}, \mathbf{F^0}) \cdot \nabla_{x,v})^{-1} (\mathbf{F} \cdot \nabla_v f^0) \, dv$$

which is an equation of the form

$$(\Delta - \lambda^2)\phi + \mathcal{K}^\lambda_{--}\phi + \mathcal{K}^\lambda_{-+}\mathbf{A} = 0, \tag{5.6.5}$$

where, for instance,

$$\mathcal{K}^\lambda_{--}\phi = \int (\lambda + (\hat{v}, \mathbf{F^0}) \cdot \nabla_{x,v})^{-1} (\nabla\phi \cdot \nabla_v f^0) \, dv,$$
$$\mathcal{K}^\lambda_{-+}\mathbf{A} = \int (\lambda + (\hat{v}, \mathbf{F^0}) \cdot \nabla_{x,v})^{-1} ((\hat{v} \times (\nabla \times \mathbf{A})) \cdot \nabla_v f^0) \, dv.$$

The rest of Maxwell's equations can be written as

$$(-\boldsymbol{\Delta} + \lambda^2)\mathbf{A} + \mathcal{K}^\lambda_{+-}\phi + \mathcal{K}^\lambda_{++}\mathbf{A} = \mathbf{0}. \tag{5.6.6}$$

(we omit the precise form of these operators here). The system (5.6.5)-(5.6.6) for ϕ and $\mathbf{A}$ turns out to be self-adjoint and is precisely of the form (5.1.1). Exhibiting linear instability, i.e. the existence of a growing mode with rate $\lambda > 0$, is equivalent to solving this system for some $\lambda > 0$. The operator in this system has the form

$$\mathcal{M}^\lambda = \mathcal{A}^\lambda + \mathcal{K}^\lambda = \begin{bmatrix} -\boldsymbol{\Delta} + \lambda^2 & 0 \\ 0 & \Delta - \lambda^2 \end{bmatrix} + \begin{bmatrix} \mathcal{K}^\lambda_{++} & \mathcal{K}^\lambda_{+-} \\ \mathcal{K}^\lambda_{-+} & \mathcal{K}^\lambda_{--} \end{bmatrix}, \quad \lambda > 0.$$

Hence now one would like to show that for some $\lambda > 0$, the operator $\mathcal{M}^\lambda$ has a nontrivial kernel. As this operator is self-adjoint for all $\lambda > 0$, its spectrum lies on the real line. We use this fact to "track" the spectrum as λ varies from 0 to $+\infty$ and find an eigenvalue that crosses through 0. By adding to $\mathcal{A}^\lambda$ the operator

$$\mathcal{W} = \begin{bmatrix} 1 + x^2 & 0 \\ 0 & -1 - x^2 \end{bmatrix}$$

and defining

$$\mathcal{M}^\lambda_\varepsilon = \mathcal{A}^\lambda + \varepsilon\mathcal{W} + \mathcal{K}^\lambda, \quad \lambda > 0, \, \varepsilon > 0$$

we obtain a family of operators with a compact resolvent. This family enjoys the properties that we studied in this chapter. For instance, natural candidates for the projection operators $\mathcal{P}_\pm$ are multiplications by the indicator functions (in the appropriate spaces) onto the (compact) support of the steady-state around which we linearise.

Let us describe the method for finding a nontrivial kernel in a nutshell. It is shown that there exist $0 < \lambda_* < \lambda^* < \infty$ (independent of n and ε) for which the corresponding approximate operators $\mathcal{M}_{\varepsilon,n}^{\lambda_*}$ and $\mathcal{M}_{\varepsilon,n}^{\lambda^*}$ have a different number of negative (and positive) eigenvalues, and therefore due to the continuous dependence of the spectrum (as a set) on the parameter λ there must exist $\lambda_* < \lambda_n < \lambda^*$ for which $\mathcal{M}_{\varepsilon,n}^{\lambda_n}$ has a nontrivial kernel. Since λ_n is a bounded sequence, one can extract a convergent subsequence converging, say, to some $\lambda_\infty \in [\lambda_*, \lambda^*]$. Theorem 5.1.1 is then invoked to show that one can also take the two limits $n \to \infty$ and $\varepsilon \to 0$ to conclude that $\mathcal{M}^{\lambda_\infty}$ has a nontrivial kernel. We refer to Chapter 6 for full details.

Chapter 6

Instabilities of the relativistic Vlasov-Maxwell system on unbounded domains

The relativistic Vlasov-Maxwell system describes the evolution of a collisionless plasma. The problem of linear instability of this system is considered in two physical settings: the so-called "one and one-half" dimensional case, and the three dimensional case with cylindrical symmetry. Sufficient conditions for instability are obtained in terms of the spectral properties of certain Schrödinger operators that act on the spatial variable alone (and not in full phase space). An important aspect of these conditions is that they do not require any boundedness assumptions on the domains, nor do they require monotonicity of the equilibrium.

Acknowledgements

The work in this chapter was done in collaboration with Jonathan Ben≈Artzi and appears in a similar form in [19].

6.1 Introduction

We obtain new linear instability results for plasmas governed by the relativistic Vlasov-Maxwell (RVM) system of equations. The main unknowns are two functions $f^{\pm} = f^{\pm}(t, \boldsymbol{x}, \boldsymbol{v}) \geq 0$ measuring the density of positively and negatively charged particles that at time $t \in [0, \infty)$ are located at the point $\boldsymbol{x} \in \mathbb{R}^d$ and have momentum $\boldsymbol{v} \in \mathbb{R}^d$. The densities $f^{\pm}$ evolve according to the Vlasov equations

$$\frac{\partial f^{\pm}}{\partial t} + \hat{\boldsymbol{v}} \cdot \nabla_{\boldsymbol{x}} f^{\pm} + \mathbf{F}^{\pm} \cdot \nabla_{\boldsymbol{v}} f^{\pm} = 0 \tag{6.1.1}$$

where $\hat{\boldsymbol{v}} = \boldsymbol{v}/\sqrt{1 + |\boldsymbol{v}|^2}$ is the relativistic velocity (the speed of light is taken to be 1 for simplicity) and where $\mathbf{F}^{\pm} = \mathbf{F}^{\pm}(t, \boldsymbol{x}, \boldsymbol{v})$ is the Lorentz force, given by

$$\mathbf{F}^{\pm} = \pm \Big(\mathbf{E} + \mathbf{E}^{ext} + \hat{\boldsymbol{v}} \times (\mathbf{B} + \mathbf{B}^{ext}) \Big)$$

with $\mathbf{E} = \mathbf{E}(t, \boldsymbol{x})$ and $\mathbf{B} = \mathbf{B}(t, \boldsymbol{x})$ being the electric and magnetic fields, respectively, and $\mathbf{E}^{ext}(t, \boldsymbol{x}), \mathbf{B}^{ext}(t, \boldsymbol{x})$ external fields. The *self-consistent* fields obey Maxwell's equations

$$\nabla \cdot \mathbf{E} = \rho, \quad \nabla \cdot \mathbf{B} = 0, \quad \nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}, \quad \nabla \times \mathbf{B} = \mathbf{j} + \frac{\partial \mathbf{E}}{\partial t},$$

where

$$\rho = \rho(t, \boldsymbol{x}) = \int (f^{+} - f^{-}) \, d\boldsymbol{v} \tag{6.1.2}$$

is the charge density and

$$\mathbf{j} = \mathbf{j}(t, \boldsymbol{x}) = \int \hat{\boldsymbol{v}} (f^{+} - f^{-}) \, d\boldsymbol{v} \tag{6.1.3}$$

is the current density. In addition to the speed of light, we have taken all other constants that typically appear in these equations (such as the particle masses) to be 1 so as to keep the notation simple.

Novelty of the results. Let us mention the main novel aspects of our instability results:

Unbounded domains: our problems are set in unbounded domains (as opposed

to domains with boundaries or periodic domains). One consequence is that the spectrum of the Laplacian (which shall appear prominently) has an essential part and there is no spectral gap.

Non-monotone equilibrium: we do not assume that the equilibrium in question is (strongly) monotone (see (6.1.7) below). Many estimates in previous works rely heavily on monotonicity assumptions.

Existence of equilibria: in Section 6.7 we prove the existence of nontrivial equilibria in the *unbounded, compactly supported* $1.5d$ case. Previously, this has been done in the periodic setting by means of a perturbation argument about the trivial solution which is a center (in the dynamical systems sense). The proof here relies on a fixed point argument.

6.1.1 Main results

For the convenience of the reader, we provide the full statements of our results here, although some necessary definitions are too cumbersome. We shall refer to the later sections for these definitions.

The physical setting. As is explained in detail below we consider two problems: the 1.5 *dimensional* case and the 3 *dimensional* case with *cylindrical symmetry*. We shall refer to these two settings as the $1.5d$ *case* and the $3d$ *case*, respectively, for brevity. In a nutshell, we consider these settings because they provide enough structure so that basic existence and uniqueness results hold and because they possess well-known conserved quantities which may be written explicitly.

The equilibrium. The conserved quantities mentioned above – the microscopic energy $e^{\pm}$ and momentum $p^{\pm}$ – are the subject of further discussion below (see (6.1.17) for the $1.5d$ case and (6.1.28) for the $3d$ case), however we stress the fact that they are functions that satisfy the *time-independent* Vlasov equations.

Hence any functions of the form

$$f^{0,\pm}(\boldsymbol{x},\boldsymbol{v}) = \mu^{\pm}(e^{\pm},p^{\pm}) \tag{6.1.4}$$

are equilibria of the corresponding Vlasov equations. The converse statement – that any equilibrium may be written in this form – is called Jeans' theorem [103] (see Remark 6.1.1 below). In Section 6.7 we prove that there exist such equilibria. When there is no room for confusion we simply write $\mu^{\pm}(e,p)$ or $\mu^{\pm}$ instead of $\mu^{\pm}(e^{\pm},p^{\pm})$.

Remark 6.1.1 (Jeans' "theorem")**.** *Jeans' "theorem" is commonly referred to as such in the literature, though it is not (strictly speaking) a "theorem". For instance, for the so-called Vlasov-Einstein system it has been shown to be false [179] while for the gravitational Vlasov-Poisson system it has indeed been proven rigorously [14]. As far as the authors know, there are no other proofs (or disproofs), though it is often easy to give a formal justification of this "theorem" by counting degrees of freedom. Indeed, if one can argue that (due to symmetries) an equilibrium $f^0(\boldsymbol{x},\boldsymbol{v})$ can have at most $d \in \{1,\dots,6\}$ degrees of freedom and find d conserved quantities (that is, d quantities that are constant along the Vlasov flow), then formally it could be argued that f^0 may be rewritten as a function of these quantities.*

We shall always assume that

$$0 \leq f^{0,\pm}(\boldsymbol{x},\boldsymbol{v}) \in C^1 \text{ have compact support } \Omega \text{ in the } \boldsymbol{x}\text{-variable.}$$

Again, the existence of such equilibria is the subject of Section 6.7. Note that in the $3d$ case, Ω must be cylindrically symmetric. In addition, we must assume that

$$\begin{aligned} &\text{there exist weight functions } w^{\pm} = c(1+|e^{\pm}|)^{-\alpha} \\ &\text{where } \alpha > \text{dimension of momentum space, and } c > 0, \end{aligned} \tag{6.1.5}$$

such that the integrability condition

$$\left(\left| \frac{\partial \mu^{\pm}}{\partial e} \right| + \left| \frac{\partial \mu^{\pm}}{\partial p} \right| \right) (e^{\pm},p^{\pm}) < w^{\pm}(e^{\pm}) \tag{6.1.6}$$

holds[1]. This implies that $\int \left(|\mu_e^{\pm}| + |\mu_p^{\pm}|\right) d\boldsymbol{x}d\boldsymbol{v} < \infty$ in both the $1.5d$ and $3d$ cases, where we have abbreviated the writing of the partial derivatives of $\mu^{\pm}$. This abbreviated notation shall be used throughout the chapter. It is often assumed that

$$\mu_e^{\pm} < 0 \quad \text{whenever } \mu^{\pm} > 0. \tag{6.1.7}$$

We call this a *strong monotonicity condition.* **We do <u>not</u> make any such assumption.** Monotonicity assumptions are natural both in the study of Vlasov systems [130, 132, 160, 125, 151] and the $2d$ Euler equations [39, 191] as they typically lead to stability. A famous exception to this rule is Penrose's result [167] often referred to as the "Penrose criterion". In many of the aforementioned works monotonicity assumptions play an important role throughout. It is therefore not always clear if such conditions can be relaxed, or altogether dropped, as this would require extensive reformulation of the existing proofs.

The main results. To facilitate the understanding of our main results we state them now, trying not obscure the big picture with technical details. Hence we attempt to only extract those aspects of the statements that are crucial for understanding, while referring to later sections for some additional definitions. First we define our precise notion of instability:

Definition 6.1.1 (Spectral instability)**.** *We say that a given equilibrium $\mu^{\pm}$ is* spectrally unstable*, if the system linearised around it has a purely growing mode solution of the form*

$$\left(e^{\lambda t} f^{\pm}(\boldsymbol{x}, \boldsymbol{v}), e^{\lambda t}\mathbf{E}(\boldsymbol{x}), e^{\lambda t}\mathbf{B}(\boldsymbol{x})\right), \quad \lambda > 0. \tag{6.1.8}$$

We also need the following definition:

Definition 6.1.2. *Given a self-adjoint operator (bounded or unbounded) $\mathcal{A}$, we denote by* $\mathrm{neg}(\mathcal{A})$ *the number of negative eigenvalues (counting multiplicity) that it has whenever there is a finite number of such eigenvalues.*

[1] This condition is one of smoothness and decay for large energies of the particle density. In the presence of a confining potential, equilibria can be shown to exist under the even more stringent condition that $\mu^{\pm}$ has compact support (see Section 6.7 for a construction in the 1.5d case.)

In our first result we obtain a sufficient condition for spectral instability of equilibria in the $1.5d$ case. The condition is expressed in terms of spectral properties of certain operators that act on functions of the spatial variable alone.

Theorem 6.1.1 (Spectral instability: $1.5d$ case)**.** *Let $f^{0,\pm}(x,\boldsymbol{v}) = \mu^{\pm}(e,p)$ be an equilibrium of the $1.5d$ system (6.1.15) satisfying (6.1.6). There exist self-adjoint Schrödinger operators $\mathcal{A}_1^0$ and $\mathcal{A}_2^0$ and a bounded operator $\mathcal{B}^0$ (all defined in (6.1.23)) acting only on functions of the spatial variable (and not the momentum variable) such that the equilibrium is spectrally unstable if 0 is not in the spectrum of $\mathcal{A}_1^0$ and*

$$\operatorname{neg}\left(\mathcal{A}_2^0 + \left(\mathcal{B}^0\right)^* \left(\mathcal{A}_1^0\right)^{-1} \mathcal{B}^0\right) > \operatorname{neg}\left(\mathcal{A}_1^0\right). \tag{6.1.9}$$

The second result provides a similar statement in the $3d$ case with cylindrical symmetry, as discussed in further detail in Section 6.1.4 below.

Theorem 6.1.2 (Spectral instability: $3d$ case)**.** *Let $f^{0,\pm}(\boldsymbol{x},\boldsymbol{v}) = \mu^{\pm}(e,p)$ be a cylindrically symmetric equilibrium of the RVM system satisfying (6.1.6). There exist self-adjoint operators $\widetilde{\mathcal{A}}_1^0$,$\widetilde{\mathcal{A}}_2^0$ and $\widetilde{\mathcal{A}}_3^0$ and a bounded operator $\widetilde{\mathcal{B}}_1^0$ (all defined in (6.1.33)) acting in the spatial variable alone (and not the momentum variable) such that the equilibrium is spectrally unstable if 0 is not an L^6-eigenvalue of $\widetilde{\mathcal{A}}_3^0$ (see Definition 6.1.3 below), 0 is not an eigenvalue of $\widetilde{\mathcal{A}}_1^0$ (0 will always lie in the essential spectrum of $\widetilde{\mathcal{A}}_1^0$, but this is not the same as 0 being an eigenvalue) and*

$$\operatorname{neg}\left(\widetilde{\mathcal{A}}_2^0 + \left(\widetilde{\mathcal{B}}_1^0\right)^* \left(\widetilde{\mathcal{A}}_1^0\right)^{-1} \widetilde{\mathcal{B}}_1^0\right) > \operatorname{neg}\left(\widetilde{\mathcal{A}}_1^0\right) + \operatorname{neg}\left(\widetilde{\mathcal{A}}_3^0\right). \tag{6.1.10}$$

Let us make precise the notion of an L^6-eigenvalue.

Definition 6.1.3 (L^6-eigenvalue)**.** *We say that $\lambda \in \mathbb{R}$ is an L^6-eigenvalue of a self-adjoint Schrödinger operator $\boldsymbol{\mathcal{A}} : H^2(\mathbb{R}^n;\mathbb{R}^m) \subset L^2(\mathbb{R}^n;\mathbb{R}^m) \to L^2(\mathbb{R}^n;\mathbb{R}^m)$ given by $\boldsymbol{\mathcal{A}} = -\boldsymbol{\Delta} + \boldsymbol{\mathcal{K}}$, if there exists a function $0 \neq \boldsymbol{u}^\lambda \in H^2_{loc}(\mathbb{R}^n;\mathbb{R}^m) \cap L^6(\mathbb{R}^n;\mathbb{R}^m)$, with $\boldsymbol{\nabla u}^\lambda \in L^2(\mathbb{R}^n;\mathbb{R}^m)^n$, such that $\boldsymbol{\mathcal{A}u}^\lambda = \lambda \boldsymbol{u}^\lambda$ in the sense of distributions. The function $\boldsymbol{u}^\lambda$ is called an L^6-eigenfunction.*

Remark 6.1.2. *We remark that L^6 is a natural space to consider in three dimensions due to the embedding $H^1(\Omega) \hookrightarrow L^6(\Omega)$ where $\Omega \subset \mathbb{R}^3$ is a bounded and smooth domain. In fact, any function which decays at infinity and whose first*

derivatives are square integrable, also belongs to $L^6(\mathbb{R}^3)$*. Therefore this is a natural condition for the potential formulation of Maxwell's equations where there is no physical reason for the potentials to be square integrable, but the condition that the fields are square integrable corresponds to the physical condition that the electromagnetic fields have finite energy.*

The proofs of these two theorems appear in Section 6.4.1 and Section 6.4.2, receptively. Let us describe the main ideas of the proofs. For brevity, we omit the $\pm$ signs distinguishing between positively and negatively charged particles in this paragraph. Since we are interested in linear instability, we linearise the Vlasov equation around f^0. The only nonlinear term is the forcing term $\mathbf{F} \cdot \nabla_v f$, so that the linearisation of (6.1.1) becomes

$$\frac{\partial f}{\partial t} + \hat{\boldsymbol{v}} \cdot \nabla_{\boldsymbol{x}} f + \mathbf{F}^0 \cdot \nabla_{\boldsymbol{v}} f = -\mathbf{F} \cdot \nabla_{\boldsymbol{v}} f^0 \tag{6.1.11}$$

where $\mathbf{F}^0$ is the equilibrium self consistent Lorentz force and $\mathbf{F}$ is the linearised Lorentz force. We make the following growing-mode ansatz:

$$\begin{aligned} &\textbf{Ansatz}\text{: the perturbations } (f, \mathbf{E}, \mathbf{B}) \text{ have} \\ &\qquad\qquad \text{time dependence } e^{\lambda t}, \text{ where } \lambda > 0. \end{aligned} \tag{6.1.12}$$

Equation (6.1.11) can therefore be written as

$$(\lambda + \mathcal{D})\, f = -\mathbf{F} \cdot \nabla_{\boldsymbol{v}} f^0 \tag{6.1.13}$$

where

$$\mathcal{D} = \hat{\boldsymbol{v}} \cdot \nabla_{\boldsymbol{x}} + \mathbf{F}^0 \cdot \nabla_{\boldsymbol{v}} \tag{6.1.14}$$

is the linearised Vlasov transport operator. We then invert (6.1.13) by applying $\lambda\,(\lambda + \mathcal{D})^{-1}$, which is an ergodic averaging operator along the trajectories of $\mathcal{D}$ (depending upon λ as a parameter), see [16, Eq. (2.10)]. Hence we obtain an expression of f in terms of a certain average of the right hand side $-\mathbf{F} \cdot \nabla_{\boldsymbol{v}} f^0$ depending upon the parameter λ (see (6.3.2) and (6.3.9)). This expression for f is substituted into Maxwell's equations through the charge and current densities, resulting in a system of (elliptic) equations for the spatial variable alone (recall that the momentum variable is integrated in the expressions for ρ and $\mathbf{j}$). The number of linearly independent equations is less than one would expect due to

the imposed symmetries. However, in both cases the equations can be written so that they form a self-adjoint system denoted $\boldsymbol{\mathcal{M}}^\lambda$ (see (6.3.5) for the $1.5d$ case and (6.3.14) for the $3d$ case) that has the general form

$$\boldsymbol{\mathcal{M}}^\lambda = \begin{bmatrix} -\Delta + 1 & 0 \\ 0 & \Delta - 1 \end{bmatrix} + \boldsymbol{\mathcal{K}}^\lambda$$

acting on the electric and magnetic potentials, where $\boldsymbol{\mathcal{K}}^\lambda$ is a uniformly bounded and symmetric family.

The problem then reduces to showing that the equation $\boldsymbol{\mathcal{M}}^\lambda \boldsymbol{u} = 0$ has a nontrivial solution for some value of $\lambda > 0$. The difficulty here is twofold: first, the spectrum of $\boldsymbol{\mathcal{M}}^\lambda$ is unbounded (not even semi-bounded) and includes essential spectrum extending to both $+\infty$ and $-\infty$. Second, for each λ, the operator $\boldsymbol{\mathcal{M}}^\lambda$ has a different spectrum: one must analyse a family of spectra that depends upon the parameter λ. In the previous Chapter 5 we address the following related problem:

Problem 6.1.1. *Consider the family of self-adjoint unbounded operators*

$$\boldsymbol{\mathcal{M}}^\lambda = \boldsymbol{\mathcal{A}} + \boldsymbol{\mathcal{K}}^\lambda = \begin{bmatrix} -\Delta + 1 & 0 \\ 0 & \Delta - 1 \end{bmatrix} + \begin{bmatrix} \mathcal{K}^\lambda_{++} & \mathcal{K}^\lambda_{+-} \\ \mathcal{K}^\lambda_{-+} & \mathcal{K}^\lambda_{--} \end{bmatrix}, \quad \lambda \in [0,1]$$

acting in (an appropriate subspace of) $L^2(\mathbb{R}^d) \oplus L^2(\mathbb{R}^d)$*, where* $\{\boldsymbol{\mathcal{K}}^\lambda\}_{\lambda\in[0,1]}$ *is a uniformly bounded, symmetric and strongly continuous family. Is it possible to construct explicit finite-dimensional symmetric approximations of* $\boldsymbol{\mathcal{M}}^\lambda$ *whose spectrum in* $(-1,1)$ *converges to that of* $\boldsymbol{\mathcal{M}}^\lambda$ *for all* λ *simultaneously?*

A solution to this problem allows us to construct finite-dimensional approximations to $\boldsymbol{\mathcal{M}}^\lambda$. We discuss this problem in Section 6.2.2. The conditions (6.1.9) and (6.1.10) appearing in the main theorems above translate into analogous conditions on the approximations, and those, in turn, guarantee the existence of a nontrivial *approximate* solution. Since the approximate problems converge (in an appropriate sense) to the original problem, this is enough to complete the proof. A crucial ingredient is the self-adjointness of all operators: this guarantees that the spectrum is restricted to the real line. The strategy is to "track" eigenvalues as a function of the parameter λ and conclude that they cross through 0 for some

value $\lambda > 0$. To do so, we require knowledge of the spectrum of the operator $\boldsymbol{\mathcal{M}}^\lambda$ for small positive λ, which is obtained from the assumptions (6.1.9) and (6.1.10), and for large λ which arises naturally from the form of the problem.

Yet even with a solution to this problem at hand, some difficulties remain. In the cylindrically symmetric case there is a geometric difficulty. Namely, cylindrical symmetries must be respected, a fact that requires a somewhat more cumbersome functional setup. In particular, the singular nature of the coordinate chart along the axis of symmetry requires special attention. To circumvent this issue we shall do all computation in Cartesian coordinates, and use carefully chosen subspaces to decompose the magnetic potential. The second difficulty is the lack of a spectral gap, which is due to the unbounded nature of the problem in physical space. As a consequence, the dependence of the spectrum of $\boldsymbol{\mathcal{M}}^\lambda$ on λ is delicate, especially as $\lambda \to 0$, and needs careful consideration.

6.1.2 Previous results

Existence theory. The main difficulty in attaining existence results for Vlasov systems is in controlling particle acceleration due to the nonlinear forcing term. Hence existence and uniqueness has only been proved under various symmetry assumptions. In [73] global existence in the $1.5d$ case was established and in [72] the cylindrically symmetric case was considered. Local existence and uniqueness is due to [203].

Stability theory. One of the important early results on (linear) stability of plasmas is that of Penrose [167]. Two notable later results are [96, 139]. We refer to [16] for additional references. The current result continues a program initiated by Lin and Strauss [130, 132] and continued by the first author [16, 17]. In [130, 132] the equilibria were always assumed to be strongly monotone, in the sense of (6.1.7). This added sign condition (which is widely used within the physics community, and is believed to be crucial for stability results) allowed them to obtain in [130] a linear *stability* criterion which was complemented by a linear *instability* criterion in [132]. Combined, these two results produced a necessary and sufficient criterion for stability in the following sense: there exists

a Schrödinger operator $\mathcal{L}^0$ acting only in the spatial variable, such that $\mathcal{L}^0 \geq 0$ implies spectral stability, and $\mathcal{L}^0 \ngeq 0$ implies spectral instability. In [16, 17] the monotonicity assumption was removed, which mainly impacted the ability to obtain stability results. The instability results are similar to the ones of Lin and Strauss, though the author only considers the $1.5d$ case with periodicity. This is due to his methods which crucially require a Poincaré inequality. We remark that our results recover all previous results when one restricts to the monotone case.

6.1.3 The $1.5d$ case

First we treat the so-called $1.5d$ case, which is the lowest dimensional setting that permits nontrivial electromagnetic fields. In this setting, the plasma is assumed to have certain symmetries in phase-space that render the distribution function to be a function of only one spatial variable x and two momentum variables $\boldsymbol{v} = (v_1, v_2)$, with v_1 being aligned with x. The only non-trivial components of the fields are the first two components of the electric field and the third component of the magnetic field: $\mathbf{E} = (E_1, E_2, 0)$ and $\mathbf{B} = (0, 0, B)$, and similarly for the equilibrium fields. The RVM system becomes the following system of scalar equations

$$\partial_t f^\pm + \hat{v}_1 \partial_x f^\pm \pm (E_1 + \hat{v}_2 B)\partial_{v_1} f^\pm \pm (E_2 - \hat{v}_1 B)\partial_{v_2} f^\pm = 0 \tag{6.1.15a}$$

$$\partial_t E_1 = -j_1 \tag{6.1.15b}$$

$$\partial_t E_2 + \partial_x B = -j_2 \tag{6.1.15c}$$

$$\partial_x E_1 = \rho \tag{6.1.15d}$$

$$\partial_t B = -\partial_x E_2 \tag{6.1.15e}$$

where ρ and j_1, j_2 are defined by (6.1.2) and (6.1.3).

6.1.3.1 Equilibrium

In Section 6.7 we prove that there exist equilibria $f^{0,\pm}(x, \boldsymbol{v})$ which can be written as functions of the energy and momentum

$$f^{0,\pm}(x, \boldsymbol{v}) = \mu^\pm(e^\pm, p^\pm) \tag{6.1.16}$$

as in (6.1.4). The energy and momentum are defined as:

$$e^{\pm} = \langle \boldsymbol{v} \rangle \pm \phi^0(x) \pm \phi^{ext}(x), \qquad p^{\pm} = v_2 \pm \psi^0(x) \pm \psi^{ext}(x) \tag{6.1.17}$$

where $\langle \boldsymbol{v} \rangle = \sqrt{1 + |\boldsymbol{v}|^2}$, and ϕ^0 and ψ^0 are the equilibrium electric and magnetic potentials[2] (both scalar, in this case), respectively:

$$\partial_x \phi^0 = -E_1^0, \qquad \partial_x \psi^0 = B^0 \tag{6.1.18}$$

and similarly ϕ^{ext} and ψ^{ext} are external electric and magnetic potentials that give rise to external fields E_1^{ext} and B^{ext}. It is a straightforward calculation to verify that $e^{\pm}$ and $p^{\pm}$ are conserved quantities of the Vlasov flow, i.e. that $\mathcal{D}_{\pm} e^{\pm} = \mathcal{D}_{\pm} p^{\pm} = 0$, where the operators $\mathcal{D}_{\pm}$ are defined below, (6.1.20).

6.1.3.2 Linearisation

Let us discuss the linearisation of the Vlasov-Maxwell system (6.1.15) about a steady-state solution $(f^{0,\pm}, \mathbf{E}^0, \mathbf{B}^0)$. Using ansatz (6.1.12) and Jeans' theorem (6.1.4), together with (6.1.17) and (6.1.18) the linearised system becomes:

$$(\lambda + \mathcal{D}_{\pm}) f^{\pm} = \mp \mu_e^{\pm} \hat{v}_1 E_1 \pm \mu_p^{\pm} \hat{v}_1 B \mp (\mu_e^{\pm} \hat{v}_2 + \mu_p^{\pm}) E_2 \tag{6.1.19a}$$

$$\lambda E_1 = -j_1 \tag{6.1.19b}$$

$$\lambda E_2 + \partial_x B = -j_2 \tag{6.1.19c}$$

$$\partial_x E_1 = \rho \tag{6.1.19d}$$

$$\lambda B = -\partial_x E_2, \tag{6.1.19e}$$

where

$$\mathcal{D}_{\pm} = \hat{v}_1 \partial_x \pm (E_1^0 + E_1^{ext} + \hat{v}_2 (B^0 + B^{ext})) \partial_{v_1} \pm (E_2^0 - \hat{v}_2 (B^0 + B^{ext})) \partial_{v_2} \tag{6.1.20}$$

are the linearised transport operators as in (6.1.14), and

$$\rho = \int (f^+ - f^-)\, d\boldsymbol{v}, \qquad j_i = \int \hat{v}_i (f^+ - f^-)\, d\boldsymbol{v}$$

[2] Note that E_2^0 vanishes for any equilibrium. This can be deduced from (6.1.15). We provide the details in appendix for completeness (see Lemma 6.A.1.)

are the charge and current densities, respectively.

We now construct electric and magnetic potentials ϕ and ψ, respectively, as in (6.1.18). Equation (6.1.19b) implies that E_1 has the same spatial support as j_1 which is a moment of $f^\pm$ which, in turn, has the same x support as $\mu^\pm$ (this can be seen from (6.1.19a) for instance). We deduce that E_1 is compactly supported in Ω and choose an electric potential $\phi \in H^2(\Omega)$ such that $E_1 = -\partial_x \phi$ in Ω and $E_1 = 0$ outside Ω. Since E_1 vanishes at the boundary of Ω, we must impose Neumann boundary conditions on ϕ, and as E_1 depends only on the derivative of ϕ we may impose that ϕ has mean zero. The magnetic potential ψ is chosen to satisfy $B = \partial_x \psi$ and $E_2 = -\lambda\psi$ (this is due to (6.1.19e)). Then the remaining Maxwell's equations (6.1.19b)-(6.1.19d) become

$$\lambda \partial_x \phi = -\lambda E_1 = j_1 \qquad \text{in } \Omega \tag{6.1.21a}$$

$$(-\partial_x^2 + \lambda^2)\psi = -\partial_x B - \lambda E_2 = j_2 \qquad \text{in } \mathbb{R} \tag{6.1.21b}$$

$$-\partial_x^2 \phi = \partial_x E_1 = \rho \qquad \text{in } \Omega \tag{6.1.21c}$$

where (6.1.21c) is complemented by the Neumann boundary condition

$$-\partial_x \phi = E_1 = 0 \text{ on } \partial\Omega.$$

The linearised Vlasov equations can now be written as

$$\begin{aligned}(\lambda + \mathcal{D}_\pm) f^\pm &= \pm \mu_e^\pm \hat{v}_1 \partial_x \phi \pm \mu_p^\pm \hat{v}_1 \partial_x \psi \pm \lambda(\mu_e^\pm \hat{v}_2 + \mu_p^\pm)\psi \\ &= \pm \mu_e^\pm \mathcal{D}_\pm \phi \pm \mu_p^\pm \mathcal{D}_\pm \psi \pm \lambda(\mu_e^\pm \hat{v}_2 + \mu_p^\pm)\psi\end{aligned} \tag{6.1.22}$$

where we have used the fact that $\mathcal{D}_\pm u = \hat{v}_1 \partial_x u$ if u is a function of x only.

Now let us specify the functional spaces that we shall use. For the scalar potential ϕ we define the space

$$L_0^2(\Omega) := \left\{ f \in L^2(\Omega) \ : \ \int_\Omega f = 0 \right\}$$

while for the magnetic potential ψ we simply use $L^2(\mathbb{R})$, the standard space of square integrable functions. We denote by $H^k(\mathbb{R})$ (resp. $H^k(\Omega)$) the usual Sobolev space of functions whose first k derivatives are in $L^2(\mathbb{R})$ (resp. $L^2(\Omega)$).

Moreover, we naturally define

$$H_0^k(\Omega) := \left\{ f \in H^k(\Omega) \ : \ \int_\Omega f = 0 \right\}$$

and the corresponding version which imposes Neumann boundary conditions

$$H_{0,n}^k(\Omega) := \left\{ f \in H_0^k(\Omega) \ : \ \partial_x f = 0 \text{ on } \partial\Omega \right\}.$$

Finally, to allow us to consider functions that do not decay at infinity we use the conditions (6.1.5) and (6.1.6) to define weighted spaces $\mathfrak{L}_\pm$ as follows: we take the closure of the smooth and compactly supported functions of $(x, \boldsymbol{v})$ (with the x support contained in Ω) under the weighted-L^2 norm given by

$$\|u\|_{\mathfrak{L}_\pm}^2 = \int_{\Omega\times\mathbb{R}^2} w^\pm |u|^2 \, d\boldsymbol{v}dx$$

and we denote the inner product by $\langle \cdot, \cdot \rangle_{\mathfrak{L}_\pm}$. In particular we can view any function $u(x) \in L^2(\Omega)$ or $L_0^2(\Omega)$ as being in $\mathfrak{L}_\pm$ by considering u as a function of $(x, \boldsymbol{v})$ which does not depend on $\boldsymbol{v}$. We can extend this to functions in $L^2(\mathbb{R})$ by multiplying them by the characteristic function $\mathbb{1}_\Omega$ of the set Ω. Hence the function $\mathbb{1}_\Omega$ itself may be regarded as an element in $\mathfrak{L}_\pm$.

6.1.3.3 The operators

Finally, we define the operators used in the statement of Theorem 6.1.1. First define the following projection operators:

Definition 6.1.4 (Projection operators)**.** *We define $\mathcal{Q}_\pm^0$ to be the orthogonal projection operators in $\mathcal{L}_\pm$ onto* $\ker(\mathcal{D}_\pm)$.

Remark 6.1.3. *Although this definition makes reference to the spaces $\mathfrak{L}_\pm$, the operators $\mathcal{Q}_\pm^0$ do not depend on the exact choice of weight functions $w^\pm$. This may be seen by writing $(\mathcal{Q}_\pm^0 h)(x, \boldsymbol{v})$ as the pointwise limit of ergodic averages along trajectories (see Remark 6.3.1 and Lemma 6.6.1).*

This allows us to define the following operators acting on functions of the spatial

variable x, not the full phase-space variables:

$$\mathcal{A}_1^0 h = -\partial_x^2 h + \int \sum_{\pm} \mu_e^{\pm} (\mathcal{Q}_{\pm}^0 - 1) h \, d\boldsymbol{v} \tag{6.1.23a}$$

$$\mathcal{A}_2^0 h = -\partial_x^2 h - \left(\sum_{\pm} \int \mu_p^{\pm} \hat{v}_2 \, d\boldsymbol{v} \right) h - \int \sum_{\pm} \hat{v}_2 \mu_e^{\pm} \mathcal{Q}_{\pm}^0 \, [\hat{v}_2 h] \, d\boldsymbol{v} \tag{6.1.23b}$$

$$\mathcal{B}^0 h = \left(\int \sum_{\pm} \mu_p^{\pm} \, d\boldsymbol{v} \right) h + \int \sum_{\pm} \mu_e^{\pm} \mathcal{Q}_{\pm}^0 \, [\hat{v}_2 h] \, d\boldsymbol{v} \tag{6.1.23c}$$

$$\left(\mathcal{B}^0 \right)^* h = \left(\int \sum_{\pm} \mu_p^{\pm} \, d\boldsymbol{v} \right) h + \int \sum_{\pm} \mu_e^{\pm} \hat{v}_2 \mathcal{Q}_{\pm}^0 h \, d\boldsymbol{v}. \tag{6.1.23d}$$

Their precise properties are discussed in Section 6.6.1. For future reference, we mention the important identity

$$\int_{\mathbb{R}} \left(\mu_p^{\pm} + \hat{v}_2 \mu_e^{\pm} \right) dv_2 = 0 \tag{6.1.24}$$

which is due to the fact that $\frac{\partial \mu^{\pm}}{\partial v_2} = \mu_e^{\pm} \hat{v}_2 + \mu_p^{\pm}$.

6.1.4 The cylindrically symmetric case

Since notation can be confusing when multiple coordinate systems are in use, we start this section by making clear what our conventions are.

Vector transformations and notational conventions. We let $\boldsymbol{x} = (x, y, z) = x\mathbf{e}_1 + y\mathbf{e}_2 + z\mathbf{e}_3$ denote the representation of the point $\boldsymbol{x} \in \mathbb{R}^3$ in terms of the standard Cartesian coordinates. We define the usual cylindrical coordinates as

$$r = \sqrt{x^2 + y^2}, \qquad \theta = \arctan(y/x), \qquad z = z$$

and the local cylindrical coordinates as

$$\mathbf{e}_r = r^{-1}(x, y, 0), \qquad \mathbf{e}_\theta = r^{-1}(-y, x, 0), \qquad \mathbf{e}_z = (0, 0, 1).$$

By *cylindrically symmetric* we mean that in what follows no quantity depends upon θ (which does not imply that the θ component is zero!). When writing $f(\boldsymbol{x})$ we mean the value of the function f at the point $\boldsymbol{x}$ in Cartesian coordinates. We shall often abuse notation and write $f(r, \theta, z)$ to mean the value of f at the point (r, θ, z) in cylindrical coordinates. A point $\boldsymbol{v} \in \mathbb{R}^3$ in momentum space shall either be expressed in Cartesian coordinates as

$$\boldsymbol{v}_{xyz} = (v_x, v_y, v_z) = (\boldsymbol{v} \cdot \mathbf{e}_1)\mathbf{e}_1 + (\boldsymbol{v} \cdot \mathbf{e}_2)\mathbf{e}_2 + (\boldsymbol{v} \cdot \mathbf{e}_3)\mathbf{e}_3$$

or in cylindrical coordinates *(depending upon the point $\boldsymbol{x} \in \mathbb{R}^3$ in physical space)* as

$$\boldsymbol{v}_{rz\theta} = (v_r, v_\theta, v_z) = (\boldsymbol{v} \cdot \mathbf{e}_r)\mathbf{e}_r + (\boldsymbol{v} \cdot \mathbf{e}_\theta)\mathbf{e}_\theta + (\boldsymbol{v} \cdot \mathbf{e}_z)\mathbf{e}_z.$$

However we shall not be too pedantic about this notation, and shall use $\boldsymbol{v}$ (rather than $\boldsymbol{v}_{xyz}$ or $\boldsymbol{v}_{rz\theta}$) when there's no reason for confusion.

A vector-valued function $\mathbf{F}$ shall be understood to be represented in Cartesian coordinates. That is, unless otherwise said, $\mathbf{F} = (F_x, F_y, F_z) = F_x\mathbf{e}_1 + F_y\mathbf{e}_2 + F_z\mathbf{e}_3$. Its expression in cylindrical coordinates shall typically be written as $\mathbf{F} = F_r\mathbf{e}_r + F_\theta\mathbf{e}_\theta + F_z\mathbf{e}_z$.

Differential operators. Partial derivatives in Cartesian coordinates are written as ∂_x, ∂_y and ∂_z, while in cylindrical coordinates they are $\partial_r, \partial_\theta$ and ∂_z. They transform in the standard manner. It is important to note that since we work in phase space, we shall require derivatives with respect to $\boldsymbol{v}$ as well. One important factor appearing in the Vlasov equation is $\hat{\boldsymbol{v}} \cdot \nabla_{\boldsymbol{x}}$, which transforms as

$$\begin{aligned}(\hat{\boldsymbol{v}} \cdot \nabla_{\boldsymbol{x}})h &= \hat{v}_x\partial_x h + \hat{v}_y\partial_y h + \hat{v}_z\partial_z h \\ &= \hat{v}_r\partial_r h + r^{-1}\hat{v}_\theta\partial_\theta h + \hat{v}_z\partial_z h \\ &= \hat{v}_r\partial_r h + r^{-1}\hat{v}_\theta(v_\theta\partial_{v_r} h - v_r\partial_{v_\theta} h) + \hat{v}_z\partial_z h.\end{aligned}$$

However the next term in the Vlasov equation transforms "neatly":

$$\begin{aligned}(\mathbf{F}\cdot\nabla_{\boldsymbol{v}})h &= F_x\partial_{v_x}h + F_y\partial_{v_y}h + F_z\partial_{v_z}h \\ &= (F_x\cos\theta + F_y\sin\theta)\partial_{v_r}h + (-F_x\sin\theta + F_y\cos\theta)\partial_{v_\theta}h + F_z\partial_{v_z}h \\ &= F_r\partial_{v_r}h + F_\theta\partial_{v_\theta}h + F_z\partial_{v_z}h.\end{aligned}$$

6.1.4.1 The Lorenz gauge

As opposed to the system (6.1.19), here we do not get a system of scalar equations. It is well known that there is some freedom in defining the electromagnetic potentials φ (we use φ in the cylindrically symmetric case rather than ϕ to avoid confusion) and $\mathbf{A}$, satisfying

$$\partial_t\mathbf{A} + \nabla\varphi = -\mathbf{E}, \qquad \nabla\times\mathbf{A} = \mathbf{B}.$$

Remark 6.1.4. *Whenever the differential operator ∇ appears without any subscript, it is understood to be $\nabla_{\boldsymbol{x}}$, that is, the operator $(\partial_x, \partial_y, \partial_z)$ acting on functions of the spatial variable in Cartesian coordinates. The same holds for the Laplacian Δ.*

We choose to impose the *Lorenz gauge* $\nabla\cdot\mathbf{A} + \frac{\partial\varphi}{\partial t} = 0$, hence transforming Maxwell's equations into the hyperbolic system

$$\frac{\partial^2}{\partial t^2}\varphi - \Delta\varphi = \rho, \tag{6.1.25a}$$

$$\frac{\partial^2}{\partial t^2}\mathbf{A} - \boldsymbol{\Delta}\mathbf{A} = \mathbf{j}. \tag{6.1.25b}$$

We remark that this is not unique to the cylindrically symmetric case, and the expressions above are written in Cartesian coordinates.

6.1.4.2 Equilibrium and the linearised system

We define the steady-state potentials $\varphi^0 : \mathbb{R}^3 \to \mathbb{R}$ and $\mathbf{A}^0 : \mathbb{R}^3 \to \mathbb{R}^3$ through

$$\nabla\varphi^0 = -\mathbf{E}^0, \qquad \nabla\times\mathbf{A}^0 = \mathbf{B}^0 \tag{6.1.26}$$

which become

$$\mathbf{E}^0 = -\partial_r \varphi^0 \mathbf{e}_r - \partial_z \varphi^0 \mathbf{e}_z, \qquad \mathbf{B}^0 = -\partial_z A^0_\theta \mathbf{e}_r + \frac{1}{r}\partial_r (r A^0_\theta)\mathbf{e}_z. \tag{6.1.27}$$

The energy and momentum may be defined (analogously to (6.1.17)) as

$$\begin{aligned} e^\pm_{cyl} &= \langle \boldsymbol{v} \rangle \pm \varphi^0(r,z) \pm \varphi^{ext}(r,z), \\ p^\pm_{cyl} &= r(v_\theta \pm A^0_\theta(r,z) \pm A^{ext}_\theta(r,z)), \end{aligned} \tag{6.1.28}$$

where we remind that $\langle \boldsymbol{v} \rangle = \sqrt{1 + |\boldsymbol{v}_{xyz}|^2}$. It is straightforward to verify that they are indeed conserved along the Vlasov flow (which is given by the integral curves of the differential operators $\widetilde{\mathcal{D}}_\pm$, defined in (6.1.30) below). To maintain simple notation we won't insist on writing the *cyl* subscript where it is clear which energy and momentum are meant. The external fields are also assumed to be cylindrically symmetric and their potentials satisfy equations analogous to (6.1.27). We recall (6.1.4), namely that any equilibrium is assumed to be of the form

$$f^{0,\pm}(\boldsymbol{x}, \boldsymbol{v}) = \mu^\pm(e^\pm, p^\pm).$$

Considering the Lorenz gauge, and applying the ansatz (6.1.12) and Jeans' theorem (6.1.4) the linearisation of the RVM system about a steady-state solution $(f^{0,\pm}, \mathbf{E}^0, \mathbf{B}^0)$ is

$$(\lambda + \widetilde{\mathcal{D}}_\pm) f^\pm = \pm(\lambda + \widetilde{\mathcal{D}}_\pm)(\mu^\pm_e \varphi + r\mu^\pm_p (\mathbf{A} \cdot \mathbf{e}_\theta)) \pm \lambda \mu^\pm_e (-\varphi + \mathbf{A} \cdot \hat{\boldsymbol{v}}) \tag{6.1.29a}$$

$$\lambda^2 \varphi - \Delta \varphi = \int (f^+ - f^-)\, d\boldsymbol{v} \tag{6.1.29b}$$

$$\lambda^2 \mathbf{A} - \boldsymbol{\Delta} \mathbf{A} = \int (f^+ - f^-) \hat{\boldsymbol{v}}\, d\boldsymbol{v} \tag{6.1.29c}$$

where

$$\begin{aligned} \widetilde{\mathcal{D}}_\pm &= \hat{\boldsymbol{v}}_{xyz} \cdot \nabla_{\boldsymbol{x}} \pm (\mathbf{E}^0 + \mathbf{E}^{ext} + \hat{\boldsymbol{v}}_{xyz} \times (\mathbf{B}^0 + \mathbf{B}^{ext})) \cdot \nabla_{\boldsymbol{v}} \\ &= \hat{v}_r \partial_r + \hat{v}_z \partial_z + (\pm E^0_r \pm E^{ext}_r \pm \hat{v}_\theta (B^0_z + B^{ext}_z) + r^{-1} \hat{v}_\theta v_\theta) \partial_{v_r} \\ &\quad + (\pm \hat{v}_z (B^0_r + B^{ext}_r) \mp \hat{v}_r (B^0_z + B^{ext}_z) + r^{-1} \hat{v}_\theta v_r) \partial_{v_\theta} \\ &\quad \pm (E^0_z + E^{ext}_z + \hat{v}_\theta (B^0_r + B^{ext}_r)) \partial_{v_z} \end{aligned} \tag{6.1.30}$$

are the linearised transport operators. The Lorenz gauge condition under the growing mode ansatz is

$$\nabla \cdot \mathbf{A} + \lambda\varphi = 0. \tag{6.1.31}$$

6.1.4.3 Functional spaces

Even more so than in the $1.5d$ case, choosing convenient functional spaces is crucial, due to the singular nature of the correspondence between Cartesian and cylindrical coordinates. We define

> $L^2_{cyl}(\mathbb{R}^3)$ = the largest closed subspace of $L^2(\mathbb{R}^3)$ comprised of functions which have cylindrical symmetry.

A short computation using cylindrical coordinates shows that the decomposition $L^2(\mathbb{R}^3) = L^2_{cyl}(\mathbb{R}^3) \oplus (L^2_{cyl}(\mathbb{R}^3))^\perp$ reduces the Laplacian. This means that the Laplacian commutes with the orthogonal projection of $L^2(\mathbb{R}^3)$ onto $L^2_{cyl}(\mathbb{R}^3)$. Hence the Laplacian is decomposed as

$$\Delta = \Delta_{cyl} + \Delta_{cyl^\perp}.$$

As we have no use for $(L^2_{cyl}(\mathbb{R}^3))^\perp$ we shall abuse notation slightly and denote Δ_{cyl} as simply Δ. We now consider vector valued functions

$$\mathbf{A} \in L^2_{cyl}(\mathbb{R}^3;\mathbb{R}^3) := (L^2_{cyl}(\mathbb{R}^3))^3.$$

We decompose such functions as

$$\begin{aligned} \mathbf{A} &= (\mathbf{A}\cdot\mathbf{e}_\theta)\mathbf{e}_\theta + ((\mathbf{A}\cdot\mathbf{e}_r)\mathbf{e}_r + (\mathbf{A}\cdot\mathbf{e}_z)\mathbf{e}_z) \\ &= \mathbf{A}_\theta + \mathbf{A}_{rz} \in L^2_\theta(\mathbb{R}^3;\mathbb{R}^3) \oplus L^2_{rz}(\mathbb{R}^3;\mathbb{R}^3). \end{aligned} \tag{6.1.32}$$

By computing with cylindrical coordinates we once again discover that this decomposition reduces the vector Laplacian $\boldsymbol{\Delta}$ on $L^2_{cyl}(\mathbb{R}^3;\mathbb{R}^3)$. Note that this reduction does not occur for $\boldsymbol{\Delta}$ on $L^2(\mathbb{R}^3;\mathbb{R}^3)$ (i.e. without the cylindrical symmetry).

We further define the corresponding Sobolev spaces $H^k_{cyl}(\mathbb{R}^3)$,$H^k_\theta(\mathbb{R}^3;\mathbb{R}^3)$, $H^k_{rz}(\mathbb{R}^3;\mathbb{R}^3)$ of functions whose first k weak derivatives lie in $L^2_{cyl}(\mathbb{R}^3)$, $L^2_\theta(\mathbb{R}^3;\mathbb{R}^3)$ and $L^2_{rz}(\mathbb{R}^3;\mathbb{R}^3)$,

respectively. Note that, because of the reductions above, Δ is self-adjoint on $L^2_{cyl}(\mathbb{R}^3)$ with domain $H^2_{cyl}(\mathbb{R}^3)$ and $\boldsymbol{\Delta}$ is self-adjoint on each of $L^2_\theta(\mathbb{R}^3;\mathbb{R}^3)$ and $L^2_{rz}(\mathbb{R}^3;\mathbb{R}^3)$ with domains $H^2_\theta(\mathbb{R}^3;\mathbb{R}^3)$ and $H^2_{rz}(\mathbb{R}^3;\mathbb{R}^3)$, respectively.

As in the $1.5d$ case we shall require certain weighted spaces $\mathfrak{N}_\pm$ that allow us to include functions that do not decay. We define $\mathfrak{N}_\pm$ as the closure of the smooth compactly supported functions $u : \mathbb{R}^3\times\mathbb{R}^3 \to \mathbb{R}$ which are cylindrically symmetric in the $\boldsymbol{x}$ variable, and have $\boldsymbol{x}$-support contained in Ω, under the norms

$$\|u\|_{\mathfrak{N}_\pm} = \int_{\mathbb{R}^3\times\Omega} w^\pm |u|^2 \, d\boldsymbol{v}d\boldsymbol{x},$$

where the weight functions $w^\pm$ are the ones introduced in (6.1.5).

6.1.4.4 The operators

We now define the operators used in the statement of Theorem 6.1.2. As in the $1.5d$ case we shall require the following definition of projection operators:

Definition 6.1.5 (Projection operators)**.** *We define $\widetilde{\mathcal{Q}}^0_\pm$ to be the orthogonal projection operators in $\mathfrak{N}_\pm$ onto* $\ker(\widetilde{\mathcal{D}}_\pm)$.

As in the 1.5d case, the operators $\widetilde{\mathcal{Q}}^0_\pm$ do not depend upon the exact choice of weights $w^\pm$. Now we are ready to define the operators of the cylindrically symmetric case. For brevity, given $\hat{\boldsymbol{v}} = (\hat{v}_r, \hat{v}_\theta, \hat{v}_z)$, we define $\hat{\boldsymbol{v}}_\theta = \hat{v}_\theta \mathbf{e}_\theta$ and $\hat{\boldsymbol{v}}_{rz} = \hat{v}_r\mathbf{e}_r + \hat{v}_z\mathbf{e}_z$. All operators act on functions of the spatial variables only: the operator $\widetilde{\mathcal{A}}^0_1$ acts on functions in $L^2_{cyl}(\mathbb{R}^3)$, $\widetilde{\mathcal{A}}^0_2$ on functions in $L^2_\theta(\mathbb{R}^3;\mathbb{R}^3)$, $\widetilde{\mathcal{A}}^0_3$ on functions in $L^2_{rz}(\mathbb{R}^3;\mathbb{R}^3)$, and $\widetilde{\mathcal{B}}^0_1$ on functions in $L^2_\theta(\mathbb{R}^3;\mathbb{R}^3)$ with range $L^2_{cyl}(\mathbb{R}^3)$. We have:

$$\widetilde{\mathcal{A}}^0_1 h = -\Delta h + \int \sum_\pm \mu^\pm_e (\widetilde{\mathcal{Q}}^0_\pm - 1)h \, d\boldsymbol{v} \tag{6.1.33a}$$

$$\widetilde{\mathcal{A}}^0_2 \boldsymbol{h} = -\boldsymbol{\Delta h} - \left(r\int \sum_\pm \mu^\pm_p \hat{v}_\theta \, d\boldsymbol{v}\right)\boldsymbol{h} - \int \sum_\pm \hat{\boldsymbol{v}}_\theta \mu^\pm_e \widetilde{\mathcal{Q}}^0_\pm [\boldsymbol{h}\cdot\hat{\boldsymbol{v}}_\theta]\, d\boldsymbol{v} \tag{6.1.33b}$$

$$\widetilde{\mathcal{A}}_3^0 \boldsymbol{h} = -\boldsymbol{\Delta} \boldsymbol{h} - \int \sum_{\pm} \hat{\boldsymbol{v}}_{rz} \mu_e^{\pm} \widetilde{\mathcal{Q}}_{\pm}^0 [\boldsymbol{h} \cdot \hat{\boldsymbol{v}}_{rz}] \, d\boldsymbol{v} \tag{6.1.33c}$$

$$\widetilde{\mathcal{B}}_1^0 \boldsymbol{h} = \int \sum_{\pm} \mu_e^{\pm} (\widetilde{\mathcal{Q}}_{\pm}^0 - 1)[\boldsymbol{h} \cdot \hat{\boldsymbol{v}}_{\theta}] \, d\boldsymbol{v} \tag{6.1.33d}$$

$$(\widetilde{\mathcal{B}}_1^0)^* h = \int \sum_{\pm} \mu_e^{\pm} \hat{\boldsymbol{v}}_{\theta} (\widetilde{\mathcal{Q}}_{\pm}^0 - 1) h \, d\boldsymbol{v} \tag{6.1.33e}$$

The precise properties of these operators are discussed in Section 6.6.2. We also mention an identity analogous to (6.1.24)

$$\int_{\mathbb{R}^3} (r\mu_p^{\pm} + \hat{v}_{\theta} \mu_e^{\pm}) \, d\boldsymbol{v} = 0 \tag{6.1.34}$$

which is due to the integrand being a perfect derivative: $\frac{\partial \mu^{\pm}}{\partial v_{\theta}} = r\mu_p^{\pm} + \hat{v}_{\theta} \mu_e^{\pm}$.

6.1.5 Organization of the chapter

In Section 6.2 we provide some necessary background, including the crucial result on approximating spectra found in Chapter 5. Then we treat the two problems – the $1.5d$ and $3d$ cases – in parallel: in Section 6.3 we formulate the two problems as an equivalent family of self-adjoint problems which we then successively solve in Section 6.4. The proofs of the main theorems are concluded in Section 6.5. In Section 6.6 we provide the rigorous treatment of the various operators appearing throughout the chapter, and in Section 6.7 we show that there exist nontrivial equilibria.

6.2 Background, Definitions and Notation

In this section we remind the reader of the various notions of convergence in Hilbert spaces in order to avoid confusion. For a Hilbert space $\mathfrak{H}$ we denote its norm and inner product by $\|\cdot\|_{\mathfrak{H}}$ and $\langle \cdot, \cdot \rangle_{\mathfrak{H}}$, respectively. When there is no ambiguity we drop the subscript. We denote the set of bounded linear operators

from a Hilbert space $\mathfrak{H}$ to a Hilbert space $\mathfrak{G}$ as $\mathfrak{B}(\mathfrak{H},\mathfrak{G})$, and when $\mathfrak{H}=\mathfrak{G}$ we simply write $\mathfrak{B}(\mathfrak{H})$. The operator norm is denoted $\|\cdot\|_{\mathfrak{H}\to\mathfrak{G}}$, where, again, when there is no ambiguity we may drop the subscript.

Definition 6.2.1 (Convergence in $\mathfrak{B}(\mathfrak{H},\mathfrak{G})$)**.** *Let $\mathcal{T},\mathcal{T}_n\in\mathfrak{B}(\mathfrak{H},\mathfrak{G})$, where $n\in\mathbb{N}$.*

(a) We say that the sequence $\mathcal{T}_n$ converges to $\mathcal{T}$ in norm (or uniformly) as $n\to\infty$ whenever $\|\mathcal{T}_n-\mathcal{T}\|_{\mathfrak{H}\to\mathfrak{G}}\to 0$ as $n\to\infty$. In this case we write $\mathcal{T}_n\to\mathcal{T}$.

(b) We say that the sequence $\mathcal{T}_n$ converges to $\mathcal{T}$ strongly as $n\to\infty$ whenever we have the pointwise convergence $\mathcal{T}_n u\to\mathcal{T}u$ in $\mathfrak{G}$ for all $u\in\mathfrak{H}$. In this case we write $\mathcal{T}_n\xrightarrow{s}\mathcal{T}$.

Now let us recall some important notions related to unbounded self-adjoint operators:

Definition 6.2.2 (Convergence of unbounded operators)**.** *Let $\mathcal{A}$ and $\mathcal{A}_n$ be self-adjoint, where $n\in\mathbb{N}$.*

(a) We say that the sequence $\mathcal{A}_n$ converges to $\mathcal{A}$ in the norm resolvent sense as $n\to\infty$ whenever $(\mathcal{A}_n-z)^{-1}\to(\mathcal{A}-z)^{-1}$ for any $z\in\mathbb{C}\setminus\mathbb{R}$. In this case we write $\mathcal{A}_n\xrightarrow{n.r.}\mathcal{A}$.

(b) We say that the sequence $\mathcal{A}_n$ converges to $\mathcal{A}$ in the strong resolvent sense as $n\to\infty$ whenever $(\mathcal{A}_n-z)^{-1}\xrightarrow{s}(\mathcal{A}-z)^{-1}$ for any $z\in\mathbb{C}\setminus\mathbb{R}$. In this case we write $\mathcal{A}_n\xrightarrow{s.r.}\mathcal{A}$.

Remark 6.2.1. *Notice that for any self-adjoint operator $\mathcal{A}$, the resolvent $(\mathcal{A}-z)^{-1}$ is a bounded operator for any $z\in\mathbb{C}\setminus\mathbb{R}$.*

6.2.1 Basic facts

The subsequent results will be used throughout the chapter without explicit reference.

Lemma 6.2.1. *Let $\mathfrak{H},\mathfrak{G}$ be Banach spaces, $\mathcal{T},\mathcal{T}_n\in\mathfrak{B}(\mathfrak{H},\mathfrak{G})$ and $u,u_n\in\mathfrak{H}$ where $n\in\mathbb{N}$, and assume that $\mathcal{T}_n\xrightarrow{s}\mathcal{T}$ and $u_n\to u$. Then $\mathcal{T}_n u_n\to\mathcal{T}u$ as*

$n \to \infty$.

Proof. We compute

$$\begin{aligned}\|\mathcal{T}_n u_n - \mathcal{T} u\| &\leq \|\mathcal{T}_n(u_n - u)\| + \|(\mathcal{T}_n - \mathcal{T})u\| \\ &\leq \left(\sup_{n\in\mathbb{N}} \|\mathcal{T}_n\|\right) \|u_n - u\| + \|(\mathcal{T}_n - \mathcal{T})u\|\,.\end{aligned}$$

This supremum is finite by the uniform boundedness principle, so the first term converges to zero since $u_n \to u$. The second term converges to zero since $\mathcal{T}_n \xrightarrow{s} \mathcal{T}$. □

Corollary 6.2.1. *If $\mathcal{T}_n \xrightarrow{s} \mathcal{T}$ and $\mathcal{S}_n \xrightarrow{s} \mathcal{S}$ as $n \to \infty$ then $\mathcal{T}_n\mathcal{S}_n \xrightarrow{s} \mathcal{T}\mathcal{S}$ as $n \to \infty$.*

The following result complements Weyl's theorem (see [109, IV, Theorem 5.35]) on the stability of the essential spectrum under a relatively compact perturbation. In our setting we know more about the perturbation than being merely relatively compact.

Lemma 6.2.2. *Let $\mathcal{A} = -\Delta + \mathcal{K} : H^2(\mathbb{R}^n) \subset L^2(\mathbb{R}^n) \to L^2(\mathbb{R}^n)$ be a self-adjoint Schrödinger operator with $\mathcal{K} \in \mathfrak{B}(L^2(\mathbb{R}^n))$ and $\mathcal{K} = \mathcal{K}\mathcal{P}$ where $\mathcal{P} : L^2(\mathbb{R}^n) \to L^2(\mathbb{R}^n)$ is the multiplication operator by the characteristic function $\mathbb{1}_\Omega$ of some bounded domain $\Omega \subset \mathbb{R}^n$. Then $\mathcal{A}$ has a finite number of negative eigenvalues (counting multiplicity).*

Proof. By Weyl's theorem (see [109, IV, Theorem 5.35]), there are at most countably many negative eigenvalues and they may only accumulate at 0. Denote these as the increasing sequence $\lambda_1 \leq \lambda_2 \leq \cdots$, where equality comes from multiplicity. As $\mathcal{A}$ is self-adjoint, the corresponding normalised eigenfunctions $e_1, e_2, \ldots$ form an orthonormal set. Let $\mathcal{E}$ be their linear span, i.e.

$$\mathcal{E} = \operatorname{span}\{e_i : i = 1, 2, \ldots\}.$$

Note that $\mathcal{E}$ is a linear subspace of $L^2(\mathbb{R}^d)$, but is not necessarily closed (at this point in the proof) as we have not taken the closed span.

We claim that there exists an injective linear map from $\mathcal{E}$ into a finite dimensional

space, and hence $\mathcal{E}$ is finite dimensional, proving the lemma. Indeed, we define the map $\mathcal{T} : \mathcal{E} \to H^2(\Omega)$ by $u \mapsto 1_\Omega u$, with image $\mathcal{T}(\mathcal{E})$. (We do not denote it $\mathcal{P}$ as their codomains differ.) $\mathcal{T}$ is manifestly linear, so it remains to check the other claimed properties.

Step 1. $\mathcal{T}$ is injective into its image. By linearity it suffices to show that $\mathcal{T}u = 0$ implies that $u = 0$ for any $u \in \mathcal{E}$. Let $u = \sum_{i=1}^m a_i e_i$ for m finite and scalar a_i be an arbitrary element of $\mathcal{E}$. Then it is enough to show that

$$u(x) = \sum_{i=1}^m a_i e_i(x) = 0 \text{ for almost every } x \in \Omega \implies a_i = 0, i = 1, \dots, m.$$

(If Ω were $\mathbb{R}^n$ then this would follow immediately from orthogonality of the e_i.) Suppose that this is not the case, and let a_j be the first non-zero coefficient in the sum and j' be the first integer above j for which $\lambda_j \neq \lambda_{j'}$. Then we have

$$\lambda_j^{-k} \mathcal{A}^k u = \sum_{i=j}^{j'-1} a_i e_i + \lambda_j^{-k} \sum_{i=j+1}^{m} \lambda_i^k a_i e_i \to \sum_{i=j}^{j'-1} a_i e_i \text{ in } L^2(\mathbb{R}^n) \text{ as } k \to \infty$$

where $\mathcal{A}^k$ is the k-th power of the operator $\mathcal{A}$, by the ordering of eigenvalues. By the assumption on $\mathcal{K}$ we have $\mathcal{A}u = 0$ almost everywhere in Ω. This implies that $v(x) = \sum_{i=j}^{j'-1} a_i e_i(x)$ is zero for almost every x in Ω. But v is an eigenfunction of $\mathcal{A}$ with eigenvalue λ_j, so that

$$\lambda_j v = \mathcal{A}v = -\Delta v + \mathcal{K}v = -\Delta v,$$

where we have used once again that $\mathcal{K} = \mathcal{K}\mathcal{P}$ and that $v = 0$ almost everywhere in Ω. As $\lambda_j < 0$ and $v \in L^2(\mathbb{R}^n)$ we must have $v = 0$ almost everywhere in $\mathbb{R}^n$, and by orthogonality of the e_i in $L^2(\mathbb{R}^n)$ this implies that $a_j = 0$, which is a contradiction.

Step 2. The image of $\mathcal{T}$ is finite dimensional. Let $v = \sum_{i=1}^m a_i \mathcal{T} e_i = \mathcal{T}u$

for scalar a_i and m finite be an arbitrary element of $\mathcal{T}(\mathcal{E})$. Then we have

$$\begin{aligned}\|\nabla v\|^2_{L^2(\Omega)} - \|\mathcal{K}\|\,\|v\|^2_{L^2(\Omega)} &\leq \left\|\nabla \sum_{i=1}^m a_i e_i\right\|^2_{L^2(\mathbb{R}^n)} - \|\mathcal{K}\| \left\|\mathcal{P}\sum_{i=1}^m a_i e_i\right\|^2_{L^2(\mathbb{R}^n)} \\ &\leq \left\langle \mathcal{A}\sum_{i=1}^m a_i e_i, \sum_{i=1}^m a_i e_i \right\rangle_{L^2(\mathbb{R}^n)} \\ &= \sum_{i=1}^m \lambda_i |a_i|^2 \,\|e_i\|^2_{L^2(\mathbb{R}^n)} \leq 0\end{aligned}$$

where on the second line we have used orthogonality of the eigenfunctions e_i and on the final line we have used that $\lambda_i < 0$. Hence any $v \in \mathcal{T}(\mathcal{E})$ obeys the bound

$$\|\nabla v\|^2_{L^2(\Omega)} \leq C\,\|v\|^2_{L^2(\Omega)}$$

and by the Poincaré inequality on the bounded domain Ω, (enlarging Ω as needed to ensure that its boundary is smooth), $\mathcal{T}(\mathcal{E})$ is finite dimensional.

As discussed above, these two steps complete the proof. □

6.2.2 Approximating strongly continuous families of unbounded operators

Here we summarise the main results of Chapter 5 on properties of approximations of strongly continuous families of unbounded operators. We shall require these results in the sequel. We refer to Chapter 5 for full details, including proofs. Let $\mathfrak{H} = \mathfrak{H}_+ \oplus \mathfrak{H}_-$ be a (separable) Hilbert space with inner product $\langle\cdot,\cdot\rangle$ and norm $\|\cdot\|$ and let

$$\mathcal{A}^\lambda = \begin{bmatrix} \mathcal{A}^\lambda_+ & 0 \\ 0 & -\mathcal{A}^\lambda_- \end{bmatrix} \quad \text{and} \quad \mathcal{K}^\lambda = \begin{bmatrix} \mathcal{K}^\lambda_{++} & \mathcal{K}^\lambda_{+-} \\ \mathcal{K}^\lambda_{-+} & \mathcal{K}^\lambda_{--} \end{bmatrix}, \quad \lambda \in [0,1]$$

be two families of operators on $\mathfrak{H}$ depending upon the parameter $\lambda \in [0,1]$ (the range $[0,1]$ of values of the parameter is, of course, arbitrary), where the family $\mathcal{A}^\lambda$ is also assumed to be defined for λ in an open neighbourhood D_0 of $[0,1]$ in the complex plane. They satisfy:

i) Sectoriality: The sesquilinear forms $\mathfrak{a}_\pm^\lambda$ corresponding to $\mathcal{A}_\pm^\lambda$ are sectorial and closed for $\lambda \in D_0$, symmetric for real λ, have dense domains $\mathfrak{D}(\mathfrak{a}_\pm^\lambda)$ independent of $\lambda \in D_0$, and $D_0 \ni \lambda \mapsto \mathfrak{a}_\pm^\lambda[u,v]$ are holomorphic for any $u, v \in \mathfrak{D}(\mathfrak{a}_\pm^\lambda)$. [In the terminology of [109], $\mathfrak{a}_\pm^\lambda$ are *holomorphic families of type (a)* and $\mathcal{A}^\lambda$ are *holomorphic families of type (B).*]

ii) Gap: $\mathcal{A}_\pm^\lambda > 1$ for every $\lambda \in [0,1]$. We let $\alpha > 1$ be a lower bound to all $\mathcal{A}_\pm^\lambda$.

iii) Bounded perturbation: $\{\mathcal{K}^\lambda\}_{\lambda\in[0,1]} \subset \mathfrak{B}(\mathfrak{H})$ is a symmetric strongly continuous family.

iv) Compactness: There exist symmetric operators $\mathcal{P}_\pm \in \mathfrak{B}(\mathfrak{H}_\pm)$ which are relatively compact with respect to the forms $\mathfrak{a}_\pm^\lambda$, satisfying $\mathcal{K}^\lambda = \mathcal{K}^\lambda\mathcal{P}$ for all $\lambda \in [0,1]$ where

$$\mathcal{P} = \begin{bmatrix} \mathcal{P}_+ & 0 \\ 0 & \mathcal{P}_- \end{bmatrix}.$$

Finally, if the family $\mathcal{A}^\lambda$ does not have a compact resolvent we assume:

v) Compactification of the resolvent: There exist holomorphic forms $\{\mathfrak{w}_\pm^\lambda\}_{\lambda\in D_0}$ of type (a) and associated operators $\{\mathcal{W}_\pm^\lambda\}_{\lambda\in D_0}$ of type (B) such that for $\lambda \in [0,1]$, $\mathcal{W}_\pm^\lambda$ are self-adjoint and non-negative, and if $\mathfrak{w}^\lambda$ is the form associated with

$$\mathcal{W}^\lambda = \begin{bmatrix} \mathcal{W}_+^\lambda & 0 \\ 0 & -\mathcal{W}_-^\lambda \end{bmatrix}, \quad \lambda \in D_0,$$

then $\mathfrak{D}(\mathfrak{w}^\lambda) \cap \mathfrak{D}(\mathfrak{a}_\pm)$ are dense for all $\lambda \in D_0$ and the inclusion $(\mathfrak{D}(\mathfrak{w}^\lambda) \cap \mathfrak{D}(\mathfrak{a}), \|\cdot\|_{\mathfrak{a}_\varepsilon^\lambda}) \to (\mathfrak{H}, \|\cdot\|)$ is compact for some $\lambda \in D_0$ and all $\varepsilon > 0$, where $\mathfrak{a}_\varepsilon^\lambda$ is the form associated with

$$\mathcal{A}_\varepsilon^\lambda := \mathcal{A}^\lambda + \varepsilon\mathcal{W}^\lambda, \quad \lambda \in D_0,\ \varepsilon \geq 0. \tag{6.2.1}$$

Define the family of (unbounded) operators $\{\mathcal{M}_\varepsilon^\lambda\}_{\lambda\in[0,1],\varepsilon\geq 0}$, acting in $\mathfrak{H}$, as

$$\mathcal{M}_\varepsilon^\lambda = \mathcal{A}_\varepsilon^\lambda + \mathcal{K}^\lambda, \quad \lambda \in [0,1].$$

For $\varepsilon > 0$, let

- $\{e_{\varepsilon,k}^\lambda\}_{k\in\mathbb{N}} \subset \mathfrak{H}$ be a complete orthonormal set of eigenfunctions of $\mathcal{A}_\varepsilon^\lambda$,
- $\mathcal{G}_{\varepsilon,n}^\lambda : \mathfrak{H} \to \mathfrak{H}$ be the orthogonal projection operators onto $\operatorname{span}(e_{\varepsilon,1}^\lambda, \ldots, e_{\varepsilon,n}^\lambda)$,
- $\widetilde{\mathcal{M}}_{\varepsilon,n}^\lambda$ be the n-dimensional operator defined as the restriction of $\mathcal{M}_\varepsilon^\lambda$ to $\mathcal{G}_{\varepsilon,n}^\lambda(\mathfrak{H})$.

Now, for $\lambda \in [0,1]$, $\varepsilon \geq 0$ and $n \in \mathbb{N}$ we define the measures (where we *always* take multiplicities into account!)

$$\nu_{\lambda,\varepsilon} = \sum_{x\in \mathrm{sp}_{\mathrm{pp}}(\mathcal{M}_\varepsilon^\lambda)\setminus \mathrm{sp}_{\mathrm{ess}}(\mathcal{M}_\varepsilon^\lambda)} \delta_x$$

and for any $\varepsilon > 0$ and $n \in \mathbb{N}$ the measures

$$\widetilde{\nu}_{\lambda,\varepsilon,n} = \sum_{x\in \mathrm{sp}(\widetilde{\mathcal{M}}_{\varepsilon,n}^\lambda)} \delta_x.$$

Consider a cutoff function ϕ_η satisfying

$$\phi_\eta(x) = \begin{cases} 1 & x \in [-1,1] \\ 0 & x \in \mathbb{R}\setminus(-1-\eta, 1+\eta) \end{cases}, \quad \phi_\eta \in C(\mathbb{R},[0,1]), \quad \eta \in (0,\alpha).$$

Finally, define the measures

$$\mu_{\lambda,\varepsilon}^\eta = \phi_\eta \nu_{\lambda,\varepsilon}$$

and

$$\widetilde{\mu}_{\lambda,\varepsilon,n}^\eta = \phi_\eta \widetilde{\nu}_{\lambda,\varepsilon,n}.$$

Recall that the space of finite positive Borel measures equipped with the topology of weak convergence is metrisable, for example with the bounded Lipschitz distance

$$d_{BL}(\mu,\nu) := \sup_{\|\psi\|_{\mathrm{Lip}}\leq 1, |\psi|\leq 1} \int \psi \,\mathrm{d}(\mu-\nu).$$

Our main result in Chapter 5 is:

Theorem 6.2.1. *The mappings* $(\lambda, \varepsilon) \mapsto \mu^{\eta}_{\lambda,\varepsilon}$ *and* $\lambda \mapsto \widetilde{\mu}^{\eta}_{\lambda,\varepsilon,n}$ *are weakly continuous and as* $n \to \infty$, $d_{BL}(\widetilde{\mu}^{\eta}_{\lambda,\varepsilon,n}, \mu^{\eta}_{\lambda,\varepsilon}) \to 0$ *uniformly in* $\lambda \in [0,1]$.

6.3 An equivalent problem

6.3.1 The $1.5d$ case

We will now reduce the linearised Vlasov-Maxwell system (6.1.19) to a self-adjoint problem in $L^2(\mathbb{R}) \times L^2_0(\Omega)$ depending continuously (in the norm resolvent sense) on the parameter $\lambda > 0$.

6.3.1.1 Inverting the linearised Vlasov equation

Rearranging the terms in (6.1.22) we obtain

$$(\lambda + \mathcal{D}_{\pm}) f^{\pm} = \pm(\lambda + \mathcal{D}_{\pm})(\mu_e^{\pm}\phi + \mu_p^{\pm}\psi) \pm \lambda \mu_e^{\pm}(-\phi + \hat{v}_2\psi), \tag{6.3.1}$$

where we use the fact that $\mu^{\pm}$ are constant along trajectories of the vector-fields $\mathcal{D}_{\pm}$. In order to obtain an expression for $f^{\pm}$ in terms of the potentials ϕ, ψ we invert the operators $(\lambda + \mathcal{D}_{\pm})$ and to do this we must study the operators $\mathcal{D}_{\pm}$.

Lemma 6.3.1. *The operators* $\mathcal{D}_{\pm}$ *on* $\mathfrak{L}_{\pm}$ *satisfy:*

(a) $\mathcal{D}_{\pm}$ *are skew-adjoint and the resolvents* $(\lambda + \mathcal{D}_{\pm})^{-1}$ *are bounded linear operators for* $\operatorname{Re}\lambda \neq 0$ *with norm bounded by* $1/|\operatorname{Re}\lambda|$.

(b) $\mathcal{D}_{\pm}$ *flip parity with respect to the variable* v_1, *i.e. if* $h(x, v_1, v_2) \in \mathfrak{D}(\mathcal{D}_{\pm})$ *is an even function of* v_1 *then* $\mathcal{D}_{\pm}h$ *is an odd function of* v_1 *and vice versa.*

(c) For real $\lambda \neq 0$ *the resolvents of* $\mathcal{D}_{\pm}$ *split as follows:*

$$(\lambda + \mathcal{D}_{\pm})^{-1} = \lambda(\lambda^2 - \mathcal{D}_{\pm}^2)^{-1} - \mathcal{D}_{\pm}(\lambda^2 - \mathcal{D}_{\pm}^2)^{-1}$$

where the first part is symmetric and preserves parity with respect to v_1, and the second part is skew-symmetric and inverts parity with respect to v_1.

Proof. Skew-adjointness follows from integration by parts, noting that $w^{\pm}$ are in the kernels of $\mathcal{D}_{\pm}$ (to be fully precise, only skew-*symmetry* follows. However, skew-adjointness is a simple extension, see e.g. [187] and in particular exercise 28 therein). The existence of bounded resolvents follows. The statement regarding parity follows directly from the formulas for $\mathcal{D}_{\pm}$ term by term. Finally, for the last part we use functional calculus formalism to compute

$$\frac{1}{\lambda + \mathcal{D}_{\pm}} = \frac{\lambda - \mathcal{D}_{\pm}}{\lambda^2 - \mathcal{D}_{\pm}^2} = \frac{\lambda}{\lambda^2 - \mathcal{D}_{\pm}^2} - \frac{\mathcal{D}_{\pm}}{\lambda^2 - \mathcal{D}_{\pm}^2}.$$

As $\mathcal{D}_{\pm}$ are skew-adjoint, $\mathcal{D}_{\pm}^2$ are self-adjoint and hence the first term is self-adjoint and the second skew-adjoint. For the parity properties we note that as $\mathcal{D}_{\pm}$ flip parity, $\mathcal{D}_{\pm}^2$ preserve parity and hence so do $\lambda^2 - \mathcal{D}_{\pm}^2$ and their inverses. □

Applying $(\lambda + \mathcal{D}_{\pm})^{-1}$ to (6.3.1) yields,

$$f^{\pm} = \pm\mu_e^{\pm}\phi \pm \mu_p^{\pm}\psi \pm \lambda(\lambda + \mathcal{D}_{\pm})^{-1}[\mu_e^{\pm}(-\phi + \hat{v}_2\psi)]. \tag{6.3.2}$$

Furthermore, using Lemma 6.3.1 we split $f^{\pm}$ into even and odd functions of v_1:

$$\begin{aligned} f_{ev}^{\pm} &= \pm\mu_e^{\pm}\phi \pm \mu_p^{\pm}\psi \pm \mu_e^{\pm}\lambda^2(\lambda^2 - \mathcal{D}_{\pm}^2)^{-1}[-\phi + \hat{v}_2\psi] \\ f_{od}^{\pm} &= \mp\mu_e^{\pm}\lambda\mathcal{D}_{\pm}(\lambda^2 - \mathcal{D}_{\pm}^2)^{-1}[-\phi + \hat{v}_2\psi] \end{aligned}$$

using the fact that ϕ, ψ and μ are all even functions of v_1. For brevity, we define operators $\mathcal{Q}_{\pm}^{\lambda} : \mathfrak{L}_{\pm} \to \mathfrak{L}_{\pm}$ as

$$\mathcal{Q}_{\pm}^{\lambda} = \lambda^2(\lambda^2 - \mathcal{D}_{\pm}^2)^{-1}, \quad \lambda > 0.$$

When $\lambda \to 0$ their strong limits exist, and are defined in Definition 6.1.4 (this convergence is proved in Lemma 6.6.1).

Remark 6.3.1. *Operators $\mathcal{Q}_{\pm}^{\lambda}$ also appeared in the prior works [130, 132, 16, 17]. In each of these $\mathcal{Q}_{\pm}^{\lambda}$ were defined as an integrated average over the characteristics*

of the operators $\mathcal{D}_\pm$. In fact, as the Laplace transform of a semigroup is the resolvent of its generator we see that the operators $\mathcal{Q}_\pm^\lambda$ in these prior works have the rule:

$$\mathcal{Q}_\pm^\lambda h = \int_{-\infty}^0 \lambda e^{\lambda s} e^{s\mathcal{D}_\pm} h \, ds = \lambda \int_0^\infty e^{-\lambda s} e^{-s\mathcal{D}_\pm} h \, ds = \lambda(\lambda + \mathcal{D}_\pm)^{-1} h.$$

Here we have defined the operators $\mathcal{Q}_\pm^\lambda$ directly from the resolvents of $\mathcal{D}_\pm$ as this makes some of its properties clearer, although both approaches have advantages. In particular we are able to split $\lambda(\mathcal{D}_\pm + \lambda)^{-1}$ into symmetric and skew-symmetric parts in Lemma 6.3.1 which simplifies some computations.

6.3.1.2 Reformulating Maxwell's equations

Now we substitute the expressions (6.3.2) into Maxwell's equations (6.1.21). This shall result in an equivalent system of equations for ϕ and ψ. Due to the integration $d\boldsymbol{v}$ we notice that $f_{od}^\pm$ and $f_{od}^\pm \hat{v}_2$ both integrate to zero, so that ρ and j_2 only depend on $f_{ev}^\pm$.

Remark 6.3.2. *It is important to note that due to the continuity equation it is possible to express either (6.1.21a) or (6.1.21b) using the remaining two equations in (6.1.21). See Lemma 6.5.4.*

Gauss' equation (6.1.21c). Gauss' equation becomes

$$\begin{aligned} -\partial_x^2 \phi &= \int (f_{ev}^+ - f_{ev}^-) \, d\boldsymbol{v} \\ &= \int \sum_\pm \left(\mu_e^\pm \phi + \mu_p^\pm \psi + \mathcal{Q}_\pm^\lambda [\mu_e^\pm (-\phi + \hat{v}_2 \psi)] \right) d\boldsymbol{v} \\ &= \int \sum_\pm (\mu_p^\pm + \mu_e^\pm \hat{v}_2) \psi \, d\boldsymbol{v} + \int \sum_\pm \mu_e^\pm (\mathcal{Q}_\pm^\lambda - 1)[-\phi + \hat{v}_2 \psi] \, d\boldsymbol{v}, \end{aligned} \tag{6.3.3}$$

where we have pulled $\mu_e^\pm$ outside the application of $\mathcal{Q}_\pm^\lambda$ as they belong to $\ker(\mathcal{D}_\pm)$.

Ampère's equation (6.1.21b)**.** Similarly, Ampère's equation becomes

$$\begin{aligned}(-\partial_x^2+\lambda^2)\psi &= \int \hat{v}_2(f_{ev}^+ - f_{ev}^-)\,d\boldsymbol{v}\\ &= \int \sum_{\pm} \hat{v}_2\left(\mu_e^{\pm}\phi + \mu_p^{\pm}\psi + \mathcal{Q}_{\pm}^{\lambda}[\mu_e^{\pm}(-\phi+\hat{v}_2\psi)]\right)\,d\boldsymbol{v}\\ &= \int \sum_{\pm} \hat{v}_2(\mu_p^{\pm}+\mu_e^{\pm}\hat{v}_2)\psi\,d\boldsymbol{v}\\ &\quad + \int \sum_{\pm} \hat{v}_2\mu_e^{\pm}(\mathcal{Q}_{\pm}^{\lambda}-1)[-\phi+\hat{v}_2\psi]\,d\boldsymbol{v}.\end{aligned} \tag{6.3.4}$$

An equivalent formulation. We write the two new expressions (6.3.3) and (6.3.4) abstractly in the compact form

$$\boldsymbol{\mathcal{M}}^{\lambda}\begin{bmatrix}\psi\\ \phi\end{bmatrix} = \begin{bmatrix}-\partial_x^2\psi+\lambda^2\psi - j_2\\ \partial_x^2\phi+\rho\end{bmatrix} = \begin{bmatrix}0\\0\end{bmatrix}, \tag{6.3.5}$$

where, for $\lambda > 0$, $\boldsymbol{\mathcal{M}}^{\lambda}$ is a self-adjoint matrix of operators mapping $L^2(\mathbb{R})\times L_0^2(\Omega) \to L^2(\mathbb{R})\times L_0^2(\Omega)$ (see Lemma 6.6.4). We claim that this operator may be written either as

$$\boldsymbol{\mathcal{M}}^{\lambda} = \begin{bmatrix}-\partial_x^2+\lambda^2 & 0\\ 0 & \partial_x^2\end{bmatrix} - \boldsymbol{\mathcal{J}}^{\lambda} \tag{6.3.6}$$

or, equivalently, as

$$\boldsymbol{\mathcal{M}}^{\lambda} = \begin{bmatrix}\mathcal{A}_2^{\lambda} & \left(\mathcal{B}^{\lambda}\right)^*\\ \mathcal{B}^{\lambda} & -\mathcal{A}_1^{\lambda}\end{bmatrix} \tag{6.3.7}$$

where the various operators appearing above are given by

$$\begin{aligned}\boldsymbol{\mathcal{J}}^{\lambda}\begin{bmatrix}h\\ g\end{bmatrix} &= -\left(\sum_{\pm}\int \mu^{\pm}\frac{1+v_1^2}{\langle\boldsymbol{v}\rangle^3}\,d\boldsymbol{v}\right)\begin{bmatrix}h\\0\end{bmatrix} +\\ &\quad + \sum_{\pm}\int \begin{bmatrix}\hat{v}_2\\-1\end{bmatrix}\mu_e^{\pm}(\mathcal{Q}_{\pm}^{\lambda}-1)\left(\begin{bmatrix}\hat{v}_2\\-1\end{bmatrix}\cdot\begin{bmatrix}h\\g\end{bmatrix}\right)\,d\boldsymbol{v}\end{aligned} \tag{6.3.8a}$$

$$\mathcal{A}_1^{\lambda}h = -\partial_x^2 h + \int\sum_{\pm}\mu_e^{\pm}(\mathcal{Q}_{\pm}^{\lambda}-1)h\,d\boldsymbol{v} \tag{6.3.8b}$$

$$\mathcal{A}_2^\lambda h = -\partial_x^2 h + \lambda^2 h - \left(\sum_\pm \int \mu_p^\pm \hat{v}_2 \, d\boldsymbol{v}\right) h - \int \sum_\pm \hat{v}_2 \mu_e^\pm \mathcal{Q}_\pm^\lambda \, [\hat{v}_2 h] \, d\boldsymbol{v} \tag{6.3.8c}$$

$$\mathcal{B}^\lambda h = \left(\int \sum_\pm \mu_p^\pm \, d\boldsymbol{v}\right) h + \int \sum_\pm \mu_e^\pm \mathcal{Q}_\pm^\lambda \, [\hat{v}_2 h] \, d\boldsymbol{v} \tag{6.3.8d}$$

$$\left(\mathcal{B}^\lambda\right)^* h = \left(\int \sum_\pm \mu_p^\pm \, d\boldsymbol{v}\right) h + \int \sum_\pm \mu_e^\pm \hat{v}_2 \mathcal{Q}_\pm^\lambda h \, d\boldsymbol{v}. \tag{6.3.8e}$$

Remark 6.3.3. *Though $\lambda > 0$ in the foregoing discussion, all operators can be defined for $\lambda = 0$, as we have already done for some (see (6.1.23)).*

The expression (6.3.7) is no more than a rewriting of (6.3.3) and (6.3.4). However the expression (6.3.6) requires some attention. In particular, to obtain it one has to use (6.1.24) as well as the integration by parts

$$\int \frac{\partial \mu^\pm}{\partial v_2} \hat{v}_2 \, d\boldsymbol{v} = -\int \mu^\pm \frac{\partial \hat{v}_2}{\partial v_2} \, d\boldsymbol{v} = -\int \mu^\pm \frac{1 + v_1^2}{\langle \boldsymbol{v} \rangle^3} \, d\boldsymbol{v}.$$

The properties of the operators appearing in (6.3.8) are discussed in details in Lemma 6.6.2 and Lemma 6.6.3. Let us briefly summarise:

- $\mathcal{A}_1^\lambda : H^2_{n,0}(\Omega) \subset L^2_0(\Omega) \to L^2_0(\Omega)$ is self-adjoint and has a purely discrete spectrum with finitely many negative eigenvalues.
- $\mathcal{A}_2^\lambda : H^2(\mathbb{R}) \subset L^2(\mathbb{R}) \to L^2(\mathbb{R})$ is self-adjoint, has essential spectrum in $[\lambda^2, \infty)$ and finitely many negative eigenvalues.
- $\mathcal{B}^\lambda : L^2(\mathbb{R}) \to L^2_0(\Omega)$ is a bounded operator, with bound independent of λ.
- $\mathcal{J}^\lambda : L^2(\mathbb{R}) \times L^2_0(\Omega) \to L^2(\mathbb{R}) \times L^2_0(\Omega)$ is a bounded symmetric operator, with bound independent of λ.

6.3.2 The cylindrically symmetric case

Our approach here is fully analogous to the one presented in Section 6.3.1 hence we shall keep it brief, omitting repetitions as much as possible. For convenience

we denote analogous operators by the same letter, but we shall add a *tilde* to any such operator in this section. Hence, e.g. the operators analogous to $\mathcal{D}_\pm$ shall be denoted $\widetilde{\mathcal{D}}_\pm$.

6.3.2.1 Inverting the linearised Vlasov equation

Recall the linearised Vlasov equation (6.1.29a)

$$(\lambda + \widetilde{\mathcal{D}}_\pm) f^\pm = \pm(\lambda + \widetilde{\mathcal{D}}_\pm)(\mu_e^\pm \varphi + r\mu_p^\pm (\mathbf{A} \cdot \boldsymbol{e}_\theta)) \pm \lambda \mu_e^\pm (-\varphi + \mathbf{A} \cdot \hat{\boldsymbol{v}}).$$

Inverting, we get the expression

$$f^\pm = \pm\mu_e^\pm \varphi \pm r\mu_p^\pm (\mathbf{A} \cdot \boldsymbol{e}_\theta) \pm \mu_e^\pm \lambda (\lambda + \widetilde{\mathcal{D}}_\pm)^{-1} (-\varphi + \mathbf{A} \cdot \hat{\boldsymbol{v}}), \tag{6.3.9}$$

and, recalling that we only care about the quantity $f^+ - f^-$, we write it for future reference:

$$f^+ - f^- = \sum_\pm \mu_e^\pm \varphi + \sum_\pm r\mu_p^\pm (\mathbf{A} \cdot \boldsymbol{e}_\theta) + \sum_\pm \mu_e^\pm \lambda (\lambda + \widetilde{\mathcal{D}}_\pm)^{-1} (-\varphi + \mathbf{A} \cdot \hat{\boldsymbol{v}}). \tag{6.3.10}$$

Lemma 6.3.2. *The operators $\widetilde{\mathcal{D}}_\pm$ on $\mathfrak{N}_\pm$ satisfy:*

(a) $\widetilde{\mathcal{D}}_\pm$ are skew-adjoint and the resolvents $(\lambda + \widetilde{\mathcal{D}}_\pm)^{-1}$ are bounded linear operators for $\operatorname{Re}\lambda \neq 0$ with norm bounded by $1/|\operatorname{Re}\lambda|$.

(b) $\widetilde{\mathcal{D}}_\pm$ flip parity with respect to the pair of variables (v_r, v_z), i.e. if $h \in \mathfrak{D}(\widetilde{\mathcal{D}}_\pm)$ is an even function of the pair (v_r, v_z) then $\widetilde{\mathcal{D}}_\pm h$ is an odd function of (v_r, v_z) and vice versa (see Remark 6.3.4 below).

(c) For real $\lambda \neq 0$ the resolvents of $\widetilde{\mathcal{D}}_\pm$ split as follows:

$$(\lambda + \widetilde{\mathcal{D}}_\pm)^{-1} = \lambda(\lambda^2 - \widetilde{\mathcal{D}}_\pm^2)^{-1} - \widetilde{\mathcal{D}}_\pm(\lambda^2 - \widetilde{\mathcal{D}}_\pm^2)^{-1} \tag{6.3.11}$$

where the first part is symmetric and preserves parity with respect to (v_r, v_z), and the second part is skew-symmetric and inverts parity with respect to (v_r, v_z).

We leave the proof, which is analogous to the proof of Lemma 6.3.1, to the

reader.

Remark 6.3.4. *For a function h that is expressed in cylindrical coordinates as $h(\boldsymbol{x}, v_r, v_z, v_\theta)$, we say that h is an even function of the pair (v_r, v_z) if it holds that $h(\boldsymbol{x}, v_r, v_z, v_\theta) = h(\boldsymbol{x}, -v_r, -v_z, v_\theta)$, where we flip the sign of both variables simultaneously. Note that this is a weaker property than both being an even function of v_r and an even function of v_z. Odd functions of (v_r, v_z) are defined similarly.*

As in the $1.5d$ case, we define averaging operators. However, in this case both the symmetric and skew-symmetric parts are required. The operators $\widetilde{\mathcal{Q}}^{\lambda}_{\pm,sym}$ and $\widetilde{\mathcal{Q}}^{\lambda}_{\pm,skew}$ map $\mathfrak{N}_\pm$ to $\mathfrak{N}_\pm$ and are defined by the rules

$$\begin{aligned}\widetilde{\mathcal{Q}}^{\lambda}_{\pm,sym} &= \lambda^2(\lambda^2 - \widetilde{\mathcal{D}}^2_\pm)^{-1}, \quad \lambda > 0\\ \widetilde{\mathcal{Q}}^{\lambda}_{\pm,skew} &= -\lambda\widetilde{\mathcal{D}}_\pm(\lambda^2 - \widetilde{\mathcal{D}}^2_\pm)^{-1}, \quad \lambda > 0.\end{aligned}$$

Note that by (6.3.11) we have $\lambda(\lambda + \widetilde{\mathcal{D}}_\pm)^{-1} = \widetilde{\mathcal{Q}}^{\lambda}_{\pm,sym} + \widetilde{\mathcal{Q}}^{\lambda}_{\pm,skew}$.

6.3.2.2 Reformulating Maxwell's equations

We now rewrite Maxwell's equations (6.1.29b)-(6.1.29c) as an equivalent self-adjoint problem using the expression (6.3.10). We start with (6.1.29b):

$$\begin{aligned}0 &= \lambda^2\varphi - \Delta\varphi - \int (f^+ - f^-)\, d\boldsymbol{v}\\ &= \lambda^2\varphi - \Delta\varphi - \int \sum_{\pm} \left(\mu_e^\pm\varphi + r\mu_p^\pm(\mathbf{A}\cdot\boldsymbol{e}_\theta) + \mu_e^\pm\lambda(\lambda + \widetilde{\mathcal{D}}_\pm)^{-1}(-\varphi + \mathbf{A}\cdot\hat{\boldsymbol{v}})\right) d\boldsymbol{v}\end{aligned} \tag{6.3.12}$$

where $\varphi \in \mathfrak{H}_\varphi$. Next, the system of equations (6.1.29c) becomes

$$\begin{aligned}0 &= \lambda^2\mathbf{A} - \boldsymbol{\Delta}\mathbf{A} - \int (f^+ - f^-)\hat{\boldsymbol{v}}\, d\boldsymbol{v}\\ &= \lambda^2\mathbf{A} - \boldsymbol{\Delta}\mathbf{A} - \int \sum_{\pm} \left(\mu_e^\pm\varphi + r\mu_p^\pm(\mathbf{A}\cdot\boldsymbol{e}_\theta) + \mu_e^\pm\lambda(\lambda + \widetilde{\mathcal{D}}_\pm)^{-1}(-\varphi + \mathbf{A}\cdot\hat{\boldsymbol{v}})\right)\hat{\boldsymbol{v}}\, d\boldsymbol{v},\end{aligned} \tag{6.3.13}$$

where $\mathbf{A} = (\mathbf{A}_\theta, \mathbf{A}_{rz}) \in L^2_\theta(\mathbb{R}^3;\mathbb{R}^3) \times L^2_{rz}(\mathbb{R}^3;\mathbb{R}^3)$. As in (6.3.5), we shall write these equations as a single system of the form

$$\widetilde{\boldsymbol{\mathcal{M}}}^\lambda \begin{bmatrix} \mathbf{A}_\theta \\ \varphi \\ \mathbf{A}_{rz} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \tag{6.3.14}$$

which is a self-adjoint operator on the space $L^2_\theta(\mathbb{R}^3;\mathbb{R}^3) \times L^2_{cyl}(\mathbb{R}^3) \times L^2_{rz}(\mathbb{R}^3;\mathbb{R}^3)$, see Lemma 6.6.8. In analogy with (6.3.7), we define

$$\widetilde{\boldsymbol{\mathcal{M}}}^\lambda = \begin{bmatrix} \widetilde{\mathcal{A}}_2^\lambda & (\widetilde{\mathcal{B}}_1^\lambda)^* & (\widetilde{\mathcal{B}}_2^\lambda)^* \\ \widetilde{\mathcal{B}}_1^\lambda & -\widetilde{\mathcal{A}}_1^\lambda & -(\widetilde{\mathcal{B}}_3^\lambda)^* \\ \widetilde{\mathcal{B}}_2^\lambda & -\widetilde{\mathcal{B}}_3^\lambda & -\widetilde{\mathcal{A}}_3^\lambda \end{bmatrix}. \tag{6.3.15}$$

With $\hat{\boldsymbol{v}} = (\hat{v}_r, \hat{v}_\theta, \hat{v}_z)$, we recall the notation $\hat{\boldsymbol{v}}_\theta = \hat{v}_\theta \boldsymbol{e}_\theta$ and $\hat{\boldsymbol{v}}_{rz} = \hat{v}_r \boldsymbol{e}_r + \hat{v}_z \boldsymbol{e}_z$ introduced before. Then the components of $\widetilde{\boldsymbol{\mathcal{M}}}^\lambda$ are now given by

$$\widetilde{\mathcal{A}}_1^\lambda h = -\Delta h + \lambda^2 h + \int \sum_\pm \mu_e^\pm (\widetilde{\mathcal{Q}}^\lambda_{\pm,sym} - 1) h \, d\boldsymbol{v} \tag{6.3.16a}$$

$$\widetilde{\mathcal{A}}_2^\lambda \boldsymbol{h} = -\boldsymbol{\Delta h} + \lambda^2 \boldsymbol{h} - \left(r \int \sum_\pm \mu_p^\pm \hat{v}_\theta \, d\boldsymbol{v} \right) \boldsymbol{h} - \int \sum_\pm \hat{\boldsymbol{v}}_\theta \mu_e^\pm \widetilde{\mathcal{Q}}^\lambda_{\pm,sym} [\boldsymbol{h} \cdot \hat{\boldsymbol{v}}_\theta] \, d\boldsymbol{v} \tag{6.3.16b}$$

$$\widetilde{\mathcal{A}}_3^\lambda \boldsymbol{h} = -\boldsymbol{\Delta h} + \lambda^2 \boldsymbol{h} - \int \sum_\pm \hat{\boldsymbol{v}}_{rz} \mu_e^\pm \widetilde{\mathcal{Q}}^\lambda_{\pm,sym} [\boldsymbol{h} \cdot \hat{\boldsymbol{v}}_{rz}] \, d\boldsymbol{v} \tag{6.3.16c}$$

$$\widetilde{\mathcal{B}}_1^\lambda \boldsymbol{h} = \int \sum_\pm \mu_e^\pm (\widetilde{\mathcal{Q}}^\lambda_{\pm,sym} - 1) [\boldsymbol{h} \cdot \hat{\boldsymbol{v}}_\theta] \, d\boldsymbol{v} \tag{6.3.16d}$$

$$(\widetilde{\mathcal{B}}_1^\lambda)^* h = \int \sum_\pm \mu_e^\pm \hat{\boldsymbol{v}}_\theta (\widetilde{\mathcal{Q}}^\lambda_{\pm,sym} - 1) h \, d\boldsymbol{v} \tag{6.3.16e}$$

$$\widetilde{\mathcal{B}}_2^\lambda \boldsymbol{h} = \int \sum_\pm \mu_e^\pm \hat{\boldsymbol{v}}_{rz} \widetilde{\mathcal{Q}}^\lambda_{\pm,skew} [\boldsymbol{h} \cdot \hat{\boldsymbol{v}}_\theta] \, d\boldsymbol{v} \tag{6.3.16f}$$

$$(\widetilde{\mathcal{B}}_2^\lambda)^* \boldsymbol{h} = -\int \sum_{\pm} \mu_e^{\pm} \hat{\boldsymbol{v}}_\theta \widetilde{\mathcal{Q}}_{\pm,skew}^\lambda \, [\hat{\boldsymbol{v}}_{rz} \cdot \boldsymbol{h}] \, d\boldsymbol{v} \tag{6.3.16g}$$

$$\widetilde{\mathcal{B}}_3^\lambda h = \int \sum_{\pm} \mu_e^{\pm} \hat{\boldsymbol{v}}_{rz} \widetilde{\mathcal{Q}}_{\pm,skew}^\lambda \, h \, d\boldsymbol{v} \tag{6.3.16h}$$

$$(\widetilde{\mathcal{B}}_3^\lambda)^* \boldsymbol{h} = -\int \sum_{\pm} \mu_e^{\pm} \hat{\boldsymbol{v}}_{rz} \widetilde{\mathcal{Q}}_{\pm,skew}^\lambda \, [\boldsymbol{h} \cdot \hat{\boldsymbol{v}}_{rz}] \, d\boldsymbol{v}. \tag{6.3.16i}$$

These are derived from (6.3.12) and (6.3.13), where some terms vanish due to parity in (v_r, v_z), (see Lemma 6.3.2(c)). In particular, in every occurrence of $\lambda(\lambda + \widetilde{\mathcal{D}}_\pm)^{-1} = \widetilde{\mathcal{Q}}_{\pm,sym}^\lambda + \widetilde{\mathcal{Q}}_{\pm,skew}^\lambda$, exactly one of these operators vanishes after integration $d\boldsymbol{v}$. In addition, we have made use of (6.1.34). We further define an operator $\widetilde{\boldsymbol{\mathcal{J}}}^\lambda$ as

$$\widetilde{\boldsymbol{\mathcal{J}}}^\lambda = \begin{bmatrix} \lambda^2 - \boldsymbol{\Delta} & 0 & 0 \\ 0 & -\lambda^2 + \Delta & 0 \\ 0 & 0 & -\lambda^2 + \boldsymbol{\Delta} \end{bmatrix} - \widetilde{\boldsymbol{\mathcal{M}}}^\lambda.$$

Let us briefly discuss these operators in further detail (their precise properties are treated in Section 6.6.2):

- The operators

$$\begin{aligned} &\widetilde{\mathcal{A}}_1^\lambda : H^2_{cyl}(\mathbb{R}^3) \subset L^2_{cyl}(\mathbb{R}^3) \to L^2_{cyl}(\mathbb{R}^3) \\ &\widetilde{\mathcal{A}}_2^\lambda : H^2_\theta(\mathbb{R}^3;\mathbb{R}^3) \subset L^2_\theta(\mathbb{R}^3;\mathbb{R}^3) \to L^2_\theta(\mathbb{R}^3;\mathbb{R}^3) \\ &\widetilde{\mathcal{A}}_3^\lambda : H^2_{rz}(\mathbb{R}^3;\mathbb{R}^3) \subset L^2_{rz}(\mathbb{R}^3;\mathbb{R}^3) \to L^2_{rz}(\mathbb{R}^3;\mathbb{R}^3) \end{aligned}$$

  are self-adjoint, have essential spectrum in $[\lambda^2, \infty)$ and a finite number of eigenvalues in $(-\infty, \lambda^2)$.

- The operators

$$\begin{aligned} &\widetilde{\mathcal{B}}_1^\lambda : L^2_\theta(\mathbb{R}^3;\mathbb{R}^3) \to L^2_{cyl}(\mathbb{R}^3) \\ &\widetilde{\mathcal{B}}_2^\lambda : L^2_\theta(\mathbb{R}^3;\mathbb{R}^3) \to L^2_{rz}(\mathbb{R}^3;\mathbb{R}^3) \\ &\widetilde{\mathcal{B}}_3^\lambda : L^2_{cyl}(\mathbb{R}^3) \to L^2_{rz}(\mathbb{R}^3;\mathbb{R}^3) \end{aligned}$$

are bounded, with bound independent of λ.

- $\widetilde{\mathcal{J}}^{\lambda}: L^2_\theta(\mathbb{R}^3;\mathbb{R}^3)\times L^2_{cyl}(\mathbb{R}^3)\times L^2_{rz}(\mathbb{R}^3;\mathbb{R}^3) \to L^2_\theta(\mathbb{R}^3;\mathbb{R}^3)\times L^2_{cyl}(\mathbb{R}^3)\times L^2_{rz}(\mathbb{R}^3;\mathbb{R}^3)$ is a bounded symmetric operator with bound independent of λ.

6.4 Solving the equivalent problem

The problem is now reduced to finding some $\lambda \in (0,\infty)$ for which the operators $\boldsymbol{\mathcal{M}}^{\lambda}$ (in the $1.5d$ case) and $\widetilde{\boldsymbol{\mathcal{M}}}^{\lambda}$ (in the cylindrically symmetric case) have non-trivial kernels (not the same λ in both cases, of course). Our method is to compare their spectrum for $\lambda = 0$ and λ very large, and use spectral continuity arguments to deduce that as λ varies an eigenvalue must cross through 0 (both operators are self-adjoint, see Lemma 6.6.4 and Lemma 6.6.8 below, hence the spectrum lies on the real axis).

6.4.1 The $1.5d$ case

6.4.1.1 Continuity of the spectrum at $\lambda = 0$

Recall the condition (6.1.9) which we require for instability:

$$\operatorname{neg}(\mathcal{A}_2^0 + (\mathcal{B}^0)^*(\mathcal{A}_1^0)^{-1}\mathcal{B}^0) > \operatorname{neg}(\mathcal{A}_1^0). \tag{6.4.1}$$

We wish to move this condition to values of λ greater than 0:

Lemma 6.4.1. *Assume that (6.4.1) holds and that zero is in the resolvent set of $\mathcal{A}_1^0$. Then there exists $\lambda_* > 0$ such that for all $\lambda \in [0,\lambda_*]$*

$$\operatorname{neg}(\mathcal{A}_2^\lambda + (\mathcal{B}^\lambda)^*(\mathcal{A}_1^\lambda)^{-1}\mathcal{B}^\lambda) > \operatorname{neg}(\mathcal{A}_1^\lambda).$$

Proof. The proof follows immediately from the following three simple steps:

Step 1. $\mathcal{A}_1^\lambda$ is invertible for small $\lambda \geq 0$. We know from Lemma 6.6.3 (below) that $\mathcal{A}_1^\lambda$ is continuous in the norm resolvent sense and has discrete spectrum. The

norm resolvent continuity implies that its spectrum varies continuously in λ, so as 0 is not in its spectrum at $\lambda = 0$ there exists λ_* such that 0 is not in the spectrum for $0 \leq \lambda \leq \lambda_*$. Hence for all such λ, $\mathcal{A}_1^\lambda$ is invertible and the operator $\mathcal{A}_2^\lambda + (\mathcal{B}^\lambda)^*(\mathcal{A}_1^\lambda)^{-1}\mathcal{B}^\lambda$ is well defined.

Step 2. $\operatorname{neg}(\mathcal{A}_1^\lambda) = \operatorname{neg}(\mathcal{A}_1^0)$ **for all** $\lambda \in [0, \lambda_*]$**.** The spectrum of $\mathcal{A}_1^\lambda$ is purely discrete and 0 is in its resolvent set. This means that none of its eigenvalues can cross 0 for small values of λ.

Step 3. $\operatorname{neg}(\mathcal{A}_2^\lambda + (\mathcal{B}^\lambda)^*(\mathcal{A}_1^\lambda)^{-1}\mathcal{B}^\lambda) \geq \operatorname{neg}(\mathcal{A}_2^0 + (\mathcal{B}^0)^*(\mathcal{A}_1^0)^{-1}\mathcal{B}^0)$ **for all** $\lambda \in [0, \lambda_*]$**.** Observe that

- $[0, \infty) \ni \lambda \mapsto \mathcal{A}_2^\lambda + (\mathcal{B}^\lambda)^*(\mathcal{A}_1^\lambda)^{-1}\mathcal{B}^\lambda$ is norm resolvent continuous,
- $\mathcal{A}_2^\lambda + (\mathcal{B}^\lambda)^*(\mathcal{A}_1^\lambda)^{-1}\mathcal{B}^\lambda$ has essential spectrum in $[\lambda^2, \infty)$,
- $\mathcal{A}_2^\lambda + (\mathcal{B}^\lambda)^*(\mathcal{A}_1^\lambda)^{-1}\mathcal{B}^\lambda$ has finitely many negative eigenvalues.

These statements follow from arguments similar to those appearing in the proof of Lemma 6.6.3(a)-(c), the last by the boundedness of the perturbation and the location of the essential spectrum (see Lemma 6.2.2). Since 0 is not in the resolvent set at $\lambda = 0$ we pick $\sigma < 0$ larger than all the (finitely many) negative eigenvalues of $\mathcal{A}_2^0 + (\mathcal{B}^0)^*(\mathcal{A}_1^0)^{-1}\mathcal{B}^0$. The continuous dependence of the spectrum (as a set) on the parameter λ implies that for small values of λ no eigenvalues cross σ and the number of negative eigenvalues can only grow as λ increases. □

6.4.1.2 Truncation

We follow the plan hinted at in Section 6.2.2: first we discretise the spectrum, then truncate. The only continuous part in the spectrum of $\boldsymbol{\mathcal{M}}^\lambda$ is due to $\mathcal{A}_2^\lambda$, hence we let $W(x)$ be a smooth positive potential function satisfying $W(x) \to \infty$ as $x \to \pm\infty$ which we shall add to $\mathcal{A}_2^\lambda$. It is well known that the Schrödinger operator $-\partial_x^2 + W$ on $L^2(\mathbb{R})$ is self-adjoint (on an appropriate domain therein) with compact resolvent (and therefore discrete spectrum). Moreover, $C_0^\infty(\mathbb{R})$ is a core for both $\partial_x^2 + W$ and ∂_x^2. Thus our approximating operator family is

$\{\boldsymbol{\mathcal{M}}_\varepsilon^\lambda\}_{\lambda\in[\lambda_*,\infty),\varepsilon\in[0,\infty)}$, where

$$\boldsymbol{\mathcal{M}}_\varepsilon^\lambda = \begin{bmatrix} \mathcal{A}_{2,\varepsilon}^\lambda & (\mathcal{B}^\lambda)^* \\ \mathcal{B}^\lambda & -\mathcal{A}_1^\lambda \end{bmatrix} = \underbrace{\begin{bmatrix} -\partial_x^2 + \varepsilon W & 0 \\ 0 & \partial_x^2 \end{bmatrix} + \begin{bmatrix} \lambda^2 & 0 \\ 0 & 0 \end{bmatrix}}_{\boldsymbol{\mathcal{A}}_\varepsilon^\lambda} - \boldsymbol{\mathcal{J}}^\lambda$$

defined on $L^2(\mathbb{R}) \times L_0^2(\Omega)$ and where λ_* is as given in Lemma 6.4.1. For $\varepsilon > 0$ this operator has discrete spectrum. As indicated in the statement of Theorem 6.2.1 and the preceding definitions, we define truncated versions using the eigenspaces of the operator $\boldsymbol{\mathcal{A}}_\varepsilon^\lambda$. As this operator is diagonal, we can choose the eigenvectors to lie in exactly one of $L^2(\mathbb{R})$ or $L_0^2(\Omega)$. We denote the nth truncation, a projection onto an eigenspace of dimension $2n$ consisting of n eigenvectors in each of $L^2(\mathbb{R})$ and $L_0^2(\Omega)$, as $\boldsymbol{\mathcal{M}}_{\varepsilon,n}^\lambda$ which is self-adjoint and defined for $\varepsilon > 0, \lambda \geq 0, n \in \mathbb{N}$. Moreover, the mapping $\lambda \mapsto \mathrm{sp}(\boldsymbol{\mathcal{M}}_{\varepsilon,n}^\lambda)$ is continuous (that is, the set of eigenvalues varies continuously). In particular, if there are $\lambda_* < \lambda^*$ for which $\mathrm{neg}(\boldsymbol{\mathcal{M}}_{\varepsilon,n}^{\lambda_*}) \neq \mathrm{neg}(\boldsymbol{\mathcal{M}}_{\varepsilon,n}^{\lambda^*})$ then there must exist $\lambda_{\varepsilon,n} \in (\lambda_*, \lambda^*)$ for which $0 \in \mathrm{sp}(\boldsymbol{\mathcal{M}}_{\varepsilon,n}^\lambda)$. We have therefore just proved:

Lemma 6.4.2. *Fix $\varepsilon > 0, n \in \mathbb{N}$. Suppose that there exist $0 < \lambda_* < \lambda^* < \infty$ such that $\mathrm{neg}(\boldsymbol{\mathcal{M}}_{\varepsilon,n}^{\lambda_*}) \neq \mathrm{neg}(\boldsymbol{\mathcal{M}}_{\varepsilon,n}^{\lambda^*})$. Then there is a $\lambda_{\varepsilon,n} \in (\lambda_*, \lambda^*)$ for which $\ker(\boldsymbol{\mathcal{M}}_{\varepsilon,n}^\lambda)$ is non-trivial.*

The next step is thus to establish estimates on $\mathrm{neg}(\boldsymbol{\mathcal{M}}_{\varepsilon,n}^\lambda)$.

6.4.1.3 The spectrum for large λ

We begin by looking at $\mathrm{neg}(\boldsymbol{\mathcal{M}}_{\varepsilon,n}^\lambda)$ when λ is large. This turns out to be relatively simple due to the block form of the untruncated operator.

Lemma 6.4.3. *There is $\lambda^* > 0$ such that for all $\lambda \geq \lambda^*$, $\varepsilon > 0$, and $n \in \mathbb{N}$, the truncated operator $\boldsymbol{\mathcal{M}}_{\varepsilon,n}^\lambda$ has spectrum composed of exactly n positive and n negative eigenvalues. In particular $\mathrm{neg}(\boldsymbol{\mathcal{M}}_{\varepsilon,n}^\lambda) = n$.*

Proof. Take $\boldsymbol{u} = (u_1, 0) \in L^2(\mathbb{R}) \times L_0^2(\Omega)$ with $u_1 \in \mathfrak{D}(\mathcal{A}_{2,\varepsilon}^\lambda)$, $\|\boldsymbol{u}\|_{L^2(\mathbb{R})\times L^2(\Omega)} = 1$

and $\boldsymbol{u}$ in the $2n$ dimensional subspace associated with the truncation. Then,

$$\begin{aligned}\left\langle \boldsymbol{\mathcal{M}}^{\lambda}_{\varepsilon,n}\boldsymbol{u},\boldsymbol{u}\right\rangle_{L^2(\mathbb{R})\times L^2_0(\Omega)} &= \left\langle \mathcal{A}^{\lambda}_{1,\varepsilon,n}u_1,u_1\right\rangle_{L^2(\mathbb{R})} = \left\langle \mathcal{A}^{\lambda}_{1,\varepsilon}u_1,u_1\right\rangle_{L^2(\mathbb{R})}\\ &= \left\langle \mathcal{A}^{\lambda}_{1}u_1,u_1\right\rangle_{L^2(\mathbb{R})} + \varepsilon\left\|\sqrt{W}u_1\right\|^2_{L^2(\mathbb{R})}.\end{aligned}$$

As the second term is non-negative we may apply Lemma 6.6.3(d) to see that, for all large enough λ (independently of n and ε), $\boldsymbol{\mathcal{M}}^{\lambda}_{\varepsilon,n}$ is positive definite on a subspace of dimension n, so has n positive eigenvalues. Performing the same computation on $\boldsymbol{u} = (0,u_2)$ in the subspace associated with the truncation and with $u_2 \in \mathfrak{D}(\mathcal{A}^{\lambda}_{1,\varepsilon})$, we obtain that for large enough λ, $\boldsymbol{\mathcal{M}}^{\lambda}_{\varepsilon,n}$ is negative definite on a subspace of dimension n. As $\boldsymbol{\mathcal{M}}^{\lambda}_{\varepsilon,n}$ has exactly $2n$ eigenvalues the proof is complete. □

6.4.1.4 The spectrum for small λ

We now consider $\mathrm{sp}(\boldsymbol{\mathcal{M}}^{\lambda_*}_{\varepsilon,n})$. We recall the result on spectra of real block matrix operators in [17]:

Lemma 6.4.4. *Let M be the real symmetric block matrix*

$$M = \begin{bmatrix} A_2 & B^T \\ B & -A_1 \end{bmatrix}$$

with A_1 invertible. Then M has the same number of negative eigenvalues as the matrix

$$N = \begin{bmatrix} A_2 + B^T A_1^{-1} B & 0 \\ 0 & -A_1 \end{bmatrix}.$$

Lemma 6.4.5. *Assume that (6.4.1) holds and that zero is in the resolvent set of $\mathcal{A}^0_1$. Then there exist $\lambda_*, \varepsilon_* > 0$ such that for all $\varepsilon \in (0,\varepsilon_*)$ there is $N > 0$ such that for all $n > N$ the operator $\boldsymbol{\mathcal{M}}^{\lambda_*}_{\varepsilon,n}$ satisfies*

$$\mathrm{neg}(\boldsymbol{\mathcal{M}}^{\lambda_*}_{\varepsilon,n}) \geq \mathrm{neg}(\mathcal{A}^0_2 + (\mathcal{B}^0)^*(\mathcal{A}^0_1)^{-1}\mathcal{B}^0) + n - \mathrm{neg}(\mathcal{A}^0_1).$$

Proof. The number λ_* is the one given in Lemma 6.4.1, and satisfies that for all $\lambda \in [0,\lambda_*]$ the kernel of $\mathcal{A}^{\lambda}_1$ is trivial. Since eigenvalues (counting multiplicity) are stable under strong resolvent perturbations (see [109, VIII.3.5.Thm 3.15.]), there

exists $\varepsilon_* > 0$ such that $\operatorname{neg}(\mathcal{A}^0_{2,\varepsilon} + (\mathcal{B}^0)^*(\mathcal{A}^0_1)^{-1}\mathcal{B}^0) \geq \operatorname{neg}(\mathcal{A}^0_2 + (\mathcal{B}^0)^*(\mathcal{A}^0_1)^{-1}\mathcal{B}^0)$ for all $\varepsilon \in [0, \varepsilon_*]$. The result then follows from Lemma 6.4.4, since $\operatorname{neg}(\boldsymbol{\mathcal{M}}^{\lambda_*}_{\varepsilon,n}) = \operatorname{neg}(\mathcal{A}^{\lambda_*}_{2,\varepsilon} + (\mathcal{B}^{\lambda_*})^*(\mathcal{A}^{\lambda_*}_1)^{-1}\mathcal{B}^{\lambda_*}) + n - \operatorname{neg}(\mathcal{A}^{\lambda_*}_1)$. □

6.4.2 The cylindrically symmetric case

For brevity we write

$$\widetilde{\boldsymbol{\mathcal{M}}}^\lambda = \begin{bmatrix} \widetilde{\mathcal{A}}^\lambda_2 & (\widetilde{\boldsymbol{\mathcal{B}}}^\lambda_4)^* \\ \widetilde{\boldsymbol{\mathcal{B}}}^\lambda_4 & -\widetilde{\boldsymbol{\mathcal{A}}}^\lambda_4 \end{bmatrix}$$

where

$$\widetilde{\boldsymbol{\mathcal{A}}}^\lambda_4 = \begin{bmatrix} \widetilde{\mathcal{A}}^\lambda_1 & (\widetilde{\mathcal{B}}^\lambda_3)^* \\ \widetilde{\mathcal{B}}^\lambda_3 & \widetilde{\mathcal{A}}^\lambda_3 \end{bmatrix} \qquad \text{and} \qquad \widetilde{\boldsymbol{\mathcal{B}}}^\lambda_4 = \begin{bmatrix} \widetilde{\mathcal{B}}^\lambda_1 \\ \widetilde{\mathcal{B}}^\lambda_2 \end{bmatrix}.$$

6.4.2.1 Continuity of the spectrum at $\lambda = 0$

Lemma 6.4.6. *Assume that (6.1.10) holds, that $\widetilde{\mathcal{A}}^0_3$ does not have 0 as an L^6-eigenvalue (see Definition 6.1.3) and that $\widetilde{\mathcal{A}}^0_1$ does not have 0 as an eigenvalue. Then there exists $\lambda_* > 0$ such that for $\lambda \in [0, \lambda_*]$,*

$$\operatorname{neg}(\widetilde{\mathcal{A}}^\lambda_2 + (\widetilde{\boldsymbol{\mathcal{B}}}^\lambda_4)^*(\widetilde{\boldsymbol{\mathcal{A}}}^\lambda_4)^{-1}\widetilde{\boldsymbol{\mathcal{B}}}^\lambda_4) > \operatorname{neg}(\widetilde{\boldsymbol{\mathcal{A}}}^\lambda_4).$$

Proof. We first note that as the mean perturbed charge is zero (this is since $\int \rho \, d\boldsymbol{x}$ is an invariant of the linearised system), it follows from direct computation on the Green's function of the Laplacian that any L^6-eigenfunction of $\widetilde{\mathcal{A}}^0_1$ will also be square integrable and so be a proper eigenfunction. Indeed, for any L^6 eigenfunction u of $\widetilde{\mathcal{A}}^0_1$, we may define ρ by $\rho = -\Delta u$ which satisfies $\int \rho \, d\boldsymbol{x} = 0$ and has compact support. Then, for $\boldsymbol{x}$ outside the support of ρ, we have

$$\begin{aligned} u(\boldsymbol{x}) &= \frac{1}{4\pi} \int \frac{\rho(\boldsymbol{y})}{|\boldsymbol{x} - \boldsymbol{y}|} \, d\boldsymbol{y} \\ &= \frac{1}{4\pi|\boldsymbol{x}|} \int \rho(\boldsymbol{y}) \, d\boldsymbol{y} + \frac{1}{4\pi|\boldsymbol{x}|^2} \int \rho(\boldsymbol{y}) \frac{|\boldsymbol{x}|(|\boldsymbol{x}| - |\boldsymbol{x} - \boldsymbol{y}|)}{|\boldsymbol{x} - \boldsymbol{y}|} \, d\boldsymbol{y}. \end{aligned}$$

The first term vanishes, and using the compact support of ρ it is easily seen that the second integral is bounded independently of $\boldsymbol{x}$. Hence $u(\boldsymbol{x})$ decays like $C|\boldsymbol{x}|^{-2}$

for large $|\boldsymbol{x}|$ and square integrability follows. Note also that $\widetilde{\mathcal{B}}_3^0 = 0$. Thus $\widetilde{\boldsymbol{\mathcal{A}}}_4^0$ having an L^6-eigenfunction of 0 contradicts our assumptions, a fact that we will later use.

We model the proof on that of Lemma 6.4.1, splitting it into 4 steps.

Step 1. $\widetilde{\boldsymbol{\mathcal{A}}}_4^\lambda$ is invertible for small $\lambda \geq 0$ when restricted to functions supported in Ω. Let $\boldsymbol{\mathcal{P}} \in \mathfrak{B}\left(L^2_{cyl}(\mathbb{R}^3) \times L^2_{rz}(\mathbb{R}^3;\mathbb{R}^3)\right)$ be multiplication by the indicator function of Ω. We claim that for all small enough $\lambda > 0$, $\boldsymbol{\mathcal{P}}(\widetilde{\boldsymbol{\mathcal{A}}}_4^\lambda)^{-1}\boldsymbol{\mathcal{P}}$ is a well defined bounded operator that is strongly continuous in $\lambda > 0$ and has a strong limit as $\lambda \to 0$. To prove this, we argue that if this were not the case, then 0 would be an L^6-eigenvalue of $\widetilde{\boldsymbol{\mathcal{A}}}_4^0$, a contradiction.

As $L^2_{cyl}(\mathbb{R}^3) \times L^2_{rz}(\mathbb{R}^3;\mathbb{R}^3)$ is a closed subspace of $L^2(\mathbb{R}^3;\mathbb{R}^4)$ we may work in the larger space to ease notation. To this end, let $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ denote the $L^2(\mathbb{R}^3;\mathbb{R}^4)$ norm and inner product. We can express $\widetilde{\boldsymbol{\mathcal{A}}}_4^\lambda$ in the form

$$\widetilde{\boldsymbol{\mathcal{A}}}_4^\lambda \boldsymbol{u} = -\boldsymbol{\Delta}\boldsymbol{u} + \lambda^2 \boldsymbol{u} + \boldsymbol{\mathcal{K}}^\lambda \boldsymbol{u}$$

where $\boldsymbol{\mathcal{K}}^\lambda$ is uniformly bounded, strongly continuous in $\lambda \geq 0$ and $\boldsymbol{\mathcal{K}}^\lambda = \boldsymbol{\mathcal{P}}\boldsymbol{\mathcal{K}}^\lambda\boldsymbol{\mathcal{P}}$.

Step 1.1. $\widetilde{\boldsymbol{\mathcal{A}}}_4^\lambda$ is bounded from below when restricted to functions supported in Ω. First we claim that there exist constants $\lambda' > 0$ and $C > 0$ such that we have the uniform lower bound

$$\left\| \mathbb{1}_\Omega \widetilde{\boldsymbol{\mathcal{A}}}_4^\lambda \boldsymbol{u}^\lambda \right\| \geq C \left\| \mathbb{1}_\Omega \boldsymbol{u}^\lambda \right\|, \qquad \forall \lambda \in (0, \lambda'] \tag{6.4.2}$$

where the constant C does not depend on λ or on $\boldsymbol{u}^\lambda$ and where $\boldsymbol{u}^\lambda$ satisfies $\widetilde{\boldsymbol{\mathcal{A}}}_4^\lambda \boldsymbol{u}^\lambda = 0$ outside Ω. Indeed, if not there would be sequences $\lambda_n \to 0$ and $\{\boldsymbol{u}_n\}_{n=1}^\infty$ with $\|\mathbb{1}_\Omega \boldsymbol{u}_n\|_{L^2} = 1$ that satisfy

$$\widetilde{\boldsymbol{\mathcal{A}}}_4^{\lambda_n} \boldsymbol{u}_n = -\boldsymbol{\Delta}\boldsymbol{u}_n + \lambda_n^2 \boldsymbol{u}_n + \boldsymbol{\mathcal{K}}^{\lambda_n}\boldsymbol{u}_n = \boldsymbol{f}_n \to 0 \tag{6.4.3}$$

as $n \to \infty$, with $\boldsymbol{f}_n$ supported in Ω. Hence,

$$\|\nabla \boldsymbol{u}_n\|^2 + \lambda_n^2 \|\boldsymbol{u}_n\|^2 + \left\langle \boldsymbol{\mathcal{K}}^{\lambda_n} \boldsymbol{u}_n, \boldsymbol{u}_n \right\rangle = \langle \boldsymbol{f}_n, \boldsymbol{u}_n \rangle \to 0 \tag{6.4.4}$$

so that $\|\nabla \boldsymbol{u}_n\|^2$ is uniformly bounded for large enough n. Therefore, there exists a subsequence (we abuse notation and keep the same sequence) such that $\nabla \boldsymbol{u}_n \rightharpoonup \boldsymbol{v}$ weakly in $L^2(\mathbb{R}^3; \mathbb{R}^4)$ for some $\boldsymbol{v} \in L^2(\mathbb{R}^3; \mathbb{R}^4)$. By the standard Sobolev inequality $\|\varphi\|_{L^6(\mathbb{R}^3)} \leq C \|\nabla \varphi\|_{L^2(\mathbb{R}^3)}$ we have a uniform bound on $\|\boldsymbol{u}_n\|_{L^6(\mathbb{R}^3;\mathbb{R}^4)}$. Therefore, by passing again to a subsequence if necessary, we have the convergence $\boldsymbol{u}_n \rightharpoonup \boldsymbol{u}$ weakly in $L^6(\mathbb{R}^3; \mathbb{R}^4)$ for some $\boldsymbol{u} \in L^6(\mathbb{R}^3; \mathbb{R}^4)$. Furthermore, by Rellich's theorem we have the strong convergence $\boldsymbol{u}_n \to \boldsymbol{u}$ in $L^2_{loc}(\mathbb{R}^3; \mathbb{R}^4)$. This implies that necessarily $\boldsymbol{v} = \nabla \boldsymbol{u}$. In particular we deduce that $\|\mathbb{1}_\Omega \boldsymbol{u}\| = 1$ so $\boldsymbol{u} \neq 0$. Passing to the limit in (6.4.3), $\boldsymbol{u}$ satisfies

$$-\boldsymbol{\Delta} \boldsymbol{u} + \boldsymbol{\mathcal{K}}^0 \boldsymbol{u} = 0$$

in the sense of distributions and by elliptic regularity $\boldsymbol{u} \in H^2_{loc}(\mathbb{R}^3; \mathbb{R}^4)$. In fact $\boldsymbol{u}$ is an L^6-eigenfunction of $\widetilde{\boldsymbol{\mathcal{A}}}_4^0$ with eigenvalue 0, which contradicts our assumptions. This proves the claim.

Step 1.2. $\widetilde{\boldsymbol{\mathcal{A}}}_4^\lambda$ is invertible for all small enough $\lambda > 0$. For any $\lambda > 0$, 0 does not lie in the essential spectrum of $\widetilde{\boldsymbol{\mathcal{A}}}_4^\lambda$ so is either an eigenvalue or in the resolvent set. Let $\lambda > 0$ be small enough so that (6.4.2) holds, then any eigenfunction $\boldsymbol{u}$ of 0 satisfies all the assumptions of the claim above, and hence $\|\mathbb{1}_\Omega \boldsymbol{u}\| \leq C^{-1} \left\| \mathbb{1}_\Omega \widetilde{\boldsymbol{\mathcal{A}}}_4^\lambda \boldsymbol{u} \right\| = 0$ so that $\boldsymbol{u} = 0$ inside Ω. Clearly this implies that $\boldsymbol{u} = 0$ in $\mathbb{R}^3$ which is a contradiction. In the same way we deduce a uniform bound C from below for the operator $\boldsymbol{\mathcal{P}}(\widetilde{\boldsymbol{\mathcal{A}}}_4^\lambda)^{-1}\boldsymbol{\mathcal{P}}$ for such small $\lambda > 0$.

Step 1.3. $\boldsymbol{\mathcal{P}}(\widetilde{\boldsymbol{\mathcal{A}}}_4^0)^{-1}\boldsymbol{\mathcal{P}}$ is well defined and bounded. Finally we give a meaning to $\boldsymbol{\mathcal{P}}(\widetilde{\boldsymbol{\mathcal{A}}}_4^0)^{-1}\boldsymbol{\mathcal{P}}$ (which is required as $\widetilde{\boldsymbol{\mathcal{A}}}_4^0$ is not invertible on the whole space). We define it to be the strong operator limit of $\boldsymbol{\mathcal{P}}(\widetilde{\boldsymbol{\mathcal{A}}}_4^\lambda)^{-1}\boldsymbol{\mathcal{P}}$ as $\lambda \to 0$. Indeed, suppose that $\boldsymbol{f}$ is fixed with support in Ω and $\lambda_n \to 0$. Then we wish to compute the limit of $\boldsymbol{\mathcal{P}}\boldsymbol{u}_n$ for $\boldsymbol{u}_n = (\widetilde{\boldsymbol{\mathcal{A}}}_4^{\lambda_n})^{-1}\boldsymbol{\mathcal{P}}\boldsymbol{f}$ as $n \to \infty$ and show that it is

independent of the sequence $\lambda_n \to 0$. Indeed $\boldsymbol{u}_n$ will satisfy

$$\widetilde{\boldsymbol{\mathcal{A}}}_4^{\lambda_n}\boldsymbol{u}_n = \lambda_n^2\boldsymbol{u}_n - \boldsymbol{\Delta}\boldsymbol{u}_n + \boldsymbol{\mathcal{K}}^{\lambda_n}\boldsymbol{u}_n = \boldsymbol{f}.$$

By the same argument as before we can extract a subsequence and limit $\boldsymbol{u} \in L^6(\mathbb{R}^3;\mathbb{R}^4)$ with convergences as in Step 1.1. In particular $\boldsymbol{\mathcal{P}}\boldsymbol{u}_n \to \boldsymbol{\mathcal{P}}\boldsymbol{u}$. We claim that the limit $\boldsymbol{u}$ is independent of the limiting sequence $\lambda_n \to 0$. Indeed, if two different limits $\boldsymbol{u}$ and $\boldsymbol{v}$ existed, then their difference $\boldsymbol{w} = \boldsymbol{u} - \boldsymbol{v} \in L^6(\mathbb{R}^3;\mathbb{R}^4)$ would solve $\widetilde{\boldsymbol{\mathcal{A}}}_4^0\boldsymbol{w} = 0$, i.e. would be an L^6-eigenfunction with eigenvalue 0, which we assumed impossible.

Finally, the uniform bound (6.4.2) implies that the approximations $\boldsymbol{\mathcal{P}}(\widetilde{\boldsymbol{\mathcal{A}}}_4^{\lambda})^{-1}\boldsymbol{\mathcal{P}}$ are uniformly bounded in operator norm for all sufficiently small positive λ. The convergence, for all $\boldsymbol{u} \in L^2(\mathbb{R}^3;\mathbb{R}^4)$, $\boldsymbol{\mathcal{P}}(\widetilde{\boldsymbol{\mathcal{A}}}_4^{\lambda})^{-1}\boldsymbol{\mathcal{P}}\boldsymbol{u} \to \boldsymbol{\mathcal{P}}(\widetilde{\boldsymbol{\mathcal{A}}}_4^{0})^{-1}\boldsymbol{\mathcal{P}}\boldsymbol{u}$ as $\lambda \to 0$ implies that the limiting operator has the same bound in operator norm.

Step 2. $\operatorname{neg}(\widetilde{\boldsymbol{\mathcal{A}}}_4^{\lambda}) = \operatorname{neg}(\widetilde{\boldsymbol{\mathcal{A}}}_4^{0})$ **for all** $\lambda \in [0, \lambda_*]$**.** $\widetilde{\boldsymbol{\mathcal{A}}}_4^{\lambda}$ is norm resolvent continuous in $\lambda \geq 0$, so the only way the number of negative eigenvalues could change is for an eigenvalue to be absorbed into the essential spectrum at 0 as $\lambda \to 0$. Assume this happens, then we have a sequence $\lambda_n \to 0$, a sequence of negative eigenvalues $\sigma_n \to 0$ and eigenfunctions $\boldsymbol{u}_n$ which satisfy

$$-\boldsymbol{\Delta}\boldsymbol{u}_n + \lambda_n^2\boldsymbol{u}_n + \boldsymbol{\mathcal{K}}^{\lambda_n}\boldsymbol{u}_n = \sigma_n\boldsymbol{u}_n.$$

By the same argument as in the previous steps, we may take subsequences and obtain a contradiction.

Step 3. $\operatorname{neg}(\widetilde{\mathcal{A}}_2^{\lambda} + (\widetilde{\boldsymbol{\mathcal{B}}}_4^{\lambda})^*(\widetilde{\boldsymbol{\mathcal{A}}}_4^{\lambda})^{-1}\widetilde{\boldsymbol{\mathcal{B}}}_4^{\lambda}) \geq \operatorname{neg}(\widetilde{\mathcal{A}}_2^{0} + (\widetilde{\boldsymbol{\mathcal{B}}}_4^{0})^*(\widetilde{\boldsymbol{\mathcal{A}}}_4^{0})^{-1}\widetilde{\boldsymbol{\mathcal{B}}}_4^{0})$ **for all** $\lambda \in [0, \lambda_*]$**.** This may be proved in the same way as Step 3 of Lemma 6.4.1.

Step 4. $\operatorname{neg}(\widetilde{\mathcal{A}}_2^{0} + (\widetilde{\boldsymbol{\mathcal{B}}}_4^{0})^*(\widetilde{\boldsymbol{\mathcal{A}}}_4^{0})^{-1}\widetilde{\boldsymbol{\mathcal{B}}}_4^{0}) > \operatorname{neg}(\widetilde{\boldsymbol{\mathcal{A}}}_4^{0})$**.** As $\widetilde{\mathcal{B}}_2^0 = 0$ and $\widetilde{\mathcal{B}}_3^0 = 0$ we have,

$$\begin{aligned}&\operatorname{neg}(\widetilde{\mathcal{A}}_2^{0} + (\widetilde{\boldsymbol{\mathcal{B}}}_4^{0})^*(\widetilde{\boldsymbol{\mathcal{A}}}_4^{0})^{-1}\widetilde{\boldsymbol{\mathcal{B}}}_4^{0})\\ &\quad = \operatorname{neg}(\widetilde{\mathcal{A}}_2^{0} + (\widetilde{\mathcal{B}}_1^{0})^*(\widetilde{\mathcal{A}}_1^{0})^{-1}\widetilde{\mathcal{B}}_1^{0}) > \operatorname{neg}(\widetilde{\mathcal{A}}_1^{0}) + \operatorname{neg}(\widetilde{\mathcal{A}}_3^{0}) = \operatorname{neg}(\widetilde{\boldsymbol{\mathcal{A}}}_4^{0})\end{aligned}$$

where the inequality is obtained from the assumption of the lemma. □

6.4.2.2 Finding a non-trivial kernel

The next few steps of the proof follow those of the $1.5d$ case, hence we only provide a short overview.

Truncation. As the domain is unbounded, each Laplacian appearing in the problem contributes an essential spectrum on $[0, \infty)$. We therefore introduce a smooth positive potential function $W : \mathbb{R}^3 \to \mathbb{R}$ satisfying $W(x) \to \infty$ as $|x| \to \infty$ and denote by $W^{\otimes n}$ the n-dimensional vector-valued function with n copies of W. Then we define

$$\widetilde{\boldsymbol{\mathcal{M}}}_{\varepsilon}^{\lambda} = \begin{bmatrix} \widetilde{\mathcal{A}}_{2,\varepsilon}^{\lambda} & (\widetilde{\boldsymbol{\mathcal{B}}}_{4}^{\lambda})^{*} \\ \widetilde{\boldsymbol{\mathcal{B}}}_{4}^{\lambda} & -\widetilde{\boldsymbol{\mathcal{A}}}_{4,\varepsilon}^{\lambda} \end{bmatrix} = \underbrace{\begin{bmatrix} -\boldsymbol{\Delta} + \varepsilon W^{\otimes 3} & 0 \\ 0 & \boldsymbol{\Delta} - \varepsilon W^{\otimes 4} \end{bmatrix} + \begin{bmatrix} \lambda^2 & 0 \\ 0 & -\lambda^2 \end{bmatrix}}_{\widetilde{\boldsymbol{\mathcal{A}}}_{\varepsilon}^{\lambda}} - \widetilde{\boldsymbol{\mathcal{J}}}^{\lambda}.$$

As above we can naturally define finite-dimensional operators $\widetilde{\boldsymbol{\mathcal{M}}}_{\varepsilon,n}^{\lambda}$, for which we can easily prove:

Lemma 6.4.7. *Fix $\varepsilon > 0, n \in \mathbb{N}$. Suppose that there exist $0 < \lambda_* < \lambda^* < \infty$ such that $\operatorname{neg}(\widetilde{\boldsymbol{\mathcal{M}}}_{\varepsilon,n}^{\lambda_*}) \neq \operatorname{neg}(\widetilde{\boldsymbol{\mathcal{M}}}_{\varepsilon,n}^{\lambda^*})$. Then there exists $\lambda_{\varepsilon,n} \in (\lambda_*, \lambda^*)$ for which $\ker(\widetilde{\boldsymbol{\mathcal{M}}}_{\varepsilon,n}^{\lambda})$ is non-trivial.*

The spectrum for large λ. This is again similar to the $1.5d$ case, in particular due to the appearance of the λ^2 terms. We have:

Lemma 6.4.8. *There is a number $\lambda^* > 0$ such that for all $\lambda \geq \lambda^*$, $\varepsilon > 0$, and $n \in \mathbb{N}$, the truncated operator $\widetilde{\boldsymbol{\mathcal{M}}}_{\varepsilon,n}^{\lambda}$ has spectrum composed of exactly n positive and n negative eigenvalues. In particular $\operatorname{neg}(\widetilde{\boldsymbol{\mathcal{M}}}_{\varepsilon,n}^{\lambda}) = n$.*

The spectrum for small λ. Again this is similar to the $1.5d$ case.

Lemma 6.4.9. *Assume that (6.1.10) holds and that zero is neither an eigenvalue of $\widetilde{\mathcal{A}}_1^0$ nor is it an L^6-eigenvalue of $\widetilde{\mathcal{A}}_3^0$. Then there exist $\lambda_*, \varepsilon_* > 0$ such that for*

all $\varepsilon \in (0, \varepsilon_*)$ *there is* $N > 0$ *such that for all* $n > N$ *the operator* $\widetilde{\boldsymbol{\mathcal{M}}}^{\lambda_*}_{\varepsilon,n}$ *satisfies*

$$\operatorname{neg}(\widetilde{\boldsymbol{\mathcal{M}}}^{\lambda_*}_{\varepsilon,n}) \geq \operatorname{neg}\left(\widetilde{\mathcal{A}}^0_2 + \left(\widetilde{\mathcal{B}}^0_1\right)^* \left(\widetilde{\mathcal{A}}^0_1\right)^{-1} \widetilde{\mathcal{B}}^0_1\right) + n - \operatorname{neg}\left(\widetilde{\mathcal{A}}^0_1\right) - \operatorname{neg}\left(\widetilde{\mathcal{A}}^0_3\right). \quad (6.4.5)$$

6.5 Proofs of the main theorems

In this section we complete the proofs of Theorem 6.1.1 and Theorem 6.1.2. In both settings – the $1.5d$ and the cylindrically symmetric – we first show that the results of Section 6.4 imply that there exists some $\lambda > 0$ such that the equivalent problems (6.3.5) and (6.3.14) have a non-trivial solution (the λ need not be the same in both cases, of course). Then we show that these non-trivial solutions lead to genuine non-trivial solutions of the linearised RVM in either case.

6.5.1 The $1.5d$ case

6.5.1.1 Existence of a non-trivial kernel of the equivalent problem

By Lemma 6.4.3 and Lemma 6.4.5 we have $0 < \lambda_* < \lambda^* < \infty$ and $\varepsilon_* > 0$ such that for any $\varepsilon < \varepsilon_*$ there is an N_ε such that for $n > N_\varepsilon$ we have,

$$\begin{aligned}\operatorname{neg}(\mathcal{M}^{\lambda_*}_{\varepsilon,n}) &\geq \operatorname{neg}(\mathcal{A}^0_2 + (\mathcal{B}^0)^*(\mathcal{A}^0_1)^{-1}\mathcal{B}^0) + n - \operatorname{neg}(\mathcal{A}^0_1) \\ &> n = \operatorname{neg}(\boldsymbol{\mathcal{M}}^{\lambda^*}_{\varepsilon,n}),\end{aligned}$$

where the strict inequality is due to the assumption (6.1.9). Fix $\varepsilon \in (0, \varepsilon_*)$. By Lemma 6.4.2 for each $n > N_\varepsilon$ there exists $\lambda_{\varepsilon,n} \in (\lambda_*, \lambda^*)$ such that $0 \in \operatorname{sp}(\boldsymbol{\mathcal{M}}^{\lambda_{\varepsilon,n}}_{\varepsilon,n})$. By compactness of the interval $[\lambda_*, \lambda^*]$ we may pass to a subsequence where $\lambda_{\varepsilon,n_k} \to \lambda_\varepsilon$ as $k \to \infty$, for some $\lambda_\varepsilon \in [\lambda_*, \lambda^*]$. By Theorem 6.2.1 we have $\widetilde{\mu}^\eta_{\lambda_{\varepsilon,n_k},\varepsilon,n_k} \rightharpoonup \mu^\eta_{\lambda_\varepsilon,\varepsilon}$ as $k \to \infty$, where $\widetilde{\mu}^\eta_{\lambda_{\varepsilon,n_k},\varepsilon,n_k}$ is the measure generated by the spectra of the approximations $\boldsymbol{\mathcal{M}}^{\lambda_{\varepsilon,n_k}}_{\varepsilon,n_k}$ and $\mu^\eta_{\lambda_\varepsilon,\varepsilon}$ is the measure generated by the spectra of $\boldsymbol{\mathcal{M}}^{\lambda_\varepsilon}_\varepsilon$. In order to avoid the continuous spectrum tending to $+\infty$ and the discrete spectrum tending to $-\infty$, the cutoff function ϕ_η must be chosen so that its support lies within $[-\frac{K}{2}, \frac{\lambda_*^2}{2}]$ where $K > 0$ is the spectral gap of the Neumann Laplacian ∂_x^2 on $L^2_0(\Omega)$. Since 0 lies in the support of all $\widetilde{\mu}^\eta_{\lambda_{\varepsilon,n_k},\varepsilon,n_k}$ it

must also lie in the support of $\mu^{\eta}_{\lambda_\varepsilon,\varepsilon}$. Furthermore, since $\phi_\eta(0) = 1$ we have that $0 \in \mathrm{sp}(\boldsymbol{\mathcal{M}}^{\lambda_\varepsilon}_{\varepsilon})$. We now repeat this argument to send $\varepsilon \downarrow 0$, obtaining $\lambda \in [\lambda_*, \lambda^*]$ with $0 \in \mathrm{sp}(\boldsymbol{\mathcal{M}}^{\lambda})$. Finally, the discreteness of the spectrum of $\boldsymbol{\mathcal{M}}^{\lambda}$ in $(-\infty, \lambda^2)$, (Lemma 6.6.4), ensures that 0 is an eigenvalue of $\boldsymbol{\mathcal{M}}^{\lambda}$, i.e. $\boldsymbol{\mathcal{M}}^{\lambda}$ has a non-trivial kernel.

6.5.1.2 Existence of a growing mode

Now that we know that there exist some $\lambda \in (0, \infty)$ and some $\boldsymbol{u} = \begin{bmatrix}\psi & \phi\end{bmatrix}^T \in H^2(\mathbb{R}) \times H^2_{0,n}(\Omega)$ that solve (6.3.5) we show that a genuine growing mode as defined in (6.1.8) really exists. To this end, we use ϕ, ψ and λ to define

$$E_1 = -\partial_x \phi \qquad E_2 = -\lambda \psi \qquad B = \partial_x \psi$$

(which lie in $H^1(\Omega)$, $H^2(\mathbb{R})$ and $H^1(\mathbb{R})$, respectively) and to define $f^{\pm}(x, v)$ as in (6.3.2):

$$f^{\pm} = \pm \mu^{\pm}_e \phi \pm \mu^{\pm}_p \psi \pm \lambda(\lambda + \mathcal{D}_{\pm})^{-1}[\mu^{\pm}_e(-\phi + \hat{v}_2 \psi)].$$

Observe that $f^{\pm}$ are both in $L^2(\mathbb{R} \times \mathbb{R}^2)$ since $\mu^{\pm}_e$ and $\mu^{\pm}_p$ are continuous functions that are compactly supported in the spatial variable which satisfy the integrability condition (6.1.6). In fact, $f^{\pm}$ are in the domains of $\mathcal{D}_{\pm}$, respectively, since $e^{\pm}$ and $p^{\pm}$ are constant along trajectories and ϕ and ψ are twice differentiable.

Lemma 6.5.1. *The functions $f^{\pm}$ solve the linearised Vlasov equations (6.3.1).*

Proof. This is almost a tautology: applying the operators $\lambda + \mathcal{D}_{\pm}$ to the expressions for $f^{\pm}$, respectively, one is left precisely with the expressions (6.3.1). □

Lemma 6.5.2. *The functions $f^{\pm}$ belong to $L^1(\mathbb{R} \times \mathbb{R}^2)$.*

Proof. Dropping the $\pm$ for brevity, the first term making up f is estimated as follows

$$\|\mu_e \phi\|_{L^1(\mathbb{R}^3)} \lesssim \|\mu_e\|_{L^2(\mathbb{R}^3)} \|\phi\|_{L^2(\mathbb{R})} \lesssim \|\mu_e\|^{1/2}_{L^\infty(\mathbb{R}^3)} \|\mu_e\|^{1/2}_{L^1(\mathbb{R}^3)} \|\phi\|_{L^2(\mathbb{R})} < \infty.$$

The other terms are estimated similarly (for the terms involving the averaging operator this may be seen by writing the ergodic average explicitly (see Re-

mark 6.3.1) or by using boundedness of the averaging operator on $\mathfrak{L}_\pm$). This implies that $f^\pm \in L^1(\mathbb{R}\times\mathbb{R}^2)$. □

We now define the charge and current densities ρ and j_i by

$$\rho = \int (f^+ - f^-)\, dv \qquad j_i = \int \hat{v}_i (f^+ - f^-)\, dv, \quad i = 1, 2.$$

Integrating $f^\pm$ in the momentum variable v alone, we obtain that $\rho \in L^1(\mathbb{R})$ as well as $j_i \in L^1(\mathbb{R})$ since $|\hat{v}_i| \leq 1$. In particular ρ, j_i are distributions on $\mathbb{R}$.

Lemma 6.5.3. *The continuity equation $\lambda\rho + \partial_x j_1 = 0$ holds in the sense of distributions.*

Proof. This follows from integrating the linearised Vlasov equations in the momentum variable. Indeed, we informally have

$$\begin{aligned}\int (\lambda + \mathcal{D}_\pm) f^\pm\, dv &= \pm \int \left[(\lambda + \mathcal{D}_\pm)(\mu_e^\pm \phi + \mu_p^\pm \psi) + \lambda \mu_e^\pm (-\phi + \hat{v}_2 \psi)\right] dv \\ &= \pm \int \lambda\psi \left(\mu_p^\pm + \mu_e^\pm \hat{v}_2\right) dv \pm \int \mathcal{D}_\pm \left(\mu_e^\pm \phi + \mu_p^\pm \psi\right) dv = 0,\end{aligned}$$

where the first term on the right hand side vanishes due to the identity (6.1.24) and the second term vanishes since $\mu^\pm$ are even in $\hat{v}_1$, whereas $\mathcal{D}_\pm = \hat{v}_1 \partial_x$ when applied to functions of x alone (recall that $\mu^\pm$ are constant along trajectories of $\mathcal{D}_\pm$, as are $\mu_e^\pm$ and $\mu_p^\pm$). We obtain the continuity equation by subtracting the "$-$" expression above from the "$+$" expression. Owing to the low regularity of $f^\pm$, ρ and j_1 this is true in a weak sense. □

Lemma 6.5.4. *Maxwell's equations (6.1.21) hold.*

Proof. Equations (6.1.21b) and (6.1.21c) hold due to (6.3.5) and the definitions of the operators (6.3.8). Indeed, from the second line of (6.3.5), we have

$$\begin{aligned}0 &= \left(\int \sum_\pm \mu_p^\pm \, d\boldsymbol{v}\right)\psi + \int \sum_\pm \mu_e^\pm \mathcal{Q}_\pm^\lambda \left[\hat{v}_2 \psi\right] d\boldsymbol{v} + \partial_x^2 \phi - \int \sum_\pm \mu_e^\pm (\mathcal{Q}_\pm^\lambda - 1)\phi \, d\boldsymbol{v} \\ &= \partial_x^2 \phi + \int \sum_\pm \left(\mu_p^\pm \psi + \mu_e^\pm \mathcal{Q}_\pm^\lambda \left[\hat{v}_2 \psi\right] - \mu_e^\pm (\mathcal{Q}_\pm^\lambda - 1)\phi\right) d\boldsymbol{v} \\ &\overset{(6.3.2)}{=} \partial_x^2 \phi + \int (f^+ - f^-)\, d\boldsymbol{v}.\end{aligned}$$

which is (6.1.21c). Similarly, (6.1.21b) is obtained from the first line of (6.3.5).

We therefore just need to show that (6.1.21a) holds. However this is a simple consequence of (6.1.21c) and the continuity equation. Indeed, we may first write

$$-\lambda\partial_x E_1 = \lambda\partial_x^2\phi \overset{(6.1.21c)}{=} -\lambda\rho \overset{\text{cont. eq.}}{=} \partial_x j_1$$

which is the derivative of (6.1.21a). Next, as $\phi \in H^2_{n,0}(\Omega)$, its derivative E_1 vanishes on $\partial\Omega$, and j_1 also vanishes there due to the compact support of the equilibrium in Ω. Thus, $-\lambda E_1$ and j_1 have the same derivative inside Ω and the same values on $\partial\Omega$, which means they must be equal. □

Lemma 6.5.5. *The charge and current densities ρ, j_1 and j_2 are elements in $L^1(\mathbb{R}) \cap L^2(\mathbb{R})$.*

Proof. This follows from Maxwell's equations and the regularity of ψ, ϕ which are in $H^2(\mathbb{R})$ and $H^2_{0,n}(\Omega)$ respectively. □

This concludes the proof of Theorem 6.1.1.

6.5.2 The cylindrically symmetric case

6.5.2.1 Existence of a non-trivial kernel of the equivalent problem

The proof of the existence of a non-trivial kernel in the cylindrically symmetric case is in complete analogy to the one in the $1.5d$ case presented in Section 6.5.1.1 and is therefore omitted.

6.5.2.2 Existence of a growing mode

Let $\lambda > 0$ and $\boldsymbol{u} = \begin{bmatrix}\mathbf{A}_\theta & \varphi & \mathbf{A}_{rz}\end{bmatrix}^T \in H^2_\theta(\mathbb{R}^3;\mathbb{R}^3) \times H^2_{cyl}(\mathbb{R}^3) \times H^2_{rz}(\mathbb{R}^3;\mathbb{R}^3)$ be such that (6.3.14) is satisfied, i.e. $\widetilde{\boldsymbol{\mathcal{M}}}^\lambda \boldsymbol{u} = \mathbf{0}$. Let $H^2_{cyl}(\mathbb{R}^3;\mathbb{R}^3) \ni \mathbf{A} = \mathbf{A}_\theta + \mathbf{A}_{rz}$ as in (6.1.32) and define

$$\mathbf{E} = -\nabla\varphi \qquad\qquad \mathbf{B} = \nabla \times \mathbf{A}$$

(which each lie in $H^1_{cyl}(\mathbb{R}^3;\mathbb{R}^3) \subseteq H^1(\mathbb{R}^3;\mathbb{R}^3)$). Furthermore, define

$$f^{\pm} = \pm\mu_e^{\pm}\varphi \pm r\mu_p^{\pm}(\mathbf{A}\cdot\boldsymbol{e}_\theta) \pm \mu_e^{\pm}\lambda(\lambda+\widetilde{\mathcal{D}}_{\pm})^{-1}(-\varphi+\mathbf{A}\cdot\hat{\boldsymbol{v}}).$$

As in the 1.5d case we begin by establishing that $f^{\pm}$ are integrable and satisfy the linearised Vlasov and continuity equations. The proof of this result is analogous to the corresponding results in the 1.5d case, so is omitted.

Lemma 6.5.6. *The functions $f^{\pm}$ solve the linearised Vlasov equations (6.1.29a) in the sense of distributions, and belong to $L^1(\mathbb{R}^3\times\mathbb{R}^3)$. Furthermore, the charge and current densities ρ and $\mathbf{j}$ defined by*

$$\rho = \int (f^+ - f^-)\,d\boldsymbol{v} \qquad \mathbf{j} = \int \hat{\boldsymbol{v}}(f^+ - f^-)\,d\boldsymbol{v},$$

belong to $L^1(\mathbb{R}^3)$ and $L^1(\mathbb{R}^3;\mathbb{R}^3)$, respectively, and satisfy the continuity equation $\lambda\rho + \nabla\cdot\mathbf{j} = 0$ in the sense of distributions.

Next we recover Maxwell's equations from (6.3.14) and the continuity equation.

Lemma 6.5.7. *Both the Lorenz gauge condition $\lambda\varphi + \nabla\cdot\mathbf{A} = 0$ (see (6.1.31)) and Maxwell's equations (6.1.25) are satisfied.*

Proof. In the same way as the 1.5d case, (6.1.25a) is obtained from the second line of (6.3.14). Similarly, (6.1.25b) is obtained from the first and third lines of (6.3.14).

It remains to show that the Lorenz gauge condition holds. Using (6.1.25a) and (6.1.25b) in the continuity equation, we have, in the sense of distributions

$$\begin{aligned} 0 &= \lambda(-\Delta+\lambda^2)\varphi + \nabla\cdot[(-\boldsymbol{\Delta}+\lambda^2)\mathbf{A}] \\ &= (-\Delta+\lambda^2)[\lambda\varphi + \nabla\cdot\mathbf{A}]. \end{aligned}$$

As $-\Delta+\lambda^2$ is invertible, this implies that $\lambda\varphi+\nabla\cdot\mathbf{A} = 0$. □

This concludes the proof of Theorem 6.1.2.

6.6 Properties of the operators

Here we gather all important properties of the operators defined in Section 6.3, as well as the operators defined in (6.1.23) and (6.1.33).

6.6.1 The $1.5d$ case

As the only dependence on λ is through the operators $\mathcal{Q}_\pm^\lambda$ we start with them:

Lemma 6.6.1. *In the respective spaces $\mathfrak{L}_\pm$, $\mathcal{Q}_\pm^\lambda$ satisfy:*

(a) $\left\|\mathcal{Q}_\pm^\lambda\right\|_{\mathfrak{B}(\mathfrak{L}_\pm)} = 1$.

(b) $\mathcal{Q}_\pm^\lambda$ *can be extended from* $\lambda > 0$ *to* $\operatorname{Re}\lambda > 0$ *as holomorphic operator valued functions. In particular they are continuous for* $\lambda > 0$ *in operator norm topology.*

(c) As $\mathbb{R} \ni \lambda \to \infty$, $\mathcal{Q}_\pm^\lambda \xrightarrow{s} 1$, *and for* $u \in \mathfrak{D}(\mathcal{D}_\pm)$, $\left\|(\mathcal{Q}_\pm^\lambda - 1)u\right\|_{\mathfrak{L}_\pm} \leq \|\mathcal{D}_\pm u\|_{\mathfrak{L}_\pm}/\lambda$.

(d) As $\lambda \to 0$, $\mathcal{Q}_\pm^\lambda$ *converge strongly to the projection operators* $\mathcal{Q}_\pm^0$ *defined in Definition 6.1.4.*

(e) For any $\lambda \geq 0$, $\mathcal{Q}_\pm^\lambda$ *are symmetric.*

Proof. $\left\|\mathcal{Q}_\pm^\lambda\right\|_{\mathfrak{B}(\mathfrak{L}_\pm)} \leq 1$ follows from $\|(\mathcal{D}_\pm + \lambda)^{-1}\|_{\mathfrak{B}(\mathfrak{L}_\pm)} \leq \frac{1}{|\lambda|}$ as $i\mathcal{D}_\pm$ is self-adjoint and the nearest point of the spectrum of $\mathcal{D}_\pm$ is 0. That $\left\|\mathcal{Q}_\pm^\lambda\right\|_{\mathfrak{B}(\mathfrak{L}_\pm)} = 1$ is proved by observing that $\mathcal{Q}_\pm^\lambda 1 = 1$. Part (b) follows from the analyticity of resolvents as functions of λ. For (c) we compute using functional calculus for $u \in \mathfrak{D}(\mathcal{D}_\pm)$:

$$\begin{aligned}\left\|\mathcal{Q}_\pm^\lambda u - u\right\|_{\mathfrak{L}_\pm} &= \left\|\left(\frac{\lambda^2}{\lambda^2 - \mathcal{D}_\pm^2} - 1\right) u\right\|_{\mathfrak{L}_\pm} = \left\|\frac{\mathcal{D}_\pm^2}{\lambda^2 - \mathcal{D}_\pm^2} u\right\|_{\mathfrak{L}_\pm} \\ &\leq \left\|\frac{\mathcal{D}_\pm}{\lambda + \mathcal{D}_\pm}\right\|_{\mathfrak{B}(\mathfrak{L}_\pm)} \left\|\frac{1}{\lambda - \mathcal{D}_\pm}\right\|_{\mathfrak{B}(\mathfrak{L}_\pm)} \|\mathcal{D}_\pm u\|_{\mathfrak{L}_\pm} \\ &\leq 1 \cdot \frac{1}{\lambda} \cdot \|\mathcal{D}_\pm u\|_{\mathfrak{L}_\pm} \to 0 \quad \text{as } \lambda \to \infty\end{aligned}$$

and deduce the strong convergence $\mathcal{Q}_\pm^\lambda \xrightarrow{s} 1$ by the density of $\mathfrak{D}(\mathcal{D}_\pm)$ in $\mathfrak{L}_\pm$.

For (d) we introduce the spectral measure (resolution of the identity) of the self-adjoint operator $-i\mathcal{D}_\pm$, which we denote by $M_\pm(\alpha)$, where $\alpha \in \mathbb{R}$. The projection onto $\ker(\mathcal{D}_\pm)$ is then $\mathcal{Q}^0_\pm = M_\pm(\{0\}) = \int_{\mathbb{R}} \chi(\alpha)\, dM_\pm(\alpha)$ where $\chi(0) = 1$ and $\chi(\alpha) = 0$ when $\alpha \neq 0$. Recall that $\lambda(\lambda + \mathcal{D}_\pm)^{-1} = \int_{\mathbb{R}} \frac{\lambda}{\lambda + i\alpha}\, dM_\pm(\alpha)$. We compute for $u \in \mathfrak{L}_\pm$,

$$\begin{aligned}\left\|\lambda(\lambda+\mathcal{D}_\pm)^{-1}u - M_\pm(\{0\})u\right\|^2_{\mathfrak{L}_\pm} &= \left\|\int_{\mathbb{R}} \left(\frac{\lambda}{\lambda+i\alpha} - \chi(\alpha)\right) dM_\pm(\alpha)u\right\|^2_{\mathfrak{L}_\pm} \\ &= \int_{\mathbb{R}} \left|\frac{\lambda}{\lambda+i\alpha} - \chi(\alpha)\right|^2 d\left\|M_\pm(\alpha)u\right\|^2_{\mathfrak{L}_\pm},\end{aligned}$$

the last equality being due to orthogonality of spectral projections. This now tends to 0 as $\lambda \to 0$ by the dominated convergence theorem. Replacing $\mathcal{D}_\pm$ with $-\mathcal{D}_\pm$, which has the same kernel, we deduce that $\lambda(\lambda - \mathcal{D}_\pm)^{-1} \xrightarrow{s} \mathcal{Q}^0_\pm$. Finally we have $\mathcal{Q}^\lambda_\pm = \lambda(\lambda - \mathcal{D}_\pm)^{-1}\lambda(\lambda + \mathcal{D}_\pm)^{-1} \xrightarrow{s} (\mathcal{Q}^0_\pm)^2 = \mathcal{Q}^0_\pm$ by the composition of strong operator convergence. To show (e) for $\lambda > 0$ we simply note that $\mathcal{D}^2_\pm$ are self-adjoint, and extend to $\lambda = 0$ by the strong operator convergence. □

These results carry through to the other operators.

Lemma 6.6.2. *The operators $\boldsymbol{\mathcal{J}}^\lambda$ and $\mathcal{B}^\lambda$ have the properties:*

(a) For all $\lambda \in [0, \infty)$, $\mathcal{B}^\lambda$ maps $L^2(\mathbb{R})$ into $L^2_0(\Omega)$ and $\boldsymbol{\mathcal{J}}^\lambda$ maps $L^2(\mathbb{R}) \times L^2_0(\Omega) \to L^2(\mathbb{R}) \times L^2_0(\Omega)$.

(b) The familes $\{\boldsymbol{\mathcal{J}}^\lambda\}_{\lambda\in[0,\infty)}$ and $\{\mathcal{B}^\lambda\}_{\lambda\in[0,\infty)}$ are both uniformly bounded in operator norm.

(c) Both $(0, \infty) \ni \lambda \mapsto \boldsymbol{\mathcal{J}}^\lambda$ and $(0, \infty) \ni \lambda \mapsto \mathcal{B}^\lambda$ are continuous in the operator norm topology.

(d) As $\lambda \to 0$, $\boldsymbol{\mathcal{J}}^\lambda \to \boldsymbol{\mathcal{J}}^0$ and $\mathcal{B}^\lambda \to \mathcal{B}^0$ in the strong operator topology.

(e) For any $\lambda \geq 0$ the operator $\boldsymbol{\mathcal{J}}^\lambda$ is symmetric.

(f) Let $\boldsymbol{\mathcal{P}}$ be the multiplication operator acting in $L^2(\mathbb{R}) \times L^2_0(\Omega)$ defined by

$$\boldsymbol{\mathcal{P}} = \begin{bmatrix} \mathbb{1}_\Omega & 0 \\ 0 & \mathbb{1}_\Omega \end{bmatrix}$$

where $\mathbb{1}_\Omega$ is the indicator function of the set Ω. Then $\boldsymbol{\mathcal{J}}^\lambda = \boldsymbol{\mathcal{J}}^\lambda \boldsymbol{\mathcal{P}}$.

Proof. Part (a) is easily verifiable. We note that due to the relation

$$\mathcal{B}^\lambda = -\begin{bmatrix} 0 & 1 \end{bmatrix} \boldsymbol{\mathcal{J}}^\lambda \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

it is sufficient to prove the results for $\boldsymbol{\mathcal{J}}^\lambda$. We observe that due to the decay assumptions (6.1.6) on $\mu^\pm$, the moment

$$-\sum_\pm \int \mu^\pm \frac{1+v_1^2}{\langle v \rangle^3}\, d\boldsymbol{v}$$

is bounded in $L^\infty(\mathbb{R})$ and is real valued, so it is a bounded symmetric multiplication operator from $L^2(\mathbb{R})$ to $L^2(\mathbb{R})$. Next we decompose the second part of $\boldsymbol{\mathcal{J}}^\lambda$ as

$$\sum_\pm \int \mu_e^\pm \mathcal{T}_\pm (\mathcal{Q}_\pm^\lambda - 1) \mathcal{T}_\pm^* \begin{bmatrix} \psi \\ \phi \end{bmatrix} d\boldsymbol{v} \tag{6.6.1}$$

where $\mathcal{T}_\pm : \mathfrak{L}_\pm \times \mathfrak{L}_\pm \to \mathfrak{L}_\pm$ is multiplication by the vector $\begin{bmatrix} \hat{v}_2 & -1 \end{bmatrix}$, and we have used the natural (and bounded) inclusions from $L^2(\mathbb{R})$ and $L_0^2(\Omega)$ into $\mathfrak{L}_\pm$. Clearly $\mathcal{T}_\pm$ are bounded and we know that $\mathcal{Q}_\pm^\lambda$ have bound 1 by Lemma 6.6.1. Finally, we note that due to the decay assumptions on $\mu_e^\pm$ and its compact support in x, that multiplication by $\mu_e^\pm$ followed by integration $d\boldsymbol{v}$ is bounded from $\mathfrak{L}_\pm$ to $L^2(\mathbb{R})$ and $L^2(\Omega)$. Therefore $\boldsymbol{\mathcal{J}}^\lambda$ has a uniform bound in operator norm. Parts (c) and (d) then follow from the corresponding results for $\mathcal{Q}_\pm^\lambda$ in Lemma 6.6.1 using (6.6.1). (e) is clear from the symmetry of $\mathcal{Q}_\pm^\lambda$ and (6.6.1). Finally (f) follows from the compact spatial support of $\mu^\pm, \mu_e^\pm, \mu_p^\pm$ inside Ω. □

Lemma 6.6.3 (Properties of $\mathcal{A}_1^\lambda$ and $\mathcal{A}_2^\lambda$)**.** *Let $0 \le \lambda < \infty$.*

(a) The operator $\mathcal{A}_1^\lambda$ is self-adjoint on $L_0^2(\Omega)$ and the operator $\mathcal{A}_2^\lambda$ is self-adjoint on $L^2(\mathbb{R})$ with the respective domains $H_{0,n}^2(\Omega)$ and $H^2(\mathbb{R})$.

(b) Both $[0,\infty) \ni \lambda \mapsto \mathcal{A}_1^\lambda$ and $[0,\infty) \ni \lambda \mapsto \mathcal{A}_2^\lambda$ are continuous in the norm resolvent topology.

(c) The spectrum of $\mathcal{A}_1^\lambda$ is purely discrete. The spectrum of $\mathcal{A}_2^\lambda$ in $(-\infty, \lambda^2)$ is discrete and made up of finitely many eigenvalues. It is continuous (possibly

with embedded eigenvalues) in $[\lambda^2, \infty)$.

(d) There exist constants $\gamma > 0$ *and* $\Lambda > 0$ *such that for all* $\lambda \geq \Lambda$, $\mathcal{A}_i^\lambda > \gamma$, $i = 1, 2$.

Proof. Clearly $-\partial_x^2$ is symmetric. The perturbative terms are symmetric as well since $\mathcal{Q}_\pm^\lambda$ are symmetric, see Lemma 6.6.1. Self-adjointness is guaranteed by standard arguments, such as the Kato-Rellich theorem.

Let us prove (b), considering first $\mathcal{A}_2^\lambda$. It is sufficient to prove that $(\mathcal{A}_2^\lambda - i)^{-1} \to (\mathcal{A}_2^\sigma - i)^{-1}$ in operator norm as $\lambda \to \sigma$ (with $\lambda, \sigma \geq 0$). We use the second resolvent identity to obtain

$$\begin{aligned}(\mathcal{A}_2^\lambda - i)^{-1} - (\mathcal{A}_2^\sigma - i)^{-1} &= (\mathcal{A}_2^\lambda - i)^{-1}(\mathcal{A}_2^\sigma - \mathcal{A}_2^\lambda)(\mathcal{A}_2^\sigma - i)^{-1} \\ &= (\mathcal{A}_2^\lambda - i)^{-1}((\sigma^2 - \lambda^2) - (\mathcal{J}_{11}^\sigma - \mathcal{J}_{11}^\lambda))(\mathcal{A}_2^\sigma - i)^{-1}\end{aligned}$$

where $\mathcal{J}_{11}^\lambda$ is the upper left component of $\boldsymbol{\mathcal{J}}^\lambda$ written in block matrix form. Hence, as the resolvents are each bounded in operator norm by 1,

$$\left\|(\mathcal{A}_2^\lambda - i)^{-1} - (\mathcal{A}_2^\sigma - i)^{-1}\right\|_{\mathfrak{B}(L^2(\mathbb{R}))} \leq |\sigma^2 - \lambda^2| + \left\|(\mathcal{J}_{11}^\sigma - \mathcal{J}_{11}^\lambda)(\mathcal{A}_2^\sigma - i)^{-1}\right\|_{\mathfrak{B}(L^2(\mathbb{R}))}.$$

It thus suffices using Lemma 6.6.2(f) to show that $(\mathcal{J}_{11}^\sigma - \mathcal{J}_{11}^\lambda)\mathcal{P}(\mathcal{A}_2^\sigma - i)^{-1} \to 0$ in operator norm, where $\mathcal{P}$ is the multiplication operator on $L^2(\mathbb{R})$ given by the indicator function of the set Ω. $\mathcal{P}$ is relatively compact with respect to $-\partial_x^2$ by the Rellich theorem, and hence also relatively compact with respect to $\mathcal{A}_2^\sigma$ as it also has the domain $H^2(\mathbb{R})$ by part (a). Hence $\mathcal{P}(\mathcal{A}_2^\sigma - i)^{-1}$ is compact, which allows us to upgrade the strong convergence $\mathcal{J}_{11}^\lambda \xrightarrow{s} \mathcal{J}_{11}^\sigma$ given by Lemma 6.6.2 to operator norm convergence. The norm resolvent continuity of $\mathcal{A}_2^\lambda$ follows. The proof for $\mathcal{A}_1^\lambda$ is analogous, but lacking the $|\sigma^2 - \lambda^2|$ term.

Part (c) is simple: both operators have a differential part (Laplacian) and a relatively compact perturbation. Hence both conclusions follow from Weyl's theorem [109, IV, Theorem 5.35]. The finiteness of the discrete spectrum below the essential part in the case of $\mathcal{A}_2^\lambda$ is a consequence of Lemma 6.2.2. For part (d), we

begin with $\mathcal{A}_1^\lambda$. Fix an arbitrary $h \in H^2_{0,n}(\Omega)$, then we compute,

$$\left\langle \mathcal{A}_1^\lambda h, h \right\rangle_{L^2_0(\Omega)} = \|\partial_x h\|^2_{L^2_0(\Omega)} - \sum_{\pm} \iint \overline{h} \mu_e^\pm (1 - \mathcal{Q}_\pm^\lambda) h \, d\boldsymbol{v} dx$$

Now we note as $h \in H^2_{0,n}(\Omega)$, h is in $\mathfrak{D}(\mathcal{D}_\pm)$ when interpreted in $\mathfrak{L}_\pm$. We now use Lemma 6.6.1(c) to estimate,

$$\begin{aligned}
\left\langle \mathcal{A}_1^\lambda h, h \right\rangle_{L^2_0(\Omega)} &\geq \|\partial_x h\|^2_{L^2_0(\Omega)} - \frac{1}{\lambda} \sum_{\pm} \left\| \mu_e^\pm / w^\pm \right\|_{L^\infty(\Omega \times \mathbb{R}^3)} \|\mathcal{D}_\pm h\|_{\mathfrak{L}_\pm} \|h\|_{\mathfrak{L}_\pm} \\
&\geq \|\partial_x h\|^2_{L^2_0(\Omega)} - \frac{C}{\sqrt{K}\lambda} \|\partial_x h\|^2_{L^2_0(\Omega)} \\
&\geq K \|h\|^2_{L^2_0(\Omega)} \left(1 - \frac{C}{\sqrt{K}\lambda} \right)
\end{aligned}$$

where K is the spectral gap of the Laplacian on the bounded domain Ω, and we have used

$$\|\mathcal{D}_\pm h\|^2_{\mathfrak{L}_\pm} = \iint w^\pm |\hat{v}_1 \partial_x h|^2 \, d\boldsymbol{v} dx \leq \|\partial_x h\|^2_{L^2_0(\Omega)} \sup_{x \in \Omega} \int w^\pm |\hat{v}_1|^2 \, d\boldsymbol{v}$$

and the natural bounded inclusions from $L^2_0(\Omega)$ into $\mathfrak{L}_\pm$. We now just take $\Lambda > C/\sqrt{K}$.

For $\mathcal{A}_2^\lambda$ the proof is easier due to the λ^2 term. For $h \in H^2(\mathbb{R})$ we compute, using the formulation (6.3.8a),

$$\begin{aligned}
\left\langle \mathcal{A}_2^\lambda h, h \right\rangle_{L^2(\mathbb{R})} &= \|\partial_x h\|^2_{L^2(\mathbb{R})} + \lambda^2 \|h\|^2_{L^2(\mathbb{R})} - \left\langle \boldsymbol{\mathcal{J}}^\lambda \begin{bmatrix} h \\ 0 \end{bmatrix}, \begin{bmatrix} h \\ 0 \end{bmatrix} \right\rangle_{L^2(\mathbb{R}) \times L^2_0(\Omega)} \\
&\geq (\lambda^2 - C') \|h\|^2_{L^2(\mathbb{R})}
\end{aligned}$$

where we have used that the uniform bound in operator norm of $\boldsymbol{\mathcal{J}}^\lambda$ given by Lemma 6.6.2. We now take $\Lambda > \sqrt{C'}$. □

Lemma 6.6.4 (Properties of $\boldsymbol{\mathcal{M}}^\lambda$)**.** *For each $\lambda \in [0, \infty)$, the operator $\boldsymbol{\mathcal{M}}^\lambda$ is self-adjoint on $L^2(\mathbb{R}) \times L^2_0(\Omega)$ with domain $H^2(\mathbb{R}) \times H^2_{0,n}(\Omega)$. For any $\lambda \geq 0$, the operator $\boldsymbol{\mathcal{M}}^\lambda$ has essential spectrum $[\lambda^2, \infty)$. The family $\{\boldsymbol{\mathcal{M}}^\lambda\}_{\lambda \in [0,\infty)}$ is continuous in the norm resolvent topology.*

Proof. The proof essentially mimics (and uses) the preceding proofs, and is therefore left for the reader. □

6.6.2 The cylindrically symmetric case

As many of the proofs are the same as in the 1.5 dimensional case above, we give the details only where they differ.

Lemma 6.6.5. *In the respective spaces* $\mathfrak{N}_\pm$, $\widetilde{\mathcal{Q}}^\lambda_{\pm,sym}$ *and* $\widetilde{\mathcal{Q}}^\lambda_{\pm,skew}$ *satisfy:*

(a) $\left\|\widetilde{\mathcal{Q}}^\lambda_{\pm,sym}\right\|_{\mathfrak{B}(\mathfrak{N}_\pm)} = 1$ *and* $\left\|\widetilde{\mathcal{Q}}^\lambda_{\pm,skew}\right\|_{\mathfrak{B}(\mathfrak{N}_\pm)} \leq \frac{1}{2}$.

(b) $\widetilde{\mathcal{Q}}^\lambda_{\pm,sym}$ *and* $\widetilde{\mathcal{Q}}^\lambda_{\pm,skew}$ *can be extended from* $\lambda > 0$ *to* $\operatorname{Re}\lambda > 0$ *as holomorphic operator valued functions. In particular they are continuous for* $\lambda > 0$ *in operator norm topology.*

(c) As $\mathbb{R} \ni \lambda \to \infty$, $\widetilde{\mathcal{Q}}^\lambda_{\pm,sym} \xrightarrow{s} 1$, *and for* $u \in \mathfrak{D}(\widetilde{\mathcal{D}}_\pm)$, *we have the bound* $\left\|(\widetilde{\mathcal{Q}}^\lambda_{\pm,sym} - 1)u\right\|_{\mathfrak{N}_\pm} \leq \left\|\widetilde{\mathcal{D}}_\pm u\right\|_{\mathfrak{N}_\pm}/\lambda$.

(d) As $\mathbb{R} \ni \lambda \to \infty$, $\widetilde{\mathcal{Q}}^\lambda_{\pm,skew} \xrightarrow{s} 0$, *and for* $u \in \mathfrak{D}(\widetilde{\mathcal{D}}_\pm)$, *we have the bound* $\left\|\widetilde{\mathcal{Q}}^\lambda_{\pm,sym} u\right\|_{\mathfrak{N}_\pm} \leq \left\|\widetilde{\mathcal{D}}_\pm u\right\|_{\mathfrak{N}_\pm}/\lambda$.

(e) As $0 < \lambda \to 0$, $\widetilde{\mathcal{Q}}^\lambda_{\pm,sym} \xrightarrow{s} \widetilde{\mathcal{Q}}^0_\pm$, *where* $\widetilde{\mathcal{Q}}^0_\pm$ *are defined in Definition 6.1.5.*

(f) As $0 < \lambda \to 0$, $\widetilde{\mathcal{Q}}^\lambda_{\pm,skew} \xrightarrow{s} 0$.

(g) For any $\lambda \geq 0$, $\widetilde{\mathcal{Q}}^\lambda_{\pm,sym}$ *are symmetric and* $\widetilde{\mathcal{Q}}^\lambda_{\pm,skew}$ *are skew-symmetric.*

Proof. The claims about $\widetilde{\mathcal{Q}}^\lambda_{\pm,sym}$ may be proved in the same way as those in Lemma 6.6.1. For (a), the spectral theorem applied to the self-adjoint operators $-i\widetilde{\mathcal{D}}_\pm$ implies that $\widetilde{\mathcal{Q}}^\lambda_{\pm,skew}$ are unitarily equivalent to a multiplication operator, so that

$$\left\|\widetilde{\mathcal{Q}}^\lambda_{\pm,skew}\right\|_{\mathfrak{B}(\mathfrak{N}_\pm)} = \left\|\frac{-i\alpha\lambda}{\lambda^2+\alpha^2}\right\|_{L^\infty_\alpha(\mathrm{sp}(i\widetilde{\mathcal{D}}_\pm))} \leq \left\|\frac{-i\alpha\lambda}{\lambda^2+\alpha^2}\right\|_{L^\infty_\alpha(\mathbb{R})} = \frac{1}{2}.$$

The proof of (b) follows, as in the proof of Lemma 6.6.1, from the holomorphicity

of the resolvent. For (d), we let $u \in \mathfrak{D}(\widetilde{\mathcal{D}}_\pm)$, and then for $\lambda > 0$ we have,

$$\begin{aligned}\left\|\widetilde{\mathcal{Q}}^\lambda_{\pm,skew}u\right\|_{\mathfrak{N}_\pm} &\leq \lambda \left\|(\lambda^2+\widetilde{\mathcal{D}}^2_\pm)^{-1}\right\|_{\mathfrak{B}(\mathfrak{N}_\pm)} \left\|\widetilde{\mathcal{D}}_\pm u\right\|_{\mathfrak{N}_\pm} \\ &\leq \frac{1}{\lambda}\left\|\widetilde{\mathcal{D}}_\pm u\right\|_{\mathfrak{N}_\pm} \to 0 \text{ as } \lambda \to \infty.\end{aligned}$$

The strong convergence to 0 then follows from the density of $\mathfrak{D}(\widetilde{\mathcal{D}}_\pm)$. For (f), we repeat the proof of Lemma 6.6.1, noting that it is shown that $\lambda(\lambda+\widetilde{\mathcal{D}}_\pm)^{-1} \xrightarrow{s} \widetilde{\mathcal{Q}}^0_\pm$ and $\widetilde{\mathcal{Q}}^\lambda_{\pm,sym} \xrightarrow{s} \widetilde{\mathcal{Q}}^0_\pm$ as $\lambda \to 0$. That $\widetilde{\mathcal{Q}}^\lambda_{\pm,skew} \xrightarrow{s} 0$ as $\lambda \to 0$ now follows from the identity, valid for all $\lambda > 0$,

$$\lambda(\lambda+\widetilde{\mathcal{D}}_\pm)^{-1} = \widetilde{\mathcal{Q}}^\lambda_{\pm,sym} + \widetilde{\mathcal{Q}}^\lambda_{\pm,skew}.$$

Finally, (g) is a consequence of Lemma 6.3.2. □

Lemma 6.6.6. *The operators $\widetilde{\boldsymbol{\mathcal{J}}}^\lambda$, and $\widetilde{\mathcal{B}}^\lambda_i$ for $i = 1,2,3,4$ have the properties:*

(a) For all $\lambda \in [0,\infty)$, we have

$$\begin{aligned}\widetilde{\mathcal{B}}^\lambda_1 &\in \mathfrak{B}\left(L^2_\theta(\mathbb{R}^3;\mathbb{R}^3), L^2_{cyl}(\mathbb{R}^3)\right), \\ \widetilde{\mathcal{B}}^\lambda_2 &\in \mathfrak{B}\left(L^2_\theta(\mathbb{R}^3;\mathbb{R}^3), L^2_{rz}(\mathbb{R}^3;\mathbb{R}^3)\right), \\ \widetilde{\mathcal{B}}^\lambda_3 &\in \mathfrak{B}\left(L^2_{cyl}(\mathbb{R}^3), L^2_{rz}(\mathbb{R}^3;\mathbb{R}^3)\right), \\ \widetilde{\boldsymbol{\mathcal{B}}}^\lambda_4 &\in \mathfrak{B}\left(L^2_\theta(\mathbb{R}^3;\mathbb{R}^3), L^2_{cyl}(\mathbb{R}^3)\times L^2_{rz}(\mathbb{R}^3;\mathbb{R}^3)\right),\end{aligned}$$

with bounds uniform in λ.

(b) Each of $(0,\infty) \ni \lambda \mapsto \widetilde{\boldsymbol{\mathcal{J}}}^\lambda$ and $(0,\infty) \ni \lambda \mapsto \widetilde{\mathcal{B}}^\lambda_i$, $i = 1,2,3,4$ are continuous in the operator norm topology.

(c) As $\lambda \to 0$, $\widetilde{\boldsymbol{\mathcal{J}}}^\lambda \to \widetilde{\boldsymbol{\mathcal{J}}}^0$, $\widetilde{\mathcal{B}}^\lambda_1 \to \widetilde{\mathcal{B}}^0_1$, $\widetilde{\mathcal{B}}^\lambda_2 \to 0$ and $\widetilde{\mathcal{B}}^\lambda_3 \to 0$ in the strong topology.

(d) For any $\lambda \geq 0$ the operator $\widetilde{\boldsymbol{\mathcal{J}}}^\lambda$ is symmetric.

(e) Let $\widetilde{\boldsymbol{\mathcal{P}}}$ be the multiplication operator acting in $L^2_{cyl}(\mathbb{R}^3) \times L^2_\theta(\mathbb{R}^3;\mathbb{R}^3) \times L^2_{rz}(\mathbb{R}^3;\mathbb{R}^3)$ defined by

$$\widetilde{\boldsymbol{\mathcal{P}}} = \begin{bmatrix} \mathbb{1}_\Omega & 0 & 0 \\ 0 & \mathbb{1}_\Omega & 0 \\ 0 & 0 & \mathbb{1}_\Omega \end{bmatrix}$$

where $\mathbb{1}_\Omega$ is the indicator function of the set Ω. Then $\widetilde{\boldsymbol{\mathcal{J}}}^\lambda = \widetilde{\boldsymbol{\mathcal{J}}}^\lambda \widetilde{\boldsymbol{\mathcal{P}}}$.

Proof. That the operators map the corresponding spaces to each other may be verified directly from (6.3.16), noting in particular the notation $\hat{\boldsymbol{v}}_\theta = \boldsymbol{e}_\theta \hat{v}_\theta$ and $\hat{\boldsymbol{v}}_{rz} = \boldsymbol{e}_r \hat{v}_r + \boldsymbol{e}_z \hat{v}_z$. As in the proof of Lemma 6.6.2, the uniform (in λ) bound on the operator norms may be obtained using the decay assumptions on the equilibrium and the uniform bound on the norms of $\widetilde{\mathcal{Q}}^\lambda_{\pm,sym}$ and $\widetilde{\mathcal{Q}}^\lambda_{\pm,skew}$ given by Lemma 6.6.5. In the same way (c) and (d) follow from the corresponding results in Lemma 6.6.5.

To show the symmetry of $\widetilde{\boldsymbol{\mathcal{J}}}^\lambda$ for $\lambda > 0$ we use the block matrix form, noting that $\widetilde{\mathcal{Q}}^\lambda_{\pm,sym}$ appears on the diagonal, and that the off diagonal entries have $\widetilde{\mathcal{B}}^\lambda_i$ and their adjoints in the appropriate configuration. Then we extend to $\lambda = 0$ by the strong convergence. As in Lemma 6.6.2 (e) follows from the spatial support properties of the equilibrium. □

Lemma 6.6.7 (Properties of $\widetilde{\mathcal{A}}^\lambda_1$, $\widetilde{\mathcal{A}}^\lambda_2$, $\widetilde{\mathcal{A}}^\lambda_3$ and $\widetilde{\boldsymbol{\mathcal{A}}}^\lambda_4$)**.** *Let $0 \leq \lambda < \infty$.*

(a) The operator $\widetilde{\mathcal{A}}^\lambda_1$ is self-adjoint on $L^2_{cyl}(\mathbb{R}^3)$, the operator $\widetilde{\mathcal{A}}^\lambda_2$ is self-adjoint on $L^2_\theta(\mathbb{R}^3;\mathbb{R}^3)$, $\widetilde{\mathcal{A}}^\lambda_3$ is self-adjoint on $L^2_{rz}(\mathbb{R}^3;\mathbb{R}^3)$ and $\widetilde{\boldsymbol{\mathcal{A}}}^\lambda_4$ is self-adjoint on $L^2_{cyl}(\mathbb{R}^3) \times L^2_{rz}(\mathbb{R}^3;\mathbb{R}^3)$ with the respective domains $H^2_{cyl}(\mathbb{R}^3)$, $H^2_\theta(\mathbb{R}^3;\mathbb{R}^3)$, $H^2_{rz}(\mathbb{R}^3;\mathbb{R}^3)$ and $H^2_{cyl}(\mathbb{R}^3) \times H^2_{rz}(\mathbb{R}^3;\mathbb{R}^3)$.

(b) The mappings $[0,\infty) \ni \lambda \mapsto \widetilde{\mathcal{A}}^\lambda_1$, $[0,\infty) \ni \lambda \mapsto \widetilde{\mathcal{A}}^\lambda_2$, $[0,\infty) \ni \lambda \mapsto \widetilde{\mathcal{A}}^\lambda_3$ and $[0,\infty) \ni \lambda \mapsto \widetilde{\boldsymbol{\mathcal{A}}}^\lambda_4$ are continuous in the norm resolvent topology.

(c) The spectra of $\widetilde{\mathcal{A}}^\lambda_1$, $\widetilde{\mathcal{A}}^\lambda_2$, $\widetilde{\mathcal{A}}^\lambda_1$ and $\widetilde{\boldsymbol{\mathcal{A}}}^\lambda_4$ in $(-\infty, \lambda^2)$ are discrete and finite. In $[\lambda^2, \infty)$ their spectra are continuous (possibly with embedded eigenvalues).

(d) There exist constants $\gamma > 0$ and $\Lambda > 0$ such that for all $\lambda \geq \Lambda$, $\widetilde{\mathcal{A}}^\lambda_i > \gamma$, $i = 1, 2, 3, 4$.

Proof. The proof for each of $\widetilde{\mathcal{A}}^\lambda_i$, $i = 1, 2, 3, 4$ is analogous to that of Lemma 6.6.3 for $\mathcal{A}^\lambda_2$. We omit the details. □

Lemma 6.6.8 (Properties of $\widetilde{\boldsymbol{\mathcal{M}}}^\lambda$)**.** *For each $\lambda \in [0,\infty)$, the operator $\widetilde{\boldsymbol{\mathcal{M}}}^\lambda$ is self-adjoint on $L^2_\theta(\mathbb{R}^3;\mathbb{R}^3) \times L^2_{cyl}(\mathbb{R}^3) \times L^2_{rz}(\mathbb{R}^3;\mathbb{R}^3)$ with domain $H^2_\theta(\mathbb{R}^3;\mathbb{R}^3) \times H^2_{cyl}(\mathbb{R}^3) \times H^2_{rz}(\mathbb{R}^3;\mathbb{R}^3)$. For any $\lambda \geq 0$, the operator $\widetilde{\boldsymbol{\mathcal{M}}}^\lambda$ has essential spectrum*

$(-\infty, -\lambda^2] \cup [\lambda^2, \infty)$. *The family* $\{\widetilde{\mathcal{M}}^\lambda\}_{\lambda\in[0,\infty)}$ *is continuous in the norm resolvent topology.*

Proof. This is again analogous to the previous proofs, and is left to the reader. □

6.7 Existence of equilibria

In this section we prove that there exist compactly supported equilibria of the $1.5d$ system which can be written in the form (6.1.16)-(6.1.17). Existence in the $3d$ case was already provided in [130]. We note that providing explicit *examples* of equilibria is a much more challenging task, which we do not pursue here. The construction below utilises the physically relevant idea of *magnetic confinement*. We mention in this context the recent result [159] where global-in-time existence and uniqueness of solutions was established in a similar setting, though with a singular magnetic potential.

Proposition 6.7.1 (Existence of confined equilibria)**.** *Let* $R > 0, \alpha > 2$ *and* $A^\pm \subset \mathbb{R}^2$ *be bounded domains. Then there are constants* $c, C > 0$ *such that if two functions* $\mu^\pm(e^\pm, p^\pm) \in C_0^1(\mathbb{R}^2)$ *with support in* $A^\pm$ *satisfy*

$$|\mu^\pm|, |\mu_e^\pm|, |\mu_p^\pm| \leq c(1 + |e^\pm|)^{-\alpha}$$

and a function $\psi^{ext} \in H^2_{loc}(\mathbb{R})$ *satisfies*

$$|\psi^{ext}(x)| \geq C(1 + |x|^2) \text{ for } |x| \geq R$$

then there are potentials $\phi^0, \psi^0 \in H^2_{loc}(\mathbb{R})$ *such that* $(\mu^\pm(e^\pm, p^\pm), \phi^0, \psi^0, \psi^{ext}, \phi^{ext} = 0)$ *is an equilibrium of the* $1.5d$ *relativistic Vlasov-Maxwell equations* (6.1.15) *with spatial support in* $[-R, R]$*, where the relationship between* (x, v_1, v_2) *and* $(e^\pm, p^\pm)$ *is as defined in* (6.1.17).

Remark 6.7.1 (Trivial solutions)**.** *Of course, the result does not say that the obtained equilibrium is not everywhere zero. This may be ruled out by choosing* $\mu^\pm$ *and* ψ^{ext} *in such a way that (for example)* $\mu^\pm(x = 0, v = 0) > 0$ *if* $\phi^0, \psi^0 \equiv 0$*. Let us sketch the argument. Recall that we write* $f^{0,\pm}(x, v) = \mu^\pm(\langle \boldsymbol{v} \rangle \pm \phi^0(x), v_2 \pm$

$\psi^0(x) \pm \psi^{ext}(x)) = \mu^\pm(e^\pm, p^\pm)$. *If* $f^{0,\pm}(x, \boldsymbol{v}) = 0$ *for all* $(x, \boldsymbol{v})$ *then* $\rho, j_i = 0$ *and* $\phi^0, \psi^0 = 0$ *for all* x. *Therefore* $e^\pm = \langle \boldsymbol{v} \rangle$ *and* $p^\pm = v_2 \pm \psi^{ext}(x)$, *and*

$$f^{0,\pm}(0,0) = \mu^\pm(1, \pm\psi^{ext}(0)).$$

The RHS is something we can ensure is positive by choosing $A^\pm$, $\mu^\pm$ *and* ψ^{ext} *appropriately. Under this appropriate choice one obtains a contradiction.*

Proof of Proposition 6.7.1. Given two elements $\rho, j_2 \in L^2(\mathbb{R})$ with compact support, we define

$$\phi^0 = G * \rho, \qquad \psi^0 = G * j_2,$$

where $G(x) = -|x|/2$ is the fundamental solution of the Laplacian in one dimension. [We note that one expects j_1 to vanish for an equilibrium, due to parity in v_1, hence it does not appear in the setup.] Thus we define $e^\pm = e^\pm(x, v_1, v_2)$ and $p^\pm = p^\pm(x, v_1, v_2)$ via the usual relations (6.1.17), which we recall for the reader's convenience:

$$e^\pm(x, \boldsymbol{v}) = \langle \boldsymbol{v} \rangle \pm \phi^0(x), \qquad p^\pm(x, \boldsymbol{v}) = v_2 \pm \psi^0(x) \pm \psi^{ext}(x)$$

(ϕ^{ext} is zero). We let $\mathcal{F} : L^2(\mathbb{R})^2 \to L^2(\mathbb{R})^2$ be the (non-linear) map defined by

$$\mathcal{F} \begin{bmatrix} \rho \\ j_2 \end{bmatrix} = \int \begin{bmatrix} 1 \\ \hat{v}_2 \end{bmatrix} (\mu^+(e^+, p^+) - \mu^-(e^-, p^-))\, d\boldsymbol{v}. \tag{6.7.1}$$

A fixed point of $\mathcal{F}$ is the charge and current densities of an equilibrium solution $(\mu^\pm(e^\pm, p^\pm), \phi^0, \psi^0, \psi^{ext}, \phi^{ext} = 0)$. We define $X \subseteq L^2(\mathbb{R})^2$ as

$$X = \{(\rho, j_2) \in L^2(\mathbb{R})^2 \ : \ \text{both supported in } [-R, R] \text{ and bounded by } C'\}$$

for a positive constant C' to be chosen. This set is clearly convex. We will show that for $c > 0$ sufficiently small and $C > 0$ sufficiently large, $\mathcal{F}$ is a compact continuous map $X \hookrightarrow X$ and thus, by the Schauder fixed point theorem, has a fixed point.

Step 1: Compact support. We show that C' and C can be chosen so that $\mathcal{F}$ maps X into functions supported in $[-R, R]$.

For $(\rho, j_2) \in X$ and $|x| > R$, we have,

$$|\phi^0(x)| = |(G * \rho)(x)| \leq C' \int_{-R}^{R} |G(x-y)|\, dy = \frac{C'}{2} \int_{-R}^{R} |x-y|\, dy = C'R|x|$$

and the same bound holds for ψ^0. This allows us to control v_2 using $e^{\pm}$ and x. Indeed,

$$|v_2| \leq \langle \boldsymbol{v} \rangle = e^{\pm} \mp \phi^0(x) \leq |e^{\pm}| + |\phi^0(x)| \leq |e^{\pm}| + C'R|x|.$$

Which gives the following lower bound for $|p^{\pm}| + |e^{\pm}|$ in terms of x:

$$\begin{aligned} |p^{\pm}| + |e^{\pm}| &= |v_2 \pm \psi^0(x) \pm \psi^{ext}(x)| + |e^{\pm}| \geq \psi^{ext}(x) - |v_2| - |\psi^0(x)| + |e^{\pm}| \\ &\geq \psi^{ext}(x) - |e^{\pm}| - 2C'R|x| + |e^{\pm}| \\ &\geq C(1 + |x|^2) - 2C'R|x|. \end{aligned}$$

By taking C' small enough and C large enough we can ensure that if $|x| > R$ then $(e^{\pm}, p^{\pm})$ lie outside any disc in $\mathbb{R}^2$, and in particular outside $A^{\pm}$, where $\mu^{\pm}$ are supported. This proves the claim.

Step 2: Uniform L^∞ bound. We show that C' and c can be chosen so that $\mathcal{F}$ maps to functions with L^∞ norm smaller than C'.

Estimating $\phi^0(x)$ for $|x| \leq R$

$$|\phi^0(x)| \leq \frac{C'}{2} \int_{-R}^{R} |x-y|\, dy = \frac{C'}{2}(x^2 + R^2),$$

we take C' small enough (recall it was already taken to be small in the previous step, hence we may require it to be even smaller) so that $|\phi^0(x)| \leq 3/4$ for $|x| \leq R$. Now the decay assumption on $\mu^{\pm}$ allows us to show a uniform bound on $|\mathcal{F}_1(\rho, j_2)(x)|$ in $|x| \leq R$:

$$\begin{aligned} |\mathcal{F}_1(\rho, j_2)(x)| &\leq \sum_{\pm} \int |\mu^{\pm}(e^{\pm}, p^{\pm})|\, d\boldsymbol{v} \leq \sum_{\pm} \int \frac{c}{(1 + |e^{\pm}|)^{\alpha}}\, d\boldsymbol{v} \\ &\leq \sum_{\pm} \int \frac{c}{(1 + \langle \boldsymbol{v} \rangle - |\phi^0(x)|)^{\alpha}}\, d\boldsymbol{v} \leq \sum_{\pm} \int \frac{c}{(1 + \langle \boldsymbol{v} \rangle - \frac{3}{4})^{\alpha}}\, d\boldsymbol{v} \quad (6.7.2) \\ &= \int \frac{2c}{(\frac{1}{4} + \langle \boldsymbol{v} \rangle)^{\alpha}}\, d\boldsymbol{v} = C''c < \infty. \end{aligned}$$

We can bound $|\mathcal{F}_2(\rho, j_2)(x)|$ in the same way as $|\hat{v}| \leq 1$. Finally we choose c so that $C''c \leq C'$.

Step 3: Uniform L^∞ bound on the derivative. We show that there is a constant C''' such that for any $(\rho, j_2) \in X$, we have $\|\partial_x \mathcal{F}_1(\rho, j_2)\|_{L^\infty[-R,R]} \leq C'''$ and $\|\partial_x \mathcal{F}_2(\rho, j_2)\|_{L^\infty[-R,R]} \leq C'''$.

We compute for $\mathcal{F}_1$, and note that $\mathcal{F}_2$ is analogous. Using the chain rule, we have

$$\begin{aligned}\partial_x \int \left(\mu^+(e^+, p^+) - \mu^-(e^-, p^-)\right) d\boldsymbol{v} = \\ (\partial_x \phi^0) \int (\mu_e^+(e^+, p^+) + \mu_e^-(e^-, p^-))\, d\boldsymbol{v} \\ + (\partial_x \psi^0 + \partial_x \psi^{ext}) \int (\mu_p^+(e^+, p^+) + \mu_p^-(e^-, p^-))\, d\boldsymbol{v}.\end{aligned}$$

The two integrals are bounded uniformly in x by the arguments in Step 2 using the corresponding assumed bounds on $\mu_e^\pm$ and $\mu_p^\pm$ respectively. As the external field ψ^{ext} lies in $H^2_{loc}(\mathbb{R})$, its derivative $\partial_x \psi^{ext}$ lies in $H^1([-R, R])$ and is bounded in $L^\infty([-R, R])$ by Morrey's inequality. It remains to bound $\partial_x \phi^0$ and $\partial_x \psi^0$ uniformly for all $x \in [-R, R]$. These are controlled directly using the Green's function $G(x)$ and uniform bounds of Step 2. Indeed,

$$|(\partial_x \phi^0)(x)| = |((\partial_x G) * \rho)(x)| \leq \frac{C'}{2} \int_{-R}^{R} |\operatorname{sign}(x - y)|\, dy \leq C'R$$

and the computation for $\partial_x \psi^0$ is identical.

Step 4: $\mathcal{F}$ is a compact continuous map from X to X. Steps 1 and 2 imply that $\mathcal{F}(X) \subseteq X$. Step 3 and the Rellich theorem imply that $\mathcal{F}(X)$ is relatively compact in X. It remains to show that $\mathcal{F}$ is continuous. This may be shown using dominated convergence and the bounds in Step 2. Indeed, suppose that $\{(\rho^n, j_2^n)\}_{n\in\mathbb{N}} \subseteq X$ is a sequence converging to $(\rho, j_2) \in X$ strongly in $L^2(\mathbb{R})^2$. We shall show that $\mathcal{F}_1(\rho^n, j_2^n) \to \mathcal{F}_1(\rho, j_2)$ in L^2, the result for $\mathcal{F}_2$ is analogous. By Step 2 and dominated convergence it is enough to show convergence pointwise, i.e. for each $x \in [-R, R]$. Next, by (6.7.2) and dominated convergence again, it is sufficient to show that the corresponding densities $\mu^\pm(e^\pm, p^\pm)$ converge pointwise in $(x, \boldsymbol{v})$. Continuity of $\mu^\pm$ reduces this to showing pointwise convergence of the

corresponding microscopic energy and momenta $e^{\pm}$ and $p^{\pm}$. The definitions of these quantities imply that it is enough to show that the corresponding electric and magnetic potentials $\phi^{0,n}$ and $\psi^{0,n}$ converge pointwise. Finally, as the potentials are (ρ^n, j_2^n) convolved with $G(x) = -|x|/2$, the convergence $(\rho^n, j_2^n) \to (\rho, j_2)$ in $L^2([-R,R])^2$ gives the required pointwise convergence.

This concludes the proof. □

6.A Appendix

Lemma 6.A.1. *Let $(f^{0,\pm}, E_1^0, E_0^2, B^0) \in C^1$ be an equilibrium of (6.1.15) with $f^{0,\pm}$ compactly supported in x and be such that $f^{0,\pm}(x, \boldsymbol{v})|\boldsymbol{v}| \to 0$ as $|\boldsymbol{v}| \to \infty$. Then E_0^2 is identically zero.*

Proof. Let $\psi(x)$ be a test function (smooth and compactly supported) and define $\Psi(x)$ as

$$\Phi(x) = \int_{-\infty}^{x} \psi(y)\, dy$$

Then by testing the Vlasov equation with Ψ (which is possible due to the compact x support of $f^{0,\pm}$), we deduce that

$$-\int_{\mathbb{R}^3} \hat{v}_1 f^{\pm}\, d\boldsymbol{v}\phi(x)dx = 0.$$

As ϕ was arbitrary we have

$$\int f^{\pm}\hat{v}_1\, d\boldsymbol{v} = 0$$

for all x (and each of $\pm$).

By multiplying the Vlasov equation by v_2 and integrating by parts (using the decay assumption on $f^{0,\pm}$ in $\boldsymbol{v}$) we obtain that

$$\mp\int_{\mathbb{R}^3} (E_2^0(x) - \hat{v}_1 B^0(x)) f^{0,\pm}\, dx d\boldsymbol{v} = 0$$

This implies that

$$\int E_2^0(x) \int f^{0,\pm}\, d\boldsymbol{v} - B^0(x) \int f^{0,\pm}\hat{v}_1\, d\boldsymbol{v}\, dx = 0$$

As the second integral of $f^{\pm}$ vanishes by the computation before, and as $f^{\pm} \geq 0$, we deduce that E_2, being a constant due to (6.1.15e), is zero. $\square$

Chapter 7

A stability estimate for solutions of the Vlasov-Poisson system with spatial density in Orlicz spaces

In this chapter we present quantitative well-posedness results for the Vlasov-Poisson system for solutions with unbounded spatial density belonging to a class of exponential Orlicz spaces. This improves and interpolates between the uniqueness proved in [134] and [143]. We also show that the conditions on the spatial density are satisfied for a class of initial data that possess exponential velocity moments. The proofs exploit the second order structure of the Newtonian dynamics of the Vlasov-Poisson system.

Acknowledgements

The work in this chapter was done in collaboration with Evelyne Miot and consists of a paper currently in preparation [95].

7.1 Introduction

The purpose of this chapter is to study uniqueness and stability issues for a class of weak solutions of the Vlasov-Poisson system in dimension $d = 2$ or $d = 3$, which reads:

$$\begin{cases} \partial_t f + v \cdot \nabla_x f + E \cdot \nabla_v f = 0, \quad (t,x,v) \in [0,T] \times \mathbb{R}^d \times \mathbb{R}^d \\ E(t,x) = (K *_x \rho)(t,x), \quad K(x) = \gamma \dfrac{x}{|x|^d} \\ \rho(t,x) = \displaystyle\int_{\mathbb{R}^d} f(t,x,v)\, dv. \end{cases} \tag{7.1.1}$$

The system (7.1.1) describes the evolution of a microscopic density $f = f(t,x,v)$ of interacting particles, that are electric particles for $\gamma = 1$ (Coulombian interaction) or stars for $\gamma = -1$ (gravitational interaction). The function ρ is called macroscopic (or spatial) density.

Existence and uniqueness of classical solutions of (7.1.1) defined on $[0,T]$ for all $T > 0$ were established by Ukai and Okabe [192] for $d = 2$ and by Pfaffelmoser [168] for $d = 3$. Arsenev [11] proved global existence of weak solutions with finite energy. Another kind of global solutions, which propagate the velocity moments, was constructed by Lions and Perthame [133]. We refer to the articles [71, 163], and to references quoted therein, for further related results. On the other hand, part of the literature is devoted to determining sufficient conditions for uniqueness. Loeper [134] established uniqueness on $[0,T]$ in the class of weak solutions such that the macroscopic density ρ is uniformly bounded: [1]

$$f \in C\left([0,T], \mathcal{M}_+(\mathbb{R}^d \times \mathbb{R}^d) - w^*\right) \quad \text{and } \rho \in L^\infty\left([0,T], L^\infty(\mathbb{R}^d)\right). \tag{7.1.2}$$

This result was extended by the second author in [143] to weak solutions satisfying

$$f \in C\left([0,T], \mathcal{M}_+(\mathbb{R}^d \times \mathbb{R}^d) - w^*\right) \quad \text{and } \sup_{[0,T]} \sup_{p \geq 1} \frac{\|\rho(t)\|_{L^p}}{p} < +\infty. \tag{7.1.3}$$

Our first result, stated in Theorem 7.1.1 below, establishes uniqueness in the class

[1] Here and throughout $\mathcal{M}_+(\mathbb{R}^d \times \mathbb{R}^d)$ denotes the space of bounded positive measures.

of solutions with macroscopic density belonging to a certain class of exponential Orlicz spaces defined in (7.1.7). These spaces interpolate the functional spaces arising in (7.1.2) and (7.1.3). More precisely, we obtain in Theorem 7.1.1 a quantitative stability estimate involving the Wasserstein distance[2] between such weak solutions, in the spirit of the method of Dobrushin [49] to establish stability estimates for mean field PDE with Lipschitz convolution Kernels K.

In the second part of this paper, we look for sufficient conditions on the initial data ensuring that any corresponding solution has spatial density belonging to the exponential Orlicz spaces defined in (7.1.7) on $[0, T]$. In Proposition 7.1.1 we prove that this holds for data with finite exponential velocity moment.

7.1.1 Preliminary definitions on Orlicz spaces and on the Wasserstein distance

Orlicz spaces. We begin by recalling some standard definitions related to Orlicz spaces. We refer the reader to e.g. [170] for a more thorough exposition.

Definition 7.1.1 (N-function)**.** *We say that a function $\phi : [0, \infty) \to [0, \infty)$ is an N-function if it is continuous, convex with $\phi(\tau) > 0$ for $\tau > 0$ and satisfies both $\lim_{\tau \to 0} \phi(\tau)/\tau = 0$ and $\lim_{\tau \to \infty} \phi(\tau)/\tau = \infty$.*

Definition 7.1.2 (Luxemburg norm)**.** *Let U be a domain of $\mathbb{R}^d$. For an N-function ϕ we define the Luxemburg norm of a function f defined on U as*

$$\|f\|_{L_\phi(U)} = \inf \left\{ \lambda > 0 : \int_U \phi(|f(x)|/\lambda)\, dx < 1 \right\}. \tag{7.1.4}$$

Remark 7.1.1. *If it holds for some constant C' that*

$$\int_U \phi(|f(x)|/\gamma)\, dx < C' \tag{7.1.5}$$

then $\|f\|_{L_\phi(U)} \leq C\gamma$, where C is an absolute constant depending only on C'.

Remark 7.1.2. *On bounded domains only the asymptotic behaviour as $\tau \to \infty$ of the N-function ϕ is important in defining the space L_ϕ. In particular, if two*

[2] See Definition 7.1.4 hereafter of the Wasserstein distance.

N-functions $\phi, \tilde{\phi}$ have the same behaviour at infinity in the sense that there are $K, \tilde{K} > 0$ such that $\phi(\tau) \leq \tilde{\phi}(\tilde{K}\tau)$ and $\tilde{\phi}(\tau) \leq \phi(K\tau)$ for all sufficiently large τ, then the norms $\|\cdot\|_{L_\phi(U)}$ and $\|\cdot\|_{L_{\tilde{\phi}}(U)}$ are equivalent for any bounded domain $U \subseteq \mathbb{R}^d$.

Definition 7.1.3 (Complementary N-function)**.** *For an N-function ϕ we define its complementary N-function $\bar{\phi}$ as*

$$\bar{\phi}(\tau) = \int_0^\tau a(s)\, ds$$

where a is the right inverse of the right derivative of ϕ.

For $\alpha \in [1, +\infty)$ we let, for $\tau \geq 0$,

$$\phi_\alpha(\tau) = \exp(|\tau|^\alpha) - 1. \tag{7.1.6}$$

The spaces $L_{\phi_\alpha}(\mathbb{R}^d)$ are exponential Orlicz spaces, and can be equivalently characterised as those functions g which lie in L^p for all $p \in [\alpha, \infty)$ and have the following norm finite:

$$\|g\|_{\phi_\alpha} = \sup_{p \geq \alpha} p^{-1/\alpha} \|g\|_{L^p(\mathbb{R}^d)}, \tag{7.1.7}$$

which is an equivalent norm to the Luxemburg norm $\|\cdot\|_{L_{\phi_\alpha}}$. This equivalence is standard and can be verified by Taylor expansion of the exponential. Note that in the $\alpha \to \infty$ limiting case we obtain the function ϕ_∞ given by $\phi_\infty(\tau) = \infty$ if $\tau > 1$ and 0 otherwise. Although ϕ_∞ is not an N-function, we will use the convention that $L_{\phi_\infty} = L^\infty$. Therefore with this convention L_{ϕ_α} indeed interpolates the functional spaces for ρ that are considered in [134] (for $\alpha = +\infty$) and [143] (for $\alpha = 1$).

Remark 7.1.3. *When working with exponential Orlicz spaces, one has the choice between working with the Luxemburg norm (7.1.2) directly, or working with L^p norms* uniformly *in p and using (7.1.7), as is done in [143] for $\alpha = 1$. We take the former approach in this work.*

Transportation distances. We next turn to some definitions concerning the notion of transportation distances.

Definition 7.1.4 (Wasserstein distance)**.** *For two measures $\mu, \nu \in \mathcal{M}_+(\mathbb{R}^d)$ with the same mass and finite first moments, we define the (Monge-Kantorovich-Rubenstein)-Wasserstein distance $W_1(\mu, \nu)$ as*

$$W_1(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} |x - y| \, d\pi(x, y),$$

where, here and throughout, $\Pi(\mu, \nu)$ denotes the set of couplings *between μ and ν, by which we mean measures in $\mathcal{M}_+(\mathbb{R}^d \times \mathbb{R}^d)$ which have marginals μ and ν respectively.*

Remark 7.1.4. *The Wasserstein distance is usually defined on probability measures (i.e. elements of $\mathcal{M}_+$ with mass 1) and metrises the weak* topology on the space of probability measures with finite first moment. In the case of the extension to general bounded positive measures given above, it should be noted that the Wasserstein distance does not metrise the weak* topology on $\mathcal{M}_+$ with finite first moment. However, given any fixed mass m, the Wasserstein distance metrises the weak* topology on measures in $\mathcal{M}_+$ of mass m with first moment finite.*

7.1.2 Main results

We are now in position to state a quantitative estimate on the Wasserstein distance between two weak solutions of (7.1.1) with macroscopic density belonging to some exponential Orlicz space:

$$\rho_j \in L^\infty([0, T]; L_{\phi_\alpha}(\mathbb{R}^d)) \text{ for } j = 1, 2 \text{ and some } \alpha \in [1, \infty]. \tag{7.1.8}$$

Theorem 7.1.1. *Let $\varepsilon > 0$ and $T > 0$. Let $f_1, f_2 \in C([0, T], \mathcal{M}_+(\mathbb{R}^d \times \mathbb{R}^d) - w^*)$ be two weak solutions of the Vlasov-Poisson system (7.1.1) with the same total mass such that (7.1.8) holds. If $(1 + T)^{1+\varepsilon} W_1(f_1(0), f_2(0)) < 1/18$, then we have the bound for $t \in [0, T^*]$:*

- *If* $\alpha = 1$ *then*

$$W_1(f_1(t), f_2(t)) \leq C W_1(f_1(0), f_2(0))^{\exp(-Ct)} \left(1 + t|\log W_1(f_1(0), f_2(0)|^2\right).$$

- *If* $\alpha > 1$ *then*

$$W_1(f_1(t), f_2(t)) \leq C W_1(f_1(0), f_2(0))^{1/\gamma} \exp(Ct^\gamma) \left(1 + t|\log W_1(f_1(0), f_2(0))|^{1+(1/\alpha)}\right),$$

where

$$\gamma = \frac{2}{1 - (1/\alpha)} \in [2, +\infty),$$

and where T^* *satisfies the lower bound*

$$T^* \geq \begin{cases} C' \log|\log W_1(f_1(0), f_2(0))| - C'^{-1} & \text{when } \alpha = 1, \\ C'\gamma|\log W_1(f_1(0), f_2(0))|^{1/\gamma} - C'^{-1} & \text{when } \alpha \neq 1. \end{cases}$$

(we set $T^* = T$ *if the right hand side is larger than* T*).*

The constants C *and* C' *depend only upon the norms of* ρ_1, ρ_2 *in* (7.1.8) *and on* ε*.*

Remark 7.1.5. *The bound is stated in a way that is easy to understand for* large t *and is suboptimal near* $t = 0$*. In particular the bound does not converge to* $W_1(f_1(0), f_2(0))$ *as* $t \to 0$[3]*. Such a bound could be obtained by a careful analysis of the proofs, but we do not present this here.*

Remark 7.1.6. *As will be clear in the proof of Theorem 5.1.1, the time* T^* *essentially corresponds to the first time at which the right hand side becomes larger or equal to* 1*.*

Remark 7.1.7. *For* $\alpha = +\infty$*, the estimate of Theorem 7.1.1 reads*

$$W_1(f_1(t), f_2(t)) \leq C W_1(f_1(0), f_2(0))^{1/2} \exp(Ct^2) \left(1 + t|\log W_1(f_1(0), f_2(0))|\right),$$

which is valid up to times of order $|\log W_1(f_1(0), f_2(0))|^{1/2}$*.*

In [49], Dobrushin considered the stability of measure-valued solutions of first

[3] They do, of course, converge to zero as $W_1(f_1(0), f_2(0)) \to 0$.

order mean-field PDE with Lipschitz convolution Kernels K and obtained the inequality

$$W_1(f_1(t), f_2(t)) \leq W_1(f_1(0), f_2(0)) \exp\left(Ct\|\nabla K\|_{L^\infty}\right).$$

The same estimate was derived by Moussa and Sueur [152] for a mixed first/second order PDE. Hauray and Jabin [81] handled the case of more singular Kernels, see also the recent work by Lazarovici and Pickl [123] on cut-off kernels and the references quoted therein.

In the present situation, we are able to address the case of the singular convolution Kernel $K = \gamma x/|x|^d$ because, in contrast with the works mentioned above, the solutions have some additional regularity - the macroscopic density belongs to L_{ϕ_α}. Nevertheless, as a consequence of the singularity of K, the growth of $W_1(f_1(t), f_2(t))$ in Theorem 7.1.1 is not linearly bounded in terms of $W_1(f_1(0), f_2(0))$.

We mention that although stability estimates are not explicitly done in [134], the computations therein involve a log-Lipschitz Grönwall estimate and would yield the inequality

$$W_2(f_1(t), f_2(t)) \leq C W_2(f_1(0), f_2(0))^{\exp(-Ct)}, \tag{7.1.9}$$

with W_2 denoting

$$W_2(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} |x - y|^2 \, d\pi(x, y) \right)^{1/2},$$

so the W_2-Wasserstein distance grows in time roughly like an exponential tower $e^{e^{ct}}$. Therefore the estimate of Theorem 7.1.1 setting $\alpha = +\infty$, which corresponds to the regularity considered in [134], improves this to stretched exponential growth of the form e^{ct^2}. This improvement is due to the second-order structure of the characteristic system (7.2.5) of ODE associated to the Vlasov-Poisson system, which was already exploited in the proof of uniqueness in [143].

Finally, we would like to point out that the same technique of exploiting the second-order structure can be applied to general measure solutions (with no regularity assumption on the spatial density), and allows the Dobrushin estimate to

be improved slightly from *Lipschitz* kernels to *log²-Lipschitz* kernels:

Theorem 7.1.2. *Let the convolution kernel K be bounded and satisfy the log²-Lipschitz property:*

$$|K(x) - K(y)| \leq C|x-y||\log(x-y)|^2 \text{ for all } |x-y| \leq 1/9. \tag{7.1.10}$$

Then the Vlasov-Poisson system (7.1.1) possesses a unique solution in the space $C([0,T]; \mathcal{M}_+(\mathbb{R}^d \times \mathbb{R}^d) - w^*)$ for any initial datum in $\mathcal{M}_+(\mathbb{R}^d \times \mathbb{R}^d)$. Moreover obeys the stability estimate, for any two solutions $f_1, f_2 \in C([0,T]; \mathcal{M}_+(\mathbb{R}^d \times \mathbb{R}^d) - w^*)$ with the same mass and satisfying $(1+T)W_1(f_1(0), f_2(0)) < 1/9$,

$$W_1(f_1(t), f_2(t)) \leq C W_1(f_1(0), f_2(0))^{\exp(-Ct)}(1 + t|\log W_1(f_1(0), f_2(0))|^2).$$

which holds for times $t \in [0, T^*]$ with T^* defined analogously to Theorem 5.1.1.

We remark that the conventional improvement of the Dobrushin estimate by replacing the Grönwall inequality with a log-Lipschitz inequality only allows one to treat log-Lipschitz kernels K, rather than the slightly weaker assumption (7.1.10).

In the second part of our analysis, we seek for initial data f_0 for which the macroscopic density indeed belongs to some exponential Orlicz space.

Proposition 7.1.1. *Let $f_0 \in L^\infty(\mathbb{R}^d \times \mathbb{R}^d)$ be such that*

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} f_0(x,v) e^{c\langle v \rangle^{d\alpha}} \, dxdv < \infty$$

for some $\alpha \in [1, \infty)$ and $c > 0$, where $\langle v \rangle = \sqrt{1+|v|^2}$. For $T > 0$, let $f \in C([0,T], \mathcal{M}_+(\mathbb{R}^d \times \mathbb{R}^d) - w^*) \cap L^\infty([0,T], L^1 \cap L^\infty(\mathbb{R}^d \times \mathbb{R}^d))$ be any solution to (7.1.1), with this initial datum, provided by [133, Theo. 1][4]. Then it satisfies

$$\sup_{t \in [0,T]} \|\rho(t)\|_{L_{\phi_\alpha}} \leq C < \infty.$$

In particular, this solution satisfies the uniqueness criterion of Theorem 7.1.1.

[4]The existence of such a solution is ensured by [133, Theo. 1] because f_0 has finite velocity moments of sufficiently large order: $\int |v|^m f_0(x,v)\, dx\, dv < +\infty$ for some $m > d^2 - d$. This is proved in [133] for $d = 3$. The case $d = 2$ is a straightforward adaptation of the case $d = 3$.

We remark that setting $\alpha = 1$, we retrieve as a special case the condition obtained in [143, Theo. 1.2] to ensure that (7.1.3) holds. For any $\alpha \in [1, \infty)$, an example of initial data that are allowed by Proposition 7.1.1 are Maxwell-Boltzmann distributions of the form

$$f_0(x, v) = e^{-c|v|^{d\alpha}} \langle v \rangle^p h_0(x, v), \quad c > 0,\ p \geq 0,\ h_0 \in L^1 \cap L^\infty.$$

The plan of the remainder of this chapter is as follows. In Section 7.2 we prove Theorem 7.1.1. In order to do so, we first establish a log-Lipschitz like estimate for the force field $E = K * \rho$ associated to a function ρ satisfying (7.1.8). Then, we introduce in (7.2.7) a notion of distance between two solutions in terms of the characteristics defined in (7.2.5), which controls the Wasserstein distance (see (7.2.8)). This quantity was used in the original proof of Dobrushin and also in [143], while the proof of [134] uses a slightly different version. Applying similar arguments as in [49], we derive a second-order differential inequality for this distance, which eventually leads to Theorem 7.1.1. In Section 7.2.4 we show how to adapt this technique to prove Theorem 7.1.2. Finally, the last Section 7.3 is devoted to the proof of Proposition 7.1.1.

7.2 Proof of Theorem 7.1.1

7.2.1 An estimate for the Newton kernel

To prove Theorem 7.1.1 we have need of the following lemma on the Newton kernel. Note that the complementary N-functions of the ϕ_α behave asymptotically (see Remark 7.1.2) like

$$\bar{\phi}_\alpha(\tau) \sim \tau \log(\tau)^{1/\alpha} \text{ as } \tau \to \infty. \tag{7.2.1}$$

Recall that Orlicz spaces obey a form of Hölders inequality (see e.g. [170])

$$\int f g \, dx \leq C \, \|f\|_{L_\phi} \, \|g\|_{L_{\bar{\phi}}}$$

for the constant $C > 1$.

Given $\alpha \in [1, \infty]$ we define the constant $\beta \in [1, 2]$ by

$$\beta = \frac{1}{\alpha} + 1. \tag{7.2.2}$$

In particular, note that $\beta = 1$ for $\alpha = +\infty$ and $\beta = 2$ as $\alpha = 1$.

Lemma 7.2.1. *Let $\alpha \in [1, \infty)$, then there exists $C = C(\alpha) > 0$ such that for all $g \in L_{\phi_\alpha} \cap L^1$ we have the estimate*

$$\int_{\mathbb{R}^d} |K(x-z) - K(y-z)|\, |g(z)|\, dz \leq C(\|g\|_{L_{\phi_\alpha}} + \|g\|_{L^1}) \psi_\alpha(|x-y|) \tag{7.2.3}$$

where ψ is defined by

$$\psi_\alpha(\tau) = \begin{cases} \tau |\log(\tau)|^\beta, & \text{for } \tau \in \left[0, \frac{1}{9}\right], \\ \frac{1}{9} \log(9)^\beta, & \text{for } \tau \geq \frac{1}{9} \end{cases} \tag{7.2.4}$$

and where β is defined by (7.2.2).

Remark 7.2.1. *In the case $\alpha = +\infty$, namely for $g \in L^1 \cap L^\infty$ the estimate of Lemma 7.2.1 is standard, see e.g. [136, Lemma 8.1] for the case $n = 2$: we have*

$$\int_{\mathbb{R}^d} |K(x-z) - K(y-z)|\, |g(z)|\, dz \leq C(\|g\|_{L^1 \cap L^\infty}) |x-y| \left(1 + |\log|x-y||\right).$$

Remark 7.2.2. *For the case $\alpha = 1$, the following variant of Lemma 7.2.1 was obtained in [143, Lemma 2.2]: for all $x, y \in \mathbb{R}^d$ with $|x-y|$ sufficiently small,*

$$\int_{\mathbb{R}^d} |K(x-z) - K(y-z)|\, |g(z)|\, dz \leq C\, p\, (\|g\|_{L^1} + \|g\|_{L^p}) |x-y|^{1-d/p}, \quad \forall p > d.$$

In particular, recalling (7.1.7) for $\alpha = 1$, this yields

$$\int_{\mathbb{R}^d} |K(x-z) - K(y-z)|\, |g(z)|\, dz \leq C \|g\|_{L_{\phi_1}} |x-y| \left(p^2 |x-y|^{-d/p}\right), \quad \forall p > d,$$

so setting $p = |\log|x-y||$ we retrieve the estimate of Lemma 7.2.1. In fact one can also prove the other cases via this method. Nevertheless, we give a direct proof of Lemma 7.2.1 below for completeness.

Proof. We set

$$\delta = \frac{1}{d}\left(1+\frac{1}{\alpha}\right) = \frac{\beta}{d} \in \left(\frac{1}{d}, \frac{2}{d}\right].$$

By standard estimates using Hölder's inequality (see e.g. [138]) it is well-known that, fixing some $p_0 > d$,

$$\int_{\mathbb{R}^d} |K(x-z) - K(y-z)|\,|g(z)|\,dz \le C(\|g\|_{L^{p_0}} + \|g\|_{L^1}) \le C(\|g\|_{L_{\phi_\alpha}} + \|g\|_{L^1}).$$

Hence, in view of the form of ψ_α, letting $R = |x-y|$ we may assume without loss of generality that $R \le 1/9$. We introduce $A = (x+y)/2$. Since $R|\log R|^\delta < 1$, we may split the integral as follows:

$$\begin{aligned}
&\int_{\mathbb{R}^d} |K(x-z) - K(y-z)|\,|g(z)|\,dz \\
&\quad = \int_{\mathbb{R}^d \setminus B(A,|\log R|^{-\delta})} |K(x-z) - K(y-z)|\,|g(z)|\,dz \\
&\quad + \int_{B(A,|\log R|^{-\delta}) \setminus B(A,R)} |K(x-z) - K(y-z)|\,|g(z)|\,dz \\
&\quad + \int_{B(A,R)} |K(x-z) - K(y-z)|\,|g(z)|\,dz \\
&\quad = I_1 + I_2 + I_3.
\end{aligned}$$

For I_1 we apply the mean value theorem to obtain the bound

$$I_1 \le C\|g\|_{L^1}\, R \sup_{u\in[x,y], z\in\mathbb{R}^d\setminus B(A,|\log R|^{-\delta})} \frac{1}{|u-z|^d} \le C\|g\|_{L^1}\, R|\log R|^{d\delta}$$

where $[x,y]$ is the line segment joining x and y, and where we have used that

$$|u-z| \ge |\log R|^{-\delta} - \frac{R}{2} \ge \frac{|\log R|^{-\delta}}{2}$$

in the considered supremum. Therefore we have obtained

$$I_1 \le C\|g\|_{L^1} R|\log R|^\beta.$$

For I_3 we apply Hölder's inequality for Orlicz spaces,

$$I_3 \le C\|g\|_{L_{\phi_\alpha}} \|1_{B(A,R)}|K(x-z) - K(y-z)|\|_{L_{\bar\phi_\alpha}} \le C\|g\|_{L_{\phi_\alpha}} \|1_{B(0,3R/2)} K\|_{L_{\bar\phi_\alpha}}$$

where we have used the fact that $\bar{\phi}_\alpha$ is increasing, that

$$|K(x-z) - K(y-z)| \leq |K(x-z)| + |K(y-z)|$$

and that $z \in B(A,R)$ implies that both $x-z$ and $y-z$ lie in $B(0,3R/2)$.

Now we set $\lambda = R|\log R|^{1/\alpha}$ and we consider the integral

$$\int_{\mathbb{R}^d} \bar{\phi}_\alpha(1_{B(0,3R/2)}|K(z)|/\lambda)\, dz.$$

By Remark 7.1.1, to show that $I_3 \leq C\|g\|_{L_{\phi_\alpha}}\lambda$ it is sufficient to show that the integral above is bounded by a constant. Furthermore, by Remark 7.1.2 using the fact that $|K(z)|/\lambda \geq 1 > 0$ on $B(0,3R/2)$ we may work with the asymptotic form (7.2.1).

Thus, we estimate

$$\begin{aligned}\int_{|z|\leq 3R/2} \frac{|K(z)|}{\lambda} \log\left(\frac{|K(z)|}{\lambda}\right)^{1/\alpha} dz &= \frac{1}{\lambda}\int_{|z|\leq 3R/2} |z|^{1-d} \log\left(\frac{|z|^{1-d}}{\lambda}\right)^{1/\alpha} dz \\ &\leq \frac{C}{\lambda}\int_0^{3R/2} |\log r|^{1/\alpha}\, dr\end{aligned}$$

where we have used the inequality

$$0 \leq \log\left(\frac{|z|^{1-d}}{\lambda}\right) \leq (d-1)|\log|z|| + |\log\lambda| \leq C|\log|z||$$

with this definition of λ.

Thus, noting that for $r \leq 3R/2 \leq 1/6$ we have

$$\frac{1}{\alpha}|\log r|^{(1/\alpha)-1} \leq \frac{1}{2}|\log r|^{1/\alpha},$$

so that

$$|\log r|^{1/\alpha} \leq 2\left(|\log r|^{1/\alpha} - \frac{1}{\alpha}|\log r|^{(1/\alpha)-1}\right),$$

we obtain

$$\int_{|z|\leq 3R/2} \frac{|K(z)|}{\lambda} \log\left(\frac{|K(z)|}{\lambda}\right)^{1/\alpha} dz \leq \frac{C}{\lambda}\int_0^{3R/2} \left(|\log r|^{1/\alpha} - \frac{1}{\alpha}|\log r|^{(1/\alpha)-1}\right) dr$$
$$= \frac{C}{\lambda}\left[r|\log r|^{1/\alpha}\right]_0^{3R/2} \leq \frac{C}{\lambda} R|\log R|^{1/\alpha}.$$

Thus we have shown that $I_3 \leq C\|g\|_{L_{\phi_\alpha}} R|\log R|^{1/\alpha}$.

Finally, we bound I_2. In the same way as for I_3, we apply Hölder's inequality for Orlicz spaces to obtain

$$I_2 \leq C\|g\|_{L_{\phi_\alpha}} \|1_{B(A,|\log R|^{-\delta})\setminus B(A,R)}|K(x-z)-K(y-z)|\|_{L_{\bar{\phi}_\alpha}}.$$

Applying the mean value theorem we obtain for $z \in B(A,|\log R|^{-\delta}) \setminus B(A,R)$

$$|K(x-z)-K(y-z)| \leq C\,R \sup_{u\in[x,y]} |u-z|^{-d}$$
$$\leq C\,R \sup_{u\in[x,y]} \frac{1}{||z-A|-|u-A||^d}$$
$$\leq C\,R|z-A|^{-d},$$

where we have used that $|z-A| \geq R$ to obtain the final inequality. Hence, by a change of variables, and since $\bar{\phi}_\alpha$ is increasing, in order to estimate I_2 it is sufficient to obtain the bound

$$\|1_{B(0,|\log R|^{-\delta})\setminus B(0,R)} R|z|^{-d}\|_{L_{\bar{\phi}_\alpha}} \leq C\,R|\log R|^{\beta}.$$

Therefore, setting $\lambda' = R|\log R|^{\beta}$, by Remark 7.1.1 it is enough to show that

$$\int_{R\leq|z|\leq|\log R|^{-\delta}} \bar{\phi}_\alpha(R|z|^{-d}/\lambda')\,dz \leq C.$$

Let $|z| \leq |\log R|^{-\delta}$, then we have by definition of λ'

$$\frac{R|z|^{-d}}{\lambda'} \geq \frac{R|\log R|^{d\delta}}{\lambda'} = 1,$$

so by Remark 7.1.2 we may instead bound the asymptotic form (7.2.1). Therefore,

we estimate

$$\int_{R\le|z|\le|\log R|^{-\delta}} \frac{R|z|^{-d}}{\lambda'}\log\left(\frac{R|z|^{-d}}{\lambda'}\right)^{1/\alpha} dz$$
$$= |\log R|^{-\beta}\int_{R\le|z|\le|\log R|^{-\delta}} |z|^{-d}\log(|z|^{-d}|\log R|^{-\beta})^{1/\alpha}\,dz\cdot$$
$$= C|\log R|^{-\beta}\int_R^{|\log R|^{-\delta}} r^{-1}|\log(r^{-d}|\log R|^{-\beta})|^{1/\alpha}\,dr$$

Since for $r \le |\log R|^{-\delta}$ we have

$$|\log(r^{-d}|\log R|^{-\beta})| = \log(r^{-d}|\log R|^{-\beta}) = d|\log r| - \beta\log|\log R| \le d|\log r|,$$

we infer that

$$\int_{R\le|z|\le|\log R|^{-\delta}} \frac{R|z|^{-d}}{\lambda'}\log\left(\frac{R|z|^{-d}}{\lambda'}\right)^{1/\alpha} dz$$
$$\le C|\log R|^{-\beta}\int_R^{|\log R|^{-\delta}} r^{-1}|\log r|^{1/\alpha}\,dr$$
$$\le C|\log R|^{-\beta}\left[-|\log r|^{(1/\alpha)+1}\right]_R^{|\log R|^{-\delta}}$$
$$\le C,$$

as we wanted, and hence we obtain $I_2 \le C\|g\|_{L_{\phi_\alpha}(\mathbb{R}^d)}R|\log R|^\beta$. Finally, putting this all together, we conclude that

$$\int_{\mathbb{R}^d} |K(x-z)-K(y-z)|\,|g(z)|\,dz$$
$$\le I_1+I_2+I_3 \le C(\|g\|_{L^1}R + \|g\|_{L_{\phi_\alpha}}R|\log R|^\beta + \|g\|_{L_{\phi_\alpha}}R|\log R|^{1/\alpha})$$

which implies the claim of the lemma. □

7.2.2 Lagrangian formulation of the Vlasov-Poisson system and the Wasserstein distance

Let $f \in C([0,T],\mathcal{M}_+(\mathbb{R}^d\times\mathbb{R}^d)-w^*)$ be a weak measure-valued solution of the Vlasov-Poisson system (7.1.1) on $[0,T]$ such that ρ belongs to $L^\infty([0,T],L^1\cap L^p(\mathbb{R}^d))$ for some $p>d$. By potential estimates it is well-known that E belongs

to $L^\infty([0,T]\times\mathbb{R}^d)$. Moreover, by Caldéron-Zygmund inequality (see e.g. see [52, Theo. 4.12]) $\nabla E \in L^\infty([0,T], L^p(\mathbb{R}^d))$. By the theory on transport equations (see [47, Theo. III2] or [7, Theo. 5.7] for more recent results on the theory), there exists a unique Lagrangian flow associated to E, namely a map $(X,V) \in L^1_{\mathrm{loc}}([0,T]\times\mathbb{R}^d\times\mathbb{R}^d;\mathbb{R}^d\times\mathbb{R}^d)$ such that for a.e. $(x,v)\in\mathbb{R}^d\times\mathbb{R}^d$, $t\mapsto(X,V)(t,x,v)$ is an absolutely continuous integral solution of the characteristic system of ODE

$$\begin{cases} \dot{X}(t,x,v) = V(t,x,v), \quad X(0,x,v) = x \\ \dot{V}(t,x,v) = E(t,X(t,x,v)), \quad V(0,x,v) = v. \end{cases} \tag{7.2.5}$$

Moreover, we have the representation[5]

$$\forall t\in[0,T], \quad f(t) = (X,V)(t)_\# f_0. \tag{7.2.6}$$

Let f_1, f_2 be two weak solutions of the Vlasov-Poisson equation (7.1.1) as in Theorem 7.1.1, then $f_j(t) = (X_j(t), V_j(t))_\# f_{j0}$ for $(X_j, V_j)(t,x,v)$ the solutions to the characteristic equations (7.2.5) associated to E_j.

Remark 7.2.3. *In fact, under the assumptions of Theorem 7.1.1, the characteristic flows are Hölder continuous as functions of (x,v). This may be deduced from a similar Grönwall type estimate to the proof of Lemma 7.2.2 below using that E_i satisfy a $\log^2$-Lipschitz bound of the form (7.1.10). This will not be needed for the proof of Theorem 7.1.1.*

Given a coupling $\pi_0 \in \Pi(f_{10}, f_{20})$ (defined in Definition 7.1.4) we define the following quantities:

$$\begin{aligned} \mathcal{X}(t) = \mathcal{X}_{\pi_0}(t) &= \int_{\mathbb{R}^{2d}\times\mathbb{R}^{2d}} |X_1(t,x,v) - X_2(t,y,w)|\, d\pi_0(x,v,y,w), \\ \mathcal{V}(t) = \mathcal{V}_{\pi_0}(t) &= \int_{\mathbb{R}^{2d}\times\mathbb{R}^{2d}} |V_1(t,x,v) - V_2(t,y,w)|\, d\pi_0(x,v,y,w). \end{aligned} \tag{7.2.7}$$

By (7.2.6), the measure $\pi_t = ((X_1(t),V_1(t));(X_2(t),V_2(t)))_\#\,\pi_0$ belongs to the space $\Pi(f_1(t), f_2(t))$. Therefore, by the Definition 7.1.4 of the Wasserstein dis-

[5] The $\#$ notation means that $f(t)(B) = f_0\left(((X,V)(t,\cdot,\cdot)^{-1}(B)\right)$ for all Borel set $B\subset\mathbb{R}^d$.

tance, we have

$$W_1(f_1(t), f_2(t)) \leq \inf_{\pi_0 \in \Pi(f_{10}, f_{20})} \left(\mathcal{X}_{\pi_0}(t) + \mathcal{V}_{\pi_0}(t)\right), \quad t \in [0, T]. \tag{7.2.8}$$

On the other hand, note the converse estimate:

$$\begin{aligned} &\inf_{\pi_0 \in \Pi(f_{10}, f_{20})} \left(\mathcal{X}_{\pi_0}(0) + \mathcal{V}_{\pi_0}(0)\right) \\ &\qquad \leq \inf_{\pi_0 \in \Pi(f_{10}, f_{20})} \int_{\mathbb{R}^{2d} \times \mathbb{R}^{2d}} \left(|x - y| + |v - w|\right) d\pi_0(x, v, y, w) \\ &\qquad \leq \inf_{\pi_0 \in \Pi(f_{10}, f_{20})} \int_{\mathbb{R}^{2d} \times \mathbb{R}^{2d}} \sqrt{2}|(x, v) - (y, w)| \, d\pi_0(x, v, y, w) \\ &\qquad = \sqrt{2} W_1(f_{1,0}, f_{2,0}). \end{aligned} \tag{7.2.9}$$

We emphasize that the quantity which would lead to (7.1.9) in [134] is instead

$$\left\{ \int_{\mathbb{R}^{2d} \times \mathbb{R}^{2d}} \left(|X_1(t, x, v) - X_2(t, y, w)|^2 + |V_1(t, x, v) - V_2(t, y, w)|^2 \right) d\pi_0(x, v, y, w) \right\}^{1/2},$$

since it controls the Wasserstein distance $W_2(f_1(t), f_2(t))$.

7.2.3 Proof of Theorem 7.1.1 completed

We will prove Theorem 7.1.1 proper with the following lemma which controls the which controls the distance involving the spatial characteristics, namely the quantity $\mathcal{X}(t)$.

We recall that $\beta = 1 + (1/\alpha)$. For a given $0 < A < 1/9$, we define the function $G_\alpha(t) = G_\alpha(t; A) = G_\alpha(t; A, c)$ as the solution to

$$G'_\alpha(t; A, c) = -c G_\alpha(t; A, c)^{\beta/2} \qquad G_\alpha(0; A, c) = |\log(A)| \tag{7.2.10}$$

for times $t \in [0, T^*]$ where $T^* = T^*(\alpha, A, c)$ is the maximal time such that $G_\alpha(\cdot; A, c) > \log(9)$ on $[0, T^*)$ (we set $T^* = T$ if T^* is larger than T). Note

that $G_\alpha(\cdot; A, c)$ is decreasing and is explicitly given by

$$G_\alpha(t; A, c) = \begin{cases} |\log(A)| \exp(-ct) & \text{when } \alpha = 1, \\ \left(|\log(A)|^{1/\gamma} - c\gamma^{-1}t\right)^\gamma & \text{when } \alpha \neq 1, \end{cases} \tag{7.2.11}$$

where γ is given by (7.1.1). Moreover, we have

$$T^*(\alpha, A, c) = \begin{cases} \dfrac{1}{c} \log\left(\dfrac{|\log A|}{\log 9}\right) & \text{when } \alpha = 1, \\ \dfrac{\gamma}{c}\left(|\log A|^{1/\gamma} - (\log 9)^{1/\gamma}\right) & \text{when } \alpha \neq 1. \end{cases} \tag{7.2.12}$$

Lemma 7.2.2. *Let $\pi_0 \in \Pi(f_{10}, f_{20})$ and (7.1.8) hold. Assume that*

$$0 \leq A := (1 + T)(\mathcal{X}(0) + \mathcal{V}(0)) < \frac{1}{9}.$$

Then the following estimate holds

$$\mathcal{X}(t) \leq \exp(-G_\alpha(t; A(t), c)), \quad t \leq T^*(\alpha, A, c),$$

where

$$A(t) = (1 + t)(\mathcal{X}(0) + \mathcal{V}(0)) \leq A$$

and where $c > 0$ is a constant depending only on α and the norms in (7.1.8).

Proof. By integrating the characteristic ODEs (7.2.5) twice we have

$$\begin{aligned} &X_1(t, x, v) - X_2(t, y, w) \\ &\quad = x - y + (v - w)t \\ &\qquad + \int_0^t \int_0^s \left[E_1(\tau, X_1(\tau, x, v)) - E_2(\tau, X_2(\tau, y, w))\right] d\tau ds. \end{aligned} \tag{7.2.13}$$

Since $f_j(t) = (X_j(t), V_j(t))\# f_{j0}$, we can evaluate the fields E_j as follows, where

we omit the τ dependence for brevity:

$$
\begin{aligned}
&E_1(X_1(x,v)) - E_2(X_2(y,w)) \\
&\quad = \int_{\mathbb{R}^{2d}} K(X_1(x,v) - X_1(x_0,v_0)) f_{10}(x_0,v_0)\, dx_0 dv_0 \\
&\qquad - \int_{\mathbb{R}^{2d}} K(X_2(x,v) - X_2(y_0,w_0)) f_{20}(y_0,w_0)\, dy_0 dw_0 \\
&\quad = \int_{\mathbb{R}^{2d}\times\mathbb{R}^{2d}} [K(X_1(x,v) - X_1(x_0,v_0)) - K(X_2(x,v) - X_2(y_0,w_0))]\, d\pi_0(x_0,v_0,y_0,w_0) \\
&\quad = \int_{\mathbb{R}^{4d}} [K(X_1(x,v) - X_1(x_0,v_0)) - K(X_2(x,v) - X_1(x_0,v_0))]\, d\pi_0(x_0,v_0,y_0,w_0) \\
&\qquad + \int_{\mathbb{R}^{4d}} [K(X_2(x,v) - X_1(x_0,v_0)) - K(X_2(x,v) - X_2(y_0,w_0))]\, d\pi_0(x_0,v_0,y_0,w_0) \\
&\quad = \int_{\mathbb{R}^{d}} [K(X_1(x,v) - z) - K(X_2(x,v) - z)] \rho_1(z)\, dz \\
&\qquad + \int_{\mathbb{R}^{4d}} [K(X_2(x,v) - X_1(x_0,v_0)) - K(X_2(x,v) - X_2(y_0,w_0))]\, d\pi_0(x_0,v_0,y_0,w_0).
\end{aligned}
$$

Thus, by applying Lemma 7.2.1 we obtain the estimate

$$
\begin{aligned}
&|E_1(X_1(x,v)) - E_2(X_2(y,w))| \le C\psi_\alpha(|X_1(x,v) - X_2(x,v)|) \\
&\qquad + \int_{\mathbb{R}^{4d}} |K(X_2(x,v) - X_1(x_0,v_0)) - K(X_2(x,v) - X_2(y_0,w_0))|\, d\pi_0(x_0,v_0,y_0,w_0).
\end{aligned}
$$

It follows that

$$
\begin{aligned}
&\int_{\mathbb{R}^{4d}} |E_1(X_1(x,v)) - E_2(X_2(y,w))| d\pi_0(x,v,y,w) \\
&\le C \int_{\mathbb{R}^{4d}} \psi_\alpha(|X_1(x,v) - X_2(x,v)|)\, d\pi_0(x,v,y,w) \\
&+ \int_{\mathbb{R}^{4d}} d\pi_0(x,v,y,w) \\
&\left(\int_{\mathbb{R}^{4d}} |K(X_2(x,v) - X_1(x_0,v_0)) - K(X_2(x,v) - X_2(y_0,w_0))|\, d\pi_0(x_0,v_0,y_0,w_0) \right) \\
&\le C \int_{\mathbb{R}^{4d}} \psi_\alpha(|X_1(x,v) - X_2(x,v)|)\, d\pi_0(x,v,y,w) \\
&+ \int_{\mathbb{R}^{4d}} \left(\int_{\mathbb{R}^{4d}} |K(z - X_1(x_0,v_0)) - K(z - X_2(y_0,w_0))| \rho_2(\tau,z)\, dz \right) d\pi_0(x_0,v_0,y_0,w_0)
\end{aligned}
$$

where we have exchanged the order of integration with $d\pi_0(x_0,v_0,y_0,w_0)$ and used that $f_1(\tau) = (X_1(\tau), V_1(\tau))\# f_{10}$ in the last inequality. Therefore, by integrating

(7.2.13) against the measure $d\pi_0(x,v,y,w)$ we obtain

$$\begin{aligned}\mathcal{X}(t) &\le \int_{\mathbb{R}^{4d}} |x-y+t(v-w)|\, d\pi_0(x,v,y,w)\\ &\quad + \int_0^t \int_0^s \int_{\mathbb{R}^{4d}} |E_1(\tau, X_1(\tau,x,v)) - E_2(\tau, X_2(\tau,y,w))| d\pi_0(x,v,y,w)\, d\tau\, ds\\ &\le [\mathcal{X}(0) + t\mathcal{V}(0)]\\ &\quad + 2C \int_0^t \int_0^s \int_{\mathbb{R}^{4d}} \psi_\alpha(|X_1(\tau,x,v) - X_2(\tau,x,v)|)\, d\pi_0(x,v,y,w)\, d\tau\, ds\end{aligned}$$

where we have applied Lemma 7.2.1 (noting Remark 7.2.1 if $\alpha = \infty$) to find the second inequality.

Using that ψ_α is concave we deduce that

$$\mathcal{X}(t) \le (\mathcal{X}(0) + \mathcal{V}(0)(1+t) + C_0 \int_0^t \int_0^s \psi_\alpha(\mathcal{X}(\tau))\, d\tau\, ds$$

for a constant C_0 depending only on α and the norms in (7.1.8). For a constant c to be determined later on, let $T^*(\alpha, A, c)$ be the corresponding time defined by (7.2.12). Let $t_0 \in [0, T^*(\alpha, A, c)]$ be fixed and set

$$\mathcal{F}(t) = (\mathcal{X}(0) + \mathcal{V}(0)(1+t_0) + C_0 \int_0^t \int_0^s \psi_\alpha(\mathcal{X}(\tau))\, d\tau\, ds \ge \mathcal{X}(t), \quad t \in [0, t_0].$$

Define $\varphi_\alpha(t) = \int_0^t \psi_\alpha(s)\, ds$ and note that $\varphi_\alpha(\tau) \le C\tau^2 |\log(\tau)|^\beta$ for $\tau \le 1/9$ and $\varphi(\tau) \le C\tau$ for $\tau \ge 1/9$. Then it holds that

$$\mathcal{F}(0) = (\mathcal{X}(0) + \mathcal{V}(0)(1+t_0), \quad \mathcal{F}'(0) = 0, \quad \mathcal{F}'(t) \ge 0,$$

and

$$\mathcal{F}''(t) = C_0 \psi_\alpha(\mathcal{X}(t)) \le C_0 \psi_\alpha(\mathcal{F}(t)) = C_0 \varphi_\alpha'(\mathcal{F}(t)).$$

Thus

$$[(\mathcal{F}'(t))^2]' = 2\mathcal{F}''(t)\mathcal{F}'(t) \le 2C_0 \varphi_\alpha'(\mathcal{F}(t))\mathcal{F}'(t) = 2C_0[\varphi_\alpha(\mathcal{F}(t))]'$$

and by integrating we deduce that

$$\mathcal{F}'(t) \le \sqrt{2C_0\, \varphi_\alpha(\mathcal{F}(t))}, \quad \text{for } t \le t_0,$$

which by definition of φ_α implies

$$\mathcal{F}'(t) \le C_1\, \mathcal{F}(t) |\log \mathcal{F}(t)|^{\beta/2}, \quad \text{for } \mathcal{F}(t) \le \frac{1}{9}, \tag{7.2.14}$$

where C_1 depends only on α and the norms in (7.1.8). Now let $y(t)$ be the solution to

$$y'(t) = C_1 y(t) |\log y(t)|^{\beta/2}, \quad y(0) = \mathcal{F}(0) = A(t_0),$$

for $t \in [0, T^*(\alpha, A, c)]$. In view of the definition (7.2.12), since $A(t_0) \le A$ we have $T^*(\alpha, A(t_0), C_1) \ge T^*(\alpha, A, C_1)$ so that $y \le 1/9$ on $[0, T^*(\alpha, A, C_1)]$. Then (7.2.14) obeys $\mathcal{F}(t) \le y(t) \le 1/9$ on its domain of definition. By applying the change of variables $y = e^{-G}$ we deduce that $G' = -C_1 G^{\beta/2}$ and that therefore

$$\mathcal{F}(t) \le y(t) = \exp(-G_\alpha(t, A(t_0), C_1)), \quad t \in [0, t_0],$$

and as t_0 was arbitrary the proof of the lemma is complete by setting $c = C_1$. □

Using this lemma we are now able to prove the main result Theorem 7.1.1.

Proof of Theorem 7.1.1. By integrating the characteristic equation (7.2.5) once we obtain

$$V_1(t, x, v) - V_2(t, y, w) = v - w + \int_0^t \left[E_1(s, X_1(s, x, v)) - E_2(s, X_2(s, y, w))\right] ds. \tag{7.2.15}$$

Letting $\pi_0 \in \Pi(f_{10}, f_{20})$ be arbitrary, in the same way as in the proof of Lemma 7.2.2 we find that

$$\mathcal{V}(t) \le \mathcal{V}(0) + C \int_0^t \psi_\alpha(\mathcal{X}(s))\, ds.$$

By (7.2.9), we may consider only couplings π_0 such that $\mathcal{X}(0)+\mathcal{V}(0) \le 2W_1(f_1(0), f_2(0))$ and, therefore, by assumption on $W_1(f_1(0), f_2(0))$ in Theorem 5.1.1

$$A := (\mathcal{X}(0) + \mathcal{V}(0))(1 + T) < \frac{1}{9}.$$

So we also have $\mathcal{X}(t) \le \exp(-G_\alpha(t, A(t), c))$ with $A(t) = (\mathcal{X}(0) + \mathcal{V}(0))(1 + t)$ by Lemma 7.2.2 and for $t \le T^*(\alpha, A, c)$. Note that by definition (7.2.12) of the time

T^*, since $W_1(f_1(0), f_2(0))(1+T) \le A$ we have

$$T^*(\alpha, A, c) \ge T^* := T^*(\alpha, W_1(f_1(0), f_2(0))(1+T), c).$$

Thus all the subsequent estimates hold for $t \in [0, T^*]$.

Thus, since ψ_α is an increasing function, we obtain, dropping the α, A and c in G for brevity,

$$\begin{aligned}\int_0^t \psi_\alpha(\mathcal{X}(s))\,ds &\le \int_0^t \psi_\alpha(\exp(-G(s)))\,ds = \int_0^t \exp(-G(s))G(s)^\beta\,ds \\ &\le \int_0^t \exp(-G(t))G(0)^\beta\,ds = t\exp(-G(t))G(0)^\beta \\ &= t\exp(-G(t))|\log A(0)|^\beta,\end{aligned}$$

where we have used that $s \mapsto G(s) = G_\alpha(s, A(s), c)$ is a decreasing function of s.

Combining the estimates for $\mathcal{X}$ and $\mathcal{V}$ we have

$$\begin{aligned}&\mathcal{X}(t) + \mathcal{V}(t) \\ &\quad\le \exp(-G_\alpha(t; A(t), c)) + \mathcal{V}(0) + Ct\exp(-G_\alpha(t; A(t), c))|\log A(0)|^\beta \\ &\quad\le \exp(-G_\alpha(t; A(t), c)) + A(t) + Ct\exp(-G_\alpha(t; A(t), c))|\log A(0)|^\beta \\ &\quad= \exp(-G_\alpha(t; A(t), c)) + \exp(-G_\alpha(0, A(t), c)) \\ &\qquad + Ct\exp(-G_\alpha(t; A(t), c))|\log A(0)|^\beta \\ &\quad\le 2\exp(-G_\alpha(t; A(t), c)) \\ &\quad+ Ct\exp(-G_\alpha(t; A(t), c))|\log A(0)|^\beta\end{aligned}$$

where we have used that $t \mapsto G_\alpha(t, \overline{A}, c)$ is decreasing for fixed $\overline{A}$ in the last line. Thus for $t \in [0, T^*]$ we obtain

$$\mathcal{X}(t) + \mathcal{V}(t) \le \exp(-G_\alpha(t; A(t), c))(2 + C\,t|\log A(0)|^\beta).$$

We set $B = W_1(f_1(0), f_2(0))$, so that $B \le A(0) \le 2B$. By taking the infimum over couplings π_0 (recall (7.2.8) and (7.2.9)) we obtain

$$W_1(f_1(t), f_2(t)) \le \exp(-G_\alpha(t; 2(1+t)B))(2 + C\,t|\log B|^\beta).$$

Now suppose $\alpha = 1$, then by the explicit formula (7.2.11) we have (recalling that $\beta = 2$ in this case)

$$\begin{aligned} W_1(f_1(t), f_2(t)) &\leq \exp(\log(2(1+t)B)\exp(-ct))(2 + C\,t\,|\log B|^2) \\ &\leq (2(1+t)B)^{\exp(-ct)}(2 + C\,t|\log B|^2) \\ &\leq CB^{\exp(-ct)}(1 + t|\log B|^2) \end{aligned}$$

where we have used that $(2(1+t))^{\exp(-ct)}$ is bounded by a constant uniformly over $t \in [0, \infty)$.

Suppose instead that $\alpha \neq 1$, then

$$\begin{aligned} W_1(f_1(t), f_2(t)) &\leq \exp(-G_\alpha(t; A(t), c))(2 + C\,t|\log B|^\beta) \\ &\leq \exp(\gamma^{-1}\log(2B(1+t)) + Ct^\gamma)(2 + C\,t|\log B|^\beta) \\ &\leq (2B(1+t))^{1/\gamma}\exp(Ct^\gamma)(2 + C\,t|\log B|^\beta) \\ &\leq CB^{1/\gamma}\exp(Ct^\gamma)(1 + t|\log B|^\beta) \end{aligned}$$

where on the last line we have used that $e^{Ct^\gamma}t^\delta \leq Ce^{C't^\gamma}$ for a larger constant $C' > C$, and on the second line we have used the lower bound

$$G_\alpha(t; A, c) \geq \gamma^{-1}\log(1/A) - Ct^\gamma \tag{7.2.16}$$

for $\alpha \neq 1$, which we will now prove. Indeed, from (7.2.11) we use convexity (noting $\gamma \geq 2$) to obtain

$$G(t) \geq G(0) - tG'(0) = G(0) - CtG(0)^{\beta/2},$$

and the desired bound follows from an application of Young's inequality, i.e.

$$G(0)^{\beta/2}t \leq \gamma^{-1}t^\gamma + (\beta/2)G(0).$$

Finally, we infer the lower bound for $T^* = T^*(\alpha, B(1+T), c)$ as follows: since $(1+T)^{1+\varepsilon}B \leq 1$, we have

$$|\log B(1+T)| \geq \frac{\varepsilon}{1+\varepsilon}|\log B|.$$

So in view of (7.2.12), we have for $\alpha = 1$

$$T^* \geq \frac{1}{c}\left(\log|\log B| + \log\left(\frac{\varepsilon}{1+\varepsilon}\right) - \log\log 9\right) \geq C'\log|\log B| - C'^{-1},$$

and for $\alpha > 1$

$$T^* \geq \frac{\gamma}{c}\left(\left(\frac{\varepsilon}{1+\varepsilon}\right)^{1/\gamma}|\log B|^{1/\gamma} - (\log 9)^{1/\gamma}\right) \geq C'\gamma|\log B|^{1/\gamma} - C'^{-1}$$

for sufficiently large constant C'. □

7.2.4 Proof of Theorem 7.1.2

To prove Theorem 7.1.2 we note that we have the following result, analogous to Lemma 7.2.1. As its proof is immediate we omit it.

Lemma 7.2.3. *Let K be bounded and satisfy (7.1.10), then for any $\mu \in \mathcal{M}_+(\mathbb{R}^d)$ with mass m, we have the inequality*

$$\int_{\mathbb{R}^d} |K(x-z) - K(y-z)|d\mu(z) \leq Cm\psi_1(|x-y|)$$

where C is the constant in (7.1.10) and ψ_1 is defined by (7.2.4).

Furthermore, we note that due to this lemma the vector fields E_i are $\log^2$-Lipschitz, and as noted in Remark 7.2.3 this is enough to define the characteristic ODEs. The proof of Theorem 7.1.2 is now entirely analogous to the proof of Theorem 7.1.1 for $\alpha = 1$, replacing Lemma 7.2.1 with this lemma. Thus we leave it to the reader.

7.3 Proof of Proposition 7.1.1

We first show that it is sufficient to propagate the exponential velocity moment.

Lemma 7.3.1. *Let $f \in L^\infty(\mathbb{R}^d \times \mathbb{R}^d)$ and $c > 0$, then*

$$\int \exp\left(\left|\int_{\mathbb{R}^d} f(x,v)\,dv\right|^\alpha / \lambda\right)\,dx \le C \int_{\mathbb{R}^d \times \mathbb{R}^d} f(x,v) e^{c\langle v\rangle^{d\alpha}}\,dxdv$$

for constants C, λ depending only upon $\|f\|_{L^\infty}$ and c.

Proof. We apply the usual 'interpolation' method: let

$$M(x) = \int_{\mathbb{R}^d} f(x,v) e^{c\langle v\rangle^{d\alpha}}\,dv,$$

then for each x we have

$$\rho(x) := \int_{\mathbb{R}^d} f(x,v)\,dv \le \int_{|v|\ge R} f(x,v)\,dv + C\,\|f\|_{L^\infty} R^d \le e^{-cR^{d\alpha}} M(x) + CR^d,$$

by Markov's inequality. We now choose $R = R(x) = c^{-1/(\alpha d)} \log(1 \vee M(x))^{1/(d\alpha)}$ which gives

$$\rho(x) \le 1 + C \log(1 \vee M(x))^{1/\alpha}.$$

Thus,

$$\begin{aligned}\int \exp(|\rho(x)|^\alpha/\lambda)\,dx &\le \int \exp(|1 + C\log(1 \vee M(x))^{1/\alpha}|^\alpha/\lambda)\,dx \\ &\le \int \exp((C/\lambda)(1 + \log(1 \vee M(x))))\,dx\end{aligned}$$

and choosing $\lambda = C$ we have

$$\int \exp(|\rho(x)|^\alpha/\lambda)\,dx \le C \int M(x) \vee 1\,dx \le C \int f(x,v) e^{c\langle v\rangle^{d\alpha}}\,dxdv. \qquad \square$$

We now prove that the exponential moment is propagated. Since f_0 has finite velocity moments of order larger than $d^2 - d$, the solution provided by [133, Theo. 1] has bounded velocity moments of order larger than $d^2 - d$ on $[0,T]$. By [133, Cor. 2] it follows that

$$\sup_{t\in[0,T]} \|E(t)\|_{L^\infty} \le C < \infty \tag{7.3.1}$$

for any finite T.

Lemma 7.3.2. *Let $f \in L^\infty(\mathbb{R}^d \times \mathbb{R}^d)$, $\alpha \geq 1$ and $c > 0$. Define*

$$M(t) = \int f(t,x,v) e^{1+c\langle v\rangle^{d\alpha}}\, dxdv.$$

Then we have the differential inequality along the Vlasov-Poisson flow

$$\frac{dM}{dt} \leq C(1 + \log(M(t))^{1-\frac{1}{d\alpha}} M(t))$$

for a constant C depending only upon c, α and $\|f\|_{L^\infty}$.

Proof. We directly compute, using the weak formulation of the Vlasov-Poisson equation

$$\begin{aligned}
\frac{dM}{dt} &= -\int_{\mathbb{R}^d\times\mathbb{R}^d} cE(t,x)\cdot\nabla_v[\langle v\rangle^{d\alpha}] f(t,x,v) e^{1+c\langle v\rangle^{d\alpha}}\, dxdv \\
&\leq C\int_{\mathbb{R}^d\times\mathbb{R}^d} |E(t,x)|\,\langle v\rangle^{d\alpha-1} f(t,x,v) e^{1+c\langle v\rangle^{d\alpha}}\, dxdv \\
&\leq C\,\|E(t)\|_{L^\infty}\int_{\mathbb{R}^d\times\mathbb{R}^d} \log(e^{1+c\langle v\rangle^{d\alpha}})^{1-\frac{1}{d\alpha}} e^{1+c\langle v\rangle^{d\alpha}} f(t,x,v)\, dxdv.
\end{aligned}$$

The claim of the lemma now follows from (7.3.1) and Jensen's inequality, using the convexity of $\tau \mapsto \tau\log(\tau)^{1-\frac{1}{d\alpha}}$ on $\tau \in (e,\infty)$. □

Proof of Proposition 7.1.1. By using Lemma 7.3.2 and solving the resulting differential inequality we deduce that

$$\sup_{t\in[0,T]}\int_{\mathbb{R}^d\times\mathbb{R}^d} f(t,x,v) e^{c\langle v\rangle^{d\alpha}}\, dxdv \leq C < \infty.$$

The claim of the proposition now follows from an application of Lemma 7.3.1. □

Chapter 8

Contraction in the Wasserstein metric for the kinetic Fokker-Planck equation on the torus

We study contraction for the kinetic Fokker-Planck operator on the torus. Solving the stochastic differential equation, we show contraction and therefore exponential convergence in the Monge-Kantorovich-Wasserstein $\mathcal{W}_2$ distance. Finally, we investigate if such a coupling can be obtained by a co-adapted coupling, and show that then the bound must depend on the square root of the initial distance.

Acknowledgements

This work in this chapter was done in collaboration with Josephine Evans and Helge Dietert and appears in a similar form in [46]. We would like to thank Clément Mouhot for the initial discussion to look into the problem, and José A. Carrillo for discussion relating to the (lack of) a gradient flow representation of the kinetic Fokker-Planck equation.

8.1 Introduction

The kinetic Fokker-Planck equation, also known as the Kramers equation, is a basic model for the spreading of a solute due to interaction with the fluid background. It is derived from Langevin dynamics, where the time scale of observation is much larger than the correlation time of the solute-fluid interactions (see e.g. [207]).

We prove contraction properties of the spatially periodic kinetic Fokker-Planck equation in the Wasserstein metric, and show to what extent the probabilistic technique of coupling can be used in such situations. This is of interest, both intrinsically, and in the broader context of analytic and probabilistic methods of proving convergence to equilibrium and contraction properties of Fokker-Planck equations which we summarise in the paragraphs below. The Monge-Kantorovich-Wasserstein (MKW) distance comes from optimal transport and is defined as

$$\mathcal{W}_2(\mu,\nu) = \inf_{\pi\in\Pi_{\mu,\nu}} \left(\int |x-y|^2 \mathrm{d}\pi(x,y) \right)^{1/2},$$

where $\Pi_{\mu,\nu}$ is the set of all couplings between μ and ν.

A common analytic technique to show contraction or convergence to equilibrium of Fokker-Planck equations is to work in a L^2 space weighted by the reciprocal of the equilibrium measure. Here, in the spatially homogeneous setting, contractivity is established by showing that the generator of the Fokker-Planck semi-group is *coercive* on this L^2 space, which implies that the generator has a spectral gap. In the spatially inhomogeneous setting, which is common in kinetic theory, the generator is, however, not *coercive* and this method fails.

This lead Villani to develop the celebrated theory of *hypocoercivity* [197] where a spectral gap and contraction of the semi-group are shown, roughly speaking, by constructing equivalent 'skew' L^2 or Sobolev norms on which the generator is coercive. This theory is well developed, and applies to a large class of SDEs [197] and also to collisional models [85]. The kinetic Fokker-Planck equation in particular has received much attention [86, 69, 149] both in the case of a spatial confining potential and in, the analytically simpler, case of spatial periodicity. These works, however, do not address the question of convergence or contraction

in the Wasserstein metric $\mathcal{W}_2$, as this distance is 'inaccessible' from these analytic tools; the closest result to this being [145] where $\mathcal{W}_1$ results are obtained by duality.

A second analytic approach to the study of the Fokker-Planck equation is the theory of gradient flows [105], in which the Fokker-Planck equation is identified with the steepest descent flow of an entropy functional in the Wasserstein space $\mathcal{W}_2$. A similar type of degeneracy, to those dealt with in hypocoercivity, causes this theory fails for the spatially inhomogeneous Fokker-Plank equation in which we are interested. Dissipation in the Wasserstein distance can also be shown for non-gradient drifts in the homogeneous setting using analytic methods [24].

A common probabilistic technique to show contraction or convergence is to construct a *coupling* between two copies of the stochastic process that realises the desired bound on the metric between the laws. In the spatially homogeneous Fokker-Planck equation, the *synchronisation* coupling, where the infinitesimal motions of the noise are coupled together, gives contraction in Wasserstein metrics when the velocity potential is strongly convex. In the spatially inhomogeneous case with a confining potential, such a straightforward coupling only establishes contraction if the confining potentials are quadratic (or a small perturbation thereof) see for example [25]. Establishing contraction in the Wasserstein metric for more general confining potentials is an open problem. In the spatially periodic case results are even more limited. In this case the synchronisation coupling does not cause the spatial distance on the torus to decay. That the spatially periodic case is more difficult in the probabilistic case is in contrast to the analytic setting, where having the spatial variable on the torus makes many computations simpler.

In this work we study the contraction properties in the Wasserstein metric of the kinetic Fokker-Planck equation with spatial variable on the torus and a quadratic velocity potential. Despite the simplicity of this equation, to the authors' knowledge this question has not been answered in the literature, and a second goal of this chapter it to understand what difficulties might explain this.

This kinetic Fokker-Planck equation describes the law of a particle moving in the phase space $\mathbb{T} \times \mathbb{R}$ whose location in the phase space is (X_t, V_t) and evolves as

$$\begin{cases} \mathrm{d}X_t = V_t \mathrm{d}t, \\ \mathrm{d}V_t = -\lambda V_t \mathrm{d}t + \mathrm{d}W_t, \end{cases} \tag{8.1.1}$$

where $\mathrm{d}W_t$ is a standard white noise and the spacial variable is in the torus $\mathbb{T} = \mathbb{R}/(2\pi L\mathbb{Z})$ of length $2\pi L$.

The corresponding measure μ_t on $\mathbb{T} \times \mathbb{R}$ evolves as

$$\partial_t \mu_t + v\partial_x \mu_t = \partial_v[\lambda v \mu_t + \frac{1}{2}\partial_v \mu_t], \tag{8.1.2}$$

where this equation is considered in the weak sense.

Solving the stochastic evolution, we are show exponential decay of the distance between two solutions.

Theorem 8.1.1. *If μ_t and ν_t are two solutions to the kinetic Fokker-Planck equation (8.1.2), then we have*

$$\mathcal{W}_2(\mu_t, \nu_t) \leq \left(e^{-\lambda t} + c\, e^{-t/4\lambda^2 L^2}\right) \mathcal{W}_2(\mu_0, \nu_0)$$

for a constant c only depending on L.

The key idea is that, after fixing the net effect of the velocity noise, the spatial variable has enough randomness left to allow such a coupling. This approach is not based on a functional inequality which is integrated over time and in fact the evolution is not a contraction semigroup. We can show the lack of coercivity directly in a straightforward way using the explicit solution to the SDE. Precisely,

Proposition 8.1.1. *The kinetic Fokker-Planck operator is not coercive in the Wasserstein-2 distance. i.e. there is not $\gamma > 0$ such that*

$$\mathcal{W}_2(\mu_t, \nu_t) \leq e^{-\gamma t} \mathcal{W}_2(\mu_0, \nu_0).$$

In order to construct a coupling showing convergence in the MKW distance, random variables (X_t^i, V_t^i) are constructed for $t \in \mathbb{R}^+$ and $i = 1, 2$ such that (X_t^1, V_t^1) has law μ_t and (X_t^2, V_t^2) has law ν_t. Then for $t \in \mathbb{R}^+$ the coupling

$((X_t^1, V_t^1), (X_t^2, V_t^2))$ gives an upper bound of the MKW distance $\mathcal{W}_2(\mu_t, \nu_t)$.

The stochastic differential equation (8.1.1) motivates us to look at couplings where (X_t^i, V_t^i) are continuous Markov processes with initial distribution μ_0 and ν_0, respectively, and whose transition semigroup is determined by (8.1.1). For such couplings we can consider a more restrictive class of couplings.

Definition 8.1.1 (co-adapted coupling)**.** *The coupling $((X_t^1, V_t^1), (X_t^2, V_t^2))$ is co-adapted if, for $i = 1, 2$, under the filtration $\mathcal{F}$ generated by the coupling $((X_t^1, V_t^1), (X_t^2, V_t^2))$, the process (X_t^i, V_t^i) is a continuous Markov process whose transition semigroup is determined by* (8.1.1)*.*

This is an important subclass of couplings, which contains many natural couplings, and an even more restrictive subclass is the class of Markovian couplings, where additionally the coupling itself is imposed to be Markovian. The existence and obtainable convergence behaviour under this restriction has already been studied in different cases, e.g. [117, 31, 37]. Note that the co-adapted coupling is equivalent to the condition that the filtration generated by (X_t^i, V_t^i) is immersed in the filtration generated by the coupling, which motivates Kendall [111] to call such couplings *immersed couplings*.

By adapting the reflection/synchronisation coupling, we can still obtain exponential convergence but with a loss in dependence on the initial data.

Theorem 8.1.2. *Given initial distributions μ_0 and ν_0, there exists a co-adapted coupling $((X_t^1, V_t^1), (X_t^2, V_t^2))$ such that*

$$\begin{aligned}\mathcal{W}_2(\mu_t, \nu_t) &\leq \left(\mathbb{E}\left[|X_t^1 - X_t^2|_{\mathbb{T}}^2 + (V_t^1 - V_t^2)^2\right]\right)^{1/2} \\ &\leq C\zeta(t)(\sqrt{\mathcal{W}_2(\mu_0, \nu_0)} + \mathcal{W}_2(\mu_0, \nu_0)),\end{aligned}$$

where

$$\zeta(t) = \begin{cases} e^{-\min(2\lambda, 1/(2\lambda^2 L^2))t} & 4L^2\lambda^3 \neq 1 \\ e^{-2\lambda t}(1+t) & 4L^2\lambda^3 = 1 \end{cases}$$

and C is a constant that depends only on λ and L.

Here we used the notation $|X_t^1 - X_t^2|_{\mathbb{T}}$ to emphasis that this is the distance on the torus $\mathbb{T}$. In fact the filtrations generated by (X^1, V^1) and (X^2, V^2) agree which

Kendall [111] calls an equi-filtration coupling.

Remark 8.1.1. *This theorem achieves the same exponential decay rate as the non-Markovian argument, except for the case $4L^2\lambda^3 = 1$, when the spatial and velocity decay rates coincide and we have an addition polynomial factor.*

In general the loss in the dependence is necessary.

Theorem 8.1.3. *Suppose there exists a function $\alpha : \mathbb{R}^+ \mapsto \mathbb{R}^+$ and a constant $\gamma > 0$ such that for all initial distributions μ_0 and ν_0 there exists a co-adapted coupling $((X_t^1, V_t^1), (X_t^2, V_t^2))$ such that*

$$\left(\mathbb{E}\left[|X_t^1 - X_t^2|_{\mathbb{T}}^2 + (V_t^1 - V_t^2)^2\right]\right)^{1/2} \leq \alpha(\mathcal{W}_2(\mu_0, \nu_0))e^{-\gamma t}.$$

Then there exists a constant C such that for $z \in (0, \pi L]$ we have the following lower bound on the dependence on the initial distance

$$\alpha(z) \geq C\sqrt{z}.$$

The idea is to focus on a drift-corrected position on the torus, which evolves as a Brownian motion. By stopping the Brownian motion at a large distance we can then prove the claimed lower bound.

This shows that a simple hypocoercivity argument on a Markovian coupling cannot work. Precisely, there cannot exist a semigroup P on the probability measures over $(\mathbb{T} \times \mathbb{R})^{\times 2}$, whose marginals behave like the solution of (8.1.1) and which satisfies $H(P_t(\pi)) \leq cH(\pi)e^{-\gamma t}$ for $H^2(\pi) = \int[(X^1 - X^2)^2 + (V^1 - V^2)^2]\mathrm{d}\pi(X^1, V^1, X^2, V^2)$. Otherwise, the Markov process associated to P would be a coupling contradicting Theorem 8.1.3.

Addendum: Subsequent to the preparation of the manuscript [46] we attempted to generalise to non-quadratic potentials, i.e. to the SDE

$$\begin{cases} dX_t = V_t, \\ dV_t = -U'(V_t)dt + dW_t, \end{cases}$$

where $U : \mathbb{R} \to \mathbb{R}$ is C^2 and strongly convex with derivative U', and (X, V) evolves

on $\mathbb{T} \times \mathbb{R}$. This was mostly a failure[1]. The explicit computations in the proof of Theorem 8.1.1 fail to work for non-quadratic potentials U. The construction used in Theorem 8.1.2 also fails and led me to question whether any such coupling at all can converge in $\mathcal{W}_2$[2]. However, surprisingly the optimality result Theorem 8.1.3 can be proved in great generality, which we present below:

Theorem 8.1.4. *Let $(V^1)_{t\in\mathbb{R}_+}, (V^2)_{t\in\mathbb{R}_+}$ be real valued stochastic processes with the same individual laws. Let $X_0^1, X_0^2 \in [0, 2\pi)$ be deterministic constants, and define the real valued stochastic processes $(X_t^1)_{t\in\mathbb{R}_+}$,$(X_t^2)_{t\in\mathbb{R}_+}$, by*

$$X_t^i := X_0^i + \int_0^t V_s^i \, ds \qquad i = 1, 2.$$

Let $\gamma > 0$ be fixed such that

$$\sqrt{\mathbb{E}\left(|V_t^1 - V_t^2|^2 + |X_t^1 - X_2^t|_{\mathbb{T}}^2\right)} \leq ce^{-\gamma t} \tag{8.1.3}$$

for some constant c. Then there is an absolute constant C depending only on γ (and not on the choice of V^1, V^2, X_0^1, X_0^2) such that

$$c \geq C\sqrt{|X_0^1 - X_0^2|_{\mathbb{T}}}.$$

Remark 8.1.2. *We do not require any assumptions on the processes V^i. They may be arbitrarily dependent on each other. They need not be Markov (or in particular jointly Markov), and need not be adapted to any filtration (or in particular the same filtration). This is due to the technique of the proof, which is analytic in nature and does not use martingales (unlike Theorem 8.1.1).*

Remark 8.1.3. *The laws of V^1, V^2 are arbitrary, so long as they are equal in distribution. This means that the proposition applies to many kinetic equations other than just the kinetic Fokker-Planck equation. These include collisional models such as the linear BGK equation and linear Boltzmann equations (so long as the collision/scattering kernels do not depend on the spatial position on the torus.) The theorem also applies to such models in any dimension by considering only the first coordinate pair.*

[1] Something one can admit in a thesis, but not in a paper.

[2] Any embarrassment from the subsequent exhibition of a simple convergent coupling falls exclusively on the author of this thesis and not upon any collaborators.

8.2 Set up

The stochastic differential equation (8.1.1) has the explicit solution, when posed in $\mathbb{R}^2$. For clarity, when we are considering X to be in $\mathbb{R}$ rather than the torus we will denote it $\hat{X}$. The explicit solution is

$$\begin{aligned} \hat{X}_t &= \hat{X}_0 + \frac{1}{\lambda}(1 - e^{-\lambda t})V_0 + \int_0^t \frac{1}{\lambda}(1 - e^{-\lambda(t-s)})\mathrm{d}W_s, \\ V_t &= e^{-\lambda t}V_0 + \int_0^t e^{-\lambda(t-s)}\mathrm{d}W_s, \end{aligned} \tag{8.2.1}$$

where W_t is the common Brownian motion. In this we separate the stochastic driving as (A_t, B_t) given by the stochastic integrals

$$\begin{aligned} A_t &= \int_0^t \frac{1}{\lambda}(1 - e^{-\lambda(t-s)})\mathrm{d}W_s, \\ B_t &= \int_0^t e^{-\lambda(t-s)}\mathrm{d}W_s, \end{aligned}$$

which evolve as a vector in $\mathbb{R}^2$ with the common Brownian motion W_t. By Itô's isometry (A_t, B_t) is a Gaussian random variable with covariance matrix $\Sigma(t)$ given by

$$\Sigma_{AA}(t) = \frac{1}{\lambda^2}\left[t - \frac{2}{\lambda}(1 - e^{-\lambda t}) + \frac{1}{2\lambda}(1 - e^{-2\lambda t})\right], \tag{8.2.2}$$

$$\Sigma_{AB}(t) = \frac{1}{\lambda^2}\left[(1 - e^{-\lambda t}) - \frac{1}{2}(1 - e^{-2\lambda t})\right], \tag{8.2.3}$$

$$\Sigma_{BB}(t) = \frac{1}{2\lambda}(1 - e^{-2\lambda t}). \tag{8.2.4}$$

From this we calculate that the conditional distribution of A_t given B_t is a Gaussian with variance $\Sigma_{AA}(t) - \Sigma_{AB}^2(t)\Sigma_{BB}^{-1}(t)$ and mean given by

$$\mu_{A|B}(t, b) = \Sigma_{AB}(t)\Sigma_{BB}^{-1}(t)b.$$

We write $g_{A|B}$ for the conditional density of A given B and g_B for the marginal density of B. Hence

$$g(t, a, b) = g_{A|B}(t, a, b)g_B(t, b) \tag{8.2.5}$$

is the joint density of A and B.

The last part of the set up is the change of variables we will need for the Markovian coupling. We define new coordinates (Y, V) by taking the drift away

$$\begin{cases} Y = X + \frac{1}{\lambda} V, \\ V = V. \end{cases} \tag{8.2.6}$$

The motivation for this change is the explicit formulas found in (8.2.1) from which we see that Y is the limit as $t \to \infty$ of X_t without additional noise. In the new variables, (8.1.1) becomes

$$\begin{cases} \mathrm{d}Y_t = \frac{1}{\lambda} \mathrm{d}W_t, \\ \mathrm{d}V_t = -\lambda V_t \mathrm{d}t + \mathrm{d}W_t, \end{cases}$$

for the common Brownian motion W_t. Note that the motion of Y_t does not depend explicitly upon V_t and is a Brownian motion on the torus.

It remains to show that these new coordinates define an equivalent norm on $\mathbb{T} \times \mathbb{R}$. This follows from the triangle inequality and we have

$$|X^1 - X^2|_{\mathbb{T}} + |V^1 - V^2| \leq |Y^1 - Y^2|_{\mathbb{T}} + \left(1 + \frac{1}{\lambda}\right) |V^1 - V^2|$$

and the other direction is similar. Thus, the two norms are equivalent up to a constant factor that depends only on λ.

8.3 Non-Markovian Coupling

We wish to estimate how much the spatial variable will spread out over time. We will then use this to construct a coupling at a fixed time t which exploits the fact that a proportion of the spatial density is distributed uniformly. In order to do this we give a lemma on the spreading of a Gaussian density wrapped on the torus.

Lemma 8.3.1. *For $\sigma^2 > 2L^2 \log(3)$ consider the Gaussian density h on $\mathbb{R}$ given*

by

$$h(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/2\sigma^2}$$

and wrap it onto the torus $\mathbb{T}$*, i.e. define the density* Qh *on* $\mathbb{T}$ *by*

$$(Qh)(x) = \sum_{n\in\mathbb{Z}} h(x + 2\pi L n). \tag{8.3.1}$$

We have the following estimate on the spatial spreading

$$Qh(x) \geq \frac{\beta}{2\pi L}$$

where

$$1 - \beta = \frac{2e^{-\sigma^2/2L^2}}{1 - e^{-\sigma^2/2L^2}} \in (0, 1).$$

Proof. We define the Fourier transform of a function on $\mathbb{T}$ to be

$$(\mathcal{F}g)(k) = \int_{\mathbb{T}} e^{ikx/L} g(x)\mathrm{d}x,$$

where

$$\int_{\mathbb{T}} g(x)\mathrm{d}x = \int_0^{2\pi L} g(x)\mathrm{d}x.$$

By the definition of Q, the Fourier transform of Qh is given by

$$\begin{aligned}(\mathcal{F}Qh)(k) &= \int_{\mathbb{T}} \sum_{n\in\mathbb{Z}} h(x + 2\pi L n) e^{ikx/L}\mathrm{d}x \\ &= \int_{\mathbb{R}} h(x) e^{ikx/L}\mathrm{d}x \\ &= \exp\left(-\frac{k^2\sigma^2}{2L^2}\right)\end{aligned}$$

where we have used the well-known Fourier transformation of a Gaussian.

By the Fourier series we find that, for any $x \in \mathbb{T}$, we have

$$Qh(x) - \frac{\beta}{2\pi L} = \frac{1}{2\pi L} \sum_{|k|\geq 1} e^{-k^2\sigma^2/2L^2 - ikx/L} + \frac{1-\beta}{2\pi L}.$$

We want this to be positive. Therefore it is sufficient to show that

$$\left| \sum_{|k|\geq 1} e^{-k^2\sigma^2/2L^2 - ikx/L} \right| \leq 1 - \beta.$$

We estimate the left hand side by

$$\left| \sum_{|k|\geq 1} e^{-k^2\sigma^2/2L^2 - ikx/L} \right| \leq 2 \sum_{k\geq 1} e^{-k\sigma^2/2L^2} = 1 - \beta$$

where the final equality follows from summing the geometric series. □

We can now use this to construct a coupling at time t. We will use this coupling to prove exponential decrease in the Wasserstein distance.

Lemma 8.3.2. *Let $t \geq 0$, be large enough so the variance of $g_{A|B}$ is greater than $2L\log(3)$, and β be such that*

$$(Qg_{A|B})(t,a,b) \geq \frac{\beta}{2\pi L},$$

where $g_{A|B}$ is defined by (8.2.5) above. Let μ_t resp. ν_t be the distribution of the solution to the Fokker-Planck equation (8.1.2) with deterministic initial data $\mu_0 = \delta_{x_0^1, v_0^1}$ and $\nu_0 = \delta_{x_0^2, v_0^2}$ respectively, at time t. Then there exists a coupling $((X_t^1, V_t^1), (X_t^2, V_t^2))$ between μ_t and ν_t satisfying

$$\mathbb{E}\left[(V_t^1 - V_t^2)^2\right] = e^{-2\lambda t}\left[(v_0^1 - v_0^2)^2\right]$$

and

$$\mathbb{E}\left[|X_t^1 - X_t^2|_{\mathbb{T}}^2\right] \leq 2(1-\beta)\left[|x_0^1 - x_0^2|_{\mathbb{T}}^2 + \frac{1}{\lambda^2}(v_0^1 - v_0^2)^2\right].$$

Proof. Let us construct such a coupling. Since we have seen that $g_{A|B}$ is Gaussian density with variance $\sigma^2 = \Sigma_{AA}(t) - \Sigma_{AB}^2(t)\Sigma_{BB}^{-1}(t)$, we can us Lemma 8.3.1 to split the distribution $Qg_{A|B}$ as

$$Qg_{A|B}(t,a,b) = \frac{\beta}{2\pi L} + (1-\beta)s(t,a,b).$$

Then by assumption s is again a probability density for the variable a on the

torus $\mathbb{T}$. We now consider the torus as a subset of $\mathbb{R}$ and then $Qg_{A|B}$ and s are probability density functions supported on $[0, 2\pi L]$. Let B be an independent random variable with density $g_B(t, b)$, let Z be an independent uniform random variable over $[0, 1]$ and let U be an independent uniform random variable over the torus. Finally let S be a random variable on $\mathbb{R}$ with density $s(t, \cdot, B)$, viewed as a density function on $\mathbb{R}$, only depending on B.

With this define the random parts A^1, A^2 of X_t^1, X_t^2 as

$$\begin{aligned} A^1 =& 1_{Z\leq\beta}\left[U - x_0^1 - \frac{1}{\lambda}(1 - e^{-\lambda t})v_0^1\right] + 1_{\beta>Z}S, \\ A^2 =& 1_{Z\leq\beta}\left[U - x_0^2 - \frac{1}{\lambda}(1 - e^{-\lambda t})v_0^2\right] + 1_{\beta>Z}S. \end{aligned}$$

We then construct $(\hat{X}_t^1, V_t^1)$ defined by

$$\begin{aligned} \hat{X}_t^1 &= x_0^1 + \frac{1}{\lambda}(1 - e^{-\lambda t})v_0^1 + A^1, \\ V_t^1 &= e^{-\lambda t}v_0^1 + B, \end{aligned}$$

and $(\hat{X}_t^2, V_t^2)$ defined by

$$\begin{aligned} \hat{X}_t^2 &= x_0^2 + \frac{1}{\lambda}(1 - e^{-\lambda t})v_0^2 + A^2, \\ V_t^2 &= e^{-\lambda t}v_0^2 + B. \end{aligned}$$

We then construct X_t^i by wrapping $\hat{X}_t^i$ onto the torus (i.e. $X_t^i \in [0, 2\pi L)$ and $X_t^i \equiv \hat{X}_t^i \mod 2\pi L$). By construction the pairs (X^i, V^i) have the right laws so they form a valid coupling.

We find

$$\mathbb{E}\left[(V_t^1 - V_t^2)^2\right] = e^{-2\lambda t}\left[(v_0^1 - v_0^2)^2\right]$$

and

$$\mathbb{E}\left[|X_t^1 - X_t^2|_{\mathbb{T}}^2\right] = (1 - \beta)\left[\left|x_0^1 - x_0^2 + \frac{1}{\lambda}(1 - e^{-\lambda t})(v_0^1 - v_0^2)\right|_{\mathbb{T}}^2\right]$$

and we can use Young's inequality to find the claimed control. □

We now put these two lemmas together to prove Theorem 8.1.1, which states exponential convergence in the MKW $\mathcal{W}_2$ distance.

Proof of Theorem 8.1.1. We first show that we can reduce to working with deterministic initial conditions. We denote $\mu_t^{x,v}$ to be the law of the solution to the SDE initialized at (x, v). Suppose we know that

$$\mathcal{W}_2(\mu_t^{x_1,v_1}, \mu_t^{x_1,v_2}) \leq \omega(t) d((x_1, v_1), (x_1, v_2)).$$

Then given any coupling π of initial measures μ_0, ν_0 we have

$$\begin{aligned}\mathcal{W}_2(\mu_t, \nu_t)^2 &\leq \int_{(\mathbb{T}\times\mathbb{R})^2} \mathcal{W}_2(\mu_t^{x_1,v_1}, \mu_t^{x_2,v_2})^2 \mathrm{d}\pi((x_1, v_1), (x_2, v_2)) \\ &\leq \omega(t)^2 \int_{(\mathbb{T}\times\mathbb{R})^2} d((x_1, v_1), (x_2, v_2))^2 \mathrm{d}\pi((x_1, v_1), (x_2, v_2)).\end{aligned}$$

Then taking an infimum over π shows that this implies

$$\mathcal{W}_2(\mu_t, \nu_t) \leq \omega(t) \mathcal{W}_2(\mu_0, \nu_0).$$

Given any initial points $((x_0^1, v_0^1), (x_0^2, v_0^2))$, we can use Lemma 8.3.2 to get a coupling $((X_t^1, V_t^1), (X_t^2, V_t^2))$ of μ_t and ν_t. From explicitly calculating the variance of the distribution of $A|B$ using (8.2.2), (8.2.3), (8.2.4), we see that the variance grows asymptotically as t/λ^2. Hence by Lemma 8.3.1 we can choose β so that $1 - \beta \to 0$ exponentially fast with rate $1/2\lambda^2 L^2$. This, combined with the control from the second lemma, shows that

$$\mathbb{E}\left[(V_t^1 - V_t^2)^2 + |X_t^1 - X_t^2|_{\mathbb{T}}^2\right] \leq \left(e^{-2\lambda t} + ce^{-t/2\lambda^2 L^2}\right)\left[(v_0^1 - v_0^2)^2 + |x_0^1 - x_0^2|_{\mathbb{T}}^2\right].$$

□

The explicit solution also allows to prove that the evolution is not a contraction semigroup.

Proof of Proposition 8.1.1. We will prove the theorem by contradiction. Suppose $\gamma > 0$ and let $a \neq b$ be two distinct points on the torus. Consider the initial measures

$$\mu_0 = \delta_{x=a}\delta_{v=0}$$

and

$$\nu_0 = \delta_{x=b}\delta_{v=0}.$$

Then the distance is $\mathcal{W}_2(\mu_0, \nu_0) = |a - b|_{\mathbb{T}}$.

At time t the spatial distribution of μ_t and ν_t, interpreted in $\mathbb{R}$, is a Gaussian with variance Σ_{AA} which by the explicit formula Equation (8.2.2) can be bounded as

$$\Sigma_{AA}(t) \leq C_A t^2$$

for a constant C_A and $t \leq 1$.

Hence for $d > 0$ and $t \leq 1$ the spatial spreading is controlled as

$$\begin{aligned}\mu_t((\mathbb{T} \setminus [a-d, a+d]) \times \mathbb{R}) &\leq \frac{2\Sigma_{AA}(t)}{d\sqrt{2\pi}} \exp\left(\frac{-d^2}{2\Sigma_{AA}^2(t)}\right) \\ &\leq C_1 \frac{t^2}{d} \exp\left(-C_2 \frac{d^2}{t^4}\right)\end{aligned}$$

for positive constants C_1 and C_2, where we have used the standard tail bound for the Gaussian distribution (see e.g. [148, Lemma 12.9]).

For any $d > 0$ small enough that $a \pm d$ and $b \pm d$ do not wrap around the torus, any coupling between μ_t and ν_t must transfer at least the mass

$$1 - \mu_t((\mathbb{T} \setminus [a-d, a+d]) \times \mathbb{R}) - \nu_t((\mathbb{T} \setminus [b-d, b+d]) \times \mathbb{R})$$

between $[a - d, a + d]$ and $[b - d, b + d]$.

Hence the Wasserstein distance is bounded by

$$\mathcal{W}_2^2(\mu_t, \nu_t) \geq (|a - b|_{\mathbb{T}} - 2d)^2 \left(1 - 2C_1 \frac{t^2}{d} \exp\left(-C_2 \frac{d^2}{t^4}\right)\right).$$

Taking $d = |a - b|_{\mathbb{T}} t^{3/2}$ for t sufficiently small, this shows that

$$\mathcal{W}_2^2(\mu_t, \nu_t) \geq |a - b|_{\mathbb{T}}^2 (1 - 2t^{3/2})^2 \left(1 - \frac{2C_1}{|a - b|_{\mathbb{T}}} \sqrt{t} \exp\left(-\frac{C_2 |a - b|_{\mathbb{T}}^2}{t}\right)\right).$$

However, for all small enough positive t, we have

$$(1 - 2t^{3/2})^2 > e^{-\gamma t/2}$$

and

$$\left(1 - \frac{2C_1}{|a-b|_{\mathbb{T}}}\sqrt{t}\exp\left(-\frac{C_2|a-b|_{\mathbb{T}}^2}{t}\right)\right) > e^{-\gamma t/2}$$

contradicting the assumed contraction. For the second estimate we use $\exp(-c/t) \leq (1+c/t)^{-1} = t/(c+t)$. □

8.4 Co-adapted couplings

8.4.1 Existence

For Theorem 8.1.2 we construct a reflection/synchronisation coupling using the drift-corrected positions Y_t^i. As the positions are on the torus we can use a reflection coupling until Y_t^1 and Y_t^2 agree. Afterwards, we use a synchronisation coupling which keeps $Y_t^1 = Y_t^2$ and reduces the velocity distance.

For a formal definition let $((X_0^1, V_0^1), (X_0^2, V_0^2))$ be a coupling between μ and ν obtaining the MKW distance (the existence of such a coupling is a standard result, see e.g. [198, Theorem 4.1.]).

We then define the evolution of this coupling in two stages. First, define (X_t^1, V_t^1) and (X_t^3, V_t^3) to be strong solutions to (8.1.1) with initial conditions $((X_0^1, V_0^1)$ and (X_0^2, V_0^2) respectively and driving Brownian motion W_t^1. Then we recall the definition of Y^i from (8.2.6), and define the stopping time $T := \inf\{t \geq 0 : Y_t^1 = Y_t^3\}$. Then we define a new process W_t^2 by

$$W_t^2 = \begin{cases} -W_t^1 & t \leq T, \\ W_t^1 - 2W_T^1 & t > T. \end{cases}$$

By the reflection principle, W^2 is a Brownian motion. We use this to define a new solution (X_t^2, V_t^2) to be the strong solution to (8.1.1) with driving Brownian motion W^2 and initial condition (X_0^2, V_0^2). Note now that $T = \inf\{t \geq 0 : Y_t^1 = Y_t^2\}$.

For the analysis we introduce the notation

$$M_t = Y_t^1 - Y_t^2,$$
$$Z_t = V_t^1 - V_t^2.$$

Then by the construction the evolution is given by

$$\mathrm{d}M_t = \frac{2}{\lambda} 1_{t \leq T} \mathrm{d}W_t^1, \tag{8.4.1}$$
$$\mathrm{d}Z_t = -\lambda Z_t \mathrm{d}t + 2 \cdot 1_{t \leq T} \mathrm{d}W_t^1, \tag{8.4.2}$$

where M_t evolves on the torus $\mathbb{T}$.

As a first step we introduce a bound for T.

Lemma 8.4.1. *The stopping time T satisfies*

$$\mathbb{P}(T > t | M_0) = \frac{4}{\pi} \sum_{k=0}^{\infty} \frac{1}{2k+1} \exp\left(-\frac{(2k+1)^2}{2\lambda^2 L^2} t \right) \sin\left(\frac{(2k+1)|M_0|_{\mathbb{T}}}{2L} \right). \tag{8.4.3}$$

Proof. As M_t evolves on the torus, T is the first exit time of a Brownian motion starting at M_0 from the interval $(0, 2\pi L)$. See [148, (7.14-7.15)], from which the claim follow after rescaling to incorporate the $2/\lambda$ factor. □

Remark 8.4.1. *The second expression in (8.4.3) is obtained by solving the heat equation on $[0, 2\pi L]$ with Dirichlet boundary conditions and initial condition δ_{M_0}.*

Lemma 8.4.2. *There exists a constant C such that for any $t > 0$ the following holds*

$$\mathbb{P}(T > t | M_0) \leq C |M_0|_{\mathbb{T}} (1 + t^{-1/2}) e^{-t/(2\lambda^2 L^2)}. \tag{8.4.4}$$

Proof. Using (8.4.3) and the inequality $\sin(x) \leq x$ for $x \geq 0$, we have

$$\begin{aligned}
\mathbb{P}(T > t | M_0) &\leq \frac{4}{\pi} e^{-t/(2\lambda^2 L^2)} \sum_{k=0}^{\infty} \frac{|M_0|_{\mathbb{T}}}{2L} \frac{2k+1}{2k+1} e^{-4k^2 t/(2\lambda^2 L^2)} \\
&\leq \frac{2}{\pi L} |M_0|_{\mathbb{T}} e^{-t/(2\lambda^2 L^2)} \left(1 + \int_0^{\infty} e^{-4u^2 t/(2\lambda^2 L^2)} \mathrm{d}u \right) \\
&= \frac{2}{\pi L} |M_0|_{\mathbb{T}} e^{-t/(2\lambda^2 L^2)} \left(1 + \sqrt{\frac{\pi}{8t/(\lambda^2 L^2)}} \right) \\
&\leq C |M_0|_{\mathbb{T}} (1 + t^{-1/2}) e^{-t/(2\lambda^2 L^2)}
\end{aligned}$$

where on the second line we have bounded the sum by an integral. □

Using these simple estimates, we now study the convergence rate of the coupling.

Lemma 8.4.3. *There exists a constants C such that for any $t \geq 0$ we have the bound*

$$\mathbb{E}\left[|M_t|_{\mathbb{T}}^2 + |Z_t|^2 \middle| (Z_0, M_0)\right] \leq |Z_0|^2 e^{-2\lambda t} + \begin{cases} C|M_0|_{\mathbb{T}} e^{-2\lambda t} & 2\lambda < 1/(2\lambda^2 L^2) \\ C|M_0|_{\mathbb{T}}(1+t)e^{-2\lambda t} & 2\lambda = 1/(2\lambda^2 L^2) \\ C|M_0|_{\mathbb{T}} e^{-t/(2\lambda^2 L^2)} & 2\lambda > 1/(2\lambda^2 L^2). \end{cases}$$

Proof. Without loss of generality we may assume that Z_0 and M_0 are deterministic in order to avoid writing the conditional expectation.

Applying ItÅD's lemma, we find from (8.4.2) that

$$\mathrm{d}|Z_t|^2 = -2\lambda|Z_t|^2 \mathrm{d}t + 4 \cdot 1_{t \leq T} Z_t \mathrm{d}W_t^1 + 2 \cdot 1_{t \leq T} \mathrm{d}t.$$

After taking expectations we see that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}|Z_t|^2 = -2\lambda \mathbb{E}|Z_t|^2 + 2\mathbb{P}(t \leq T). \tag{8.4.5}$$

By explicitly solving (8.4.5) and using Lemma 8.4.2, we obtain

$$\begin{aligned} \mathbb{E}|Z_t|^2 &= |Z_0|^2 e^{-2\lambda t} + 2e^{-2\lambda t} \int_0^t e^{2\lambda s} \mathbb{P}(s \leq T)\, \mathrm{d}s \\ &\leq |Z_0|^2 e^{-2\lambda t} + C|M_0|_{\mathbb{T}} e^{-2\lambda t} \underbrace{\int_0^t e^{(2\lambda - 1/(2\lambda^2 L^2))s}(1 + s^{-1/2})\, \mathrm{d}s}_{=:I_t}. \end{aligned}$$

Let us bound I_t. As the integrand is locally integrable, we have for a constant C

$$I_t \leq C\left(1 + \int_0^t e^{(2\lambda - 1/(2\lambda^2 L^2))s}\, \mathrm{d}s\right).$$

Here the $s^{-1/2}$ term can be bounded by 1 for $s > 1$ and for $s \leq 1$ the additional contribution can be absorbed into the constant. To bound the remaining integral we consider three cases:

- $2\lambda < 1/(2\lambda^2 L^2)$: The integral (and I_t) are uniformly bounded, $I_t \leq C$.

- $2\lambda = 1/(2\lambda^2 L^2)$: The integrand is equal to 1 and $I_t \leq C(1+t)$.

- $2\lambda > 1/(2\lambda^2 L^2)$: The integrand grows and $I_t \leq C(1 + e^{(2\lambda - 1/(2\lambda^2 L^2))t})$.

In each case we multiply I_t by $e^{-2\lambda t}$ to obtain the decay rate. In the first two cases this gives the dominant term with $|M_0|_{\mathbb{T}}$ (as opposed to $|Z_0|$) dependence, while in the last case it is lower order than the $e^{-t/(2\lambda^2 L^2)}$ decay we obtain from $\mathbb{E}|M_t|_{\mathbb{T}}^2$ below.

Next let us consider $\mathbb{E}|M_t|_{\mathbb{T}}^2$. Using the finite diameter of the torus we have the simple estimate

$$\mathbb{E}|M_t|_{\mathbb{T}}^2 \leq \pi^2 L^2 \mathbb{P}(T > t).$$

For $t \geq 1$ (say), we can use Lemma 8.4.2, to obtain

$$\mathbb{E}|M_t|_{\mathbb{T}}^2 \leq C|M_0|_{\mathbb{T}} e^{-t/(2\lambda^2 L^2)} \quad \text{for } t \geq 1.$$

This leaves the case when $t \leq 1$ where (8.4.4) blows up. We instead use the martingale property of M_t. Without loss of generality we may assume that $M_0 \in [0, \pi L]$. Then as M_t is stopped at T we know that $M_t \in [0, 2\pi L]$ for all $t \geq 0$. Hence, for any $t \geq 0$,

$$\mathbb{E}|M_t|_{\mathbb{T}}^2 \leq \mathbb{E}|M_t|^2 \leq 2\pi L \mathbb{E} M_t = 2\pi L M_0 = 2\pi L |M_0|_{\mathbb{T}}$$

by the martingale property. Combining the $t \leq 1$ and $t \geq 1$ estimates we have

$$\mathbb{E}|M_t|_{\mathbb{T}}^2 \leq C|M_0|_{\mathbb{T}} e^{-t/(2\lambda^2 L^2)} \quad \text{for } t \geq 0.$$

This together with the bound for $\mathbb{E}|Z_t|^2$ provides the claimed bounds of the lemma and completes its proof. □

Proof of Theorem 8.1.2. By the equivalence of the norms from (X, V) and (Y, V),

we see that

$$\begin{aligned}\mathbb{E}\left(|X_t^1 - X_t^2|_{\mathbb{T}}^2 + |V_t^1 - V_t^2|^2\right) &\leq \left(1 + \frac{1}{\lambda}\right)\mathbb{E}\left(|M_t|_{\mathbb{T}}^2 + |Z_t|^2\right)\\ &\leq C'\zeta(t)\mathbb{E}(|M_0|_{\mathbb{T}} + |Z_0|^2)\\ &\leq C\zeta(t)\mathbb{E}\left(\left(|X_0^1 - X_0^2|_{\mathbb{T}}^2 + |V_0^1 - V_0^2|^2\right)^{1/2} + \left(|X_0^1 - X_0^2|_{\mathbb{T}}^2 + |V_0^1 - V_0^2|^2\right)\right).\end{aligned}$$

Here we used Lemma 8.4.3 to go between the first and second line, and to find the exponentially decreasing term ζ. The constants C and C' come from the constants in equivalence of norms. □

8.4.2 Optimality

In order to show Theorem 8.1.3, we focus on the drift-corrected positions Y_t^1 and Y_t^2 which behave like time-rescaled Brownian motion on the torus. For their quadratic distance we prove the following decay bound.

Proposition 8.4.1. *Suppose there exist functions $\alpha : (0, \pi L] \mapsto \mathbb{R}^+$ and $\zeta : [0, \infty) \mapsto \mathbb{R}^+$ with $\zeta \in L^1([0, \infty))$, such that, for any $z \in (0, \pi L]$ there exist two standard Brownian motions W_t^1 and W_t^2 on the torus $\mathbb{T} = \mathbb{R}/(2\pi L\mathbb{Z})$ with respect to a common filtration such that $|W_0^1 - W_0^2| = z$, and for $t \in \mathbb{R}^+$ it holds that*

$$\mathbb{E}[|W_t^1 - W_t^2|_{\mathbb{T}}^2] \leq (\alpha(z))^2\zeta(t).$$

Then with a constant c only depending on L, the function α satisfies the bound

$$\alpha(z) \geq c\,\|\zeta\|_{L^1([0,\infty))}^{-1/2}\sqrt{z}.$$

From this Theorem 8.1.3 follows easily.

Proof of Theorem 8.1.3. Fix $z \in (0, \pi L]$ and consider the initial distributions $\mu = \delta_{X=0}\delta_{V=0}$ and $\nu = \delta_{X=z}\delta_{V=0}$. Between μ and ν, there is only one coupling and $\mathcal{W}_2(\mu, \nu) = z$.

If there exists a co-adapted coupling $((X_t^1, V_t^1), (X_t^2, V_t^2))$ satisfying the bound, then Y_{t/λ^2}^1 and Y_{t/λ^2}^2 are Brownian motions on the torus with a common filtration.

Moreover,

$$\mathbb{E}[|Y_t^1 - Y_t^2|_{\mathbb{T}}^2] \leq C\,\mathbb{E}[|X_t^1 - X_t^2|_{\mathbb{T}}^2 + |V_t^1 - V_t^2|^2]$$

for a constant C only depending on λ. Hence we can apply Proposition 8.4.1 to find the claimed lower bound for α. □

For the proof of Proposition 8.4.1, we first prove the following lemma.

Lemma 8.4.4. *Given two Brownian motions W_t^1 and W_t^1 on the torus with a common filtration, then there exists a numerical constant c such that*

$$\mathbb{E}[|W_t^1 - W_t^2|_{\mathbb{T}}^2] \geq c\,e^{-2t/L^2}\mathbb{E}[|W_0^1 - W_0^2|_{\mathbb{T}}^2].$$

Proof. The natural (squared) metric $|x-y|_{\mathbb{T}}^2$ on the torus is not a global smooth function of $x, y \in \mathbb{R}$ as it takes $x, y \bmod 2\pi L$. Therefore we introduce the equivalent metric

$$d_{\mathbb{T}}^2(x,y) = L^2 \sin^2\left(\frac{x-y}{2L}\right),$$

which is a smooth function of $x, y \in \mathbb{R}$. Moreover, the constants of equivalence are independent of L, i.e. there exist numerical constants c_1 and c_2 such that

$$c_1|x-y|_{\mathbb{T}}^2 \leq d_{\mathbb{T}}^2(x,y) \leq c_2|x-y|_{\mathbb{T}}^2.$$

Now consider H_t defined by

$$H_t = L\sin\left(\frac{W_t^1 - W_t^2}{2L}\right)\exp\left(\frac{[W^1 - W^2]_t}{4L^2}\right).$$

As W_t^1 and W_t^2 are Brownian motions, their quadratic variation is controlled as $[W^1 - W^2]_t \leq 4t$. By ItÅĐ's lemma

$$\mathrm{d}H_t = \frac{1}{2}\cos\left(\frac{W_t^1 - W_t^2}{2L}\right)\exp\left(\frac{[W^1 - W^2]_t}{4L^2}\right)\mathrm{d}(W^1 - W^2)_t.$$

Therefore we may bound the quadratic variation of H by

$$\begin{aligned}[H]_t &= \int_0^t \frac{1}{4}\cos^2\left(\frac{W_t^1 - W_t^2}{2L}\right)\exp\left(\frac{[W^1 - W^2]_t}{2L^2}\right)\mathrm{d}[W^1 - W^2]_t \\ &\leq \int_0^t \exp\left(\frac{2t}{L^2}\right)\mathrm{d}t \\ &< \infty.\end{aligned}$$

Therefore, as also $|H_0| \leq L$, the local martingale H_t is a true martingale and by Jensen's inequality

$$\mathbb{E}[|H_t|^2] \geq \mathbb{E}[|H_0|^2].$$

Using the equivalence of two metrics, we thus find the required bound

$$\begin{aligned}\mathbb{E}[|W_t^1 - W_t^2|_{\mathbb{T}}^2] &\geq c_2^{-1}\mathbb{E}\left[|H_t|^2\exp\left(-\frac{[W^1 - W^2]_t}{2L^2}\right)\right] \\ &\geq c_2^{-1}\mathbb{E}\left[|H_0|^2\right]\exp\left(-\frac{2t}{L^2}\right) \\ &\geq c_1c_2^{-1}\mathbb{E}[|W_0^1 - W_0^2|_{\mathbb{T}}^2]\exp\left(-\frac{2t}{L^2}\right).\end{aligned}$$

□

With this we approach the final proof.

Proof of Proposition 8.4.1. Fix $a \in (0,1)$, let $z \in (0, \pi L]$ be given, and by symmetry assume without loss of generality that $W_0^1 - W_0^2 = |W_0^1 - W_0^2| = z$. Then define the stopping time

$$T = \inf\{t \geq 0 : W_t^1 - W_t^2 \notin (az, \pi L)\}.$$

The distance can be directly bounded as

$$\mathbb{E}[|W_t^1 - W_t^2|_{\mathbb{T}}^2] \geq \mathbb{P}[T \geq t](az)^2.$$

As ζ is integrable, it must decay along a subsequence of times and thus T must be almost surely finite.

As W_t^1 and W_t^2, considered on $\mathbb{R}$, are continuous martingales, their difference is also a continuous martingale. By the construction of the stopping time, the stopped martingale $(W^1 - W^2)_{t\wedge T}$ is bounded by πL and the optional stopping

theorem implies

$$\mathbb{P}[W_T^1 - W_T^2 = \pi L] = \frac{z - az}{\pi L - az}.$$

Since Brownian motions satisfy the strong Markov property, we find together with Lemma 8.4.4

$$\begin{aligned}
\mathbb{E}\int_0^\infty |W_t^1 - W_t^2|_{\mathbb{T}}^2 \mathrm{d}t &\geq \mathbb{E}\int_T^\infty |W_t^1 - W_t^2|_{\mathbb{T}}^2 \mathrm{d}t \\
&\geq \mathbb{P}[W_T^1 - W_T^2 = \pi L]\mathbb{E}\left[\int_T^\infty |W_t^1 - W_t^2|_{\mathbb{T}}^2 \,\mathrm{d}t \;\middle|\; W_T^1 - W_T^2 = \pi\right] \\
&\geq \mathbb{P}[W_T^1 - W_T^2 = \pi L] c\,(\pi L)^2 \int_0^\infty e^{-2t/L^2} \mathrm{d}t \\
&\geq \frac{z - az}{\pi L - az} c\,(\pi L)^2 \frac{L^2}{2} \\
&\geq C_a z
\end{aligned}$$

for a constant C_a only depending on a and L, where the strong Markov property and then the lemma are applied on the second line.

On the other hand, integrating the assumed bound gives

$$\mathbb{E}\int_0^\infty |W_t^1 - W_t^2|_{\mathbb{T}}^2 \mathrm{d}t \leq (\alpha(z))^2 \int_0^\infty \zeta(t)\mathrm{d}t \leq (\alpha(z))^2 \|\zeta\|_{L^1([0,\infty))}.$$

Hence

$$C_a z \leq (\alpha(z))^2 \|\zeta\|_{L^1([0,\infty))}$$

which is the claimed result. □

8.5 Proof of Theorem 8.1.4

Proof of Theorem 8.1.4. Let $(\Omega, \mathbb{P})$ denote the probability space. Without loss of generality, we may assume that $|X_0^1 - X_0^2|_{\mathbb{T}} = X_0^1 - X_0^2$. By (8.1.3) and an application of the Borel Cantelli lemma, we see that

$$V_t^1 - V_t^2 \to 0, \qquad |X_t^1 - X_t^2|_{\mathbb{T}} \to 0 \qquad \text{a.s. as } t \to \infty.$$

By continuity in time of X^1, X^2, the process $\tilde{X}_t := X_t^1 - X_t^2$ converges almost surely as $t \to \infty$ to a limit $\tilde{X}_\infty$, which satisfies $\tilde{X}_\infty \in 2\pi\mathbb{Z}$ almost surely. Due to (8.1.3), we can also express

$$\tilde{X}_\infty = X_0^1 - X_0^2 + \int_0^\infty V_t^1 - V_t^2 \, dt,$$

where the integral converges absolutely in $L^1(\Omega)$. Hence,

$$\mathbb{E}\tilde{X}_\infty = X_0^1 - X_0^2 + \int_0^\infty \mathbb{E}V_t^1 - \mathbb{E}V_t^2 \, dt = X_0^1 - X_0^2 = |X_0^1 - X_0^2|_{\mathbb{T}}$$

as V_t^1, V_t^2 have the same individual laws.

Let N be the number of times $\tilde{X}_t$ crosses an interval of the form $[(2k+1)\pi, (2k+3/2)\pi]$ for some $k \in \mathbb{Z}$ as t goes from 0 to ∞. We will assume that N is a.s. finite. The case where N is infinite is an easy adaptation. As $\tilde{X}_\infty \in 2\pi\mathbb{Z}$ a.s. and $\tilde{X}_0 = d \in [0, \pi]$, N is at least as large as $|\tilde{X}_\infty|/2\pi$. In particular,

$$\mathbb{E}N \geq \frac{1}{2\pi}\mathbb{E}|\tilde{X}_\infty| \geq \frac{1}{2\pi}\mathbb{E}\tilde{X}_\infty = \frac{1}{2\pi}|X_0^1 - X_0^2|_{\mathbb{T}}$$

Observe that, by Fubini's theorem and (8.1.3)

$$\mathbb{E}\int_0^\infty |V_t^1 - V_t^2|^2 + |X_t^1 - X_t^2|_{\mathbb{T}}^2 \, dt \leq Cc^2,$$

where C depends only on γ. Letting $\underline{t}_n, \bar{t}_n$ be the start and end times of the nth crossing of an interval as in the definition of N, we have the lower bound

$$\sum_{n=1}^N \int_{\underline{t}_n}^{\bar{t}_n} |V_t^1 - V_t^2|^2 + |X_t^1 - X_t^2|_{\mathbb{T}}^2 \, dt \leq \int_0^\infty |V_t^1 - V_t^2|^2 + |X_t^1 - X_t^2|_{\mathbb{T}}^2 \, dt.$$

We claim that each term in the sum on the left hand side is almost surely bounded below by an absolute deterministic constant C'. With this claim the proposition follows by taking expectations and combining the above displays. Indeed,

$$\begin{aligned}\frac{C'}{2\pi}|X_0^1 - X_0^2|_{\mathbb{T}} \leq C'\mathbb{E}N &\leq \mathbb{E}\sum_{n=1}^N \int_{\underline{t}_n}^{\bar{t}_n} |V_t^1 - V_t^2|^2 + |X_t^1 - X_t^2|_{\mathbb{T}}^2 \, dt \\ &\leq \mathbb{E}\int_0^\infty |V_t^1 - V_t^2|^2 + |X_t^1 - X_t^2|_{\mathbb{T}}^2 \, dt \leq Cc^2,\end{aligned}$$

and we finish by taking square roots.

We now prove the claim. Denote the time interval as $[\tau, \tau + T]$. The integral considered is then

$$I = \int_\tau^{\tau+T} |V_t^1 - V_t^2|^2 + |X_t^1 - X_t^2|_{\mathbb{T}}^2 \, dt.$$

We bound this integral below in two ways. First, which is optimal when T is large, we have the bound

$$I \geq \int_\tau^{\tau+T} |X_t^1 - X_t^2|_{\mathbb{T}}^2 \, dt \geq \int_\tau^{\tau+T} \left(\frac{\pi}{2}\right)^2 \, dt \geq \frac{\pi^2 T}{4}.$$

Second, which is optimal when T is small, we have the bound

$$\begin{aligned} I &\geq \int_\tau^{\tau+T} |V_t^1 - V_t^2|^2 \, dt = T \int_0^T \frac{1}{T} |V_t^1 - V_t^2|^2 \, dt \geq T \left(\int_\tau^{\tau+T} \frac{1}{T} |V_t^1 - V_t^2| \, dt \right)^2 \\ &= \frac{1}{T} \left(\int_\tau^{\tau+T} |V_t^1 - V_t^2| \, dt \right)^2 \geq \frac{1}{T} |\tilde{X}_{\tau+T} - \tilde{X}_\tau|^2 = \frac{\pi^2}{4T}, \end{aligned}$$

where we have used Jensen's inequality on the first line. The claim is proved by noting that $\min(T, 1/T) = 1$. This completes the proof of the theorem. □

Chapter 9

Convergence Along Mean Flows

We develop a technique of multiple scale asymptotic expansions along mean flows and a corresponding notion of weak multiple scale convergence. These are applied to homogenize convection dominated parabolic equations with rapidly oscillating, locally periodic coefficients and $\mathcal{O}(\varepsilon^{-1})$ mean convection term. Crucial to our analysis is the introduction of a fast time variable, $\tau = t/\varepsilon$, not apparent in the heterogeneous problem. The effective diffusion coefficient is expressed in terms of the average of Eulerian cell solutions along the orbits of the mean flow in the fast time variable. To make this notion rigorous, we use the theory of ergodic algebras with mean value.

Acknowledgements

The work in this chapter was done in collaboration with Harsha Hutridurga and Jeffrey Rauch. We would like to thank Grégoire Allaire for his fruitful suggestions during the preparation of this article. The authors would also like to thank Mariapia Polambaro for helpful discussions regarding Euclidean motions and for bringing to our attention the work of P-E. Jabin and A. Tzavaras [99]. This chapter appears in a similar form in [88]. We would like to thank the referee of [88] for helpful comments.

9.1 Introduction

This chapter studies the homogenization of parabolic equations of convection-diffusion type with locally periodic (in space), rapidly oscillating coefficients. This work addresses the self-similar diffusive scaling in these equations, i.e. for an unknown scalar density $u^\varepsilon(t,x)$, we consider the Cauchy problem for a convection-diffusion equation with large convection term:

$$\frac{\partial u^\varepsilon}{\partial t} + \frac{1}{\varepsilon}\mathbf{b}\left(x,\frac{x}{\varepsilon}\right)\cdot\nabla u^\varepsilon - \nabla\cdot\left(\mathbf{D}\left(x,\frac{x}{\varepsilon}\right)\nabla u^\varepsilon\right) = 0 \qquad \text{for } (t,x)\in\,]0,T[\times\mathbb{R}^d \tag{9.1.1}$$

with $0<\varepsilon\ll 1$ the scale of heterogeneity. This scaling corresponds to the long-term behaviour which can be described in terms of the effective or homogenized limit of the above scaled system.

It has remained a largely open problem to determine the homogenized limit of the scaled equation (9.1.1). This present work gives a partial answer to this question in the sense that we homogenize the non-homogeneous equation with locally periodic coefficients under some structural assumptions on the flows associated with certain vector fields. This is achieved by the introduction of a new notion of weak convergence in L^p spaces with $1<p<\infty$.

Under no diffuse scaling, i.e. with no large convection term, homogenization of such equations is classical. In such a scenario, we can either employ the method of asymptotic expansions (see for instance, the monographs [20, 175]) which provides us with the approximation

$$u^\varepsilon(t,x)\approx u_0(t,x)+\varepsilon u_1\left(t,x,\frac{x}{\varepsilon}\right)+\varepsilon^2 u_1\left(t,x,\frac{x}{\varepsilon}\right)+\cdots \tag{9.1.2}$$

or employ a weak convergence approach of the two-scale convergence method introduced by G. Nguetseng in [154] and further developed by G. Allaire in [1]. The cornerstone result of the two-scale convergence method is that, up to extraction of a subsequence, any uniformly (w.r.t. ε) bounded sequence $\{u^\varepsilon\}$ in some L^p

space with $1 < p < \infty$ satisfies

$$\lim_{\varepsilon \to 0} \iint\limits_{(0,T)\times\mathbb{R}^d} u^\varepsilon(t,x)\psi\left(t,x,\frac{x}{\varepsilon}\right)\,\mathrm{d}x\,\mathrm{d}t = \iiint\limits_{(0,T)\times\mathbb{R}^d\times\mathbb{T}^d} u_0(t,x,y)\psi(t,x,y)\,\mathrm{d}y\,\mathrm{d}x\,\mathrm{d}t$$

for some $u_0(t,x,y) \in L^p((0,T)\times\mathbb{R}^d\times\mathbb{T}^d)$ called the weak two-scale limit and for any smooth $\psi(t,x,y)$ which is periodic in the y variable.

Any weak convergence approach to homogenize a partial differential equation would involve passing to the limit (as the heterogeneity length scale tends to zero) in the weak formulation associated to the partial differential equation. This would require passing to the limit in products of weakly converging sequences. The main feature of the two-scale convergence method is that the particular choice of test functions allows us to pass to the limit in such products. If $u^\varepsilon(t,x)$ weakly two-scale converges to $u_0(t,x,y) \in L^p((0,T)\times\mathbb{R}^d\times\mathbb{T}^d)$ and if the coefficient function $a(t,x,y)$, which is periodic in the y variable, is admissible (roughly speaking, continuous or approximable by continuous functions in a certain sense - see Definition 9.3.7 for precise statement), then the product has the convergence

$$a\left(t,x,\frac{x}{\varepsilon}\right)u^\varepsilon(t,x) \rightharpoonup \int\limits_{\mathbb{T}^d} a(t,x,y)u_0(t,x,y)\,\mathrm{d}y \qquad \text{as } \varepsilon \to 0,$$

in the sense of distributions.

In recent years, there have been numerous publications in the mathematics literature dedicated to generalize the notion of two-scale convergence (originally developed to handle periodic structures) to address the homogenization of partial differential equations with coefficients that belong to some ergodic algebras. Typically, all these works are about the study of the limiting behaviour (as $\varepsilon \to 0$) of the integral

$$\int\limits_{\mathbb{R}^d} v^\varepsilon(x)\psi\left(x,\frac{x}{\varepsilon}\right)\,\mathrm{d}x$$

when $\{v^\varepsilon\}$ is a uniformly bounded sequence in some Lebesgue space L^p with $1 < p < \infty$ and $\psi(x,y)$ belongs to certain ergodic algebra in the y variable. The notion of *algebras with mean value* play a crucial role in these theories. This

notion goes back to the work of Zhikov and Krivenko [206] in the early 1980's (also see the book of Jikov, Kozlov and Oleinik [205] for a pedagogical exposition). We cite some of the references in this context which we have consulted in developing our theory: [35, 155, 156, 158].

With regard to the homogenization of the scaled equation (9.1.1), the known results are when the rapidly oscillating coefficients are purely periodic, i.e. of the type $\mathbf{b}\left(\frac{x}{\varepsilon}\right)$, $\mathbf{D}\left(\frac{x}{\varepsilon}\right)$. The case when the fluid field $\mathbf{b}(\cdot)$ is of zero mean was treated in [20, 142] using two-scale asymptotic expansions of the form (9.1.2). They do not prove convergence. Over two decades ago, to address the case of fluid field $\mathbf{b}(\cdot)$ with non-zero mean, G. Papanicolaou suggested in [166] a modified two-scale asymptotic expansion where the coefficients in the expansion are taken along rapidly moving coordinates:

$$u^\varepsilon(t,x) \approx u_0\left(t, x - \frac{\mathbf{b}^* t}{\varepsilon}\right) + \varepsilon u_1\left(t, x - \frac{\mathbf{b}^* t}{\varepsilon}, \frac{x}{\varepsilon}\right) + \varepsilon^2 u_2\left(t, x - \frac{\mathbf{b}^* t}{\varepsilon}, \frac{x}{\varepsilon}\right) + \cdots \tag{9.1.3}$$

The constant $\mathbf{b}^* \in \mathbb{R}^d$ is the mean field associated with $\mathbf{b}(\cdot)$. Note that the case $\mathbf{b}^* = 0$ coincides with the classical expansion (9.1.2). We cite the works in [6, 5, 3] where the above expansion with drift is employed in homogenizing reactive transport models in periodic porous media.

Analogous to the two-scale convergence method, Marušić-Paloka and Piatnitski introduced a notion of weak convergence in [140] called the two-scale convergence with drift (see [2] for a pedagogical exposition of this method) characterizing the limit

$$\lim_{\varepsilon\to 0} \iint\limits_{(0,T)\times\mathbb{R}^d} u^\varepsilon(t,x)\psi\left(t, x - \frac{\mathbf{b}^* t}{\varepsilon}, \frac{x}{\varepsilon}\right) \mathrm{d}x\,\mathrm{d}t$$

where $\psi(t,x,y)$ is periodic in the y variable and as usual the family $\{u^\varepsilon\}$ is uniformly bounded (w.r.t. ε) in some L^p space with $1 < p < \infty$.

Neither the modified two-scale expansion (9.1.3) nor the notion of two-scale convergence with drift seem capable of treating equation (9.1.1) with locally periodic, rapidly oscillating coefficients, i.e. when $\mathbf{b}$ depends upon both x and y. We cite the work of P-E. Jabin and A. Tzavaras [99] which treats the homogenization

of (9.1.1) with locally periodic fluid field $\mathbf{b}(x, y)$ and diffusion coefficient being unity. They treat a special case when the mean field $\bar{\mathbf{b}}(x)$ of the locally periodic fluid field $\mathbf{b}(x, y)$ vanishes, i.e. $\bar{\mathbf{b}}(x) \equiv 0$ for all x. They introduce a notion of *kinetic decomposition* to address this problem. As far as the authors are aware, the techniques of [99] are not capable of addressing the case of non-zero mean field.

In this work, we introduce a new multiple scale expansion

$$u^\varepsilon(t,x) \approx u_0\left(t, \Phi_{-t/\varepsilon}(x)\right) + \varepsilon u_1\left(t, \frac{t}{\varepsilon}, \Phi_{-t/\varepsilon}(x), \frac{x}{\varepsilon}\right) + \varepsilon^2 u_2\left(t, \frac{t}{\varepsilon}, \Phi_{-t/\varepsilon}(x), \frac{x}{\varepsilon}\right) + \cdots \tag{9.1.4}$$

which we call *multiple scale expansion along mean flows*. The coefficient functions u_i in (9.1.4) are taken on rapidly moving coordinates $\Phi_{-t/\varepsilon}(x)$ which is the flow associated with the mean field $\bar{\mathbf{b}}(x)$ of the locally periodic fluid field $\mathbf{b}(x, y)$. A novelty of our method is the introduction of the fast time variable $\tau := t/\varepsilon$. The main assumption in this work is on the Jacobian matrix $J(\tau, x)$ associated with the flow $\Phi_\tau(x)$.

Assumption: *There is a uniform constant C such that $|J(\tau, x)| \leq C$ for all $(\tau, x) \in \mathbb{R} \times \mathbb{R}^d$.*

The above assumption is trivially satisfied in all the previously known works on the homogenization of (9.1.1) because the Jacobian matrix associated with the flows in all these works is the identity.

Under this assumption, we derive a homogenized diffusion equation for the zeroth order approximation u_0 in (9.1.4) with an explicit expression for the effective diffusion coefficient. The diffusion equation for u_0 is in Lagrangian coordinates because of the structure of the asymptotic expansion. The effect of Lagrangian stretching on the gradient of the scalar density u^ε, i.e. creating large gradients has been widely studied in the literature in the case of non-oscillating coefficients (see for e.g. [83, 21, 40, 79]). If the above assumption is not made on the Jacobian matrix, we cannot expect a nontrivial limit as the large gradients can drive the solution to zero quickly. The mathematical model considered in this article is one of the simplified models for turbulent diffusion studied widely in the physics and mathematics literature – for further details consult [137, Section 2].

Taking inspiration from the work of Marušić-Paloka and Piatnitski [140], we devise a weak convergence approach which involves the characterization of the limit

$$\lim_{\varepsilon\to 0} \iint\limits_{(0,T)\times\mathbb{R}^d} u^\varepsilon(t,x)\psi\left(t,\Phi_{-t/\varepsilon}(x),\frac{t}{\varepsilon},\frac{x}{\varepsilon}\right)\,\mathrm{d}x\,\mathrm{d}t$$

with a uniformly bounded family $\{u^\varepsilon\}$ in some L^p space $(1<p<+\infty)$ and the test function $\psi(t,\tau,x,y)$ being periodic in the y variable and belongs to an *ergodic algebra with mean value* in the τ variable. We call this notion of convergence *weak* Σ*-convergence along flows*.

To use this new notion of convergence, our strategy is to use test functions of the form $\psi\left(t,\Phi_{-t/\varepsilon}(x),\frac{t}{\varepsilon},\frac{x}{\varepsilon}\right)$ in the weak formulation of the scaled problem (9.1.1). This weak formulation would have terms involving the Jacobian matrix associated with the flow $\Phi_{-t/\varepsilon}(x)$. Note that the Jacobian matrix depends on the fast time variable, i.e. appears as $J\left(\frac{t}{\varepsilon},x\right)$, because of the chosen time scale in the flow. Our strategy, hence, is to consider test functions that belong to some ergodic algebra in the fast time variable τ.

Inspired by the notion of *admissible* functions introduced by G. Allaire in [1] and further clarified by M. Radu in her PhD thesis [169], we introduce a notion of *admissible* functions adapted to the weak Σ-convergence along flows (see Definition 9.3.8). Another novelty of our approach is to consider the *flow-representation* of functions (see Section 9.2.2 for precise definition). Our main result is to show that if the flow-representations of the fluid field $\mathbf{b}(x,y)$, the diffusion matrix $\mathbf{D}(x,y)$, the Jacobian matrix $J(\tau,x)$ are admissible, then we can derive the effective limit diffusion equation. These assumptions on the coefficients and the Jacobian matrix are very essential for our analysis as is evident from the counterexamples that are constructed in Section 9.5 of this chapter.

The main homogenization result of this article is Theorem 9.4.1. We summarize this result below (consult Theorem 9.4.1 in Section 9.4 for precise statement).

Theorem. *Let $\Phi_\tau(x)$ be the flow associated with the mean field $\bar{\mathbf{b}}(x)$. Suppose the associated Jacobian matrix $J(\tau,x)$ is a uniformly bounded function of τ and*

x variables. Let the flow-representations of the coefficients in (9.1.1) and that of the Jacobian matrix belong to certain ergodic algebra with mean value. Then the solution family $u^\varepsilon(t,x)$ weakly Σ-converges along the flow Φ_τ to the unique solution of the homogenized equation

$$\frac{\partial u_0}{\partial t} - \nabla_x \cdot \Big(\mathfrak{D}(x) \nabla_x u_0 \Big) = 0$$

where the effective diffusion matrix $\mathfrak{D}(x)$ is given in terms of certain averages of solutions to cell problems and the averages are taken along the orbits of the mean flow.

Outline of the chapter:

- In Section 9.2, we introduce the method of *multiple scale asymptotic expansions along mean flows* to derive the effective equation for the scaled equation (9.2.4a)-(9.2.4b). This result is recorded as Proposition 9.2.1 which gives an explicit expression for the effective diffusion matrix.

- Section 9.3 introduces the new notion of weak multiple scale convergence. In Subsections 9.3.1 through 9.3.5, we recall enough of the theory of algebras with mean value. The notion of Σ-convergence along flows is introduced in Section 9.3.6. The main compactness result with regard to this new notion of convergence is given by Theorem 9.3.2. In Section 9.3.8, we obtain compactness results on the gradient sequences (in the sense of corrector results in homogenization).

- Section 9.4 deals with the homogenization result. The main result of this section is Theorem 9.4.1. The main assumptions made on the coefficients and the Jacobian matrix are explained in Section 9.4.2.

- Section 9.5 provides some discussion on the assumptions made on the coefficients and the Jacobian matrix. In particular, we give some examples of fluid fields with bounded Jacobians and show that unbounded growth in the Jacobian matrix can lead to trivial and singular behaviour of the limit u_0. We also provide an explicit example of an equation, where the assumptions on the flow-representation of the coefficients do not hold, leading to two different homogenized equations in the $\varepsilon \to 0$ limit.

- In Section 9.6, we perform asymptotic analysis on some explicit convection-diffusion models which highlights the effectiveness of this new approach in addressing the large convection terms. Finally, in Section 9.7, we give some concluding remarks.

9.2 Asymptotic expansion along flows

9.2.1 Mathematical model

Let $\mathbf{b}(x,y) : \mathbb{R}^d \times \mathbb{T}^d \to \mathbb{R}^d$ be a prescribed time-independent fluid field which is incompressible in both the x and y variables, i.e.

$$\nabla_x \cdot \mathbf{b}(x,y) = \nabla_y \cdot \mathbf{b}(x,y) = 0 \qquad \text{for a.e. } (x,y) \in \mathbb{R}^d \times \mathbb{T}^d. \tag{9.2.1}$$

Define the associated mean field as

$$\bar{\mathbf{b}}(x) := \int_{\mathbb{T}^d} \mathbf{b}(x,y)\,\mathrm{d}y. \tag{9.2.2}$$

NOTATION: For any matrix $\mathbf{B}$, its transpose is denoted by ${}^\top\mathbf{B}$.

Let $\mathbf{D}(x,y) \in L^\infty(\mathbb{R}^d \times \mathbb{T}^d; \mathbb{R}^{d\times d})$ be a given time-independent symmetric (i.e. $\mathbf{D} = {}^\top\mathbf{D}$) matrix-valued diffusion coefficient which is assumed to be uniformly coercive, i.e.

$$\begin{aligned} &\exists\, \lambda, \Lambda > 0 \ \text{ s.t. } \ \lambda|\xi|^2 \le {}^\top\xi\, \mathbf{D}(x,y)\, \xi \le \Lambda|\xi|^2, \\ &\qquad\qquad \text{holds for all } \xi \in \mathbb{R}^d \text{ and for a.e. } (x,y) \in \mathbb{R}^d \times \mathbb{T}^d. \end{aligned} \tag{9.2.3}$$

Let $0 < \varepsilon \ll 1$ be the scale of heterogeneity. Let us consider a scaled Cauchy problem with rapidly oscillating coefficients for an unknown scalar density $u^\varepsilon(t,x)$:

$[0,T[\times\mathbb{R}^d \to [0,\infty)$.

$$\frac{\partial u^\varepsilon}{\partial t} + \frac{1}{\varepsilon}\mathbf{b}\left(x,\frac{x}{\varepsilon}\right)\cdot\nabla u^\varepsilon - \nabla\cdot\left(\mathbf{D}\left(x,\frac{x}{\varepsilon}\right)\nabla u^\varepsilon\right) = 0 \qquad \text{for } (t,x)\in\,]0,T[\times\mathbb{R}^d, \tag{9.2.4a}$$

$$u^\varepsilon(0,x) = u^{in}(x) \qquad \text{for } x\in\mathbb{R}^d. \tag{9.2.4b}$$

The *two-scale expansions with drift* method (see [140, 50, 6, 2, 5, 3]) employs the asymptotic expansion for the unknown density:

$$u^\varepsilon(t,x) = \sum_{i=0}^{\infty}\varepsilon^i u_i\left(t, x-\frac{\mathbf{b}^* t}{\varepsilon}, \frac{x}{\varepsilon}\right), \tag{9.2.5}$$

where the drift velocity $\mathbf{b}^*\in\mathbb{R}^d$ is a constant and the coefficient functions $u_i(t,x,y)$ are assumed to be periodic in the y variable. Remark that the coefficient functions u_i in (9.2.5) are written in moving coordinates. To be precise, consider the ordinary differential equation

$$\dot{x} = \mathbf{b}^*; \qquad x(0) = x. \tag{9.2.6}$$

Denote by $\Phi_\tau(x)$ the flow associated with (9.2.6). The flow evaluated at the time instant $(-t/\varepsilon)$ is nothing but the moving coordinates taken in the asymptotic expansion (9.2.5), i.e.

$$\Phi_{-t/\varepsilon}(x) = x - \frac{\mathbf{b}^* t}{\varepsilon}.$$

The idea of considering the asymptotic expansion along moving coordinates was mentioned by G. Papanicolaou in a survey paper [166]. It should be noted that the *two-scale expansions with drift* method can handle the homogenization of convection-diffusion equation (9.2.4a) only when the fluid field is purely periodic, i.e. $\mathbf{b}(x,y)\equiv\mathbf{b}(y)$. In that case, the constant drift velocity is taken to be

$$\mathbf{b}^* = \fint_{\mathbb{T}^d}\mathbf{b}(y)\,\mathrm{d}y.$$

Taking cues from the constant drift scenario, consider the autonomous system

$$\dot{x} = \bar{\mathbf{b}}(x); \qquad x(0) = x. \tag{9.2.7}$$

Again, denoting the flow associated with (9.2.7) by $\Phi_\tau(x)$, we postulate the asymptotic expansion in the spirit of (9.2.5):

$$u^\varepsilon(t,x) = \sum_{i=0}^{\infty} \varepsilon^i u_i \left(t, \frac{t}{\varepsilon}, \Phi_{-t/\varepsilon}(x), \frac{x}{\varepsilon}\right). \tag{9.2.8}$$

Note that the coefficient functions $u_i(t,\tau,x,y)$ in (9.2.8) depend on an additional variable τ which we shall call the *fast time variable*. We assume that the coefficient functions $u_i(t,\tau,x,y)$ are periodic in the y variable. The structural assumption on the coefficients $u_i(t,\tau,x,y)$ with regard to the τ variable is a little bit subtle. We shall assume that the coefficient functions, as a function of τ, belong to an *ergodic algebra with mean value*. This shall guarantee the existence of certain weak* limits. This will be made more rigorous in a later stage of the article (see Section 9.3). The authors of [26] also introduced a fast time variable in their asymptotic expansion. However, they do not consider the expansion along moving coordinates as is the case in (9.2.8). Also, the authors of [26] assume that the coefficient functions decay exponentially in the fast time variable.

9.2.2 Flow representation

We introduce a notion of *flow representation* that is very central to our analysis. The choice of considering rapidly moving coordinates in the expansion (9.2.8) is equivalent to expressing the convection-diffusion equation (9.2.4a) in Lagrangian coordinates. This necessitates the consideration of the coefficient functions in the convection-diffusion equation in Lagrangian coordinates. Essentially, the flow representation takes into account the underlying flow structure associated with the mean field $\bar{\mathbf{b}}(x)$.

To be precise, consider a function $f : \mathbb{R}^d \to \mathbb{R}$ and a flow $\Phi_\tau(x) : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$. The flow representation of f is given by the function $\widetilde{f} : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$ defined

as

$$\widetilde{f}(\tau, x) := f(\Phi_\tau(x)).$$

We shall use the following convention for their flow representations when we encounter locally periodic functions, i.e. functions of the form $f(x, y) : \mathbb{R}^d \times \mathbb{T}^d \to \mathbb{R}$

$$\widetilde{f}(\tau, x, y) := f(\Phi_\tau(x), y).$$

The following observations are obvious for the flow representations:

$$\begin{aligned}
&\widetilde{f}(0, x) = f(x);\\
&\widetilde{f}(\tau, \Phi_{\tau'}(x)) = f(\Phi_{\tau+\tau'}(x)) \qquad \text{for any } \tau, \tau' \in \mathbb{R};\\
&\widetilde{f}(\tau, x) = f(x) \qquad \text{with the convention } x := \Phi_{-\tau}(x).
\end{aligned}$$

When we encounter vector-valued functions, it should be noted that their flow representations are taken component-wise. It should also be noted that the flow Φ_τ used in giving the flow representation of a function can be any one-parameter group of transformation and need not be associated with any vector field.

Remark 9.2.1. *The one parameter group U^τ defined by $(U^\tau f)(x) := \widetilde{f}(\tau, x)$ is generated (at least formally, i.e. without regard to functional spaces) by the skew-symmetric operator $\bar{\mathbf{b}}(x) \cdot \nabla$.*

9.2.3 Flows associated with vector fields

Let $J(\tau, x)$ denote the Jacobian matrix of the flow Φ_τ generated by (9.2.7), i.e.

$$J(-\tau, x) = \begin{bmatrix} \frac{\partial \Phi_\tau^1}{\partial x_1} & \cdots & \frac{\partial \Phi_\tau^1}{\partial x_d} \\ \vdots & & \vdots \\ \frac{\partial \Phi_\tau^d}{\partial x_1} & \cdots & \frac{\partial \Phi_\tau^d}{\partial x_d} \end{bmatrix} = \left(\frac{\partial \Phi_\tau^i}{\partial x_j} \right)_{i,j=1}^{d}. \tag{9.2.9}$$

We have used the convention that $J(\tau, x)$ is the Jacobian of the *backwards* flow $\Phi_{-\tau}(x)$ to ease notation as it is this that appears throughout. The flow represen-

tation of the Jacobian matrix function $J : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is defined by

$$\widetilde{J}(\tau, \Phi_{-\tau}(x)) = \widetilde{J}(\tau, x) = J(\tau, x).$$

In order to ensure the validity of the proposed asymptotic expansion (9.2.8) we make the assumption of uniform boundedness on the Jacobian matrix:

Assumption 9.2.1. *There is a constant C such that $|J(\tau, x)| \leq C$ for all $\tau \in \mathbb{R}$ and $x \in \mathbb{R}^d$.*

To finish this subsection we record some facts regarding the change of variables. Although these are well known, we provide a proof in Appendix 9.A for the convenience of the reader.

Lemma 9.2.1. *Let $\bar{\mathbf{b}} \in C^1(\mathbb{R}^d)$, then the following hold:*

(i) $\nabla_x \cdot {}^\top\widetilde{J}(\tau, x) = 0$ *in the sense of distributions.*

(ii) $\nabla_x \cdot \left(\widetilde{J}(\tau, x)\widetilde{f}(\tau, x, y)\right) = 0$ *in the sense of distributions, for any vector field $f(x, y)$ which is of null-divergence in the x variable, i.e. $\nabla_x \cdot f(x, y) = 0$.*

(iii) For any $\phi, \varphi \in C_c^\infty(\mathbb{R}^d; \mathbb{R})$ we have the integration by parts formula:

$$\int_{\mathbb{R}^d} \phi(x) \left({}^\top\widetilde{J}(\tau, x)\nabla_x\varphi(x)\right) \mathrm{d}x = -\int_{\mathbb{R}^d} \varphi(x) \left({}^\top\widetilde{J}(\tau, x)\nabla_x\phi(x)\right) \mathrm{d}x.$$

(iv) For any $\tau \in \mathbb{R}$ and $x \in \mathbb{R}^d$ it holds that

$$\bar{\mathbf{b}}(x) = \bar{\mathbf{b}}\left(\Phi_{-\tau}(x)\right) = J(\tau, x)\bar{\mathbf{b}}(x) = \widetilde{J}(\tau, x)\widetilde{\bar{\mathbf{b}}}(\tau, x). \tag{9.2.10}$$

9.2.4 Multiple scale expansion along mean flows

We present a strategy to formally arrive at an effective equation for (9.2.4a)-(9.2.4b) by using the asymptotic expansion (9.2.8) postulated earlier. In the case of constant drift (9.2.5), we have the following chain rules for differentiating the

coefficient functions in the space and times variables:

$$\nabla_x \left(u_i \left(t, x - \frac{\mathbf{b}^* t}{\varepsilon}, \frac{x}{\varepsilon} \right) \right) = \nabla_x u_i \left(t, x - \frac{\mathbf{b}^* t}{\varepsilon}, \frac{x}{\varepsilon} \right) + \frac{1}{\varepsilon} \nabla_y u_i \left(t, x - \frac{\mathbf{b}^* t}{\varepsilon}, \frac{x}{\varepsilon} \right),$$

$$\frac{\partial}{\partial t} \left(u_i \left(t, x - \frac{\mathbf{b}^* t}{\varepsilon}, \frac{x}{\varepsilon} \right) \right) = \frac{\partial u_i}{\partial t} \left(t, x - \frac{\mathbf{b}^* t}{\varepsilon}, \frac{x}{\varepsilon} \right) - \frac{1}{\varepsilon} \mathbf{b}^* \cdot \nabla_x u_i \left(t, x - \frac{\mathbf{b}^* t}{\varepsilon}, \frac{x}{\varepsilon} \right).$$

Remark that the above simple expression for the derivative is because of the Jacobian matrix being the identity for the change of variables:

$$x \mapsto x - \frac{\mathbf{b}^* t}{\varepsilon}.$$

However, for the change of variables

$$x \mapsto \Phi_{-t/\varepsilon}(x)$$

where the flow Φ_τ is associated with (9.2.7), the associated chain rules for differentiating the coefficient functions in the asymptotic expansion (9.2.8) with respect to the space and time variables shall be

$$\begin{aligned} \nabla_x \left(u_i \left(t, \frac{t}{\varepsilon}, \Phi_{-t/\varepsilon}(x), \frac{x}{\varepsilon} \right) \right) = \; & {}^\top \widetilde{J} \left(\frac{t}{\varepsilon}, \Phi_{-t/\varepsilon}(x) \right) \nabla_X u_i \left(t, \frac{t}{\varepsilon}, \Phi_{-t/\varepsilon}(x), \frac{x}{\varepsilon} \right) \\ & + \frac{1}{\varepsilon} \nabla_y u_i \left(t, \frac{t}{\varepsilon}, \Phi_{-t/\varepsilon}(x), \frac{x}{\varepsilon} \right), \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial t} \left(u_i \left(t, \frac{t}{\varepsilon}, \Phi_{-t/\varepsilon}(x), \frac{x}{\varepsilon} \right) \right) = \; & \frac{\partial u_i}{\partial t} \left(t, \frac{t}{\varepsilon}, \Phi_{-t/\varepsilon}(x), \frac{x}{\varepsilon} \right) + \frac{1}{\varepsilon} \frac{\partial u_i}{\partial \tau} \left(t, \frac{t}{\varepsilon}, \Phi_{-t/\varepsilon}(x), \frac{x}{\varepsilon} \right) \\ & - \frac{1}{\varepsilon} \bar{\mathbf{b}} \left(\Phi_{-t/\varepsilon}(x) \right) \cdot \nabla_X u_i \left(t, \frac{t}{\varepsilon}, \Phi_{-t/\varepsilon}(x), \frac{x}{\varepsilon} \right). \end{aligned}$$

The strategy of any asymptotic expansion method in homogenization is to substitute the postulated expansion into the model equation and solve a cascade of equations for obtaining the coefficient functions in the asymptotic expansion. All the equations in this cascade obtained by this approach have a similar structure. Next, we state a standard Fredholm type result which guarantees the solvability of such equations provided the source terms satisfy a compatibility condition.

Lemma 9.2.2. *Let $x \in \mathbb{R}^d$ be a fixed parameter. Suppose $g(x, \cdot) \in L^2(\mathbb{T}^d)$ be the*

source term in the boundary value problem:

$$\mathbf{b}(x,y)\cdot\nabla_y f - \nabla_y\cdot(\mathbf{D}(x,y)\nabla_y f) = g(x,y) \qquad \text{in } \mathbb{T}^d. \tag{9.2.11}$$

Then there exists a unique solution $f \in H^1(\mathbb{T}^d)/\mathbb{R} := \{f \in H^1(\mathbb{T}^d) : \int_{\mathbb{T}^d} f \,\mathrm{d}y = 0\}$ *to* (9.2.11) *if and only if the source term satisfies*

$$\int_{\mathbb{T}^d} g(x,y)\,\mathrm{d}y = 0. \tag{9.2.12}$$

Next, we record a formal result on the homogenized equation for the scaled equation with rapidly oscillating coefficients (9.2.4a)-(9.2.4b).

Proposition 9.2.1 (formal result)**.** *Under Assumption 9.2.1 and the assumption* (9.2.8)*, the solution to the Cauchy problem* (9.2.4a)-(9.2.4b) *formally satisfies*

$$u^\varepsilon(t,x) \approx u_0\left(t, \Phi_{-t/\varepsilon}(x)\right) + \varepsilon u_1\left(t, \frac{t}{\varepsilon}, \Phi_{-t/\varepsilon}(x), \frac{x}{\varepsilon}\right) \tag{9.2.13}$$

where the first order corrector u_1 *can be written as*

$$u_1(t,x,\tau,y) = \widetilde{\omega}(\tau,x,y)\cdot {}^\top\widetilde{J}(\tau,x)\nabla_x u_0(t,x) \tag{9.2.14}$$

and the zeroth order term u_0 *satisfies the homogenized diffusion equation*

$$\frac{\partial u_0}{\partial t} = \nabla_x\cdot\left(\mathfrak{D}(x)\nabla_x u_0\right) \qquad \text{for } (t,x)\in\,]0,T[\times\mathbb{R}^d, \tag{9.2.15a}$$

$$u_0(0,x) = u^{in}(x) \qquad \text{for } x\in\mathbb{R}^d. \tag{9.2.15b}$$

The effective diffusion coefficient is given by

$$\mathfrak{D}(x) = \lim_{\ell\to\infty}\frac{1}{2\ell}\int_{-\ell}^{+\ell}\widetilde{J}(\tau,x)\mathfrak{B}(\tau,x)^\top\widetilde{J}(\tau,x)\,\mathrm{d}\tau, \tag{9.2.16}$$

where the elements of the matrix $\mathfrak{B}$ *are given by*

$$\begin{aligned}\mathfrak{B}_{ij}(\tau,x) &= \int_{\mathbb{T}^d} \widetilde{\mathbf{D}}(\tau,x,y)\Big(\nabla_y\widetilde{\omega}_j(\tau,x,y)+\mathbf{e}_j\Big)\cdot\Big(\nabla_y\widetilde{\omega}_i(\tau,x,y)+\mathbf{e}_i\Big)\,\mathrm{d}y\\ &+\int_{\mathbb{T}^d}\Big(\widetilde{\mathbf{b}}(\tau,x,y)\cdot\nabla_y\widetilde{\omega}_i(\tau,x,y)\Big)\,\widetilde{\omega}_j(\tau,x,y)\,\mathrm{d}y\\ &+\int_{\mathbb{T}^d}\widetilde{\mathbf{D}}(\tau,x,y)\nabla_y\widetilde{\omega}_j(\tau,x,y)\cdot\mathbf{e}_i\,\mathrm{d}y-\int_{\mathbb{T}^d}\widetilde{\mathbf{D}}(\tau,x,y)\nabla_y\widetilde{\omega}_i(\tau,x,y)\cdot\mathbf{e}_j\,\mathrm{d}y\end{aligned} \tag{9.2.17}$$

for $i,j\in\{1,\cdots,d\}$*. Furthermore, the components of* ω *satisfy the cell problem*

$$\mathbf{b}(x,y)\cdot\Big(\nabla_y\omega_i+\mathbf{e}_i\Big)-\nabla_y\cdot\Big(\mathbf{D}(x,y)\Big(\nabla_y\omega_i+\mathbf{e}_i\Big)\Big)=\bar{\mathbf{b}}(x)\cdot\mathbf{e}_i\qquad\text{in }\mathbb{T}^d, \tag{9.2.18}$$

for each $i\in\{1,\cdots,d\}$ *and with the standard canonical basis* $(\mathbf{e}_i)_{1\le i\le d}$ *in* $\mathbb{R}^d$*.*

Remark 9.2.2. *Even though we postulate an infinite sum in the asymptotic expansion* (9.2.8)*, we compute only the zeroth and first order coefficients as in* (9.2.13)*. Our goal is to obtain an evolution equation for the zeroth order approximation, i.e. the homogenized equation* (9.2.15a)*. For this purpose, the first order approximation* (9.2.13) *suffices. Continuing the expansion for higher order coefficients in the asymptotic expansion* (9.2.8) *in the spirit of the theory of matching asymptotics is out of the scope of this present article.*

Remark 9.2.3. *The dispersion effects are evident from the expression* (9.2.16) *of the effective diffusion in the homogenized equation, i.e. the effective diffusion coefficients depend on the convective velocity. This is because of the strong convection in the scaled convection-diffusion equation* (9.2.4a)*.*

Remark 9.2.4. *The cell problem* (9.2.18) *in Proposition 9.2.1 is given in fixed spatial coordinate, i.e.* $\omega\equiv\omega(x,y)$*. As our analysis essentially considers the asymptotic expansion in moving coordinates along flows, we can recast the cell problem* (9.2.18) *along flows, i.e. for the flow representation of the cell solutions*

$\widetilde{\omega}(\tau, x, y)$:

$$\begin{aligned}\widetilde{\mathbf{b}}(\tau, x, y) \cdot \Big(\nabla_y \widetilde{\omega}_i(\tau, x, y) + \mathbf{e}_i\Big) \\ - \nabla_y \cdot \Big(\widetilde{\mathbf{D}}(\tau, x, y)\left(\nabla_y \widetilde{\omega}_i(\tau, x, y) + \mathbf{e}_i\right)\Big) = \bar{\widetilde{\mathbf{b}}}(\tau, x) \cdot \mathbf{e}_i.\end{aligned} \tag{9.2.19}$$

The above problem is posed on $\mathbb{T}^d$. The spatial variable x and the fast time variable τ are treated as parameters.

Remark 9.2.5. *Even though the molecular diffusion matrix $\mathbf{D}(x, y)$ is assumed to be symmetric, the effective diffusion matrix in the homogenized matrix is not symmetric as is evident from the expression (9.2.17) for $\mathfrak{B}(\tau, x)$. The symmetric part of $\mathfrak{B}$ is given by*

$$\mathfrak{B}_{ij}^{\mathrm{sym}} = \int_{\mathbb{T}^d} \widetilde{\mathbf{D}}(\tau, x, y)\Big(\nabla_y \widetilde{\omega}_j(\tau, x, y) + \mathbf{e}_j\Big) \cdot \Big(\nabla_y \widetilde{\omega}_i(\tau, x, y) + \mathbf{e}_i\Big)\, \mathrm{d}y$$

and the skew-symmetric part of the matrix $\mathfrak{B}$ is given by

$$\begin{aligned}\mathfrak{B}_{ij}^{\mathrm{asym}} = &\int_{\mathbb{T}^d} \widetilde{\omega}_j(\tau, x, y)\left(\bar{\widetilde{\mathbf{b}}}(\tau, x) - \widetilde{\mathbf{b}}(\tau, x, y)\right) \cdot \mathbf{e}_i \, \mathrm{d}y \\ &- \int_{\mathbb{T}^d} \widetilde{\mathbf{D}}(\tau, x, y) \nabla_y \widetilde{\omega}_i(\tau, x, y) \cdot \Big(\nabla_y \widetilde{\omega}_j(\tau, x, y) + \mathbf{e}_j\Big)\, \mathrm{d}y,\end{aligned}$$

where we have used the cell problem for flow representations (9.2.19) to arrive at the above simplified expression for the skew-symmetric part. The contribution of the non-symmetric part of the effective diffusion to the dynamics of the homogenized equation (9.2.15a) is because of the fact that the effective diffusion coefficient $\mathfrak{D}$ is space dependent. In a purely periodic setting, i.e. when $\mathbf{b}(x, y) \equiv \mathbf{b}(y)$ and $\mathbf{D}(x, y) \equiv \mathbf{D}(y)$, the skew-symmetric part of the effective diffusion matrix does not contribute to the dynamics of the homogenized equation.

Remark 9.2.6. *The expression (9.2.16) for the effective diffusion involves the averaging in the fast time variable. In this section dealing with formal derivation of the homogenized limit, we admit that the limits in the expression of the effective diffusion exist and are finite. In Section 9.3, we introduce a notion of weak convergence in some Lebesgue function spaces which proves that these limits indeed exist and are finite under certain assumptions on the coefficients. Note that some*

of these assumptions are required: in Counterexample 9.5.2 in Section 9.5, we provide an explicit example where these limits do not exist, and in fact multiple limit equations can be obtained on different sequences $\varepsilon \to 0$.

Remark 9.2.7. *An interesting feature in the expression (9.2.16) is that the integrands are all in their flow representations. This suggests that the effective diffusion is the cumulative effect of the convection and diffusion effects averaged along the flows.*

Proof of Proposition 9.2.1. The equations at different orders of ε obtained by inserting the asymptotic expansion (9.2.8) in the scaled equation (9.2.4a) are

$$\begin{aligned}
\mathcal{O}(\varepsilon^{-2}): \quad & \widetilde{\mathbf{b}} \cdot \nabla_y u_0 - \nabla_y \cdot \left(\widetilde{\mathbf{D}} \nabla_y u_0\right) && = 0, \\
\mathcal{O}(\varepsilon^{-1}): \quad & \widetilde{\mathbf{b}} \cdot \nabla_y u_1 - \nabla_y \cdot \left(\widetilde{\mathbf{D}} \nabla_y u_1\right) && = \nabla_y \cdot \left(\widetilde{\mathbf{D}}^{\top} \widetilde{J} \nabla_x u_0\right) + {}^{\top}\widetilde{J} \nabla_x \cdot \left(\widetilde{\mathbf{D}} \nabla_y u_0\right) \\
& && \quad + \left(\widetilde{\widetilde{\mathbf{b}}} - \widetilde{\mathbf{b}}\right) \cdot \left({}^{\top}\widetilde{J} \nabla_x u_0\right) - \frac{\partial u_0}{\partial \tau}, \\
\mathcal{O}(\varepsilon^{0}): \quad & \widetilde{\mathbf{b}} \cdot \nabla_y u_2 - \nabla_y \cdot \left(\widetilde{\mathbf{D}} \nabla_y u_2\right) && = -\frac{\partial u_0}{\partial t} - \frac{\partial u_1}{\partial \tau} + \left(\widetilde{\widetilde{\mathbf{b}}} - \widetilde{\mathbf{b}}\right) \cdot \left({}^{\top}\widetilde{J} \nabla_x u_1\right) \\
& && \quad + {}^{\top}\widetilde{J} \nabla_x \cdot \left(\widetilde{\mathbf{D}} \left({}^{\top}\widetilde{J} \nabla_x u_0 + \nabla_y u_1\right)\right) \\
& && \quad + \nabla_y \cdot \left(\widetilde{\mathbf{D}}^{\top} \widetilde{J} \nabla_x u_1\right),
\end{aligned} \tag{9.2.20}$$

where the flow representation of the Jacobian matrix and coefficients are used. Note that the relation (9.2.10) is needed, for example, to show the right hand side of the $\mathcal{O}(\varepsilon^{-2})$ equation is zero. We remark that all the equations in (9.2.20) have the same structure as the boundary value problem (9.2.11) addressed in Lemma 9.2.2 which says that the solvability of these equations is subject to satisfying the compatibility condition (9.2.12).

The compatibility condition (9.2.12) is trivially satisfied for the equation of $\mathcal{O}(\varepsilon^{-2})$ in (9.2.20). Further, the equation of $\mathcal{O}(\varepsilon^{-2})$ in (9.2.20) implies that u_0 is independent of y, i.e.

$$u_0(t, \tau, x, y) \equiv u_0(t, \tau, x).$$

So the term involving $\nabla_y u_0$ in the equation of $\mathcal{O}(\varepsilon^{-1})$ vanishes. To check if the right hand side of the equation of $\mathcal{O}(\varepsilon^{-1})$ in (9.2.20) satisfies the compatibility

condition (9.2.12), consider

$$\int_{\mathbb{T}^d} \nabla_y \cdot \left(\widetilde{\mathbf{D}}(\tau, x, y)^\top \widetilde{J} \nabla_x u_0 \right) \mathrm{d}y \\ + \int_{\mathbb{T}^d} \left(\widetilde{\bar{\mathbf{b}}}(\tau, x) - \widetilde{\mathbf{b}}(\tau, x, y) \right) \cdot \left({}^\top \widetilde{J} \nabla_x u_0 \right) \mathrm{d}y - \int_{\mathbb{T}^d} \frac{\partial u_0}{\partial \tau} \mathrm{d}y.$$

The first integral in the previous expression vanishes by integration by parts. The second integral in the previous expression vanishes as well, thanks to the definition (9.2.2) of the mean field $\bar{\mathbf{b}}(x)$, and as neither J nor u_0 depend upon y. In the third integral, since u_0 is independent of the y variable, in order to satisfy the compatibility condition, we should have that u_0 is independent of the fast time variable. Hence we have $u_0(t, \tau, x, y) \equiv u_0(t, x)$.

The linearity of equations in (9.2.20) implies that we can separate the variables in the first order corrector as in (9.2.14). The function $\widetilde{\omega}(\tau, x, y)$ is the flow representation of the function $\omega = (\omega_i)_{1 \leq i \leq d}$ whose components solve the cell problem (9.2.19) (see Remark 9.2.4).

Finally, we write the compatibility condition for the equation of $\mathcal{O}(\varepsilon^0)$ in (9.2.20):

$$\int_{\mathbb{T}^d} \frac{\partial u_0}{\partial t} \mathrm{d}y + \int_{\mathbb{T}^d} \frac{\partial u_1}{\partial \tau} \mathrm{d}y \\ = \int_{\mathbb{T}^d} \left(\widetilde{\bar{\mathbf{b}}}(\tau, x) - \widetilde{\mathbf{b}}(\tau, x, y) \right) \cdot \left({}^\top \widetilde{J} \nabla_x u_1 \right) \mathrm{d}y \\ + \int_{\mathbb{T}^d} {}^\top \widetilde{J} \nabla_x \cdot \left(\widetilde{\mathbf{D}}(\tau, x, y) \left({}^\top \widetilde{J} \nabla_x u_0 + \nabla_y u_1 \right) \right) \mathrm{d}y.$$

The previous expression contains terms that depend on the fast time variable τ. We propose to average the above equation in the τ variable. The left hand side becomes:

$$\frac{\partial u_0}{\partial t} + \lim_{\ell \to \infty} \frac{1}{2\ell} \int_{-\ell}^{+\ell} \int_{\mathbb{T}^d} \frac{\partial u_1}{\partial \tau} \mathrm{d}y \, \mathrm{d}\tau, \tag{9.2.21}$$

and the right hand side averages to

$$\begin{aligned}
&\lim_{\ell\to\infty}\frac{1}{2\ell}\int_{-\ell}^{+\ell}\int_{\mathbb{T}^d}\left(\widetilde{\bar{\mathbf{b}}}(\tau,x)-\widetilde{\mathbf{b}}(\tau,x,y)\right)\cdot\left({}^{\top}\widetilde{J}(\tau,x)\nabla_x u_1\right)\,\mathrm{d}y\,\mathrm{d}\tau\\
&+\lim_{\ell\to\infty}\frac{1}{2\ell}\int_{-\ell}^{+\ell}\int_{\mathbb{T}^d}{}^{\top}\widetilde{J}(\tau,x)\nabla_x\cdot\left(\widetilde{\mathbf{D}}(\tau,x,y)\left({}^{\top}\widetilde{J}(\tau,x)\nabla_x u_0\right)\right)\,\mathrm{d}y\,\mathrm{d}\tau \qquad (9.2.22)\\
&+\lim_{\ell\to\infty}\frac{1}{2\ell}\int_{-\ell}^{+\ell}\int_{\mathbb{T}^d}{}^{\top}\widetilde{J}(\tau,x)\nabla_x\cdot\left(\widetilde{\mathbf{D}}(\tau,x,y)\nabla_y u_1\right)\,\mathrm{d}y\,\mathrm{d}\tau.
\end{aligned}$$

The second term on in (9.2.21) is zero. The first term in (9.2.22) can be successively written as

$$\begin{aligned}
&\lim_{\ell\to\infty}\frac{1}{2\ell}\int_{-\ell}^{+\ell}\int_{\mathbb{T}^d}\left(\widetilde{\bar{\mathbf{b}}}(\tau,x)-\widetilde{\mathbf{b}}(\tau,x,y)\right)\cdot\left({}^{\top}\widetilde{J}(\tau,x)\nabla_x u_1\right)\,\mathrm{d}y\,\mathrm{d}\tau\\
&=\lim_{\ell\to\infty}\frac{1}{2\ell}\int_{-\ell}^{+\ell}\int_{\mathbb{T}^d}\widetilde{J}(\tau,x)\left(\widetilde{\bar{\mathbf{b}}}(\tau,x)-\widetilde{\mathbf{b}}(\tau,x,y)\right)\\
&\qquad\qquad\cdot\nabla_x\left(\widetilde{\omega}(\tau,x,y)\cdot{}^{\top}\widetilde{J}(\tau,x)\nabla_x u_0(t,x)\right)\,\mathrm{d}y\,\mathrm{d}\tau
\end{aligned}$$

which is equal to

$$\begin{aligned}
&\nabla_x\cdot\lim_{\ell\to\infty}\frac{1}{2\ell}\int_{-\ell}^{+\ell}\int_{\mathbb{T}^d}\widetilde{J}(\tau,x)\left(\widetilde{\bar{\mathbf{b}}}(\tau,x)-\widetilde{\mathbf{b}}(\tau,x,y)\right)\\
&\qquad\qquad{}^{\top}\widetilde{\omega}(\tau,x,y)^{\top}\widetilde{J}(\tau,x)\nabla_x u_0(t,x)\,\mathrm{d}y\,\mathrm{d}\tau,
\end{aligned}$$

where we are able to move the x derivative thanks to Lemma 9.2.1.(ii). The second term in (9.2.22) evaluates to

$$\begin{aligned}
&\lim_{\ell\to\infty}\frac{1}{2\ell}\int_{-\ell}^{+\ell}\int_{\mathbb{T}^d}{}^{\top}\widetilde{J}(\tau,x)\nabla_x\cdot\left(\widetilde{\mathbf{D}}(\tau,x,y)\left({}^{\top}\widetilde{J}(\tau,x)\nabla_x u_0\right)\right)\,\mathrm{d}y\,\mathrm{d}\tau\\
&=\nabla_x\cdot\lim_{\ell\to\infty}\frac{1}{2\ell}\int_{-\ell}^{+\ell}\int_{\mathbb{T}^d}\widetilde{J}(\tau,x)\widetilde{\mathbf{D}}(\tau,x,y)^{\top}\widetilde{J}(\tau,x)\nabla_x u_0\,\mathrm{d}y\,\mathrm{d}\tau.
\end{aligned}$$

The third term in (9.2.22) can be successively written as

$$\begin{aligned}
&\lim_{\ell\to\infty}\frac{1}{2\ell}\int_{-\ell}^{+\ell}\int_{\mathbb{T}^d}{}^{\top}\widetilde{J}(\tau,x)\nabla_x\cdot\left(\widetilde{\mathbf{D}}(\tau,x,y)\nabla_y u_1\right)\,\mathrm{d}y\,\mathrm{d}\tau\\
&=\nabla_x\cdot\lim_{\ell\to\infty}\frac{1}{2\ell}\int_{-\ell}^{+\ell}\int_{\mathbb{T}^d}\widetilde{J}(\tau,x)\widetilde{\mathbf{D}}(\tau,x,y)\nabla_y\left(\widetilde{\omega}(\tau,x,y)\cdot{}^{\top}\widetilde{J}(\tau,x)\nabla_x u_0(t,x)\right)\,\mathrm{d}y\,\mathrm{d}\tau\\
&=\nabla_x\cdot\lim_{\ell\to\infty}\frac{1}{2\ell}\int_{-\ell}^{+\ell}\int_{\mathbb{T}^d}\widetilde{J}(\tau,x)\widetilde{\mathbf{D}}(\tau,x,y)^{\top}\nabla_y\widetilde{\omega}(\tau,x,y)^{\top}\widetilde{J}(\tau,x)\nabla_x u_0(t,x)\,\mathrm{d}y\,\mathrm{d}\tau.
\end{aligned}$$

Again, we are allowed to move the x derivative past the Jacobian because of Lemma 9.2.1.(i). Considering all the above observations, the compatibility condition (9.2.21)=(9.2.22) can be rewritten as a diffusion equation for $u_0(t,x)$, i.e. (9.2.15a)-(9.2.15b). The expression of the effective diffusion coefficient is given by

$$\begin{aligned}
\mathfrak{D}(x)=&\lim_{\ell\to\infty}\frac{1}{2\ell}\int_{-\ell}^{+\ell}\int_{\mathbb{T}^d}\widetilde{J}(\tau,x)\left(\widetilde{\bar{\mathbf{b}}}(\tau,x)-\widetilde{\mathbf{b}}(\tau,x,y)\right){}^{\top}\widetilde{\omega}(\tau,x,y)^{\top}\widetilde{J}(\tau,x)\,\mathrm{d}y\,\mathrm{d}\tau\\
&+\lim_{\ell\to\infty}\frac{1}{2\ell}\int_{-\ell}^{+\ell}\int_{\mathbb{T}^d}\widetilde{J}(\tau,x)\widetilde{\mathbf{D}}(\tau,x,y)^{\top}\widetilde{J}(\tau,x)\,\mathrm{d}y\,\mathrm{d}\tau\\
&+\lim_{\ell\to\infty}\frac{1}{2\ell}\int_{-\ell}^{+\ell}\int_{\mathbb{T}^d}\widetilde{J}(\tau,x)\widetilde{\mathbf{D}}(\tau,x,y)^{\top}\nabla_y\widetilde{\omega}(\tau,x,y)^{\top}\widetilde{J}(\tau,x)\,\mathrm{d}y\,\mathrm{d}\tau.
\end{aligned}$$

Moving the y integration inside, the expression for the effective diffusion becomes

$$\begin{aligned}
&\mathfrak{D}(x)=\\
&\lim_{\ell\to\infty}\frac{1}{2\ell}\int_{-\ell}^{+\ell}\widetilde{J}(\tau,x)\left(\int_{\mathbb{T}^d}\left(\widetilde{\bar{\mathbf{b}}}(\tau,x)-\widetilde{\mathbf{b}}(\tau,x,y)\right){}^{\top}\widetilde{\omega}(\tau,x,y)\,\mathrm{d}y\right){}^{\top}\widetilde{J}(\tau,x)\,\mathrm{d}\tau\\
&+\lim_{\ell\to\infty}\frac{1}{2\ell}\int_{-\ell}^{+\ell}\widetilde{J}(\tau,x)\left(\int_{\mathbb{T}^d}\left\{\widetilde{\mathbf{D}}(\tau,x,y)+\widetilde{\mathbf{D}}(\tau,x,y)^{\top}\nabla_y\widetilde{\omega}(\tau,x,y)\right\}\mathrm{d}y\right){}^{\top}\widetilde{J}(\tau,x)\,\mathrm{d}\tau\\
&=\lim_{\ell\to\infty}\frac{1}{2\ell}\int_{-\ell}^{+\ell}\widetilde{J}(\tau,x)\mathfrak{B}(\tau,x)^{\top}\widetilde{J}(\tau,x)\,\mathrm{d}\tau,
\end{aligned}$$

where the elements of $\mathfrak{B}$ are given by

$$\mathfrak{B}_{ij}(\tau,x)=\int\limits_{\mathbb{T}^d}\Big\{\Big(\widetilde{\overline{\mathbf{b}}}_i(\tau,x)-\widetilde{\mathbf{b}}_i(\tau,x,y)\Big)\widetilde{\omega}_j(\tau,x,y)+\widetilde{\mathbf{D}}_{ij}(\tau,x,y)\\ +\widetilde{\mathbf{D}}(\tau,x,y)\nabla_y\widetilde{\omega}_j(\tau,x,y)\cdot\mathbf{e}_i\Big\}\,\mathrm{d}y$$

for each $i,j\in\{1,\cdots,d\}$. To simplify the expression for the matrix $\mathfrak{B}$, we test the equation (9.2.19) for the flow-representation $\widetilde{\omega}_i$ by $\widetilde{\omega}_j$ and deduce

$$\int\limits_{\mathbb{T}^d}\Big(\widetilde{\overline{\mathbf{b}}}_i(\tau,x)-\widetilde{\mathbf{b}}_i(\tau,x,y)\Big)\widetilde{\omega}_j(\tau,x,y)\,\mathrm{d}y=\\ \int\limits_{\mathbb{T}^d}\Big(\widetilde{\mathbf{b}}(\tau,x,y)\cdot\nabla_y\widetilde{\omega}_i(\tau,x,y)\Big)\,\widetilde{\omega}_j(\tau,x,y)\,\mathrm{d}y\\ +\int\limits_{\mathbb{T}^d}\widetilde{\mathbf{D}}(\tau,x,y)\nabla_y\widetilde{\omega}_j(\tau,x,y)\cdot\nabla_y\widetilde{\omega}_i(\tau,x,y)\,\mathrm{d}y\\ +\int\limits_{\mathbb{T}^d}\widetilde{\mathbf{D}}(\tau,x,y)\nabla_y\widetilde{\omega}_j(\tau,x,y)\cdot\mathbf{e}_i\,\mathrm{d}y.$$

Using the above equation, we can rewrite the elements of the matrix $\mathfrak{B}$ as in (9.2.17). □

Remark 9.2.8. *The solution ω_i to the cell problem (9.2.18) is unique up to addition of constants in the y-variable, i.e. up to addition of a function $\eta(t,\tau,x)$. However, any such function would not contribute to the expression of the effective diffusion. It is evident from the equation (9.2.21)-(9.2.22). So, for our purposes at hand, we shall not dwell on characterizing $\eta(t,\tau,x)$. It should be noted that the first order corrector obtained in (9.2.13) essentially is considering the oscillations in the space variable. We have not characterized the first order corrector with regard to the fast time variable. This shall be the focus of future publications.*

Proposition 9.2.2. *The homogenized equation (9.2.15a)-(9.2.15b) has a unique solution such that*

$$u_0(t,x)\in C([0,T];L^2(\mathbb{R}^d));\qquad \nabla_x u_0(t,x)\in[L^2((0,T)\times\mathbb{R}^d)]^d.$$

Proof. The elements of the symmetric part of the matrix $\mathfrak{B}$ are given by

$$\mathfrak{B}_{ij}^{\mathrm{sym}}(\tau, x) = \int_{\mathbb{T}^d} \widetilde{\mathbf{D}}(\tau, x, y)\Big(\nabla_y \widetilde{\omega}_j(\tau, x, y) + \mathbf{e}_j\Big) \cdot \Big(\nabla_y \widetilde{\omega}_i(\tau, x, y) + \mathbf{e}_i\Big)\, \mathrm{d}y.$$

It is positive definite because, for all $\xi \in \mathbb{R}^d$ we have

$$^{\top}\xi\, \mathfrak{B}^{\mathrm{sym}}\, \xi \geq \lambda \int_{\mathbb{T}^d} |\nabla_y \widetilde{\omega}_\xi + \xi|^2\, \mathrm{d}y = \lambda \int_{\mathbb{T}^d} |\nabla \widetilde{\omega}_\xi|^2 + 2\xi \cdot \nabla_y \widetilde{\omega}_\xi + |\xi|^2\, \mathrm{d}y \geq \lambda |\xi|^2,$$

where $\widetilde{\omega}_\xi := \widetilde{\omega} \cdot \xi$, and the last inequality follows from the vanishing of the second of the three terms in the integrand due to y-periodicity of $\widetilde{\omega}$.

Take the effective diffusion matrix $\mathfrak{D}$ and consider, for $\xi \neq 0$,

$$\begin{aligned}
^{\top}\xi\, \mathfrak{D}\, \xi &= \lim_{\ell \to \infty} \frac{1}{2\ell} \int_{-\ell}^{+\ell} {}^{\top}\xi \widetilde{J}(\tau, x) \mathfrak{B}(\tau, x)^{\top} \widetilde{J}(\tau, x)\, \xi\, \mathrm{d}\tau \\
&= \lim_{\ell \to \infty} \frac{1}{2\ell} \int_{-\ell}^{+\ell} {}^{\top}\Big[{}^{\top}\widetilde{J}(\tau, x)\, \xi\Big] \mathfrak{B}(\tau, x)^{\top} \widetilde{J}(\tau, x)\, \xi\, \mathrm{d}\tau \\
&\geq \lambda \lim_{\ell \to \infty} \frac{1}{2\ell} \int_{-\ell}^{+\ell} \Big|{}^{\top}\widetilde{J}(\tau, x)\xi\Big|^2\, \mathrm{d}\tau \geq C\lambda |\xi|^2 > 0.
\end{aligned}$$

This holds, thanks firstly to the positive definite property of the matrix $\mathfrak{B}$, and secondly to uniform bounds from below on the Jacobian matrix, i.e.

$$\begin{aligned}
C^{-1}|\xi| &= C^{-1}|J(\tau, x)^{-1} J(\tau, x)\xi| \\
&= C^{-1}|J(-\tau, \Phi_{-\tau}(x)) J(\tau, x)\xi| \leq |J(\tau, x)\xi| \leq C|\xi|
\end{aligned}$$

for the uniform constant C given by Assumption 9.2.1, and where we have used that the flow is autonomous to express the inverse of the Jacobian in terms of the Jacobian at a different point. Thus we have shown that the effective diffusion coefficient $\mathfrak{D}(x)$ is positive definite. On the other hand, $\mathfrak{D}(x)$ is uniformly bounded from above. Then, it is a standard process to prove existence and uniqueness for (9.2.15a)-(9.2.15b) (cf. [118] if necessary). □

9.3 Σ-convergence along flows

This section puts forth a new notion of convergence in L^p-spaces (with $1 < p < \infty$) which gives a rigorous justification of (at least) the first two terms in the asymptotic expansion along mean flows (9.2.8) postulated in Section 9.2, i.e. to justify the approximation (9.2.13) in Proposition 9.2.1. This work is inspired from the seminal works of G. Nguetseng [154] and G. Allaire [1]. In Section 9.2, we have formally derived the homogenized limit and obtained an explicit expression for the effective diffusion (9.2.16). As mentioned in Remark 9.2.6, there was an inherent assumption that the limits in the fast time variable exist and are finite.

The works [154, 1] are in the context of periodic homogenization. G. Allaire does mention in [1] that it would be interesting to extend the two-scale convergence theory from the periodic setting to the more general almost-periodic setting (see p.1484 in [1]). This has been addressed in the past one and a half decade [35, 155, 156, 158, 176]. In all these new developments, a central role is played by the notion of *algebra with mean value* introduced by Zhikov and Krivenko in [206].

In this section we present the abstract framework of Σ-*convergence along flows*. In subsections 9.3.1-9.3.5 we develop enough of the theory of *algebras with mean value* for our later purposes. As we do not aim to extend this theory beyond what already exists, we shall not give the theory in full generality and we refer the reader to existing literature (e.g. [35, 155, 156, 157, 176, 8], see also [67] for an introductory exposition and [205] for a pedagogical exposition) for a more complete presentation and full proofs. In subsections 9.3.6-9.3.8 we introduce the new concept of Σ-*convergence along flows* and prove compactness results.

9.3.1 Algebras with mean value

We shall denote the space of bounded uniformly continuous functions on $\mathbb{R}$ by $\mathrm{BUC}(\mathbb{R})$.

Definition 9.3.1 (Algebra with mean value)**.** *An* algebra with mean value *(or*

algebra w.m.v., *in short) is a Banach sub-algebra* $\mathcal{A}$ *of* $\mathrm{BUC}(\mathbb{R})$ *such that the following hold:*

(i) $\mathcal{A}$ *contains the constants.*

(ii) $\mathcal{A}$ *is translation invariant, i.e. for every* $f \in \mathcal{A}$ *and* $a \in \mathbb{R}$, $f(\cdot - a) \in \mathcal{A}$.

(iii) Any $f \in \mathcal{A}$ *possesses a mean value* $M(f)$*, by which we mean that*

$$f\left(\frac{\cdot}{\varepsilon}\right) \rightharpoonup M(f) \text{ in } L^\infty(\mathbb{R})\text{-weak* as } \varepsilon \to 0.$$

Note that the mean value can be equivalently expressed as

$$M(f) = \lim_{\ell \to \infty} \frac{1}{2\ell} \int_{-\ell}^{+\ell} f(\tau)\, \mathrm{d}\tau$$

and that this limit exists for any $f \in \mathcal{A}$.

The theory of algebra w.m.v. is developed for the Banach space of bounded uniformly continuous functions on $\mathbb{R}^d$, i.e. in any arbitrary dimension. As this current work considers a fast time variable (i.e. in one dimension), we recall all the essential notions in this theory with emphasis on one dimension.

9.3.2 Gelfand representation theory

Definition 9.3.2 (Spectrum of a Banach algebra)**.** *Given a commutative Banach algebra* $\mathcal{A}$ *with an identity* $1 \in \mathcal{A}$*, we define its* spectrum $\Delta(\mathcal{A})$ *as the set of algebra homeomorphisms, i.e. the maps* $s : \mathcal{A} \to \mathbb{C}$ *such that*

1. *s is linear, i.e. for all* $f, g \in \mathcal{A}, \lambda \in \mathbb{C}$, $s(f+g) = s(f) + s(g)$ *and* $s(\lambda f) = \lambda s(f)$,
2. *s is multiplicative, i.e. for all* $f, g \in \mathcal{A}$, $s(fg) = s(f)s(g)$,
3. *s preserves the identity, i.e.* $s(1) = 1$.

The elements $s \in \Delta(\mathcal{A})$ *are called the* characters *of* $\mathcal{A}$.

As $\Delta(\mathcal{A}) \subset \mathcal{A}'$, the topological dual of $\mathcal{A}$, we equip $\Delta(\mathcal{A})$ with the weak* subspace topology induced by $\mathcal{A}'$. This makes $\Delta(\mathcal{A})$ a compact Hausdorff space by the Banach-Alaoglu theorem.

Of central importance to the study of Banach algebras is the Gelfand transform. We denote by $C(\Delta(\mathcal{A}))$, the space of complex-valued continuous functions on $\Delta(\mathcal{A})$.

Definition 9.3.3 (Gelfand transform)**.** *The* Gelfand transform *is the map* $\mathcal{G} : \mathcal{A} \to C(\Delta(\mathcal{A}))$ *defined by* $\mathcal{G}(f)(s) = s(f)$.

NOTATION: For brevity, we denote $\mathcal{G}(f)$ as $\widehat{f}$.

The importance of the spectrum and Gelfand transform is in the following result, which allows us to replace the analysis of functions in $C(\mathbb{R})$ with functions on a *compact* space.

Theorem 9.3.1 (Gelfand-Naimark)**.** *Let* $\mathcal{A}$ *be a* $\mathcal{C}^*$ *algebra. Then* $\mathcal{G}$ *is an isometric isomorphism of* $\mathcal{A}$ *into* $C(\Delta(\mathcal{A}))$.

The mean value operator M is a bounded linear functional on $\mathcal{A}$. By identifying $\mathcal{A}$ with $C(\Delta(\mathcal{A}))$ using the Gelfand transform, and applying the Riesz representation theorem we arrive at the following proposition, the observation of which forms the basis of Nguetseng's formalism of *Homogenization Structures* [155, 156].

Proposition 9.3.1. *Let* $\mathcal{A}$ *be an algebra w.m.v.. Then the mean value operator* M *is represented by a Radon probability measure* β *on* $\Delta(\mathcal{A})$*, i.e. for all* $f \in \mathcal{A}$ *we have:*

$$M(f) = \int_{\Delta(\mathcal{A})} \widehat{f}(s)\, \mathrm{d}\beta(s). \tag{9.3.1}$$

This allows us to introduce the space $L^2(\Delta(\mathcal{A})) := L^2(\Delta(\mathcal{A}), \mathrm{d}\beta)$. It follows from Definition 9.3.1(iii) that β is invariant under the action of translation operator $f(\cdot) \mapsto f(\cdot + t)$ for any $t \in \mathbb{R}$. Note that β may not be supported on the whole of $\Delta(\mathcal{A})$. Indeed, in the Example 9.3.2 below it is a Dirac mass at a single point.

9.3.3 Examples of algebras with mean value

To give some intuition for these objects we provide some examples.

Example 9.3.1 (Periodic functions)**.** *Let $\mathcal{A}$ be the set of continuous functions from $\mathbb{R}$ to $\mathbb{C}$ which are periodic with period L. Then the characters $s \in \Delta(\mathcal{A})$ are the maps defined by $s_t(f) = f(t)$ for $t \in \mathbb{R}/(L\mathbb{Z})$, so that $\Delta(\mathcal{A})$ can be identified with the torus of length L. The Gelfand transform takes $f \in \mathcal{A} \subset C(\mathbb{R})$ to its representative on the torus. The mean value operator M is given by*

$$M(f) = \int_{\Delta(\mathcal{A})} \widehat{f}(s)\,\mathrm{d}\beta(s) = \frac{1}{L}\int_0^L f(\tau)\,\mathrm{d}\tau.$$

Example 9.3.2 (Functions that converge at infinity)**.** *Let $\mathcal{A}$ be the space of continuous functions $f : \mathbb{R} \to \mathbb{C}$ that converge to a limit at infinity, i.e. $\lim_{|\tau|\to\infty} f(\tau)$ exists. Then the spectrum $\Delta(\mathcal{A})$ are the point evaluation maps $s_t(f) = f(t)$ for $t \in \mathbb{R}\cup\{\infty\}$, and the spectrum can be identified with $\bar{\mathbb{R}}$ the one point compactification of $\mathbb{R}$. Under this identification, the Gelfand transform takes a function $f \in \mathcal{A}$ to a function $\widehat{f} : \bar{\mathbb{R}} \to \mathbb{C}$ with $\widehat{f}(t) = f(t)$ for $t \in \mathbb{R}$ and $\widehat{f}(\infty) = \lim_{|\tau|\to\infty} f(\tau)$. The mean value operator M acts by $M(f) = \widehat{f}(\infty)$.*

Example 9.3.3 (Almost-periodic functions)**.** *Let $\mathsf{T}(\mathbb{R})$ denote the set of all trigonometric polynomials, i.e. all $f(t)$ that are finite linear combinations of the functions in the set*

$$\Big\{\cos(kt), \sin(kt) : k \in \mathbb{R}\Big\}.$$

The space of almost-periodic functions in the sense of Bohr [23] is the closure of $\mathsf{T}(\mathbb{R})$ in the supremum norm.

A function $f(t) \in L^2_{loc}(\mathbb{R})$ is called almost-periodic in the sense of Besicovitch if there is a sequence in $\mathsf{T}(\mathbb{R})$ that converges to u in the Besicovitch semi-norm (given by (9.3.3) below).

A function $f(t) \in BUC(\mathbb{R})$ is said to be almost-periodic if the set of translates

$$\Big\{f(\cdot - a) : a \in \mathbb{R}\Big\} \tag{9.3.2}$$

is relatively compact in $BUC(\mathbb{R})$.

All the above three definitions of *almost-periodic functions* are equivalent [205].

We also give the example of weakly almost periodic functions due to Eberlein [53].

Example 9.3.4 (Weakly almost-periodic functions)**.** *A function* $f(t) \in BUC(\mathbb{R})$ *is weakly almost periodic if the set of translates* (9.3.2) *is relatively weakly compact in* $BUC(\mathbb{R})$.

Readers are to consult [176] for more information on the space of weakly almost-periodic functions.

9.3.4 Besicovitch spaces

Definition 9.3.4 (Besicovitch space)**.** *For an algebra w.m.v.* $\mathcal{A}$ *the corresponding* Besicovitch space $\mathcal{B}^2 = \mathcal{B}^2_{\mathcal{A}}$ *is the abstract completion of* $\mathcal{A}$ *with respect to the Besicovitch semi-norm:*

$$\|f\|^2_{\mathcal{B}^2_{\mathcal{A}}} = \limsup_{\ell\to\infty} \frac{1}{2\ell} \int_{-\ell}^{+\ell} |f(\tau)|^2 \, \mathrm{d}\tau. \tag{9.3.3}$$

Note that the elements of $\mathcal{B}^2$ are equivalence classes of functions that are indistinguishable under (9.3.3). The mean value operator M extends to a bilinear form $M(fg)$ on $\mathcal{B}^2_{\mathcal{A}}$. It is a standard result (see e.g. [176]) that the Gelfand transform is an isometric isomorphism between $\mathcal{B}^2$ and $L^2(\Delta(\mathcal{A}))$. Note that $\mathcal{B}^2_{\mathcal{A}}$ inherits the *translation invariance* (in the sense of Definition 9.3.1(iii)) from $\mathcal{A}$.

Definition 9.3.5 (Ergodic algebra w.m.v.)**.** *An algebra w.m.v.* $\mathcal{A}$ *is said to be* ergodic *if any* $f \in \mathcal{B}^2_{\mathcal{A}}$ *satisfying*

$$\|f(\cdot) - f(\cdot - a)\|_{\mathcal{B}^2_{\mathcal{A}}} = 0 \qquad \text{for all } a \in \mathbb{R}$$

is equivalent in $\mathcal{B}^2_{\mathcal{A}}$ *to a constant.*

It is easy to see that the constant in Definition 9.3.5 must be $M(f)$.

Remark 9.3.1. *All of the examples of algebras w.m.v. given in Section 9.3.3 are ergodic.*

For our purposes the importance of *ergodicity* of an algebra w.m.v. is the following lemma, whose proof may be found in [8].

Lemma 9.3.1. *Let $\mathcal{A}$ be an ergodic algebra w.m.v. and $f \in \mathcal{B}^2_{\mathcal{A}}$ have the property that, for any $g \in \mathcal{A}$ with $\frac{\mathrm{d}g}{\mathrm{d}\tau} \in \mathcal{A}$ we have:*

$$M\left(f\frac{\mathrm{d}g}{\mathrm{d}\tau}\right) = \int_{\Delta(\mathcal{A})} \widehat{f}(s)\widehat{\frac{\mathrm{d}g}{\mathrm{d}\tau}}(s)\,\mathrm{d}\beta(s) = 0$$

where the first equality is automatic. Then $f = M(f)$ in $\mathcal{B}^2_{\mathcal{A}}$ and equivalently $\widehat{f} = M(f)$ β-almost everywhere.

9.3.5 Product algebras and vector valued algebras

We wish to consider continuous functions $f(\tau, y)$ for which heuristically 'f is in $\mathcal{A}$ as a function of τ' and 'f is in $C(\mathbb{T}^d)$ as a function of y'. To make sense of this, we recall that the tensor product $\mathcal{A} \otimes C(\mathbb{T}^d)$ is defined by

$$\mathcal{A} \otimes C(\mathbb{T}^d) := \left\{ \sum_{i=1}^{N} f_i g_i : N \in \mathbb{N}, f_1, \ldots, f_N \in \mathcal{A}, \text{ and } g_1, \ldots, g_N \in C(\mathbb{T}^d) \right\}$$

and we define $\mathcal{A} \odot C(\mathbb{T}^d)$ as the closure of $\mathcal{A} \otimes C(\mathbb{T}^d)$ in the Banach algebra $\mathrm{BUC}(\mathbb{R} \times \mathbb{T}^d)$. Note that by construction $\mathcal{A} \otimes C(\mathbb{T}^d)$ is dense in $\mathcal{A} \odot C(\mathbb{T}^d)$. More discussion of product algebras may be found in [155, 156].

We will often need to use vector valued algebras of functions mapping to $\mathbb{C}^d$. This poses essentially no additional complications; we refer the reader to e.g. [8] for details.

9.3.6 Σ-convergence along flows

Throughout this section we shall consider a flow $\Phi_\tau(x) : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$. One might think of $\Phi_\tau(x)$ as the flow of an autonomous ODE

$$\dot{x} = \bar{\mathbf{b}}(x),$$

as was considered in Section 9.2, but this assumption will not be needed in this section. We will, however, make the following assumptions on the flow $\Phi_\tau(x)$.

Assumption 9.3.1. *We assume that the flow $\Phi_\tau(x)$ satisfies the following:*

(i) *$\Phi_\tau(x)$ is continuously differentiable from $\mathbb{R} \times \mathbb{R}^d$ to $\mathbb{R}^d$.*

(ii) *$\Phi_\tau(x)$ satisfies the group property, i.e. $\Phi_t(\Phi_s(x)) = \Phi_{t+s}(x)$ for all $t, s \in \mathbb{R}$ and $x \in \mathbb{R}^d$.*

(iii) *The Jacobian J of $\Phi_\tau(x)$ defined by (9.2.9) is an* uniformly bounded *function of τ, locally uniformly in x, i.e. for any compact $K \subset \mathbb{R}^d$ we have*

$$\sup_{x \in K} \sup_{\tau \in \mathbb{R}} |J(\tau, x)| < \infty.$$

(iv) *For any $\tau \in \mathbb{R}$, $\Phi_\tau(x)$ is volume preserving, i.e. $\det(J(\tau, x)) = 1$.*

We now define the notion of weak Σ-convergence along flows, which generalizes the notion of two-scale convergence with drift introduced in [140] and also the notion of Σ-convergence introduced in [155].

Definition 9.3.6 (weak Σ-convergence along flow)**.** *Let $\mathcal{A}$ be an algebra w.m.v.. Suppose $\Phi_\tau(x)$ be a flow satisfying Assumption 9.3.1 and let $u^\varepsilon(t, x)$ be a sequence in $L^2((0, T) \times \mathbb{R}^d)$. We say that u^ε weakly Σ-converges along $\Phi_\tau(x)$ to a limit $u_0(t, x, s, y) \in L^2((0, T) \times \mathbb{R}^d \times \Delta(\mathcal{A}) \times \mathbb{T}^d)$ if, for any smooth test function $\psi(t, x, \tau, y)$ which is periodic in the y variable and belongs to $\mathcal{A}$ in the τ variable,*

we have

$$\lim_{\varepsilon\to 0}\iint\limits_{(0,T)\times\mathbb{R}^d} u^\varepsilon(t,x)\psi\left(t,\Phi_{-t/\varepsilon}(x),\frac{t}{\varepsilon},\frac{x}{\varepsilon}\right)\mathrm{d}x\,\mathrm{d}t = \iiiint\limits_{(0,T)\times\mathbb{R}^d\times\Delta(\mathcal{A})\times\mathbb{T}^d} u_0(t,x,s,y)\widehat{\psi}(t,x,s,y)\,\mathrm{d}y\,\mathrm{d}\beta(s)\,\mathrm{d}x\,\mathrm{d}t, \tag{9.3.4}$$

where $\widehat{\psi}=\mathcal{G}(\psi)$ *is the Gelfand transform of* ψ *(Definition 9.3.3),* β *is given by* (9.3.1) *and* $\mathcal{A}\odot C(\mathbb{T}^d)$ *is defined in Section 9.3.5.*

NOTATION: We denote the weak Σ-convergence along flow $\Phi_\tau(x)$ by $u^\varepsilon \xrightarrow{\Sigma-\Phi_\tau} u_0$.

CONVENTION: Whenever the limit (9.3.4) holds, we call u_0 *the* Σ-Φ_τ *weak limit of* u^ε.

Remark 9.3.2. *The test functions in* (9.3.4) *are taken along rapidly moving coordinates in their second variable. This is analogous to the choice of test functions in the theory of two-scale convergence with drift [140, 2]. Note also that the the* Σ-Φ_τ *weak limit of the family* $u^\varepsilon(t,x)$ *depends on the choice of the flow* $\Phi_\tau(x)$*. It should be noted that when* $\Phi_\tau(x)=x$ *for all* $\tau\in\mathbb{R}$ *and for each* $x\in\mathbb{R}^d$*, i.e. when the test functions in* (9.3.4) *are taken on a fixed coordinate system, the weak convergence given in Definition 9.3.6 coincides with the notion of weak* Σ*-convergence with regard to the product algebra* $\mathcal{A}\odot C(\mathbb{T}^d)$ *developed in [158, 176].*

Remark 9.3.3. *Definition 9.3.6 makes sense even for test functions* $\psi(t,x,\tau)$ *without oscillations in space, and in this case the limit* u_0 *will be a function of* (t,x,s) *only.*

9.3.7 Compactness

To show that the Definition 9.3.6 is not empty, we give the following weak-compactness result, which is the main result of this section.

Theorem 9.3.2. *Let* $\mathcal{A}$ *be an algebra w.m.v.. Suppose* $\Phi_\tau(x)$ *be a flow satisfying Assumption 9.3.1 and let* $u^\varepsilon(t,x)$ *be a uniformly (with respect to* ε*) bounded sequence in* $L^2((0,T)\times\mathbb{R}^d)$*. Then there exists a subsequence (still denoted* u^ε*) and*

a limit $u_0(t,x,s,y) \in L^2((0,T)\times\mathbb{R}^d\times\Delta(\mathcal{A})\times\mathbb{T}^d)$ *such that*

$$u_\varepsilon \xrightarrow{\Sigma-\Phi_\tau} u_0$$

in the sense of Definition 9.3.6.

To prove the above theorem, we will follow the method of Casado-Díaz and Gayte [35], as we would like to consider algebras which are not separable. To that end, we will need the following result from [35].

Theorem 9.3.3 (Casado-Díaz and Gayte, Theorem 2.1. [35])**.** *Let* X *be a subspace (not necessarily closed) of a reflexive space* Y *and let* $f_n : X \to \mathbb{R}$ *be a sequence of linear functionals (not necessarily continuous). Assume there exists a constant* $C > 0$ *which satisfies*

$$\limsup_{n\to\infty} |f_n(x)| \le C\|x\|, \qquad \forall x \in X.$$

Then there exists a subsequence n_k *and a functional* $f \in Y'$ *such that*

$$\lim_{k\to\infty} f_{n_k}(x) = f(x), \qquad \forall x \in X.$$

We will also need the following lemma, which is the main novel part of the proof.

Lemma 9.3.2. *Let* $\mathcal{A}$ *be an algebra w.m.v. and let* $\Phi_\tau(x)$ *be a flow satisfying Assumption 9.3.1. Take* $\varphi(t,x,\tau,y) \in L^2((0,T)\times\mathbb{R}^d;\mathcal{A}\odot C(\mathbb{T}^d))$*. Then*

$$\lim_{\varepsilon\to 0} \iint_{(0,T)\times\mathbb{R}^d} \left|\varphi\left(t,\Phi_{-t/\varepsilon}(x),\frac{t}{\varepsilon},\frac{x}{\varepsilon}\right)\right|^2 \mathrm{d}x\,\mathrm{d}t = \iiiint_{(0,T)\times\mathbb{R}^d\times\Delta(\mathcal{A})\times\mathbb{T}^d} |\widehat{\varphi}(t,x,s,y)|^2 \,\mathrm{d}y\,\mathrm{d}\beta(s)\,\mathrm{d}x\,\mathrm{d}t.$$

Proof. By density in $L^2((0,T)\times\mathbb{R}^d;\mathcal{A}\odot C(\mathbb{T}^d))$ of functions of the form

$$\sum_{j=1}^N g_j(t)h_j(x)f_j(\tau)e^{in_j\cdot y}$$

where $g_j \in C^\infty(0,T)$, $h_j \in C_c^\infty(\mathbb{R}^d)$, $f_j \in \mathcal{A}$ and $n_j \in \mathbb{Z}^d$, and linearity it suffices

to show that

$$\lim_{\varepsilon\to 0} \iint\limits_{(0,T)\times\mathbb{R}^d} g(t)h\left(\Phi_{-t/\varepsilon}(x)\right) f\left(\frac{t}{\varepsilon}\right) e^{in\cdot x/\varepsilon}\,\mathrm{d}x\,\mathrm{d}t$$
$$= \left(\int_0^T g(t)\,\mathrm{d}t\right)\left(\int_{\mathbb{R}^d} h(x)\,\mathrm{d}x\right)\left(\int_{\Delta(\mathcal{A})} \widehat{f}(s)\,\mathrm{d}\beta(s)\right) 1_{n=0}$$

for g, h, f, n in the spaces above, and $1_{n=0}$ is one when $n = 0$ and zero otherwise.

We first consider $n = 0$, in which case the only x dependence of the integrand is through h. By Fubini's theorem we may do the x integration first. By the coordinate change: $x = \Phi_{-t/\varepsilon}(x)$, which has determinant 1 by the Assumption 9.3.1.(iv), we have

$$\iint\limits_{(0,T)\times\mathbb{R}^d} g(t)h\left(\Phi_{-t/\varepsilon}(x)\right) f\left(\frac{t}{\varepsilon}\right)\,\mathrm{d}x\,\mathrm{d}t$$
$$= \int_0^T g(t) f\left(\frac{t}{\varepsilon}\right)\left(\int_{\mathbb{R}^d} h(x)\,\mathrm{d}x\right)\,\mathrm{d}t = \left(\int_{\mathbb{R}^d} h(x)\,\mathrm{d}x\right)\left(\int_0^T g(t) f\left(\frac{t}{\varepsilon}\right)\,\mathrm{d}t\right),$$

so it suffices to show that the last integral converges to the required limit. By the definition of the mean value operator, $f(\cdot/\varepsilon)$ converges $L^\infty(\mathbb{R})$-weak* to $M(f)$. As $g \in L^1(0,T)$ this completes the proof for $n = 0$, noting the identification of M with β (Proposition 9.3.1).

Now suppose that $n \neq 0$. As in the $n = 0$ case we perform the x integration first with t fixed, but this time we do not change coordinates. Define the (formally) self-adjoint differential operator $L_n = -in \cdot \nabla_x$. Then we have the relation

$$e^{in\cdot x/\varepsilon} = \frac{\varepsilon}{|n|^2} L_n(e^{in\cdot x/\varepsilon}).$$

Substituting this into the x integral and integrating by parts yields

$$\begin{aligned}\int_{\mathbb{R}^d} h\left(\Phi_{-t/\varepsilon}(x)\right) e^{in\cdot x/\varepsilon}\,\mathrm{d}x &= \frac{\varepsilon}{|n|^2}\int_{\mathbb{R}^d} h\left(\Phi_{-t/\varepsilon}(x)\right) L_n(e^{in\cdot x/\varepsilon})\,\mathrm{d}x \\ &= \frac{\varepsilon}{|n|^2}\int_{\mathbb{R}^d} e^{in\cdot x/\varepsilon} L_n\left(h\left(\Phi_{-t/\varepsilon}(x)\right)\right)\,\mathrm{d}x \\ &= \frac{-i\varepsilon}{|n|^2}\int_{\mathbb{R}^d} e^{in\cdot x/\varepsilon} n\cdot {}^{\top}\tilde{J}\left(\frac{t}{\varepsilon},\Phi_{-t/\varepsilon}(x)\right)\nabla_X h(\Phi_{-t/\varepsilon}(x))\,\mathrm{d}x.\end{aligned}$$

Consider the integrand on the last line. By assumption h is smooth with compact support, K say, so by Assumption 9.3.1.(iii) we may estimate

$$\left|\int_{\mathbb{R}^d} {}^{\top}\tilde{J}(t/\varepsilon,\Phi_{-t/\varepsilon}(x))\nabla_X h(\Phi_{-t/\varepsilon}(x))e^{in\cdot x/\varepsilon}\,\mathrm{d}x\right| \le C_K\|\nabla h\|_{L^\infty(\mathbb{R}^d)}|\Phi^{-1}_{-t/\varepsilon}(K)|,$$

where $|\Phi^{-1}_{-t/\varepsilon}(K)|$ is the Lebesgue measure of the set inside the modulus sign. As Φ is volume preserving (Assumption 9.3.1.(iv)) this is equal to the Lebesgue measure of K and is finite. Therefore, the x integral has the bound

$$\left|\int_{\mathbb{R}^d} h(\Phi_{-t/\varepsilon}(x))e^{in\cdot x/\varepsilon}\,\mathrm{d}x\right| \le C\varepsilon$$

for some constant C. Using this bound in the full t,x integral yields

$$\begin{aligned}&\left|\iint_{(0,T)\times\mathbb{R}^d} g(t)h\left(\Phi_{-t/\varepsilon}(x)\right) f\left(\frac{t}{\varepsilon}\right) e^{in\cdot x/\varepsilon}\,\mathrm{d}x\,\mathrm{d}t\right| \\ &\qquad \le C\varepsilon\int_0^T |g(t)||f(t/\varepsilon)|\,\mathrm{d}t \le C\varepsilon\|f\|_{L^\infty(\mathbb{R})}\|g\|_{L^1([0,T])}.\end{aligned}$$

This completes the $n\neq 0$ case and the proof of the lemma. □

Remark 9.3.4. *It is evident from the above proof that the uniform bound upon the Jacobian (Assumption 9.3.1.(iii)) is needed only for test functions that depend upon the fast spatial variable y. As a consequence, an analogous compactness result for convergence against test functions depending only upon (t,x,τ) can be obtained without this assumption (see Remark 9.3.3). However, Assump-*

tion 9.3.1.(iii) is needed to identify the Σ-Φ_τ limit of gradient sequences (Proposition 9.3.3 below).

The weak Σ-convergence along flows is not limited to bounded sequences in L^2. Our main result, Theorem 9.3.2, generalises straightaway to bounded sequences in L^p with $1 < p < +\infty$.

We are now ready to prove the compactness result.

Proof of Theorem 9.3.2. Let $Y = L^2((0,T) \times \mathbb{R}^d \times \Delta(\mathcal{A}) \times \mathbb{T}^d)$ and X be the vector subspace of Gelfand transforms of functions in $L^2((0,T) \times \mathbb{R}^d; \mathcal{A} \odot C(\mathbb{T}^d))$. Now define the linear functionals $F^\varepsilon : X \subset Y \to \mathbb{R}$, by

$$F^\varepsilon(\hat{\varphi}) = \iint\limits_{(0,T)\times\mathbb{R}^d} u^\varepsilon(t,x)\varphi\left(t, \Phi_{-t/\varepsilon}(x), \frac{t}{\varepsilon}, \frac{x}{\varepsilon}\right) \mathrm{d}x\,\mathrm{d}t, \qquad \hat{\varphi} \in X.$$

By the Cauchy-Schwarz inequality, the L^2 boundedness of $\{u^\varepsilon\}$ and Lemma 9.3.2 we have

$$\begin{aligned} |F^\varepsilon(\hat{\varphi})| &\leq \left(\sup_\varepsilon \|u^\varepsilon\|_{L^2((0,T)\times\mathbb{R}^d)}\right) \left\|\varphi\left(t, \Phi_{-t/\varepsilon}(x), \frac{t}{\varepsilon}, \frac{x}{\varepsilon}\right)\right\|_{L^2((0,T)\times\mathbb{R}^d)} \\ &\leq C\|\hat{\varphi}(t,x,y,s)\|_{L^2((0,T)\times\mathbb{R}^d\times\mathbb{T}^d\times\Delta(\mathcal{A}))}. \end{aligned}$$

By Theorem 9.3.3, we may pass to a subsequence (still indexed by ε) for which

$$F^\varepsilon(\hat{\varphi}) \to F(\hat{\varphi}) \text{ as } \varepsilon \to 0, \qquad \forall \varphi \in L^2((0,T) \times \mathbb{R}^d; \mathcal{A} \odot C(\mathbb{T}^d))$$

where $F \in Y'$. Note that $Y = L^2((0,T)\times\mathbb{R}^d\times\Delta(\mathcal{A})\times\mathbb{T}^d)$ is a Hilbert space, (but is in general non-separable). Therefore, by the Riesz representation theorem, F is represented by

$$F(\hat{\varphi}) = \iiiint\limits_{(0,T)\times\mathbb{R}^d\times\Delta(\mathcal{A})\times\mathbb{T}^d} u_0(t,x,s,y)\hat{\varphi}(t,x,s,y)\,\mathrm{d}y\,\mathrm{d}\beta(s)\,\mathrm{d}x\,\mathrm{d}t$$

for some $u_0 \in L^2((0,T) \times \mathbb{R}^d \times \mathbb{T}^d \times \Delta(\mathcal{A}))$, which is the desired limit. □

As is classical in the theory of two-scale convergence, we have the following result shedding some light on the product of two sequences that converge in the sense

of Σ-convergence along flows.

Theorem 9.3.4 (Limit of the product). *Let u^ε and v^ε be two families in $L^2((0,T)\times\mathbb{R}^d)$ such that*

$$u^\varepsilon \xrightarrow{\Sigma-\Phi_\tau} u_0(t,x,s,y); \qquad v^\varepsilon \xrightarrow{\Sigma-\Phi_\tau} v_0(t,x,s,y).$$

Assume further that

$$\lim_{\varepsilon\to 0}\|u^\varepsilon\|_{L^2((0,T)\times\mathbb{R}^d)} = \|u_0\|_{L^2((0,T)\times\mathbb{R}^d\times\Delta(\mathcal{A})\times\mathbb{T}^d)}.$$

Then, we have

$$u^\varepsilon(t,x)\,v^\varepsilon(t,x) \rightharpoonup \iint\limits_{\Delta(\mathcal{A})\times\mathbb{T}^d} u_0(t,x,s,y)\,v_0(t,x,s,y)\,\mathrm{d}\beta(s)\,\mathrm{d}y$$

in the sense of distributions.

The proof of Theorem 9.3.4 is by a density argument. These arguments are similar to the ones found in [1] (see p.1488 in [1] to be precise). As the proof can be given mutatis mutatndis, we skip the proof of Theorem 9.3.4.

Next, we recall the notion of *admissible test functions* given by M. Radu [169] in the context of two-scale convergence:

Definition 9.3.7. *Let $\varphi \in L^2(\Omega\times\mathbb{T}^d)$ be a function that can be approximated by a sequence of functions $\varphi_n \in C^\infty(\Omega; C^\infty(\mathbb{T}^d))$ such that for $n\to\infty$:*

- $\|\varphi_n - \varphi\|_{L^2(\Omega\times\mathbb{T}^d)} \to 0.$
- $\sup\limits_{\varepsilon>0}\left\|(\varphi_n-\varphi)\left(x,\dfrac{x}{\varepsilon}\right)\right\|_{L^2(\Omega)} \to 0.$

Then φ is said to be an admissible test function.

Inspired by the above definition, we introduce the notion of *admissible test functions* suitable for the notion of weak Σ-convergence along flows.

Definition 9.3.8 (Admissible test functions). *A function $\psi(t,x,\tau,y)$ which is periodic in the y variable and belongs to a certain algebra w.m.v. $\mathcal{A}$ in the τ variable is said to be an admissible test function if it can be approximated by*

a sequence of functions $\psi_n(t,x,\tau,y) \in C((0,T)\times\mathbb{R}^d; \mathcal{A}\odot C(\mathbb{T}^d))$ *such that for* $n\to\infty$*:*

- $\left\|\widehat{\psi}-\widehat{\psi}_n\right\|_{L^2((0,T)\times\mathbb{R}^d\times\Delta(\mathcal{A})\times\mathbb{T}^d)}\to 0.$
- $\sup_{\varepsilon>0}\left\|(\psi-\psi_n)\left(t,\Phi_{-t/\varepsilon}(x),\frac{t}{\varepsilon},\frac{x}{\varepsilon}\right)\right\|_{L^2((0,T)\times\mathbb{R}^d)}\to 0.$

The following result says that having coefficients that are 'admissible' in the sense of Definition 9.3.8 enables us to pass to the limit in the product sequence.

Lemma 9.3.3. *Let* $\mathcal{A}$ *be an algebra w.m.v. and* $\Phi_\tau(x)$ *be a flow satisfying Assumption 9.3.1. Let the family* $u^\varepsilon(t,x)\subset L^2((0,T)\times\mathbb{R}^d)$ *be such that*

$$u^\varepsilon \xrightarrow{\Sigma-\Phi_\tau} u_0(t,x,s,y).$$

Finally, let $a(t,x,\tau,y)$ *be admissible in the sense of Definition 9.3.8. Then, for any smooth test function* $\psi(t,x,\tau,y)$ *which is periodic in the* y *variable and which belongs to* $\mathcal{A}$ *as a function of the* τ *variable, we have*

$$\begin{aligned}\lim_{\varepsilon\to 0}\iint_{(0,T)\times\mathbb{R}^d} u^\varepsilon(t,x)a\left(t,\Phi_{-t/\varepsilon}(x),\frac{t}{\varepsilon},\frac{x}{\varepsilon}\right)\psi\left(t,\Phi_{-t/\varepsilon}(x),\frac{t}{\varepsilon},\frac{x}{\varepsilon}\right)\,\mathrm{d}x\,\mathrm{d}t\\ =\iiiint_{(0,T)\times\mathbb{R}^d\times\Delta(\mathcal{A})\times\mathbb{T}^d} u_0(t,x,s,y)\widehat{a}(t,x,s,y)\widehat{\psi}(t,x,s,y)\,\mathrm{d}y\,\mathrm{d}\beta(s)\,\mathrm{d}x\,\mathrm{d}t\end{aligned}$$

The proof of Lemma 9.3.3 is by a density argument (this is inherent in the definition of admissibility). These arguments are similar to the ones found in [169] (see p.6 in [169] to be precise). As the proof can be given mutatis mutatndis, we skip the details.

9.3.8 Additional bounds on derivatives

We first establish conditions under which the Σ-Φ_t limit does not depend upon y. The following result follows the flavour of standard two-scale convergence (see e.g. [154, 1]), where gradient bounds imply that the two-scale limit is independent of the fast spatial variable. Here the proof is slightly complicated by the flow Φ, but

is otherwise the same.

Proposition 9.3.2. *Let $\mathcal{A}$ be an algebra w.m.v., Φ a flow satisfying the Assumption 9.3.1, and $u^\varepsilon \xrightarrow{\Sigma-\Phi_\tau} u_0$ in the sense of Definition 9.3.6. Then if*

$$\sup_\varepsilon \|\nabla u^\varepsilon\|_{L^2((0,T)\times\mathbb{R}^d)} < \infty,$$

then u_0 does not depend on y, i.e. $u_0(t,s,x,y) = u_0(t,s,x)$.

Proof. Let $\Psi(t,x,\tau,y) \in [C_c^1((0,T)\times\mathbb{R}^d\times\mathbb{T}^d;\mathcal{A})]^d$, then by the uniform bound on ∇u^ε in L^2 and Lemma 9.3.2 we have

$$\sup_\varepsilon \iint_{(0,T)\times\mathbb{R}^d} \nabla u^\varepsilon(t,x)\cdot\Psi\left(t,\Phi_{-t/\varepsilon}(x),\frac{t}{\varepsilon},\frac{x}{\varepsilon}\right)\mathrm{d}x\,\mathrm{d}t \le C < \infty \tag{9.3.5}$$

for some constant C depending on Ψ. By integration by parts the integral on the left hand side is equal to

$$-\frac{1}{\varepsilon}\iint_{(0,T)\times\mathbb{R}^d} u^\varepsilon(t,x)\nabla_y\cdot\Psi\left(t,\Phi_{-t/\varepsilon}(x),\frac{t}{\varepsilon},\frac{x}{\varepsilon}\right)\mathrm{d}x\,\mathrm{d}t + \mathcal{O}(1)$$

where the order 1 term comes from the gradient hitting $\Phi_{-t/\varepsilon}(x)$ which are bounded due to Assumption 9.3.1.(iii) on the Jacobian of the flow. Multiplying this by ε, using the convergence $u^\varepsilon \xrightarrow{\Sigma-\Phi_\tau} u_0$ and comparing to the bound (9.3.5), we have

$$\iiiint_{(0,T)\times\mathbb{R}^d\times\Delta(\mathcal{A})\times\mathbb{T}^d} u_0(t,x,s,y)\nabla_y\cdot\widehat{\Psi}(t,x,s,y)\,\mathrm{d}\beta(s)\,\mathrm{d}y\,\mathrm{d}x\,\mathrm{d}t = 0.$$

Noting that, by linearity and as it acts in a different variable, the Gelfand transform commutes with ∇_y, we deduce that u_0 is orthogonal (in the $L^2(\mathbb{T}^d)$ sense) to all y-divergences and is hence independent of y. □

To obtain the $\Sigma-\Phi$ limit of the gradient sequence ∇u^ε, we require that the Jacobian of the flow lie in the algebra.

Proposition 9.3.3 (Two-scale limit for the gradient sequence)**.** *Let $\mathcal{A}$ be an algebra w.m.v., Φ a flow satisfying Assumption 9.3.1 and $J(\tau,\Phi_\tau(x)) \in C(\mathbb{R}^d;\mathcal{A})$.*

Suppose that

$$u^\varepsilon \xrightarrow{\Sigma-\Phi_\tau} u_0 \qquad \textit{and} \qquad \nabla u^\varepsilon \xrightarrow{\Sigma-\Phi_\tau} v_0 \qquad \textit{as } \varepsilon \to 0$$

for a sequence u^ε in the sense of Definition 9.3.6. Then we have

$$v_0 = {}^\top\widehat{\widetilde{J}}(s,x)\nabla_x u_0 + \nabla_y u_1$$

for some $u_1(t,x,s,y) \in L^2(\mathbb{R}_+ \times \mathbb{R}^d \times \Delta(\mathcal{A}); H^1(\mathbb{T}^d))$.

Remark 9.3.5. *The above result differs from the classical result for two-scale convergence (see e.g. [1]) and two-scale convergence with constant drift [140, 2], in the presence of the Jacobian of the flow, which depends on the fast time variable, in the limit. If the flow Φ is taken to be a constant drift flow $\Phi_\tau(x) = x + \mathbf{b}^*\tau$ then the Jacobian is the identity matrix.*

Proof of Proposition 9.3.3. Note that u_0 is independent of y by Proposition 9.3.2. We test against $\Psi(t,x,\tau,y) \in [C_c^1(\mathbb{R}_+ \times \mathbb{R}_d \times \mathbb{T}^d; \mathcal{A})]^d$ which satisfy $\nabla_y \cdot \Psi = 0$. By integration by parts we obtain

$$\begin{aligned}
&\iint_{(0,T)\times\mathbb{R}^d} \nabla u^\varepsilon(t,x) \cdot \Psi\left(t, \Phi_{-t/\varepsilon}(x), \frac{t}{\varepsilon}, \frac{x}{\varepsilon}\right) \mathrm{d}x\,\mathrm{d}t \\
&= -\iint_{(0,T)\times\mathbb{R}^d} u^\varepsilon(t,x)\nabla_x \cdot \left(\Psi\left(t, \Phi_{-t/\varepsilon}(x), \frac{t}{\varepsilon}, \frac{x}{\varepsilon}\right)\right) \mathrm{d}x\,\mathrm{d}t \\
&= -\frac{1}{\varepsilon}\iint_{(0,T)\times\mathbb{R}^d} u^\varepsilon(t,x)\nabla_y \cdot \Psi\left(t, \Phi_{-t/\varepsilon}(x), \frac{t}{\varepsilon}, \frac{x}{\varepsilon}\right) \mathrm{d}x\,\mathrm{d}t \\
&\quad - \iint_{(0,T)\times\mathbb{R}^d} u^\varepsilon(t,x) \sum_{i=1}^d \left({}^\top J\left(\frac{t}{\varepsilon}, x\right)\nabla_x \Psi_i\left(t, \Phi_{-t/\varepsilon}(x), \frac{t}{\varepsilon}, \frac{x}{\varepsilon}\right)\right)_i \mathrm{d}x\,\mathrm{d}t \\
&= -\iint_{(0,T)\times\mathbb{R}^d} u^\varepsilon(t,x) \sum_{i=1}^d \left({}^\top\tilde{J}\left(\frac{t}{\varepsilon}, \Phi_{-t/\varepsilon}(x)\right)\nabla_x \Psi_i\left(t, \Phi_{-t/\varepsilon}(x), \frac{t}{\varepsilon}, \frac{x}{\varepsilon}\right)\right)_i \mathrm{d}x\,\mathrm{d}t.
\end{aligned}$$

By the convergences $\nabla u^\varepsilon \xrightarrow{\Sigma-\Phi_\tau} v_0$ and $u^\varepsilon \xrightarrow{\Sigma-\Phi_\tau} u_0$ we may pass to the limit

the first and last line respectively to obtain

$$\iiint_{(0,T)\times\mathbb{R}^d\times\Delta(\mathcal{A})\times\mathbb{T}^d} v_0(t,x,s,y)\cdot\widehat{\Psi}(t,x,s,y)\,\mathrm{d}\beta(s)\,\mathrm{d}y\,\mathrm{d}x\,\mathrm{d}t$$
$$= -\iiint_{(0,T)\times\mathbb{R}^d\times\Delta(\mathcal{A})\times\mathbb{T}^d} u_0(t,x)\sum_{i=1}^{d}\left(\widehat{{}^{\top}\tilde{J}}(s,x)\widehat{\nabla_X\Psi_i}(t,x,s,y)\right)_i \mathrm{d}\beta(s)\,\mathrm{d}y\,\mathrm{d}x\,\mathrm{d}t$$

The Gelfand transform is with regard to the s-variable. Hence we have the commutation: $\widehat{\nabla_X\Psi_i} = \nabla_X\widehat{\Psi}_i$. This observation and an integration by parts in the x-variable yields

$$0 = \iiint_{(0,T)\times\mathbb{R}^d\times\Delta(\mathcal{A})\times\mathbb{T}^d}\left(v_0(t,x,s,y) - \widehat{{}^{\top}\tilde{J}}(s,x)\nabla_X u_0(t,x)\right)\cdot\widehat{\Psi}(t,x,s,y)\,\mathrm{d}\beta(s)\,\mathrm{d}y\,\mathrm{d}x\,\mathrm{d}t.$$

Thus the bracketed expression in the above integrand is orthogonal (in the $L^2(\mathbb{T}^d)$ sense) to y-divergence free vector fields, and is hence equal to the y-gradient of some function u_1. Basic Fourier analysis in $\mathbb{T}^d$ tells us that u_1 is bounded in $L^2(\mathbb{R}_+\times\mathbb{R}^d\times\Delta(\mathcal{A});H^1(\mathbb{T}^d))$. This completes the proof of the Proposition. □

9.4 Homogenization Result

This section is dedicated to the rigorous derivation of the homogenized equation for (9.2.4a)-(9.2.4b) using the Σ-convergence along flows developed in Section 9.3.

9.4.1 Qualitative analysis

The compactness results (Theorem 9.3.2, Proposition 9.3.3) of previous section demand uniform (with respect to ε) estimates on the solution family $\{u^\varepsilon(t,x)\}$ and on the family of derivatives (in space) of the solution family $\{\nabla u^\varepsilon(t,x)\}$.

Lemma 9.4.1. *Suppose the fluid field $\mathbf{b}(x,y)\in L^\infty(\mathbb{R}^d\times\mathbb{T}^d;\mathbb{R}^d)$ is incompressible in both x and y variables, i.e. satisfying (9.2.1). Suppose the molecular diffusion tensor $\mathbf{D}(x,y)\in L^\infty(\mathbb{R}^d\times\mathbb{T}^d;\mathbb{R}^{d\times d})$ is uniformly coercive, i.e. satisfying (9.2.3).*

Suppose the initial data $u^{in}(x) \in L^2(\mathbb{R}^d)$. Then we have uniform (with respect to ε) a priori estimates on the solutions to (9.2.4a)-(9.2.4b) given by

$$\|u^\varepsilon\|_{L^\infty([0,T];L^2(\mathbb{R}^d))} + \|\nabla u^\varepsilon\|_{L^2((0,T)\times\mathbb{R}^d)} \leq C\|u^{in}\|_{L^2(\mathbb{R}^d)}, \tag{9.4.1}$$

for any arbitrary time $T > 0$. The constant C in (9.4.1) is independent of ε and the time instant T.

As the proof of the above lemma is very classical and follows the energy method, we shall skip the details. Now, we state the following result showing that our model problem (9.2.4a)-(9.2.4b) is well-posed.

Proposition 9.4.1. *Suppose the fluid field $\mathbf{b}(x,y) \in L^\infty(\mathbb{R}^d \times \mathbb{T}^d; \mathbb{R}^d)$ is incompressible in both x and y variables, i.e. satisfying (9.2.1). Suppose further that the diffusion tensor $\mathbf{D}(x,y) \in L^\infty(\mathbb{R}^d \times \mathbb{T}^d; \mathbb{R}^{d\times d})$ is uniformly coercive, i.e. satisfying (9.2.3). Suppose the initial data $u^{in} \in L^2(\mathbb{R}^d)$. Then, for any fixed $\varepsilon > 0$, there exists a unique solution $u^\varepsilon \in L^2((0,T); H^1(\mathbb{R}^d)) \cap C^1((0,T); L^2(\mathbb{R}^d))$ to (9.2.4a)-(9.2.4b).*

For any fixed $\varepsilon > 0$, we can use the a priori bounds (9.4.1) and the Galerkin method to prove the above result. As this approach is very well-established (see Chapter 7 in [57] if necessary), we shall skip the proof of the above result as well.

Remark 9.4.1. *The regularity of the coefficients in (9.2.4a) considered in Lemma 9.4.1 and Proposition 9.4.1 are quite weak. We shall impose some stronger regularity assumptions on the fluid field $\mathbf{b}(x,y)$ when we get to the homogenization result later in this section.*

We denote the difference between the mean-field and the locally periodic fluid field by

$$\mathcal{F}(x,y) := \bar{\mathbf{b}}(x) - \mathbf{b}(x,y), \qquad \text{for } (x,y) \in \mathbb{R}^d \times \mathbb{T}^d. \tag{9.4.2}$$

Remark 9.4.2. *We specialise the main result to the 3 dimensional case. All the arguments to follow can be cast in the language of differential forms to generalize the theory to dimensions $d \geq 2$, but to simplify presentation and increase accessibility of the proof, we leave this extension to the reader.*

The null-divergence assumption on the fluid field $\mathbf{b}(x,y)$ in the y variable implies that $\mathcal{F}(x,y)$ is divergence free in the y-variable. Helmholtz decomposition of vector fields on the torus $\mathbb{T}^3$ yields the following result.

Lemma 9.4.2. *There exists* $\Upsilon(x,y) \in [L^2(\mathbb{R}^3; H^1(\mathbb{T}^3))]^3$ *such that*

$$\mathcal{F}(x,y) = \nabla_y \times \Upsilon(x,y); \qquad \text{with} \int_{\mathbb{T}^3} \Upsilon(x,y)\,\mathrm{d}y = 0.$$

Under the scaling $y = x/\varepsilon$, we have the chain rule:

$$\nabla_x \times \left(\Upsilon\left(x, \frac{x}{\varepsilon}\right)\right) = \nabla_x \times \Upsilon\left(x, \frac{x}{\varepsilon}\right) + \frac{1}{\varepsilon}\nabla_y \times \Upsilon\left(x, \frac{x}{\varepsilon}\right). \tag{9.4.3}$$

Hence by Lemma 9.4.2, we have

$$\mathcal{F}\left(x, \frac{x}{\varepsilon}\right) = \varepsilon\, \nabla_x \times \left(\Upsilon\left(x, \frac{x}{\varepsilon}\right)\right) - \varepsilon\, \nabla_x \times \Upsilon\left(x, \frac{x}{\varepsilon}\right). \tag{9.4.4}$$

9.4.2 Assumptions

In this subsection we shall make precise the assumptions on the fluid field $\mathbf{b}(x,y)$, the mean field $\bar{\mathbf{b}}(x)$ and the Jacobian matrix $J(\tau,x)$ associated with the flow Φ_τ. Throughout, we will assume that $\mathcal{A}$ is a fixed given ergodic algebra w.m.v.. See Section 9.5 for further discussions on the assumptions made here.

Assumption 9.4.1. *The fluid field* $\mathbf{b}(x,y)$ *belongs to* $C^1(\mathbb{R}^3 \times \mathbb{T}^3; \mathbb{R}^3)$ *and its flow-representation belongs to* $\mathcal{A}$ *as follows:*

$$\tilde{\mathbf{b}}(\tau,x,y) = \mathbf{b}(\Phi_\tau(x), y) \in [C^1(\mathbb{R}^3 \times \mathbb{T}^3; \mathcal{A})]^3.$$

Remark 9.4.3. *The mean-field* $\bar{\mathbf{b}}(x)$ *is nothing but the* y*-average of the fluid field* $\mathbf{b}(x,y)$*. The regularity hypothesis in Assumption 9.4.1 implies that* $\bar{\mathbf{b}}(x) \in C^1(\mathbb{R}^3; \mathbb{R}^3)$*. Furthermore, the linearity of* $\mathcal{A}$ *implies that the flow-representation of the mean-field belongs to* $\mathcal{A}$ *as follows:*

$$\tilde{\bar{\mathbf{b}}}(\tau,x) = \bar{\mathbf{b}}(\Phi_\tau(x)) \in [C^1(\mathbb{R}^3; \mathcal{A})]^3.$$

Assumption 9.4.2. *The field* $\nabla_x \times \mathcal{F}(x,y) \in C(\mathbb{R}^3 \times \mathbb{T}^3; \mathbb{R}^3)$ *and its flow-representation belongs to* $\mathcal{A}$ *as follows:*

$$\widetilde{\nabla_x \times \mathcal{F}}(\tau, x, y) = \nabla_x \times \mathcal{F}\left(\Phi_\tau(x), y\right) \in [C(\mathbb{R}^3 \times \mathbb{T}^3; \mathcal{A})]^3.$$

Remark 9.4.4. *Lemma 9.4.2 implies the flow-representation of* Υ *is given by a convolution in the* y*-variable of a Greens function and the flow-representation of* $\mathcal{F}$*. The linearity of* $\mathcal{A}$ *allows us deduce that the flow-representation of* $\Upsilon(x,y)$ *belongs to* $\mathcal{A}$ *as follows:*

$$\tilde{\Upsilon}(\tau, x, y) = \Upsilon(\Phi_\tau(x), y) \in [C^1(\mathbb{R}^3 \times \mathbb{T}^3; \mathcal{A})]^3.$$

Remark 9.4.5. *Observe that* $\zeta := \nabla_x \times \Upsilon$ *solves the equation* $\nabla_y \times \zeta = \nabla_x \times \mathcal{F}$*. Hence a similar argument as in Remark 9.4.4 and the hypothesis in Assumption 9.4.2 implies that* ζ *belongs to* $\mathcal{A}$ *as follows:*

$$\tilde{\zeta}(\tau, x, y) = \widetilde{\nabla_x \times \Upsilon}(\tau, x, y) = \nabla_x \times \Upsilon(\Phi_\tau(x), y) \in [C(\mathbb{R}^3 \times \mathbb{T}^3; \mathcal{A})]^3.$$

Assumption 9.4.3. *The molecular diffusion matrix* $\mathbf{D}(x,y) \in [L^\infty(\mathbb{R}^3; C(\mathbb{T}^3))]^{3\times 3}$ *and its flow-representation belongs to* $\mathcal{A}$ *as follows:*

$$\tilde{\mathbf{D}}(\tau, x, y) = \mathbf{D}(\Phi_\tau(x), y) \in [L^\infty(\mathbb{R}^3; C(\mathbb{T}^3) \odot \mathcal{A})]^{3\times 3}.$$

Assumption 9.4.4. *The Jacobian matrix associated with the flow* Φ *has the regularity* $J(\tau, x) \in [L^\infty(\mathbb{R}^3; \mathcal{A})]^{3\times 3}$ *and its flow-representation belongs to* $\mathcal{A}$ *as follows:*

$$\tilde{J}(\tau, x) = J(\tau, \Phi_\tau(x)) \in [L^\infty(\mathbb{R}^3; \mathcal{A})]^{3\times 3}.$$

Remark 9.4.6. *Assumption 9.4.3 is trivially satisfied if the molecular diffusion is purely periodic and bounded, i.e.* $\mathbf{D}(x,y) \equiv \mathbf{D}(y) \in [L^\infty(\mathbb{T}^3)]^{3\times 3}$*. Similarly, that the flow representation of* $\mathbf{b}$ *belongs to* $\mathcal{A}$ *as in Assumption 9.4.1 follows from Assumption 9.4.4 if the fluid field* $\mathbf{b}$ *has the special form:*

$$\mathbf{b}(x,y) = \bar{\mathbf{b}}(x) + \mathbf{b}^1(y).$$

This may be seen from using Lemma 9.2.1.(iv) to write

$$\tilde{\mathbf{b}}(\tau, x, y) = \tilde{\bar{\mathbf{b}}}(\tau, x) + \mathbf{b}^1(y) = \left(\tilde{J}(\tau, x)\right)^{-1} \bar{\mathbf{b}}(x) + \mathbf{b}^1(y).$$

Remark 9.4.7. *The assumptions on the coefficients and the Jacobian matrix (Assumption 9.4.1 - Assumption 9.4.4) ensure that they are admissible test functions in the sense of Definition 9.3.8.*

Next, we state our main result on the homogenization of the scaled convection-diffusion equation (9.2.4a)-(9.2.4b).

Theorem 9.4.1. *Let Φ_τ be the flow associated with the three dimensional autonomous system (9.2.7). Suppose $u^\varepsilon(t, x)$ be the family of solutions associated with the scaled convection-diffusion equation (9.2.4a)-(9.2.4b). Suppose $u_0(t, x)$ be the Σ-Φ_τ limit associated with the solution family. Suppose that the following assumptions hold:*

- *The fluid field $\mathbf{b}(x, y)$ satisfies the Assumption 9.4.1.*
- *The molecular diffusion $\mathbf{D}(x, y)$ satisfies the Assumption 9.4.3.*
- *The Jacobian matrix $J(\tau, x)$ satisfies the Assumption 9.4.4.*
- *The field $\mathcal{F}(x, y)$ given by (9.4.2) satisfies Assumption 9.4.2.*

Then the limit $u_0(t, x)$ solves the weak formulation

$$\begin{aligned} &- \iint\limits_{(0,T)\times\mathbb{R}^3} u_0(t, x) \frac{\partial \psi}{\partial t}(t, x) \, \mathrm{d}x \, \mathrm{d}t \\ &+ \iint\limits_{(0,T)\times\mathbb{R}^3} \mathfrak{D}(x) \nabla_x u_0(t, x) \cdot \nabla_x \psi(t, x) \, \mathrm{d}x \, \mathrm{d}t \\ &- \int\limits_{\mathbb{R}^3} u^{in}(x) \psi(0, x) \, \mathrm{d}x = 0, \end{aligned} \tag{9.4.5}$$

where the effective diffusion matrix $\mathfrak{D}(x)$ is given by

$$\mathfrak{D}(x) = \int\limits_{\Delta(\mathcal{A})} \hat{\tilde{J}}(s, x) \mathfrak{B}(s, x)^\top \hat{\tilde{J}}(s, x) \, \mathrm{d}\beta(s) \tag{9.4.6}$$

with the elements of the matrix $\mathfrak{B}(s,x)$ *given by*

$$\begin{aligned}\mathfrak{B}_{ij}(s,x) = &\int_{\mathbb{T}^3} \widehat{\widetilde{\mathbf{D}}}(s,x,y)\left(\nabla_y \widehat{\widetilde{\omega}}_j(s,x,y) + \mathbf{e}_j\right)\cdot\left(\nabla_y \widehat{\widetilde{\omega}}_i(s,x,y) + \mathbf{e}_i\right)\,\mathrm{d}y \\ &+ \int_{\mathbb{T}^3} \left(\widehat{\widetilde{\mathbf{b}}}(s,x,y)\cdot\nabla_y \widehat{\widetilde{\omega}}_i(s,x,y)\right)\widehat{\widetilde{\omega}}_j(s,x,y)\,\mathrm{d}y \\ &+ \int_{\mathbb{T}^3} \widehat{\widetilde{\mathbf{D}}}(s,x,y)\nabla_y \widehat{\widetilde{\omega}}_j(s,x,y)\cdot\mathbf{e}_i\,\mathrm{d}y \\ &- \int_{\mathbb{T}^3} \widehat{\widetilde{\mathbf{D}}}(s,x,y)\nabla_y \widehat{\widetilde{\omega}}_i(s,x,y)\cdot\mathbf{e}_j\,\mathrm{d}y,\end{aligned} \tag{9.4.7}$$

where the ω_i *are the solutions to the cell problem* (9.2.18).

Remark 9.4.8. *The weak formulation* (9.4.5) *corresponds to solving the homogenized equation*

$$\frac{\partial u_0}{\partial t} - \nabla_x\cdot\left(\mathfrak{D}(x)\nabla_x u_0(t,x)\right) = 0 \qquad \text{in } (0,T)\times\mathbb{R}^d, \tag{9.4.8}$$

$$u_0(0,x) = u^{in}(x) \qquad \text{in } \mathbb{R}^d. \tag{9.4.9}$$

This equation is identical to that given in the formal homogenisation result Proposition 9.2.1, which may be seen from the identity (9.3.1). *In particular, the diffusion coefficient* $\mathfrak{D}$ *may be computed using* (9.2.16)-(9.2.17), *and the limits therein exist and are finite.*

Remark 9.4.9. *Recall that for any function* $f(x,y)$, *the flow representation*

$$\widetilde{f}(\tau,x,y) = f\left(\Phi_\tau(x),y\right) = f\left(x,y\right) \quad \text{for } \tau\in\mathbb{R}$$

with the convention $x := \Phi_{-\tau}(x)$. *Taking the Gelfand transform of a flow representation should be understood in the abstract as follows:*

$$\widehat{\widetilde{f}}(s,x,y) = s\left(\widetilde{f}(\tau,x,y)\right) \quad \text{for } s\in\Delta(\mathcal{A}).$$

Theorem 9.4.1 asserts that a Σ-Φ_τ of the solution family $u^\varepsilon(t,x)$ solves the homogenized equation (9.4.8)-(9.4.9). The rest of the section is devoted to proving

this result. In Lemma 9.4.3, we first prove that we can extract subsequences off the solution family $\{u^\varepsilon\}$ and the gradient sequence $\{\nabla u^\varepsilon\}$ such that the extracted subsequences admit Σ-Φ_τ limits. Lemma 9.4.3 also proves that the Σ-Φ_τ limit u_0 is independent of the s variable.

Inspired by the structure of the Σ-Φ_τ limits in Lemma 9.4.3, we make a particular choice of the test functions (in Section 9.4.4) in the weak formulation of the scaled convection-diffusion (9.2.4a)-(9.2.4b).

As we need to pass to the limit in some singular terms in the weak formulation, we prove the limit behaviour of the those singular terms in Lemma 9.4.4. In Section 9.4.6, we derive the cell problem. Finally, in Section 9.4.7, we give the proof of Theorem 9.4.1.

Remark 9.4.10. *Even though the Σ-Φ_τ compactness results are up to extraction of a subsequence, the entire sequence u^ε does converge to the Σ-Φ_τ limit u_0 as the homogenized equation is uniquely solvable (Proposition 9.2.2).*

9.4.3 Σ-compactness along the flow Φ_τ

Lemma 9.4.3. *Let $u^\varepsilon(t,x)$ be the family of solutions to (9.2.4a)-(9.2.4b). Then, there exists a sub-sequence (which we still index by ε) and two limits $u_0(t,x) \in L^2((0,T);H^1(\mathbb{R}^3))$, $u_1(t,x,s,y) \in L^2((0,T)\times\mathbb{R}^3\times\Delta(\mathcal{A});H^1(\mathbb{T}^3))$ such that*

$$u^\varepsilon \xrightarrow{\Sigma-\Phi_\tau} u_0(t,x), \tag{9.4.10}$$

$$\nabla u^\varepsilon \xrightarrow{\Sigma-\Phi_\tau} \widehat{{}^\top\widetilde{J}}(s,x)\nabla_x u_0(t,x) + \nabla_y u_1(t,x,s,y). \tag{9.4.11}$$

Proof. The a priori bounds from Lemma 9.4.1 give us the necessary uniform bounds (with respect to ε) so that the result of Proposition 9.3.3 implies the existence of $u_0(t,x,s)$ and $u_1(t,x,s,y)$ such that (9.4.10) and (9.4.11) hold. To prove that u_0 is independent of the s variable, we shall consider the weak formulation of the ε-problem (9.2.4a)-(9.2.4b) with the test function $\varepsilon\varphi\left(t,\Phi_{-t/\varepsilon}(x),\frac{t}{\varepsilon}\right)$ such that $\varphi(T,\cdot,\cdot)=0$ and $\varphi(t,x,\cdot),\frac{\partial\varphi}{\partial\tau}(t,x,\cdot)\in\mathcal{A}$. The weak formulation of interest

shall be

$$
\begin{aligned}
&-\varepsilon \iint\limits_{(0,T)\times\mathbb{R}^3} u^\varepsilon(t,x)\frac{\partial\varphi}{\partial t}\left(t,\Phi_{-t/\varepsilon}(x),\frac{t}{\varepsilon}\right)\,\mathrm{d}x\,\mathrm{d}t-\varepsilon\int\limits_{\mathbb{R}^3} u^{in}(x)\varphi(0,x,0)\,\mathrm{d}x\\
&+\iint\limits_{(0,T)\times\mathbb{R}^3} u^\varepsilon(t,x)\bar{\mathbf{b}}\left(\Phi_{-t/\varepsilon}(x)\right)\cdot\nabla_x\varphi\left(t,\Phi_{-t/\varepsilon}(x),\frac{t}{\varepsilon}\right)\,\mathrm{d}x\,\mathrm{d}t\\
&-\iint\limits_{(0,T)\times\mathbb{R}^3} u^\varepsilon(t,x)\frac{\partial\varphi}{\partial\tau}\left(t,\Phi_{-t/\varepsilon}(x),\frac{t}{\varepsilon}\right)\,\mathrm{d}x\,\mathrm{d}t\\
&-\iint\limits_{(0,T)\times\mathbb{R}^3} u^\varepsilon(t,x)\mathbf{b}\left(x,\frac{x}{\varepsilon}\right)\cdot{}^{\top}\tilde{J}\left(\frac{t}{\varepsilon},\Phi_{-t/\varepsilon}(x)\right)\nabla_x\varphi\left(t,\Phi_{-t/\varepsilon}(x),\frac{t}{\varepsilon}\right)\,\mathrm{d}x\,\mathrm{d}t\\
&+\varepsilon\iint\limits_{(0,T)\times\mathbb{R}^3}\mathbf{D}\left(x,\frac{x}{\varepsilon}\right)\nabla u^\varepsilon(t,x)\cdot{}^{\top}\tilde{J}\left(\frac{t}{\varepsilon},\Phi_{-t/\varepsilon}(x)\right)\nabla_x\varphi\left(t,\Phi_{-t/\varepsilon}(x),\frac{t}{\varepsilon}\right)\,\mathrm{d}x\,\mathrm{d}t=0.
\end{aligned}
\tag{9.4.12}
$$

The first and second terms on the left hand side of the above expression are of $\mathcal{O}(\varepsilon)$. The third and the fifth terms in the weak formulation (9.4.12) together become, using (9.4.2) and (9.4.4),

$$
\begin{aligned}
&\iint\limits_{(0,T)\times\mathbb{R}^3} u^\varepsilon(t,x)\left(\bar{\mathbf{b}}\left(x\right)-\mathbf{b}\left(x,\frac{x}{\varepsilon}\right)\right)\cdot{}^{\top}\tilde{J}\left(\frac{t}{\varepsilon},\Phi_{-t/\varepsilon}(x)\right)\nabla_x\varphi\left(t,\Phi_{-t/\varepsilon}(x),\frac{t}{\varepsilon}\right)\,\mathrm{d}x\,\mathrm{d}t\\
&=\iint\limits_{(0,T)\times\mathbb{R}^3} u^\varepsilon(t,x)\mathcal{F}\left(x,\frac{x}{\varepsilon}\right)\cdot{}^{\top}\tilde{J}\left(\frac{t}{\varepsilon},\Phi_{-t/\varepsilon}(x)\right)\nabla_x\varphi\left(t,\Phi_{-t/\varepsilon}(x),\frac{t}{\varepsilon}\right)\,\mathrm{d}x\,\mathrm{d}t\\
&=\varepsilon\iint\limits_{(0,T)\times\mathbb{R}^3} u^\varepsilon(t,x)\nabla_x\times\left(\Upsilon\left(x,\frac{x}{\varepsilon}\right)\right)\cdot{}^{\top}\tilde{J}\left(\frac{t}{\varepsilon},\Phi_{-t/\varepsilon}(x)\right)\nabla_x\varphi\left(t,\Phi_{-t/\varepsilon}(x),\frac{t}{\varepsilon}\right)\,\mathrm{d}x\,\mathrm{d}t\\
&-\varepsilon\iint\limits_{(0,T)\times\mathbb{R}^3} u^\varepsilon(t,x)\nabla_x\times\Upsilon\left(x,\frac{x}{\varepsilon}\right)\cdot{}^{\top}\tilde{J}\left(\frac{t}{\varepsilon},\Phi_{-t/\varepsilon}(x)\right)\nabla_x\varphi\left(t,\Phi_{-t/\varepsilon}(x),\frac{t}{\varepsilon}\right)\,\mathrm{d}x\,\mathrm{d}t,
\end{aligned}
$$

continuing, this is equal to

$$
\begin{aligned}
&=\varepsilon\iint\limits_{(0,T)\times\mathbb{R}^3}\Upsilon\left(x,\frac{x}{\varepsilon}\right)\cdot\left(\nabla_x u^\varepsilon(t,x)\times{}^{\top}\tilde{J}\left(\frac{t}{\varepsilon},\Phi_{-t/\varepsilon}(x)\right)\nabla_x\varphi\left(t,\Phi_{-t/\varepsilon}(x),\frac{t}{\varepsilon}\right)\right)\,\mathrm{d}x\,\mathrm{d}t\\
&+\varepsilon\iint\limits_{(0,T)\times\mathbb{R}^3}\Upsilon\left(x,\frac{x}{\varepsilon}\right)\cdot\left(u_\varepsilon(t,x)\nabla_x\times{}^{\top}\tilde{J}\left(\frac{t}{\varepsilon},\Phi_{-t/\varepsilon}(x)\right)\nabla_x\varphi\left(t,\Phi_{-t/\varepsilon}(x),\frac{t}{\varepsilon}\right)\right)\,\mathrm{d}x\,\mathrm{d}t\\
&-\varepsilon\iint\limits_{(0,T)\times\mathbb{R}^3} u^\varepsilon(t,x)\nabla_x\times\Upsilon\left(x,\frac{x}{\varepsilon}\right)\cdot{}^{\top}\tilde{J}\left(\frac{t}{\varepsilon},\Phi_{-t/\varepsilon}(x)\right)\nabla_x\varphi\left(t,\Phi_{-t/\varepsilon}(x),\frac{t}{\varepsilon}\right)\,\mathrm{d}x\,\mathrm{d}t
\end{aligned}
$$

where we have used the Greens formula for the Curl operator. The second term on the far right hand side of the above expression vanishes because

$$\begin{aligned}\nabla_x \times {}^{\top}\tilde{J}\left(\frac{t}{\varepsilon}, \Phi_{-t/\varepsilon}(x)\right) \nabla_X \varphi\left(t, \Phi_{-t/\varepsilon}(x), \frac{t}{\varepsilon}\right) \\ = \nabla_x \times {}^{\top}J\left(\frac{t}{\varepsilon}, x\right) \nabla_X \varphi\left(t, \Phi_{-t/\varepsilon}(x), \frac{t}{\varepsilon}\right) \\ = \nabla_x \times \nabla_x \left(\varphi\left(t, \Phi_{-t/\varepsilon}(x), \frac{t}{\varepsilon}\right)\right) = 0,\end{aligned}$$

as the curl of a gradient is zero. (This is the reason that Υ was chosen in this particular manner). In the rest of the terms, using the flow-representation, we have

$$\begin{aligned}\varepsilon \iint_{(0,T)\times\mathbb{R}^3} \tilde{\Upsilon}\left(\frac{t}{\varepsilon}, \Phi_{-t/\varepsilon}(x), \frac{x}{\varepsilon}\right) \\ \cdot \left(\nabla_x u^{\varepsilon}(t,x) \times {}^{\top}\tilde{J}\left(\frac{t}{\varepsilon}, \Phi_{-t/\varepsilon}(x)\right) \nabla_X \varphi\left(t, \Phi_{-t/\varepsilon}(x), \frac{t}{\varepsilon}\right)\right) \mathrm{d}x\,\mathrm{d}t \\ - \varepsilon \iint_{(0,T)\times\mathbb{R}^3} u^{\varepsilon}(t,x) \widetilde{\nabla_x \times \Upsilon}\left(\frac{t}{\varepsilon}, \Phi_{-t/\varepsilon}(x), \frac{x}{\varepsilon}\right) \\ \cdot {}^{\top}\tilde{J}\left(\frac{t}{\varepsilon}, \Phi_{-t/\varepsilon}(x)\right) \nabla_X \varphi\left(t, \Phi_{-t/\varepsilon}(x), \frac{t}{\varepsilon}\right) \mathrm{d}x\,\mathrm{d}t.\end{aligned}$$

These are of $\mathcal{O}(\varepsilon)$. Using the flow-representation for the molecular diffusion $\mathbf{D}$, the final term on the left hand side of the weak formulation is also of $\mathcal{O}(\varepsilon)$. Hence, passing to the limit as ε tends to zero in the weak formulation yields

$$\iiint_{(0,T)\times\mathbb{R}^3\times\Delta(\mathcal{A})} u_0(t,x,s) \widehat{\frac{\partial\varphi}{\partial\tau}}(t,x,s)\,\mathrm{d}\beta(s)\,\mathrm{d}x\,\mathrm{d}t = 0.$$

Upon using Lemma 9.3.1, we deduce that the u_0 is independent of the s-variable. □

9.4.4 Choice of test functions

We consider the weak formulation of the convection-diffusion equation (9.2.4a)-(9.2.4b) with test function $\psi^{\varepsilon}(t,x)$ such that $\psi^{\varepsilon}(T,x) = 0$. We split the weak

formulation into four terms:

$$\mathcal{I}_{time} + \mathcal{I}_{convect} + \mathcal{I}_{diffuse} + \mathcal{I}_{initial} = 0$$

which are respectively given by

$$\begin{aligned} &- \iint\limits_{(0,T)\times\mathbb{R}^3} u^\varepsilon(t,x)\frac{\partial\psi^\varepsilon}{\partial t}(t,x)\,\mathrm{d}x\,\mathrm{d}t + \frac{1}{\varepsilon}\iint\limits_{(0,T)\times\mathbb{R}^3} \mathbf{b}\left(x,\frac{x}{\varepsilon}\right)\cdot\nabla u^\varepsilon(t,x)\psi^\varepsilon(t,x)\,\mathrm{d}x\,\mathrm{d}t \\ &+ \iint\limits_{(0,T)\times\mathbb{R}^3} \mathbf{D}\left(x,\frac{x}{\varepsilon}\right)\nabla u^\varepsilon(t,x)\cdot\nabla\psi^\varepsilon(t,x)\,\mathrm{d}x\,\mathrm{d}t - \int\limits_{\mathbb{R}^3} u^{in}(x)\psi^\varepsilon(0,x)\,\mathrm{d}x = 0. \end{aligned} \tag{9.4.13}$$

The choice of the family of test functions $\psi^\varepsilon(t,x)$ is as follows:

$$\psi^\varepsilon(t,x) = \psi\left(t,\Phi_{-t/\varepsilon}(x)\right) + \varepsilon\psi_1\left(t,\Phi_{-t/\varepsilon}(x),\frac{t}{\varepsilon},\frac{x}{\varepsilon}\right) \tag{9.4.14}$$

where

$$\psi\in C^1((0,T)\times\mathbb{R}^3) \qquad \text{and} \qquad \psi_1\in C^1((0,T)\times\mathbb{R}^3;C^1(\mathbb{T}^3)\odot\mathcal{A}) \tag{9.4.15}$$

which are compactly supported in space. We shall treat term by term. To begin with, let us consider the term with the partial time derivative:

$$\begin{aligned} \mathcal{I}_{time} = &- \iint\limits_{(0,T)\times\mathbb{R}^3} u^\varepsilon(t,x)\frac{\partial\psi}{\partial t}\left(t,\Phi_{-t/\varepsilon}(x)\right)\,\mathrm{d}x\,\mathrm{d}t \\ &+ \frac{1}{\varepsilon}\iint\limits_{(0,T)\times\mathbb{R}^3} u^\varepsilon(t,x)\bar{\mathbf{b}}(\Phi_{-t/\varepsilon}(x))\cdot\nabla_x\psi\left(t,\Phi_{-t/\varepsilon}(x)\right)\,\mathrm{d}x\,\mathrm{d}t \\ &- \iint\limits_{(0,T)\times\mathbb{R}^3} u^\varepsilon(t,x)\frac{\partial\psi_1}{\partial\tau}\left(t,\Phi_{-t/\varepsilon}(x),\frac{t}{\varepsilon},\frac{x}{\varepsilon}\right)\,\mathrm{d}x\,\mathrm{d}t \\ &+ \iint\limits_{(0,T)\times\mathbb{R}^3} u^\varepsilon(t,x)\bar{\mathbf{b}}(\Phi_{-t/\varepsilon}(x))\cdot\nabla_x\psi_1\left(t,\Phi_{-t/\varepsilon}(x),\frac{t}{\varepsilon},\frac{x}{\varepsilon}\right)\,\mathrm{d}x\,\mathrm{d}t \\ &+ \mathcal{O}(\varepsilon). \end{aligned}$$

Now, for the convection term:

$$\mathcal{I}_{convect} = -\frac{1}{\varepsilon}\iint\limits_{(0,T)\times\mathbb{R}^3} u^\varepsilon(t,x)\mathbf{b}\left(x,\frac{x}{\varepsilon}\right)\cdot {}^\top\tilde{J}\left(\frac{t}{\varepsilon},\Phi_{-t/\varepsilon}(x)\right)\nabla_X\psi\left(t,\Phi_{-t/\varepsilon}(x)\right)\,\mathrm{d}x\,\mathrm{d}t$$
$$+\iint\limits_{(0,T)\times\mathbb{R}^3}\mathbf{b}\left(x,\frac{x}{\varepsilon}\right)\cdot\nabla u^\varepsilon(t,x)\psi_1\left(t,\Phi_{-t/\varepsilon}(x),\frac{t}{\varepsilon},\frac{x}{\varepsilon}\right)\,\mathrm{d}x\,\mathrm{d}t$$

Next, for the diffusion term:

$$\mathcal{I}_{diffuse} = \iint\limits_{(0,T)\times\mathbb{R}^3}\mathbf{D}\left(x,\frac{x}{\varepsilon}\right)\nabla u^\varepsilon(t,x)\cdot {}^\top\tilde{J}\left(\frac{t}{\varepsilon},\Phi_{-t/\varepsilon}(x)\right)\nabla_X\psi\left(t,\Phi_{-t/\varepsilon}(x)\right)\,\mathrm{d}x\,\mathrm{d}t$$
$$+\iint\limits_{(0,T)\times\mathbb{R}^3}\mathbf{D}\left(x,\frac{x}{\varepsilon}\right)\nabla u^\varepsilon(t,x)\cdot\nabla_y\psi_1\left(t,\Phi_{-t/\varepsilon}(x),\frac{t}{\varepsilon},\frac{x}{\varepsilon}\right)\,\mathrm{d}x\,\mathrm{d}t+\mathcal{O}(\varepsilon).$$

Finally, for the term involving initial data:

$$\mathcal{I}_{initial} = -\int\limits_{\mathbb{R}^3} u^{in}(x)\psi(0,x)\,\mathrm{d}x+\mathcal{O}(\varepsilon).$$

By using the flow-representation and Lemma 9.2.1.(iv) we notice that

$$\bar{\mathbf{b}}\left(\Phi_{-t/\varepsilon}(x)\right)=\tilde{J}\left(\frac{t}{\varepsilon},\Phi_{-t/\varepsilon}(x)\right)\bar{\mathbf{b}}(x)=\tilde{J}\left(\frac{t}{\varepsilon},\Phi_{-t/\varepsilon}(x)\right)\tilde{\bar{\mathbf{b}}}\left(\frac{t}{\varepsilon},\Phi_{-t/\varepsilon}(x)\right). \tag{9.4.16}$$

Again using the flow-representation, we have

$$\mathbf{b}\left(x,\frac{x}{\varepsilon}\right)=\tilde{\mathbf{b}}\left(\frac{t}{\varepsilon},\Phi_{-t/\varepsilon}(x),\frac{x}{\varepsilon}\right),\qquad \mathbf{D}(x,y)=\tilde{\mathbf{D}}\left(\frac{t}{\varepsilon},\Phi_{-t/\varepsilon}(x),\frac{x}{\varepsilon}\right). \tag{9.4.17}$$

The observations (9.4.16)-(9.4.17), combined with the Σ-compactness result along the flow Φ_t (Lemma 9.4.3) will allow us to pass to the limit as $\varepsilon\to 0$ in all but two singular terms.

9.4.5 Singular terms

We record below a result giving the limit of the singular terms in the weak formulation.

Lemma 9.4.4. *Under Assumption 9.4.2 on the field $\mathcal{F}(x,y)$ and for ψ satisfying (9.4.15), we have*

$$\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\iint\limits_{(0,T)\times\mathbb{R}^3} u^\varepsilon(t,x)\tilde{J}\left(\frac{t}{\varepsilon},\Phi_{-t/\varepsilon}(x)\right)\left[\bar{\mathbf{b}}(x)-\mathbf{b}\left(x,\frac{x}{\varepsilon}\right)\right]\cdot\nabla_X\psi\left(t,\Phi_{-t/\varepsilon}(x)\right)\,\mathrm{d}x\,\mathrm{d}t$$

$$=\iiiint\limits_{\mathbb{R}_+\times\mathbb{R}^d\times\Delta(\mathcal{A})\times\mathbb{T}^d}\left(\hat{\tilde{\mathbf{b}}}(s,x)-\hat{\tilde{\mathbf{b}}}(s,x,y)\right)\cdot\left(u_1(t,x,s,y)^\top\hat{\tilde{J}}(s,x)\nabla_X\hat{\psi}(t,x)\right)\,\mathrm{d}y\,\mathrm{d}\beta(s)\,\mathrm{d}x\,\mathrm{d}t. \tag{9.4.18}$$

Proof. Consider the singular terms in the weak formulation:

$$\frac{1}{\varepsilon}\iint\limits_{(0,T)\times\mathbb{R}^3} u^\varepsilon(t,x)\tilde{J}\left(\frac{t}{\varepsilon},\Phi_{-t/\varepsilon}(x)\right)\left[\bar{\mathbf{b}}(x)-\mathbf{b}\left(x,\frac{x}{\varepsilon}\right)\right]\cdot\nabla_X\psi\left(t,\Phi_{-t/\varepsilon}(x)\right)\,\mathrm{d}x\,\mathrm{d}t. \tag{9.4.19}$$

Using the observation (9.4.4) on the field $\mathcal{F}\left(x,\frac{x}{\varepsilon}\right)$, we rewrite the singular terms (9.4.19) successively as follows:

$$\begin{aligned}
&\iint\limits_{(0,T)\times\mathbb{R}^3}\nabla_x\times\left(\Upsilon\left(x,\frac{x}{\varepsilon}\right)\right)\cdot\left(u_\varepsilon(t,x)^\top\tilde{J}\left(\frac{t}{\varepsilon},\Phi_{-t/\varepsilon}(x)\right)\nabla_X\psi\left(t,\Phi_{-t/\varepsilon}(x)\right)\right)\,\mathrm{d}x\,\mathrm{d}t\\
&-\iint\limits_{(0,T)\times\mathbb{R}^3}\nabla_x\times\Upsilon\left(x,\frac{x}{\varepsilon}\right)\cdot\left(u_\varepsilon(t,x)^\top\tilde{J}\left(\frac{t}{\varepsilon},\Phi_{-t/\varepsilon}(x)\right)\nabla_X\psi\left(t,\Phi_{-t/\varepsilon}(x)\right)\right)\,\mathrm{d}x\,\mathrm{d}t\\
&=\iint\limits_{(0,T)\times\mathbb{R}^3}\Upsilon\left(x,\frac{x}{\varepsilon}\right)\cdot\left(\nabla_x u_\varepsilon(t,x)\times{}^\top\tilde{J}\left(\frac{t}{\varepsilon},\Phi_{-t/\varepsilon}(x)\right)\nabla_X\psi\left(t,\Phi_{-t/\varepsilon}(x)\right)\right)\,\mathrm{d}x\,\mathrm{d}t\\
&+\iint\limits_{(0,T)\times\mathbb{R}^3}\Upsilon\left(x,\frac{x}{\varepsilon}\right)\cdot\left(u_\varepsilon(t,x)\nabla_x\times\left({}^\top\tilde{J}\left(\frac{t}{\varepsilon},\Phi_{-t/\varepsilon}(x)\right)\nabla_X\psi\left(t,\Phi_{-t/\varepsilon}(x)\right)\right)\right)\,\mathrm{d}x\,\mathrm{d}t\\
&-\iint\limits_{(0,T)\times\mathbb{R}^3}\nabla_x\times\Upsilon\left(x,\frac{x}{\varepsilon}\right)\cdot\left(u_\varepsilon(t,x)^\top\tilde{J}\left(\frac{t}{\varepsilon},\Phi_{-t/\varepsilon}(x)\right)\nabla_X\psi\left(t,\Phi_{-t/\varepsilon}(x)\right)\right)\,\mathrm{d}x\,\mathrm{d}t,
\end{aligned}$$

where we have used the Green's formula for the curl operator. Note that the second term on the right hand side of the previous expression is zero because

$${}^\top\tilde{J}\left(\frac{t}{\varepsilon},\Phi_{-t/\varepsilon}(x)\right)\nabla_X\psi\left(t,\Phi_{-t/\varepsilon}(x)\right)=\nabla_x\left(\psi\left(t,\Phi_{-t/\varepsilon}(x)\right)\right)$$

and because of the fact that curl of a gradient is zero. Next, using the flow-representation, the singular terms (9.4.19) simplify to the following expression:

$$\begin{aligned}
&\iint_{(0,T)\times\mathbb{R}^3} \tilde{\Upsilon}\left(\frac{t}{\varepsilon}, \Phi_{-t/\varepsilon}(x), \frac{x}{\varepsilon}\right) \\
&\qquad\qquad \cdot \left(\nabla_x u_\varepsilon(t,x) \times {}^{\top}\tilde{J}\left(\frac{t}{\varepsilon}, \Phi_{-t/\varepsilon}(x)\right) \nabla_X \psi\left(t, \Phi_{-t/\varepsilon}(x)\right)\right) \mathrm{d}x\,\mathrm{d}t \\
&- \iint_{(0,T)\times\mathbb{R}^3} \widetilde{\nabla_x \times \Upsilon}\left(\frac{t}{\varepsilon}, \Phi_{-t/\varepsilon}(x), \frac{x}{\varepsilon}\right) \\
&\qquad\qquad \cdot \left(u_\varepsilon(t,x)^{\top}\tilde{J}\left(\frac{t}{\varepsilon}, \Phi_{-t/\varepsilon}(x)\right) \nabla_X \psi\left(t, \Phi_{-t/\varepsilon}(x)\right)\right) \mathrm{d}x\,\mathrm{d}t.
\end{aligned}$$

Thanks to the Assumption 9.4.2 (see Remark 9.4.4 and Remark 9.4.5), we can pass to the limit, as $\varepsilon \to 0$, in the previous expression using Σ-convergence along the flow Φ_t yielding

$$\begin{aligned}
&\iiiint_{(0,T)\times\mathbb{R}^3\times\Delta(\mathcal{A})\times\mathbb{T}^3} \hat{\tilde{\Upsilon}}(s,x,y) \cdot \left(\widehat{{}^{\top}\tilde{J}}(s,x)\nabla_X u_0(t,x) \times \widehat{{}^{\top}\tilde{J}}(s,x)\nabla_X \hat{\psi}(t,x)\right) \mathrm{d}y\,\mathrm{d}\beta(s)\,\mathrm{d}x\,\mathrm{d}t \\
&+ \iiiint_{(0,T)\times\mathbb{R}^3\times\Delta(\mathcal{A})\times\mathbb{T}^3} \hat{\tilde{\Upsilon}}(s,x,y) \cdot \left(\nabla_y u_1(t,x,s,y) \times \widehat{{}^{\top}\tilde{J}}(s,x)\nabla_X \hat{\psi}(t,x)\right) \mathrm{d}y\,\mathrm{d}\beta(s)\,\mathrm{d}x\,\mathrm{d}t \\
&- \iiiint_{(0,T)\times\mathbb{R}^3\times\Delta(\mathcal{A})\times\mathbb{T}^3} \widehat{\nabla_x \times \Upsilon}(s,x,y) \cdot \left(u_0(t,x)\widehat{{}^{\top}\tilde{J}}(s,x)\nabla_X \hat{\psi}(t,x)\right) \mathrm{d}y\,\mathrm{d}\beta(s)\,\mathrm{d}x\,\mathrm{d}t.
\end{aligned} \tag{9.4.20}$$

By construction, Υ is of zero average in the y variable (Lemma 9.4.2). Hence the first term in (9.4.20) vanishes. In the second term of (9.4.20), we use the Green's formula for the curl operator in y variable leading to the following expression:

$$\iiiint_{(0,T)\times\mathbb{R}^3\times\Delta(\mathcal{A})\times\mathbb{T}^3} \nabla_y \times \hat{\tilde{\Upsilon}}(s,x,y) \cdot \left(u_1(t,x,s,y)^{\top}\hat{\tilde{J}}(s,x)\nabla_X \hat{\psi}(t,x)\right) \mathrm{d}y\,\mathrm{d}\beta(s)\,\mathrm{d}x\,\mathrm{d}t.$$

Again by Lemma 9.4.2, we have

$$\nabla_y \times \hat{\tilde{\Upsilon}}(s,x,y) = \hat{\tilde{\mathcal{F}}}(s,x,y) = \hat{\tilde{\mathbf{b}}}(s,x) - \hat{\tilde{\mathbf{b}}}(s,x,y).$$

Hence, the second term of (9.4.20) is the same as

$$\iiiint_{(0,T)\times\mathbb{R}^3\times\Delta(\mathcal{A})\times\mathbb{T}^3} \left(\widehat{\widetilde{\mathbf{b}}}(s,x) - \widehat{\widetilde{\mathbf{b}}}(s,x,y)\right) \cdot \left(u_1(t,x,s,y)\widehat{{}^{\top}\widetilde{J}}(s,x)\nabla_x\widehat{\psi}(t,x)\right) \mathrm{d}y\,\mathrm{d}\beta(s)\,\mathrm{d}x\,\mathrm{d}t. \tag{9.4.21}$$

Regarding the third term in (9.4.20), remark that

$$\widehat{\widetilde{\nabla_x \times \Upsilon}}(s,x,y) = \widehat{{}^{\top}\widetilde{J}}(s,x)\left(\nabla_x \times \widehat{\widetilde{\Upsilon}}(s,x,y)\right).$$

Hence the third term in (9.4.20) rewrites as (upon using Green's formula):

$$\iiiint_{(0,T)\times\mathbb{R}^3\times\Delta(\mathcal{A})\times\mathbb{T}^3} \widehat{\widetilde{\Upsilon}}(s,x,y)\cdot\nabla_x \times \left(u_0(t,x)\widehat{\widetilde{J}}(s,x)\widehat{{}^{\top}\widetilde{J}}(s,x)\nabla_x\widehat{\psi}(t,x)\right) \mathrm{d}y\,\mathrm{d}\beta(s)\,\mathrm{d}x\,\mathrm{d}t.$$

Thanks again to the construction that Υ is of zero average in the y variable (Lemma 9.4.2), the above expression vanishes. Hence the only non-zero term in the limit expression is (9.4.21). This is nothing but the limit in (9.4.18). □

9.4.6 Cell problem

Now, we record a result to give the limit equation for the weak formulation involving the test function ψ_1, i.e. to derive the cell problem.

Proposition 9.4.2. *Let Φ_τ be the flow associated with the autonomous system (9.2.7). Under Assumption 9.4.1 on the fluid field $\mathbf{b}(x,y)$, Assumption 9.4.3 on the molecular diffusion $\mathbf{D}(x,y)$ and Assumption 9.4.4 on the Jacobian matrix $J(\tau,x)$, the Σ-Φ_τ limit $u_1(t,x,s,y)$ obtained in Lemma 9.4.3 can be written as*

$$u_1(t,x,s,y) = \widehat{\widetilde{\omega}}(s,x,y)\cdot\widehat{{}^{\top}\widetilde{J}}(s,x)\nabla_x u_0(t,x), \tag{9.4.22}$$

where the components of $\omega(x,y)\in[L^\infty(\mathbb{R}^3;H^1(\mathbb{T}^3))]^d$ with $\int_{\mathbb{T}^3} w\,\mathrm{d}y = 0$ solve the

cell problem:

$$\mathbf{b}(x,y)\cdot(\nabla_y\omega_i+\mathbf{e}_i)-\nabla_y\cdot(\mathbf{D}(x,y)\,(\nabla_y\omega_i+\mathbf{e}_i))=\bar{\mathbf{b}}(x)\cdot\mathbf{e}_i \qquad \text{in } \mathbb{T}^3, \tag{9.4.23}$$

for each $i\in\{1,2,3\}$*, where* $\{\mathbf{e}_i\}_{i=1}^3$ *denote the canonical basis in* $\mathbb{R}^3$ *and* x *is viewed as a parameter.*

Proof. Taking $\psi\equiv 0$ in the weak formulation (9.4.13) and passing to the limit in the sense of Σ-convergence along the flow Φ_τ, we obtain

$$\begin{aligned}
&-\iiiint_{(0,T)\times\mathbb{R}^3\times\Delta(\mathcal{A})\times\mathbb{T}^3} u_0(t,x)\widehat{\frac{\partial\psi_1}{\partial\tau}}(t,x,s,y)\,\mathrm{d}y\,\mathrm{d}\beta(s)\,\mathrm{d}x\,\mathrm{d}t\\
&+\iiiint_{(0,T)\times\mathbb{R}^3\times\Delta(\mathcal{A})\times\mathbb{T}^3} u_0(t,x)\left\{\widehat{\tilde{J}}(s,x)\widehat{\tilde{\mathbf{b}}}(s,x)\cdot\nabla_x\widehat{\psi_1}(t,x,s,y)\right\}\mathrm{d}y\,\mathrm{d}\beta(s)\,\mathrm{d}x\,\mathrm{d}t\\
&+\iiiint_{(0,T)\times\mathbb{R}^3\times\Delta(\mathcal{A})\times\mathbb{T}^3}\left\{\widehat{\tilde{\mathbf{b}}}(s,x,y)\cdot\widehat{{}^{\top}\tilde{J}}(s,x)\nabla_x u_0(t,x)\right\}\widehat{\psi_1}(t,x,s,y)\,\mathrm{d}y\,\mathrm{d}\beta(s)\,\mathrm{d}x\,\mathrm{d}t\\
&+\iiiint_{(0,T)\times\mathbb{R}^3\times\Delta(\mathcal{A})\times\mathbb{T}^3}\left\{\widehat{\tilde{\mathbf{b}}}(s,x,y)\cdot\nabla_y u_1(t,x,s,y)\right\}\widehat{\psi_1}(t,x,s,y)\,\mathrm{d}y\,\mathrm{d}\beta(s)\,\mathrm{d}x\,\mathrm{d}t\\
&+\iiiint_{(0,T)\times\mathbb{R}^3\times\Delta(\mathcal{A})\times\mathbb{T}^3}\widehat{\tilde{\mathbf{D}}}(s,x,y)\widehat{{}^{\top}\tilde{J}}(s,x)\nabla_x u_0(t,x)\cdot\nabla_y\widehat{\psi_1}(t,x,s,y)\,\mathrm{d}y\,\mathrm{d}\beta(s)\,\mathrm{d}x\,\mathrm{d}t\\
&+\iiiint_{(0,T)\times\mathbb{R}^3\times\Delta(\mathcal{A})\times\mathbb{T}^3}\widehat{\tilde{\mathbf{D}}}(s,x,y)\nabla_y u_1(t,x,s,y)\cdot\nabla_y\widehat{\psi_1}(t,x,s,y)\,\mathrm{d}y\,\mathrm{d}\beta(s)\,\mathrm{d}x\,\mathrm{d}t=0.
\end{aligned}$$

The first term in the previous equation vanishes as u_0 is independent of the s-variable (Lemma 9.4.3). Finally, performing an integration by parts in the x variable in the second integral of the previous equation leads to

$$\begin{aligned}
&\iiiint_{(0,T)\times\mathbb{R}^3\times\Delta(\mathcal{A})\times\mathbb{T}^3}\widehat{\widetilde{\mathbf{D}}}(s,x,y)\nabla_y u_1(t,x,s,y)\cdot\nabla_y\widehat{\psi_1}(t,x,s,y)\,\mathrm{d}y\,\mathrm{d}\beta(s)\,\mathrm{d}x\,\mathrm{d}t\\
&+\iiiint_{(0,T)\times\mathbb{R}^3\times\Delta(\mathcal{A})\times\mathbb{T}^3}\widehat{\widetilde{\mathbf{D}}}(s,x,y)\widehat{{}^{\top}\tilde{J}}(s,x)\nabla_x u_0(t,x)\cdot\nabla_y\widehat{\psi_1}(t,x,s,y)\,\mathrm{d}y\,\mathrm{d}\beta(s)\,\mathrm{d}x\,\mathrm{d}t\\
&+\iiiint_{(0,T)\times\mathbb{R}^3\times\Delta(\mathcal{A})\times\mathbb{T}^3}\left\{\widehat{\tilde{\mathbf{b}}}(s,x,y)\cdot\nabla_y u_1(t,x,s,y)\right\}\widehat{\psi_1}(t,x,s,y)\,\mathrm{d}y\,\mathrm{d}\beta(s)\,\mathrm{d}x\,\mathrm{d}t
\end{aligned}\tag{9.4.24}$$

$$
\begin{aligned}
&+ \iiiint\limits_{(0,T)\times\mathbb{R}^3\times\Delta(\mathcal{A})\times\mathbb{T}^3} \left\{ \left(\widehat{\widetilde{\mathbf{b}}}(s,x,y) - \widehat{\bar{\widetilde{\mathbf{b}}}}(s,x) \right) \cdot \widehat{{}^{\top}\widetilde{J}}(s,x) \nabla_x u_0(t,x) \right\} \\
&\qquad\qquad \widehat{\psi_1}(t,x,s,y)\,\mathrm{d}y\,\mathrm{d}\beta(s)\,\mathrm{d}x\,\mathrm{d}t \\
&= 0,
\end{aligned}
\tag{9.4.25}
$$

where we have used Lemma 9.2.1.(ii) and that $\mathbf{b}$ is of zero x-divergence. The weak formulation (9.4.24)-(9.4.25) is associated with the following PDE for $u_1(t,x,s,y)$ in $\mathbb{T}^3$:

$$
\begin{aligned}
\widehat{\widetilde{\mathbf{b}}}(s,x,y) \cdot \left(\nabla_y u_1 + \widehat{{}^{\top}\widetilde{J}}(s,x) \nabla_x u_0 \right) - \nabla_y \cdot \left(\widehat{\widetilde{\mathbf{D}}}(s,x,y) \left(\nabla_y u_1 + \widehat{{}^{\top}\widetilde{J}}(s,x) \nabla_x u_0 \right) \right) \\
= \widehat{\bar{\widetilde{\mathbf{b}}}}(s,x) \cdot \widehat{{}^{\top}\widetilde{J}}(s,x) \nabla_x u_0 .
\end{aligned}
$$

The above PDE is linear and hence separation of variables may be performed as in (9.4.22). Taking (9.4.22) and undoing the flow-representation, yields the cell problem (9.4.23). □

9.4.7 Homogenized problem

Proof of Theorem 9.4.1. Taking $\psi_1 \equiv 0$ in the weak formulation (9.4.13) and passing to the limit in the sense of Σ-convergence along Φ_τ, we obtain

$$
\begin{aligned}
&- \iint\limits_{(0,T)\times\mathbb{R}^3} u_0(t,x) \frac{\partial \psi}{\partial t}(t,x)\,\mathrm{d}x\,\mathrm{d}t - \int\limits_{\mathbb{R}^3} u^{in}(x)\psi(0,x)\,\mathrm{d}x \\
&+ \iiiint\limits_{(0,T)\times\mathbb{R}^3\times\Delta(\mathcal{A})\times\mathbb{T}^3} \left(\widehat{\bar{\widetilde{\mathbf{b}}}}(s,x) - \widehat{\widetilde{\mathbf{b}}}(s,x,y) \right) \\
&\qquad\qquad \cdot \left(u_1(t,x,s,y) \widehat{{}^{\top}\widetilde{J}}(s,x) \nabla_x \psi(t,x) \right) \mathrm{d}y\,\mathrm{d}\beta(s)\,\mathrm{d}x\,\mathrm{d}t \\
&+ \iiiint\limits_{(0,T)\times\mathbb{R}^3\times\Delta(\mathcal{A})\times\mathbb{T}^3} \widehat{\widetilde{\mathbf{D}}}(s,x,y) \widehat{{}^{\top}\widetilde{J}}(s,x) \nabla_x u_0(t,x) \cdot \widehat{{}^{\top}\widetilde{J}}(s,x) \nabla_x \psi(t,x)\,\mathrm{d}y\,\mathrm{d}\beta(s)\,\mathrm{d}x\,\mathrm{d}t \\
&+ \iiiint\limits_{(0,T)\times\mathbb{R}^3\times\Delta(\mathcal{A})\times\mathbb{T}^3} \widehat{\widetilde{\mathbf{D}}}(s,x,y) \nabla_y u_1(t,x,s,y) \cdot \widehat{{}^{\top}\widetilde{J}}(s,x) \nabla_x \psi(t,x)\,\mathrm{d}y\,\mathrm{d}\beta(s)\,\mathrm{d}x\,\mathrm{d}t = 0
\end{aligned}
\tag{9.4.26}
$$

where we have used Lemma 9.4.4 for the singular terms. Substituting (9.4.22) for $u_1(t,s,x,y)$ in the second term of the above equation yields

$$\iiiint\limits_{(0,T)\times\mathbb{R}^3\times\Delta(\mathcal{A})\times\mathbb{T}^3} \widehat{\widetilde{J}}(s,x)\left(\widehat{\widetilde{\mathbf{b}}}(s,x)-\widehat{\widetilde{\mathbf{b}}}(s,x,y)\right)\left(\widehat{\widetilde{\omega}}(s,x,y)\cdot {}^{\top}\widehat{\widetilde{J}}(s,x)\nabla_x u_0(t,x)\right)$$
$$\cdot\nabla_x\psi(t,x)\,\mathrm{d}y\,\mathrm{d}\beta(s)\,\mathrm{d}x\,\mathrm{d}t$$
$$=\iiiint\limits_{(0,T)\times\mathbb{R}^3\times\Delta(\mathcal{A})\times\mathbb{T}^3}\left(\widehat{\widetilde{J}}(s,x)\left(\widehat{\widetilde{\mathbf{b}}}(s,x)-\widehat{\widetilde{\mathbf{b}}}(s,x,y)\right)^{\top}\widehat{\widetilde{\omega}}(s,x,y)^{\top}\widehat{\widetilde{J}}(s,x)\nabla_x u_0(t,x)\right)$$
$$\cdot\nabla_x\psi(t,x)\,\mathrm{d}y\,\mathrm{d}\beta(s)\,\mathrm{d}x\,\mathrm{d}t.$$

Substituting (9.4.22) for $u_1(t,s,x,y)$ in the fourth term on the left hand side of (9.4.26) yields

$$\iiiint\limits_{(0,T)\times\mathbb{R}^3\times\Delta(\mathcal{A})\times\mathbb{T}^3}\widehat{\widetilde{J}}(s,x)\widehat{\widetilde{\mathbf{D}}}(s,x,y)\nabla_y\left(\widehat{\widetilde{\omega}}(s,x,y)\cdot {}^{\top}\widehat{\widetilde{J}}(s,x)\nabla_x u_0(t,x)\right)$$
$$\cdot\nabla_x\psi(t,x)\,\mathrm{d}y\,\mathrm{d}\beta(s)\,\mathrm{d}x\,\mathrm{d}t$$
$$=\iiiint\limits_{(0,T)\times\mathbb{R}^3\times\Delta(\mathcal{A})\times\mathbb{T}^3}\left(\widehat{\widetilde{J}}(s,x)\widehat{\widetilde{\mathbf{D}}}(s,x,y)^{\top}\nabla_y\widehat{\widetilde{\omega}}(s,x,y)^{\top}\widehat{\widetilde{J}}(s,x)\nabla_x u_0(t,x)\right)$$
$$\cdot\nabla_x\psi(t,x)\,\mathrm{d}y\,\mathrm{d}\beta(s)\,\mathrm{d}x\,\mathrm{d}t.$$

Hence the limit weak formulation (9.4.26) rewrites as

$$-\iint\limits_{(0,T)\times\mathbb{R}^3}u_0(t,x)\frac{\partial\psi}{\partial t}(t,x)\,\mathrm{d}x\,\mathrm{d}t+\iint\limits_{(0,T)\times\mathbb{R}^3}\mathfrak{D}(x)\nabla_x u_0(t,x)\cdot\nabla_x\psi(t,x)\,\mathrm{d}x\,\mathrm{d}t$$
$$-\int\limits_{\mathbb{R}^3}u^{in}(x)\psi(0,x)\,\mathrm{d}x=0,$$

where the expression for the diffusion matrix $\mathfrak{D}(x)$ is given by

$$\mathfrak{D}(x)=\int\limits_{\Delta(\mathcal{A})}\widehat{\widetilde{J}}(s,x)\left(\int\limits_{\mathbb{T}^3}\left(\widehat{\widetilde{\mathbf{b}}}(s,x)-\widehat{\widetilde{\mathbf{b}}}(s,x,y)\right)^{\top}\widehat{\widetilde{\omega}}(s,x,y)\,\mathrm{d}y\right){}^{\top}\widehat{\widetilde{J}}(s,x)\,\mathrm{d}\beta(s)$$
$$+\int\limits_{\Delta(\mathcal{A})}\widehat{\widetilde{J}}(s,x)\left(\int\limits_{\mathbb{T}^3}\widehat{\widetilde{\mathbf{D}}}(s,x,y)\,\mathrm{d}y\right){}^{\top}\widehat{\widetilde{J}}(s,x)\,\mathrm{d}\beta(s)$$

$$+ \int_{\Delta(\mathcal{A})} \widehat{\widehat{J}}(s,x) \left(\int_{\mathbb{T}^3} \widehat{\widehat{\mathbf{D}}}(s,x,y)^\top \nabla_y \widehat{\widehat{\omega}}(s,x,y)\,\mathrm{d}y \right)^\top \widehat{\widehat{J}}(s,x)\,\mathrm{d}\beta(s).$$

Using the cell problem, we arrive at the desired expression for the effective diffusion. As the computations are exactly similar to the ones present in the proof of Proposition 9.2.1, we skip the details. □

9.5 Discussion of assumptions

In this section we discuss the assumptions (detailed in Section 9.4.2) of the homogenization result (Theorem 9.4.1) and give both the examples where they are satisfied and counterexamples where the failure of these assumptions can lead either to trivial or non-unique limits. We remark that the main obstacle to obtaining homogenization results in our setting is, in fact, not related to the oscillating coefficients, but rather to deriving an effective equation in Lagrangian coordinates.

9.5.1 Bounds on the Jacobian

The main restriction on the fluid flow is Assumption 9.4.4 which implies that the Jacobian of the flow is *uniformly bounded in time*. This is a highly non-generic assumption, but is needed for the validity of the posited asymptotic expansion (9.1.4). Indeed, if the Jacobian is not uniformly bounded in time, then, for example, the right hand side of the $\mathcal{O}(\varepsilon^0)$ equation in the cascade (9.2.20) may grow to be of $\mathcal{O}(\varepsilon^{-1})$ for sufficiently large values of the fast time variable τ, breaking the formal expansion. First we shall give some examples of mean fluid fields which obey this assumption, which although restrictive, still covers a large class of vector fields.

Example 9.5.1 (Constant drift)**.** *The most obvious example is the constant drift flow* $\bar{\mathbf{b}}(x) = \mathbf{b}^*$ *for a constant vector* $\mathbf{b}^* \in \mathbb{R}^d$*. In this case the Jacobian matrix is the identity for all times. This case falls under the regime of* two-scale convergence with drift *studied in [166, 5].*

Example 9.5.2 (Euclidean motions)**.** *Euclidean motions are the composition of a translation and a rigid rotation. An autonomous flow consists of Euclidean motions if and only if the vector field is given by* $\bar{\mathbf{b}}(x) = \mathbf{A}x + \mathbf{b}^*$ *for a constant skew-symmetric matrix* $\mathbf{A}$ *and a constant vector* $\mathbf{b}^*$*. The associated Jacobian matrix is an orthogonal matrix and hence of norm* 1*.*

Example 9.5.3 (Asymptotically constant drift)**.** *Let the mean flow* $\bar{\mathbf{b}}$ *in dimension* $d \geq 2$ *be given by*

$$\bar{\mathbf{b}}(x) = \begin{cases} \mathbf{b}^* & \text{when } x_1 < -R, \\ \mathbf{c}(x) & \text{when } x_1 \in [-R, R], \\ \mathbf{b}^{**} & \text{when } x_1 > R, \end{cases}$$

where $R > 0$*,* $\mathbf{e}_1 \cdot \mathbf{b}^*, \mathbf{e}_1 \cdot \mathbf{b}^{**} > 0$ *and* $\mathbf{c}(x)$ *is chosen to make* $\bar{\mathbf{b}}$ *continuously differentiable and divergence free. To ensure that the Jacobian of the flow is uniformly bounded in time we require that any integral curve spends only finite time* T *in* $\{x_1 \in [-R, R]\}$*, which implies that the Jacobian is norm bounded by* $C \exp(T\|\nabla \mathbf{c}\|_{L^\infty})$*. This can easily be achieved by requiring that* $\mathbf{e}_1 \cdot \mathbf{c}(x) \geq c > 0$*.*

We remark also that the Jacobian in each of these examples belongs to some algebra w.m.v., specifically the Jacobian in Examples 9.5.1 and 9.5.3 belong to the algebra of functions that converge at infinity (see Example 9.3.2), and the Jacobian in Example 9.5.2 belongs to the algebra of almost periodic functions (see Example 9.3.3).

9.5.2 Necessity of uniformly bounded Jacobian

The assumption of uniform bounds on the Jacobian is not a mere technical assumption. We illustrate this with a counterexample, which we have made as simple as possible to allow explicit calculations.

Counterexample 9.5.1 (Blow-up of the Jacobian for a shear flow)**.** *Consider the simplest example of a shear flow:*

$$\bar{\mathbf{b}}(x_1, x_2) = \begin{bmatrix} x_2 \\ 0 \end{bmatrix}.$$

An easy computation gives that the flow Φ generated by this vector field and its Jacobian are given by

$$\Phi_\tau(x_1, x_2) = \begin{bmatrix} x_1 + \tau x_2 \\ x_2 \end{bmatrix}, \qquad J(-\tau, x) = \begin{bmatrix} 1 & \tau \\ 0 & 1 \end{bmatrix}.$$

In particular, the Jacobian grows linearly in time.

Consider the parabolic problem on $]0,T[\times\mathbb{R}^2$ given by

$$\frac{\partial u^\varepsilon}{\partial t} + \frac{1}{\varepsilon}\bar{\mathbf{b}}(x_1, x_2) \cdot \nabla u^\varepsilon - \Delta u^\varepsilon = 0 \tag{9.5.1}$$

(Note that this example does not have oscillating coefficients.) The posited asymptotic expansion (9.1.4) becomes in this case

$$u^\varepsilon(t, x_1, x_2) \;\approx\; u_0\left(t, \frac{t}{\varepsilon}, x_1 - \frac{x_2 t}{\varepsilon}, x_2\right) \;+\; \varepsilon\, u_1\left(t, \frac{t}{\varepsilon}, x_1 - \frac{x_2 t}{\varepsilon}, x_2\right) \;+\; \cdots$$

and the cascade of equations (9.2.20) becomes

$$\begin{aligned}
\mathcal{O}(\varepsilon^{-2}): \quad & 0 = 0, \\
\mathcal{O}(\varepsilon^{-1}): \quad & 0 = -\frac{\partial u_0}{\partial \tau}, \\
\mathcal{O}(\varepsilon^{0}): \quad & 0 = -\frac{\partial u_0}{\partial t} - \frac{\partial u_1}{\partial \tau} + (1+\tau^2)\frac{\partial^2 u_0}{\partial X_1^2} - 2\tau\frac{\partial^2 u_0}{\partial X_1 \partial X_2} + \frac{\partial^2 u_0}{\partial X_2^2},
\end{aligned}$$

where the $\mathcal{O}(\varepsilon^{-2})$ equation is trivial due to the lack of oscillating coefficients. But this cascade is only valid for times $\tau \ll \varepsilon^{-1/2}$, as at this value of τ the $(1+\tau^2)$ coefficient in the posited $\mathcal{O}(\varepsilon^0)$ equation jumps order. To be a valid asymptotic expansion for the parabolic problem (9.5.1), we require it to be valid for $\tau \in [0, T/\varepsilon]$, i.e. up to $\mathcal{O}(\varepsilon^{-1})$ values of τ. This means that the posited asymptotic expansion cannot be correct.

Indeed, the problem (9.5.1) can be explicitly solved using the Fourier transform. Let (ξ_1, ξ_2) be Fourier variables corresponding to (x_1, x_2). Then an easy computation yields

$$\hat{u}^\varepsilon(t, \xi_1, \xi_2) = \exp\left(\int_0^t -|\xi_1|^2 - \left|\xi_2 - \frac{s}{\varepsilon}\xi_1\right|^2 \mathrm{d}s\right)\hat{u}^\varepsilon(0, \xi_1, \xi_2),$$

(where we have abused notation and used $\hat{\cdot}$ to denote the Fourier instead of Gelfand transform). The integrand in the above exponential converges pointwise to $-\infty$ as $\varepsilon \to 0$ so long as $s\xi_1 \neq 0$. Therefore, $\hat{u} \to 0$ almost everywhere in $[0,T] \times \mathbb{R}^2$ as $\varepsilon \to 0$, and it follows from dominated convergence and Plancherel's theorem that $u^\varepsilon \to 0$ strongly in $L^2([0,T] \times \mathbb{R}^2)$. So, not only is the asymptotic expansion not correct, but the limit as $\varepsilon \to 0$ is trivial.

This counterexample illustrates a general phenomenon for shear flows, where the convection enhances the diffusion (see for example [58] where this is considered in detail for the more complicated case of cats eye flows, and [79, 40, 21] where conditions under which the solution converges strongly to zero are studied). As a consequence of this enhancement, the time scale on which diffusion is observed is different and one should not expect to obtain a non-trivial limit in the scaling we consider. We give a partial result to this effect below. The authors shall address this problem in a forthcoming publication [89].

Proposition 9.5.1. *Let the assumptions of Proposition 9.4.1 hold and u^ε be the solution to (9.2.4a)-(9.2.4b). Let v^ε be the solution in Lagrangian coordinates, i.e. $v^\varepsilon(t,x) = u^\varepsilon(t, \Phi_{t/\varepsilon}(x))$. Let $\xi \in \mathbb{R}^d$ be a unit vector, and suppose that for some (non-empty) open set $A \subset \mathbb{R}^d$ we have, for $x \in A$,*

$$\lim_{\tau \to \infty} |{}^\top \widetilde{J}(\tau, x)\xi| = \infty. \tag{9.5.2}$$

Then for any v_0 a $L^2((0,T); H^1(\mathbb{R}^d))$-weak limit of v^ε, we have $\xi \cdot \nabla_X v_0 = 0$ on $(0,T) \times A$.

Remark 9.5.1. *As the set A is independent of the choice of initial data u^{in}, the initial data can be chosen so that $v_0 \notin C([0,T]; L^2(\mathbb{R}^d))$, and in particular so that v_0 does not solve a 'nice' parabolic PDE with this initial datum.*

Proof. Without loss of generality we can assume that A is bounded, and by applying Egorov's theorem it is sufficient to prove the claim for A measurable with the limit (9.5.2) uniform on A. By converting (9.4.1) to Lagrangian coordinates, we have the estimate

$$\iint_{(0,T)\times\mathbb{R}^d} |{}^\top \widetilde{J}(t/\varepsilon, x) \nabla_X v^\varepsilon(t,x)|^2 \, \mathrm{d}x\mathrm{d}t \leq C$$

with C only depending on u^{in}. Let $t_0 \in (0,T)$ be arbitrary, and define

$$\theta(\varepsilon) = \inf_{X\in A,\tau\geq t_0/\varepsilon} |^\top \widetilde{J}(\tau,x)\xi|^2,$$

so that $\theta(\varepsilon)\to\infty$ as $\varepsilon\to 0$. Then we have

$$\begin{aligned}
\iint\limits_{(t_0,T)\times A} |\xi\cdot\nabla_X v^\varepsilon(t,x)|^2\,\mathrm{d}x\mathrm{d}t &\leq \theta(\varepsilon)^{-1} \iint\limits_{(t_0,T)\times A} |^\top\widetilde{J}(t/\varepsilon,x)\xi|^2|\xi\cdot\nabla_X v^\varepsilon(t,x)|^2\,\mathrm{d}x\mathrm{d}t\\
&\leq \theta(\varepsilon)^{-1} \iint\limits_{(t_0,T)\times A} |^\top\widetilde{J}(t/\varepsilon,x)\nabla_X v^\varepsilon(t,x)|^2\,\mathrm{d}x\mathrm{d}t\\
&\leq \theta(\varepsilon)^{-1} \iint\limits_{(0,T)\times\mathbb{R}^d} |^\top\widetilde{J}(t/\varepsilon,x)\nabla_X v^\varepsilon(t,x)|^2\,\mathrm{d}x\mathrm{d}t\\
&\leq C\theta(\varepsilon)^{-1}.
\end{aligned}$$

That $\xi\cdot\nabla_X v_0 = 0$ on (t_0,A) follows from upper semi-continuity under weak convergence. That this holds for $(0,T)\times A$ follows as t_0 was arbitrary. □

9.5.3 Flow representations of coefficients

The main assumptions upon the coefficients $\mathbf{b}(x,y)$ and $\mathbf{D}(x,y)$ is their flow representations $\tilde{\mathbf{b}}(\tau,x,y)$ and $\tilde{\mathbf{D}}(\tau,x,y)$ belong to some fixed algebra w.m.v. $\mathcal{A}$. The reason we require this is to ensure that we obtain a single unique homogenized equation. This is in contrast to the uniform bounds on the Jacobian, which as described above, we require in order to be sure that we can obtain *any* non-trivial limit. We illustrate the non-uniqueness phenomenon with the following counterexample. We remark that, again, the difficulty is present without any rapid spatial oscillations, or complicated mean flows.

Counterexample 9.5.2 (Non-uniqueness of the limit)**.** *Consider the* $1+1$ *dimensional parabolic problem on* $]0,T[\times\mathbb{R}$ *given by*

$$\frac{\partial u^\varepsilon}{\partial t} + \frac{1}{\varepsilon}\frac{\partial u^\varepsilon}{\partial x} - \frac{\partial}{\partial x}\left(\mathbf{D}(x)\frac{\partial u^\varepsilon}{\partial x}\right) = 0, \tag{9.5.3}$$

where $\mathbf{D}(x)$ *is given by*

$$\mathbf{D}(x) = \begin{cases} 1 & \text{if } |x| \in [2^{(2n)^2}, 2^{(2n+1)^2}) \text{ for some integer } n \geq 0, \\ 2 & \text{otherwise.} \end{cases} \tag{9.5.4}$$

(Note that although this function $\mathbf{D}$ *is not continuous, the example could be easily modified to have* $\mathbf{D} \in C^\infty$*.) The corresponding mean flow field* $\bar{\mathbf{b}}$ *and its flow and Jacobian are given by*

$$\bar{\mathbf{b}}(x) = 1, \qquad \Phi_\tau(x) = x + \tau, \qquad J(\tau, x) = 1,$$

i.e. we are in the constant drift case. The posited asymptotic expansion (9.1.4) *becomes*

$$u^\varepsilon(t, x_1, x_2) \approx u_0\left(t, \frac{t}{\varepsilon}, x - \frac{t}{\varepsilon}\right) + u_1\left(t, \frac{t}{\varepsilon}, x - \frac{t}{\varepsilon}\right) + \cdots$$

and the cascade of equations (9.2.20) *is*

$$\begin{aligned} \mathcal{O}(\varepsilon^{-2}): \quad 0 &= 0, \\ \mathcal{O}(\varepsilon^{-1}): \quad 0 &= -\frac{\partial u_0}{\partial \tau}, \\ \mathcal{O}(\varepsilon^{0}): \quad 0 &= -\frac{\partial u_0}{\partial t} - \frac{\partial u_1}{\partial \tau} + \frac{\partial}{\partial X}\left(\tilde{\mathbf{D}}(\tau, X)\frac{\partial u_0}{\partial X}\right), \end{aligned}$$

where the simplicity of the equations is due to the lack of fast spatial oscillations, and the flow representation of $\mathbf{D}$ *is given by*

$$\tilde{\mathbf{D}}(\tau, X) = \mathbf{D}(X + \tau).$$

Unlike in the previous Counterexample 9.5.1, there is nothing obviously wrong with this asymptotic expansion. The problem comes when we try to average the $\mathcal{O}(\varepsilon^0)$ *equation in the fast time variable* τ*. Consider the limit*

$$(M\tilde{\mathbf{D}})(X) = \lim_{l \to \infty} \frac{1}{2l} \int_{-l}^{l} \tilde{\mathbf{D}}(\tau, X)\, \mathrm{d}\tau.$$

We claim that this limit does not exist. Indeed, let $l_n = 2^{(2n)^2}$ *then for* $n \geq 1$,

$$\left| \frac{1}{2l_n} \int_{-l_n}^{l_n} \tilde{\mathbf{D}}(\tau, 0)\, \mathrm{d}\tau - 2 \right| \leq \frac{2 \cdot 2^{(2n-1)^2}}{2 \cdot 2^{(2n)^2}} = 2^{-4n+1} \to 0 \text{ as } n \to \infty, \tag{9.5.5}$$

as the contribution from $|\tau| \in [2^{(2n-1)^2}, 2^{(2n)^2})$ *dominates. Similarly, for* $l'_n = 2^{(2n-1)^2}$ *we obtain*

$$\lim_{n\to\infty} \frac{1}{2l'_n} \int_{-l'_n}^{l'_n} \tilde{\mathbf{D}}(\tau, 0)\, \mathrm{d}\tau = 1.$$

Thus the limit $l \to \infty$ *depends upon the choice of sequence.*

We remark that, as a consequence, $\tilde{\mathbf{D}}$*, cannot belong to any algebra w.m.v., and therefore none of our results apply to this problem.*

One might think that the failure of the asymptotic expansion in the above counterexample could be rectified by some smarter choice of expansion. However, this is not the case. For this counterexample it is impossible to obtain a unique homogenised equation as we will now show.

Proposition 9.5.2. *Let* $v^\varepsilon(t, x) = u^\varepsilon(t, x + t/\varepsilon)$ *where* u^ε *solves the problem* (9.5.3) *in Counterexample 9.5.2 with initial data* $u^{in} \in L^2(\mathbb{R})$*. Then there are two sequences* $\varepsilon \to 0$ *and* $\varepsilon' \to 0$ *such* $v^\varepsilon \rightharpoonup u_0$ *and* $v^{\varepsilon'} \rightharpoonup u'_0$ *weakly in* $L^2([0, T] \times \mathbb{R})$*, which solve the homogenised problems*

$$\frac{\partial u_0}{\partial t} - \frac{\partial^2 u_0}{\partial x^2} = 0, \qquad \text{and} \qquad \frac{\partial u'_0}{\partial t} - 2\frac{\partial^2 u'_0}{\partial x^2} = 0, \tag{9.5.6}$$

with $u_0(0, x) = u'_0(0, x) = u^{in}(x)$*, which are different equations.*

Proof. Without loss of generality let $T = 1$. Define $\varepsilon_n = 1/l_n$ and $\varepsilon'_n = 1/l'_n$ for $l_n = 2^{(2n)^2}$ and $l'_n = 2^{(2n-1)^2}$ as in Counterexample 9.5.2. We will first show that the following strong convergences

$$\tilde{\mathbf{D}}(t/\varepsilon_n, x) \to 2, \qquad \text{and} \qquad \tilde{\mathbf{D}}(t/\varepsilon'_n, x) \to 1,$$

hold in $L^2_{loc}([0, 1] \times \mathbb{R})$ as $n \to \infty$. Indeed, let $K \in (0, \infty)$ be arbitrary, then,

$$\int_0^1 \int_{-K}^{K} |\tilde{\mathbf{D}}(t/\varepsilon_n, x) - 2|^2 \, \mathrm{d}t\mathrm{d}x = \int_{-K}^{K} \int_0^{T_n(X)} |\mathbf{D}(x + t/\varepsilon_n) - 2|^2 \, \mathrm{d}t\mathrm{d}x + 0 \tag{9.5.7}$$

where $T_n(x)$ is chosen as the solution of $x + T_n(x)/\varepsilon_n = 2^{(2n-1)^2}$ (or 0 if this is negative) so that $\mathbf{D}(x + t/\varepsilon) = 2$ for $t \in [T_n, 1]$. Hence

$$T_n(x) = 2^{(2n-1)^2-4n^2} + x\varepsilon_n \leq 2^{-4n+1} + K\varepsilon_n \to 0 \text{ as } n \to \infty,$$

and the convergence of (9.5.7) to zero follows easily. The proof that $\tilde{\mathbf{D}}(t/\varepsilon'_n)$ converges to 1 is similar, where instead $T_n(x)$ is chosen so that $\mathbf{D}(x + t/\varepsilon'_n) = 1$ for $t \in [T_n(y), 1]$.

We will now show the convergence v^{ε_n} to u_0. The argument for $v^{\varepsilon'_n}$ is analogous, using instead the convergence of $\tilde{\mathbf{D}}(t/\varepsilon'_n) \to 2$, and we leave it to the reader. Straight forward estimates allow us to pass to a subsequence n_k on which $v_{\varepsilon_{n_k}}(t, x)$ and $\frac{\partial v^{\varepsilon_{n_k}}}{\partial X}$ converge $L^2([0, 1] \times \mathbb{R})$-weak to limits u_0 and $\frac{\partial u_0}{\partial X}$ as $k \to \infty$. Uniqueness of solutions of the equation for u_0 will later show that $v^{\varepsilon_n} \to u_0$ as $n \to \infty$, i.e. the original sequence converges. We abuse notation and keep the original sequence. Writing (9.5.3) in $(t, x) = (t, x - t/\varepsilon)$ coordinates, multiplying by a test function $\varphi(t, x)$ and integrating by parts, we obtain

$$\begin{aligned}\int_{-\infty}^{\infty} \varphi(0, x)u^{in}(0, x)\,\mathrm{d}x - \int_0^1 \int_{-\infty}^{\infty} \frac{\partial \varphi}{\partial t}(t, x)v^{\varepsilon_n}(t, x)\,\mathrm{d}t\mathrm{d}x \\ + \int_0^1 \int_{-\infty}^{\infty} \tilde{\mathbf{D}}(t/\varepsilon_n, x)\frac{\partial \varphi}{\partial x}(t, x)\frac{\partial v^{\varepsilon_n}}{\partial x}(t, x)\,\mathrm{d}x\mathrm{d}t = 0.\end{aligned}$$

By the weak convergences of $v^{\varepsilon_n} \to u_0$, $\frac{\partial v^{\varepsilon_n}}{\partial X} \to \frac{\partial u_0}{\partial X}$, the strong convergence $\tilde{\mathbf{D}}(t/\varepsilon_n, x) \to 2$ and the compact support of φ we can pass to the limit as $n \to \infty$ in each of these terms to obtain the weak formulation of the equation (9.5.6) for u_0. □

We remark that, although the above counterexample features bad behaviour in the diffusion coefficient, similar examples could be constructed where the undesirable behaviour is in the drift term $\bar{\mathbf{b}}$ (or $\mathbf{b}$) or the Jacobian J. The issue here is the appearance of the spatial scale $x = \mathcal{O}(\varepsilon^{-1})$ in the problem due to the $\mathcal{O}(\varepsilon^{-1})$ mean drift. Such a scale is not present when the average convection is zero, i.e. $\bar{\mathbf{b}} = 0$, even in the convection dominated regime. This additional spatial scale is exploited in the choice of diffusion coefficient (9.5.4), which exhibits different behaviour at a sequence of spatial scales tending to infinity.

Next we show that this bad behaviour is a problem only at infinity, in the sense that if the trajectories of Φ are bounded, then our assumptions always hold.

Proposition 9.5.3. *Let Assumption 9.2.1 hold. Then exactly one of the following hold:*

(i) Φ_τ *has bounded orbits, i.e. for any* x *the set* $\{\Phi_\tau(x) : \tau \in \mathbb{R}\}$ *is bounded.*

(ii) Φ_τ *converges to infinity, i.e. for any* x *we have* $|\Phi_\tau(x)| \to \infty$ *as* $|\tau| \to \infty$.

Let $\mathcal{AP}$ *denote the algebra of almost-periodic functions (Example 9.3.3). In each respective case the following also holds:*

(i) Φ_t *is uniformly almost-periodic, i.e.* $\Phi_t(x) \in [C(\mathbb{R}^d; \mathcal{AP})]^d$. *For every* $f(x, y) \in C(\mathbb{R}^d \times \mathbb{T}^d)$, *the flow-representation is uniformly almost-periodic, i.e.* $\tilde{f} \in C(\mathbb{R}^d \times \mathbb{T}^d; \mathcal{AP})$. *If additionally* $J(\tau, x)$ *is uniformly continuous on* $\mathbb{R} \times K$ *for each compact set* $K \subset \mathbb{R}^d$, *then* J *and* $\tilde{J}$ *are uniformly almost-periodic, i.e.* $J, \tilde{J} \in [C(\mathbb{R}^d; \mathcal{AP})]^{d \times d}$.

(ii) Let $f \in C(\mathbb{R}^d \times \mathbb{T}^d)$ *converge to a limit as* $|x| \to \infty$, *i.e.* $\lim_{|x| \to \infty} f(x, y)$ *exists and is finite for each* $y \in \mathbb{T}^d$. *Then for each* $x, y \in \mathbb{R}^d \times \mathbb{T}^d$, *the flow-representation* $\tilde{f}(\cdot, x, y)$ *belongs to the algebra of functions that converge at infinity (Example 9.3.2).*

Remark 9.5.2. *The almost-periodicity of the Jacobian of an (locally) uniformly almost-periodic flow is a subtle issue as there are uniformly almost-periodic functions whose derivative is not uniformly almost-periodic. The assumption in (i) that* J *is uniformly continuous is to side steps this issue.*

To prove this proposition we have need a definition.

Definition 9.5.1 (Equicontinuous flow)**.** *A one-parameter group* ϕ_τ *of homeomorphisms of* $K \subseteq \mathbb{R}^d$ *is* equicontinuous *if for any* $\varepsilon > 0$ *and* $x \in K$ *there is a* $\delta = \delta(x, \varepsilon)$ *such that whenever* $|x' - x| \leq \delta$ *and* $x' \in K$ *it holds that* $|\phi_\tau(x) - \phi_\tau(x')| \leq \varepsilon$ *for all* $\tau \in \mathbb{R}$.

Proof of Proposition 9.5.3. We first prove the dichotomy. Let $x \in \mathbb{R}^d$ be fixed. We first claim that either $|\Phi_\tau(x)| \to \infty$ as $|\tau| \to \infty$ or its orbit is bounded. Suppose $|\Phi_\tau(x)| \not\to \infty$ as $|\tau| \to \infty$, then there must be a compact set K containing

x and a sequence of times τ_n with $|\tau_n| \to \infty$ as $n \to \infty$ and $\Phi_{\tau_n}(x) \in K$. Without loss of generality let $0 < \tau_1 < \tau_2 \cdots$. By integrating Assumption 9.2.1, for any $n \geq 1$ it holds that

$$\sup_{\tau_n \leq \tau \leq \tau_{n+1}} |\Phi_\tau(x) - \Phi_{\tau-\tau_n}(x)| \leq C|\Phi_{\tau_n}(x) - x| \leq C \operatorname{diam}(K)$$

and hence the forward orbit is bounded. We now claim that the backwards orbit is also bounded. Indeed, let $s > 0$ be arbitrary, then, using Assumption 9.2.1 once more we have

$$|\Phi_{-s}(x) - x| = |\Phi_{-s}(x) - \Phi_{-s}(\Phi_s(x))| \leq C|x - \Phi_s(x)|$$

and the right hand side is bounded uniformly in $s > 0$.

We have shown that the dichotomy holds for some fixed x, but this together with Assumption 9.2.1 imply that the same dichotomy holds for all x. Indeed, consider the orbit starting from an arbitrary x', then

$$\sup_{\tau \in \mathbb{R}} |\Phi_\tau(x) - \Phi_\tau(x')| \leq C|x - x'| < \infty$$

which we obtain by again integrating Assumption 9.2.1. This implies that if the orbit of x is bounded (resp. converges to infinity) then the orbit of x' is bounded (resp. converges to infinity).

We now prove the claims, starting with (i). Let $R > 0$ be arbitrary, then the set

$$K_R = \overline{\{\Phi_\tau(x) : \tau \in \mathbb{R}, |x| \leq R\}}$$

is invariant under Φ_τ and compact. Moreover, the K_R are nested and cover $\mathbb{R}^d$. It is thus sufficient to prove the claims on K_R. Note that $(\Phi_\tau, K, |\cdot|)$ is a compact dynamical system, and Assumption 9.2.1 implies that it is *equicontinuous* in the sense of the above definition. It is a classical result of topological dynamical systems (see e.g. [56]) that for compact dynamical systems the property of equicontinuity is equivalent to being *uniformly almost-periodic*, in the sense that $\Phi_\tau(x) \in [C(K_R; \mathcal{AP})]^d$. Now suppose that $f \in C(\mathbb{R}^d \times \mathbb{T}^d)$, then f is uniformly continuous on $K_R \times \mathbb{T}^d$ as this set is compact. Moreover, as K_R is invariant under Φ_τ the function $\tilde{f}(\tau, x, y)$ restricted to $x \in K_R$ depends only on f restricted to

$K_R \times \mathbb{T}^d$. Hence $\tilde{f}(\tau, x, y) = f(\Phi_\tau(x), y)$ (restricted to $x \in K_R$) is the composition of a uniformly continuous function and a uniformly almost-periodic function, and is uniformly almost-periodic. Finally, suppose that J is uniformly continuous on $\mathbb{R} \times K_R$, then the difference quotients defined for any unit vector $\xi \in \mathbb{R}^d$, by

$$J_h(\tau, x)\xi = \frac{\Phi_{-\tau}(x + h\xi) - \Phi_{-\tau}(x)}{|h|}$$

converge in $[C(K_{R/2} \times \mathbb{R})]^d$ as $\mathbb{R} \ni h \to 0$ to $J(\tau, x)\xi$. As both terms in the difference quotient are uniformly almost-periodic the limit is also. That $\tilde{J}$ is uniformly almost-periodic can be proved in the same way.

Now we prove the claim for (ii). Let $f(x, y) \in C(\mathbb{R}^d \times \mathbb{T}^d)$ be as assumed and converge to $g(y)$ as $|x| \to \infty$. Clearly $\tilde{f}(\tau, x, y) \to g(y)$ as $|\tau| \to \infty$, it only remains to show that $\tilde{f}$ is uniformly continuous. To this end, note that f is continuous on $\overline{\mathbb{R}^d} \times \mathbb{T}$ where $\overline{\mathbb{R}^d}$ is the one-point compactification of $\mathbb{R}^d$. As this set is compact, f is uniformly continuous on this set. Moreover, as $\Phi_\tau(x)$ is uniformly continuous from $\mathbb{R} \times \mathbb{R}^d$ to $\overline{\mathbb{R}^d}$, the composition $\tilde{f}$ is uniformly continuous, which completes the proof of the proposition. □

9.6 Applications to other models

In this section, we consider some explicit models and perform the asymptotic analysis using the Σ-Φ_τ convergence.

9.6.1 Lagrangian coordinates

For a smooth fluid field $\bar{\mathbf{b}}(x) \in C^1(\mathbb{R}^d; \mathbb{R}^d)$ and diffusion coefficient $\mathbf{D}(x) \in L^\infty(\mathbb{R}^d; \mathbb{R}^{d \times d})$, consider the Cauchy problem with large convection term

$$\frac{\partial u^\varepsilon}{\partial t} + \frac{1}{\varepsilon}\bar{\mathbf{b}}(x) \cdot \nabla u^\varepsilon - \nabla \cdot \Big(\mathbf{D}(x)\nabla u^\varepsilon\Big) = 0 \qquad \text{for } (t, x) \in]0, T[\times \mathbb{R}^d. \quad (9.6.1)$$

Let $\Phi_\tau(x)$ be the flow associated with the vector field $\bar{\mathbf{b}}(x)$. As Remark 9.3.3 suggests, we consider the Σ-Φ_τ convergence with no oscillations in space, i.e.

with test functions $\psi\left(t, \Phi_{-t/\varepsilon}(x), \frac{t}{\varepsilon}\right)$:

$$\lim_{\varepsilon\to 0} \iint\limits_{(0,T)\times\mathbb{R}^d} u^\varepsilon(t,x)\psi\left(t, \Phi_{-t/\varepsilon}(x), \frac{t}{\varepsilon}\right) \mathrm{d}x\,\mathrm{d}t = \iiint\limits_{(0,T)\times\mathbb{R}^d\times\Delta(\mathcal{A})} u_0(t,x,s)\psi(t.x,s)\,\mathrm{d}\beta(s)\,\mathrm{d}x\,\mathrm{d}t. \tag{9.6.2}$$

An argument similar to the proof of Lemma 9.4.3 implies that the above limit function u_0 is independent of the s variable. As done earlier in Section 9.4, we need to pass to the limit (as $\varepsilon \to 0$) in the weak formulation

$$\begin{aligned}
&-\iint\limits_{(0,T)\times\mathbb{R}^d} u^\varepsilon(t,x)\frac{\partial\psi}{\partial t}\left(t, \Phi_{-t/\varepsilon}(x)\right) \mathrm{d}x\,\mathrm{d}t - \int\limits_{\mathbb{R}^d} u^{in}(x)\psi(0,x)\,\mathrm{d}x \\
&+\iint\limits_{(0,T)\times\mathbb{R}^d} \widetilde{\mathbf{D}}\left(\frac{t}{\varepsilon}, \Phi_{-t/\varepsilon}(x)\right)\nabla u^\varepsilon(t,x)\cdot {}^\top\widetilde{J}\left(\frac{t}{\varepsilon}, \Phi_{-t/\varepsilon}(x)\right)\nabla_x\psi\left(t, \Phi_{-t/\varepsilon}(x)\right) \mathrm{d}x\,\mathrm{d}t = 0.
\end{aligned}$$

From the Assumption 9.4.3 on the diffusion coefficient $\mathbf{D}(x)$ and the Assumption 9.4.4 on the Jacobian matrix $J(\tau,x)$, it follows that the product expression ${}^\top\mathbf{D}(\tau,x)^\top J(\tau,x)\nabla_x\psi(t,x)$ is an admissible test function in the sense of Definition 9.3.8. Hence, passing to the limit yields

$$\begin{aligned}
&-\iint\limits_{(0,T)\times\mathbb{R}^d} u_0(t,x)\,\mathrm{d}x\,\mathrm{d}t - \int\limits_{\mathbb{R}^d} u^{in}(x)\,\mathrm{d}x \\
&\quad+\iiint\limits_{(0,T)\times\mathbb{R}^d\times\Delta(\mathcal{A})} \widehat{\widetilde{J}}(s,x)\widehat{\widetilde{\mathbf{D}}}(s,x)^\top\widehat{\widetilde{J}}(s,x)\nabla_x u_0(t,x)\cdot\nabla_x\psi(t,x)\,\mathrm{d}\beta(s)\,\mathrm{d}x\,\mathrm{d}t = 0.
\end{aligned}$$

Remark 9.6.1. *In the above computation, passing to the limit as $\varepsilon \to 0$ using Σ-Φ_τ convergence amount to arrive at a limit equation which is in Lagrangian coordinates*

$$\frac{\partial u_0}{\partial t} - \nabla_x\cdot\left(\mathfrak{D}(x)\nabla_x u_0\right) = 0$$

where the diffusion coefficients are given by

$$\mathfrak{D}(x) = \int\limits_{\Delta(\mathcal{A})} \widehat{\widetilde{J}}(s,x)\widehat{\widetilde{\mathbf{D}}}(s,x)^\top\widehat{\widetilde{J}}(s,x)\,\mathrm{d}\beta(s).$$

9.6.2 Fluid field with $\mathcal{O}(\varepsilon)$ perturbation

In a next transport model, we consider a smooth fluid field with a particular structure

$$\mathbf{b}\left(x, \frac{x}{\varepsilon}\right) = \mathbf{h}\left(\frac{x}{\varepsilon}\right) + \varepsilon \mathbf{h}^1\left(x, \frac{x}{\varepsilon}\right). \tag{9.6.3}$$

The convection-diffusion equation the we consider is

$$\begin{aligned} &\frac{\partial u^\varepsilon}{\partial t} + \frac{1}{\varepsilon}\mathbf{h}\left(\frac{x}{\varepsilon}\right)\cdot\nabla u^\varepsilon + \mathbf{h}^1\left(x, \frac{x}{\varepsilon}\right)\cdot\nabla u^\varepsilon - \nabla\cdot\left(\mathbf{D}\left(\frac{x}{\varepsilon}\right)\nabla u^\varepsilon\right) = 0 \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{for } (t,x) \in (0,T)\times\mathbb{R}^d. \end{aligned} \tag{9.6.4}$$

As only the field $\mathbf{h}\left(\frac{x}{\varepsilon}\right)$ is of $\mathcal{O}(\varepsilon^{-1})$ in (9.6.4), we need to consider the flow associated with the mean field

$$\mathbf{h}^* := \int_{\mathbb{T}^d} \mathbf{h}(y)\,\mathrm{d}y, \qquad \text{i.e. } \Phi_\tau(x) = x + \mathbf{h}^*\tau.$$

This suggests the use of *two-scale convergence with drift* [140, 2]. The solution family u^ε satisfies the uniform a priori bounds:

$$\|u^\varepsilon\|_{L^2((0,T)\times\mathbb{R}^d)} \le C; \qquad \|\nabla u^\varepsilon\|_{L^2((0,T)\times\mathbb{R}^d)} \le C.$$

The compactness results in *two-scale convergence with drift* theory implies the existence of $u_0 \in L^2((0,T); H^1(\mathbb{R}^d))$ and $u_1 \in L^2((0,T)\times\mathbb{R}^d; H^1(\mathbb{T}^d))$ such that, on a subsequence,

$$\begin{aligned} &u^\varepsilon \xrightarrow{2-scale-\mathbf{h}^*} u_0(t,x); \\ &\nabla u^\varepsilon \xrightarrow{2-scale-\mathbf{h}^*} \nabla_x u_0(t,x) + \nabla_y u_1(t,x,y). \end{aligned} \tag{9.6.5}$$

The idea is indeed to pass to the limit in the weak formulation with

$$\psi\left(t, x - \frac{\mathbf{h}^* t}{\varepsilon}\right) + \varepsilon\psi_1\left(t, x - \frac{\mathbf{h}^* t}{\varepsilon}, \frac{x}{\varepsilon}\right)$$

as test function which vanish at time instant $t = T$.

$$
\begin{aligned}
&- \iint\limits_{(0,T)\times\mathbb{R}^d} u^\varepsilon(t,x)\frac{\partial\psi}{\partial t}\left(t, x-\frac{\mathbf{h}^* t}{\varepsilon}\right)\mathrm{d}x\,\mathrm{d}t \\
&+ \iint\limits_{(0,T)\times\mathbb{R}^d} u^\varepsilon(t,x)\mathbf{h}^*\cdot\nabla_x\psi_1\left(t, x-\frac{\mathbf{h}^* t}{\varepsilon}, \frac{x}{\varepsilon}\right)\mathrm{d}x\,\mathrm{d}t \\
&+ \frac{1}{\varepsilon}\iint\limits_{(0,T)\times\mathbb{R}^d} u^\varepsilon(t,x)\left(\mathbf{h}^* - \mathbf{h}\left(\frac{x}{\varepsilon}\right)\right)\cdot\nabla_x\psi\left(t, x-\frac{\mathbf{h}^* t}{\varepsilon}\right)\mathrm{d}x\,\mathrm{d}t \\
&+ \iint\limits_{(0,T)\times\mathbb{R}^d} \mathbf{h}^1\left(x,\frac{x}{\varepsilon}\right)\cdot\nabla u^\varepsilon(t,x)\psi\left(t, x-\frac{\mathbf{h}^* t}{\varepsilon}\right)\mathrm{d}x\,\mathrm{d}t \\
&+ \iint\limits_{(0,T)\times\mathbb{R}^d} \mathbf{D}\left(\frac{x}{\varepsilon}\right)\nabla u^\varepsilon(t,x)\cdot\left(\nabla_x\psi\left(t, x-\frac{\mathbf{h}^* t}{\varepsilon}\right) + \nabla_y\psi_1\left(t, x-\frac{\mathbf{h}^* t}{\varepsilon}, \frac{x}{\varepsilon}\right)\right)\mathrm{d}x\,\mathrm{d}t \\
&+ \mathcal{O}(\varepsilon) = 0.
\end{aligned}
$$

We can pass to the limit in almost all the terms in above expression except for the fourth term on the left hand side. This is essentially because the product $\mathbf{h}^1(x,y)\,\psi(t,x)$ does not form an admissible test function in the sense of Definition 9.3.8. However, if we consider the flow representation of the fluid field $\mathbf{h}^1(x,y)$ and assume that $\widetilde{\mathbf{h}^1}(\cdot,x,y) \in \mathcal{A}$ for certain ergodic algebra w.m.v. $\mathcal{A}$, then the product $\widetilde{\mathbf{h}^1}(\tau,x,y)\psi(t,x)$ forms an admissible test function in the sense of Definition 9.3.8. Thus, using the notion of weak Σ-Φ_τ convergence, we can prove the following result. The proof of which is a simple adaptation of the calculations already present in Section 9.4. Hence is left to the reader.

Theorem 9.6.1. *Suppose the flow representation of the fluid field $\mathbf{h}(x,y)$ belongs to an ergodic algebra w.m.v. $\mathcal{A}$. The two-scale with drift $\mathbf{h}^*$ limits for the solution family u^ε obtained in (9.6.5) satisfy the homogenized equation*

$$
\frac{\partial u_0}{\partial t} + \mathfrak{h}(x)\cdot\nabla u_0 - \nabla\cdot\left(\mathfrak{D}\nabla u_0\right) = 0 \qquad \text{for } (t,x)\in\,]0,T[\times\mathbb{R}^d, \tag{9.6.6}
$$

where the convective field in the homogenized equation is given by

$$
\mathfrak{h}(x) = \iint\limits_{\Delta(\mathcal{A})\times\mathbb{T}^d} \widehat{\widetilde{\mathbf{h}^1}}(s,x,y)\,\mathrm{d}\beta(s)\,\mathrm{d}y \tag{9.6.7}
$$

and the effective diffusion coefficient in the homogenized equation is given by

$$\mathfrak{D}_{ij} = \int_{\mathbb{T}^d} \mathbf{D}(y)\Big(\nabla_y \omega_j(y) + \mathbf{e}_j\Big) \cdot \Big(\nabla_y \omega_i(y) + \mathbf{e}_i\Big)\, \mathrm{d}y$$

for $i, j \in \{1, \cdots, d\}$ *where the* ω_i *solve the cell problem*

$$\mathbf{h}(y) \cdot (\nabla_y \omega_i + \mathbf{e}_i) - \nabla_y \cdot (\mathbf{D}(y)\,(\nabla_y \omega_i + \mathbf{e}_i)) = \mathbf{h}^* \cdot \mathbf{e}_i \qquad \text{in } \mathbb{T}^d.$$

Remark that, due to particular choice of the fluid field (9.6.3), the field $\mathbf{h}^1(x, \frac{x}{\varepsilon})$ only contributes to the homogenized equation (9.6.6) via the convective field (9.6.7) and not the effective diffusion coefficient. Only the fluid field of $\mathcal{O}(\varepsilon^{-1})$ contribute the dispersive effects in the effective diffusion coefficient.

Remark also that even in the constant drift scenario, the previously known *two-scale convergence with drift* developed in [140] has a handicap in dealing with coefficients that depend on the macroscopic variable. Hence, the notion of weak convergence developed in this work generalizes the known multiple scale techniques (in the spirit of two-scale convergence of Nguetseng and Allaire) in homogenization theory to a great extent.

9.7 Conclusion

The structural assumption of periodicity (in the y variable) on the fluid field $\mathbf{b}(x, y)$ made in the previous sections is for the sake of simplicity. We can indeed develop a theory of Σ-convergence along flows (similar to the theory developed in Section 9.3) under the assumption that the oscillations in space belong to certain *ergodic algebra with mean value*. To be precise, suppose $\mathbf{b}(x, y)$ a smooth fluid field which belongs to an ergodic algebra w.m.v. (say $\mathcal{A}_1$) in the y variable. By the definition of algebra w.m.v. (precisely, property (iii) in Definition 9.3.1), $\mathbf{b}(x, \cdot) \in \mathcal{A}_1$ possesses a mean value, i.e.

$$\mathbf{b}\left(x, \frac{x}{\varepsilon}\right) \rightharpoonup M\mathbf{b}(x) \quad \text{ in } L^\infty(\mathbb{R}^d)\text{-weak}^* \text{ as } \varepsilon \to 0.$$

In this scenario, we take the mean field $\bar{\mathbf{b}}(x) = M\mathbf{b}(x)$ and consider the flow Φ_τ associated with this mean field. To extend the notion of Σ-convergence along flows (Definition 9.3.6), we need to essentially characterize the limit

$$\lim_{\varepsilon\to 0} \iint\limits_{(0,T)\times\mathbb{R}^d} u^\varepsilon(t,x)\psi\left(t,\frac{t}{\varepsilon},\Phi_{-t/\varepsilon}(x),\frac{x}{\varepsilon}\right)\,\mathrm{d}x\,\mathrm{d}t$$

where the test function $\psi(t,\tau,x,y)$ belongs to an ergodic algebra w.m.v. (say $\mathcal{A}_2$) as a function of the fast time variable τ and belongs to an ergodic algebra w.m.v. $\mathcal{A}_1$ as a function of the y variable. To prove compactness result, in the spirit of Theorem 9.3.2, the approach is to consider the differentiation theory developed in the context of *algebras w.m.v.* developed in [155, 156, 35, 176]. We also need to approach it use the reiterated homogenization techniques as in [157]. The effective diffusion matrix obtained under the periodicity assumption (see (9.4.6)-(9.4.7)) is given in terms of the cell solutions obtained by solving elliptic problems on a torus. In this general setting, however, the expressions for effective diffusion shall involve solutions to some variational problems solved on the spectrum of the algebra w.m.v., i.e. $\Delta(\mathcal{A}_1)$ (cf. the works of Nguetseng [155, 156]). This potential theory of Σ-convergence along flows in a more general setting is quite intricate and is left for future investigations.

As is evident from Section 9.5.3, even in the constant drift case, one can only homogenize the convection-diffusion problems in strong convection regime provided the flow representation of the diffusion matrix belongs to an algebra w.m.v., i.e. satisfies Assumption 9.4.3.

All along this article, we have considered time-independent coefficients. This resulted in the study of autonomous ordinary differential systems (see (9.2.7)). Considering flows associated with non-autonomous systems would be interesting. But, the authors believe that the analysis would be very complicated and it remains to be checked if we can get compactness results (in the spirit of Theorem 9.3.2) for non-autonomous flows.

The assumption that the Jacobian matrices are bounded functions of the fast time variable is quite non-generic (see Section 9.5). To lift this assumption would require an enormous amount of work in the theory of Banach algebras. The main difficulty is the appearance of new time scales (as is evident from the shear

flow case considered in Counterexample 9.5.1). This problem largely remains to be solved. A partial result in this direction shall be given by the authors in a forthcoming publication [89].

Finally, the assumption of incompressibility on the fluid field has ensured that the associated flows are volume preserving (see (iv) in Assumption 9.3.1). This property of the flows has played an intricate role in our analysis, notably the proof of Lemma 9.3.2. It is worth mentioning [22] where they have treated the homogenization of convection-diffusion problem in strong convection regime where the fluid field is given by an harmonic potential. In the context of purely periodic fluid fields, there are works that consider compressible flows and perform the homogenization of convection-diffusion problems in strong convection regime (see [50, 4]). The approach is to employ a factorization principle to factor out oscillations from the solution via principal eigenfunctions of an associated spectral problem and to cancel any exponential decay in time of the solution using the principal eigenvalue of the same spectral problem. This approach has not been attempted in the literature for locally periodic coefficients.

9.A Appendix

In this section, we give the proof of Lemma 9.2.1 on some basic facts on the flows.

Proof of Lemma 9.2.1. We prove each claim in turn.

(i) Let $\varphi(x) \in C_c^\infty(\mathbb{R}^d;\mathbb{R})$ be an arbitrary test function and let the index i be arbitrary. By the chain rule,

$$\frac{\partial}{\partial x_i}\Big(\varphi\left(\Phi_{-\tau}(x)\right)\Big) = \sum_{j=1}^{d} \frac{\partial \varphi}{\partial X_j}\left(\Phi_{-\tau}(x)\right) \frac{\partial \Phi^j_{-\tau}}{\partial x_i}(x).$$

Integrating over $\mathbb{R}^d$ yields:

$$0 = \int_{\mathbb{R}^d} \frac{\partial}{\partial x_i}\left(\varphi\left(\Phi_{-\tau}(x)\right)\right) \mathrm{d}x = \int_{\mathbb{R}^d} \sum_{j=1}^{d} \frac{\partial \varphi}{\partial X_j}\left(\Phi_{-\tau}(x)\right) \frac{\partial \Phi^j_{-\tau}}{\partial x_i}(x)\, \mathrm{d}x.$$

Making the change of variables: $x = \Phi_{-\tau}(x)$, the above expression can be successively written as

$$0 = \int_{\mathbb{R}^d} \sum_{j=1}^{d} \frac{\partial \varphi}{\partial x_j}(x) \frac{\partial \Phi_{-\tau}^j}{\partial x_i} \left(\Phi_\tau(x)\right) \mathrm{d}x = \int_{\mathbb{R}^d} \nabla_x \varphi(x) \cdot \left(\widetilde{J}_{ji}(\tau, x)\right)_{j=1}^{d} \mathrm{d}x,$$

i.e. each column of $\widetilde{J}$ is divergence free in the sense of distributions, proving the claim.

(ii) We compute

$$\nabla_x \cdot \left(\widetilde{J}(\tau, x)\widetilde{f}(\tau, x)\right) = \widetilde{f}(\tau, x) \cdot \left(\nabla_x \cdot {}^\top\widetilde{J}(\tau, x)\right) + \widetilde{J}(\tau, x) : \nabla_x \widetilde{f}(\tau, x),$$

where : is the Frobenius inner product. The first term on the right hand side vanishes thanks to (i). For the second term, we use the flow representation to obtain

$$\nabla_x \cdot \left(\widetilde{J}(\tau, x)\widetilde{f}(\tau, x)\right) = \widetilde{J}(\tau, x) J(-\tau, x) : \nabla_x f(\Phi_\tau(x), y).$$

Thanks to the autonomy of the flow, the left side of the Frobenius product is the identity matrix. Therefore the above display vanishes as f is divergence free.

(iii) Performing an integration by parts, we have:

$$\begin{aligned}
&\int_{\mathbb{R}^d} \phi(x) \left({}^\top\widetilde{J}(\tau, x) \nabla_x \varphi(x)\right) \mathrm{d}x \\
&\quad = -\int_{\mathbb{R}^d} \phi(x)\varphi(x) \left(\nabla_x \cdot {}^\top\widetilde{J}(\tau, x)\right) \mathrm{d}x - \int_{\mathbb{R}^d} \varphi(x) \left({}^\top\widetilde{J}(\tau, x) \nabla_x \phi(x)\right) \mathrm{d}x.
\end{aligned}$$

The first term on the right hand side of the previous expression vanishes, thanks to (i). Hence, we have proved the result.

(iv) Consider the time derivatives for the i-th component:

$$\frac{\mathrm{d}}{\mathrm{d}\tau} \bar{\mathbf{b}}_i \left(\Phi_{-\tau}(x)\right) = -\bar{\mathbf{b}} \left(\Phi_{-\tau}(x)\right) \cdot \nabla_x \bar{\mathbf{b}}_i \left(\Phi_{-\tau}(x)\right) \tag{9.A.1}$$

and

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\tau}\left[\sum_{j=1}^{d} J_{ij}(\tau,x)\bar{\mathbf{b}}_j(x)\right] &= \frac{\mathrm{d}}{\mathrm{d}\tau}\left[\sum_{j=1}^{d}\frac{\partial\Phi^i_{-\tau}}{\partial x_j}(x)\bar{\mathbf{b}}_j(x)\right] \\ &= -\sum_{j=1}^{d}\frac{\partial}{\partial x_j}\Big(\bar{\mathbf{b}}_i\left(\Phi_{-\tau}(x)\right)\Big)\bar{\mathbf{b}}_j(x).\end{aligned} \tag{9.A.2}$$

The relation in (9.A.2) can be continued as

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\tau}\left[\sum_{j=1}^{d} J_{ij}(\tau,x)\bar{\mathbf{b}}_j(x)\right] &= -\sum_{j,k=1}^{d}\frac{\partial\bar{\mathbf{b}}_i}{\partial x_k}\left(\Phi_{-\tau}(x)\right)\frac{\partial\Phi^k_{-\tau}}{\partial x_j}(x)\bar{\mathbf{b}}_j(x) \\ &= -\nabla_x\bar{\mathbf{b}}_i\left(\Phi_{-\tau}(x)\right)\cdot\Big(J(\tau,x)\bar{\mathbf{b}}(x)\Big).\end{aligned} \tag{9.A.3}$$

Fix $x\in\mathbb{R}^d$ and define

$$g_i(\tau) := \bar{\mathbf{b}}_i\left(\Phi_{-\tau}(x)\right) - \left[\sum_{j=1}^{d} J_{ij}(\tau,x)\bar{\mathbf{b}}_j(x)\right]. \tag{9.A.4}$$

Then, from (9.A.1) and (9.A.3), we have:

$$\frac{\mathrm{d}}{\mathrm{d}t}g_i(\tau) = -\nabla_x\bar{\mathbf{b}}_i\left(\Phi_{-\tau}(x)\right)\cdot g(\tau).$$

As $g(0)=0$, a Grönwall type argument yields $g(\tau)=0$. Hence the result. □

Bibliography

[1] Grégoire Allaire. Homogenization and two-scale convergence. *SIAM J. Math. Anal.*, 23(6):1482–1518, 1992.

[2] Grégoire Allaire. Periodic homogenization and effective mass theorems for the Schrödinger equation. In *Quantum transport*, volume 1946 of *Lecture Notes in Math.*, pages 1–44. Springer, Berlin, 2008.

[3] Grégoire Allaire and Harsha Hutridurga. Homogenization of reactive flows in porous media and competition between bulk and surface diffusion. *IMA J. Appl. Math.*, 77(6):788–815, 2012.

[4] Grégoire Allaire and Harsha Hutridurga. On the homogenization of multi-component transport. *Discrete Contin. Dyn. Syst. Ser. B*, 20(8):2527–2551, 2015.

[5] Grégoire Allaire, Andro Mikelić, and Andrey Piatnitski. Homogenization approach to the dispersion theory for reactive transport through porous media. *SIAM J. Math. Anal.*, 42(1):125–144, 2010.

[6] Grégoire Allaire and Anne-Lise Raphael. Homogenization of a convection-diffusion model with reaction in a porous medium. *C. R. Math. Acad. Sci. Paris*, 344(8):523–528, 2007.

[7] Luigi Ambrosio and Gianluca Crippa. Existence, uniqueness, stability and differentiability properties of the flow associated to weakly differentiable vector fields. In *Transport equations and multi-D hyperbolic conservation*

laws, volume 5 of *Lecture Notes of the Unione Matematica Italiana*, pages 3–57. Springer, Berlin, 2008.

[8] Luigi Ambrosio, Hermano Frid, and Jean Silva. Multiscale Young measures in homogenization of continuous stationary processes in compact spaces and applications. *J. Funct. Anal.*, 256(6):1962–1997, 2009.

[9] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition, 2008.

[10] Kenneth J. Arrow, Leonid Hurwicz, and Hirofumi Uzawa. *Studies in linear and non-linear programming*. Standford University Press, 1958.

[11] Alekseĭ Arsenev. Global existence of a weak solution of Vlasov's system of equations. *USSR Computational Mathematics and Mathematical Physics*, 15:131–143, 1975.

[12] Baños, David R. and Duedahl, Sindre and Meyer-Brandis, Thilo and Proske, Frank. Construction of Malliavin differentiable strong solutions of SDEs under an integrability condition on the drift without the Yamada-Watanabe principle, 2015.

[13] Zhong-Zhi Bai. On semi-convergence of hermitian and skew-hermitian splitting methods for singular linear systems. *Computing*, 89(3-4):171–197, 2010.

[14] Jürgen Batt, W. Faltenbacher, and E Horst. Stationary spherically symmetric models in stellar dynamics. *Arch. Ration. Mech. Anal.*, 93(2):159–183, 1986.

[15] Jacob Bedrossian and Nader Masmoudi. Inviscid damping and the asymptotic stability of planar shear flows in the 2D Euler equations. *Publications mathématiques de l'IHÉS*, 122(1):195–300, 2015.

[16] Jonathan Ben-Artzi. Instability of nonmonotone magnetic equilibria of the relativistic Vlasov-Maxwell system. *Nonlinearity*, 24(12):3353–3389, dec 2011.

[17] Jonathan Ben-Artzi. Instability of nonsymmetric nonmonotone equilibria of the Vlasov-Maxwell system. *J. Math. Phys.*, 52(12):123703, 2011.

[18] Jonathan Ben-Artzi and Thomas Holding. Approximations of Strongly Continuous Families of Unbounded Self-Adjoint Operators. *Comm. Math. Phys.*, 345(2):615–630, 2016.

[19] Jonathan Ben-Artzi and Thomas Holding. Instabilities of the relativistic vlasov–maxwell system on unbounded domains. *SIAM journal on mathematical analysis*, 49(5):4024–4063, 2017.

[20] Alain Bensoussan, Jacques-Louis Lions, and George Papanicolaou. *Asymptotic analysis for periodic structures*. AMS Chelsea Publishing, Providence, RI, 2011. Corrected reprint of the 1978 original [MR0503330].

[21] Henri Berestycki, François Hamel, and Nikolai Nadirashvili. Elliptic eigenvalue problems with large drift and applications to nonlinear propagation phenomena. *Comm. Math. Phys.*, 253(2):451–480, 2005.

[22] Adrien Blanchet, Jean Dolbeault, and Michał Kowalczyk. Stochastic Stokes' drift, homogenized functional inequalities, and large time behavior of Brownian ratchets. *SIAM J. Math. Anal.*, 41(1):46–76, 2009.

[23] Harald Bohr. *Almost Periodic Functions*. Chelsea Publishing Company, New York, N.Y., 1947.

[24] François Bolley, Ivan Gentil, and Arnaud Guillin. Convergence to equilibrium in Wasserstein distance for Fokker-Planck equations. *J. Funct. Anal.*, 263(8):2430–2457, 2012.

[25] François Bolley, Arnaud Guillin, and Florent Malrieu. Trend to equilibrium and particle approximation for a weakly selfconsistent Vlasov-Fokker-Planck equation. *M2AN Math. Model. Numer. Anal.*, 44(5):867–884, 2010.

[26] Alain Bourgeat, Mladen Jurak, and Andrey L. Piatnitski. Averaging a transport equation with small diffusion and oscillating velocity. *Math. Methods Appl. Sci.*, 26(2):95–117, 2003.

[27] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.

[28] Stephen Boyd, Lin Xiao, Almir Mutapcic, and Jacob Mattingley. Notes on decomposition methods. Lecture Notes, Stanford University, 2007.

[29] W. Braun and K. Hepp. The Vlasov dynamics and its fluctuations in the $1/N$ limit of interacting classical particles. *Comm. Math. Phys.*, 56(2):101–113, 1977.

[30] Robert Brown. A brief account of microscopical observations made in the months of June, July and August, 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies. 1828.

[31] Krzysztof Burdzy and Wilfrid S. Kendall. Efficient Markovian couplings: examples and counterexamples. *Ann. Appl. Probab.*, 10(2):362–409, 2000.

[32] Zhao C., Topcu U., Li N., and Low S.H. Design and stability of load-side primary frequency control in power systems. *IEEE Trans. Automat. Contr.*, 59(5):1177–1189, 2014.

[33] José A. Carrillo, Young-Pil Choi, and Maxime Hauray. The derivation of swarming models: Mean-field limit and Wasserstein distances. In *Collective Dynamics from Bacteria to Crowds*, volume 553 of *CISM International Centre for Mechanical Sciences*, pages 1–46. Springer Vienna, 2014.

[34] José A. Carrillo, Young-Pil Choi, and Sergio Pérez. A review on attractive-repulsive hydrodynamics for consensus in collective behavior. *arXiv:1605.00232*, 2016.

[35] Juan Casado-Díaz and Inmaculada Gayte. The two-scale convergence method applied to generalized Besicovitch spaces. *R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci.*, 458(2028):2925–2946, 2002.

[36] Carlo Cercignani. *The Boltzmann equation and its applications*, volume 67 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1988.

[37] Mufa Chen. Optimal Markovian couplings and applications. *Acta Math. Sinica (N.S.)*, 10(3):260–275, 1994. A Chinese summary appears in Acta Math. Sinica **3**8 (1995), no. 4, 575.

[38] Ashish Cherukuri, Enrique Mallada, and Jorge Cortés. Asymptotic convergence of constrained primal–dual dynamics. *Systems & Control Letters*, 87:10–15, 2016.

[39] Antoine Choffrut and Vladimír Šverák. Local Structure of The Set of Steady-State Solutions to The 2d Incompressible Euler Equations. *Geom. Funct. Anal.*, 22(1):136–201, jan 2012.

[40] Peter Constantin, Alexander Kiselev, Leonid Ryzhik, and Andrej Zlatoš. Diffusion and mixing in fluid flow. *Ann. of Math. (2)*, 168(2):643–674, 2008.

[41] Christopher W. Curtis and Bernard Deconinck. On the convergence of Hill's method. *Math. Comput.*, 79(269):169–169, jan 2010.

[42] E. Brian Davies and M. Plum. Spectral pollution. *IMA J. Numer. Anal.*, 24(3):417–438, jul 2004.

[43] Bernard Deconinck and J Nathan Kutz. Computing spectra of linear operators using the Floquet-Fourier-Hill method. *J. Comput. Phys.*, 219(1):296–321, 2006.

[44] Giacomo Della Riccia. Equicontinuous semi-flows (one-parameter semigroups) on locally compact or complete metric spaces. *Mathematical systems theory*, 4(1):29–34, 1970.

[45] Helge Dietert. Stability and bifurcation for the Kuramoto model. *Journal de Mathématiques Pures et Appliquées*, 105(4):451–489, 2016.

[46] Helge Dietert, Jo Evans, and Thomas Holding. Convergence to equilibrium for the kinetic Fokker-Planck equation on the torus. *Prepr. arXiv1506.06173*, pages 1–12, Submitted.

[47] R. J. DiPerna and P.-L. Lions. Ordinary differential equations, transport theory and Sobolev spaces. *Invent. Math.*, 98(3):511–547, 1989.

[48] V. Dobrić and J.E. Yukich. Asymptotics for transportation cost in high dimensions. *Journal of Theoretical Probability*, 8(1):97–118, 1995.

[49] Roland L. Dobrushin. Vlasov equations. *Functional Analysis and Its Applications*, 13(2):115–123, 1979.

[50] Patrizia Donato and Andrey Piatnitski. Averaging of nonstationary parabolic operators with large lower order terms. In *Multi scale problems*

and asymptotic analysis, volume 24 of *GAKUTO Internat. Ser. Math. Sci. Appl.*, pages 153–165. Gakkōtosho, Tokyo, 2006.

[51] Florian Dörfler, John Simpson-Porco, and Francesco Bullo. Breaking the hierarchy: distributed control & economic optimality in microgrids. *IEEE Transactions on Control of Network Systems*, 2016. To appear.

[52] Javier Duoandikoetxea. *Fourier analysis*, volume 29 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2001. Translated and revised from the 1995 Spanish original by David Cruz-Uribe.

[53] William F. Eberlein. Abstract ergodic theorems and weak almost periodic functions. *Trans. Amer. Math. Soc.*, 67:217–240, 1949.

[54] David E. Edmunds and Hans Triebel. *Function Spaces, Entropy Numbers, Differential Operators*. Cambridge University Press, 1996. Cambridge Books Online.

[55] Albert Einstein. The theory of the brownian movement. *Annalen der Physik*, 17:549, 1905.

[56] Robert Ellis. *Lectures on topological dynamics*. W. A. Benjamin, Inc., New York, 1969.

[57] Lawrence C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2010.

[58] Albert Fannjiang and George Papanicolaou. Convection enhanced diffusion for periodic flows. *SIAM J. Appl. Math.*, 54(2):333–408, 1994.

[59] Ennio Fedrizzi and Franco Flandoli. Hölder Flow and Differentiability for SDEs with Nonregular Drift. *Stochastic Analysis and Applications*, 31(4):708–736, 2013.

[60] Ennio Fedrizzi, Franco Flandoli, Enrico Priola, and Julien Vovelle. Regularity of Stochastic Kinetic Equations. *arXiv preprint arXiv:1606.01088*, 2016.

[61] Diego Feijer and Fernando Paganini. Stability of primal-dual gradient dynamics and applications to network optimization. *Automatica J. IFAC*, 46(12):1974–1981, 2010.

[62] Wei-Jie Feng, Lei Wang, and Qing-Guo Wang. A family of multi-path congestion control algorithms with global stability and delay robustness. *Automatica*, 50(12):3112 – 3122, 2014.

[63] Franco Flandoli, M. Gubinelli, and Enrico Priola. Well-posedness of the transport equation by stochastic perturbation. *Inventiones mathematicae*, 180(1):1–53, 2010.

[64] Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.

[65] Nicolas Fournier and Maxime Hauray. Propagation of chaos for the Landau equation with moderately soft potentials, 2015.

[66] Nicolas Fournier, Maxime Hauray, and Stéphane Mischler. Propagation of chaos for the 2d viscous vortex model. *Journal of the European Mathematical Society*, 16(7):1423–1466, 2014.

[67] Jan Francu. Outline of Nguetseng's approach to non-periodic homogenization. *Mathematics for Applications*, 1(2), 2012.

[68] Robert M. Freund. Penalty and barrier methods for constrained optimization. *Lecture notes for nonlinear programming. MIT.*, 2004.

[69] Sébastien Gadat and Laurent Miclo. Spectral decompositions and L^2-operator norms of toy hypocoercive semi-groups. *Kinet. Relat. Models*, 6(2):317–372, 2013.

[70] Isabelle Gallagher, Laure Saint-Raymond, and Benjamin Texier. *From Newton to Boltzmann: hard spheres and short-range potentials*. Zurich Lectures in Advanced Mathematics. European Mathematical Society (EMS), Zürich, 2013.

[71] Ingenuin Gasser, Pierre-Emmanuel Jabin, and Benoit Perthame. Regularity and propagation of moments in some nonlinear Vlasov systems. *Proc. Roy. Soc. Edinburgh Sect. A*, 130(6):1259–1273, 2000.

[72] Robert T. Glassey and Jack W. Schaeffer. Control of velocities generated in a two-dimensional collisionless plasma with symmetry. *Transp. Theory Stat. Phys.*, 17(5-6):467–560, 1988.

[73] Robert T. Glassey and Jack W. Schaeffer. On the 'one and one-half dimensional' relativistic Vlasov-Maxwell system. *Math. Methods Appl. Sci.*, 13(2):169–179, aug 1990.

[74] Robert T. Glassey and Jack W. Schaeffer. The 'Two and One-Half Dimensional' Relativistic Vlasov Maxwell System. *Commun. Math. Phys.*, 185(2):257–284, 1997.

[75] Robert T. Glassey and Walter A. Strauss. Large velocities in the relativistic Vlasov-Maxwell equations. *J. Fac. Sci. Univ. Tokyo*, 36(3):615–627, 1989.

[76] François Golse, Clément Mouhot, and Valeria Ricci. Empirical measures and Vlasov Hierarchies. *Kinetic & Related Models*, 6(4), 2013.

[77] François Golse. On the dynamics of large particle systems in the mean field limit. In *Macroscopic and large scale phenomena: coarse graining, mean field limits and ergodicity*, volume 3 of *Lecture Notes in Applied Mathematics and Mechanics*, pages 1–144. Springer, [Cham], 2016.

[78] Peter M. Gruber and Jörg .M. Wills. *Handbook of Convex Geometry*. Convex Geometry. North-Holland, 1993.

[79] François Hamel and Nikolai Nadirashvili. Extinction versus persistence in strong oscillating flows. *Arch. Ration. Mech. Anal.*, 195(1):205–223, 2010.

[80] Anders C Hansen. On the approximation of spectra of linear operators on Hilbert spaces. *J. Funct. Anal.*, 254(8):2092–2126, apr 2008.

[81] Maxime Hauray and Pierre-Emmanuel Jabin. Particle approximation of Vlasov equations with singular forces: Propagation of chaos. *Ann. ENS*, 2013.

[82] Maxime Hauray and Samir Salem. Propagation of chaos for the Vlasov-Poisson-Fokker-Planck system in 1D. *arXiv preprint arXiv:1510.06260*, 2015.

[83] Peter H. Haynes and Jacques Vanneste. What controls the decay of passive scalars in smooth flows? *Phys. Fluids*, 17(9):097103, 16, 2005.

[84] Hermann von Helmholtz. Über Integrale der hydrodynamischen Gleichungen, welche den Wirbelbewegungen entsprechen. *J. Reine Angew. Math.*, 55:25–55, 1858.

[85] Frédéric Hérau. Hypocoercivity and exponential time decay for the linear inhomogeneous relaxation Boltzmann equation. *Asymptot. Anal.*, 46(3-4):349–359, 2006.

[86] Frédéric Hérau and Francis Nier. Isotropic hypoellipticity and trend to equilibrium for the Fokker-Planck equation with a high-degree potential. *Arch. Ration. Mech. Anal.*, 171(2):151–218, 2004.

[87] Thomas Holding. Propagation of chaos for Hölder continuous interaction kernels via Glivenko-Cantelli. *Prepr. arXiv:1608.02877*.

[88] Thomas Holding, Harsha Hutridurga, and Jeffrey Rauch. Convergence along mean flows. *SIAM journal on mathematical analysis*, 49(1):222–271, 2017.

[89] Thomas Holding, Harsha Hutridurga, and Jeffrey Rauch. Time scale analysis in passive transport with strong two dimensional Hamiltonian flows. In preparation.

[90] Thomas Holding and Ioannis Lestas. On the emergence of oscillations in distributed resource allocation. In *52nd IEEE Conference on Decision and Control*, December 2013.

[91] Thomas Holding and Ioannis Lestas. Stability and instability in primal-dual algorithms for multi-path routing. In *54th IEEE Conference on Decision and Control*, pages 1–6, December 2015.

[92] Thomas Holding and Ioannis Lestas. Stability and instability in gradient dynamics. *In preparation*, 2016.

[93] Thomas Holding and Ioannis Lestas. Stability and instability in gradient dynamics: Part II - The subgradient method. *In preparation*, 2016.

[94] Thomas Holding and Ioannis Lestas. On the emergence of oscillations in distributed resource allocation. *Automatica*, 85:22–33, 2017.

[95] Thomas Holding and Evelyne Miot. Uniqueness and continuous dependence for the Vlasov-Poisson system with spatial density in Orlicz spaces. In preparation.

[96] Darryl D. Holm, Jerrold E. Marsden, Tudor Ratiu, and Alan Weinstein. Nonlinear stability of fluid and plasma equilibria. *Phys. Rep.*, 123(1&2):1–116, 1985.

[97] Leonid Hurwicz. The design of mechanisms for resource allocation. *The American Economic Review*, 63(2):1–30, May 1973.

[98] Pierre-Emmanuel Jabin. A review of the mean field limits for Vlasov equations. *Kinet. Relat. Models*, 7(4):661–711, 2014.

[99] Pierre-Emmanuel Jabin and Athanasios E. Tzavaras. Kinetic decomposition for periodic homogenization problems. *SIAM J. Math. Anal.*, 41(1):360–390, 2009.

[100] Pierre-Emmanuel Jabin and Zhenfu Wang. Mean Field Limit and Propagation of Chaos for Vlasov Systems with Bounded Forces. *arXiv preprint arXiv:1511.03769*, 2015.

[101] Pierre-Emmanuel Jabin and Zhenfu Wang. Mean Field Limit and Propagation of Chaos for Vlasov Systems with Bounded Potentials. *In preparation*, 2016.

[102] Pierre-Emmanuel Jabin and Zhenfu Wang. Mean Field Limit for Stochastic Particle Systems. In N. Bellomo, P. Degond, and E. Tadmor, editors, *Active Particles Volume 1, Theory, Methods and Applications.* Birkhauser-Springer, To appear.

[103] J. H Jeans. On the theory of star-streaming and the structure of the universe. *Mon. Not. R. Astron. Soc.*, 76:70–84, 1915.

[104] Mathew A. Johnson and Kevin Zumbrun. Convergence of Hill's Method for Nonselfadjoint Operators. *SIAM J. Numer. Anal.*, 50(1):64–78, jan 2012.

[105] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker-Planck equation. *SIAM J. Math. Anal.*, 29(1):1–17, 1998.

[106] Mark Kac. Foundations of kinetic theory. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, number 3, pages 171–197, 1955.

[107] Koushik Kar, Saswati Sarkar, and Ros Tassiulas. Optimization based rate control for multipath sessions. Technical report, Univ. of Maryland, Inst. Systems Research, 2001. http://hdl.handle.net/1903/6225.

[108] Andreas Kasis, Eoin Devane, and Ioannis Lestas. On the stability and optimality of primary frequency regulation with load-side participation. In *54th IEEE Conference on Decision and Control*, 2015. arXiv:1602.02800.

[109] Tosio Kato. *Perturbation Theory for Linear Operators*. Springer-Verlag, 1995.

[110] Frank Kelly, Aman Maulloo, and David Tan. Rate control in communication networks: shadow prices, proportional fairness and stability. *Journal of the Operational Research Society*, 49(3):237–252, March 1998.

[111] Wilfrid S. Kendall. Coupling, local times, immersions. *Bernoulli*, 21(2):1014–1046, 2015.

[112] Hassan K. Khalil. *Nonlinear Systems*. Prentice Hall, 2002.

[113] Yuri L. Klimontovich. *Statistical theory of non-equilibrium processes in a plasma*. Pergamon Press, Oxford, New York, Pergamon Press [1967], [1st english ed.] edition, 1967.

[114] Nicholas A. Krall and Alvin W. Trivelpiece. *Principles of plasma physics*. Number v. 0-911351 in International series in pure and applied physics. McGraw-Hill, 1973.

[115] K Kumar, M N N Namboodiri, and S Serra-Capizzano. Perturbation of operators and approximation of spectrum. *Proc. - Math. Sci.*, 124(2):205–224, 2014.

[116] H. Kunita. *Stochastic Flows and Stochastic Differential Equations*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 1997.

[117] Kazumasa Kuwada. Characterization of maximal Markovian couplings for diffusion processes. *Electron. J. Probab.*, 14:no. 25, 633–662, 2009.

[118] Ol'ga A. Ladyženskaja, Vsevolod A. Solonnikov, and Nina N. Ural′ceva. *Linear and quasilinear equations of parabolic type*. Translated from the Russian by S. Smith. Translations of Mathematical Monographs, Vol. 23. American Mathematical Society, Providence, R.I., 1968.

[119] Lev D. Landau. On the vibrations of the electronic plasma. *J. Phys. USSR*, 10(25):574, 1946.

[120] Oscar E. Lanford, III. Time evolution of large classical systems. In *Dynamical systems, theory and applications (Rencontres, Battelle Res. Inst., Seattle, Wash., 1974)*, pages 1–111. Lecture Notes in Phys., Vol. 38. Springer, Berlin, 1975.

[121] Paul Langevin. Sur la théorie du mouvement brownien. *Comptes Rendus de l'Académie des Sciences*, 146(530-533):530, 1908.

[122] Ronald Larsen. *Banach algebras*. Marcel Dekker, Inc., New York, 1973. An introduction, Pure and Applied Mathematics, No. 24.

[123] Dustin Lazarovici and Peter Pickl. A mean-field limit for the Vlasov-Poisson system. *arXiv preprint arXiv:1502.04608*, 2015.

[124] Gyo Lee and Jui Choi. A survey of multipath routing for traffic engineering.

[125] Mohammed Lemou, Florian Méhats, and Pierre Raphaël. Orbital stability of spherical galactic models. *Invent. Math.*, 187(1):145–194, apr 2011.

[126] Ioannis Lestas and Glenn Vinnicombe. Combined control of routing and flow: a multipath routing approach. In *43rd IEEE Conference on Decision and Control*, December 2004.

[127] Michael Levitin and Eugene Shargorodsky. Spectral pollution and second-order relative spectra for self-adjoint operators. *IMA J. Numer. Anal.*, 24(3):393–416, jul 2004.

[128] Mathieu Lewin and Eric Sere. Spectral pollution and how to avoid it. *Proc. London Math. Soc.*, 100(3):864–900, dec 2009.

[129] Xiaojun Lin and Ness B Shroff. Utility maximization for communication networks with multipath routing. *Automatic Control, IEEE Transactions on*, 51(5):766–781, 2006.

[130] Zhiwu Lin and Walter A. Strauss. Linear stability and instability of relativistic Vlasov-Maxwell systems. *Commun. Pure Appl. Math.*, 60(5):724–787, may 2007.

[131] Zhiwu Lin and Walter A. Strauss. Nonlinear stability and instability of relativistic Vlasov-Maxwell systems. *Commun. Pure Appl. Math.*, 60(6):789–837, jun 2007.

[132] Zhiwu Lin and Walter A. Strauss. A sharp stability criterion for the Vlasov-Maxwell system. *Invent. Math.*, 173(3):497–546, apr 2008.

[133] Pierre-Louis Lions and Benoît Perthame. Propagation of moments and regularity for the 3-dimensional Vlasov-Poisson system. *Invent. Math.*, 105(1):415–430, dec 1991.

[134] Gregoire Loeper. Uniqueness of the solution to the Vlasov-Poisson system with bounded density. *Journal de Mathématiques Pures et Appliquées*, 86 (9), no.1:68–79, 2006.

[135] John Lygeros, Karl Henrik Johansson, Slobodan N. Simić, Jun Zhang, and S. Shankar Sastry. Dynamical properties of hybrid automata. *IEEE Trans. Automat. Control*, 48(1):2–17, 2003.

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

[137] Andrew J. Majda and Peter R. Kramer. Simplified models for turbulent diffusion: theory, numerical modelling, and physical phenomena. *Phys. Rep.*, 314(4-5):237–574, 1999.

[138] C. Marchioro and M. Pulvirenti. *Mathematical Theory of Incompressible Non-viscous Fluids*. Applied mathematical sciences. Springer-Verlag, 1994.

[139] Carlo Marchioro and Mario Pulvirenti. A note on the nonlinear stability of a spatially symmetric vlasov-possion flow. *Math. Methods Appl. Sci.*, 8(1):284–288, jun 1986.

[140] Eduard Marušić-Paloka and Andrey L. Piatnitski. Homogenization of a nonlinear convection-diffusion equation with rapidly oscillating coefficients and strong convection. *J. London Math. Soc. (2)*, 72(2):391–409, 2005.

[141] James Clerk Maxwell. V. Illustrations of the dynamical theory of gases. -Part I. On the motions and collisions of perfectly elastic spheres. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 19(124):19–32, 1860.

[142] D. W. McLaughlin, G. C. Papanicolaou, and O. R. Pironneau. Convection of microstructure and related problems. *SIAM J. Appl. Math.*, 45(5):780–797, 1985.

[143] Evelyne Miot. A uniqueness criterion for unbounded solutions to the Vlasov-Poisson system. *Comm. Math. Phys.*, 346 (2):469–492, 2016.

[144] Stéphane Mischler and Clément Mouhot. Kac's program in kinetic theory. *Inventiones mathematicae*, 193(1):1–147, 2013.

[145] Stéphane Mischler and Clément Mouhot. Exponential stability of slowly decaying solutions to the kinetic Fokker-Planck equation. *ArXiv e-prints*, December 2014.

[146] Salah-Eldin A. Mohammed, Torstein K. Nilssen, and Frank N. Proske. Sobolev differentiable stochastic flows for SDEs with singular coefficients:

Applications to the transport equation. *Ann. Probab.*, 43(3):1535–1576, 05 2015.

[147] P.D. Moral. *Mean Field Simulation for Monte Carlo Integration*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press, 2013.

[148] Peter Mörters and Yuval Peres. *Brownian Motion*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2010.

[149] Clément Mouhot and Lukas Neumann. Quantitative perturbative study of convergence to equilibrium for collisional kinetic models in the torus. *Nonlinearity*, 19(4):969–998, 2006.

[150] Clément Mouhot and Cédric Villani. Landau damping. *J. Math. Phys.*, 51(1):015204, 2010.

[151] Clément Mouhot and Cédric Villani. On Landau Damping. *Acta Math.*, 207:29–201, 2011.

[152] Ayman Moussa and Franck Sueur. On a Vlasov-Euler system for 2D sprays with gyroscopic effects. *Asymptot. Anal.*, 81(1):53–91, 2013.

[153] Angelia Nedić and Asuman Ozdaglar. Subgradient methods for saddle-point problems. *Journal of Optimization Theory and Applications*, 142(1):205–228, 2009.

[154] Gabriel Nguetseng. A general convergence result for a functional related to the theory of homogenization. *SIAM J. Math. Anal.*, 20(3):608–623, 1989.

[155] Gabriel Nguetseng. Homogenization structures and applications. I. *Z. Anal. Anwendungen*, 22(1):73–107, 2003.

[156] Gabriel Nguetseng. Homogenization structures and applications. II. *Z. Anal. Anwendungen*, 23(3):483–508, 2004.

[157] Gabriel Nguetseng, Mamadou Sango, and Jean Louis Woukeng. Reiterated ergodic algebras and applications. *Comm. Math. Phys.*, 300(3):835–876, 2010.

[158] Gabriel Nguetseng and Nils Svanstedt. Σ-convergence. *Banach J. Math. Anal.*, 5(1):101–135, 2011.

[159] Toan T. Nguyen, Truyen V. Nguyen, and Walter A. Strauss. Global magnetic confinement for the 1.5D Vlasov-Maxwell system. *Kinet. Relat. Model.*, 8(1):153–168, oct 2015.

[160] Toan T. Nguyen and Walter A. Strauss. Linear Stability Analysis of a Hot Plasma in a Solid Torus. *Arch. Ration. Mech. Anal.*, 211(2):619–672, oct 2014.

[161] Richard Nickl and Benedikt M. Pötscher. Bracketing metric entropy rates and empirical central limit theorems for function classes of Besov-and Sobolev-type. *Journal of Theoretical Probability*, 20(2):177–199, 2007.

[162] Fernando Paganini and Enrique Mallada. A unified approach to congestion control and node-based multipath routing. *IEEE/ACM Trans. Netw.*, 17(5):1413–1426, October 2009.

[163] Christophe Pallard. Space moments of the Vlasov-Poisson system: propagation and regularity. *SIAM J. Math. Anal.*, 46(3):1754–1770, 2014.

[164] Daniel Pérez Palomar and Mung Chiang. A tutorial on decomposition methods for network utility maximization. *IEEE Journal on Selected Areas in Communications*, 24(8):1439–1451, 2006.

[165] D. Papadaskalopoulos and G. Strbac. Decentralized participation of flexible demand in electricity markets-part i: Market mechanism. *IEEE Transactions On Power Systems*, 28:3658–3666, 2013.

[166] George C. Papanicolaou. Diffusion in random media. In *Surveys in applied mathematics, Vol. 1*, volume 1 of *Surveys Appl. Math.*, pages 205–253. Plenum, New York, 1995.

[167] Oliver Penrose. Electrostatic Instabilities of a Uniform Non-Maxwellian Plasma. *Phys. Fluids*, 3(2):258–265, 1960.

[168] K. Pfaffelmoser. Global classical solutions of the Vlasov-Poisson system in three dimensions for general initial data. *J. Differ. Equ.*, 95(2):281–303, feb 1992.

[169] Maria Radu. *Homogenization techniques*. PhD thesis, Diplomarbeit, University of Heidelberg, Faculty of Mathematics, 1992.

[170] M.M. Rao and Z.D. Ren. *Applications Of Orlicz Spaces*. Monographs and textbooks in pure and applied mathematics. Taylor & Francis, 2002.

[171] Michael Reed and Barry Simon. *Methods of modern mathematical physics volume 4: Analysis of operators*. 1978.

[172] Michael Reed and Barry Simon. *Methods of Modern Mathematical Physics Volume 1: Functional Analysis*. Academic Press Inc, 1981.

[173] Dean Richert and Jorge Cortés. Robust distributed linear programming. *IEEE Trans. Automat. Control*, 60(10):2567–2582, 2015.

[174] Walter Rudin. *Real and complex analysis*. McGraw-Hill Book Co., New York, third edition, 1987.

[175] Enrique Sánchez-Palencia. *Non-homogeneous media and vibration theory*. Springer, 1980.

[176] Mamadou Sango, Nils Svanstedt, and Jean Louis Woukeng. Generalized Besicovitch spaces and applications to deterministic homogenization. *Nonlinear Anal.*, 74(2):351–379, 2011.

[177] Jack Schaeffer. Global existence of smooth solutions to the Vlasov-Poisson system in three dimensions. *Comm. Partial Differential Equations*, 16(8-9):1313–1335, 1991.

[178] Jack W. Schaeffer. Global existence of smooth solutions to the vlasov poisson system in three dimensions. *Commun. Partial Differ. Equations*, 16(8-9):1313–1335, jan 1991.

[179] Jack W. Schaeffer. A Class of Counterexamples to Jeans' Theorem for the Vlasov-Einstein System. *Commun. Math. Phys.*, 204(2):313–327, jul 1999.

[180] Jeff S. Shamma and Gurdal Arslan. Dynamic fictitious play, dynamic gradient play, and distributed convergence to nash equilibria. *IEEE Transactions on Automatic Control*, 50(3):312–327, March 2005.

[181] Barry Simon. Fifty years of eigenvalue perturbation theory. *Bull. Am. Math. Soc.*, 24(2):303–320, apr 1991.

[182] Herbert Spohn. *Large scale dynamics of interacting particles*. Springer Science & Business Media, 2012.

[183] Rayadurgam Srikant. *The mathematics of Internet congestion control*. Birkhauser, 2004.

[184] Michael Strauss. A new approach to spectral approximation. *J. Funct. Anal.*, 267:3084–3103, mar 2014.

[185] Alain-Sol Sznitman. Topics in propagation of chaos. In Paul-Louis Hennequin, editor, *Ecole d'Été de Probabilités de Saint-Flour XIX - 1989*, volume 1464 of *Lecture Notes in Mathematics*, pages 165–251. Springer Berlin Heidelberg, 1991.

[186] Michel Talagrand. New concentration inequalities in product spaces. *Inventiones mathematicae*, 126(3):505–563, 1996.

[187] Terence Tao. The spectral theorem and its converses for unbounded symmetric operators (from the blog "What's New?" http://terrytao.wordpress.com/2011/12/20/the-spectral-theorem-and-its-converses-for-unbounded-symmetric-operators/), 2011.

[188] Luc Tartar. *The general theory of homogenization*, volume 7 of *Lecture Notes of the Unione Matematica Italiana*. Springer-Verlag, Berlin; UMI, Bologna, 2009. A personalized introduction.

[189] Hans Triebel. *Theory of function spaces III*. Monographs in Mathematics. Springer, Basel, 2006.

[190] Cameron Tropea, Alexander Yarin, and John F. Foss. *Springer Handbook of Experimental Fluid Mechanics*. Number v. 1 in Springer Handbook of Experimental Fluid Mechanics. Springer, 2007.

[191] Bruce Turkington. Statistical equilibrium measures and coherent states in two-dimensional turbulence. *Communications on Pure and Applied Mathematics*, 52(7):781–809, jul 1999.

[192] Seiji Ukai and Takayoshi Okabe. On classical solutions in the large in time of two-dimensional Vlasov's equation. *Osaka J. Math.*, 15(2):245–261, 1978.

[193] Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer New York, 1996.

[194] Hal R. Varian. *Microeconomic Analysis*. W.W.Norton & Company, 1996.

[195] Mark Veraar and Tuomas Hytonen. On besov regularity of brownian motions in infinite dimensions. *Probability and Mathematical Statistics Vol. 28, Fasc. 1 (2008), pp. 143–162*, 2008.

[196] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed sensing*, pages 210–268. Cambridge Univ. Press, Cambridge, 2012.

[197] Cédric Villani. Hypocoercivity. *Mem. Amer. Math. Soc.*, 202(950):iv+141, 2009.

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

[199] Thomas Voice. Stability of multi-path dual congestion control algorithms. *IEEE/ACM Trans. Netw.*, 15(6):1231–1239, December 2007.

[200] Marian von Smoluchowski. Zur kinetischen theorie der brownschen molekularbewegung und der suspensionen. *Annalen der Physik*, 326(14):756–780, 1906.

[201] Feng-Yu Wang and Xicheng Zhang. Degenerate SDE with H\" older-Dini Drift and Non-Lipschitz Noise Coefficient. *arXiv preprint arXiv:1504.04450*, 2015.

[202] Zheng Wang and Jon Crowcroft. Analysis of shortest-path routing algorithms in a dynamic network environment. *ACM SIGCOMM Computer Communication Review*, 22(2):63–71, 1992.

[203] Stephen Wollman. An existence and uniqueness theorem for the Vlasov-Maxwell system. *Commun. Pure Appl. Math.*, 37(4):457–462, jul 1984.

[204] D. Zhang and A. Nagurney. On the stability of projected dynamical systems. *J. Optim. Theory Appl.*, 85(1):97–124, 1995.

[205] Vasiliĭ V. Zhikov, Sergeĭ M. Kozlov, and Ol'ga A. Oleĭnik. *Homogenization of differential operators and integral functionals*. Springer-Verlag, Berlin, 1994. Translated from the Russian by G. A. Yosifian [G. A. Iosif′yan].

[206] Vasiliĭ V. Zhikov and E. V. Krivenko. Averaging of singularly perturbed elliptic operators. *Mat. Zametki*, 33(4):571–582, 1983.

[207] Robert Zwanzig. *Nonequilibrium Statistical Mechanics*. Oxford University Press, USA, 2001.