Energies, gradient flows, and large deviations: a modelling point of view

Mark A. Peletier

November 3, 2011
Version 0.1

1 Introduction

1.1 Modelling

Modelling is the art of taking a real-world situation and constructing some mathematics to describe it. It's an art rather than a science, because it involves choices that can not be rationally justified. Personal preferences are important, and 'taste' plays a major role.

Having said that, some choices are more rational than others. And we often also need to explain our choices to ourselves and to our colleagues. These notes are about such issues in an area that has recently increased in importance, that of *gradient flows*, and especially *Wasserstein* gradient flows. The notes are organized around a central question, a question that one might ask for any given model, but that we will ask specifically for gradient flows:

What are the modelling assumptions that underlie such-and-such model?

In these notes we will provide the background and theory that we will need to give a good answer to this question.

Gradient flows are evolutionary systems driven by an energy, in the sense that the energy decreases along solutions, *as fast as possible*. The property 'as fast as possible' requires specification, which is typically done by choosing a *dissipation mechanism*. From a modelling point of view the two choices of (a) the driving energy and (b) the dissipation mechanism completely determine the system—I will give more detail below. That means that in order to justify any given model, we need to explain three things:

1. Why do we choose this energy?
2. Why this dissipation?
3. Why this *combination* of energy and dissipation?

Just to illustrate how these are non-trivial questions, here is an example. The diffusion equation,

$$\partial_t u = \Delta u \qquad \text{in } \mathbb{R}^d, \tag{1}$$

is a gradient flow in many different ways:

1. with energy $\int |\nabla u|^2$ and the L^2-norm as dissipation mechanism;

2. with energy $\int u^2$ and the (homogeneous) H^{-1}-norm as dissipation mechanism;

3. with energy $\|u\|_{H^s}^2$ (the homogeneous H^s-norm or seminorm) and the H^{s-1} (semi-)norm as dissipation;

4. with energy $\int u \log u$ and the Wasserstein metric as dissipation;

5. with energy $\int [u \log u + (1-u)\log(1-u)]$ and a modified Wasserstein metric as dissipation;

and there are many more. This shows that the equation itself determines neither the energy nor the dissipation; those are really choices that result from the modelling context. For some of the pairs above we will see this in these notes (numbers 1, 4, and 5, to be precise). Note that each of the 'energies' above is a Lyapunov function of equation (1)—they decrease along solutions—and therefore each might reasonably serve with any of the five dissipation mechanisms mentioned above (or any other). This shows how not only the choices of energy and dissipation, but also of the *combination* requires explaining.

We limit ourselves to systems that can be modelled by gradient flows because these systems have recently become much more important. A major step was taken in 1997 by the introduction by Jordan, Kinderlehrer, and Otto of *Wasserstein gradient flows* [JKO97, JKO98], which we discuss in detail in these notes. Since 1997 it has become clear that a large number of systems can be written as Wasserstein gradient flows, and an even larger number are gradient flows that incorporate a Wasserstein-type dissipation term together with other dissipations. The box on this page shows some of these.

1.2 Gradient flows

In the next section we will describe some examples that we address in these notes. Before that we need to explain why gradient flows are determined by an energy and a dissipation, and how these two components together determine the evolution. For simplicity we now take the case of $\mathbb{R}^n$; generalizations are discussed later.

The gradient flow in the Euclidean space $\mathcal{Z} = \mathbb{R}^n$ of a functional $\mathcal{E}\colon \mathbb{R}^n \to \mathbb{R}$ is the evolution in $\mathbb{R}^n$ given by

$$\dot{z}^i(t) = -\partial_i \mathcal{E}(z(t)). \tag{2}$$

Many diffusive systems are Wasserstein gradient flows
A surprisingly large number of well-known diffusive partial differential equations have the structure of a gradient flow with respect to the Wasserstein metric (or a related metric). These are just a few examples:

convection and nonlinear diffusion [JKO98]	$\partial_t \rho = \operatorname{div} \rho \nabla \big[U'(\rho) + V\big]$
nonlocal pbs ([AGS05, CMV03, CMV06] *et al.*)	$\partial_t \rho = \operatorname{div} \rho \nabla \big[U'(\rho) + V + W * \rho\big]$
thin-film equation [Ott98, GO01]	$\partial_t \rho = -\partial_x \big[\rho \partial_x^3 \rho\big]$
DLSS and related [DLSS91, JM09, MMS09]	$\partial_t \rho = -\operatorname{div} \rho \nabla \big(\rho^{\alpha-1} \Delta \rho^\alpha\big), \quad 1/2 \le \alpha \le 1$
a reactive moving-boundary problem [PP10]	$\partial_t \rho = \Delta \rho$ in $\Omega(t)$, with $\partial_n \rho = -\rho v_n$ and $v_n = f(\rho)$ on $\partial\Omega(t)$
the Cahn-Hilliard equation	$\partial_t \rho = -\operatorname{div} D(\rho) \nabla \big[\Delta \rho + f(\rho)\big]$
two-phase flow [Ott99]	$\partial_t \rho = -\operatorname{div} \rho u,\ \operatorname{div} u = 0,\ u = f(\rho)\big[\nabla p + \rho \boldsymbol{e}_z\big]$

Although this equation appears to involve only an energy $\mathcal{E}$, a dissipation mechanism is hidden, and this can be seen in two equivalent ways.

1. The first argument rests on geometry. Equation (2) can be written in a geometrically more correct way as
$$\dot{z}^i(t) = -\sum_j g^{ij}\partial_j \mathcal{E}(z(t)), \tag{3}$$
with $g^{ij} = \delta^{ij}$. This form is 'more correct' in the sense that the first form is not invariant under changes of coordinates, as the difference between the upper and lower indices indicates; by the rules of coordinate changes in Riemannian manifolds, the second form *is* coordinate-invariant. Because of the difference between vectors and covectors, we use different names and notation for the *(Fréchet) derivative* or *differential* ∂_i (also denoted $\mathcal{E}'$ or $\operatorname{diff}\mathcal{E}$) and the *gradient*, which is the object $g^{ij}\partial_j\mathcal{E}$ (also written as $\operatorname{grad}\mathcal{E}$).

   The form (3) shows how the energy $\mathcal{E}$ and the metric g together determine the evolution of z. In this context the metric g is called the dissipation metric, since it characterizes the decrease ('dissipation') of energy along a solution:
$$\partial_t \mathcal{E}(z(t)) = \sum_i \partial_i \mathcal{E}(z(t))\dot{z}^i(t) = -\sum_{i,j} g^{ij}\partial_i\mathcal{E}(z(t))\partial_j\mathcal{E}(z(t)) = -\sum_{i,j} g_{ij}\dot{z}^i(t)\dot{z}^j(t),$$
where we write g_{ij} for the inverse of g^{ij}. Note how the last term above is the squared length of $\dot{z}$ in the g-metric.

2. The necessity of the distinction between derivative and gradient can also be explained in terms of function spaces. If $\mathcal{Z}$ is a Hilbert space, then the derivative $\mathcal{E}'(z)$ of a differentiable functional $\mathcal{E} : \mathcal{Z} \to \mathbb{R}$ at $z \in \mathcal{Z}$ is a bounded linear mapping from $\mathcal{Z}$ to $\mathbb{R}$, and therefore an element of the topological dual $\mathcal{Z}'$. Suppose we seek a curve $z \in C^1([0,T];\mathcal{Z})$ that solves in some sense the equation $\dot{z} = -\mathcal{E}'(z)$; we then run into the problem that for given t, $\dot{z}(t) \in \mathcal{Z}$, while $\mathcal{E}'(z(t)) \in \mathcal{Z}'$, and therefore these two objects can not be equated to each other. A natural way to solve this definition problem is use Riesz' Lemma and the inner product in $\mathcal{Z}$ to convert $\mathcal{E}'(z) \in \mathcal{Z}'$ to an object $\operatorname{grad}\mathcal{E}(z) \in \mathcal{Z}$ satisfying
$$\forall y \in \mathcal{Z}: \qquad (\operatorname{grad}\mathcal{E}(z), y)_{\mathcal{Z}} = {}_{\mathcal{Z}'}\langle \mathcal{E}'(z), y\rangle_{\mathcal{Z}}. \tag{4}$$
Here we use the notation $(\cdot,\cdot)_{\mathcal{Z}}$ for the inner product in $\mathcal{Z}$, and ${}_{\mathcal{Z}'}\langle\cdot,\cdot\rangle_{\mathcal{Z}}$ for the duality product between $\mathcal{Z}'$ and $\mathcal{Z}$. The equation
$$\dot{z} = -\operatorname{grad}\mathcal{E}(z) \tag{5}$$
now makes sense, since both sides are elements of $\mathcal{Z}$. In terms of coordinates, $(x,y)_{\mathcal{Z}} = g_{ij}x^iy^j$, which implies that equations (5) and (3) are the same.

   Note that while the derivative $\mathcal{E}'$ is independent of the choice of inner product, the gradient $\operatorname{grad}\mathcal{E}$ does depend on the inner product.

1.3 Guiding examples and questions

We now describe some examples that will serve as beacons: these will be the key examples that we will want to explain.

The diffusion equation (1) is both simple and has many gradient-flow structures, and therefore it is an excellent first example. We will want to understand the modelling arguments that lead to three of the structures mentioned above. Let us formulate some of the questions that we want to address.

- Equation (1) is the L^2-gradient flow of the H^1-seminorm $\mathcal{E}(u) := \frac{1}{2}\int |\nabla u|^2$, in the sense described above: for each t, the time derivative $\partial_t u$ satisfies the weak characterization

$$\forall v \in L^2(\mathbb{R}^d), \qquad (\partial_t u, v)_{L^2(\mathbb{R}^d)} = -\langle \mathcal{E}'(u), v\rangle. \tag{6}$$

  Indeed, if we write the terms explicitly as

$$\forall v \in L^2(\mathbb{R}^d), \qquad \int_{\mathbb{R}^d} v\partial_t u = -\int_{\mathbb{R}^d} \nabla u \nabla v, \tag{7}$$

  then we recognize here a weak formulation of (1).

  Question 1: What is the modelling background of this gradient-flow structure?

  Note how (7) contains a mathematical inconsistency: the right-hand side is not well defined for all $v \in L^2(\mathbb{R}^d)$. While the expression (7) can simply be corrected by passing to a dense subset of L^2 for which the right-hand side is meaningful, the better solution to this problem involves a discussion of Fréchet subdifferentials and semigroups; see e.g. [Bre73, RS06]. This type of issue will arise many times, but since we want to focus on modelling issues we will typically disregard it.

- Equation (1) is also the Wasserstein gradient flow of $\mathsf{Ent}(u) = \int u \log u$, which we call the *entropy*. We will define and discuss the Wasserstein metric in much more detail below; since there is no corresponding inner-product structure, the concept of a gradient flow as described above will be generalized below. For the moment it is sufficient to define a *Wasserstein gradient flow* of any functional $\mathcal{E}$ as the evolution equation

$$\partial_t u = \operatorname{div} u \nabla \frac{\delta \mathcal{E}}{\delta u}, \tag{8}$$

  where we use the physicist's notation $\delta\mathcal{E}/\delta u$ for the *variational derivative* (or L^2-gradient) of $\mathcal{E}$, i.e. the function w satisfying

$$\forall v \in L^2, \qquad \langle \mathcal{E}'(u), v\rangle = (w, v)_{L^2}.$$

  Question 2: What is the modelling background of this structure?

  An important sub-question here is

  Question 3: What do the entropy Ent and the Wasserstein distance mean, in modelling terms?

Let us check that equation (1) is indeed the Wasserstein gradient flow, in this sense, of the entropy $\mathsf{Ent}(u) = \int u \log u$. The variational derivative of Ent is $\delta \mathsf{Ent}/\delta u = \log u + 1$, so that equation (8) becomes

$$\partial_t u = \operatorname{div} u \nabla(\log u + 1) = \Delta u.$$

- In a similar way we might define what I call the *Kawasaki-Wasserstein gradient flow*[1] as the equation

$$\partial_t u = \operatorname{div} u(1-u)\nabla \frac{\delta \mathcal{E}}{\delta u}. \tag{9}$$

Then equation (1) is the gradient flow of the mixing entropy $\mathsf{Ent}_{\mathrm{mix}}(u) = \int [u \log u + (1-u)\log(1-u)]$, since $\delta \mathsf{Ent}_{\mathrm{mix}}/\delta u = \log u - \log(1-u)$ and

$$\partial_t u = \operatorname{div} u(1-u)\nabla[\log u - \log(1-u)] = \Delta u.$$

Question 4: What is the modelling background of this structure? How does this modified Wasserstein gradient flow arise?

These three gradient-flow systems all have the diffusion equation (1) as the corresponding equation. The fact that equation (1) can be written as a Wasserstein gradient flow was actually first published in [JKO97, JKO98], as a consequence of a stronger statement:

- The *advection-diffusion equation* (or *Fokker-Planck equation*)

$$\partial_t u = \Delta u + \operatorname{div} u \nabla \Phi, \tag{10}$$

where $\Phi : \mathbb{R}^d \to \mathbb{R}$ is a spatially dependent potential, is the Wasserstein gradient flow of the functional

$$\mathcal{F}(u) := \int_{\mathbb{R}^d} \big[u \log u + u\Phi\big]. \tag{11}$$

This can be checked again by (8). We will come back to this example below in much more detail.

Question 5: What is $\mathcal{F}$ in (11)? Why does it drive the evolution in (10)?

These are the questions that we will be concerned with in these lecture notes. There are many more examples of gradient-flow systems, such as

- The Allen-Cahn and Cahn-Hilliard equations, which are gradient flows of the energy $u \mapsto \int \frac{1}{2}|\nabla u|^2 + W(u)$ with the L^2 and H^{-1} metrics, respectively:

$$(AC) \qquad \partial_t u = \Delta u - W'(u),$$
$$(CH) \qquad \partial_t u = -\Delta(\Delta u - W'(u)).$$

[1] The name arises from the fact that this structure arises from microscopic models with Kawasaki exchange dynamics; see e.g. [KOV89].

- Nonlocal Cahn-Hilliard-type equations such as the H^{-1}-gradient flow of a nonlocal energy [Cho01]
$$u \mapsto \int \Big[\frac{1}{2}|\nabla u|^2 + W(u)\Big]\,dx + \|u - \fint u\|_{H^{-1}}^2,$$
where $\fint u$ is the average of u. The resulting equation is
$$\partial_t u = -\Delta(\Delta u - W'(u)) - u + \fint u.$$

- Phase-separation problems based on convolution energies, such as the Kawasaki-Wasserstein gradient flow of the energy [GL98]
$$u \mapsto \int f(u)\,dx - \frac{1}{2}\iint J(x-x')u(x)u(x')\,dxdx'$$
with $J \geq 0$, $\int J = 1$, and $f(u) = u\log u + (1-u)\log(1-u)$ (so that the first term in the energy is $\mathsf{Ent}_{\mathrm{mix}}$). The correspdonding equation is
$$\partial_t u = \Delta u + \operatorname{div} u(1-u)\nabla J * u.$$

- Various free-boundary problems, such as the Stefan problem [RS06], fourth-order thin-film equations [GO01, Ott98], and a crystal-dissolution problem [PP10].

- Various equations of viscous flow, such as the Stokes equations and flow through porous media.

- If we allow for generalizations to non-quadratic ψ and ψ^* (see Section 5) then the large theory of rate-independent systems is also (formally) of this form [**?**]. This applies to many problems in elasto-plasticity, hysteresis in phase transitions, delamination, friction, and other mechanical situations

In these lecture notes we do not have the time to discuss all the possible types of gradient flows and their modelling considerations. Instead we focus on the most conspicuous open problems, which are those surrounding the Wasserstein distance and Wasserstein gradient flows. The numbered questions above serve as a good indication of where we want to go.

1.4 Overview

The questions formulated above are also the questions that have been driving my own interest in these matters over the last few years. The surprising thing here (at least to me) is that the answer to most of these questions can be found only by broadening our view, and taking into account that most of these equations arise as the continuum limit of particle systems. As it turns out, *the gradient-flow structures themselves* ***also*** *arise from these particle systems.* And in order to understand this, it turns out that one has to consider these particle systems as *stochastic* particle systems, and one has to study the occurrence of *rare events*, or in probabilistic parlance, *large deviations.*

This last statement is so important that I'm going to state it again, as it is the moral of this course:

The origin of the gradient-flow structures can only be understood by including the large-deviation behaviour of underlying microscopic particle systems.

In order to understand how to *model* such Wasserstein-type gradient flows, therefore, one has to understand the underlying stochastic particle systems.

The rest of these notes are structured as follows. In Section 2 we first discuss entropy, we introduce large deviations, and then we use the concept to give an answer to one of the central questions: Why does the entropy $\mathsf{Ent}(u) = \int u \log u$ appear as the driving energy in the Wasserstein-entropy gradient-flow formulation of (1)? (see 3).

In Section 4 we discuss the first example in detail: the diffusion equation as the Wasserstein gradient-flow of entropy.

1.5 Prerequisites, notation, and assumptions

In these notes I assume an understanding of the basic concepts of abstract measure theory, such as measurable spaces, absolute continuity, the total variation norm $\|\cdot\|_{TV}$ and the Radon-Nikodym derivative.

We will work with a measurable space (Ω, Σ) consisting of a topological set Ω with a σ-algebra Σ that contains the Borel σ-algebra $\mathcal{B}(\Omega)$. We always assume that Ω is complete, metrizable, and separable. $\mathcal{M}(\Omega)$ is the set of all (finite or infinite) signed measures on $(\Omega, \mathcal{B}(\Omega))$, and $\mathcal{P}(\Omega)$ is the set of all non-negative measures $\mu \in \mathcal{M}(\Omega)$ with $\mu(\Omega) = 1$. For $\rho \in \mathcal{P}(\Omega)$, $L^2(\rho)$ is the set of measurable functions on Ω with finite $L^2(\rho)$-norm:

$$\|f\|^2_{L^2(\rho)} := \int_\Omega f^2 \, d\rho.$$

A sequence of measures $\mu_n \in \mathcal{M}(\Omega)$ is said to converge *narrowly* to $\mu \in \mathcal{M}(\Omega)$ if

$$\int_\Omega f \, d\mu_n \overset{n\to\infty}{\longrightarrow} \int_\Omega f \, d\mu \qquad \text{for all } f \in C_b(\Omega).$$

The Lebesgue measure on $\mathbb{R}^d$ is indicated by $\mathcal{L}^d$.

The *push-forward* of a measure μ by a Borel measurable mapping $\varphi : \Omega \to \Omega$ is the measure $\varphi_\# \mu$ defined by

$$\varphi_\#\mu(A) := \mu(\varphi^{-1}(A)) \qquad \text{for any } A \in \mathcal{B}(\Omega).$$

It satisfies the identity

$$\int_\Omega f(y) \, \varphi_\#\mu(dy) = \int_\Omega f(\varphi(x)) \, \mu(dx) \qquad \text{for all } f \in C_b(\Omega).$$

In these notes Brownian particles have transition probability $(4\pi t)^{d/2} \exp(-|x-y|^2/4t)$, i.e. the corresponding generator is Δ rather than $\frac{1}{2}\Delta$. A subscript t indicates a time slice at time t.

2 Entropy and large deviations

In this section we focus on the concepts of *entropy* and *large-deviation theory*, and their interconnections. The main modelling insights that we want to establish are

M1 *Entropy is best understood as the rate functional of the empirical measure of a large number of identical particles; it describes the probability of that system* (Section 2.4);

M2 *Entropy arises from the indistinguishability of the particles* (Section 2.2);

M3 *Free energy arises from the tilting of a particle system by a heat bath* (Section 2.5).

M3 will also answer part of Question 5.

We start by introducing entropy (Sections 2.1 and 2.2) and large-deviation theory (Section 2.3).

2.1 Entropy

What is entropy? This question has been asked an unimaginable number of times, and has received a wide variety of answers. Here I will not try to summarize the literature, but only mention that I personally like the treatments in [LY97, Eva01].

Instead we define one version of entropy, the *relative* entropy of two probability measures. Fix a probability space (Ω, Σ).

Definition 1. *Let $\mu, \nu \in \mathcal{M}(\Omega)$, with $\mu, \nu \geq 0$. The* relative entropy *of μ with respect to ν is*

$$\mathcal{H}(\mu|\nu) := \begin{cases} \int_\Omega f \log f \, \mathrm{d}\nu & \text{if } \mu \ll \nu \text{ and } f = \dfrac{\mathrm{d}\mu}{\mathrm{d}\nu} \\ +\infty & \text{otherwise.} \end{cases}$$

With this definition we see that the entropy Ent that we defined above can be written as $\mathsf{Ent}(\rho) = \mathcal{H}(\rho|\mathcal{L}^d)$, where $\mathcal{L}^d$ is the Lebesgue measure on $\mathbb{R}^d$.

Before we discuss the interpretation of this object in the next section, we first mention a number of properties.

Theorem 2. *1. If $\mu(\Omega) = \nu(\Omega)$, then $\mathcal{H}(\mu|\nu) \geq 0$, and $\mathcal{H}(\mu|\nu) = 0$ if and only if $\mu = \nu$;*

2. If $\mu(\Omega) = \nu(\Omega)$, then $2\|\mu - \nu\|_{TV}^2 \leq \mathcal{H}(\mu|\nu)$ (Csiszár-Kullback-Pinsker inequality);

3. $\mathcal{H}$ is invariant under transformations of the underlying space, i.e. if $\varphi : \Omega \to \Omega$ is a one-to-one measurable mapping, and $\varphi_\# \mu$ and $\varphi_\# \nu$ are the push-forwards of μ and ν, then $\mathcal{H}(\mu|\nu) = \mathcal{H}(\varphi_\# \mu | \varphi_\# \nu)$.

Proof. The first part of the theorem can be understood by writing $\mathcal{H}$ as

$$\mathcal{H}(\mu|\nu) = \int (f \log f - f + 1) \, d\nu \qquad \text{if } f = \frac{\mathrm{d}\mu}{\mathrm{d}\nu},$$

and using the fact that $g(s) = s \log s - s + 1$ is non-negative and only zero at $s = 1$. For the Csiszár-Kullback-Pinsker inequality we refer to [Csi67, Kul67].

To prove the invariance under transformations, note that for all $\omega \in \Omega$,

$$\frac{\mathrm{d}\varphi_{\#}\mu}{\mathrm{d}\varphi_{\#}\nu}(\varphi(\omega)) = \frac{\mathrm{d}\mu}{\mathrm{d}\nu}(\omega),$$

so that

$$\mathcal{H}(\mu|\nu) = \int \log \frac{\mathrm{d}\mu}{\mathrm{d}\nu}(\omega)\, \mu(d\omega) = \int \log \frac{\mathrm{d}\varphi_{\#}\mu}{\mathrm{d}\varphi_{\#}\nu}(\varphi(\omega))\, \mu(d\omega) = \\ = \int \log \frac{\mathrm{d}\varphi_{\#}\mu}{\mathrm{d}\varphi_{\#}\nu}(\omega')\, \varphi_{\#}\mu(d\omega') = \mathcal{H}(\varphi_{\#}\mu|\varphi_{\#}\nu).$$

□

The properties in this theorem are relevant for the central role that this relative entropy plays. The non-negativity and Csiszár-Kullback-Pinsker inequality show that when μ and ν have the same mass (e.g. if $\mu, \nu \in \mathcal{P}(\Omega)$) then $\mathcal{H}$ acts like a measure of distance between μ and ν. It's not a distance function, since it is not symmetric ($\mathcal{H}(\mu|\nu) \neq \mathcal{H}(\nu|\mu)$), but via the Csiszár-Kullback-Pinsker inequality it does generate the topology of total variation on the space of probability measures.

The fact that $\mathcal{H}$ is invariant under transformations of space is essential for the modelling, for the following reason. Every modelling process involves a choice of coordinates, and this choice can often be made in many different ways. Nevertheless, if we believe that there is a well-defined energy that drives the evolution, then the value of this energy should not depend on which set of coordinates we have chosen to describe it in.

Note that Ent is *not* invariant under change of coordinates. This implies that this functional corresponds to a specific choice of coordinates.

2.2 Entropy as measure of degeneracy

This mathematical definition of the relative entropy $\mathcal{H}$ above does not explain why it might appear in a model. One way to give an interpretation to the relative entropy $\mathcal{H}$ is by a counting argument, that we explain here for the case of a finite state space. It involves *microstates* and *macrostates*, where multiple microstates correspond to a single macrostate. The result will be that the entropy characterizes two concepts: one is the degree of degeneracy, that is the number of microstates that corresponds to a given macrostate, and the other is the probability of observing a macrostate, given a probability distribution over microstates. The two are closely connected.

Take a finite state space I consisting of $|I|$ elements. If $\mu \in \mathcal{P}(I)$, then μ is described by $|I|$ numbers μ_i, and the relative entropy with respect to μ is

$$\mathcal{H}(\rho|\mu) = \sum_{i \in I} \rho_i \log \frac{\rho_i}{\mu_i}, \qquad \text{for } \rho \in \mathcal{P}(I).$$

Consider N particles on the lattice described by I, i.e. consider a mapping $x : \{1, \ldots, N\} \to I$. We think of x as the *microstate*. Define an *empirical measure* $\rho \in \mathcal{P}(I)$ by

$$k_i := \#\big\{j \in \{1, \ldots, N\} : x(j) = i\big\}, \qquad \rho_i = \frac{k_i}{N}. \tag{12}$$

In going from x to ρ there is loss of information; multiple mappings x produce the same empirical measure ρ. The degree of degeneracy, the number of unique mappings x that correspond to a given ρ, is $N!\left(\prod_{i\in I} k_i!\right)^{-1}$. The fact that the particles are identical, *indistinguishable*, is important here—this is required for the description in terms of integers k_i. Because of this loss of information, we think of ρ as the *macrostate*.

We now determine the behaviour of this 'degree of degeneracy' in the limit $N \to \infty$. Using Stirling's formula in the form

$$\log n! = n \log n - n + o(n) \qquad \text{as } n \to \infty,$$

we estimate

$$\begin{aligned}
\log N! \left(\prod_{i\in I} k_i!\right)^{-1} &= \log N! - \sum_{i\in I} \log k_i! \\
&= N \log N - N - \sum_{i\in I} (k_i \log k_i - k_i) + o(N) \\
&= -N \sum_{i\in I} \rho_i \log \rho_i + o(N) \qquad \text{as } N \to \infty.
\end{aligned}$$

One interpretation of the relative entropy therefore is as follows. Take for the moment μ_i to be the uniform measure, i.e. $\mu_i = |I|^{-1}$; then

$$\mathcal{H}(\rho|\mu) = \sum_{i\in I} \rho_i \log \rho_i + \log |I|.$$

Then

$$\mathcal{H}(\rho|\mu) = -\lim_{N\to\infty} \frac{1}{N} \log \#\text{realizations of } \rho \; + \; \log |I|. \tag{13}$$

This shows that if the number of microscopic realizations x of the macroscopic object ρ is large, then $\mathcal{H}(\rho|\mu)$ is small, and vice versa.[2] This is the interpretation in terms of a counting argument.

We now switch to the probabilistic point of view. If we allocate particles at random with the same, independent, probability for each microstate x, then the probability of obtaining each microstate is $|I|^{-N}$, and the probability of a macrostate ρ satisfies

$$\log \text{Prob}(\rho) = \log |I|^{-N} N! \left(\prod_{i\in I} k_i!\right)^{-1} = -N\mathcal{H}(\rho|\mu) + o(N) \qquad \text{as } N \to \infty.$$

We can do the same with non-equal probabilties: we place each particle at an $i \in I$ with probability μ_i. Then the probability of a microstate x is

$$\prod_{j=1}^{N} \mu_{x(j)},$$

[2]There is a problem here, though; the ρ in the right-hand side can not be independent of N, since it is a vector with components of the form k/N, while the ρ in the left-hand side expression can not depend on N. This contradiction will be resolved when we discuss large deviations in the next section.

and now the probability of a macrostate ρ satisfies

$$\begin{aligned}\log \mathrm{Prob}(\rho) &= \log\Big(\prod_{j=1}^{N} \mu_{x(j)}\Big) N! \,\Big(\prod_{i\in I} k_i!\Big)^{-1} \\ &= \sum_{j=1}^{N} \log \mu_{x(j)} + \log N! \,\big(\prod_{i\in I} k_i!\big)^{-1} \\ &= N \sum_{i\in I} \rho_i \log \mu_i - N \sum_{i\in I} \rho_i \log \rho_i + o(N) \\ &= -N\mathcal{H}(\rho|\mu) + o(N) \qquad\qquad \text{as } N \to \infty.\end{aligned}$$

The common element in both points of view is the degeneracy, the number of microstates that is mapped to a single macrostate. Of course, this degeneracy only arises if the particles can not be distinguished from each other. Therefore I like to summarize this section like this:

> *Entropy arises from the indistinguishability of the particles in an empirical measure.*

We will return to this issue in Section 2.4.

2.3 Large deviations

For our purposes, the main role of the relative entropy is in the characterization of *large deviations of empirical measures*.

Large deviations are best explained by an example. We toss a balanced coin n times, and we call S_n the number of heads. Well-known properties of S_n are[3]

- $\frac{1}{n} S_n \overset{n\to\infty}{\longrightarrow} \frac{1}{2}$ almost surely (the law of large numbers)
- $\frac{2}{\sqrt{n}}\big(S_n - \frac{n}{2}\big) \overset{n\to\infty}{\longrightarrow} Z$ in law, where Z is a standard normal random variable (the central limit theorem).

The second (which contains the first) states that S_n is typically $n/2$ plus a random deviation of order $O(1/\sqrt{n})$. Deviations of this size from the expectation are called normal. Large deviations are those that are larger than normal, such as for instance the event that $S_n \geq an$ for some $a > 1/2$. Such large-deviation events have a probability that vanishes as $n \to \infty$, and a *large-deviation principle* characterizes exactly how fast it vanishes. A typical example is

$$\text{For any } a \geq \frac{1}{2}, \qquad \mathrm{Prob}(S_n \geq an) \sim e^{-nI(a)} \qquad \text{as } n \to \infty, \tag{14}$$

where

$$I(a) := \begin{cases} a\log a + (1-a)\log(1-a) + \log 2 & \text{if } 0 \leq a \leq 1 \\ +\infty & \text{otherwise} \end{cases}$$

The characterization (14) states that the probability of such a rare event decays exponentially in n, for large n, for each $a \geq 1/2$. The function I is called the *rate function*, since it characterizes the constant in the exponential decay.

[3] This section draws heavily from the introduction in [dH00].

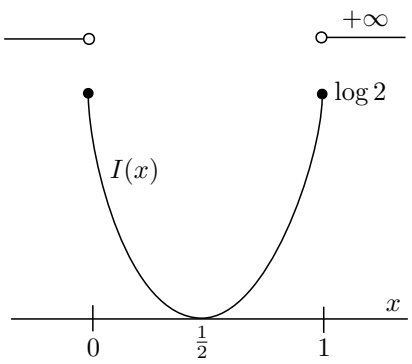

Figure 1: The function I

In order to explain what exactly the symbol $\sim$ in (14) means we give a precise definiton of a large-deviation principle. As always we assume we are working in a complete metrizable separable space Ω with a σ-algebra Σ that contains the Borel sets.

Definition 3. *A sequence $\mu_n \in \mathcal{P}(\Omega)$ satisfies a* large-deviation principle *with speed n and rate function I iff*

$$\begin{aligned} &\forall O \subset \Omega \text{ open}, && \liminf_{n\to\infty} \frac{1}{n} \log \mu_n(O) \geq -\inf_O I \\ &\forall C \subset \Omega \text{ closed}, && \liminf_{n\to\infty} \frac{1}{n} \log \mu_n(C) \leq -\inf_C I. \end{aligned}$$

Let us make a few remarks.

- The definition of the large-deviation principle is rather cumbersome. We often write it, formally, as
$$\mathrm{Prob}(X_n \approx x) \sim e^{-nI(x)},$$
which is intended to mean exactly the same as Definition 3.

- The characterization (14) can be deduced from Definition 3 as follows. Define
$$\mu_n(A) := \mathrm{Prob}(S_n \in nA).$$
Then $\mathrm{Prob}(S_n \geq an) = \mu_n([a,\infty))$, and note that $\inf_{[a,\infty)} I = I(a)$ whenever $a \geq 1/2$. Supposing that the large-deviation principle has been proved for μ_n with rate function I (see e.g. [dH00, Th. I.3]), we then find that for all $1/2 \leq a < 1$
$$\begin{aligned} -I(a) = -\inf_{(a,\infty)} I \;&\leq\; \liminf_{n\to\infty} \mathrm{Prob}(S_n > an) \\ &\leq\; \limsup_{n\to\infty} \mathrm{Prob}(S_n \geq an) \;\leq\; -\inf_{[a,\infty)} I \;=\; -I(a). \end{aligned}$$
Therefore, if $1/2 \leq a < 1$ then the liminf and limsup coincide, and we have that
$$\lim_{n\to\infty} \frac{1}{n} \log \mathrm{Prob}(S_n \geq an) = \lim_{n\to\infty} \frac{1}{n} \log \mathrm{Prob}(S_n > an) = -I(a).$$
This is the precise version of (14).

- In the example of the coin, the two inequalities in Definition 3 reduced to one and the same equality at all points between 0 and 1, by the continuity of I at those points. In general a rate function need not be continuous, as the example of I above shows; neither is the function I unique. However, we can always assume that I is lower semi-continuous, and this condition makes I unique.
- Looking back at the discussion in Section 2.2, we see that for instance the limit (13) is a large-deviation description, at least formally. We also remarked there that the characterization (13) can not be true as it stands. In Definition 3 we see how this is remedied: instead of a single macrostate ρ, we consider open and closed sets of macrostates, which may contain states ρ of the form k/N for different values of N.

Remark 4 The definition of the large-deviation principle has close ties to two other concepts of convergence.

- A sequence of probability measures μ_n converges *narrowly* (in duality with continuous and bounded functions) to μ if

$$\begin{aligned} &\forall O \subset \Omega \text{ open}, && \liminf_{n\to\infty} \mu_n(O) \geq \mu(O) \\ &\forall C \subset \Omega \text{ closed}, && \liminf_{n\to\infty} \mu_n(C) \leq \mu(C). \end{aligned}$$

  Apparently, the large-deviation principle corresponds to a statement like 'the measures $\frac{1}{n}\log \mu_n$ converge narrowly'.

- The definition in terms of two inequalities also recalls the definition of Gamma-convergence, and indeed we have the equivalence (shown to me by Lorenzo Bertini)

$$\mu_n \text{ satisfies a large-deviation principle with rate function } I \quad \Longleftrightarrow \quad \frac{1}{n}\mathcal{H}(\,\cdot\,|\mu_n) \xrightarrow{\Gamma} \hat{I},$$

  where $\hat{I}(\nu) := \int I \, d\nu$. A proof is given in the Appendix. Write this.

A property of large deviations that will come back later is the following. Suppose that we have a large-deviation result for a sequence of probability measures μ_n on a space $\mathcal{X}$ with rate functional I. Suppose that we now *tilt* the probability distribution μ_n by a functional $F : \mathcal{X} \to \mathbb{R}$, by

$$\tilde{\mu}_n(A) = \frac{\displaystyle\int_A e^{-nF(x)}\, \mu_n(dx)}{\displaystyle\int_{\mathcal{X}} e^{-nF(x)}\, \mu_n(dx)}.$$

This increases the probability of events x with lower $F(x)$, with respect to events x with higher $F(x)$, with a similar exponential rate (the prefactor n) as a large-deviation result.

The large-deviation behaviour of $\tilde{\mu}_n$ is now given by

Theorem 5 (Varadhan's Lemma)**.** *Let $F : \mathcal{X} \to \mathbb{R}$ be continuous and bounded from below. Then $\tilde{\mu}_n$ satisfies a large deviation principle with rate function*

$$\tilde{I}(x) := I(x) + F(x) - \inf_{\mathcal{X}}(I + F).$$

The final term in this expression is only a normalization constant that makes sure that $\inf \tilde{I} = 0$. The important part is that $\tilde{I}$ is the sum of the two functions I and F. In words: if we modify a probability distribution by tilting it with an exponential factor e^{-nF}, then that tilting function F ends up being *added* to the original rate function I.

2.4 Entropy as large-deviation rate function

Now to the question why relative entropy appears in the context of thermodynamics. Consider the following situation. We place n independent particles in a space $\mathcal{X}$ according to a distribution $\mu \in \mathcal{P}(\mathcal{X})$, i.e. the probability that the particle is placed in a set $A \subset \mathcal{X}$ is $\mu(A)$. We now consider the empirical measure of these n particles, which is the measure

$$\rho_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i},$$

where X_i is the position of the i^{th} particle. The empirical measure ρ_n is a random element of $\mathcal{P}(\mathcal{X})$, and the law of large numbers gives us that with probability one ρ_n converges weakly (in the sense of measures) to the law μ. (This is of course the standard way of determining μ if one only has access to the sample points X_i).

In this situation the large deviations of ρ_n are given by Sanov's theorem (see e.g. [DZ98, Th. 6.2.10]). The random measure ρ_n satisfies a large-deviation principle with rate n and rate function

$$I(\rho) := \mathcal{H}(\rho|\mu),$$

or in the shorthand notation that we used earlier,

$$\boxed{\text{Prob}(\rho_n \approx \rho) \sim e^{-n\mathcal{H}(\rho|\mu)} \qquad \text{as } n \to \infty.} \tag{15}$$

This is such an important result that I state it separately:

> *The relative entropy is the rate functional of the empirical measure of a large number of identical particles.*

Note the stress on 'empirical measure': the appearance of the relative entropy is intimately linked to the fact that we are considering empirical measures. Section 2.2 gives an insight into why this is: when passing from a vector of positions to the corresponding empirical measure, there is loss of information, since particles at the same position are indistinguishable.

2.5 Free energy and the Boltzmann distribution

In many books one encounters in various forms the following claim. Take a system of particles living in a space $\mathcal{X}$, and introduce an 'energy' $E : \mathcal{X} \to \mathbb{R}$ (in Joules, J) depending on the position $x \in \mathcal{X}$. Bring the system of particles into contact with a 'heat bath' of temperature T. Then the probability distribution of the particles will be given by

$$\text{Prob}(A) = \frac{\displaystyle\int_A e^{-E(x)/kT}\, dx}{\displaystyle\int_{\mathcal{X}} e^{-E(x)/kT}\, dx}. \tag{16}$$

This is known as the *Boltzmann distribution*, or *Boltzmann statistics*, and the *Boltzmann constant* k has the value $1.4 \cdot 10^{-23}\,\text{J/K}$. (Note that it only exists if the exponentials are integrable, which is equivalent to sufficient growth of E for large x. We will assume this for this discussion).

Where does this distribution come from? The concept of entropy turns out to give us the answer.

Since we need a system and a heat bath, we take two systems, called S and S_B (for 'bath'). Both are probabilistic systems of particles; S consists of n independent particles $X_i \in \mathcal{X}$, with probability law $\mu \in \mathcal{P}(\mathcal{X})$; similarly S_B consists of m independent particles $Y_j \in \mathcal{Y}$, with law $\nu \in \mathcal{P}(\mathcal{Y})$. The total state space of the system is therefore $\mathcal{X}^n \times \mathcal{Y}^m$.

The *coupling* between these systems will be via an *energy constraint*. We assume that there are energy functions $e : \mathcal{X} \to \mathbb{R}$ and $e_B : \mathcal{Y} \to \mathbb{R}$, and we will constrain the joint system to be in a state of fixed total energy, i.e. we will only allow states in $\mathcal{X}^n \times \mathcal{Y}^m$ that satisfy

$$\sum_{i=1}^{n} e(X_i) + \sum_{j=1}^{m} e_B(Y_j) = \text{constant}. \tag{17}$$

The physical interpretation of this is that energy (in the form of heat) may flow freely from one system to the other, but no other form of interaction is allowed.

Similar to the example above, we describe the total states of systems S and S_B by empirical measures $\rho_n = \frac{1}{n}\sum_i \delta_{X_i}$ and $\zeta_m = \frac{1}{m}\sum_j \delta_{Y_j}$. We define the average energies $E(\rho_n) := \frac{1}{n}\sum_i e(X_i) = \int_{\mathcal{X}} e \, d\rho_n$ and $E_B(\zeta_m) := \int_{\mathcal{Y}} e_B \, d\zeta_m$, so that the energy constraint above reads $nE(\rho_n) + mE_B(\zeta_m) = \text{constant}$.

By Section 2.4, each of the systems *separately* satisfies a large-deviation principle with rate functions $I(\rho) = \mathcal{H}(\rho|\mu)$ and $I_B(\zeta) = \mathcal{H}(\zeta|\nu)$. However, instead of using the explicit formula for I_B, we are going to assume that I_B can be written as a function of the energy E_B of the heat bath alone, i.e. $I_B(\zeta) = \tilde{I}_B(E_B(\zeta))$. For the coupled system we derive a joint large-deviation principle by choosing that (a) $m = nN$ for some large $N > 0$, and (b) the constant in (17) scales as n, i.e.

$$nE(\rho_n) + nNE_B(\zeta_{nN}) = n\overline{E} \qquad \text{for some } \overline{E}.$$

The joint system satisfies then a large-deviation principle[4]

$$\text{Prob}\Big((\rho_n, \zeta_{nN}) \approx (\rho, \zeta) \;\Big|\; E(\rho_n) + NE_B(\zeta_{nN}) = \overline{E}\Big) \sim \exp\big(-nJ(\rho, \zeta)\big),$$

with rate functional

$$J(\rho, \zeta) := \begin{cases} \mathcal{H}(\rho|\mu) + N\tilde{I}_B(E_B(\zeta)) + \text{constant} & \text{if } E(\rho) + NE_B(\zeta) = \overline{E}, \\ +\infty & \text{otherwise.} \end{cases}$$

Here the constant is chosen to ensure that $\inf J = 0$.

The functional J can be reduced to a functional of ρ alone,

$$J(\rho) = \mathcal{H}(\rho|\mu) + N\tilde{I}_B\left(\frac{\overline{E} - E(\rho)}{N}\right) + \text{constant}.$$

In the limit of large N, one might approximate

$$N\tilde{I}_B\left(\frac{\overline{E} - E(\rho)}{N}\right) \approx N\tilde{I}_B(\overline{E}) - \tilde{I}_B'(\overline{E})E(\rho).$$

[4]This statement is formal; I haven't yet worked out how to formulate this rigorously for general state spaces.

The first term above is absorbed in the constant, and we find

$$J(\rho) \approx \mathcal{H}(\rho|\mu) - E(\rho)\tilde{I}'_B(\overline{E}) + \text{constant}.$$

We expect that I'_B is negative, since larger energies typically lead to higher probabilities and therefore smaller values of I_B. Now we simply define $kT := -1/\tilde{I}'_B(\overline{E})$, and we find

$$J(\rho) \approx \mathcal{H}(\rho|\mu) + \frac{1}{kT}E(\rho) + \text{constant}.$$

Compare this to the typical expression for free energy $E - TS$; if we interpret S as $-k\mathcal{H}(\cdot|\mu)$, then we find the expression above, up to a factor kT.

Note that the right-hand side can be written as $\mathcal{H}(\rho|\tilde{\mu})$, where $\tilde{\mu}$ is the tilted distribution

$$\tilde{\mu}(A) = \frac{\displaystyle\int_A e^{-e(x)/kT}\,\mu(dx)}{\displaystyle\int_{\mathcal{X}} e^{-e(x)/kT}\,\mu(dx)}$$

Here we recognize the expression (16) for the case when μ is the Lebesgue measure.

This derivation shows that the effect of the heat bath is to *tilt* the system S: a state ρ of S with larger energy $E(\rho)$ implies a smaller energy E_B of S_B, which in turn reduces the probability of ρ. The role of temperature T is that of an exchange rate, since it characterizes the change in probability (as measured by the rate function I_B) per unit of energy. When T is large, the exchange rate is low, and then larger energies incur only a small probabilistic penalty. When temperature is low, then higher energies are very expensive, and therefore more rare. From this point of view, the Boltzmann constant k is simply the conversion factor that converts our Kelvin temperature scale for T into the appropriate 'exchange rate' scale.

In books on thermodynamics one often encounters the identity (or definition) $T = dS/dE$. This is formally the same as our definition of kT as $-dI_B/dE$, if one interprets I_B as an entropy and adopts the convention to multiply the non-dimensional quantity I_B with $-k$.

One consequence of the discussion above is that the expression

$$\mathcal{H}(\rho|\mu) + \frac{1}{kT}E(\rho) + \text{constant}$$

is the rate function for a system of particles in contact with a heat bath. This is the same expression as (11), if we identify Φ with E/kT, and this answers the first part of Question 5, 'what is $\mathcal{F}$?' Answer: this 'free energy' is the rate function for a system of particles, in contact with a heat bath of temperature T, and with energy function $\Phi = E$. It's good to note that the Boltzmann distribution is the unique minimizer of the free energy.

Yet another insight that this calculation gives is the following. I was often puzzled by the fact that in defining a free energy (e.g. $E - TS$, or $kT\,\mathsf{Ent} + E$), one adds two rather different objects: the energy of a system seems to be a completely different type of object than the entropy. This derivation of Boltzmann statistics shows that it's not exactly energy and entropy that one's adding; it really is more like adding two entropies ($\mathcal{H}$ and I_B, in the notation above). The fact that we write the second entropy as a constant times energy follows from the coupling and the approximation allowed by the assumption of a large heat bath.

3 Wasserstein distance and Wasserstein gradient flows

In this section we introduce the Wasserstein distance function (Section 3.1) and Wasserstein gradient flows (Section 3.2), in preparation for the discussion of Section 4.

3.1 The Wasserstein distance function

The natural set on which to define the Wasserstein distance is the set of probability measures with finite second moments,

$$\mathcal{P}_2(\mathbb{R}^d) := \left\{ \rho \in \mathcal{P}(\mathbb{R}^d) : \int_{\mathbb{R}^d} |x|^2 \, d\rho < \infty \right\}.$$

The *Wasserstein distance* between two such probability measures $\rho_{0,1} \in \mathcal{P}_2(\mathbb{R}^d)$ is [Vil03]

$$d(\rho_0, \rho_1)^2 := \inf_{q \in \Gamma(\rho_0, \rho_1)} \int_{\mathbb{R}^d \times \mathbb{R}^d} |x - y|^2 \, q(dxdy), \tag{18}$$

where $\Gamma(\rho_0, \rho_1)$ is the set of all *couplings* of ρ_0 and ρ_1, i.e. those q with marginals ρ_0 and ρ_1:

$$\text{for any } A \subset \mathbb{R}^d, \qquad q(A \times \mathbb{R}^d) = \rho_0(A) \qquad \text{and} \qquad q(\mathbb{R}^d \times A) = \rho_1(A).$$

It is also known as a type of 'earth-mover's distance', by its interpretation in terms of the minimal cost of transport of a given pile of sand ρ_0 to fill a given hole ρ_1, where the unit cost of transport of a grain of sand from x to y is the squared distance $|x - y|^2$.

The distance function d is a complete separable metric on the space $\mathcal{P}_2(\mathbb{R}^d)$. Convergence in Wasserstein metric is equivalent to the combination of (a) narrow convergence of measures (i.e. in duality with continuous and bounded functions) and (b) convergence of the second marginals [AGS05, Remark 8.1.5-7].

The Wasserstein distance function has an alternative formulation due to Benamou and Brenier [BB00],

$$d(\rho_0, \rho_1)^2 = \inf \left\{ \int_0^1 \|\partial_t \rho_t\|^2_{-1,\rho_t} \, dt : \rho : [0,1] \to \mathcal{P}_2(\mathbb{R}^d), \ \rho_t\big|_{t=0,1} = \rho_{0,1} \right\},$$

where for fixed t the *local Wasserstein norm* at ρ_t of the 'tangent' $\partial_t \rho_t$ is given by

$$\|\partial_t \rho_t\|^2_{-1,\rho_t} := \inf \left\{ \int_{\mathbb{R}^d} |v|^2 \, d\rho_t : v \in L^2(\rho_t; \mathbb{R}^d), \ \partial_t \rho_t + \operatorname{div} \rho_t v = 0 \right\}. \tag{19}$$

If the distributional derivative $\partial_t \rho_t$ at time t has no representation as $\operatorname{div} \rho_t v$ for any $v \in L^2(\rho_t)$, then the norm is given the value $+\infty$.

The infimum in (19) is achieved, and the corresponding vector field v is an element of the set

$$\overline{\{\nabla \varphi : \varphi \in C_c^\infty(\mathbb{R}^d)\}}^{L^2(\rho_t)}. \tag{20}$$

The distributional derivative $\partial_t \rho_t$ and the corresponding velocity field v are two different ways of describing the same tangent vector. We often need to switch from one to the other, and so we introduce the notation

$$(s, v) \in \operatorname{Tan}_\rho \qquad \Longleftrightarrow \qquad s + \operatorname{div} \rho v = 0 \quad \text{and} \quad v \in \overline{\{\nabla \varphi : \varphi \in C_c^\infty(\mathbb{R}^d)\}}^{L^2(\rho)}.$$

This optimal vector field allows us to define a local metric tensor on the tangent space, the corresponding local inner product:

$$(s_1, s_2)_{-1,\rho_t} := \int_{\mathbb{R}^d} v_1 \cdot v_2 \, d\rho_t \qquad \text{where } (s_i, v_i) \in \mathrm{Tan}_{\rho_t}, \tag{21}$$

such that the norm (19) satisfies $\|s\|^2_{-1,\rho} = (s, s)_{-1,\rho}$.

3.2 Wasserstein gradient flows

Turning to Wasserstein *gradient flows*, there is a simple, but formal, way of describing these. This way is obtained by taking the Hilbert-space concept of a gradient flow (as in e.g. (4–5) or (6)) and taking for the inner product the local metric tensor $(\cdot, \cdot)_{-1,\rho_t}$ defined in (21). In this way the Wasserstein gradient flow of a functional $\mathcal{E}$ defined on $\mathcal{P}_2(\mathbb{R}^d)$ is formally characterized by

$$(\partial_t \rho_t, s)_{-1,\rho_t} = -\langle \mathcal{E}'(\rho_t), s \rangle \qquad \text{for all times } t \text{ and tangent vectors } s. \tag{22}$$

In Section 1.3 we claimed that the diffusion equation (1) is the Wasserstein gradient flow of the entropy $\mathsf{Ent}(\rho) := \int_{\mathbb{R}^d} \rho(x) \log \rho(x) \, dx$ (we often write $\rho(x)$ for the Lebesgue density of a measure ρ). Let us verify this, at least formally.

Finiteness of the entropy $\mathsf{Ent}(\rho)$ implies that ρ is Lebesgue absolutely continuous. If we take a tangent vector s that is also Lebesgue absolutely continuous, then the action of the derivative $\mathsf{Ent}'(\rho)$ on a tangent vector s is given by

$$\langle \mathsf{Ent}'(\rho), s \rangle := \lim_{h \to 0} \frac{\mathsf{Ent}(\rho + hs) - \mathsf{Ent}(\rho)}{h} = \int_{\mathbb{R}^d} (\log \rho(x) + 1) s(x) \, dx.$$

The left-hand side of (22) is equal to

$$\int_{\mathbb{R}^d} v_1 \cdot v_2 \, d\rho_t,$$

where $(\partial_t \rho_t, v_1), (s, v_2) \in \mathrm{Tan}_{\rho_t}$. Assume we can write $v_i = \nabla \varphi_i$ for some $\varphi_i \in C_c^\infty$; the definition (20) assures that this is possible at least in approximation. Then the integral above becomes

$$\int_{\mathbb{R}^d} \nabla \varphi_1 \cdot \nabla \varphi_2 \, d\rho_t,$$

which we rewrite by partial integration to

$$-\int_{\mathbb{R}^d} \varphi_2 \, \mathrm{div}\, \rho_t \nabla \varphi_1 \overset{(\partial_t \rho_t, \nabla \varphi_1) \in \mathrm{Tan}_{\rho_t}}{=} \int \varphi_2 \partial_t \rho_t.$$

For fixed t and s we then rewrite (22) as

$$\begin{aligned} \int \varphi_2 \partial_t \rho_t &= \int (\log \rho_t + 1) s \\ &= -\int (\log \rho_t + 1) \, \mathrm{div}\, \rho_t \nabla \varphi_2 \\ &= \int \nabla (\log \rho_t + 1) \cdot \nabla \varphi_2 \, d\rho_t \\ &= \int \nabla \rho_t \cdot \nabla \varphi_2. \end{aligned} \tag{23}$$

This equality is to hold for all tangent vectors $s = -\operatorname{div} \rho_t \nabla \varphi_2$, and therefore for all functions φ_2. Therefore (23), like (7), is a weak formulation of the diffusion equation (1).

A very similar calculation shows how the expression (8) describes a Wasserstein gradient flow for an arbitrary functional $\mathcal{F}$, using the fact that, at least formally,

$$\langle \mathcal{F}'(\rho_t), \partial_t \rho_t \rangle = \int_{\mathbb{R}^d} \frac{\partial \mathcal{F}}{\partial \rho}(\rho_t)\, \partial_t \rho_t.$$

4 Example: the diffusion equation

The aim of this section is to develop the following modelling concepts:

M1 *The Wasserstein distance characterizes the mobility of empirical measures of systems of Brownian particles* (Section 4.1);

M2 *The Wasserstein gradient flow of the entropy arises from the large-deviation behaviour of a system of Brownian particles* (Section 4.2).

This will also answer Questions 2 and 3.

4.1 A corresponding microscopic model: independent Brownian particles

The calculations in the previous section might be sufficient to convince you that the diffusion equation *is* the Wasserstein gradient flow of the entropy, but they do not explain *why* this is the case, which is one of the main issues of these notes. We now turn to a more microscopic model, in terms of stochastic particles, to motivate this.

The model is a system of n independent Brownian particles $X_{n,i}$ in $\mathbb{R}^d$ ($i = 1, \dots, n$). Each particle has a transition kernel[5]

$$p_h(x, y) := \frac{1}{(4\pi h)^{d/2}} e^{-|x-y|^2/4h}, \tag{24}$$

which describes the probability of finding the particle at y after time h conditioned on it being at x at time zero. For the moment we focus on just the two times $t = 0$ and $t = h$. The particles are not identically distributed; while their transition probabilities are the same, their initial position differs, as follows. Given an initial datum ρ^0, we choose a sequence $x_n = (x_{n,i})_i$, for $i = 1 \dots n$ and $n = 1, 2, \dots$, such that

$$\rho^{0,n} := \frac{1}{n} \sum_{i=1}^{n} \delta_{x_{n,i}} \longrightarrow \rho^0 \qquad \text{as } n \to \infty.$$

For each n, we then give particle $X_{n,i}$ a deterministic initial position $x_{n,i}$. In words, therefore, this is a collection of particles that start at positions constructed such as to reproduce ρ^0, and jump over time h to new positions as described by p_h.

In Section 2.4 we showed how the entropy Ent is the large-deviation rate functional of the empirical measure of a similar system of particles in a static situation. Here we want to work towards the dynamic situation. Mirroring the earlier setup, we consider the empirical measure ρ^n of the particles at time h,

$$\rho^n := \frac{1}{n} \sum_{i=1}^{n} \delta_{X_{n,i}(h)}.$$

For this particle system a large-deviation result states that [Léo07, PR11]

$$\mathrm{Prob}(\rho^n \approx \rho) \sim e^{-nI_h(\rho|\rho^0)}, \tag{25}$$

[5] In the probability literature Brownian particles usually have generator $\frac{1}{2}\Delta$, and the exponent reads $-|x-y|^2/2h$; here we take generator Δ and therefore exponent $-|x-y|^2/4h$.

where the rate functional I is given by

$$I_h(\rho|\rho^0) = \inf_{q \in \Gamma(\rho,\rho^0)} \mathcal{H}(q|Q), \tag{26}$$

for the measure Q given by

$$Q(dxdy) = p_h(x,y)\rho^0(dx)dy.$$

It was proved in [Léo07] that

$$4hI_h(\,\cdot\,|\rho^0) \xrightarrow{\Gamma} d(\,\cdot\,,\rho^0)^2 \qquad \text{as } h \to 0. \tag{27}$$

(The type of convergence is Γ- or Gamma-convergence, a natural concept of convergence for functionals. See [DM93] for an introduction.)

Although this is a simple convergence result, it is actually the first result that indicates to us *why* the Wasserstein distance plays a role in the gradient-flow structure of the diffusion equation. The argument runs as follows. For the moment, disregard the distinction between $4hI_h$ and d^2, or between I_h and $d^2/4h$. The rate functional I_h characterizes the probability distribution of the system of particles: high I_h implies low probability, and vice versa. If I_h is essentially the same as $d^2/4h$, then apparently the Wasserstein distance tells us the same thing: high values of d have low probability, and low values have high probability.

Therefore the value of $d(\rho,\rho^0)$ is a characterization of the distance that the particles have travelled between time zero and time h:

> *The Wasserstein distance characterizes the mobility of empirical measures of systems of Brownian particles.*

We'll see another interpretation of this fact in Section 6.1 below. Also compare this to the statement on page 14.

Before continuing we briefly describe why (27) is true. Using the definition of the relative entropy, we write (assuming all measures are Lebesgue absolutely continuous for simplicity)

$$\begin{aligned}\mathcal{H}(q|Q) &= \int_{\mathbb{R}^d\times\mathbb{R}^d} q(x,y)\big[\log q(x,y) - \log p_h(x,y) - \log\rho^0(x)\big]\,dxdy \\ &= \mathsf{Ent}(q) + \int_{\mathbb{R}^d\times\mathbb{R}^d} q(x,y)\Big[\frac{d}{2}\log 4\pi h + \frac{1}{4h}|x-y|^2 - \log\rho^0(x)\Big]\,dxdy.\end{aligned}$$

Formally, the largest term on the right-hand side is the term

$$\frac{1}{4h}\int_{\mathbb{R}^d\times\mathbb{R}^d} q(x,y)|x-y|^2\,dxdy,$$

where the integral is the same as in the definition of $d(\rho^0,\pi_2 q)^2$, where $\pi_2 q(y) = \int_{\mathbb{R}^d} q(x,y)\,dx$ is the projection onto the second variable.

Indeed, as $h \to 0$, it is shown in [ADPZ10] under some restrictions that this term dominates the right-hand side, and that the other terms diverge as $h \to 0$, but only logarithmically; therefore they vanish after multiplication by h, as in (27). In addition, in the limit $h \to 0$ the optimal q in the infimum (26) converges narrowly to the optimum in (18).

This 'proof' explains for instance why there is a relationship between the Wasserstein distance and a system of Brownian particles: the expression $|x-y|^2$ for the cost in (18) is exactly the same as the exponent $|x-y|^2$ in the Gaussian transition probability p_h (24). This exponent itself can ultimately be traced back to the central limit theorem.

4.2 Scaling the ladder: the next step up

The insight gained in the previous section is useful, but not good enough to explain the gradient-flow structure completely. As we discussed in the Introduction, three questions need to be answered. The first question for the Entropy-Wasserstein gradient flow, 'why Wasserstein?', has now received an answer, but the other two questions, 'why entropy?' and 'why this combination?' are still open. Therefore we need to sharpen up our analysis.

The convergence result (27) implies that I_h is similar to $d^2/4h$ for small h. As we shall see, we need what is essentially the next order in this asymptotic development in h, which is the $O(h^0)$ term. In [ADPZ10] it was shown that (27) indeed can be improved to give

$$I_h(\,\cdot\,|\rho^0) - \frac{1}{4h} d(\,\cdot\,,\rho^0)^2 \xrightarrow{\Gamma} \frac{1}{2}\,\mathsf{Ent}(\,\cdot\,) - \frac{1}{2}\,\mathsf{Ent}(\rho^0) \qquad \text{as } h \to 0,$$

under some technical restrictions. This suggests the asymptotic development

$$I_h(\rho|\rho^0) \approx \frac{1}{4h} d(\rho,\rho^0)^2 + \frac{1}{2}\,\mathsf{Ent}(\rho) - \frac{1}{2}\,\mathsf{Ent}(\rho^0) + o(1) \qquad \text{as } h \to 0. \tag{28}$$

This expression connects to a well-known approximation scheme for gradient flows, which we now discuss.

4.3 Time-discrete approximation

Gradient flows have a classical time-discrete approximation scheme that goes back at least as far as De Giorgi. For instance, for the case of the Wasserstein gradient flow of the free-energy functional $\mathcal{F}$ defined in (11), this takes the following form. Given an initial datum ρ^0 and a time step $h > 0$, iteratively construct a sequence of approximations ρ^k by the minimization problem

Given ρ^{k-1}, let ρ^k be a minimizer of

$$\rho \mapsto \mathcal{F}(\rho) + \frac{1}{2h} d(\rho,\rho^{k-1})^2. \tag{29}$$

Here d is the Wasserstein metric defined in (18). The piecewise-constant interpolation $t \mapsto \rho_t$ defined by $\rho_t := \rho^{\lfloor t/h \rfloor}$ then converges narrowly to the solution ρ of (10) with initial datum ρ^0 [JKO98].

This origin of this approximation scheme is best recognized by formulating it for a simple gradient flow in $\mathbb{R}^n$ of an energy $\mathcal{E}$. Minimizing the corresponding functional

$$u \mapsto \mathcal{E}(u) + \frac{1}{2h}|u - u^{k-1}|^2$$

leads to the stationarity condition for u

$$\nabla\mathcal{E}(u) + \frac{u - u^{k-1}}{h} = 0,$$

which is the backward-Euler discretization of the gradient-flow equation $\dot{u} = -\nabla\mathcal{E}$.

This time-discrete approximation shows how the expression (28) connects to the Wasserstein gradient flow of the entropy Ent, since (28) is exactly one-half of (29) (with $\mathcal{F} = \mathsf{Ent}$). Need to straighten out the free energy

4.4 Wrapping up the modelling arguments

With all the pieces in place we can now describe the whole argument.

1. The starting point is the assumption that our intended gradient flow is the macroscopic, or upscaled version, of a set of particles with Brownian mobility.

2. The asymptotic expression for the rate function for the empirical measure after time h, starting from ρ^0, is (see (28))

$$\rho \mapsto \frac{1}{4h} d(\rho^0, \rho)^2 + \frac{1}{2}\,\mathsf{Ent}(\rho) - \frac{1}{2}\,\mathsf{Ent}(\rho^0). \tag{30}$$

   We think of this as the result of two separate effects. The first, the Wasserstein distance $d(\rho^0, \rho)^2/4h$, describes the probability that the empirical measure transitions from ρ^0 to ρ, as discussed in Section 4.1.

3. The two terms $\frac{1}{2}\,\mathsf{Ent}(\rho) - \frac{1}{2}\,\mathsf{Ent}(\rho^0)$, as the entropy difference between initial and final time, represent the degree to which the final distribution ρ has a higher probability of being chosen than the initial distribution ρ^0—*because* of indistinguishibility. This is in the same sense as the characterization (15) and the discussion of Section 2.2. Two things are different, in comparison to (15): the factor 1/2 and the absence of a reference measure.

   - The factor 1/2 I can't explain, to be honest, other than by remarking 'that it works'—in the sense that the resulting equation (1) is the right one if and only if the prefactor of Ent is 1/2. If anyone has any suggestions I'd be very interested.
   - The lack of a reference measure is a consequence of the fact that, for a Brownian motion without potential landscape, all positions x are equal. The natural reference measure therefore is translation-invariant. This rules out a probability measure on $\mathbb{R}^d$, but the Lebesgue measure $\mathcal{L}^d$ is a natural choice: and $\mathcal{H}(\rho|\mathcal{L}^d) = \mathsf{Ent}(\rho)$. Note that if $a > 0$, then $\mathcal{H}(\rho|a\mathcal{L}^d) = \mathsf{Ent}(\rho) - \log a$; therefore multiplying the Lebesgue measure by a doesn't change the *difference* $\frac{1}{2}\mathcal{H}(\rho|a\mathcal{L}^d) - \frac{1}{2}\mathcal{H}(\rho^0|a\mathcal{L}^d) = \frac{1}{2}\,\mathsf{Ent}(\rho) - \frac{1}{2}\,\mathsf{Ent}(\rho^0)$.

4. The fact that the two contributions are *added* can be understood with Theorem 5, Varadhan's Lemma. The large-deviation behaviour of Brownian particles *without* taking into account the indistinguishibility is given by $d(\rho^0, \rho)^2/4h$; if we tilt the distribution by adding in the indistinguishibility, described by the difference of the entropies, then the result is the sum of the two expressions.

5. We now finish with the remark that the expression (30) is the time-discrete approximation of a gradient flow driven by Ent and with the Wasserstein distance d as the dissipation, as described in Section 4.3.

This overview provides the answer to Question 2, 'What is the modelling background of the Wasserstein-Entropy gradient-flow structure?'.

5 Intermezzo: Continuous-time gradient-flow formulations

The discussion of the previous section is based upon a discrete-time view: the large-deviation rate function concerned empirical measures at time zero and time $h > 0$, and the time-discrete approximation of the gradient-flow similarly connected states at time 0 and time h.

A parallel way to connect large-deviation principles and gradient flows makes use of a time-continuous description, which we now describe. Consider a linear space $\mathcal{Z}$ and an energy functional $\mathcal{E} : \mathcal{Z} \to \mathbb{R}$. If $\mathcal{Z}$ happens to be a Hilbert space, then the Fréchet derivative $\mathcal{E}'$ can be represented as the gradient $\operatorname{grad}\mathcal{E}$ (see (4)). Then we can compute that along any curve $z \in C^1([0,T];\mathcal{Z})$,

$$\begin{aligned}
\partial_t \mathcal{E}(z(t)) \;=\; \langle \mathcal{E}'(z), \dot z\rangle \quad &= \quad (\operatorname{grad}\mathcal{E}(z), \dot z)_{\mathcal{Z}} \\
&\overset{(*)}{\geq} \quad -\|\operatorname{grad}\mathcal{E}(z)\|_{\mathcal{Z}}\|\dot z\|_Z \\
&\overset{(**)}{\geq} \quad -\frac{1}{2}\|\operatorname{grad}\mathcal{E}(z)\|^2_{\mathcal{Z}} - \frac{1}{2}\|\dot z\|^2_Z \\
&\overset{(***)}{=} \quad -\frac{1}{2}\|\mathcal{E}'(z)\|^2_{\mathcal{Z}'} - \frac{1}{2}\|\dot z\|^2_Z.
\end{aligned}$$

The identity $(***)$ is a consequence of the definition of the gradient.

The inequality $(*)$ is the Cauchy-Schwarz inequality, and becomes an identity if and only if $\operatorname{grad}\mathcal{E}$ is equal to $\lambda\dot z$ for some $\lambda < 0$; Young's inequality $(**)$ becomes an identity if and only if the two norms are equal, implying $\lambda = -1$. Therefore identity is achieved in this inequality iff $\operatorname{grad}\mathcal{E}(z) = -\dot z$, i.e. if z is a gradient-flow solution.

This can be generalized by introducing a pair of convex function $\psi : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ and its Legendre transform, or convex dual, $\psi^* : \mathcal{Z}' \times \mathcal{Z} \to \mathbb{R}$. In ψ and ψ^* we think of the second variable as a parameter, and the convexity and duality only refers to the first parameter. Therefore

$$\forall\, s \in \mathcal{Z},\, \xi \in \mathcal{Z}',\; z \in \mathcal{Z} : \qquad \psi(s;z) + \psi^*(\xi;z) \geq \langle s, z\rangle, \tag{31}$$

with equality iff $s \in \partial\psi^*(\xi;z) \iff \xi \in \partial\psi(s;z)$. For smooth convex functions, $\partial\psi$ is simply the derivative of ψ. (See [Roc72] for a general treatise on convex functions, or [Bre11, Section 1.4] for a quick introduction). The example of the Hilbert space above is of this type, with $\psi(s;z) = \psi(s) = \frac{1}{2}\|s\|^2_Z$ and $\psi^*(\xi;z) = \psi(\xi) = \frac{1}{2}\|\xi\|^2_{\mathcal{Z}'}$.

With (31) we find similarly that

$$\partial_t \mathcal{E}(z(t)) = \langle \mathcal{E}'(z), \dot z\rangle \geq -\psi(\dot z; z) - \psi^*(-\mathcal{E}'(z); z), \tag{32}$$

with equality if and only if

$$\dot z \in \partial\psi^*(-\mathcal{E}'(z); z). \tag{33}$$

This latter equation is the generalization of the gradient-flow equation $\dot z = -\operatorname{grad}\mathcal{E}(z)$ to the case of general ψ.

We will mostly use this property in an integrated form. Integrate (32) in time to find that the expression

$$\mathcal{E}(z(T)) - \mathcal{E}(z(0)) + \int_0^T \big[\psi(\dot z; z) + \psi^*(-\mathcal{E}'(z); z)\big]\, dt$$

is non-negative for any curve z. In addition, if the expression is zero, then at almost all $t \in [0,T]$ the equation (33) is satisfied. This provides the definition of a gradient-flow solution that we will use extensively below:

Definition 6. *The curve $z : [0, T] \to \mathcal{Z}$ is a gradient-flow solution (associated with energy $\mathcal{E}$ and dissipation potential ψ) if and only if*

$$\mathcal{E}(z(T)) - \mathcal{E}(z(0)) + \int_0^T \left[\psi(\dot{z}; z) + \psi^*(-\mathcal{E}'(z); z)\right] dt \leq 0. \tag{34}$$

Note that although this definition seems to require that $\dot{z}$ and $-\mathcal{E}'$ are well-defined, actually only the objects $\psi(\dot{z}; z)$ and $\psi^*(-\mathcal{E}'(z); z)$ need to be well-defined, as we shall see in the next section.

6 Examples of continuous-time large deviations and gradient flows

In this section we again study the connection between gradient flows and large-deviation principles, but now in a continuous-time context. The main modelling insights are

M1 *The continuous-time large-deviation rate functional* ***is*** *the gradient flow* (Section 6.1);

M2 *The Kawasaki-Wasserstein gradient flow of the mixing entropy arises from the simple symmetric exclusion process (Question 4)* (Section 6.2).

6.1 Continuous-time large deviations for the system of Brownian particles

To take the important case of the Wasserstein metric, the corresponding objects are

$$\begin{aligned} \psi(\partial_t\rho;\rho) &= \frac{1}{2}\|\partial_t\rho\|^2_{-1,\rho} && \text{where the norm } \|\cdot\|_{-1,\rho} \text{ is defined in (19),} \\ \psi^*(\xi;\rho) &= \frac{1}{2}\int_{\mathbb{R}^d} |\nabla w|^2\, d\rho && \text{if } w \text{ is related to } \xi \text{ by } (w,\zeta)_{L^2(\mathbb{R}^d)} = \langle \xi,\zeta\rangle \ \forall \zeta. \end{aligned}$$

With this description, the gradient flow *is the same as* the large-deviation rate functional, as we shall now see.

Taking the same system of particles as in Section 4.1, the continuous-time large-deviation principle for that system of particles is as follows. Fix a final time $T > 0$ and consider again the empirical measure $t \mapsto \rho_n(t)$, now considered as a time-parametrized curve of measures. Then the probability that the *entire curve* $\rho_n(\cdot)$ is close to some other $\rho(\cdot)$ is characterized as (see [KO90] or [FK06, Th. 13.37])

$$\text{Prob}(\rho_n \approx \rho) \sim \exp[-nI(\rho)],$$

where now

$$I(\rho) := \frac{1}{2}\int_0^T \|\partial_t\rho_t - \Delta\rho_t\|^2_{-1,\rho_t}\, dt. \tag{35}$$

Note that I is nonnegative and its minimum of zero is achieved if and only if ρ is a solution of the diffusion equation (1).

This rate function I has the structure of the left-hand side in (34). Expanding the square in (35), and using the property that

$$(\partial_t\rho_t, -\Delta\rho_t)_{-1,\rho_t} = \frac{d}{dt}\,\mathsf{Ent}(\rho_t),$$

we find that

$$I(\rho) = \mathsf{Ent}(\rho_T) - \mathsf{Ent}(\rho_0) + \frac{1}{2}\int_0^T \Big[\|\partial_t\rho_t\|^2_{-1,\rho_t} + \|\Delta\rho_t\|^2_{-1,\rho_t}\Big]\, dt.$$

The first term is $\psi(\partial_t\rho_t)$; the final term is equal to $\psi^*(\mathsf{Ent}'(\rho_t);\rho_t)$, since (assuming all measures are absolutely continuous)

$$\Delta\rho = \operatorname{div}\rho\nabla\log\rho \quad\Longrightarrow\quad \|\Delta\rho\|^2_{-1,\rho} = \int |\nabla\log\rho(x)|^2\rho(x)\, dx,$$

and

$$\langle \mathsf{Ent}'(\rho), \zeta \rangle = \int (\log \rho(x) + 1)\zeta \, dx \quad \Longrightarrow \quad \psi^*(\mathsf{Ent}'(\rho); \rho) = \frac{1}{2} \int |\nabla(\log \rho(x) + 1)|^2 \rho(x) \, dx.$$

In this sense, now, the large-deviation rate function I of the system of particles *is the same as* the gradient-flow formulation (as given by Definition 6).

Let's connect this to the appearance of the Wasserstein metric as we discussed in Section 4. In that discrete-time context we saw that the Wasserstein distance characterizes the mobility of empirical measures of Brownian particles, through the large-deviation result (25) and the convergence result (27). Here we see that the Wasserstein geometry also characterizes the large deviations of the full time-parametrized path: in (35) the deviation from the deterministic limit $\partial_t \rho_t = \Delta \rho_t$ is measured (and penalized) in the Wasserstein norm $\| \cdot \|_{-1,\rho}$.

6.2 Simple symmetric exclusion process

Question 4 asks 'where does the Kawasaki-Wasserstein gradient-flow formulation of the diffusion equation come from?' Here we give an answer in terms of the symmetric simple exclusion process.

Consider a periodic lattice $\mathbb{T}_n = \{0, 1/n, 2/n, \dots (n-1)/n\}$ and its continuum limit, the flat torus $\mathbb{T} = \mathbb{R}/\mathcal{Z}$. Each lattice site contains zero or one particle; each particle attempts to jump from to a neighbouring site with rate $n^2/2$, and they succeed if the target site is empty. We define the configuration $\rho_n : \mathbb{T}_n \to \{0, 1\}$ such that $\rho_n(k/n) = 1$ if there is a particle at site k/n, and zero otherwise. For this system the large deviations are characterized by the rate function [KOV89]

$$I(\rho) := \frac{1}{2} \int_0^T \|\partial_t \rho - \partial_{xx} \rho\|^2_{\rho(1-\rho)} \, dt, \tag{36}$$

where the norm $\| \cdot \|_{\rho(1-\rho)}$ is given by a modified version of (19), as follows (and again assuming absolute continuity of ρ):

$$\|s\|_{-1,\rho(1-\rho)} := \inf \left\{ \int_{\mathbb{T}} |v|^2 \, \rho(1-\rho) \, dx : v \in L^2(\rho(1-\rho)), \ s + \operatorname{div} \rho(1-\rho) v = 0 \right\}.$$

This functional can be written as

$$I(\rho) = \mathsf{Ent}_{\mathrm{mix}}(\rho_T) - \mathsf{Ent}_{\mathrm{mix}}(\rho_0) + \frac{1}{2} \int_0^T \left[\|\partial_t \rho_t\|^2_{-1,\rho_t(1-\rho_t)} + \|\partial_{xx} \rho_t\|^2_{-1,\rho_t(1-\rho_t)} \right] dt,$$

where the mixing entropy $\mathsf{Ent}_{\mathrm{mix}}$ is defined as in the introduction as

$$\mathsf{Ent}_{\mathrm{mix}}(\rho) := \int_{\mathbb{T}} \big[\rho \log \rho + (1-\rho) \log(1-\rho) \big].$$

The expression $-\partial_{xx} \rho$ is the '$\rho(1-\rho)$'-Wasserstein gradient of $\mathsf{Ent}_{\mathrm{mix}}$, since

$$-\partial_{xx} \rho = -\partial_x \left(\rho(1-\rho) \partial_x \log \frac{\rho}{1-\rho} \right) = -\partial_x \left(\rho(1-\rho) \partial_x \frac{\delta \, \mathsf{Ent}_{\mathrm{mix}}}{\delta \rho}(\rho) \right).$$

Therefore I is again of the form (34), and the fact that the equation $\partial_t \rho = \partial_{xx} \rho$ is (also) the gradient flow of $\mathsf{Ent}_{\mathrm{mix}}$ with respect to this '$\rho(1-\rho)$'-Wasserstein norm $\| \cdot \|_{-1,\rho(1-\rho)}$ can be traced back to the large deviations of this simple symmetric exclusion process.

7 The H^1-L^2-gradient flow

Question 1 of these notes was 'where does the H^1-L^2 gradient-flow formulation of the diffusion equation come from?' (see page 4). I don't have a convincing model for this, but actually the absense of a good model is even more interesting than its existence might have been.

In this section I describe my best shot. It's an unreasonable, unconvincing model, but it's the best way I have of connecting this gradient flow to a model.

In this model the energy and dissipation are mechanical, not probabilistic. The continuum system is the limit of discrete systems, consisting of n spheres in a viscous fluid, arranged in a horizontal row and spaced at a distance from each other. Each sphere is constrained to move only vertically, so that the degrees of freedom are the n vertical positions u_i.

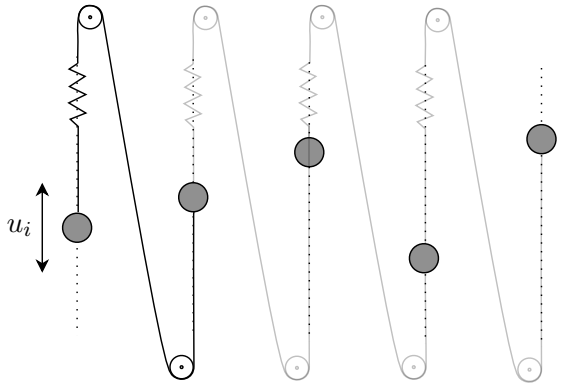

Figure 2: The model of this section: spheres in a viscous fluid, only moving up and down, and connected by elastic springs. The spring measures the difference in displacement $u_i - u_{i+1}$.

The energy of this discrete system arises from springs that connect the spheres, in such a way that the energy of the spring connecting sphere i to sphere $i+1$ is $k(u_i - u_{i+1})^2/2$. Therefore the total energy is

$$\mathcal{E}_n(u_1, \dots, u_n) = \sum_{i=1}^{n-1} \frac{k}{2}(u_i - u_{i+1})^2.$$

Movement of the spheres requires displacement of the surrounding fluid, and therefore dissipates energy. The dissipation rate corresponding to a single sphere moving with velocity v is assumed to be $cv^2/2$, where $c > 0$ depends on the fluid. (Stokes' law gives $c = 6\pi R\eta$, where R is the radius of the sphere and η the viscosity of the fluid). For a system of moving spheres the total dissipation rate is therefore

$$\sum_{i=1}^{n} \frac{c}{2}\dot{u}_i^2.$$

We now construct a gradient flow by putting energy and dissipation together, resulting in the set of equations

$$\begin{aligned} c\dot{u}_1 &= k(u_2 - u_1) \\ c\dot{u}_i &= k(u_{i+1} - 2u_i + u_{i-1}) \qquad \text{for } i = 2, \dots, n-1 \\ c\dot{u}_n &= k(u_{n-1} - u_n) \end{aligned}$$

We can now take the continuum limit, either in the energy and dissipation, or in the equations above; both yield the same limit, which is the L^2-gradient flow of the energy

$$\mathcal{E}(u) := \int |\nabla u|^2,$$

and the diffusion equation (1). However, the important modelling choices are already clear in the discrete case. These two choices, viscous dissipation and interparticle coupling, are very reasonable in themselves, but the unnatural restriction to vertical movement and the just-as-unnatural spring arrangement make this model a very strange one.

References

[ADPZ10] S. Adams, N. Dirr, M. A. Peletier, and J. Zimmer. From a large-deviations principle to the Wasserstein gradient flow: A new micro-macro passage. *Arxiv preprint arxiv:1004.4076*, 2010. Accepted for publication in Communications in Mathematical Physics.

[AGS05] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Lectures in mathematics ETH Zürich. Birkhäuser, 2005.

[BB00] J.-D. Benamou and Y. Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numer. Math.*, 84:375–393, 2000.

[Bre73] H. Brezis. *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*. North Holland, 1973.

[Bre11] H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer, New York, 2011.

[Cho01] R. Choksi. Scaling laws in microphase separation of diblock copolymers. *J. Nonlinear Sci.*, 11:223–236, 2001.

[CMV03] J. A. Carrillo, R. J. McCann, and C. Villani. Kinetic equilibration rates for granular media and related equations: Entropy dissipation and mass transportation estimates. *Revista Matematica Iberoamericana*, 19(3):971–1018, 2003.

[CMV06] J. A. Carrillo, R. J. McCann, and C. Villani. Contractions in the 2-Wasserstein length space and thermalization of granular media. *Archive for Rational Mechanics and Analysis*, 179:217–263, 2006.

[Csi67] I. Csiszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318, 1967.

[dH00] F. den Hollander. *Large Deviations*. American Mathematical Society, Providence, RI, 2000.

[DLSS91] B. Derrida, JL Lebowitz, ER Speer, and H. Spohn. Fluctuations of a stationary nonequilibrium interface. *Physical review letters*, 67(2):165–168, 1991.

[DM93] G. Dal Maso. *An Introduction to* Γ*-Convergence*, volume 8 of *Progress in Nonlinear Differential Equations and Their Applications*. Birkhäuser, Boston, first edition, 1993.

[DZ98] A. Dembo and O. Zeitouni. *Large deviations techniques and applications*. Springer Verlag, 1998.

[Eva01] L. C. Evans. Entropy and partial differential equations. Technical report, UC Berkeley, 2001.

[FK06] J. Feng and T.G. Kurtz. *Large deviations for stochastic processes*, volume 131 of *Mathematical Surveys and Monographs*. American Mathematical Society, 2006.

[GL98] G. Giacomin and J. L. Lebowitz. Phase segregation dynamics in particle systems with long range interactions II: Interface motion. *SIAM J. Appl. Math*, 58:1707–1729, 1998.

[GO01] L. Giacomelli and F. Otto. Variational formulation for the lubrication approximation of the Hele-Shaw flow. *Calculus of Variations and Partial Differential Equations*, 13(3):377–403, 2001.

[JKO97] R. Jordan, D. Kinderlehrer, and F. Otto. Free energy and the Fokker-Planck equation. *Physica D: Nonlinear Phenomena*, 107(2-4):265–271, 1997.

[JKO98] R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the Fokker-Planck Equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

[JM09] A. Jüngel and D. Matthes. A review on results for the Derrida-Lebowitz-Speer-Spohn equation. In *Proceedings of the EQUADIFF07 Conference*, 2009.

[KO90] C. Kipnis and S. Olla. Large deviations from the hydrodynamical limit for a system of independent Brownian particles. *Stochastics and stochastics reports*, 33(1-2):17–25, 1990.

[KOV89] C. Kipnis, S. Olla, and SRS Varadhan. Hydrodynamics and large deviation for simple exclusion processes. *Communications on Pure and Applied Mathematics*, 42(2):115–137, 1989.

[Kul67] S. Kullback. A lower bound for discrimination information in terms of variation. *IEEE Transactions on Information Theory*, 13(1):126–127, 1967.

[Léo07] C. Léonard. A large deviation approach to optimal transport. *Arxiv preprint arXiv:0710.1461*, 2007.

[LY97] E.H. Lieb and J. Yngvason. A guide to entropy and the second law of thermodynamics. *Notices of the AMS*, 45(5), 1997.

[MMS09] D. Matthes, R. J. McCann, and G. Savaré. A family of nonlinear fourth order equations of gradient flow type. *Arxiv preprint arXiv:0901.0540*, 2009.

[Ott98] F. Otto. Lubrication approximation with prescribed nonzero contact angle. *Communications in Partial Differential Equations*, 23(11):63–103, 1998.

[Ott99] F. Otto. Evolution of microstructure in unstable porous media flow: a relaxational approach. *Communications on Pure and Applied Mathematics*, 52(7):873–915, 1999.

[PP10] J. W. Portegies and M. A. Peletier. Well-posedness of a parabolic moving-boundary problem in the setting of Wasserstein gradient flows. *Interfaces and Free Boundaries*, 12(2):121–150, 2010.

[PR11] M. Peletier and M. Renger. Variational formulation of the fokker-planck equation with decay: a particle approach. *Arxiv preprint arXiv:1108.3181*, 2011.

[Roc72] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1972.

[RS06] R. Rossi and G. Savaré. Gradient flows of non convex functionals in Hilbert spaces and applications. *ESAIM: Control, Optimization and Calculus of Variations*, 12:564–614, 2006.

[Vil03] C. Villani. *Topics in Optimal Transportation*. American Mathematical Society, 2003.