Electron. Commun. Probab. **25** (2020), article no. 52, 1–14.
https://doi.org/10.1214/20-ECP330
ISSN: 1083-589X

ELECTRONIC COMMUNICATIONS in PROBABILITY

On the long-time behaviour of McKean-Vlasov paths*

K. Bashiri†

Abstract

It is well-known that, in a certain parameter regime, the so-called *McKean-Vlasov evolution* $(\mu_t)_{t\in[0,\infty)}$ admits exactly three *stationary states*. In this paper we study the *long-time behaviour* of the flow $(\mu_t)_{t\in[0,\infty)}$ in this regime. The main result is that, for any initial measure μ_0, the flow $(\mu_t)_{t\in[0,\infty)}$ converges to a stationary state as $t\to\infty$ (see Theorem 1.2). Moreover, we show that if the energy of the initial measure is below some critical threshold, then the limiting stationary state can be identified (see Proposition 1.3). Finally, we also show some topological properties of the basins of attraction of the McKean-Vlasov evolution (see Proposition 1.4). The proofs are based on the representation of $(\mu_t)_{t\in[0,\infty)}$ as a *Wasserstein gradient flow*.

Some results of this paper are not entirely new. The main contribution here is to show that the *Wasserstein framework* provides short and elegant proofs for these results. However, up to the author's best knowledge, the statement on the topological properties of the basins of attraction (Proposition 1.4) is a new result.

Keywords: Wasserstein gradient flows; McKean-Vlasov evolution; ergodicity; basin of attraction.
MSC2020 subject classifications: 35B40; 35B30; 35K55; 49J40; 60G10.
Submitted to ECP on January 24, 2020, final version accepted on June 26, 2020.
Supersedes arXiv:2001.09106.

1 Introduction

In this paper we study the *ergodicity* and the *energy landscape* of the flow $(\mu_t)_{t\in[0,\infty)}$ of marginal laws associated to the stochastic differential equation given by

$$dx_t \;=\; -\Psi'(x_t)\,dt + J\int_{\mathbb{R}} z\,d\mu_t(z)\,dt + \sqrt{2}\,dB_t. \tag{1.1}$$

Here, the *single-site potential* $\Psi:\mathbb{R}\to\mathbb{R}$ and the *interaction strength* $J\in\mathbb{R}$ satisfy Assumption 1.1 below, and B is a one-dimensional Brownian motion. This flow $(\mu_t)_{t\in[0,\infty)}$ is often called *McKean-Vlasov evolution* in the literature.

In order to understand the main motivation for this paper, we recall five well-known facts.

*K.B. is partially supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 211504053 – SFB 1060 and by Germany's Excellence Strategy – GZ 2047/1, projekt-id 390685813 – "Hausdorff Center for Mathematics" at Bonn University.

†Institut für Angewandte Mathematik, Rheinische Friedrich-Wilhelms-Universität, Endenicher Allee 60, 53115 Bonn, Germany. E-mail: bashiri@iam.uni-bonn.de

(i) Let $(\mathcal{P}_2(\mathbb{R}), W_2)$ be the *Wasserstein space*; see Section 2.1 below. Then, we know from [1, Chapter 11] and [6] that $(\mu_t)_{t\in[0,\infty)}$ can be represented as a so-called *Wasserstein gradient flow* (again see Section 2.1) for the functional $\mathcal{F} : \mathcal{P}_2(\mathbb{R}) \to (-\infty, \infty]$, which is defined by

$$\mathcal{F}(\mu) \;=\; \int_{\mathbb{R}} \log(\rho) d\mu + \int_{\mathbb{R}} \Psi \, d\mu - \frac{J}{2} \left(\int_{\mathbb{R}} z \, d\mu(z) \right)^2 \tag{1.2}$$

if $\mu \in \mathcal{P}_2(\mathbb{R})$ has a Lebesgue density ρ, and $\mathcal{F}(\mu) = \infty$ otherwise. Moreover, in [1, 11.2.8] it is shown that for all $\mu \in \overline{D(\mathcal{F})} = \mathcal{P}_2(\mathbb{R})$ (where $D(\mathcal{F}) = \{\mu \in \mathcal{P}_2(\mathbb{R}) \,|\, \mathcal{F}(\mu) < \infty\}$), there exists a unique Wasserstein gradient flow for $\mathcal{F}$ with initial value μ. In this paper, we denote this gradient flow by $(S[\mu](t))_{t\in(0,\infty)}$.

(ii) Consider the system of $N \in \mathbb{N}$ *mean-field interacting diffusions* given by

$$dx_i^N(t) = -\Psi'\left(x_i^N(t)\right) dt + \frac{J}{N} \sum_{j=0}^{N-1} x_j^N(t) \, dt + \sqrt{2} \, dB_i(t) \quad \text{for } 0 \le i \le N-1, \tag{1.3}$$

where $B^N = (B_i)_{i=0,\dots,N-1}$ is an N-dimensional Brownian motion. Let $(L^N(t))_{t\in[0,\infty)}$ denote the corresponding *empirical distribution process*, i.e.,

$$L^N(t) \;=\; \frac{1}{N} \sum_{i=0}^{N-1} \delta_{x_i^N(t)} \quad \text{for all } t \in [0, \infty). \tag{1.4}$$

Then, ever since the classic papers [8] and [9], it is known that the process $(L^N(t))_{t\in[0,\infty)}$ converges weakly to the deterministic McKean-Vlasov evolution as $N \to \infty$.

(iii) Already in the paper [7] it was conjectured that the process $(L^N(t))_{t\in[0,\infty)}$ exhibits *metastable behaviour*[1]. It is a long outstanding problem to verify this conjecture rigorously. Although some progress in this direction was established in the paper [4], there are still many open and challenging questions.

(iv) It is well-known that, in order to analyse the metastable behaviour of a stochastic system, it is essential to have deep knowledge on the underlying *energy landscape* of the system and its *ergodicity*, i.e., its possible convergence towards stationary measures.

(v) In order to study curves and other objects that belong to the infinite-dimensional *space of probability measures*, the *Wasserstein formalism* provides a natural and convenient framework. Indeed, ever since the seminal papers [12] and [15], it is known that the Wasserstein formalism provides the structure of a *Riemannian manifold* on the space of probability measures. We refer to [2, p. 421] or [4, Section 1.4] for more arguments that speak in favour of the Wasserstein formalism.

We now formulate the main motivation for this paper. Combining the facts (ii), (iii) and (iv), we see that, in order to understand the metastable behaviour of $(L^N(t))_{t\in[0,\infty)}$, it is essential to study the *ergodicity* and the *energy landscape* of the McKean-Vlasov evolution. Moreover, from fact (i) we see that the energy landscape associated to $(\mu_t)_{t\in[0,\infty)}$ is determined by the functional $\mathcal{F}$ and its *basins of attraction*; see Proposition 1.4 for the precise definition of the latter. This is the main motivation why we study the ergodicity of $(\mu_t)_{t\in[0,\infty)}$ and the basins of attraction of $\mathcal{F}$. Finally, fact (v) explains why we use the Wasserstein setting as the framework for this paper.

We make the following assumptions throughout this paper.

[1] We refer the reader with no background in metastability to the monumental monographs in this subject given by [5] and [14].

Assumption 1.1. *(1) There is a splitting $\Psi = \Psi_c + \Psi_b$ for some $\Psi_c, \Psi_b \in C^2(\mathbb{R})$, and there are constants $0 < c, c' < \infty$ such that $\Psi_c'' \geq c$ and $|\Psi_b| + |\Psi_b'| + |\Psi_b''| \leq c'$ on $\mathbb{R}$.*

(2) There exist $\epsilon, c'' \in (0,\infty)$ such that $\Psi(z) \geq c''(|z|^{2+\epsilon} - 1)$ for all $z \in \mathbb{R}$.

(3) $\Psi(z) = \Psi(-z)$ for all $z \in \mathbb{R}$.

(4) $1/J < \int_{\mathbb{R}} z^2 \, \mathrm{e}^{-\Psi(z)} \, dz / (\int \mathrm{e}^{-\Psi(z)} \, dz)$.

(5) $z \mapsto \Psi'(z)$ is convex on $[0,\infty)$.

In particular, Assumption 1.1 is fulfilled if Ψ is a polynomial of degree 2ℓ for some $\ell \in \mathbb{N} \cap [2,\infty)$ such that Assumption 1.1 (4) and Assumption 1.1 (5) are satisfied. In Section 6 we briefly discuss the assumptions we make in this paper.

An important observation in Lemma 2.5 is that, as an immediate consequence of Assumption 1.1, the system (1.1) admits exactly three *stationary points* at some measures $\mu^-, \mu^0, \mu^+ \in \mathcal{P}_2(\mathbb{R})$, which are defined in (2.18); see Lemma 2.5 for more details. We also mention here that, as we will see in Lemma 2.4, the measures μ^- and μ^+ are the *global minimizers* of the functional $\mathcal{F}$.

We now formulate the main result of this paper in the following theorem.

Theorem 1.2. *Suppose Assumption 1.1. Let $\mu \in \mathcal{P}_2(\mathbb{R})$. Then, there exists $\mu^* \in \{\mu^-, \mu^0, \mu^+\}$ such that*

$$\lim_{t\to\infty} W_2(S[\mu](t), \mu^*) \;=\; 0 \quad \textit{and} \quad \lim_{t\to\infty} \mathcal{F}(S[\mu](t)) = \mathcal{F}(\mu^*). \tag{1.5}$$

Proof. The proof is postponed to Section 5. □

As a by-product of the proof of Theorem 1.2, we obtain the following two propositions, which are interesting on their own. The first one shows that inside the valleys of the set $\{\mu \in \mathcal{P}_2(\mathbb{R}) \,|\, \mathcal{F}(\mu) \leq \mathcal{F}(\mu^0)\}$ the convergence of the gradient flows for $\mathcal{F}$ is determined by the sign of the mean of the initial value.

Proposition 1.3. *Suppose Assumption 1.1. Let $\mu \in \mathcal{P}_2(\mathbb{R})$ be such that $\int_{\mathbb{R}} z \, d\mu(z) \neq 0$ and $\mathcal{F}(\mu) \leq \mathcal{F}(\mu^0)$. Then,*

$$\lim_{t\to\infty} \mathcal{F}(S[\mu](t)) \;=\; \mathcal{F}(\mu^-) \;=\; \mathcal{F}(\mu^+), \tag{1.6}$$

and

$$\lim_{t\to\infty} W_2(S[\mu](t), \mu^-) \;=\; 0 \quad \textit{if} \int_{\mathbb{R}} z \, d\mu(z) < 0 \quad \textit{and} \tag{1.7}$$

$$\lim_{t\to\infty} W_2(S[\mu](t), \mu^+) \;=\; 0 \quad \textit{if} \int_{\mathbb{R}} z \, d\mu(z) > 0. \tag{1.8}$$

Proof. The proof is postponed to Section 3. □

The second by-product is the following proposition, which provides useful informations on the energy landscape determined by $\mathcal{F}$.

Proposition 1.4. *Suppose Assumption 1.1. Let $\mathcal{B}^-$, $\mathcal{B}^0$ and $\mathcal{B}^-$ be the* basins of attraction *of the stationary measures μ^-, μ^0 and μ^+, respectively. That is,*

$$\begin{aligned} \mathcal{B}^- &= \{\mu \in \mathcal{P}_2(\mathbb{R}) \mid \lim_{t\to\infty} W_2(S[\mu](t), \mu^-) = 0\}, \\ \mathcal{B}^+ &= \{\mu \in \mathcal{P}_2(\mathbb{R}) \mid \lim_{t\to\infty} W_2(S[\mu](t), \mu^+) = 0\}, \quad \textit{and} \\ \mathcal{B}^0 &= \{\mu \in \mathcal{P}_2(\mathbb{R}) \mid \lim_{t\to\infty} W_2(S[\mu](t), \mu^0) = 0\}. \end{aligned} \tag{1.9}$$

Then, $\mathcal{B}^-$ and $\mathcal{B}^+$ are open subsets of the metric space[2] $(\mathcal{P}_2(\mathbb{R}), W_2)$, and $\mathcal{B}^0$ is a closed subset of $(\mathcal{P}_2(\mathbb{R}), W_2)$.

[2] It is shown in [17, 6.18] that $(\mathcal{P}_2(\mathbb{R}), W_2)$ is even a *Polish space*, i.e. a complete separable metric space.

Proof. The claim follows from Proposition 4.1 and Corollary 5.1 below. □

The results of this paper are not completely new. Indeed, Theorem 1.2 and Proposition 1.3 have already been obtained in the paper [16][3]. The proofs in [16] are based on methods from the theory of partial differential equations. The main contributions of this paper are that we use the Wasserstein framework to prove these results (which provides shorter proofs than in [16]), and that the results hold in the stronger topology of the Wasserstein distance (whereas the results in [16] are formulated in terms of the weak topology). However, to our knowledge, Proposition 1.4 is a new result. It is expected that this proposition will become useful in the study of the metastable behaviour of the system (1.3) via the Wasserstein framework. The latter is left for future research.

This paper is organized as follows. First, we recall some elements of the construction of *Wasserstein gradient flows* in Section 2.1. Then, in Section 2.2, we compare $\mathcal{F}$ with the functional $\bar{H}$, which appeared in [4]. In Section 2.3 we characterize the stationary measures, and in Section 2.4 we show a useful symmetry property of the McKean-Vlasov evolution. In Chapter 3 we first show some compactness property of the gradient flows for $\mathcal{F}$, and then use this property to prove Proposition 1.3. In Chapter 4 we prove the main part of Proposition 1.4. In Chapter 5 we provide the proof of Theorem 1.2 and state some immediate consequences of this theorem for the set $\mathcal{B}^0$. Finally, in Section 6 we briefly discuss the assumptions we make in this paper.

2 Preliminaries

2.1 Wasserstein gradient flows

In this section, we briefly recall some elements of the construction of *Wasserstein gradient flows*. For simplicity, we restrict all definitions to the functional $\mathcal{F}$ from (1.2). For more general functionals and for the details, we refer to [1].

Let $\mathcal{P}_2(\mathbb{R})$ denote the space of all probability measures on $\mathbb{R}$, whose second moment is finite. We equip $\mathcal{P}_2(\mathbb{R})$ with the *Wasserstein distance* W_2, which, for $\mu,\nu\in\mathcal{P}_2(\mathbb{R})$ is defined by

$$W_2(\mu,\nu)^2 := \inf_{\gamma\in\mathrm{Cpl}(\mu,\nu)}\int_{\mathbb{R}^2}|y-y'|^2\,d\gamma(y,y'), \tag{2.1}$$

where $\mathrm{Cpl}(\mu,\nu)$ denotes the space of all probability measures on $\mathbb{R}^2$ that have μ and ν as marginals.

Let $(\mu_t)_{t\in[0,\infty)}$ be a curve of probability measures such that $\mu_t\in\mathcal{P}_2(\mathbb{R})$ for all $t\in[0,\infty)$. Then we say that $(\mu_t)_{t\in[0,\infty)}$ is *absolutely continuous* if there exists $m\in\mathrm{L}^2_{\mathrm{loc}}((0,\infty))$ such that

$$W_2(\mu_s,\mu_t) \;\le\; \int_s^t m(r)\,dr \quad \text{for all } 0<s<t<\infty. \tag{2.2}$$

We denote the set of all absolutely continuous curves in $(\mathcal{P}_2(\mathbb{R}),W_2)$ by $\mathcal{AC}((0,\infty);\mathcal{P}_2(\mathbb{R}))$. It is shown in [1, 1.1.2] that for all $(\mu_t)_{t\in[0,\infty)}\in\mathcal{AC}((0,\infty);\mathcal{P}_2(\mathbb{R}))$, there exists $|\mu'|\in\mathrm{L}^2_{\mathrm{loc}}((0,\infty))$, called the *metric derivative* of $(\mu_t)_{t\in[0,\infty)}$, such that

$$|\mu'|(t) \;=\; \lim_{s\to t}\frac{W_2(\mu_s,\mu_t)}{|s-t|} \quad \text{for almost every } t\in(0,\infty). \tag{2.3}$$

[3]Note that the interaction term in [16] is of polynomial form, whereas in the present paper we restrict to the linear interaction from the system (1.1). Hence, the setting in [16] is more general than here.

Another important object is the *metric slope* (cf. [1, 1.2.4]) of $\mathcal{F}$, which is defined by

$$|\partial\mathcal{F}|(\mu) \ = \ \limsup_{\nu\to\mu}\left(\frac{\mathcal{F}(\mu)-\mathcal{F}(\nu)}{W_2(\mu,\nu)}\right)^+ \qquad \text{for } \mu\in D(\mathcal{F}), \tag{2.4}$$

and $|\partial\mathcal{F}|(\mu)=\infty$ for $\mu\in\mathcal{P}_2(\mathbb{R})\setminus D(\mathcal{F})$.

We are now in the position to define the notion of *Wasserstein gradient flows for* $\mathcal{F}$. There are several different and equivalent ways to do this; some of them are listed in [1, Chapter 11]. In this paper, we choose the definition as a *curve of maximal slope* (cf. [1, 1.3.2]).

Definition 2.1. *We say that a curve* $(S[\mu](t))_{t\in[0,\infty)}\in\mathcal{AC}((0,\infty);\mathcal{P}_2(\mathbb{R}))$ *is a* (Wasserstein) gradient flow for $\mathcal{F}$ *with initial value* $\mu\in\mathcal{P}_2(\mathbb{R})$ *if* $\lim_{t\downarrow 0}W_2(S[\mu](t),\mu)=0$*, and if the map* $t\mapsto\mathcal{F}(S[\mu](t))$ *is locally absolutely continuous in* $(0,\infty)$ *with*

$$\frac{d}{dt}\mathcal{F}(S[\mu](t)) = -|\partial\mathcal{F}|^2(S[\mu](t)) = -\left|(S[\mu])'\right|^2(t) \quad \textit{for almost every } t\in(0,\infty). \tag{2.5}$$

We conclude this section with some useful properties of $\mathcal{F}$ and Wasserstein gradient flows for $\mathcal{F}$, which we use many times in this paper.

Lemma 2.2. *Suppose Assumption 1.1. Then the following statements are true.*

(i) (Lower bound on $\mathcal{F}$) *There exists* $c>0$ *such that*

$$\mathcal{F}(\mu) \ \geq \ c\left(\int_{\mathbb{R}}|x|^{2+\epsilon}\,d\mu(x)-1\right) \qquad \textit{for all } \mu\in\mathcal{P}_2(\mathbb{R}). \tag{2.6}$$

(ii) (λ-convexity of $\mathcal{F}$) *There exists* $\lambda<0$ *such that* $\mathcal{F}$ *is* λ-convex along generalized geodesics *in the sense of [1, 4.0.1].*

(iii) (Existence) *For each* $\mu\in\mathcal{P}_2(\mathbb{R})$*, there exists a gradient flow* $(S[\mu](t))_{t\in[0,\infty)}$ *for* $\mathcal{F}$*.*

(iv) (Energy identity) *Let* $\mu\in D(\mathcal{F})$*. Then, for all* $t\in(0,\infty)$*,*

$$0 \ = \ \mathcal{F}(S[\mu](t))-\mathcal{F}(\mu)+\frac{1}{2}\int_0^t\left(|\partial\mathcal{F}|^2(S[\mu](r))+|(S[\mu])'|^2(r)\right)dr. \tag{2.7}$$

(v) (Regularization estimate) *Let* $\mu\in\mathcal{P}_2(\mathbb{R})$*. Then,*

$$\mathcal{F}(S[\mu](t)) \ \leq \ \mathcal{F}(\nu)+\frac{\lambda}{2(\mathrm{e}^{\lambda t}-1)}W_2(\mu,\nu)^2 \quad \textit{for all } \nu\in D(\mathcal{F}) \textit{ and } t\in(0,\infty). \tag{2.8}$$

(vi) (contraction and semigroup property) *Let* $\mu,\nu\in\mathcal{P}_2(\mathbb{R})$*. Then,*

$$W_2(S[\mu](t),S[\nu](t)) \ \leq \ \mathrm{e}^{-\lambda t}\,W_2(\mu,\nu) \quad \textit{for all } t\in(0,\infty). \tag{2.9}$$

In particular, the semigroup property $S[S[\mu](h)](t)=S[\mu](t+h)$ *holds for all* $t,h>0$*.*

Proof. Part *(ii)* is proven in [1, Section 9.3] or [4, 3.35], part *(iii)* in [1, 11.1.3 and 11.2.8], part *(iv)* in [1, 2.3.3 and 4.0.4], part *(v)* in [1, 4.3.2][4] and part *(vi)* is proven in [1, (11.2.2)].

It remains to show part *(i)*. Let $\mu\in D(\mathcal{F})$, since otherwise the claim is trivial. In the following let $C>0$ denote a constant which does not depend on μ, and may change from

[4]Note that there is a typo in [1, (4.3.2)]: It must be $\frac{\mathrm{e}^{\lambda T}-1}{\lambda}$ instead of $\frac{\mathrm{e}^{\lambda T}-1}{T}$.

line to line. We proceed as in the proof of [4, 3.34] and use Assumption 1.1 (2) to observe that

$$\begin{aligned}\mathcal{F}(\mu) &\geq -C + \frac{1}{4}\int_{\mathbb{R}^2} \Psi(x)\, d\mu(x) + \frac{1}{2}\int_{\mathbb{R}^2} \left(\frac{1}{4}\big(\Psi(x)+\Psi(\bar{x})\big) - Jx\bar{x}\right) d\mu(x)d\mu(\bar{x}) \\ &\geq -C + \frac{c''}{4}\int_{\mathbb{R}} |x|^{2+\epsilon}\, d\mu(x) + \frac{1}{2}\int_{\mathbb{R}^2} \left(\frac{c''}{4}\big(|x|^{2+\epsilon}+|\bar{x}|^{2+\epsilon}\big) - Jx\bar{x}\right) d\mu(x)d\mu(\bar{x}).\end{aligned} \tag{2.10}$$

Note that, as a consequence of the classic Young inequalities, for all $x, \bar{x} \in \mathbb{R}$ and all $\alpha > 0$, $|x\bar{x}| \leq |x|^2/2 + |\bar{x}|^2/2$ and $|x|^{2+\epsilon} \geq \alpha|x|^2 - C_\alpha$ for some constant $C_\alpha > 0$ (which only depends on α and ϵ). Then, by choosing α large enough, we can show that the last term on the right-hand side of (2.10) is greater or equal to $-C_\alpha \frac{c''}{4}$. This concludes the proof. □

2.2 Macroscopic Hamiltonians

In this section we first introduce and recall some facts about the function $\bar{H} : \mathbb{R} \to \mathbb{R}$, which was the object of investigation in the paper [4]. Then, in Lemma 2.4, we show the relation between $\mathcal{F}$ and $\bar{H}$, and infer from that useful analytic facts about $\mathcal{F}$.

Let the function $\varphi^* : \mathbb{R} \to \mathbb{R}$ be defined by

$$\varphi^*(\sigma) \;=\; \log \int_{\mathbb{R}} \mathrm{e}^{\sigma z - \Psi(z)}\, dz \quad \text{for } \sigma \in \mathbb{R}. \tag{2.11}$$

Let $\varphi : \mathbb{R} \to \mathbb{R}$ be the *Legendre transform* of φ^*, i.e.,

$$\varphi(m) \;=\; \sup_{\sigma\in\mathbb{R}} \left(\sigma m - \varphi^*(\sigma)\right) \quad \text{for } m \in \mathbb{R}. \tag{2.12}$$

It is then well-known from standard properties of Legendre transforms (see for instance [13, III.2.5], [4, A.1.1] or [10, Lemma 41]) that for all $m, \sigma \in \mathbb{R}$,

$$\varphi'(m)m - \varphi^*(\varphi'(m)) = \varphi(m), \qquad (\varphi^*)'(\varphi'(m)) = m \quad \text{and} \quad (\varphi^*)'(\sigma) = \int_{\mathbb{R}} z\, d\mu^\sigma, \tag{2.13}$$

where, for $\sigma \in \mathbb{R}$, the probability measure $\mu^\sigma \in \mathcal{P}_2(\mathbb{R})$ is defined by

$$d\mu^\sigma(z) \;=\; \mathrm{e}^{-\varphi^*(\sigma)+\sigma z - \Psi(z)}\, dz \;=\; \frac{\mathrm{e}^{\sigma z - \Psi(z)}}{\int_{\mathbb{R}} \mathrm{e}^{\sigma\bar{z} - \Psi(\bar{z})}\, d\bar{z}}\, dz. \tag{2.14}$$

Finally, we define the function $\bar{H} : \mathbb{R} \to \mathbb{R}$ by

$$\bar{H}(z) \;=\; \varphi(z) \;-\; \frac{J}{2}z^2 \quad \text{for } z \in \mathbb{R}. \tag{2.15}$$

Remark 2.3. The function $\bar{H}$ played the role of the *macroscopic Hamiltonian* in [4], where the metastable behaviour of the system (1.3) was studied. It is important to notice that in [4] the *empirical mean* was chosen to be the *macroscopic order parameter*. Recall from fact (i) and (ii) of the introduction that the functional $\mathcal{F}$ appears as the macroscopic Hamiltonian of the system (1.3) by choosing the *empirical distribution* as the macroscopic order parameter; see [4, Section 1.4] for more details on this.

Moreover, as it is shown in [4, 3.4], under Assumption 1.1, the function $\bar{H}$ admits exactly three critical points, which are located at $-m^\star$, 0 and $m^\star$ for some $m^\star > 0$. Furthermore, $\bar{H}''(0) < 0$, $\bar{H}''(m^\star) = \bar{H}''(-m^\star) > 0$, and $\bar{H}(0) > \bar{H}(m^\star) = \bar{H}(-m^\star)$. That is, $\bar{H}$ has a local maximum at 0, and the two global minima of $\bar{H}$ are located at $\pm m^\star$.

In the following let $\mathrm{m}[\mu] = \int_{\mathbb{R}} z d\mu(z)$ denote the mean of a probability measure $\mu \in \mathcal{P}_2(\mathbb{R})$. We have the following relation[5] between the macroscopic Hamiltonians $\mathcal{F}$ and $\bar{H}$.

Lemma 2.4. *Suppose Assumption 1.1. Then, for all $m \in \mathbb{R}$, we have that*

$$\mathcal{F}(\mu) \;>\; \mathcal{F}(\mu^{\varphi'(m)}) \quad \text{for all } \mu \in \mathcal{P}_2(\mathbb{R}) \text{ such that } \mathrm{m}[\mu] = m \text{ and } \mu \neq \mu^{\varphi'(m)}, \tag{2.16}$$

and,

$$\bar{H}(m) \;=\; \min_{\mu \in \mathcal{P}_2(\mathbb{R}), \mathrm{m}[\mu]=m} \mathcal{F}(\mu) \;=\; \mathcal{F}(\mu^{\varphi'(m)}). \tag{2.17}$$

Moreover, let

$$\mu^- := \mu^{\varphi'(-m^\star)}, \qquad \mu^0 := \mu^{\varphi'(0)} \quad \text{and} \quad \mu^+ := \mu^{\varphi'(m^\star)}. \tag{2.18}$$

Then, $\mathcal{F}$ admits exactly two global minima, one at μ^- and one at μ^+, and we have that $\mathcal{F}(\mu^-) = \mathcal{F}(\mu^+) < \mathcal{F}(\mu^0)$.

Proof. If $\mathcal{F}(\mu) = \infty$, then (2.16) is trivially satisfied. So we assume that $\mathcal{F}(\mu) < \infty$. In the following let $\mathcal{H}(\cdot|\cdot)$ denote the *relative entropy functional* (see e.g. [1, 9.4.1]), and let $\mathrm{m}[\mu] = m$. Then, by using (2.13) and by denoting the Lebesgue density of μ by ρ,

$$\begin{aligned} \mathcal{F}(\mu) \;&=\; \int_{\mathbb{R}} \log(\rho\, \mathrm{e}^{\Psi}) d\mu - \frac{J}{2} m^2 \;=\; \mathcal{H}(\mu \,|\, \mu^{\varphi'(m)}) + \varphi'(m) m - \varphi^*(\varphi'(m)) - \frac{J}{2} m^2 \\ &=\; \mathcal{H}(\mu \,|\, \mu^{\varphi'(m)}) + \bar{H}(m). \end{aligned} \tag{2.19}$$

Since $\mathcal{H}(\mu^{\varphi'(m)} \,|\, \mu^{\varphi'(m)}) = 0$ and $\mathcal{H}(\mu \,|\, \mu^{\varphi'(m)}) > 0$ if $\mu \neq \mu^{\varphi'(m)}$, (2.19) implies that

$$\mathcal{F}(\mu) > \bar{H}(m) \text{ if } \mu \neq \mu^{\varphi'(m)} \quad \text{and} \quad \mathcal{F}(\mu^{\varphi'(m)}) \;=\; \bar{H}(m). \tag{2.20}$$

From (2.20) we immediately infer (2.16) and (2.17). Finally, (2.17) and Remark 2.3 imply the last two claims. □

2.3 Stationary points of the McKean-Vlasov evolution

In this section we characterize the *stationary points* of the McKean-Vlasov evolution[6], where we say that $\mu \in \mathcal{P}_2(\mathbb{R})$ is *stationary* if

$$S[\mu](t) = \mu \quad \text{for all } t \in (0, \infty), \tag{2.21}$$

or equivalently,

$$|(S[\mu])'|(t) \;=\; 0 \quad \text{for almost every } t \in (0, \infty). \tag{2.22}$$

Lemma 2.5. *Suppose Assumption 1.1. Let $\mu \in \mathcal{P}_2(\mathbb{R})$. Then, the following statements are equivalent.*

(i) *μ is stationary.*

(ii) $|\partial \mathcal{F}|(\mu) = 0$.

(iii) $\mu \in \{\mu^-, \mu^0, \mu^+\}$.

[5] See also [13, Section IV.2] for a more general result.
[6] See also [11] for similar results.

Proof. $(i) \Rightarrow (ii)$. Suppose that μ is stationary. Then, combining (2.5) (which holds true even if $\mu \notin D(\mathcal{F})$) and (2.22), we infer that $|\partial\mathcal{F}|(S[\mu](t)) = |(S[\mu])'|(t) = 0$ for almost every $t \in (0,\infty)$. Then, by the lower semi-continuity of $|\partial\mathcal{F}|$ (see [1, 2.4.10]) and the fact that $\lim_{t\downarrow 0} W_2(\mu, S[\mu](t)) = 0$, we conclude that $|\partial\mathcal{F}|(\mu) \;\leq\; \liminf_{t\downarrow 0} |\partial\mathcal{F}|(S[\mu](t)) \;=\; 0.$

$(ii) \Rightarrow (iii)$. Using [1, 10.4.13], we have that the Lebesgue density ρ of μ belongs to the Sobolev space $W^{1,1}_{loc}(\mathbb{R})$[7]. Let $m = \mathrm{m}[\mu]$. Then, by using again [1, 10.4.13],

$$|\partial\mathcal{F}|^2(\mu) = \int_{\mathbb{R}} \left| \frac{\partial_z \rho(z)}{\rho(z)} + \Psi'(z) - Jm \right|^2 d\mu(z) = \int_{\mathbb{R}} \left| \frac{\partial_z \left(\rho(z) \mathrm{e}^{\Psi(z) - Jmz}\right)}{\rho(z)\mathrm{e}^{\Psi(z)-Jmz}} \right|^2 d\mu(z). \tag{2.23}$$

Since $|\partial\mathcal{F}|(\mu) = 0$, (2.23) implies that the map $z \mapsto \rho(z)\mathrm{e}^{\Psi(z)-Jmz}$ is constant μ-almost everywhere. Therefore, for μ-a.e. $z, z' \in \mathbb{R}$,

$$\rho(z) \;=\; \rho(z')\, \mathrm{e}^{\Psi(z') - Jmz'}\, \mathrm{e}^{-\Psi(z)+Jmz}. \tag{2.24}$$

By fixing $z' \in \mathbb{R}$ and by using the definition of φ^* and that $\int_{\mathbb{R}} \rho(z)\, dz = 1$, (2.24) implies that

$$\rho(z) \;=\; \mathrm{e}^{-\varphi^*(Jm)}\, \mathrm{e}^{-\Psi(z)+Jmz}. \tag{2.25}$$

In particular, combining (2.13) and (2.25) yields that $m = (\varphi^*)'(Jm)$. And by using the second claim in (2.13), we infer that $\bar{H}'(m) = 0$. However, in Remark 2.3 we have seen that there are only three solutions to this equation. This implies that

$$m \;\in\; \{-m^\star, 0, m^\star\}. \tag{2.26}$$

Combining (2.25) and (2.26) yields part *(iii)*.

$(iii) \Rightarrow (ii)$. Combining the representation (2.23) with the definition of the measures μ^-, μ^0 and μ^+ yields part *(ii)*.

$(ii) \Rightarrow (i)$. From [1, 2.4.15 and 2.4.16], we have that for all $t > 0$,

$$|\partial\mathcal{F}|(S[\mu](t)) \;\leq\; \mathrm{e}^{-\lambda t}|\partial\mathcal{F}|(\mu) \;=\; 0, \tag{2.27}$$

where the parameter λ was introduced in Lemma 2.2. Again, using that $|(S[\mu])'|(t) = |\partial\mathcal{F}|(S[\mu](t))$ for almost every $t \in (0,\infty)$, (2.27) yields part *(i)*. □

2.4 Symmetry property

In this section we show that gradient flows for $\mathcal{F}$ admit a useful symmetry property. In the following we denote by $f_\# \mu$ the image measure of a measure μ under a Borel map f.

Lemma 2.6. *Let $\varsigma : \mathbb{R} \to \mathbb{R}$ be defined by $\varsigma(z) = -z$, and let $\mu \in \mathcal{P}_2(\mathbb{R})$. Then,*

$$S[\varsigma_\#\mu](t) = \varsigma_\# S[\mu](t) \quad \textit{for all } t \in (0,\infty). \tag{2.28}$$

Proof. First note that

$$\mathcal{F}(\nu) = \mathcal{F}(\varsigma_\#\nu) \quad \text{for all } \nu \in \mathcal{P}_2(\mathbb{R}), \tag{2.29}$$

and therefore,

$$|\partial\mathcal{F}|(\nu) = |\partial\mathcal{F}|(\varsigma_\#\nu) \quad \text{for all } \nu \in \mathcal{P}_2(\mathbb{R}). \tag{2.30}$$

[7] More precisely, in [1, 10.4.13] it is shown that $L_F(\rho) \in W^{1,1}_{loc}(\mathbb{R})$ for some function $L_F : [0,\infty) \to [0,\infty)$, which is defined right after [1, (10.4.17)]. However, in our case we have that $L_F(z) = z$ for all $z \in [0,\infty)$.

Moreover, for all $\nu \in \mathcal{AC}((0,\infty);\mathcal{P}_2(\mathbb{R}))$ and $0 < s < t < \infty$,

$$W_2(\nu_s,\nu_t) \;=\; W_2(\varsigma_\#(\varsigma_\#\nu_s),\varsigma_\#(\varsigma_\#\nu_t)) \;\leq\; W_2(\varsigma_\#\nu_s,\varsigma_\#\nu_t) \;\leq\; W_2(\nu_s,\nu_t). \tag{2.31}$$

Therefore, $W_2(\nu_s,\nu_t) = W_2(\varsigma_\#\nu_s,\varsigma_\#\nu_t)$, and we have that the metric derivatives coincide, i.e.,

$$|\nu'|(t) = |(\varsigma_\#\nu)'|(t) \text{ for almost every } t \in (0,\infty) \text{ and for all } \nu \in \mathcal{AC}((0,\infty);\mathcal{P}_2(\mathbb{R})). \tag{2.32}$$

Then, by combining (2.5) (which holds true even if $\mu \notin D(\mathcal{F})$), (2.29), (2.30) and (2.32), we have that for almost every $t \in (0,\infty)$,

$$\begin{aligned} \frac{d}{dt}\mathcal{F}(\varsigma_\# S[\mu](t)) \;&=\; \frac{d}{dt}\mathcal{F}(S[\mu](t)) = -\left|(S[\mu])'\right|^2(t) = -\left|(\varsigma_\# S[\mu])'\right|^2(t), \quad \text{and} \\ \frac{d}{dt}\mathcal{F}(\varsigma_\# S[\mu](t)) \;&=\; \frac{d}{dt}\mathcal{F}(S[\mu](t)) = -|\partial\mathcal{F}|^2(S[\mu](t)) = -|\partial\mathcal{F}|^2(\varsigma_\# S[\mu](t)). \end{aligned} \tag{2.33}$$

Moreover, by using the same arguments as in (2.31), we infer that

$$\lim_{t\downarrow 0} W_2(\varsigma_\# S[\mu](t),\varsigma_\#\mu) \;=\; \lim_{t\downarrow 0} W_2(S[\mu](t),\mu) \;=\; 0. \tag{2.34}$$

Combining (2.33) and (2.34) yields that the curve $(\varsigma_\# S[\mu](t))_{t\in(0,\infty)}$ is the gradient flow for $\mathcal{F}$ with initial value $\varsigma_\#\mu$. □

3 Convergence in the valleys

In this chapter we first show some compactness property of the McKean-Vlasov paths in Lemma 3.1. Then, we use this result to prove Proposition 1.3.

Lemma 3.1. *Suppose Assumption 1.1. Let $\mu \in D(\mathcal{F})$. Then, there exist a sequence $(t_k)_k$ and $\mu^* \in \{\mu^-,\mu^0,\mu^+\}$ such that $\lim_{k\to\infty} t_k = \infty$,*

$$\lim_{k\to\infty} W_2(S[\mu](t_k),\mu^*) \;=\; 0 \quad \textit{and} \quad \lim_{t\to\infty} \mathcal{F}(S[\mu](t)) \;=\; \mathcal{F}(\mu^*). \tag{3.1}$$

Proof. In the following let $\mu_t = S[\mu](t)$. We prove this lemma in three steps.

Step 1. [There exists a subsequence $(t_n)_n$ such that $\lim_{n\to\infty} |\partial\mathcal{F}|(\mu_{t_n}) = 0$.]
Note that the sequence $(\mathcal{F}(\mu_t))_{t\in[0,\infty)}$ is a continuous, monotone and bounded sequence of real numbers by (2.5) and (2.6). Therefore, it converges, as $t \to \infty$, to a number $L^* \in \mathbb{R}$. In particular, by (2.5),

$$\int_0^\infty |\partial\mathcal{F}|^2(\mu_r)\,dr \;=\; -\int_0^\infty \frac{d}{dr}\mathcal{F}(\mu_r)\,dr \;=\; -L^* + \mathcal{F}(\mu) \;<\; \infty. \tag{3.2}$$

This implies the claim of Step 1.

Step 2. [$\lim_{k\to\infty} W_2(\mu_{t_{n_k}},\mu^*)$ for some $\mu^* \in \{\mu^-,\mu^0,\mu^+\}$ and a subsubsequence $(t_{n_k})_k$.]
By (2.6), the monotonicity of $t \mapsto \mathcal{F}(\mu_t)$ and the fact that $\mu_0 = \mu \in D(\mathcal{F})$, we have that

$$\sup_{n\in\mathbb{N}} \int_{\mathbb{R}} |x|^{2+\epsilon}\,d\mu_{t_n}(x) \;\leq\; \sup_{n\in\mathbb{N}} \left(\frac{1}{c}\mathcal{F}(\mu_{t_n}) + 1\right) \;\leq\; \frac{1}{c}\mathcal{F}(\mu) + 1 \;<\; \infty. \tag{3.3}$$

Using [17, 6.8 (iii)], this implies that there exist a further subsequence $(t_{n_k})_k$ and $\mu^* \in \mathcal{P}_2(\mathbb{R})$ such that $\lim_{k\to\infty} W_2(\mu_{t_{n_k}},\mu^*)$. It remains to show that $\mu^* \in \{\mu^-,\mu^0,\mu^+\}$. In order to do this, we use the lower semi-continuity of $|\partial\mathcal{F}|$ ([1, 2.4.10]) and Step 1 to observe that

$$|\partial\mathcal{F}|(\mu^*) \;\leq\; \liminf_{k\to\infty} |\partial\mathcal{F}|(\mu_{t_{n_k}}) \;=\; 0. \tag{3.4}$$

Combining this with Lemma 2.5 yields the claim of Step 2.

Step 3. $[\lim_{t\to\infty}\mathcal{F}(\mu_t)=\mathcal{F}(\mu^*).]$

First note that by the lower semi-continuity of $\mathcal{F}$ (see [3, 3.35] or [1, Section 9.3]), we have that

$$L^* = \lim_{t\to\infty}\mathcal{F}(\mu_t) = \lim_{k\to\infty}\mathcal{F}(\mu_{t_{n_k}}) \geq \mathcal{F}(\mu^*). \tag{3.5}$$

To show the other inequality, we use [1, 2.4.9], and observe that for all $k\in\mathbb{N}$,

$$|\partial\mathcal{F}|(\mu_{t_{n_k}}) \;\geq\; \left(\frac{\mathcal{F}(\mu_{t_{n_k}})-\mathcal{F}(\mu^*)}{W_2(\mu_{t_{n_k}},\mu^*)} + \frac{\lambda}{2}W_2(\mu_{t_{n_k}},\mu^*)\right)^+, \tag{3.6}$$

where the parameter λ was introduced in Lemma 2.2. Note that (3.6) is equivalent to

$$W_2(\mu_{t_{n_k}},\mu^*)\,|\partial\mathcal{F}|(\mu_{t_{n_k}}) \;\geq\; \left(\mathcal{F}(\mu_{t_{n_k}})-\mathcal{F}(\mu^*) + \frac{\lambda}{2}W_2^2(\mu_{t_{n_k}},\mu^*)\right)^+. \tag{3.7}$$

Taking the limit as $k\to\infty$ on both sides, and using Step 1 and Step 2, implies that

$$0 \;\geq\; (L^*-\mathcal{F}(\mu^*))^+. \tag{3.8}$$

We conclude that $L^*\leq\mathcal{F}(\mu^*)$. □

With this compactness result in hand, we are able to prove Proposition 1.3.

Proof of Proposition 1.3. In the following let $\mu_t = S[\mu](t)$. It suffices to consider only the case that $\mathrm{m}[\mu]<0$. We know from Lemma 3.1 that there exists a subsequence $(\mu_{t_k})_k$ such that

$$\lim_{k\to\infty}W_2(\mu_{t_k},\mu^*)=0 \quad\text{and}\quad \lim_{t\to\infty}\mathcal{F}(\mu_t)=\mathcal{F}(\mu^*) \quad \text{for some } \mu^*\in\{\mu^-,\mu^0,\mu^+\}. \tag{3.9}$$

We first show that $\mu^*=\mu^-$ (which implies (1.6)), and then show that $\lim_{t\to\infty}W_2(\mu_t,\mu^-)=0$ (which implies (1.7)).

Step 1. $[\mu^*=\mu^-.]$

We show that the cases $\mu^*=\mu^+$ or $\mu^*=\mu^0$ lead to contradictions. First suppose that $\mu^*=\mu^+$. Since the map $t\mapsto\mathrm{m}[\mu_t]$ is continuous and since $\mathrm{m}[\mu_0]=\mathrm{m}[\mu]<0$, we have that there exists $t'\in(0,\infty)$ such that $\mathrm{m}[\mu_{t'}]=0$. Then, by the monotonicity of $t\mapsto\mathcal{F}(\mu_t)$ and by Lemma 2.4,

$$\mathcal{F}(\mu^0) \;\geq\; \mathcal{F}(\mu) \;\geq\; \mathcal{F}(\mu_{t'}) \;\geq\; \mathcal{F}(\mu^0). \tag{3.10}$$

Hence, $\mathcal{F}(\mu_{t'})=\mathcal{F}(\mu)=\mathcal{F}(\mu^0)$. Combining this with (2.7), implies that $\int_0^{t'}|\partial\mathcal{F}|^2(\mu_r)\,dr = 0$. This in turn yields that $|\partial\mathcal{F}|(\mu_r)=0$ for almost every $r\in(0,t')$. Then, by the lower semi-continuity of $|\partial\mathcal{F}|$ ([1, 2.4.10]), we infer that $|\partial\mathcal{F}|(\mu)\;\leq\;\liminf_{r\downarrow 0}|\partial\mathcal{F}|(\mu_r)\;=\;0$. Therefore, we have that

$$|\partial\mathcal{F}|(\mu)\;=\;0 \quad\text{and}\quad \mathcal{F}(\mu)\;=\;\mathcal{F}(\mu^0). \tag{3.11}$$

By Lemma 2.4 and Lemma 2.5, (3.11) implies that $\mu=\mu^0$. This yields to a contradiction, since $\mathrm{m}[\mu]<0$. The case $\mu^*=\mu^0$ is treated analogously.

Step 2. $[\lim_{t\to\infty}W_2(\mu_t,\mu^-)=0.]$

Let $(\mu_{s_n})_{n\in\mathbb{N}}$ be any subsequence of $(\mu_t)_{t\in[0,\infty)}$. Using the same compactness argument from Step 2 of the proof of Lemma 3.1, we know that there exists a further subsequence $(\mu_{s_{n_k}})_{k\in\mathbb{N}}$ such that $\lim_{k\to\infty}W_2(\mu_{s_{n_k}},\mu')=0$ for some $\mu'\in\mathcal{P}_2(\mathbb{R})$. In order to show the claim of Step 2, it remains to show that $\mu'=\mu^-$. First we notice that

$$\mathcal{F}(\mu') \;\leq\; \liminf_{k\to\infty}\mathcal{F}(\mu_{s_{n_k}}) \;=\; \lim_{t\to\infty}\mathcal{F}(\mu_t) \;=\; \mathcal{F}(\mu^-). \tag{3.12}$$

In view of Lemma 2.4, this implies that either $\mu' = \mu^-$ or $\mu' = \mu^+$. We now use similar arguments as in Step 1 to show that the latter case yields to a contradiction. So suppose that $\mu' = \mu^+$. Then, since $\lim_{k\to\infty} W_2(\mu_{s_{n_k}}, \mu^+) = 0$ and $\mathrm{m}[\mu^+] > 0$, there exists $s' \in (0,\infty)$ such that $\mathrm{m}[\mu_{s'}] > 0$. By the continuity of the map $t \mapsto \mathrm{m}[\mu_t]$ and since $\mathrm{m}[\mu_0] = \mathrm{m}[\mu] < 0$, there must be a $t' \in (0, s')$ such that $\mathrm{m}[\mu_{t'}] = 0$. Now we use the same arguments as in Step 1 to conclude (3.11). This in turn implies that $\mu = \mu^0$, which yields to a contradiction, since $\mathrm{m}[\mu] < 0$. This concludes the proof. □

4 Basin of attraction

Proposition 4.1. *Suppose Assumption 1.1. Recall the definition of $\mathcal{B}^-$ and $\mathcal{B}^+$ from* (1.9)*. Then, $\mathcal{B}^-$ and $\mathcal{B}^+$ are open subsets of $\mathcal{P}_2(\mathbb{R})$.*

Proof. In view of Lemma 2.6, it suffices to show the claim only for $\mathcal{B}^-$. In the following abbreviate $\Delta := \mathcal{F}(\mu^0) - \mathcal{F}(\mu^-)$ and recall the definition of $\lambda < 0$ from Lemma 2.2.

Let $\nu \in \mathcal{B}^-$. That is, $\nu \in \mathcal{P}_2(\mathbb{R})$ (in particular, it may be that $\nu \notin D(\mathcal{F})$) and we have that $\lim_{t\to\infty} W_2(S[\nu](t), \mu^-) = 0$. Let $h \in (0,\infty)$. Note that, by using (2.8) and the semigroup property (see Lemma 2.2), we have that

$$S[\nu](h) \in D(\mathcal{F}) \quad \text{and} \quad \lim_{t\to\infty} W_2(S[S[\nu](h)](t), \mu^-) = \lim_{t\to\infty} W_2(S[\nu](t+h), \mu^-) = 0. \tag{4.1}$$

Then, applying Lemma 3.1 with $\mu = S[\nu](h)$ implies that

$$\mathcal{F}(\mu^-) \;=\; \lim_{t\to\infty} \mathcal{F}(S[S[\nu](h)](t)) \;=\; \lim_{t\to\infty} \mathcal{F}(S[\nu](t+h)) \;=\; \lim_{t\to\infty} \mathcal{F}(S[\nu](t)). \tag{4.2}$$

Therefore, we have that

$$\lim_{t\to\infty} W_2(S[\nu](t), \mu^-) = 0 \quad \text{and} \quad \lim_{t\to\infty} \mathcal{F}(S[\nu](t)) \;=\; \mathcal{F}(\mu^-). \tag{4.3}$$

This implies that there exists some $t' > 0$ such that for all $t \geq t'$,

- $W_2(S[\nu](t), \mu^-) \;\leq\; \frac{1}{4} m^\star$,
- $\mathcal{F}(S[\nu](t)) \;\leq\; \mathcal{F}(\mu^-) + \frac{1}{4}\Delta$, and
- $\mathrm{e}^{\lambda t} \;=\; \mathrm{e}^{-|\lambda| t} \;\leq\; \frac{1}{2}$.

Set

$$\delta \;:=\; \min\left\{ \mathrm{e}^{2\lambda t'} \, \frac{m^\star}{4} \, , \, \sqrt{\mathrm{e}^{2\lambda t'} \, \frac{1}{|\lambda|} \, \frac{\Delta}{4}} \right\}. \tag{4.4}$$

We now show that $B_\delta(\nu) = \{\mu \in \mathcal{P}_2(\mathbb{R}) \,|\, W_2(\mu, \nu) < \delta\} \subset \mathcal{B}^-$. Let $\mu \in B_\delta(\nu)$. We have to show that $\lim_{t\to\infty} S[\mu](t) = \mu^-$. In view of Proposition 1.3, it suffices to show that

(i) $\mathrm{m}[S[\mu](2t')] < 0$, and that
(ii) $\mathcal{F}(S[\mu](2t')) \;\leq\; \mathcal{F}(\mu^0)$.

In order to show (i), note that by the contraction estimate (2.9) and the definition of t' and δ,

$$W_2(S[\mu](2t'), \mu^-) \;\leq\; W_2(S[\nu](2t'), \mu^-) + \mathrm{e}^{-2\lambda t'} \delta \;\leq\; \frac{m^\star}{2}. \tag{4.5}$$

This implies claim (i). To show claim (ii), we use the regularization estimate (2.8), and obtain that

$$\mathcal{F}(S[\mu](2t')) \;\leq\; \mathcal{F}(S[\nu](t')) + |\lambda| \, W_2(S[\nu](t'), S[\mu](t'))^2 \;\leq\; \mathcal{F}(\mu^-) + \frac{1}{2}\Delta \;<\; \mathcal{F}(\mu^0). \tag{4.6}$$

This concludes the proof of claim (ii). □

5 Proof of Theorem 1.2

Proof of Theorem 1.2. In the following let $\mu_t = S[\mu](t)$ for all $t \in [0,\infty)$. First suppose that $\mu \in D(\mathcal{F})$. We know from Lemma 3.1 that there exists a subsequence $(\mu_{t_k})_k$ and $\mu^* \in \{\mu^-, \mu^0, \mu^+\}$ such that

$$\lim_{k\to\infty} W_2(\mu_{t_k}, \mu^*) = 0 \quad \text{and} \quad \lim_{t\to\infty} \mathcal{F}(\mu_t) = \mathcal{F}(\mu^*). \tag{5.1}$$

Let $(\mu_{s_n})_{n\in\mathbb{N}}$ be a subsequence of $(\mu_t)_{t\in[0,\infty)}$. As in Step 2 of the proof of Lemma 3.1, we infer the existence of a further subsequence, still denoted by $(\mu_{s_n})_{n\in\mathbb{N}}$, such that

$$\lim_{n\to\infty} W_2(\mu_{s_n}, \nu^*) = 0 \quad \text{for some } \nu^* \in \mathcal{P}_2(\mathbb{R}). \tag{5.2}$$

It remains to show that $\nu^* = \mu^*$. We divide the proof into the three cases $\mu^* = \mu^-$, $\mu^* = \mu^0$ and $\mu^* = \mu^+$.

Case 1. $[\mu^* = \mu^-.]$
As in (3.12), we infer that $\mathcal{F}(\nu^*) \le \mathcal{F}(\mu^-)$. By Lemma 2.4, this implies that either $\nu^* = \mu^- = \mu^*$ or $\nu^* = \mu^+$. It remains to show that the latter case leads to a contradiction. Note that by (5.1) and (5.2),

- there exists $T > 0$ such that $\mathcal{F}(\mu_t) \in [\mathcal{F}(\mu^-), \mathcal{F}(\mu^0))$ for all $t \ge T$,
- there exists $N \in \mathbb{N}$ such that $s_n \ge T$ and $\mathrm{m}[\mu_{s_n}] > 0$ for all $n \ge N$, and
- there exists $K \in \mathbb{N}$ such that $t_k > s_N$ and $\mathrm{m}[\mu_{t_k}] < 0$ for all $k \ge K$.

In particular, we have that

$$\mathcal{F}(\mu_t) < \mathcal{F}(\mu^0) \text{ for all } t \in [s_N, t_K], \qquad \mathrm{m}[\mu_{s_N}] > 0, \quad \text{and} \quad \mathrm{m}[\mu_{t_K}] < 0. \tag{5.3}$$

Hence, there exists $t' \in [s_N, t_K]$ such that $\mathcal{F}(\mu_{t'}) < \mathcal{F}(\mu^0)$ and $\mathrm{m}[\mu_{t'}] = 0$. This contradicts Lemma 2.4.

Case 2. $[\mu^* = \mu^+.]$
This case is treated in the same way as Case 1.

Case 3. $[\mu^* = \mu^0.]$
In this case we have that $\mathcal{F}(\nu^*) \le \mathcal{F}(\mu^0)$. There are three subcases given by $\mathrm{m}[\nu^*] = 0$, $\mathrm{m}[\nu^*] > 0$ and $\mathrm{m}[\nu^*] < 0$.

Case 3.1. $[\mathrm{m}[\nu^*] = 0.]$
By Lemma 2.4, the combination of $\mathcal{F}(\nu^*) \le \mathcal{F}(\mu^0)$ and $\mathrm{m}[\nu^*] = 0$ yields that $\nu^* = \mu^0 = \mu^*$.

Case 3.2. $[\mathrm{m}[\nu^*] < 0.]$
From Proposition 1.3 we know that $\nu^* \in \mathcal{B}^-$. Hence, by Proposition 4.1, there exists $\delta > 0$ such that $B_\delta(\nu^*) \subset \mathcal{B}^-$. In particular, by (5.2), there exists $N \in \mathbb{N}$ such that $\mu_{s_N} \in \mathcal{B}^-$. This contradicts (5.1). Indeed, the fact that $\mu_{s_N} \in \mathcal{B}^-$ implies that

$$\lim_{t\to\infty} \mu_{s_N + t} \;=\; \lim_{t\to\infty} S[\mu_{s_N}](t) \;=\; \mu^- \quad \text{in } \mathcal{P}_2(\mathbb{R}), \tag{5.4}$$

which contradicts the fact that $\lim_{k\to\infty} \mu_{t_k} \;=\; \mu^* \;=\; \mu^0 \quad \text{in } \mathcal{P}_2(\mathbb{R})$.

Case 3.3. $[\mathrm{m}[\nu^*] > 0.]$
This case is treated in the same way as Case 3.2. This concludes the proof of this theorem for the case $\mu \in D(\mathcal{F})$.

Now let $\mu \in \mathcal{P}_2(\mathbb{R}) \setminus D(\mathcal{F})$. Let $h \in (0,\infty)$. Applying the regularization estimate (2.8) yields that $S[\mu](h) \in D(\mathcal{F})$. Hence, we have proven the claims of this theorem for $S[\mu](h)$. That is, there exists $\mu^* \in \{\mu^-, \mu^0, \mu^+\}$ such that $\lim_{t\to\infty} W_2(S\big[S[\mu](h)\big](t), \mu^*) = 0$ and $\lim_{t\to\infty} \mathcal{F}(S\big[S[\mu](h)\big](t)) = \mathcal{F}(\mu^*)$. Therefore, since $\lim_{t\to\infty} W_2(S\big[S[\mu](h)\big](t), \mu^*) = \lim_{t\to\infty} W_2(S[\mu](t), \mu^*)$ and $\lim_{t\to\infty} \mathcal{F}(S\big[S[\mu](h)\big](t)) = \lim_{t\to\infty} \mathcal{F}(S[\mu](t))$, we conclude the proof of this theorem also for the case $\mu \in \mathcal{P}_2(\mathbb{R}) \setminus D(\mathcal{F})$. □

In the following corollary, we state some consequences of Theorem 1.2 for the set $\mathcal{B}^0$.

Corollary 5.1. *(i) $\mathcal{B}^0$ is closed,*

(ii) $\mathcal{B}^0 \supset \{ \mu \in \mathcal{P}_2(\mathbb{R}) \mid \mu$ is symmetric, i.e. $\varsigma_\# \mu = \mu \}$, and

(iii) $\mu^0 \in \partial \mathcal{B}^0$.

Proof. To show part *(i)*, we simply use Proposition 4.1 and that, by Theorem 1.2, $\mathcal{P}_2(\mathbb{R}) = \mathcal{B}^- \cup \mathcal{B}^0 \cup \mathcal{B}^+$. Part *(ii)* is a straightforward consequence of Theorem 1.2 and Lemma 2.6. Finally, to show part *(iii)*, we use that by Proposition 1.3, $\mu^{\varphi'(-\eta)} \in \mathcal{B}^-$ and $\mu^{\varphi'(\eta)} \in \mathcal{B}^+$ for all $\eta > 0$, and that $\lim_{\eta \downarrow 0} W_2(\mu^{\varphi'(-\eta)}, \mu^0) = \lim_{\eta \downarrow 0} W_2(\mu^{\varphi'(\eta)}, \mu^0) = 0$. ☐

6 Some comments on the assumptions in this paper

In this chapter we briefly discuss the assumptions we make in this paper.

We first discuss Assumption 1.1. Assumption 1.1 (1) ensures that $z \mapsto \mathrm{e}^{-\Psi(z)}$ is integrable and that $\Psi'' \geq \tilde{\lambda}$ for some $\tilde{\lambda} \in \mathbb{R}$. The latter condition implies Lemma 2.2 (ii), which is an essential ingredient in order to apply the Wasserstein gradient flow theory for the functional $\mathcal{F}$; see [1, Section 10.4]. Assumption 1.1 (2) implies that the absolute moments of order $2+\epsilon$ of the McKean-Vlasov evolution are uniformly bounded; see (2.6) and (3.3). This uniform boundedness in turn implies some compactness property, which is an essential ingredient for the proofs; see, e.g., the proof of Lemma 3.1. Assumption 1.1 (3) implies some symmetry properties of $\mathcal{F}$ that simplify our analysis considerably; see the Lemmas 2.4 and 2.6. We believe that Assumptions 1.1 (3) is not essential and can be circumvented. Assumption 1.1 (4) is the main reason why the system (1.1) admits exactly the three stationary states μ^-, μ^0, μ^+. Indeed, as it can be seen in [4, 3.4], if Assumption 1.1 (4) is not true, then the macroscopic Hamiltonian $\bar{H}$ admits at most one critical point. This implies that the system (1.1) admits at most one stationary state, since the critical points of $\bar{H}$ determine the stationary states of the system (1.1); see the proof of Lemma 2.5. Assumption 1.1 (5) is a technical assumption taken from [4, Chapter 3], where it is used in the proof of the fact that $\bar{H}$ admits exactly three critical points. We believe that also Assumptions 1.1 (5) is not essential and can be circumvented. The latter is left for future research.

We finally note that the Wasserstein gradient flow theory holds in a much more generality than it is used here. Hence, one may wonder if the results of this paper can be extended to more general settings. Unfortunately, our arguments rely on the facts that the system (1.1) is one-dimensional and that the interaction energy in (1.2) is quadratic. It is also left for future research to generalize the results of this paper to multi-dimensional settings and for more general interaction energies.

References

[1] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition, 2008. MR-2401600

[2] S. Arnrich, A. Mielke, M. A. Peletier, G. Savaré, and M. Veneroni. Passing to the limit in a Wasserstein gradient flow: from diffusion to reaction. *Calc. Var. Partial Differential Equations*, 44(3-4):419–454, 2012. MR-2915328

[3] K. Bashiri and A. Bovier. Gradient flow approach to local mean-field spin systems. *Stochastic Process. Appl.*, 130(3):1461–1514, 2020. MR-4058280

[4] K. Bashiri and G. Menz. Metastability in a continuous mean-field model at low temperature and strong interaction. 2019. Preprint, arXiv:1910.11828v3.

[5] A. Bovier and F. den Hollander. *Metastability*, volume 351 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer, Cham, 2015. A potential-theoretic approach. MR-3445787

[6] J. A. Carrillo, R. J. McCann, and C. Villani. Kinetic equilibration rates for granular media and related equations: entropy dissipation and mass transportation estimates. *Rev. Mat. Iberoamericana*, 19(3):971–1018, 2003. MR-2053570

[7] D. A. Dawson and J. Gärtner. Large deviations and tunnelling for particle systems with mean field interaction. *C. R. Math. Rep. Acad. Sci. Canada*, 8(6):387–392, 1986. MR-0866922

[8] D. A. Dawson and J. Gärtner. Large deviations from the McKean-Vlasov limit for weakly interacting diffusions. *Stochastics*, 20(4):247–308, 1987. MR-0885876

[9] J. Gärtner. On the McKean-Vlasov limit for interacting diffusions. *Math. Nachr.*, 137:197–248, 1988. MR-0968996

[10] N. Grunewald, F. Otto, C. Villani, and M. G. Westdickenberg. A two-scale approach to logarithmic Sobolev inequalities and the hydrodynamic limit. *Ann. Inst. Henri Poincaré Probab. Stat.*, 45(2):302–351, 2009. MR-2521405

[11] S. Herrmann and J. Tugaut. Non-uniqueness of stationary measures for self-stabilizing processes. *Stochastic Process. Appl.*, 120(7):1215–1246, 2010. MR-2639745

[12] R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the Fokker-Planck equation. *SIAM J. Math. Anal.*, 29(1):1–17, 1998. MR-1617171

[13] P. E. Müller. Limiting properties of a continuous local mean-field interacting spin system: Hydrodynamic limit, propagation of chaos, energy landscape and large deviations. PhD thesis, Rheinische Friedrich-Wilhelms-Universität Bonn, 2016.

[14] E. Olivieri and M. E. Vares. *Large deviations and metastability*, volume 100 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 2005. MR-2123364

[15] F. Otto. The geometry of dissipative evolution equations: the porous medium equation. *Comm. Partial Differential Equations*, 26(1-2):101–174, 2001. MR-1842429

[16] J. Tugaut. Convergence to the equilibria for self-stabilizing processes in double-well landscape. *Ann. Probab.*, 41(3A):1427–1460, 2013. MR-3098681

[17] C. Villani. *Optimal transport*, volume 338 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2009. Old and new. MR-2459454

Acknowledgments. The author would like to thank Anton Bovier, Matthias Erbar and André Schlichting for numerous useful discussions. Moreover, he would like to thank the anonymous referees for reading the paper with great care and for their valuable comments.