

BARRIERS OF THE MCKEAN–VLASOV ENERGY VIA A MOUNTAIN PASS THEOREM IN THE SPACE OF PROBABILITY MEASURES

RISHABH S. GVALANI AND ANDRÉ SCHLICHTING

ABSTRACT. We show that the empirical process associated with a system of weakly interacting diffusion processes exhibits a form of noise-induced metastability. The result is based on an analysis of the associated McKean–Vlasov free energy, which, for suitable attractive interaction potentials, has at least two distinct global minimisers at the critical parameter value $\beta = \beta_c$. On the torus, one of these states is the spatially homogeneous constant state, and the other is a clustered state. We show that a third critical point exists at this value. As a result, we obtain that the probability of transition of the empirical process from the constant state scales like $\exp(-N\Delta)$, with Δ the energy gap at $\beta = \beta_c$. The proof is based on a version of the mountain pass theorem for lower semicontinuous and λ -geodesically convex functionals on the space of probability measures $\mathcal{P}_2(M)$ equipped with the 2-Wasserstein metric, where M is a complete, connected, and smooth Riemannian manifold.

1. INTRODUCTION

In recent years, a lot of progress has been made in understanding the convergence of interacting particle systems to their hydrodynamic or mean-field limits at the level of the convergence of gradient flows (cf. [ADPZ11, ADPZ13, DPZ13, Fat16, EFLS16, FS16, KJZ19]). These limits are described by dissipative evolution equations which are driven by some macroscopic free energy with respect to some metric. This gradient flow structure allows for a characterisation of the stationary states of the system in terms of critical points and minimisers of the free energy. Hence, the free energy landscape and the underlying metric encode some of the system’s dynamical properties. In many applications, the free energy is usually a lower semicontinuous (l.s.c) function with the space of probability measures $\mathcal{P}_2(M)$ as its domain. Here $\mu \in \mathcal{P}_2(M)$ represents the distribution of particle positions on some base manifold M . The appropriate metric

Date: Submitted May, 28 2019. arXiv: 1905.11823.

Key words and phrases. Free energy barrier, large deviations, McKean–Vlasov equation, mountain pass theorem, optimal transport, space of probability measures.

RSG is funded by an Imperial College President’s PhD Scholarship, partially through EPSRC Award Ref. 1676118. AS is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany’s Excellence Strategy EXC 2047 – 390685813, the *Hausdorff Center for Mathematics*, as well as the Collaborative Research Center 1060 – 211504053, *The Mathematics of Emergent Effects* at the Universität Bonn.

for the gradient flow is usually the 2-Wasserstein metric and its variants. For example, in [BB20], the authors derive a local mean-field model as the gradient flow of the macroscopic free energy with respect to a modified Wasserstein metric.

For macroscopic models originating from interacting particle systems, the free energy can exhibit multiple local minima corresponding to distinguished stationary states of the macroscopic system. In this case, one may want to understand typical transition times and transition states between two such distinct states in the presence of noise. A typical example of this is a classical particle moving in \mathbb{R}^d along the gradient of some potential $V \in C^2(\mathbb{R}^d; \mathbb{R})$, i.e.

$$(1.1) \quad \dot{x}(t) = -\nabla V(x),$$

with $x(0) = x_0 \in \mathbb{R}^d$. Let us assume that V has exactly two distinct global minima $x_1, x_2 \in \mathbb{R}^d$, which are also the stationary points of (1.1). If one considers these to be the states of interest, then a relevant question is how does the particle transition from one to the other under the influence of noise. To understand this, one considers the stochastic differential equation (SDE)

$$(1.2) \quad dX_t = -\nabla V(X_t) dt + \sqrt{2\beta^{-1}} dB_t,$$

where B_t is a \mathbb{R}^d -valued Wiener process and $\beta > 0$ is a parameter representing the strength of the noise in the system. In the setting of the above SDE, the question can be reframed as follows: given $X_0 = x_1$, what is the probability that in some finite time $T > 0$, we have that $X_T = x_2$. This question is answered, at least for $\beta \gg 1$, by the Freidlin–Wentzell theorem. In particular, it tells us that the family of processes $\{X_t^\beta\} \in C([0, T]; \mathbb{R})$ with $X_0 = x_1$ satisfy a large deviations principle with good rate function $S : C([0, T]; \mathbb{R}^d) \rightarrow \mathbb{R} \cup \{+\infty\}$ given by

$$S(f) := \frac{1}{2} \int_0^T |\dot{f}(t) + \nabla V(f(t))|^2 dt,$$

whenever the above integral is finite and $+\infty$ otherwise. As a consequence of the above result, we have that, for any closed and measurable $\Gamma \subset C([0, T]; \mathbb{R}^d)$

$$\limsup_{\beta \rightarrow +\infty} 2\beta^{-1} \log \mathbb{P}(X_t^\beta \in \Gamma) \leq - \inf_{f \in \Gamma} S(f).$$

If we pick $\Gamma = \{f \in C([0, T]; \mathbb{R}^d) : f(0) = x_1, f(T) = x_2\}$, we obtain an upper bound on the probability that the process reaches x_2 given that it starts at x_1 . Setting $T^* = \arg \max_{t \in [0, T]} (V(f(t)) - V(f(0)))$, we can obtain the following lower bound for $f \in \Gamma$,

$$\begin{aligned} S(f) &\geq \frac{1}{2} \int_0^{T^*} |\dot{f}(t) + \nabla V(f(t))|^2 dt \geq \int_0^{T^*} \dot{f}(t) \cdot \nabla V(f(t)) dt \\ &= V(f(T^*)) - V(f(0)) \geq \inf_{f \in \Gamma} (V(f(T^*)) - V(f(0))) =: c - V(f(0)). \end{aligned}$$

It turns out that $c > 0$ is in fact a critical value of V , i.e. there exists $x_3 \in \mathbb{R}^d$ such that $V(x_3) = c$ and $\nabla V(x_3) = 0$. The reader will recognise this as the finite-dimensional version of the well-known mountain pass theorem. Setting $\Delta := V(x_3) - V(x_1)$, we see that for β sufficiently large

$$\mathbb{P}(X_t^\beta \in \Gamma) \lesssim \exp(-\beta \Delta/2).$$

Thus, the probability of the process reaching the new phase/state in time $T > 0$ goes exponentially with β with the rate given by the difference between the energies of the saddle point and the initial phase. Thus, we can see that the process finds the path of least resistance to reach the new phase in agreement with the fundamental tenet of large deviations theory that “*an unlikely event will happen in the most likely of the possible unlikely ways.*” These transitions correspond to the phenomenon of noise-induced metastability, i.e. the process is stable around x_1 for $\beta \gg 1$, but there is an exponentially small probability of it transitioning to x_2 .

The purpose of this paper is to obtain results in a similar flavour but in an infinite-dimensional setting. Specifically, we are interested in understanding how related phenomena, i.e. noise-induced transitions, occur in systems governed by the Wasserstein gradient flow of some free energy I , especially those that arise as mean-field limits of interacting particle systems. We consider the following system of N interacting SDEs on \mathbb{T}_L^d (the d -dimensional torus of side length $L > 0$)

$$dX_t^i = -\frac{1}{N} \sum_{j=1}^N \nabla W(X_t^i - X_t^j) dt + \sqrt{2\beta^{-1}} dB_t^i$$

$$\text{Law}(\overline{X}_0^N) = \prod_{i=1}^N \nu(x_i) \quad \overline{X}_t^N = (X_t^1, \dots, X_t^N)$$

where $\beta > 0$ is a parameter, $W \in C^2(\mathbb{T}_L^d)$ is an interaction potential which is even along every coordinate, and B_t^i are \mathbb{T}_L^d -valued independent Wiener processes.

Let $\mu^{(N)}(t) := N^{-1} \sum_{i=1}^N \delta_{X_t^i}$, then it is well known (cf. [Szn91]) that $\mu^{(N)}(t)$ as a measure-valued random variable converges in law to $\mu = \mu(x, t)$ for each $t > 0$, where μ is a weak solution of the following PDE

$$(1.3) \quad \partial_t \mu = \nabla \cdot (\mu \nabla (\beta^{-1} \log \mu + W \star \mu)) \quad \text{with} \quad \mu(x, 0) = \nu(x).$$

The above PDE is commonly referred to as the McKean–Vlasov equation and can be rewritten as W_2 -gradient flow

$$\partial_t \mu = \nabla \cdot \left(\mu \nabla \frac{\delta I}{\delta \mu} \right),$$

where $I : \mathcal{P}(\mathbb{T}_L^d) \rightarrow \mathbb{R} \cup \{+\infty\}$ is the associated free energy. Its domain is the space of absolutely continuous measures and for those it is given by

$$(1.4) \quad I(\mu) = \beta^{-1} \int \log\left(\frac{d\mu}{dx}\right) d\mu + \frac{1}{2} \iint W(x-y) d\mu(y) d\mu(x),$$

where $\frac{d\mu}{dx}$ denotes the density of μ with respect to the Lebesgue measure dx on \mathbb{T}_L^d . The first term in (1.4) is referred to as the entropy and the second as the interaction energy. The function I is referred to as the free energy of the system. The balance between entropy and interaction energy in terms of β determines what the minimisers of I look like. For β smaller than some critical value β_c , the normalised Lebesgue measure, $\mu^L(dx) = L^{-d} dx$ is the unique minimiser of the free energy. Above the value, β_c , a new minimiser of the free energy, which is not μ^L , emerges. The change in the structure of the set of minimisers of I is called a phase transition and is observed in many models from the physical sciences [LP66, Sin82, Daw83, Shi87, GP18, FV18].

This operator $\nabla \cdot (\mu \nabla_{\frac{\delta}{\delta\mu}}(\cdot))$ can be formally thought of as a gradient in the space of probability measures on \mathbb{T}_L^d equipped with the 2-Wasserstein mass transportation distance, which is defined as follows

$$W_2^2(\mu, \nu) := \inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathbb{T}_L^d \times \mathbb{T}_L^d} d_{\mathbb{T}_L^d}^2(x, y)^2 d\pi(x, y),$$

where $\Pi(\mu, \nu)$ is the set of all couplings between μ and ν and $d_{\mathbb{T}_L^d}(\cdot, \cdot)$ is the distance on \mathbb{T}_L^d . We note that $\mathcal{P}(\mathbb{T}_L^d)$ equipped with W_2 is a complete, separable metric space. For μ, ν absolutely continuous with respect to dx , the definition of the metric can be recast into the form discussed in Theorem 3.4. This notion of a gradient flow can be made rigorous and is an extremely active field of research. However, the present work relies on quite classical results (cf. [CEMS01, McC01, McC97, AGS08]). Indeed, the solutions of the McKean-Vlasov PDE are curves of maximal slope of the McKean-Vlasov energy I with respect to W_2 (see [DS14]). Comparing this with the toy model discussed further up in the introduction, we see that the PDE has a gradient structure in W_2 and so the functional I will play a similar role to the potential V in (1.1). The distinct phases/states are then characterised by the global minima of the functional I over $\mathcal{P}(\mathbb{T}_L^d)$. The role of the SDE in (1.2) is then played by the empirical process $\mu^{(N)}$ and that of the parameter β is played by N . In this context, we also refer to some recent progress in the understanding of singular SPDEs related to the fluctuations of the empirical process around its mean-field limit [FG19, KLvR20, CSZ19, CSZ20].

Understanding such noise-induced transitions requires two ingredients: a version of the mountain pass theorem in the space of probability measures $\mathcal{P}_2(M)$ equipped with the Wasserstein metric and an appropriate large deviations principle for the underlying

particle system. We focus on the first ingredient noting that the second ingredient is usually application-specific. Our main result in this direction is as follows.

Theorem 1.1. *Assume M is a complete, connected, and smooth Riemannian manifold. Let $I : \mathcal{P}_2(M) \rightarrow \mathbb{R} \cup \{+\infty\}$ be a proper, l.s.c, and λ -geodesically convex functional. Suppose $\mu, \nu \in \mathcal{P}_2(M) \cap D(I)$, Γ is the set of all continuous curves $\gamma : [0, 1] \rightarrow \mathcal{P}_2(M)$ (where $\mathcal{P}_2(M)$ is equipped with the 2-Wasserstein metric, W_2) with $\gamma(0) = \mu$ and $\gamma(1) = \nu$, and the function $\Upsilon : \Gamma \rightarrow \mathbb{R}$ is defined by:*

$$\Upsilon(\gamma) = \sup_{t \in [0, 1]} I(\gamma(t)).$$

Let $c = \inf_{\gamma \in \Gamma} \Upsilon(\gamma)$ and $c_1 = \max\{I(\mu), I(\nu)\}$. If $c > c_1$ and I satisfies (MPS) (see Assumption (2.2)), then c is a critical value of I , that is there exists a $\eta \in \mathcal{P}_2(M)$ with $I(\eta) = c$ such that $|\partial I|(\eta) = |dI|(\eta) = 0$ (see Definition 2.1 and Definition 4.5).

The proof utilises the notion of the weak metric slope $|dI|$ first introduced in [Kat94]. The main advantage over previous results in this direction is that we can apply the result to l.s.c functionals on $\mathcal{P}_2(M)$ as long as they are λ -convex by working with the extension of the function to its epigraph based on ideas discussed by Degiovanni and Marzocchi [DM94] originating from work in [dGMT80]. In fact for λ -convex functionals one can identify the usual (strong) metric slope $|\partial I|$ and $|dI|$. We focus on the case in which the metric is W_2 although the results generalise for W_p or other variants of the metric.

Our first result shows that the abstract mountain pass Theorem 1.1 can be applied to the McKean–Vlasov free energy I , as defined in (1.4), after verifying the necessary regularity assumptions.

Theorem 1.2. *Assume W and β are such that there exist two measures $\mu, \nu \in \mathcal{P}(\mathbb{T}_L^d)$ such that μ is a strict local minimum of the McKean–Vlasov free energy I (cf. (1.4)) and $I(\nu) \leq I(\mu)$. Then, there exists $\mu^* \in \mathcal{P}(\mathbb{T}_L^d)$, distinct from μ and ν , such that $|\partial I|(\mu^*) = |dI|(\mu^*) = 0$. Additionally, $I(\mu^*) = c$, where c is given by*

$$c = \inf_{\gamma \in \Gamma} \sup_{t \in [0, 1]} I(\gamma(t)),$$

where $\Gamma = \{C([0, 1]; \mathcal{P}(\mathbb{T}_L^d)) : \gamma(0) = \mu, \gamma(1) = \nu\}$.

The proof of the above result can be found in Section 5. Furthermore, in Section 5, by relying on results from [CGPS20], we will show that we can establish explicit conditions on the interaction potential W such that two distinct global minimisers $\{\mu^L, \bar{\mu}\}$ of the free energy I (1.4) exist. This happens at a so-called *discontinuous transition point* $\beta_c > 0$. This provides us with a scenario in which we can apply Theorem 1.2. Hereby,

$\mu^L := L^{-d} dx$ is the uniform state and $\bar{\mu}$ is a clustered state. The existence of the associated saddle point can be found in Corollary 5.7.

Remark 1.3. The abstract mountain pass theorem in Theorem 1.1 holds whenever one can find two measures, not necessarily critical points, in the domain of some $I : \mathcal{P}_2(M) \rightarrow \mathbb{R} \cup \{+\infty\}$, such that the barrier value c exceeds the maximum of their energies. We have chosen to apply the result in Theorem 1.2 at a strict local minimum of the McKean–Vlasov free energy, I , because in this setting it is clear that the barrier value exceeds the value at the local minimum.

The fact, from Corollary 5.7, that the free energy functional I has an energy barrier at $\beta = \beta_c$ allows us to study escape probabilities for the underlying particle system using results which were first proved by Dawson and Gärtner [DG87]. We refer the reader to [ADPZ11, Rey18, GPY13] for further discussions of the connections between large deviations theory and theory of gradient flows.

Theorem 1.4. *Assume W and β_c are such that there exist at least two distinct minimisers $\{\mu^L, \bar{\mu}\}$ of I . It follows then that the underlying empirical process $\mu^{(N)} \in \mathcal{C}_T$ with initial i.i.d uniformly distributed particles satisfies*

$$\mathbb{P}(\mu^{(N)}(T) \in \overline{B}_\varepsilon^{W_2}(\bar{\mu}), \mu^{(N)}(0) = \mu_0^{(N)}) \leq \exp\left(-N(\Delta - O(\varepsilon^2) - o_T(1))\right)$$

for N sufficiently large, where $\overline{B}_\varepsilon^{W_2}(\bar{\mu})$ is the closed ball of size $\varepsilon > 0$ around $\bar{\mu}$ in W_2 , $\Delta := I(\mu^*) - I(\mu^L)$ and μ^* is the critical point defined in Corollary 5.7.

The above result says that the probability of the empirical process reaching the clustered state, $\bar{\mu}$, from the uniform state, μ^L , in time $T > 0$ becomes exponentially small as the number of particles increases, as long as the system is at a discontinuous transition point. In light of the recent results in [FG19], we expect the above bound to hold for a stochastic version of the McKean–Vlasov PDE in the regime of vanishing noise, i.e. Freidlin–Wentzell-type large deviations.

Remark 1.5. We note that we have considered the McKean–Vlasov system on the torus in this paper which on the one hand has the advantage that the space $\mathcal{P}(\mathbb{T}_L^d)$ equipped with the 2-Wasserstein distance is compact, while on the other hand we can build on the characterisation of critical points and phase transitions from the work in [CGPS20]. Thus, we can extract a lot of information about the structure of stationary states in this setting. On the torus, the normalised Lebesgue measure $\mu^L = L^{-d} dx$ is always an invariant measure for the McKean–Vlasov dynamics (1.3) and it is the unique minimiser of the free energy before the critical temperature. Thus, linearisation arguments provide us with a lot of useful information that may not be readily available for the diffusions on \mathbb{R}^d . An extension to \mathbb{R}^d , with some suitable confinement potential, seems to be

possible. The critical points and phase transitions are studied for specific choices of W , for example, in [Tam84] and [Tug14]. For the case of diffusions on \mathbb{R}^d with a bi-stable confinement and Curie–Weiss-type quadratic interaction, large deviations, escape probabilities, and tunnelling results can be found in [DG86, DG87, DG89] while a study of the basins of attraction of the different stationary states is the content of [Bas20].

Remark 1.6. Even though the Dawson–Gärtner large deviations principle provides an exponential lower bound on the above probability, it is not clear how this can be compared to the energy barrier $\exp(-N(\Delta - O(\varepsilon^2)))$ for a general model. However, such a lower bound could be obtained, for example, in the following setting: for all $\varepsilon > 0$, there exist two points $\mu_0^*, \bar{\mu}^*$ in a neighborhood of μ^* such that μ_0^* is connected to μ^L and $\bar{\mu}^*$ is connected to $\bar{\mu}$ through a heteroclinic orbit under the flow of the McKean–Vlasov PDE. However, it is unlikely that such heteroclinic connections exist at this level of generality in the choice of W . A first step in this direction would be the characterisation of μ^* for specific choices of W . In analogy to the situation in finite-dimensional Hamiltonian dynamical systems, it may be possible to use a version of weak KAM theory in the Wasserstein space of probability measures to construct such heteroclinic orbits. We refer the reader to [FK09, FN12, AF14] for the construction of solutions to Hamiltonian–Jacobi PDEs in the space of probability measures. These ideas may help in obtaining a generalisation of weak KAM theory to dissipative evolution equations in the space of probability measures.

Outline. The paper is organised as follows: In Section 2, we introduce the notion of the weak metric slope and metric critical points that we will use throughout the paper and a version of the mountain pass theorem due to Katriel that holds for continuous functions on metric spaces. In Section 3, we briefly recall some results due to McCann on optimal transport on Riemannian manifolds. In Section 4, we compare the notion of the weak metric slope with the notion of (strong) metric slope used in the gradient flows community and show that, under the assumption of λ -convexity of I , the two are equivalent. We conclude the section by proving Theorem 1.1. In the final Section 5, we discuss as a specific application of the result: the McKean–Vlasov model. We state and extend some results from [CGPS20] on the structure of the set of minimisers of I and their phase transitions. We proceed by showing the existence of mountain pass at the point of discontinuous phase transition, thus proving Theorem 1.2. Finally, we introduce the precise form of the large deviations principle due to Dawson and Gärtner and complete the proof of Theorem 1.4. In particular, these results imply a kind of noise-induced metastability for the underlying particle system.

2. CRITICAL POINTS IN METRIC SPACES

We will assume throughout this section that (\mathcal{X}, d) is a complete metric space. We start with the definition of the weak metric slope for some real-valued continuous function defined on \mathcal{X} . The notion goes back to Ioffe and Schwartzman [IS96] who provided the definition in the Banach space setting.

Definition 2.1 (δ -regular points, weak metric slope and critical points [Kat94]). Let $x \in \mathcal{X}$, and $I : \mathcal{X} \rightarrow \mathbb{R}$ be a continuous function defined in a neighbourhood of x . Given $\delta > 0$, x is said to be a δ -regular point of I if there is a neighbourhood U of x , a constant $\alpha > 0$, and a continuous mapping $\psi : U \times [0, \alpha] \rightarrow \mathcal{X}$ such that for all $(u, t) \in U \times [0, \alpha]$, it holds:

- (1) $d(\psi(u, t), u) \leq t$.
- (2) $I(u) - I(\psi(u, t)) \geq \delta t$.

If this is the case, ψ is called a δ -regularity mapping for I at x and x is called a *regular point* of I .

The *weak metric slope* of I at x is given by the extended real number

$$|dI|(x) = \sup\{\delta \in (0, \infty) : I \text{ is } \delta\text{-regular at } x\}.$$

If x is not δ -regular for any $\delta > 0$, then x is called a *critical point* of I with $|dI|(x) = 0$.

Assumption 2.2 (Weak metric Palais–Smale condition). A function $I : \mathcal{X} \rightarrow \mathbb{R}$ is said to satisfy the *weak metric Palais–Smale condition* (MPS) if any *Palais sequence*, that is $\{u_n\}_{n \in \mathbb{N}} \in \mathcal{X}$ with $I(u_n) \rightarrow c \in \mathbb{R}$ and $|dI|(u_n) \rightarrow 0$, possesses a convergent subsequence.

Given this notion, we have the following generalisation of the Ambrosetti–Rabinowitz mountain pass theorem due to Katriel [Kat94].

Theorem 2.3. *Let \mathcal{X} be a path-connected metric space and $I : \mathcal{X} \rightarrow \mathbb{R}$ be continuous. For $u_0, u_1 \in \mathcal{X}$ let Γ be the set of all continuous curves $\gamma : [0, 1] \rightarrow \mathcal{X}$ with $\gamma(0) = u_0$ and $\gamma(1) = u_1$, and the function $\Upsilon : \Gamma \rightarrow \mathbb{R}$ is given by*

$$\Upsilon(\gamma) = \sup_{t \in [0, 1]} I(\gamma(t)).$$

Let $c = \inf_{\gamma \in \Gamma} \Upsilon(\gamma)$ and $c_1 = \max\{I(u_0), I(u_1)\}$. If $c > c_1$ and I satisfies (MPS), then c is a critical value of I .

For the application of the mountain pass theorem to the energies encountered in gradient flows, we need a working definition of the weak slope if I is only lower semicontinuous. We also need to deal with the fact that Theorem 2.3 holds only for continuous functions. Consider the example $I : \mathbb{R} \rightarrow \mathbb{R}$ with $I(x) = x + 1$ for $x < 0$ and $I(x) = x$ for $x \geq 0$. Then I is l.s.c, and it is easy to verify that I has a critical point at $x = 0$ in the sense

of Definition 2.1. However, this seems to be in some sense pathological for identifying mountain pass points as I does not attain the value of the barrier at $x = 0$, i.e. $I(0) = 0$. We like to use a theory of critical points for l.s.c. functionals which handles such critical points in our generalisation of the mountain pass theorem. One possible resolution for this example would be to view the function I as a multivalued map at $x = 0$ with values $[0, 1]$ from which the energy barrier is identified as the maximum of those values. Another resolution for this issue was suggested by Degiovanni and Marzocchi [DM94], who, using notions developed in [dGMT80], proposed a generalisation based on the so-called *epigraph extension*, \mathcal{G}_I , a continuous function associated to I and defined on its epigraph (cf. Definition 2.4). This idea also helps us overcome the difficulty that Theorem 2.3 holds only for continuous functions.

Definition 2.4 (Extension to the epigraph). Let $I : \mathcal{X} \rightarrow \mathbb{R} \cup \{+\infty\}$ be a proper l.s.c functional and denote by $\text{epi}(I) = \{(u, \xi) \in \mathcal{X} \times \mathbb{R} : I(u) \leq \xi\}$ its epigraph, which is equipped with the *graph metric* $d_{\text{epi}}((u, \xi), (v, \zeta)) = \sqrt{d(u, v)^2 + |\xi - \zeta|^2}$. The *epigraph extension* of I is the functional $\mathcal{G}_I : \text{epi}(I) \rightarrow \mathbb{R} \cup \{+\infty\}$ defined by

$$\mathcal{G}_I(u, \xi) = \xi, \quad (u, \xi) \in \text{epi}(I).$$

It is now straightforward to check that \mathcal{G}_I is a continuous function with respect to d_{epi} and that $|d\mathcal{G}_I|(u, \xi) \leq 1$ for all $(u, \xi) \in \text{epi}(I)$. Let us also point out that the epigraph of a lower semicontinuous function is closed [BP12, Proposition 2.5]. It turns out, that the notion of the weak slope for l.s.c functions on \mathcal{X} based on the epigraph extension is suitable for applications to mountain pass theorems in metric spaces.

Definition 2.5. Let $I : \mathcal{X} \rightarrow \mathbb{R} \cup \{+\infty\}$ be a proper l.s.c function. Define its domain as

$$D(I) := \{x \in \mathcal{X} : I(x) < +\infty\}.$$

Then the *weak metric slope* at $x \in D(I)$ is defined by

$$|dI(x)| = \begin{cases} \frac{|d\mathcal{G}_I(x, I(x))|}{\sqrt{1 - |d\mathcal{G}_I(x, I(x))|^2}} & \text{if } |d\mathcal{G}_I(x, I(x))| < 1 \\ +\infty & \text{if } |d\mathcal{G}_I(x, I(x))| = 1. \end{cases}$$

Again, $x \in D(I)$ is called *critical point* of I if $(x, I(x)) \in \text{epi}(I)$ is a critical point of $|d\mathcal{G}_I|(u, I(u))$.

In the case when I is continuous the above definition is equivalent to Definition 2.1. Indeed, it holds by [DM94, Proposition 2.3], that in this case

$$|d\mathcal{G}_I(x, I(x))| = \begin{cases} \frac{|dI(x)|}{\sqrt{1 + |dI(x)|^2}} & \text{if } |dI(x)| < \infty \\ 1 & \text{if } |dI(x)| = \infty \end{cases} \quad \text{and} \quad |d\mathcal{G}_I(x, \xi)| = 1 \text{ if } I(x) < \xi.$$

Hence, the Definition 2.5 is a generalisation of the weak metric slope from Definition 2.1 to lower semicontinuous functionals. However, this definition is, in general, hard to verify. For this reason, we state without a proof a result from [DM94] that provides a lower bound on $|dI|$.

Proposition 2.6 ([DM94, Proposition 2.5]). *Let $I : \mathcal{X} \rightarrow \mathbb{R} \cup \{+\infty\}$ be a proper, l.s.c functional and for $b \in \mathbb{R}$ let $D(I)_b = \{x \in D(I) : I \leq b\}$. If for some $x \in D(I)$ there exist constants $\delta > 0, b > I(x), \alpha > 0$, a neighbourhood U of x , and a mapping $\Psi : (U \cap D(I)_b) \times [0, \alpha] \rightarrow \mathcal{X}$ such that for all $(u, t) \in (U \cap D(I)_b) \times [0, \alpha]$ it holds that*

$$d(\Psi(u, t), u) \leq t \quad \text{and} \quad I(u) - I(\Psi(u, t)) \geq \delta t.$$

Then, $|dI|(x) \geq \delta$.

Although we can apply the mountain pass Theorem 2.3 to the function \mathcal{G}_I , we do not know if the critical point we obtain is of the form $(x, I(x))$, i.e. we have no information about how $|d\mathcal{G}_I|$ behaves away from $(x, \xi) \in \text{epi}(I)$ such that $\xi > I(x)$. Degiovanni and Marzocchi [DM94] provide some intuition in the case in which I is a functional defined on a Banach space and consists of convex l.s.c part plus a C^1 perturbation. The critical point, in this case, is defined relative to the metric generated by the norm. Another more abstract approach could be to establish a mountain pass theorem for multivalued maps, as indicated in the discussion surrounding Definition 2.4.

In the study of gradient flows, this problem can be treated differently. In Section 4, we show how the notion of λ -convexity ensures that critical points in the sense of Definition 2.5 are actually of the form $(x, I(x))$ (cf. Lemma 4.2). Furthermore, the notion of weak metric slope for l.s.c functions introduced in Definition 2.5 ensures that the critical points obtained in Theorem 1.1 attain the barrier value. Before discussing this in further detail, we first cover some preliminaries on optimal transport on Riemannian manifolds.

3. OPTIMAL TRANSPORT ON MANIFOLDS

Let M be a complete, connected, and smooth Riemannian manifold equipped with a metric given in local coordinates by g_{ij} . We denote the geodesic distance between $x, y \in M$ by $d_M(x, y)$ and the Riemannian volume element by $\text{dvol}(x) = \sqrt{\det g_{ij}(x)} \, dx$ in local coordinates. For $x \in M$, we denote the inner product on the tangent space $T_x M$ by $\langle \cdot, \cdot \rangle$. Let $c(x, y) := d_M(x, y)^2/2$ denote the cost function. We denote by $\mathcal{P}_2(M)$ the space of Borel probability measures on M with finite second moment. Given $\mu, \nu \in \mathcal{P}_2(M)$, the 2-Wasserstein distance, $W_2(\cdot, \cdot)$, between them is defined as

$$W_2(\mu, \nu) := \inf_{\pi \in \Pi(\mu, \nu)} \int_{M \times M} d_M^2(x, y) \, d\pi(x, y),$$

where $\Pi(\mu, \nu)$ is the set of all couplings between μ and ν . We first discuss the existence of minimising geodesics in $\mathcal{P}_2(M)$ in the absence of any regularity assumptions on the initial and final measures. Given a curve $\mu \in C([0, 1]; \mathcal{P}_2(M))$, we define its length to be

$$L(\mu) := \sup_{N \in \mathbb{N}} \sup_{0=t_0 < \dots < t_N=1} \sum_{i=0}^{N-1} W_2(\mu(t_i), \mu(t_{i+1})).$$

We now state the following result from [Vil09, Corollary 7.2.2]:

Proposition 3.1. *The space $\mathcal{P}_2(M)$ is a geodesic metric space, i.e. for any two $\mu_0, \mu_1 \in \mathcal{P}_2(M)$ there exists a minimising geodesic between them. That is to say, there exists a curve $\mu \in C([0, 1]; \mathcal{P}_2(M))$ between μ_0 and μ_1 , such that*

$$W_2(\mu_0, \mu_1) = \min \left\{ L(\bar{\mu}) : \bar{\mu} \in C([0, 1]; \mathcal{P}_2(M)), \bar{\mu}(0) = \mu_0, \bar{\mu}(1) = \mu_1 \right\} = L(\mu).$$

Furthermore, the curve μ can be reparametrised to have unit speed, that is

$$W_2(\mu(t), \mu(s)) = |t - s| W_2(\mu_0, \mu_1),$$

for all $s, t \in [0, 1]$. Such a reparametrised curve $\mu \in C([0, 1]; \mathcal{P}_2(M))$ is called a unit speed minimising geodesic between μ_0 and μ_1 .

Remark 3.2. We will use the above proposition extensively to prove the mountain pass theorem in $\mathcal{P}_2(M)$, i.e. Theorem 1.1. Note, however, that the result of Proposition 3.1 holds even if M is replaced by some \mathcal{M} , where \mathcal{M} is a complete, separable, and locally compact, length space (as is any complete, connected, Riemannian manifold, by the Hopf–Rinow theorem). However, we will refrain from working at this level of generality and will instead remind the reader when any stated results can be generalised to \mathcal{M} .

We now introduce the following definition following [CEMS01].

Definition 3.3. Let A, B be compact subsets of M . The set $\mathcal{I}^c(A, B)$ of *c-concave functions* is the set of functions $\phi : A \rightarrow \mathbb{R} \cup \{-\infty\}$ not identically $-\infty$, for which there exists a function $\psi : B \rightarrow \mathbb{R} \cup \{-\infty\}$ such that

$$\phi(x) = \inf_{y \in B} (c(x, y) - \psi(y)), \quad \forall x \in A.$$

The function ϕ is called the *c-transform* of ψ and abbreviate it as $\phi = \psi^c$.

We have the following main result on the well-posedness of the Monge problem from [McC01].

Theorem 3.4. *Let M be a complete Riemannian manifold. Fix two Borel probability measures $\mu \ll \text{vol}$ and ν on M and two compact subsets $A, B \subset M$ containing the supports of μ and ν , respectively. Then there exists a $\phi \in \mathcal{I}^c(A, B)$ such that the map*

$$F(x) := \exp_x(-\nabla \phi(x)) \quad \text{is a pushforward of } \mu \text{ to } \nu.$$

Furthermore, F is the unique minimiser of the quadratic cost $\int_M c(x, G(x)) d\mu(x)$ among all Borel maps $G : M \rightarrow M$ pushing μ forward to ν apart from variations on sets of μ -measure zero. It follows then that the W_2 transportation distance between μ and ν takes the following form

$$W_2^2(\mu, \nu) = \int_M d_M(x, F(x))^2 d\mu(x).$$

The natural extension on McCann's notion of displacement interpolation [McC97] to the manifold setting is given in the following definition.

Definition 3.5 (Optimal interpolant). Let M be a complete Riemannian manifold. Fix two Borel probability measures $\mu \ll \text{vol}$ and ν on M and two compact subsets $A, B \subset M$ containing the supports of μ and ν , respectively. We define the *optimal interpolant* to be the map $t \mapsto \mu(t)$ for $t \in [0, 1]$ such that $\mu(t) = (F_t)_\# \mu$ and $F_t = \exp_x(-t\nabla\phi(x))$. Here $\phi \in \mathcal{I}^c(\overline{A}, B)$ is the so-called Kantorovich potential between μ and ν from Theorem 3.4.

We are finally in a position to conclude this section with the following results from [CEMS01] about the properties of the optimal interpolant.

Lemma 3.6. *Let M be a complete Riemannian manifold. Fix two Borel probability measures $\mu \ll \text{vol}$ and ν on M and two compact subsets $A, B \subset M$ containing the supports of μ and ν , respectively. Then the following two results hold*

- (a) *Optimality of the optimal interpolant. The map F_t defined in Definition 3.5 is the minimiser of the quadratic cost between $\mu(t)$ and μ among all maps pushing forward μ to $\mu(t)$ for all $t \in [0, 1]$.*
- (b) *Absolute continuity of the interpolant. If μ and ν are compactly supported absolutely continuous with respect to the Riemannian volume, then so is their optimal interpolant $t \rightarrow \mu(t)$ for all $t \in [0, 1]$.*

4. A MOUNTAIN PASS THEOREM IN $\mathcal{P}_2(M)$

We now turn to the question of obtaining a notion of mountain passes for l.s.c functions. We fix our metric space to be $\mathcal{X} = \mathcal{P}_2(M)$, where M is now a complete connected smooth Riemannian manifold, and we equip it with the $d = W_2$ transportation distance which makes it a complete, separable metric space [Vil09]. The functionals under consideration satisfy a geodesic λ -convexity assumption introduced in the following definition.

Definition 4.1 (Geodesic λ -convexity). A proper l.s.c function $I : \mathcal{P}_2(M) \rightarrow \mathbb{R} \cup \{+\infty\}$ is said to be λ -geodesically convex for some $\lambda \in \mathbb{R}$, if for any $\mu_0, \mu_1 \in \mathcal{P}_2(M) \cap D(I)$ it holds that $I(\mu(t))$, where $\mu \in C([0, 1]; \mathcal{P}_2(M))$ is any unit speed minimising geodesic between μ_0 and μ_1 (cf. Proposition 3.1), satisfies

$$I(\mu(t)) \leq (1-t)I(\mu_0) + tI(\mu_1) - \frac{\lambda}{2}t(1-t)W_2^2(\mu_0, \mu_1) \quad \forall t \in [0, 1].$$

The following lemma, whose proof is similar in spirit to [DM94, Theorem 3.13], shows that the weak metric slope of \mathcal{G}_I is non-zero for geodesically λ -convex functionals for $(\mu, \xi) \in \text{epi}(I)$ such that $\xi > I(\mu)$. In particular, any critical point of \mathcal{G}_I , if present, satisfies $\xi = I(\mu)$.

Lemma 4.2. *Let $I : \mathcal{P}_2(M) \rightarrow \mathbb{R} \cup \{+\infty\}$ be a proper, l.s.c, and λ -geodesically convex function. Then, it holds for all $\mu \in \mathcal{P}_2(M)$ and $\xi \in \mathbb{R}$ that*

$$|d\mathcal{G}_I|(\mu, \xi) = 1 \quad \text{if } \xi > I(\mu).$$

In particular, any critical point (μ, ξ) of \mathcal{G}_I satisfies $\xi = I(\mu)$.

Proof. Let $(\mu_1, \xi) \in \text{epi}(I)$ be such that $\xi = I(\mu_1) + 2\varepsilon$ for some $\varepsilon > 0$. We define for any $\delta > 0$ the map $\Psi : B_\delta^{d_{\text{epi}}}(\mu_1, \xi) \times [0, \varepsilon] \rightarrow \text{epi}(I)$ as follows

$$(4.1) \quad \Psi((\mu_0, \alpha), t) = \left(\mu\left(\frac{t}{\Lambda}\right), \alpha - \frac{t}{\Lambda} \left(\alpha - \frac{|\lambda|}{2} W_2^2(\mu_0, \mu_1) - I(\mu_1) \right) \right),$$

where

$$\Lambda = \sqrt{W_2^2(\mu_0, \mu_1) + \left| \left(\alpha - \frac{|\lambda|}{2} W_2^2(\mu_0, \mu_1) - I(\mu_1) \right) \right|^2}$$

and $\mu(\cdot)$ is a unit speed minimising geodesic between μ_0 and μ_1 . We need to first verify that $t/\Lambda \in [0, 1]$. Since $\xi \geq I(\mu_1) + 2\varepsilon$, we find $\delta_0 = \delta_0(\varepsilon)$ such that for all $\delta \in (0, \delta_0)$ it holds

$$(4.2) \quad \varepsilon \leq \xi - I(\mu_1) - \frac{|\lambda|}{2} \delta^2 - 2\delta.$$

The above estimate yields $\Lambda \geq \varepsilon$ implying $\frac{t}{\Lambda} \in [0, 1]$ and so $\mu(\frac{t}{\Lambda})$ is well-defined, provided that $\delta \in (0, \delta_0)$. We also have, from Proposition 3.1, that

$$d_{\text{epi}}\left(\Psi((\mu_0, \alpha), t), (\mu_0, \alpha)\right) = t.$$

Thus, the map Ψ satisfies condition (1) of Definition 2.1. We still have to check that $\Psi((\mu_0, \alpha), t) \in \text{epi}(I)$. Indeed we have from the definition of λ -geodesic convexity

$$\begin{aligned} I\left(\mu\left(\frac{t}{\Lambda}\right)\right) &\leq I(\mu_0) + \frac{t}{\Lambda}(I(\mu_1) - I(\mu_0)) - \frac{\lambda}{2} \frac{t}{\Lambda} \left(1 - \frac{t}{\Lambda}\right) W_2^2(\mu_0, \mu_1) \\ &\leq \alpha - \frac{t}{\Lambda}(\alpha - I(\mu_1)) + \frac{|\lambda|}{2} \frac{t}{\Lambda} W_2^2(\mu_0, \mu_1) \\ &= \alpha - \frac{t}{\Lambda} \left(\alpha - \frac{|\lambda|}{2} W_2^2(\mu_0, \mu_1) - I(\mu_1) \right). \end{aligned}$$

Finally, we can proceed from (4.1) to the following estimate

$$\mathcal{G}_I(\Psi((\mu_0, \alpha), t)) = \alpha - \frac{t}{\Lambda} \left(\alpha - \frac{|\lambda|}{2} W_2^2(\mu_0, \mu_1) - I(\mu_1) \right)$$

$$\leq \mathcal{G}_I((\mu_0, \alpha)) - t \frac{\xi - I(\mu_1) - \delta - \delta^2 \frac{|\lambda|}{2}}{\sqrt{\delta^2 + \left| \xi - I(\mu_1) + \delta + \delta^2 \frac{|\lambda|}{2} \right|^2}}.$$

Thanks to (4.2), we can make δ arbitrarily small and obtain that $|d\mathcal{G}_I|(\mu_1, \xi) \geq 1$ from Definition 2.1 (2). Since $|d\mathcal{G}_I|(\mu_1, \xi) \leq 1$ by Definition 2.4, the result follows. \square

Having showed that the weak metric slope of \mathcal{G}_I is a constant equal to one for all points $(\mu, \xi) \in \text{epi}(I)$ such that $\xi > I(\mu)$, we investigate how the critical points of I defined through the weak metric slope relate to other relevant notions. Specifically, we compare it to the notion of critical point derived from the strong metric slope used in theory of gradient flows [AGS08]. This theory makes rigorous the notion of the Wasserstein gradient discussed in the introduction. We briefly introduce some terminology. Let $I : \mathcal{P}_2(M) \rightarrow \mathbb{R} \cup \{+\infty\}$ be a proper, l.s.c, and λ -geodesically convex.

Definition 4.3 (Absolutely continuous curves). A curve $\mu : [a, b] \subset \mathbb{R} \rightarrow \mathcal{P}_2(M)$ is said to belong to $AC^p([a, b]; \mathcal{P}_2(M))$ for some $p \in [1, +\infty]$ if there exists $m \in L^p([a, b])$ such that

$$(4.3) \quad W_2(\mu(s), \mu(t)) \leq \int_s^t m(r) \, dr, \quad a \leq s \leq t \leq b.$$

If $p = 1$, then μ is said to be an *absolutely continuous curve*.

Theorem 4.4 (Metric derivative). If $\mu : [a, b] \rightarrow \mathcal{P}_2(M)$ is an absolutely continuous curve then the limit

$$|\mu'| (t) = \lim_{s \rightarrow t} \frac{W_2(\mu(s), \mu(t))}{|t - s|},$$

exists for a.e. t and is called the metric derivative of μ . Additionally, $|\mu'| \in L^1([a, b])$ and is admissible as an m in (4.3). In fact it is the minimal admissible m , i.e.

$$|\mu'| (t) \leq m(t)$$

for t a.e. where m satisfies (4.3).

Now we introduce the notion of the (strong) metric slope.

Definition 4.5 (Metric slope). The metric slope $|\partial I|$ of I at $\mu \in \mathcal{P}_2(M)$ is defined as

$$|\partial I|(\mu) = \begin{cases} \limsup_{\nu \rightarrow \mu} \frac{(I(\mu) - I(\nu))_+}{W_2(\mu, \nu)} & \mu \in D(I) \\ +\infty & \text{otherwise.} \end{cases}$$

The metric slope is defined with a positive part, since we are interested in (negative) gradient flows decreasing the energy functional I (see [AGS08, Chapter 10]). Finally, we are in a position to define the notion of a curve of maximal slope.

Definition 4.6 (Curves of maximal slope). A curve $\mu \in AC^2([0, +\infty); \mathcal{P}_2(M))$ is a *curve of maximal slope* of the function I if the following energy dissipation inequality is

satisfied

$$\frac{1}{2} \int_s^t |\mu'|^2(r) \, dr + \frac{1}{2} \int_s^t |\partial\phi|^2(\mu_r) \, dr \leq I(\mu(s)) - I(\mu(t)),$$

for all $0 \leq s \leq t < +\infty$. A curve μ is a *stationary curve* of maximal slope if it is a curve of maximal slope and $\mu(t) = \mu(s)$ for all $s, t \in [0, +\infty)$.

We have the following straightforward corollary.

Corollary 4.7. *A curve of maximal slope μ of a function I is stationary if and only if $|\partial I|(\mu) = 0$.*

Using all these notions we can finally compare the weak metric slope with the metric slope.

Lemma 4.8 (Equivalence of the two notions of slope). *Let $I : \mathcal{P}_2(M) \rightarrow \mathbb{R} \cup \{+\infty\}$ be a proper, l.s.c, and λ -geodesically convex functional. Then for $\mu \in D(I)$ it holds that $|dI|(\mu) = |\partial I|(\mu)$.*

Proof. We first show that $|dI|(\mu) \leq |\partial I|(\mu)$. Let \mathcal{G}_I be the continuous extension to the epigraph and let Ψ be a δ -regularity mapping for the point $(\mu, I(\mu))$ with $\mu \in D(I)$, that is by Definition 2.1

$$\frac{I(\mu) - \Psi((\mu, I(\mu)), t)}{d_{\text{epi}}((\mu, I(\mu)), \Psi((\mu, I(\mu)), t))} \geq \delta.$$

At the same time, we can choose $\Psi((\mu, I(\mu)), t)$ as the approximating sequence in Definition 4.5 of the strong metric slope and obtain

$$|\partial \mathcal{G}_I|(\mu, I(\mu)) \geq \delta.$$

Taking the supremum over all such δ , we have the bound $|\partial \mathcal{G}_I|(\mu, I(\mu)) \geq |d \mathcal{G}_I|(\mu, I(\mu))$. This yields only the comparison of the two different slopes of the epigraph extension \mathcal{G}_I . To obtain the comparison of the slopes of the functional I itself, we first assume that $\mu \in D(I)$ is not a local minimum, and $|\partial I|(\mu) < +\infty$. Then there exists a sequence $(\nu_n, I(\nu_n)) \in \text{epi}(I)$ such that it converges to $(\mu, I(\mu))$ and such that $I(\nu_n) \leq I(\mu)$ for all n sufficiently large. Using this as the approximating sequence in Definition 4.5, we obtain

$$\begin{aligned} |\partial \mathcal{G}_I|(\mu, I(\mu)) &= \limsup_{(\nu_n, I(\nu_n)) \rightarrow (\mu, I(\mu))} \frac{(I(\mu) - I(\nu_n))_+}{\sqrt{|I(\mu) - I(\nu_n)|^2 + W_2^2(\mu, \nu_n)}} \\ &= \limsup_{(\nu_n, I(\nu_n)) \rightarrow (\mu, I(\mu))} \frac{(I(\mu) - I(\nu_n))_+}{\sqrt{(I(\mu) - I(\nu_n))_+^2 + W_2^2(\mu, \nu_n)}} \\ &= \frac{|\partial I|(\mu)}{\sqrt{1 + |\partial I|(\mu)^2}} \end{aligned}$$

When $\mu \notin D(I)$ or μ is local minimum both $|\partial I|(\mu)$ and $|\partial \mathcal{G}_I|(\mu)$ are $+\infty$ and 0 respectively. Using Definition 2.5 of the weak metric slope, we have $|dI|(\mu) \leq |\partial I|(\mu)$.

To prove the other inequality, we first assume that $|\partial I|(\mu) =: \varepsilon_0 > 0$. Then, for any $\varepsilon \in (0, \varepsilon_0)$ exists $\delta = \delta(\varepsilon) > 0$ by Definition 4.5 such that there exists $\nu \in B_\delta(\mu)$ with

$$I(\nu) < I(\mu) - \varepsilon W_2(\mu, \nu).$$

Choose such a ν and set $\delta' = W_2(\mu, \nu) < \delta$. Since I is l.s.c, we find for any $n \in \mathbb{N}, n \geq 2$ some $\alpha = \alpha(\delta', n) > 0$ such that for all $\eta \in B_\alpha(\mu)$ it holds that

$$I(\mu) - I(\eta) \leq \frac{\delta'}{n}.$$

We define $\alpha' = \alpha'(\delta', n) = \min\{\alpha, \delta'/n\}$ and define a map $\Psi : B_{\alpha'}(\mu) \times [0, \alpha'] \rightarrow \mathcal{P}_2(M)$ as follows

$$\Psi(\eta, t) = \gamma_{\eta, \nu} \left(\frac{t}{W_2(\nu, \eta)} \right)$$

where $\gamma_{\eta, \nu}(\cdot)$ is any unit speed minimising geodesic between η and ν (cf. Proposition 3.1). Again, we have to check that $0 \leq t/W_2(\nu, \eta) \leq 1$. We have from the definition of α' by the triangle inequality

$$\frac{n-1}{n} \delta' \leq -W_2(\mu, \eta) + W_2(\mu, \nu) \leq W_2(\nu, \eta) \leq W_2(\mu, \eta) + W_2(\mu, \nu) \leq \frac{n+1}{n} \delta'.$$

Thus, it follows that $0 \leq t/W_2(\eta, \nu) \leq 1$. Also, by construction holds $W_2(\eta, \Psi(\eta, t)) = t$. Now, by the λ -geodesically convexity of I , we obtain the following estimate

$$\begin{aligned} I(\Psi(\eta, t)) &\leq I(\eta) + \frac{t}{W_2(\eta, \nu)} (I(\nu) - I(\eta)) + \frac{|\lambda|}{2} t \left(1 - \frac{t}{W_2(\eta, \nu)} \right) W_2(\eta, \nu) \\ &\leq I(\eta) + \frac{t}{W_2(\eta, \nu)} (I(\nu) - I(\mu)) + \frac{t}{W_2(\eta, \nu)} (I(\mu) - I(\eta)) + \frac{|\lambda|}{2} t \delta' \frac{n+1}{n} \\ &< I(\eta) + t \left(-\varepsilon \left(\frac{n}{n+1} \right) + \frac{1}{n-1} + \delta' |\lambda| \right). \end{aligned}$$

We can pick $\delta' > 0$ to be as small and n as large as we want and conclude $I(\Psi(\eta, t)) \leq I(\eta) - \varepsilon t$. It follows from Proposition 2.6 that $|dI|(\mu) \geq \varepsilon$. Thus, since $\varepsilon \in (0, \varepsilon_0)$ is arbitrary, we have that $|dI|(\mu) \geq \varepsilon_0 = |\partial I|(\mu)$ for all positive values. For the case in which $|dI|(\mu) = 0$, assume that $|\partial I|(\mu) = \varepsilon_0 > 0$. But we have shown that $|\partial I|(\mu) = |\partial I|(\mu) = \varepsilon_0 > 0$ which would be a contradiction. Thus, we have $|\partial I|(\mu) = |dI|(\mu)$. \square

Proposition 4.9. *Let $I : \mathcal{P}_2(M) \rightarrow \mathbb{R} \cup \{+\infty\}$ be a proper, l.s.c, and λ -geodesically convex functional. Then $\text{epi}(I)$ is complete and path-connected.*

Proof. Since $\mathcal{P}_2(M) \times \mathbb{R}$ is complete, we have for any convergent sequence $(\mu_n, \xi_n) \in \text{epi}(I)$ converging to some $(\mu, c) \in \mathcal{P}_2(M) \times \mathbb{R}$, that $I(\mu) \leq \liminf I(\mu_n) \leq \liminf \xi_n = c$. Thus, $\text{epi}(I) \subset \mathcal{P}_2(M) \times \mathbb{R}$ is closed and thus complete.

Let $(\mu_0, \alpha), (\mu_1, \beta) \in \text{epi}(I)$. Then $(\mu(\cdot), (1-t)\alpha + t\beta - \frac{\lambda}{2}t(1-t)W_2^2(\mu_0, \mu_1))$, where $\mu \in C([0, 1]; \mathcal{P}_2(M))$ is any unit speed minimising geodesic between μ_0 and μ_1 (cf. Proposition 3.1), is a continuous path (with respect to d_{epi}) between them which lies entirely in $\text{epi}(I)$. \square

We conclude this section with the proof of Theorem 1.1.

Proof of Theorem 1.1. Denote by Γ_{epi} the set of all continuous curves $\gamma_{\text{epi}} : [0, 1] \rightarrow \text{epi}(I)$ with $\gamma_{\text{epi}}(0) = (\mu, I(\mu))$ and $\gamma_{\text{epi}}(1) = (\nu, I(\nu))$. We can identify any $\gamma_{\text{epi}} \in \Gamma_{\text{epi}}$ with a $\tilde{\gamma} \in \Gamma$ by projecting onto the first factor, i.e. $(t \mapsto (\mu(t), \xi(t))) \mapsto (t \mapsto \mu(t))$. Because of the definition of the epigraph and \mathcal{G}_I , we have that

$$(4.4) \quad \inf_{\gamma_{\text{epi}} \in \Gamma_{\text{epi}}} \max_{t \in [0, 1]} \mathcal{G}_I(\gamma_{\text{epi}}(t)) \geq \inf_{\gamma_{\text{epi}} \in \Gamma_{\text{epi}}} \sup_{t \in [0, 1]} I(\tilde{\gamma}(t)) \geq \inf_{\gamma \in \Gamma} \sup_{t \in [0, 1]} I(\gamma(t)) = c.$$

Now, we prove the inequality holds the other way as well. Note that, for every $\varepsilon > 0$, there must exist some $\gamma \in \Gamma$ such that

$$c \leq \sup_{t \in [0, 1]} I(\gamma(t)) \leq c + \frac{\varepsilon}{2}.$$

Consider a partition $\mathfrak{P}_\delta = \{t_i\}_{i=0, \dots, N}$ of $[0, 1]$ having mesh size $\delta > 0$. Consider the family of curves $\{\gamma^{\delta, i}\}_{i=0, \dots, N-1} \subset C([0, 1]; \mathcal{P}_2(M))$ associated to the partition \mathfrak{P}_δ , where $\gamma^{\delta, i}(\tau) = \mu^i(\tau)$, $\tau \in [0, 1]$ and μ^i is a unit speed minimising geodesic between $\gamma(t_i)$ and $\gamma(t_{i+1})$. We can now construct a family of curves $\{(\gamma^{\delta, i}, \xi^{\delta, i})\}_{i=0, \dots, N-1} \subset C([0, 1]; \text{epi}(I))$ such that

$$\xi^{\delta, i}(\tau) := (1-\tau)I(\gamma(t_i)) + \tau(I(\gamma(t_{i+1}))) + \frac{|\lambda|}{2}\tau(1-\tau)W_2^2(\gamma(t_i), \gamma(t_{i+1})).$$

The fact that I is λ -geodesically convex ensures that $(\gamma^{\delta, i}(\tau), \xi^{\delta, i}(\tau)) \in \text{epi}(I)$ for all $\tau \in [0, 1]$ and $i = 0, \dots, N-1$. We can now concatenate these curves as follows

$$\gamma_{\text{epi}}^\delta(t) = (\gamma^\delta(t), \xi^\delta(t)) = (\gamma^{\delta, i}(Nt - i), \xi^{\delta, i}(Nt - i)) \quad i/N \leq t < (i+1)/N,$$

such that the curve $\gamma_{\text{epi}}^\delta \in \Gamma_{\text{epi}}$. It follows then that

$$\begin{aligned} \max_{t \in [0, 1]} \mathcal{G}_I(\gamma_{\text{epi}}^\delta) &\leq \sup_{t \in [0, 1]} I(\gamma(t)) + \max_{i=0, \dots, N-1} \frac{|\lambda|}{2} W_2^2(\gamma(t_i), \gamma(t_{i+1})) \\ &\leq c + \frac{\varepsilon}{2} + \max_{i=0, \dots, N-1} \frac{|\lambda|}{2} W_2^2(\gamma(t_i), \gamma(t_{i+1})). \end{aligned}$$

Note that the curve $\gamma \in C([0, 1]; \mathcal{P}_2(M))$ is uniformly continuous, since $[0, 1]$ is compact. Thus, we can find a $\delta > 0$ small enough such that

$$\frac{|\lambda|}{2} W_2^2(\gamma(s), \gamma(t)) \leq \frac{\varepsilon}{2},$$

for all $|s - t| \leq \delta$ and $s, t \in [0, 1]$. Thus, we have that

$$\max_{t \in [0, 1]} \mathcal{G}_I(\gamma_{\text{epi}}^\delta) \leq c + \varepsilon.$$

Using the inequality in (4.4), it follows that

$$c \leq \inf_{\gamma_{\text{epi}} \in \Gamma_{\text{epi}}} \max_{t \in [0, 1]} \mathcal{G}_I(\gamma_{\text{epi}}(t)) \leq c + \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, it follows that

$$\inf_{\gamma_{\text{epi}} \in \Gamma_{\text{epi}}} \max_{t \in [0, 1]} \mathcal{G}_I(\gamma_{\text{epi}}(t)) = c = \inf_{\gamma \in \Gamma} \sup_{t \in [0, 1]} I(\gamma(t)).$$

Also we have that $c_1 = \max\{I(\mu), I(\nu)\} = \max\{\mathcal{G}_I(\mu, I(\mu)), \mathcal{G}_I(\nu, I(\nu))\}$ and from the above identity that $\inf_{\gamma_{\text{epi}} \in \Gamma_{\text{epi}}} \max_{t \in [0, 1]} \mathcal{G}_I(\gamma_{\text{epi}}(t)) = c > c_1$. Furthermore, if I satisfies (MPS), it follows that \mathcal{G}_I satisfies it as well. Let (μ_n, ξ_n) be a Palais sequence. Since $|d\mathcal{G}_I|(\mu_n, \xi_n) \rightarrow 0$ it follows from Lemma 4.2 that for n large enough the sequence must be of the form $(\mu_n, I(\mu_n))$ and that $|dI|(\mu_n) \rightarrow 0$. Since $\mathcal{G}_I((\mu_n, I(\mu_n))) = I(\mu_n) \rightarrow c$, it follows that μ_n is a Palais sequence for I . Thus, we can construct a subsequence which converges to some $\mu^* \in \mathcal{P}_2(M)$ and by extension to $(\mu^*, c) \in \text{epi}(I)$. Finally we can apply Theorem 2.3 to \mathcal{G}_I to extract the existence of a critical point $(\eta, c) \in \text{epi}(I)$ such that $|d\mathcal{G}_I|(\eta, c) = 0$ and $\mathcal{G}_I((\eta, c)) = c = \inf_{\gamma_{\text{epi}} \in \Gamma_{\text{epi}}} \Upsilon(\gamma)$. However, the contraposition of Lemma 4.2 implies that $c = I(\eta)$ if $|d\mathcal{G}_I|(\eta, c) < 1$, from which it follows that $|dI|(\eta) = |d\mathcal{G}_I|(\eta, I(\eta)) = 0$. Thus, η is critical point of I with critical value c . Also, since I is λ -geodesically convex it follows from Lemma 4.8 that $|\partial I|(\eta) = 0$. \square

Remark 4.10. We remark that a similar regularisation argument to the one used in the above proof, i.e. using the curves γ^δ , can also be used to prove that

$$\inf_{\gamma \in \Gamma} \sup_{t \in [0, 1]} I(\gamma(t)) = \inf_{\gamma \in \Gamma_{AC}} \max_{t \in [0, 1]} I(\gamma(t)),$$

where $\Gamma_{AC} = \Gamma \cap AC([0, 1]; \mathcal{P}_2(M))$.

Remark 4.11. Since all we have used in Definition 4.1, Lemma 4.2, Lemma 4.8, and Proposition 4.9, is the existence of a unit speed minimising geodesic between two points μ and ν , it follows that these results hold true if M is replaced by \mathcal{M} , a complete, separable, and locally compact length space. Furthermore, the abstract mountain pass theorem, i.e. Theorem 1.1, continues to hold true if $\mathcal{P}_2(M)$ is replaced by $\mathcal{P}_2(\mathcal{M})$.

5. APPLICATION TO THE MCKEAN-VLASOV MODEL

This section is devoted to the analysis of the McKean-Vlasov free energy I (1.4). As an immediate consequence of Theorem 1.1, we obtain Theorem 1.2.

Proof of Theorem 1.2. The functional $I : \mathcal{P}(\mathbb{T}_L^d) \rightarrow \mathbb{R}$ is proper and l.s.c and since $\|D^2W\|_{L^\infty(\mathbb{T}_L^d)} \leq C$ it is also λ -geodesically convex. The space $\mathcal{P}(\mathbb{T}_L^d)$ is compact and thus I trivially satisfies (MPS). Since μ_0 is a strict local minimum of I , it follows that there exists an $R > 0$, such that, for all $0 < r < R$, μ_0 is the unique minimiser of I in $B_r^{W_2}(\mu_0)$. Thus, we have that $I(\mu) > I(\mu_0)$ for all $\mu \in \partial B_r^{W_2}(\mu_0)$, $0 < r < R$. Since $\partial B_r^{W_2}(\mu)$ is compact (because $\mathcal{P}(\mathbb{T}_L^d)$ equipped with the W_2 metric is compact) and I is l.s.c, it follows that the minimum of $I(\mu) - I(\mu_0)$ must be attained by some $\mu_2^r \in \partial B_r^{W_2}(\mu_0)$. Thus, we have that

$$I(\mu) - I(\mu_0) \geq I(\mu_2^r) - I(\mu_0) =: \delta(r) > 0,$$

for all $\mu \in \partial B_r^{W_2}(\mu_0)$ and all $0 < r < R$. Let us set $r < \min\{R, W_2(\mu_0, \mu_1)\}$. Since any curve $\gamma \in \Gamma$ must pass through $\partial B_r^{W_2}(\mu_0)$, it follows that

$$\inf_{\gamma \in \Gamma} \sup_{t \in [0,1]} I(\gamma(t)) \geq I(\mu_0) + \delta(r) = \max\{I(\mu_0), I(\mu_1)\} + \delta(r).$$

□

At this level of generality, one still needs to find at least one strict local minimum to apply Theorem 1.2. Hence, as a next step, we would like to provide conditions on the interaction potential W and the parameter values β at which we can find two measures μ_0 and μ_1 which satisfy the assumptions of Theorem 1.2. We do this by using the results of [CGPS20] to argue that one can find potentials W and parameter values β_c such that the free energy I has two distinct minimisers and one of them, μ^L , is a strict local minimum of I . Such parameter values are referred to as discontinuous transition points of I (cf. Definition 5.1). In the second part, we formulate the large deviations results and complete the proof of Theorem 1.4. Let us recall some of the main definitions and results about the free energy functional from [CP10] and [CGPS20].

Definition 5.1 (Transition point). A parameter value $\beta_c > 0$ is said to be a *transition point* of I from the uniform measure $\mu^L(dx) = dx/L^d$ if it satisfies the following conditions:

- (1) For $0 < \beta < \beta_c$, μ^L is the unique minimiser of I .
- (2) For $\beta = \beta_c$, μ^L is a minimiser of I .
- (3) For $\beta > \beta_c$, there exists $\mu_\beta \in \mathcal{P}(\mathbb{T}_L^d) \setminus \{\mu^L\}$, such that μ_β is a minimiser of I .

Additionally, a transition point $\beta_c > 0$ is said to be a *continuous transition point* of I if:

- (1) For $\beta = \beta_c$, μ^L is the unique minimiser of I .
- (2) Given any family of minimisers $\{\mu_\beta | \beta > \beta_c\}$, we have that

$$\limsup_{\beta \downarrow \beta_c} \|\mu_\beta - \mu^L\|_{TV} = 0.$$

A transition point β_c which is not continuous is said to be *discontinuous*.

In thermodynamics, continuous phase transitions correspond to second-order ones similar to those seen in the theory of magnetisation and spin systems [Daw83, Shi87, GP18], whereas discontinuous phase transitions correspond to first-order ones similar to the ones observed in nucleation processes or phase transformation from liquid to vapour [LP66].

One can show that if β_c is discontinuous, i.e. if either one of the two conditions in Definition 5.1 are violated, then it must be the case that condition (1) is violated (cf. Theorem 5.2 (b)). This is the key idea we will use to obtain a set of conditions under which we can apply the result of Theorem 1.2. The original statement of these definitions and the proof of the above statement can be found in [CP10]. We summarise the main results about the free energy functional in the theorem below. The proofs can be found in [CP10] and [CGPS20, Theorems 5.11 and 5.19]. The conditions are expressed in terms of the Fourier coefficients of W denoted by]

(5.1)

$$\hat{W}(k) = \int_{\mathbb{T}_L^d} e_k(x) W(x) dx \quad \text{with} \quad e_k = L^{-d/2} \exp\left(\frac{2\pi i}{L} k \cdot x\right) \quad \text{for} \quad k \in \mathbb{Z}^d.$$

Theorem 5.2. *Assume that $W \in C^2(\mathbb{T}_L^d)$ and $\beta > 0$.*

- (a) *The free energy function $I : \mathcal{P}(\mathbb{T}_L^d) \rightarrow \mathbb{R} \cup \{+\infty\}$ always has a minimiser $\mu \in \mathcal{P}(\mathbb{T}_L^d)$ such that $\mu \ll dx$ with a positive and smooth density.*
- (b) *If there exists $k \in \mathbb{Z}^d \setminus \{0\}$ such that $\hat{W}(k) < 0$, then there exists a $\beta_c > 0$ such that β_c is a transition point of I . Furthermore, if the transition point $\beta_c > 0$ is discontinuous, then there exist at least two minimisers of the free energy I over $\mathcal{P}(\mathbb{T}_L^d)$ at $\beta = \beta_c$, such that one is $\mu^L = L^{-d} dx$ and the other is some $\bar{\mu} \in \mathcal{P}(\mathbb{T}_L^d)$. Additionally, if β_c is discontinuous then $\beta_c < \frac{L^{d/2}}{|\min_{k \in \mathbb{Z}^d, k \neq 0} \hat{W}(k)|}$.*
- (c) *For W with $\beta_c > 0$ a transition point, let the set K^δ be given for any $\delta > 0$ by*

$$K^\delta := \left\{ k \in \mathbb{Z}^d, k \neq 0 : \hat{W}(k) \leq \min_{k \in \mathbb{Z}^d, k \neq 0} \hat{W}(k) + \delta \right\}.$$

Let $\delta_ > 0$ be the smallest δ for which there exist distinct $k^a, k^b, k^c \in K^{\delta_*}$ with $k^a = k^b + k^c$, if such points exist, else $\delta_* = \infty$. If δ_* is sufficiently small, then β_c is a discontinuous transition point.*

- (d) *Let $\{W_n\}_{n \in \mathbb{N}} \in C^2(\mathbb{T}_L^d)$, with $\beta_{c,n} > 0$ the associated transition points, be a sequence of interaction potentials such that $\delta_* \rightarrow 0$ as $n \rightarrow \infty$. Assume there exists $N \in \mathbb{N}$ and a positive constant $C > 0$ such that for all $n > N$, $|\min_{k \in \mathbb{Z}^d, k \neq 0} \hat{W}_n(k)| > C\delta_*^\gamma$ for any $\gamma < \frac{1}{2}$. Then for n sufficiently large, $\beta_{c,n}$ is a discontinuous transition point and $\beta_{c,n} < \frac{L^{d/2}}{|\min_{k \in \mathbb{Z}^d, k \neq 0} \hat{W}_n(k)|}$.*

The above result provides conditions when to expect a discontinuous transition point. The case of the discontinuous transition point is particularly interesting for us as it implies the existence of a parameter value β_c at which there are two distinct minimisers and hints at a possible scenario in which the mountain pass theorem could be applied. To provide more intuition we show in the following lemma that any potential that under rescaling localises sufficiently fast but loses mass sufficiently slow will eventually exhibit a discontinuous transition point for the associated free energy I .

Lemma 5.3. *Let $W \in C^2(\mathbb{T}_L^d)$ be a compactly supported interaction potential with support strictly contained in \mathbb{T}_L^d and $\int_{\mathbb{T}_L^d} W \, dx < 0$. Assume further, that for some $\epsilon_1 > 0$ and all $\epsilon \in (0, \epsilon_1]$, it holds that*

$$(5.2) \quad L^{-d/2} \int_{\mathbb{T}_L^d} W(x) e^{i \frac{2\pi \epsilon k \cdot x}{L}} \, dx \geq L^{-d/2} \int_{\mathbb{T}_L^d} W \, dx := -C \quad \text{for all } k \in \mathbb{Z}^d.$$

Consider the rescaled potential, $W_\epsilon(x) = f(\epsilon)W(x/\epsilon)$ and positive function $f : (0, \epsilon_1] \rightarrow \mathbb{R}_+$. If $\epsilon^\ell \lesssim f(\epsilon) \lesssim \epsilon^m$ as $\epsilon \rightarrow 0$ for $m > -d - 2$, $\ell \geq -d$, $\ell < \frac{m-d}{2} + 1$ (along with the natural restriction $\ell \geq m$), then for ϵ small enough, the associated free energy I possesses a discontinuous transition point at some $\beta_c < \frac{L^{d/2}}{|\min_{k \in \mathbb{Z}^d, k \neq 0} \hat{W}_\epsilon(k)|}$.

Proof. We proceed by checking that the conditions of Theorem 5.2(d) hold for this class of potentials. We first check that for ϵ small enough, W_ϵ has at least one negative Fourier mode. Let $V := \text{supp } W$ and $V_\epsilon := \text{supp } W_\epsilon$. We have for $k \in \mathbb{Z}^d$,

$$(5.3) \quad \begin{aligned} \hat{W}_\epsilon(k) &= L^{-d/2} \int_{\mathbb{T}_L^d} W_\epsilon(x) e^{i \frac{2\pi k \cdot x}{L}} \, dx \\ &= L^{-d/2} f(\epsilon) \int_{V_\epsilon} W(x/\epsilon) e^{i \frac{2\pi k \cdot x}{L}} \, dx \\ &= L^{-d/2} f(\epsilon) \epsilon^d \int_V W(x) e^{i \frac{2\pi \epsilon k \cdot x}{L}} \, dx. \end{aligned}$$

Since $e^{i \frac{2\pi \epsilon k \cdot x}{L}} \rightarrow 1$ uniformly on V as $\epsilon \rightarrow 0$, it follows that eventually $\hat{W}_\epsilon(k) < 0$ for ϵ sufficiently small since $\int_V W(x) e^{i \frac{2\pi \epsilon k \cdot x}{L}} \, dx \rightarrow \left(\int_{\mathbb{T}_L^d} W \, dx \right) < 0$. Using (5.2) and (5.3), we can now obtain the following bound

$$\min_{k \in \mathbb{Z}^d, k \neq 0} \hat{W}_\epsilon(k) \geq -C f(\epsilon) \epsilon^d.$$

Since W is even along every coordinate we have that

$$\begin{aligned} \int_{\mathbb{T}_L^d} W(x) e^{i \frac{2\pi \epsilon k \cdot x}{L}} \, dx &= \int_{\mathbb{T}_L^d} W(x) \cos\left(\frac{2\pi \epsilon k \cdot x}{L}\right) \, dx \\ &= \int_{\mathbb{T}_L^d} W(x) \left(1 + \left(\frac{2\pi |k| \epsilon x}{L} \right)^2 + O(\epsilon^4) \right) \, dx \end{aligned}$$

Fix some $k^* \in \mathbb{Z}^d$. The above expansion tells us that we can find some ϵ_1 sufficiently small and some $C_1 > 0$ independent of ϵ such that

$$\hat{W}_\epsilon(k^*), \hat{W}_\epsilon(2k^*) \leq f(\epsilon)\epsilon^d(-C + C_1\epsilon^2) \quad \text{for all } \epsilon < \epsilon_1.$$

We can thus obtain the following bound

$$(5.4) \quad -Cf(\epsilon)\epsilon^d \leq \min_{k \in \mathbb{Z}^d, k \neq 0} \hat{W}_\epsilon(k) \leq f(\epsilon)\epsilon^d(-C + C_1\epsilon^2) \quad \text{for all } \epsilon < \epsilon_1.$$

Combining the two of them we derive

$$\left. \begin{aligned} \hat{W}_\epsilon(k^*) - \min_{k \in \mathbb{Z}^d, k \neq 0} \hat{W}_\epsilon(k) &\leq C_1 f(\epsilon) \epsilon^{2+d} \\ \hat{W}_\epsilon(2k^*) - \min_{k \in \mathbb{Z}^d, k \neq 0} \hat{W}_\epsilon(k) &\leq C_1 f(\epsilon) \epsilon^{2+d} \end{aligned} \right\} \quad \text{for all } \epsilon < \epsilon_1,$$

which tells us that $k^*, 2k^* \in K^{C_1 f(\epsilon) \epsilon^{2+d}}$ and that $\delta_* \leq C_1 f(\epsilon) \epsilon^{2+d}$. Thus, $\delta_* \lesssim \epsilon^{m+d+2}$ and since $m > -d-2$, $\delta_* \rightarrow 0$ as $\epsilon \rightarrow 0$. Furthermore, using (5.4) we can deduce

$$\left| \min_{k \in \mathbb{Z}^d, k \neq 0} \hat{W}_\epsilon(k) \right| \geq f(\epsilon)\epsilon^d(C - C_1\epsilon^2) \geq C_2 \epsilon^{\ell+d}.$$

The fact that $\ell \geq -d$ tells us that

$$\left| \min_{k \in \mathbb{Z}^d, k \neq 0} \hat{W}_\epsilon(k) \right| \geq C_3 \delta_*^{\frac{\ell+d}{m+d+2}}.$$

We now use the assumption that $l < \frac{m-d}{2} + 1$ and apply Theorem 5.2(d), to obtain the desired result. \square

The choice $f(\epsilon) = \epsilon^{-d}$ is an admissible scaling for the function f in Lemma 5.3. Now that we have a set of concrete conditions under which we can expect there to be two distinct minimisers at a particular parameter value, we can try to apply the mountain pass theorem. To apply Theorem 1.1, it is sufficient to show that μ^L is a strict local minima at parameter values β_c and that this property is uniform, i.e. we can find a ball $B_r^{W_2}(\mu^L)$ around μ^L in W_2 such that $I(\mu) \geq I(\mu^L) + \delta$ for all $\mu \in \partial B_r^{W_2}(\mu^L)$ and some $\delta > 0$. In order to show this, we need the following comparison of W_2 with the homogeneous negative Sobolev space $\dot{H}^{-1}(\mathbb{T}_L^d)$, which we identify with all formal Fourier series of μ as defined in (5.1) given by $\sum_{k \in \mathbb{Z}^d \setminus \{0\}} \hat{\mu}(k) e_k$ such that $\sum_{k \in \mathbb{Z}^d \setminus \{0\}} \frac{1}{|k|^2} |\hat{\mu}(k)|^2 < \infty$. Note that the functions $\{e_k\}_{k \in \mathbb{Z}^d \setminus \{0\}}$ form an orthogonal basis for $\dot{H}^{-1}(\mathbb{T}_L^d)$ with respect to the inner product defined by duality with the homogeneous space $\dot{H}^1(\mathbb{T}_L^d)$. We have that $\langle \mu, f \rangle_{\dot{H}^{-1}, \dot{H}^1} = \sum_{k \in \mathbb{Z}^d \setminus \{0\}} \hat{\mu}(k) \hat{f}(-k)$. Also, the inner product on $\dot{H}^1(\mathbb{T}_L^d)$ is defined as $(f, g)_{\dot{H}^1} = \sum_{k \in \mathbb{Z}^d \setminus \{0\}} |k|^2 \hat{f}(k) \hat{g}(-k)$. It is easy to check then that the Riesz representation of any $\mu \in \dot{H}^{-1}(\mathbb{T}_L^d)$ is given by $\sum_{k \in \mathbb{Z}^d \setminus \{0\}} \frac{1}{|k|^2} \hat{\mu}(k) e_k \in \dot{H}^1(\mathbb{T}_L^d)$.

Lemma 5.4 (Comparison of \dot{H}^{-1} with W_2). *Let $\mu_0, \mu_1 \in \mathcal{P}(\mathbb{T}_L^d) \cap L^\infty(\mathbb{T}_L^d)$. Then the following estimate holds*

$$\|\mu_0 - \mu_1\|_{\dot{H}^{-1}(\mathbb{T}_L^d)} \leq \left(\max \left[\|\mu_0\|_{L^\infty(\mathbb{T}_L^d)}, \|\mu_1\|_{L^\infty(\mathbb{T}_L^d)} \right] \right)^{1/2} W_2(\mu_0, \mu_1)$$

Proof. The proof follows the argument in [Loe06, Proposition 2.1]. Let $\mu(\cdot)$ be the optimal interpolant between μ_0 and μ_1 from Theorem 3.4. Then by the Benamou–Brenier formulation of the optimal transport problem, there exists a vector field $[0, 1] \ni t \mapsto v(t) \in L^2(\mu(t); \mathbb{R}^d)$ such that the pair $(\mu(t), v(t))$ satisfies

$$\partial_t \mu + \nabla \cdot (\mu v) = 0, \quad \text{for } t \in [0, 1],$$

in the sense of distributions. Now, we consider the sequence of parameterised problems given by

$$\Delta \Psi_t = \mu(t) - L^{-d} \quad \text{for } t \in [0, 1].$$

Note that $\|\mu(t)\|_{L^\infty(\mathbb{T}_L^d)} \leq \max \left\{ \|\mu_0\|_{L^\infty(\mathbb{T}_L^d)}, \|\mu_1\|_{L^\infty(\mathbb{T}_L^d)} \right\}$ [Vil09, Corollary 17.19], and thus the above equation has a unique weak solution in $\dot{H}^1(\mathbb{T}_L^d)$ for all $t \in [0, 1]$. We know that $\int_{\mathbb{T}_L^d} |v(t)|^2 \mu(t) dx = W_2^2(\mu(t), \mu_0) = t^2 W_2^2(\mu_0, \mu_1) < \infty$. From this it follows that $\mu(t)v(t) \in L^2(\mathbb{T}_L^d; \mathbb{R}^d)$ and thus $\nabla \cdot (\mu(t)v(t)) \in \dot{H}^{-1}(\mathbb{T}_L^d)$. Differentiating with respect to t we have

$$\Delta \partial_t \Psi_t = -\nabla \cdot (\mu(t)v(t)).$$

It follows then that $\partial_t \Psi_t \in \dot{H}^1(\mathbb{T}_L^d)$. Multiplying by $\partial_t \Psi_t$ and integrating by parts with respect to the space variable and then integrating with respect to time, we obtain

$$\begin{aligned} \|\nabla \Psi_1 - \nabla \Psi_0\|_{L^2(\mathbb{T}_L^d)} &\leq \|\mu(t)\|_\infty^{1/2} W_2(\mu_0, \mu_1) \\ &\leq \left(\max \left[\|\mu_0\|_{L^\infty(\mathbb{T}_L^d)}, \|\mu_1\|_{L^\infty(\mathbb{T}_L^d)} \right] \right)^{1/2} W_2(\mu_0, \mu_1) \end{aligned}$$

Since Ψ_t is precisely the Riesz representation of $\mu(t)$ in $\dot{H}^1(\mathbb{T}_L^d)$, the claimed estimate holds. \square

Remark 5.5. For $d = 1$, we remark that the $\dot{H}^{-1}(\mathbb{T})$ -norm and $W_2(\cdot, \cdot)$ -distance are comparable in both directions. Indeed, from the Kantorovich–Rubinstein dual formulation of the W_1 distance, we have that

$$\begin{aligned} W_1(\mu, \nu) &= \sup_{\text{Lip}(\varphi) \leq 1} \int_{\mathbb{T}} \varphi d(\mu - \nu) \\ &\leq L^{1/2} \sup_{\|\varphi\|_{\dot{H}^1(\mathbb{T})} \leq 1} \int_{\mathbb{T}} \varphi d(\mu - \nu) = L^{1/2} \|\mu - \nu\|_{\dot{H}^{-1}(\mathbb{T})}. \end{aligned}$$

Furthermore, since \mathbb{T} is compact, we have that $W_2(\mu, \nu) \leq C\sqrt{W_1(\mu, \nu)}$. Thus, we have that

$$W_2(\mu, \nu) \leq C_1 \|\mu - \nu\|_{\dot{H}^{-1}(\mathbb{T})}^{1/2}.$$

Furthermore, by using [MM13, Lemma 4.1 (ii) and (iii)] we obtain that for $d = 1$,

$$\|\mu - \nu\|_{\dot{H}^{-1}(\mathbb{T})} \leq C_2 \sqrt{W_2(\mu, \nu)}.$$

Thus, for $d = 1$, the argument in the proof Lemma 5.4 is not needed.

The following lemma establishes the strictness of local minima in W_2 for discontinuous transition points.

Lemma 5.6. *Assume $W \in C^2(\mathbb{T}_L^d)$ with $\beta_c > 0$ a discontinuous transition point. Then, for $\beta \leq \beta_c$, the measure $\mu^L = L^{-d} dx$ is a strict local minimum of I .*

Proof. By the definition of β_c from Definition 5.1, we have that for $\beta \leq \beta_c$, μ^L is a minimiser of I . The proof for $\beta < \beta_c$ is obvious. The idea of the proof is based on the fact that any minimiser of the free energy must be a solution of $T(\mu) = \mu - F(\mu) = 0$ (cf. [CGPS20, Proposition 2.4]), where $F : \dot{H}^{-1}(\mathbb{T}_L^d) \rightarrow \dot{H}^{-1}(\mathbb{T}_L^d)$ is the map given by

$$F(\mu) = \exp\left(-\beta W \star \mu - \log \int_{\mathbb{T}_L^d} \exp(-\beta W \star \mu) dx\right).$$

By the result of Theorem 5.2(a), all minimisers are smooth, and hence it is sufficient to consider the fixed point map F on the space $\dot{H}^{-1}(\mathbb{T}_L^d)$. Note that here we identify a measure $\mu \in \mathcal{P}(\mathbb{T}_L^d)$ with the formal Fourier series $\sum_{k \in \mathbb{Z}^d \setminus \{0\}} \hat{\mu}(k) e_k$. It is possible to check now that, for $\beta \leq \beta_c$, $DT(\mu^L) : \dot{H}^{-1}(\mathbb{T}_L^d) \rightarrow \dot{H}^{-1}(\mathbb{T}_L^d)$ is a bounded, linear, isomorphism. Indeed, we have that

$$DT(\mu^L)(\eta) = \eta - \beta \mu^L(W \star \eta) - \beta \mu^L \int_{\mathbb{T}_L^d} W \star \eta d\mu^L$$

The above operator is bounded on $\dot{H}^{-1}(\mathbb{T}_L^d)$ since $W \in C^2(\mathbb{T}_L^d) \subset H^1(\mathbb{T}_L^d)$. Diagonalising $DT(\mu^L)$ using $\{e_k\}_{k \in \mathbb{Z}^d \setminus \{0\}}$, we obtain

$$DT(\mu^L)e_k = (1 - \beta L^{-d/2} \hat{W}(k)) e_{-k}.$$

It follows that if $\beta \leq L^{d/2} / \min_{k \in \mathbb{Z}^d \setminus \{0\}} \hat{W}(k)$, then the above map is a bijection. That it is an injection is clear from the fact that if $DT(\mu^L)\eta_1 = DT(\mu^L)\eta_2$ then $\hat{\eta}_1(k) = \hat{\eta}_2(k)$ for all $k \in \mathbb{Z}^d \setminus \{0\}$. It is also surjective since for any $\eta \in \dot{H}^{-1}(\mathbb{T}_L^d)$, we have that $\sum_{k \in \mathbb{Z}^d \setminus \{0\}} \frac{\hat{\eta}(k)}{1 - \beta L^{-d/2} \hat{W}(-k)} e_{-k}$ maps to η under $DT(\mu^L)$. We know from Theorem 5.2(b) that β_c is lesser than this value and hence the result. Now, by the inverse function theorem, there exists for some $\varepsilon > 0$ an ε -open ball $B_\varepsilon^{\dot{H}^{-1}}(\mu^L)$ around μ^L in $\dot{H}^{-1}(\mathbb{T}_L^d)$ such that it is the unique solution of $T(\mu) = 0$ in this ball. This tells us that μ^L is the

unique minimiser of the free energy in $B_\varepsilon^{\dot{H}^{-1}}(\mu^L)$ at $\beta = \beta_c$. Note further that we have the following bounds for all $\mu \in B_\varepsilon^{\dot{H}^{-1}}(\mu^L)$

$$\mu^L \exp\left(-2\beta\|W\|_{\dot{H}^1(\mathbb{T}_L^d)}\|\mu\|_{\dot{H}^{-1}(\mathbb{T}_L^d)}\right) \leq F(\mu) \leq \mu^L \exp\left(2\beta\|W\|_{\dot{H}^1(\mathbb{T}_L^d)}\|\mu\|_{\dot{H}^{-1}(\mathbb{T}_L^d)}\right).$$

Additionally we have that $\|\mu - \mu^L\|_{\dot{H}^{-1}(\mathbb{T}_L^d)} < \varepsilon$ from which it follows that

$$\frac{\mu^L}{C} \leq F(\mu) \leq C \mu^L \quad \text{with} \quad C := \exp\left(2\beta\|W\|_{\dot{H}^1(\mathbb{T}_L^d)}(\|\mu^L\|_{\dot{H}^{-1}(\mathbb{T}_L^d)} + \varepsilon)\right),$$

for all $\mu \in B_\varepsilon^{\dot{H}^{-1}}(\mu^L)$. Consider the set

$$\mathcal{I} := \left\{ \mu \in \dot{H}^{-1}(\mathbb{T}_L^d) \cap \mathcal{P}(\mathbb{T}_L^d) \cap L^\infty(\mathbb{T}_L^d) : \frac{\mu^L}{C} \leq \mu \leq C \mu^L \right\}.$$

Then, any minimiser of I must lie in \mathcal{I} by construction. Additionally, for all $\mu \in \mathcal{I}$ we have from Lemma 5.4 for some fixed constant $C_0 = C_0(\mu^L, C)$ the bound

$$\|\mu^L - \mu\|_{\dot{H}^{-1}(\mathbb{T}_L^d)} \leq C_0 W_2(\mu^L, \mu).$$

We can thus pick a ball $B_r^{W_2}(\mu^L)$ with $r > 0$ sufficiently small such that $\|\mu - \mu^L\|_{\dot{H}^{-1}(\mathbb{T}_L^d)} < \varepsilon$ for all $\mu \in B_r^{W_2}(\mu^L) \cap \mathcal{I}$. Since all minimizers lie in \mathcal{I} , it thus follows that we can find a ball in W_2 for which μ^L is the unique minimiser of I . Thus, μ^L is a strict local minima in $\mathcal{P}(\mathbb{T}_L^d)$ equipped with the Wasserstein metric.

The boundary of the ball $\partial B_r^{W_2}(\mu^L)$ is a compact set in $(\mathcal{P}(\mathbb{T}_L^d), W_2)$, since \mathbb{T}_L^d is compact. Hence, the l.s.c. functional I attains its minimiser on this set, say μ^* . Setting $\delta = I(\mu^*) - I(\mu^L) > 0$ concludes the estimate in the lemma. \square

We can now prove the existence of a mountain pass point in the presence of a discontinuous phase transition.

Corollary 5.7. *Assume $W \in C^2(\mathbb{T}_L^d)$ with $\beta_c > 0$ a discontinuous transition point, i.e. there exist at least two distinct minimisers of I at $\beta = \beta_c$ such that one is μ^L and the other is some $\bar{\mu} \in \mathcal{P}(\mathbb{T}_L^d)$. It follows then that there exists a $\mu^* \in \mathcal{P}(\mathbb{T}_L^d)$ distinct from μ^L and $\bar{\mu}$ such that $|\partial I|(\mu^*) = |dI|(\mu^*) = 0$. Additionally, $I(\mu^*) = c$ with*

$$c = \inf_{\gamma \in \Gamma} \sup_{t \in [0,1]} I(\gamma(t)),$$

where $\Gamma = \{C([0,1]; \mathcal{P}(\mathbb{T}_L^d)) : \gamma(0) = \mu^L, \gamma(1) = \bar{\mu}\}$.

Proof. We can directly apply Theorem 1.2, once we show that μ^L is a strict local minimum of I , which is established in Lemma 5.6. Thus, the result follows. \square

Remark 5.8. We have chosen to apply Theorem 1.2 at a discontinuous transition point β_c because we know from Lemma 5.6 that μ^L is a strict local minimum. Furthermore, since $\beta_c > 0$ is discontinuous, we know that $\min_{\mathcal{P}(\mathbb{T}_L^d)} I(\mu) = I(\mu^L) = I(\bar{\mu})$. However,

we expect, by the continuity of the minimum value of the free energy in β (cf. [CP10, Proposition 2.4]), that the result of Theorem 1.2 holds in a small neighbourhood of the critical value β_c .

The situation is different at a continuous transition point $\beta_c > 0$, where by Definition 5.1, the minimisers are unique at $\beta = \beta_c$. In this situation the mountain pass argument of Theorem 1.2 can only be applied either by finding another measure $\mu \in D(I)$ (possibly a critical point) such the barrier value between μ^L and μ exceeds the maximum of their energies or by showing that for $\beta > \beta_c$ we can find measures satisfying the assumptions of Theorem 1.2. In principle, this may be possible for specific choices of W , but proving the existence of new critical points is usually non-constructive (cf. [CGPS20, Theorems 5.11 and 5.19]). Thus, extracting any information about the value of their free energy is a challenging problem.

We turn now, to the large deviations principle of the underlying particle system and the study of escape probabilities. We start by stating, without proof, the reformulated version of the main result from [DG87]. We also refer the reader to [FK06, Example 9.35, Section 13.3] for a discussion and proof of such large deviations principles for weakly interacting diffusions.

Theorem 5.9. *Let $\mathcal{P}^{(N)}(\mathbb{T}_L^d)$ be the space of empirical probability measures on \mathbb{T}_L^d , that is*

$$\mathcal{P}^{(N)}(\mathbb{T}_L^d) := \left\{ \mu \in \mathcal{P}(\mathbb{T}_L^d) : \mu = \frac{1}{N} \sum_{i=1}^N \delta_{x_i}, x_i \in \mathbb{T}_L^d \right\}.$$

Assume that $\mu_0^{(N)} \in \mathcal{P}^{(N)}(\mathbb{T}_L^d)$ is such that there exists $\mu_0 \in \mathcal{P}(\mathbb{T}_L^d)$ with $W_2(\mu_0^{(N)}, \mu_0) \rightarrow 0$ as $N \rightarrow \infty$. Denote by \mathcal{C}_T the space $C([0, T]; \mathcal{P}(\mathbb{T}_L^d))$, equipped with the topology of uniform convergence.

(a) *For all open subsets G of \mathcal{C}_T holds*

$$\liminf_{N \rightarrow \infty} N^{-1} \log \mathbb{P}(\mu^{(N)}(\cdot) \in G, \mu^{(N)}(0) = \mu_0^{(N)}) \geq - \inf_{\mu(\cdot) \in G, \mu(0) = \mu_0} S(\mu(\cdot)).$$

(b) *For all closed subsets F of \mathcal{C}_T holds*

$$\limsup_{N \rightarrow \infty} N^{-1} \log \mathbb{P}(\mu^{(N)}(\cdot) \in F, \mu^{(N)}(0) = \mu_0^{(N)}) \leq - \inf_{\mu(\cdot) \in F, \mu(0) = \mu_0} S(\mu(\cdot)),$$

(c) *For each compact subset K of $\mathcal{P}(\mathbb{T}_L^d)$ and $s \geq 0$ is the set*

$$\Phi_K(s) = \{ \mu(\cdot) \in \mathcal{C}_T : S(\mu(\cdot)) \leq s, \mu(0) \in K \},$$

compact.

Here $S : \mathcal{C}_T \rightarrow \mathbb{R} \cup \{+\infty\}$ is the action or rate functional given for $\mu \in AC^2([0, T]; \mathcal{P}(\mathbb{T}_L^d))$ by

$$S(\mu(\cdot)) := \frac{1}{2} \int_0^T \|\partial_t \mu - \nabla \cdot (\mu \nabla (\beta^{-1} \log \mu + W \star \mu))\|_{\dot{H}^{-1}(\mathbb{T}_L^d, \mu)}^2 dt,$$

and by $+\infty$ otherwise.

We are interested in using the above result to understand the probability of the empirical process escaping from the uniform state μ^L and reaching the clustered state $\bar{\mu}$ in time $T > 0$.

Theorem 5.10. *Assume $W \in C^2(\mathbb{T}_L^d)$ with β_c a discontinuous transition point, i.e. there exist at least two distinct minimisers of I at $\beta = \beta_c$ such that one is μ^L and the other is some $\bar{\mu} \in \mathcal{P}(\mathbb{T}_L^d)$. It follows then that the underlying empirical process $\mu^{(N)} \in \mathcal{C}_T$ with initial i.i.d uniformly distributed particles satisfies*

$$\mathbb{P}(\mu^{(N)}(T) \in \overline{B}_\varepsilon^{W_2}(\bar{\mu}), \mu^{(N)}(0) = \mu_0^{(N)}) \leq \exp\left(-N(\Delta - O(\varepsilon^2) - o_T(1))\right)$$

for N sufficiently large, where $\overline{B}_\varepsilon^{W_2}(\bar{\mu})$ is the closed ball of size $\varepsilon > 0$ around $\bar{\mu}$, $\Delta := I(\mu^*) - I(\mu^L)$, where μ^* is the critical point defined in Corollary 5.7. Here, $o_T(1)$ is a constant which vanishes as $N \rightarrow \infty$ with a rate depending on the time interval $T > 0$.

Proof. In order to prove this result we need to relate the rate functional S with the energy functional I . We can assume without loss of generality that $\mu \in AC^2([0, T]; H^1(\mathbb{T}_L^d) \cap \mathcal{P}(\mathbb{T}_L^d))$ since $S(\mu) = +\infty$ otherwise. It follows that there exists $\phi \in L^2([0, T]; \dot{H}^1(\mathbb{T}_L^d, \mu))$ [San15, Theorem 5.14] such that

$$\partial_t \mu = \nabla \cdot (\mu \nabla \phi),$$

where the above equation is satisfied in $\dot{H}^{-1}(\mathbb{T}_L^d, \mu)$. Thus, for any $\mu \in AC^2([0, T]; H^1(\mathbb{T}_L^d) \cap \mathcal{P}(\mathbb{T}_L^d))$ we can rewrite the rate functional, using the chain rule for gradient flows discussed in [AGS08, Section 10.1.2 E., Lemma 8.1.2] (see also [FN12]), as follows

$$\begin{aligned} S(\mu) &= \frac{1}{2} \int_0^T \left\| \partial_t \mu - \nabla \cdot (\mu \nabla (\beta_c^{-1} \log \mu + W \star \mu)) \right\|_{\dot{H}^{-1}(\mathbb{T}_L^d, \mu)}^2 dt \\ &= \frac{1}{2} \int_0^T \left\| \phi - \beta_c^{-1} \log \mu - W \star \mu \right\|_{\dot{H}^1(\mathbb{T}_L^d, \mu)}^2 dt \\ &= \frac{1}{2} \int_0^T \|\phi\|_{\dot{H}^1(\mathbb{T}_L^d, \mu)}^2 dt + \frac{1}{2} \int_0^T \left\| \beta_c^{-1} \log \mu + W \star \mu \right\|_{\dot{H}^1(\mathbb{T}_L^d, \mu)}^2 dt \\ &\quad + \int_0^T \left\langle \beta_c^{-1} \log \mu + W \star \mu, \phi \right\rangle_{\dot{H}^1(\mathbb{T}_L^d, \mu)} dt \\ &= \frac{1}{2} \int_0^T \|\phi\|_{\dot{H}^1(\mathbb{T}_L^d, \mu)}^2 dt + \frac{1}{2} \int_0^T \left\| \beta_c^{-1} \log \mu + W \star \mu \right\|_{\dot{H}^1(\mathbb{T}_L^d, \mu)}^2 dt \end{aligned}$$

$$+ \int_0^T \langle (\beta_c^{-1} \log \mu + W \star \mu), \partial_t \mu \rangle_{\dot{H}^1(\mathbb{T}_L^d, \mu), \dot{H}^{-1}(\mathbb{T}_L^d, \mu)} dt.$$

We choose the closed subset $F = \{\mu \in \mathcal{C}_T : \mu(T) \in \overline{B}_\varepsilon^{W_2}(\bar{\mu}), \mu(0) = \mu^L\}$ and we set $T^* = \arg \max_{t \in [0, T]} (I(\mu(t)) - I(\mu^L))$ if it is uniquely defined or pick any one if it is not. We can then rewrite the rate functional as follows

$$\begin{aligned} S(\mu) &= \frac{1}{2} \int_0^{T^*} \|\phi\|_{\dot{H}^1(\mathbb{T}_L^d, \mu)}^2 dt + \frac{1}{2} \int_0^{T^*} \|\beta_c^{-1} \log \mu + W \star \mu\|_{\dot{H}^1(\mathbb{T}_L^d, \mu)}^2 dt \\ &\quad + \int_0^{T^*} \langle (\beta_c^{-1} \log \mu + W \star \mu), \partial_t \mu \rangle_{\dot{H}^1(\mathbb{T}_L^d, \mu), \dot{H}^{-1}(\mathbb{T}_L^d, \mu)} dt \\ &\quad + \frac{1}{2} \int_{T^*}^T \|\partial_t \mu - \nabla \cdot (\mu \nabla (\beta_c^{-1} \log \mu + W \star \mu))\|_{\dot{H}^{-1}(\mathbb{T}_L^d, \mu)}^2 dt \\ &\geq \int_0^{T^*} \langle (\beta_c^{-1} \log \mu + W \star \mu), \partial_t \mu \rangle_{\dot{H}^1(\mathbb{T}_L^d, \mu), \dot{H}^{-1}(\mathbb{T}_L^d, \mu)} dt \\ &= \max_{t \in [0, T]} (I(\mu(t)) - I(\mu^L)). \end{aligned}$$

Note that we have again applied the chain rule for gradient flows from [AGS08, Section 10.1.2 E]. The estimate implies the lower bound

$$\inf_{\mu \in F} S(\mu) \geq \inf_{\mu \in F \cap AC^2} \max_{t \in [0, T]} (I(\mu(t)) - I(\mu^L)).$$

At this point, we cannot apply Theorem 1.2 directly, since F contains curves with varying endpoints not necessarily critical points. To handle this case, we define

$$\begin{aligned} F_{\text{epi}} &= \left\{ (\mu(\cdot), \xi(\cdot)) \in C([0, T]; \text{epi}(I)) : (\mu(0), \xi(0)) = (\mu^L, I(\mu^L)), \right. \\ &\quad \left. (\mu(T), \xi(T)) \in \bigcup_{\mu \in \overline{B}_\varepsilon^{W_2}(\bar{\mu})} (\mu, I(\mu)) \right\} \end{aligned}$$

If $\mu \in F \cap AC^2$, then the function $t \mapsto I(\mu(t))$ is absolutely continuous by [AGS08, Section 10.1.2 E]. Thus, the curve $(\mu(\cdot), I(\mu(\cdot)))$ lies in the set F_{epi} with $\max_{t \in [0, T]} (I(\mu(t)) - I(\mu^L)) = \max_{t \in [0, T]} (\mathcal{G}_I(\mu(t), I(\mu(t))) - \mathcal{G}_I(\mu^L, I(\mu^L)))$. Thus, we have that

$$\inf_{\mu \in F \cap AC^2} \max_{t \in [0, T]} (I(\mu(t)) - I(\mu^L)) \geq \inf_{\mu \in F_{\text{epi}}} \max_{t \in [0, T]} (\mathcal{G}_I(\mu(t), \xi(t)) - \mathcal{G}_I(\mu^L, I(\mu^L)))$$

Now, we argue that if ε is small enough the above quantity can be made arbitrarily close to Δ . For doing so, we define $\delta > 0$ such that $\inf_{\mu \in F_{\text{epi}}} \max_{t \in [0, T]} (\mathcal{G}_I(\mu(t), \xi(t)) - \mathcal{G}_I(\mu^L, I(\mu^L))) = \Delta - 2\delta$. Then, we find $(\tilde{\mu}(t), \xi(t)) \in F_{\text{epi}}$ with $\max_{t \in [0, T]} (\mathcal{G}_I(\tilde{\mu}(t), \xi(t)) - \mathcal{G}_I(\mu^L, I(\mu^L))) \leq \Delta - \delta$ from which it follows that $I(\tilde{\mu}(T)) - I(\mu^L) \leq \Delta - \delta$. Let

$$\begin{aligned} \Gamma_{\text{epi}} &= \left\{ (\mu(\cdot), \xi(\cdot)) \in C([0, T]; \text{epi}(I)) : (\mu(0), \xi(0)) = (\mu^L, I(\mu^L)), \right. \\ &\quad \left. (\mu(T), \xi(T)) = (\bar{\mu}, I(\bar{\mu})) \right\} \subset F_{\text{epi}} \end{aligned}$$

We know that $\inf_{\mu \in \Gamma_{\text{epi}}} \max_{t \in [0, T]} (\mathcal{G}_I(\mu(t), \xi(t)) - \mathcal{G}_I(\mu^L, I(\mu^L))) = \Delta$. Thus, if we take any continuous curve $(\mu(s), \xi(s))$ in $\text{epi}(I)$ from $(\tilde{\mu}(T), I(\tilde{\mu}(T)))$ to $(\bar{\mu}, I(\bar{\mu}))$ parametrised by $s \in [0, 1]$, $\mathcal{G}_I(\cdot, \cdot)$ must exceed or be equal to $I(\bar{\mu}) + \Delta$ at some $s \in [0, 1]$. Indeed, if this would not be the case then we could concatenate $(\tilde{\mu}(t), \xi(t))$ and $(\mu(s), \xi(s))$ to obtain, after reparametrisation, a new curve $[0, 1] \ni t \mapsto (\mu(t), \xi(t))$ in $\text{epi}(I)$ from $(\mu^L, I(\mu^L))$ to $(\bar{\mu}, I(\bar{\mu}))$ such that $\max_{t \in [0, 1]} \mathcal{G}_I(\mu(t), \xi(t)) < \Delta$, a contradiction, since this curve is also an element of Γ_{epi} .

We pick the curve $(\mu(\cdot), I(\mu(\cdot)))$ where $\mu \in C([0, 1]; \mathcal{P}_2(M))$ is a unit speed minimizing geodesic between $\tilde{\mu}(T)$ and $\bar{\mu}$, as defined in Proposition 3.1. Let t' be the time at which $I(\mu(t'))$ exceeds $I(\bar{\mu}) + \Delta$. By λ -geodesic convexity of I we have

$$I(\bar{\mu}) + \Delta \leq I(\mu(t')) \leq (1 - t')I(\tilde{\mu}(T)) + t'I(\bar{\mu}) + \frac{|\lambda|}{2}t'(1 - t')\varepsilon^2$$

Bounding $I(\tilde{\mu}(T))$ by $I(\bar{\mu}) + \Delta - \delta$, we obtain,

$$\begin{aligned} I(\bar{\mu}) + \Delta &\leq I(\bar{\mu}) + (1 - t')\Delta - (1 - t')\delta + \frac{|\lambda|}{2}t'(1 - t')\varepsilon^2 \\ &\leq I(\bar{\mu}) + \Delta - (1 - t')\delta + \frac{|\lambda|}{2}(1 - t')\varepsilon^2. \end{aligned}$$

From this it follows that

$$\delta \leq \frac{|\lambda|}{2}\varepsilon^2.$$

Thus, we obtain

$$\inf_{\mu \in F} S(\mu) \geq \inf_{\mu \in F_{\text{epi}}} \max_{t \in [0, T]} (\mathcal{G}_I(\mu(t), \xi(t)) - \mathcal{G}_I(\mu^L, I(\mu^L))) = \Delta - 2\delta \geq \Delta - |\lambda|\varepsilon^2$$

Finally, we can apply the result of Theorem 5.9 (b), to obtain that

$$\limsup_{N \rightarrow \infty} N^{-1} \log \mathbb{P}(\mu^{(N)}(\cdot) \in F, \mu^{(N)}(0) = \mu_0^{(N)}) \leq - \inf_{\mu(\cdot) \in F, \mu(0) = \mu^L} S(\mu(\cdot)) \leq -\Delta + |\lambda|\varepsilon^2,$$

where we use that $W_2(\mu_0^{(N)}, \mu^L) \rightarrow 0$ as $N \rightarrow \infty$ is implied by the strong law of large numbers. We set

$$a_N = N^{-1} \log \mathbb{P}(\mu^{(N)}(\cdot) \in F, \mu^{(N)}(0) = \mu_0^{(N)}).$$

It follows that

$$a_N \leq \sup_{N_1 \geq N} a_{N_1} = \left(\sup_{N_1 \geq N} a_{N_1} - \limsup_{N \rightarrow \infty} a_N \right) + \limsup_{N \rightarrow \infty} a_N \leq C_{N,T} - \Delta + |\lambda|\varepsilon^2,$$

where $C_{N,T} = \left(\sup_{N_1 \geq N} a_{N_1} - \limsup_{N \rightarrow \infty} a_N \right) = o_T(1)$. Plugging in the expression for a_N , we obtain

$$\mathbb{P}(\mu^{(N)}(\cdot) \in F, \mu^{(N)}(0) = \mu_0^{(N)}) \leq e^{-N(\Delta - |\lambda|\varepsilon^2 - o_T(1))}$$

The result then follows from the above estimate and the definition of the set F . \square

Remark 5.11. To obtain a result that is uniform in $T > 0$ would require something stronger than the Dawson–Gärtner large deviations principle in Theorem 5.9. The approach of quasi-potentials discussed in [DG86, DG89] may be the correct idea to use to obtain such a result. However, this would require much more information about the structure of the non-trivial minimiser and its basin of attraction. Since this is not the focus of this work, we refer to [Bas20] for a first step in this direction for a particular choice of the interaction (and confining) potential. We hope to treat the general case in a future work.

Similarly, the $O(\varepsilon^2)$ appearing in the exponent $\exp(-N(\Delta - O(\varepsilon^2)))$ can be removed if one can show that the minimiser $\bar{\mu}$ is a local basin of attraction for the McKean–Vlasov dynamics, i.e. there exists some $\varepsilon > 0$ such that all measures in $\overline{B}_\varepsilon^{W_2}(\bar{\mu})$ converge to $\bar{\mu}$ under the flow of the McKean–Vlasov PDE as $t \rightarrow \infty$. In this case we can choose the continuous curve between $\tilde{\mu}(T)$ and $\bar{\mu}$ (in the proof of Theorem 5.10) to be the solution of the McKean–Vlasov PDE starting $\tilde{\mu}(T)$. This solution does not increase the energy and thus the $O(\varepsilon^2)$ error from the λ -convexity argument will not appear in the exponent. Such a characterisation of $\bar{\mu}$ is expected under more specific assumptions on the potential W .

Acknowledgements. The authors would like to thanks José A. Carrillo and Greg Pavliotis for useful discussions during the course of this work. The authors would also like to thank the anonymous referees for their careful reading of the draft manuscript and their useful comments and suggestions.

RSG also acknowledges the hospitality of RWTH Aachen University. AS acknowledges the hospitality of Imperial College London. Part of this work was carried out at the workshop “Nonlocal differential equations in collective behaviour” held at the American Institute of Mathematics, San José and at the “Junior Trimester Programme in Kinetic Theory” held at the Hausdorff Research Institute for Mathematics, Bonn. RSG and AS are grateful to both institutes for their hospitality.

REFERENCES

- [ADPZ11] 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. *Comm. Math. Phys.*, 307(3):791–815, 2011.
- [ADPZ13] S. Adams, N. Dirr, M. Peletier, and J. Zimmer. Large deviations and gradient flows. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 371(2005):20120341, 17, 2013.
- [AF14] L. Ambrosio and J. Feng. On a class of first order Hamilton-Jacobi equations in metric spaces. *J. Differential Equations*, 256(7):2194–2245, 2014.

- [AGS08] 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, 2nd edition, 2008.
- [Bas20] K. Bashiri. On the basin of attraction of McKean-Vlasov paths. *arXiv e-prints*, page arXiv:2001.09106, January 2020.
- [BB20] K. Bashiri and A. Bovier. Gradient flow approach to local mean-field spin systems. *Stochastic Process. Appl.*, 130(3):1461–1514, 2020.
- [BP12] V. Barbu and T. Precupanu. *Convexity and Optimization in Banach Spaces*. Springer Netherlands, 2012.
- [CEMS01] D. Cordero-Erausquin, R. J. McCann, and M. Schmuckenschläger. A Riemannian interpolation inequality à la Borell, Brascamp and Lieb. *Invent. Math.*, 146(2):219–257, 2001.
- [CGPS20] J. A. Carrillo, R. S. Gvalani, G. A. Pavliotis, and A. Schlichting. Long-time behaviour and phase transitions for the McKean-Vlasov equation on the torus. *Arch. Ration. Mech. Anal.*, 235(1):635–690, 2020.
- [CP10] L. Chayes and V. Panferov. The McKean-Vlasov equation in finite volume. *J. Stat. Phys.*, 138(1-3):351–380, 2010.
- [CSZ19] F. Cornalba, T. Shardlow, and J. Zimmer. A regularized Dean-Kawasaki model: derivation and analysis. *SIAM J. Math. Anal.*, 51(2):1137–1187, 2019.
- [CSZ20] F. Cornalba, T. Shardlow, and J. Zimmer. From weakly interacting particles to a regularised Dean-Kawasaki model. *Nonlinearity*, 33(2):864–891, 2020.
- [Daw83] D. A. Dawson. Critical dynamics and fluctuations for a mean-field model of cooperative behavior. *J. Statist. Phys.*, 31(1):29–85, 1983.
- [DG86] 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.
- [DG87] D. A. Dawson and J. Gärtner. Large deviations from the McKean-Vlasov limit for weakly interacting diffusions. *Stochastics*, 20(4):247–308, 1987.
- [DG89] D. A. Dawson and J. Gärtner. Large deviations, free energy functional and quasi-potential for a mean field model of interacting diffusions. *Mem. Amer. Math. Soc.*, 78(398):iv+94, 1989.
- [dGMT80] E. de Giorgi, A. Marino, and M. Tosques. Problems of evolution in metric spaces and maximal decreasing curve. *Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur. (8)*, 68(3):180–187, 1980.
- [DM94] M. Degiovanni and M. Marzocchi. A critical point theory for nonsmooth functionals. *Ann. Mat. Pura Appl. (4)*, 167:73–100, 1994.
- [DPZ13] M. H. Duong, M. A. Peletier, and J. Zimmer. GENERIC formalism of a Vlasov-Fokker-Planck equation and connection to large-deviation principles. *Nonlinearity*, 26(11):2951–2971, 2013.
- [DS14] S. Danieri and G. Savaré. Lecture Notes on Gradient Flows and Optimal Transport. In H. Pajot, Y. Ollivier, and C. Villani, editors, *Optimal Transportation*, pages 100–144. Cambridge University Press, Sep 2014.
- [EFLS16] M. Erbar, M. Fathi, V. Laschos, and A. Schlichting. Gradient flow structure for McKean-Vlasov equations on discrete spaces. *Discrete Contin. Dyn. Syst.*, 36(12):6799–6833, 2016.
- [Fat16] M. Fathi. A gradient flow approach to large deviations for diffusion processes. *J. Math. Pures Appl. (9)*, 106(5):957–993, 2016.
- [FG19] B. Fehrman and B. Gess. Large deviations for conservative stochastic PDE and non-equilibrium fluctuations. *arXiv e-prints*, page arXiv:1910.11860, October 2019.

- [FK06] J. Feng and T. G. Kurtz. *Large deviations for stochastic processes*, volume 131 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2006.
- [FK09] J. Feng and M. Katsoulakis. A comparison principle for Hamilton-Jacobi equations related to controlled gradient flows in infinite dimensions. *Arch. Ration. Mech. Anal.*, 192(2):275–310, 2009.
- [FN12] J. Feng and T. Nguyen. Hamilton-Jacobi equations in space of measures associated with a system of conservation laws. *J. Math. Pures Appl. (9)*, 97(4):318–390, 2012.
- [FS16] M. Fathi and M. Simon. The gradient flow approach to hydrodynamic limits for the simple exclusion process. In *From particle systems to partial differential equations. III*, volume 162 of *Springer Proc. Math. Stat.*, pages 167–184. Springer, [Cham], 2016.
- [FV18] S. Friedli and Y. Velenik. *Statistical mechanics of lattice systems*. Cambridge University Press, Cambridge, 2018. A concrete mathematical introduction.
- [GP18] S. N. Gomes and G. A. Pavliotis. Mean Field Limits for Interacting Diffusions in a Two-Scale Potential. *J. Nonlinear Sci.*, 28(3):905–941, 2018.
- [GPY13] J. Garnier, G. Papanicolaou, and T.-W. Yang. Large deviations for a mean field model of systemic risk. *SIAM J. Financial Math.*, 4(1):151–184, 2013.
- [IS96] A. Ioffe and E. Schwartzman. Metric critical point theory. I. Morse regularity and homotopic stability of a minimum. *J. Math. Pures Appl. (9)*, 75(2):125–153, 1996.
- [Kat94] G. Katriel. Mountain pass theorems and global homeomorphism theorems. *Ann. Inst. H. Poincaré Anal. Non Linéaire*, 11(2):189–209, 1994.
- [KJZ19] M. Kaiser, R. L. Jack, and J. Zimmer. A variational structure for interacting particle systems and their hydrodynamic scaling limits. *Commun. Math. Sci.*, 17(3):739–780, 2019.
- [KLvR20] V. Konarovskiy, T. Lehmann, and M. von Renesse. On Dean-Kawasaki dynamics with smooth drift potential. *J. Stat. Phys.*, 178(3):666–681, 2020.
- [Loe06] G. Loeper. Uniqueness of the solution to the Vlasov-Poisson system with bounded density. *J. Math. Pures Appl. (9)*, 86(1):68–79, 2006.
- [LP66] J. L. Lebowitz and O. Penrose. Rigorous treatment of the van der Waals-Maxwell theory of the liquid-vapor transition. *J. Mathematical Phys.*, 7:98–113, 1966.
- [McC97] R. J. McCann. A convexity principle for interacting gases. *Adv. Math.*, 128(1):153–179, 1997.
- [McC01] R. J. McCann. Polar factorization of maps on Riemannian manifolds. *Geom. Funct. Anal.*, 11(3):589–608, 2001.
- [MM13] S. Mischler and C. Mouhot. Kac’s program in kinetic theory. *Invent. Math.*, 193(1):1–147, 2013.
- [Rey18] J. Reygner. Equilibrium large deviations for mean-field systems with translation invariance. *Ann. Appl. Probab.*, 28(5):2922–2965, 2018.
- [San15] F. Santambrogio. *Optimal transport for applied mathematicians*, volume 87 of *Progress in Nonlinear Differential Equations and their Applications*. Birkhäuser/Springer, Cham, 2015. Calculus of variations, PDEs, and modeling.
- [Shi87] M. Shiino. Dynamical behavior of stochastic systems of infinitely many coupled nonlinear oscillators exhibiting phase transitions of mean-field type: H theorem on asymptotic approach to equilibrium and critical slowing down of order-parameter fluctuations. *Phys. Rev. A*, 36(5):2393–2412, 1987.
- [Sin82] Y. G. Sinai. *Theory of phase transitions: rigorous results*, volume 108 of *International Series in Natural Philosophy*. Pergamon Press, Oxford-Elmsford, N.Y., 1982. Translated from the Russian by J. Fritz, A. Krámli, P. Major and D. Szász.

- [Szn91] A.-S. Sznitman. Topics in propagation of chaos. In *École d'Été de Probabilités de Saint-Flour XIX—1989*, volume 1464 of *Lecture Notes in Math.*, pages 165–251. Springer, Berlin, 1991.
- [Tam84] Y. Tamura. On asymptotic behaviors of the solution of a nonlinear diffusion equation. *J. Fac. Sci. Univ. Tokyo Sect. IA Math.*, 31(1):195–221, 1984.
- [Tug14] J. Tugaut. Phase transitions of McKean-Vlasov processes in double-wells landscape. *Stochastics*, 86(2):257–284, 2014.
- [Vil09] C. Villani. *Optimal Transport: Old and New*, volume 338. Springer Berlin Heidelberg, 2009.

DEPARTMENT OF MATHEMATICS, IMPERIAL COLLEGE LONDON, LONDON SW7 2AZ
Email address: rishabh.gvalani14@imperial.ac.uk

INSTITUT FÜR ANGEWANDTE MATHEMATIK, UNIVERSITÄT BONN
Email address: schlichting@iam.uni-bonn.de