Birth–death dynamics for sampling: global convergence, approximations and their asymptotics

Yulong Lu\(^1\), Dejan Slepčev\(^2\) and Lihan Wang\(^2,*\)

\(^1\) School of Mathematics, University of Minnesota, Twin Cities, Minneapolis, MN 55455, United States of America
\(^2\) Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213, United States of America

E-mail: lihanw@andrew.cmu.edu

Received 7 November 2022; revised 14 August 2023
Accepted for publication 13 September 2023
Published 26 September 2023

Recommended by Dr Yao Yao

Abstract
Motivated by the challenge of sampling Gibbs measures with nonconvex potentials, we study a continuum birth–death dynamics. We improve results in previous works (Liu \textit{et al} 2023 \textit{Appl. Math. Optim.} \textbf{87} 48; Lu \textit{et al} 2019 arXiv:1905.09863) and provide weaker hypotheses under which the probability density of the birth–death governed by Kullback–Leibler divergence or by \(\chi^2\) divergence converge exponentially fast to the Gibbs equilibrium measure, with a universal rate that is independent of the potential barrier. To build a practical numerical sampler based on the pure birth–death dynamics, we consider an interacting particle system, which is inspired by the gradient flow structure and the classical Fokker–Planck equation and relies on kernel-based approximations of the measure. Using the technique of \(\Gamma\)-convergence of gradient flows, we show that on the torus, smooth and bounded positive solutions of the kernelised dynamics converge on finite time intervals, to the pure birth–death dynamics as the kernel bandwidth shrinks to zero. Moreover we provide quantitative estimates on the bias of minimisers of the energy corresponding to the kernelised dynamics.

\(^*\) Author to whom any correspondence should be addressed.
Finally we prove the long-time asymptotic results on the convergence of the asymptotic states of the kernelised dynamics towards the Gibbs measure.

Keywords: spherical Hellinger metric, gradient flow, statistical sampling, birth–death dynamics

Mathematics Subject Classification numbers: 35Q62, 82C21, 62D05, 45M05

(Some figures may appear in colour only in the online journal)

1. Introduction

Sampling from a given target probability distribution has diverse applications, including Bayesian statistics, machine learning, statistical physics, and many others. In practice, the measure $\pi$ is often the Gibbs measure corresponding to potential $V: \mathbb{R}^d \rightarrow \mathbb{R}$:

$$
\pi(x) = \frac{1}{Z} e^{-V(x)}, \quad \text{for } x \in \mathbb{R}^d,
$$

where $Z$ is the typically unknown normalisation constant.

Some of the most popular methods for sampling such distributions are based on Markov chain Monte Carlo (MCMC) approach. Much of the research works on MCMC have been devoted to designing Markov chains that are ergodic with respect to the target probability measure and enjoy fast mixing properties. Popular sampling methods include Langevin MCMC [35, 67], Hamiltonian Monte Carlo [3, 9, 62, 68], bouncy particle and zigzag samplers [8, 10], and affine-invariant ensemble MCMC [34], and Stein variational gradient descent (SVGD) [53].

When the potential function $V$ is strongly convex, these sampling methods perform quite well; we refer to recent literature [23, 28, 56, 74, 77] and references therein for understanding their convergence and computational complexity. However, the efficiency of these sampling methods are hampered by the multi-modality of $\pi$ (corresponding to a non-convex $V$) as it takes exponentially long time for the sampler to hop from one mode to another. Many diffusion-based samplers suffer from such metastability issue, and numerous techniques have been proposed to alleviate this issue, including in particular parallel and simulated tempering [59, 61, 73] and adaptive biasing methods [24, 36, 43, 47, 75].

The sampling problem can be recast as an optimisation problem on the space of probability measures [7, 77]. Indeed, inspired by the seminal work of Jordan et al [38], the Fokker–Planck equation associated to the (overdamped) Langevin dynamics can be viewed as the Wasserstein gradient flow of the Kullback–Leibler (KL)-divergence

$$
\mathcal{F}(\rho) := KL(\rho\|\pi) = \int_{\mathbb{R}^d} \log \frac{\rho}{\pi} \, d\rho,
$$

which suggests that the Langevin dynamics can be seen as the steepest descent flow of the KL-divergence, along which the initial distribution flows towards the target distribution. The gradient-flow perspective provides a way towards building new sampling dynamics by designing new objective functions or new metrics for the manifolds of probability measures. For example, the underdamped Langevin dynamics can be viewed as a Nesterov’s accelerated gradient descent on the space of probability measures [58].

Langevin dynamics converge exponentially fast to the Gibbs measure under the assumption that the target measure satisfies the logarithmic Sobolev (Log-Sobolev) inequality [6]. Unfortunately, the convergence rate is limited by the optimal constant in the Log-Sobolev
inequality, which can be very small when the target measure is multimodal. This is because it can take exponentially long time for the dynamics to overcome the energy barriers and hop between the multiple modes.

1.1. Birth–death dynamics, and their long-time convergence

The issues described above prompt the following question:

Can one construct a gradient-flow dynamics for sampling that achieves a potential-independent convergence rate?

Recent work [57] by Lu, Nolen and one of the authors gives an affirmative answer to the question by proposing the following birth–death dynamics for sampling

$$\partial_t \rho_t = -\rho_t \log \frac{\rho_t}{\pi} + \rho_t \int_{\mathbb{R}^d} \frac{\rho_t}{\pi} \log \frac{\rho_t}{\pi} \, dx.$$ (BD)

An equation of similar type on discrete space, known as replicator equation, also appears in information geometry literature, see [5, chapter 6] and references therein. Note that the dynamics (BD) is agnostic to the normalisation constant. It has been shown that (BD) is a gradient flow of the KL-divergence with respect to the spherical Hellinger distance $d_{SH}$ defined by (10) (see [45, section 3]). Furthermore the authors show an exponential rate of convergence for initial data such that $\rho_0$ is bounded below by a positive multiple of $\pi$ and $\text{KL}(\rho_0|\pi) \leq 1$. More recent work [51] established exponential convergence of (BD) if $\frac{\rho_0}{\pi}$ is bounded from above and below.

In theorem 2.4 we improve both results by showing that the solution $\rho_t$ contracts to the equilibrium with a uniform (potential-independent) rate from any $\rho_0$ that is only bounded from below, but not necessarily above with respect to $\pi$. The condition $\text{KL}(\rho_0|\pi) \leq 1$ is no longer required. This has an important practical consequence, namely it shows that it is sufficient to start the dynamics with an initial density that is more spread out than $\pi$, which is easy to guarantee for many target measures $\pi$. The removal of upper bound also allows us to choose a sufficient wide round Gaussian to satisfy the pointwise lower bound for a large class of $\pi$ in $\mathbb{R}^d$.

In section 3 we also investigate birth–death dynamics

$$\partial_t \rho_t = -\rho_t \left( \frac{\rho_t}{\pi} - \int \frac{\rho_t}{\pi} \, d\rho_t \right)$$ (BD2)

that arises as the spherical Hellinger gradient flow of $\chi^2$-divergence which is at the basis of the algorithm proposed in [50]. In particular we prove that the $\chi^2$-divergence between the dynamics and the target measure converges to zero exponentially fast, with a rate independent of the potential, see theorem 3.1. This complements the result of [50], which proved the convergence of the reverse $\chi^2$-divergence.

1.2. Approximations to (BD) that allow for discrete measures

Since the equation (BD) is not well-defined when $\rho$ is a discrete measure, it is unclear whether one can build interacting particle sampling schemes directly based on it. A principled way to

\footnote{In information geometry literature [1, 5], the spherical Hellinger distance is also known as the Fisher–Rao distance. On the other hand, in other works, for example [20], the terminology ‘Fisher–Rao distance’ refers to the Hellinger distance, which is defined on positive measures. To avoid confusion and emphasise the fact that these two distances are defined on different spaces, we avoid using ‘Fisher–Rao distance’ altogether in this work.}
build particle approximations to (BD) and (BD2) is to define new dynamics which approximate (BD) and (BD2) and are well-defined for discrete measures. A further desirable property for these dynamics is to retain the (spherical) Hellinger gradient flow structure.

In doing so we are inspired by the work of Carrillo et al [16], who considered the Wasserstein gradient flow of the regularised KL-divergence as an approximation of the Fokker–Planck equation. In particular they introduced the regularised energy

$$F_\varepsilon(\rho) = \int \rho \log (K_\varepsilon * \rho) - \int \rho \log \pi = \int \rho \log (K_\varepsilon * \rho) + \int \rho V.$$  

The gradient flow of $F_\varepsilon$ with respect to Wasserstein metric, studied in [16], is

$$\partial_t \rho_t^{(\varepsilon)} = \nabla \cdot \left( \rho \nabla \frac{\delta F_\varepsilon}{\delta \rho} \right), \tag{1}$$

where $\frac{\delta F_\varepsilon}{\delta \rho}$ is the functional derivative of $F_\varepsilon$ defined by

$$\frac{\delta F_\varepsilon}{\delta \rho} = \log \left( \frac{K_\varepsilon * \rho}{\pi} \right) + K_\varepsilon * \left( \frac{\rho}{K_\varepsilon * \rho} \right). \tag{2}$$

For discrete initial data, i.e. $\rho_0 = \sum m_i \delta_{x_i}$, the solution remains a discrete measure for any positive time, where the evolution of particles is given by a system of ordinary differential equations (ODEs). It heuristically provides a deterministic particle approximation of the Fokker–Planck equation, though rigorous convergence analysis remains open.

The gradient flow of $F_\varepsilon$ with respect to the spherical Hellinger distance is the equation

$$\partial_t \rho_t^{(\varepsilon)} = -\rho_t^{(\varepsilon)} \left[ \log \left( \frac{K_\varepsilon * \rho_t^{(\varepsilon)}}{\pi} \right) + K_\varepsilon * \left( \frac{\rho_t^{(\varepsilon)}}{K_\varepsilon * \rho_t^{(\varepsilon)}} \right) - \int \log \left( \frac{K_\varepsilon * \rho_t^{(\varepsilon)}}{\pi} \right) \rho_t^{(\varepsilon)} - 1 \right]. \tag{BD_\varepsilon}$$

Note that the right hand side is well-defined even if $\rho^{(\varepsilon)}$ is a discrete measure. This suggests the possibility of approximating (BD_\varepsilon) with interacting particles. Indeed, we will introduce and discuss in section 4.5 a particle-based jump process whose mean-field limit is heuristically characterised by equation (BD_\varepsilon) and present some numerical experiments on its application for sampling in section 5.

**Bias of global minimisers of $F_\varepsilon$.** For sufficiently small $\varepsilon$, we expect $F_\varepsilon$ to be only a small perturbation of $F$, and hence the global minimisers of $F_\varepsilon$ should also be a perturbation of $\pi$. In section 4.1 we prove that such bias is at most of order $\varepsilon$, improving on the qualitative $\Gamma$-convergence result in [16]. More precisely, we show that for any minimiser $\pi_\varepsilon$ of $F_\varepsilon$, the Wasserstein distance between $\pi_\varepsilon$ and $\pi$ is no more than the order $O(\varepsilon)$. The optimality of such upper bound is demonstrated by numerical experiments.

**Convergence of the dynamics, on finite time intervals.** In section 4.3 we investigate the convergence of the solutions of (BD_\varepsilon) towards solutions of the pure birth–death dynamics (BD). More precisely we use the $\Gamma$-convergence of gradient flows to show that on arbitrary finite time intervals, smooth solutions of (BD_\varepsilon) with initial condition $\rho_0$ bounded below on a bounded domain converge to solutions of (BD) as $\varepsilon \to 0$. For unbounded domains there are substantial difficulties to handle the decay of $\pi$ at infinity, and proving $\Gamma$-convergence of gradient flows in such setting remains an open problem.
Convergence of the dynamics, asymptotic states. We note that since \( \Gamma \)-convergence of gradient flows is in general stated for finite time intervals, and thus does not directly imply that the asymptotic states of the dynamics (BD\(_e\)) converge towards the asymptotic state of (BD), namely \( \pi \). This is a general issue for the convergence of gradient flows and is of practical interest. Namely in many applications one uses approximate gradient flows with aim to approximate the limiting state of the original gradient flow. In section 4.4 we investigate the relationship between \( \Gamma \)-convergence of gradient flows and the convergence of asymptotic states. Specifically in proposition 4.19 we prove convergence of asymptotic states in the general setting of gradient flows in metric spaces. We apply the result in two settings, one is the setting of the gradient flows of this paper (theorem 4.21) and the other (theorem 4.20) is the convergence of two-layer neural networks studied in [37], which we discuss at the end of this section. In particular, we prove in theorem 4.21 that \( \pi \) must be the only possible limit of the asymptotic states \( \rho_{\infty}^{(e)} \) in \( W_2 \). The proof is general and relies on the fact that \( \pi \) is the unique minimiser of the KL divergence.

Application to convergence of asymptotic states for two-layer neural networks. In [37], the authors considered the problem of learning a strongly concave function \( f \) using bump-like neurons where the width of the kernels \( \delta \). As the number of neurons approach infinity, the process of stochastic gradient descent with noise \( \tau \) converges to the Wasserstein gradient flow of the following entropy-regularised risk functional

\[
F^{\delta}(\rho^{\delta}) = \int_{\Omega} \left( \frac{1}{2} (K^{\delta} * \rho^{\delta} - f)^2 + \tau \rho^{\delta} \log \rho^{\delta} \right) \, dx.
\]  

(3)

Here \( \Omega \) is a smooth, convex and compact domain. More precisely, the gradient flow equation writes

\[
\partial_t \rho_t^{\delta} = \nabla : (\rho_t^{\delta} \nabla \Psi) + \tau \Delta \rho_t^{\delta}, \quad \text{with} \quad \Psi = -K^{\delta} * f + K^{\delta} * K^{\delta} * \rho_t^{\delta}.
\]  

(4)

The authors proved that as \( \delta \to 0 \), with suitable initial and boundary conditions, the solution of (4) converges strongly in \( L^2 \) to the solution of the limiting gradient flow

\[
\partial_t \rho_t = \nabla : (\rho_t \nabla (\rho_t - f)) + \tau \Delta \rho_t,
\]  

(5)

which is the Wasserstein gradient flow of the limiting functional

\[
F(\rho) = \int_{\Omega} \left( \frac{1}{2} (\rho - f)^2 + \tau \rho \log \rho \right) \, dx.
\]  

(6)

Moreover, since (6) is displacement convex with respect to Wasserstein geodesics, (6) has a unique minimiser and (5) converges exponentially to that minimiser as \( t \to \infty \). The work [37], however, does not provide results regarding the long-time behaviour of the regularised gradient flow (4), which is the numerically approximated dynamics. Our proposition 4.19 provides the tools to prove convergence of limiting states, resulting in theorem 4.20 below. For more results regarding the long-time convergence of such dynamics arising from the training of neural networks, we refer the readers to the recent works [19, 21].

In section 5 we provide two numerical experiments, one on a toy example of two-dimensional Gaussian mixture, another on a real-world Bayesian classification problem, to demonstrate the effectiveness of the birth–death algorithm. In both examples we observe
that birth–death sampler allows significantly faster mixing of particles compared to Langevin dynamics or SVGD. More specifically, in the multimodal example 5.1, one can see in figure 3 that, once a high-probability mode is discovered, birth–death sampler will facilitate movement of particles towards this newly discovered mode, which helps overcoming the issue of metastability. In the real-world example 5.2 where the non-convexity is not strong and SVGD works well, birth–death sampler can reach the equilibrium in an extremely short time.

1.3. Contributions

We highlight the major contributions of the present paper as follows:

- We prove that the pure birth–death dynamics (BD) converges globally to its unique equilibrium measure $\pi$ with a uniform rate with respect to KL-divergence, improving the results in [51, 57]. See theorem 2.4 for the precise statements. Using similar techniques, we also investigate the algorithm proposed in [50], and prove that their time-rescaled infinite-particle equation converges in $\chi^2$-divergence converges exponentially with a rate independent of the potential, see theorem 3.1.
- We show that under suitable conditions, any global minimiser of the regularised energy (32) $\pi_\varepsilon$ is $O(\varepsilon)$ close to $\pi$ under $W_2$ distance. The precise statement can be found in theorem 4.2.
- We show in theorem 4.16 that smooth solutions of the kernelised dynamics (BD$\varepsilon$) with densities bounded above and below on torus $\Gamma$-converges to the pure birth–death dynamics (BD) within any finite time-horizon in the limit of small kernel width. As a corollary, this justifies the convergence of the density $\rho^\varepsilon_t$ of the kernelised dynamics with width $\varepsilon$ to the target measure $\pi$ as $\varepsilon \to 0$ and $t \to \infty$.
- Finally, we show in theorem 4.21 that on torus, the long-time limit of (BD$\varepsilon$) converges with respect to Hausdorff distance corresponding to the Wasserstein distance.

1.4. Related works

The pure birth–death dynamics (BD) is a gradient flow on relative entropy with respect to the the spherical Hellinger distance we define in (10). The mass-conservative metric (10) was introduced in [11, 40, 45] and used in the study equations modeling fluids and population dynamics. Later it was applied in the analysis of training process of neural networks [69, 76]. In the context of statistical sampling, the spherical Hellinger metric was first applied by [57] to accelerate Langevin dynamics for sampling, where a local exponential convergence (theorem 2.3) was proved. The paper [30] uses the idea of birth–death dynamics to improve the training of normalising flows that learn the target distribution. The recent paper [50] constructs an ensemble MCMC algorithm whose mean field evolution is given by the spherical Hellinger gradient flow of the $\chi^2$-divergence (see (BD$\varepsilon$)). Convergence to the equilibrium was also established therein on a finite state space. The construction of birth–death dynamics (BD) is also related to the recent study on unbalanced optimal transport and associated gradient flows. In particular, the unbalanced transportation metric, called the Hellinger–Kantorovich metric, which interpolates between two-Wasserstein and Hellinger metric, was defined and studied in [20, 39, 48] and allows for transport between measures with different masses.

Ensemble-based sampling methods have also been widely studied in recent years, which is another important motivation of our work. Ensemble-based sampling allows for global view of the particle configurations and enables for the particles to exchange information. One of the most successful sampling methods in this category is the affine invariant sampler introduced by Goodman and Weare [34]; see also [29]. Ensemble-based sampling are also related
to sequential Monte Carlo [25] and importance sampling [12, 66]. In a continuous-time point of view, ensemble-based samplers can be developed via interacting particle systems, examples of which include ensemble Kalman methods [32, 65], consensus based sampling [13] (which also has ideas from optimisation [14, 64]), and ensemble Langevin dynamics [34].

We are particularly interested in sampling approaches that are defined as gradient flows of a functional measuring the difference from the target measure. There are a variety of functionals and metrics considered. Blob particle method [16] is the Wasserstein gradient flow of the regularisation of KL divergence that allows for discrete measures where it becomes an interacting particle system. Such viewpoint has been applied to sampling purposes in [21] where the authors considered the Wasserstein gradient flow for regularised $\chi^2$ energy. A different approach to create gradient-flow based interacting-particle systems for sampling is the SVGD introduced in [53]. There the authors consider the gradient flow of the standard KL-divergence with respect to a metric (now known as the Stein geometry) which requires smoothness of the velocities. Thus the gradient-flow velocity makes sense even when considered for particle measures [27, 52, 55]. The work [18] provides a new perspective on SVGD by viewing it as a kernelised gradient flow of the $\chi^2$-divergence. A further direction of research considers Wasserstein gradient flows of the distance to the target measure in a very weak metric that is well defined for particles; in particular kernelised Stein discrepancy [42]. Recently the work [44] considered using Wasserstein gradient flow for variational inference, where they use Gaussian mixtures to approximate the target density with the evolution of mean and variance governed by gradient flows.

2. Pure birth–death dynamics governed by relative entropy

Let us first introduce the Benamou–Brenier formulation of the Hellinger distance (the distance plays an important role in information geometry, see for example [1, 5]) on (not necessarily probability) measures

$$d^2_H(\rho_0, \rho_1) = \inf_{(\rho_t, u_t)} \int_0^1 \int_{\mathbb{R}^d} u_t^2 \, d\rho_t \, dt,$$

where $(\rho_t, u_t)$ satisfies the equation

$$\frac{\partial \rho_t}{\partial t} = -\rho_t u_t.$$

If measures $\rho_0, \rho_1 \ll \lambda$ for some probability measure $d\lambda(x)$, then one can explicitly compute the minimal cost in (7) and obtain

$$d^2_H(\rho_0, \rho_1) = 4 \int_{\mathbb{R}^d} \left( \sqrt{\frac{d\rho_1}{d\lambda}} - \sqrt{\frac{d\rho_0}{d\lambda}} \right)^2 \, d\lambda.$$

Moreover, this expression does not depend on the specific choice of $\lambda$. Indeed, substituting $u_t = -\frac{\partial \rho_t}{\partial \rho_t \, d\lambda}$ into (7), we have

$$\int_0^1 \int_{\mathbb{R}^d} u_t^2 \, d\rho_t \, dt = \int_0^1 \int_{\mathbb{R}^d} \left( \frac{\partial}{\partial t} \frac{d\rho_t}{d\lambda} \right)^2 \, d\rho_t \, dt = 4 \int_0^1 \int_{\mathbb{R}^d} \left( \frac{\partial}{\partial t} \sqrt{\frac{d\rho_t}{d\lambda}} \right)^2 \, d\lambda(x) \, dt \geq 4 \int_{\mathbb{R}^d} \left( \int_0^1 \frac{\partial}{\partial t} \sqrt{\frac{d\rho_t}{d\lambda}} \, dt \right)^2 \, d\lambda(x) = 4 \int_{\mathbb{R}^d} \left( \sqrt{\frac{d\rho_1}{d\lambda}} - \sqrt{\frac{d\rho_0}{d\lambda}} \right)^2 \, d\lambda.$$

5737
Equality is obtained when
\[
\sqrt{\frac{d\rho_t}{d\lambda}} = (1 - t) \sqrt{\frac{d\rho_0}{d\lambda}} + t \sqrt{\frac{d\rho_1}{d\lambda}}.
\]
(9)

From the expression (8) we can derive immediately that for probability measures \(d_H(\rho_0, \rho_1) \leq 2\sqrt{2}\). We note that
\[
d_H^2(\rho_0, \rho_1) = r_0r_1d_H^2(\rho_0, \rho_1) + 4(r_0 - r_1)^2
\]
and hence \(d_H\) is a cone geodesic distance on the space of positive measures satisfying [45, (2.1)]. Thus from theorem 2.2 and corollary 2.3 of [45] (see also [5, chapter 2]) it follows that the spherical Hellinger distance, which is obtained by restricting the configurations and paths to probability measures and considering the same path lengths, is given by\(^4\)
\[
d_{SH}(\rho_0, \rho_1) = 2\arccos \left( 1 - \frac{d_H^2(\rho_0, \rho_1)}{8} \right) = 4\arcsin \left( \frac{d_H(\rho_0, \rho_1)}{4} \right).
\]
(10)
This implies that \(d_{SH} \leq \pi\) and furthermore \(d_{SH}(\rho_0, \rho_1) = \pi\) if and only if \(\rho_0\) and \(\rho_1\) are orthogonal measures. From the definition (10) one can also observe immediately
\[
d_{SH}(\rho_0, \rho_1) \geq d_H(\rho_0, \rho_1) \quad \text{and} \quad \lim_{d_H(\rho_0, \rho_1) \to 0} \frac{d_{SH}(\rho_0, \rho_1)}{d_H(\rho_0, \rho_1)} = 1.
\]
(11)
Furthermore \((\mathcal{P}(\mathbb{R}^d), d_{SH})\) is a geodesic metric space and [45, theorem 2.7] identifies the geodesics, based on geodesics w.r.t \(d_H\) in the cone of positive measures.

The distances \(d_H, d_{SH}\) metrise strong convergences of measures, which we present in the lemma below.

**Lemma 2.1.** Suppose \(\{\rho_n\}^\infty_{n=1}\) and \(\rho\) are measures on \(\mathbb{R}^d\) and are all absolutely continuous with respect to some measure \(\lambda\). Suppose also that \(\rho\) has finite total mass. Then
\[
\lim_{n \to \infty} d_H(\rho_n, \rho) = 0 \iff \frac{d\rho_n}{d\lambda} \overset{L^1(\lambda)}{\to} \frac{d\rho}{d\lambda}.
\]
(12)
As a consequence of (11), if we further assume \(\rho_n, \rho\) are probability measures on \(\mathbb{R}^d\), then
\[
\lim_{n \to \infty} d_{SH}(\rho_n, \rho) = 0 \iff \frac{d\rho_n}{d\lambda} \overset{L^1(\lambda)}{\to} \frac{d\rho}{d\lambda}.
\]
(13)

**Proof.** Thanks to (11), it suffices prove (12), which is a direct consequence of the following inequalities (see [72, theorem 2.1]): for any \(\rho_1, \rho_2 \ll \lambda\),
\[
d_H^2(\rho_1, \rho_2) \leq \|\rho_1 - \rho_2\|_{L^1(\lambda)} \leq \left( \sqrt{\|\rho_1\|_{L^1(\lambda)}} + \sqrt{\|\rho_2\|_{L^1(\lambda)}} \right) d_H(\rho_1, \rho_2).
\]
\[\square\]

\(^4\) The paper [40] gives an alternative definition of \(d_{SH}\) using Benamou–Brenier formulation \(d_{SB}^2(\rho_0, \rho_1) = \inf_{(\rho_n, u_n)} \int_0^1 \int_0^1 (u_t - \int_0^t \rho_n) d\lambda dt\) with geodesic equation \(\partial_t \rho_n = -\rho_n(u_t - \int_0^t \rho_n)\). As we do not use this formulation, we just remark that showing the equivalence is straightforward. In particular by definition the distances \(d_{SB} \geq d_{SH}\). On the other hand it is direct to check that for the geodesic paths w.r.t \(d_{SB}\), identified in [45, theorem 2.7], the length w.r.t \(d_{SB}\) is the same as w.r.t \(d_{SH}\).
Before presenting the main results of this section, let us state the general assumptions of \( \pi \) that we assume throughout this work.

**Assumption 1.** The invariant measure \( \pi \) and initial condition \( \rho_0 \) are absolutely continuous with respect to the Lebesgue measure and have density functions \( \pi(x), \rho_0(x) \). Let

\[
\Omega := \{ x \in \mathbb{R}^d, \pi(x) > 0 \}.
\]

We require that \( \rho_0 > 0 \) in \( \Omega \) and \( \rho_0 = 0 \) in \( \Omega^c \).

The pure birth–death dynamics \((\text{BD})\) is the \( d_{\text{SH}} \)-gradient flow of relative entropy \( \text{KL}(\cdot | \pi) \). Under sufficient regularity hypotheses, for any energy functional \( G \), the \( d_{\text{SH}} \)-gradient flow of \( G \) has the form

\[
\partial_t \rho_t = -\rho_t \left( \frac{\delta G}{\delta \rho} - \int_{\mathbb{R}^d} \frac{\delta G}{\delta \rho} \, d\rho_t \right).
\]

(14)

Here \( \frac{\delta G}{\delta \rho} \) is the first variation density of \( G \) at \( \rho \) \([2, 15]\).

The following lemma shows the well-posedness of \((\text{BD})\) whenever \( \rho_0 > 0 \) on \( \Omega \). In addition, the proof reveals the deeper structure of \((\text{BD})\) which indicates exponential convergence to \( \pi \).

We prove the convergence rate in theorem 2.4. We note that similar discussion is carried out in the proof of theorem 2.1 in \([51]\).

**Lemma 2.2.** Suppose \( \rho_0, \pi \) satisfies assumption 1, and \( \text{KL}(\rho_0 | \pi) < \infty \), then there exists a unique solution of \((\text{BD})\) \( \rho \in C^1([0, \infty), L^1(\Omega) \cap \mathcal{P}(\Omega)) \), where the differentiability is with respect to \( L^1 \) norm and which dissipates the \( \text{KL} \)-divergence.

**Proof.** Assume that \( \rho \in C^1([0, \infty), L^1(\Omega) \cap \mathcal{P}(\Omega)) \) is a \( \text{KL} \)-dissipating solution of \((\text{BD})\). Then \( \rho = 0 \) a.e. in space and time outside \( \Omega \). In \( \Omega \) we have for \( \rho \)-a.s. \( x \), the function \( \eta_t = \log \frac{\rho}{\pi} \) satisfies the equation

\[
\partial_t \eta_t(x) = -\eta_t(x) + \text{KL}(\rho_t | \pi).
\]

(15)

Note that \( \text{KL}(\rho_t | \pi) \) depends only on \( t \) and is bounded. Thus, using the theory of linear ODEs, should a solution of \((15)\) exist, it has to be of the form

\[
\eta_t(x) = \eta_0(x) e^{-t} + \psi_t.
\]

Taking exponential, we have

\[
\rho_t(x) = \pi(x) \left( \frac{\rho_0(x)}{\pi(x)} \right) e^{-t} \Psi_t
\]

(16)

where \( \Psi_t = e^{\psi_t} \). Since the solution of \((\text{BD})\) must be a probability density for all \( t \), we have \( \Psi_t^{-1} = \int_{\Omega} \left( \frac{\rho_0}{\pi} \right)^{e^{-t}} \, d\pi \), which is uniquely determined by \( \rho_0 \) and \( \pi \). Also, \( \Psi_t \) must be finite and positive since by Hölder’s inequality, \( \rho_0^{e^{-t}} \pi^{1-e^{-t}} \in L^1(\Omega) \) and \( \int_{\Omega} \rho_0^{e^{-t}} \pi^{1-e^{-t}} \leq 1 \). This means the equation \((\text{BD})\) has at most one solution. Finally we can verify the existence of a solution by substituting the expression \((16)\) into \((\text{BD})\). Direct computation also verifies that \( \text{KL} \) divergence is nonincreasing.

Before stating our result we recall the convergence result of \([57]\).
Theorem 2.3 [57, theorem 3.3]. Suppose \( \rho_0, \pi \) satisfies assumption 1. Let \( \rho_t \) be the solution of (BD) with the initial condition \( \rho_0 \) satisfying \( \text{KL}(\rho_0|\pi) \leq 1 \) and that
\[
\inf_{x \in \Omega} \frac{\rho_0(x)}{\pi(x)} \geq e^{-M} \tag{17}
\]
for some \( M > 0 \). Then
\[
\text{KL}(\rho_t|\pi) \leq e^{-(2-3\delta)(t-t_*)} \text{KL}(\rho_0|\pi)
\]
for every \( \delta \in (0, 1/4) \) and all \( t \geq t_* := \log(M/\delta^3) \).

We note that the above theorem requires the condition \( \text{KL}(\rho_0|\pi) \leq 1 \); some result would still hold for \( \text{KL}(\rho_0|\pi) < 2 \) but not for larger bounds. The result in [51] removes the \( \text{KL}(\rho_0|\pi) \leq 1 \) condition, but they also requires a pointwise upper bound for \( \rho_0 \).

We now state our main results that improve the conditions for convergence above by removing the requirement that \( \text{KL}(\rho_0|\pi) \leq 1 \) with the only assumption (18). Furthermore our second result establishes that \( \text{KL}(\rho_t|\pi) \) contracts exponentially fast to 0 at all times \( t \geq 0 \). We remark that asymptotically our rate becomes slower than the one in theorem 2.3; once our bounds ensure that \( \text{KL}(\rho_t|\pi) \leq 1 \) one can apply the results of the above theorem.

Theorem 2.4. Under the assumptions of lemma 2.2, and let \( \rho_t \) satisfy the pure birth–death dynamics (BD) with initial condition \( \rho_0 \in L^1(\Omega) \cap \mathcal{P}(\Omega) \). Then for any \( \rho_0 \) satisfying
\[
\inf_{x \in \Omega} \frac{\rho_0(x)}{\pi(x)} \geq e^{-M}, \tag{18}
\]
for some constant \( M \), we have for all \( t > 0 \)
\[
\text{KL}(\rho_t|\pi) \leq Me^{-t} + e^{-t+Me^{-t}} \text{KL}(\rho_0|\pi), \tag{19}
\]
as well as
\[
\text{KL}(\rho_t|\pi) \leq \exp\left(-\int_0^t \lambda(s) \, ds\right) \text{KL}(\rho_0|\pi), \text{ with } \lambda(t) = \frac{M^2e^{-2t}}{9eMe^{-t}(eMe^{-t} - Me^{-t} - 1)}. \tag{20}
\]

Proof. We first prove (19). Recall from the proof of lemma 2.2 that
\[
\rho_t = \pi \left( \frac{\rho_0}{\pi} \right)^{e^{-t}} \Psi_t
\]
for some \( \Psi_t > 0 \). Moreover, notice that under the condition (18),
\[
\frac{1}{\Psi_t} = \int_\Omega \pi \left( \frac{\rho_0}{\pi} \right)^{e^{-t}} \geq e^{-Me^{-t}},
\]
which implies $\Psi_t \leq e^{Me^{-t}}$. Therefore

$$\text{KL}(\rho_0|\pi) = \Psi_t \int_\pi \left( \rho_0 \right)^{e^{-t}} \left( \log \Psi_t + e^{-t} \log \rho_0 \right)$$

$$= \log \Psi_t + \Psi_t e^{-t} \int_\pi \rho_0 \left( \frac{\rho_0}{\pi} \right)^{e^{-t}} \log \rho_0$$

$$\leq Me^{-t} + e^{-t+Me^{-t}} \int_\pi \rho_0 \log \rho_0 = Me^{-t} + e^{-t+Me^{-t}} \text{KL}(\rho_0|\pi), \quad (21)$$

where in the last inequality we used that if $\rho_0 \geq \pi$ then $\left( \frac{\rho_0}{\pi} \right)^{e^{-t}} \leq \frac{\rho_0}{\pi}$ and $\log \frac{\rho_0}{\pi} \geq 0$; meanwhile if $\rho_0 < \pi$ then $\left( \frac{\rho_0}{\pi} \right)^{e^{-t}} \geq \frac{\rho_0}{\pi}$ and $\log \frac{\rho_0}{\pi} \leq 0$.

We now prove (20). The proof strategy is a modification of [40, lemmas 2.11 and 2.12] and [41, theorem 4.1]. We divide the proof into three steps.

**Step 1:** proof of

$$\int_\Omega \rho \log \frac{\rho}{\pi} \, dx \leq \frac{e^M - M - 1}{M^2} \int_\Omega \rho \log^2 \frac{\rho}{\pi} \, dx. \quad (22)$$

The key of this step is to prove that for any $r \geq e^{-M}$,

$$\frac{\log r - 1 + \frac{1}{2}}{\log^2 r} \leq \frac{e^M - M - 1}{M^2}. \quad (23)$$

Let $\varphi(r) = \frac{\log r + 1}{2}$, then $\varphi'(r) = \frac{1}{2} \log r$. Now let $\psi(r) = r \log r - r - 2 + 2r$, then $\psi'(r) = 1 - \log r + \frac{1}{2} \leq 0$, so for any $r \geq 1$, $\psi(r) \leq \psi(1) = 0$, and therefore $\varphi'(r) \leq 0$; on the other hand, when $r \leq 1$, $\psi(r) \geq \psi(1) = 0$, which again yields $\varphi'(r) \leq 0$, and thus when $r \in (e^{-M}, \infty)$, the maximum of $\varphi(r)$ is attained at $r = e^{-M}$, which finishes the proof of (23). Thus taking $r = \frac{\rho}{\pi}$ for any $\rho$ satisfying (18), we have

$$\frac{\rho \log \frac{\rho}{\pi} - \rho + \pi}{\rho \log^2 \frac{\rho}{\pi}} \leq \frac{e^M - M - 1}{M^2},$$

which indicates (22) after integration.

**Step 2:** we strengthen (22) into

$$\int_\Omega \rho \log \frac{\rho}{\pi} \, dx \leq \frac{9 \rho_0^M (e^M - M - 1)}{M^2} \int_\Omega \rho \left( \log \frac{\rho}{\pi} - \int_\Omega \rho \log \frac{\rho}{\pi} \, dx \right)^2 \, dx. \quad (24)$$

Let $a = \int \rho \log \frac{\rho}{\pi} \, dx > 0$. If $\mathbb{P}_\pi \left( e^{-M} \leq \frac{\rho}{\pi} \leq e^{\frac{1}{2}} \right) \geq \frac{1}{2}$, then, noticing $\log \frac{\rho}{\pi} - a \leq -\frac{a}{2} < 0$ for $\frac{\rho}{\pi} < e^{\frac{1}{2}}$, we obtain

$$\int_\Omega \rho \left( \log \frac{\rho}{\pi} - a \right)^2 \, dx \geq e^{-M} \int_{e^{-M} \leq \frac{\rho}{\pi} \leq e^{\frac{1}{2}}} \left( \log \frac{\rho}{\pi} - a \right)^2 \, dx$$

$$\geq \frac{a^2}{4} e^{-M} \mathbb{P}_\pi \left( e^{-M} \leq \frac{\rho}{\pi} \leq e^{\frac{1}{2}} \right) \geq \frac{a^2}{8} e^{-M}. \quad (25)$$

5741
Otherwise we must have $\mathbb{P}_\pi \left( \frac{\rho}{\pi} > e^\xi \right) \geq \frac{1}{2}$. Notice for probability densities we have $\hat{\rho} > \pi \rho d\pi \geq \hat{\rho} > e^\xi \pi \rho d\pi \geq e^{\xi^2 - 1} \geq \frac{a}{4}$.

We can estimate the l.h.s. by

$$ \int_{\rho > \pi} \rho - \pi d\pi \geq \int_{\rho > e^{\xi^2}} \rho - \pi d\pi \geq (e^{\xi^2} - 1) \mathbb{P}_\pi \left( \frac{\rho}{\pi} > e^\xi \right) \geq \frac{1}{2} (e^{\xi^2} - 1) \geq \frac{a}{4}. $$

On the other hand,

$$ \int_{\rho < \pi} \pi - \rho d\pi \leq \int_{\rho < \pi} \pi \log \frac{\pi}{\rho} d\pi, $$

which means that

$$ \int_{\rho < \pi} \pi \log \frac{\pi}{\rho} d\pi \geq \frac{1}{4}. $$

Hence, using that $\mathbb{P}_\pi (\rho < \pi) \leq \frac{1}{2}$, we obtain

$$ \int_{\Omega} \rho \left( \log \frac{\rho}{\pi} - a \right)^2 d\rho \geq e^{-M} \int_{\rho < \pi} \pi \left( \log \frac{\pi}{\rho} + a \right)^2 d\rho $$

$$ \geq e^{-M} \left( \int_{\rho < \pi} \pi \log \frac{\pi}{\rho} d\rho \right)^2 \geq \frac{a^2}{8} e^{-M}. \quad (26) $$

Thus, combining (25) and (26), in any case we obtain

$$ \left( \int_{\Omega} \rho \log \frac{\rho}{\pi} d\rho \right)^2 \leq 8e^M \int_{\Omega} \rho \left( \log \frac{\rho}{\pi} - \int_{\Omega} \rho \log \frac{\rho}{\pi} d\rho \right)^2 d\rho. $$

Finally, by further combining (22), we arrive at

$$ \int_{\Omega} \rho \log \frac{\rho}{\pi} d\rho \leq \frac{e^{M} - M - 1}{M^2} \left( \int_{\Omega} \rho \left( \log \frac{\rho}{\pi} - \int_{\Omega} \rho \log \frac{\rho}{\pi} d\rho \right)^2 d\rho + \left( \int_{\Omega} \rho \log \frac{\rho}{\pi} d\rho \right)^2 \right) $$

$$ \leq \frac{(1 + 8e^M)}{M^2} \left( \frac{e^{M} - M - 1}{M^2} \right) \int_{\Omega} \rho \left( \log \frac{\rho}{\pi} - \int_{\Omega} \rho \log \frac{\rho}{\pi} d\rho \right)^2 d\rho. $$

Step 3: proof of exponential convergence. From the proof of lemma 2.2 we have $\Psi_t \geq 1$. By taking infimum on both sides of (16), we obtain

$$ \inf_{x \in \Omega} \frac{\rho_t(x)}{\pi(x)} = \Psi_t \inf_{x \in \Omega} \left( \frac{\rho_0(x)}{\pi(x)} \right)^{-t} \geq e^{-Me^{-t}}, \quad (27) $$
In other words, $\rho_t$ satisfies (18) with $Me^{-t}$ playing the role of $M$. Therefore, a combination of direct time differentiation and (24) yields

$$\frac{d}{dt} \text{KL}(\rho_t|\pi) = -\int_\Omega \rho_t \left( \log \frac{\rho_t}{\pi} - \int_\Omega \rho_t \log \frac{\rho_t}{\pi} \ dx \right)^2 \ dx \leq -\lambda (t) \text{KL}(\rho_t|\pi).$$

The convergence result (20) therefore directly follows from a Gronwall inequality. □

**Remark 2.5.** The condition (18) can be relaxed or modified if we add conditions on $\pi$. One such modification, inspired by [17, proposition 3.23], is that, suppose for some $p \in [1, \infty)$ we have $M_p(\pi) := \int_\Omega |x|^p \, d\pi(x) < \infty$, and in $\Omega$, we have

$$\frac{\rho_0(x)}{\pi(x)} \geq e^{-M(1+|x|^p)},$$

then along (BD), we have convergence

$$\text{KL}(\rho_t|\pi) \leq Me^{-t} (1 + M_p(\pi)) + \exp \left( -t + Me^{-t} (1 + M_p(\pi)) \right) \text{KL}(\rho_0|\pi).$$

The proof follows closely along that of (19), with the difference being

$$\frac{1}{\Psi_t} = \int_\Omega \left( \frac{\rho_0(x)}{\pi(x)} \right)^{e^{-t}} \, d\pi \geq \int_\Omega \left( e^{-M(1+|x|^p)} \right)^{e^{-t}} \, d\pi = \int_\Omega e^{-Me^{-t}(1+|x|^p)} \, d\pi \geq \exp \left( -\int Me^{-t}(1+|x|^p) \, d\pi \right) = \exp \left( -Me^{-t}(1 + M_p(\pi)) \right),$$

and we finish the proof after substituting this into (21). This new assumption (28) covers almost all reasonable scenarios with $\rho_0$ being Gaussian, as long as $\pi$ has second moment. The upper bound in the assumption of [17] is unnecessary. As is suggested in [57], the optimal asymptotic convergence rate should be $e^{-2t}$, which is proved in [26] under a different set of assumptions.

**Remark 2.6.** Combining Langevin dynamics with the birth–death dynamics would result in dynamics with convergence rate that is at least the maximum of the Log-Sobolev constant of $\pi$ and the rates obtained in theorem 2.4. That is, suppose $\pi$ satisfies a Log-Sobolev inequality with constant $C_{LSI}$, then for the dynamics

$$\partial_t \rho_t = -\rho_t \log \frac{\rho_t}{\pi} + \rho_t \int_\Omega \rho_t \log \frac{\rho_t}{\pi} \ dx + \nabla \cdot \left( \rho_t \nabla \log \frac{\rho_t}{\pi} \right),$$

as long as $\rho_0$ satisfies (18), we have convergence

$$\text{KL}(\rho_t|\pi) \leq \min \left\{ \exp \left( -\int_0^t \tilde{\lambda}(s) \, ds \right) \text{KL}(\rho_0|\pi), Me^{-t} + e^{-t+Me^{-t}} \text{KL}(\rho_0|\pi) \right\},$$

with

$$\tilde{\lambda}(t) = \max \left\{ C_{LSI}, 9e^{Me^{-t}} \frac{M^2 e^{-2t}}{e^{Me^{-t}} - Me^{-t} - 1} \right\}.$$

Convergence rate of $C_{LSI}$ is guaranteed even without the condition (18).
At the end of this section, we would like to use two examples to illustrate that the pointwise lower bound condition (18) is not numerically restrictive. In particular, if $V$ has at least quadratic growth at infinity, one can choose $\rho_0$ to be any sufficiently wide round Gaussian to satisfy (18).

Example 2.7. Suppose $V(x)$ is strongly convex and $\frac{3}{2} |x|^2 \leq V(x) \leq \frac{4}{2} |x|^2$. Then we can pick $\rho_0(x) = \left( \frac{m}{2\pi} \right)^{\frac{d}{2}} \exp(-\frac{m}{2} |x|^2)$, and therefore

$$\inf_{x \in \mathbb{R}^d} \frac{\rho_0}{\pi} = \inf_{x \in \mathbb{R}^d} Z \left( \frac{m}{2\pi} \right)^{\frac{d}{2}} \exp \left( V(x) - \frac{m}{2} |x|^2 \right) \geq Z \left( \frac{m}{2\pi} \right)^{\frac{d}{2}} \geq \left( \frac{m}{L} \right)^{\frac{d}{2}},$$

which means $\rho_0$ satisfies (18) with $M = \frac{d}{2} \log \frac{L}{m}$. Moreover, after a time of $t \geq \log M = \log(d \log \frac{L}{m})$, the convergence rate becomes $O(1)$.

Example 2.8. Let us consider the double well potential $V(x) = \frac{1}{2} (1 - |x|^2)^2$. We pick $\rho_0 = \left( \frac{4}{\pi \varepsilon} \right)^{\frac{d}{2}} \exp^{-\frac{d}{2} |x|^2}$, then

$$\frac{\rho_0}{\pi} = Z \left( \frac{4}{\pi \varepsilon} \right)^{\frac{d}{2}} \exp \left( \frac{1}{2 \varepsilon} (|x|^2 - 2)^2 - \frac{3}{2} \varepsilon \right) \geq \left( \frac{1}{2 \pi \varepsilon} \right)^{\frac{d}{2}} Z \exp \left( -\frac{3}{2} \varepsilon \right).$$

Notice that

$$Z = \sigma (2^d - 1) \int_0^\infty r^{d-1} \exp \left( -\frac{1}{\varepsilon} (1 - r)^2 \right) dr \geq \frac{\pi^{\frac{d}{2}}}{d \sqrt{\pi} d \left( \frac{d}{2\pi} \right)^{\frac{d}{2}} \exp \left( -\frac{1}{\varepsilon} \right),$$

which means $\rho_0$ satisfies (18) with $M = \frac{d}{2} \log \frac{d}{\pi} + \frac{d}{2}$, and therefore the burn-in time needed is $O(\log M) = O(\log d + \log \frac{1}{\varepsilon})$.

3. Pure birth–death dynamics governed by chi-squared divergence

In this section we consider the spherical Hellinger gradient flow with $\chi^2(\rho|\pi) := \int_\Omega \left( \frac{\rho}{\pi} - 1 \right)^2 d\pi$ as the energy functional:

$$\partial_t \rho_t = -\rho_t \left( \frac{\rho_t}{\pi} - \int_\Omega \frac{\rho_t}{\pi} d\rho_t \right).$$

This is the dynamics appeared in [50, (3.6)]. There the authors first derive a related family of dynamics, [50, (3.1)], as the continuum limit of the ensemble Monte-Carlo sampling schemes they introduced. For the dynamics [50, (3.1)], with kernel $Q = 1d$, they prove exponential decay of the ‘reverse’ $\chi^2$-divergence. In the time scaling we consider, this can be stated as $\chi^2(\pi|\rho_t) \lesssim e^{-t}$, where $\rho_t$ is the rescaling in time for which $\frac{d\rho_t}{dt} = \chi^2(\rho_t|\pi) + 1$. Since $\chi^2(\rho_t|\pi) \rightarrow 0$ as $t \rightarrow \infty$, $\frac{d\rho_t}{dt} \rightarrow 1$ as $t \rightarrow \infty$. Thus the exponential rates of [50, theorem 1] implies asymptotic exponential rates of order $e^{-t}$ for $\chi^2(\pi|\rho_t)$. However, due to the non-symmetry of the $\chi^2$-divergence, the convergence result on $\chi^2(\pi|\rho_t)$ from [50] does not directly imply a convergence result for $\chi^2(\rho_t|\pi)$, although the later is a more natural Lyapunov function for the gradient flow (BD2) since it is the underlying energy. In the next theorem, we show that $\chi^2(\rho_t|\pi)$ also contracts exponentially fast and provide a quantitative estimate for the convergence rate. The authors of [50] also note that the dynamics we consider, (BD2), is formally spherical Hellinger gradient flow.
Theorem 3.1. Let \( \rho_0, \pi \) satisfy assumption 1, and let \( \rho_t \in C^1([0,\infty), L^1(\Omega) \cap \mathcal{P}(\Omega)) \) be the solution of (BD2) with initial condition \( \rho_0 \). Then, for any initial probability density \( \rho_0(x) \) such that

\[
\inf_{x \in \Omega} \frac{\rho_0(x)}{\pi(x)} \geq e^{-M}
\]  

for some \( M > 1 \), we have exponential convergence to equilibrium along the dynamics (BD2)

\[
\chi^2(\rho_t|\pi) \leq \exp \left( -\int_0^t \lambda(s) \, ds \right) \chi^2(\rho_0|\pi),
\]

with

\[
\lambda(t) = \frac{2}{(9 + 8(e^M - 1)e^{-t})(1 + (e^M - 1)e^{-t})}.
\]  

Proof. The proof is similar to that of (20) in theorem 2.4. The core step is the following functional inequality which holds for any \( \rho \) satisfying \( \inf \rho \pi \geq e^{-M} \),

\[
\int_\Omega \frac{\rho^2}{\pi} \, dx \leq e^{M}(1 + 8e^M) \int_\Omega \rho \left( \frac{\rho}{\pi} - \int_\Omega \frac{\rho^2}{\pi} \, dx \right)^2 \, dx.
\]  

Let \( a = \int \frac{\rho}{\pi} \, dx > 1 \). If \( \mathbb{P}_\pi \left( e^{-M} \leq \frac{\rho}{\pi} \leq \frac{a + 1}{2} \right) \geq \frac{1}{2} \), then

\[
\int_\Omega \rho \left( \frac{\rho}{\pi} - a \right)^2 \, dx \geq e^{-M} \int_{e^{-M} \leq \frac{\rho}{\pi} \leq \frac{a + 1}{2}} \left( \frac{\rho}{\pi} - a \right)^2 \, dx \geq \frac{(a - 1)^2}{4e^M} \mathbb{P}_\pi \left( e^{-M} \leq \frac{\rho}{\pi} \leq \frac{a + 1}{2} \right) \geq \frac{(a - 1)^2}{8e^M}.
\]

Otherwise \( \mathbb{P}_\pi \left( \frac{\rho}{\pi} \geq \frac{a + 1}{2} \right) \geq \frac{1}{2} \), which means

\[
\int_{\rho < \pi} \pi - \rho \, dx = \int_{\rho > \pi} \rho - \pi \, dx \geq \int_{\rho > \frac{a + 1}{2}} (\rho - \pi) \, dx \geq \frac{a - 1}{2} \mathbb{P}_\pi \left( \frac{\rho}{\pi} \geq \frac{a + 1}{2} \right) \geq \frac{a - 1}{4}.
\]

Therefore

\[
\int_\Omega \rho \left( \frac{\rho}{\pi} - a \right)^2 \, dx \geq e^{-M} \int_{\rho < \pi} \pi \left( \frac{\rho}{\pi} - a \right)^2 \, dx \geq \frac{\left( \int_{\rho < \pi} (1 - \frac{\rho}{\pi}) \, dx \right)^2}{e^M \mathbb{P}_\pi (\rho < \pi)} \geq \frac{(a - 1)^2}{8e^M}.
\]

To conclude,

\[
\int_\Omega \pi \left( \frac{\rho}{\pi} - 1 \right)^2 \, dx \leq e^M \int_\Omega \rho \left( \frac{\rho}{\pi} - 1 \right)^2 \, dx = e^M \left( \int_\Omega \rho \left( \frac{\rho}{\pi} - a \right)^2 \, dx + (a - 1)^2 \right) \leq e^M (1 + 8e^M) \int_\Omega \rho \left( \frac{\rho}{\pi} - a \right)^2 \, dx.
\]
This finishes the proof of (31). Now let us return to the dynamics (BD2). Taking time derivative, we have for $e^{-M(t)} = \inf \frac{\rho}{\pi}$,

$$\frac{d}{dt} \int_{\Omega} \pi \left( \frac{\rho_t}{\pi} - 1 \right)^2 dx = -2 \int_{\Omega} \rho_t \left( \frac{\rho_t}{\pi} - \int_{\Omega} \frac{\rho_t^2}{\pi} dx \right)^2 dx \leq \frac{2}{e^{M(t)} (1 + 8e^{M(t)})} \int_{\Omega} \pi \left( \frac{\rho_t}{\pi} - 1 \right)^2 dx.$$

By Gronwall inequality, this means the dynamics converge exponentially with instantaneous rate $\lambda(t) = \frac{2}{e^{M(t)} (1 + 8e^{M(t)})}$. Finally, notice that

$$\frac{d}{dt} \frac{\rho}{\pi} = -\frac{\rho^2}{\pi^2} + \frac{\rho}{\pi} \int_{\Omega} \frac{\rho^2}{\pi} dx \geq -\frac{\rho^2}{\pi^2} + \frac{\rho}{\pi}.$$

Therefore, solving the ODE, one has

$$e^{-M(t)} \geq \frac{e^t}{e^M + e - 1},$$

which gives the convergence rate in (30).

Notice that one has $\lim_{t \to \infty} \lambda(t) = \frac{2}{9}$, which we believe is suboptimal based on the results in theorem 2.4 (i) as well as the observation made in [50]. On the other hand, if $M \gg 1$, then the instantaneous convergence rate is $O(1)$ only when $e^{M-t} = O(1)$, which means $t \geq O(M)$. Hence the waiting time of (BD2) is longer than that of (BD), which is $O(\log M)$.

4. Kernelised dynamics and its particle approximations

In this section we investigate a particle-based approximation to the dynamics (BD). We first introduce a nonlocal approximation of (BD) that is based on regularising the relative entropy:

$$\frac{d}{dt} \rho^{(e)} = -\rho^{(e)} \left[ \log \left( \frac{K_e \ast \rho^{(e)}}{\pi} \right) + K_e \ast \left( \frac{\rho^{(e)}}{K_e \ast \rho^{(e)}} \right) - \int \log \left( \frac{K_e \ast \rho^{(e)}}{\pi} \right) d\rho^{(e)} - 1 \right].$$

(BD_ε)

It is the spherical Hellinger gradient flow of the regularised entropy

$$\mathcal{F}_\varepsilon(\rho) = \int \log (K_e \ast \rho) d\rho - \int \log \pi d\rho = \int \log (K_e \ast \rho) d\rho + \int Vd\rho + C,$$

(32)

where $C = \log \left( \int \exp(-V(x)) dx \right)$. We first study the energy $\mathcal{F}_\varepsilon$ on the whole space. In sections 4.2 and 4.3 we study the well posedness and the convergence as $\varepsilon \to 0$ of the gradient flow on the torus.

We now state the conditions on the kernel $K_e$ that we require for our results.
Assumption 2. The kernel $K_\varepsilon(x-y)$ is of the form $K_\varepsilon(x-y) = \varepsilon^{-d}K(\frac{x-y}{\varepsilon})$, where $K \in C^\infty(\mathbb{R}^d) \cap L^\infty(\mathbb{R}^d)$ satisfies the following:

(i) $\int_{\mathbb{R}^d} K \, dx = 1$ and $K$ is positive definite, in the sense that for any function $f \in C^\infty_c(\mathbb{R}^d)$,
$$
\int K(x-y)f(x)f(y) \, dx \, dy \geq 0.
$$

(ii) $K$ is radially symmetric, i.e. $K(x) = K(|x|)$, and $M_2(K) := \int_{\mathbb{R}^d} |x|^2 K(x) \, dx < \infty$.

Assumption 2 also indicates that there exists a kernel $\xi \geq 0$ such that $K = \xi \ast \xi$ and $\int_{\mathbb{R}^d} \xi = 1$, namely $\xi = \sqrt{K}$. One example that satisfies assumption 2 is the Gaussian kernel $K(x) = (2\pi)^{-d/2} e^{-\frac{|x|^2}{2}}$, in which case $\xi = K^{\frac{1}{2}}$.

4.1. Quantitative distance between minimisers

The $\Gamma$-convergence of $F_\varepsilon$, defined in (32), to relative entropy $KL(\rho|\pi)$ and the convergence of minimisers are already proved in [16], which we restate in the following theorem 4.1.

**Theorem 4.1 ([16]).** Suppose $\pi = \exp(-V) \in \mathcal{P}_2(\mathbb{R}^d)$. Let $K_\varepsilon$ satisfy assumption 2.

(i) ([16, theorem 4.1]) As $\varepsilon \to 0$, $F_\varepsilon$ defined in (32) $\Gamma$-converges to $F(\rho) := KL(\rho|\pi)$, in the sense that for any sequence $\rho_\varepsilon(\cdot) \rightharpoonup \rho$, we have $\liminf_{\varepsilon \to 0} F_\varepsilon(\rho_\varepsilon(\cdot)) \geq F(\rho)$. Moreover, $\limsup_{\varepsilon \to 0} F_\varepsilon(\rho) \leq F(\rho)$.

(ii) ([16, theorem 4.5]) Suppose in addition that $K$ is Gaussian, and that there exists a constant $C$ such that $V(x) \geq C(|x|^2 - 1)$, then minimisers of $F_\varepsilon$ over $\mathcal{P}_2(\mathbb{R}^d)$ exist. Moreover, for any sequence $(\pi_\varepsilon)_\varepsilon$ such that $\pi_\varepsilon \in \mathcal{P}_2(\mathbb{R}^d)$ is a minimiser of $F_\varepsilon$, we have, up to a subsequence, $\pi_\varepsilon \rightharpoonup \pi$.

However, [16] did not prove quantitatively how close the minimisers $\pi_\varepsilon$ are to $\pi$. In the case $V$ is sufficiently regular and has between quadratic and quartic growth, we can prove a more quantitative result. We would like to comment here that, while our assumption on $V$ is slightly stronger than that of theorem 4.1, we do not require $K$ to be Gaussian in our following theorem.

**Theorem 4.2.** Suppose $K_\varepsilon$ satisfies assumption 2 with some $\xi \in \mathcal{P}_2(\mathbb{R}^d)$. Suppose $\pi$ satisfies a Talagrand inequality [63] with some constant $m > 0$:
$$
W_2(\rho, \pi) \leq \sqrt{\frac{2}{m}} KL(\rho|\pi)
$$
for any probability measure $\rho$. Suppose also that
$$
||D^2V(x)|| \leq L(1 + |x|^2) \text{ and } ||D^4V(x)|| \leq L
$$
for some $L > 1$. Then, for any $0 < \varepsilon < \sqrt{\frac{m}{4AM(\xi)}}$, let $\pi_\varepsilon$ be a minimiser of $F_\varepsilon$ defined in (32), we have for some constant $C = C(\pi, \xi)$ independent of $\varepsilon$,
$$
W_2(\pi_\varepsilon, \pi) \leq C\varepsilon,
$$
\[
0 \geq \mathcal{F}_\varepsilon (\pi_\varepsilon) \geq -Ce^2.
\]  

(36)

We remark that analogous result holds if we consider the energy on the torus \(\mathbb{T}^d\). The integrals considered are on the torus, while for evaluating convolutions, all of the functions are extended periodically to \(\mathbb{R}^d\).

**Proof.** To get started, we have for arbitrary \(\rho\),

\[
\mathcal{F}_\varepsilon (\rho) - KL (\rho|\pi) = -KL (\rho|K_\varepsilon * \rho) \leq 0.
\]

(37)

In particular, taking \(\rho = \pi_\varepsilon\) where \(\pi_\varepsilon\) is any minimiser of \(\mathcal{F}_\varepsilon\), we have

\[
\mathcal{F}_\varepsilon (\pi_\varepsilon) \leq \mathcal{F}_\varepsilon (\pi) \leq KL (\pi|\pi) = 0.
\]

(38)

On the other hand, since \(K_\varepsilon = \xi_\varepsilon * \xi_\varepsilon\) and \(x \mapsto \log x\) is concave, we use Jensen inequality to obtain

\[
\log (K_\varepsilon * \rho) = \log (\xi_\varepsilon * \xi_\varepsilon * \rho) \geq \xi_\varepsilon * \log (\xi_\varepsilon * \rho)
\]

(39)

We then use the regularity of \(V\) and Talagrand inequality:

\[
\mathcal{F}_\varepsilon (\rho) = \int \rho \log (K_\varepsilon * \rho) - \int \rho \log \pi
\]

\[
\geq \int (\xi_\varepsilon * \rho) \log (\xi_\varepsilon * \rho) - \int \rho \log \pi
\]

\[
= KL (\xi_\varepsilon * \rho|\pi) + \int (\xi_\varepsilon * \rho - \rho) \log \pi
\]

\[
\geq \frac{m}{2} W^2_2 (\xi_\varepsilon * \rho, \pi) + \int \rho(x) \int \xi_\varepsilon (x-y) (V(x) - V(y)) \, dy \, dx.
\]

(40)

For the first term, we can use triangle inequality

\[
W^2_2 (\xi_\varepsilon * \rho, \pi) \geq |W^2_2 (\rho, \pi) - W^2_2 (\xi_\varepsilon * \rho, \pi)| \geq W^2_2 (\rho, \pi) - \sqrt{M_2 (\xi)\varepsilon}.
\]

(41)

For the second term, by Taylor expansion, we have\(^5\)

\[
V(y) - V(x) = \nabla V(x) \cdot (y-x) + \frac{1}{2} (y-x)^\top D^2 V(x) (y-x) + \frac{1}{6} (y-x)^3 : D^3 V(x)
\]

\[
+ \frac{1}{24} (y-x)^4 : D^4 V(\zeta)
\]

\(^5\) Here we use the short hand notation \((y-x)^3 : D^3 V(x) := \sum_{i,j,k} (y_i-x_i)(y_j-x_j)(y_k-x_k) \partial_{ijk} V(x)\), similarly for the fourth derivative term.
for some $\zeta \in [x, y]$. Then, we appeal to the symmetry of the kernel $\xi$ and derive,

$$
\int \rho(x) \int \xi \left( (x-y)(V(x) - V(y)) \right) \, dy \, dx \\
= \int \rho(x) \int \xi \left( (x-y) \left( \frac{1}{2} (y-x) \top D^2 V(x) (y-x) + \frac{1}{24} (y-x)^4 : D^4 V(\xi) \right) \right) \, dy \, dx \\
\geq - \frac{L}{2} \int \left( 1 + |x|^2 \right) \xi \left( (x-y) |y-x|^2 \rho(x) - \frac{L}{24} \int \rho(x) \xi (x-y) |y-x|^4 \right) \, dx \\
= - \frac{L \varepsilon^2}{2} M_2(\xi) \int \rho(x) \left( 1 + |x|^2 \right) \, dx - \frac{L \varepsilon^4}{24} M_4(\xi) \\
\geq - L \varepsilon^2 M_2(\xi) \left( M_2(\pi) + W^2_2(\rho, \pi) + \frac{1}{2} \right) - \frac{L \varepsilon^4}{24} M_4(\xi).
$$

(42)

Here in the last step, we use that for any $\gamma(x, y)$ being a coupling between $\rho$ and $\pi$,

$$
\int |\gamma|^2 \, d\gamma \leq 2 \inf_{\gamma} \int |x-y|^2 \, d\gamma(x, y) + 2 \int |y|^2 \, d\gamma(y) = 2 W^2_2(\rho, \pi) + 2 M_2(\pi).
$$

If $W_2(\pi_\varepsilon, \pi) \leq \varepsilon \sqrt{M_2(\xi)}$ then the theorem is immediate. Otherwise, we combine (40)--(42) for $\rho = \pi_\varepsilon$ to obtain that there exists a constant $C = C(\pi, \xi) > 0$ independent of $\varepsilon$ such that

$$
0 \geq J_\varepsilon(\pi_\varepsilon) \geq \frac{m}{2} \left( W_2^2(\pi_\varepsilon, \pi) - \sqrt{M_2(\xi)} \varepsilon \right)^2 - L \varepsilon^2 W^2_2(\pi_\varepsilon, \pi) M_2(\xi) - C \varepsilon^2,
$$

which leads to (35) after completing the square, and (36) after optimising the right-hand side above with $W_2(\pi_\varepsilon, \pi)$.

\[ \Box \]

**Remark 4.3.** We are not able to prove the sharpness of the error estimate (35). However, we demonstrate via numerical experiments that the error bound $O(\varepsilon)$ above seems to be optimal. Consider the target measure $\pi(x) \propto e^{-V(x)}$ on the one-dimensional torus $T = \mathbb{R}/2\pi\mathbb{Z}$ with potential $V(x) = \sin x + 2 \sin 2x$. We solve the equations (BD) and (BD$_\varepsilon$) using the finite difference method and use numerical integration to compute the evolution of $\text{KL}(\rho_\varepsilon^1 | \pi)$ for various values of $\varepsilon$. The left side of figure 1 shows that for a fixed $\varepsilon > 0$, $\text{KL}(\rho_\varepsilon^1 | \pi)$ approaches to a positive constant as $t \to \infty$. The right side of figure 1 plots the function $\varepsilon \mapsto \text{KL}(\rho_\varepsilon^2 | \pi)$ for $T = 15$, which indicates that $\text{KL}(\rho_\varepsilon^2 | \pi) \approx O(\varepsilon^2)$ as $\varepsilon \to 0$. This is consistent with the scaling in (35) in view of the Talagrand inequality (33).

### 4.2. Well-posedness of gradient flows (BD$_\varepsilon$)

In this section we develop the well-posedness of (BD$_\varepsilon$) considered on the torus $\mathbb{T}^d$ and establish the elements of its gradient flow structure relevant for the convergence argument. Due to the smoothness of the terms of the equation it is easy to show well-posedness in the classical sense of ODE in Banach spaces. More precisely we first show local well-posedness in the spaces of measures on the torus which are bounded from above and below by a positive constant. We then show $L^\infty$ bounds that allow us to extend the solution up to some large $T_\varepsilon \xrightarrow{\varepsilon \to 0} \infty$. After
establishing the existence and uniqueness of classical $L^1$ solutions, we show that both energies $\mathcal{F}$ and $\mathcal{F}_\varepsilon$ are $\lambda$-geodesically convex with respect to $d_{\text{SH}}$, in the restricted sense described in definition 4.9. We then characterise the subdifferential of $\mathcal{F}_\varepsilon$ with respect to the $d_{\text{SH}}$ geometry. These allow us to show that the classical solutions coincide with gradient flow solutions, as well as the curves of maximal slope.

For $C > 1$ let

$$\mathcal{P}_C = \left\{ \rho \in \mathcal{P}(T^d) \cap L^1(T^d), \text{and } \frac{1}{C} \leq \rho(x) \leq C, \text{a.e. in } T^d \right\}.$$  \hspace{1cm} (43)

**Lemma 4.4.** Assume $K_\varepsilon$ satisfies assumption 2, and $\pi(x) \propto e^{-V(x)}$ where $V$ is a $C^2$ function of $T^d$. For any $C > 1$ and $\rho_0 \in \mathcal{P}_C$, there exists $T > 0$, independent of $\rho_0$ and a unique $\rho(\varepsilon) \in C^1([0, T], (\mathcal{P}_C, \| \cdot \|_{L^1}))$ solving (BD$$_\varepsilon$). Moreover $\rho(\varepsilon) \in C^1([0, T], (\mathcal{P}_C, \cdot \cdot \cdot \|_{L^1}))$. Finally if $\rho_0 \in C^1(T^d)$ then $\rho(\varepsilon) \in C^1(T^d)$ for all $t \in [0, T]$.

**Proof.** The existence and uniqueness follow from the classical existence of ODE in Banach spaces. In particular we claim that the right-hand side of (BD$$_\varepsilon$),

$$A(\rho(\varepsilon)) := \rho(\varepsilon) \left[ \log \left( \frac{K_\varepsilon \ast \rho(\varepsilon)}{\pi} \right) + K_\varepsilon \ast \left( \frac{\rho(\varepsilon)}{K_\varepsilon \ast \rho(\varepsilon)} \right) - \int \log \left( \frac{K_\varepsilon \ast \rho(\varepsilon)}{\pi} \right) d\rho(\varepsilon) - 1 \right]$$

is Lipschitz continuous in $L^1$ norm on $\mathcal{P}_C$ and is uniformly bounded in $L^\infty$. That is, there exist $L, M \in (0, \infty)$ such that for all $\rho, \sigma \in \mathcal{P}_C$,

$$\|A(\rho) - A(\sigma)\|_{L^1(T^d)} \leq L\|\rho - \sigma\|_{L^1(T^d)} \quad \text{and} \quad \|A(\rho)\|_{L^\infty(T^d)} \leq M.$$
\[ \left\| K_\varepsilon \left( \frac{\sigma}{K_\varepsilon \ast \sigma} - \frac{\rho}{K_\varepsilon \ast \rho} \right) \right\|_{L^1} \leq \left\| \frac{\sigma}{K_\varepsilon \ast \sigma} - \frac{\rho}{K_\varepsilon \ast \rho} \right\|_{L^1} + \left\| \frac{\rho K_\varepsilon (\sigma - \rho)}{K_\varepsilon \ast \sigma K_\varepsilon \ast \rho} \right\|_{L^1} \leq (C + C^2) \left\| \rho - \sigma \right\|_{L^1}. \]

Combining these facts provides the desired constants \( L \) and \( M \). Thus there exists a \( T > 0 \) and \( \rho^{(\varepsilon)} \in C^1([0, T], L^1(\mathbb{T}^d)) \) such that \( \rho_t^{(\varepsilon)} \in \mathcal{P}_C \) for all \( t \in [0, T] \), \( \rho^{(\varepsilon)}(0) = \rho_0 \), and \( \partial_t \rho_t^{(\varepsilon)} = A(\rho_t^{(\varepsilon)}) \) for all \( t \in [0, T] \), where the derivative is taken in \( L^1(\mathbb{T}^d) \). The fact that \( \rho^{(\varepsilon)} \in C^1([0, T], (\mathcal{P}_C, \| \cdot \|_{L^1})) \) follows from \( \rho^{(\varepsilon)} \in C^1([0, T], (\mathcal{P}_C, \| \cdot \|_{L^1})) \) and the \( L^\infty \) boundedness of \( \rho_t^{(\varepsilon)} \) and \( A(\rho_t^{(\varepsilon)}) \).

The proof that the solution is in \( C^1(\mathbb{T}^d) \) follows in the standard way. Namely we first observe that, once we know that \( \rho^{(\varepsilon)} \) is a solution in \( \mathcal{P}_C \) then it satisfies \( \partial_t \rho^{(\varepsilon)} = h(x, t) \rho^{(\varepsilon)} \) where \( h \) is continuous and bounded. This implies that \( \rho(t) \in C(\mathbb{T}^d) \) for all \( t \in [0, T] \). To show the \( C^1 \) regularity one needs to show that \( u = \partial_t \rho^{(\varepsilon)} \) satisfies an integral equation. Note that taking a derivative of (BDy) gives that \( u \) satisfies a linear integral equation. Obtaining the existence of the solution and showing that it is the derivative of \( \rho^{(\varepsilon)} \) is straightforward and we do not present the details to conserve space.

The next few lemmas aim to establish \( L^\infty \) upper and lower bounds for \( \rho_t^{(\varepsilon)} \), which allows us to extend the local existence theory to long time intervals.

**Lemma 4.5.** Suppose that \( K \in C^\infty_c(B_1) \) satisfies assumption 2, and that \( \omega = \log \rho \) is \( L \)-Lipschitz continuous, then for any fixed \( x \), we have

\[ e^{-L \varepsilon} \leq \frac{K_\varepsilon \ast \rho(x)}{\rho(x)} \leq e^{L \varepsilon}. \]  

**Proof.**

\[
\frac{K_\varepsilon \ast \rho(x)}{\rho(x)} = \int_{|y-x| \leq \varepsilon} K_\varepsilon(x-y) \exp(w(y) - w(x)) \, dy \
\leq \int_{|y-x| \leq \varepsilon} K_\varepsilon(x-y) e^{L \varepsilon} \, dy = e^{L \varepsilon}.
\]

Similarly, we have \( \frac{K_\varepsilon \ast \rho(x)}{\rho(x)} \geq e^{-L \varepsilon}. \]

**Lemma 4.6.** Let \( \varepsilon > 0 \). Assume \( \rho_t^{(\varepsilon)} \in C^1([0, T], \mathcal{P}_C(\mathbb{T}^d)) \) is a solution of (BDy) with initial condition \( \rho_0 \in C^1(\mathbb{T}^d) \cap \mathcal{P}_C(\mathbb{T}^d) \) for some \( C > 1 \), and let \( w_t^{(\varepsilon)} = \log \rho_t^{(\varepsilon)} \). Then there exists constant \( \overline{C} \) such that for any \( t \in [0, \min\{T, T_\varepsilon\}] \),

\[ \left\| \nabla w_t^{(\varepsilon)} \right\|_{L^\infty(\mathbb{T}^d)} \leq \varepsilon(1 + 2e^{c_0}) \left( \left\| \nabla w_0 \right\|_{L^\infty(\mathbb{T}^d)} + \frac{\| \nabla V \|_{L^\infty(\mathbb{T}^d)}}{1 + 2e^{c_0}} \right) - \left( \| \nabla V \|_{L^\infty(\mathbb{T}^d)} + 1 + 2e^{c_0} \right), \]

where \( T_\varepsilon \) is defined in the proof and satisfies that \( T_\varepsilon \to \infty \) as \( \varepsilon \to 0 \).

**Proof.** We start with the observation that \( w_t^{(\varepsilon)} \) satisfies the equation

\[
\partial_t w_t^{(\varepsilon)} = -\log \frac{K_\varepsilon \ast \rho_t^{(\varepsilon)}}{\pi} - K_\varepsilon \ast \frac{\rho_t^{(\varepsilon)}}{K_\varepsilon \ast \rho_t^{(\varepsilon)}} + \mathcal{F}_\varepsilon \left( \rho_t^{(\varepsilon)} \right) + 1.
\]

5751
Taking spatial derivative, we have
\[
\partial_t \partial_t w_i^{(e)} = -\frac{K_i \partial_i \rho_i^{(e)}}{\rho_i^{(e)}} - \partial_t V - \frac{\rho_i^{(e)}}{K_i \rho_i^{(e)}} \frac{\partial_i \rho_i^{(e)}}{\rho_i^{(e)}} - \partial_t \frac{\rho_i^{(e)} \partial_i \rho_i^{(e)}}{K_i \rho_i^{(e)}} - \partial_t \frac{\rho_i^{(e)} \partial_i \rho_i^{(e)}}{K_i \rho_i^{(e)}}
\]
\[
= -\frac{K_i \rho_i^{(e)} \partial_i \rho_i^{(e)}}{K_i \rho_i^{(e)} - \rho_i^{(e)} - \rho_i^{(e)} K_i \rho_i^{(e)}} \partial_i \rho_i^{(e)} - \partial_t V - \frac{\rho_i^{(e)} \partial_i \rho_i^{(e)}}{K_i \rho_i^{(e)}} \left( \frac{\rho_i^{(e)}}{K_i \rho_i^{(e)}} \right)^2.
\]
\[(46)\]

Now fix an index $i$, and let $L_\varepsilon(t) = \sup_{x \in [\varepsilon]} |\partial_t w_i^{(e)}|$, we have
\[
|\partial_t \partial_t w_i^{(e)}| \leq |\partial_t V| + \left(1 + 2K_\varepsilon \frac{\rho_i^{(e)}}{K_\varepsilon \rho_i^{(e)}}\right) L_\varepsilon(t) \leq \|\nabla V\|_{L^\infty([\varepsilon])} + \left(1 + 2e^{L_\varepsilon(t)}\right) L_\varepsilon(t).
\]

After taking supremum on the left side above, this turns into
\[
L'_\varepsilon(t) \leq \|\nabla V\|_{L^\infty([\varepsilon])} + \left(1 + 2e^{L_\varepsilon(t)}\right) L_\varepsilon(t).
\]
\[(47)\]

Let $\tilde{L}_\varepsilon = \varepsilon L_\varepsilon$, then
\[
\frac{d}{dt} \tilde{L}_\varepsilon(t) \leq \varepsilon \|\nabla V\|_{L^\infty([\varepsilon])} + \tilde{L}_\varepsilon(t) \left(1 + 2e^{L_\varepsilon(t)}\right).
\]

Notice that for every $\varepsilon > 0$, the solution $z_\varepsilon$ to the ODE problem
\[
\frac{d}{dt} z_\varepsilon(t) = \varepsilon \|\nabla V\|_{L^\infty([\varepsilon])} + z_\varepsilon(t) \left(1 + 2e^{L_\varepsilon(t)}\right), \quad z_\varepsilon(0) = \varepsilon L(0)
\]
blows up at the finite time
\[
\tau_\varepsilon := \int_{z_\varepsilon(0)}^\infty \frac{1}{\varepsilon \|\nabla V\|_{L^\infty([\varepsilon])} + z(1 + 2e^{L_\varepsilon(t)})} \, dz < \infty.
\]

However, as $\varepsilon \to 0$, one has that $z_\varepsilon(0) \to 0$ and $\varepsilon \|\nabla V\|_{L^\infty([\varepsilon])} \to 0$. Therefore,
\[
\tau_\varepsilon \to \int_0^\infty \frac{1}{z(1 + 2e^{L_\varepsilon(t)})} \, dz = \infty.
\]

Notice that $z_\varepsilon(t)$ is an increasing function of both $t$ and $\varepsilon$. Let $T_\varepsilon = \frac{1}{\varepsilon} \tau_\varepsilon$. Let $\tilde{\varepsilon} = e(t \leq 1 \tau_\varepsilon)$. Then $L_\varepsilon(t) \leq \tilde{\varepsilon}$ for all $t \leq \min\{T, T_\varepsilon\}$. Taking account of above in the differential inequality (47), one obtains from Gronwall’s inequality that for all $t \in [0, \min\{T, T_\varepsilon\}]$,
\[
L(t) \leq e^{\left(1 + 2e^{\tilde{\varepsilon}}\right)} \left(L(0) + \frac{\|\nabla V\|_{L^\infty([\varepsilon])}}{1 + 2e^{\tilde{\varepsilon}}}\right) = \frac{\|\nabla V\|_{L^\infty([\varepsilon])}}{1 + 2e^{\tilde{\varepsilon}}},
\]
which is precisely (45).
Lemma 4.7. Under the same conditions as lemma 4.6, suppose that for all \( t \in [0, T] \), \( w_i^{(\epsilon)} = \log \rho_i^{(\epsilon)} \) is \( L \)-Lipschitz continuous. Then

\[
\sup_{\pi_t} \frac{\rho_i^{(\epsilon)}}{\pi_t} \leq \exp \left( (1 - e^{-\tau}) (L \epsilon + KL (\rho_0 | \pi) + 1) + e^{-\tau} \log \frac{\rho_0}{\pi} \right)
\]

(48)

and

\[
\inf_{\pi_t} \frac{\rho_i^{(\epsilon)}}{\pi_t} \geq \exp \left( - (1 - e^{-\tau}) (L \epsilon + e^L \epsilon + C_\pi \epsilon^2) + e^{-\tau} \log \frac{\rho_0}{\pi} \right)
\]

(49)

Here \( C_\pi \) is the constant depending on \( \pi \) that appears in theorem 4.2, such that \( F_\pi (\pi_\epsilon) \geq -C_\pi \epsilon^2 \).

To applying this lemma one can use the Lipschitz estimate of lemma 4.6 and take \( L \) to be the right hand side of (45) for \( t = T_\epsilon \).

Proof. Notice that

\[
\partial_t \log \frac{\rho_i^{(\epsilon)}}{\pi} = - \log \frac{K \ast \rho_i^{(\epsilon)}}{\pi} - K \ast \frac{\rho_i^{(\epsilon)}}{K \ast \rho_i^{(\epsilon)}} + F_\pi \left( \rho_i^{(\epsilon)} \right) + 1
\]

(44), (37)

\[
\leq - \log \frac{\rho_i^{(\epsilon)}}{\pi} + L \epsilon + KL (\rho_0 | \pi) + 1.
\]

Using Gronwall’s inequality, we have

\[
\log \frac{\rho_i^{(\epsilon)}}{\pi} \leq (1 - e^{-\tau}) (L \epsilon + KL (\rho_0 | \pi) + 1) + e^{-\tau} \log \frac{\rho_0}{\pi}
\]

The proof of lower bound is similar. Since

\[
\partial_t \log \frac{\rho_i^{(\epsilon)}}{\pi} \geq - \log \frac{\rho_i^{(\epsilon)}}{\pi} - L \epsilon - e^L \epsilon - C_\pi \epsilon^2,
\]

we obtain

\[
\log \frac{\rho_i^{(\epsilon)}}{\pi} \geq e^{-\tau} \log \frac{\rho_0}{\pi} - (1 - e^{-\tau}) (L \epsilon + e^L \epsilon + C_\pi \epsilon^2).
\]
In order to study the convergence of \((\text{BD}_\varepsilon)\) to \((\text{BD})\) as \(\varepsilon \to 0\), we will rely on their gradient flow structure, which we investigate next.

**Definition 4.9.** Let \(\mathcal{G}\) be a functional defined on \(\mathcal{P}_C\). We say \(\mathcal{G}(\rho)\) is \(\lambda\)-geodesically convex in \(\mathcal{P}_C\) with respect to \(d_{\text{SH}}\) if, for any \(\rho_0, \rho_1 \in \mathcal{P}_C\), let \((\tilde{\rho}_t)_{t \in [0,1]}\) be the \(d_{\text{SH}}\)-geodesics connecting \(\rho_0\) to \(\rho_1\), then

\[
\mathcal{G}(\rho_1) - \mathcal{G}(\rho_0) - \frac{d}{dt} \mathcal{G}(\tilde{\rho}_t) \bigg|_{t=0} \geq \frac{\lambda}{2} d_{\text{SH}}^2(\rho_0, \rho_1). \tag{50}
\]

Note that the above definition does not require \(\mathcal{P}_C\) itself to be a geodesically convex set of \(d_{\text{SH}}\). As long as \(\rho\) is bounded above and below away from 0, both the relative entropy \(F\) and regularised entropy \(\rho\) of \(d\) are geodesically semiconvex with respect to \(d_{\text{SH}}\), which is why we introduced the restricted submanifold \(\mathcal{P}_C\). For general results regarding displacement convexity with respect to Hellinger–Kantorovich distance, we refer the readers to [49].

**Lemma 4.10.** Let \(\varepsilon > 0\) and \(K_{\varepsilon}\) satisfy assumption 2. For any \(C > 1\), if \(\pi \in \mathcal{P}_C\), then both \(F(\rho) = \int_{\mathcal{X}} \rho \log \frac{\rho}{\pi}\) and \(F_\varepsilon(\rho)\) are \(\lambda\)-geodesically convex with respect to \(d_{\text{SH}}\) in \(\mathcal{P}_C\), for some \(\lambda = \lambda(C, \varepsilon) \in \mathbb{R}\).

**Proof.** Consider \(\rho_0, \rho_1 \in \mathcal{P}_C\) and \(\rho_0 \neq \rho_1\). Let us recall the \(d_{\text{SH}}\)-geodesics from \(\rho_0\) to \(\rho_1\) given in the expression in [45, lemma 2.7]:

\[
\rho_t = \frac{\tilde{\rho}_t^\beta}{r_{\beta_t}}, \text{ with } \beta_t = \frac{\sin \left( t d_{\text{SH}}(\rho_0, \rho_1) / 2 \right)}{\sin \left( t d_{\text{SH}}(\rho_0, \rho_1) / 2 \right) + \sin \left( (1 - t) d_{\text{SH}}(\rho_0, \rho_1) / 2 \right)} \text{ and } r_t = \int_{\mathcal{X}} \tilde{\rho}_t, \tag{51}
\]

and \(\tilde{\rho}_t\) is the \(d_{\text{SH}}\)-geodesics given in explicit form (9). Since \(d_{\text{SH}}(\rho_0, \rho_1) \leq 2\sqrt{2}\), we can obtain from (10) that \(d_{\text{SH}}(\rho_0, \rho_1) \leq \pi\), hence \(\beta_t\) is well-defined and increases from 0 to 1. We may also obtain

\[
r_t = \int_{\mathcal{X}} \left( t \sqrt{\rho_t} + (1 - t) \sqrt{\rho_0} \right)^2 = 1 - \frac{(1 - t)}{4} d_{\text{SH}}^2(\rho_0, \rho_1) \in \left[ \frac{1}{2}, 1 \right], \tag{52}
\]

which indicates \(\rho_t \in \mathcal{P}_{2C}\) for all \(t \in [0,1]\). Thus, one can explicitly calculate that for \(F(\rho) = \text{KL}(\rho|\pi)\),

\[
\frac{d^2}{dt^2} F(\rho_t) = \frac{d^2}{dt^2} \int_{\mathcal{X}^t} \rho_t \log \frac{\rho_t}{\pi} = \frac{d}{dt} \int_{\mathcal{X}^t} \partial_t \rho_t \left( \log \frac{\rho_t}{\pi} + 1 \right) = \int_{\mathcal{X}^t} \left( \log \frac{\rho_t}{\pi} \frac{d^2}{dt^2} \rho_t + \frac{1}{\rho_t} \left( \frac{d}{dt} \rho_t \right)^2 \right).
\]

The contribution from the second term is already non-negative; therefore in view of the point-wise bounds of \(\rho_t\) and \(\pi\), we only need to prove

\[
\int_{\mathcal{X}^t} \left| \frac{d^2}{dt^2} \rho_t \right| \leq d_{\text{SH}}^2(\rho_0, \rho_1). \tag{53}
\]

Taking second derivative on (51), using chain rule and quotient rule, we obtain

\[
\frac{d^2}{dt^2} \rho_t = \left( \frac{\tilde{\rho}_t''}{r_t} - \frac{2 \tilde{\rho}_t' r_t'}{r_t^2} + \frac{\tilde{\rho}_t'' r_t'}{r_t^3} \right) \bigg|_{s=\beta_t} \left( \beta_t' \right)^2 + \left( \frac{\tilde{\rho}_t'}{r_t} - \frac{\tilde{\rho}_t'' r_t'}{r_t^3} \right) \bigg|_{s=\beta_t} \beta_t''.
\]
We complete the proof of (53) by combining the following facts:

\[
\int \hat{\rho}' = r_s \in \left[ \frac{1}{2}, 1 \right],
\]

\[
\int |\hat{\rho}'| = |r_s' | = 2 \int (\sqrt{\rho_0} - \sqrt{\rho_0})^2 \leq \frac{1}{2} d_{SH}^2 (\rho_0, \rho_1),
\]

\[
|\hat{\rho}'| = 2 \int \sqrt{\rho_0} (\sqrt{\rho_0} - \sqrt{\rho_0}) \leq 2 \left( \int \rho_0 \right)^{\frac{1}{2}} \left( \int (\sqrt{\rho_0} - \sqrt{\rho_0})^2 \right)^{\frac{1}{2}} \leq \sqrt{\rho_0} d_{SH} (\rho_0, \rho_1),
\]

\[
|r_s' | = \left| \frac{2 \sqrt{1 - \rho_0}}{4} \right| d_{SH} (\rho_0, \rho_1) \leq \frac{1}{4} d_{SH} (\rho_0, \rho_1).
\]

\[
\beta' = \frac{d_{SH} (\rho_0, \rho_1) \sin (d_{SH} (\rho_0, \rho_1) / 2)}{2 (\sin (td_{SH} (\rho_0, \rho_1) / 2) + \sin ((1 - t) d_{SH} (\rho_0, \rho_1) / 2))},
\]

\[
|\beta'| = \frac{d_{SH}^2 (\rho_0, \rho_1) \sin (d_{SH} (\rho_0, \rho_1) / 2)}{2 \sin (td_{SH} (\rho_0, \rho_1) / 2) + \sin ((1 - t) d_{SH} (\rho_0, \rho_1) / 2)} \left( \cos (td_{SH} (\rho_0, \rho_1) / 2) - \cos ((1 - t) d_{SH} (\rho_0, \rho_1) / 2) \right) \leq C d_{SH} (\rho_0, \rho_1).
\]

We now turn our attention to the regularised energy \( F_{\varepsilon} \). Taking second derivative, following the same calculation as above, one has

\[
\frac{d^2}{dr^2} \int_{\mathbb{T}^2} F_{\varepsilon} (\rho_0) = \int_{\mathbb{T}^2} \log \frac{K_{x} \ast \rho_0}{\pi} \frac{d^2}{dr^2} \rho_0 + \int_{\mathbb{T}^2} \frac{\rho_0}{K_{x} \ast \rho_0} K_{x} \ast \frac{d^2}{dr^2} \rho_0 + 2 \int_{\mathbb{T}^2} \frac{d\rho_0}{K_{x} \ast \rho_0} \frac{d\rho_0}{dr} - C \int_{\mathbb{T}^2} \frac{d\rho_0}{K_{x} \ast \rho_0} \frac{d\rho_0}{dr} + C \int_{\mathbb{T}^2} \frac{d\rho_0}{K_{x} \ast \rho_0} \frac{d\rho_0}{dr}.
\]

Here the second term on the right side of the second line can be bounded using assumption \( 2 \) on \( K_{x} \) and Jensen’s inequality:

\[
\int_{\mathbb{T}^2} \frac{d\rho_0}{K_{x} \ast \rho_0} \frac{d\rho_0}{dr} = \int_{\mathbb{T}^2} \left( \xi_0 \ast \frac{d\rho_0}{dr} \right)^2 \leq \int_{\mathbb{T}^2} \xi_0 \ast \left( \frac{d\rho_0}{dr} \right)^2 = \int_{\mathbb{T}^2} \left( \frac{d\rho_0}{dr} \right)^2.
\]

The third term on the right side of the second line can be treated in the identical way. In view of (53), it remains to show

\[
\int_{\mathbb{T}^2} \left( \frac{d\rho_0}{dr} \right)^2 \leq d_{SH}^2 (\rho_0, \rho_1).
\]

which is true since \( \rho_0, \rho_1 \in \mathcal{P}_{C} \) and we may apply the bounds in (54) to

\[
\frac{d}{dr} \rho_0 = \left( \frac{\beta'}{r_s} - \frac{\beta'}{r_s^2} \right) \bigg|_{r_s = \beta'}.
\]

\[
\square
\]

5755
We now introduce the elements needed to consider (BD$_c$) as $d_{SH}$ gradient flows. The space of all probability densities on $T^d$ equipped with $d_{SH}$ is a complete Riemannian manifold with corners. The explicit formulas for geodesics (51) ensure that for all $\rho_0, \rho_1 \in \mathcal{P}_C$ and all $t \in [0, 1]$ the geodesics $\rho_t \in \mathcal{P}_C$. The tangent spaces are already identified in the earlier work [31]: for $\rho \in \mathcal{P}_C$, the tangent space at $\rho$ with respect to $d_{SH}$ can be identified with

$$T_{SH, \rho} = L^2_0(\rho) := \left\{ \phi \in L^2(\rho) : \int \phi \, d\rho = 0 \right\}. \quad (56)$$

We now compute the geometric logarithm (inverse of the exponential map) on the space of probability measures endowed with $d_{SH}$ geometry and a point $\rho \in \mathcal{P}_C$ for some $C$. In other words we compute the tangent vector to the unit-time geodesic connecting two measures.

**Lemma 4.11.** For $\rho, \mu \in \mathcal{P}_C$, we have

$$\ln^{d_{SH}}_{\rho} \mu = \frac{d_{SH}(\rho, \mu) / 2}{\sin (d_{SH}(\rho, \mu) / 2)} \left[ 2 \left( \frac{\mu}{\rho} - 1 \right) + \frac{d_{SH}(\rho, \mu)^2}{4} \right]. \quad (57)$$

**Proof.** Let $\{\rho_t\} \in [0, 1]$ be the $d_{SH}$-geodesics connecting from $\rho$ to $\mu$. We refer the readers to the expression in (51), then apply the chain rule and we obtain

$$\ln^{d_{SH}}_{\rho} \mu = \frac{d}{d\rho} \ln r |_{t=0} = \frac{\ddot{\rho} \dot{\rho} - \ddot{\rho} \dot{\rho}}{\rho^2} \beta_0' \left( \frac{d_{SH}(\rho, \mu)^2}{\sin (d_{SH}(\rho, \mu) / 2)} \left[ 2 \left( \frac{\mu}{\rho} - 1 \right) + \frac{d_{SH}(\rho, \mu)^2}{4} \right] \right). \quad (58)$$

We note that due to the remark after (10), $\sin (d_{SH}(\rho, \mu) / 2) > 0$.

Note that the tangent spaces are Hilbert spaces, which allows us to define the subdifferentials as the classical Fréchet subdifferentials:

**Definition 4.12.** (Subdifferential) Let $G : \mathcal{P}(T^d) \to \mathbb{R}$ be proper and lower semicontinuous with respect to $d_{SH}$. Assume $G$ be proper and lower semicontinuous with respect to $d_{SH}$. Assume $G$ be proper and lower semicontinuous with respect to $d_{SH}$. Assume $\rho > 0$ on $T^d$, $G(\rho) < \infty$, and $\zeta \in L^2_0(\rho)$. We say that $\zeta$ is in the subdifferential of $G$ at $\rho$, and write $\zeta \in \partial_{SH} G(\rho)$, if for any $\phi \in L^2_0(\rho)$ such that $(1 + \phi) \rho \geq 0,

$$G((1 + \phi) \rho) - G(\rho) \geq \int_{T^d} \zeta \phi \, dx + o(\|\phi\|_{L^2(\rho)}). \quad (59)$$

We note that $\phi$ belongs to the Fréchet differential if the inequality above is replaced by equality.

We note that $o(\|\phi\|_{L^2(\rho)}) = o(d_{SH}(\rho, (1 + \phi) \rho))$.

**Lemma 4.13.** Consider $\rho \in \mathcal{P}_C$ for some $C > 1$. Suppose $G : \mathcal{P} \to \mathbb{R}$ be proper and lower semicontinuous. Assume that there exists $\zeta \in L^\infty(T^d)$ such that for all $h \in L^2(\rho)$ which is essentially bounded from below (i.e. $h \in L^\infty$), the first variation of $G$ in direction $h$ exists and is of the form $\frac{d}{d\rho} \left[ \frac{d_{SH}}{\rho} \right] [h] = \int \zeta h \, dx$. Furthermore assume that $G$ is $d_{SH}$ $\lambda$-geodesically convex in $\mathcal{P}_C$ for some $\lambda \in \mathbb{R}$. Then

$$\partial_{SH} G(\rho) = \left\{ \zeta - \int_{T^d} \zeta \rho \right\}. \quad (60)$$
Proof. Let us first show that any element of \( \partial_{\text{SH}} G(\rho) \) must be equal to \( \zeta - \int_{\mathcal{T}} \zeta \rho \). Assume that \( \xi \in \partial_{\text{SH}} G(\rho) \), namely it satisfies (59). For \( \phi \in L^0_\mu(\rho) \cap L^\infty(\mathbb{T}^d) \) and \( t > 0 \) consider \( t\phi \) playing the role of \( \phi \) in (59), dividing by \( t \) and taking the limit as \( t \to 0 \) provides

\[
\int \zeta \phi \rho \geq \int \xi \phi \rho.
\]

Doing the same for \( t < 0 \) yields

\[
\int \zeta \phi \rho \leq \int \xi \phi \rho.
\]

Hence \( \int \zeta \phi \rho = \int \xi \phi \rho \) for all \( \phi \in L^0_\mu(\rho) \cap L^\infty(\mathbb{T}^d) \). Thus \( \xi = \zeta - \int_{\mathcal{T}} \zeta \rho \) since \( \xi \in L^0_\mu(\rho) \).

Let us now show that \( \zeta - \int_{\mathcal{T}} \zeta \rho \in \partial_{\text{SH}} G(\rho) \). Let \( \{\rho_t\}_{t \in [0,1]} \) be the \( d_{\text{SH}} \) geodesics connecting \( \rho \) and \( (1 + \phi)\rho \). Note that, by (57), \( \frac{d}{dt} \rho_t(0) \) is essentially bounded from below. Using that \( \zeta \in L^\infty \) we obtain

\[
\frac{d}{dt} \bigg|_{t=0} G(\rho) = \int \zeta \ln \frac{(1 + \phi)\rho}{\rho} \rho \, d\xi
\]

\[
= \frac{d_{\text{SH}}(\rho, (1 + \phi)\rho) / 2}{\ln(d_{\text{SH}}(\rho, (1 + \phi)\rho))} \int \frac{(1 + \phi)\rho}{\rho} \, d\xi + o(d_{\text{SH}}(\rho, (1 + \phi)\rho))
\]

\[
= \int \zeta \phi \rho + o(d_{\text{SH}}(\rho, (1 + \phi)\rho)) = \int \left( \zeta - \int_{\mathcal{T}} \zeta \rho \right) \phi \rho + o(\|\phi\|_{L^0(\rho)})
\]

Combining this with the \( \lambda \) convexity of \( G \) implies that \( \zeta - \int_{\mathcal{T}} \zeta \rho \in \partial_{\text{SH}} G(\rho) \).

The above lemma allows us to identify the subgradients of \( F \) and \( F_\varepsilon \). Since the subgradients are singletons we identify each set with its only element. That is, for \( \rho \in \mathcal{P}_C \),

\[
\partial_{\text{SH}} F(\rho) = \log \frac{\rho}{\pi} - \int_{\mathbb{R}^d} \log \frac{\rho}{\pi} \rho
\]

\[
\partial_{\text{SH}} F_\varepsilon(\rho) = \log \left( \frac{K_\varepsilon * \rho}{\pi} \right) + K_\varepsilon * \left( \frac{\rho}{K_\varepsilon * \rho} \right) - \int \log \left( \frac{K_\varepsilon * \rho}{\pi} \right) \rho - 1.
\]

The last step to rigorously identify (BD) and (BD\(_x\)) as \( d_{\text{SH}} \)-gradient flows is to show that they are curves of maximal slope. Let us recall that for a curve \( \rho_t \), the metric derivative is given by

\[
|\partial_t \rho_t| = \lim_{s \to t} \frac{d_{\text{SH}}(\rho_s, \rho_t)}{|s - t|},
\]

and the metric slope is defined as [2, (10.0.9)]

\[
|\partial_{\text{SH}} G(\rho)| = \limsup_{d_{\text{SH}}(\rho, \nu) \to 0} \frac{(G(\rho) - G(\nu))_+}{d_{\text{SH}}(\nu, \rho)}.
\]

We say that a path \( \rho \in AC([0, T], (\mathcal{P}_C, d_{\text{SH}})) \) is a gradient flow solution of (14) if \( \partial_t \rho_t = -\rho_t \partial_{\text{SH}} G(\rho_t) \) for a.e. \( t \in [0, T] \). We say that \( \rho_t \in AC([0, T], (\mathcal{P}_C, d_{\text{SH}})) \) is a curve of maximal slope for functional \( G \) if \( \frac{d}{dt} G(\rho) \leq -\frac{1}{2} |\partial_t \rho|^2 - \frac{1}{2} |\partial_{\text{SH}} G(\rho)|^2 \) where \( |\partial_t \rho| \) is the \( d_{\text{SH}} \) metric derivative and \( |\partial_{\text{SH}} G(\rho)| \) is the metric slope. Note that in definition 4.12 we already require \( \partial_{\text{SH}} G(\rho) \in L^\infty(\mathbb{T}^d) \), and \( \rho_t \in \mathcal{P}_C \subset L^\infty(\mathbb{T}^d) \), hence \( \partial_t \rho_t \) is well-defined in the classical sense.
Lemma 4.14. Consider $\rho \in \mathcal{P}_C$ for some $C > 1$ and suppose that the assumptions of lemma 4.13 are satisfied. Assume furthermore that $\partial_{\text{SH}} \mathcal{G}(\mu)$ is continuous in $L^\infty(\mathbb{T}^d)$ in an $L^\infty$-neighbourhood of $\rho$. Then the metric slope satisfies

$$|\partial_{\text{SH}} \mathcal{G}(\rho)| = \|\partial_{\text{SH}} \mathcal{G}(\rho)\|_{L^2(\rho)}.$$  

**Proof.** Let $\zeta = \partial_{\text{SH}} \mathcal{G}(\rho)$. Let $\mu_t = (1 - t \zeta) \rho$. The fact that $|\partial_{\text{SH}} \mathcal{G}(\rho)| \leq \|\partial_{\text{SH}} \mathcal{G}(\rho)\|_{L^2(\rho)}$ follows by considering the limit along $\mu_t \to \rho$. To show the opposite inequality note that

$$\mathcal{G}(\rho) - \mathcal{G}(\mu_t) \geq \int \zeta (\rho - \mu_t) + o(d_{\text{SH}}(\rho, \mu_t)) = t \int \zeta \zeta + o(d_{\text{SH}}(\rho, \mu_t))$$

where $\zeta = \partial_{\text{SH}} \mathcal{G}(\mu_t)$. Noting that $d_{\text{SH}}(\rho, \mu_t) = t\|\zeta\|_{L^2(\rho)} + o(d_{\text{SH}}(\rho, \mu_t))$ and using the continuity of $\zeta_t$ implies that

$$\limsup_{t \to 0^+} \frac{\mathcal{G}(\rho) - \mathcal{G}(\mu_t)}{d_{\text{SH}}(\rho, \mu_t)} \geq \|\zeta\|_{L^2(\rho)}.$$  

\[\Box\]

With the above preparation, we are able to rigorously identify (BD$_\epsilon$) as the $d_{\text{SH}}$ gradient flow of $\mathcal{F}_\epsilon$. The existence results of lemma 2.2 and theorem 4.8 imply that during the interval of existence $\partial_{\text{BD}_\epsilon}(\rho) = -\partial_{\text{BD}_\epsilon} \mathcal{F}_\epsilon(\rho(\epsilon))$. Furthermore the solutions are in $AC([0, T], (\mathcal{P}_C, d_{\text{SH}}))$, which can be verified by direct check based on solution formula of lemma 2.2 for $\rho_0$, and by $\rho(\epsilon) \in C([0, T], L^2(\mathbb{T}^d))$ using theorem 4.8. Thus $\rho$ and $\rho(\epsilon)$ are gradient flows of the respective equations, and consequently curves of maximal slope.

Due to the semiconvexity of the functionals, the solutions also satisfy an evolution variational inequality (chapter 11 of [2] for Wasserstein gradient flows, and [60] for general metric spaces). This implies that gradient flow solutions are unique. In particular while our notion of convexity is restricted, the proof of quantitative stability of theorem 11.1.4 of [2] carries over. Thus gradient flow solutions coincide with the solutions of lemma 2.2 and theorem 4.8, respectively.

4.3. $\Gamma$-convergence of gradient flows

The next question is whether the dynamics (BD$_\epsilon$) converges in any sense to the idealised dynamics (BD) as $\epsilon \to 0$. The natural notion to study is the $\Gamma$-convergence of gradient flows à la Sandier-Serfaty [70, 71]. We will show a proof for $\mathbb{T}^d$ with a compactly supported kernel. The following theorem, essentially a rephrase of [71, theorem 2] but adapted to our setting, summarises the conditions we need to verify for the $\Gamma$-convergence of gradient flow.

**Theorem 4.15 [71, theorem 2].** Let $\rho(\epsilon)$ be solutions of (BD$_\epsilon$) that are curves of maximal slopes. Suppose the initial conditions are well-prepared, in the sense that as $\epsilon \to 0$,

$$\rho_0(\epsilon) \xrightarrow{d_{\text{SH}}} \rho_0, \text{ and } \mathcal{F}_\epsilon(\rho_0(\epsilon)) \to \mathcal{F}(\rho_0).$$

Suppose also that as $\epsilon \to 0$, we have $\rho(\epsilon) \xrightarrow{d_{\text{SH}}} \nu_t$ for almost every $t$, as well as the following conditions:

(i) $\liminf_{\epsilon \to 0} \int_0^t |\partial_{\text{SH}} \rho(\epsilon)|^2 \text{d}s \geq \int_0^t |\partial_\nu \nu_s|^2 \text{d}s,$
liminf \limits_{\varepsilon \to 0} \int_0^T |\partial_{\SH} F_\varepsilon(\rho^{(\varepsilon)}_t)|^2 \, dt \geq \int_0^T |\partial_{\SH} F(\nu_t)|^2 \, dt,

then \nu_t must be the solution of gradient flow (BD) with energy functional \( F \) and initial condition \( \nu_0 = \rho_0 \).

**Theorem 4.16.** Under the same assumptions as theorem 4.8, for any fixed \( T, (\rho^{(\varepsilon)}_t)_{0 \leq t \leq T} \) converges in \( d_{\SH} \) to \( (\rho_t)_{0 \leq t \leq T} \) the solution of (BD) with the same initial condition \( \rho^{(\varepsilon)}_0 = \rho_0 \). In particular, \( \lim_{\varepsilon \to 0} F_\varepsilon(\rho^{(\varepsilon)}_t) = F(\rho_t) \).

**Proof.** The plan of our proof is to first use Arzela–Ascoli theorem to identify the limiting sequence \( \nu_t \), then verify the three conditions in theorem 4.15. Notice \( \nabla \rho^{(\varepsilon)}_t = \rho^{(\varepsilon)}_t \nabla w^{(\varepsilon)}_t \), hence the combination of lemmas 4.6 and 4.7, as well as \( \pi \in \mathcal{P}_C \), yield that \( \nabla \rho^{(\varepsilon)}_t \) is uniformly bounded for any \( t \in [0, T] \), when \( \varepsilon \) is sufficiently small. Thus, since \( \rho^{(\varepsilon)}_t \) is also uniformly bounded (see lemma 4.7), we may invoke Arzela–Ascoli Theorem, so that there exists a subsequence, still denoted as \( \rho^{(\varepsilon)}_t \), that converges uniformly to some function \( \nu_t \) as \( \varepsilon \to 0 \).

We now verify the three conditions of theorem 4.15. The proof of (ii) is straightforward, since by \( \Gamma \)-convergence of the energy functional [16, theorem 4.1] (note that convergence in \( d_{\SH} \) implies convergence in \( W_2 \) on \( \mathbb{T}^d \)), \( \liminf_{\varepsilon \to 0} F_\varepsilon(\rho^{(\varepsilon)}_t) \geq F(\nu_t) \), and trivially \( \rho^{(\varepsilon)}_t \in \mathcal{P}_\Lambda(\mathbb{T}^d) \) on bounded domain.

We now prove (i). The proof is standard and follows from the arguments in [22, theorem 5.6]. Assume without loss of generality that there exists \( 0 \leq C < \infty \) so that

\[
C = \liminf_{\varepsilon \to 0} \int_0^T |\partial_{\SH} \rho^{(\varepsilon)}_t|^2 \, dt.
\]

Choose a subsequence \( |\partial_{\SH} \rho^{(\varepsilon)}_t| \) so that \( \lim_{\varepsilon \to 0} \int_0^T |\partial_{\SH} \rho^{(\varepsilon)}_t|^2 \, dt = C \). Then \( |\partial_{\SH} \rho^{(\varepsilon)}_t| \) is bounded in \( L^2(0, T) \) so, up to a further subsequence, it is weakly convergent to some \( v(t) \in L^2(0, T) \). Consequently, for any \( 0 \leq s_0 \leq s_1 \leq T \),

\[
\lim_{\varepsilon \to 0} \int_{s_0}^{s_1} |\partial_{\SH} \rho^{(\varepsilon)}_t| \, dt = \int_{s_0}^{s_1} v(t) \, dt.
\]

By taking limits in the definition of the metric derivative and using the lower semi-continuity of \( d_{\SH} \) with respect to weak-* convergence (see the proof in [39, theorem 5] for \( d_{\HI} \), which also applies to \( d_{\SH} \) using conic structure (10)),

\[
d_{\SH} \left( \tilde{\rho}^{(s)}_{s_0}, \tilde{\rho}^{(s)}_{s_1} \right) \leq \int_{s_0}^{s_1} |\partial_{\SH} \tilde{\rho}^{(s)}_t| \, dt \Rightarrow d_{\SH}(\nu_{s_0}, \nu_{s_1}) \leq \int_{s_0}^{s_1} v(t) \, dt.
\]

By [2, theorem 1.1.2], this implies that \( |\partial_{\nu_t}| \leq V(t) \) for a.e. \( t \in (0, T) \). Thus, by the lower semicontinuity of the \( L^2(0, T) \) norm with respect to weak convergence,

\[
\liminf_{\varepsilon \to 0} \int_0^T |\partial_{\SH} \rho^{(\varepsilon)}_t|^2 \, dt = \lim_{\varepsilon \to 0} \int_0^T |\partial_{\SH} \rho^{(\varepsilon)}_t|^2 \, dt \geq \int_0^T v(t)^2 \, dt \geq \int_0^T |\partial_{\nu_t}|^2 \, dt.
\]
and we have that for \( \varepsilon = 0 \) and \( \varepsilon \to 0 \), \( K_x \ast \rho_t^{(e)}(x) \) is uniformly bounded in \( z \) and \( \varepsilon \), which implies that \( \rho_t^{(e)}(x) \) converges uniformly to \( \nu_t \) as \( \varepsilon \to 0 \). Indeed,

\[
|K_x \ast \rho_t^{(e)}(x) - \nu_t(x)| = \left| \int_{\mathbb{T}^d} K_x(x - y) \left( \rho_t^{(e)}(y) - \nu_t(x) \right) \, dy \right|
\]

\[
\leq \left| \int_{\mathbb{T}^d} K_x(x - y) \left( \rho_t^{(e)}(y) - \rho_t^{(e)}(x) \right) \, dy \right|
\]

\[
+ \left| \int_{\mathbb{T}^d} K_x(x - y) \left( \rho_t^{(e)}(x) - \nu_t(x) \right) \, dy \right|
\]

\[
\leq \sup_{z \in \mathbb{T}^d} |\nabla \rho_t^{(e)}(z)| \int_{\mathbb{T}^d} K_x(x - y) |x - y| \, dy + \sup_{z \in \mathbb{T}^d} |\rho_t^{(e)}(z) - \nu_t(z)|.
\]

The first term goes to zero since \( |\nabla \rho_t^{(e)}(z)| \) is uniformly bounded in \( z \) and \( \varepsilon \), while the integral is \( \varepsilon M_1(K) \to 0 \), while the second term also goes to zero due to uniform convergence of \( \rho_t^{(e)}(x) \) to \( \nu_t \). Moreover, by (49), \( \nu_t(x) \) is bounded and away from zero for all \( x \in \mathbb{T}^d \) and \( t \in [0,T] \), which implies that \( \frac{\nu_t(x)}{K_x \ast \rho_t^{(e)}(x)} \) and consequently \( K_x \ast \frac{\rho_t^{(e)}}{K_x \ast \rho_t^{(e)}} \) converge uniformly to 1 as \( \varepsilon \to 0 \). Consequently, the inequality

\[
\liminf_{\varepsilon \to 0} \int_0^T \left( \log \frac{K_x \ast \rho_t^{(e)}}{\pi} + K_x \ast \left( \frac{\rho_t^{(e)}}{K_x \ast \rho_t^{(e)}} \right) - \int \rho_t^{(e)} \log \frac{K_x \ast \rho_t^{(e)}}{\pi} - 1 \right)^2
\]

\[
\geq \int_0^T \left( \log \frac{\nu_t}{\pi} - \int \nu_t \log \frac{\nu_t}{\pi} \right)^2
\]

holds by Fatou’s lemma. Now, by [2, corollary 2.4.10], since \( \mathcal{F} \) and \( \mathcal{F}_e \) are geodesically semi-convex on \( P_C, |\partial \text{sh}_{\mathcal{F}}(\rho)| \) and \( |\partial \text{sh}_{\mathcal{F}_e}(\rho^{(e)})| \) are strong upper gradients. By lemmas 4.13 and 4.14 we have that for \( \mathcal{G} = \mathcal{F} \) or \( \mathcal{F}_e \),

\[
|\partial \text{sh}_{\mathcal{G}}(\rho)| = \left( \int_{\mathbb{T}^d} \rho \frac{\delta G}{\delta \rho} - \int_{\mathbb{T}^d} \rho \frac{\delta \bar{G}}{\delta \rho} \right)^2
\]

Hence, in view of (62), we can verify condition (iii) of theorem 4.15. This allows us now to conclude: since all three conditions of theorem 4.15 are now fulfilled, by \( \Gamma \)-convergence, \( \nu_t \) must be a solution of (BD) with initial condition \( \rho_0 \), and therefore by the uniqueness result established in lemma 2.2, must be \( \rho_t \).

\[ \square \]

**Remark 4.17.** We would like to comment here that the above strategy does not apply to the whole space \( \mathbb{R}^d \), since we could not assume that \( \nabla V \in L^\infty(\mathbb{R}^d) \) or \( u_i^{(e)} \) being globally Lipschitz continuous on \( \mathbb{R}^d \). Moreover, the ratio \( \frac{\delta G}{\delta \rho} \) may be very close to 0 at infinity, unlike lemma 4.5 which says the ratio is always \( 1 \pm O(\varepsilon) \), making it difficult to compare (BD) with (BD). Finally \( \mathcal{F}_e \) might not be displacement semiconvex when \( \rho \) is close to zero, which is unavoidable in the whole space.

**Remark 4.18.** If the dynamics (BD) has an initial condition \( \rho_0 = \sum_{j=1}^N m_j(0) \delta_{x_j} \) for some \( x_j \in \mathbb{R}^d, i = 1, \ldots, N, m_i(0) > 0, \sum_{j=1}^N m_i(0) = 1 \), then (BD) has a solution of the form \( \rho_t = \sum_{j=1}^N m_j(t) \delta_{x_j} \) where the masses \( m_i(t) \) satisfy the following ODE:
The above ODE is obviously well-posed, since it is a finite-dimensional ODE with a trapping region, which is the probability simplex. On the other hand, although we are unable to prove it, we believe the long-time well-posedness of (BD) with a smooth initial condition is true with more careful estimates. More specifically, we believe the solution can be approximated by certain minimising movement scheme, at least on a compact domain as shown [46] in a different setting. We leave careful investigations along this line for future research.

4.4. Convergence of asymptotic sets

While \( F_\varepsilon \) defined in (32) may not have a unique minimiser, and the dynamics (BD) may not converge to a unique probability distribution as \( t \to \infty \), the \( \Gamma \)-convergence of regularised gradient flows (BD) to (BD) and the long-time convergence of the limiting gradient flow (BD) guarantees that the long-time limiting set of (BD) is close to that of (BD), which is \( \pi \). The goal of this section is to discuss some sufficient conditions where convergence of asymptotic sets of gradient flows hold.

We first present below proposition 4.19, which we state for general gradient flows in metric spaces. In particular the lemma can be applied to gradient flows in Wasserstein metric and has interesting consequences for two-layer neural network training which we will discuss in theorem 4.20.

**Proposition 4.19.** Let \( E_\varepsilon \) and \( E \) be energy functionals, and let \( (\rho_\varepsilon^{(\varepsilon)})_{0 \leq t < T_\varepsilon} \) and \( (\rho_t)_{0 \leq t < \infty} \) be continuous curves in a metric space, with metric \( d \) and maximal existence time \( T_\varepsilon \) and \( \infty \) respectively, such that \( E_\varepsilon \) is nonincreasing along \( \rho_\varepsilon^{(\varepsilon)} \) and \( E \) is nonincreasing along \( \rho_t \). Assume the following conditions hold:

(i) \( E \) has a unique minimiser \( \pi \), and \( \rho_t \) converges to \( \pi \) as \( t \to \infty \) in the sense that \( E(\rho_t) \to E(\pi) \).

(ii) As \( \varepsilon \to 0 \), \( \liminf_{\varepsilon \to 0} E_\varepsilon(\mu_\varepsilon) \geq E(\mu) \) for all sequences \( \mu_\varepsilon \overset{d}{\to} \mu \).

(iii) As \( \varepsilon \to 0 \), we have \( T_\varepsilon \to \infty \). Moreover, for every \( t \geq 0 \), we have that \( \rho_\varepsilon^{(\varepsilon)} \to \rho_t \) in \( d \) and \( E_\varepsilon(\rho_\varepsilon^{(\varepsilon)}) \to E(\rho_t) \).

(iv) The sub-level sets of \( E_\varepsilon \) are uniformly precompact in the following sense: there exists \( \varepsilon_0 > 0 \) such that for any \( M \in (0, \infty) \) the set \( \bigcup \{ E_\varepsilon^{-1}(\varepsilon, M) : 0 < \varepsilon < \varepsilon_0 \} \) is precompact.

Then we have the following:

(a) For any \( \varepsilon > 0 \) and any time sequence \( (T_\varepsilon)_\varepsilon \) such that \( T_\varepsilon < T_\varepsilon \) and \( \lim_{\varepsilon \to 0} T_\varepsilon = \infty \), we have \( \rho_{T_\varepsilon}^{(\varepsilon)} \overset{d}{\to} \pi \) as \( \varepsilon \to 0 \).
(b) If $T^*_\varepsilon = \infty$ for any sufficiently small $\varepsilon$, then let
\[
\mathcal{A}_\varepsilon := \left\{ \rho(\varepsilon) \mid 0 \leq t_1 < t_2 < \ldots < t_n < \ldots \text{s.t. } \lim_{n \to \infty} t_n = \infty \text{ and } \lim_{n \to \infty} \rho(\varepsilon)_n = \rho(\varepsilon)_\infty \right\}
\]
be the $\omega$-limit set of $\rho(\varepsilon)$, we have
\[
\mathcal{A}_\varepsilon \to \{ \pi \} \text{ as } \varepsilon \to 0
\]
with respect to Hausdorff distance corresponding to $d$.

**Proof.** We only prove (a) since the proof for (b) is identical. Fix $\delta > 0$, then there exists a $T > 0$ such that
\[
E(\rho_T) \leq E(\pi) + \delta.
\]
By assumption (iii), there exists some $\varepsilon_0 \geq \varepsilon_1 > 0$ such that for all $\varepsilon < \varepsilon_1$, we have $T^*_\varepsilon > T_\varepsilon > T$, and
\[
E(\rho(\varepsilon)_T) \leq E(\rho(\varepsilon)_T) + \delta.
\]
To prove the claim we argue by contradiction. Assume that $\limsup_{\varepsilon \to 0} d(\rho(\varepsilon), \pi) \geq \lambda > 0$, then along a subsequence (not relabelled) $\varepsilon \to 0$, there exists $\rho(\varepsilon)$ such that $d(\rho(\varepsilon), \pi) > \lambda/2$.

Using that $E(\varepsilon)$ is nonincreasing along $\rho(\varepsilon)$, we have
\[
E(\rho(\varepsilon)_T) \leq E(\rho(\varepsilon)_T) \leq E(\pi) + 2\delta.
\]
Using the compactness assumption (iv) it follows that $\rho(\varepsilon)_T$ is a convergent subsequence in metric $d$ along a further subsequence as $\varepsilon \to 0$ for some $\sigma$. From lower-semicontinuity assumption (ii) follows that
\[
E(\sigma) \leq \liminf_{\varepsilon \to 0} E(\rho(\varepsilon)_T) \leq E(\pi) + 2\delta.
\]
Since $\delta$ is arbitrary, we can take $\delta \to 0$ to obtain
\[
E(\sigma) \leq E(\pi),
\]
which in turn gives $\sigma = \pi$ since $\pi$ is the unique minimiser of $E$. On the other hand,
\[
d(\sigma, \pi) = \lim_{\varepsilon \to 0} d(\rho(\varepsilon)_T, \pi) \geq \frac{\lambda}{2}.
\]
Contradiction. □

An application of proposition 4.19 is the following theorem 4.20. Here let us recall the setting of two-layer neural network training in [37]: the goal is to learn a concave function $f$ using a neural network, which is achieved by minimising the risk functional in domain $\Omega$ with noise level $\tau > 0$:
\[
F^\delta(\rho) = \int_\Omega \left( \frac{1}{2} (K_\delta * \rho - f)^2 + \tau \rho \log \rho \right) dx.
\]
As $\delta \to 0$, $F^\delta$ is close to the limiting functional

$$F(\rho) = \int_\Omega \left( \frac{1}{2} (\rho - f)^2 + \tau \rho \log \rho \right) \, dx,$$

(67)

which has a unique minimiser. In the regime where the number of neurons approach infinity, the process of stochastic gradient descent is characterised by the following equation

$$\partial_t \rho^\delta_t = \nabla \cdot (\rho^\delta_t \nabla \Psi) + \tau \Delta \rho^\delta_t,$$

with $\Psi = -K^\delta * f + K^\delta * K^\delta * \rho^\delta_t$, (68)

which is the Wasserstein gradient flow of (66). Heuristically, as $\delta \to 0$, the solution $\rho^\delta_t$ converges to the solution $\rho_t$ of the viscous porous-medium equation:

$$\partial_t \rho_t = -\nabla \cdot (\rho_t \nabla f) + \Delta \rho_t^2 + \tau \Delta \rho_t.$$  

(69)

Observe that equation (69) is the Wasserstein gradient flow of $F$.

**Theorem 4.20.** Let $\Omega \subset \mathbb{R}^d$ be a convex compact set with a $C^2$-boundary. Assume that $f \in C^\infty(\Omega; \mathbb{R}_+)$ and that $f$ is uniformly concave, i.e. there exists $\alpha > 0$ such that $\nabla^2 f(x) y = -\alpha |y|^2$ for any $x \in \Omega$ and $y \in \mathbb{R}^d$. Under the conditions of [37], let $\rho^\delta_t$ be the solution of (68). Then as $\delta \to 0$, the $\omega$-limit set of $\rho^\delta_t$ converges in Hausdorff metric with respect to $W_2$ to the unique minimiser of $F$ defined in (67).

**Proof.** We need to verify all the conditions in proposition 4.19. The first condition (i) holds using Wasserstein displacement convexity of the energy functional (67), condition (ii) can be proved easily using arguments from [16, theorem 4.1], which is also done in [21, theorem 5.1], and condition (iv) holds trivially on bounded domain. As for (iii), it is proven in [37, lemma E.2] that $T^\delta = \infty$ for all $\delta$, and in [37, theorem 5.2] that as $\delta \to 0$, $\rho^\delta(t) \overset{L^1}{\longrightarrow} \rho(t)$ strongly for almost every $t$. The proof relies on showing the tightness of sequence $\{\rho^\delta_t\} \subseteq [0,T]$ on the space $C([0,T]; P(\Omega))$ as well as the uniqueness of the weak solution of (69). Moreover, for the regularised energy (66) convergence as $\delta \to 0$ is proved in lemma F.3 for the first term, and in the proof of theorem F.8 for the entropy term. Therefore, we can appeal to proposition 4.19 (ii) to prove that the asymptotic sets must also be consistent as $\delta \to 0$. \hfill \Box

In our setting of kernelised birth–death dynamics, the assumptions (i), (ii), (iii) of proposition 4.19, with $d$ being the Wasserstein metric, are verified by lemma 2.2 and theorem 2.4, theorems 4.1 and 4.16 (since $d_{W_2}$ convergence implied convergence in Wasserstein metric) respectively, while assumption (iv) holds trivially on $\mathbb{T}^d$ due to compactness. Therefore we have the following theorem:

**Theorem 4.21.** Under the assumptions of theorem 4.16, for any $\epsilon$ and time sequence $(T_\epsilon)_\epsilon$, such that $\lim_{\epsilon \to 0} T_\epsilon = \infty$ and $(BD_\epsilon)$ is well-posed up to time $T_\epsilon$, we have

$$\rho_{T_\epsilon} \overset{\epsilon \to 0}{\longrightarrow} \pi \text{ in } W_2,$$

and

$$\lim_{\epsilon \to 0} F_\epsilon (\rho_{T_\epsilon}^{(\epsilon)}) = 0.$$

4.5. **Particle based schemes**

One possible idea for particle approximation is to consider particle solutions to $(BD_\epsilon)$, in analogy with the blob method for the Fokker–Planck equation [16]. For discrete measure initial data $\sum_{i=1}^N m_i \delta_{x_i}$, the equation $(BD_\epsilon)$ becomes an ODE system for the masses (64). While formally this provides a deterministic particle-based algorithm that converges to approximation of
there are a number of challenges. Namely, the support of the measure does not change
(i.e. particles do not move), and since masses of particles can become very uneven, this affects
the quality of approximation.

Instead of working to overcome these challenges (which remains an intriguing direction) we
will consider a random, jump, particle process whose mean field limit is the equation \((\text{BD}_e)\).
In the idealised birth–death dynamics with infinitely many particles \((\text{BD})\) (with similar modi-
fications to \((\text{BD}_2)\)), each particle has a jump rate

\[
\Lambda (x, \rho) = \delta \mathcal{F} \delta \rho - \int \delta \mathcal{F} \partial \rho = \log \left( \frac{\rho}{\pi} \right) - \int \log \left( \frac{\rho}{\pi} \right) \rho.
\]

If \(\Lambda > 0\) then the particle jumps out of position \(x\), and if \(\Lambda < 0\) then particle jumps into \(x\), both
with rate \(|\Lambda|\). The issue with implementing such algorithm on the level of particle measures is
that the pointwise density \(\rho\) is not available.

However if one considers the energy \(\mathcal{F}_e\) instead of \(\mathcal{F}\), then the jump rates become

\[
\Lambda_e (x, \rho_t (\varepsilon)) = \frac{\delta \mathcal{F}_e}{\delta \rho_t (\varepsilon)} - \int \frac{\delta \mathcal{F}_e}{\partial \rho_t (\varepsilon)} \rho_t (\varepsilon) = \log \left( \frac{K_e * \rho_t (\varepsilon)}{\pi} \right) + K_e * \left( \frac{\rho_t (\varepsilon)}{K_e * \rho_t (\varepsilon)} \right) - \int \log \left( \frac{K_e * \rho_t (\varepsilon)}{\pi} \right) \rho_t (\varepsilon) - 1. \tag{70}
\]

This interpretation of \((\text{BD}_e)\) allows us to construct a finite particle approximation of the dynam-
ics \((\text{BD}_e)\), that is, if we let \(\rho_t (\varepsilon) = \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i}\), then the particle at location \(x_i\) is removed or added
with rate

\[
\Lambda (x_i) = \log \left( \frac{1}{N} \sum_{j=1}^{N} K_e \left( x_i - x_j \right) \right) + \sum_{j=1}^{N} \frac{K_e \left( x_i - x_j \right)}{\sum_{\ell=1}^{N} K_e \left( x_i - x_\ell \right)} - \log \pi (x_i)
- \frac{1}{N} \sum_{\ell=1}^{N} \log \left( \frac{1}{N} \sum_{j=1}^{N} K_e \left( x_i - x_j \right) \right) - 1 + \frac{1}{N} \sum_{\ell=1}^{N} \log \pi (x_\ell). \tag{71}
\]

This interpretation is where our sampling algorithm, as well as the one in [57], are based on. To
preserve the number of particles, at each birth–death step we uniformly add or remove another
particle if the one in \(x_i\) is remover or added, respectively.

Before we discuss some properties and modifications to this jump dynamics, let us remark
that there is an alternative way to create a jump process whose mean field limit, as \(\varepsilon \to 0\) is
expected to approach the pure birth–death process, \((\text{BD})\), namely simply replacing \(\rho\) by \(\rho * K_e\)
in the rates for birth and death in \((\text{BD})\). This is the approach considered in [57]. The jump rates
for such process are

\[
\hat{\Lambda}_e (x, \rho_t (\varepsilon)) = \log \left( \frac{K_e * \rho_t (\varepsilon)}{\pi} \right) - \int \log \frac{K_e * \rho_t (\varepsilon)}{\pi} \rho_t (\varepsilon) \tag{72}
\]

The expected mean field limit for fixed \(\varepsilon > 0\) would be

\[
\hat{\partial} \rho_t (\varepsilon) \equiv -\rho_t \left( \log \left( K_e * \rho_t (\varepsilon) \right) - \log \pi - \int \log \frac{K_e * \rho_t (\varepsilon)}{\pi} \rho_t (\varepsilon) \right). \tag{73}
\]
A downside of the dynamics (73) is that it is unclear if it possesses a gradient flow structure or a Lyapunov functional that approximates KL divergence, which is why we are modifying the jump rates to be (71).

An alternative ensemble based sampling, where the birth–death process was achieved via jumps, has recently been introduced and studied in [50], where the jump rate was

$$\Theta_{\varepsilon} = \frac{K_{\varepsilon} * \rho^{(c)}}{\pi} - \int K_{\varepsilon} * \rho^{(c)} \rho^{(c)},$$

(74)

The limit of the mean field dynamics as $\varepsilon \to 0$ is the spherical Hellinger gradient flow of the $\chi^2$ divergence, (BD2). In particular the birth–death part of their dynamics relates to (BD2) in the same way as the dynamics of (73) relates to Hellinger gradient flow of KL divergence, (BD). Unlike our choice of (71), the rate in (74) does depend on the normalisation constant of $\pi$, which [50] can avoid by a rescaling of time.

We note that a serious issue with jump processes discussed above is that the support of $\pi$ can avoid by a rescaling of time. In our numerical experiments in section 5 we combine the jump process of $\Lambda_{\varepsilon}$ with unadjusted Langevin algorithm (ULA). The latter sampler is responsible for exploring new territory, especially high density regions in the state space. In effect we move each particles by a gradient descent plus sampling the Gaussian centered at the particle. We did not add a Metropolis step in our experiments.

5. Numerical Examples

Example 5.1 (A toy example). This example is a modification of [57, example 2]. We consider a two-dimensional Gaussian mixture model with four components, i.e. $\pi(x, y) = \sum_{i=1}^{4} w_i \times \mathcal{N}(m_i, \Sigma_i)$ and initial particles sampled from Gaussian $\mathcal{N}(m_0, \Sigma_0)$ where the parameters are given by

$$[w_1, w_2, w_3, w_4] = [0.5, 0.1, 0.1, 0.3], \quad m_1 = [0, 2], \quad m_2 = [-3, 5], \quad m_3 = [0, 8], \quad m_4 = [3, 5].$$

$$\Sigma_1 = \Sigma_3 = \begin{pmatrix} 0.8 & 0 \\ 0 & 0.01 \end{pmatrix}, \quad \Sigma_2 = \Sigma_4 = \begin{pmatrix} 0.01 & 0 \\ 0 & 1 \end{pmatrix}, \quad m_0 = [0, 8], \quad \Sigma_0 = \begin{pmatrix} 0.3 & 0 \\ 0 & 0.3 \end{pmatrix}.$$

Morally speaking, each of the four Gaussian components of $\pi$ are essentially supported on a very narrow domain with little intersection between each other. At the beginning, all particles are concentrated near the top Gaussian centered at $m_0 = m_3$, which is a metastable region, so if the particles follow the overdamped Langevin dynamics, it will take an extremely long time for each particle to escape any certain Gaussian, and with many particles, it is numerically intractable to observe a significant amount of particles to be present in all metastable regions.

Our algorithm birth-death Langevin sampler (BDLS)-KL is an implementation on a modification of (BD2), that is,

$$\partial_{t} \rho_{i}^{(c)} = \nabla \cdot \left( \rho_{i}^{(c)} \nabla \log \frac{\rho_{i}^{(c)}}{\pi} \right) - \rho_{i}^{(c)} \left( \log \left( \frac{K_{\varepsilon} * \rho_{i}^{(c)}}{\pi} \right) + K_{\varepsilon} * \frac{\rho^{(c)}}{K_{\varepsilon} * \rho^{(c)}} \right) - \int \rho_{i}^{(c)} \log \left( \frac{K_{\varepsilon} * \rho_{i}^{(c)}}{\pi} \right) - 1 \right).$$

(75)
Nonlinearity 36 (2023) 5731
Y Lu et al

Figure 2. Gaussian mixture example. Left: error of observable \( f(x, y) = x^2/3 + y^2/5 \); centre: MMD with kernel \( K(x, y) = (2\pi)^{-d/2} e^{-|x-y|^2/(2\varepsilon^2)} \); right: observable error and MMD for Langevin dynamics (ULA) and SVGD up to \( T = 100 \). Both left and center plots are averaged over 30 experiments. Both birth–death algorithms based on KL and \( \chi^2 \) converge much faster to equilibrium as \( t \) gets larger.

We simulate (75) using a ‘splitting scheme’, that is alternating between an unadjusted Langevin step and a birth–death step. More precisely, we use approximate the density \( \rho(t) \) with a finite sum of Diracs with equal weights, i.e. \( \rho(t) \approx \frac{1}{N} \sum_{i=1}^{N} \delta_{\varepsilon x_i(t)} \), and at each time step we first perform a Langevin move for all particles and then a birth–death move with jump rates given by (71). The Fokker–Planck term is necessary in our algorithm due to the fact that the pure birth–death dynamics does not find new locations. For the algorithm BDLS-chi2, we replace the energy functional by regularized \( \chi^2 \)-divergence, that is \( \frac{1}{2} \int \rho K \exp, \frac{1}{\pi} \), and everything else is identical to BDLS-KL.

We compare these two birth–death based sampling methods with the unadjusted overdamped Langevin dynamics (ULA) as well as SVGD [53]. We choose \( N = 800 \) particles, \( \Delta t = 10^{-3} \) and kernel bandwidth \( \varepsilon = 0.2 \) for all algorithms and compare their error of estimating \( \mathbb{E}_\pi f \) with \( f(x, y) = x^2/3 + y^2/5 \), as well as their maximum mean discrepancy (MMD) [4], which can be computed explicitly for Gaussian mixtures.

Figure 2 shows that the algorithms based on birth–death dynamics converge to equilibrium much faster than Langevin dynamics or SVGD in terms of both observable error and MMD distance. More precisely, algorithms based on birth–death dynamics reach equilibrium at \( T \approx 10 \), while for the other two algorithms, although they also eventually converge to equilibrium, it is not achieved even at \( T \approx 100 \). Figure 3 provides a more intuitive explanation on the fact that birth–death based algorithms are significantly better at penetrating energy barriers and overcoming metastability.

Example 5.2 (Real-world data). We also tested the birth–death method on the Bayesian logistic regression for binary classification using the same setting as [33, 42, 53], which assigns the hidden regression weights \( w \) with a precision parameter \( \alpha \in \mathbb{R}_+ \), and that we impose Gaussian prior \( p(w|\alpha) = \mathcal{N}(w, \alpha^{-1}\text{Id}) \) on \( w \) and \( p(\alpha) = \text{Exp}(0.01) \). The observables \( y \in \{-1, 1\} \) are generated by \( P(y=1|x, w) = (1 + \exp(-w^\top x))^{-1} \). The inference is applied on posterior \( p(w, \log \alpha||x, y) \). We compare the performance of our algorithm with SVGD in terms of accuracy and log-likelihood. We would like to comment here that the kernel bandwidth of SVGD is chosen using the median trick, while for birth–death, since the bandwidth is proportional to the bias, we choose the bandwidth to be 0.1, 0.5 or 1, whichever provides the best performance. The results are shown in table 1, showing that when the dimension is not too large and running time is relatively long, both birth–death sampler (with Langevin steps) and SVGD perform similarly well.
Figure 3. Gaussian mixture example. Top: position of particles at $T = 3$; bottom: position of particles at $T = 10$. Algorithms based on birth–death are better at attracting particles into under-explored regions.
Table 1. Bayesian logistic regression for binary classification. For both algorithms, \( N = 500 \), time step size \( \Delta t = 10^{-3} \), final time \( T = 15 \).

| Dataset     | Dimension | BDLS accuracy | BDLS log-likelihood | SVGD accuracy | SVGD log-likelihood |
|-------------|-----------|---------------|---------------------|---------------|---------------------|
| Banana      | 3         | 0.583         | -0.690              | 0.585         | -0.686              |
| Breast_cancer | 10       | 0.714         | -0.604              | 0.714         | -0.586              |
| Diabetes    | 9         | 0.763         | -0.527              | 0.753         | -0.529              |
| Flare_solar | 10        | 0.683         | -0.578              | 0.685         | -0.600              |
| German      | 21        | 0.687         | -0.598              | 0.680         | -0.597              |
| Heart       | 14        | 0.840         | -0.376              | 0.850         | -0.379              |
| Image       | 19        | 0.817         | -0.433              | 0.815         | -0.434              |
| Ringnorm    | 21        | 0.760         | -0.521              | 0.760         | -0.501              |
| Thyroid     | 6         | 0.933         | -0.250              | 0.933         | -0.294              |
| Titanic     | 4         | 0.780         | -0.586              | 0.785         | -0.566              |
| Twonorm     | 21        | 0.973         | -0.069              | 0.975         | -0.112              |
| Waveform    | 22        | 0.773         | -0.466              | 0.773         | -0.469              |

Figure 4. Bayesian classification problem with dataset 'Image'. The birth–death Langevin algorithm reaches the desired accuracy and log-likelihood much faster than SVGD.

We also compare the behaviour of both criteria, accuracy and log-likelihood, as time evolves between \( t \in [0, 5] \). The results are shown in figure 4. One can observe that birth–death Langevin sampler reaches the desired accuracy at an extremely short time \( t \approx 0.3 \), which SVGD cannot achieve before \( t = 5 \). This indicates that one can run BDLS for a much shorter time to reach equilibrium, significantly alleviating the computational cost issue of running a system of many particles. This is the spirit of our theorem 2.4.

We would like to comment that for the dataset ‘splice’ where \( d = 61 \), the performance of birth–death sampler is significantly worse, even with \( N = 2000 \), which is not entirely surprising due to the limitations of kernel density estimation. It is an interesting open question to design a sampling algorithm which can inherit the fast convergence properties of spherical Hellinger gradient flows and be robust in high dimension settings.
Data availability statement

The data cannot be made publicly available upon publication because they are not available in a format that is sufficiently accessible or reusable by other researchers. The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgment

The authors would like to thank Katy Craig and Léonard Monsaingeon for helpful discussions. L W would like to thank Rishabh Gvalani for discussions during revision. Y L thanks NSF for the support via the Award DMS-2107934. D S is grateful to NSF for support via grant DMS 2206069. Part of this work was done while the authors were visiting the Simons Institute for the Theory of Computing. The authors thank the institute for hospitality. D S and L W would also like to thank CNA at CMU for support.

ORCID iDs

Yulong Lu https://orcid.org/0000-0003-2914-414X
Dejan Slepečev https://orcid.org/0000-0002-7600-1144
Lihan Wang https://orcid.org/0000-0002-9130-0505

References

[1] Amari S-I 2016 Information Geometry and its Applications vol 194 (Springer)
[2] Ambrosio L, Gigli N and Savaré G 2008 Gradient Flows in Metric Spaces and in the Space of Probability Measures (Springer)
[3] Andersen H C 1980 Molecular dynamics simulations at constant pressure and/or temperature J. Chem. Phys. 72 2384–93
[4] Arbel M, Korba A, Salim A and Gretton A 2019 Maximum mean discrepancy gradient flow Advances in Neural Information Processing Systems vol 32
[5] Ay N, Jost J, Van Lê H and Schwachhöfer L 2017 Information Geometry vol 64 (Springer)
[6] Bakry D, Gentil I and Ledoux M 2014 Analysis and Geometry of Markov Diffusion Operators vol 103 (Springer)
[7] Bernton E 2018 Langevin Monte Carlo and JKO splitting Proc. 31st Conf. On Learning Theory (PMLR) pp 1777–98
[8] Bierkens J, Fearnhead P and Roberts G 2019 The zig-zag process and super-efficient sampling for Bayesian analysis of big data Ann. Stat. 47 1288–320
[9] Bou-Rabee N and Sanz-Serna J M 2017 Randomized Hamiltonian Monte Carlo Ann. Appl. Probab. 27 2159–94
[10] Bouchard-Côté A, Vollmer S J and Doucet A 2018 The bouncy particle sampler: a nonreversible rejection-free Markov chain Monte Carlo method J. Am. Stat. Assoc. 113 855–67
[11] Breiner Y and Vorotnikov D 2020 On optimal transport of matrix-valued measures SIAM J. Math. Anal. 52 2849–73
[12] Bunch P and Godsill S 2016 Approximations of the optimal importance density using Gaussian particle flow importance sampling J. Am. Stat. Assoc. 111 748–62
[13] Carrillo J, Hoffmann F, Stuart A and Vaes U 2022 Consensus-based sampling Stud. Appl. Math. 148 1069–140
[14] Carrillo J A, Choi Y-P, Totzeck C and Tse O 2018 An analytical framework for consensus-based global optimization method Math. Models Methods Appl. Sci. 28 1037–66
[15] Carrillo J A, McCann R J and Villani C 2006 Contractions in the 2-Wasserstein length space and thermalization of granular media Arch. Ration. Mech. Anal. 179 217–63
[16] Carrillo J A, Craig K and Patacchini F S 2019 A blob method for diffusion Calc. Var. PDE 58 53
[17] Chen Y, Huang D Z, Huang J, Reich S and Stuart A M 2023 Gradient flows for sampling: mean-field models, Gaussian approximations and affine invariance (arXiv:2302.11024)
[18] Chizat L, Peyré G, Schmitzer B and Vialard F-X 2018 An interpolating distance between optimal transport and Fisher–Rao metrics Found. Comput. Math. 18 1–44
[19] Craig K, Elamvazhuthi K, Haberland M and Turanova O 2022 A blob method method for inhomogeneous diffusion with applications to multi-agent control and sampling (arXiv:2202.12927)
[20] Craig K and Topaloglu I 2016 Convergence of regularized nonlocal interaction energies SIAM J. Math. Anal. 48 34–60
[21] Dalalyan A S 2017 Theoretical guarantees for approximate sampling from smooth and log-concave densities J. R. Stat. Soc. B 79 651–76
[22] Darve E and Pohorille A 2001 Calculating free energies using average force J. Chem. Phys. 115 9169–83
[23] Del Moral P, Doucet A and Jasra A 2006 Sequential Monte Carlo samplers J. R. Stat. Soc. B 68 411–36
[24] Domingo-Erich C and Pooladian A-A 2023 An explicit expansion of the Kullback-Leibler divergence along its Fisher–Rao gradient flow (arXiv:2302.12229)
[25] Duncan A, Nüsken N and Szpruch L 2019 On the geometry of Stein variational gradient descent (arXiv:1912.00894)
[26] Durmus A and Moulines E 2017 Nonasymptotic convergence analysis for the unadjusted Langevin algorithm Ann. Appl. Probab. 27 1551–87
[27] Foreman-Mackey D, Hogg D W, Lang D and Goodman J 2013 emcee: the MCMC hammer Publ. Astron. Soc. Pac. 125 306
[28] Gershman S, Hoffman M D and Blei D M 2012 Nonparametric variational inference Proc. 29th Int. Conf. on Machine Learning
[29] Goodman J and Weare J 2010 Ensemble samplers with affine invariance Commun. Appl. Math. Comput. Sci. 5 65–80
[30] Hénin J and Chipot C 2004 Overcoming free energy barriers using unconstrained molecular dynamics simulations J. Chem. Phys. 121 2904–14
[31] Javanmard A, Mondelli M and Montanari A 2020 Analysis of a two-layer neural network via displacement convexity Ann. Stat. 48 3619–42
[32] Jordan R, Kinderlehrer D and Otto F 1998 The variational formulation of the Fokker–Planck equation SIAM J. Math. Anal. 29 1–17
[33] Kondratyev S, Monsaingeon L and Vorotnikov D 2016 A new optimal transport distance on the space of finite radon measures Adv. Differ. Equ. 21 1117–64
[34] Lambert M, Chewi S, Bach F, Bonnabel S and Rigollet P 2022 Variational inference via Wasserstein gradient flows (arXiv:2205.15902)
[45] Laschos V and Mielke A 2019 Geometric properties of cones with applications on the Hellinger-Kantorovich space and a new distance on the space of probability measures J. Funct. Anal. 276 3529–76
[46] Laschos V and Mielke A 2022 Evolutionary variational inequalities on the Hellinger-Kantorovich and spherical Hellinger-Kantorovich spaces’ (arXiv:2207.09815)
[47] Lelièvre T, Roussel M and Stoltz G 2008 Long-time convergence of an adaptive biasing force method Nonlinearity 21 1155
[48] Lelièvre T, Rousset M and Stoltz G 2008 Long-time convergence of an adaptive biasing force method Nonlinearity 21 1155
[49] Liero M, Mielke A and Savaré G 2018 Optimal entropy-transport problems and a new Hellinger–Kantorovich distance Invent. Math. 211 969–1117
[50] Lindsey M, Wearn J and Zhang A 2022 Ensemble Markov chain Monte Carlo with teleporting walkers SIAM/ASA J. Uncertain. Quantification 10 860–85
[51] Liu L, Majka M B and Szpruch Ł 2023 Polyak–lojasiewicz inequality on the space of measures and convergence of mean-field birth-death processes Appl. Math. Optim. 87 48
[52] Liu J and Wang L 2022 On explicit L2-convergence rate estimate for piecewise deterministic Markov processes in MCMC algorithms Ann. Appl. Probab. 32 1333–61
[53] Lu J, Lu Y and Nolen J 2019 Accelerating Langevin sampling with birth-death (arXiv:1905.09863)
[54] Ma Y-A, Chatterji N S, Cheng X, Flammarion N, Bartlett P L and Jordan M I 2021 Is there an analog of Nesterov acceleration for gradient-based MCMC? Bernoulli 27 1942–92
[55] Marinari E and Parisi G 1992 Simulated tempering: a new Monte Carlo scheme Europhys. Lett. 19 451
[56] Muratori M and Savaré G 2020 Gradient flows and evolution variational inequalities in metric spaces. I: structural properties J. Funct. Anal. 278 108347
[57] Neal R M 2001 Annealed importance sampling Stat. Comput. 11 125–39
[58] Neal R M 2011 MCMC using hamiltonian dynamics Handbook of Markov Chain Monte Carlo vol 2 p 2
[59] Otto F and Villani C 2000 Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality J. Funct. Anal. 173 361–400
[60] Pinnau R, Totzeck C, Tse O and Martin S 2017 A consensus-based model for global optimization and its mean-field limit Math. Models Methods Appl. Sci. 27 183–204
[61] Reich S 2011 A dynamical systems framework for intermittent data assimilation BIT Numer. Math. 51 235–49
[62] Reich S 2013 A guided sequential Monte Carlo method for the assimilation of data into stochastic dynamical systems Recent Trends in Dynamical Systems (Springer) pp 205–20
[63] Roberts G O and Tweedie R L 1996 Exponential convergence of Langevin distributions and their discrete approximations Bernoulli 2 341–63
[64] Rossky P J, Doll J D and Friedman H L 1978 Brownian dynamics as smart Monte Carlo simulation J. Chem. Phys. 69 4628–33
[65] Rotskoff G M, Jelassi S, Bruna J and Vanden-Eijnden E 2019 Global convergence of neuron birth-death dynamics 36th Int. Conf. on Machine Learning (Int. Machine Learning Society (IMLS)) pp 9689–98
[66] Sander E and Serfaty S 2004 Gamma-convergence of gradient flows with applications to Ginzburg-Landau Commun. Pure Appl. Math. 57 1627–72
[67] Serfaty S 2011 Gamma-convergence of gradient flows on Hilbert and metric spaces and applications Discrete Contin. Dyn. Syst. 31 1427
[68] Steerneman T 1983 On the total variation and Hellinger distance between signed measures; an application to product measures Proc. Am. Math. Soc. 88 684–8
[69] Swendsen R H and Wang J-S 1986 Replica Monte Carlo simulation of spin-glasses Phys. Rev. Lett. 57 2607
[74] Vempala S and Wibisono A 2019 Rapid convergence of the unadjusted Langevin algorithm: isoperimetry suffices \textit{Advances in Neural Information Processing Systems} vol 32

[75] Wang F and Landau D P 2001 Efficient, multiple-range random walk algorithm to calculate the density of states \textit{Phys. Rev. Lett.} \textbf{86} 2050

[76] Wei C, Lee J D, Liu Q and Ma T 2019 Regularization matters: generalization and optimization of neural nets vs their induced kernel \textit{Advances in Neural Information Processing Systems} vol 32

[77] Wibisono A 2018 Sampling as optimization in the space of measures: the Langevin dynamics as a composite optimization problem \textit{Conf. on Learning Theory} (PMLR) pp 2093–3027