Weierstraß-Institut
für Angewandte Analysis und Stochastik
Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

Hellinger–Kantorovich gradient flows: Global exponential decay of entropy functionals

Alexander Mielke[1], Jia-Jie Zhu[2]

submitted: February 17, 2025

[1] Weierstraß-Institut
Mohrenstr. 39
10117 Berlin
and
Institut für Mathematik
Humboldt-Universität zu Berlin
Rudower Chaussee 25
12489 Berlin
Germany
E-Mail: alexander.mielke@wias-berlin.de

[2] Weierstraß-Institut
Mohrenstr. 39
10117 Berlin
Germany
E-Mail: jia-jie.zhu@wias-berlin.de

No. 3176
Berlin 2025

2020 *Mathematics Subject Classification.* 49Q22, 35Q49, 28A33, 49J40, 47J30.

Key words and phrases. Optimal transport, gradient flow, Otto–Wasserstein, Hellinger, Fisher–Rao, unbalanced transport, optimization, calculus of variations, statistical inference, sampling.

This project was partially supported by the Deutsche Forschungsgemeinschaft (DFG) through the Berlin Mathematics Research Center MATH+ (EXC-2046/1, project ID: 390685689).

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Hellinger–Kantorovich gradient flows: Global exponential decay of entropy functionals

Alexander Mielke, Jia-Jie Zhu

Abstract

We investigate a family of gradient flows of positive and probability measures, focusing on the Hellinger–Kantorovich (HK) geometry, which unifies transport mechanism of Otto–Wasserstein, and the birth-death mechanism of Hellinger (or Fisher–Rao). A central contribution is a complete characterization of global exponential decay behaviors of entropy functionals under Otto–Wasserstein and Hellinger-type gradient flows. In particular, for the more challenging analysis of HK gradient flows on positive measures—where the typical log-Sobolev arguments fail—we develop a specialized shape-mass decomposition that enables new analysis results. Our approach also leverages the Polyak–Łojasiewicz-type functional inequalities and a careful extension of classical dissipation estimates. These findings provide a unified and complete theoretical framework for gradient flows and underpin applications in computational algorithms for statistical inference, optimization, and machine learning.

1 Introduction

We adopt a perspective rooted in the series of works from the 1990s that pioneered the study of Otto-Wasserstein gradient flows, as eloquently articulated by [Ott01]:

> The merit of the right gradient flow formulation of a dissipative evolution equation is that it separates *energetics* and *kinetics*: The energetics endow the state space with a *functional*, the kinetics endow the state space with a (Riemannian) *geometry* via the metric tensor.

In essence, the seminal works such as [Ott01, JKO98] enabled a systematic perspective of studying the PDE such as the type

$$\partial_t \mu = -\operatorname{div}\left(\mu \nabla \frac{\delta F}{\delta \mu}[\mu]\right)$$

as gradient flows of the energy functional F, where $\frac{\delta F}{\delta \mu}[\mu]$ is its first variation. The solution μ_t can be viewed as the dynamics and the solution paths of the measure optimization problem $\min_{\mu \in \mathcal{P}(\mathbb{R}^d)} F(\mu)$ in the Wasserstein space of probability measures with finite second moment, denoted by $\big(\mathcal{P}(\mathbb{R}^d), W_2\big)$. This perspective has been instrumental in advancing the theory of computational algorithms for statistical inference and, more recently, machine learning.

DOI 10.20347/WIAS.PREPRINT.3176 Berlin 2025

Statistical sampling and particle approximation For example, suppose a statistician wishes to generate samples from a distribution π, whose density is in the form $\pi(x) = \frac{1}{\int \mathrm{e}^{-V(x)}\,\mathrm{d}x}\mathrm{e}^{-V(x)}$, where V is referred to as the potential energy function. This can be cast in the Bayesian inference framework, that infers the posterior distribution π of some model parameters. In such applications, one can rely on the fact that π is the invariant distribution of the system associated with the Langevin stochastic differential equation $\mathrm{d}X_t = -\nabla V(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}Z_t$, where Z_t is the standard Brownian motion. Then, by numerically simulating the SDE, computational algorithms can be designed to generate samples that approximate those of π. From the PDE perspective, this Langevin SDE describes the same dynamical system as the deterministic drift-diffusion Fokker-Planck PDE

$$\partial_t \mu = -\operatorname{div}\left(\mu \nabla \left(V + \log \mu\right)\right) \tag{1}$$

for probability measure μ, which is the *gradient-flow equation* of the Otto-Wasserstein gradient flow of the KL divergence as driving energy functional, $F(\mu) = \mathrm{D}_{\mathrm{KL}}(\mu|\pi)$. Then, the rigorous analysis developed in the applied analysis context can be used to study the computational algorithms.

Variational inference and information geometry In practice, the exact posterior distribution π is often intractable, and one can resort to approximate *variational inference* methods [JG*99, WaJ08, BKM17]. Different from the Langevin sampler approach, this amounts to finding the approximate posterior probability measure by parameterizing μ with some parameter $\eta \in E \subset \mathbb{R}^n$, resulting in the optimization problem

$$\min_{\eta \in E \subset \mathbb{R}^n} \mathrm{D}_{\mathrm{KL}}(\mu_\eta|\pi). \tag{2}$$

The parameterized distribution μ_η can be chosen from certain families of distributions, e.g., the family of Gaussian or its mixtures. In such cases, an efficient approach is the natural gradient descent [Ama98, AmN00, KhN18, HB*13, KhR23] on η that respects the geometry of the parameterized probability space. In practice, the update rule is a Riemannian gradient descent scheme

$$\eta^{k+1} \leftarrow \operatorname*{argmin}_{\eta \in E} \nabla_\eta F(\mu_{\eta^k})(\eta - \eta^k) + \frac{1}{2\tau}(\eta - \eta^k)^\top \mathbb{G}_{\mathsf{FR}}(\eta^k)(\eta - \eta^k), \tag{3}$$

where $F = \mathrm{D}_{\mathrm{KL}}(\cdot|\pi)$ in the KL variational inference context and $\nabla_\eta F(\mu_\eta)$ its Euclidean gradient with respect to η. The matrix $\mathbb{G}_{\mathsf{FR}}(\nu) := \int \mu_\nu(x) \cdot \left(\nabla_\nu \log \mu_\nu\,(x)\right)\left(\nabla_\nu \log \mu_\nu\,(x)\right)^\top \mathrm{d}x$ is referred to as the Fisher information matrix, which can be seen as a Riemannian tensor on (the tangent bundle of) $E \subset \mathbb{R}^n$ and hence induces the Fisher-Rao distance over some family of distributions. It is closely related to the Hellinger distance, which is a central topic in this paper: the Fisher-Rao distance can be viewed as a restriction of the Hellinger geodesic distance to the submanifold of cerntain families of distributions, e.g., Gaussian; see Remark 2.5 for more details.

Optimization and mirror descent In the optimization literature, there is a class of algorithms that uses the Bregman divergence as the underlying geometry under the name of mirror descent, e.g., [BeT03, NeY83, DH*16]. If the Bregman divergence is chosen as the KL divergence, this approach is termed the *entropic mirror descent*, i.e., an optimization algorithm solving, at the k-th iteration,

$$\nu^{k+1} \leftarrow \operatorname*{argmin}_{\nu \in \Delta^d} \nabla_\nu F(\nu^k)(\nu - \nu^k) + \frac{1}{\tau}\mathrm{D}_{\mathrm{KL}}(\nu|\nu^k). \tag{4}$$

for $\Delta^d := \mathcal{P}(\Omega)$ where Ω is a finite discrete set of cardinality d. In essence, they can be used to solve the optimization problem $\min\limits_{\mu\in\mathcal{P}(\mathbb{R}^d)} F(\mu)$ in practice by (i) either considering the probability measure on a finite domain [BeT03], or (ii) by considering a particle approximation of the measure $\hat{\nu} = \sum_{i=1}^N w_i\delta_{x_i}$, where δ_{x_i} is the Dirac measure at the particle location x_i and w_i is the weight, see e.g., [Chi22, DH*16]. Furthermore, the extension to non-gradient flows, albeit finite-dimensional, has been studied in the optimization literature, e.g., [WWJ16, KBB15]

The above applications are deeply connected to the gradient flow theory in various geometries, i.e., not just the Otto-Wasserstein geometry. Specifically, the Hellinger type gradient flows, which is the focus of this paper, plays a crucial role and possesses distinct properties when compared with Otto-Wasserstein. In this paper, we advance the state-of-the-art analysis of gradient flows over positive and probability measures using tools such as the Polyak-Lojasiewicz functional inequality.

Analysis of Polyak-Lojasiewicz functional inequalities Historically, the celebrated Bakry-Émery theorem [BaÉ85] provides a powerful strategy for analyzing the convergence of the Otto-Wasserstein gradient flow under the KL-divergence energy. The Bakry-Émery condition implies a key functional inequality, the Logarithmic-Sobolev inequality (LSI)

$$\left\|\nabla \log \frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right\|^2_{L^2(\mu)} \geq c_{\mathrm{LSI}} \cdot \mathrm{D}_{\mathrm{KL}}(\mu|\pi) \text{ for some } c_{\mathrm{LSI}} > 0. \tag{LSI}$$

The LSI can be viewed as a special case of the (Polyak-)Łojasiewicz inequality specialized to the Otto-Wasserstein gradient flow of the KL energy $\mathrm{D}_{\mathrm{KL}}(\mu|\pi)$. It provides a powerful tool for characterizing the convergence of the dynamics, e.g., governed by the Langevin SDE. From the optimization perspective, this is equivalent to analyzing the optimization dynamics of the problem

$$\min_{\mu\in A\subset\mathcal{P}} \mathrm{D}_{\mathrm{KL}}(\mu|\pi) \text{ in the space of } (\mathcal{P}, W_2).$$

The main goal of this paper is to extend this type of functional inequalities, and consequently the convergence analysis, to a broader class of gradient flows beyond the now-standard setting of Otto-Wasserstein flows in $(\mathcal{P}, W_2)$ and the KL energy functional D_{KL}. We now briefly elaborate on those two aspects of our contribution.

Generalizing the energy functional: from KL to φ_p-divergences The KL-divergence is by no means the only entropy-type divergence that possesses interesting properties. For example, some works by [CG*20, LSW23] also consider the χ^2-divergence as the driving energy for machine learning applications. [Zhu24] shows that many existing machine learning algorithms are performing the forward KL (also referred to as the inclusive KL therein) minimization via kernelized Wasserstein gradient flows. For this reason, we first generalize the KL-divergence energy in (2) to a commonly used family of divergence functional, the *φ-divergence* [Csi67]. We now define this divergence on non-negative measures $\mathcal{M}^+(\Omega) \times \mathcal{M}^+(\Omega)$ as

$$\mathrm{D}_\varphi(\mu|\nu) := \begin{cases} \int \varphi(\frac{\mathrm{d}\mu}{\mathrm{d}\nu})\,\mathrm{d}\nu & \text{if } \mu \ll \nu \text{ (i.e. } \mu = f\nu \text{ for some } f \in L^1(\nu)), \\ +\infty, & \text{otherwise.} \end{cases} \tag{5}$$

where $\varphi : [0, +\infty) \to [0, +\infty]$ is a convex entropy generator function that satisfies

$$\varphi(1) = \varphi'(1) = 0,\ \varphi''(1) = 1. \tag{6}$$

We delve specifically into the concrete instantiations of the Łojasiewicz inequality for the following power entropy generator functions.

$$\varphi_p(s) := \frac{1}{p(p-1)}\left(s^p - ps + p - 1\right), \quad p \in \mathbb{R}\setminus\{0,1\}, \tag{7}$$

which satisfies (6) and $\varphi_p''(s) = s^{p-2}$. Using the property $\varphi_p(s) = s\varphi_{1-p}(1/s)$, we obtain the symmetry $D_{\varphi_p}(\rho|\pi) = D_{\varphi_{1-p}}(\pi|\rho)$. Many commonly used divergences can be recovered using the power entropy, e.g.,

$$\begin{aligned} &\text{KL: } \varphi_1(s) := s\log s - s + 1, && \text{fwd-KL: } \varphi_0(s) := s - 1 - \log s, \\ &\chi^2: \ \varphi_2(s) = \tfrac{1}{2}(s-1)^2, && \text{rev-}\chi^2: \varphi_{-1}(s) = \tfrac{1}{2}\left(s + \tfrac{1}{s} - 2\right). \end{aligned} \tag{8}$$

We refer to the resulting divergence functional $\mathrm{D}_{\varphi_\mathrm{p}}$ as the φ_p*-divergence* or the $p-$relative entropy (cf. [OhT11]). Slightly abusing the terminology due to a scaling factor, we refer to the power-like entropy generated by $\varphi_{\frac{1}{2}}$ as the squared Hellinger distance, i.e.,

$$\frac{1}{2}\mathrm{He}^2(\mu,\nu) = \int \varphi_{\frac{1}{2}}\left(\frac{\delta\mu}{\delta\nu}\right)\mathrm{d}\nu. \tag{9}$$

We plot the corresponding entropy generator functions in Figure 1. Alternatively, one may use the

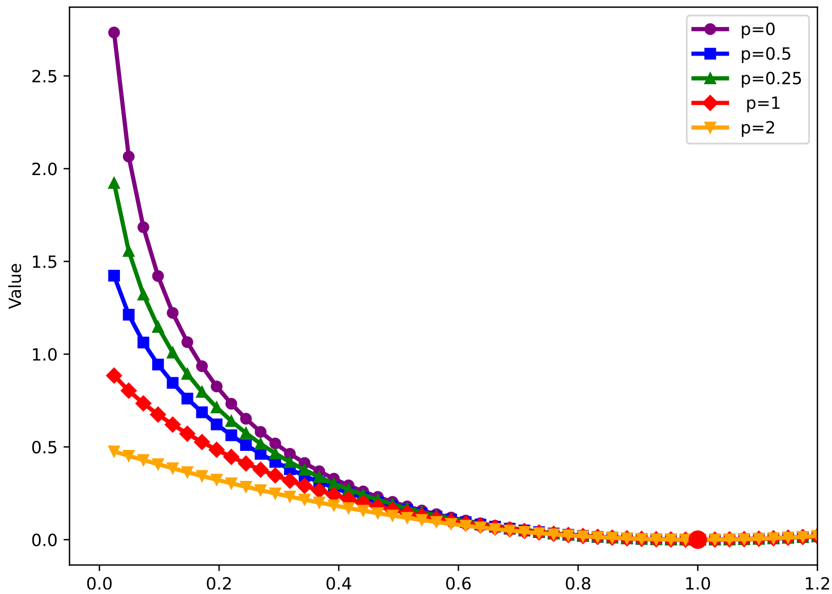

Figure 1: The plot illustrates the power-like entropy generator functions $\varphi_p(s)$ for $s \in [0, 1.2]$ and different p: purple $p = 0$ (forward KL), green $p = 0.25$, blue $p = 0.5$ (Hellinger), red $p = 1$ (KL), orange $p = 2$ (χ^2). The large red dot represents the equilibrium at $s = 1$ where $\varphi_p(1) = \varphi_p{}'(1) = 0$.

Hellinger integrals to define the α-divergence $\mathrm{D}_\alpha(\mu|\nu) := \frac{4}{1-\alpha^2}(1 - \int \mu^{\frac{1+\alpha}{2}}\nu^{\frac{1-\alpha}{2}})$ with $\alpha \in (-1,1)$, from which one also obtains the KL, forward KL, and the Hellinger as special cases (for $\alpha \to 1$, $\alpha \to -1$, and $\alpha = 0$, respectively).

Generalizing the flow geometry: from Otto-Wasserstein to Hellinger-Kantorovich In addition to generalizing the energy functional, we extend the analysis of gradient flows beyond the standard Otto-Wasserstein geometry. Similar to (LSI) in that case, one can examine the validity of such Łojasiewicz

Gradient-flow geometry	Geod. convexity of φ_p-divergence
Otto-Wasserstein (Bakry-Émery)	$p \in [1,2]$ and (BE) with $c > 0 \implies$ geod. c-cvx
Otto-Wasserstein (McCann cond.)	$p \in [\frac{d-1}{d}, \infty) \implies$ geod. cvx
HK [LMS23a] over $\mathcal{M}^+$	$p \in [\frac{d}{d+2}, \frac{1}{2}] \cup (1,\infty) \implies$ geod. cvx

Table 1: Geodesic convexity of φ_p-divergence

Gradient-flow geometry	Specialized Łojasiewicz-type inequality ($\alpha, \beta > 0$)
Hellinger	$\left\Vert \frac{\delta F}{\delta\mu}[\mu]\right\Vert^2_{L^2_\mu} \ge c\cdot(F(\mu)-F_{\inf})$
Spherical Hellinger	$\left\Vert \frac{\delta F}{\delta\mu}[\mu] - \int \frac{\delta F}{\delta\mu}[\mu]\,\mathrm{d}\mu\right\Vert^2_{L^2_\mu} \ge c\cdot(F(\mu)-F_{\inf})$
Hellinger-Kantorovich (WFR)	$\alpha\left\Vert \nabla\frac{\delta F}{\delta\mu}[\mu]\right\Vert^2_{L^2_\mu} + \beta\left\Vert \frac{\delta F}{\delta\mu}[\mu]\right\Vert^2_{L^2_\mu} \ge c\cdot(F(\mu)-F_{\inf})$
Spher. Hellinger-Kantorovich	$\alpha\left\Vert \nabla\frac{\delta F}{\delta\mu}[\mu]\right\Vert^2_{L^2_\mu} + \beta\left\Vert \frac{\delta F}{\delta\mu}[\mu] - \int\frac{\delta F}{\delta\mu}[\mu]\,\mathrm{d}\mu\right\Vert^2_{L^2_\mu} \ge c\,(F(\mu)-F_{\inf})$

Table 2: Łojasiewicz inequalities for different gradient flows, where $F_{\inf} := \inf_\mu F(\mu)$.

type functional inequalities when considering general energy functional F in other gradient-flow geometries; see Table 2. Our main topics of study are gradient flows in the Hellinger-Kantorovich (HK) geometry, which independently discovered by a few groups of researchers [CP*18a, CP*18b, LMS18, KMV16, GaM17]. It is an infimal convolution (inf-convolution) of the Hellinger and Wasserstein distances over positive measures $\mathcal{M}^+$. Intuitively, gradient flows in the HK and spherical HK (SHK) geometry combine the dissipation mechanisms of the Otto-Wasserstein flow, i.e., the transport of mass, and the Hellinger flow, i.e., the creation-destruction or birth-death of mass. It is often referred to as the unbalanced transport geometry and possesses a richer structure and more advantageous properties than either of the pure flows alone, and is the focus of this paper. At the same time, the analysis of the HK and SHK gradient flows is more involved than the pure Hellinger or the Otto-Wasserstein gradient flows.

On one hand, the Otto-Wasserstein geometry describes the transport dynamics that can easily handle the change of support of measures. However, it suffers from slow asymptotic convergence in practical applications. For example, the behavior of its gradient flow of the KL divergence depends crucially on the log-Sobolev constant. The reason is that, in the (quadratic) Otto-Wasserstein setting, the transport over large distances (e.g., of outliers in a point cloud) has an over-proportional cost. In contrast, Hellinger type gradient flows enjoy fast asymptotic convergence because mass can be destroyed and created at other places instead without any transport. To understand the distinct nature of the two gradient flows, consider an intuitive example of particle gradient descent method where the measure is approximated using two particles, i.e., $\mu = w_1\delta_{x_1} + w_2\delta_{x_2}$ with $w_1 + w_2 = 1$ and $w_i > 0$. The objective function is an asymmetric double-well potential. The minimization is initialized as solid dots in the illustration in Figure 2. In this case, it is easy to see that the gradient descent for each individual particle, induced by the Otto-Wasserstein gradient flow, will get stuck in the local minimum, as illustrated as dashed dots. On the other hand, the birth-death process, induced by the (spherical) Hellinger gradient flow, can easily teleport the mass to the right well, but it does not allow the change of the location of the particles. An intuitive idea is to consider the *Hellinger-Kantorovich* (HK) gradient flows to combine the strengths of both the Hellinger and Otto-Wasserstein geometries, while overcoming their weaknesses as illustrated in Figure 2.

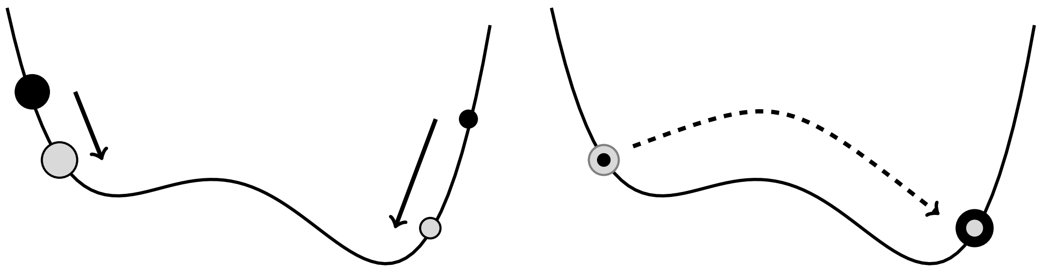

Figure 2: The two figures illustrate the conceptual advantage of combining the Otto-Wasserstein and the Hellinger gradient flows. On the left, the particles are transported by the gradient descent enabled by the Otto-Wasserstein gradient flow, where masses do not change. On the right, the dashed arrow represent the "teleportation" of mass enabled by the Hellinger gradient flow, where the positions do not change. The size of the dots represents the amount of mass of the particles.

Overview of the main results In making the above intuition precise, this paper advances the theory for the HK and SHK gradient flows by establishing a few new and precise analysis results. We provide complete and nontrivial answers to the following open question:

> For the commonly used entropy functionals, e.g., (reverse) KL divergence, forward KL, Hellinger distance, χ^2-divergence, reverse χ^2, what is the convergence behavior of gradient flows in geometries such as the Hellinger-Kantorovich space of positive measures $(\mathcal{M}^+, \mathsf{HK})$, or the spherical Hellinger-Kantorovich space of probability measures $(\mathcal{P}, \mathsf{SHK})$? Similar to the Bakry-Émery Theorem and (LSI) in the standard setting of $(\mathcal{P}, \mathrm{D}_{\mathrm{KL}}, W_2)$, can we establish precise conditions for global convergence of the gradient flows in all the geometries mentioned above?

A few major results are summarized in Table 1. In addition and more concretely, we first establish analysis results for the Łojasiewicz inequality for the pure Hellinger gradient flows. As an example, we show that there is no global Łojasiewicz inequality for the Hellinger flow of the KL energy over positive measures $\mathcal{M}^+$. The global Łojasiewicz inequality is significantly more nontrivial to establish than the local version since we need to create enough metric slope for the Hellinger gradient flow to escape the initial birth from zero mass; see the illustration in Figure 5 and Remark 3.9 for technical details. Our result captures the fundamental nature of the Hellinger gradient flows in contrast to the LSI for the Otto-Wasserstein. Going deeper, the analysis of the HK and SHK flows of φ-divergence is more involved. We systematically extract explicit conditions for global convergence of gradient flows under φ_p-divergence energy functional as defined in (5) and (7). Previously, the geodesic convexity of energy functionals has been under scrutiny in the gradient flow literature; see a summary of the implication on the φ_p-divergence in Table 1. However, when studying the convergence of gradient flows geodesic convexity is a sufficient but not necessary condition. For this reason, this paper establishes a few new functional inequalities that are weaker than geodesic convexity, but still sufficient to guarantee global exponential decay of the energy functional; see Table 1 for the precise statements and references to the corresponding theorems for convenience. In particular, the standard LSI, when considered for positive measures $\mathcal{M}^+$, does not hold globally and must be generalized. This adds to the difficulty of establishing the global convergence of the HK gradient flows. Nonetheless, using a novel shape-mass decomposition analysis technique, we were able to establish global convergence to equilibrium along the HK gradient flow for the KL divergence as driving energy, see Theorem 5.8.

Gradient-flow geometry	Global exp. decay, & fcn. ineq. for φ_p-divergence
Otto-Wasserstein on $\mathcal{P}$	• $\Omega \subsetneq \mathbb{R}^d$ bounded Lipschitz, $p \geq 1-\frac{1}{d} \implies$ Ł with $c_* > 0$ • $\Omega = \mathbb{R}^d, p \in [1,2]$ and (BE) $\implies$ Ł with $c_* = 2c_{\mathrm{BE}}$
Otto-Wasserstein on $\mathcal{M}^+$ (Prop. 5.2, 5.3)	$\nexists\, c_* > 0$ for Ł; see (LSI-$\mathcal{M}^+$)
Hellinger on $\mathcal{M}^+$ (Prop. 3.7)	$p \in (-\infty, \frac{1}{2}] \iff$ Ł with $c_* = \frac{1}{1-p}$
Spherical Hellinger on $\mathcal{P}$ (Thm. 4.1)	$p \in (-\infty, \frac{1}{2}] \iff$ Ł with $c_* = M_p := \begin{cases} \frac{1}{1-p} & \text{for } p \leq \frac{1}{3} \\ \frac{p(7-12p)}{1-p} & \text{for } p \in [\frac{1}{3}, \frac{1}{2}] \end{cases}$
Hellinger-Kantorovich on $\mathcal{M}^+$ (Thm. 5.8)	• $p \in (-\infty, \frac{1}{2}] \implies$ Ł with $c_* = \frac{1}{1-p}$ • $p > \frac{1}{2} \implies$ there exists no with $c_* > 0$ • $p = 1$ and (LSI) $\implies$ No Ł; exp. decay is possible
Spherical Hellinger-Kantorovich on $\mathcal{P}$ (Thm. 4.4)	In general, decay rate $c_* = \max\{\alpha\, c_{\text{Ł-W}}\,,\ \beta\, M_p\}$ (see Thm. 4.4). Specifically: • $p \in (-\infty, \frac{1}{2}] \implies$ Ł with $c_* = \frac{1+2p}{1-p}$ • $\Omega \subsetneq \mathbb{R}^d$ bounded Lipschitz, $p \in (-\infty, 1/2] \cup [1-\frac{1}{d}, \infty)$ $\implies$ Ł with $c_* > 0$ • $\Omega = \mathbb{R}^d,\ p \in [1,2]$ and (Ł-W) $\implies$ Ł with $c_* = 2c_{\text{Ł-W}}$

Table 3: Summary of results for Łojasiewicz inequalities for φ_p-divergence energy functional $F(\mu) = \int \varphi_p(\mathrm{d}\mu/\mathrm{d}\pi)\,\mathrm{d}\pi$ in different dissipation geometries; see (7) for the definition of φ_p. The $p = 1$ case, the KL divergence, corresponds to the well-known logarithmic Sobolev inequality (LSI). Remarkably, using the (S)He or (S)HK geometry, the dimension restriction $p > 1-\frac{1}{d}$ in the Wasserstein setting can be circumvented.

Other related works While there are a few works analyzing gradient flows in the unbalanced transport geometry, such analysis, while valuable in its own rights, has not yet been able to capture the true strength of the (S)HK gradient flows that this paper showcases. In [LSW23], the focus is the regime under a uniform lower bound of the initial density ratio ($\mathrm{d}\mu_0/\ \mathrm{d}\pi$); see [LSW23, Theorem 2.3]. Various type of assumptions on the initial measure also exist in the literature, such as [DoP23, RJ*19]. From this paper's perspective, such characterizations are *local* and in contrast to this paper's *global* analysis. We also refer to [Chi22] for a different type of analysis where he shows that the HK gradient flows of certain functionals, under the assumption of a dense initialization condition, converge to states that satisfy a local Łojasiewicz inequality. A few recent works have also applied the spherical Hellinger-Kantorovich gradient flow with the KL divergence energy functional to practical statistical inference problems [YWR24, LC*22]. [GD*24] considered the sampling problem using the unbalanced transport gradient flow of the so-called maximum-mean discrepancy functional. They also exploited a Łojasiewicz type inequality to establish the convergence. Furthermore, this gradient flow is later shown by [Zhu24] to be a kernel approximation to the Hellinger-Kantorovich gradient flow of the forward KL divergence (i.e., φ_0-divergence). After the initial preprint version of this paper first appeared on the author's website on January 21, 2024 (`https://jj-zhu.github.io/file/ZhuMielke24AppKerEntFR.pdf`), the preprint [CC*24] appeared on `arXiv.org` on July 22, 2024. It contains an insightful but different analysis of the convergence of the pure spherical Hellinger flow (referred to as Fisher-Rao therein); see the discussion in Section 4.

Organization of the paper In Section 2, we provide background on gradient systems and optimal transport, with a focus on the dynamic formulation and geodesics. Section 3 is dedicated to the analysis of evolutionary behaviors in the gradient systems using the celebrated Polyak-Łojasiewicz inequalities. There, we introduce the standard log-Sobolev inequality for the pure Otto-Wasserstein gradient flow of the KL-divergence energy. Then, we provide novel results on the pure Hellinger gradient flows. In Section 4, we study the unbalanced transport gradient flows restricted to the probability measures, i.e., the spherical Hellinger-Kantorovich gradient flows. There, we were able to establish the global exponential decay of a large class of entropy functionals. In Section 5, we turn to the Hellinger-Kantorovich gradient flows over the positive measures, where the analysis of functional inequalities is more involved. Nonetheless, using a novel shape-mass decomposition in Section 5.3, we were able to establish the exponential decay when the LSI is not applicable. Additional proofs are given in Section A.

Notation We use the notation $\mathcal{P}(\bar\Omega), \mathcal{M}^+(\bar\Omega)$ to denote the space of probability and positive measures on the closure of a open set $\Omega \subset \mathbb{R}^d$ with Lipschitz boundary. The base space symbol Ω is often dropped if there is no ambiguity in the context. In this paper, the first variation of a functional F at $\mu \in \mathcal{M}^+$ is defined as a function $\frac{\delta F}{\delta \mu}[\mu]$

$$\frac{\mathrm{d}}{\mathrm{d}\epsilon} F(\mu + \epsilon \cdot v)|_{\epsilon=0} = \int \frac{\delta F}{\delta \mu}[\mu](x)\,\mathrm{d}v(x) \tag{10}$$

for any perturbation in measure v such that $\mu+\epsilon\cdot v \in \mathcal{M}^+$. The Fréchet (sub-)differential in a Banach space $(X, \|\cdot\|_X)$ is defined as a set in the dual space

$$\partial F(\mu) := \{\xi \in X^* \mid F(\nu) \geq F(\mu) + \langle \xi, \nu{-}\mu\rangle_X + o\left(\|\mu{-}\nu\|_X\right) \text{ for } \nu \to \mu\},$$

where the small-o notation signifies that the term vanishes more rapidly than the term inside the parentheses. When $\partial F(\mu)$ is a singleton, i.e., $\partial F(\mu) = \{\xi\} \subset X^*$, we simply write $\mathrm{D}F(\mu) := \xi \in X^*$. For simplicity, we carry out the Fenchel-conjugation calculation in the un-weighted L^2 space. Replacing that with duality pairing in the weighted L^2_ρ space does not alter the results. Common acronyms, such as partial differential equation (PDE) and integration by parts (IBP), are used without further specifications. We often omit the time index t to lessen the notational burden, e.g., the measure at time t, $\mu(t,\cdot)$, is written as μ. In formal calculation, we often use measures and their density interchangeably, i.e., $\int f \cdot \mu$ means the integral w.r.t. the measure μ. This is based on the standard rigorous generalization from flows over continuous measures to discrete measures [AGS05].

After publishing the first version of this manuscript, we got notice of very similar work in [KoV19, KoV20b, KoV20a].

2 Preliminaries

2.1 Gradient-flow systems and geodesics

Intuitively, a gradient flow describes a dynamical system that is driven towards the fastest dissipation of certain energy, through a geometric structure measuring dissipation. In this work, we restrict ourselves to the case that the dissipation law is linear and consequently can be given in terms of a (pseudo) Riemannian metric. Such a system is called a *gradient system*. For example, the dynamical system described by an ODE in Euclidean space, namely $\dot{u}(t) = -\nabla F(u(t))$, $u(t) \in \mathbb{R}^d$, is a simple gradient system.

In this paper, we take the perspective of variational modeling and principled mathematical analysis, i.e., we study the underlying dynamical systems modeled as gradient systems specified by the underlying space X, energy functional F, and the dissipation geometry specified by the *dissipation potential* functional $\mathcal{R}$. Given a smooth state space X, a dissipation potential is a function on the tangent bundle $\mathrm{T}X$, i.e. $\mathcal{R} = \mathcal{R}(u, \dot{u})$, where, for all $u \in X$, the functional $\mathcal{R}(u, \cdot)$ is non-negative, convex, lower semi-continuous, and satisfies $\mathcal{R}(u, 0) = 0$, see [Mie23] for more details and motivation. We denote by

$$\mathcal{R}^*(u, \xi) = \sup\left\{ \langle \xi, v\rangle - \mathcal{R}(u, v) \,\middle|\, v \in \mathrm{T}_u X \right\} \tag{11}$$

the (partial) Legendre transform of $\mathcal{R}$ and call it the *dual dissipation potential*. Throughout this work, we will only consider the case that $\mathcal{R}(u, \cdot)$ is quadratic, i.e.

$$\mathcal{R}(u, \dot{u}) = \frac{1}{2}\langle \mathbb{G}(u)\dot{u}, \dot{u}\rangle \quad \text{or equivalently} \quad \mathcal{R}^*(u, \xi) = \frac{1}{2}\langle \xi, \mathbb{K}(u)\xi\rangle.$$

Definition 2.1 (Gradient system) *A triple $(X, F, \mathcal{R})$ is called a generalized gradient system, if X is a manifold or a subset of a Banach space, $F : X \to \mathbb{R}$ is a differentiable functional, and $\mathcal{R}$ is a dissipation potential. The associated gradient-flow equation has the primal and dual form*

$$0 = \mathrm{D}_{\dot{u}}\mathcal{R}(u, \dot{u}) + \mathrm{D}F(u) \quad \Longleftrightarrow \quad \dot{u} = \mathrm{D}_\xi \mathcal{R}^*\big(u, -\mathrm{D}F(u)\big). \tag{12}$$

If $\mathcal{R}$ is quadratic, we simply call $(X, F, \mathcal{R})$ a gradient system *and obtain the gradient flow equations*

$$0 = \mathbb{G}(u)\dot{u} + \mathrm{D}F(u) \quad \Longleftrightarrow \quad \dot{u} = -\mathbb{K}(u)\mathrm{D}F(u). \tag{13}$$

$\mathbb{G} = \mathbb{K}^{-1}$ *is called the Riemannian tensor, and* $\mathbb{K} = \mathbb{G}^{-1}$ *is called the Onsager operator.*

Both forms of (12) and (13) have their advantages, but we will often use the form with $\mathcal{R}^*$ and $\mathbb{K}$, because they have an additive structure in the cases of interest.

Of particular interest to this paper is the gradient flow in the Hellinger space of positive measures $(\mathcal{M}^+, \mathsf{He})$, also called the *Hellinger-Kakutani* space (cf. [Kak48, LMS18, LaM19]), which is the gradient system that generates the following reaction equation as its *gradient flow equation* in the primal form of (12),

$$\partial_t \mu = -\mu \cdot \frac{\delta F}{\delta \mu}\,[\mu]\,. \tag{14}$$

This ODE is a consequence of the Hellinger dissipation geometry detailed in Example 2.4. Alternatively, one can also view the whole right-hand side as the Hellinger metric gradient induced by the Onsager operator $\mathbb{K}_{\mathsf{He}}(\mu)\xi = \mu \cdot \xi$.

The Hellinger gradient system is a special case of general gradient flows in metric spaces, which has gained significant attention in recent machine learning literature due to the study of the Otto-Wasserstein gradient flow, originated from the seminal works of Otto and colleagues, e.g., [Ott96, JKO98, Ott01]. Rigorous characterizations of general metric gradient systems have been carried out in PDE literature, for which we refer to [AGS05] for complete treatments and [San15, Pel14, Mie23] for a first-principles introduction. To get a concrete intuition, the gradient structure of the following two classical PDEs will become relevant in later discussions about Hellinger and Otto-Wasserstein respectively.

Example 2.2 (Classical PDE: Allen-Cahn and Cahn-Hilliard) *Recall the Allen-Cahn PDE*

$$\partial_t \rho = \Delta \rho - V'(\rho), \tag{15}$$

and the Cahn-Hilliard PDE

$$\partial_t \rho = \Delta \left(-\Delta \rho + V'(\rho) \right). \tag{16}$$

They are the gradient flows of the energy functional $F(\mu) = \int \big(\frac{1}{2}|\nabla \rho|^2 + V(\rho)\big)\,\mathrm{d}x$ *in two different Hilbert space geometries, where* V *is a potential function, e.g., the double-well potential* $V(r) = \frac{1}{4}(1-r^2)^2$*. The Allen-Cahn equation is the Hilbert-space gradient-flow equation of the energy* F *in unweighted* L^2*, i.e.,* $\mathbb{K}_{\mathrm{AC}} = 1$*, with dissipation potentials*

$$\mathcal{R}_{\mathrm{AC}}(\mu, \dot u) = \frac{1}{2}\|\dot u\|^2_{L^2} \quad \textit{and} \quad \mathcal{R}^*_{\mathrm{AC}}(\mu, \xi) = \frac{1}{2}\|\xi\|^2_{L^2}. \tag{17}$$

Cahn-Hilliard is the gradient flow of F *in unweighted* H^{-1}*, i.e.,* $\mathbb{K}_{\mathrm{AC}} = -\Delta$*, with dissipation potentials*

$$\mathcal{R}_{\mathrm{CH}}(\rho, \dot u) = \frac{1}{2}\|\dot u\|^2_{H^{-1}}, \quad \textit{and} \quad \mathcal{R}^*_{\mathrm{CH}}(\rho, \xi) = \frac{1}{2}\|\nabla\xi\|^2_{L^2}. \tag{18}$$

Geodesics and their Hamiltonian formulation. For many considerations of gradient flows, the geodesic curves play an important role. These curves are obtained as minimizers of the length of all curves connecting two points:

$$\gamma_{\mu_0\to\mu_1} \in \underset{\mu}{\operatorname{argmin}} \int_0^1 \langle \mathbb{G}(\mu(s))\dot\mu(s), \dot\mu(s)\rangle\,\mathrm{d}s \left(= \underset{\mu}{\operatorname{argmin}} \int_0^1 \langle \xi(s), \mathbb{K}(\mu(s))\xi(s)\,\mathrm{d}s \right) \\ \left(\text{ subject to } \dot\mu(s) = \mathbb{K}(\mu(s))\xi(s) \right), \tag{19}$$

where $s \mapsto \mu(s)$ has to be absolutely continuous, satisfy $\mu(0) = \mu_0$ and $\mu(1) = \mu_1$.

In the sense of classical mechanics, the dissipation potential $\mathcal{R}(\mu, \dot\mu) = \frac{1}{2}\langle \mathbb{G}(\mu)\dot\mu, \dot\mu\rangle$ plays the role of a "Lagrangian" $L(\mu, \dot\mu) = \mathcal{R}(\mu, \dot\mu)$, and the dual dissipation potential $\mathcal{R}^*(\mu, \xi) = \frac{1}{2}\langle \xi, \mathbb{K}(\mu)\xi\rangle$ as a "Hamiltonian" $H(\mu, \xi) = \mathcal{R}^*(\mu, \xi)$. Then, minimizing the integral of L is equivalent to solving the Hamiltonian system

$$\begin{cases} \dot\mu = \partial_\xi H(\mu, \xi) = \partial_\xi \mathcal{R}^*(\mu, \xi) = \mathbb{K}(\mu)\xi, \\ \dot\xi = -\mathrm{D}_\mu H(\mu, \xi) = -\mathrm{D}_\mu \mathcal{R}^*(\mu, \xi), \end{cases} \qquad \mu(0) = \mu_0,\ \mu(1) = \mu_1. \tag{H}$$

Here, the conditions for u at $s = 0$ and $s = 1$ indicate that finding geodesic curves leads to solving a two-point boundary value problem.

The theory for geodesics becomes particularly interesting in the case that $\mathcal{R}^*$ is affine in the state μ. Because, then, $\mathrm{D}_\mu \mathcal{R}^*(\mu, \xi)$ no longer depends on μ and the system (H) decouples in the sense that the equation for ξ no longer depends on μ. This particular case occurs in the Otto-Wasserstein, Hellinger, and consequently Hellinger-Kantorovich (a.k.a. Wasserstein-Fisher-Rao) space. This structure allows for the derivation of the following characterizations of the geodesic curves and static formulations of the associated Riemannian distances.

Example 2.3 (Otto-Wasserstein geodesics in Hamiltonian formulation) *In the case of the Otto-Wasserstein geometry, the dual dissipation potential takes the simple form*

$$\mathcal{R}^*_{\mathsf{Otto}}(\mu,\xi) = \frac{1}{2}\|\nabla\xi\|^2_{L^2_\mu} = \int \frac{1}{2}|\nabla\xi|^2 \mathrm{d}\mu.$$

The Onsager operator is given by $\mathbb{K}_{\mathsf{Otto}}(\mu)\xi = -\operatorname{div}(\mu\nabla\xi)$ *and the geodesic curves are characterized by*

$$\begin{cases} \dot\mu = -\operatorname{div}\left(\mu\nabla\xi\right), \\ \dot\xi = -\dfrac{1}{2}|\nabla\xi|^2. \end{cases} \tag{Geod-W}$$

Here, the first equation is the continuity equation which implies that μ *is transported along the vector field* $(t,x)\mapsto\nabla\xi(t,x)$*, and the second equation is the Hamilton-Jacobi equation, which is notably independent of* μ*. The Hopf-Lax formula then gives the explicit characterization of the solution*

$$\xi(s,x) = \inf_y \left\{\xi(0,y) + \frac{1}{2s}|x-y|^2\right\},$$

yielding the celebrated dual Kantorovich formulation of the Wasserstein distance. See [AGS05] for details.

The main focus of this paper is to study the Hellinger type gradient flows, generated by the geometry of the Hellinger distance between two nonnegative measures $\mu,\nu\in\mathcal{M}^+$,

$$\mathsf{He}^2(\mu_0,\mu_1) = 4\cdot\int\left(\sqrt{\frac{\delta\mu_0}{\delta\gamma}} - \sqrt{\frac{\delta\mu_1}{\delta\gamma}}\right)^2 \mathrm{d}\gamma \tag{20}$$

for a reference measure γ such that $\mu_0,\mu_1 << \gamma$. It is straightforward to show that this definition formally coincides with (9) with the precise scaling factor. A unique feature of the Hellinger distance (20) is that it allows the two measures to have disjoint supports in contrast to divergences such as KL and χ^2. We recall its dynamic formulation below using the reaction equation; see also [GaM17], [LMS18].

$$\mathsf{He}^2(\mu_0,\mu_1) = \min\left\{\int_0^1 \|\xi\|^2_{L^2_\mu}\,\mathrm{d}t \;\middle|\; \dot\mu = -\mu\cdot\xi,\ \mu(0)=\mu_0,\ \mu(1)=\mu_1\right\}. \tag{21}$$

If we add a correction term to the reaction dynamics, i.e., $\dot\mu = -\mu\cdot\left(\xi - \mu\cdot\int\mu\xi\right)$, we obtain the spherical Hellinger distance [LaM19] over the probability space $\mathcal{P}$, instead of positive measures $\mathcal{M}^+$.

$$\mathsf{SHe}^2(\mu_0,\mu_1) = \min\left\{\int_0^1 \|\xi\|^2_{L^2_\mu}\,\mathrm{d}t \;\middle|\; \dot\mu = -\mu\cdot\left(\xi - \int\xi\,\mathrm{d}\mu\right),\ \mu(0)=\mu_0,\ \mu(1)=\mu_1\right\}. \tag{22}$$

The spherical Hellinger distance, also termed the Bhattacharya distance by [Rao45b] after its first occurrence in [Bha42], can be calculated explicitly, namely

$$\mathsf{SHe}^2(\mu_0,\mu_1) = 4\arcsin\left(\frac{1}{4}\mathsf{He}(\mu_0,\mu_1)\right)$$

see [LaM19], but note the different scaling there. Subsequently, we refer to the above as the pure Hellinger and pure spherical Hellinger distances, i.e., without the transport aspect of Otto-Wasserstein.

Example 2.4 (Hellinger geodesics in Hamiltonian formulation) *For the Hellinger distance in* (21)*, the primal and dual dissipation potential takes the form*

$$\begin{aligned}\mathcal{R}_{\mathsf{He}}(\mu,\dot u) &= \frac{1}{2}\left\|\frac{\delta \dot u}{\delta\mu}\right\|^2_{L^2_\mu} = \frac{1}{2}\int\left|\frac{\mathrm{d}\dot u}{\mathrm{d}\mu}\right|^2\mathrm{d}\mu,\\ H(\mu,\xi) &= \mathcal{R}^*_{\mathsf{He}}(\mu,\xi) = \frac{1}{2}\big\|\xi\big\|^2_{L^2_\mu} = \int\frac{1}{2}\xi^2\mathrm{d}\mu,\end{aligned} \tag{23}$$

where $\frac{\mathrm{d}\dot u}{\mathrm{d}\mu}$ *denotes the Radon-Nikodym derivative between measures. The Onsager operator is given by* $\mathbb{K}_{\mathsf{He}}(\mu)\xi = \xi\mu$ *and the geodesic curves are characterized by*

$$\begin{cases}\dot\mu = -\mu\xi,\\ \dot\xi = -\frac{1}{2}|\xi|^2.\end{cases} \tag{Geod-FR}$$

Remarkably, different from the Hamilton-Jacobi setting of Otto-Wasserstein, this system can be solved in the explicit form

$$\xi(s,x) = \frac{\xi(0,x)}{1{+}s\xi(0,x)/2} \quad \textit{and} \quad \mu(s,\mathrm{d}x) = \big(1{+}s\xi(0,x)/2\big)^2\mu_0(\mathrm{d}x),$$

where we have used the initial condition $\mu(0)=\mu_0$*. Applying the terminal condition* $\mu(1)=\mu_1$*, we arrive at the explicit representation of the Hellinger geodesic*

$$\gamma_{\mu_0\to\mu_1}(s) = \big((1{-}s)\sqrt{\mu_0} + s\sqrt{\mu_1}\big)^2 = (1{-}s)^2\mu_0 + 2s(1{-}s)\sqrt{\mu_0\mu_1} + s^2\mu_1. \tag{24}$$

see [LMS16, Eqn. (2.8)] or [LaM19] for details. Finally, using the explicit solution for $\xi(s,x)$ *above, one can show that the Hellinger geodesic distance indeed admits the formula* (20)*. Formally, one can also obtain a static dual Kantorovich type formulation*

$$\frac{1}{2}\mathsf{He}^2(\mu_0,\mu_1) = \sup_{(2+\phi)(2-\psi)=4}\left\{\int\psi\,\mathrm{d}\mu_1 - \int\phi\,\mathrm{d}\mu_0\right\}.$$

Remark 2.5 ("Hellinger" versus "Fisher-Rao") *In the literature, the popular naming of "Fisher-Rao" has been used to describe the infinite-dimensional geometry over probability and positive measures. However, the name "Hellinger" distance was introduced after a paper of Kakutani in 1948 (based on his work [Hel09]) and has been used largely since the early 1960s, and even by Rao in 1963. We refer to [Mie24, Sec. 5] for some historical remarks. Nevertheless, starting from 2016 some authors such as [BBM16, GaM17, San17] used the name "Fisher-Rao" instead, and it is now very popular in imaging and machine learning. However, many such uses of the name "Fisher-Rao" are an abuse of the naming convention because it should be used in the sense of Rao's original definition in [Rao45a] as a way to characterize the distance of measures within a given submanifold of measures. Thus, the Fisher-Rao distance depends on the submanifold and is given by the length of the shortest curve within the submanifold, where length is measured in the Hellinger metric.*

In the present paper, the spherical Hellinger distance SHe *can be understood as a type of Fisher-Rao distance with respect to the submanifold* $\mathcal{P}(\Omega)$ *as a submanifold of* $\mathcal{M}^+(\Omega)$*. Another type of the Fisher-Rao distance occurs, for instance, if one chooses the submanifold of exponential family distributions.*

2.2 Unbalanced optimal transport: Hellinger-Kantorovich

As we have seen in the previous subsection, the Otto-Wasserstein geometry gives us the transport type dynamics, while the Hellinger geometry provides the birth-death, also reaction, mass creation or destruction, type dynamics. A few groups of researchers, including [CP*18a, CP*18b, LMS18, KMV16, GaM17], proposed the Hellinger-Kantorovich (HK) geometry, which is the combination of the Hellinger and Wasserstein distances. We refer to their works for the details and provide below a self-contained introduction to the HK geometry and gradient flow.

The optimal transport problem of Kantorovich must be generalized for the transport between measures of different mass to become admissible. The construction is as follows: In addition to the initial and target measures μ_0 and μ_1, one considers measures π_0 and π_1 between which classical optimal transport happens. Then, the mismatch between μ_0 and π_0 and between μ_1 and π_1 is penalized using a divergence functional Ψ, e.g., the KL divergence. This is then called *unbalanced transport*, defined using the *entropy-transport* functional

$$\mathsf{ET}_{c,\Psi}\left(\Pi|\mu_0,\mu_1\right) := \left\{ \int c(x_0,x_1)\,\mathrm{d}\Pi\left(x_0,x_1\right) + \Psi(\pi_0|\mu_0) + \Psi(\pi_1|\mu_1) \,\middle|\, \pi_0(\mathrm{d}x_0) := \Pi\left(\mathrm{d}x_0,\Omega\right),\ \pi_1(\mathrm{d}x_1) := \Pi\left(\Omega,\mathrm{d}x_1\right) \right\},$$

c is a cost function of transport, e.g., the squared Euclidean distance. In general, functionals defined using this type of inf-convolution do *not* generate a (squared) distance on $\mathcal{M}^+(\Omega)$. And even if it is a distance, it may not be a geodesic distance. It was the main achievement of [LMS16, LMS18] that the HK distance, defined as a geodesic distance in the sense of the dynamic Benamou-Brenier sense, via

$$\mathsf{HK}^2(\mu_0,\mu_1) = \min\left\{ \int_0^1 \alpha\|\nabla\xi\|^2_{L^2_\mu} + \beta\|\xi\|^2_{L^2_\mu}\,\mathrm{d}t \,\middle|\, \dot\mu = \alpha\,\mathrm{div}(\mu\cdot\nabla\xi) - \beta\mu\xi,\ \mu(0)=\mu_0,\ \mu(1)=\mu_1 \right\}, \tag{25}$$

can be characterized as an unbalanced transport problem as shown below, if c and Ψ are chosen in a very particular way. Different choices of $\alpha, \beta > 0$ allow us to tune the relative strength of the two geometries, trading off the transport and the birth-death mechanisms.

Theorem 2.6 (Logarithmic-Entropy-Transport definition of HK) [LMS16, Thm. 8] *The Hellinger-Kantorovich distance over positive measures $\mathcal{M}^+$ has the equivalent characterization as the optimal value of the* logarithmic-entropy-transport *(LET) problem*

$$\mathsf{HK}^2\left(\mu_0,\mu_1\right) := \inf_{\Pi\in\mathcal{M}^+(\Omega\times\Omega)} \mathsf{ET}_{c,\Psi}\left(\Pi|\mu_0,\mu_1\right), \tag{26}$$

where functional Ψ is the (scaled) KL divergence $\Psi(u|v) := \frac{1}{\beta}\,\mathrm{D}_{KL}(u|v)$ and the transport cost is

$$c(x_0,x_1) := \begin{cases} \dfrac{-2}{\beta}\log\left(\cos\left(\sqrt{\dfrac{\beta}{4\alpha}}\,|x_0-x_1|\right)\right) & \text{for } |x_0-x_1| < \pi\sqrt{\dfrac{\alpha}{\beta}}, \\ +\infty & \text{otherwise.} \end{cases}$$

Intuitively, the HK geometry combines the mechanisms of the Otto-Wasserstein geometry, i.e., the transport of mass, and the Hellinger geometry, i.e., the birth-death of mass. It possesses a richer structure and more advantageous properties than either of the pure geometries alone.

The gradient flow in the HK geometry generates the gradient flow equation, which is the following reaction-diffusion PDE.

Example 2.7 (Reaction-diffusion PDE) *The gradient-flow equation of the HK gradient system over positive measures $(\mathcal{M}^+, F, \mathsf{HK})$ corresponds to the reaction-diffusion PDE*

$$\dot\mu = -\alpha \cdot \mathbb{K}_{\mathsf{Otto}}(\mu)\,\frac{\delta F}{\delta\mu}[\mu] - \beta\cdot\mathbb{K}_{\mathsf{He}}(\mu)\,\frac{\delta F}{\delta\mu}[\mu] = \alpha\,\mathrm{div}\left(\mu\nabla\frac{\delta F}{\delta\mu}[\mu]\right) - \beta\mu\frac{\delta F}{\delta\mu}[\mu]\,. \tag{27}$$

The HK geometry and gradient flows are defined over the space of positive measures $\mathcal{M}^+$. For many machine learning applications, it is often more convenient to only work with probability measures. The restriction of the HK geometry to the space of probability measures $\mathcal{P}$ is discussed in [LaM19], referred to as the spherical Hellinger-Kantorovich (SHK) geometry. In this paper's context, we establish the following explicit formula.

Proposition 2.8 (Explicit formula for $\mathsf{SHK}_{\alpha,\beta}$) *For $\alpha, \beta > 0$ we have the formula for the spherical Hellinger-Kantorovich distance,*

$$\mathsf{SHK}_{\alpha,\beta}(\mu_0,\mu_1) = \frac{4}{\sqrt{\beta}}\,\arcsin\Big(\frac{\sqrt{\beta}}{4}\,\mathsf{HK}_{\alpha,\beta}(\mu_0,\mu_1)\Big). \tag{28}$$

Proof. In [LaM19], the passage from $\mathsf{HK}_{\alpha,\beta}$ to $\mathsf{SHK}_{\alpha,\beta}$ is discussed in detail by showing how the geodesics of $(\mathcal{P}(\Omega), \mathsf{SHK})$ and $(\mathcal{M}^+(\Omega), \mathsf{HK})$ can be transformed into each other. Under the assumption that $\beta = 4$, which is used in the scaling assumption (2.1) and (2.2) therein, it has been shown that

$$\mathsf{SHK}_{\alpha,4}(\mu_0,\mu_1) = \arccos\left(1 - \frac{1}{2}\,\mathsf{HK}_{\alpha,4}(\mu_0,\mu_1)^2\right) = 2\arcsin\left(\frac{1}{2}\mathsf{HK}_{\alpha,4}(\mu_0,\mu_1)\right),$$

where the first identity follows from [LaM19, Thm. 2.2] and the second from the trigonometric identity $\sin\sigma = \sqrt{(1-\cos(2\sigma))/2}$.

It now remains to apply the simple scaling $\mathsf{HK}^2_{\alpha,\beta} = \frac{4}{\beta}\mathsf{HK}_{4\alpha/\beta,4}$ and $\mathsf{SHK}^2_{\alpha,\beta} = \frac{4}{\beta}\mathsf{SHK}_{4\alpha/\beta,4}$, and the assertion follows. ■

Similarly, the spherical Hellinger distance $\mathsf{SHe} = \mathsf{SHK}_{0,1}$, also known as the Bhattacharya distance, is related to the Hellinger distance by

$$\mathsf{SHe}(\rho_0,\rho_1) = 4\arcsin\left(\frac{1}{4}\,\mathsf{He}(\rho_0,\rho_1)\right).$$

Recall our scaling of He in (9) with $\mathsf{He}(0,\mu) = 2\mu(\Omega)$, while some other works use $\widetilde{\mathsf{He}} = \mathsf{HK}_{0,4}$ giving $\widetilde{\mathsf{He}}(0,\mu) = \mu(\Omega)$. We also remind the reader of the use of the notation ρ for the probability measure instead of the positive measure μ.

The associated Onsager operator (inverse of the Riemannian metric tensor $\mathbb{G}_{\mathsf{SHe}}$) is given by restricting that of the Hellinger to the probability measures, namely

$$\mathbb{K}_{\mathsf{SHe}}(\rho)\eta = \beta\,\rho\Big(\eta - \int\eta\,\mathrm{d}\rho\Big). \tag{29}$$

Using that relation, we obtain the Onsager operator (inverse of the Riemannian metric tensor) for the spherical Hellinger-Kantorovich (SHK) geometry

$$\mathbb{K}_{\mathsf{SHK}}(\rho)\eta = -\alpha \operatorname{div}(\rho\nabla\eta) + \beta\rho\left(\eta - \int \rho\eta \,\mathrm{d}x\right),$$

and the SHK gradient flow equation

$$\dot\rho = -\mathbb{K}_{\mathsf{SHK}}(\rho)\frac{\delta F}{\delta\mu}[\mu] = \alpha \operatorname{div}\left(\rho\nabla\frac{\delta F}{\delta\mu}[\mu]\right) - \beta\rho\left(\frac{\delta F}{\delta\mu}[\mu] - \int \rho\frac{\delta F}{\delta\mu}[\mu]\,\mathrm{d}x\right). \tag{30}$$

3 Functional inequalities: Otto-Wasserstein and Hellinger

Functional inequalities are the building blocks for the analysis of many computational algorithms, such as for sampling and optimization over probability measures. The main goal of this section is to develop an intuition for the Łojasiewicz type inequalities for the Otto-Wasserstein and Hellinger type gradient-flow geometries.

3.1 Otto-Wasserstein gradient flow over probability measures $\mathcal{P}$

Our starting point is the differential *energy dissipation balance* relation of gradient flow systems,

$$\frac{\mathrm{d}}{\mathrm{d}t}F(\mu(t)) = \langle \mathrm{D}F, \dot\mu\rangle = -\Big(\mathcal{R}(\mu,\dot\mu) + \mathcal{R}^*(\mu, -\mathrm{D}F)\Big) =: -\mathcal{I}(\mu(t)). \tag{31}$$

where the functionals $\mathcal{R}$ and $\mathcal{R}^*$ are the primal and dual dissipation potentials discussed in Section 2.1. We refer to the quantity $\mathcal{I}$ as the *dissipation* of energy F. It was also referred to, in some contexts, as entropy production. The letter $\mathcal{I}$ is due to Fisher's information while the letter $\mathcal{R}$ is due to the Helmholtz-Rayleigh dissipation principle [Ray73]. From this, we introduce the following version of the Łojasiewicz condition. Note that, in the definition of the functional $\mathcal{I}$, it is understood that $\dot\mu$ is replaced by $\mathrm{D}_\xi\mathcal{R}^*(\mu, -\mathrm{D}F(\mu))$ to obtain a functional of μ alone. As we are in the quadratic case, we always have $\mathcal{I}(\mu) = \mathcal{R}^*(\mu, -\mathrm{D}F(\mu))$.

Definition 3.1 (Polyak-Łojasiewicz inequality for generalized gradient systems) *We say that the Polyak-Łojasiewicz inequality holds if*

$$\mathcal{R}(\mu,\dot\mu) + \mathcal{R}^*(\mu,-\mathrm{D}F) = \mathcal{I}(\mu) \ge c\cdot(F(\mu(t)) - F_*) \quad \textit{with } F_* = \inf_\mu F(\mu). \tag{Ł}$$

holds for some constant $c > 0$.

For conciseness, this paper does not analyze more general Łojasiewicz inequalities, i.e., no higher order powers on the right-hand side, due to the relevance of (Ł) to computational algorithms in machine learning and optimization; cf. [KNS20]. We simply refer to it as the Łojasiewicz inequality in the rest of the paper. We refer to articles such as [OtV00, BlB18] for a wider scope of related inequalities. An immediate consequence of (Ł) is that the energy of the gradient system converges exponentially via Grönwall's lemma, i.e.,

$$\text{(Ł)} \implies F(\mu(t)) - F_* \le \mathrm{e}^{-c\cdot t}\left(F(\mu(0)) - F_*\right).$$

Therefore, on the formal level, the intuition of the analysis is to produce the Łojasiewicz type relations in the succinct form of $\mathcal{I} \geq c \cdot (F - F^*)$.

Concretely, in the Otto-Wasserstein gradient flows and the Fokker-Planck PDEs, energy dissipation can be easily calculated

$$\mathcal{I}(\mu) = -\frac{\mathrm{d}}{\mathrm{d}t} F(\mu) \overset{\text{(along WGF)}}{=} \int \mu \left| \nabla \frac{\delta F}{\delta \mu} [\mu] \right|^2. \tag{32}$$

As an already well-known example, we now formally check the inequality (Ł) for the Otto-Wasserstein gradient system with the KL-divergence, i.e., $(\mathcal{P}(\mathbb{R}^d), \mathrm{D}_{\mathrm{KL}}(\cdot|\pi), W_2)$, where D_{KL} is defined in (5), (7). We calculate the dissipation

$$-\mathcal{I}(\mu) = \frac{\mathrm{d}}{\mathrm{d}t} \mathrm{D}_{\mathrm{KL}}(\mu|\pi) = \left\langle \log \frac{\mathrm{d}\mu}{\mathrm{d}\pi}, -\operatorname{div}\left(\mu \nabla \log \frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right) \right\rangle_{L^2} \overset{\text{(IBP)}}{=} -\left\| \nabla \log \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \right\|^2_{L^2_\mu}.$$

Specializing the Łojasiewicz inequality (Ł) to this setting, we arrive at the *logarithmic Sobolev inequality* (LSI)

$$\left\| \nabla \log \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \right\|^2_{L^2(\mu)} \geq c \cdot \mathrm{D}_{\mathrm{KL}}(\mu|\pi), \tag{LSI}$$

which needs to hold for some $c > 0$. By Grönwall's lemma, the entropy decays exponentially, i.e., $\mathrm{D}_{\mathrm{KL}}(\mu|\pi) \leq \mathrm{e}^{-c \cdot t} \mathrm{D}_{\mathrm{KL}}(\mu(0)\|\pi)$. (LSI) is a special case of the (Polyak)-Łojasiewicz inequality for the Otto-Wasserstein geometry and the more general φ-divergence energies, namely

$$\left\| \nabla \varphi' \left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi} \right) \right\|^2_{L^2_\mu} \geq c \cdot \mathrm{D}_\varphi(\mu|\pi). \tag{Ł-W}$$

In particular, we will exhaustively investigate the φ_p-divergence energy functional case. The inequality (Ł-W) reads

$$\frac{1}{(p-1)^2} \int \mu \left| \nabla \left(\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi} \right)^{p-1} \right) \right|^2 \mathrm{d}x \geq c \cdot \mathrm{D}_{\varphi_p}(\mu|\pi) \tag{33}$$

For $p = 1$, i.e., the choice of φ_{KL} (φ_1-divergence or the 1-relative entropy), recovers the (LSI), which has already been intensely investigated in the literature. The Bakry-Émery theorem [BaÉ85] gives a sufficient condition for the logarithmic Sobolev inequality (LSI) to hold along the solution of the Fokker-Planck equations: the target probability measure π satisfies the Bakry-Émery condition, if $\pi \propto \exp(-V)$ for the potential function V that satisfies

$$\nabla^2 V \geq c_{\mathrm{BE}} \cdot \mathrm{Id},\ c_{\mathrm{BE}} > 0. \tag{BE}$$

Moreover, following [BaÉ85], [AM*01] provided an elementary proof of the Bakry-Émery theorem for general φ-divergence energies that satisfies

$$\varphi(1) = \varphi'(1) = 0,\ \varphi''(1) > 0 \quad \text{and} \quad (\varphi'''(s))^2 \leq \frac{1}{2} \varphi''(s) \varphi^{(4)}(s). \tag{34}$$

Their results state that if (34) holds, the Otto-Wasserstein gradient flow with the corresponding φ-divergence energy converges exponentially. That is, the following sufficient relation holds

$$\text{(BE)} + \text{(34)} \implies \text{(Ł-W)} : \text{Łojasiewicz for Otto-Wasserstein} \implies \text{exp. decay}. \tag{35}$$

First, we slightly modify this result for the φ_p-divergence energy functional and the case of domain $\Omega = \mathbb{R}^d$. The proof is straightforward by plugging in the definition of the φ-divergence into (34) and using the relation (35).

Theorem 3.2 (Functional inequality for pure Otto-Wasserstein: $\mathbb{R}^d$) *Suppose the Bakry-Émery condition* (BE) *holds for the target measure π. Then, under the φ_p-divergence energy for $p \in [1,2]$, the Łojasiewicz inequality* (Ł-W) *holds for the Otto-Wasserstein gradient flow with the constant $2c_{BE}$.*

In addition, when the domain is an open and bounded subset of $\mathbb{R}^d$, we no longer need Bakry-Émery or LSI type conditions when working with sufficiently smooth measures.

Theorem 3.3 (Functional inequality for pure Otto-Wasserstein: bounded domain) *Assume that $\Omega \subset \mathbb{R}^d$ is an open and bounded Lipschitz domain and that $\pi \in \mathrm{L}^\infty(\Omega)$ is bounded from below by a positive constant. Then, for all $p \geq 1 - \frac{2}{d}$ there exists a positive constant $c > 0$ such that the Łojasiewicz inequality* (33) *holds for all sufficiently smooth measure $\mu \in \mathcal{P}(\Omega)$.*

Proof. For the case $\pi = c_0 \cdot \mathrm{d}x$, the result is established in [MiM18, Sec. 3] as well as the master thesis of the second author. The general case follows by estimating π from above and from below and by applying the result to $r = \frac{\mathrm{d}\rho}{\mathrm{d}\pi}$. ■

The important question lingering is when and if the Łojasiewicz inequality holds for other gradient flows and other energy functionals, i.e., a theory mirroring the Bakry-Émery results but going beyond the standard Otto-Wasserstein geometry. Our starting point is replacing the Otto-Wasserstein geometry of the gradient flows with the Hellinger geometry.

3.2 Hellinger gradient flow over $\mathcal{M}^+$

By a derivation similar to the Otto-Wasserstein setting, we find the energy dissipation for the Hellinger gradient flow

$$\mathcal{I}(\mu) = -\frac{\mathrm{d}}{\mathrm{d}t} F(\mu(t)) \overset{\text{(along HeGF)}}{=} \int \mu \left| \frac{\delta F}{\delta \mu}[\mu] \right|^2. \tag{36}$$

In the settings other than Otto-Wasserstein gradient flow, however, the Łojasiewicz inequality (Ł) cannot be expected to hold globally for arbitrary geometry in general. We now show that this is precisely the case for Hellinger. Consider the Hellinger gradient flow with the KL entropy energy functional, i.e., $F(\mu) = \mathrm{D}_{\mathrm{KL}}(\mu|\pi)$. Then, the specialized Łojasiewicz inequality asks for the existence of some $c > 0$ such that

$$\left\| \log \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \right\|^2_{L^2_\mu} \geq c \cdot \mathrm{D}_{\mathrm{KL}}(\mu \| \pi). \tag{37}$$

Lemma 3.4 (No global Łojasiewicz condition in Hellinger flows of KL) *There exists no $c > 0$ such that* (37) *holds globally for positive measures $\mu \in \mathcal{M}^+$, i.e., the gradient system $(\mathcal{M}^+, \mathrm{D}_{\mathrm{KL}}(\cdot|\pi), \mathrm{He})$ does not satisfy the global Łojasiewicz condition for any positive constant.*

For a counter-example, consider $\mu_r = r\pi$ in (37). Then we have $\mathrm{D}_{\mathrm{KL}}(\mu_r|\pi) \to \mu_r(\Omega)$ for $r \to 0^+$, but $\| \log r \|_{L^2_{\mu_r}} \| = r \log^2 r \cdot \pi(\Omega) \to 0$. See Figure 3 and the caption for an illustration of Lemma 3.4. Despite this lack of the global Łojasiewicz condition in general, a local condition can be satisfied trivially around the equilibrium measure $\mu = \pi$. However, from this paper's perspective, we are not interested in the local version for the reason explained next.

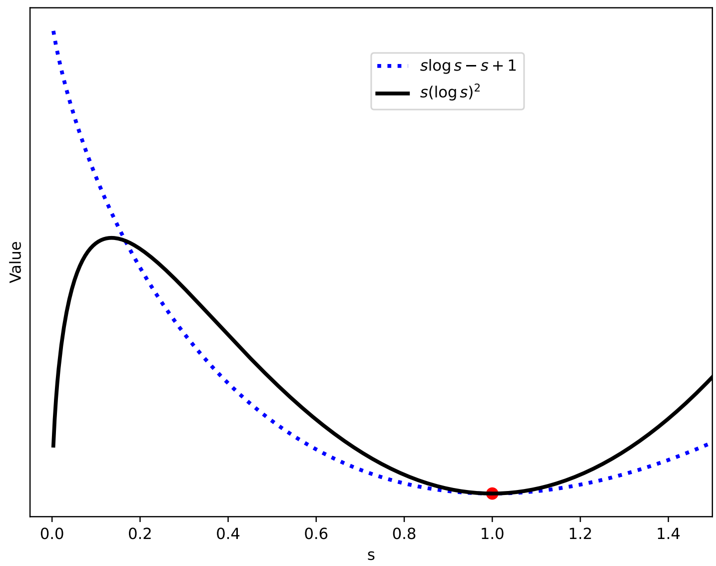

Figure 3: The plot illustrates the lack of global Łojasiewicz inequality as in Lemma 3.4. We plot the KL-entropy generator function $\varphi(s) = s\log s - s + 1$. The blue dotted curve represents the KL-entropy generator $\varphi(s)$. The function $s|\log s|^2$ is plotted in solid black. The Łojasiewicz inequality condition is satisfied locally around the equilibrium $s = 1$ (red dot). However, it can never be satisfied in a neighborhood around $s = 0$.

Example 3.5 (Birth escaping zero in the Hellinger geometry) *Suppose we wish to minimize the energy $F(\mu) = \mathrm{D}_\varphi(\mu|\pi)$ starting from the initial measure μ_0. It is possible that the measure μ_0 does not have the full support as the target measure π; see Figure 4 (left), i.e., $\operatorname{supp}(\mu_0) \subsetneq \operatorname{supp}(\pi)$. In addition, many variational inference methods, e.g., [KhN18, LC*22] use Gaussian densities to approximate the target measure. In such cases, the measures share the support as in Figure 4 (right), i.e., $\operatorname{supp}(\mu_0) = \operatorname{supp}(\pi)$, but the density ratio can be arbitrarily close to zero. For example, Figure 4 (right) depicts a Gaussian mixture distribution as the initial μ_0 that has very little mass near $x = 2$. This can be quite likely in high dimensions. The most difficult part of the minimization is to* escape the near-zero region with enough metric slope provided by the energy. *For example, the reaction dynamics $\dot\mu = -\mu\frac{\delta F}{\delta\mu}[\mu]$ implies that a significant growth field is needed to escape when μ is near zero, i.e., the* birth process*. Our theory precisely characterizes this escape threshold via the global Łojasiewicz condition, e.g., in Corollary 3.8. In contrast, the local convergence behavior near the equilibrium is much easier to capture; see Figure 3, Figure 5. Therefore, we place the focus of our analysis on the global Łojasiewicz condition without delving into local equilibrium behavior. Finally, we note that our analysis is for general positive measures. The submanifolds of parameterized probability distributions, such as Gaussian densities (i.e. Fisher-Rao geometry), are not considered in this paper.*

While the above results show that the Hellinger flow of the KL-entropy cannot satisfy the global Łojasiewicz, we now show a positive result for the case when the energy functional is the squared Hellinger distance, $F(\mu) = \frac{1}{2}\mathsf{He}^2(\mu,\pi) = \int\left(\sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\pi}} - 1\right)^2 \mathrm{d}\pi$. First, note that the first variation of the squared Hellinger distance is $\dfrac{\delta}{\delta\mu}\left(\dfrac{1}{2}\mathsf{He}^2(\mu,\pi)\right) = f'(\mu) = 2 - 2\sqrt{\mathrm{d}\pi/\mathrm{d}\mu}$. Specializing the

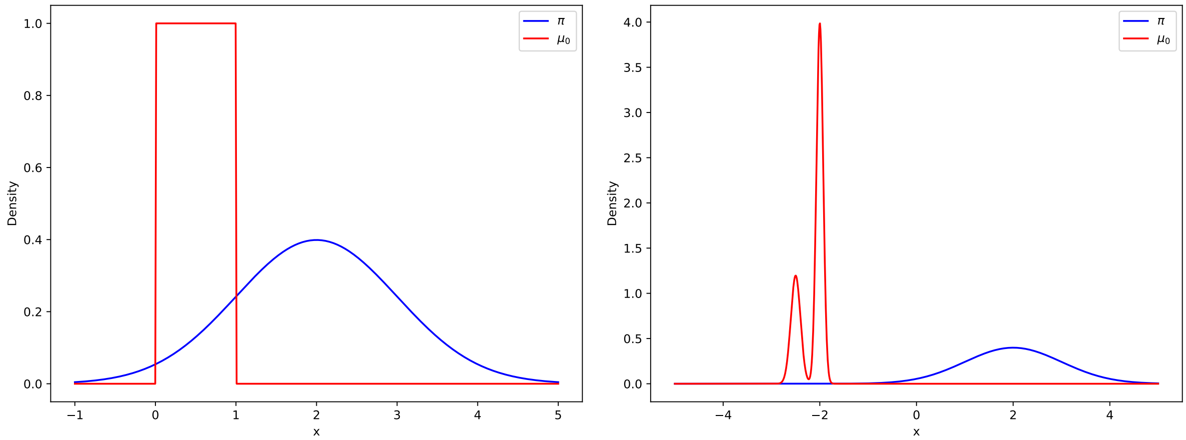

Figure 4: Illustration of Example 3.5: birth escaping (near) zero with initial densities μ_0 (red) and target densities π (blue, Gaussian).

Łojasiewicz inequality to this setting requires

$$4 \cdot \left\| 1 - \sqrt{\frac{\mathrm{d}\pi}{\mathrm{d}\mu}} \right\|_{L^2_\mu}^2 \geq \frac{c}{2} \mathsf{He}^2(\mu|\pi). \tag{38}$$

It can be easily checked by definition that we have the unconditional satisfaction of the global Łojasiewicz inequality in this case:

Lemma 3.6 (Global Łojasiewicz with Hellinger energy) *The Łojasiewicz inequality* (38) *holds for the Hellinger gradient system* $(\mathcal{M}^+, \frac{1}{2}\mathsf{He}^2(\cdot, \pi), \mathsf{He})$ *globally for* $c = 2$.

Going beyond the Hellinger energy, we are now ready to extract some general principles. The natural question is whether relations such as Bakry-Émery and (LSI) exist for the Hellinger geometry. To answer that, we first establish the condition for global Łojasiewicz condition for the class of φ_p-divergence energy (5). We observe that $\mathrm{D}_{\varphi_p}(\mu|\pi) \geq 0$ with equality if and only if $\mu = \pi$ (in the sense of measures). Moreover, $\mu \mapsto \mathrm{D}_{\varphi_p}(\mu|\pi)$ is convex, and the Fréchet subdifferential is given by

$$\partial \mathrm{D}_{\varphi_p}(\mu|\pi) = D\mathrm{D}_{\varphi_p}(\mu|\pi) = \varphi_p' \left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi} \right) = \frac{1}{p{-}1} \left(\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi} \right)^{p-1} - 1 \right), \quad D\mathrm{D}_{\varphi_1}(\mu|\pi) = \log \left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi} \right).$$

Proposition 3.7 (Global Łojasiewicz for Hellinger gradient flow of relative entropy) *Given the Hellinger gradient system with* φ*-divergence energy, i.e.,* $(\mathcal{M}^+, \mathrm{D}_\varphi(\cdot|\pi), \mathsf{He})$*. If* $\varphi : (0, \infty) \to [0, \infty)$ *is a convex entropy generator function satisfying*

$$\varphi(1) = \varphi'(1) = 0, \varphi''(1) > 0 \quad \textit{and} \quad \exists\, c_* > 0 \textit{ such that} \,\forall\, s > 0 : \ s\big(\varphi'(s)\big)^2 \geq c_* \varphi(s), \tag{39}$$

then the Łojasiewicz inequality holds globally, i.e.,

$$\left\| \varphi' \left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi} \right) \right\|_{L^2_\mu}^2 \geq c_* \, \mathrm{D}_\varphi(\mu|\pi). \tag{Ł-He}$$

Proof of Proposition 3.7. As previously calculated, the first variation of the φ-divergence is given by $\frac{\delta}{\delta\mu}\mathrm{D}_\varphi(\mu|\pi) = \varphi'\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right)$. Thus, using the Hellinger metric, we obtain the dissipation relation $\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{D}_\varphi(\mu|\pi) = -\mathcal{I}(\mu)$ with

$$\mathcal{I}(\mu) = \left\|\varphi'\left(\tfrac{\mathrm{d}\mu}{\mathrm{d}\pi}\right)\right\|^2_{L^2_\mu} = \int_\Omega \left(\varphi'\left(\tfrac{\mathrm{d}\mu}{\mathrm{d}\pi}\right)\right)^2 \mathrm{d}\mu = \int_\Omega \left(\varphi'\left(\tfrac{\mathrm{d}\mu}{\mathrm{d}\pi}\right)\right)^2 \tfrac{\mathrm{d}\mu}{\mathrm{d}\pi}\,\mathrm{d}\pi.$$

Now exploiting the assumption (39) for estimating the integrand, we immediately obtain (Ł-He). ■

Because of the simple point-wise estimate in the above proof, it is also clear that condition (39) is *necessary and sufficient* for the Łojasiewicz estimate (Ł-He).

Corollary 3.8 (Hellinger gradient flows: necessary sufficient condition) *The Łojasiewicz inequality* (Ł-He) *for the Hellinger gradient system with the power-like entropy* φ_p (7) *energy,* $(\mathcal{M}^+, \mathrm{D}_{\varphi_p}, \mathsf{He})$*, holds globally if and only if* $p \le \frac{1}{2}$*. Furthermore, the constant is* $c_* = 1/(1-p)$ *in that case.*

Therefore, the Hellinger gradient flow under the φ_p*-divergence energy functional decays exponentially globally, i.e.,*

$$\mathrm{D}_{\varphi_p}(\mu(t)|\pi) \le \mathrm{e}^{-\frac{t}{(1-p)}} \cdot \mathrm{D}_{\varphi_p}(\mu(0)|\pi)$$

if and only if $p \le \frac{1}{2}$*.*

This decay result is also referred to as global exponential convergence in energy. In short, for the φ-divergence energy functional,

$$\text{(39)} \iff \text{(Ł-He)} \implies \text{exp. decay}$$

In particular, Corollary 3.8 shows that the energy functionals, for which the globally Łojasiewicz estimate holds, include the squared Hellinger ($p = \frac{1}{2}$), the forward KL ($p = 0$), the reverse χ^2 ($p = -1$), and the fractional-power entropies between those. On the negative side, it states that the Łojasiewicz estimate does not hold globally for many commonly used entropy functionals such as the KL ($p = 1$) and χ^2 ($p = 2$).

Remark 3.9 (Metric slope and entropy power threshold $p = \frac{1}{2}$**)** *The relevance of the threshold* $p = 1/2$ *can be seen from two perspectives. First, we observe that* $\mu = 0$ *is a steady-state solution for the gradient systems* $\big(\mathcal{M}^+, \mathrm{D}_{\varphi_p}(\cdot|\pi), \mathsf{He}\big)$ *for* $p > 1/2$*. However, if* $\mu(t) = 0$ *is a solution, then it cannot converge exponentially to the equilibrium measure* π*. The point is that the* Hellinger metric slope*, defined as*

$$|\partial \mathrm{D}_{\varphi_p}|_{\mathsf{He}}(0) := \limsup_{\mu\to 0} \frac{\big(\mathrm{D}_{\varphi_p}(0) - \mathrm{D}_{\varphi_p}(\mu)\big)_+}{\mathsf{He}(0,\mu)},$$

can be calculated explicitly as the following:

Lemma 3.10 *The Hellinger metric slope of the* φ_p*-divergence energy functional at* $\mu = 0$ *is given by*

$$|\partial \mathrm{D}_{\varphi_p}|_{\mathsf{He}}(0) = \begin{cases} 0 & \text{for } p > 1/2, \\ 1 & \text{for } p = 1/2, \\ \infty & \text{for } p < 1/2. \end{cases}$$

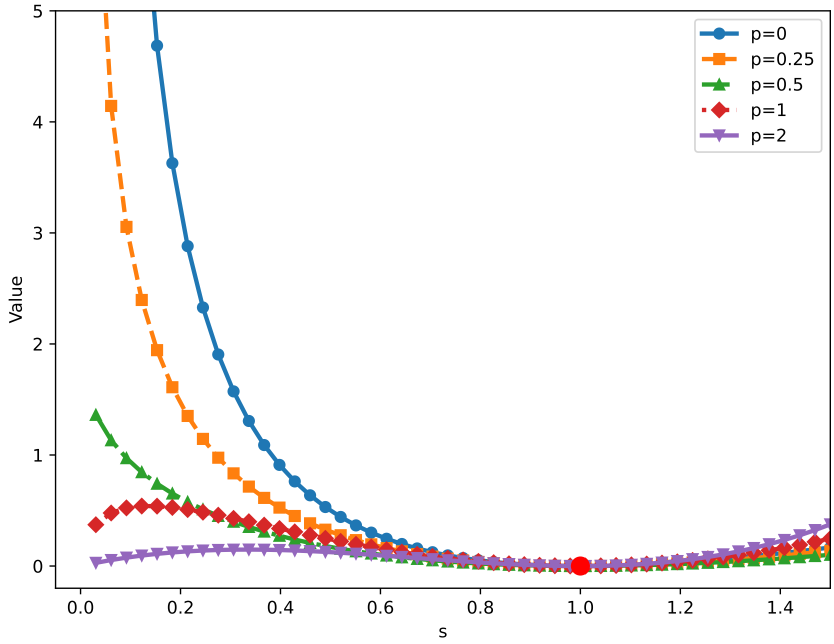

Figure 5: The plot illustrates the left-hand side $s({\varphi_p}'(s))^2$ of the Łojasiewicz inequality (39) for the Hellinger geometry for $s \in [0, 1.2]$ and different p: purple $p = 2$ (χ^2), red $p = 1$ (KL), green $p = 0.5$ (Hellinger), orange $p = 0.25$, blue $p = 0$ (forward KL), The red dot represents the equilibrium at $s = 1$, where $\phi'(s) = 0$. This plot provides insights into the slopes of the power-like entropies in the Hellinger gradient flow. Indeed, Proposition 3.7 discusses the relation of the corresponding curves $\varphi_p(s)$ in Figure 1 with those here. We observe the threshold $p = 0.5$ (Hellinger; green) where the behavior near $s = 0$ jumps. See the main text, especially Remark 3.9, for analysis.

In the case $p > 0$ where $\mathrm{D}_{\varphi_p}(0) < \infty$ the curve $t \mapsto \mu(t) = 0$ can still be considered a solution of the gradient-flow equation, however, the exponential decay only applies to the curves of maximal slopes (see, e.g., [AGS05]) satisfying the dissipation balance

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{D}_{\varphi_p}(\mu(t)|\pi) = -\frac{1}{2}|\mu'|_{\mathsf{He}}(t)^2 - \frac{1}{2}|\partial \mathrm{D}_{\varphi_p}|_{\mathsf{He}}(\mu(t))^2.$$

We refer to [LaM23, Section 2] for a more detailed discussion.

A second way to see the importance of the threshold $p \leq \frac{1}{2}$ involves the results in [OtV00] showing that geodesic Λ-convexity of a functional implies the Łojasiewicz inequality with $c_Ł = 2\Lambda$. This is analogous to the finite-dimensional case in [KNS20]. For the condition of geodesic Λ-convexity for functionals $\mathrm{D}_\varphi(\mu|\pi) = \int_\Omega \varphi(\frac{\mathrm{d}\mu}{\mathrm{d}\pi})\,\mathrm{d}\pi$ in the Hellinger geometry, it can be shown that $\Lambda := \inf_{w\geq 0}\Big\{w\,\varphi''(w) + \frac{1}{2}\,\varphi'(w)\Big\}$. This gives the same result when considering the p-power family φ_p. But for general φ, we may have $2\Lambda \lneqq c_Ł$. The geodesic Λ-convexity in the HK geometry has been established in [LMS23a, Theorem 7.2]. However, one can only obtain a Łojasiewicz type inequality with constant zero by directly applying the results in [LMS23a], which is not sufficient for exponential convergence.

3.2.1 Explicit solution of the Hellinger gradient flow equation

To further characterize the phenomena regarding the HK gradient flow mathematically, we now delve deeper into the gradient-flow equation for the Hellinger gradient systems $\big(\mathcal{M}^+, F, \mathsf{He}\big)$ with $F =$

$\mathrm{D}_{\varphi_p}(\cdot|\pi)$, which is the reaction equation

$$\dot{\mu} = -\beta \cdot \mu \, \frac{\delta \mathrm{D}_{\varphi_p}(\mu|\pi)}{\delta \mu}[\mu] = -\beta \cdot \mu \, \varphi_p'(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}). \tag{40}$$

A delicate situation arises when considering the gradient flow equation (40): in general, not all solutions of (40) will converge to the desired equilibrium π. The reason is the degeneracy of the Hellinger Onsager operator $\mathbb{K}_{\mathsf{He}}(\mu)\xi = \mu \cdot \xi$ at $\mu = 0$.

Example 3.11 (Hellinger gradient flow of KL) *Taking the driving energy to be the KL divergence ($p = 1$) in* (40)*, we obtain the gradient flow equation*

$$\dot{\mu} = -\beta\mu \, \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right). \tag{KL-He}$$

This ODE can be explicitly solved with elementary arguments, yielding the following result.

Proposition 3.12 *The ODE* (KL-He) *admits the unique solution, for all $x \in \Omega$ and $t \geq 0$,*

$$\mu(t,x) = \pi(x)\Big(\frac{\mathrm{d}\mu(0,\cdot)}{\mathrm{d}\pi}(x)\Big)^{\mathrm{e}^{-\beta t}} \tag{41}$$

Furthermore, we have $\mu(t,x) \to \pi(x)$ as $t \to \infty$ if and only if $\mu(0,x) \gneqq 0$.

In other words, for some location x' with zero initial density $\mu(0,x') = 0$, the solution gets stuck and no new mass is born. This precisely corresponds the illustration in Figure 3. In addition to the KL divergence functional ($p = 1$), similar problems occur for the Hellinger gradient flow of the D_{φ_p}-divergence with $p \in (0,1)$, because solutions starting with $\mu(0,x) = 0$ may satisfy $\mu(t,x) = 0$ for $t \in [0,\tau(x)]$ and $\mu(t,x) > 0$ for $t > \tau(x)$, where $\tau(x)$ can be chosen arbitrarily. However, for the interesting case of φ_p-divergence with $p \leq 1/2$, the notion of *curves of maximal slope* selects the unique solution with $\mu(t,x) > 0$ for $t > 0$.

3.2.2 Exponential decaying Lyapunov functions for Hellinger gradient flows

Clearly, the driving energy $\mathrm{D}_{\varphi_p}(\,\cdot\,|\pi)$ itself decays along solutions because it is the driving energy of the gradient system. Furthermore, an examination of the simple structure of the gradient flow equation (40) implies that $\dot{\rho} \geq 0$ for $\frac{\mathrm{d}\rho}{\mathrm{d}\pi} \in [0,1]$ and $\dot{\rho} \leq 0$ for $\frac{\mathrm{d}\rho}{\mathrm{d}\pi} \geq 1$. Hence, the divergence functional $\mathrm{D}_{\varphi_q}(\,\cdot\,|\pi)$ with any q is non-increasing along solutions. We have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{D}_{\varphi_q}(\mu(t)|\pi) = -\beta \int_\Omega \underbrace{\mu\varphi_p'\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right)\varphi_q'\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right)}_{\geq 0} \mathrm{d}x \;\leq\; 0,$$

i.e., $\mathrm{D}_{\varphi_q}(\,\cdot\,|\pi)$ is a Lyapunov functional for the Hellinger gradient flow of the φ_p-divergence.

We now show that the divergence $\mathrm{D}_{\varphi_q}(\cdot|\pi)$, for some $q \neq p$, decays exponentially along the gradient flow solutions. Set an auxiliary constant depending on p and q as

$$m_{p,q} := \inf\Big\{ \frac{r\varphi_p'(r)\varphi_q'(r)}{\varphi_q(r)} \;\Big|\; r > 0 \text{ and } r \neq 1 \Big\} \geq 0.$$

Proposition 3.13 (Lyapunov functionals for $(\mathcal{M}^+, \mathrm{D}_{\varphi_p}, \mathrm{He})$) *For $p, q \in \mathbb{R}$, we have*

$$m_{p,q} \gneqq 0 \iff p \le \max\{0, \min\{1, 1-q\}\}.$$

Assume that the initial condition $\mu(0)$ satisfies $\mathrm{D}_{\varphi_q}(\mu(0)|\pi) < \infty$ and $m_{p,q} > 0$. Then D_{φ_q} decays exponentially along the solutions of the gradient flow for $(\mathcal{M}^+, D_{\varphi_p}(\cdot|\pi), \mathrm{He})$, namely

$$\mathrm{D}_{\varphi_q}(\mu(t)|\pi) \le \mathrm{e}^{-\beta m_{p,q} t} \mathrm{D}_{\varphi_q}(\mu(0)|\pi) \quad \textit{for } t \ge 0.$$

That is, the φ_q-divergence is an exponentially decaying Lyapunov functional for the Hellinger gradient flow of the φ_p-divergence.

Proof. The technical characterization of the region with $m_{p,q} > 0$ is given in Lemma A.1.

For the decay estimate we simply observe that

$$\begin{aligned} -\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{D}_{\varphi_q}(\mu(t)|\pi) &= \beta \int_\Omega \varphi_p'\big(\tfrac{\mu}{\pi}\big)\, \varphi_q'\big(\tfrac{\mu}{\pi}\big)\, \mathrm{d}\mu = \beta \int_\Omega \varphi_p'\big(\tfrac{\mu}{\pi}\big)\, \varphi_q'\big(\tfrac{\mu}{\pi}\big)\, \tfrac{\mu}{\pi}\, \mathrm{d}\pi \\ &\ge \int_\Omega m_{p,q} \varphi_q\big(\tfrac{\mu}{\pi}\big)\, \mathrm{d}\pi = \mathrm{D}_{\varphi_q}(\mu(t)). \end{aligned}$$

Now, the desired result follows by Grönwall's estimate. ■

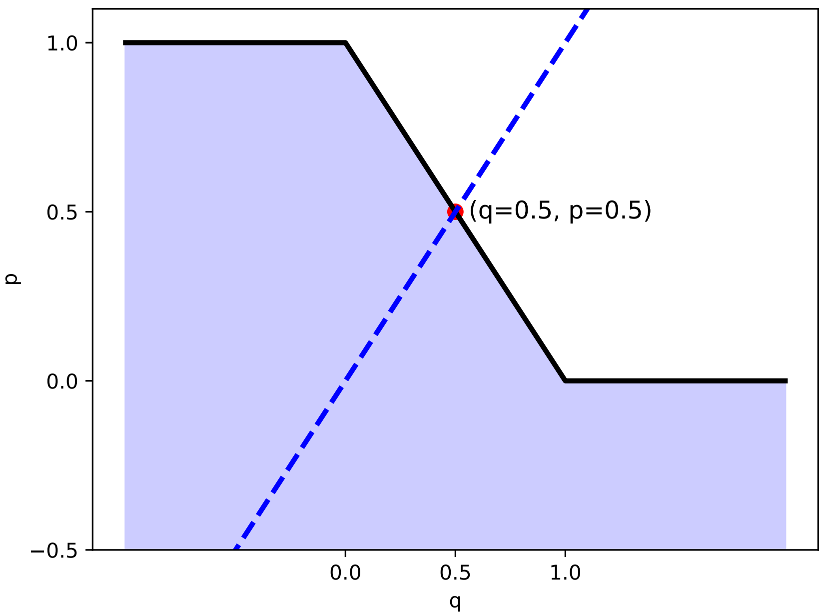

Figure 6: The plot illustrates the p and q values that satisfy the condition $p \le \max\{0, \min\{1, 1-q\}\}$. The shaded area represents the region where $p \le \max\{0, \min\{1, 1-q\}\} \iff m_{p,q} > 0$, i.e., the φ_q-divergence converges exponentially for the Hellinger gradient flow of the φ_p-divergence. See Proposition 3.13 and Lemma A.1 for the details. Furthermore, we observe that the shaded area contains the part of the line $p = q$ for $q \le \frac{1}{2}$. This shows that Proposition 3.13 generalizes the result in Corollary 3.8 to exponentially decaying Lyapunov functionals. In this case, when $q > \frac{1}{2}$, the intersection is empty and hence our result no longer implies D_{φ_p} itself decays exponentially.

The proof above and Lemma A.1 can further explain why it is easier to have decay estimates for the divergences $\mathrm{D}_{\varphi_q}(\mu(t)|\pi)$ with $q\leq 0$. Since, if $q>0$, the condition $\mathrm{D}_{\varphi_q}(\mu(0)|\pi)<\infty$ will impose strict positivity of the density $\mu(0,x)>0$ a.e. with respect to π. For example, the forward KL divergence ($q=0$) does not impose this condition.

Example 3.14 (Lypunov functional for Hellinger gradient flows of KL) *We again consider the driving energy of a Hellinger gradient flow to be the KL divergence ($p=1$). Due to Proposition 3.13, the forward KL ($q=0$) divergence, $\mathrm{D}_{\varphi_0}(\cdot|\pi)=\mathrm{D}_{KL}(\pi|\cdot)$, and the reverse χ^2 ($q=-1$) divergence, $\mathrm{D}_{\varphi_{-1}}(\cdot|\pi)=\mathrm{D}_{\chi^2}(\pi|\cdot)$, are both exponentially decaying Lyapunov candidates for this system, i.e., they decay exponentially along the Hellinger gradient flow of the KL divergence, given the finite initialization condition in Proposition 3.13. See also Figure 6 for the relation between p and q.*

4 Spherical HK gradient flows of probability measures $\mathcal{P}$

Our main goal of this section is to advance the state-of-the-art analysis for the spherical Hellinger-Kantorovich (a.k.a., spherical Wasserstein-Fisher-Rao) space and gradient flows of probability measures $\mathcal{P}$. Its properties differ from the HK (a.k.a. WFR) geometry over positive measures $\mathcal{M}^+$. Remarkably, we are able to establish a global Polyak-Łojasiewicz inequality for the φ_p-divergence energy when $p\in(-\infty,\frac{1}{2}]\cup[1,\infty)$, which showcases the advantages of the SHK geometry over the pure Otto-Wasserstein and Hellinger geometries. This is due to the flexibility of SHK by combining the strengths of the Otto-Wasserstein and the spherical Hellinger geometries over probability measures. We note that a detailed and insightful analysis of the gradient flow for $(\mathcal{M}^+(\Omega),\mathrm{D}_\varphi(\cdot|\pi),\mathsf{SHe})$ is also contained in [CC*24]. In particular, sufficient conditions for geodesic convexity are presented, and a necessary and sufficient condition for the Łojasiewicz inequality (called "gradient dominance" therein) are derived. Our results are different in the sense that we look for general Lyapunov functions such as $\mathrm{D}_{\varphi_q}(\cdot|\pi)$, where $q\neq p$ is allowed, whereas their focus is on the decay of the sum of certain two entropy functionals.

4.1 Pure spherical Hellinger gradient flow of probability measures

The gradient-flow equation for the gradient system $(\mathcal{P}(\Omega),\mathrm{D}_{\varphi_p}(\,\cdot\,|\pi),\mathsf{SHe})$ takes the form

$$\dot\rho=-\beta\rho\left(\varphi_p'\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right)-\int_\Omega\varphi_p'\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right)\mathrm{d}\rho\right),\tag{42}$$

where we have used the letter $\rho\in\mathcal{P}$ for probability measure instead of the positive measure $\mu\in\mathcal{M}^+$. The following result establishes a Polyak-Łojasiewicz inequality for the pure SHe gradient flow, which reads

$$\int_\Omega\varphi_p'\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right)\mathbb{K}_{\mathsf{SHe}}(\rho)\varphi_p'\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right)\mathrm{d}x\geq\beta\,M_p\,D_p(\rho|\pi)\text{ for some }M_p>0\tag{43}$$

for the case $p\in[0,1/2]$ that leads to exponential decay of solutions for $t>0$. Recall the definition of the spherical Hellinger Onsager operator $\mathbb{K}_{\mathsf{SHe}}$ in (29) and use $\int_\Omega\mathrm{d}\rho=1$, we establish the following result.

Theorem 4.1 (Ł-SHe estimate) *Assume $p \in (-\infty, 1/2]$ and that $\rho \in \mathcal{P}(\Omega)$ satisfies $\rho(x) > 0$ a.e. with respect to π. Then, the following functional inequality holds,*

$$\int_\Omega \frac{\mathrm{d}\rho}{\mathrm{d}\pi} \left(\varphi_p'\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right)\right)^2 \mathrm{d}\pi - \left(\int_\Omega \frac{\mathrm{d}\rho}{\mathrm{d}\pi}\varphi_p'\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right)\mathrm{d}\pi\right)^2 \geq M_p \int_\Omega \varphi_p\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right)\mathrm{d}\pi \qquad \text{(Ł-SHe)}$$

with $M_p = \begin{cases} \frac{1}{1-p} & \text{for } p \leq \frac{1}{3}, \\ \frac{p(7-12p)}{1-p} & \text{for } p \in [\frac{1}{3}, \frac{1}{2}]. \end{cases}$ *For $p > 1/2$ the best possible constant is $M_p = 0$.*

For $p > 1/2$, we do not have convergence because there are multiple steady states, namely $\rho_{\text{steady}}(x) = \frac{1}{\pi(A)}\mathbf{1}_A(x)$ for arbitrary sets $A \subset \Omega$ with $\pi(A) > 0$.

Proof of Theorem 4.1. We use the abbreviation $r = \frac{\mathrm{d}\rho}{\mathrm{d}\pi}$ such that $r \geq 0$ a.e. We treat the case $p \in (0, 1/2]$ first. We now use the definition of φ_p (7) and the abbreviation $I_\alpha(r) = \int_\Omega r^\alpha \,\mathrm{d}\pi$. Since $\rho, \pi \in \mathcal{P}(\Omega)$ we have

$$I_0(\rho) = I_1(\rho) = 1 \quad \text{and} \quad \mathrm{D}_{\varphi_p}(\rho|\pi) = \frac{1 - I_p(r)}{p(1-p)} \geq 0, \tag{44}$$

which implies $I_p(r) \leq 1$.

For the left-hand side of (Ł-SHe), we obtain, after some major cancellations, the simple relation

$$\text{LHS} = \frac{1}{(1-p)^2}\big(I_{2p-1}(r) - I_p(r)^2\big).$$

Moreover, Hölder's inequality gives

$$I_{\alpha+\beta}(r) \leq I_{\alpha/\theta}(r)^\theta I_{\beta/(1-\theta)}(r)^{1-\theta} \quad \text{for } \alpha, \beta \in \mathbb{R} \text{ and } \theta \in (0,1).$$

We use $I_0(\rho) = 1$ and choose $\alpha,\ \beta$, and θ such that $\alpha + \beta = 0$, $\alpha/\theta = p$, and $\beta/(1-\theta) = 2p - 1$. This gives $\theta = (1-2p)/(1-p) \in [0,1]$ and $\alpha = -\beta = p(1-2p)/(1-p)$, and we find $I_{2p-1}^{p/(1-p)} I_p^{(1-2p)/(1-p)} \geq I_0(r) = 1$ and conclude

$$\text{LHS} \geq \frac{I_p(r)^{2-1/p} - I_p(r)^2}{(1-p)^2} \geq \frac{A_p}{(1-p)^2}\big(1 - I_p(r)\big) \quad \text{with } A_p = \begin{cases} 1/p & \text{for } p \in (0, \frac{1}{3}], \\ 7-12p & \text{for } p \in [\frac{1}{3}, \frac{1}{2}]. \end{cases}$$

The last estimate follows from the fact that $y \mapsto g_p(y) := (y^{2-1/p} - y^2)/(1-y)$ satisfies $g_p(y) \to 1/p$ for $y \nearrow 1$. Moreover, for $p \in (0, \frac{1}{3}]$ we have $g_p'(y) \leq 0$ which gives $g_p(y) \geq 1/p$. For $p \in [\frac{1}{3}, \frac{1}{2}]$ the result can be similarly checked or numerically verified.

The case $p = 0$ is easier, because $\int_\Omega r\varphi_0'(r)\,\mathrm{d}\xi = \int_\Omega (r-1)\,\mathrm{d}\xi = \int_\Omega \mathrm{d}\rho - \int_\Omega \mathrm{d}\pi = 0$. Hence, we have

$$\text{LHS} = \int_\Omega r\big(1 - \frac{1}{r}\big)^2 \mathrm{d}\pi \quad \text{and} \quad \mathrm{D}_{\varphi_0}(\rho|\pi) = \int_\Omega -\log r \,\mathrm{d}\pi.$$

We further process the left-hand side,

$$\int_\Omega r\big(1 - \frac{1}{r}\big)^2 \mathrm{d}\pi = \int_\Omega \big(r - 2 + \frac{1}{r}\big)\,\mathrm{d}\pi = \int_\Omega \left(\frac{1}{r} - 1\right)\mathrm{d}\pi,$$

where the last equality follows from $\int_\Omega r \,\mathrm{d}\pi = 1$. Then, the desired estimate with $M_0 = 1$ follows from the elementary inequality $\frac{1}{r} - 1 \geq -\log r$ for $r > 0$, i.e.,

$$\text{LHS} = \int_\Omega \big(\frac{1}{r} - 1\big)\,\mathrm{d}\pi \geq \int_\Omega \big(-\log r\big)\,\mathrm{d}\pi = \mathrm{D}_{\varphi_0}(\rho|\pi).$$

In the case of $p < 0$, we use an argument similar to the case of $0 < p < 1/2$, but taking into account $I_p(\rho) \geq 1$. Using Hölder's inequality, we have $I_p(\rho) \leq I_0(\rho)^{1-\theta} I_{p/\theta}(\rho)^\theta = I_{p/\theta}(\rho)^\theta$. Choosing $\theta = -p/(1{-}2p) \in [0, \frac{1}{2})$ gives $I_{2p-1}(\rho) \geq I_p(\rho)^{2+1/|p|}$. By the convexity of the function $y^{2+1/|p|} - y^2$ and a Taylor expansion around $y = 1$, we have $y^{2+1/|p|} - y^2 \geq \frac{1}{|p|}(y{-}1)$ for $y \geq 1$. With $y = I_p(\rho) \geq 1$, we find

$$\text{LHS} = \frac{I_{2p-1}(\rho){-}I_p(\rho)^2}{(p-1)^2} \geq \frac{I_p(\rho)^{2+1/|p|}{-}I_p(\rho)^2}{(p-1)^2} \geq \frac{I_p(\rho){-}1}{|p|(p{-}1)^2} = \frac{1}{1{-}p}\, D_p(\rho|\pi),$$

which is the desired result.

For $p > 1/2$ we consider the measure ρ_ε such that $\rho_\varepsilon(x) = \varepsilon \cdot \pi(x),\ \varepsilon > 0$ on A_ε and $\rho_\varepsilon(x) = 2 \cdot \pi(x)$ on $X \setminus A_\varepsilon$. Since $I_1(\rho_\varepsilon) = 1$ must be satisfied, we obtain $\pi(A_\varepsilon) = 1/(2{-}\varepsilon),\ 1 - \pi(A_\varepsilon) = (1{-}\varepsilon)/(2{-}\varepsilon)$. Moreover, for $q > 0$ we have $I_q(\rho_\varepsilon) \to 2^{q-1}$ for $\varepsilon \to 0$. Thus, for $\varepsilon \to 0$, we obtain

$$\mathrm{D}_{\varphi_p}(\rho_\varepsilon|\pi) \to (2^p{-}1)/(p^2 - p) > 0, \quad I_p(\rho_\varepsilon) \to 2^{p-1}, \quad I_{2p-1}(\rho_\varepsilon) \to 2^{2p-2},$$

where the last relation uses the assumption $p > 1/2$. Thus, LHS$(\rho_\varepsilon) \to 0$ for $\varepsilon \to 0$, and the ratio LHS$(\rho_\varepsilon)/\mathrm{D}_{\varphi_p}(\rho_\varepsilon|\pi) \to 0$, i.e., this ratio cannot be lower bounded by a positive constant $M_p > 0$. Hence, the statement is proved. ■

Remark 4.2 (Hellinger flow of forward KL is mass-preserving) *From the proof of Theorem 4.1, we observe that the Łojasiewicz inequality for the spherical Hellinger gradient flow of the forward KL energy (φ_p with $p = 0$) is contained in the case for the (non-spherical) Hellinger Łojasiewicz* (Ł-He). *However, it must be noted that those two gradient flows are not the same: in the case of (non-spherical) Hellinger, the mass is only preserved when starting in the probability subspace $\mathcal{P}(\Omega)$, but not otherwise. In fact, the total mass can be explicitly calculated with elementary arguments as $Z(t) = 1 + e^{-\beta t}(Z(0) - 1)$. In contrast, the SHK flows can be extended to the outside of $\mathcal{P}(\Omega)$ to a mass-preserving flow. This is often done for the Otto-Wasserstein flow on positive measures; see Section 5.*

Corollary 4.3 (Exponential Decay of D_{φ_p}-divergence along SHe gradient flow)
Assume $p \in (-\infty, \frac{1}{2}]$ and consider an initial datum $\rho(0) \in \mathcal{P}(\Omega)$ with $\mathrm{D}_{\varphi_p}(\rho(0)|\pi) < \infty$. Then, the solution ρ of (42) *with $\rho(t,x) > 0$ for all $t > 0$ a.e. with respect to π satisfies an exponential decay estimate with constant $M_p > 0$ from Theorem 4.1, namely*

$$\mathrm{D}_{\varphi_p}(\rho(t)|\pi) \leq \mathrm{e}^{-\beta M_p t} \mathrm{D}_{\varphi_p}(\rho(0)|\pi) \ \textit{for all}\ t > 0.$$

Proof. This follows directly from (43) and a Grönwall estimate. ■

4.2 Spherical Hellinger-Kantorovich space and gradient flows

Finally, we apply our results for the SHe flows above to obtain the *global functional inequality and hence exponential convergence of the spherical Hellinger-Kantorovich gradient flow over probability measures* $\mathcal{P}$.

The specialized Łojasiewicz inequality for the SHK gradient flow reads

$$\int \rho \left(\alpha \left| \nabla \frac{\delta F}{\delta \rho}[\rho] \right|^2 + \beta \left| \frac{\delta F}{\delta \rho}[\rho] \right|^2 \right) - \beta \left(\int \rho \cdot \frac{\delta F}{\delta \rho}[\rho] \right)^2 \geq c_* \left(F(\rho) - \inf_{\nu \in \mathcal{M}^+} F(\nu) \right). \quad (45)$$

We establish the following result.

Theorem 4.4 (Functional inequality for spherical Hellinger-Kantorovich) *The SHK Łojasiewicz inequality* (45) *holds globally with a positive constant for the spherical Hellinger-Kantorovich (a.k.a., Wasserstein-Fisher-Rao) gradient flow over probability measures for the* φ_p *divergence energy for* $p \in (-\infty, \frac{1}{2}]$ *with* $c_* = c_{\mathsf{SHe}} = \beta M_p > 0$.

Furthermore, if the Otto-Wasserstein-Łojasiewicz inequality (Ł-W) *with reference measure* π *holds with* $c_{\text{Ł-W}} > 0$ *for all probability measures* $\mathcal{P}(\Omega)$*, then the SHK Łojasiewicz inequality holds with* $c_* = \alpha\, c_{\text{Ł-W}} > 0$*. Consequently, the SHK gradient flow converges globally with exponential decay rate* $c_* = \max\{\alpha\, c_{\text{Ł-W}},\ \beta\, M_p\}$.

Note that, for bounded Lipschitz domains Ω and $\pi \in L^\infty(\Omega)$ with $\inf_\Omega \pi(x) > 0$, the Otto-Wasserstein-Łojasiewicz inequality (Ł-W) indeed holds with $c_{\text{Ł-W}} > 0$ for $p > 1 - \frac{1}{d}$, see [MiM18, Sec. 3.1]. If the domain $\Omega = \mathbb{R}^d$, the Łojasiewicz inequality holds for the SHK gradient flows of φ_p-divergence energy for $p \in (-\infty, \frac{1}{2}] \cup [1, 2]$ given that (Ł-W) holds.

Remark 4.5 *This theorem showcases the strength of the SHK gradient flows. For dimension* $d \leq 4$*, the Łojasiewicz inequality holds for SHK gradient flows of all* φ_p*–divergence energy! For* $d \geq 5$*, we still have the generous interval* $p \in (-\infty, 1/2] \cup [1 - \frac{1}{d}, \infty)$*, which improves significantly from the pure Otto-Wasserstein and the pure (spherical) Hellinger geometries.*

A direct consequence is the following qualitative statement that applies to a large family of practical energy functionals

Corollary 4.6 *The SHK gradient flows converge exponentially globally for the following energy functionals: KL divergence* ($p = 1$) *under LSI, forward KL divergence* ($p = 0$) *unconditionally,* χ^2*-divergence* ($p = 2$) *under a Łojasiewicz inequality, reverse* χ^2*-divergence* ($p = -1$) *unconditionally, and the Hellinger distance* ($p = 1/2$).

5 Hellinger-Kantorovich gradient flows of positive measures $\mathcal{M}^+$

Unlike the spherical counterpart, the HK gradient flows over positive measures $\mathcal{M}^+$ are more challenging to treat. This is due to the absence of the global LSI type inequalities for the Otto-Wasserstein flows over positive measures $\mathcal{M}^+$, which we discuss in Section 5.1. Subsequently, we provide the analysis for the HK gradient flow over positive measures $\mathcal{M}^+$. Concretely, we establish global convergence results for the φ_p-divergence energy for $p \in (-\infty, 1/2]$, as well as for the KL divergence energy ($p = 1$)using a novel analysis via a shape-mass decomposition.

5.1 The loss of LSI on positive measures $\mathcal{M}^+$ and a sufficient condition

For the Wasserstein distance, the McCann condition (see, e.g., [AGS05]) shows that $\mathrm{D}_{\varphi_p}(\cdot|\,\mathrm{d}x)$ (i.e., the reference measure is Lebesgue) is geodesically convex only for $p \geq (d-1)/d$ where d is the dimension. In [LMS23a], necessary and sufficient conditions for the geodesic convexity of entropy functionals with respect to the HK distance were derived. The upper threshold $p = 1/2$ was also observed in the sense that densities with $p \in [p_*, 1/2] \cup (1, \infty)$ lead to geodesically convex p-divergences, where $p_* = 1/3$ for space dimension $d = 1$ and $p_* = 1/2$ for $d = 2$. For $d \geq 3$ only the range $p > 1$ is admitted. However, only the convexity constant $\Lambda = 0$ has been shown for all $p > 1$ (i.e., not strongly convex).

To improve on the state-of-the-art analysis, we first provide our result on the HK Łojasiewicz in the following corollary. The Łojasiewicz inequality in the HK geometry over positive measures $\mathcal{M}^+$ reads, for $\alpha, \beta > 0$,

$$\int \left(\alpha\left|\nabla\frac{\delta F}{\delta\mu}[\mu]\right|^2 + \beta\left|\frac{\delta F}{\delta\mu}[\mu]\right|^2\right) \mathrm{d}\mu \geq \beta c_* \left(F(\mu) - \inf_{\nu\in\mathcal{M}^+} F(\nu)\right). \tag{46}$$

Corollary 5.1 (A sufficient condition for HK flow) *For φ_p-divergence energy $F(\mu) = \mathrm{D}_{\varphi_p}(\mu|\pi)$ with $p \in (-\infty, \frac{1}{2}]$, the Łojasiewicz inequality* (46) *holds globally over positive measures $\mathcal{M}^+$ with the constant $c_* = \dfrac{1}{1-p}$.*

In relating those results to previous geodesic convexity results for the HK gradient flows in Table 1, we first note that geodesic convexity implies Łojasiewicz inequality but only with a non-negative constant $c \geq 0$. As the dimension increases, [LMS23a]'s result and the McCann condition have an increasing power threshold for the value of p. For dimension $d \geq 3$, their intervals no longer overlap with the threshold of $p \leq \frac{1}{2}$ for the global Łojasiewicz in the Hellinger geometry. Yet, we are able to provide a further Łojasiewicz result that is weaker than [LMS23a]'s geodesic convexity condition; see Table 1. In previous works such as [LMS23b], it has been suggested that the Łojasiewicz inequality for the HK geometry holds whenever the Łojasiewicz inequalities for the Otto-Wasserstein (LSI) and Hellinger both hold. However, we next show that such a strategy cannot result in a global Łojasiewicz inequality.

First, if $p \leq \frac{1}{2}$, Corollary 5.1 has established the Łojasiewicz inequality (Ł-He) with a constant $c \geq \frac{1}{1-p}$. This directly results in the Łojasiewicz inequality in the HK geometry. If $p > \frac{1}{2}$, different from the pure Fisher-Rao case, it does not automatically imply the absence of the HK Łojasiewicz inequality. This can be seen by first assuming a Łojasiewicz condition for the Otto-Wasserstein dissipation (Ł-W) with a constant condition $c_W > 0$. Then, since the Hellinger dissipation quantity is always non-negative along gradient flows,

$$\int \mu \cdot \left(\alpha\left|\nabla\frac{\delta F}{\delta\mu}[\mu]\right|^2 + \beta\left|\frac{\delta F}{\delta\mu}[\mu]\right|^2\right) \geq \alpha c_W \cdot \left(F(\mu) - \inf_{\nu\in\mathcal{M}^+} F(\nu)\right) + \beta \cdot 0,$$

which yields the HK Łojasiewicz with the constant c_W. Now, it may seem that the Łojasiewicz inequality in the HK geometry can be established in this manner. However, the situation is more nuanced due to the Otto-Wasserstein dissipation over positive measures $\mathcal{M}^+$, instead of the probability measure space $\mathcal{P}$. In such cases, Łojasiewicz inequality *cannot* hold globally for the Otto-Wasserstein flow.

Proposition 5.2 (No Łojasiewicz for Otto-Wasserstein flows over $\mathcal{M}^+$) *Given the φ-divergence energy functional $\mathrm{D}_\varphi(\cdot|\pi)$. Then, there exists no global Łojasiewicz inequality* (Ł-W) *for the Otto-Wasserstein gradient flow over the positive measures $\mathcal{M}^+$.*

Proof of proposition 5.2. For any non-negative target measure $\pi \in \mathcal{M}^+$, we pick the measure μ to be a scalar multiple of π, i.e., $\mu = Z\,\pi$ with $Z > 0$. See the illustration in Figure 7. The Radon-Nikodym derivative is a constant $\frac{\mathrm{d}\mu}{\mathrm{d}\pi} \equiv Z$. Then, the Łojasiewicz inequality reads

$$0 = \left\| \nabla \left(\varphi'\Big(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\Big) \right) \right\|^2_{L^2_\mu} \geq c \cdot \mathrm{D}_\varphi(\mu|\pi) \text{ for some } c > 0,$$

which cannot hold whenever $\mathrm{D}_\varphi(\mu|\pi) = \varphi(Z)\pi(\Omega) > 0$. ■

The intuition for the above proposition is that the Otto-Wasserstein flow of $\mathrm{D}_\varphi(\cdot|\pi)$ "gets stuck" when the density ratio is constant $\frac{\mathrm{d}\mu}{\mathrm{d}\pi} \equiv Z$ since the metric slope is zero. See also the illustration in Figure 7. This result shows that we cannot hope to rely on the Otto-Wasserstein dissipation to establish the Łojasiewicz inequality in the HK flow of positive measures.

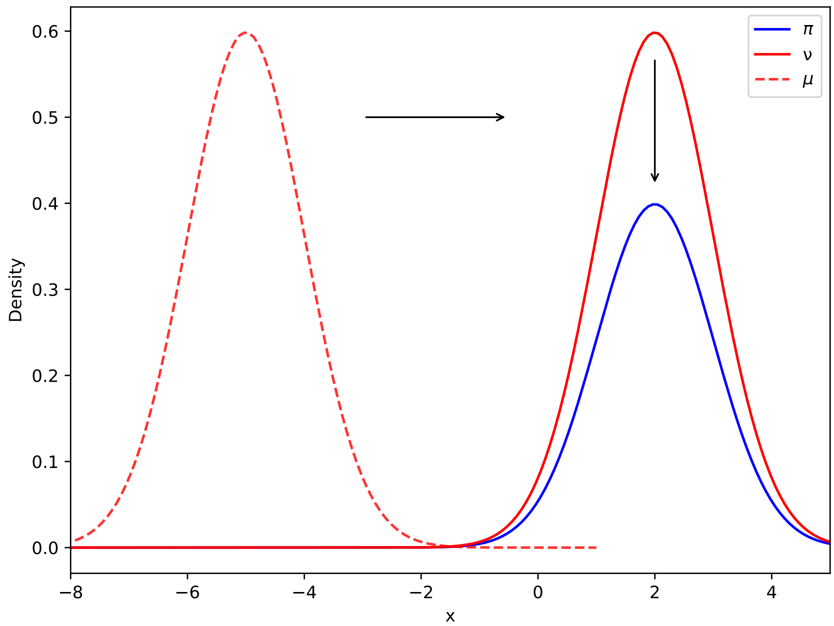

Figure 7: See Proposition 5.2 for the details of the functional inequality and Otto-Wasserstein flow of positive measures. In this plot, the density ratio $\frac{\mathrm{d}\nu}{\mathrm{d}\pi}$ is a constant $Z > 0$. Hence, there is no "Otto-Wasserstein gradient" to drive the curve from ν towards π. In the opposite regime, the density ratio $\frac{\mathrm{d}\mu}{\mathrm{d}\pi}$ has many close-to-zero locations. Hence, there is not enough "Hellinger gradient" to drive the curve from μ towards π.

5.2 A special case: HK gradient flows of KL divergence energy

In contrast to the global convergence results for the SHK gradient flows in Section 4, results such as Proposition 5.2 might hint a pessimism about the HK gradient flows. However, in this section, we show that the HK gradient flows of KL divergence (i.e. φ_p for $p = 1$) driving energy have a special property that still guarantees the global convergence.

To further understand the idea behind Proposition 5.2, we first show a straightforward extension of the (LSI) over probability measures to positive measures. Without loss of generality, we assume that the target measure π is a probability measure, i.e., $\pi(\Omega) = 1$.

DOI 10.20347/WIAS.PREPRINT.3176

Proposition 5.3 (Generalized log-Sobolev inequality on $\mathcal{M}^+$) *Suppose the logarithmic Sobolev inequality* (LSI) *holds with a positive constant $c_{LSI\text{-}\mathcal{P}} > 0$ when restricted to probability measures (i.e. μ and π are probability measures). Then, the following inequality holds for the Otto-Wasserstein gradient flow over the positive measures $\mathcal{M}^+$:*

$$\int \left|\nabla \log \frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right|^2 \mathrm{d}\mu \geq c_{LSI\text{-}\mathcal{P}} \cdot \Big(\mathrm{D}_{\mathrm{KL}}(\mu|\pi) - (z\log z - z + 1)\Big), \tag{LSI-$\mathcal{M}^+$}$$

where $z := \mu(\Omega)$ is the total mass of the measure μ. Moreover, we have

$$\int \left|\nabla \log \frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right|^2 \mathrm{d}\mu \geq c_{LSI\text{-}\mathcal{P}} \cdot \mathrm{D}_{\mathrm{KL}}(\mu|z\cdot\pi). \tag{47}$$

The intuition here is that the Otto-Wasserstein gradient flow, viewed as a mass-preserving flow with total mass $\mu(\Omega)$, satisfies the LSI type inequality. This is illustrated in Figure 7.

Proof of Proposition 5.3. We have the logarithmic Sobolev inequality (LSI) for the probability measures $\tilde{\mu} := \frac{1}{z}\cdot\mu$ where $z := \mu(\Omega)$ is the mass of μ,

$$\int \frac{\mathrm{d}\tilde{\mu}}{\mathrm{d}\pi}\left(\nabla \log \frac{\mathrm{d}\tilde{\mu}}{\mathrm{d}\pi}\right)^2 \mathrm{d}\pi \geq c_{\mathrm{LSI}\text{-}\mathcal{P}} \cdot \mathrm{D}_{\mathrm{KL}}(\tilde{\mu}|\pi).$$

Expanding the KL divergence, we have

$$\begin{aligned}\mathrm{D}_{\mathrm{KL}}(\tilde{\mu}|\pi) = \int \frac{\mathrm{d}\tilde{\mu}}{\mathrm{d}\pi}\log\frac{\mathrm{d}\tilde{\mu}}{\mathrm{d}\pi}\mathrm{d}\pi &= \int \frac{1}{z}\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\left(\log\frac{\mathrm{d}\mu}{\mathrm{d}\pi} - \log z\right)\mathrm{d}\pi \\ &= \frac{1}{z}\left(\int \frac{\mathrm{d}\mu}{\mathrm{d}\pi}\log\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\mathrm{d}\pi - z + 1\right) - \log z + 1 - \frac{1}{z} \\ &= \frac{1}{z}\mathrm{D}_{\mathrm{KL}}(\mu|\pi) - \frac{1}{z}(z\log z - z + 1).\end{aligned}$$

Combining the above relation with property of the Sobolev norm,

$$\begin{aligned}\int \frac{\mathrm{d}\mu}{\mathrm{d}\pi}\left(\nabla \log \frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right)^2 \mathrm{d}\pi &= z\cdot\int \frac{\mathrm{d}\tilde{\mu}}{\mathrm{d}\pi}\left(\nabla \log \frac{\mathrm{d}\tilde{\mu}}{\mathrm{d}\pi}\right)^2 \mathrm{d}\pi \\ &\geq c_{\mathrm{LSI}\text{-}\mathcal{P}} \cdot (\mathrm{D}_{\mathrm{KL}}(\mu|\pi) - (z\log z - z + 1)) \\ &= c_{\mathrm{LSI}\text{-}\mathcal{P}} \cdot (\mathrm{D}_{\mathrm{KL}}(\mu|\pi) - \mathrm{D}_{\mathrm{KL}}(\mu(\Omega)\cdot\pi|\pi)).\end{aligned}$$

For the last part of the result, we rewrite the right-hand side of the inequality above using the relation $z\log z - z + 1 = \mathrm{D}_{\mathrm{KL}}(\mu(\Omega)\cdot\pi|\pi)$. Recall a generalized Pythagorean inequality for the KL divergence that reads

$$\mathrm{D}_{\mathrm{KL}}(\mu|\pi) \geq \mathrm{D}_{\mathrm{KL}}(\mu|\pi^*) + \mathrm{D}_{\mathrm{KL}}(\pi^*|\pi),$$

where π^* is the information projection of π onto the positive measures of total mass z

$$\pi^* \in \operatorname{arginf}\left\{\mathrm{D}_{\mathrm{KL}}(\gamma|\pi) \mid \gamma \in \mathcal{M}^+, \gamma(\Omega) = z\right\}.$$

By Jensen's inequality,

$$\mathrm{D}_{\mathrm{KL}}(\gamma|\pi) = \int \varphi_{\mathrm{KL}}\,\mathrm{d}\pi \geq \varphi_{\mathrm{KL}}\left(\int \mathrm{d}\pi\right) = \varphi_{\mathrm{KL}}(z),$$

where the inequality holds when $\pi^* = z \cdot \pi$. Therefore,

$$\mathrm{D}_{\mathrm{KL}}(\mu|\pi) \geq \mathrm{D}_{\mathrm{KL}}(\mu|z\cdot\pi) + \mathrm{D}_{\mathrm{KL}}(z\cdot\pi|\pi). \tag{48}$$

Combining the results above, we obtain the desired inequality

$$\int \frac{\mathrm{d}\mu}{\mathrm{d}\pi}(\nabla \log \frac{\mathrm{d}\mu}{\mathrm{d}\pi})^2 \,\mathrm{d}\pi \geq c_{\text{LSI-}\mathcal{P}} \cdot (\mathrm{D}_{\mathrm{KL}}(\mu|\pi) - \mathrm{D}_{\mathrm{KL}}(\mu(\Omega)\cdot\pi|\pi)) \geq C \cdot \mathrm{D}_{\mathrm{KL}}(\mu|\mu(\Omega)\cdot\pi). \tag{49}$$

Thus, Proposition 5.3 is established. ■

The insight from Proposition 5.3 also provides us an exponentially decaying Lyapunov functional along the Otto-Wasserstein flow over the $\mathcal{M}^+$. Noting the property of the KL divergence $\mathrm{D}_{\mathrm{KL}}(\mu|z\cdot\pi) = z\cdot\mathrm{D}_{\mathrm{KL}}\left(\frac{1}{z}\cdot\mu|\pi\right)$, we find

$$\begin{aligned} -\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{D}_{\mathrm{KL}}(\mu|z\cdot\pi) &= \frac{1}{z}\int \frac{\mathrm{d}\mu}{\mathrm{d}\pi}\cdot\left(\nabla\left(\log\frac{\mathrm{d}\mu}{\mathrm{d}\pi} - \log z\right)\right)^2 \mathrm{d}(z\cdot\pi) \\ &= \int \frac{\mathrm{d}\mu}{\mathrm{d}\pi}\cdot\left(\nabla\log\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right)^2 \mathrm{d}\pi \overset{\text{Prop. 5.3}}{\geq} c_{\text{LSI-}\mathcal{P}}\cdot\mathrm{D}_{\mathrm{KL}}(\mu|z\cdot\pi) \end{aligned} \tag{50}$$

Then, by Grönwall's lemma, the Lyapunov functional $\mathrm{D}_{\mathrm{KL}}(\mu|z\cdot\pi)$ decays exponentially along the mass-preserving Otto-Wasserstein gradient flow.

Corollary 5.4 ($\mathrm{D}_{\mathrm{KL}}(\mu|\mu(\Omega)\cdot\pi)$ is Lyapunov for Otto-Wasserstein-$\mathcal{M}^+$) *For the mass-preserving Otto-Wasserstein gradient flow over the positive measures $\mathcal{M}^+$ with the KL divergence energy $F(\mu) = \mathrm{D}_{\mathrm{KL}}(\mu|\pi)$, the Lyapunov functional $\mathrm{D}_{\mathrm{KL}}(\mu|z\cdot\pi)$ (where $z := \mu(\Omega)$ is the total mass of the measure μ) decays exponentially along the flow, i.e.,*

$$\mathrm{D}_{\mathrm{KL}}(\mu(t)|z\cdot\pi) \leq \mathrm{e}^{-c_{\text{LSI-}\mathcal{P}}\cdot t}\mathrm{D}_{\mathrm{KL}}(\mu(0)|z\cdot\pi) \text{ and } \mathrm{D}_{\mathrm{KL}}(\frac{1}{z}\mu(t)|\pi) \leq \mathrm{e}^{-c_{\text{LSI-}\mathcal{P}}\cdot t}\mathrm{D}_{\mathrm{KL}}(\frac{1}{z}\mu(0)|\pi)$$

for $t \geq 0$ and the LSI constant $c_{\text{LSI-}\mathcal{P}}$ as in Proposition 5.3.

This result, combined with the Pythagorean type relation (48), shows the Lyapunov functional $\mathrm{D}_{\mathrm{KL}}(\mu|z\cdot\pi)$ decays towards the lower bound $\mathrm{D}_{\mathrm{KL}}(z\cdot\pi|\pi) = z\log z - z + 1$. Furthermore, it implies that, while μ itself does not converge to π due to Proposition 5.2, the *shape* $\frac{1}{\mu(\Omega)}\mu$ does converge to the target. Using a similar idea, we next analyze the Hellinger (He) and the spherical Hellinger (SHe) geometries.

We exploit a special property of the KL divergence, namely, for any constant $Z \in \mathbb{R}^+$, the SHe flow of the KL divergence energy $\mathrm{D}_{\mathrm{KL}}(\cdot|Z\pi)$ is independent of Z. Yet, in the He flow of positive measures, the constant Z in the minimization has an impact. This idea can be easily seen by calculating the gradient flow equation of the He flow

$$\dot{\rho} = -\rho\log\left(\frac{\mathrm{d}\rho}{\mathrm{d}(Z\pi)}\right) = -\rho\left(\log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right) - \log Z\right),$$

i.e., the growth field is indeed affected by the scalar Z. In comparison, the scalar Z is canceled for the SHe flow of probability measures

$$\dot{\rho} = -\rho\left(\log\left(\frac{\mathrm{d}\rho}{\mathrm{d}(Z\pi)}\right) - \int\rho\log\left(\frac{\mathrm{d}\rho}{\mathrm{d}(Z\pi)}\right)\right) = -\rho\left(\log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right) - \int\rho\log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right)\right).$$

Since the Otto-Wasserstein flow is always mass-conserving, this difference in He and SHe is the key for our analysis next, which we term the *shape-mass analysis*.

5.3 Shape-mass analysis: global KL decay of HK gradient flows

Our starting point is to carefully compare the HK and SHK gradient flows. For the convenience, we remember below the associated gradient-flow equations of HK and SHK flows under the KL energy

$$\dot{\mu} = \alpha \operatorname{div}\big(\nabla\mu + \frac{\mathrm{d}\mu}{\mathrm{d}\pi}\nabla\pi\big) - \beta\mu\log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right), \tag{HK-KL}$$

$$\dot{\rho} = \alpha \operatorname{div}\big(\nabla\rho + \frac{\mathrm{d}\rho}{\mathrm{d}\pi}\nabla\pi\big) - \beta\rho\left(\log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right) - \int_{\mathbb{R}^d}\rho\log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right)\mathrm{d}x\right). \tag{SHK-KL}$$

For the clarity of the analysis, we use the symbol μ for the positive measure in the HK flow and ρ for the probability measure in the SHK flow. We exploit the following simple observation of those two equations.

Theorem 5.5 (Relation between solutions to HK and SHK equations) *If $t \mapsto \mu(t)$ solves* (HK-KL)*, then $t \mapsto \rho(t) = \frac{1}{z(t)}\mu(t)$ with $z(t) = \int_{\mathbb{R}^d}\mu(t,x)\,\mathrm{d}x$ solves* (SHK-KL)*. Moreover, if $t \mapsto \rho(t)$ solves* (SHK-KL)*, then $t \mapsto \mu(t) = \kappa(t)\rho(t)$ solves* (HK-KL) *for suitable functions $t \mapsto \kappa(t)$ independent of the variable x. Furthermore, $\kappa(t)$ is the solution to the following equation of mass*

$$\dot{z} = -\beta z\log z - \beta z\int_{\mathbb{R}^d}\rho\log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right)\mathrm{d}x. \tag{Mass-HK}$$

Proof of Theorem 5.5. The first part of the proposition is straightforward. To derive the mass equation, suppose μ is a solution to the HK equation (HK-KL). Applying the chain rule to the time derivative $\dot{\mu} = \dot{z}\rho + z\dot{\rho}$, where the shape-mass decomposition $\mu = z\rho$ is used. Plug this into the HK equation (HK-KL),

$$\dot{z}\rho = \alpha\operatorname{div}\left(\nabla\mu + \frac{\mathrm{d}\mu}{\mathrm{d}\pi}\nabla\pi\right) - \beta\mu\log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right) - z\dot{\rho}.$$

Since the shape ρ is a probability measure, we have $\int\rho\,\mathrm{d}x = 1, \int\dot{\rho}\,\mathrm{d}x = 0$. Then, we integrate both sides of the above equation to obtain

$$\dot{z} = \alpha\int\operatorname{div}\left(\nabla\mu + \frac{\mathrm{d}\mu}{\mathrm{d}\pi}\nabla\pi\right) - \beta\int\mu\log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right)\mathrm{d}x \overset{\text{IBP}}{=} -\beta z\log z - \beta\int\rho\log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right)\mathrm{d}x,$$

which is the desired mass equation (Mass-HK). ■

This observation reveals the key to the following shape-mass analysis we will present: consider a general solution $t \mapsto \mu(t)$ of (HK-KL) and write it in the form $\mu(t) = z(t)\rho(t)$ with the normalized density $\rho(t) \in \mathcal{P}(\Omega)$ describing the shape and $z(t) > 0$ the total mass. Using this observation, We can further extend the SHK analysis to general target measure $\pi \in \mathcal{M}^+$. The ρ-equation (SHK-KL) is mass-preserving and invariant under the change of variable from π to $\gamma\pi$ with $\gamma > 0$. In that case, one expects convergence to the steady state $\gamma_\pi\pi \in \mathcal{P}(\Omega)$, where γ_π is a normalizing constant. Hence, we now denote the shape-mass decomposition of the target $\pi = z_*\pi_*$ where $\pi_* \in \mathcal{P}(\Omega)$. Then, when starting from a solution $t \mapsto \rho(t)$ of the mass-preserving flow (SHK-KL) and assuming $\pi_* \in \mathcal{P}(\Omega)$, $z_0 > 0$, and $z_* > 0$, the mass equation (Mass-HK) reads

$$\dot{z} = \beta\, z\left(\log z_* - \mathrm{D}_{\mathrm{KL}}(\rho|\pi) - \log z\right), \qquad z(0) = z_0.$$

Then $t \mapsto \mu(t) = z(t)\rho(t)$ is a solution of (HK-KL) with the initial condition $\rho(0) = z_0\rho(0)$ for the energy functional $\mathrm{D}_{\mathrm{KL}}(\,\cdot\,|z_*\pi_*)$.

To provide a general decay estimate for solutions of $(\mathcal{M}^+(\Omega), \mathcal{H}_{\mathrm{B}}, \mathsf{HK}_{\alpha,\beta})$, we now use the shape-mass decomposition $\mu(t,x) = z(t)\rho(t,x)$. As shown in Proposition 5.2 and Proposition 5.3, LSI cannot hold globally over $\mathcal{M}^+$. Therefore, we use the standard log-Sobolev inequality but restricted to the probability measures, which is the same as in (LSI) and recalled here for convenience: for $\pi_* = \gamma_\pi \, \pi \in \mathcal{P}(\Omega)$,

$$\exists\, c_{\mathsf{LSI}} > 0 \; \forall\, \rho \in \mathcal{P}(\Omega) : \quad \int_\Omega \rho \big|\nabla \log(\rho/\pi_*)\big|^2 \mathrm{d}x \geq c_{\mathsf{LSI}} \mathrm{D}_{\mathrm{KL}}(\rho|\pi_*). \tag{LSI-$\mathcal{P}$}$$

In the following result, we can see two contributions to the convergence of $\mu(t) = z(t)\rho(t)$ to $\pi = z_*\pi_*$, where $z_* := \pi(\Omega)$ is the total mass of the target measure and π_* is a probability measure, a.k.a. the shape. We now detail the results of the shape-mass analysis for the HK-KL gradient flow.

We first provide the convergence of the mass variable $z(t)$ to the target mass z_*.

Proposition 5.6 (Solution of the mass equation) *The equation of mass* (Mass-HK) *admits the explicit solution*

$$z(t) = z_* \left(\frac{z_0}{z_*}\right)^{\mathrm{e}^{-\beta t}} \mathrm{e}^{-h(t)}. \tag{51}$$

where $h(t) = \int_0^t \mathrm{e}^{-\beta(t-s)} \mathrm{D}_{KL}(\rho(s)|\pi_*)\,\mathrm{d}s$ is an auxiliary function.
If $\mathrm{D}_{KL}(\rho(s)|\pi_*) \to 0$ for $t \to \infty$, then $h(t) \to 0$ and $z(t) \to z_*$.

Setting $H_0 = \mathrm{D}_{\mathrm{KL}}(\rho(0)|\pi_*)$ and $\widehat{\alpha} = \alpha c_{\mathrm{LSI}\text{-}\mathcal{P}}$, we now deliver the convergence of the shape $\rho(t)$ to the target shape π_* and the mass $z(t)$ to the target mass z_*.

Proposition 5.7 (Shape and mass convergence) *The shape, i.e. the normalized probability measure $\rho(t) = \frac{1}{z(t)}\mu(t)$, converges to the target π_* exponentially in KL divergence along the HK gradient flow, i.e.,*

$$\mathrm{D}_{KL}(\rho(t)|\pi_*) \leq \mathrm{e}^{-\widehat{\alpha} t} H_0. \tag{shape convergence}$$

The mass variable $z(t)$ converges to the target mass z_* exponentially, i.e.,

$$|z(t) - z_*| \leq \max\{z_0, z_*\} \left|\log\left(\frac{z_0}{z_*}\right)\right| \mathrm{e}^{-\beta t} + H_0 \frac{\mathrm{e}^{-\widehat{\alpha} t} - \mathrm{e}^{-\beta t}}{\beta - \widehat{\alpha}}. \tag{mass convergence}$$

Note that the convergence rate of the shape $\rho(t)$ to the limiting shape π_* is dominated by the transport part alone, with an exponential decay rate $\widehat{\alpha} = \alpha C_{\mathsf{LSI}}$. The total mass can only be changed by the growth through the Hellinger dissipation. Hence, the decay rate is simply β, but it may be delayed by $\mathrm{e}^{-\widehat{\alpha} t}$ if the shape converges only slowly.

Combining the results of Proposition 5.6 and Proposition 5.7, we can now provide the global exponential decay analysis for the HK-KL gradient flow in the sense of the Hellinger distance.

Theorem 5.8 (Convergence to equilibrium via shape-mass analysis) *The following convergence estimate in the Hellinger distance holds*

$$\mathsf{He}(\mu(t),\pi) \le \left(\frac{\max\{z_0^{1/2}, z_*^{1/2}\}}{2}\, H_0^{1/2} + z_*^{1/2} \left(g\left(\left(\frac{z_0}{z_*}\right)^{1/2} \right) + \frac{1}{\widehat\alpha} \right) \right) \mathrm{e}^{-\gamma t} \quad \textit{for } t>0, \tag{52}$$

where $\gamma = \min\left\{\beta, \widehat\alpha/2\right\}$ *and* $g(a) = \max\{\log(1/a), a{-}1\} \ge 0.$

Before delving into the proof, we highlight that the singularity of $g(a) = \log(1/a)$ (for $a < 1$) is needed to cover the case that, for a very small initial mass z_0, it takes a long time to build up enough mass to see the exponential decay to the limiting profile.

The above results imply that we cannot have a global Łojasiewicz inequality for the HK gradient flow over the positive measures $\mathcal{M}^+$. However, the last theorem shows that, for the KL divergence as driving energy, global exponential decay is still guaranteed. An exception is the case that we start with $\mu = 0$, which remains an unstable steady state of the flow.

Proof of Theorem 5.8. We use the shape-mass decomposition $\mu(t) = z(t)\rho(t)$ and $\pi = z_*\pi_*$ with $\pi_*, \rho(t) \in \mathcal{P}(\mathbb{R}^d)$. We first estimate the convergence of ρ to π_* via

$$-\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{D}_{\mathrm{KL}}(\rho|\pi_*) = \int_{\mathbb{R}^d} \Big(\alpha\rho \big|\nabla \log(\rho/\pi_*)\big|^2 + \beta\rho\big(\log(\rho/\pi_*) - \int \rho\log(\rho/\pi_*)\big)^2 \Big)\,\mathrm{d}x \\ \ge \widehat\alpha \mathrm{D}_{\mathrm{KL}}(\rho|\pi_*) + \beta\cdot 0,$$

where we ignored the term due to the spherical Hellinger geometry since it's non-negative. Thus, we obtain

$$\mathrm{D}_{\mathrm{KL}}(\rho(t)|\pi_*) \le \mathrm{e}^{-\widehat\alpha t} H_0 \quad \text{with } \widehat\alpha = \alpha c_{\mathsf{LSI}\text{-}\mathcal{P}} \text{ and } H_0 = \mathrm{D}_{\mathrm{KL}}(\rho(0)|\pi_*).$$

Next, we use the relation for $z(t)$ with $z_0 = \int_{\mathbb{R}^d} \mu(0,x)\,\mathrm{d}x$:

$$z(t) = z_*\big(z_0/z_*\big)^{\mathrm{e}^{-\beta t}}\, \mathrm{e}^{-h(t)} \text{ with } h(t) = \int_0^t \mathrm{e}^{-\beta(t-s)} \mathrm{D}_{\mathrm{KL}}(\rho(s)|\pi_*)\,\mathrm{d}s.$$

Using the estimate for $\mathrm{D}_{\mathrm{KL}}(\rho(t)|\pi_*)$ we have $0 \le h(t) \le H(t) := H_0\big(\mathrm{e}^{-\widehat\alpha t}{-}\mathrm{e}^{-\beta t}\big)/(\beta{-}\widehat\alpha)$. Moreover, for all $a, t > 0$ we have

$$\Big|a^{\mathrm{e}^{-\beta t}} - 1\Big| \le g(a)\,\mathrm{e}^{-\beta t} \quad \text{where } g(a) = \sup_{x\in(0,1)} \tfrac{|a^x-1|}{x} = \max\{\log(1/a), a{-}1\}.$$

Using $\mathrm{e}^{-h(t)} \le 1$, we find, for $\sigma \in \,]0,1]$, the estimate

$$\begin{aligned} |z(t)^\sigma {-} z_*^\sigma| &\le \Big|z_*^\sigma (z_0/z_*)^{\sigma\,\mathrm{e}^{-\beta t}} \mathrm{e}^{-h(t)\sigma} - z_*^\sigma \mathrm{e}^{-h(t)\sigma}\Big| + \Big|z_*^\sigma \mathrm{e}^{-h(t)\sigma} - z_*^\sigma\Big| \\ &\le z_*^\sigma\Big|(z_0/z_*)^{\sigma\,\mathrm{e}^{-\beta t}} {-}1\Big| + z_*^\sigma \sigma H(t) \le z_*^\sigma\, g\big(z_0^\sigma/z_*^\sigma\big)\,\mathrm{e}^{-\beta t} + z_*^\sigma\sigma H(t). \end{aligned}$$

We estimated the last term on the first line by $|\mathrm{e}^{-x}{-}1| \le x$ for all $x > 0$, using $x = \sigma h(t)$.

For the full estimate, we use the classical bound $4\mathsf{He}(\rho,\pi)^2 = 2\mathcal{D}_{\phi_{1/2}}(\rho|\pi) \le \mathrm{D}_{\mathrm{KL}}(\rho|\pi) \le H_0 \mathrm{e}^{-\widehat\alpha t}$. With $z(t) \le \max\{z_0, z_*\}$, we are now able to establish (52) as follows:

$$\begin{aligned} \mathsf{He}(\mu,\pi) &= \mathsf{He}(z(t)\rho, z_*\pi_*) \le \mathsf{He}(z(t)\rho, z(t)\pi_*) + \mathsf{He}(z(t)\pi_*, z_*\pi_*) \\ &= \sqrt{z}\,\mathsf{He}(\rho,\pi) + |\sqrt{z}{-}\sqrt{z_*}| \\ &\le \max\{z_0^{1/2}, z_*^{1/2}\}\frac{H_0^{1/2}}{2}\mathrm{e}^{\widehat\alpha t/2} + z_*^{1/2} g\big((z_0/z_*)^{1/2}\big)\mathrm{e}^{-\beta t} + \frac{z_*^{1/2}}{2} H(t). \end{aligned}$$

Moreover, we establish $H(t) \leq (2/\widehat{\alpha})\,\mathrm{e}^{-\gamma t}$ with $\gamma = \min\{\beta, \widehat{\alpha}/2\}$ as follows: For $\beta \leq \widehat{\alpha}/2$, we have

$$\frac{\mathrm{e}^{-\widehat{\alpha}t} - \mathrm{e}^{-\beta t}}{\beta - \widehat{\alpha}} = \mathrm{e}^{-\beta t}\int_0^t \mathrm{e}^{(\beta-\widehat{\alpha})s}\,\mathrm{d}s \leq \mathrm{e}^{-\beta t}\int_0^t \mathrm{e}^{-\widehat{\alpha}s/2}\,\mathrm{d}s \leq \frac{2}{\widehat{\alpha}}\,\mathrm{e}^{-\beta t}$$

and for $\beta \geq \widehat{\alpha}/2$ we estimate as follows:

$$\frac{\mathrm{e}^{-\widehat{\alpha}t} - \mathrm{e}^{-\beta t}}{\beta - \widehat{\alpha}} = \int_0^t \mathrm{e}^{-\beta(t-s)}\mathrm{e}^{\widehat{\alpha}s}\,\mathrm{d}s \leq \int_0^t \mathrm{e}^{-\widehat{\alpha}(t-s)/2}\mathrm{e}^{-\widehat{\alpha}s}\,\mathrm{d}s \leq \frac{2}{\widehat{\alpha}}\,\mathrm{e}^{-\widehat{\alpha}t/2}.$$

Putting together the results, the desired estimate (52) is established. ■

A decay estimate for $\mathrm{D}_{\mathrm{KL}}(\mu(t)|\pi)$ similar to (52) can also be derived by using the decomposition $\mathrm{D}_{\mathrm{KL}}(\mu(t)|\pi) = \mathrm{D}_{\mathrm{KL}}(z(t)\rho(t)|z_*\pi_*) = z(t)\,\mathrm{D}_{\mathrm{KL}}(\rho(t)|\pi_*) + z_*\lambda(z(t)/z_*)$ with $\lambda(r) = r\log r - r + 1$. We omit the elementary proof to avoid redundancy.

Remark 5.9 (Beyond the KL energy functional) *It is tempting to generalize the above analysis to general φ-divergence energy $F(\mu) = \mathrm{D}_\varphi(\mu|\pi)$, beyond the KL case. However, we now present the following observation that such a generalization is difficult.*

Using a similar shape-mass decomposition $\mu = z\rho$ as in the proof of Theorem 5.5, we extract the equation

$$\dot{\rho} = \alpha\,\mathrm{div}\left(\rho\nabla\frac{\delta F}{\delta\mu}[\mu]\right) - \beta\rho\left(\frac{\delta F}{\delta\mu}[\mu] + \frac{\dot{z}}{z}\right). \tag{53}$$

We integrate both sides and again use the fact that ρ remains a probability measure along the mass-preserving flow. Noting the simple relation $\frac{\delta F}{\delta\mu}[\mu] = \frac{\delta F}{\delta\mu}[z\rho]$, we obtain

$$\frac{\dot{z}}{z} = -\int \rho\cdot\frac{\delta F}{\delta\mu}[z\rho]. \tag{54}$$

Therefore, the shape equation (53) *can be rewritten as*

$$\dot{\rho} = \alpha\,\mathrm{div}\left(\rho\nabla\frac{\delta F}{\delta\mu}[z\rho]\right) - \beta\rho\left(\frac{\delta F}{\delta\mu}[z\rho] - \int \rho\cdot\frac{\delta F}{\delta\mu}[z\rho]\right)$$

Specialized to the φ_p-divergence energy $F(\mu) = \mathrm{D}_{\varphi_p}(\mu|\pi)$, we have

$$\rho = \alpha\,\mathrm{div}\left(\rho\nabla\varphi_p'\left(z\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right)\right) - \beta\rho\left(\varphi_p'\left(z\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right) - \int\rho\cdot\varphi_p'\left(z\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right)\right) \tag{55}$$

This shape equation (55) *reveals the insight about the SHK flow of the φ_p-divergence energy. If $p = 1$, i.e., the KL divergence energy, the shape equation* (55) *simplifies to the energy equation of* (SHK-KL), *which is the observation of Theorem 5.5 and Section 5.2.*

In the case of $p \neq 1$, the shape equation (55) *is* not *the SHK gradient flow equation by itself – the shape and mass variables are coupled. Therefore, the observation of Theorem 5.5 does not hold for other φ_p-divergence energies than the KL. For example, in the case of $p = 2$, the shape equation* (55) *reads*

$$\dot{\rho} = \alpha\cdot z\,\mathrm{div}\left(\rho\nabla\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right) - \beta\cdot z\rho\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi} - \int\rho\cdot\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right)$$

where the right-hand side has a coupled mass variable z. Hence, it is not the SHK gradient flow equation. In this sense, our shap-mass analysis is specifically designed for the HK-KL gradient flow.

A Further proofs and technical results

We need the following technical properties to prepare for the characterization of the Lyapunov functional of the HK gradient flows.

Lemma A.1 *The constant $m_{p,q}$ satisfies the following estimates:*
(a) $m_{p,q} > 0$ if and only if $p \le \widehat{p}(q)$ with $\widehat{p}(q) := \max\big\{0, \min\{1, 1{-}q\}\big\}$,
(b) For $p \le 1/2$ we have $m_{p,p} = 1/(1{-}p)$.
(c) For $p \in [0,1]$ we have $m_{p,1-p} = \min\{1/p, 1/(1{-}p)\} \in [1,2]$.

Proof. We define $N_{p,q}(r) = r\varphi_p'(r)\varphi_q'(r)/\phi_q(r)$ which can be continuously extended at $r = 1$ by the value $N_{p,q}(1) = 2$. Thus, using the continuity and positivity of $N_{p,q} : (0,\infty) \to (0,\infty)$, we obtain $m_{p,q} > 0$ if and only if the two limits for $N_{p,q}(0) = \lim_{r\to 0} N_{p,q}(r)$ and $N_{p,q}(\infty) = \lim_{r\to\infty} N_{p,q}(\infty)$ are positive as well.

The asymptotic behavior for $r \to \infty$ is easily discussed:

$$N_{p,q}(\infty) = \begin{cases} \infty & \text{for } p \ge 1, \\ \frac{\max\{q,1\}}{1-p} & \text{for } p < 1. \end{cases}$$

To see this, we first observe that $r\varphi_q'(r)/\phi_q(r) \to \max\{1, q\}$ for $r \to \infty$. Second, we have $\varphi_p'(r) \to \infty$ for ≥ 1 and $\varphi_p'(r) \to 1/(1{-}p)$ for $p < 1$.

The determination of $N_{p,q}(0)$ needs a more detailed case-by-case study, but is elementary. We obtain

$$N_{p,q}(0) = \begin{cases} 0 & \text{for } q \ge 1 \text{ and } p \ge \max\{0, 1{-}q\}, \\ \frac{q}{q-1} & \text{for } q > 1 \text{ and } p = 0, \\ \frac{1}{1-q} & \text{for } q \in \,]0,1] \text{ and } p = 1{-}q, \\ \frac{-q}{p-1} & \text{for } q < 0 \text{ and } p > 0, \\ \infty & \text{otherwise.} \end{cases}$$

With this, part (a) is established.

To see part (b) we observe $\frac{\mathrm{d}}{\mathrm{d}r}N_{p,p}(r) \le 0$ and find $m_{p,p} = N_{p,p}(\infty) = 1/(1{-}p)$.

Similarly, for part (c) one shows $\frac{\mathrm{d}}{\mathrm{d}r}N_{p,1-p}(r) \le 0$ for $p \in [0, 1/2]$ giving $m_{p,1-p} = N_{p,1-p}(\infty) = 1/(1{-}p)$. Moreover, for $p \in [1/2, 1]$ one shows $\frac{\mathrm{d}}{\mathrm{d}r}N_{p,1-p}(r) \ge 0$, which implies $m_{p,1-p} = N_{p,1-p}(0)$ $= 1/p$. ■

Proof of Corollary 3.8. According to the previous result, we need to find $c_* = c_p$ which is given via

$$\frac{1}{c_p} = \sup_{0<s\neq 1} \Phi(s) \quad \text{with } \Phi(s) := \frac{\varphi_p(s)}{s(\varphi_p'(s))^2}.$$

Observe that $\varphi_{1/2}(s) = 2(\sqrt{s} - 1)^2$ implies $\Phi_{1/2} \equiv 1/2$, and hence $c_{1/2} = 2$.

The derivative of the power-like entropy generator (7) is

$$\varphi_p'(s) = \frac{s^{p-1} - 1}{p-1} \text{ for } p \in \mathbb{R} \setminus \{0,1\}, \quad \varphi_0'(s) = 1 - \frac{1}{s}, \quad \varphi_1'(s) = \log s.$$

For general $p \in \mathbb{R}$, an explicit calculation yields

$$\Phi(s) = \frac{p-1}{p} \cdot \frac{s^p - ps + p - 1}{s(s^{p-1}-1)^2},$$

we easily verify that Φ is continuous at the $s = 1$ and hence continuous on $(0, \infty)$. Moreover, we have $\Phi(s) \to \max\{0, 1-p\}$ for $s \to \infty$. For $s \to 0$ we obtain $\Phi(s) \to \infty$ for $p > 1/2$ and $\Phi(s) \to 0$ for $p < 1/2$.

Thus, we conclude $\sup \Phi = \infty$ for $p > 1/2$. For $p \leq 1/2$ a closer inspection shows that $\sup \Phi = 1-p$. and hence $c_p = 1/(1-p)$ as stated. ■

Proof of Corollary 5.1. By our threshold condition, for $p \in (-\infty, \frac{1}{2}]$, the constant $c_{\mathsf{He}} = \frac{1}{1-p}$ satisfies

$$\left\| \frac{\delta F}{\delta \mu} [\mu] \right\|^2_{L^2_\mu} \geq c_{\mathsf{He}} \cdot (F(\mu) - F(\pi)) .$$

Since the dissipation of the Otto-Wasserstein part is always non-negative, we trivially have

$$\alpha \left\| \nabla \frac{\delta F}{\delta \mu} [\mu] \right\|^2_{L^2_\mu} + \beta \left\| \frac{\delta F}{\delta \mu} [\mu] \right\|^2_{L^2_\mu} \geq \beta c_{\mathsf{He}} \cdot (F(\mu) - F(\pi)) + 0,$$

which is the desired statement. ■

References

[AGS05] L. Ambrosio, N. Gigli, and G. Savare, *Gradient flows: in metric spaces and in the space of probability measures*, Springer Science & Business Media, 2005.

[AM*01] A. Arnold, P. Markowich, G. Toscani, and A. Unterreiter: *On convex Sobolev inequalities and the rate of convergence to equilibrium for Fokker-Planck type equations*. Comm. Partial Diff. Eqns. **26**:1-2 (2001) 43–100.

[Ama98] S.-I. Amari: *Natural gradient works efficiently in learning*. Neural computation **10**:2 (1998) 251–276.

[AmN00] S.-i. Amari and H. Nagaoka, *Methods of information geometry*, vol. 191, American Mathematical Soc., 2000.

[BaÉ85] D. Bakry and M. Émery, *Diffusions hypercontractives*, Séminaire de Probabilités XIX 1983/84 (J. Azéma and M. Yor, eds.), vol. 1123, Springer Berlin Heidelberg, Berlin, Heidelberg, 1985, pp. 177–206.

[BBM16] M. Bauer, M. Bruveris, and P. W. Michor: *Uniqueness of the Fisher-Rao metric on the space of smooth densities*. Bull. Lond. Math. Soc. **48**:3 (2016) 499–506.

[BeT03] A. Beck and M. Teboulle: *Mirror descent and nonlinear projected subgradient methods for convex optimization*. Operations Research Letters **31**:3 (2003) 167–175.

[Bha42] A. Bhattacharya: *On discrimination and divergence*. Proc. Indian Sci. Congress **Part III** (1942) 13.

[BKM17] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe: *Variational inference: A review for statisticians*. Journal of the American Statistical Association **112**:518 (2017) 859–877.

[BlB18] A. Blanchet and J. Bolte: *A family of functional inequalities: Łojasiewicz inequalities and displacement convex functions*. Journal of Functional Analysis **275**:7 (2018) 1650–1673.

[CC*24] J. A. Carrillo, Y. Chen, D. Z. Huang, J. Huang, and D. Wei: *Fisher-Rao gradient flow: Geodesic convexity and functional inequalities*. arXiv:2407.15693 (2024) .

[CG*20] S. Chewi, T. L. Gouic, C. Lu, T. Maunu, and P. Rigollet: *SVGD as a kernelized Wasserstein gradient flow of the Chi-squared divergence*. preprint **arXiv:2006.02509** (2020) .

[Chi22] L. Chizat: *Sparse optimization on measures with over-parameterized gradient descent*. Mathematical Programming **194**:1-2 (2022) 487–532.

[CP*18a] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard: *An interpolating distance between optimal transport and Fisher–Rao metrics*. Foundations of Computational Mathematics **18**:1 (2018) 1–44.

[CP*18b] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard: *Unbalanced optimal transport: Dynamic and Kantorovich formulation*. J. Functional Analysis **274**:11 (2018) 3090–3123 (en).

[Csi67] I. Csiszár: *On information-type measure of difference of probability distributions and indirect observations*. Studia Sci. Math. Hungar. **2** (1967) 299–318.

[DH*16] B. Dai, N. He, H. Dai, and L. Song, *Provable Bayesian inference via particle mirror descent*, Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, PMLR, May 2016, pp. 985–994.

[DoP23] C. Domingo-Enrich and A.-A. Pooladian: *An explicit expansion of the Kullback-Leibler divergence along its Fisher-Rao gradient flow*. Preprint **arXiv:2302.12229** (2023) .

[GaM17] T. O. Gallouët and L. Monsaingeon: *A JKO splitting scheme for Kantorovich–Fisher–Rao gradient flows*. SIAM Journal on Mathematical Analysis **49**:2 (2017) 1100–1130.

[GD*24] E. Gladin, P. Dvurechensky, A. Mielke, and J.-J. Zhu: *Interaction-force transport gradient flows*. arXiv preprint arXiv:2405.17075 (2024) .

[HB*13] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley: *Stochastic variational inference*. Journal of Machine Learning Research (2013) .

[Hel09] E. Hellinger: *Neue Begründung der Theorie quadratischer Formen von unendlich vielen Veränderlichen.*. J. reine angew. Mathematik **1909**:136 (1909) 210–271.

[JG*99] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul: *An introduction to variational methods for graphical models*. Machine Learning **37**:2 (1999) 183–233.

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

[Kak48] S. Kakutani: *On equivalence of infinite product measures.* Annals of Mathematics **49**:1 (1948) 214–224.

[KBB15] W. Krichene, A. Bayen, and P. L. Bartlett, *Accelerated mirror descent in continuous and discrete time*, Advances in Neural Information Processing Systems, vol. 28, Curran Associates, Inc., 2015.

[KhN18] M. E. Khan and D. Nielsen, *Fast yet simple natural-gradient descent for variational inference in complex models*, 2018 International Symposium on Information Theory and Its Applications (ISITA), IEEE, 2018, pp. 31–35.

[KhR23] M. E. Khan and H. Rue: *The Bayesian learning rule*. Preprint **arXiv:2107.04562** (2023) .

[KMV16] S. Kondratyev, L. Monsaingeon, and D. Vorotnikov: *A new optimal transport distance on the space of finite Radon measures*. Adv. Differ. Eqns. **21**:11/12 (2016) 1117–1164.

[KNS20] H. Karimi, J. Nutini, and M. Schmidt: *Linear convergence of gradient and proximal-gradient methods under the Polyak-Lojasiewicz condition*. Preprint **arXiv:1608.04636** (2020) .

[KoV19] S. Kondratyev and D. Vorotnikov: *Spherical Hellinger-Kantorovich gradient flows*. SIAM J. Math. Analysis **51**:3 (2019) 2053–2084.

[KoV20a] ______: *Convex Sobolev inequalities related to unbalanced optimal transport*. J. Diff. Eqns. **268** (2020) 3705–3724.

[KoV20b] ______: *Nonlinear Fokker-Planck equations with reactions as gradient flows of the free energy*. J. Funct. Analysis **278**:108310 (2020) 1–40.

[LaM19] V. Laschos and A. Mielke: *Geometric properties of cones with applications on the Hellinger–Kantorovich space, and a new distance on the space of probability measures*. J. Funct. Analysis **276**:11 (2019) 3529–3576.

[LaM23] ______: *Evolutionary variational inequalites on the Hellinger-Kantorovich and the spherical Hellinger-Kantorovich spaces*. Preprint **arXiv:2207.09815v3** (2023) .

[LC*22] M. Lambert, S. Chewi, F. Bach, S. Bonnabel, and P. Rigollet: *Variational inference via Wasserstein gradient flows*. Preprint **arXiv:2205.15902** (2022) .

[LMS16] M. Liero, A. Mielke, and G. Savaré: *Optimal transport in competition with reaction – the Hellinger–Kantorovich distance and geodesic curves*. SIAM J. Math. Anal. **48**:4 (2016) 2869–2911.

[LMS18] M. Liero, A. Mielke, and G. Savaré: *Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures*. Inventiones mathematicae **211**:3 (2018) 969–1117 (en).

[LMS23a] M. Liero, A. Mielke, and G. Savaré: *Fine properties of geodesics and geodesic λ-convexity for the Hellinger–Kantorovich distance*. Arch. Rat. Mech. Analysis **247**:112 (2023) 1–73.

[LMS23b] L. Liu, M. B. Majka, and L. Szpruch: *Polyak-Lojasiewicz inequality on the space of measures and convergence of mean-field birth-death processes*. Applied Mathematics & Optimization **87**:3 (2023) 48.

[LSW23] Y. Lu, D. Slepčev, and L. Wang: *Birth–death dynamics for sampling: global convergence, approximations and their asymptotics*. Nonlinearity **36**:11 (2023) 5731.

[Mie23] A. Mielke: *An introduction to the analysis of gradients systems*. Preprint arXiv:2306.05026 (2023) .

[Mie24] ______: *On EVI flows in the (spherical) Hellinger-Kantorovich space*. Oberwolfach Reports **21**:1 (2024) 313–318.

[MiM18] A. Mielke and M. Mittnenzweig: *Convergence to equilibrium in energy-reaction–diffusion systems using vector-valued functional inequalities*. J. Nonlinear Science **28**:2 (2018) 765–806.

[NeY83] A. S. Nemirovskij and D. B. Yudin, *Problem complexity and method efficiency in optimization*, Wiley-Interscience, 1983.

[OhT11] S.-i. Ohta and A. Takatsu: *Displacement convexity of generalized relative entropies*. Advances in Mathematics **228**:3 (2011) 1742–1787.

[Ott96] F. Otto, *Double degenerate diffusion equations as steepest descent*, Preprint no. 480, SFB 256, University of Bonn, 1996.

[Ott01] ______: *The geometry of dissipative evolution equations: The porous medium equation*. Comm. Partial Diff. Eqns. **26**:1-2 (2001) 101–174.

[OtV00] F. Otto and C. Villani: *Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality*. Journal of Functional Analysis **173**:2 (2000) 361–400.

[Pel14] M. A. Peletier: *Variational Modelling: Energies, gradient flows, and large deviations*. arXiv:1402.1990 [math-ph] (2014) , arXiv: 1402.1990.

[Rao45a] C. R. Rao: *Information and accuracy attainable in the estimation of statistical parameters*. Bulletin of the Calcutta Mathematical Society **37**:3 (1945) 81–91.

[Rao45b] C. R. Rao: *Information and the accuracy attainable in the estimation of statistical parameters*. Bull. Calcutta Math. Soc. **37** (1945) 81–91.

[Ray73] J. W. S. B. Rayleigh, *Some general theorems relating to vibrations*, London Mathematical Society, 1873.

[RJ*19] G. Rotskoff, S. Jelassi, J. Bruna, and E. Vanden-Eijnden, *Neuron birth-death dynamics accelerates gradient descent and converges asymptotically*, International conference on machine learning, PMLR, 2019, pp. 5508–5517.

[San15] F. Santambrogio: *Optimal transport for applied mathematicians*. Birkäuser, NY **55**:58-63 (2015) 94, Publisher: Springer.

[San17] ______: *{Euclidean, metric, and Wasserstein} gradient flows: An overview*. Bulletin of Mathematical Sciences **7**:1 (2017) 87–154.

[WaJ08] M. J. Wainwright and M. I. Jordan: *Graphical models, exponential families, and variational inference*. Foundations and Trends® in Machine Learning **1**:1–2 (2008) 1–305.

[WWJ16] A. Wibisono, A. C. Wilson, and M. I. Jordan: *A variational perspective on accelerated methods in optimization*. Proceedings of the National Academy of Sciences **113**:47 (2016) .

[YWR24] Y. Yan, K. Wang, and P. Rigollet: *Learning Gaussian mixtures using the Wasserstein-Fisher-Rao gradient flow*. The Annals of Statistics **52**:4 (2024) 1774–1795.

[Zhu24] J.-J. Zhu: *Inclusive KL minimization: A Wasserstein-Fisher-Rao gradient flow perspective*. arXiv preprint arXiv:2411.00214 (2024) .