Approximation, Kernelization, and Entropy-Dissipation of Gradient Flows: from Wasserstein to Fisher-Rao

Jia-Jie Zhu ZHU@WIAS-BERLIN.DE
Alexander Mielke MIELKE@WIAS-BERLIN.DE
Weierstrass Institute for Applied Analysis and Stochastics
Mohrenstrasse 39
10117 Berlin, Germany

Abstract

Motivated by various machine learning applications, we present a principled investigation of gradient flow dissipation geometry, emphasizing the Fisher-Rao type gradient flows and the interplay with Wasserstein space. Using the dynamic Benamou-Brenier formulation, we reveal a few precise connections between those flow dissipation geometries and commonly used machine learning tools such as Stein flows, kernel discrepancies, and nonparametric regression. In addition, we present analysis results in terms of Łojasiewicz type functional inequalities, with an explicit threshold condition for a family of entropy dissipation along the Fisher-Rao flows. Finally, we establish rigorous evolutionary Γ-convergence for the Fisher-Rao type gradient flows obtained via regression, justifying the approximation beyond static point-wise convergence.

Keywords: optimal transport, kernel methods, gradient flow, partial differential equation, Wasserstein, Fisher-Rao, Hellinger, optimization, calculus of variations, variational inference, sampling, generative models,

Contents

1 Introduction **2**

2 Preliminaries **10**
2.1 Gradient-flow systems and geodesics 10
2.2 Reproducing kernel Hilbert space 13

3 Fisher-Rao setting **15**
3.1 Kernelization and approximation of Fisher-Rao gradient flows 15
3.2 MMD as de-kernelized Fisher-Rao distance 17
3.3 Flattened Fisher-Rao, Allen-Cahn, and φ-divergences 19

4 Wasserstein setting **20**
4.1 Gradient structure for the (regularized) Stein gradient flow 20
4.2 De-Stein: de-kernelized Wasserstein geometry 21
4.3 Flattened Wasserstein, Cahn-Hilliard, and Sobolev discrepancy 22

5 Wasserstein-Fisher-Rao setting **22**
5.1 Kernelization and approximation of the WFR gradient flow 23

5.2 De-kernelized Wasserstein-Fisher-Rao and Kernel Sobolev-Fisher 24
5.3 Flattened Wasserstein-Fisher-Rao and kernel-Sobolev-Fisher 24

6 Analysis **25**
6.1 Entropy dissipation and functional inequalities 25
6.2 Kernel discrepancies and entropy dissipation in kernelized geometries 32
6.3 Explicit connection between nonparametric regression and Rayleigh Principle 34
6.4 Energy dissipation in kernel-approximate flows: formal arguments 38
6.5 Evolutionary Γ-convergence at the approximation limit 39

7 Other related works **41**

8 Further proofs **42**

9 Discussion **46**

1 Introduction

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

> 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 (Otto; Jordan et al., 1998) 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, i.e., the solution paths of the measure optimization problem

$$\min_{\mu \in \mathcal{P}(\mathbb{R}^d)} F(\mu) \tag{1.1}$$

in the Wasserstein space $\big(\mathcal{P}(\mathbb{R}^d), W_p\big)$.

While much recent research in machine learning applications has predominantly focused on modifying the *energy functionals* (i.e., F above) of the pure Wasserstein gradient flows, e.g., (Arbel et al., 2019; Chewi et al., 2020; Glaser et al., 2021; Korba et al., 2021; Carrillo et al., 2019; Lu et al., 2023; Javanmard et al., 2020; Craig et al., 2023), we delve into various approximations and kernelizations of the gradient flow *geometry*, beyond the confines of the pure Wasserstein and Fisher-Rao. In doing so, our investigation also reveals precise relations between various previously proposed geometries over probability measures, such as the Stein distance (Duncan et al., 2019; Liu and Wang, 2019), kernel Stein discrepancy (Liu et al.), Sobolev discrepancy (Mroueh and Rigotti, 2020), and maximum mean discrepancy (Gretton et al., 2012).

By working with the different gradient flow geometries and leaving the energy functional to be chosen for specific applications, we provide a measure optimization (1.1) framework to be adapted to various applications beyond the confines of ad-hoc energy functionals, evidenced in the following examples.

Sampling in Stein geometry Suppose a statistician wishes to generate samples from a probability distribution π, whose density is in the form $\pi(x) = \dfrac{1}{\int \mathrm{e}^{-V(x)}\,\mathrm{d}x} e^{-V(x)}$, where V is the potential energy. In this case, using the fact that π is the invariant distribution of the Langevin stochastic differential equation

$$\mathrm{d}X_t = -\nabla V(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}Z_t, \tag{1.2}$$

where Z_t is the standard Brownian motion. From a PDE perspective, this Langevin SDE (1.2) describes the same dynamical system as the deterministic drift-diffusion Fokker-Planck PDE

$$\partial_t \mu = -\operatorname{div}\left(\mu \nabla \left(V + \log \mu\right)\right) \tag{1.3}$$

for probability measure μ, which is also the gradient-flow equation of a Wasserstein gradient flow. Therefore, instead of relying on the stochastic simulation of (1.2), one can forward simulate the deterministic PDE (1.3). Liu and Wang (2019) have proposed a deterministic discrete-time update algorithm referred to as Stein variational gradient descent (SVGD). This algorithm has been related to the Stein PDE by Duncan et al. (2019)

$$\partial_t \mu = -\operatorname{div}\left(\mu \mathcal{K}_\mu \nabla \left(V + \log \mu\right)\right) \tag{1.4}$$

where $\mathcal{K}_\mu$ is the integral operator. The gradient flow equation (1.4) can be viewed as the kernelization of the pure Wasserstein gradient flow equation (1.3).

Variational inference and natural gradient descent One major topic in machine learning research is inferring the posterior distribution of the model parameters $\theta \in \Theta$, given the observed data, i.e., finding $\pi(\theta|\text{Data})$. In practice, the exact posterior distribution is often intractable, and one must resort to approximate variational inference methods (Jordan et al., 1999; Wainwright and Jordan, 2008; Blei et al., 2017). This amounts to finding the approximate posterior probability measure μ by solving

$$\min_{\mu \in \mathcal{P}} \mathrm{D}_{\mathrm{KL}}(\mu|\pi). \tag{1.5}$$

A large class of variational inference methods is based on optimizing a parameterized distribution μ_η, e.g., the Gaussian distribution family. Then, the optimization problem (1.5) is solved by minimizing with respect to the parameter η, often picked as the natural parameters of exponential families. In such cases, an efficient approach is the natural gradient descent (Amari, 1998; Khan and Nielsen, 2018; Hoffman et al., 2013; Khan and Rue, 2023) on η that respects the geometry of the parameterized probability space. In practice, the update rule is an Riemannian gradient descent scheme

$$\eta^{k+1} \leftarrow \operatorname*{argmin}_{\eta} \nabla_\eta F(\mu_{\eta^k})(\eta - \eta^k) + \frac{1}{2\tau}(\eta - \eta^k)^\top G(\eta^k)(\eta - \eta^k), \tag{1.6}$$

where $F(\mu_\eta) = \mathrm{D}_{\mathrm{KL}}(\mu_\eta|\pi)$ in the variational inference context and $\nabla_\eta F(\mu_{\eta^k})$ its gradient with respect to η. The matrix $G(\eta^k) = \int \mu_{\eta^k}(x) \cdot \left(\nabla_\eta \log \mu_{\eta^k}(x)\right) \left(\nabla_\eta \log \mu_{\eta^k}(x)\right)^\top \mathrm{d}x$ is referred to as the Fisher information matrix. In this paper's context, the last (proximal) term in the update rule (1.6) is a quadratic approximation to the squared Fisher-Rao distance between the probability measures when η and η^k are close, i.e.,

$$(\eta - \eta^k)^\top G(\eta^k)(\eta - \eta^k) \approx \mathsf{FR}^2(\mu_{\eta^k}, \mu_\eta).$$

Therefore, those methods correspond to gradient flows in the Fisher-Rao geometry, which is a central topic in this paper. See (Chen et al., 2023) for more details on this connection.

Deep generative models A recent application of gradient flows is generative models. One particular relevant class of algorithms is the score-based deep diffusion generative models (Song et al., 2020; Song and Ermon, 2020; Ho et al., 2020; Sohl-Dickstein et al., 2015; De Bortoli, 2023; Oko et al., 2023). The goal of the so-called *score-matching* task is to compute the vector field $\nabla \log \mu_t$ to simulate a backward SDE

$$\mathrm{d}X_t = \left(X_t + 2\nabla \log \mu_t\left(X_t\right)\right)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t, \tag{1.7}$$

The term $\nabla \log \mu_t\left(X_t\right)$ can be approximated via regression in practice,

$$\inf_{f \in \mathcal{F}} \int_0^T \|f(\cdot, t) - \nabla \log \mu_t\|_{L^2_{\mu_t}}^2 \, \mathrm{d}t, \tag{1.8}$$

where μ_t is the state distribution of a diffusion process at time t, e.g., Ornstein–Uhlenbeck process. Another class of generative models that has shown improved efficiency and stability is the flow-based generative models (Lipman et al., 2022). They learn the solution u to the ODE $\dot{u} = -v_t(u)$ for some velocity field v_t by solving the regression problem with explicit target velocity v_t

$$\inf_{f \in \mathcal{F}} \int_0^T \|f(\cdot, t) - v_t\|_{L^2_{\mu_t}}^2 \, \mathrm{d}t. \tag{1.9}$$

Furthermore, they observed that, by choosing v_t to be the velocity field for the optimal transport between Gaussian distributions, they obtained more efficient and stable training than previous generative models.

The hope of those learning algorithms is to approximate some vector field of the original target flows. Note that in practice, f is often parameterized using a time-dependent neural network and training is further done over various initial conditions. From this paper's perspective, it is crucial to note that the *new flow following the learned velocity field, in the above formulations, has a new geometry that is different from the original flow even if the training error is close to zero.*

Motivated by those applications, we first study in detail the gradient flows in the Fisher-Rao geometry, generated by the Fisher-Rao distance, also known as the Hellinger distance, between two nonnegative measures μ, ν. It is defined as

$$\mathsf{FR}^2(\mu, \nu) = 4 \cdot \int \left(\sqrt{\frac{\delta \mu}{\delta \gamma}} - \sqrt{\frac{\delta \nu}{\delta \gamma}}\right)^2 \mathrm{d}\gamma \tag{1.10}$$

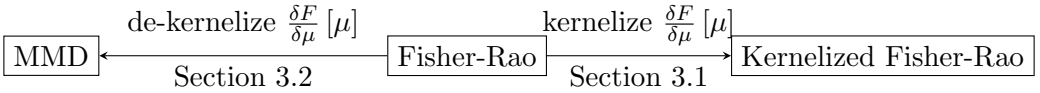
for a reference measure γ, $\mu, \nu << \gamma$. Recall its dynamic formulation [1] (see, e.g., (Gallouët and Monsaingeon, 2017), (Liero et al., 2018))

$$\mathsf{FR}^2(\mu,\nu) = \min_{\mu,\xi_t} \left\{ \int_0^1 \|\xi_t\|_{L^2_\mu}^2 \,\mathrm{d}t \;\middle|\; \dot{\mu} = -\mu\cdot\xi_t,\ \mu(0)=\mu,\ \mu(1)=\nu \right\}. \tag{1.11}$$

Using the tools from kernel methods and the Benaou-Brenier dynamic formulation, we first investigate new gradient systems centered around the Fisher-Rao geometry. Our emphasis is on the precise relation in terms of the kernelization of gradient flows, stated in Definition 3.1. Motivated by the kernelization in the Stein geometry (see, e.g., (1.4)), we provide the geodesic and gradient structures of the *kernelized Fisher-Rao* geometry in Section 3, whose gradient flow equation is a reaction equation with a kernelized growth field

$$\dot{\mu} = -\mu \mathcal{K}_\mu \frac{\delta F}{\delta \mu}[\mu].$$

Furthermore, we find that the kernelization of the kernel MMD (Gretton et al., 2012), commonly used in machine learning applications, results precisely in the pure Fisher-Rao geometry. We summarize the relations below.



The arrows denote kernelization operations of the gradient flow by the operator $\mathcal{K}_\rho^{\frac{1}{2}}$, as in Definition 3.1. Analogously, for the Wasserstein setting, we first continue the work of Duncan et al. (2019) on Stein geometry, of which we briefly provide the gradient structure in Section 4.1. We then establish the kernelization and de-kernelization relation in the diagram below.

For example, we de-kernelize the Wasserstein geometry to obtain the De-Stein distance, which results in a flat distance in the form of a weak norm

$$\text{De-Stein}^2(\mu,\nu) = \sup_{\zeta} \left\{ \int \zeta \,\mathrm{d}(\mu-\nu) - \frac{1}{4}\|\nabla\zeta\|_{\mathcal{H}}^2 \right\},$$

which is the transport analog of the MMD.

Motivated by the approximation of velocity fields in generative models via score-matching (1.8) and flow-based models (1.9), we investigate the gradient flow structure that generates the

1. One may replace the dynamics by $\dot{\mu} = -\mu\cdot\xi_t + \mu\cdot\int\mu\xi_t$ for a flow over probability measures, i.e., spherical Hellinger (Laschos and Mielke, 2019), instead of non-negative measures. We do not consider this flow in this paper mainly due to technicality and additional steps in the analysis. However, many of our results generalize directly to the spherical version.

reaction equation whose growth field is obtained by nonparametric regression, e.g., the kernel ridge regression,

$$\dot{\mu_t} = -\mu_t \cdot r_t, \quad r_t = \operatorname*{argmin}_{f\in\mathcal{F}}\left\{ \left\| f - \frac{\delta F}{\delta \mu}[\mu_t] \right\|^2_{L^2_{\mu_t}} + \lambda \|f\|^2_{\mathcal{F}} \right\}. \tag{1.12}$$

Note that the nonparametric regression can be cast in the more general maxium likelihood estimation (MLE) form (6.26). We refer to the (gradient) system that generates the equation above as an *approximate* Fisher-Rao (gradient) system.

The analog in the Wasserstein setting has been studied under the name of (regularized) Stein gradient flow by Duncan et al. (2019); He et al. (2022). Combining the Fisher-Rao and Wasserstein settings, we find the approximate Wasserstein-Fisher-Rao flow

$$\begin{aligned} \dot{\mu_t} = \operatorname{div}(\mu_t \cdot v_t) - \mu_t \cdot r_t, \quad r_t &= \operatorname*{argmin}_{f\in\mathcal{F}}\left\{ \left\| f - \frac{\delta F}{\delta \mu}[\mu_t] \right\|^2_{L^2_{\mu_t}} + \lambda \|f\|^2_{\mathcal{F}} \right\}, \\ v_t &= \operatorname*{argmin}_{f\in\mathcal{F}}\left\{ \| f - \nabla \frac{\delta F}{\delta \mu}[\mu_t] \|^2_{L^2_{\mu_t}} + \lambda \|f\|^2_{\mathcal{F}} \right\}. \end{aligned}$$

The connection between nonparametric regression and gradient flows can be seen in the Helmholtz-Rayleigh Principle (Rayleigh, 1873); see Section 6.3. The intuition is that the principle of

$$\min\left\{ \text{energy} + \text{dissipation potential} \right\}$$

is equivalent to the nonparametric regression formulation (1.12) agnostic of the dissipation geometry, i.e., in both the Fisher-Rao and Wasserstein settings. This implies different training objectives, such as matching the score function or the log density, can be unified using the same formalism of the Rayleigh Principle.

Historically, the entropy dissipation method is used in proving the Bakry-Émery theorem, e.g., in (Arnold et al., 2001). It gives rise to a sufficient condition closely related to the Polyak-Łojasiewicz inequality in optimization

$$\mathcal{R}(\mu, \dot{\mu}) + \mathcal{R}^*(\mu, -\mathrm{D}F) \geq c \cdot \left(F(\mu(t)) - \inf_{\mu} F(\mu) \right), \tag{1.13}$$

where the dissipation potential terms $\mathcal{R}, \mathcal{R}^*$ must be adapted to the specific gradient-flow geometry. Beyond standard KL-dissipation in the Wasserstein geometry, researchers have also studied specialized versions of the log-Sobolev inequality (LSI), e.g., the Stein-log-Sobolev inequality (Duncan et al., 2019, Lemma 35). In Table 1, we provide generalized Łojasiewicz inequalities specialized to the various geometries studied in this paper. In addition, one immediately obtains other functional inequalities, such as the LSI-type by setting $F(\mu) = \mathrm{D}_{\mathrm{KL}}(\mu|\pi)$ in Wasserstein type geometries.

A commonly used energy functional in applications is the *φ-divergences* (Csiszár, 1967)

$$\mathrm{D}_\varphi(\mu|\nu) := \int \varphi(\sigma(x))\, \mathrm{d}\nu, \quad \sigma = \frac{\mathrm{d}\mu}{\mathrm{d}\nu}. \tag{1.14}$$

Gradient system	Łojasiewicz-type inequality ($c > 0$)
Kernelized FR	$\|\mathcal{K}_\mu^{\frac{1}{2}} \frac{\delta F}{\delta \mu}[\mu]\|_{L^2_\mu}^2 \geq c \cdot (F(\mu(t)) - \inf_\mu F(\mu))$
Kernel-approx. FR	$\|(\mathcal{K}_\mu + \lambda \operatorname{Id})^{-\frac{1}{2}} \mathcal{K}_\mu \frac{\delta F}{\delta \mu}[\mu]\|_{L^2_\mu}^2 \geq c \cdot (F(\mu(t)) - \inf_\mu F(\mu))$
MMD	$\|\frac{\delta F}{\delta \mu}[\mu]\|_{\mathcal{H}}^2 \geq c \cdot (F(\mu(t)) - \inf_\mu F(\mu))$
Stein	$\|\mathcal{K}_\mu^{\frac{1}{2}} \nabla \frac{\delta F}{\delta \mu}[\mu]\|_{L^2_\mu}^2 \geq c \cdot (F(\mu(t)) - \inf_\mu F(\mu))$
Kernel-approx. W	$\|(\mathcal{K}_\mu^{\frac{1}{2}} + \lambda \operatorname{Id})^{-1} \mathcal{K}_\mu \nabla \frac{\delta F}{\delta \mu}[\mu]\|_{L^2_\mu}^2 \geq c \cdot (F(\mu(t)) - \inf_\mu F(\mu))$
WFR	$\|\frac{\delta F}{\delta \mu}[\mu]\|_{L^2_\mu}^2 + \|\nabla \frac{\delta F}{\delta \mu}[\mu]\|_{L^2_\mu}^2 \geq c \cdot (F(\mu(t)) - \inf_\mu F(\mu))$
De-Stein	$\|\nabla \frac{\delta F}{\delta \mu}[\mu]\|_{\mathcal{H}}^2 \geq c \cdot (F(\mu(t)) - \inf_\mu F(\mu))$
Kernelized WFR	$\|\mathcal{K}_\mu^{\frac{1}{2}} \frac{\delta F}{\delta \mu}[\mu]\|_{L^2_\mu}^2 + \|\mathcal{K}_\mu^{\frac{1}{2}} \nabla \frac{\delta F}{\delta \mu}[\mu]\|_{L^2_\mu}^2 \geq c \cdot (F(\mu(t)) - \inf_\mu F(\mu))$
K-approx. WFR	$\|(\mathcal{K}_\mu + \lambda \operatorname{Id})^{-\frac{1}{2}} \mathcal{K}_\mu \nabla \frac{\delta F}{\delta \mu}[\mu]\|_{L^2_\mu}^2 + \|(\mathcal{K}_\mu + \lambda \operatorname{Id})^{-\frac{1}{2}} \mathcal{K}_\mu \frac{\delta F}{\delta \mu}[\mu]\|_{L^2_\mu}^2 \geq c \cdot (F(\mu(t)) - \inf_\mu F(\mu))$

Table 1: Summary of the Łojasiewicz inequalities for different gradient systems

where $\varphi : [0, +\infty) \to [0, +\infty]$ is a convex entropy generator function. We delve specifically into the concrete instantiations of the Łojasiewicz inequality for the following power-like entropy generator (see e.g. (Liero et al., 2018)).

$$\begin{aligned} \varphi_p(s) &:= \frac{1}{p(p-1)}\big(s^p - p(s-1) - 1\big), \quad p \in \mathbb{R} \setminus \{0, 1\}, \\ \varphi_0(s) &:= s - 1 - \log s, \quad \varphi_1(s) := \varphi_{\mathrm{KL}} = s \log s - s + 1. \end{aligned} \tag{1.15}$$

Note that by this definition and our scaling of the Fisher-Rao distance (1.10), we have

$$\frac{1}{2}\mathsf{FR}^2(\mu, \nu) = \int \varphi_{\frac{1}{2}}\left(\frac{\delta \mu}{\delta \nu}\right) \mathrm{d}\nu. \tag{1.16}$$

Slightly abusing the terminology, we still refer to power-like entropy generated by $\varphi_{\frac{1}{2}}$ as the squared FR distance. We plot the corresponding entropy generator functions in Figure 1. Alternatively, one may use Hellinger's integral to define the α-divergence $\mathrm{D}_\alpha(\mu|\nu) := \frac{4}{1-\alpha^2}(1 - \int \mu^{\frac{1+\alpha}{2}} \nu^{\frac{1-\alpha}{2}})$, from which one obtains the KL, reverse KL, and the Fisher-Rao as special cases.

First, we examine a few concrete cases of the Łojasiewicz inequality for the Fisher-Rao and Wasserstein geometry, e.g., the lack of global Łojasiewicz inequality for the Fisher-Rao flow of KL energy, and the unconditionally satisfied Łojasiewicz inequality for the squared Fisher-Rao energy. From the perspective of optimization, the global Łojasiewicz inequality is arguably more nontrivial to characterize than the local version since we need to create enough "slope" for the gradient flow to escape the initial birth from zero mass. See the illustration in Figure 4. To that end, we extract an explicit condition for global convergence in terms of a power threshold when the energy is chosen as the power-like entropy functional φ_p in (1.15). See the summary in Table 2 for our current knowledge of the global Łojasiewicz inequality in the Fisher-Rao and Wasserstein setting. Our perspective here established that

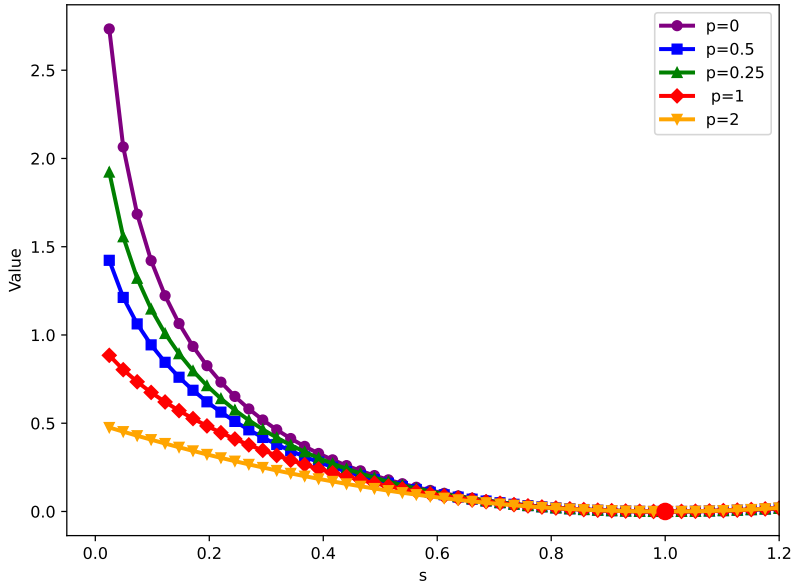


Figure 1: The plot illustrates the power-like entropy generator functions φ_p for different values of p. The purple curve represents $p = 0$ (reverse KL), the green curve represents $p = 0.25$, the blue curve represents $p = 0.5$ (FR), the red curve represents $p = 1$ (KL), and the orange curve represents $p = 2$ (χ^2). The large red dot represents the equilibrium at $s = 1$. The plot helps visualize the behavior of the functions for different values of p and provides insights into their convexity and slopes.

Order p of entropy D_{φ_p}	Gradient-flow geometry	Functional inequality
$[1, 2]$	Wasserstein (Bakry-Émery)	(BE), $c > 0 \implies$ Ł, $c > 0$
$[\frac{d-1}{d}, \infty)$	Wasserstein (McCann cond.)	geod. cvx $\implies$ Ł with $c \geq 0$
$[\frac{d-1}{d}, \frac{1}{2}] \cup (1, \infty)$	WFR (Liero et al., 2023)	geod. cvx $\implies$ Ł with $c \geq 0$
$(-\infty, \frac{1}{2}]$	Fisher-Rao (Corollary 6.5)	(Ł-FR) with $c_* = \dfrac{1}{1-p}$
$[\frac{d-1}{d}, \frac{1}{2}]$	WFR (Corollary 6.7)	Ł with $c_* \geq \dfrac{1}{1-p}$

Table 2: Summary of Łojasiewicz inequalities for power-like entropy energy functionals in different geometries. The Bakry-Émery case for $p = 1$ is the well-known (LSI).

1) the Fisher-Rao energy functional creates enough slope for the birth of mass from zero in the Fisher-Rao geometry; 2) the reverse KL entropy (i.e., 0-th order power-like entropy) creates a singularity near zero and thus a large slope for mass creation.

In the kernelized geometry, it is clear that energies that fail our threshold condition will not satisfy the global Łojasiewicz inequality, as discussed in Lemma 6.8 and Corollary 6.9

for bounded kernels. While sufficient conditions, likely placing restrictions on the kernels, are unclear, we uncover a few interesting kernel discrepancy functionals that are interaction energies obtained by dissipating entropies along the kernelized FR and Stein flow. Two prominent cases are the MMD (Gretton et al., 2012) and the KSD (Liu et al.). They are both generated by the dissipation in kernelized geometries of the reverse KL entropy.

Finally, we concern ourselves with the quality of the approximation in our approximate gradient flows. For example, in regression problems that appeared in generative models, e.g., the flow-matching problem (1.9) and the score-matching problem (1.8), does the learned flow exist, is it a gradient flow? If so, what is the gradient structure, e.g., energy and dissipation geometry? Furthermore, in nonparametric regression (1.12), one typically bounds the prediction error, i.e., quantities such as $\|r_t - \log \mu_t\|_{L^2_{\mu_t}}$ for a fixed time t. However, in gradient flows, we are also interested in the behavior of the system that follows the flow $\dot{\mu_t} = -\mu_t \cdot r_t$, e.g., its variational structure, solution existence, and convergence behavior. For those reasons, we establish the evolutionary Γ-convergence in the kernel-approximate Fisher-Rao geometries, i.e., we use tools from the calculus of variations and functional analysis instead of statistical bounds. In a nutshell, we establish

$$\text{the system generating } \dot{\mu_t} = -\mu_t \cdot r_t \xrightarrow{\Gamma} \text{Fisher-Rao gradient system.}$$

Thus, we provide a rigorous justification for the quality of approximation using nonparametric regression in the Fisher-Rao flows. This also differs from the perspective of local regression and local smoothing analyzed in the works by, e.g., Lu et al. (2023); Carrillo et al. (2019).

Organization of the paper In Section 2, we provide background on gradient systems and optimal transport, with a focus on dynamic formulation and geodesics. Then, we provide background on reproducing kernel Hilbert spaces (RKHS). Those two topics are married through our concrete investigation in the next four sections. In Section 3, 4, and 5, we provide gradient structures for a few new and existing gradient systems of interest. They are motivated by two types of geometries, namely, the Fisher-Rao and Wasserstein space. We characterize their precise relations with other kenelized and approximating geometries. Section 6 is dedicated to the analysis of evolutionary behaviors in the gradient systems using the celebrated Polyak-Łojasiewicz inequalities. Then, we analyze the approximation quality, by proving the evolutionary Γ-convergence of the kernel-approximate Fisher-Rao gradient systems. Additional proofs are given in Section 8. In Section 9, we conclude the paper and mention several future directions and implications on practical computational and learning algorithms.

Notation We use the notation $\mathcal{P}(\bar{\Omega}), \mathcal{M}^+(\bar{\Omega})$ to denote the space of probability and nonnegative measures on the closure of a set $\Omega \subset \mathbb{R}^d$. The base space symbol Ω is often dropped if there is no ambiguity in the context. We express the standard integral operator weighted by measure ρ as a weighted convolution $\mathcal{K}_\rho : L^2(\rho) \to L^2(\rho), f \mapsto \int k(x, \cdot) f(x) \,\mathrm{d}\rho(x)$; cf. Theorem 2.2. The measure ρ is omitted if it is the Lebesgue measure. In this paper, the first variation of a functional F at $\mu \in \mathcal{M}^+$ is defined as a function $\frac{\delta F}{\delta \mu}[\mu]$

$$\frac{\mathrm{d}}{\mathrm{d}\epsilon} F(\mu + \epsilon \cdot v)|_{\epsilon=0} = \int \frac{\delta F}{\delta \mu}[\mu](x) \,\mathrm{d}v(x) \tag{1.17}$$

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

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

where the small-o notation signifies that the term vanishes more rapidly than the term inside the parentheses. We use superscripts for differential derivatives to emphasize the corresponding space of those operations, i.e., we distinguish between $\mathrm{D}^X F$ and $\mathrm{D}^Y F$. 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. The infimal convolution (inf-convolution) of two functions f, g on Banach spaces is defined as $(f \square g)(x) = \inf_y \{f(y) + g(x - y)\}$. In our 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 (Ambrosio et al., 2005).

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 to the case that the dissipation law is linear, which means it 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 ordinary differential equation in the Euclidean space $\dot{u}(t) = -\nabla F(u(t)), u(t) \in \mathbb{R}^d$ 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 potential $\mathcal{R}$. 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, and satisfies $\mathcal{R}(u, 0) = 0$. We denote by

$$\mathcal{R}^*(u, \xi) = \sup\left\{ \langle \xi, v \rangle - \mathcal{R}(u, v) \,\middle|\, v \in \mathrm{T}_u X \right\} \tag{2.1}$$

the (partial) Legendre transform of $\mathcal{R}$ and call it the *dual dissipation potential*. Throughout this work, we will restrict to 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 function, and $\mathcal{R}$ is a dissipation potential. The associated gradient-flow equation has the primal and dual form*

$$0 = \mathrm{D}_{\dot{u}} \mathcal{R}(u, \dot{u}) + \mathrm{D}F(u) \quad \Longleftrightarrow \quad \dot{u} = \mathrm{D}_\xi \mathcal{R}^*\big(u, -\mathrm{D}F(u)\big). \tag{2.2}$$

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

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

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

Of particular interest to this paper is the gradient flow in the Fisher-Rao metric space, also called *Hellinger-Kakutani* or simple *Hellinger space* in (Liero et al., 2018; Laschos and Mielke, 2019), which is the gradient system that generates the following reaction equation as its *gradient flow equation*

$$\partial_t \mu = -\mu \cdot \frac{\delta F}{\delta \mu}[\mu], \tag{2.3}$$

where $\frac{\delta F}{\delta \mu}[\mu]$ is the first variation (1.17). Alternatively, one can also view the whole r.h.s as the Fisher-Rao metric gradient using the weighted tangent space L^2_μ.

The Fisher-Rao gradient system is a special case of general gradient flow in metric spaces Ambrosio et al. (2005), which has gained significant attention in recent machine learning literature due to the study of Wasserstein gradient flow (WGF), originated from the seminal works of Otto and colleagues, e.g., Otto (1996); Jordan et al. (1998); Otto. Rigorous characterizations of general metric gradient systems have been carried out in PDE literature, for which we refer to Ambrosio et al. (2005) for complete treatments and Santambrogio (2015); Peletier (2014); Mielke (2023) 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 Fisher-Rao and Wasserstein respectively.

Example 2.1 (Classical PDE: Allen-Cahn and Cahn-Hilliard) *Recall the Allen-Cahn PDE*

$$\partial_t \mu = \Delta \mu - \nabla V, \tag{2.4}$$

and the Cahn-Hilliard PDE

$$\partial_t \mu = \Delta\left(-\Delta \mu + \nabla V\right). \tag{2.5}$$

They are the gradient flows of the energy functional $F(\mu) = \frac{1}{2}\int |\nabla \mu|^2 + \int V(\mu)$ in two different Hilbert spaces, where V is a potential function, e.g., the double-well potential $V(x) = \frac{1}{4}(1-x^2)^2$. Allen-Cahn is the Hilbert-space gradient flow of the energy F in unweighted L^2, i.e.,

$$\mathcal{R}_{\mathrm{AC}}(\mu, \dot{u}) = \frac{1}{2}\|\dot{u}\|^2_{L^2}, \mathcal{R}^*_{\mathrm{AC}}(\mu, \xi) = \frac{1}{2}\|\xi\|^2_{L^2}. \tag{2.6}$$

Cahn-Hilliard is the gradient flow of F in unweighted H^{-1}, i.e.,

$$\mathcal{R}_{\mathrm{CH}}(\rho, \dot{u}) = \frac{1}{2}\|\dot{u}\|^2_{H^{-1}}, \mathcal{R}^*_{\mathrm{CH}}(\rho, \xi) = \frac{1}{2}\|\nabla \xi\|^2_{L^2}. \tag{2.7}$$

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:

$$\gamma_{u_0\to u_1} \in \operatorname*{argmin}_u \int_0^1 \frac{1}{2}\langle \mathbb{G}(u(s))\dot{u}(s), \dot{u}(s)\rangle \,\mathrm{d}s.$$

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

$$\left.\begin{cases} \dot{u} = \partial_\xi H(u,\xi) = \partial_\xi \mathcal{R}^*(u,\xi) = \mathbb{K}(u)\xi, \\ \dot{\xi} = -\mathrm{D}_u H(u,\xi) = -\mathrm{D}_u \mathcal{R}^*(u,\xi), \end{cases}\right\}, \quad u(0) = u_0,\ u(1) = u_1. \tag{H}$$

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 linear in the state u, because then $\mathrm{D}_u\mathcal{R}^*(u,\xi)$ no longer depends on u. This means that the system (H) decouples in the sense that the equation for ξ no longer depends on u. This particular case occurs in the Wasserstein, Fisher-Rao, and consequently 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 2.2 (Wasserstein geodesics in Hamiltonian formulation) *For the Wasserstein case, the dual dissipation potential takes the form* [2]

$$H(\mu,\xi) = \mathcal{R}^*_{W_2}(\mu,\xi) = \frac{1}{2}\|\nabla\xi\|^2_{L^2_\mu} = \int \frac{1}{2}|\nabla\xi|^2 \mathrm{d}\mu.$$

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

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

Here, the first equation is the continuity equation that 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*

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

yielding the celebrated dual Kantorovich formulation of the Wasserstein distance. See Ambrosio et al. (2005) for details.

2. For ease of calculation, we always consider the $\frac{1}{2}$ scaling for quadratic dissipation potentials. That is, this case corresponds to the geodesics of the $\frac{1}{2}W_2^2$.

Example 2.3 (Fisher-Rao (or Hellinger) geodesics in Hamiltonian formulation) *For the Fisher-Rao case in* (1.11)*, the primal and dual dissipation potential takes the form*

$$\begin{aligned}\mathcal{R}_{\mathsf{FR}}(\mu,\dot{u}) &= \frac{1}{2}\|\frac{\delta \dot{u}}{\delta \mu}\|^2_{L^2_\mu},\\ H(\mu,\xi) &= \mathcal{R}^*_{\mathsf{FR}}(\mu,\xi) = \frac{1}{2}\big\|\xi\big\|^2_{L^2_\mu} = \int \frac{1}{2}\xi^2 \mathrm{d}\mu.\end{aligned} \tag{2.8}$$

The Onsager operator is given by $\mathbb{K}(\mu)\xi = \xi\mu$ *and the geodesic curves are characterized by*

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

Different from the Hamilton-Jacobi setting, 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 already used the initial condition $\mu(0)=\mu_0$*. Applying the final condition* $\mu(1)=\mu_1$ *we arrive at the explicit representation of the Fisher-Rao geodesic*

$$\omega(s) = \big((1-s)\sqrt{\mu_0} + s\sqrt{\mu_1}\big)^2 = (1-s)^2\mu_0 + 2s(1-s)\sqrt{\mu_0\mu_1} + s^2\mu_1. \tag{2.9}$$

See (Laschos and Mielke, 2019) for details. Formally, one can also obtain a static dual Kantorovich type formulation using the closed-form solution above

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

2.2 Reproducing kernel Hilbert space

We first remember some basic facts about the reproducing kernel Hilbert spaces (RKHS), which are a class of Hilbert spaces that are widely used in approximation theory (Wendland, 2004; Cucker and Zhou, 2007) and machine learning (Steinwart and Christmann, 2008).

In this paper, we refer to a bi-variate function $k : \Omega \times \Omega \to \mathbb{R}$ as a symmetric positive definite kernel if k is symmetric and, for all $n \in \mathbb{N}, \alpha_1, \ldots, \alpha_n \in \mathbb{R}$ and all $x_1, \ldots, x_n \in \Omega$, we have $\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j k(x_j, x_i) \geq 0$. If the inequality is strict, then k is called strictly positive definite. Here, the space Ω can be a subset of $\mathbb{R}^d$. k is a reproducing kernel if it satisfies the reproducing property, i.e., for all $x \in X$ and all functions in a Hilbert space $f \in \mathcal{H}$, we have $f(x) = \langle f, k(\cdot, x)\rangle_{\mathcal{H}}$. Furthermore, the space $\mathcal{H}$ is an RKHS if the Dirac map $\delta_x : \mathcal{H} \mapsto \mathbb{R}, \delta_x(f) := f(x)$ is continuous. It can be shown that there is a one-to-one correspondence between the RKHS $\mathcal{H}$ and the reproducing kernel k. The following fact regarding the RKHS, whose statement is adapted from (Steinwart and Christmann, 2008, Theorem 4.27), is instrumental to this paper.

Theorem 2.2 (Integral operator) *Suppose the kernel is square-integrable* $\|k\|^2_{L^2_\rho} := \int k(x,x)d\rho(x) < \infty$ *w.r.t. a probability measure* ρ*. Then the inclusion from the the associated RKHS* $\mathcal{H}$ *to* L^2_ρ*,* $\mathrm{Id} : \mathcal{H} \to L^2_\rho$*, is continuous. Moreover, its adjoint is the operator*

$\mathcal{T}_{k,\rho} : L^2_\rho \to \mathcal{H}$ *defined by*

$$\mathcal{T}_{k,\rho} g(x) := \int k\left(x, x'\right) g\left(x'\right) d\rho\left(x'\right), \quad g \in L^2_\rho$$

$\mathcal{T}_{k,\rho}$ *is Hilbert-Schmidt (i.e., singular values are square-summable). The integral operator*

$$\mathcal{K}_\rho := \mathrm{Id} \circ \mathcal{T}_{k,\rho}, L^2(\rho) \to L^2(\rho)$$

is compact, positive, self-adjoint, and nuclear (i.e., singular values are summable).

We define the power of the integral operator $\mathcal{K}_\rho$ as, for $\alpha > 0$, $\mathcal{K}_\rho^\alpha := \sum_{i=1}^\infty \lambda_i^\alpha \langle \cdot, \phi_i \rangle_{L^2_\rho} \phi_i$, where λ_i and ϕ_i are the eigenvalues and eigenfunctions of $\mathcal{K}_\rho$ given by the spectral theorem. The image of the square root integral operator is the RKHS, i.e., $\mathcal{H} = \mathcal{K}_\rho^{\frac{1}{2}}(L^2_\rho)$ for some probability measure ρ. See, e.g., (Cucker and Zhou, 2007, Chapter 4). Then, for any $g \in \mathcal{H}$, $\exists f \in L^2_\rho$ such that $g = \mathcal{K}_\rho^{\frac{1}{2}} f, \|g\|_\mathcal{H} = \|f\|_{L^2_\rho}$. Therefore, we can conveniently write down some formal relations between the RKHS and L^2 norm useful for our later analysis

$$\|g\|^2_\mathcal{H} = \|\mathcal{K}_\rho^{-\frac{1}{2}} g\|^2_{L^2_\rho} = \langle g, \mathcal{K}_\rho^{-1} g \rangle_{L^2_\rho}, \quad \langle g, \mathcal{T}_{k,\rho} f \rangle_\mathcal{H} = \langle \mathrm{Id}\, g, f \rangle_{L^2_\rho}, \tag{2.10}$$

where $\mathcal{K}_\rho^{-\frac{1}{2}}$ denotes the inverse of $\mathcal{K}_\rho^{\frac{1}{2}}$. The power of the kernel integral operator is prominently manifested in the following nonparametric regression problem.

Lemma 2.3 (Kernel ridge regression estimator) *Given the target function $\xi \in L^2_\rho$, the kernel ridge regression (KRR) problem for $\lambda > 0$*

$$\inf_g \left\{ \|g - \xi\|^2_{L^2_\rho} + \lambda \|g\|^2_\mathcal{H} \right\}, \tag{2.11}$$

admits the closed-form solution

$$g^* = (\mathcal{K}_\rho + \lambda\, \mathrm{Id})^{-1} \mathcal{K}_\rho \xi. \tag{2.12}$$

To set the stage for our derivation later, we establish below an alternative optimization formulation of the KRR solution.

Lemma 2.4 (Alternative optimization problem of KRR estimation) *The KRR solution* (2.12) *coincides with the solution of the optimization problem*

$$\inf_f \left\{ \langle f - \xi, \mathcal{K}_\rho (f - \xi) \rangle_{L^2(\rho)} + \lambda \|f\|^2_{L^2(\rho)} \right\}.$$

One prominent applications of kernel methods to machine learning is the kernel maximum mean discrepancy (MMD) (Gretton et al., 2012), for measuring the discrepancy between probability measures. It is a special case of the integral probability metric (IPM) family that is defined as a weak norm

$$\mathrm{MMD}(\mu, \nu) := \sup_{\|f\|_\mathcal{H} \le 1} \int f \,\mathrm{d}(\mu - \nu), \tag{2.13}$$

where $\mathcal{H}$ is the RKHS associated with the kernel k. MMD is a metric on the space of probability measures if the kernel is positive definite. Furthermore, its advatage lies in its simple structure as a Hilbert space norm

$$\mathrm{MMD}^2(\mu,\nu) = \|\mathcal{K}(\mu-\nu)\|_{\mathcal{H}}^2 = \int\int k(x,x')\,\mathrm{d}(\mu-\nu)(x)\,\mathrm{d}(\mu-\nu)(x'). \tag{2.14}$$

This linear-in-measure form allows for efficient computation of the MMD via Monte Carlo sampling. It can also be viewed as a form of interaction energy (Ambrosio et al., 2005) that dissipates in gradient-flow geometry, e.g., the Wasserstein flow of the MMD energy (Arbel et al., 2019). One of our contributions is to provide a precise relation between the MMD and the Fisher-Rao geometry from two different perspectives in Theorem 3.7 and Proposition 6.10.

3 Fisher-Rao setting

This section addresses one of the main subjects of this paper, the Fisher-Rao-type gradient flow geometry. We first study the kernelized Fisher-Rao geometry in Section 3.1 and provide the gradient structure of the resulting kernelized Fisher-Rao gradient flow. Its growth field is an approximation to that of the Fisher-Rao. In Section 3.2, we perform the inverse operation to "de-kernelize" the Fisher-Rao geometry. Consequently, we obtain a flat (in the sense of Riemannian manifold) gradient-flow geometry, which we show is equivalent to the MMD.

3.1 Kernelization and approximation of Fisher-Rao gradient flows

In the machine learning literature, the term *kernelization* has been used in many contexts. We first make precise what kernelization entails in this paper through the following operation on the dual dissipation potentials of gradient systems.

Definition 3.1 (Kernelization of gradient systems) *Given a gradient system defined by $(X, F, \mathcal{R}^*)$, where $\mathcal{R}^*$ is the dual dissipation potential, we say its force-kernelization counterpart is $(X, F, \mathcal{R}^*_{F-k})$, where the force-kernelized dual dissipation potential is defined by*

$$\mathcal{R}^*_{F-k}(u,\xi) := \mathcal{R}(u, \mathcal{K}_u^{\frac{1}{2}}\,\xi). \tag{3.1}$$

If the original $\mathcal{R}^*$ depends on the generalized force ξ only through its gradient $\nabla\xi$, denoted by $\widetilde{\mathcal{R}^*}(\nabla\xi) = \mathcal{R}^*(\xi)$. Then, its velocity-kernelization is $(X, F, \mathcal{R}^*_{V-k})$,

$$\mathcal{R}^*_{V-k}(u,\xi) := \widetilde{\mathcal{R}^*}(u, \mathcal{K}_u^{\frac{1}{2}}\,\nabla\xi). \tag{3.2}$$

Equivalently for the Fisher-Rao and Wasserstein type flows, we can define kernelization using the operator $\mathcal{T}_{k,u}$ in Theorem 2.2 as a change the dissipation potentials

$$\text{Fisher-Rao: } \mathcal{R}^*(\rho,\xi) = \frac{1}{2}\langle\xi,\xi\rangle_{L^2_\rho} \longrightarrow \mathcal{R}^*_k(\rho,\xi) = \frac{1}{2}\langle\mathcal{T}_{k,\rho}\xi, \mathcal{T}_{k,\rho}\xi\rangle_{\mathcal{H}},$$
$$\text{Wasserstein: } \mathcal{R}^*(\rho,\xi) = \frac{1}{2}\langle\nabla\xi,\nabla\xi\rangle_{L^2_\rho} \longrightarrow \mathcal{R}^*_k(\rho,\xi) = \frac{1}{2}\langle\mathcal{T}_{k,\rho}\nabla\xi, \mathcal{T}_{k,\rho}\nabla\xi\rangle_{\mathcal{H}}.$$

Note that velocity-kernelization of Wasserstein gradient flow, under the name of Stein geometry, has already been investigated, e.g., by Duncan et al. (2019). As overviewed in Section 1, we will construct systems in the kernelization relation illustrated in Figure 1.

Using the above definition for kernelization, we now construct a new geometry by force-kernelizing the Fisher-Rao gradient system.

Definition 3.2 (Dynamic formulation of kernelized Fisher-Rao distance) *The kernelized Fisher-Rao distance is defined by the following dynamic formulation*

$$\mathsf{FR}_k^2(\mu,\nu)=\min_{\mu,\xi_t}\left\{\int_0^1 \|\mathcal{K}_\mu\xi_t\|_{\mathcal{H}}^2\,\mathrm{d}t \;\middle|\; \dot\mu=-\mu\mathcal{K}_\mu\xi_t,\ \mu(0)=\mu,\ \mu(1)=\nu\right\}. \tag{3.3}$$

From the Hamiltonian perspective in (H), we can derive the geodesic equation

$$\begin{cases}\dot\mu=-\mu\xi,\\ \dot\xi=-\xi\cdot\mathcal{K}_{\mu_t}\xi.\end{cases} \tag{Geod-FR-k}$$

However, is important to note that the geodesic equation above is only the necessary condition for optimality. That is, we have not proved whether its solution exists, in contrast to the Hamiltonian system (Geod-FR). This is due to the coupling introduced by the state-dependent integral operator $\mathcal{K}_{\mu_t}$. Furthermore, as a technical point, the integral operator used here is defined w.r.t a non-negative measure rather than a probability measure; see, e.g., (Conway, 1985). In the gradient structure of the kernelized Fisher-Rao gradient system, the corresponding primal and dual dissipation potentials are

$$\mathcal{R}_{\mathsf{FR}_k}(\rho,u)=\frac{1}{2}\|\frac{\delta u}{\delta\rho}\|_{\mathcal{H}}^2,\quad \mathcal{R}^*_{\mathsf{FR}_k}(\rho,\xi)=\frac{1}{2}\langle\xi,\mathcal{K}_\rho\xi\rangle_{L^2_\rho}=\frac{1}{2}\|\mathcal{K}_\rho\xi\|_{\mathcal{H}}^2. \tag{3.4}$$

Therefore, the gradient-flow equation of the FR_k gradient system $(\mathcal{M}^+,F,\mathsf{FR}_k)$ is the reaction equation *kernelized growth field*

$$\dot\mu=-\mu\mathcal{K}_\mu\frac{\delta F}{\delta\mu}[\mu]. \tag{3.5}$$

Going beyond kernelization, we derive the approximation to the original Fisher-Rao dynamics by constructing the following regularized dissipation geometry, i.e., adding the kernelized Fisher-Rao dissipation potential $\mathcal{R}_{\mathsf{FR}_k}$ (3.4) to that of the pure Fisher-Rao $\mathcal{R}_{\mathsf{FR}}$ (2.8)

$$\begin{aligned}\mathcal{R}_{\lambda\text{-}\mathsf{FR}_k}(\rho,\dot u)&:=\mathcal{R}_{\mathsf{FR}}+\lambda\cdot\mathcal{R}_{\mathsf{FR}_k}=\frac{1}{2}\|\frac{\delta\dot u}{\delta\rho}\|_{L^2_\rho}^2+\frac{\lambda}{2}\|\frac{\delta\dot u}{\delta\rho}\|_{\mathcal{H}}^2=\frac{1}{2}\langle\frac{\delta\dot u}{\delta\rho},\mathcal{K}_\rho^{-1}\left(\mathcal{K}_\rho+\lambda\,\mathrm{Id}\right)\frac{\delta\dot u}{\delta\rho}\rangle_{L^2_\rho},\\ \mathcal{R}^*_{\lambda\text{-}\mathsf{FR}_k}(\rho,\xi)&=\frac{1}{2}\langle\xi,\left(\mathcal{K}_\rho+\lambda\,\mathrm{Id}\right)^{-1}\mathcal{K}_\rho\xi\rangle_{L^2_\rho}.\end{aligned} \tag{3.6}$$

Then, we obtain the following approximate gradient-flow equation with an approximate growth field given by the kernel ridge regression solution $(\mathcal{K}_\rho+\lambda\,\mathrm{Id})^{-1}\mathcal{K}_\rho\frac{\delta F}{\delta\mu}[\mu]$.

Proposition 3.3 (Kernel-approximate Fisher-Rao gradient flow) *The generalized gradient system $(\mathcal{M}^+,F,\mathcal{R}_{\lambda\text{-}\mathsf{FR}_k})$ generates the gradient flow equation where the approximate growth field r is given by the KRR solution*

$$\dot\mu_t=-\mu_t\cdot r_t,\quad r_t=\operatorname*{argmin}_f\left\{\left\|f-\frac{\delta F}{\delta\mu}[\mu_t]\right\|_{L^2_{\mu_t}}^2+\lambda\|f\|_{\mathcal{H}}^2\right\}. \tag{3.7}$$

Specifically, the closed-form solution for the growth field is

$$\dot{\mu} = -\mu \cdot (\mathcal{K}_\mu + \lambda \operatorname{Id})^{-1} \mathcal{K}_\mu \frac{\delta F}{\delta \mu}[\mu]. \tag{3.8}$$

Remark 3.4 (Relation between KRR and infimal convolution) *The above optimization problem* (3.7) *is not equivalent to the infimal convolution of $\mathcal{R}_\lambda$-FR_k and $\mathcal{R}^*_{\lambda\text{-}\mathsf{FR}_k}$ defined in* (3.6)*. However, it is easy to check using Lemma 2.4 that*

$$r_t = \operatorname*{argmin}_g \left\{ \|g\|^2_{L^2_\mu} + \frac{1}{\lambda} \|\mathcal{K}_\mu^{1/2}(g - \xi)\|^2_{L^2_\mu} \right\}.$$

This quadratic form matches the definition of $\mathcal{R}^*_{\lambda\text{-}\mathsf{FR}_k}$ (3.6) *as an infimal convolution, i.e., $\mathcal{R}^*_{\lambda\text{-}\mathsf{FR}_k}(\mu, \xi) = \inf_g \left\{ \frac{1}{2}\|g\|^2_{L^2_\mu} + \frac{1}{2\lambda}\|\mathcal{K}_\mu^{1/2}(g-\xi)\|^2_{L^2_\mu} \right\}$.*

3.2 MMD as de-kernelized Fisher-Rao distance

We now take a different direction from the previous subsection to de-kernelize the Fisher-Rao geometry (1.11). The result is somewhat surprising to us: the resulting gradient-flow geometry is equivalent to the MMD geometry.

We now develop the gradient flow structure of the MMD[3]. The primal and dual dissipation potentials are trivial due to the flatness of the MMD geometry

$$\mathcal{R}_{\mathrm{MMD}}(u) = \frac{1}{2}\|\mathcal{K}_\rho \frac{\delta \mu}{\delta \rho}\|^2_{\mathcal{H}} = \frac{1}{2}\|\mathcal{K}\mu\|^2_{\mathcal{H}}, \quad \mathcal{R}^*_{\mathrm{MMD}}(\xi) = \frac{1}{2}\langle \xi, \mathcal{K}_\rho^{-1}\xi \rangle_{L^2_\rho} = \frac{1}{2}\langle \xi, \mathcal{K}^{-1}\xi \rangle_{L^2}. \tag{3.9}$$

It is important to note that the MMD dissipation potentials are *state-independent*, i.e., they are not functions of the measure ρ since $\rho \cdot \mathcal{K}_\rho^{-1} v = \mathcal{K}^{-1} v$.

Proposition 3.5 (Gradient flow equation in the MMD geometry) *The gradient-flow equation in the MMD geometry is given by*

$$\dot{\mu} = -\mathcal{K}^{-1} \frac{\delta F}{\delta \mu}[\mu]. \tag{3.10}$$

Proposition 3.5 gives the intuition that the MMD gradient-flow equation is equivalent to a *reaction equation with the de-kernelized growth field* $\dot{\mu} = -\mu \mathcal{K}_\mu^{-1} \frac{\delta F}{\delta \mu}[\mu]$. It is worth noting that the gradient-flow equation can be stated in the dual space, $\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{K}\mu = -\frac{\delta F}{\delta \mu}[\mu]$, where $\mathcal{K}\mu$ is the kernel-mean embedding (Smola et al., 2007).

We now derive the main result of this section using the dynamic formulation

$$\mathrm{MMD}^2(\mu, \nu) = \min \left\{ \int_0^1 \|\xi_t\|^2_{\mathcal{H}} \mathrm{d}t \;\middle|\; \dot{u} = -\mathcal{K}^{-1}\xi_t, u(0) = \mu, u(1) = \nu \right\}. \tag{3.11}$$

Because of its flat structure, the adjoint equation for the MMD is simply $\dot{\xi} = 0$. It is also easy to verify the following lemma.

3. We follow the naming convention of the Wasserstein gradient flow and refer to the gradient flows in the MMD dissipation geometry as the MMD gradient flow. This is distinct from the setting considered in (Arbel et al., 2019), which is a WGF with the MMD energy.

Lemma 3.6 (Unconstrained dual formulation of squared-MMD) *The squared-MMD admits the unconstrained dual representation*

$$\mathrm{MMD}^2(\mu,\nu) = \sup_{h\in\mathcal{H}} \int h \,\mathrm{d}(\mu-\nu) - \frac{1}{4}\|h\|_{\mathcal{H}}^2. \tag{3.12}$$

The MMD geodesic curve is simply the straight line between μ and ν, $u(t) = (1-t)\mu + t\nu$. Therefore, when both μ and ν are probability measures, the solution along the MMD geodesic remains a probability measure. This is significantly simplified compared to the Fisher-Rao setting as remarked in footnote 1. To be consistent with the FR setting, we also consider the MMD between non-negative measures instead of only probability measures.

Summarizing, we present our main result regarding the MMD-Fisher-Rao relation.

Theorem 3.7 *The dynamic formulation of the force-kernelized (Definition 3.1) squared MMD* (3.11) *coincides with that of the squared Fisher-Rao distance* (1.11).

Riemannian metric perspective Using the perspective in Section 2.1, we show another perspective of the kernelization of Fisher-Rao and MMD following Definition **??**. Using the dissipation geometry (2.8), one can easily show that the Fisher-Rao Riemannian tensor and the Onsager operator are

$$\mathbb{G}_{\mathsf{FR}}(\nu) = \frac{1}{\nu}\cdot, \quad \mathbb{K}_{\mathsf{FR}}(\nu) = \nu\cdot.$$

Using the RKHS-L^2 relation (2.10), the state-independent counterparts for the MMD are

$$\mathbb{G}_{\mathrm{MMD}} = \mathcal{K}, \quad \mathbb{K}_{\mathrm{MMD}} = \mathcal{K}^{-1}.$$

Therefore, we conclude the following relation following Theorem 3.7.

Corollary 3.8 (Kernelization of Fisher-Rao Riemannian tensor) *The MMD and Fisher-Rao Riemannian tensors and Onsager operators are related by the kernelization*

$$\mathbb{G}_{\mathrm{MMD}} = \mathcal{K}_\nu \circ \mathbb{G}_{\mathsf{FR}}(\nu), \quad \mathbb{K}_{\mathrm{MMD}} = \mathbb{K}_{\mathsf{FR}}(\nu) \circ \mathcal{K}_\nu^{-1}.$$

We note that the naming convention of kernelization is consistent with the Stein-Wasserstein relation.

Therefore, in applications such as (1.6), one considers the Fisher-Rao minimizing movement

$$\min_{\mu\in\mathcal{M}^+} F(\mu) + \frac{1}{2\tau}\langle \mu-\nu, \mathbb{G}_{\mathsf{FR}}(\nu)(\mu-\nu)\rangle_{L^2},$$

which is also the theoretical basis for applications such as distributionally robust optimization proposed by Ben-Tal et al. (2013). Here μ is a given non-negative measure. In such cases, kernelization can be used to construct the MMD minimizing movement

$$\min_{\mu\in\mathcal{M}^+} F(\mu) + \frac{1}{2\tau}\langle \mu-\nu, \mathbb{G}_{\mathrm{MMD}}(\mu-\nu)\rangle_{L^2}.$$

This has been subsequently studied by Zhu et al. (2021) to take advantage of MMD's many favorable properties for practical optimization.

3.3 Flattened Fisher-Rao, Allen-Cahn, and φ-divergences

Motivated by the relation between the MMD and the Fisher-Rao distance, we now discuss another class of divergences via an analogous construction from the dynamic formulation. This amounts to changing the state-dependent dissipation potential (**??**) to the state-independent

$$\mathcal{R}_{\overline{\mathsf{FR}}}(u) = \frac{1}{2}\|\frac{\delta u}{\delta \omega}\|_{L^2_\omega}^2,\ \mathcal{R}^*_{\overline{\mathsf{FR}}}(\xi) = \frac{1}{2}\|\xi\|_{L^2_\omega}^2, \tag{3.13}$$

for a fixed reference measure ω. Like the MMD and the classical Allen-Cahn (i.e. L^2), the flatten disspation potentials are state-independent. Using the dissipation potential (3.13), We obtain the dynamic formulation

$$\overline{\mathsf{FR}}^2_\omega(\mu,\nu) = \min_{u,\xi}\left\{\int_0^1 \|\xi_t\|_{L^2_\omega}^2\,\mathrm{d}t \;\middle|\; \dot u = \omega\cdot\xi_t,\ u(0)=\mu,\ u(1)=\nu\right\}. \tag{3.14}$$

Similar to the MMD setting, the adjoint equation of the Hamiltonian dynamics simplifies to $\dot\zeta_t = 0$, resulting in the following static formulation in Proposition 3.9. The proof is omitted since it is a slight modification of the proof in (Otto and Villani, 2000, Section 3) that of the de-kernelized Wasserstein distance we show later. Similar to the linearized optimal transport (Wang et al., 2013) and generalized geodesics (Ambrosio et al., 2005), it is natural to consider a reference measure ω along the *Fisher-Rao geodesic* between μ and ν. By doing so, we recover the following connections with the φ-divergences.

Proposition 3.9 (Static formulation of flattened Fisher-Rao distance) *The flattened Fisher-Rao distance* (3.14) *is equivalent to the static formulation*

$$\overline{\mathsf{FR}}^2_\omega(\mu,\nu) = \sup_\zeta\left\{\int \zeta\,\mathrm{d}(\mu-\nu) - \frac{1}{4}\|\zeta\|_{L^2_\omega}^2\right\}. \tag{3.15}$$

If the reference measure is the Lebesgue measure $\omega = \Lambda$, then the flattened Fisher-Rao distance coincides with the L^2 norm $\overline{\mathsf{FR}}^2_\omega(\mu,\nu) = \|\mu-\nu\|_{L^2}^2$. The resulting graient flow is the L^2 Hilbert space gradient flow (classical Allen-Cahn, Example 2.1).

Furthermore, Suppose the reference measure ω is chosen along the Fisher-Rao geodesic between μ and ν given in (2.9)*, i.e., $\omega(s) = \big((1-s)\sqrt{\mu}+s\sqrt{\nu}\big)^2$, $s\in[0,1]$. Then, $\overline{\mathsf{FR}}^2_{\omega(s)}(\mu,\nu)$ conincides with,*

1. *if $s=0$, the χ^2-divergence $\mathrm{D}_{\chi^2}(\mu|\nu) = \|\frac{\mathrm{d}\mu}{\mathrm{d}\nu}-1\|_{L^2_\nu}^2$;*
2. *if $s=1$, the reverse χ^2-divergence $\mathrm{D}_{\chi^2}(\nu|\mu) = \|\frac{\mathrm{d}\nu}{\mathrm{d}\mu}-1\|_{L^2_\mu}^2$;*
3. *if $s=\frac{1}{2}$, the squared Fisher-Rao (Hellinger) distance itself $\mathsf{FR}^2(\mu,\nu) = 4\|\sqrt{\mu}-\sqrt{\nu}\|_{L^2}^2$.*

Remark 3.10 (Fisher-Rao geodesic and flatness) *The third case demonstrates a particular nice property of the Fisher-Rao geometry. The Fisher-Rao geometry is not flat and possesses a geodesic structure* (Geod-FR)*. Yet its geodesic distance can be computed by a flat distance $\overline{\mathsf{FR}}_{\omega(\frac{1}{2})}(\mu,\nu)$ characterized in Proposition 3.9.*

4 Wasserstein setting

We now apply some of our results in the Fisher-Rao type gradient flows from the previous section to the Wasserstein type flows, exploiting the similarity in their dissipation geometry. First, we revisit the Stein gradient flow in Section 4.1, where we establish the gradient flow structure and dissipation potentials for Stein and its regularized version. In Section 4.2, we derive a new Wasserstein-type gradient-flow geometry by drawing the parallel to the relation between Fisher-Rao and MMD. We further show a related Cahn-Hilliard dissipation geometry.

4.1 Gradient structure for the (regularized) Stein gradient flow

The Stein geometry (Liu and Wang, 2019; Duncan et al., 2019) has been studied in the context of sampling for statistical inference. We first write down explicitly the gradient structure for the Stein gradient system by providing the dissipation potentials for the Stein geometry, implied in the dynamic formulation of (Duncan et al., 2019)

$$R_{\mathrm{Stein}}(\rho,u)=\frac{1}{2}\|u\|^2_{\mathrm{Stein},\rho},\quad \|u\|^2_{\mathrm{Stein},\rho}:=\inf\left\{\|\mathcal{K}_\rho v\|^2_{\mathcal{H}}:u=-\operatorname{div}(\rho\cdot\ \mathcal{K}_\rho v)\right\}. \tag{4.1}$$

We refer to the $\|u\|_{\mathrm{Stein},\rho}$ as the primal Stein norm dual dissipation potential. By Fenchel-duality, we find the velocity-kernelized dual dissipation potential of the Stein gradient flow

$$R^*_{\mathrm{Stein}}(\rho,\xi)=\frac{1}{2}\langle\nabla\xi,\mathcal{K}_\rho\nabla\xi\rangle_{L^2_\rho}=\frac{1}{2}\|\mathcal{K}_\rho\nabla\xi\|^2_{\mathcal{H}}. \tag{4.2}$$

Using this gradient structure, we obtain Stein (variational) gradient-flow equation

$$\partial_t\mu=\operatorname{div}(\mu\cdot\mathcal{K}_\mu\nabla\frac{\delta F}{\delta\mu}[\mu]). \tag{4.3}$$

Regularized Stein gradient flow as approximation to Wasserstein gradient flow Following a similar route as in the kernelized Fisher-Rao setting, we consider the regularized primal and dual dissipation potential

$$\begin{aligned}\mathcal{R}_{\lambda\text{-Stein}}(\rho,u)&:=\mathcal{R}_{W_2}(\rho,u)+\lambda R_{\mathrm{Stein}}(\rho,u)=\frac{1}{2}\|u\|^2_{H^{-1}(\rho)\cdot}+\frac{\lambda}{2}\|u\|^2_{\mathrm{Stein},\rho}\ \text{ for } u=\operatorname{div}(\rho v)\\&=\frac{1}{2}\left(\langle v,v\rangle_{L^2_\rho}+\langle v,\mathcal{K}_\rho^{-1}v\rangle_{L^2_\rho}\right)=\frac{1}{2}\langle v,\mathcal{K}_\rho^{-1}(\mathcal{K}_\rho+\lambda\,\mathrm{Id})v\rangle_{L^2_\rho},\end{aligned} \tag{4.4}$$

$$\mathcal{R}^*_{\lambda\text{-Stein}}(\rho,\xi)=\frac{1}{2}\langle\nabla\xi,(\mathcal{K}_\rho+\lambda\,\mathrm{Id})^{-1}\,\mathcal{K}_\rho\nabla\xi\rangle_{L^2_\rho}, \tag{4.5}$$

resulting in the following approximate Wasserstein gradient system.

Proposition 4.1 (Kernel-approximate Wasserstein gradient flow with KRR velocity field) *The generalized gradient system $(\mathcal{P},F,\lambda\text{-Stein})$ generates the gradient flow equation where the approximate growth field r is given by the KRR solution*

$$\dot\mu_t=\operatorname{div}(\mu_t\cdot v_t),\quad v_t=\operatorname*{argmin}_f\left\{\|f-\nabla\frac{\delta F}{\delta\mu}[\mu_t]\,\|^2_{L^2_{\mu_t}}+\lambda\|f\|^2_{\mathcal{H}}\right\}. \tag{4.6}$$

Specifically, the closed-form solution for the velocity field is

$$v_t=(\mathcal{K}_\mu+\lambda\,\mathrm{Id})^{-1}\mathcal{K}_\mu\nabla\frac{\delta F}{\delta\mu}[\mu]\,. \tag{4.7}$$

It is worth noting that the above approximation fits the setting of learning the diffusion model (1.7), where the velocity is approximated using KRR. Later, we rigorously justify this approximation.

4.2 De-Stein: de-kernelized Wasserstein geometry

As we have witnessed in Section 3.2, the MMD bears the intuition of the *de-kernelized and flat Fisher-Rao distance*. What then is the de-kernelized (flat) Wasserstein geometry, as MMD is to Fisher-Rao? Our starting point is the following de-kernelized H^{-1}-type norm

$$\|u\|^2_{\text{De-Stein}} := \inf\left\{\|v\|^2_{\mathcal{H}} : u = -\operatorname{div}(\rho\cdot\ \mathcal{K}_\rho^{-1}v)\right\} = \inf\left\{\|v\|^2_{\mathcal{H}} : u = -\operatorname{div}(\cdot\ \mathcal{K}^{-1}v)\right\}.$$

Similar to the gradient structure of MMD, this quantity no longer depends on the measure ρ. As a consequence of the above formulation, the primal and dual dissipation potential for the approximation system are *state-independent*

$$R_{\text{De-Stein}}(u) = \frac{1}{2}\|u\|^2_{\text{De-Stein}}, \quad R^*_{\text{De-Stein}}(\xi) = \frac{1}{2}\|\mathcal{K}^{-\frac{1}{2}}\nabla\xi\|^2_{L^2} = \frac{1}{2}\|\nabla\xi\|^2_{\mathcal{H}}.$$

Therefore, like the MMD, this geometry is a flat geometry and its gradient flow equation is

$$\partial_t\mu = -\operatorname{div}(\ \mathcal{K}^{-1}\nabla\frac{\delta F}{\delta\mu}[\mu]).$$

We can now write down the dynamic formulation of a new distance, which we term the *de-kernelized Wasserstein* (De-Stein) distance[4].

$$\text{De-Stein}^2(\mu,\nu) = \inf\left\{\int_0^1 \|\nabla\xi_t\|^2_{\mathcal{H}}\,\mathrm{d}t \,\middle|\, \partial_t u = -\operatorname{div}(\mathcal{K}^{-1}\nabla\xi_t), u(0)=\mu, u(1)=\nu\right\}. \quad (4.8)$$

The Hamiltonian formulation for De-Stein is

$$\begin{cases}\dot\mu = -\operatorname{div}(\mathcal{K}^{-1}\nabla\xi),\\ \dot\xi = 0.\end{cases}$$

where the Hamilton-Jacobi equation in the Wasserstein geometry is replaced by the "static" adjoint variable ξ due to the state-independent dissipation potential (4.8). More concretely, following the derivation of the relation between the static Kantorovich dual formulation and dynamic Benamou–Brenier formulation (see, e.g., (Otto and Villani, 2000)), we now derive the static defition of the metric. Different from the Wasserstein and Stein setting, we only need one static (i.e., time-independent) test function, because of the adjoint (geodesic) equation in the Hamiltonian dynamics $\partial_t\zeta = 0$. Consequently, we obtain a simple static formulation as a weak norm, similar to an IPM.

Proposition 4.2 (Static dual formulation of the De-Stein distance) *The dynamic formulation of the De-Stein distance* (4.8) *is equivalent to the static dual formulation*

$$\text{De-Stein}^2(\mu,\nu) = \sup_\zeta\left\{\int\zeta\,\mathrm{d}(\mu-\nu) - \frac{1}{4}\|\nabla\zeta\|^2_{\mathcal{H}}\right\}. \quad (4.9)$$

4. Due to the presence of static dual representation in Proposition 4.2, this distance should be more appropriately termed kernel-Sobolev distance. However, similar terms have already been used in the machine learning literature to denote a heuristically regularized H^{-1} geometry.

The intuition is self-evident — compared with the static dual of MMD, the regularization by the RKHS norm $\|\zeta\|_{\mathcal{H}}$ is replaced by the RKHS norm of its gradient. The Euler-Lagrange equation of the optimization problem in (4.9) is $\frac{1}{2}\operatorname{div}\left(-\mathcal{K}^{-1}\nabla\zeta\right) = \mu - \nu$. Plugging it back into (4.9) and integrating by parts, we find

$$\text{De-Stein}^2(\mu,\nu) = \inf_{v}\left\{\|v\|_{\mathcal{H}}^2 \text{ s. t. } \operatorname{div}\left(-\mathcal{K}^{-1}v\right) = 2(\mu-\nu)\right\},$$

which is simply a kernel-weighted H^{-1} norm.

4.3 Flattened Wasserstein, Cahn-Hilliard, and Sobolev discrepancy

Mirroring the development of the flattened Fisher-Rao setting in Section 3.3, we now focus on the a similar construction in the Wasserstein setting. Similar to the De-Stein setting, we obtain the state-independent dissipation potentials

$$R_{\overline{W}_2}(u) = \frac{1}{2}\|u\|_{H_\omega^{-1}}^2, \quad R_{\overline{W}_2}^*(\xi) = \frac{1}{2}\|\nabla\xi\|_{L_\omega^2}^2,$$

where the reference measure ω is fixed. Its dynamic formulation is given by

$$\overline{W}_\omega^2(\mu,\nu) = \min\left\{\int_0^1 \|\nabla\xi_t\|_{L_\omega^2}^2\,\mathrm{d}t \middle| \partial_t u = -\operatorname{div}(\omega\cdot\nabla\xi_t), u(0)=\mu, u(1)=\nu\right\}. \tag{4.10}$$

The adjoint equation $\partial\zeta_t = 0$ implies the static dual formulation

$$\overline{W}_\omega^2(\mu,\nu) = \sup_{\zeta}\left\{\int\zeta\,\mathrm{d}(\mu-\nu) - \frac{1}{4}\|\nabla\zeta\|_{L_\omega^2}^2\right\}. \tag{4.11}$$

Similar to the setting after Proposition 4.2, we find static dual formulation is equivalent to

$$\overline{W}_\omega^2(\mu,\nu) = \inf_{\xi}\left\{\|\xi\|_{L_\omega^2}^2 \text{ s. t. } \frac{1}{2}\operatorname{div}(\xi\cdot\omega) = \mu-\nu\right\},$$

which concincides with the weighted H_ω^{-1} norm. If the reference measure ω is chosen as the Lebesgue measure, then $\frac{1}{2}\Delta\zeta = (\mu-\nu)$. Consequently, we obtain the static formulation

$$\overline{W}^2(\mu,\nu) = \|\mu-\nu\|_{H^{-1}}^2 = -\int(\mu-\nu)\,\Delta^{-1}\,(\mu-\nu)\,\mathrm{d}x, \tag{4.12}$$

which is equivalent to the classical Cahn-Hilliard H^{-1} Hilbert space in Example 2.1. If the reference measure ω is chosen as $\omega=\nu$, the flattened Wasserstein distance is the H_ν^{-1} norm, which is equivalent to the Sobolev discrepancy proposed by Mroueh et al. (2019).

5 Wasserstein-Fisher-Rao setting

This section applies our framework to the Wasserstein-Fisher-Rao gradient flow, also referred to as Hellinger-Kantorovich by Liero et al. (2018) for a better accounting of the historical developments. We first recall an elementary fact regarding duality of the inf-convolution of functionals.

Lemma 5.1 (Dissipation potential with inf-convolution) *Suppose the effective primal dissipation potential $\mathcal{R}$ is given by the inf-convolution of two dissipation potentials $\mathcal{R}_1, \mathcal{R}_2$, i.e., $\mathcal{R}(\mu, \cdot) = \mathcal{R}_1(\mu, \cdot) \square \mathcal{R}_2(\mu, \cdot)$, where $\square$ denotes inf-convolution. The resulting gradient-flow equation generated by the gradient system $(X, F, \mathcal{R})$ is given by*

$$\dot{\mu} = \partial \mathcal{R}_1^*(\mu, -\mathrm{D}^X F) + \partial \mathcal{R}_2^*(\mu, -\mathrm{D}^X F), \tag{5.1}$$

Note that the two differentials $\mathrm{D}^X F$ in (5.1) must be taken w.r.t. the same space X.

Example 5.1 (Hellinger-Kantorovich) *In the setting of the Wasserstein-Fisher-Rao (Hellinger-Kantorovich) distance (Liero et al., 2018; Chizat et al., 2019), the gradient-flow equation* (5.1) *corresponds to the reaction-diffusion PDE*

$$\dot{\mu} = \mathcal{R}^*_{W_2}(\mu, -\frac{\delta F}{\delta \mu}[\mu]) + \partial \mathcal{R}^*_{D_{\text{Hellinger}}}(\mu, -\frac{\delta F}{\delta \mu}[\mu]) = \operatorname{div}(\mu \nabla \frac{\delta F}{\delta \mu}[\mu]) - \mu \frac{\delta F}{\delta \mu}[\mu]\,. \tag{5.2}$$

5.1 Kernelization and approximation of the WFR gradient flow

To produce a Stein-type geometry for the WFR distance, we consider the primal and dual dissipation potentials as in Lemma 5.1, where two dissipation potentials are obtained from the Stein and kernelized Fisher-Rao

$$\mathcal{R}_{\text{K-WFR}} = \mathcal{R}_{\text{Stein}} \square \mathcal{R}_{\mathsf{FR}_k}, \quad \mathcal{R}^*_{\text{K-WFR}} = \mathcal{R}_{\text{Stein}} + \mathcal{R}_{\mathsf{FR}_k}.$$

Our starting point is therefore the kernelized reaction-diffusion equation

$$\dot{\mu} - \operatorname{div}\left(\mu \cdot \mathcal{K}_\mu \nabla \frac{\delta F}{\delta \mu}[\mu]\right) = -\mu \cdot \mathcal{S}_\mu \frac{\delta F}{\delta \mu}[\mu]\,. \tag{5.3}$$

where $\mathcal{S}_\mu$ is the integral operator associated with the another kernel $s(\cdot, \cdot)$ that may be different from k. We find the following dynamic formulation of the *kernelized Wasserstein-Fisher-Rao* distance with the kernerlized reaction-diffusion equation

$$k\text{-}\mathsf{WFR}^2(\mu, \nu) = \min\Big\{ \int_0^1 \|\mathcal{K}_{u_t} \nabla \xi_t\|^2_{\mathcal{H}} + \|\mathcal{S}_{u_t} \zeta_t\|^2_{\mathcal{H}} \,\mathrm{d}t \Big| \\ \dot{u}_t - \operatorname{div}(u_t \cdot \mathcal{K}_{u_t} \nabla \xi_t) = -u_t \cdot \mathcal{S}_{u_t} \zeta_t, u(0) = \mu, u(1) = \nu \Big\}. \tag{5.4}$$

Going beyond kernelization, we now construct the kernel-approximate WFR geometry by considering both inf-convolution and additive regularization

$$\mathcal{R}_{\text{RK-WFR}}(\rho, \cdot) := \mathcal{R}_{\lambda\text{-Stein}}(\rho, \cdot) \square \mathcal{R}_{\lambda\text{-}\mathsf{FR}_k}(\rho, \cdot), \tag{5.5}$$

$$\mathcal{R}^*_{\text{RK-WFR}}(\rho, \cdot) = \mathcal{R}^*_{\lambda\text{-Stein}}(\rho, \cdot) + \mathcal{R}^*_{\lambda\text{-}\mathsf{FR}_k}(\rho, \cdot), \tag{5.6}$$

where $\mathcal{R}_{\lambda\text{-Stein}}(\rho, \cdot)$ is defined in (4.4) and $\mathcal{R}_{\lambda\text{-}\mathsf{FR}_k}(\rho, \cdot)$ is defined in (3.6). We summarize the result below by performing the calculation using the inf-convolution rules for both the kernelized version, i.e., Stein-type metric for the Wasserstein-Fisher-Rao case, and the regularized kernel-Wasserstein-Fisher-Rao gradient flow.

Corollary 5.2 (Kernel-approximate WFR gradient flow) *The generalized gradient system* $(\mathcal{M}^+, F, \mathsf{WFR})$ *generates the reaction-diffusion equation where the velocity and growth field* v, r *are given by the nonparametric regression solutions*

$$\dot{\mu}_t = \operatorname{div}(\mu_t \cdot v_t) - \mu_t \cdot r_t,$$
$$r_t = \operatorname*{argmin}_f \left\{ \left\| f - \frac{\delta F}{\delta \mu}[\mu_t] \right\|^2_{L^2_{\mu_t}} + \lambda \|f\|^2_{\mathcal{H}} \right\}, \; v_t \quad = \operatorname*{argmin}_f \left\{ \|f - \nabla \frac{\delta F}{\delta \mu}[\mu_t]\|^2_{L^2_{\mu_t}} + \lambda \|f\|^2_{\mathcal{H}} \right\}. \tag{5.7}$$

5.2 De-kernelized Wasserstein-Fisher-Rao and Kernel Sobolev-Fisher

Similar to the De-Stein and MMD, we now consider the de-kernelized Wasserstein-Fisher-Rao (D-WFR) distance by de-kernelizing the reaction-diffusion equation (5.2)

$$\dot{\mu} - \operatorname{div}\left(\mu \cdot \mathcal{K}_\mu^{-1} \nabla \frac{\delta F}{\delta \mu}[\mu] \right) = -\mu \cdot \mathcal{K}_\mu^{-1} \frac{\delta F}{\delta \mu}[\mu]. \tag{5.8}$$

We find the dynamic formulation and the regularized version

$$\text{D-}\mathsf{WFR}^2(\mu, \nu) = \inf \left\{ \int_0^1 \|\nabla \xi_t\|^2_{\mathcal{H}} + \|\zeta_t\|^2_{\mathcal{H}} \, \mathrm{d}t \middle| \right.$$
$$\left. \dot{u}_t - \operatorname{div}\left(u_t \cdot \mathcal{K}_{u_t}^{-1} \nabla \xi_t \right) = -u_t \cdot \mathcal{S}_{u_t}^{-1} \zeta_t, u(0) = \mu, u(1) = \nu \right\}, \tag{5.9}$$

The important insight here is:

Proposition 5.3 (De-kernelized Wasserstein-Fisher-Rao) *The De-kernelized Wasserstein-Fisher-Rao distance* (5.9) *coincides with the inf-convolution of the MMD and the De-Stein distance, i.e.,* $\text{D-}\mathsf{WFR}^2(\mu, \nu) := \inf_{\pi \in \mathcal{M}^+} \text{De-Stein}^2(\mu, \pi) \square \, \text{MMD}^2(\pi, \nu)$.

Furthermore, it admits the static dual formulation

$$\text{D-}\mathsf{WFR}^2(\mu, \nu) = \sup_\zeta \left\{ \int \zeta \, \mathrm{d}(\mu - \nu) - \frac{1}{4} \|\nabla \zeta\|^2_{\mathcal{H}} - \frac{1}{4} \|\zeta\|^2_{\mathcal{H}} \right\}. \tag{5.10}$$

Compared with the MMD (3.12), the test function in the de-kernelized WFR is additioanlly reguarlized by the RKHS norm of the gradient, $\|\nabla \zeta\|^2_{\mathcal{H}}$.

5.3 Flattened Wasserstein-Fisher-Rao and kernel-Sobolev-Fisher

Using the WFR-type inf-convolution with the flattened Fisher-Rao (Section 3.3) and the flattened Wasserstein (Section 4.3), we obtain the following flattened Wasserstein-Fisher-Rao formulation

$$\mathsf{WFR}^2_\omega(\mu, \nu) := \min \left\{ \int_0^1 \|\nabla \xi_t\|^2_{L^2_{\omega_1}} + \|\zeta_t\|^2_{L^2_{\omega 2}} \, \mathrm{d}t \middle| \right.$$
$$\left. \dot{u}_t - \operatorname{div}\left(\omega_1 \nabla \xi_t \right) = -\omega_2 \zeta_t, u(0) = \mu, u(1) = \nu \right\}, \tag{5.11}$$

The flattened geometry is the inf-convotion of the weighted L^2_ω and H^{-1}_ω norms, with the static formulation

$$\mathsf{WFR}^2_\omega(\mu,\nu) = \sup_\zeta \left\{ \int \zeta \,\mathrm{d}(\mu-\nu) - \frac{1}{4}\|\nabla\zeta\|^2_{L^2_\omega} - \frac{1}{4}\|\zeta\|^2_{L^2_\omega} \right\}. \tag{5.12}$$

In the simplest case that ω_1, ω_2 are both the Lebesgue measure, we recover the inf-convolution of the Allen-Cahn (L^2) and Cahn-Hilliard (H^{-1}). Choosing $\omega_1 = \omega_2 = \nu$, this distance becomes the Sobolev-Fisher discrepancy (Mroueh et al., 2019). Furthermore, those authors proposed a regularized version with a heuristic RKHS norm regularization, of which we give the precise characterization.

Example 5.2 (Regularized Kernel-Sobolev-Fisher discrepancy) *The regularized kernel Sobolev-Fisher discrepancy proposed by Mroueh and Rigotti (2020)*

$$\mathrm{KSF}^2(\mu,\nu) = \sup_\zeta \left\{ \int \zeta \,\mathrm{d}(\mu-\nu) - \frac{1}{4}\|\nabla\zeta\|^2_{L^2_\nu} - \frac{1}{4}\|\zeta\|^2_{L^2_\nu} - \frac{a}{2}\|\zeta\|^2_{\mathcal{H}} \right\}, \tag{5.13}$$

for $a > 0$, is equivalent to the following dynamic formulation

$$\min\left\{ \int_0^1 \|\nabla\xi_t\|^2_{L^2_\nu} + \|\zeta_t\|^2_{L^2_\nu} + \frac{1}{2a}\|\kappa_t\|^2_{\mathcal{H}}\,\mathrm{d}t \,\middle|\, \dot{u}_t - \mathrm{div}\,(\nabla\xi_t) = -\zeta_t - \mathcal{K}^{-1}\kappa_t, u(0)=\mu, u(1)=\nu \right\}.$$

It is the inf-convolution of three dissipaton geometries, the MMD, H^{-1}_ν norm, and L^2_ν norm.

6 Analysis

6.1 Entropy dissipation and functional inequalities

In machine learning applications, functional inequalities are building blocks for analysis of many algorithms, for example, sampling and optimization via particle schemes. The aim of this section is to develop an intuition for the Łojasiewicz type inequalities for the various gradient-flow geometries studied in this paper.

Historically, the celebrated Bakry-Émery theorem (Bakry and Émery, 1985) 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

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

Our starting point here is the differential energy dissipation balance (EDB) relation of generalized gradient flow systems,

$$\frac{\mathrm{d}}{\mathrm{d}t}F(\mu(t)) = \langle \mathrm{D}F, \dot{\mu}\rangle = -\Big(\mathcal{R}(\mu,\dot{\mu}) + \mathcal{R}^*(\mu,-\mathrm{D}F)\Big) =: -\mathcal{I}(\mu(t)). \tag{6.1}$$

where the $\mathcal{R}, \mathcal{R}^*$ quantities are the dissipation potentials discussed in the previous sections. We refer to the quantity $\mathcal{I}$ as the *energy dissipation*. From this, we impose the following version of the Łojasiewicz condition.

Definition 6.1 (Łojasiewicz inequality for generalized gradient systems) *We say that the Łojasiewicz inequality holds if*

$$\mathcal{R}(\mu,\dot{\mu})+\mathcal{R}^*(\mu,-\mathrm{D}F)\geq c\cdot\left(F(\mu(t))-\inf_{\mu}F(\mu)\right)^{\alpha}. \tag{Ł}$$

holds for some $c>0,\alpha>0$.

For conciseness, this paper only focuses on the case of $c>0,\alpha=1$, i.e., the Polyak-Łojasiewicz inequality due to its relevance to machine learning and optimization, simply referred to as the Łojasiewicz inequality in the rest of the paper. We refer to articles such as (Otto and Villani, 2000; Blanchet and Bolte) for a wider scope of related inequalities.

An immediate consequence of (Ł) is that the energy of the generalized gradient system converges exponentially via Grönwall's lemma, i.e.,

$$F(\mu(t))-\inf_{\mu}F(\mu)\leq e^{-c\cdot t}\left(F(\mu(0))-\inf_{\mu}F(\mu)\right).$$

Therefore, on the formal level, the intuition of the analysis is to produce the Łojasiewicz type relations in the form of

$$\mathcal{I}\geq c\cdot(F_t-F^*). \tag{6.2}$$

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

$$\mathcal{I}_F^W(\mu(t))=-\frac{\mathrm{d}}{\mathrm{d}t}F(\mu(t))\overset{\text{(along WGF)}}{=}\int\mu\left|\nabla\frac{\delta F}{\delta\mu}[\mu]\right|^2. \tag{6.3}$$

Carrying out the similar derivation, we find the entropy dissipation for the Fisher-Rao gradient flow

$$\mathcal{I}_F^{\mathsf{FR}}(\mu(t))=-\frac{\mathrm{d}}{\mathrm{d}t}F(\mu(t))\overset{\text{(along FRGF)}}{=}\int\mu\left|\frac{\delta F}{\delta\mu}[\mu]\right|^2. \tag{6.4}$$

As a toy example, we check the inequality (Ł) for the well-documented case: Wasserstein gradient flow of the KL-relative entropy.

Example 6.1 (Łojasiewicz for Wasserstein geometry with KL-entropy energy) *Consider the Wasserstein gradient system with the KL entropy energy function $(\mathcal{P}(\bar{\Omega}),\mathrm{D}_{\mathrm{KL}}(\cdot\|\pi),W_2^2)$, where $\mathrm{D}_{\mathrm{KL}}(\mu|\pi)=\int\log(\frac{\mathrm{d}\mu}{\mathrm{d}\pi})\mu$. Recall the energy dissipation balance relation*

$$-\mathcal{I}(\mu(t))=\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{D}_{\mathrm{KL}}(\mu_t\|\pi)=-\frac{1}{2}|\mu'|_{W_2}(t)^2-\frac{1}{2}|\nabla^-\mathrm{D}_{\mathrm{KL}}(\cdot\|\pi)|_{W_2}(\mu(t))^2.$$

Now, suppose the measure π is such that (Ł) *holds, i.e.,*

$$|\mu'|_{W_2}(t)^2+|\nabla^-\mathrm{D}_{\mathrm{KL}}(\cdot\|\pi)|_{W_2}(\mu(t))^2\geq c\cdot\mathrm{D}_{\mathrm{KL}}(\mu(t)\|\pi),$$

which is equivalent to a Logarithmic Sobolev inequality (LSI). The above relations imply exponential convergence via Grönwall's lemma.

An alternative to using analysis concepts, such as metric slope and speed, is the following formal calculation.

$$-\mathcal{I}(\mu(t)) = \frac{\mathrm{d}}{\mathrm{d}t}\mathrm{D}_{\mathrm{KL}}(\mu|\pi) = \langle \mathrm{D}^{L^2}\mathrm{D}_{\mathrm{KL}}(\mu|\pi), \dot{\mu}\rangle_{L^2} = \langle \log\frac{\mu}{\pi}, -\operatorname{div}(\mu\nabla\log\frac{\mu}{\pi})\rangle_{L^2} \overset{(IBP)}{=} -\|\nabla\log\frac{\mu}{\pi}\|^2_{L^2_\mu}. \quad (6.5)$$

Specializing the Łojasiewicz inequality (Ł) and form (6.2) to the Wasserstein-KL example above, we find the LSI for some $c > 0$

$$\|\nabla\log\frac{\mu(t)}{\pi}\|^2_{L^2(\mu(t))} \geq c\cdot \mathrm{D}_{\mathrm{KL}}(\mu(t)\|\pi). \quad \text{(LSI)}$$

By Grönwall's lemma, the entropy decays exponentially $\mathrm{D}_{\mathrm{KL}}(\mu(t)\|\pi) \leq \mathrm{e}^{-c\cdot t}\mathrm{D}_{\mathrm{KL}}(\mu(0)\|\pi)$. However, it is important to note that the Łojasiewicz inequality (Ł), of which the LSI is a special case, cannot be expected to hold globally for arbitrary geometry in general, e.g., the Fisher-Rao geometry. Consider the Fisher-Rao gradient flow with the KL entropy energy, i.e., $F(\mu) = \mathrm{D}_{\mathrm{KL}}(\mu|\pi)$. Then, the specialized Łojasiewicz condition reads, for some $c > 0$,

$$\|\log\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\|^2_{L^2_\mu} \geq c\cdot \mathrm{D}_{\mathrm{KL}}(\mu(t)\|\pi). \quad (6.6)$$

Lemma 6.2 (No global Łojasiewicz condition in Fisher-Rao flows of KL) *There exists no $0 < c < \infty$ such that (6.6) holds along the Fisher-Rao flow of the KL-entropy, i.e., the gradient system $(\mathcal{M}^+, \mathrm{D}_{\mathrm{KL}}(\cdot|\pi), \mathsf{FR})$ does not satisfy the global Łojasiewicz condition for any constant.*

See Figure 2 for an illustration. 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 stated below.

Example 6.2 (Birth escaping zero: focus on global instead of local) *Our focus on the global Łojasiewicz condition is mathematically motivated by the following reason, illustrated in Figure 3. Suppose we wish to minimize the energy $F(\mu) = \mathrm{D}_\varphi(\mu|\pi)$ starting from the initial distribution μ_0. It is very common that the distribution μ_0 does not have the full support of the target distribution π as in Figure 3 (left), i.e.,* $\mathrm{supp}(\mu_0) \subsetneq \mathrm{supp}(\pi)$. *It is also possible that they share the support as in Figure 3 (right), i.e.,* $\mathrm{supp}(\mu_0) = \mathrm{supp}(\pi)$, *but the density ratio is nearly zero at many points. For example, Figure 3 (right) depicts a Gaussian mixture distribution as the initial μ_0 that has very little mass near $x = 2$. The most difficult part of the minimization is to* escape the near-zero region with enough slopes 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 6.5. In contrast, the local convergence behavior near the equilibrium is much easier to capture, see Figure 2, Figure 4. Therefore, we place our current scope on the global Łojasiewicz condition without delving into the detailed equilibrium behavior.*

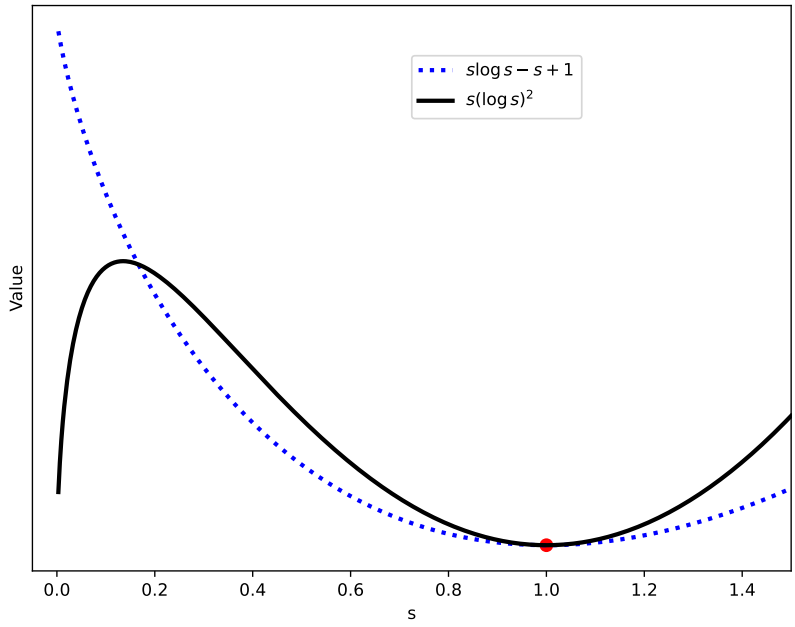


Figure 2: The plot illustrates the lack of global Łojasiewicz inequality as in Lemma 6.2. We plot the KL-entropy generator function $\varphi(s) = s\log s - s + 1$. The blue dotted curve represents the KL-entropy generator $\varphi(s)$. The function $s|\log s|^2$ is plotted in solid black. The Łojasiewicz inequality condition is satisfied locally around the equilibrium $s = 1$ (red dot). However, it can never be satisfied in a neighborhood around $s = 0$.

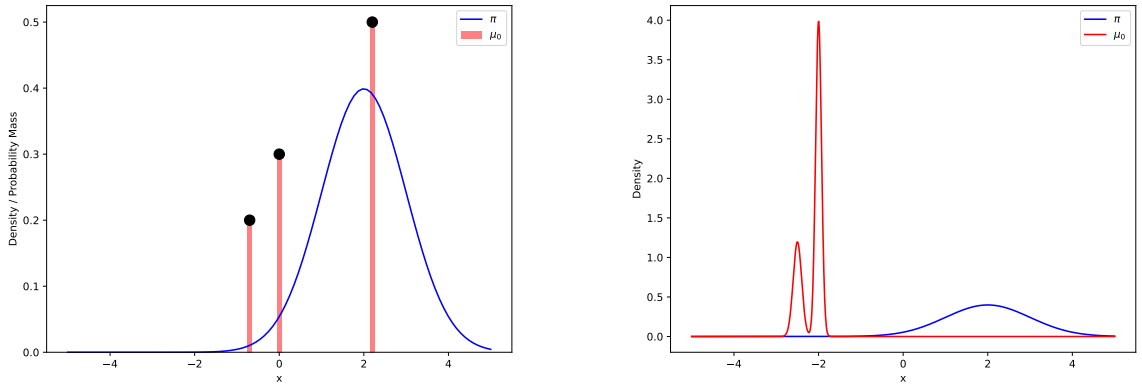


Figure 3: Illustration of Example 6.2: birth escaping zero.

While Lemma 6.2 shows that the Fisher-Rao flow of the KL-entropy cannot satisfy the global Łojasiewicz, we now show a positive result for the case when the energy functional is a nicer one: the squared Fisher-Rao distance $F(\mu) = \frac{1}{2}\mathsf{FR}^2(\mu, \pi)$. First, note that the

first variation of the squared Fisher-Rao distance is $\frac{\delta}{\delta\mu}\frac{1}{2}\mathsf{FR}^2(\mu,\pi) = f'(\mu) = 2 - 2\sqrt{\frac{\mathrm{d}\pi}{\mathrm{d}\mu}}$. Specializing the Łojasiewicz inequality to this setting,

$$4 \cdot \|1 - \sqrt{\frac{\mathrm{d}\pi}{\mathrm{d}\mu}}\|_{L^2_\mu}^2 \geq c \cdot \frac{1}{2}\mathsf{FR}^2(\mu|\pi). \tag{6.7}$$

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

Lemma 6.3 (Global Łojasiewicz with Fisher-Rao energy) *The Łojasiewicz inequality* (6.7) *holds for the Fisher-Rao gradient system* $(\mathcal{M}^+, \frac{1}{2}\mathsf{FR}^2(\cdot,\pi), \mathsf{FR})$ *globally for any* $c \in (-\infty, 2]$.

From the discussion above, we have seen the nice property of the Fisher-Rao distance. We are now ready to extract some general principles. Historically, following Bakry and Émery (1985), Arnold et al. (2001) provided an elementary proof of the Bakry-Émery theorem, i.e., for the convex entropy function φ satisfying

$$\varphi(1) = \varphi'(1) = 0,\ \varphi''(1) > 0 \quad \text{and} \quad \big(\varphi'''(s)\big)^2 \leq \frac{1}{2}\varphi''(s)\varphi^{(4)}(s), \tag{6.8}$$

the Wasserstein gradient flow with the corresponding φ-divergence energy converges exponentially. That is, the following sufficient relation holds

$$(\mathrm{BE}) + (6.8) \implies \text{Łojasiewicz for Wasserstein} \implies \text{exp. convergence.}$$

The natural question is whether such relation exists for the Fisher-Rao geometry. To answer that, we first establish the condition for global Łojasiewicz condition for a class of entropies.

Proposition 6.4 (Global Łojasiewicz for FR flow of φ-divergence energy) *Given the Fisher-Rao gradient system with φ-divergence energy, i.e.,* $(\mathcal{M}^+, \mathrm{D}_\varphi(\cdot\|\pi), \mathsf{FR})$. *If* $\varphi : (0,\infty) \to [0,\infty)$ *is a convex entropy generator function satisfying*

$$\varphi(1) = \varphi'(1) = 0, \varphi''(1) > 0 \quad \textit{and} \quad \exists\, c_* > 0 \textit{ such that } \forall\, s > 0 :\ s\big(\varphi'(s)\big)^2 \geq c_*\varphi(s), \tag{6.9}$$

then the Łojasiewicz inequality holds along the Fisher-Rao gradient flow , i.e.,

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

Proof of Proposition 6.4. The first-variations of the φ-divergence is given by (Ambrosio et al., 2005) $\frac{\delta}{\delta\mu}\mathrm{D}_\varphi(\mu|\pi) = \varphi'\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right)$. Thus, using the Fisher-Rao metric, we obtain the dissipation relation $\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{D}_\varphi(\mu|\pi) = -\mathcal{I}_\varphi^{\mathsf{FR}}(\mu)$ with

$$\mathcal{I}_\varphi^{\mathsf{FR}}(\mu) = \left\|\varphi'\big(\tfrac{\mathrm{d}\mu}{\mathrm{d}\pi}\big)\right\|_{L^2_\mu}^2 = \int_X \Big(\varphi'\big(\tfrac{\mathrm{d}\mu}{\mathrm{d}\pi}\big)\Big)^2\,\mathrm{d}\mu = \int_X \Big(\varphi'\big(\tfrac{\mathrm{d}\mu}{\mathrm{d}\pi}\big)\Big)^2\,\tfrac{\mathrm{d}\mu}{\mathrm{d}\pi}\,\mathrm{d}\pi.$$

Now exploiting the assumption (6.9) for estimating the integrand, we immediately obtain (Ł-FR). ■

In short, for the φ-divergence energy,

$$(6.9) \iff (\text{Ł-FR}) \implies \text{exp. convergence}$$

Because of the simple point-wise estimate in the above proof, it is also clear that condition (6.9) is *necessary and sufficient* for the Łojasiewicz estimate (Ł-FR).

Corollary 6.5 (Necessary sufficient condition for power-like entropy) *The Łojasiewicz inequality* (Ł-FR) *for the Fisher-Rao gradient system with the power-like entropy* φ_p (1.15) *energy,* $(\mathcal{M}^+, \mathrm{D}_{\varphi_p}, \mathsf{FR})$*, holds globally if and only if* $p \leq \frac{1}{2}$*. Furthermore, the constant is* $c_* = 1/(1-p)$ *in that case.*

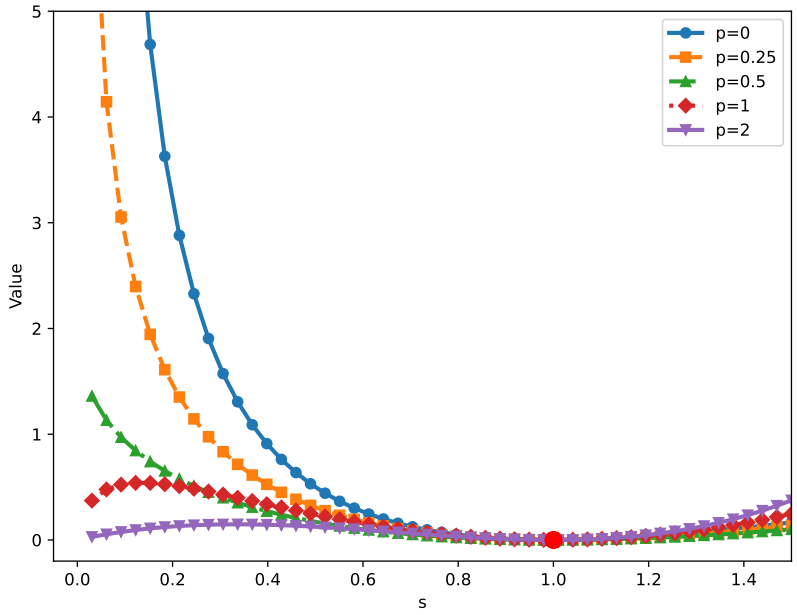


Figure 4: The plot illustrates the left-hand side Łojasiewicz inequality (Ł-FR) for the Fisher-Rao geometry. The purple curve represents $p = 0$ (reverse KL), the green curve represents $p = 0.25$, the blue curve represents $p = 0.5$ (FR), the red curve represents $p = 1$ (KL), and the orange curve represents $p = 2$ (χ^2). The red dot represents the equilibrium at $s = 1$. This plot provides insights into the slopes of the power-like entropies in the Fisher-Rao gradient flow. We observe the threshold $p = 0.5$ (FR; green) where the behavior near $s = 0$ jumps. See the main text, especially Remark 6.6, for analysis.

In particular, Corollary 6.5 shows that the energies for which the globally Łojasiewicz estimate holds include the squared Fisher-Rao ($p = \frac{1}{2}$), the reverse KL ($p = 0$), and the power-like entropies between those. On the negative side, it states that the Łojasiewicz

estimate does not hold globally for many commonly used entropy energy functionals such as the KL ($p = 1$) and χ^2 ($p = 2$).

Remark 6.6 (Entropy power threshold $p = \frac{1}{2}$) *The relevance of the threshold $p = 1/2$ can be seen on two ways. First, we observe that $\mu = 0$ is a steady state solution for the gradient systems $\big(\mathcal{M}^+, \mathrm{D}_{\phi_p}(\cdot|\pi), \mathsf{FR}\big)$ 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 metric slope*

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

can be calculated explicitly and satisfies the relation

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

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 satisfying

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

We refer to (Laschos and Mielke, 2023, Section 2) for a more detailed discussion.

A second way to see the importance of the threshold $p \le \frac{1}{2}$ involves the results in Otto and Villani (2000), showing that geodesic Λ-convexity of a functional implies the Łojasiewicz inequality with $c_{\text{Łoj}} = 2\Lambda$. For the condition of geodesic Λ-convexity for functionals $\mathrm{D}_\varphi(\mu|\pi) = \int_X \varphi(\frac{\mathrm{d}\mu}{\mathrm{d}\pi})\,\mathrm{d}\pi$ in the FR geometry, it can be shown that

$$\Lambda := \inf_{w \ge 0} \Big\{ w\,\varphi''(w) + \frac{1}{2}\,\varphi'(w) \Big\}.$$

This gives the same result when considering the p-power family φ_p. But for general φ, we may have $2\Lambda \lneqq c_{\text{Łoj}}$.

For the Wasserstein distance, the McCann condition (see, e.g., (Ambrosio et al., 2005)) shows that $\mathrm{D}_{\varphi_p}(\cdot|\,\mathrm{d}x)$ is geodesically convex only for $p \ge (d-1)/d$ where d is the dimension. In Liero et al. (2023), necessary and sufficient conditions for the geodesic convexity of entropy functionals with respect to the Wasserstein-Fisher-Rao 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 $d \ge 3$ only the range $p \ge 1$ is admitted. However, the convexity constant Λ equals 0 for all $p \ge 1$. In relating those results to ours, we first note that geodesic convexity implies Łojasiewicz inequality but only with a non-negative constant $c \ge 0$. As the dimension increases, Liero et al. (2023)'s result and the McCann condition have an increasing power threshold for the value of p. For dimension $d \ge 3$, their intervals no longer overlap with our threhold of $p \le \frac{1}{2}$ for the global Łojasiewicz in the Fisher-Rao geometry.

By combining the McCann condition and our result, we obtain the following corollary. Note that it was implied by Liero et al. (2023)'s geodesic convexity condition, but we are able to provide a positive constant lower bound instead of a non-negative one.

Corollary 6.7 *For power-like entropy energy with order $p \in [\frac{d-1}{d}, \frac{1}{2}]$, the global Łojasiewicz inequality is satisfied for the Wasserstein-Fisher-Rao gradient system with a constant $c_* \geq \frac{1}{1-p}$.*

We further compare our results with the literature in Table 2 for greater clarity.

6.2 Kernel discrepancies and entropy dissipation in kernelized geometries

We now consider entropy dissipation in the kernelized Fisher-Rao gradient flow. First, as in the pure Fisher-Rao geometry in Lemma 6.2, the KL-entropy energy does not have "enough slope" near the zero.

Lemma 6.8 *Suppose the kernel k is bounded, i.e., $\|k\|_\infty < \infty$. Then, there exists no constant $0 < c < \infty$ such that*

$$\int \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \log \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \cdot \mathcal{K}_\mu \log \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \, \mathrm{d}\pi \geq c \cdot \mathrm{D}_{\mathrm{KL}}(\mu(t) \| \pi). \tag{6.10}$$

Therefore, there is no global Łojasiewicz inequality for the kernelized Fisher-Rao gradient flow of the KL-entropy.

By Corollary 6.5, we find

Corollary 6.9 (Necessary condition for Łojasiewicz in kernelized FR flow of entropy) *Suppose the global Łojasiewicz inequality holds for kernelized-Fisher flow of power-like entropy φ_p (1.15) energy. If the kernel is bounded, then $p \leq \frac{1}{2}$.*

A sufficient condition for the Łojasiewicz type results in kernelized geometries is unclear at the moment. However, one contribution of this section is to show that the entropy dissipation in the kernelized gradient flows generates a class of discrepancies meaningful for machine learning applications, in the form of interaction energy. Similar to the pure Fisher-Rao and Wasserstein dissipation in (6.4), (6.3), we obtain the following.

Proposition 6.10 (Kernel discrepancies via entropy dissipation in kernelized FR) *The dissipation of energy F in the kernelized Fisher-Rao gradient flow (FR_k) is an interaction energy and a kernel discrepancy between measures:*

$$\mathcal{I}_F^{\mathsf{FR}_k}(\mu) = \int \int \frac{\delta F}{\delta \mu}[\mu](x) \; k(x,y) \; \frac{\delta F}{\delta \mu}[\mu](y) \, \mathrm{d}\mu(x) \, \mathrm{d}\mu(y). \tag{6.11}$$

If the energy is the φ-divergence energy, i.e., $F(\mu) = \mathrm{D}_\varphi(\mu\|\pi)$, then, the dissipation is

$$\mathcal{I}_\varphi^{\mathsf{FR}_k}(\mu|\pi) = \int \int \frac{\mathrm{d}\mu}{\mathrm{d}\pi}(x) \cdot \varphi'\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}(x)\right) k(x,y) \frac{\mathrm{d}\mu}{\mathrm{d}\pi}(y) \cdot \varphi'\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}(y)\right) \mathrm{d}\pi(x) \, \mathrm{d}\pi(y). \tag{6.12}$$

Specifically, for the power-like entropy (1.15)*, we find*

$$\mathcal{I}_p^{\mathsf{FR}_k}(\mu|\pi) = \frac{1}{(p-1)^2}\int\int \left(\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}(x)\right)^{p-1} - 1\right) k(x,y) \left(\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}(y)\right)^{p-1} - 1\right) \mathrm{d}\mu(x)\,\mathrm{d}\mu(y). \tag{6.13}$$

Specifically, we find the commonly used entropy dissipation as kernel discrepancies:

1. $p = 0$, the reverse KL energy, we obtain $\mathrm{MMD}^2(\mu,\pi)$ as its dissipation

$$\mathcal{I}_0^{\mathsf{FR}_k}(\mu|\pi) = \int\int (\mu(x) - \pi(x))\, k(x,y)\, (\mu(y) - \pi(y))\, \mathrm{d}x\,\mathrm{d}y, \tag{6.14}$$

2. $p = \frac{1}{2}$, the squared Fisher-Rao energy,

$$\mathcal{I}_{\frac{1}{2}}^{\mathsf{FR}_k}(\mu|\pi) = 4\int\int \left(\sqrt{\mu(x)} - \sqrt{\pi(x)}\right) k(x,y) \left(\sqrt{\mu(y)} - \sqrt{\pi(y)}\right) \mathrm{d}\sqrt{\mu}(x)\,\mathrm{d}\sqrt{\mu}(y), \tag{6.15}$$

3. $p = 1$, the KL-entropy energy,

$$\mathcal{I}_1^{\mathsf{FR}_k}(\mu|\pi) = \int\int \log\frac{\mathrm{d}\mu}{\mathrm{d}\pi}(x) k(x,y) \log\frac{\mathrm{d}\mu}{\mathrm{d}\pi}(y)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y), \tag{6.16}$$

4. $p = 2$, the χ^2-divergence energy,

$$\mathcal{I}_2^{\mathsf{FR}_k}(\mu|\pi) = \int\int \left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}(x) - 1\right)^2 k(x,y) \left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}(y) - 1\right)^2 \mathrm{d}\mu(x)\,\mathrm{d}\mu(y). \tag{6.17}$$

The MMD enjoys a computational advantage as it allows Monte Carlo estimation and does not require the measures μ, π to have common support. (6.14) reveals the insight that this is due to the reverse KL-entropy dissipation structure. Furthermore, we obtain an interesting insight regarding optimization

Lemma 6.11 *The gradient flow equation generated by the pure Fisher-Rao gradient system with the squared MMD energy* $\left(\mathcal{M}^+, \mathrm{MMD}^2(\cdot,\pi), \mathsf{FR}\right)$ *conincides with that of the kernelized Fisher-Rao gradient system with the reverse KL-entropy energy* $(\mathcal{M}^+, \mathrm{D}_{\varphi_0}(\cdot|\pi), \mathsf{FR}_k)$*, i.e.,*

$$\dot{\mu} = -\mu \cdot \mathcal{K}\left(\mu - \pi\right).$$

This means that solving the optimization problem $\min_\mu \mathrm{MMD}^2(\mu,\pi)$ using *mirror descent* in the pure Fisher-Rao geometry is equivalent to $\min_\mu \mathrm{D}_{\varphi_0}(\mu|\pi)$ in the kernelized Fisher-Rao geometry.

It is already established that the dissipation of the KL-divergence in the Stein geometry results in the Stein-Fisher information (Duncan et al., 2019), also referred to as the squared kernel Stein discrepancy (KSD) (Liu et al.). First, we generalize this result.

Proposition 6.12 (Kernel discrepancies via entropy dissipation in Stein) *Energy dissipation in the Stein gradient flow follows*

$$\mathcal{I}_F^{Stein}(\mu(t)) = \int\int \nabla \frac{\delta F}{\delta \mu}[\mu](x)\ k(x,y)\ \nabla \frac{\delta F}{\delta \mu}[\mu](y)\, \mathrm{d}\mu(x)\, \mathrm{d}\mu(y). \tag{6.18}$$

For the φ-divergence energy, i.e., $F(\mu) = \mathrm{D}_\varphi(\mu\|\pi)$,

$$\mathcal{I}_\varphi^{\mathrm{Stein}}(\mu|\pi) = \int\int \frac{\mathrm{d}\mu}{\mathrm{d}\pi}(x) \nabla \varphi'\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}(x)\right) k(x,y) \frac{\mathrm{d}\mu}{\mathrm{d}\pi}(y) \nabla \varphi'\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}(y)\right) \mathrm{d}\pi(x)\, \mathrm{d}\pi(y). \tag{6.19}$$

In the case of the power-like entropy (1.15)*,*

$$\mathcal{I}_p^{\mathrm{Stein}}(\mu|\pi) = -\int\int \left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}(x)\right)^{p-1} \nabla \frac{\mathrm{d}\mu}{\mathrm{d}\pi}(x) k(x,y) \nabla \frac{\mathrm{d}\mu}{\mathrm{d}\pi}(y) \left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}(y)\right)^{p-1} \mathrm{d}\pi(x)\, \mathrm{d}\pi(y). \tag{6.20}$$

Specifically, we find

1. $p = 0$, the reverse KL energy, we obtain $\mathrm{KSD}^2(\pi|\mu)$, i.e., squared reverse KSD

$$\mathcal{I}_0^{\mathrm{Stein}}(\mu|\pi) = -\int\int \nabla \log \frac{\mathrm{d}\mu}{\mathrm{d}\pi}(x) k(x,y) \nabla \log \frac{\mathrm{d}\mu}{\mathrm{d}\pi}(y)\, \mathrm{d}\pi(x)\, \mathrm{d}\pi(y). \tag{6.21}$$

2. $p = \frac{1}{2}$, the squared Fisher-Rao energy,

$$\mathcal{I}_{\frac{1}{2}}^{\mathrm{Stein}}(\mu|\pi) = -4\int\int \nabla \sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\pi}(x)} k(x,y) \nabla \sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\pi}(y)}\, \mathrm{d}\pi(x)\, \mathrm{d}\pi(y). \tag{6.22}$$

3. $p = 1$, the KL-entropy energy, we obtain $\mathrm{KSD}^2(\mu|\pi)$, a.k.a. the Stein-Fisher information

$$\mathcal{I}_1^{\mathrm{Stein}}(\mu|\pi) = -\int\int \nabla \log \frac{\mathrm{d}\mu}{\mathrm{d}\pi}(x) k(x,y) \nabla \log \frac{\mathrm{d}\mu}{\mathrm{d}\pi}(y)\, \mathrm{d}\mu(x)\, \mathrm{d}\mu(y). \tag{6.23}$$

4. $p = 2$, the χ^2-divergence energy,

$$\mathcal{I}_2^{\mathrm{Stein}}(\mu|\pi) = -\int\int \nabla \frac{\mathrm{d}\mu}{\mathrm{d}\pi}(x) k(x,y) \nabla \frac{\mathrm{d}\mu}{\mathrm{d}\pi}(y)\, \mathrm{d}\mu(x)\, \mathrm{d}\mu(y). \tag{6.24}$$

Furthermore, while researchers previously have drawn connections between the MMD and the KSD (Arbel et al., 2019; Mroueh et al., 2019), ours is a new connection from the perspective of the principled dynamics – they are both generated by the 0-th order power (reverse KL) entropy dissipation in the kernelized geometries.

6.3 Explicit connection between nonparametric regression and Rayleigh Principle

In this section, we uncover a connection between the entropy dissipation in gradient flows and the nonparametric regression

$$\operatorname*{argmin}_{f\in\mathcal{F}} \left\{ \int (f(x) - y(x))^2\, \mathrm{d}\mu(x) + \lambda \|f\|_{\mathcal{F}}^2 \right\}, \tag{6.25}$$

where y is some target functions such as the generalized force $\frac{\delta F}{\delta \mu}[\mu]$ for Fisher-Rao and $\nabla \frac{\delta F}{\delta \mu}[\mu]$ for Wasserstein. This can be further unified under the general MLE formulation with negative log-likelihood

$$\operatorname*{argmin}_{f \in \mathcal{F}} \mathrm{NLL}(f; y, \mu), \tag{6.26}$$

where NLL is the negative-log-likelihood, e.g., $\frac{1}{2\sigma^2}\|f - y\|^2_{L^2_\mu} + \lambda \|f\|^2_{L^2_\mu}$ with variance σ^2. We now give an explicit characterization that connects gradient flow dissipation geometry with nonparametric regression. Since Fisher-Rao and Wasserstein type flows are based on L^2 type geometries, we only focus on the least-squares type losses (6.25).

Consider the Fisher-Rao type approximate flows $\dot{\mu}^f_t = -\mu^f_t \cdot f_t$ where the approximate growth field f is obtained by solving the nonparametric regression. Here, the target of the regression is $\mathrm{D}F$, the derivative of energy F in the respective geoemtry, e.g., in the sense of Fréchet.

$$\begin{aligned}\|f - \mathrm{D}F\|^2_{L^2_{\mu^f}} + \lambda \|f\|^2_{\mathcal{F}} &= \|f\|^2_{L^2_{\mu^f}} - 2\langle f, \mathrm{D}F\rangle_{L^2_{\mu^f}} + \|\mathrm{D}F\|^2_{L^2_{\mu^f}} + \lambda \|f\|^2_{\mathcal{F}} \\ &= 2 \cdot \frac{\mathrm{d}}{\mathrm{d}t} F(\mu^f_t) + \|f\|^2_{L^2_{\mu^f}} + \lambda \|f\|^2_{\mathcal{F}} + \|\mathrm{D}F\|^2_{L^2_{\mu^f}}.\end{aligned}$$

In the Wasserstein type approximate flows $\dot{\mu}_t = \mathrm{div}(\mu_t \cdot f_t)$, a similar derivation with IBP yields

$$\begin{aligned}\|f - \nabla \mathrm{D}F\|^2_{L^2_{\mu^f}} + \lambda \|f\|^2_{\mathcal{F}} &= \|f\|^2_{L^2_{\mu^f}} - 2\langle f, \nabla \mathrm{D}F\rangle_{L^2_{\mu^f}} + \|\mathrm{D}F\|^2_{L^2_{\mu^f}} + \lambda \|f\|^2_{\mathcal{F}} \\ &\overset{\text{(IBP)}}{=} 2 \cdot \frac{\mathrm{d}}{\mathrm{d}t} F(\mu^f_t) + \|f\|^2_{L^2_{\mu^f}} + \lambda \|f\|^2_{\mathcal{F}} + \|\nabla \mathrm{D}F\|^2_{L^2_{\mu^f}}.\end{aligned}$$

Combining those two cases, we obtain the similar formulation for the Wasserstein-Fisher-Rao type flows,

$$\begin{aligned}&\|g - \mathrm{D}F\|^2_{L^2_{\mu^{g,h}}} + \|h - \nabla \mathrm{D}F\|^2_{L^2_{\mu^{g,h}}} + \lambda(\|g\|^2_{\mathcal{F}} + \|h\|^2_{\mathcal{F}}) \\ &\quad = 2 \cdot \frac{\mathrm{d}}{\mathrm{d}t} F(\mu^{g,h}_t) + \|g\|^2_{L^2_{\mu^{g,h}}} + \|h\|^2_{L^2_{\mu^{g,h}}} + \lambda(\|g\|^2_{\mathcal{F}} + \|h\|^2_{\mathcal{F}}) + \|\mathrm{D}F\|^2_{L^2_{\mu^f}} + \|\nabla \mathrm{D}F\|^2_{L^2_{\mu^{g,h}}}.\end{aligned}$$

By setting the product $f = g \otimes h$, we observe that the formal calculation in different geometries above results in the same optimization objectives on the right-hand sides independent of the dissipation geometries.

Proposition 6.13 (Nonparametric regression as Rayleigh Principle) *In approximate Fisher-Rao, Wasserstein, and Wasserstein-Fisher-Rao flows, the nonparametric regression* (6.25) *is equivalent to the optimization problem*

$$\operatorname*{argmin}_{f \in \mathcal{F}} \left\{ \frac{\mathrm{d}}{\mathrm{d}t} F(\mu^f_t) + \frac{1}{2}\|f\|^2_{L^2_\mu} + \frac{\lambda}{2}\|f\|^2_{\mathcal{F}} \right\}. \tag{6.27}$$

Furhermore, solving the nonparametric regression (6.25) *is equivalent to the minimization problem in the form of the Rayleigh Principle (Rayleigh, 1873)*

$$\operatorname*{argmin}_{f\in\mathcal{F}} \underbrace{\langle \mathrm{D}F, \frac{\mathrm{d}}{\mathrm{d}t}\mu_t^f\rangle}_{\textit{energy}} + \underbrace{\frac{1}{2}\|f\|_{L^2_\mu}^2 + \frac{\lambda}{2}\|f\|_{\mathcal{F}}^2}_{\textit{dissipation potential}}.$$

Therefore, the approximation of the gradient flow is equivalent to searching for the approximate growth or velocity field $f \in \mathcal{F}$ that maximally dissipates the energy F, while regularized by its function class $\mathcal{F}$, L^2 norm, and optionally a regularization term.

Remark 6.14 (Helmholtz-Rayleigh Principle) *While our discussion around Proposition 6.13 is formal, it can be made mathematically rigorous in the form of the (Helmholtz-)Rayleigh Principle (Rayleigh, 1873), also known as the maximum dissipation principle; see (Mielke, 2015, Proposition 5.2.1) for the rigorous statement. Our notation for the dissipation geometry $\mathcal{R}, \mathcal{R}^*$ is also due to the Rayleigh dissipation function.*

As discussed in the previous subsection, the 0-th order power entropy dissipation generates the MMD. Using Proposition 6.13, we show further connection to tools such as ISM and IPM, explained in the following examples. First, Proposition 6.13 specialized to the Wasserstein setting gives a connection to generative models.

Example 6.3 (Wasserstein case as implicit score-matching) *A standard technique for solving a score-matching problem is the implicit score-matching (ISM) for solving regression*

$$\operatorname*{argmin}_{f\in\mathcal{F}} \|f - \nabla \log \frac{\mu}{\pi}\|_{L^2_\mu}^2. \tag{6.28}$$

Apply Proposition 6.13 (with $\lambda = 0$) in the Wasserstein setting with the KL entropy energy, we recover the ISM result (Hyvarinen; Vincent, 2011)

$$\operatorname*{argmin}_{f\in\mathcal{F}} \left\{ -\langle f, \nabla \log \frac{\mu}{\pi}\rangle_{L^2_\mu}^2 + \|f\|_{L^2_\mu}^2 \overset{(IBP)}{=} \int \left(f^2 + \operatorname{div} f + f\cdot \nabla \log \pi\right) \mu \right\}. \tag{6.29}$$

In practice, the estimator $\hat{f}$ in (6.29) *can be fitted using, e.g., neural networks or random Fourier features, and used as the velocity field to update the particle locations, i.e., performing Langevin update $X_{t+1} \leftarrow X_t + \tau \cdot \hat{f}(X_t)$. This technique has also been applied with neural network approximation for sampling by Dong et al. (2022).*

Going beyond Wasserstein, specializing Proposition 6.13 to the Fisher-Rao setting results in a commonly used tool in machine learning applications.

Example 6.4 (Reverse KL dissipation in approximate FR^2 and two-sample test) *Let the energy functional be the reverse KL (0-th order power) entropy dissipation $F(\mu) = \mathrm{D}_0(\mu|\pi)$. Energy-dissipation formulation of nonparametric regression* (6.27) *is equivalent to the weak-norm formulation*

$$\sup_{f\in\mathcal{F}} \left\{ \int f \,\mathrm{d}\,(\mu - \pi) - \frac{1}{2}\|f\|_{L^2_\mu}^2 - \frac{\lambda}{2}\|f\|_{\mathcal{F}}^2 \right\}$$

We have already seen this type of weak-norms in the de-kernelized and flattened geometries. Weak-norm formulations are also commonly used in machine learning applications such as generative models (Nowozin et al., 2016), two-sample testing (Gretton et al., 2012), and robust learning under distributional shifts (Zhu et al., 2021).

Another implication of Proposition 6.13 is that the Stein gradient flow can not be cast in the form (6.27) since kernel smoothing is not an *M-estimator*, i.e., it cannot be written as the solution of an optimization problem like (6.25). However, it is possible to cast it as a local regression; cf. (Spokoiny, 2016; Tsybakov, 2009) for details.

Example 6.5 (Approximation in the Stein setting as local regression and MLE) *The particle approximation scheme for the Stein PDE* (1.4) *was first proposed as the SVGD algorithm by Liu and Wang (2019). We now cast the approximation in the Stein setting as local regression and thus local MLE*

$$\dot{\mu_t} = \operatorname{div}(\mu_t \cdot v_t), \quad v_t(x) = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left\{ \int \mu(x') k(x' - x) \left| \theta - \nabla \frac{\delta F}{\delta \mu} [\mu_t] (x') \right|^2 \mathrm{d}x' \right\}, \tag{6.30}$$

where the regression is now locally weighted using the shift-invariant kernel $k(x' - x)$. This problem admits a closed-form solution as a kernelized velocity

$$v_t(x) = \int \mu(x') \frac{k(x' - x)}{\int \mu(x') k(x' - x) \, \mathrm{d}x'} \nabla \frac{\delta F}{\delta \mu} [\mu_t] (x') \, \mathrm{d}x'. \tag{6.31}$$

Given data samples $\{x_i\}_{i=1}^N$, we obtain the kernel smoothing method known as the Nadaraya-Watson estimator $\hat{v}(x) = \dfrac{1}{N} \sum_{i=1}^N \dfrac{k(x_i - x)}{\sum_{i=1}^N k(x_i - x)} \cdot \nabla \dfrac{\delta F}{\delta \mu} [\mu_t] (x_i)$, which is the SVGD velocity with a normalized kernel.

Example 6.6 (Nadaraya-Watson estimator in the kernelized Fisher-Rao setting) *Similar to Stein, in the kernelized Fisher-Rao setting, we have*

$$\dot{\mu_t} = -\mu_t \cdot r_t, \quad r_t(x) = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left\{ \int \mu(x') k(x' - x) \left| \theta - \frac{\delta F}{\delta \mu} [\mu_t] (x') \right|^2 \mathrm{d}x' \right\}, \tag{6.32}$$

with the closed-form solution as a kernelized growth field

$$r_t(x) = \int \mu(x') \frac{k(x' - x)}{\int \mu(x') k(x' - x) \, \mathrm{d}x'} \frac{\delta F}{\delta \mu} [\mu_t] (x') \, \mathrm{d}x', \tag{6.33}$$

and a Nadaraya-Watson estimator $\hat{r}(x) = \dfrac{1}{N} \sum_{i=1}^N \dfrac{k(x_i - x)}{\sum_{i=1}^N k(x_i - x)} \cdot \dfrac{\delta F}{\delta \mu} [\mu_t] (x_i)$.

In particular, setting the energy as the 0-th order power entropy dissipation $F(\mu) = \mathrm{D}_0(\mu|\pi)$ in (6.33)*, we obtain growth field $\int \frac{k(x'-x)}{\int \mu(x') k(x'-x) \, \mathrm{d}x'} (\mu(x') - \pi(x')) \, \mathrm{d}x'$. Given two samples $\{x_i\}_{i=1}^N \sim \mu$ and $\{x'_i\}_{i=1}^M \sim \pi$, A sample based estimator of the growth field is the difference between two kernel density estimators*

$$\hat{r}(x) = \frac{1}{\sum_{i=1}^N k(x_i - x)} \cdot \left(\frac{1}{N} \sum_{i=1}^N k(x_i - x) - \frac{1}{M} \sum_{i=1}^M k(x'_i - x) \right).$$

6.4 Energy dissipation in kernel-approximate flows: formal arguments

To set the stage for the evolutionary Γ-convergence results in Section 6.5, we now analyze the energy dissipatioin in the kernel-approximate gradientn flows. First, we examine the kernel-regularized Fisher-Rao setting studied in Section 3.1. For the RKHS approximation in this paper, we may use the standard approximation theory characterization, e.g., (Cucker and Zhou, 2007, Chapter 8), which was applied to the Stein setting by He et al. (2022). One difference is that we do not make the assumption that $\frac{\delta F}{\delta \mu}[\mu] \in \mathrm{range}(\mathcal{K}_\mu{}^s)$. In the Wasserstein or the Fisher-Rao setting, the regularity of $\frac{\delta F}{\delta \mu}[\mu]$ or $\nabla \frac{\delta F}{\delta \mu}[\mu]$ is determined by the gradient system and geodesic structure introduced in Section 2. It is not clear whether the range condition $\xi \in \mathrm{Range}(\mathcal{K}_\mu^s)$ can be satisfied in any meaningful gradient systems. Furthermore, since $\mathcal{K}_\mu$ is a compact operator, hence the quantity $\|\mathcal{K}_\mu^{-\frac{s}{2}}\, \xi\|_{L^2_\mu}^2$ is unbounded. Instead, we rely on regularization to avoid unbounded estimate. We emphasize that the following arguments are merely formal and we provide a rigorous justification in Section 6.5.

Proposition 6.15 (Energy dissipation of kernel-approximate Fisher-Rao flows) *For $0 \le s \le 1$, the energy dissipation satisfies*

$$\frac{\mathrm{d}}{\mathrm{d}t} F(\mu(t)) \le \underbrace{-\|\frac{\delta F}{\delta \mu}[\mu_t]\|_{L^2_\mu}^2}_{\textit{FR dissipation}} + \underbrace{\lambda^s \|\, (\mathcal{K}_\mu + \lambda\, \mathrm{Id})^{-\frac{s}{2}}\, \frac{\delta F}{\delta \mu}[\mu_t]\, \|_{L^2_\mu}^2}_{\textit{approximation error}}.$$

Note that the estimate $\|\, (\mathcal{K}_\mu + \lambda\, \mathrm{Id})^{-\frac{s}{2}}\, \xi\|_{L^2_\mu}^2$ is finite for any fixed λ.

The implication of the above estimate can be seen in the following scenario. Suppose the Łojasiewicz condition in the pure Fisher-Rao geoemtry holds, i.e.,

$$\|\frac{\delta F}{\delta \mu}[\mu]\, \|_{L^2_\mu}^2 \ge c \cdot \left(F(\mu) - \inf_\mu F(\mu) \right). \tag{6.34}$$

Note that we have already justified the global Łojasiewicz inequality holds for the power-like entropy energy satisfying our threshold condition. By the generalized Gronwall's lemma, the energy decay of the approximate Fisher-Rao gradient system satisfies the following estimate. For $0 \le s \le 1$,

$$F(\mu(T)) - \inf_\mu F(\mu) \le \underbrace{\mathrm{e}^{-c\cdot T} \cdot \left(F(\mu(0)) - \inf_\mu F(\mu) \right)}_{\text{FR flow}} + \underbrace{\int_0^T \mathrm{e}^{-c\cdot(T-t)} \cdot \lambda^s \|\, (\mathcal{K}_{\mu_t} + \lambda\, \mathrm{Id})^{-\frac{s}{2}}\, \frac{\delta F}{\delta \mu}[\mu_t]\, \|_{L^2_{\mu_t}}^2 \,\mathrm{d}t}_{\text{approximation error}}. \tag{6.35}$$

As mentioned, those are merely formal arguments and do not justify the asymptotic convergence behavior. However, based on our characterization of the entropy dissipation in the pure Fisher-Rao geometry, we expect the approximation such as in Proposition 6.15 to be close. Next, we give a rigorous justification, at the rigor level of applied analysis, of convergence as $\lambda \to 0$ for the kernel-approximate Fisher-Rao flows.

6.5 Evolutionary Γ-convergence at the approximation limit

As the regularization parameter $\lambda \to 0$, techniques from approximation theory and statistics can be used to show consistency for *fixed time t point-wise*. That is, the regression problems such as (3.7) (4.6) yield the velocity v_t or growth r_t that converge to the target counterpart, e.g., $v_t \overset{\lambda\to 0^+}{\longrightarrow} \nabla \frac{\delta F}{\delta \mu}[\mu_t]$ for fixed t. However, this point-wise convergence *does not directly imply the convergence or the existence* of the approximating gradient systems. For example, it has not been shown that the approximate system generating $\dot{\mu}_t = -\mu_t \cdot r_t$ in (3.7) converges to the pure Fisher-Rao gradient system.

Different from the point-wise convergence of regression problems, we now rigorously justify this approximation limit of the gradient systems using evolutionary Γ-convergence in this section. We focus on the kernel approximation to the Fisher-Rao gradient system in Section 3.1; see (3.7), (3.6). Note that the rigor level of this subsection's analysis is elevated above the rest of the paper, i.e., not merely formal arguments. A rigorous result for the approximate Wasserstein(-Fisher-Rao) setting is beyond our current scope due to the technicality involved.

For general curves $u : [0,T] \to \mathbf{X}$, where $\mathbf{X}$ denotes the state space, we define the dissipation functionals $\mathfrak{D}_\lambda$ as

$$\mathfrak{D}_\lambda(u) := \int_0^T \left(\mathcal{R}_\lambda(u(t), \dot{u}(t)) + \mathcal{R}^*_\lambda\left(u(t), -\mathrm{D}F(u(t))\right)\right) \mathrm{d}t,$$
$$\mathfrak{D}_0(u) = \int_0^T \left(\mathcal{R}(u(t), \dot{u}(t)) + \mathcal{R}^*\left(u(t), -\mathrm{D}F(u(t))\right)\right) \mathrm{d}t.$$

The energy-dissipation principle (EDP) states that, under suitable technical assumptions (cf. (Mielke, 2023, Theorem 3.9)), $u : [0,T] \to \mathbf{X}$ is a solution to the gradient-flow equation (2.2) if and only if it satisfies the following energy-dissipation inequality:

$$F(u(t)) + \mathfrak{D}_\lambda(u) \le F(u(0)). \tag{6.36}$$

Thus, $\mathfrak{D}_\lambda$ intrinsically encodes the gradient-flow dynamics.

Definition 6.16 (EDP-convergence) *A sequence of gradient systems $(\mathbf{X}, F, \mathcal{R}_\lambda)$ is said to* converge in the sense of the energy-dissipation principle *(EDP-converge) to $(\mathbf{X}, F, \mathcal{R})$, shortly written as $(\mathbf{X}, F, \mathcal{R}_\lambda) \overset{\text{EDP}}{\longrightarrow} (\mathbf{X}, F, \mathcal{R})$, if $\mathfrak{D}_\lambda$ Γ-converges to $\mathfrak{D}_0$ with bounded energies for all $T > 0$, i.e.,*

$$(\Gamma\text{inf}) \quad u_\lambda \to u \text{ and } \sup_{\lambda>0, 0\le t\le T} F(u_\lambda(t)) < \infty \implies \liminf_{\lambda\to 0^+} \mathfrak{D}_\lambda(u_\lambda) \ge \mathfrak{D}_0(u), \tag{6.37}$$

$$(\Gamma\text{sup}) \quad \forall\, \widehat{u} \in L^2([0,T];\mathbf{X}) \ \exists\ \widehat{u}_\lambda \text{ with } \sup_{\lambda>0,\, 0\le t\le T} F(\widehat{u}_\lambda(t)) < \infty, :$$
$$\widehat{u}_\lambda \to \widehat{u} \text{ and } \limsup_{\lambda\to 0} \mathfrak{D}_\lambda(\widehat{u}_\lambda) \le \mathfrak{D}_0(\widehat{u}). \tag{6.38}$$

Recall that the dissipation geometry of the regularized-approximate Fisher-Rao gradient system in Section 3.1.

$$\mathcal{R}_{\lambda\text{-}\mathsf{FR}_k}(\rho, u) = \frac{1}{2}\langle \frac{\delta u}{\delta \rho}, \left(\mathrm{Id} + \lambda \mathcal{K}_\rho^{-1}\right) \frac{\delta u}{\delta \rho}\rangle_{L^2_\rho}, \quad \mathcal{R}^*_{\lambda\text{-}\mathsf{FR}_k}(\rho, \xi) = \frac{1}{2}\langle \xi, \left(\mathcal{K}_\rho + \lambda\, \mathrm{Id}\right)^{-1} \mathcal{K}_\rho \xi\rangle_{L^2_\rho}.$$

A useful observation is that $\mathcal{R}_{\lambda\text{-}\mathsf{FR}_k}(\rho,u)$ is decreasing with decreasing λ. In fact, it is even affine. Since the Legrendre transform is anti-monotone, $\mathcal{R}_{\lambda\text{-}\mathsf{FR}_k}(\rho,u)$ is increasing for λ decreasing to 0. To avoid technicalities, we do not show full EDP-convergence, but only the Γ-liminf estimate (6.37), this means we stay in the Γ-convergence framework of Serfaty (2011). This will be enough to conclude that solutions μ_λ of the regularized gradient-flow equation of $(\mathcal{M}^+,F,\mathcal{R}_{\lambda\text{-}\mathsf{FR}_k})$ converge to solutions μ of the pure Fisher-Rao gradient-flow equation of $(\mathcal{M}^+,F,\mathcal{R}_{\mathsf{FR}})$, see Corollary 6.18.

Theorem 6.17 (Γ-convergence of the kernel-approx. FR gradient systems) *Assume that the functional $F:\mathcal{M}^+\to\mathbb{R}$ satisfies the following assumptions:*

$$\text{the } \lambda\text{-}\mathsf{FR}_k\text{-dissipation } \mu\mapsto \mathcal{R}_{\lambda\text{-}\mathsf{FR}_k}\big(\mu,\tfrac{\delta F}{\delta\mu}(\mu)\big) \text{ is weakly lower semicontinuous.} \tag{6.39}$$

Then, the dissipation functional $\mathfrak{D}_\lambda$ for the regularized approximate Fisher-Rao gradient system $(\mathcal{M}^+,F,\mathcal{R}_{\lambda\text{-}\mathsf{FR}_k})$ satisfies the Γ-liminf estimate (6.37).

Proof. As in Serfaty (2011) we decompose $\mathfrak{D}_\lambda=\mathfrak{D}_\lambda^{\mathrm{rate}}+\mathfrak{D}_\lambda^{\mathrm{slope}}$ into a rate part and a slope part:

$$\mathfrak{D}_\lambda^{\mathrm{rate}}(\mu)=\int_0^T \mathcal{R}_{\lambda\text{-}\mathsf{FR}_k}(\mu,\dot\mu)\,\mathrm{d}t \quad\text{and}\quad \mathfrak{D}_\lambda^{\mathrm{slope}}(\mu)=\int_0^T \mathcal{R}^*_{\lambda\text{-}\mathsf{FR}_k}\big(\mu,\tfrac{\delta F}{\delta\mu}(\mu)\big)\,\mathrm{d}t.$$

(A) Extraction of a converging subsequence. By standard arguments, it is sufficient to consider a family $(\mu_\lambda)_{\lambda>0}$ of curves with $\mathfrak{D}_\lambda(\mu_\lambda)\le C<\infty$. Using $\mathcal{R}_{\lambda\text{-}\mathsf{FR}_k}\ge\mathcal{R}_{\mathsf{FR}}$ we obtain $\int_0^T|\mu_\lambda'|^2_{\mathsf{FR}}(t)\,\mathrm{d}t\le C$. This implies that there exists a subsequence (not relabeled) and a limit curve μ_0 with

$$\int_0^T|\mu_0'|^2_{\mathsf{FR}}(t)\,\mathrm{d}t\le C \quad\text{and}\quad \forall\, t\in[0,T]:\ \mu_\lambda(t)\rightharpoonup\mu_0(t), \tag{6.40}$$

where weak convergence is meant in the sense of testing with continuous functions.

(B) Γ-liminf estimate for $\mathfrak{D}_\lambda^{\mathrm{rate}}$. For this we exploit $\mathcal{R}_{\lambda\text{-}\mathsf{FR}_k}\ge\mathcal{R}_{\mathsf{FR}}=\frac12|\mu'|^2_{\mathsf{FR}}$:

$$\liminf_{\lambda\downarrow0}\mathfrak{D}_\lambda^{\mathrm{rate}}(\mu_\lambda)\ge\liminf_{\lambda\downarrow0}\mathfrak{D}_0^{\mathrm{rate}}(\mu_\lambda)=\liminf_{\lambda\downarrow0}\int_0^T\frac12|\mu_\lambda'|(t)^2\,\mathrm{d}t\ge\int_0^T\frac12|\mu_0'|(t)^2\,\mathrm{d}t=\mathfrak{D}_0^{\mathrm{rate}}(\mu_0).$$

(C) Γ-liminf estimate for $\mathfrak{D}_\lambda^{\mathrm{slope}}$. To treat the slope term we use the monotonicity $\mathcal{R}^*_{\lambda\text{-}\mathsf{FR}_k}\ge\mathcal{R}^*_{\delta\text{-}\mathsf{FR}k}$ for $0<\lambda\le\delta$ and the weak lower semi-continuity assumed in (6.39). For fixed $\delta>0$ we have

$$\liminf_{\lambda\downarrow0}\mathcal{R}^*_{\lambda\text{-}\mathsf{FR}_k}(\mu_\lambda,\tfrac{\delta F}{\delta\mu}(\mu_\lambda))\ge\liminf_{\lambda\downarrow0}\mathcal{R}^*_{\delta\text{-}\mathsf{FR}k}(\mu_\lambda,\tfrac{\delta F}{\delta\mu}(\mu_\lambda))\ge\mathcal{R}^*_{\delta\text{-}\mathsf{FR}k}(\mu_0,\tfrac{\delta F}{\delta\mu}(\mu_0)),$$

where the last estimate follows because of (6.39). Integration over $t\in[0,T]$, Fatou's lemma yields

$$\liminf_{\lambda\downarrow0}\mathfrak{D}_\lambda^{\mathrm{slope}}(\mu_\lambda)\ge\int_0^T\liminf_{\lambda\downarrow0}\mathcal{R}^*_{\lambda\text{-}\mathsf{FR}_k}(\mu_\lambda,\tfrac{\delta F}{\delta\mu}(\mu_\lambda))\,\mathrm{d}t\ge\mathfrak{D}_\delta^{\mathrm{slope}}(\mu_0)\xrightarrow{\delta\downarrow0}\mathfrak{D}_0^{\mathrm{slope}}(\mu_0),$$

where the last convergence follows by the monotone convergence principle.

Hence, the desired Γ-liminf estimate for $\mathfrak{D}_\lambda = \mathfrak{D}_\lambda^{\text{rate}} + \mathfrak{D}_\lambda^{\text{slope}}$ is established. ∎

The above result allows us to conclude that the solution of the approximate Fisher-Rao gradient flows converges to that of the pure Fisher-Rao.

Corollary 6.18 (Convergence of Gradient flow solutions) *Let μ_λ be a sequence of solutions to the regularized Fisher-Rao gradient system $(\mathcal{M}^+, F, \mathcal{R}_{\lambda\text{-}\mathsf{FR}_k})$ in the sense of energy-dissipation balance. Assume that the assumptions of Theorem 6.17 are satisfied. Suppose that for all $t \in [0,T]$ we have $\mu_\lambda(t) \rightharpoonup \mu(t)$ and that $F(\mu_\lambda(0)) \to F(\mu(0)) < \infty$.*

Then, $\mu : [0,T] \to (\mathcal{M}^+, \mathsf{FR})$ is absolutely continuous and a solution to the Fisher-Rao gradient system $(\mathcal{M}^+, F, \mathsf{FR})$.

Proof. By Theorem 6.17 we know that μ satisfies

$$\int_0^T |\dot\mu|_{\mathsf{FR}}(t)^2 \,\mathrm{d}t < \infty \quad \text{and} \quad F(\mu(T)) + \mathfrak{D}_0(\mu) \le F(\mu(0)).$$

The last relation follows by the EDP for μ_λ, namely $F(\mu_\lambda(T)) + \mathfrak{D}_\lambda(\mu_\lambda) \le F(\mu_\lambda(0))$ (see (6.36)) and the limit passage $\lambda \downarrow 0$. Now exploiting the EDP for $\lambda = 0$ we see that μ is a solution for $(\mathcal{M}^+, F, \mathcal{R}_{\mathsf{FR}})$. ∎

Thus far, we have answered the question we posed earlier: the approximate flow, using a regression formulation such as (1.12), is indeed a gradient flow that converges to the target gradient-flow system such as the Fisher-Rao gradient flow, in the sense of evolutionary Γ-convergence. This provides the mathematical basis for "learning" the flow for machine learning applications.

7 Other related works

There is a well-studied line of works using regularized energies of the Wasserstein gradient flow Carrillo et al. (2019) and Fisher-Rao gradient flow Lu et al. (2023). While closely related, our work differs significantly in 1) the focus on modifying the dissipation geometry and the gradient structure instead of the energy – our flows use the same energy as the target flow. 2) the focus on regression type approximation (3.7) and (4.6) instead of relying on convolution and letting the kernel bandwidth go to zero. Point 2) is important as it lets us use straightforward replacement of the kernel machines with deep neural networks. Furthermore, Section 6.3 shows a direct connection between our regression formulation and the Rayleigh Principle for gradient flows. Related to kernel methods and Wasserstein flows, Arbel et al. (2019) studied the Wasserstein gradient flow of the MMD. Similarly, Glaser et al. (2021) studied the WGF of an interpolation between KL-divergence and the MMD. Compared to those works, we have investigated the dynamics of generalized gradient flows (not necessarily Wasserstein) and their dissipation geometries, which should not be confused with the WGF with different energy objectives. There also exist works such as Divol et al. (2022) that characterizes the statistical error in the setting of approximating the optimal transport map using the static formulation when computing the Wasserstein distance. Marzouk et al. (2023) also took a nonparametric regression perspective for ODE learning. Although their results do not concern gradient flows. Compared with works such

as Liu et al. (2023); Lu et al. (2023), our discussion in Example 6.2 distinguishes our results regarding the pure FR dynamics from theirs. Furthermore, our analysis covers a spectrum of power-like entropies with an explicit convergence threshold of $p = \frac{1}{2}$. Lastly, there exists a large body of works that use gradient flows such as Langevin or birth-death dynamics to analyze neural network training, which we do not exhaust here.

8 Further proofs

Proof of Lemma 3.6. By definition, $\mathrm{MMD}^p(\mu,\nu) = \|\mathcal{K}\mu - \mathcal{K}\nu\|_{\mathcal{H}}^p$. We introduce the auxiliary variable $f = \int k(x,\cdot)\,\mathrm{d}\mu(x)$, then apply the Lagrange duality to the constrained optimization problem

$$\inf_{f\in\mathcal{H}} \|f - \mathcal{K}\nu\|_{\mathcal{H}}^p \ \text{ s. t. } \ f = \int k(x,\cdot)\,\mathrm{d}\mu(x).$$

Finally, we associate the equality constraint $f = \int k(x,\cdot)\,\mathrm{d}\mu(x)$ with the dual variable $h \in \mathcal{H}$. By the first order optimality condition,

$$2\cdot(f - \mathcal{K}\nu) = -h, \ \ f = \mathcal{K}\nu - \frac{1}{2}h. \tag{8.1}$$

Hence,

$$\mathrm{MMD}^2(\mu,\nu) = \sup_{h\in\mathcal{H}} \int h\,\mathrm{d}(\nu-\mu) + \|\frac{h}{2}\|_{\mathcal{H}}^2 - \frac{1}{2}\|h\|_{\mathcal{H}}^2 \tag{8.2}$$

$$= \sup_{h\in\mathcal{H}} \int h\,\mathrm{d}(\nu-\mu) - \frac{1}{4}\|h\|_{\mathcal{H}}^2. \tag{8.3}$$

The optimizing h^* can be further obtained by directly solving the quadratic program. ∎

For aesthetic reasons, we need the following lemma whose derivation is an exercise in convex analysis.

Lemma 8.1 *For a scaling parameter $\tau > 0$, we find*

$$\frac{1}{2\tau}\cdot\mathrm{MMD}^2(\mu,\nu) = \sup_{h\in\mathcal{F}} \int h\,\mathrm{d}(\mu-\nu) - \frac{\tau}{2}\|h\|_{\mathcal{H}}^2. \tag{8.4}$$

Proof of Theorem 3.7. We derive the force-kernelized MMD gradient flow, i.e., we replace the differential ξ_t with its kernelization $\mathcal{K}_{\mu_t}^{\frac{1}{2}}\xi_t$ by Definition 3.1. The boundary conditions of (3.11) are trivially equivalent to $\mu(0) = \nu, \mu(1) = \mu$. Following the dynamic definition of the Stein distance, the cost of the trajectory optimization problem in the MMD formulation $\frac{1}{2}\|\xi_t\|_{\mathcal{H}}^2 = \frac{1}{2}\langle\xi_t, \mathcal{K}^{-1}\xi_t\rangle_{L^2_\mu}$, as well as the dual dissipation potential, should be replaced with an additional (weighted) $\mathcal{K}_{\mu_t}$ operation by

$$\frac{1}{2}\langle\mathcal{K}^{-\frac{1}{2}}\mathcal{K}_{\mu_t}^{\frac{1}{2}}\xi_t, \mathcal{K}^{-\frac{1}{2}}\mathcal{K}_{\mu_t}^{\frac{1}{2}}\xi_t\rangle_{L^2} = \frac{1}{2}\|\xi_t\|_{L^2(\mu_t)}^2.$$

Using this gradient structure, the MMD gradient-flow equation is obtained as

$$\mathcal{K}\dot{\mu} = -\mathcal{K}_{\mu_t}\xi_t.$$

Since the convolutional operator $\mathcal{K}$ is positive definite, this gradient-flow equation is equivalent to the reaction equation in (1.11), i.e., $\dot{\mu} = -\mu_t \cdot \xi_t$. Therefore, we have recovered the Fisher-Rao geodesics. ∎

Proof of Proposition 3.9. If $s = 0$, then $\omega(0) = \nu$. By Fenchel-duality, the linearized FR can be written as

$$\begin{aligned}\sup_{\zeta} \int \zeta \,\mathrm{d}(\mu - \nu) - \frac{1}{4}\|\zeta\|_{L^2_\nu}^2 &= \sup_{\zeta} \int \left(\zeta \cdot (\frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1) - \frac{1}{4}\zeta^2\right) \mathrm{d}\nu \\ &= \int \left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1\right)^2 \mathrm{d}\nu = \mathrm{D}_{\chi^2}(\mu|\nu),\end{aligned}$$

hence the equivalence to the χ^2-divergence. The case of $s = 1, \omega(1) = \mu$ is similar.

If $s = \frac{1}{2}$, then $\omega(\frac{1}{2}) = \frac{1}{4}(\sqrt{\mu} + \sqrt{\nu})^2$. We find

$$\begin{aligned}&\sup_{\zeta} \int \left(\zeta \cdot \frac{\sqrt{\mu} - \sqrt{\nu}}{\sqrt{\mu} + \sqrt{\nu}} - \frac{1}{16}\zeta^2\right) \cdot \left(\sqrt{\mu} + \sqrt{\nu}\right)^2 \mathrm{d}x \\ &\qquad = \int 4\left(\frac{\sqrt{\mu} - \sqrt{\nu}}{\sqrt{\mu} + \sqrt{\nu}}\right)^2 \cdot \left(\sqrt{\mu} + \sqrt{\nu}\right)^2 \mathrm{d}x = 4\|\sqrt{\mu} - \sqrt{\nu}\|_{L^2}^2 = \mathsf{FR}^2(\mu, \nu). \quad (8.5)\end{aligned}$$

∎

Proof of Proposition 4.2. Let $\nabla\zeta_t \in \nabla C_0^\infty$ be the test function with zero boundary condition. Taking time derivative along the flow solution,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\int \zeta_t\mu_t &\overset{\text{(product)}}{=} \int \partial_t\zeta_t\mu_t + \int \zeta_t\partial_t\mu_t \overset{\text{(dynamics)}}{=} \int \partial_t\zeta_t\mu_t - \int \zeta_t \,\mathrm{div}(\mathcal{K}^{-1}\nabla\xi) \\ &\overset{\text{(IBP)}}{=} \int \partial_t\zeta_t\mu_t + \int\int \mathcal{K}^{-\frac{1}{2}}\nabla\zeta_t\mathcal{K}^{-\frac{1}{2}}\nabla\xi \,\mathrm{d}x\,\mathrm{d}t. \quad (8.6)\end{aligned}$$

Completing the squares for the last term

$$\begin{aligned}\int\int \mathcal{K}^{-\frac{1}{2}}\nabla\zeta_t\mathcal{K}^{-\frac{1}{2}}\nabla\xi \,\mathrm{d}x\,\mathrm{d}t &= \int \mathrm{d}t\Big[\int \mathcal{K}^{-\frac{1}{2}}\nabla\zeta_t\mathcal{K}^{-\frac{1}{2}}\nabla\xi \,\mathrm{d}x \\ &\quad - \frac{1}{2}\left(\|\mathcal{K}^{-\frac{1}{2}}\nabla\zeta_t\|_{L^2}^2 + \|\mathcal{K}^{-\frac{1}{2}}\nabla\xi_t\|_{L^2}^2\right) + \frac{1}{2}\left(\|\mathcal{K}^{-\frac{1}{2}}\nabla\zeta_t\|_{L^2}^2 + \|\mathcal{K}^{-\frac{1}{2}}\nabla\xi_t\|_{L^2}^2\right)\Big] \\ &= \frac{1}{2}\int \mathrm{d}t\left[-\|\nabla\zeta - \nabla\xi\|_{\mathcal{H}}^2 + \|\nabla\zeta_t\|_{\mathcal{H}}^2 + \|\nabla\xi_t\|_{\mathcal{H}}^2\right].\end{aligned}$$

Integrating (8.6) w.r.t. time t and rearranging the terms

$$\begin{aligned}&\frac{1}{2}\int \|\nabla\xi_t\|_{\mathcal{H}}^2 \,\mathrm{d}t \\ &= \int \zeta_1 \,\mathrm{d}\mu_1 - \zeta_0 \,\mathrm{d}\mu_0 - \int\int \partial_t\zeta_t \cdot \mu_t \,\mathrm{d}x\,\mathrm{d}t - \frac{1}{2}\int \|\nabla\zeta_t\|_{\mathcal{H}}^2 \,\mathrm{d}t + \frac{1}{2}\int \|\nabla\zeta - \nabla\xi\|_{\mathcal{H}}^2 \,\mathrm{d}t. \quad (8.7)\end{aligned}$$

We now consider the duality in the optimization problem of the BB-formula, i.e.,

$$\inf_{\xi,\mu} \frac{1}{2} \int \|\nabla \xi_t\|_{\mathcal{H}}^2 \,\mathrm{d}t = \sup_{\zeta} \inf_{\xi,\mu} (\text{RHS of (8.7)})$$

The crucial feature of (8.7) is that the term $\frac{1}{2}\int \|\nabla \zeta_t\|_{\mathcal{H}}^2 \,\mathrm{d}t$ is independent of the measure μ_t. Therefore, in order for the infimum w.r.t. μ to be finite, we require the condition $\partial_t \zeta_t \le 0$ to hold. At optimality, we recover the adjoint equation in the Hamiltonian dynamics

$$\partial_t \zeta = 0,$$

i.e., ζ is a time-independent (static) function. Hence, *the optimization problem is greatly simplified to a static setting*, which is in contrast to the Wasserstein and Stein settings. We find

$$\frac{1}{2} \cdot \inf_{\xi,\mu} \int \|\nabla \xi_t\|_{\mathcal{H}}^2 \,\mathrm{d}t = \sup_{\zeta} \left\{ \int \zeta \,\mathrm{d}(\mu_1 - \mu_0) - \frac{1}{2} \|\nabla \zeta\|_{\mathcal{H}}^2 \right\}$$

Noting the scaling in the dual formulation from Lemma 8.1, we find that the transport cost coincides with the static Kantorovich dual formulation. ∎

Proof of Lemma 6.2. This can be seen easily by calculating the KL-entropy dissipation

$$\begin{aligned} \mathcal{I}(\mu(t)) = -\frac{\mathrm{d}}{\mathrm{d}t} \mathrm{D}_{\mathrm{KL}}(\mu(t)\|\pi) &= \langle \log \frac{\mathrm{d}\mu}{\mathrm{d}\pi}, \mathcal{K}_\mu \log \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \rangle_{L^2_\mu} \\ &= \int \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \log \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \cdot \mathcal{K}_\mu \log \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \,\mathrm{d}\pi \le \|k\|_\infty \cdot \int \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \left(\log \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \right)^2 \mathrm{d}\pi. \end{aligned} \tag{8.8}$$

Similar to the pure Fisher-Rao geometry in Figure 2, this last quantity decays to zero as $\frac{\mathrm{d}\mu}{\mathrm{d}\pi} \to 0^+$ while the KL-entropy itself does not. Hence, by the analogous argument in Lemma 6.2, no global Łojasiewicz condition can hold. ∎

Proof of Corollary 6.5. According to the previous result we need to find $c_* = c_p$ which is given via

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

Obeserve 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 (1.15) is

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

For general $p \in \mathbb{R}$, explicityly calculating

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

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

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

Proof of Corollary 6.7. Łojasiewicz inequality in the WFR geometry reads

$$\|\frac{\delta F}{\delta\mu}[\mu]\|_{L^2_\mu}^2+\|\nabla\frac{\delta F}{\delta\mu}[\mu]\|_{L^2_\mu}^2\ge c\cdot(F(\mu)-F(\pi))\,. \tag{8.9}$$

By our power threhold condition, for $p\in[\frac{d-1}{d},\frac{1}{2}]$, there exists a constant $c_{\mathsf{FR}}=\frac{1}{1-p}$ such that

$$\|\frac{\delta F}{\delta\mu}[\mu]\|_{L^2_\mu}^2\ge c_{\mathsf{FR}}\cdot(F(\mu)-F(\pi))\,.$$

Combining this with the McCann condition which contributes a non-negative constant c_W, we find

$$\|\frac{\delta F}{\delta\mu}[\mu]\|_{L^2_\mu}^2+\|\nabla\frac{\delta F}{\delta\mu}[\mu]\|_{L^2_\mu}^2\ge(c_{\mathsf{FR}}+c_W)\cdot(F(\mu)-F(\pi))\,.$$

We find $c_{\mathsf{FR}}+c_W=:c_*\ge\frac{1}{1-p}$, which is the desired positive constant. ∎

Proof of Proposition 6.10 and Proposition 6.12. For the φ-divergence energy dissipation in kernelized Fisher-Rao gradient flow,

$$\begin{aligned}\mathcal{I}_\varphi^{\mathsf{FR}_k}(\mu|\pi)=-\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{D}_\varphi(\mu\|\pi)&=-\langle\varphi'\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right),\mathcal{K}_\mu\varphi'\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right)\rangle_{L^2_\mu}\\&\overset{\text{(IBP)}}{=}\int\int\frac{\mathrm{d}\mu}{\mathrm{d}\pi}(x)\cdot\varphi'\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}(x)\right)k(x,y)\frac{\mathrm{d}\mu}{\mathrm{d}\pi}(y)\cdot\varphi'\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}(y)\right)\mathrm{d}\pi(x)\,\mathrm{d}\pi(y).\end{aligned}$$

For the φ-divergence energy dissipation in Stein gradient flow,

$$\begin{aligned}\mathcal{I}_\varphi^{\mathrm{Stein}}(\mu(t)|\pi)=-\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{D}_\varphi(\mu(t)\|\pi)&=-\langle\varphi'\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right),\mathrm{div}\left(\mu\mathcal{K}_\mu\nabla\varphi'\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right)\right)\rangle_{L^2}\\&\overset{\text{(IBP)}}{=}\int\int\frac{\mathrm{d}\mu}{\mathrm{d}\pi}(x)\nabla\varphi'\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}(x)\right)k(x,y)\frac{\mathrm{d}\mu}{\mathrm{d}\pi}(y)\nabla\varphi'\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}(y)\right)\mathrm{d}\pi(x)\,\mathrm{d}\pi(y).\end{aligned}$$

Other calculation is straightforward. ∎

Proof of Proposition 6.15. For constant $0 \le s \le 1$ and a set of orthonormal bases $\{e_j\}$ of L^2_μ,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left(F(\mu(t)) - \inf_\mu F(\mu)\right) &= -\langle \frac{\delta F}{\delta \mu}[\mu], \dot{u}\rangle_{L^2_{\mu_t}} = -\langle \frac{\delta F}{\delta \mu}[\mu], (\mathcal{K}_\mu, +\lambda \,\mathrm{Id})^{-1}\mathcal{K}_\mu, \frac{\delta F}{\delta \mu}[\mu]\rangle_{L^2_\mu} \\ &= -\|\frac{\delta F}{\delta \mu}[\mu]\|^2_{L^2_\mu} + \sum_j \frac{\lambda}{\sigma_j+\lambda}\cdot\left|\langle \frac{\delta F}{\delta \mu}[\mu], e_j\rangle\right|^2 = -\|\frac{\delta F}{\delta \mu}[\mu_t]\|^2_{L^2_\mu} \\ + \lambda^s \sum_j \left(\frac{\lambda}{\sigma_j+\lambda}\right)^{1-s}\cdot \frac{\left|\langle \frac{\delta F}{\delta \mu}[\mu_t], e_j\rangle\right|^2}{(\sigma_j+\lambda)^s} &\le -\|\frac{\delta F}{\delta \mu}[\mu_t]\|^2_{L^2_\mu} + \lambda^s \|\left(\mathcal{K}_\mu + \lambda\,\mathrm{Id}\right)^{-\frac{s}{2}}\ \frac{\delta F}{\delta \mu}[\mu_t]\,\|^2_{L^2_\mu}.\end{aligned} \tag{8.10}$$

■

9 Discussion

Historically, geometries over probability measures such as optimal transport and information geometry had a tremendous impact on computational algorithms in optimization and statistical inference. It is our hope that the new geometric structure studied in this paper can motivate many downstream applications and further investigations.

For example, in this paper, we used kernel approximation due to its simplicity and well-established approximation theory. In modern machine learning, practitioners often use nonlinear models such as deep neural networks and scalable approximations of kernel machines. With our nonparametric regression formulation, it is straightforward to use those learning models.

In terms of generative models, we observe a direct correspondence of our nonparametric regression formulation in Proposition 6.13 to generative models such as the flow-matching Lipman et al. (2022) and score-matching Song et al. (2020) algorithms. The advantage of our formulation is that it does not designate a specific geometry, i.e., matching of quantities such as score, velocity, force, and growth can be expressed in the same formalism of the Rayleigh Principle.

As a final remark, we wish to comment on the asymptotic behavior of the functional inequality for the approximate systems, e.g., in Table 1. Although we have provided statements such as in Proposition 6.15 with an upper estimate for a finite regularization parameter λ, we have not yet proved that there always exists a finite λ such that the flow in the approximate geometry satisfies the functional inequality. However, motivated by the Γ-convergence in Section 6.5, our hope is that the approximate Fisher-Rao gradient flow with nice energy, e.g., power-like entropy with $p \le \frac{1}{2}$, can satisfy a Łojasiewicz inequality for some small $\lambda > 0$. As a consequence, we expect the resulting approximate flows with a small regularization parameter λ to exhibit a similar convergence behavior as the target flows.

Acknowledgement We thank Anna Korba for the helpful discussion on Stein geometry.

References

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

L. Ambrosio, N. Gigli, and G. Savare. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Springer Science & Business Media, 2005.

M. Arbel, A. Korba, A. Salim, and A. Gretton. Maximum Mean Discrepancy Gradient Flow. *arXiv:1906.04370 [cs, stat]*, Dec. 2019. URL `http://arxiv.org/abs/1906.04370`. arXiv: 1906.04370.

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. 2001.

D. Bakry and M. Émery. Diffusions hypercontractives. In J. Azéma and M. Yor, editors, *Séminaire de Probabilités XIX 1983/84*, volume 1123, pages 177–206. Springer Berlin Heidelberg, Berlin, Heidelberg, 1985. ISBN 978-3-540-15230-9 978-3-540-39397-9. doi: 10.1007/BFb0075847.

A. Ben-Tal, D. den Hertog, A. De Waegenaere, B. Melenberg, and G. Rennen. Robust Solutions of Optimization Problems Affected by Uncertain Probabilities. *Management Science*, 59(2):341–357, Feb. 2013. ISSN 0025-1909, 1526-5501. doi: 10.1287/mnsc.1120.1641. URL `http://pubsonline.informs.org/doi/abs/10.1287/mnsc.1120.1641`.

A. Blanchet and J. Bolte. A family of functional inequalities: Łojasiewicz inequalities and displacement convex functions. 275(7):1650–1673. ISSN 0022-1236. doi: 10.1016/j.jfa.2018.06.014. URL `https://www.sciencedirect.com/science/article/pii/S0022123618302465`.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877, Apr. 2017. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2017.1285773.

J. A. Carrillo, K. Craig, and F. S. Patacchini. A blob method for diffusion. *Calculus of Variations and Partial Differential Equations*, 58:1–53, 2019.

Y. Chen, D. Z. Huang, J. Huang, S. Reich, and A. M. Stuart. Gradient Flows for Sampling: Mean-Field Models, Gaussian Approximations and Affine Invariance, Nov. 2023.

S. Chewi, T. L. Gouic, C. Lu, T. Maunu, and P. Rigollet. SVGD as a kernelized Wasserstein gradient flow of the chi-squared divergence, June 2020.

L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. Unbalanced Optimal Transport: Dynamic and Kantorovich Formulation. *arXiv:1508.05216 [math]*, Feb. 2019. URL `http://arxiv.org/abs/1508.05216`. arXiv: 1508.05216.

J. B. Conway. A course in functional analysis. *Graduate Texts in Mathematics*, 1985.

K. Craig, K. Elamvazhuthi, M. Haberland, and O. Turanova. A blob method for inhomogeneous diffusion with applications to multi-agent control and sampling. *Mathematics of Computation*, 92(344):2575–2654, Nov. 2023. ISSN 0025-5718, 1088-6842. doi: 10.1090/mcom/3841.

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

F. Cucker and D. X. Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge University Press, Mar. 2007. ISBN 978-1-139-46286-0. Google-Books-ID: d8wmcLiuDtgC.

V. De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis, May 2023.

V. Divol, J. Niles-Weed, and A.-A. Pooladian. Optimal transport map estimation in general function spaces. *arXiv preprint arXiv:2212.03722*, 2022.

H. Dong, X. Wang, Y. Lin, and T. Zhang. Particle-based variational inference with preconditioned functional gradient flow. *arXiv preprint arXiv:2211.13954*, 2022.

A. Duncan, N. Nüsken, and L. Szpruch. On the geometry of stein variational gradient descent. *arXiv preprint arXiv:1912.00894*, 2019.

T. O. Gallouët and L. Monsaingeon. A jko splitting scheme for kantorovich–fisher–rao gradient flows. *SIAM Journal on Mathematical Analysis*, 49(2):1100–1130, 2017.

P. Glaser, M. Arbel, and A. Gretton. KALE Flow: A Relaxed KL Gradient Flow for Probabilities with Disjoint Support. In *Neural Information Processing Systems*, June 2021.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Y. He, K. Balasubramanian, B. K. Sriperumbudur, and J. Lu. Regularized stein variational gradient flow. *arXiv preprint arXiv:2211.07861*, 2022.

J. Ho, A. Jain, and P. Abbeel. Denoising Diffusion Probabilistic Models. *arXiv:2006.11239 [cs, stat]*, Dec. 2020. URL `http://arxiv.org/abs/2006.11239`. arXiv: 2006.11239.

M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 2013.

A. Hyvarinen. Estimation of Non-Normalized Statistical Models by Score Matching.

A. Javanmard, M. Mondelli, and A. Montanari. Analysis of a two-layer neural network via displacement convexity. *The Annals of Statistics*, 48(6), Dec. 2020. ISSN 0090-5364. doi: 10.1214/20-AOS1945.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An Introduction to Variational Methods for Graphical Models. *Machine Learning*, 37(2):183–233, Nov. 1999. ISSN 1573-0565. doi: 10.1023/A:1007665907178.

R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the Fokker–Planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998. Publisher: SIAM.

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

M. E. Khan and H. Rue. The Bayesian Learning Rule, June 2023.

A. Korba, P.-C. Aubin-Frankowski, S. Majewski, and P. Ablin. Kernel Stein Discrepancy Descent. In *Proceedings of the 38th International Conference on Machine Learning*, pages 5719–5730. PMLR, July 2021.

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):3529–3576, 2019. doi: 10.1016/j.jfa.2018.12.013.

V. Laschos and A. Mielke. Evolutionary Variational Inequalites on the Hellinger-Kantorovich and the spherical Hellinger-Kantorovich spaces. *Submitted*, 2023. arXiv:2207.09815v3.

M. Liero, A. Mielke, and G. Savaré. Optimal Entropy-Transport problems and a new Hellinger–Kantorovich distance between positive measures. *Inventiones mathematicae*, 211(3):969–1117, Mar. 2018. ISSN 0020-9910, 1432-1297. doi: 10.1007/s00222-017-0759-8. URL `http://link.springer.com/10.1007/s00222-017-0759-8`.

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):1–73, 2023. doi: 10.1007/s00205-023-01941-1.

Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

L. Liu, M. B. Majka, and L. Szpruch. Polyak-\L ojasiewicz inequality on the space of measures and convergence of mean-field birth-death processes. *Applied Mathematics & Optimization*, 87(3):48, June 2023. ISSN 0095-4616, 1432-0606. doi: 10.1007/s00245-022-09962-0. URL `http://arxiv.org/abs/2206.02774`. arXiv:2206.02774 [math].

Q. Liu and D. Wang. Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm. *arXiv:1608.04471 [cs, stat]*, Sept. 2019. URL `http://arxiv.org/abs/1608.04471`. arXiv: 1608.04471.

Q. Liu, J. D. Lee, and M. I. Jordan. A kernelized stein discrepancy for goodness-of-fit tests and model evaluation. URL `http://arxiv.org/abs/1602.03253`.

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

Y. Marzouk, Z. Ren, S. Wang, and J. Zech. Distribution learning via neural differential equations: a nonparametric statistical perspective, Sept. 2023. URL `http://arxiv.org/abs/2309.01043`. arXiv:2309.01043 [cs, math, stat].

A. Mielke. Variational approaches and methods for dissipative material models with multiple scales. In S. Conti and K. Hackl, editors, *Analysis and Computation of Microstructure in Finite Plasticity*, volume 78 of *Lect. Notes Appl. Comp. Mechanics*, chapter 5, pages 125–155. Springer, 2015. doi: 10.1007/978-3-319-18242-1_5.

A. Mielke. An introduction to the analysis of gradients systems. *arXiv preprint arXiv:2306.05026*, 2023.

Y. Mroueh and M. Rigotti. Unbalanced Sobolev Descent, Sept. 2020. URL `http://arxiv.org/abs/2009.14148`. arXiv:2009.14148 [cs, stat].

Y. Mroueh, T. Sercu, and A. Raj. Sobolev Descent. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, pages 2976–2985. PMLR, Apr. 2019.

S. Nowozin, B. Cseke, and R. Tomioka. f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL `https://proceedings.neurips.cc/paper/2016/hash/cedebb6e872f539bef8c3f919874e9d7-Abstract.html`.

K. Oko, S. Akiyama, and T. Suzuki. Diffusion models are minimax optimal distribution estimators. *arXiv preprint arXiv:2303.01861*, 2023.

F. Otto. The geometry of dissipative evolution equations: The porous medium equation.

F. Otto. *Double degenerate diffusion equations as steepest descent.* 1996.

F. Otto and C. Villani. Generalization of an inequality by talagrand and links with the logarithmic sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.

M. A. Peletier. Variational Modelling: Energies, gradient flows, and large deviations. *arXiv:1402.1990 [math-ph]*, Feb. 2014. URL `http://arxiv.org/abs/1402.1990`. arXiv: 1402.1990.

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

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

S. Serfaty. Gamma-convergence of gradient flows on Hilbert spaces and metric spaces and applications. 31(4):1427–1451, 2011.

A. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert Space Embedding for Distributions. In M. Hutter, R. A. Servedio, and E. Takimoto, editors, *Algorithmic Learning Theory*, pages 13–31, Berlin, Heidelberg, 2007. Springer. ISBN 978-3-540-75225-7. doi: 10.1007/978-3-540-75225-7_5.

J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

Y. Song and S. Ermon. Generative Modeling by Estimating Gradients of the Data Distribution. *arXiv:1907.05600 [cs, stat]*, Oct. 2020. URL `http://arxiv.org/abs/1907.05600`. arXiv: 1907.05600.

Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

V. Spokoiny. Nonparametric estimation: parametric view. 2016.

I. Steinwart and A. Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, NY, 2009. ISBN 978-0-387-79051-0 978-0-387-79052-7. doi: 10.1007/b13794. URL `https://link.springer.com/10.1007/b13794`.

P. Vincent. A Connection Between Score Matching and Denoising Autoencoders. *Neural Computation*, 23(7):1661–1674, July 2011. ISSN 0899-7667, 1530-888X. doi: 10.1162/NECO_a_00142.

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

W. Wang, D. Slepčev, S. Basu, J. A. Ozolek, and G. K. Rohde. A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *International journal of computer vision*, 101:254–269, 2013.

H. Wendland. *Scattered Data Approximation*. Cambridge University Press, Dec. 2004. ISBN 978-1-139-45665-4.

J.-J. Zhu, W. Jitkrittum, M. Diehl, and B. Schölkopf. Kernel Distributionally Robust Optimization: Generalized Duality Theorem and Stochastic Approximation. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pages 280–288. PMLR, Mar. 2021. URL `https://proceedings.mlr.press/v130/zhu21a.html`. ISSN: 2640-3498.