Ricci curvature for parametric statistics via optimal transport

Wuchen Li[1] · **Guido Montúfar[1]**

Abstract

We define the notion of a Ricci curvature lower bound for parametrized statistical models. Following the seminal ideas of Lott–Sturm–Villani, we define this notion based on the geodesic convexity of the Kullback–Leibler divergence in a Wasserstein statistical manifold, that is, a manifold of probability distributions endowed with a Wasserstein metric tensor structure. Within these definitions, which are based on Fisher information matrix and Wasserstein Christoffel symbols, the Ricci curvature is related to both, information geometry and Wasserstein geometry. These definitions allow us to formulate bounds on the convergence rate of Wasserstein gradient flows and information functional inequalities in parameter space. We discuss examples of Ricci curvature lower bounds and convergence rates in exponential family models.

Keywords Ricci curvature · Information projection · Wasserstein statistical manifold · Fokker–Planck equation on parameter space

1 Introduction

The Ricci curvature lower bound on *sample space* plays a crucial role in various fields, including heat semi-groups [4] and differential geometry (Brunn–Minkowski inequality) [38]. In particular, it provides sharp bounds for convergence rates of diffusion processes [4] and functional inequalities [34]. In recent years, optimal transport contributed a viewpoint that connects Ricci curvature and information functionals. In this study, optimal transport, in particular the L^2-Wasserstein metric, introduces a Riemannian structure in probability density space, named density manifold [22,33]. The Ricci curvature lower bound in sample space is equivalent to the geodesic convexity

✉ Wuchen Li
  wcli@math.ucla.edu

[1] University of California Los Angeles, Los Angeles, CA, USA

🖄 Springer

of the Kullback–Leibler divergence in the density manifold[1]. Following this angle, Lott–Sturm–Villani [28,37] defined the Ricci curvature on non-smooth metric sample spaces, and Erbar–Maas [16] introduced it on discrete sample spaces. It is worth mentioning that the discrete optimal transport is introduced by Chow et.al. [8], Maas [29], and Mielke [30]. These studies and associated Ricci curvature lower bounds on discrete sample spaces have many applications; see [9,10,13–15,17,18,36].

In statistics and machine learning, we often are interested in constructing, or selecting, a density that models the behavior of some observed data, according to some quality criterion. Very often we restrict the search to a subset of densities, as this allows us to handle large state spaces and also to incorporate prior knowledge into our search. Parametrized statistical models are a ubiquitous and powerful approach. In this paper, we develop the theory of Ricci curvature lower bounds for this situation. The Ricci curvature lower bound governs the dissipation rates of the relative entropy. In the context of learning, this corresponds to the rates of convergence of gradient descent methods for minimizing the Kullback–Leibler (KL) divergence and computing information projections.

The Wasserstein metric tensor of a statistical manifold (a parametrized set of probability densities) has been defined in [24]. A statistical manifold endowed with a Wasserstein metric tensor structure is called Wasserstein statistical manifold. We define the Ricci curvature lower bound via geodesic convexity of the KL divergence on a Wasserstein statistical manifold. We obtain a definition of the Ricci curvature that connects Wasserstein geometry [38] and information geometry [2,3], much in the spirit of [23,24], and take a natural further step towards connecting the two fields, in particular, relating notions from learning applications and the geometry of the statistical models. We focus on discrete sample spaces, which allows us to present a clear picture of the relations deriving from this theory, and leave the details of continuous settings for future work.

We consider a discrete statistical model described by a tuple (Θ, I, p) consisting of a parameter space Θ, a discrete sample space (or state space) $I = \{1, \ldots, n\}$, and a parametrization $p \colon \Theta \to \mathcal{P}(I)$. Here $\mathcal{P}(I)$ denotes the set of all probability distributions on I. We say that (Θ, I, p) has Ricci curvature lower bound $\kappa \in \mathbb{R}$ with respect to a given reference measure q, if and only if, for any $\theta \in \Theta$, it holds that

$$G_F(\theta) + \sum_{a \in I} \left(d_{\theta\theta} p_a(\theta) \log \frac{p_a(\theta)}{q_a} - \Gamma^{W,a}(\theta) \frac{d}{d\theta_a} \mathrm{D}_{\mathrm{KL}}(p(\theta)\|q) \right) \succeq \kappa G_W(\theta).$$

Here G_F is the Fisher-Rao metric tensor, G_W is the L^2-Wasserstein metric tensor, $d_{\theta\theta} p$ is the second differential of the parameterization, $\Gamma^W(\theta)$ are the Christoffel symbols of the Wasserstein statistical manifold, and $\mathrm{D}_{\mathrm{KL}}(p(\theta)\|q) = \sum_{i=1}^n p_i(\theta) \log \frac{p_i(\theta)}{q_i}$ is the KL divergence. This definition depends on the reference measure q. We emphasize that the geometry of statistical manifold comes into the picture of Ricci curvature through both Fisher information matrix and Wasserstein Christoffel symbols. In statistics and learning applications, the reference measure will play the role of a target or empirical

---

[1] Geodesic convexity is a synthetic definition. If a function f on manifold (M, g) is second differentiable, then f is λ-geodesic convex whenever $\mathrm{Hess}_M f \succeq \lambda g$.
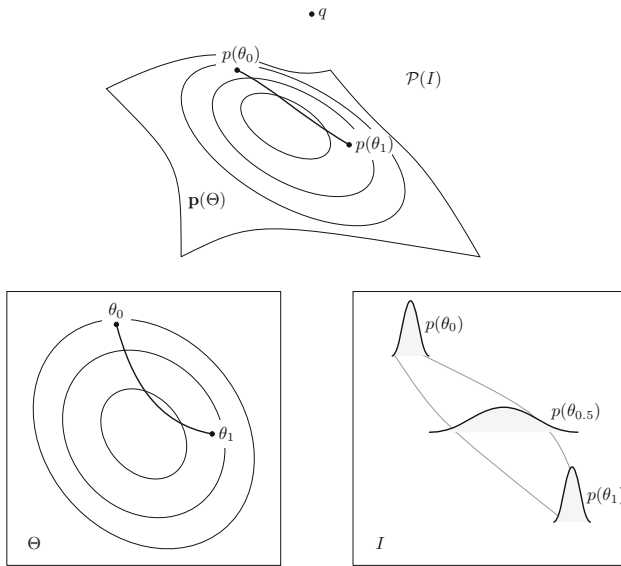
Fig. 1 Our discussion involves a state space I, a parameter space Θ, and a parametrized set $p(\Theta)$ in the space $\mathcal{P}(I)$ of probability distributions on I. For a reference measure $q \in \mathcal{P}(I)$, a positive Ricci curvature lower bound implies that the Wasserstein geodesic connecting two distributions, $p(\theta_0)$ and $p(\theta_1)$, 'bends' towards q. The figure depicts the geodesic as a thick curve, together with the level sets of $D_{KL}(p(\cdot)\|q)$, in Θ and $p(\Theta)$. In terms of the state space I, when q is uniform, a decrease of the KL divergence with respect to q corresponds to an increase of the entropy, meaning that along the geodesic, the 'volume' of states under the distributions 'bulges'. This corresponds to the synthetic notion of positive curvature in sample space. Note how the geodesics are constrained to lie within the model $p(\Theta)$, which in general does not contain q. See Definition 7, Theorem 8, Proposition 9, and Figs. 2 and 3 for more details

data distribution. A schematic illustration of the spaces and relations that we consider is provided in Fig. 1.

The Ricci curvature on discrete state spaces has been studied by many groups. (i) Ollivier [31] introduces a discrete Ricci curvature via L_1-Wasserstein metric. Many inequalities on graphs are shown under this setting; see, e.g., [19,21,32]. (ii) Lin–Yau et al. [25,26] also define a Ricci curvature lower bound by heat semi-groups and Bakry–Emery Γ_2 operators. (iii) Erbar–Maas introduce the Ricci curvature lower bound in [16] by means of equivalence relations with Lott–Sturm–Villani in the Wasserstein probability manifold, under which several information functional inequalities are established. This notion has been studied extensively in [13–15,17,18]. However, the notion of a Ricci curvature lower bound on the parameter space of a statistical manifold has not been studied so far. Parametrized Wasserstein probability sub-manifolds were not introduced until recently in [7,24]. There are several related infinite dimensional Wasserstein sub-manifolds considered in [6,35]. Our definition of the Ricci curvature lower bound for parametrized statistical models is close in spirit to the definitions by Lott–Sturm–Villani and Erbar–Maas.

This paper is organized as follows. In Sect. 2, we briefly review the connections between Ricci curvature, optimal transport, and KL divergence. We further demon-

strate these connections in the context of information projections. In Sect. 3, we introduce Wasserstein statistical manifolds. This is intended as a short review of the definitions from [24]. We derive the Fokker–Planck equation on parameter space, which is the Wasserstein gradient flow of the KL divergence. The main technical contributions of this paper are contained in Sect. 4. We describe the convergence rate of the Fokker–Planck equation in terms of a Ricci curvature lower bound. Further, we use the notion of Ricci curvature lower bound to establish information functional inequalities. We also discuss methods to estimate the Ricci curvature lower bound. In Sect. 5, we present numerical evaluations on small exponential families. These allow us to illustrate the notions introduced in the paper, and gain more intuition about their meaning.

2 Ricci curvature and information projections

In this section, we review the connection of optimal transport and information theory put forward in Villani's book [38], and we further connect with the notion of information projections described by Csiszár–Shields [11]. In later sections we will develop these connections for the case of parametric statistical models.

2.1 Wasserstein geometry

Consider a continuous measure space (Ω, g^{Ω}, q). Here Ω is a finite dimensional compact smooth Riemannian manifold without boundary, g^{Ω} is its metric tensor, dx is the volume from of Ω, and $q \in C^{\infty}(\Omega)$ is the measure volume form with $\int_{\Omega} q(x) = 1$, $q(x) > 0$. The Ricci curvature tensor on (Ω, g^{Ω}, q) refers to

$$\text{Ric} = \text{Ric}_{\Omega} - \text{Hess}_{\Omega} \log q, \tag{1}$$

where Ric_{Ω} denotes the Ricci curvature on Ω and Hess_{Ω} is the Hessian operator on Ω. Note that this notion of curvature depends on the reference measure q. Later in our discussion, the reference measure will play the role of a target or empirical data distribution.

On the one hand, optimal transport, in particular the L^2-Wasserstein metric, introduces an infinite-dimensional Riemannian structure in density space. Consider the set of smooth and strictly positive densities

$$\mathcal{P}_{+}(\Omega) = \left\{ \rho \in C^{\infty}(\Omega) \colon \rho(x) > 0, \ \int_{\Omega} \rho(x) dx = 1 \right\}.$$

The tangent space of $\mathcal{P}_{+}(\Omega)$ at $\rho \in \mathcal{P}_{+}(\Omega)$ is given by

$$T_{\rho}\mathcal{P}_{+}(\Omega) = \left\{ \sigma \in C^{\infty}(\Omega) \colon \int_{\Omega} \sigma(x) dx = 0 \right\}.$$

Definition 1 *(L^2-Wasserstein metric tensor)* Define the inner product $g_\rho \colon T_\rho \mathcal{P}_+(\Omega) \times T_\rho \mathcal{P}_+(\Omega) \to \mathbb{R}$ by

$$g_\rho(\sigma_1, \sigma_2) = \int_\Omega \sigma_1(x)(-\Delta_\rho)^\dagger \sigma_2(x) dx,$$

where $\Delta_\rho^\dagger \colon T_\rho \mathcal{P}_+(\Omega) \to T_\rho \mathcal{P}_+(\Omega)$ is the inverse of the elliptical operator $\Delta_\rho = \nabla \cdot (\rho \nabla)$. Here ∇ and $\nabla \cdot$ are the gradient and divergence operators in Ω, respectively.

Following [22], we call $(\mathcal{P}_+(\Omega), g)$ a Wasserstein density manifold or a Wasserstein manifold for short. The metric tensor introduces a variational formulation of a metric function. More precisely, the square of the L^2-Wasserstein metric function is equal to the geometric energy (action) of geodesics in the Wasserstein manifold. For any $\rho_0, \rho_1 \in \mathcal{P}_+(\Omega)$, the L^2-Wasserstein metric function is defined as

$$\begin{aligned}
W(\rho_0, \rho_1)^2 &= \inf \ \left\{ \int_0^1 g_{\rho_t}(\partial_t \rho_t, \partial_t \rho_t) dt : \rho_t \in \mathcal{P}_+(\Omega), t \in [0, 1] \right\} \\
&= \inf \ \left\{ \int_0^1 \|\nabla \Phi_t\|^2 \rho_t dx dt : \Phi_t \in C^\infty(\Omega), t \in [0, 1] \right\},
\end{aligned}$$

where the second formula is often named Benamou-Breiner formula [5]. It is the geometric action functional in term of cotangent bundle; see details in [23].

One can extend the definitions from $\mathcal{P}_+(\Omega)$ to the set $\mathcal{P}_2(\Omega)$ of Borel probability measures with finite second moments. It is well known that the L^2-Wasserstein metric defines a metric function on $\mathcal{P}_2(\Omega)$, and hence $(\mathcal{P}_2(\Omega), W)$ forms a length space. See related analytical treatments in [38].

2.2 Wasserstein gradient flow of the KL divergence

On the other hand, information theory considers a particular functional on density space, namely the KL divergence. Given a smooth reference measure $q \in \mathcal{P}_+(\Omega)$, the KL divergence of a given ρ with respect to q is defined by

$$D_{KL}(\rho \| q) = \int_\Omega \rho(x) \log \frac{\rho(x)}{q(x)} dx.$$

Notice that the KL divergence is precisely the free energy. Indeed, if we write $q(x) = \frac{1}{K} e^{-V(x)}$ with $K = \int_\Omega e^{-V(x)} dx$, we see that

$$\begin{aligned}
D_{KL}(\rho \| q) &= \int_\Omega \rho(x) \log \rho(x) dx + \int_\Omega V(x) \rho(x) dx + \log K \\
&= -H(\rho) + \mathbb{E}_\rho[V(X)] + \log K,
\end{aligned}$$

where $H(\rho) = -\int_\Omega \rho(x) \log \rho(x) \ dx$ is the Boltzmann–Shannon entropy, X is a random variable satisfying the law of density ρ and $\mathbb{E}$ is the expectation operator.

The Ricci curvature on sample space is related both to the KL divergence and the L^2-Wasserstein metric tensor. This interaction starts with the gradient flow of the KL divergence in the Wasserstein manifold $(\mathcal{P}_+(\Omega), g)$ [20]. It describes the time evolution of the density following the negative Wasserstein gradient of the KL divergence:

$$
\begin{aligned}
\frac{\partial \rho_t}{\partial t} &= - \operatorname{grad}_W D_{KL}(\rho \| q) \\
&= \nabla \cdot \left(\rho_t \nabla \left(\log \frac{\rho_t}{q} + 1 \right) \right) \\
&= \nabla \cdot (\rho_t \nabla V) + \Delta \rho_t.
\end{aligned}
\tag{2}
$$

The second line is by Definition 1 of the Wasserstein metric tensor. The last equality holds since $q(x) = \frac{1}{K} e^{-V(x)}$ and $\nabla \cdot (\rho \nabla \log \rho) = \nabla \cdot (\nabla \rho) = \Delta \rho$.

It is worth noting that there are several perspectives based on (2). Firstly, the flow (2) is a well-known dynamics called the *Fokker–Planck equation* (FPE). It describes the probability transition equation of drift diffusion process

$$
\dot{X}_t = -\nabla V(X_t) + \sqrt{2} \dot{B}_t,
$$

where B_t is the canonical Brownian motion in sample space. Secondly, along the flow (2), the KL divergence converges to zero. I.e. ρ_t converges to the minimizer of the KL divergence, known as the Gibbs measure, $q(x) = \frac{1}{K} e^{-V(x)}$. This reminds of iterative methods for computing information projections [11] in statistics and machine learning. In this context, one seeks to reproduce the behavior of a teacher system in terms of a model. To this end, the learning rule proceeds by adjusting the model parameters so as to maximize the likelihood function given the observations, which is equivalent to minimizing the divergence, for instance using Wasserstein gradient descent. The flow is the continuous limit of the gradient descent learning rule. We shall go to this connection shortly, in Sect. 2.5.

2.3 Dissipation rates and the Ricci curvature lower bound

As it turns out, the Ricci curvature lower bound governs the exponential dissipation rate of (2) towards the Gibbs measure q. In the setting of learning, this corresponds precisely to the exponential rate of convergence of the learning dynamics. To see this, the following calculations in dynamical system are used. One can find the convergence rate of (2) by comparing the ratio between the first and second time derivatives along the flow. By some computations, the first time derivative of the KL divergence along the flow is found to be equal to

$$
\begin{aligned}
-\frac{d}{dt} D_{KL}(\rho_t \| q) &= g_{\rho_t}(\partial_t \rho_t, \partial_t \rho_t) \\
&= \int_\Omega \Gamma \left(\log \frac{\rho_t}{q}, \log \frac{\rho_t}{q} \right) \rho_t \, dx,
\end{aligned}
$$

while the second time derivative is given by

$$\frac{d^2}{dt^2} D_{KL}(\rho_t \| q) = \mathrm{Hess}_W D_{KL}(\rho_t \| q) (\partial_t \rho_t, \partial_t \rho_t)$$

$$= \int_\Omega \Gamma_2 \left(\log \frac{\rho_t}{q}, \log \frac{\rho_t}{q} \right) \rho_t \, dx \ .$$

Here Hess_W is the Hessian operator with respect to the Wasserstein metric tensor, and Γ and Γ_2 are the Bakry–Emery operators defined by

$$\Gamma(f, f) = \int_\Omega g^\Omega(\nabla f, \nabla f) \, dx,$$

and

$$\Gamma_2(f, f) = (\mathrm{Ric}_\Omega - \mathrm{Hess}_\Omega \log q)(\nabla f, \nabla f) + \mathrm{tr}(\mathrm{Hess}_\Omega f, \mathrm{Hess}_\Omega f),$$

where Ric_Ω is the Ricci curvature tensor on Ω, Hess_Ω is the Hessian operator on Ω, and tr is the trace operator. By the above formulas, the ratio between $\frac{d}{dt} D_{KL}(\rho_t \| q)$ and $\frac{d^2}{dt^2} D_{KL}(\rho_t \| q)$ relates to the integral version of Γ, Γ_2, i.e. the expectation values of the operators Γ, Γ_2. Notice that $\mathrm{tr}(\mathrm{Hess}_\Omega f, \mathrm{Hess}_\Omega f) \geq 0$. Classical results [38] show that the lower bound of Ricci curvature governs the smallest ratio between $\frac{d}{dt} D_{KL}(\rho_t \| q)$ and $\frac{d^2}{dt^2} D_{KL}(\rho_t \| q)$, which further gives the exponential convergence rate of (2). In addition, the above computation demonstrate that the lower bound of the Ricci curvature, informally speaking, is equivalent to the smallest eigenvalue of the Hessian operator of the KL divergence in the Wasserstein manifold.

Theorem 2 *Given $\kappa \in \mathbb{R}$ and $q(x) \in \mathcal{P}_+(\Omega)$, the following statements are equivalent.*

(i) *κ is a Ricci curvature lower bound of (Ω, g^Ω, q). I.e. κ is the largest number for which, uniformly over Ω,*

$$Ric = Ric_\Omega - Hess_\Omega \log q \succeq \kappa g^\Omega;$$

(ii) *$\Gamma_2(f, f) \geq \kappa \Gamma(f, f)$, for any $f \in C^\infty(\Omega)$;*

(iii) *For any constant speed geodesic ρ_t, $t \in [0, 1]$, connecting ρ_0 and ρ_1 in $(\mathcal{P}_2(\Omega), W)$,*

$$D_{KL}(\rho_t \| q) \leq (1 - t) D_{KL}(\rho_0 \| q) + t D_{KL}(\rho_1 \| q) - \frac{\kappa}{2} t(1 - t) W(\rho_0, \rho_1)^2.$$

Theorem 2 opens the door to define a notion of Ricci curvature lower bound on sample space via its equivalent statements. In the literature, Bakry–Emery [4] define the Ricci curvature lower bound by applying (ii) for smooth Riemannian sample spaces, while Lott–Sturm–Villani [28,37] define it using (iii) for non-smooth metric sample spaces, and Erbar–Maas [16] define it by (iii) in a discrete sample space. In this paper, we shall define the notion of Ricci curvature lower bound for parametric statistics

taking an approach based on (iii), known as the geodesic convexity property of the KL divergence.

2.4 Learning in a parametrized model

In statistics and machine learning applications one often considers a parametrized set $\{p(\cdot; \theta) : \theta \in \mathbb{R}^d\}$ of candidate probability distributions from which one wishes to choose one to model the distribution of some given data.

One motivation for using parametrized models is reducing the dimensionality associated with large state spaces. For instance, we may be considering a state space consisting of images presented as arrays of pixel intensities, corresponding to $\{0, 1\}^n$ with n easily in the order of thousands. In this case, storing a probability distribution as a vector $p \in \mathbb{R}^{2^n}$ of individual probabilities $p(x)$, $x \in \{0, 1\}^n$, is an impossibility. With a parametric model, instead of storing the probability vector, we store only a parameter vector $\theta \in \mathbb{R}^d$, with a more manageable d, and fix a mapping that allows us to recover individual values $p(x; \theta)$ of the probability distribution for a given x, or, in other cases, which allows us to generate samples from $p(\cdot; \theta)$ that we can also use to estimate expectation values of interest. Reducing the dimensionality is useful not only in terms of storage, but also in a statistical sense, in relation to overfitting. Without going into details, the richer the class of hypotheses, with more free parameters, the more prone we are to fitting statistical nuisances, instead of capturing the true general behavior of the data. By working with a parametrized model, we can incorporate priors into the learning system and limit its vulnerability to overfitting.

When working with a parametrized model, obtaining the best possible hypothesis, e.g. the maximizer of the likelihood, is usually a non trivial problem and one has to resort to iterative methods. A relevant question then is the computational effort needed for this. In particular, one is interested in the number of iterations needed until reaching a solution that is within ϵ of the best possible. The Ricci curvature can be regarded as a way to obtain bounds on the convergence rate of gradient optimization of the KL divergence for a given a target distribution, uniformly over the start distribution. We elaborate on this in the next paragraph. The situation is illustrated in Figs. 1, 2, and 3.

2.5 I-projections

In the context of information theory and statistics, Csiszár-Shields [11] define the *I-projection* of a distribution Q onto a non-empty closed convex set $\mathcal{N}$ of distributions as the point $P^* \in \mathcal{N}$ such that

$$D(P^* \| Q) = \min_{P \in \mathcal{N}} D(P \| Q).$$

The notion of I-projection considers the minimization of the KL divergence with respect to the first argument, but it is also relevant in the context of maximum likelihood estimation, where the minimization is with respect to the second argument. Given an empirical data distribution P, a maximum likelihood estimate over a set $\mathcal{E}$ is a point $P^* \in \overline{\mathcal{E}}$ (the closure of $\mathcal{E}$), with

Fig. 2 For a distribution Q in an exponential family $\mathcal{E}$ and a distribution P in an orthogonal linear family $\mathcal{N}$, the Pythagorean relation holds: $D(P\|Q) = D(P\|P^*) + D(P^*\|Q)$, where P^* is the unique intersection point of $\mathcal{E}$ and $\mathcal{N}$
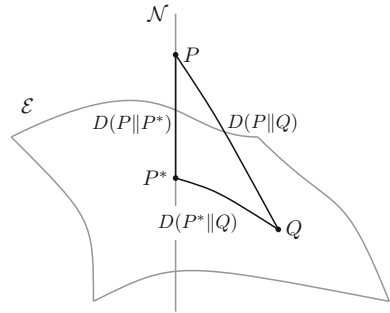
$$D(P\|P^*) = \inf_{Q \in \mathcal{E}} D(P\|Q).$$

If we consider an exponential family model $\mathcal{E} = \{p \propto Q_r \exp(\theta^{\mathsf{T}} F) \colon \theta \in \mathbb{R}^d\}$ on a finite state space I, with sufficient statistics $F \colon I \to \mathbb{R}^d$ and reference measure $Q_r \in \mathcal{P}_+(I)$, then the maximum likelihood estimator P^* for the target distribution P can be obtained as the I-projection of $Q \in \mathcal{E}$ onto the orthogonal linear family defined by $\mathcal{N} = \{p \colon \sum_x F(x)p(x) = \sum_x F(x)P(x)\}$. We have namely that

$$P^* = \operatorname{argmin}_{Q' \in \mathcal{E}} D(P\|Q') = \operatorname{argmin}_{P \in \mathcal{N}} D(P\|Q).$$

This is a consequence of the well known Pythagorean relation [2,11] illustrated in Fig. 2. Iterative numerical methods to compute I-projections, and more generally, to optimize the likelihood, include gradient optimization, Newton's method, the Fisher scoring algorithm [27], the Fisher natural gradient [1], and generalized iterative scaling [12].

Csiszár and Shields [11] consider iterative methods for computing I-projections, and obtain upper bounds on the divergence along the resulting parameter trajectories, which describe the convergence to the optimum value. For two sets of distributions, $\mathcal{P}$ and $\mathcal{Q}$, together with two functions $D(\cdot, \cdot) \colon \mathcal{P} \times \mathcal{Q} \to \mathbb{R}$ and $\delta(\cdot, \cdot) \colon \mathcal{P} \times \mathcal{P} \to \mathbb{R}$, satisfying certain conditions, they describe an iterative algorithm (alternating divergence minimization) which iterates $p_n \in \mathcal{P}$ and $q_n \in \mathcal{Q}$, and give an upper bound of the form

$$D(p_{n+1}, q_n) - D_{\min} \le \delta(p_\infty, p_n) - \delta(p_\infty, p_{n+1}). \tag{3}$$

In this paper we are in the special setting where $\mathcal{Q} = \{q\}$ and $\mathcal{P}$ is the set of all densities. There is a natural connection between (3) and the Fokker–Planck-equation (2). Indeed, setting D as the KL divergence and $q_n = q = p_\infty$, $p_n = \rho_t$, $p_{n+1} = \rho_{t+\Delta t}$, where Δt is the step size, we demonstrate in Proposition 9 that we can substitute

$$\delta(q, p) = \frac{1}{2\kappa \Delta t} D_{\mathrm{KL}}(p\|q),$$

where κ is the Ricci curvature lower bound that we will define later on. Strictly speaking, for this correspondence, we need to assume that $\kappa > 0$, which is a natural

Fig. 3 Illustration of the Ricci curvature lower bound κ in connection to the geodesic convexity of the KL divergence and the rate of convergence of the information projection flow. Here $p(\theta_t)$ is a Wasserstein geodesic connecting $p(\theta_0)$ and $p(\theta_1)$. When $q = p(\theta_1)$, the KL divergence $D_{KL}(p(\theta_t)\|q)$ is monotonically decreasing
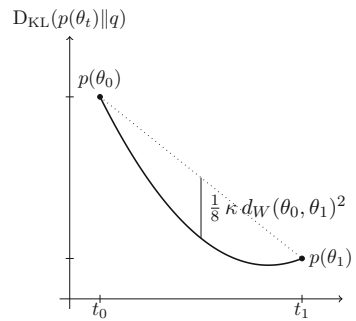
requirement similar to requiring that the KL divergence is geodesic convex in set $\mathcal{P}$. Each step dissipation in (3) then gives

$$
\begin{aligned}
D_{KL}(\rho_{t+\Delta t}\|q) - D_{\min} &\leq \delta(p_\infty, p_n) - \delta(p_\infty, p_{n+1}) \\
&= \frac{1}{2\kappa\,\Delta t}\{D_{KL}(\rho_t\|q) - D_{KL}(\rho_{t+\Delta t}\|q)\} \\
&= -\frac{1}{2\kappa}\frac{d}{dt}D_{KL}(\rho_t\|q) + o(\Delta t).
\end{aligned}
$$

In other words, the Fokker–Planck equation is a monotone information projection flow, in which the dissipation quantity is governed by the difference of relative entropy divided by twice the Ricci curvature lower bound. In the limit where Δt goes to zero,

$$
D_{KL}(\rho_t\|q) - D_{\min} \leq -\frac{1}{2\kappa}\frac{d}{dt}D_{KL}(\rho_t\|q).
$$

Grönwall's inequality then implies that this I-projection flow (2) converges to the minimizer at the rate of $e^{-2\kappa t}$, i.e.

$$
D_{KL}(\rho_t\|q) - D_{\min} \leq e^{-2\kappa t}\Big(D_{KL}(\rho_0\|q) - D_{\min}\Big).
$$

The above shows that the learning rate for the Fokker–Planck equation is linear, whose lower bound is governed by κ. Following these connections, we will pursue the definition of the Ricci curvature lower bound on parameter space. The convergence rate, in relation to the Ricci curvature lower bound and the geodesic convexity of the KL divergence, is illustrated schematically in Fig. 3. More details will be provided in Proposition 9.

3 Wasserstein statistical manifolds

In preparation for the definitions and results on the Ricci curvature that we will present in the next section, we briefly review the definition of a Wasserstein statistical manifold with discrete sample space from [24], and present the Fokker–Planck equation on parameter space.

3.1 Wasserstein geometry on the probability simplex

We recall the definition of discrete probability simplex with L^2-Wasserstein Riemannian metric. Consider the discrete sample space $I = \{1, \ldots, n\}$. The probability simplex on I is the set

$$\mathcal{P}(I) = \left\{ (p_1, \ldots, p_n) \in \mathbb{R}^n : \sum_{i=i}^{n} p_i = 1, \quad p_i \geq 0 \right\}.$$

Here $p = (p_1, \ldots, p_n)$ is a probability vector with coordinates p_i corresponding to the probabilities assigned to each node $i \in I$. The probability simplex $\mathcal{P}(I)$ is a manifold with boundary. We denote the interior by $\mathcal{P}_+(I)$. This consists of the strictly positive probability distributions, with $p_i > 0$ for all $i \in I$. To simplify the discussion, we will focus on the interior $\mathcal{P}_+(I)$. For the studies related to the boundary $\partial\mathcal{P}(I)$, we refer the reader to [23].

Next we define the L^2-Wasserstein metric tensor on $\mathcal{P}_+(I)$, which also encodes the metric tensor of discrete states I. We need to give a ground metric notion on sample space. We do this in terms of a simply connected undirected graph with weighted edges, $G = (I, E, \omega)$, where I is the vertex set, $E \subseteq \binom{I}{2}$ is the edge set, and $\omega = (\omega_{ij})_{i,j \in I} \in \mathbb{R}^{n \times n}$ is a matrix of edge weights satisfying

$$\omega_{ij} = \begin{cases} \omega_{ji} > 0, & \text{if } (i, j) \in E \\ 0, & \text{otherwise} \end{cases}.$$

The set of neighbors (adjacent vertices) of i is denoted by $N(i) = \{j \in V : (i, j) \in E\}$. The normalized volume form on node $i \in I$ is given by $d_i = \frac{\sum_{j \in N(i)} \omega_{ij}}{\sum_{i=1}^{n} \sum_{i' \in N(i)} \omega_{ii'}}$.

The graph structure $G = (I, E, \omega)$ induces a graph Laplacian matrix function.

Definition 3 (*Weighted Laplacian matrix*) Given an undirected weighted graph $G = (I, E, \omega)$, with $I = \{1, \ldots, n\}$, the matrix function $L(\cdot) : \mathbb{R}^n \to \mathbb{R}^{n \times n}$ is defined by

$$L(p) = D^{\mathsf{T}} \Lambda(p) D, \quad p = (p_i)_{i=1}^{n} \in \mathbb{R}^n,$$

where

- $D \in \mathbb{R}^{|E| \times n}$ is the discrete gradient operator defined by

$$D_{(i,j) \in E, k \in V} = \begin{cases} \sqrt{\omega_{ij}}, & \text{if } i = k, i > j \\ -\sqrt{\omega_{ij}}, & \text{if } j = k, i > j \\ 0, & \text{otherwise} \end{cases},$$

- $-D^{\mathsf{T}} \in \mathbb{R}^{n \times |E|}$ is the oriented incidence matrix, and

- $\Lambda(p) \in \mathbb{R}^{|E| \times |E|}$ is a weight matrix depending on p,

$$\Lambda(p)_{(i,j) \in E, (k,l) \in E} = \begin{cases} \frac{1}{2}(\frac{1}{d_i} p_i + \frac{1}{d_j} p_j) & \text{if } (i,j) = (k,l) \in E \\ 0 & \text{otherwise} \end{cases}.$$

The Laplacian matrix function $L(p)$ is the discrete analog of the weighted Laplacian operator $-\nabla \cdot (\rho \nabla)$ from Definition 1. Here it is used as the inverse of metric tensor for the discrete probability simplex.

Remark 1 There are several other choices of $\Lambda(p)$ proposed in [29]. It is worth mentioning that the canonical choice of $\Lambda(p)$ in discrete settings is unknown yet. We leave the study for other choices of $\Lambda(p)$ to interested readers.

We are now ready to present the L^2-Wasserstein metric tensor. Consider the tangent space of $\mathcal{P}_+(I)$ at p,

$$T_p \mathcal{P}_+(I) = \left\{ (\sigma_i)_{i=1}^n \in \mathbb{R}^n : \sum_{i=1}^n \sigma_i = 0 \right\}.$$

Denote the space of *potential functions* on I by $\mathcal{F}(I) = \mathbb{R}^n$, and consider the quotient space

$$\mathcal{F}(I)/\mathbb{R} = \{[\Phi] \mid (\Phi_i)_{i=1}^n \in \mathbb{R}^n\},$$

where $[\Phi] = \{(\Phi_1 + c, \dots, \Phi_n + c) : c \in \mathbb{R}\}$ are functions defined up to addition of constants.

We introduce an identification map via the weighted Laplacian matrix $L(p)$:

$$\mathbf{V} : \mathcal{F}(I)/\mathbb{R} \to T_p \mathcal{P}_+(I), \qquad \mathbf{V}_\Phi = L(p)\Phi.$$

We know that $L(p)$ has only one simple zero eigenvalue with eigenvector $c(1, 1, \dots, 1)$, for any $c \in \mathbb{R}$. This is true since for $(\Phi_i)_{i=1}^n \in \mathbb{R}^n$,

$$\Phi^\mathsf{T} L(p)\Phi = (D\Phi)^\mathsf{T} \Lambda(p)(D\Phi) = \sum_{(i,j) \in E} \omega_{ij} (\Phi_i - \Phi_j)^2 \left(\frac{1}{2} \left(\frac{1}{d_i} p_i + \frac{1}{d_j} p_j \right) \right) = 0,$$

implies $\Phi_i = \Phi_j$, $(i, j) \in E$. It the graph is connected, as we assume, then $(\Phi_i)_{i=1}^n$ is a constant vector. Thus $\mathbf{V} : \mathcal{F}(I)/\mathbb{R} \to T_p \mathcal{P}_+(I)$ is a well defined map, linear, and one to one. I.e., $\mathcal{F}(I)/\mathbb{R} \cong T_p^* \mathcal{P}_+(I)$, where $T_p^* \mathcal{P}_+(I)$ is the cotangent space of $\mathcal{P}_+(I)$. This identification induces the following inner product on $T_p \mathcal{P}_+(I)$.

Definition 4 (*L^2-Wasserstein metric tensor*) The inner product $g_p : T_p \mathcal{P}_+(I) \times T_p \mathcal{P}_+(I) \to \mathbb{R}$ takes any two tangent vectors $\sigma_1 = \mathbf{V}_{\Phi_1}$ and $\sigma_2 = \mathbf{V}_{\Phi_2} \in T_p \mathcal{P}_+(I)$ to

$$g_p(\sigma_1, \sigma_2) = \sigma_1^\mathsf{T} \Phi_2 = \sigma_2^\mathsf{T} \Phi_1 = \Phi_1^\mathsf{T} L(p)\Phi_2. \tag{4}$$

In other words,

$$g_p(\sigma_1, \sigma_2) := \sigma_1^\mathsf{T} L(p)^\dagger \sigma_2, \quad \text{for any} \quad \sigma_1, \sigma_2 \in T_p \mathcal{P}_+(I),$$

where $L(p)^\dagger$ is the pseudo inverse of $L(p)$.

Following the inner product (4), the Wasserstein metric (distance function) $W \colon \mathcal{P}_+(I) \times \mathcal{P}_+(I) \to \mathbb{R}$ is defined by

$$W(p^0, p^1)^2 := \inf_{p(t), \Phi(t)} \left\{ \int_0^1 \Phi(t)^\mathsf{T} L(p(t)) \Phi(t) dt \right\}. \tag{5}$$

Here the infimum is taken over pairs $(p(t), \Phi(t))$ with $p \in H^1((0, 1), \mathbb{R}^n)$ and $\Phi \colon [0, 1] \to \mathbb{R}^n$ measurable, satisfying

$$\frac{d}{dt} p(t) - L(p(t)) \Phi(t) = 0, \quad p(0) = p^0, \quad p(1) = p^1.$$

3.2 Wasserstein statistical manifold

We next consider a statistical model defined by a triplet (Θ, I, p). Here, $I = \{1, \ldots, n\}$ is the sample space, Θ is the parameter space, which is an open subset of $\mathbb{R}^d, d \leq n-1$, and $p \colon \Theta \to \mathcal{P}_+(I)$ is the parametrization function,

$$p(\theta) = (p_i(\theta))_{i=1}^n, \quad \theta \in \Theta.$$

We define a Riemannian metric g^W on Θ as the pull-back of metric g on $\mathcal{P}_+(I)$. In other words, we require that $p \colon (\Theta, g^W) \to (\mathcal{P}_+(I), g)$ is an isometric embedding:

$$g_\theta^W(a, b) := g_{p(\theta)}^W (d_\theta p(\theta)(a), d_\theta p(\theta)(b))$$
$$= (d_\theta p(\theta)(a))^\mathsf{T} L(p(\theta))^\dagger (d_\theta p(\theta)(b)), \quad \text{for all } a, b \in T_\theta(\Theta).$$

Since $dp(\theta)(a) = \left(\sum_{j=1}^d \frac{\partial p_i(\theta)}{\partial \theta_j} a_j \right)_{i=1}^n = J_\theta p(\theta) a$, we arrive at the following definition.

Definition 5 (L^2-*Wasserstein metric tensor on parameter space*) For any pair of tangent vectors $a, b \in T_\theta \Theta = \mathbb{R}^d$, define

$$G_W(\theta) := J_\theta p(\theta)^\mathsf{T} L(p(\theta))^\dagger J_\theta p(\theta), \tag{6}$$

and

$$g_\theta^W(a, b) := a^\mathsf{T} G_W(\theta) b,$$

where $J_\theta(p(\theta)) = (\frac{\partial p_i(\theta)}{\partial \theta_j})_{1 \leq i \leq n, 1 \leq j \leq d} \in \mathbb{R}^{n \times d}$ is the Jacobi matrix of the parametrization p.

This inner product is consistent with the restriction of the Wasserstein metric g^W to $p(\theta)$. We will assume that $\text{rank}(J_\theta\, p(\theta)) = d$, so that the parametrization p_i is locally injective and the metric tensor g^W is positive definite. We call (Θ, I, p), together with the induced Riemannian metric g^W, *Wasserstein statistical manifold* (WSM).

In this case, the constrained Wasserstein distance function $d_W : \Theta \times \Theta \to \mathbb{R}_+$ is given by the geometric action energy

$$
d_W(\theta_0, \theta_1)^2
$$
$$
= \inf_{\theta(t)\in C^1([0,1],\Theta)} \left\{ \int_0^1 \dot{\theta}(t)^\mathsf{T} G_W(\theta(t))\dot{\theta}(t)dt : \theta(0) = \theta_0,\ \theta(1) = \theta_1 \right\}. \quad (7)
$$

The above formula can be viewed as Benamou-Breiner formula within probability models. When working on the full probability simplex, with $\theta = p$, the metric function d_W corresponds precisely to the metric function W given in (5).

3.3 Fokker–Planck equation on parameter space

We next derive the Fokker–Planck equation on parameter space by Wasserstein gradient flow of KL divergence.

Given a reference measure $q \in \mathcal{P}_+(I)$, consider the Kullback–Leibler divergence (relative entropy) on parameter space

$$
\mathrm{D}_{\mathrm{KL}}(p(\theta)\|q) = \sum_{i=1}^{n} p_i(\theta) \log \frac{p_i(\theta)}{q_i}.
$$

Proposition 6 (Fokker–Planck equation on parameter space) *The gradient flow for the negative Boltzmann–Shannon entropy in (Θ, g) is*

$$
\frac{d\theta}{dt} = -\left(J_\theta\, p(\theta)^\mathsf{T} L(p(\theta))^\dagger J_\theta\, p(\theta)\right)^\dagger J_\theta\, p(\theta)^\mathsf{T} \log \frac{p(\theta)}{q}. \quad (8)
$$

Proof Notice that the Riemannian gradient operator of function $F \in C^\infty(\theta)$ in (Θ, g_W) is defined as

$$
\mathrm{grad}_W F(\theta) = G_W(\theta)^\dagger \nabla_\theta F(\theta),
$$

and the gradient flow satisfies

$$
\frac{d\theta}{dt} = -\mathrm{grad}_W F(\theta).
$$

Thus, the gradient flow of entropy on (Θ, g_W) satisfies

$$
\begin{aligned}
\frac{d\theta}{dt} &= -\operatorname{grad}_W D_{KL}(p(\theta)\|q) \\
&= -G_W(\theta)^\dagger \nabla_\theta D_{KL}(p(\theta)\|q) \\
&= -\left(J_\theta p(\theta)^\mathsf{T} L(p(\theta))^\dagger J_\theta p(\theta)\right)^\dagger J_\theta p(\theta)^\mathsf{T} \left(\log \frac{p(\theta)}{q} + \vec{1}\right),
\end{aligned}
$$

where ∇_θ represents the Euclidean gradient operator, $\log \frac{p(\theta)}{q} = (\log \frac{p_i(\theta)}{q_i})_{i=1}^n$ and $\vec{1} = (1, \ldots, 1) \in \mathbb{R}^n$. Since $p(\theta)^\mathsf{T}\vec{1} = 1$, we have $J_\theta p(\theta)^\mathsf{T}\vec{1} = \vec{0}$. This completes the proof. $\qquad\square$

Remark 2 Consider the full probability set with continuous sample space Ω. Denote the probability $p_i(t)$ by a density $\rho(t, x) \in \mathcal{P}(\Omega)$, then the equation (8) recovers exactly the FPE (2).

We next study the convergence properties of the Fokker–Planck equation on parameter space. In other words, how fast does the solution of (8) converge to its equilibrium? As in the full probability space, we define the concept of Ricci curvature lower bound on parameter space to give the bound of the convergence rate for (8).

4 Ricci curvature lower bound on parameter space

This section contains the main contributions of this paper. We define the Ricci curvature lower bound on parameter space and prove equivalent conditions for this definition, which connect information geometry and Wasserstein geometry. In addition, we present several information functional inequalities on parameter space. Finally, we give a simple guide for computing these quantities in practice.

4.1 Ricci curvature lower bound on parameter space

Definition 7 We say (Θ, I, p) has the Ricci curvature lower bound $\kappa \in \mathbb{R}$ if for any constant speed geodesic θ_t, $t \in [0, 1]$, connecting θ_0, θ_1 in (Θ, g^W), it holds that

$$
D_{KL}(p(\theta_t)\|q) \leq (1 - t)D_{KL}(p(\theta_0)\|q) + tD_{KL}(p(\theta_1)\|q) - \frac{\kappa}{2}t(1 - t)d_W(\theta_0, \theta_1)^2.
$$

In this case we also write

$$
\operatorname{Ric}(\Theta, I, p) \geq \kappa.
$$

If (Θ, g_W) forms a compact smooth Riemannian manifold and $p(\theta)$ is smooth. Then κ is the smallest eigenvalue of the Hessian of the KL divergence over the Wasserstein statistical manifold, i.e.

$$\mathrm{Hess}_W \mathrm{D_{KL}}(p(\theta)\|q) \succeq \kappa G_W(\theta),$$

for any $\theta \in \Theta$.

Definition 7 is based on the definition of geodesic convexity in geometry. It is a more general definition than the one in terms of the Hessian operator begin bounded below by κ. The reason is as follows. On the one hand, the probability set is a manifold with boundary. Suitable regularity studies are needed to take care of the boundary when using the Hessian operator [23]. On the other hand, not all parameterizations $p(\theta)$ are twice differentiable.

Definition 7 shares the same spirit of Lott–Sturm–Villani and Erbar–Maas. If $p(\Theta) = \mathcal{P}(I)$ is the whole probability simplex, then $\mathrm{Ric}(\Theta, I, p)$ is the Ricci curvature bound on discrete sample space. Our definition extends this idea to a statistical manifold. In other words, $\mathrm{Ric}(\Theta, I, p)$ inherits properties from both probability submanifold and Ricci curvature bound on sample space. Note that $\mathrm{Ric}(\Theta, I, p)$ is different from the Ricci curvature on (Θ, g^W), in which the former represents how the changes ratio KL divergence takes effect on the parameterized sample space, while the later reflects the curvature on the set of probability model itself.

We next given an equivalent condition for Definition 7. It naturally connects Ricci curvature (R), Information geometry (I) and Wasserstein geometry (W). We call it Ricci–Information–Wasserstein (RIW) condition.

Theorem 8 (RIW condition) *Assume Θ is a compact set. $\mathrm{Ric}(\Theta, I, p) \geq \kappa$ holds if and only if for any $\theta \in \Theta$,*

$$G_F(\theta) + \sum_{a \in I}\left(d_{\theta\theta}\, p_a(\theta) \log \frac{p_a(\theta)}{q} - \Gamma^{W,a}(\theta)\frac{d}{d\theta_a}\mathrm{D_{KL}}(p(\theta)\|q)\right) \succeq \kappa G_W(\theta), \quad (9)$$

where $G_F(\theta) = (G_F(\theta)_{ab})_{1\leq a,b\leq d}$ is the Fisher-Rao metric tensor

$$G_F(\theta)_{ab} = \sum_{i \in I}\frac{d \log p_i(\theta)}{d\theta_a}\frac{d \log p_i(\theta)}{d\theta_b}p_i(\theta), \quad (10)$$

$\Gamma^{W,k} = (\Gamma^{W,k}_{ij})_{1\leq i,j\leq n}$ *is the Wasserstein Christoffel symbol*

$$\Gamma^{W,k}_{ij} = \frac{1}{2}\sum_{l=1}^{n}(G_{W,kl})^{-1}\left(\nabla_{\theta_i}G_{W,jl} + \nabla_{\theta_j}G_{W,il} - \nabla_{\theta_l}G_{W,ij}\right),$$

and $G_W(\theta)$ is the Wasserstein metric tensor defined in (6).

Proof Let θ_t be a constant speed geodesic, i.e. $\ddot\theta_t + \Gamma^W(\dot\theta_t, \dot\theta_t) = 0$ with $\theta_0 = \theta \in \Theta$ and $\dot\theta_0 = a \in T_\theta\Theta$. Consider the Taylor expansion

$$\mathrm{D_{KL}}(p(\theta_t)\|q) = \mathrm{D_{KL}}(p(\theta)\|q) + \frac{d}{dt}\bigg|_{t=0}\mathrm{D_{KL}}(p(\theta_t)\|q)t + \frac{1}{2}\frac{d^2}{dt^2}\bigg|_{t=0}$$

$$\mathrm{D_{KL}}(p(\theta_t)\|q)t^2 + o(t^2).$$

Then the Hessian operator on Riemannian manifold (Θ, g^W) forms

$$
\begin{aligned}
&\mathrm{Hess}_W \mathrm{D_{KL}}(p(\theta_t)\|q)(\dot\theta_t, \dot\theta_t)\\
&= \frac{d^2}{dt^2}\mathrm{D_{KL}}(p(\theta_t)\|q)\\
&= \frac{d}{dt}(d_\theta \mathrm{D_{KL}}(p(\theta_t)\|q)^\mathsf{T}\dot\theta_t)\\
&= \dot\theta_t^\mathsf{T} d_{\theta\theta}\mathrm{D_{KL}}(p(\theta_t)\|q)\dot\theta_t - d_\theta \mathrm{D_{KL}}(p(\theta_t)\|q)^\mathsf{T}\Gamma^W(\dot\theta_t, \dot\theta_t)\\
&= \dot\theta_t^\mathsf{T} d_{\theta\theta}\mathrm{D_{KL}}(p(\theta_t)\|q)\,\dot\theta_t - \dot\theta_t^\mathsf{T}\left(\sum_{k\in I}\frac{d}{d\theta_k}\mathrm{D_{KL}}(p(\theta_t)\|q)\Gamma^{W,k}\right)\dot\theta_t.
\end{aligned}
$$

In addition,

$$
\begin{aligned}
d_\theta \mathrm{D_{KL}}(p(\theta_t)\|q) &= \sum_{i=1}^{n}\Big(d_\theta p_i(\theta_t)\log p_i(\theta_t) + p_i(\theta_t)d_\theta \log p_i(\theta_t) - d_\theta p_i(\theta_t)\log q_i\Big)\\
&= \sum_{i=1}^{n}\Big(d_\theta p_i(\theta_t)\log p_i(\theta_t) - d_\theta p_i(\theta_t)\log q_i\Big),
\end{aligned}
$$

where $\sum_{i=1}^{n} p_i(\theta_t)d_\theta \log p_i(\theta_t) = \sum_{i=1}^{n} p_i(\theta_t)\frac{1}{p_i(\theta_t)}J_\theta p_i(\theta_t) = 0$, since $\sum_{i=1}^{n} p_i(\theta) = 1$. Thus

$$
\begin{aligned}
d_{\theta\theta}\mathrm{D_{KL}}(p(\theta_t)\|q) &= \sum_{i=1}^{n} d_{\theta\theta} p_i(\theta_t)\log\frac{p_i(\theta_t)}{q} + \sum_{i=1}^{n}\frac{1}{p_i(\theta_t)}d_\theta p_i(\theta_t)d_\theta p_i(\theta_t)^\mathsf{T}\\
&= \sum_{i=1}^{n} d_{\theta\theta} p_i(\theta_t)\log\frac{p_i(\theta_t)}{q} + G_F(\theta),
\end{aligned}
$$

where G_F denotes the Fisher-Rao metric tensor: $G_F(\theta_t) = \sum_{i=1}^{n}\frac{1}{p_i(\theta_t)}d_\theta p_i$ $(\theta_t)d_\theta p_i(\theta_t)^\mathsf{T} = \sum_{i=1}^{n} d_\theta \log p_i(\theta_t)d_\theta \log p_i(\theta_t)^\mathsf{T} p_i(\theta)$, with the fact $\frac{1}{p_i(\theta_t)}d_\theta p_i(\theta_t) = d_\theta \log p_i(\theta_t)$.

Thus $\mathrm{Hess}_W\mathrm{D_{KL}}(p(\theta)\|q) \succeq \kappa G_W(\theta)$ is equivalent to (9). This concludes the proof. $\qquad\square$

Remark 3 If we replace I by the continuous sample space (Ω, g^Ω) and consider the full probability simplex, then RIW condition (9) is equivalent to the integral version of Bakry–Emery condition. See details in [23, Proposition 19].

4.2 Entropy dissipation on parameter space

With the Ricci curvature lower bound in hand, we can prove the following convergence properties of Fokker–Planck equations on parameter space.

Proposition 9 (Bakry–Emery condition on parameter space) *Assume Θ is a compact set. If $Ric(\Theta, I, p) \geq \kappa > 0$, then there exists a unique equilibrium $\theta^* \in \Theta$, with*

$$\theta^* = \arg \min_{\theta \in \Theta} D_{KL}(p(\theta) \| q).$$

In addition, for any initial condition $\theta_0 \in \Theta$, the solution $\theta(t)$ of (8) converges to θ^* exponentially fast, with

$$D_{KL}(p(\theta_t) \| q) - D_{KL}(p(\theta^*) \| q) \leq e^{-2\kappa t} \Big(D_{KL}(p(\theta_0) \| q) - D_{KL}(p(\theta^*) \| q) \Big), \quad \text{for all } t. \tag{11}$$

Remark 4 This result will apply for any geometry defined on Θ, whenever $\kappa > 0$ is the smallest eigenvalue of the corresponding Hessian operator of the divergence function.

Proof The proof comes from the classical study of gradient flow in Riemannian manifold (Θ, g_W). Since $\text{Hess}_W D_{KL}(p(\theta) \| q) \geq \kappa > 0$, the $D_{KL}(p(\theta) \| q)$ is κ-geodesics convex in (Θ, g_W). Thus $\theta(t)$ converges to the unique equilibrium θ^*, which is also the unique minimizer of KL divergence.

We next investigate how fast $\theta(t)$ converges to θ^*. The speed of convergence is obtained by comparing the first and second derivatives of the KL divergence w.r.t. time t along (8). We have

$$\frac{d}{dt} D_{KL}(p(\theta_t) \| q) = -g_W \big(\text{grad}_W D_{KL}(p(\theta_t) \| q), \text{grad}_W D_{KL}(p(\theta_t) \| q) \big),$$

and

$$\frac{d^2}{dt^2} D_{KL}(p(\theta_t) \| q) = 2\text{Hess}_W D_{KL}(p(\theta_t) \| q) \big(\text{grad}_W D_{KL}$$
$$(p(\theta_t) \| q), \text{grad}_W D_{KL}(p(\theta_t) \| q) \big).$$

From $Ric(\Theta, I, p) \geq \kappa > 0$, then $\text{Hess}_W D_{KL}(p(\theta) \| q) \geq \kappa > 0$, i.e.

$$\frac{d^2}{dt^2} D_{KL}(p(\theta_t) \| q) \geq -2\kappa \frac{d}{dt} D_{KL}(p(\theta_t) \| q), \quad \text{for all } t \geq 0. \tag{12}$$

Then by integrating the above formula over $[t, +\infty)$, one obtains

$$\frac{d}{dt}[D_{KL}(p(\theta^*) \| q) - D_{KL}(p(\theta_t) \| q)] \geq -2\kappa[D_{KL}(p(\theta^*) \| q) - D_{KL}(p(\theta_t) \| q)].$$

Proceed with the Grönwall's inequality, the result is proved. □

4.3 Functional inequalities on parameter space

In literature [34], the convergence rate of FPE is used to prove several functional inequalities, including Log-Sobolev, Talagrand and HWI inequalities. The HWI inequality is a relation between the relative entropy (H), Wasserstein metric (W), relative Fisher information functional ($\mathcal{I}$). We shall derive the counterparts of these inequalities on parameter space.

Here the Log-Sobolev inequality describes a relationship between relative entropy and relative Fisher information functional on parameter space. Here the relative Fisher information functional is defined by

$$\mathcal{I}(p(\theta)\|q) := g_W(\text{grad}_W D_{KL}(p(\theta_t)\|q), \text{grad}_W D_{KL}(p(\theta_t)\|q)). \tag{13}$$

Here we formulate (13) as follows:

$$\mathcal{I}(p(\theta)\|q) = \log \frac{p(\theta)}{q}^\mathsf{T} J_\theta(p(\theta))\big(J_\theta p(\theta)^\mathsf{T} L(p(\theta))^\dagger J_\theta p(\theta)\big)^\dagger J_\theta p(\theta)^\mathsf{T} \log \frac{p(\theta)}{q}.$$

We compare (13) with the one in continuous sample space and full probability space:

$$\mathcal{I}(\rho\|q) = \int_\Omega g^\Omega \left(\nabla \log \frac{\rho}{q}, \nabla \log \frac{\rho}{q}\right) \rho \, dx.$$

We note that the functional (13) is different from the commonly known Fisher information matrix (10) in parameter space. It contains the ground metric structure in the sample space, which is inherited from the L^2-Wasserstein metric tensor $L(p)^\dagger$. In other words, when applying the Fisher information in full probability set into parameter space, the following two angles arrive. Here (13) keeps the differential structure of sample space and project the differential of KL divergence into the parameter space, while Fisher information matrix (10) replaces the differential structures of sample space to the ones in parameters.

In the following, we derive inequalities based on (13).

Proposition 10 (Functional inequalities on parameter space) *Consider a statistical manifold* (Θ, I, p). *The following inequalities hold.*

(i) *If* $Ric(\Theta, I, p) \geq \kappa > 0$, *then the Logarithmic Sobolev inequality on parameter space*

$$D_{KL}(p(\theta)\|q) - D_{KL}(p(\theta^*)\|q) \leq \frac{1}{2\kappa} \mathcal{I}(p(\theta)\|q), \tag{14}$$

holds for any $\theta \in \Theta$.

(ii) *If* $Ric(\Theta, I, p) \geq \kappa > 0$, *then the Talagrand inequality on parameter space*

$$\frac{\kappa}{2} d_W(\theta, \theta^*)^2 \leq D_{KL}(p(\theta)\|q) - D_{KL}(p(\theta^*)\|q),$$

holds for any $\theta \in \Theta$.

(iii) If $Ric(\Theta, I, p) \geq \kappa \in \mathbb{R}$ (κ not necessarily positive), then the HWI inequality on parameter space

$$D_{KL}(p(\theta)\|q) - D_{KL}(p(\theta^*)\|q) \leq \sqrt{\mathcal{I}(p(\theta)\|q)}d_W(\theta, \theta^*) - \frac{\kappa}{2}d_W(\theta, \theta^*)^2,$$

holds for any $\theta \in \Theta$.

Proof Here we mainly follow the heuristic arguments in [34]. In finite dimensional parameter space, these approaches are rigorous. We demonstrate the proofs for the completeness of paper.

(i) The proof follows Proposition 9. Consider the Fokker–Planck equation (2) with initial condition $\theta(0) = \theta$. The dissipation along gradient flow of entropy gives

$$\begin{aligned} \mathcal{I}(p(\theta_t)\|q) &= -\frac{d}{dt}D_{KL}(p(\theta_t)\|q) \\ &= g_W\big(\text{grad}_W D_{KL}(p(\theta_t)\|q), \text{grad}_W D_{KL}(p(\theta_t)\|q)\big). \end{aligned} \tag{15}$$

Since (12) holds, by integrating over time $t \in [0, \infty)$, we have

$$-\frac{d}{dt}D_{KL}(p(\theta_t)\|q)|_{t=0}^{\infty} \geq -2\kappa[D_{KL}(p(\theta^*)\|q) - D_{KL}(p(\theta_0)\|q)].$$

From (15) and $\theta(0) = \theta$, we have

$$\mathcal{I}(p(\theta)\|q) - \mathcal{I}(p(\theta^*)\|q) \leq 2\kappa[D_{KL}(p(\theta)\|q) - D_{KL}(p(\theta^*)\|q)],$$

where we use the fact $\text{grad}_W D_{KL}(p(\theta)\|q) = 0$, so that $\mathcal{I}(p(\theta^*)\|q) = 0$. It proves the result.

(ii) Consider $\theta(t)$ satisfy the FPE (2) on parameter space with $\theta(0) = \theta$. Since $Ric(\Theta, I, p) \geq \kappa > 0$, then $\lim_{t\to\infty} \theta(t) = \theta^*$. Define

$$\Psi(t) = d_W(\theta, \theta(t)) + \sqrt{\frac{2}{\kappa}}\sqrt{D_{KL}(p(\theta_t)\|q) - D_{KL}(p(\theta^*)\|q)}.$$

Thus $\Psi(0) = \sqrt{\frac{2}{\kappa}}\sqrt{D_{KL}(p(\theta)\|q) - D_{KL}(p(\theta^*)\|q)}$ and $\Psi(\infty) = \lim_{t\to\infty}\Psi(t) = d_W(\theta, \theta^*)$. We claim that $\Psi(t)$ is nondecreasing. If so, then $\Psi(0) \leq \Psi(\infty)$, which proves the result.

To show $\Psi(t)$ is nondecreasing, we shall prove that

$$\frac{d^+}{dt}\Psi(t) = \limsup_{h\to 0+} \frac{\Psi(t+h) - \Psi(t)}{h} \leq 0.$$

Here we assume $\theta(t) \neq \theta^*$, otherwise $\Psi(t+h) = \Psi(t)$ for any h, which shows the upper derivative zero.

On the one hand, by triangle inequality,

$$|d_W(\theta, \theta_t) - d_W(\theta, \theta_{t+h})| \leq d_W(\theta_t, \theta_{t+h}),$$

so that

$$\lim_{h \to 0+} \sup \frac{d_W(\theta_t, \theta_{t+h})}{h} = \sqrt{g_W(\mathrm{grad}_W D_{\mathrm{KL}}(p(\theta_t)\|q), \mathrm{grad}_W D_{\mathrm{KL}}(p(\theta_t)\|q))}$$

$$= \sqrt{\mathcal{I}(p(\theta_t)\|q)}. \tag{16}$$

On the other hand, since $\theta(t) \neq \theta^*$, then

$$\sqrt{\frac{2}{\kappa}} \frac{d}{dt} \sqrt{D_{\mathrm{KL}}(p(\theta_t)\|q) - D_{\mathrm{KL}}(p(\theta^*)\|q)}$$

$$= -\frac{g_W(\mathrm{grad}_W D_{\mathrm{KL}}(p(\theta_t)\|q), \mathrm{grad}_W D_{\mathrm{KL}}(p(\theta_t)\|q))}{\sqrt{2\kappa(D_{\mathrm{KL}}(p(\theta_t)\|q) - D_{\mathrm{KL}}(p(\theta^*)\|q))}}$$

$$= -\frac{\mathcal{I}(p(\theta_t)\|q)}{\sqrt{2\kappa(D_{\mathrm{KL}}(p(\theta_t)\|q) - D_{\mathrm{KL}}(p(\theta^*)\|q))}}.$$

From (14), we have

$$\sqrt{\frac{2}{\kappa}} \frac{d}{dt} \sqrt{D_{\mathrm{KL}}(p(\theta_t)\|q) - D_{\mathrm{KL}}(p(\theta^*)\|q)} \leq -\sqrt{\mathcal{I}(p(\theta_t)\|q)}. \tag{17}$$

From (16) and (17), we have $\frac{d^+}{dt}\Psi(t) = \lim\sup_{h\to 0+} \frac{\Psi(t+h)-\Psi(t)}{h} \leq 0$, which finishes the proof.

(iii) From the definition of $\mathrm{Ric}(\Theta, I, p) \geq \kappa$, then $\mathrm{Hess}_W D_{\mathrm{KL}}(p(\theta)\|q) \succeq \kappa G_W$. Denote θ_t be a geodesic curve of least energy in manifold (Θ, g_W), joining $\theta_0 = \theta$ and $\theta_1 = \theta^*$. Thus

$$d_W(\theta, \theta^*) = \sqrt{g_W\left(\frac{d\theta_t}{dt}, \frac{d\theta_t}{dt}\right)}.$$

From the Taylor expansion on the (Θ, g_W), we have

$$D_{\mathrm{KL}}(p(\theta^*)\|q) = D_{\mathrm{KL}}(p(\theta)\|q) + \frac{d}{dt}|_{t=0} D_{\mathrm{KL}}(p(\theta_t)\|q)$$

$$+ \int_0^1 (1-t) \frac{d^2}{dt^2} D_{\mathrm{KL}}(p(\theta_t)\|q) dt.$$

We note that

$$\frac{d}{dt}|_{t=0} D_{\mathrm{KL}}(p(\theta_t)\|q)$$

$$= g_W\left(\mathrm{grad}_W D_{\mathrm{KL}}(p(\theta_t)\|q), \frac{d\theta_t}{dt}\right)|_{t=0}$$

$$\geq -\sqrt{g_W\left(\mathrm{grad}_W D_{KL}(p(\theta_t)\|q), \mathrm{grad}_W D_{KL}(p(\theta_t)\|q)\right)|_{t=0}}\sqrt{g_W\left(\frac{d\theta_t}{dt}, \frac{d\theta_t}{dt}\right)|_{t=0}}$$

$$= -\sqrt{\mathcal{I}(p(\theta)\|q)}d_W(\theta, \theta^*),$$

and

$$\int_0^1 (1-t)\frac{d^2}{dt^2}D_{KL}(p(\theta_t)\|q)dt$$

$$= \int_0^1 (1-t)g_W\left(\mathrm{Hess}_W D_{KL}(p(\theta_t)\|q)\cdot\frac{d\theta_t}{dt}, \frac{d\theta_t}{dt}\right)dt$$

$$\geq \int_0^1 \kappa(1-t)g_W\left(\frac{d\theta_t}{dt}, \frac{d\theta_t}{dt}\right)dt$$

$$= \frac{\kappa}{2}d_W(\theta, \theta^*)^2.$$

Combining the above formulas, we prove the result. □

4.4 Computing the Ricci curvature lower bound and convergence rate

In this section, we design an algorithm for Ricci curvature lower bound κ.

We first approximate κ by RIW condition in Theorem 8. In other words, we compute formulas for (9) via

$$\kappa = \text{Smallest eigenvalue of } G_W(\theta)^{-1}$$

$$\times \left\{ G_F(\theta) + \sum_{a\in I}\left(d_{\theta\theta}p_a(\theta)\log\frac{p_a(\theta)}{q} - \Gamma^{W,a}(\theta)\frac{d}{d\theta_a}D_{KL}(p(\theta)\|q)\right)\right\}.$$

where $d_{\theta\theta}p_a(\theta)$, $\Gamma^{W,a}(\theta)$, $\frac{d}{d\theta_a}D_{KL}(p(\theta)\|q)$ are computed by numerical differentiation.

In practice, we also compute a uniform convergence rate $K \geq \kappa$ as the smallest ratio of $\frac{d}{dt}D_{KL}(p(\theta_t)\|q)$ and $\frac{d^2}{dt^2}D_{KL}(p(\theta_t)\|q)$ along the gradient flow (8) for any initial conditions. I.e.

$$K = \min_{\theta_0\in\Theta}\frac{1}{2T}\frac{D_{KL}(p(\theta_{2T})\|q) - 2D_{KL}(p(\theta_T)\|q) + D_{KL}(p(\theta_0)\|q)}{D_{KL}(p(\theta_T)\|q) - D_{KL}(p(\theta_0)\|q)},$$

where T is a given short time, θ_T is the solution of (11) with initial condition θ_0.

Convergence rate

Input: Sample initial conditions $\{\theta_0^s\}_{s=1}^{|S|}$;
  Target distribution q;
  A suitable initial step size $h > 0$;
  A short terminal time $T > 0$.
Output: Approximation K of the uniform convergence rate;

---

for $s \in \{1, \ldots, |S|\}$
  for $k = 1, 2, \ldots, 2T/h$
    $\theta_{k+1}^s = \theta_k^s - h G_W(\theta_k^s)^{-1} \nabla_\theta D_{KL}(p(\theta_k^s) \| q)$;
  end
end
$K = \min_{s \in \{1, \ldots, |S|\}} \frac{1}{2T} \frac{D_{KL}(p(\theta_{2T}^s)\|q) - 2 D_{KL}(p(\theta_T^s)\|q) + D_{KL}(p(\theta_0^s)\|q)}{D_{KL}(p(\theta_T^s)\|q) - D_{KL}(p(\theta_0^s)\|q)}.$

---

5 Examples

In this section, we illustrate some of the concepts introduced in the previous sections by means of evaluating them on a simple class of exponential family models. We illustrate the effects from the choice of the ground metric on sample space in relation to the choice of the statistical model, and the relationships between the Ricci curvature lower bound and the rates of convergence in learning.

Example 1 *(Ricci curvature for a one-dimensional exponential family on three states)*
We study how the Ricci curvature changes with the choice of a probability model and with the choice of the ground metric on sample space. In order to obtain a picture as complete as possible, we consider the small setting of three states and one dimensional exponential families.

Consider the sample space $I = \{1, 2, 3\}$ with a fully connected graph with edges $E = \{(1, 2), (2, 3), (1, 3)\}$, and weights $\omega = (\omega_{12}, \omega_{23}, \omega_{13})$. The probability simplex is a triangle

$$\mathcal{P}(I) = \left\{ (p_i)_{i=1}^3 \in \mathbb{R}^3 \; : \; \sum_{i=1}^3 p_i = 1, \quad p_i \geq 0 \right\}.$$

We consider statistical manifolds of the form

$$p(\theta) = \frac{1}{Z(\theta)} (e^{\theta c_1}, e^{\theta c_2}, e^{\theta c_3}),$$

with sufficient statistic $c = (c_1, c_2, c_3) \in \mathbb{R}^3$, parameter $\theta \in \Theta = [\theta_{\min}, \theta_{\max}] \subset \mathbb{R}^1$, and partition function $Z(\theta) = \sum_{i=1}^3 e^{\theta c_i}$. These are exponential families specified by the choice of the sufficient statistic c. Here, addition of constants is immaterial. Multiplicative scaling by non-zero numbers does not change the model. For better comparability, we always choose c to have norm one.

In particular, these models can be indexed by the projective line, which for simplicity we can represent by a half circle, or an angle.

We fix a uniform reference measure $q = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. The KL divergence then takes the form

$$D_{KL}(p\|q) = \sum_{i=1}^{3} p_i \log \frac{p_i}{q_i} = \sum_{i=1}^{3} p_i \log p_i + \log 3.$$

We evaluate the Ricci curvature lower bound for 30 different exponential families and 10 different choices of the ground metric. We choose the sufficient statistics as evenly spaced points on a radius 1 half circle, and set the parameter domain as $\Theta = [-2, 2]$.

The results are shown in Fig. 4. The left panel estimates K as the minimum rate of convergence of the Wasserstein gradient flow of the KL divergence, over a grid of 10 different initial conditions on the parameter domain. As can be seen, the convergence is faster, the better ω connects the end points of the exponential family.

The right panel estimates κ as the minimum eigenvalue of the Hessian operator of the KL divergence over a grid of parameter values in the domain. Figure 5 gives a direct comparison of the estimates obtained from convergence rates and the Hessian. As can be seen, the Hessian is always a lower bound of the convergence rate, which reflects Proposition 9.

If the parameter domain is smaller, the Hessian gives a closer bound to the rate of convergence. If, on the contrary, the parameter domain is larger, the gaps between the Hessian and the convergence rates tend to be larger. Larger parameters correspond to distributions closer to the boundary of the simplex. We illustrate these effects in the Appendix, where we provide figures with different choices of Θ (Figs. 6, 7), and also comparing the Hessian and rates of convergence at individual parameter values (Fig. 8).

6 Discussion

To summarize, we introduced a notion of Ricci curvature lower bound for parametric statistical models and illustrated its possible relevance in the context of parameter estimation and learning. This notion is based on the geodesic convexity of the KL divergence in Wasserstein geometry. Following the program from [24], we hope that this paper continues to strengthen the interactions between information geometry and Wasserstein geometry.

The Ricci curvature lower bound depends on the target distribution, the statistical model, and the ground metric on sample space. We think that this notion can serve to capture the general properties of learning in different models, and hence that it can serve to guide the design of statistical models (e.g., the graph of a graphi-
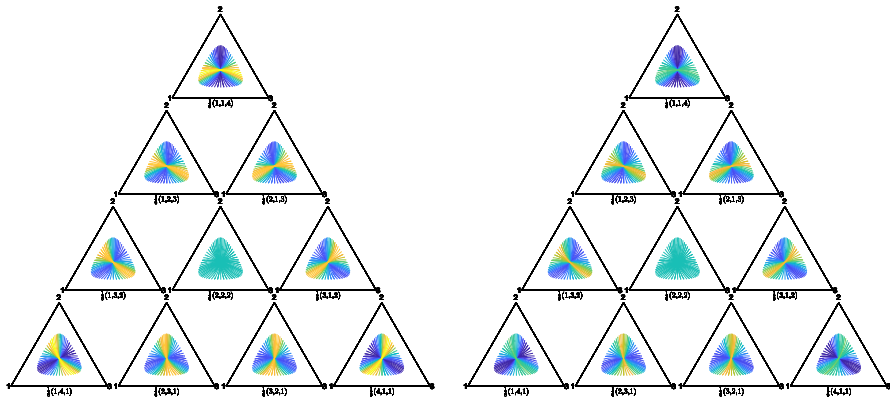
Fig. 4 Lower bound on the Ricci curvature for one-dimensional exponential families on three states. Each simplex corresponds to a different choice of $\omega = (\omega_{12}, \omega_{23}, \omega_{13})$, indicated at the bottom. Within each simplex there are 30 different exponential families (which are curves) with sufficient statistics of norm one and parameter domain $\Theta = [-1, 1]$. The color of each exponential family corresponds to the value of K estimated as the minimum convergence rate (left panel), and the value of κ as the minimum eigenvalue of the Hessian (right panel), over the parameter domain. Blue corresponds to lower and yellow to higher values. We give a direct comparison of K and κ in Fig. 5
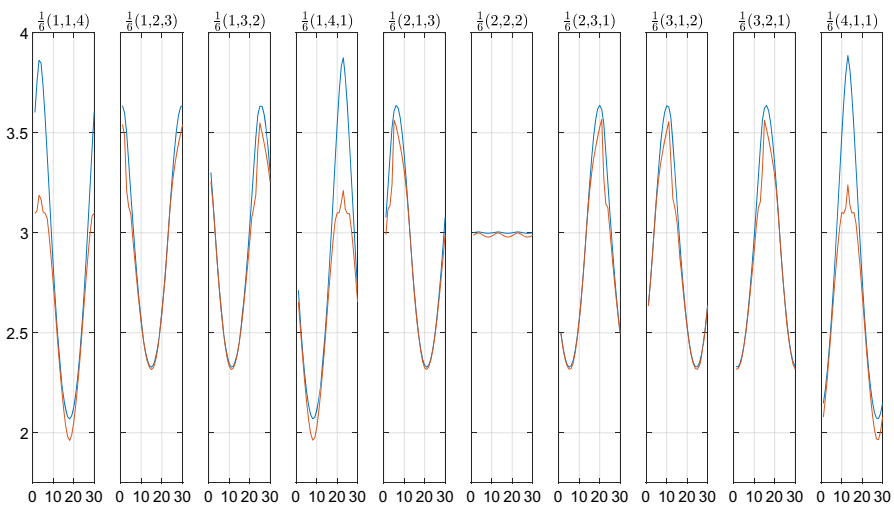

Fig. 5 This figure compares the values of K and κ from Fig. 4. Each subplot corresponds to one choice of ω, indicated at the top, with x axis corresponding to the 30 different exponential families. As can be seen, the curvature κ obtained as the smallest Hessian eigenvalue (red) is, indeed, always a lower bound of the convergence rate K (blue)

cal model or the connectivity structure of a neural network) and the ground metric. Our experiments show that an adequate choice of the two, in conjunction, can significantly increase the rates of convergence in learning. On the other hand, the Ricci curvature depends on both, the information and the Wasserstein metric tensors. An interesting question arises; namely to find the statistical interpolation of such a connection.

We note that the Ricci curvature lower bound is a global notion over the probability model. This is advantageous to provide a uniform analysis, but it can also lead to difficulties, especially when the models include points near the boundary of the simplex, where the behavior is not as regular. Our experiments indicate that restricting the parameter domain to a region bounded away from the boundary of the simplex allows us to closely track the rates of convergence. Another challenge is that, being a global quantity, the computation can be challenging. Nonetheless, we point out that computing the curvature in terms of the Hessian is much cheaper than estimating the learning rates empirically. We have focused on discrete sample spaces, which allowed us to obtain an intuitive and transparent picture of the relationships that derive form this theory. We expect that the derivations extend naturally to the case of continuous sample spaces. We also consider it worthwhile to extend the entropy method to probability models. In other words, to compute the Hessian in statistical manifolds along gradient flows. We believe that the entropy method would be very helpful in estimating the convergence rate of related natural gradient methods.

Another interesting line of investigation is the following. Our definitions are based on the KL divergence and the Wasserstein and Fisher metric tensors. In principle, it is possible to derive analogous definitions for other metric structures. In particular, one can consider the family of f-divergences. Such an analysis could allow us to compare different learning paradigms.

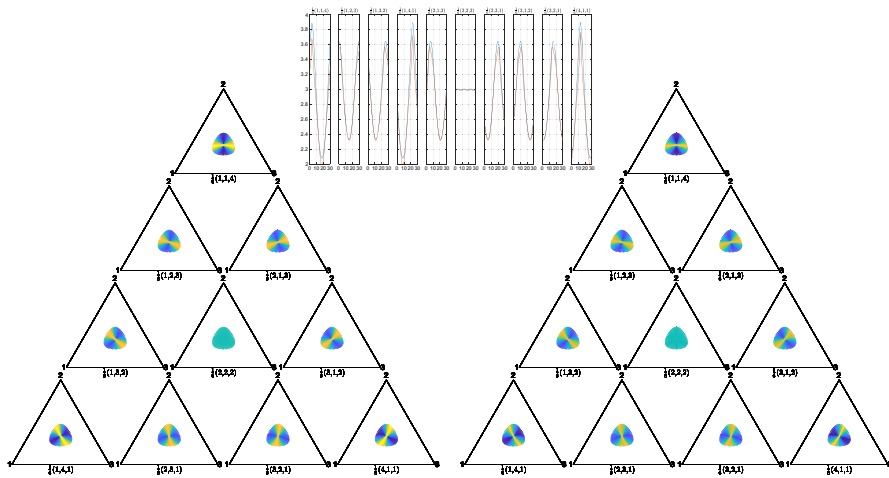# Appendix A. Additional figures to Example 1

Fig. 6 Similar to Fig. 4 but with $\Theta = [-1/2, 1/2]$. Note how on this tight parameter domain around $\theta = 0$ (the value of the reference measure), the Ricci curvature lower bound gives a very close lower bound on the minimum rate of convergence for each of the models. The middle shows the direct comparison of the two values across the 30 exponential families. The minimum rate of convergence is shown in blue, and the Hessian in red
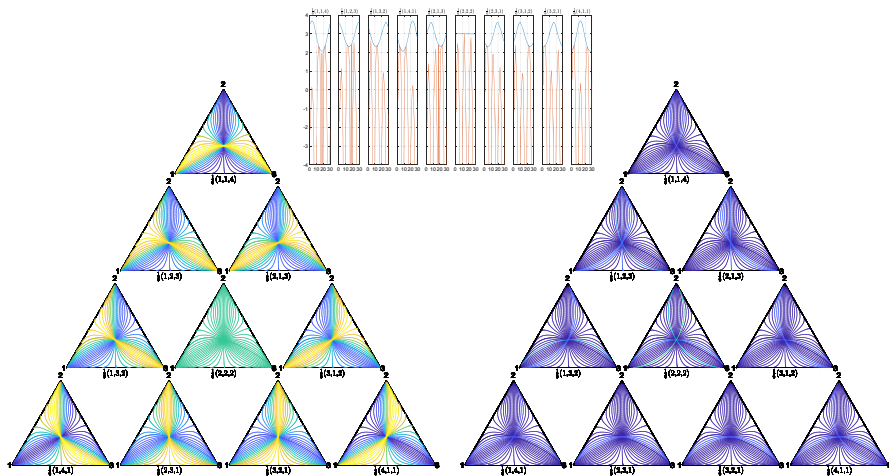

Fig. 7 Similar to Fig. 4, but with a larger parameter domain $\Theta = [-4, 4]$. On this relatively large parameter domain, the models contain points close to the boundary of the simplex, where the Hessian (and the Ricci curvature) can have large oscillations. In turn, we observe larger gaps to the minimum rate of convergence, compared with Fig. 6
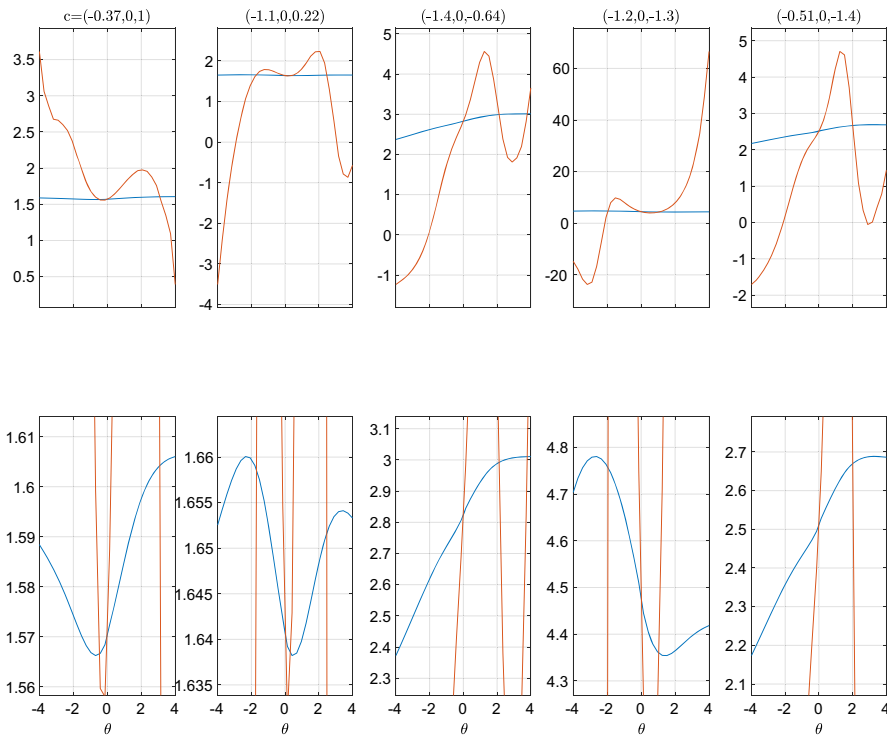
LI, MONTÚFAR

Fig. 8 Convergence rates and minimum Hessian eigenvalue at individual parameter choices. Here we fixed the ground metric $\omega = (\omega_{12}, \omega_{23}, \omega_{13}) = (1/2, 1/2, 0)$. Each subplot corresponds to one exponential family, with sufficient statistic indicated at the top. Within a region around $\theta = 0$ (the value of the reference measure), the minimum of the Hessian is closer to the convergence rates. In fact, the Hessian eigenvalue intersects the rate of convergence at $\theta = 0$. The Hessian at $\theta = 0$ is the asymptotic rate of convergence. The lower row zooms in the y axis of the upper row. For these exponential families, the convergence rates do not vary much across choices of the initial parameter value

References

1. Amari, S.: Natural gradient works efficiently in learning. Neural Comput. **10**(2), 251–276 (1998)
2. Amari, S.: Information Geometry and Its Applications. Number volume 194 in Applied Mathematical Sciences. Springer, Japan (2016)
3. Ay, N., Jost, J., Lê, H.V., Schwachhöfer, L.J.: Information Geometry. Springer, Cham (2017)
4. Bakry, D., Émery, M.: Diffusions hypercontractives. Séminaire de probabilités de Strasbourg **19**, 177–206 (1985)
5. Benamou, J.-D., Brenier, Y.: A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem. Numer. Math. **84**(3), 375–393 (2000)
6. Carlen, E.A., Gangbo, W.: Constrained steepest descent in the 2-Wasserstein metric. Ann. Math. **157**(3), 807–846 (2003)
7. Chen, Y., Li, W.: Natural gradient in Wasserstein statistical manifold. (2018)
8. Chow, S.-N., Huang, W., Li, Y., Zhou, H.: Fokker–Planck equations for a free energy functional or Markov process on a graph. Arch. Ration. Mech. Anal. **203**(3), 969–1008 (2012)
9. Chow, S.-N., Li, W., Lu, J., Zhou, H.: Population games and Discrete optimal transport. arXiv:1704.00855 [math] (2017)

10. Chow, S.-N., Li, W., Zhou, H.: A discrete Schrodinger equation via optimal transport on graphs. arXiv:1705.07583 [math] (2017)
11. Csiszár, I., Shields, P.C.: Information theory and statistics: a tutorial. Commun. Inf. Theory **1**(4), 417–528 (2004)
12. Darroch, J.N., Ratcliff, D.: Generalized iterative scaling for log-linear models. Ann. Math. Stat. **43**(5), 1470–1480, 10 (1972)
13. Erbar, M., Fathi, M.: Poincaré, modified logarithmic Sobolev and isoperimetric inequalities for Markov chains with non-negative Ricci curvature. J. Funct. Anal. **274**(11), 3056–3089 (2018)
14. Erbar, M., Henderson, C., Menz, G., Tetali, P.: Ricci curvature bounds for weakly interacting Markov chains. Electron. J. Probab. **22**, 23 (2017)
15. Erbar, M., Kopfer, E.: Super Ricci flows for weighted graphs. arXiv:1805.06703 [math] (2018)
16. Erbar, M., Maas, J.: Ricci Curvature of finite Markov Chains via convexity of the entropy. Arch. Ration. Mech. Anal. **206**(3), 997–1038 (2012)
17. Erbar, M., Maas, J., Tetali, P.: Discrete Ricci curvature bounds for Bernoulli–Laplace and Random transposition models. Annales de la faculté des sciences de Toulouse Mathématiques **24**(4), 781–800 (2015)
18. Fathi, M., Maas, J.: Entropic Ricci curvature bounds for discrete interacting systems. Ann. Appl. Probab. **26**(3), 1774–1806 (2016)
19. Hua, B., Jost, J., Liu, S.: Geometric analysis aspects of infinite semiplanar graphs with non-negative curvature. Journal für die reine und angewandte Mathematik (Crelles Journal) **2015**(700), 1–36 (2015)
20. Jordan, R., Kinderlehrer, D., Otto, F.: The variational formulation of the Fokker–Planck equation. SIAM J. Math. Anal. **29**(1), 1–17 (1998)
21. Jost, J., Liu, S.: Ollivier's Ricci Curvature, local clustering and curvature-dimension inequalities on graphs. Discr. Comput. Geom. **51**(2), 300–322 (2014)
22. Lafferty, J.D.: The density manifold and configuration space quantization. Trans. Am. Math. Soc. **305**(2), 699–741 (1988)
23. Li, W.: Geometry of probability simplex via optimal transport. arXiv:1803.06360 [math] (2018)
24. Li, W., Montúfar, G.: Natural gradient via optimal transport. Inf. Geom. **1**(2), 181–214 (2018)
25. Lin, Y., Lu, L., Yau, S.-T.: Ricci curvature of graphs. Tohoku Math. J. **63**(4), 605–627 (2011)
26. Lin, Y., Yau, S.-T.: Ricci curvature and eigenvalue estimate on locally finite graphs. Math. Res. Lett. **17**(2), 343–356 (2010)
27. Longford, N.T.: A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. Biometrika **74**(4), 817–827, 12 (1987)
28. Lott, J., Villani, C.: Ricci Curvature for metric-measure spaces via optimal transport. Ann. Math. **169**(3), 903–991 (2009)
29. Maas, J.: Gradient flows of the entropy for finite Markov chains. J. Funct. Anal. **261**(8), 2250–2292 (2011)
30. Mielke, A.: Geodesic convexity of the relative entropy in reversible markov chains. Calc. Var. Part. Differ. Equ. **48**(1), 1–31 (2013)
31. Ollivier, Y.: Ricci curvature of Markov chains on metric spaces. J. Funct. Anal. **256**(3), 810–864 (2009)
32. Ollivier, Y., Villani, C.: A curved Brunn–Minkowski inequality on the discrete hypercube, or: what is the ricci curvature of the discrete hypercube? SIAM J. Discr. Math. **26**(3), 983–996 (2012)
33. Otto, F.: The geometry of dissipative evolution equations: the porous medium equation. Commun. Partial Differ. Equ. **26**(1–2), 101–174 (2001)
34. Otto, F., Villani, C.: Generalization of an Inequality by Talagrand and links with the logarithmic sobolev inequality. J. Funct. Anal. **173**(2), 361–400 (2000)
35. Simon Eberle, B.N., Schlichting, A.: Gradient flow formulation and longtime behaviour of a constrained Fokker–Planck equation. Nonlinear Anal. **158**, 142167 (2017)
36. Solomon, J., Rustamov, R.M., Guibas, L.J., Butscher, A.: Continuous-flow graph transportation distances. arXiv:1603.06927 (2016)
37. Sturm, K.-T.: On the geometry of metric measure spaces. Acta Math. **196**(1), 65–131 (2006)
38. Villani, C.: Optimal Transport: Old and New. Number 338 in Grundlehren der mathematischen Wissenschaften. Springer, Berlin (2009)