THE PUBLISHING HOUSE
OF THE ROMANIAN ACADEMY

PROCEEDINGS OF THE ROMANIAN ACADEMY, Series A
Volume 26, Number 1/2025, pp. 27–34
DOI: 10.59277/PRA-SER.A.26.1.04

CONVEXITY OF KULLBACK-LEIBLER FUNCTIONAL AND THE DEFINITION OF HESSIAN FOR REAL MAPS DEFINED ON FINITE MARKOV SPACES

Alireza BAHRAINI

Sharif University of Technology, Dept. of Mathematical Sciences, P.O.Box 11155-9415, Tehran, Iran.
E-mail: bahraini@sharif.edu

Abstract. In order to study the convexity of Kullback-Leibler functional in the presence of discrete variables, one needs to study the convexity of a linear functional defined on some type of Wasserstein space associated with these discrete spaces. Inspired by an observation in continuous case and by employing the Wasserstein spaces associated with Markov chains as introduced by Erbar-Maas [8], we define the Hessian of a real map defined on a Hypercube as the underlying Markov structure. The convexity of linear functionals in the discrete case can be computed through the Hessian matrices that can be derived through this relationship.

Keywords: finite Markov Spaces, Erbar-Maas Wasserstein space, geodesic convexity.
Mathematics Subject Classification (MSC2020): 39B62, 49Q20.

1. INTRODUCTION

Optimization in the space of probability measures is of special interest both in mathematics, statistics and machine learning. Many classical PDEs can be identified as minimization problems of specific functionals defined over Wasserstein type measure spaces [2, 4, 12]. In machine learning and statistics numerous problems including variational Bayasian inference, maximum likelihood estimations, generative adversarial networks, are stated as minimization over the space of probability distributions [5, 6].

The gradient flow of these probability functionals and their convergence rates towards critical points are governed by the geodesic convexity computed in terms of the geometry induced by Wasserstein metrics. This has led to novel numerical methods to solve PDEs of Fokker-Planck types, as well as new connections between the geometry of the finite dimensional underlying space and the associated Wasserstein probability space. A typical example consists of entropy functional and its role in developing a synthetic notion of Ricci curvature's lower bound for continuous and discrete spaces [9, 11].

Our goal in this note is to introduce a notion of convexity for maps defined on finite Markov spaces by applying similar ideas describing characteristics of an object of finite dimensional nature in terms of the associated Wasserstein space. In fact in the Kullback-Leibler (KL) functional $D_{KL}(p||q) = \int p \log p - \int p \log q$ we have two terms the entropy functional $\int p \log p$ and the simpler term $\int p \log q$. As mentioned above the convexity coefficient of the entropy term has been already discussed. We aim at providing some tools which enable us to investigate the geodesic convexity of the second term in the discrete case. KL functional is widely used in many statistical problems in particular in Mean Field Variational Inference (MFVI). The study of its critical points and convexity coefficients is of essential importance for exploring the rate of convergence of the corresponding mean field equations algorithms towards their fixed point. To this end we first observe in

section 2 that the convexity of a map $f : M \to \mathbb{R}$ defined over a Riemannian manifold (M,g) is equivalent to the geodesic convexity of the functional $\mathscr{L}^f : \mathscr{P}_2(\mathscr{M}) \to \mathbb{R}$ defined on the L^2-Wasserstein space $\mathscr{P}_2(\mathscr{M})$ by

$$\mathscr{L}^f : \mathscr{P}_2(\mathscr{M}) \longrightarrow \mathbb{R}, \qquad \mathscr{L}^f(\rho) := \int_{\mathscr{M}} f \mathrm{d}\rho,$$

In the present note we introduce the Hessian of a real map defined on a hypercube. It turns out that one of the natural graphical structures for which a rich class of convex maps exists consists of the hypercube. This will lead to the uniqueness of solutions to mean field equations of the Ising model. By applying this observation one can compute new partition functions which had remained intractable. The main tool for us is the theory of Wasserstein spaces developed by Maas-Erbar for finite Markov spaces [9].

2. CONVEXITY OF REAL MAPS ON FINITE MARKOV SPACES THROUGH WASSERSTEIN SPACE

A subset $S \subset M$ of a Remannian manifold (M,g) is geodsically convex if for all $x,y \in S$ there exists a geodsic $\gamma : [0,1] \to M$, such that $\gamma(0) = x$ and $\gamma(1) = y$ while $\gamma(t) \in S$ for all $t \in [0,1]$. A function $f : S \to \mathbb{R}$ is geodesically convex if S is geodesically convex and for all geodesic segments $\gamma : [0,1] \to M$ such that $\gamma(0) = x$, $\gamma(1) = y$ and $\gamma(t) \in S$ for all $t \in [0,1]$, the map $f \circ \gamma : [0,1] \to R$ is convex.

We recall the following lemma whose proof is standard:

LEMMA 1. *Let $\mathscr{M}$ be a Riemannian manifold, and let $f : \mathscr{M} \to \mathbb{R}$ be a real function. Then f is geodesically convex iff Hessian of f,* Hess$f$*, is semi-positive definite. Also for C^2 maps $\psi, \phi : \mathscr{M} \to \mathbb{R}$, and a C^1 vector field ν, we have*

$$\langle \nabla \langle \nabla\psi, \nabla\phi \rangle, \nu \rangle = \mathrm{Hess}\phi(\nabla\psi, \nu) + \mathrm{Hess}\psi(\nabla\phi, \nu). \tag{1}$$

The following basic observation for convex maps defined on a Riemannian manifold $\mathscr{M}$ motivates for our definition for convexity of maps defined on discrete spaces.

PROPOSITION 1. *Let $\mathscr{M}$ be a geodesically complete Riemannian manifold, and let*

$$f : \mathscr{M} \to \mathbb{R},$$

be a differentiable real function with compact support. The functional $\mathscr{L}^f$ defined by

$$\mathscr{L}^f : \mathscr{P}_2(\mathscr{M}) \to \mathbb{R}, \qquad \mathscr{L}^f(\mu) = \int_{\mathscr{M}} f \mathrm{d}\mu.$$

is geodesically convex on $\mathscr{P}_2(\mathscr{M})$ iff, Hess(f) *is semi-positive definite every where on $\mathscr{M}$.*

Proof. Here we provide an informal proof without treating regularity which makes use of Hessian. We avoid regularity discussion and we refer to [7] for a rigorous proof using MacCann-Brenier theorem. Assume that $\mathscr{L}^f$ is geodesically convex on $\mathscr{P}_2(\mathscr{M})$. For a geodesic in $\mathscr{P}_2(\mathscr{M})$ given by the continuity equation

$$\begin{cases} \frac{\partial}{\partial_t}\mu_t + \nabla.\big(\mu_t \nabla\psi_t\big) = 0, \\ \frac{\partial}{\partial_t}\psi_t + \frac{|\nabla\psi_t|^2}{2} = 0, \end{cases}$$

we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{L}^f(\mu_t) = \frac{\mathrm{d}}{\mathrm{d}t}\int f(x)\mathrm{d}\mu_t = \int \langle \nabla f, \nabla\psi_t \rangle \mathrm{d}\mu_t.$$

By applying lemma 1, we get

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathscr{L}^f(\mu_t) &= \frac{\mathrm{d}}{\mathrm{d}t}\left[\int \langle \nabla f(x), \nabla \psi_t\rangle \,\mathrm{d}\mu_t\right] \\ &= \int \left\langle \nabla f, -\nabla \frac{|\nabla \psi_t|^2}{2}\right\rangle \mathrm{d}\mu_t + \int \langle \nabla \langle \nabla f, \nabla \psi_t\rangle, \nabla \psi_t\rangle \,\mathrm{d}\mu_t \\ &= \frac{-1}{2}\int \langle \nabla \langle \nabla \psi_t, \nabla \psi_t\rangle, \nabla f\rangle \,\mathrm{d}\mu_t + \int \langle \nabla \langle \nabla f, \nabla \psi_t\rangle, \nabla \psi_t\rangle \,\mathrm{d}\mu_t \\ &\overset{(1)}{=} \frac{-1}{2}\int 2\mathrm{Hess}\psi(\nabla \psi_t, \nabla f)\mathrm{d}\mu_t + \int \mathrm{Hess} f(\nabla \psi_t, \nabla \psi_t)\mathrm{d}\mu_t \\ &\quad + \int \mathrm{Hess}\psi(\nabla f, \nabla \psi_t)\mathrm{d}\mu_t = \int \mathrm{Hess} f(\nabla \psi_t, \nabla \psi_t)\mathrm{d}\mu_t. \end{aligned} \tag{2}$$

Hence, if f is convex, one can deduce that

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathscr{L}^f(\mu_t) = \int \mathrm{Hess} f(\nabla \psi_t, \nabla \psi_t)\mathrm{d}\mu_t \geq 0.$$

Conversely, assume that $\mathscr{L}^f$ is geodesically convex. Consider a geodesic in $\mathscr{P}_2(\mathscr{M})$ with the initial condition $(\mu_0, , \nabla\psi_0)$, where $\mu_0 = \delta_{x_0}, \nabla\psi_0 = v$, $x_0 \in \mathscr{M}$, and $v \in T_{x_0}\mathscr{M}$. Then from (2), it can be seen that

$$0 \leq \int \mathrm{Hess} f(\nabla \psi_0, \nabla \psi_0)\mathrm{d}\mu_0 = \int \mathrm{Hess} f(v, v)\mathrm{d}\delta_{x_0}(y) = \mathrm{Hess} f(x_0)(v, v).$$

This means that $\mathrm{Hess} f$ is semi-positive definite.

□

Inversely, this theorem reduces the study of the convexity of $\mathscr{L}^f$ on $\mathscr{P}_2(\mathscr{M})$ to that of the Hessian of the map f at all the points of $\mathscr{M}$. In order to carry out the same procedure in discrete case we employ the theory developed by Maas and Erbar for discrete finite Markov spaces [8].

They consider an irreducible Markov kernel $Q : \chi \times \chi \to \mathbb{R}_+$, on a finite set χ. Satisfying the condition that for every $x, y \in \chi$, there exists a sequence $\{x_i\}_{i=0}^n \in \chi$, such that

$$x_0 = x, \;\; x_n = y, \qquad \text{and} \qquad Q(x_{i-1}, x_i) > 0 \quad \forall i \in \{1, 2, \ldots, n\}.$$

The unique stationary probability measure π on χ associated with the Markov kernel Q, is assumed to be reversible, i.e. the detailed balance equation holds:

$$Q(x, y)\pi(x) = Q(y, x)\pi(y) \quad \forall x, y \in \chi.$$

Maas and Erbar define the space of all probability densities on χ with respect to π as follows:

$$\mathscr{P}_\pi(\chi) = \left\{ \rho : \chi \to \mathbb{R}_+ \,\middle|\, \sum_{x \in \chi} \pi(x)\rho(x) = 1 \right\}. \tag{3}$$

For $\rho_0, \rho_1 \in \mathscr{P}_\pi(\chi)$, Wasserstein-Mass metric $\mathscr{W}$ on $\mathscr{P}_\pi(\chi)$ is defined by [9]

$$\mathscr{W}(\rho_0, \rho_1)^2 = \inf_{\rho, \psi} \left\{ \frac{1}{2}\int_0^1 \sum_{x, y \in \chi} \big(\psi_t(x) - \psi_t(y)\big)^2 \Lambda\big(\rho_t(x), \rho_t(y)\big) Q(x, y)\pi(x)\mathrm{d}t \right\}, \tag{4}$$

where the infimum is taken over all piece-wise C^1 curves $\rho : [0, 1] \longrightarrow \mathscr{P}_\pi(\chi)$, and the functions $\psi : [0, 1] \times \chi \longrightarrow \mathbb{R}$ satisfies a certain continuity equation. The map Λ is taken to be an arithmetic average for our case. The space $(\mathscr{P}_\pi(\chi), \mathscr{W})$ has the structure of a complete Riemannian manifold.

PROPOSITION 2 [8]. *For any $\bar{\rho} \in \mathscr{P}_\pi(\chi)$ and $\bar{\psi} \in \mathbb{R}^\chi$, on a sufficiently small interval around 0, there exists a unique constant speed geodesic such that $\rho_0 = \bar{\rho}$ and initial tangent vector $\nabla\psi_0 = \nabla\bar{\psi}$ satisfying*

$$\begin{cases} \partial_t \rho_t(x) = -\sum_{y\in\chi} \big(\psi_t(y) - \psi_t(x)\big)\Lambda\big(\rho_t(x),\rho_t(y)\big)Q(x,y), \\ \partial_t \psi_t(x) = -\frac{1}{2}\sum_{y\in\chi} \big(\psi_t(x) - \psi_t(y)\big)^2 \partial_1\Lambda\big(\rho_t(x),\rho_t(y)\big)Q(x,y). \end{cases} \tag{5}$$

Definition of convex functions on Markov spaces. Let (χ, Q, π) be a Markov triple on a finite space χ as described above. Let $f : \chi \to \mathbb{R}$ be a real valued function on χ. Inspired by the observation given in Proposition 1, we want to introduce a notion of convexity for f. To this end, consider the functional L^f_π

$$\begin{aligned} &L^f_\pi : \mathscr{P}_\pi(\chi) \longrightarrow \mathbb{R} \\ &L^f_\pi(\rho) = \sum_{w\in\mathscr{X}} f(w)\rho(w). \end{aligned} \tag{6}$$

In the discrete case, we are interested in geodesic convexity of L^f_π on a totally geodesic submanifold of $\mathscr{P}_\pi(\chi)$ passing through all the vertices of the simplex $\mathscr{P}_\pi(\chi)$ (see Definition 1). In the case where χ has only two points, $\mathscr{P}_\pi(\chi)$ will become a 1-simplex in $\mathbb{R}^2$. We start by identifying geodesically convex functionals L^f_π on this 1-simplex.

Let us denote by $\mathscr{P}(\chi)$, the space of probability measures on χ, which consists of all maps $\hat{\rho} : \chi \to [0,1]$ satisfying $\sum_{x\in\chi}\hat{\rho}(x) = 1$. Let $\mathscr{H}$ be the map defined by

$$\mathscr{H} : \mathscr{P}_\pi(\chi) \longrightarrow \mathscr{P}(\chi), \qquad (\mathscr{H}(\rho))(x) = \pi(x)\rho(x).$$

Since $\mathscr{H}$ is a bijection map, we can use it to equipe $\mathscr{P}(\chi)$ with a Riemannian metric $\mathscr{G}$ with respect to which $\mathscr{H}$ is an isometry. The Riemannian metric on $\mathscr{P}_\pi(\chi)$ is defined in [8].

Given a real function $f : \chi \to \mathbb{R}$, let $L^f : \mathscr{P}(\chi) \to \mathbb{R}$ be defined by

$$L^f(\hat{\rho}) = \sum_{x\in\chi} \hat{\rho}(x) f(x).$$

Our aim is to study the geodesic convexity of L^f with respect to $\mathscr{G}$.

Based on the observation of Proposition 1, we introduce a notion of convexity for real functions $f : \chi \to \mathbb{R}$ on the discrete Markov space (χ, Q, π). We denote by $\mathscr{G}_\pi$ the Erbar-Maas metric on $\mathscr{P}_\pi(\chi)$, and we set $\mathscr{G} := ((\mathscr{H})^{-1}) * \mathscr{G}_\pi$, which is a Riemannian metric on $\mathscr{P}(\chi)$.

Definition 1. For a real $\lambda \in \mathbb{R}$, a map $f : \chi \to \mathbb{R}$ is called λ-convex, if there exists a totally geodesic submanifold $\mathscr{A} \subset \mathscr{P}(\chi)$ containing all the vertices of the simplex $\mathscr{P}(\chi)$ such that $L^f : \mathscr{A} \to \mathbb{R}$ is λ-convex with respect to $\mathscr{G}$. This is equivalent to say that $L^{\pi f}_\pi : \mathscr{A}_\pi \to \mathbb{R}$ to be λ-convex with respect to $\mathscr{G}_\pi$, where $\mathscr{A}_\pi \subset \mathscr{P}_\pi(\chi)$ is a totally geodesic submanifold containing all the vertices of $\mathscr{P}_\pi(\chi)$.

2.1. Tensorisation and hypercube structure

Let (χ_i, Q_i, π_i) , for $1 \le i \le n$, be irreducible and reversible Markov chains. Let $Q_{(i)}$ denote the lift of Q_i to the product space $\chi = \chi_1 \times \cdots \times \chi_n$, defined for $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ by

$$Q_{(i)}(x,y) = \begin{cases} Q_i(x_i, y_i), & \text{If } x_j = y_j \text{ for all } j \neq i, \\ 0, & \text{Otherwise .} \end{cases}$$

For $\alpha = (\alpha_1, \ldots, \alpha_n)$ satisfying $\sum_{i=1}^n \alpha_i = 1, \quad \alpha_i \ge 0, \quad 1 \le i \le n$, consider a weighted product Markov chain

(χ, Q_α, π), such that

$$Q_\alpha := \sum_{i=1}^{n} \alpha_i Q_{(i)},$$

and where the corresponding reversible probability measure is given by $\pi = \pi_1 \otimes \cdots \otimes \pi_n$.

Let (χ, Q, π) denote the above tensorized Markov space where we are setting $Q := Q_\alpha$ for $\alpha = (\frac{1}{n}, \ldots, \frac{1}{n})$, therefore

$$Q(x,y) = \sum_{i=1}^{n} \frac{1}{n} Q_{(i)}(x,y). \tag{7}$$

Let $\rho_t^i \in \mathscr{P}_{\pi_i}(\chi_i)$ be a geodesic and for any $x = (x_1, \ldots, x_n) \in \chi$, we define

$$\rho_t(x) = \left(\rho^1_{\frac{t}{n}} \times \cdots \times \rho^n_{\frac{t}{n}}\right)(x) := \prod_{i=1}^{n} \rho^i_{\frac{t}{n}}(x_i). \tag{8}$$

It is not difficult to see that $\rho_t \in \mathscr{P}_\pi(\chi)$.

Also for ψ_t^i satisfying the geodesic equation (5), we set

$$\psi_t(x) := \sum_{i=1}^{n} \psi^i_{\frac{t}{n}}(x_i). \tag{9}$$

Our claim is that $t \to \rho_t$ is a geodesic in $\mathscr{P}_\pi(\chi)$:

THEOREM 1. *The pair (ρ_t, ψ_t) satisfies the system of equations* (5)*, which means that $t \to \rho_t$ defines a geodesic in $\mathscr{P}_\pi(\chi)$.*

Proof. From (8) and (5) we have

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\rho_t(x) &= \frac{\mathrm{d}}{\mathrm{d}t}\left(\rho^1_{\frac{t}{n}} \times \cdots \times \rho^n_{\frac{t}{n}}\right)(x_1, \ldots, x_n) = \frac{\mathrm{d}}{\mathrm{d}t}\prod_{i=1}^{n} \rho^i_{\frac{t}{n}}(x_i) = \sum_{j=1}^{n} \frac{\mathrm{d}}{\mathrm{d}t}\rho^j_{\frac{t}{n}}(x_j) \prod_{i\neq j} \rho^i_{\frac{t}{n}}(x_i) \\
&= -\frac{1}{n}\sum_{j=1}^{n} \prod_{i\neq j} \rho^i_{\frac{t}{n}}(x_i) \sum_{y_j \in \chi_j} \left(\psi^j_{\frac{t}{n}}(y_j) - \psi^j_{\frac{t}{n}}(x_j)\right) \Lambda\left(\rho^j_{\frac{t}{n}}(y_j), \rho^j_{\frac{t}{n}}(x_j)\right) Q_j(x_j, y_j) \\
&= -\frac{1}{n}\sum_{j=1}^{n} \sum_{y_j \in \chi_j} \left(\psi^j_{\frac{t}{n}}(y_j) - \psi^j_{\frac{t}{n}}(x_j)\right) \Lambda\left(\rho^j_{\frac{t}{n}}(y_j) \prod_{i\neq j} \rho^i_{\frac{t}{n}}(x_i), \rho^j_{\frac{t}{n}}(x_j) \prod_{i\neq j} \rho^i_{\frac{t}{n}}(x_i)\right) Q_j(x_j, y_j) \\
&= -\sum_{y \in \chi, y \sim x} \left(\psi_t(y) - \psi_t(x)\right) \Lambda\left(\rho_t(y), \rho_t(x)\right) Q(x,y).
\end{aligned} \tag{10}$$

Here, in the third line, we used property $\Lambda(lt, ls) = l\Lambda(t,s)$, for $l > 0$ and $s, t \geq 0$, and, in the last line, we used the definition of ψ_t in (9) and Q in (7). Also, $y \sim x$ means y is a neighborhood of x with respect to the hypercube structure over χ.

From (9), it can also be seen that

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\psi_t &= -\frac{1}{2n}\sum_{i=1}^{n} \sum_{y_i \in \chi_i} \left(\psi^i_{\frac{t}{n}}(y_i) - \psi^i_{\frac{t}{n}}(x_i)\right)^2 \partial_1 \Lambda\left(\rho^i_{\frac{t}{n}}(x_i), \rho^i_{\frac{t}{n}}(y_i)\right) Q_i(x_i, y_i) \\
&= -\frac{1}{2}\sum_{y \sim x} \left(\psi_t(y) - \psi_t(x)\right)^2 \partial_1 \Lambda\left(\prod_{i=1}^{n} \rho^i_{\frac{t}{n}}(x_i), \prod_{i=1}^{n} \rho^i_{\frac{t}{n}}(y_i)\right) Q(x,y).
\end{aligned} \tag{11}$$

(10) and (11) are nothing but the geodesic equations for the pair (ρ_t, ψ_t) on $\mathscr{P}_\pi(\chi)$. Note that here again, we used the derivative of the relation $\Lambda(lt, ls) = l\Lambda(t,s)$, for $l > 0$ and $s, t \geq 0$, and the definition of Q.

Consider the spaces $\mathscr{A} \subset \mathscr{P}(\chi)$ and $\mathscr{A}_\pi \subset \mathscr{P}_\pi(\chi)$ respectively defined by

$$\mathscr{A} = \{\prod_{i=1}^{n} \rho_i | \rho_i \in \mathscr{P}(\chi_i)\}, \qquad \mathscr{A}_\pi = \{\prod_{i=1}^{n} \hat{\rho}_i | \hat{\rho}_i \in \mathscr{P}_\pi(\chi_i)\}.$$

COROLLARY 1. *The spaces $\mathscr{A}$ and $\mathscr{A}_\pi$ are totally geodesic subspaces of $\mathscr{P}(\chi)$ and $\mathscr{P}_\pi(\chi)$, respectively.*

3. HESSIAN OF A REAL MAP DEFINED ON HYPERCUBE

Now consider n two-point Markov chains (χ_k, Q_k, π_k), where as before $\chi_k = \{\pm 1\}$, for $k = 1, \ldots, n$. Let (χ, Q, π) denote the tensorized product space. If for $k = 1, ..., n$ we set $U := Q_k(+1,-1)$, and $V := Q_k(-1,+1)$, then by stationarity condition (3) $\pi_k(+1) = \frac{V}{U+V}$, and $\pi_k(-1) = \frac{U}{U+V}$,. We assume that $U > V$, also Λ is assumed to be arithmetic mean then from the continuity equation (10) one can deduce that

$$\frac{d}{dt} L^f(\rho_t) = \sum_x \Gamma(f, \psi_t)(x) \rho_t(x) \pi(x)$$

where $\Gamma(f,g)(x) := \frac{1}{2} \sum_y Q(x,y) \left(f(x) - f(y)\right) \left(g(x) - g(y)\right)$. A second derivation along a geodesic ρ_t leads to

$$\frac{d^2}{dt^2} L^f(\rho_t) = \sum_x \left(\Gamma\left(\Gamma(f, \psi_t), \psi_t \right) - \frac{1}{2} \Gamma\left(f, \Gamma(\psi_t)\right) \right)(x) \rho_t(x) \pi(x)$$

Hence if we define the Hessian of f as follows

$$Hess(f)[\nabla\phi, \nabla\psi] = \Gamma\left(\Gamma\left(f, \phi\right), \psi\right) - \frac{1}{2} \Gamma\left(f, \Gamma(\phi, \psi)\right)$$

then by taking $\rho_t(x) = \delta_{x_0}$ it can be seen that L^f is geodesically convex iff $Hess(f)$ is positive definite at all $x \in \chi$. We have

$$\begin{aligned}
\Gamma(\Gamma(f,\phi),\psi)(x) &= \frac{1}{2} \sum_y Q(x,y) \left(\Gamma(f,\phi)(x) - \Gamma(f,\phi)(y)\right) \left(\psi(x) - \psi(y)\right) \\
&= \frac{1}{2} \sum_y Q(x,y) \left(\frac{1}{2} \sum_z Q(x,z) \left(f(x) - f(z)\right) \left(\phi(x) - \phi(z)\right) \right. \\
&\left. - \frac{1}{2} \sum_z Q(y,z) \left(f(y) - f(z)\right) \left(\phi(y) - \phi(z)\right) \right) \left(\psi(x) - \psi(y)\right) \\
&= \frac{1}{4} \sum_{y,z} Q(x,y) Q(x,z) \left(f(x) - f(z)\right) \left(\phi(x) - \phi(z)\right) \left(\psi(x) - \psi(y)\right) \\
&- \frac{1}{4} \sum_{y,z} Q(x,y) Q(y,z) \left(f(y) - f(z)\right) \left(\phi(y) - \phi(z)\right) \left(\psi(x) - \psi(y)\right)
\end{aligned}$$

and

$$\begin{aligned}
\Gamma(\Gamma(\phi,\psi),f)(x) &= \frac{1}{4} \sum_{y,z} Q(x,y) Q(x,z) \left(\phi(x) - \phi(z)\right) \left(\psi(x) - \psi(y)\right) \left(f(x) - f(y)\right) \\
&- \frac{1}{4} \sum_{y,z} Q(x,y) Q(y,z') \left(\phi(y) - \phi(z')\right) \left(\psi(y) - \psi(z')\right) \left(f(x) - f(y)\right)
\end{aligned}$$

Now at a fixed $x \in \chi$ and for $k = 1, ..., n$ we define $\phi_k : \chi \to \mathbb{R}$ by $\phi_k(y) = \begin{cases} 0 & \text{if } y_k = x_k \\ 1 & \text{Otherwise} \end{cases}$. Then if r_k

for $k = 1,...,n$ denotes the operator $r_k(\theta_1,...,\theta_k,...,\theta_n) = (\theta_1,...,-\theta_k,...,\theta_n)$, then since $(\phi(r_k(x)) - \phi(r_l \circ r_k(x)))(\psi(r_k(x)) - \psi(r_l \circ r_k(x))) = 0$ we obtain

$$\begin{aligned}
[\mathrm{Hess}(f)[\nabla\phi_k,\nabla\phi_l](x) + \mathrm{Hess}(f)[\nabla\phi_l,\nabla\phi_k](x) & \\
&= \left[\frac{1}{8}Q(x,r_k(x))Q(x,r_l(x))\,(2f(x) - f\circ r_k(x) - f\circ r_l(x))\right. \\
&- \frac{1}{4}Q(x,r_k(x))Q(r_k(x),r_l\circ r_k(x))(f(r_k(x)) - f(r_l\circ r_k(x))) \\
&\left. - \frac{1}{4}Q(x,r_l(x))Q(r_l(x),r_k\circ r_l(x))(f(r_l(x)) - f(r_k\circ r_l(x)))\right] \\
&= \frac{UV}{n^2}\left[\frac{1}{4}f(x) - \frac{3}{8}f(r_k(x)) - \frac{3}{8}f(r_l(x)) + \frac{1}{2}f(r_k\circ r_l(x))\right]
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{Hess}(f)[\nabla\phi_k,\nabla\phi_k](x) &= \frac{1}{4}Q(x,r_k(x))Q(x,r_k(x))\,(f(x) - f(r_k(x))) \\
&- \frac{1}{4}Q(r_k(x),x)Q(r_k(x),x)\,(f(x) - f(r_k(x))) \\
&= \frac{1}{\pi_k(x_k)}\frac{UV(U-V)}{4n^2(U+V)}\big[f(x_1,...,x_{k-1},+1,x_{k+1},...,x_n) \\
&\qquad - f(x_1,...,x_{k-1},-1,x_{k+1},...,x_n)\big]
\end{aligned}$$

We set $H_{kl} = \frac{1}{2}[Hess(f)[\nabla\phi_k,\nabla\phi_l](x) + Hess(f)[\nabla\phi_l,\nabla\phi_k](x)$ and we define the matrix $H := [H_{k,l}]_{1\le k,l\le n}$.

Definition 2. We call H the Hessian of the map f at the point $(\theta_1,...,\theta_n)$.

We have proved that

THEOREM 2. *The map $f : \chi \to \mathbb{R}$ is strictly convex If H is positive definite at all $(\theta_1,...,\theta_n)$.*

THEOREM 3. *Given a strictly convex function $f : \chi \to \mathbb{R}$ on a Markov chain (χ, Q, π) in the sense of Definition 1, then f admits a unique minimum on χ.*

Proof. L^f is linear on the simplex $\mathscr{P}(\chi)$. Hence the minimum of L^f occurs at a unique vertex of the simplex $\mathscr{P}(\chi)$. On the other hand, by definition 1 we know that $L^f|_{\mathscr{A}}$ is geodesically strictly convex. So the minimum of L^f can ony occur on one of the vertices of $\mathscr{A}$, and moreover it is unique. The vertices of $\mathscr{P}(\chi)$ are of the form δ_x for some $x \in \chi$, where δ_x denotes the Dirac probability distribution on χ concentrated at x. Moreover, by definition of L^f we know that $L^f(\delta_x) = f(x)$. This proves that the minimum of f can only occur at a single point in χ.

□

3.1. Examples of convex maps on hypercube and the Ising model

Consider a linear map $f : \chi \to \mathbb{R}$ given by the following relation

$$f(\theta_1,...,\theta_n) = \sum_{k=1}^{n} a_k\theta_k$$

Then the Hessian of f at all the points $(\theta_1,...,\theta_n)$ is diagonal and it is positive definite exactly when the diagonal coefficients are positive. It is not difficult to see that this occurs when $a_k > 0$ for all k.

The quadratic map $f:\chi :\to \mathbb{R}$ defined by

$$f(\theta_1,...,\theta_n)=\sum_{k<l}J_{kl}\theta_k\theta_l,$$

is never convex. To see this, we observe that for $k\neq l$ $H_{kl}=4J_{kl}$ and

$$H_{kk}=f(\theta_1,...,\theta_{k-1},1,\theta_{k+1},...,\theta_n)-f(\theta_1,...,\theta_{k-1},-1,\theta_{k+1},...,\theta_n)=2\sum_{j\neq k}J_{kj}\theta_j.$$

There always exists choices for $\theta_1,..,\theta_n$, for which the above quantity H_{kk} is negative.

As is explained in [10], mean field equations in statistical physics can be obtained by variational method applied to the Kullback-Leibler divergence, where $\mathscr{A}\subset\mathscr{P}(\chi)$ consists of the space of factorized probability distributions. More precisely in an Ising model with the energy function $\mathscr{H}(\theta_1,...,\theta_n)=\sum_{k<l}J_{kl}\theta_k\theta_l+\sum_{k=1}^{n}a_k\theta_k$, where J_{kl}s are the coupling constants representing the strength of interaction between the atoms k and l and a_k's the external field of the system, the so called mean field equation can be derived through a variational problem of the form $\min_{q\in\mathscr{A}}D_{KL}(q||p)$. Now due to the fact that hyper cubes' Ricci curvature is non-negative according to [9] one can derive the following theorem,

THEOREM 3. *If the energy function $\mathscr{H}(\theta_1,...,\theta_n)=\sum_{k<l}J_{kl}\theta_k\theta_l+\sum_{k=1}^{n}a_k\theta_k$ is convex, then the $D_{KL}(.||p):\mathscr{A}\to\mathbb{R}$ is convex, and therefore the optimization problem $\min_{q\in\mathscr{A}}D_{KL}(q||p)$ admits a unique solution. Here p denote the probability distribution $p(\theta_1,...,\theta_n)=\frac{e^{\mathscr{H}(\theta_1,...,\theta_n)}}{\sum_\theta e^{\mathscr{H}(\theta_1,...,\theta_n)}}$.*

ACKNOWLEDGEMENTS

The author would like to thank the anonymous referee for valuable comments and suggestions on this paper.

REFERENCES

[1] Bahraini A. Mean field variational Bayesian inference for Gaussian mixture model rigorous uncertainty quantification. Submitted. 2024.

[2] Jordan R, Kinderlehrer D, Otto F. The variational formulation of the Fokker-Planck equation. SIAM J. Math. Anal. 1998; 29(1): 1–17. DOI: 10.1137/S0036141096303359.

[3] Otto F. The geometry of dissipative evolution equations: the porous medium equation. Comm. Partial Differential Equations. 2001; 26(1-2): 101–174. DOI: 10.1081/PDE-100002243.

[4] Ambrosio L, Gigli N, Savare G. Gradient flows in metric spaces and in the Wasserstein space of probability measures. Lectures in Mathematics. ETH Zürich: Birkhäuser Basel; 2005. DOI: 10.1007/b137080.

[5] Blei DM, Kucukelbir A, McAuliffe JD, Variational inference: a review for statisticians. J. Amer. Statist. Assoc. 2017; 112(518): 859–877.

[6] Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks. In: Precup D, Teh YW. Proceedings of the 34th International Conference on Machine Learning. Proc. of Machine Learning Research (PMLR), vol. 70. 2017, pp. 214–223.

[7] Bahraini A, Sadeghi S. Optimal transport and variational Bayesian inference. International Journal of Approximate Reasoning. 2023; 162: 109022.

[8] Erbar M, Maas J. Ricci curvature of finite Markov chains via convexity of the entropy. Arch. Rational. Mech. Anal. 2012; 206: 997–1038. DOI: 10.1007/s00205-012-0554-z.

[9] Maas J. Gradient flows of the entropy for finite Markov chains. Journal of Functional Analysis. 2011; 261(8): 2250–2292. DOI: 10.1016/j.jfa.2011.06.009.

[10] Opper M, Winther O. From naive mean field theory to the TAP equations. In: Advanced mean field methods: theory and practice. MIT Press; 2001, ch. 2, pp. 7–20.

[11] Lott J, Villani C. Ricci curvature for metric-measure spaces via optimal transport. Ann. of Math. 2009; 169(3): 903–991.

[12] Otto F. The geometry of dissipative evolution equations: the porous medium equation. Comm. Partial Differential Equations. 2001; 26(1-2): 101–174. DOI: 10.1081/PDE-100002243.

Received September 22, 2024