MATHEMATICAL ENGINEERING TECHNICAL REPORTS

A Jacobian Inequality for Gradient Maps on the Sphere and Its Application to Directional Statistics

Tomonari SEI

METR 2009–23 June 2009

DEPARTMENT OF MATHEMATICAL INFORMATICS
GRADUATE SCHOOL OF INFORMATION SCIENCE AND TECHNOLOGY
THE UNIVERSITY OF TOKYO
BUNKYO-KU, TOKYO 113-8656, JAPAN

WWW page: http://www.keisu.t.u-tokyo.ac.jp/research/techrep/index.html

A Jacobian inequality for gradient maps on the sphere and its application to directional statistics

Tomonari SEI

June 4, 2009

Abstract

In the field of optimal transport theory, an optimal map is known to be a gradient map of a potential function satisfying cost-convexity. In this paper, the Jacobian determinant of a gradient map is shown to be log-concave with respect to a convex combination of the potential functions when the underlying manifold is the sphere and the cost function is the distance squared. The proof uses the non-negative cross-curvature property of the sphere recently established by Kim and McCann. As an application to statistics, a new family of probability densities on the sphere is defined in terms of cost-convex functions. The log-concave property of the likelihood function follows from the inequality.

1 Introduction

In recent years, the theory of optimal transport has been actively studied. In particular, properties of the optimal transport map on Riemannian manifolds are well established. The existence and uniqueness theorem for the optimal transport map on Riemannian manifolds was proved by McCann (2001); this result extended the pioneering work of Brenier (1991) for the Euclidean case. He showed that optimal transport is given by the gradient map of a so-called cost-convex function. On the other hand, for statistical data analysis on Euclidean space, it is useful to consider convex combinations of convex functions in order to construct various probability density functions (Sei (2006)). In this paper, we show that when the underlying space is the sphere, the convex combination of cost-convex functions is actually cost-convex (Lemma 1) and the Jacobian determinant of the resultant gradient map is log-concave with respect to the convex combination (Theorem 1). This result is an extension of the Jacobian interpolation inequality shown by Cordero-Erausquin et al. (2001). We refer to our Jacobian inequality as *the Jacobian inequality* throughout this paper, for simplicity.

Our result is related to the regularity theory of optimal transport maps. Here we consider some recent studies in this field. Ma et al. (2005) showed that regularity of the transport map for general cost functions on Euclidean space is assured if a geometrical quantity called the *cost-sectional curvature* is positive. Conversely, Loeper (2005) showed that non-negativity of the cost-sectional curvature is necessary for regularity. He also showed that non-negativity of the cost-sectional curvature implies non-negativity of the usual sectional curvature if the cost function is the squared distance on a Riemannian manifold. However, the converse does not hold (Kim (2007)). Comprehensive assessment on the theory of optimal transport has been published (Villani (2009)). Kim and McCann (2007) and Kim and McCann (2008) defined *the cross-curvature* and showed that the sphere S^n has almost positive cross-curvature. In general, the cost-sectional curvature is non-negative if the cross-curvature is non-negative. In the present paper, we use the non-negative cross-curvature property of the sphere to prove our main results.

We show that our Jacobian inequality opens several doors for applications to directional statistics. In this field, a family of probability densities is used to analyze given directional data, such as locations on the earth. For example, a test on the directional character of given data is constructed via families of probability density functions on the sphere. Directional statistics has a long history since Fisher (1953) and a comprehensive text on this subject has been published (Mardia and Jupp (2000)).

We define a probability density function on the sphere by the gradient maps of cost-convex functions. Although, in the context of optimal transport, one usually considers push-forward of probability densities, we construct a family of densities by means of *pull-back* of probability densities. This follows from the fact that a pull-back density has an explicit expression for the likelihood function needed for statistical analysis. The density function does not need any special functions such as the modified Bessel function, which usually appear in directional statistics. Furthermore, the Jacobian inequality implies that the likelihood function is log-concave with respect to the statistical parameters. This property is reasonable for computation of the maximum likelihood estimator. We propose more specific models and show graphical images of each probability density. In terms of analysis of real data, we present the result of density estimation for some astronomical data.

This paper is organized as follows. In Section 2, we present basic notation and state our main theorem. In Section 3, we construct a family of probability density functions on the sphere and apply them to directional statistics. All mathematical proofs of the main theorem and lemmas are given in Section 4. Finally we present a discussion in Section 5.

2 Main theorem

Let S^n be the n-dimensional unit sphere. The tangent space at $x \in S^n$ is denoted by $T_x S^n$. The geodesic distance (arc length) between x and y in S^n is denoted by $d(x, y)$. The cost function is $c(x, y) = (1/2)d(x, y)^2$. If one uses Euclidean coordinates in $\mathbb{R}^{n+1}$ to express S^n, then $d(x, y) = \cos^{-1}(x^\top y)$, where the range of $\cos^{-1}$ is $[0, \pi]$.

The *c-transform* ϕ^c of a function $\phi : S^n \to \mathbb{R}$ is defined by

$$\phi^c(y) \;=\; \sup_{x \in S^n} \{-c(x, y) - \phi(x)\}\,. \tag{1}$$

The function ϕ is said to be cost-convex, or *c-convex*, if $(\phi^c)^c = \phi$. Examples of c-convex functions will be given in Section 3. By compactness of S^n, a function ϕ is c-convex if and only if for any $x \in S^n$ there exists some (not necessarily unique) $y \in S^n$ such that $c(x, y) + \phi(x) = \inf_{z \in S^n}\{c(z, y) + \phi(z)\}$.

The image of the exponential map of $v \in T_x S^n$ at $x \in S^n$, denoted by $\exp_x(v)$, is the end point of the geodesic starting at x with the initial vector v. More explicitly, if one uses Euclidean coordinates in $\mathbb{R}^{n+1}$ to express S^n and $T_x S^n$, the exponential map is written as $\exp_x(v) = (\cos|v|)x + (\sin|v|)(v/|v|)$, where $|v|$ denotes the Euclidean norm of the vector v. The exponential map $\exp_x$ is a diffeomorphism from $\{v \in T_x S^n \mid |v| < \pi\}$ to $S^n \setminus \{x'\}$, where x' is the antipodal point of x.

The following lemma is a consequence of the non-negative cross-curvature property of the sphere established by Kim and McCann (2008). See Section 4 for a proof.

Lemma 1 (Convex combination of c-convex functions)**.** If ϕ_0 and ϕ_1 are c-convex, then for each $t \in [0, 1]$ the function $(1 - t)\phi_0(x) + t\phi_1(x)$ of x is also c-convex.

We define $G_\phi(x) = \exp_x(\nabla\phi(x))$ as long as ϕ is differentiable at x, where ∇ is the gradient operator. Following Delanoë and Loeper (2006), we call $G_\phi : S^n \to S^n$ the *gradient map* associated with the *potential function* ϕ. The map G_ϕ is differentiable at x if $|\nabla\phi(x)| < \pi$ and ϕ has its Hessian at x. It is known that any c-convex ϕ on any compact Riemannian manifold is Lipschitz and therefore differentiable almost everywhere. Furthermore, ϕ has a Hessian almost everywhere in the Alexandrov sense, and therefore $G_\phi(x)$ is differentiable almost everywhere (see McCann (2001) and Cordero-Erausquin et al. (2001)). These technical facts on differentiability are important for the theory of optimal transport. However, we will not need them because, for statistical applications, we can assume from the beginning that $G_\phi(x)$ is differentiable except at a finite set of points (see Section 3).

For any c-convex functions ϕ_0 and ϕ_1, by Lemma 1, the convex combination $\phi_t(x) = (1-t)\phi_0(x) + t\phi_1(x)$ is c-convex. We define an interpolation of gradient maps by

$$F_t(x) \;=\; G_{\phi_t}(x) \;=\; \exp_x(\nabla\phi_t(x)), \quad t \in [0,1].$$

Assume that for each $i \in \{0,1\}$, $|\nabla\phi_i(x)| < \pi$ and $\phi_i(x)$ has its Hessian at x. Then it is easy to see that $|\nabla\phi_t(x)| < \pi$ and $\phi_t(x)$ has its Hessian defined at x. We define the Jacobian determinant $J_t(x) = \mathrm{Jac}(F_t(x)) = \det(dF_t/dx)$ with respect to any orthonormal basis on T_xS^n and $T_{F_t(x)}S^n$ with suitable orientations.

The following theorem is our main result.

Theorem 1 (Jacobian inequality)**.** Let ϕ_0 and ϕ_1 be two c-convex functions. Let x be a point in S^n such that, for each $i = 0, 1$, $|\nabla\phi_i(x)| < \pi$ and ϕ_i has its Hessian defined at x. Then the Jacobian determinant $J_t(x)$ defined above is log-concave with respect to t. It is equivalent to the inequality

$$\log J_t(x) \;\geq\; (1-t)\log J_0(x) + t\log J_1(x).$$

We refer to the above inequality as the *Jacobian inequality* in this paper.

Remark 1. This theorem is an extension of the result obtained by Cordero-Erausquin et al. (2001). They showed a similar inequality under the additional assumption that $\phi_0 \equiv 0$, as a corollary of a stronger inequality related to the geometric-arithmetic inequality. It is not known whether the stronger one holds for our case $\phi_0 \not\equiv 0$ (see also Remark 3). □

3 Application to directional statistics

3.1 Probability densities induced by gradient maps

In Sei (2006), the author proposed a family of probability density functions in terms of gradient maps on Euclidean space, where a probability density is constructed as a pull-back of some fixed measure (typically Gaussian) pulled by a gradient map. The notion can be directly extended to probability density functions on the sphere.

For statistical application, we will consider only c-convex functions ϕ such that the gradient map G_ϕ is an isomorphism on S^n and ϕ has its Hessian defined everywhere except at a finite set of points. We define some related terminology.

Definition 1 (Wrapping potential function)**.** We say that a function ϕ is a *wrapping potential function* if ϕ is c-convex, ϕ has its Hessian defined everywhere except for a finite set of points and G_ϕ is an isomorphism on S^n. Let $W(S^n)$ be the set of all wrapping potential functions.

We have the following lemma.

Lemma 2. If ϕ_0 and ϕ_1 are in $W(S^n)$, then the interpolation $\phi_t = (1-t)\phi_0 + t\phi_1$ is also in $W(S^n)$.

We construct a probability density function for each $\phi \in W(S^n)$. Let U be a random variable on S^n distributed uniformly. Then, since $x \mapsto G_\phi(x)$ is surjective, we can define a random variable on S^n by $X = G_\phi^{-1}(U)$. The probability density function of X with respect to the uniform measure is $p_\phi(x) = \mathrm{Jac}(G_\phi(x))$, where the symbol Jac refers to the Jacobian determinant. In other words, we define $p_\phi(x)$ by the *pull-back* measure of the uniform measure pulled by the gradient map G_ϕ. It is distinct from the push-forward measure typically used in other applications of optimal transport.

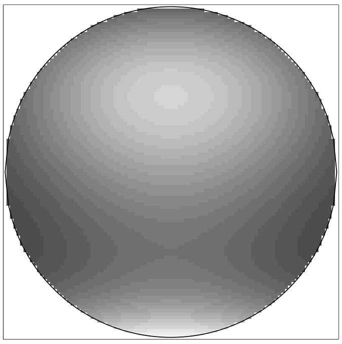
At this point, we describe the exact sampling method of the probability density function $p_\phi(x)$. A sampling procedure is important if one needs to calculate expectations by the Monte Carlo method. From the definition, it is clear that the random variable $X = G_\phi^{-1}(U)$ with a uniformly random variable U on S^n has density $p_\phi(x)$. Hence if we can generate U and solve the equation $G_\phi(X) = U$ effectively, we obtain a random sample X. Indeed, U is quite easily generated, for example, by normalization of a standard Gaussian sample in $\mathbb{R}^{n+1}$. To solve $G_\phi(X) = U$, it is sufficient to find the unique minimizer of the function $c(x, U) + \phi(x)$ with respect to x since the following lemma holds.

Lemma 3. [Lemma 7 of McCann (2001)] If ϕ is c-convex and $u = G_\phi(x_0)$ is defined at $x_0 \in S^n$, then the unique minimizer of $c(x, u) + \phi(x)$ with respect to x is x_0.

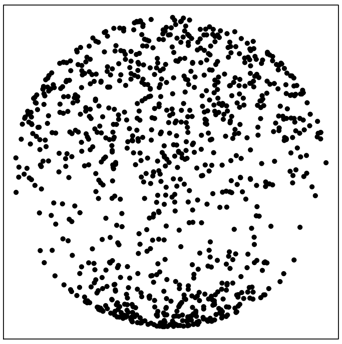
Thus our task is to solve the (deterministic) minimization problem. Although the minimization problem of $c(x, U) + \phi(x)$ is not convex in the usual sense, the objective function has no local minimum, by c-convexity. Hence the problem is efficiently solved by generic optimization packages. An example of sampling is illustrated in Figure 1.

3.2 Spherical gradient model

We consider a finite-dimensional set of probability densities on the sphere. In statistics, a finite-dimensional set of probability densities is called a *statistical model*. An unknown parameter θ that parameterizes the density functions is estimated from observed data points $x(1), \ldots, x(N) \in S^n$. One of the most important estimators is the maximum likelihood estimator that maximizes the likelihood function $\prod_{t=1}^{N} p(x(t)|\theta)$ with respect to θ.



(a) A density function on the sphere.



(b) Samples.

Figure 1: Exact sampling. (a) a density function (a white region indicates high density) and (b) 2000 sampled data points. The c-convex function used is $\phi(x) = 0.5\cos(2d(x, e_1)) + 0.5\cos(3d(x, e_2))$ for $x \in S^2$, where e_1 and e_2 denote unit vectors along the horizontal and vertical axes, respectively. See Subsection 3.3 for details. Only the northern hemisphere is drawn. The number of points on the northern hemisphere was 967 in this experiment. The program code was written in R and the computational time for sampling was about ten seconds.

We construct a new statistical model using c-convex functions. Recall that the set $W(S^n)$ of wrapping potential functions is a convex space (Lemma 2). We can consider a finite-dimensional subspace as follows. Let $\phi_{(i)} \in W(S^n)$ for $i = 1, \ldots, p$. Define

$$\phi_\theta(x) \;=\; \sum_{i=1}^{p} \theta_i \phi_{(i)}(x),$$

where $\theta = (\theta_i)_{i=1}^p$ ranges over a convex subset Θ of $\mathbb{R}^p$ such that $\phi_\theta \in W(S^n)$ for any $\theta \in \Theta$. By Lemma 2 and the elementary fact that $0 \in W(S^n)$, we can use the simplex $\{\theta \mid \theta_i \geq 0, \sum_{i=1}^p \theta_i \leq 1\}$ as Θ. Let $p(x|\theta)$ be the probability density function induced by $\phi_\theta(x) \in W(S^n)$, that is,

$$p(x|\theta) \;=\; \mathrm{Jac}(G_{\phi_\theta}(x)). \tag{2}$$

We call the family (2) *the spherical gradient model.*

The maximum likelihood estimator for the spherical gradient model (2) is reasonably computed by the following corollary of Theorem 1.

Corollary 1. Define $p(x|\theta)$ by (2). Then, for any data points $x(1), \ldots, x(N) \in S^n$, the likelihood function $\prod_{k=1}^N p(x(k)|\theta)$ is log-concave with respect to θ.

3.3 Examples

We give some examples of the spherical gradient model (2). Recall $d(x, y)$ denotes the length between x and y on S^n. All the examples are combinations of *rotationally symmetric* functions $f(d(x, z))$, where $z \in S^n$ and $f \in C^2([0, \pi])$. The k-th derivative of f is denoted by $f^{(k)}$. The following lemma is fundamental.

Lemma 4. Assume that $f^{(1)}(0) = f^{(1)}(\pi) = 0$ and $f^{(2)}(r) > -1$ for almost all $r \in [0, \pi]$. Then for each $z \in S^n$ the function $f(d(x, z))$ of x is in $W(S^n)$.

Let $\mathcal{F}$ be the set of functions on $[0, \pi]$ that satisfy the assumption in Lemma 4. Choose p pairs $\{(f_i, z_i)\}_{i=1}^p$ from $\mathcal{F} \times S^n$. Then we can define the spherical gradient model (2) with

$$\phi_\theta(x) = \sum_{i=1}^{p} \theta_i f_i(d(x, z_i)) \quad \theta = (\theta_i)_{i=1}^p \in \Theta, \tag{3}$$

where Θ is a convex subset of $\mathbb{R}^p$ such that $\phi_\theta \in W(S^n)$ for all $\theta \in \Theta$.

Remark 2. If $p = 1$, the resultant density $p(x|\theta)$ is a function of $d(x, z)$ for some $z \in S^n$. In directional statistics, such a probability density function is called rotationally symmetric. □

We briefly touch on known distributions on the sphere in statistics. A very well-known distribution on the sphere is the *von Mises-Fisher distribution* defined by

$$p(x|\mu) = \left(\frac{|\mu|}{2}\right)^{(n+1)/2} \frac{1}{\Gamma((n+1)/2) I_{(n+1)/2-1}(|\mu|)} \exp(\mu^\top x) \tag{4}$$

in Euclidean coordinates of $\mathbb{R}^{n+1}$, where $\mu \in \mathbb{R}^{n+1}$ and I_ν denotes the modified Bessel function of the first kind and order ν. A more general distribution is the *Fisher-Bingham distribution* defined by

$$p(x|\mu, A) = \frac{1}{a(\mu, A)} \exp\left(\mu^\top x + x^\top A x\right), \tag{5}$$

where $a(\mu, A)$ is a normalizing factor to ensure that $\int p(x|\mu, A)dx = 1$. We remark that (5) can also be written as a function of $d(x, z_i) = \cos^{-1}(x^\top z_i)$ for a finite number of z_i. See Mardia and Jupp (2000) for details.

We return to our spherical gradient model (2) with (3). The following explicit formula due to a general expression (11) is useful for practical implementation:

$$
\begin{aligned}
p(x|\theta) &= (\sin|v_\theta|/|v_\theta|)^{n-1}\det\left(xx^\top + H_\theta + \sum_{i=1}^p \theta_i K_i\right),\\
v_\theta &= -\sum_{i=1}^p \theta_i f_i'(\alpha_i)e_i, \quad \alpha_i = \cos^{-1}(x^\top z_i), \quad e_i = \frac{z_i - x\cos\alpha_i}{\sin\alpha_i},\\
H_\theta &= e_\theta e_\theta^\top + \frac{\alpha_\theta\cos\alpha_\theta}{\sin\alpha_\theta}\left(I - xx^\top - e_\theta e_\theta^\top\right), \quad e_\theta = v_\theta/|v_\theta|, \quad \alpha_\theta = |v_\theta|,\\
K_i &= f_i''(\alpha_i)e_i e_i^\top + \frac{f_i'(\alpha_i)\cos\alpha_i}{\sin\alpha_i}\left(I - xx^\top - e_i e_i^\top\right),
\end{aligned}
$$

where Euclidean coordinates in $\mathbb{R}^{n+1}$ are used. We remark that the above formula needs no special function, unlike the von Mises-Fisher distribution (4) or the Fisher-Bingham distribution (5).

We give examples of pairs (f_i, z_i). Recall that $W(S^n)$ is the set of all wrapping potential functions.

Example 1 (Linear potential)**.** Let $f_i(\xi) = \cos(\xi)$ for all i. We use Euclidean coordinates in $\mathbb{R}^{n+1}$ to express S^n. Then $\phi_\theta(x) = \sum_{i=1}^p \theta_i \cos(d(x, z_i)) = \sum_{i=1}^p \theta_i x^\top z_i$ is in $W(S^n)$ as long as $\sum_{i=1}^p |\theta_i| \le 1$. We deduce that a potential function $\phi_\mu(x) := \mu^\top x$ is in $W(S^n)$ if $|\mu| \le 1$. The parameter μ determines the direction and magnitude of concentration. That is, the resultant density function takes larger values at x when $-\mu/|\mu|$ is closer to x and $|\mu|$ is larger, where the negative sign of $-\mu/|\mu|$ is needed because our model is defined by the pull-back measure. We call ϕ_μ the linear potential and the resultant statistical model *the linear-potential model.* This model is rotationally-symmetric (see Remark 2). An example is given in Figure 2 (a).

Example 2 (Quadratic potential)**.** Consider $f_i(\xi) = \cos(\xi)$ for $i = 1, \ldots, p_1$ and $f_i(\xi) = \cos(2\xi)/4$ for $i = p_1 + 1, \ldots, p$. Then the potential can be written as

$$
\phi_\theta(x) = \sum_{i=1}^{p_1} \theta_i x^\top z_i + \sum_{i=p_1+1}^{p} \frac{\theta_i}{4}\left\{2(x^\top z_i)^2 - 1\right\}.
$$

Let $\mu \in \mathbb{R}^{n+1}$ and $A \in \mathrm{Sym}(\mathbb{R}^{n+1})$. Let $|A|_1$ denote the trace norm of A defined by the sum of absolute eigenvalues of A. This is actually a norm because $|A|_1 = \max_{-I\preceq B\preceq I}\mathrm{tr}[AB]$. Then we deduce that a potential function

$$
\phi_{\mu,A}(x) = x^\top\mu + \frac{1}{2}x^\top A x \tag{6}
$$

is in $W(S^n)$ if (μ, A) satisfies $|\mu| + |A|_1 \le 1$. We call the model *the quadratic-potential model.* Various numerical examples of the quadratic potential model are given in

Figure 2. Note that the representation of A includes redundancy because $x^\top x = 1$. It will be tractable if one sets $\mathrm{tr}A = 0$. However, in general this restriction strictly reduces the size of the set. For example, the matrix $A = \mathrm{diag}(0.2, 0, -0.8)$ has norm $|A|_1 = 1$ but the trace-adjusted one $B = \mathrm{diag}(0.4, 0.2, -0.6)$ has $|B|_1 = 1.2 > 1$.

Example 3 (High-frequency potential)**.** As a generalization of the above examples, we consider $f_i(\xi) = k_i^{-2}\cos(k_i\xi)$ for a positive integer k_i. If $Z = (z_1, \ldots, z_p) \in (S^n)^p$ and $K = (k_1, \ldots, k_p) \in \mathbb{Z}_{>0}^p$ are given, we obtain a potential

$$\phi_\theta(x) \;=\; \sum_{i=1}^{p} \theta_i k_i^{-2}\cos(k_i d(x, z_i)). \tag{7}$$

We call this model the *high-frequency model*. Various numerical examples of the high-frequency model are given in Figure 3. The density function used in Figure 1 belongs to this class.

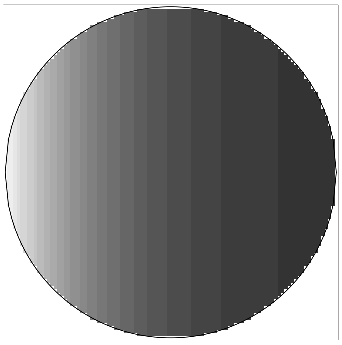
3.4 An actual data set

Here we give a brief analysis of some astronomical data. The data consist of the locations of 188 stars of magnitude brighter than or equal to 3.0. The data is available from the Bright Star Catalog (5th Revised Ed.) distributed from the Astronomical Data Center. We simply compare the quadratic model and the null model (uniform distribution) by using Akaike's Information Criterion (AIC). In general, AIC for a statistical model is defined by the sum of (-2) times the maximum log-likelihood and 2 times the parameter dimension. It is recommended to select the statistical model minimizing AIC from a set of candidates. See Akaike (1974) for details of AIC.

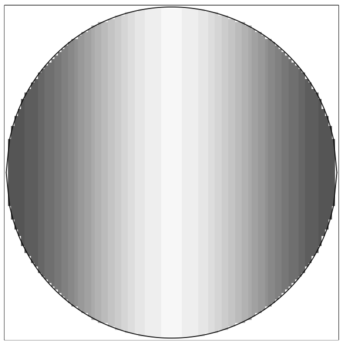
The estimated parameter for the quadratic model is

$$\hat{\mu} \;=\; (0.010, 0.017, 0.091)^\top \quad \text{and} \quad \hat{A} \;=\; 0.173\hat{z}_1\hat{z}_1^\top - 0.250\hat{z}_2\hat{z}_2^\top,$$

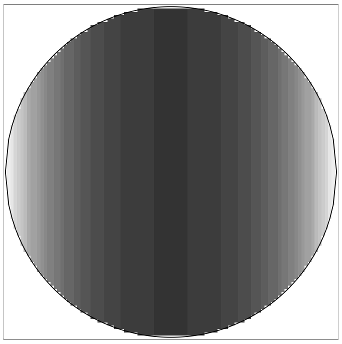
where $\hat{z}_1 = (0.731, 0.048, -0.681)^\top$ and $\hat{z}_2 = (0.544, 0.562, 0.623)^\top$. The maximum log-likelihood is 12.5. Since the number of unknown parameters is 8, AIC is -9.0. On the other hand, the likelihood of the null model (uniform distribution) is zero and AIC is also zero. Therefore, we select the quadratic model from the two candidates. Figure 4 shows the observed data and the estimated density.



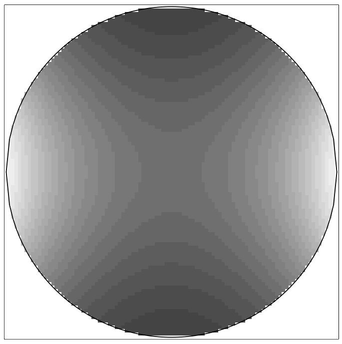
(a) $\mu = e_1$.



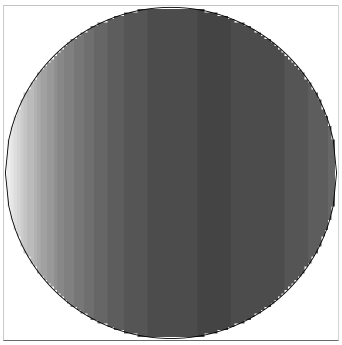
(b) $A = e_1 e_1^\top$.



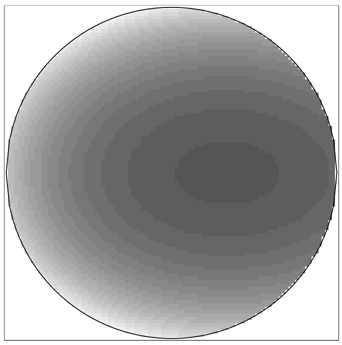
(c) $A = -e_1 e_1^\top$.



(d) $A = -0.5 e_1 e_1^\top + 0.5 e_2 e_2^\top$.

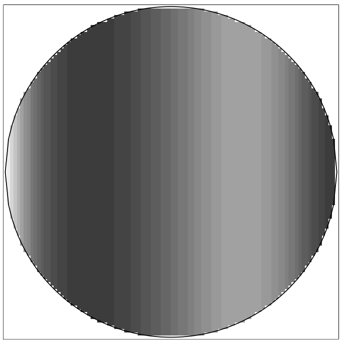


(e) $\mu = 0.5 e_1, A = -0.5 e_1 e_1^\top$.

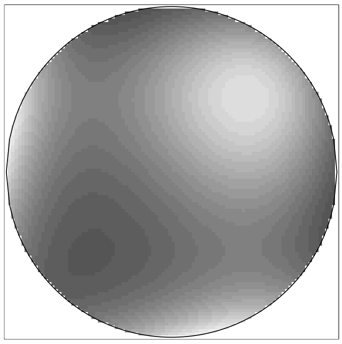


(f) $\mu = e_1/3, A = (-e_2 e_2^\top + e_3 e_3^\top)/3$.

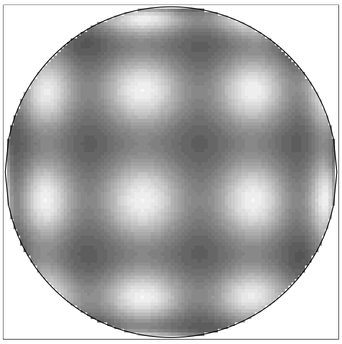
Figure 2: The quadratic-potential model. The white regions indicate high density. The figures represent (a) Concentration, (b) Negative dipole, (c) Positive dipole, (d) Complementary dipoles, (e) Unbalanced dipole and (f) General.



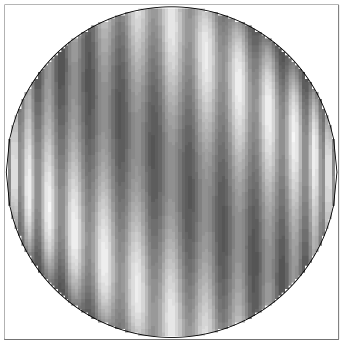
(a) $Z = (e_1), k = (3), \theta = (1)$.



(b) $Z = (e_1, e_2), K = (3, 3), \theta = (0.5, 0.5)$.

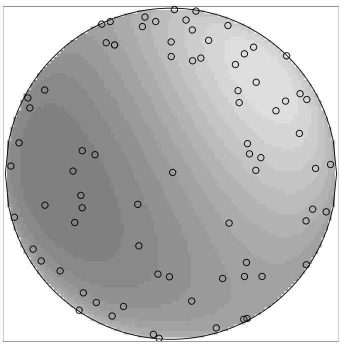


(c) $Z = (e_1, e_2), K = (9, 9), \theta = (0.5, 0.5)$.

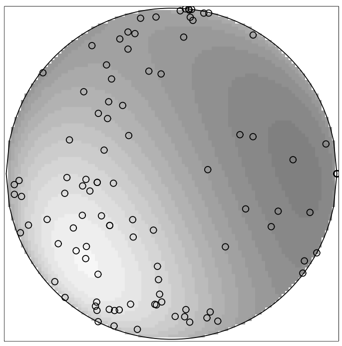


(d) $Z = (e_1, (e_1 + e_2)/\sqrt{2})$,
$K = (30, 4), \theta = (0.5, 0.5)$.

Figure 3: High-frequency spherical gradient models. White regions indicate high density.



(a) Northern hemisphere.



(b) Southern hemisphere.

Figure 4: The observed data points and the estimated density for the astronomic data. Both hemispheres are viewed from the northern side. White regions indicate high density. The points are the observed data.

4 Proofs

4.1 Proofs of Lemma 1

We use the following lemma due to Proposition 6 of McCann (2001). The lemma can also proved by direct calculation for the sphere. Recall $c(x, y) = d(x, y)^2/2$.

Lemma 5 (Inverse of the exponential map)**.** Let $x, y \in S^n$ and assume $d(x, y) < \pi$. Then $\nabla_x c(x, y) = -\exp_x^{-1}(y)$, where ∇_x denotes the gradient operator with respect to x.

We first recall the cross-curvature non-negativity and the time-convex-sliding-mountain property of the sphere established by Kim and McCann (2008). For simplicity, the definitions below are specialized for the sphere. For a given triplet $(x, y, z) \in (S^n)^3$ with $d(x, z) < \pi$ and $d(y, z) < \pi$, the curve

$$\{\exp_z((1-t)\exp_z^{-1}(x) + t\exp_z^{-1}(y)) \mid t \in [0, 1]\}$$

is called a *c-segment* connecting x and y with respect to z. We denote the c-segment by $[x, y]_t(z)$ in this paper. For given $x, y \in S^n$ with $d(x, y) < \pi$, let σ_s and τ_t be smooth curves such that $\sigma_0 = x$ and $\tau_0 = y$. We assume that either $\sigma_s = [\sigma_0, \sigma_1]_s(y)$ or $\tau_t = [\tau_0, \tau_1]_t(x)$. Note that only one of the two curves is assumed to be a c-segment. Then *the cross-curvature* $\mathcal{S}$ is well defined by

$$\mathcal{S}(x, y)(\xi, \eta) = -\frac{d^2}{ds^2}\frac{d^2}{dt^2}c(\sigma_s, \tau_t)\bigg|_{s=0,t=0},$$

where $\xi = d\sigma_s/ds|_{s=0}$ and $\eta = d\tau_t/dt|_{t=0}$. For a given quadruplet $(x, z, y_0, y_1) \in (S^n)^4$, *the sliding mountain* is defined by a function

$$t \mapsto c(z, [y_0, y_1]_t(z)) - c(x, [y_0, y_1]_t(z)). \tag{8}$$

We use the following fact proved by Kim and McCann (2008).

Lemma 6 (Cross-curvature non-negativity)**.** For the sphere, the cross-curvature $\mathcal{S}(x, y)(\xi, \eta)$ is non-negative for any (x, y, ξ, η) with $d(x, y) < \pi$.

Although the following lemma is essentially due to Kim and McCann (2008), we derive it from Lemma 6 for completeness.

Lemma 7 (Time-convex-sliding-mountain)**.** Let z be a point in S^n and let y_0 and y_1 be two points in S^n different from the antipodal point of z. Then for any $x \in S^n$ the sliding-mountain (8) is convex with respect to $t \in [0, 1]$.

Proof. Denote the sliding-mountain (8) by $f(t)$. Fix $t \in (0,1)$ and denote $y = [y_0, y_1]_t(z)$ for simplicity. We first assume y is not the antipodal point of x and prove $d^2 f(t)/dt^2 \geq 0$. Let $\sigma_s = [z,x]_s(y)$ and $\tau_u = [y, y_1]_u(z)$. Note that σ_s is a c-segment with respect to $\tau_0 = y$. Then from Lemma 6, we have

$$-\frac{d^2}{ds^2}\frac{d^2}{du^2}c(\sigma_s, \tau_u)\bigg|_{u=0} \geq 0 \tag{9}$$

for each $s \in [0,1]$. On the other hand, by Lemma 5,

$$\begin{aligned} \frac{d}{ds}c(\sigma_s, \tau_u)\bigg|_{s=0} &= -\langle \xi, \exp_z^{-1}(\tau_u)\rangle \\ &= -\langle \xi, (1-u)\exp_z^{-1}(y) + u\exp_z^{-1}(y_1)\rangle, \end{aligned}$$

where $\xi = (d\sigma_s/ds)|_{s=0}$ and $\langle \cdot, \cdot \rangle$ denotes the inner product on $T_z S^n$. We obtain

$$\frac{d}{ds}\frac{d^2}{du^2}c(\sigma_s, \tau_u)\bigg|_{s=0,u=0} = 0. \tag{10}$$

Integrating both sides of (9) with respect to s twice and using (10), we have

$$\left\{\frac{d^2}{du^2}c(z, \tau_u) - \frac{d^2}{du^2}c(x, \tau_u)\right\}\bigg|_{u=0} \geq 0.$$

Since $\tau_u = [y, y_1]_u(z) = [y_0, y_1]_{t+u(1-t)}(z)$, we have $d^2 f(t)/dt^2 \geq 0$. Next we assume that y is the antipodal point of x. By assumption, y is not the antipodal point of z. By direct calculation, we have

$$\lim_{s \to t+0}\frac{df(s)}{ds} - \lim_{s \to t-0}\frac{df(s)}{ds} = 2\pi\left|\frac{d[y_0, y_1]_t(z)}{dt}\right| \geq 0.$$

Therefore $f(t)$ is convex over $t \in [0,1]$. □

We apply Lemma 7 to prove Lemma 1 as follows:

Proof of Lemma 1. For any point $x \in S^n$, we denote the antipodal point of x by x'. Since ϕ_0 and ϕ_1 are c-convex, there exist functions ϕ_0^c and ϕ_1^c such that $\phi_i(x) = \sup_y \{-c(x,y) - \phi_i^c(y)\}$ $(i = 0, 1)$. Then we have

$$\begin{aligned} \phi_t(x) &= (1-t)\phi_0(x) + t\phi_1(x) \\ &= \sup_{y_0}\{-(1-t)c(x,y_0) - (1-t)\phi_0^c(y_0)\} + \sup_{y_1}\{-tc(x,y_1) - t\phi_1^c(y_1)\} \\ &= \sup_{y_0 \neq x'}\sup_{y_1 \neq x'}\{-(1-t)c(x,y_0) - tc(x,y_1) - (1-t)\phi_0^c(y_0) - t\phi_1^c(y_1)\}, \end{aligned}$$

where the last equality follows from continuity of c and ϕ_i^c $(i = 0, 1)$. Now we consider a c-segment $[y_0, y_1]_t(z)$ and denote it by $y_t(z)$ for simplicity. From Lemma 7, we have

$$\begin{aligned} & -(1-t)c(x, y_0) - tc(x, y_1) \\ & \quad = \sup_{z \neq y_0', y_1'} \left\{ -c(x, y_t(z)) + c(z, y_t(z)) - (1-t)c(z, y_0) - tc(z, y_1) \right\}, \end{aligned}$$

where the supremum of the right hand side is attained at $z = x$. Hence

$$\begin{aligned} \phi_t(x) &= \sup_{y_0, y_1 \neq x'} \sup_{z \neq y_0', y_1'} \{ -c(x, y_t(z)) + c(z, y_t(z)) - (1-t)c(z, y_0) - tc(z, y_1) \\ & \qquad -(1-t)\phi_0^c(y_0) - t\phi_1^c(y_1) \} \\ &= \sup_{w} \left\{ -c(x, w) - \xi(w) \right\}, \end{aligned}$$

where ξ is defined by an infimum convolution

$$\begin{aligned} \xi(w) &:= \inf_{(y_0, y_1, z) | y_0, y_1 \neq x', y_t(z) = w} \{ -c(z, w) + (1-t)c(z, y_0) + tc(z, y_1) \\ & \qquad +(1-t)\phi_0^c(y_0) + t\phi_1^c(y_1) \}. \end{aligned}$$

Since ϕ_t is written in the form of a c-transform, it is c-convex. □

4.2 Proof of Theorem 1

For each c-convex function ϕ, let $\Omega(\phi)$ be the set of points x such that $|\nabla\phi(x)| < \pi$ and ϕ has its Hessian defined at x. If ϕ is a wrapping potential function (Definition 1), then $S^n \setminus \Omega(\phi)$ consists only of a finite set of points.

Lemma 8. If ϕ is c-convex, then $|\nabla\phi(x)| < \pi$ except for at most one $x \in \Omega(\phi)$. Furthermore, if $|\nabla\phi(x)| \geq \pi$ for some x, then $G_\phi(y) = G_\phi(x)$ for any $y \in \Omega(\phi)$.

Proof. Let ϕ be c-convex. Assume that there exists $x \in \Omega(\phi)$ such that $|\nabla\phi(x)| \geq \pi$. In general, any c-convex function on a compact Riemannian manifold is Lipschitz continuous with Lipschitz constant less than or equal to the diameter of the manifold (Lemma 2 of McCann (2001)). Since the diameter of the sphere S^n is π, we have $|\nabla\phi(x)| = \pi$. Hence $G_\phi(x)$ is the antipodal point x' of x. We now prove that $G_\phi(y) = x'$ for all $y \in \Omega(\phi)$. We use 2-monotonicity of the gradient map:

$$d^2(x, G_\phi(x)) + d^2(y, G_\phi(y)) \leq d^2(x, G_\phi(y)) + d^2(y, G_\phi(x)),$$

where $d(x, y)$ is the distance between x and y. The above inequality follows from Lemma 3. Let $a = d(x, G_\phi(y))$, $b = d(y, x')$ and $c = d(y, G_\phi(y))$. Then we have

$\pi^2+c^2 \le a^2+b^2$. By the triangle inequality with respect to the triangle $(x, y, G_\phi(y))$, we have $c \ge |a+b-\pi|$. Therefore

$$\begin{aligned} 0 &\ge \pi^2 + c^2 - a^2 - b^2 \\ &\ge \pi^2 + (a+b-\pi)^2 - a^2 - b^2 \\ &= 2(\pi - a)(\pi - b). \end{aligned}$$

This implies $a = \pi$ or $b = \pi$; equivalently, $G_\phi(y) = x'$ or $y = x$. Hence we have $G_\phi(y) = x'$ for any $y \in \Omega(\phi)$. Then $|\nabla\phi(y)| < \pi$ for any $y \neq x$ from the definition of G_ϕ. □

We proceed to the proof of Theorem 1. Fix two c-convex functions ϕ_0 and ϕ_1 and let $\phi_t = (1-t)\phi_0 + t\phi_1$ for $t \in [0,1]$. Let $x \in \Omega(\phi_0) \cap \Omega(\phi_1)$. Then it is easy to see that $x \in \Omega(\phi_t)$ for any $t \in [0,1]$. Recall that the gradient map of ϕ_t is denoted by $F_t(x) = \exp_x(\nabla\phi_t(x))$. Note that $F_t(x)$ is a c-segment $[F_0(x), F_1(x)]_t(x)$. We prepare some notation to represent an explicit formula of the Jacobian determinant of $F_t(x)$. Let $\sigma_t(x)$ be the Jacobian determinant of the exponential map at $\nabla\phi_t(x)$, i.e. $\sigma_t(x) = \det\{d(\exp_x(v))/dv\}|_{v=\nabla\phi_t(x)}$, where the determinant is calculated with respect to any orthonormal bases. Denote the Hessian operator at x by Hess_x and let $H_t(x) = (\mathrm{Hess}_x c(x,y))_{y=F_t(x)}$. Then, by Cordero-Erausquin et al. (2001), the Jacobian determinant of $F_t(x)$ is

$$J_t(x) = \sigma_t(x) \det\left(H_t(x) + \mathrm{Hess}_x\phi_t\right). \tag{11}$$

Lemma 9. Let $x \in \Omega(\phi_0) \cap \Omega(\phi_1)$. The matrix-valued function $H_t(x)$ is concave with respect to t:

$$H_t(x) \succeq (1-t)H_0(x) + tH_1(x) \quad \text{for any } t \in [0,1],$$

where $A \succeq B$ means that $A - B$ is non-negative definite.

Proof. Since $F_t(x)$ is a c-segment $[F_0(x), F_1(x)]_t(x)$, Lemma 7 implies that

$$\begin{aligned} &c(w, F_t(x)) - c(x, F_t(x)) \\ &\quad \ge (1-t)\{c(w, F_0(x)) - c(x, F_0(x))\} + t\{c(w, F_1(x)) - c(x, F_1(x))\} \end{aligned}$$

for all $w \in S^n$. By taking the Hessian with respect to w at $w = x$, we obtain

$$\mathrm{Hess}_w c(w, F_t(x))|_{w=x} \succeq (1-t)\mathrm{Hess}_w c(w, F_0(x))|_{w=x} + t\mathrm{Hess}_w c(w, F_1(x))|_{w=x}.$$

This means $H_t(x) \succeq (1-t)H_0(x) + tH_1(x)$. □

Lemma 10 (Jacobian-ratio inequality)**.** Let $x \in \Omega(\phi_0) \cap \Omega(\phi_1)$. Then the following inequality holds:

$$\left(\frac{J_t(x)}{\sigma_t(x)}\right)^{1/n} \geq (1-t)\left(\frac{J_0(x)}{\sigma_0(x)}\right)^{1/n} + t\left(\frac{J_1(x)}{\sigma_1(x)}\right)^{1/n}. \tag{12}$$

Proof. By the formula (11), it is sufficient to prove that $\det^{1/n}(H_t + \mathrm{Hess}_x \phi_t)$ is concave with respect to t. Indeed, by Lemma 9 and the geometric-arithmetic inequality on $\det^{1/n}$, we obtain

$$\begin{aligned}
&\det{}^{1/n}(H_t + \mathrm{Hess}_x \phi_t) \\
&\geq \det{}^{1/n}\{(1-t)H_0 + tH_1 + \mathrm{Hess}_x \phi_t\} \\
&= \det{}^{1/n}\{(1-t)(H_0 + \mathrm{Hess}_x \phi_0) + t(H_1 + \mathrm{Hess}_x \phi_1)\} \\
&\geq (1-t)\det{}^{1/n}(H_0 + \mathrm{Hess}_x \phi_0) + t\det{}^{1/n}(H_1 + \mathrm{Hess}_x \phi_1).
\end{aligned}$$

Hence $\det^{1/n}(H_t + \mathrm{Hess}_x \phi_t)$ is concave. □

Remark 3. If $\phi_0 \equiv 0$, the inequality (12) is similar to the Jacobian inequality, due to Cordero-Erausquin et al. (2001). They showed that if $\phi_0 \equiv 0$,

$$J_t(x)^{1/n} \geq (1-t)v_{1-t}(F_1(x), x)^{1/n} + t[v_t(x, F_1(x))]^{1/n} J_1(x)^{1/n}, \tag{13}$$

where $v_t(x, y)$ denotes the volume distortion coefficient (see Cordero-Erausquin et al. 2001 for details). The inequality (13) is crucial to prove a Brunn-Minkowskii-type inequality on manifolds. However, since the inequality (13) is only established for the special case $\phi_0 \equiv 0$, it is not sufficient for our statistical application. Unfortunately, (13) is not implied from (12). In fact, if $\phi_0(x) \equiv 0$, then $J_0(x) = 1$ and $\sigma_0(x) = 1$, and the inequality (12) reduces to

$$J_t(x)^{1/n} \geq (1-t)\sigma_t(x)^{1/n} + t\left(\frac{\sigma_t(x)}{\sigma_1(x)}\right)^{1/n} J_1(x)^{1/n}.$$

This inequality is weaker than (13) because $v_{1-t}(F_1(x), x) > 1 > \sigma_t(x)$ and $v_t(x, F_1(x)) = \sigma_t(x)/\sigma_1(x)$. □

Lemma 11. For any $x \in \Omega(\phi_0) \cap \Omega(\phi_1)$, $\log \sigma_t(x)$ is concave with respect to t.

Proof. For the unit sphere S^n, the Jacobian determinant of the exponential map is given by $(\sin|v|/|v|)^{n-1}$. Therefore $\sigma_t(x) = (\sin|\nabla\phi_t(x)|/|\nabla\phi_t(x)|)^{n-1}$. Since the function $[0, \pi] \ni \rho \mapsto \log(\sin\rho/\rho)$ is decreasing and concave, and since the map $t \mapsto |\nabla\phi_t(x)|$ is convex with respect to t, we deduce that the composite map $\log \sigma_t(x) = \log(\sin|\nabla\phi_t(x)|/|\nabla\phi_t(x)|)$ is concave. □

Proof of Theorem 1. By Lemma 10 and Lemma 11, the functions $\log(J_t(x)/\sigma_t(x))$ and $\log \sigma_t(x)$ are concave with respect to t. Hence $\log J_t(x)$ is also concave. □

4.3 Proof of Lemma 2

Recall that $W(S^n)$ is the set of c-convex functions ϕ such that the gradient map G_ϕ is an isomorphism on S^n and ϕ has its Hessian defined everywhere except at a finite set of points.

Lemma 12. Let ϕ be a c-convex function and differentiable. Then G_ϕ is injective if and only if $c(x, G_\phi(x)) + c(z, G_\phi(z)) < c(x, G_\phi(z)) + c(z, G_\phi(x))$ for any $x \neq z$.

Proof. In general, by Lemma 3, 2-monotonicity

$$c(x, G_\phi(x)) + c(z, G_\phi(z)) \ \leq\ c(z, G_\phi(x)) + c(x, G_\phi(z))$$

holds for any x and z, where equality holds if and only if $G_\phi(x) = G_\phi(z)$. The result follows immediately. □

Lemma 13. Let ϕ_0 and ϕ_1 be members of $W(S^n)$. Then, for any $t \in [0, 1]$, the gradient map $F_t(x) = \exp_x(\nabla\phi_t(x))$ is injective.

Proof. Put $h_t(x, z) = c(x, F_t(x)) + c(z, F_t(z)) - c(x, F_t(z)) - c(z, F_t(x))$. By Lemma 12, it is sufficient to show that $h_t(x, z) < 0$ for any $t \in [0, 1]$ and $x \neq z$. By the assumption and Lemma 12, we have $h_0(x, z) < 0$ and $h_1(x, z) < 0$. On the other hand, by Lemma 7, $h_t(x, z) \leq (1 - t)h_0(x, z) + th_1(x, z)$. Hence we obtain $h_t(x, z) < 0$. □

Proof of Lemma 2. Assume that ϕ_0 and ϕ_1 are members of $W(S^n)$. From Lemma 13, F_t is injective. On the other hand, Lemma 8 implies that $|\nabla\phi_0(x)| < \pi$ and $|\nabla\phi_1(x)| < \pi$ for all $x \in S^n$. Then $\nabla\phi_0(x) = \exp_x^{-1}(F_0(x))$ and $\nabla\phi_1(x) = \exp_x^{-1}(F_1(x))$ are continuous. This implies F_t is continuous. Hence, by compactness and connectedness of S^n, F_t must be an isomorphism on S^n. Twice differentiability of ϕ_t follows immediately from that of ϕ_0 and ϕ_1. □

4.4 Proof of Lemma 4

Fix z and let $\phi(x) = f(d(x, z))$, for simplicity. We first prove c-convexity of ϕ. It is sufficient to show that for each $x_0 \in S^n$ there exists some $y \in S^n$ such that $c(x_0, y) + \phi(x_0) = \inf_{x \in S^n}\{c(x, y) + \phi(x)\}$. Thus we investigate the point minimizing $c(x, y) + \phi(x)$ for each fixed y. Denote the antipodal points of y and z by y' and z', respectively. If x is different from y', then the gradient vector of $c(x, y) + \phi(x)$ with respect to x is

$$\nabla_x\{c(x, y) + \phi(x)\} \ =\ \nabla_x c(x, y) + \nabla_x c(x, z)\frac{f^{(1)}(\sqrt{2c(x, z)})}{\sqrt{2c(x, z)}},$$

where ∇_x denotes the gradient operator with respect to x. Note that the above expression makes sense for $x = z$ and $x = z'$ because $f^{(1)}(0) = f^{(1)}(\pi) = 0$ and $f \in C^2([0, \pi])$. By Lemma 5, we know that $\nabla_x c(x, y) = -\exp_x^{-1}(y)$ and $\nabla_x c(x, z) = -\exp_x^{-1}(z)$. Hence the gradient vector $\nabla_x \{c(x, y) + \phi(x)\}$ vanishes only if x lies on a great circle C that passes through y and z. Since the exceptional point y' is also included in C, we deduce that the point minimizing $c(x, y) + \phi(x)$ must belong to C. We fix a circular coordinate $\xi \in (-\pi, \pi]$ representing a point on C such that y corresponds to $\xi = 0$. Let ξ and ζ be the coordinates of x and z. We assume $\zeta \in [0, \pi]$ without loss of generality. Then the function $c(x, y) + \phi(x)$ can be written as

$$h(\xi) \ := \ c(x, y) + \phi(x) \ = \ \frac{\xi^2}{2} + f(\min\{|\xi - \zeta|, |\xi - \zeta + 2\pi|\}).$$

By the assumption for f, one can easily check that the second derivative of h is $h^{(2)}(\xi) \geq 0$ (> 0 a.e.) as long as $\xi \neq \pi$. Furthermore, we obtain $h^{(1)}(\pi - 0) > h^{(1)}(-\pi + 0)$. Thus $\xi = \pi$ is not a point minimizing h. Furthermore, the point minimizing h is unique because h is strictly convex over $(-\pi, \pi)$. We denote the minimizer by $\xi_0 \in (-\pi, \pi]$ and the corresponding point in S^n by x_0. If y revolves along a great circle C passing through z, then x_0 must continuously revolve along C. Since y can belong any great circle passing through z, we deduce that for each point x_0 there exists some $y \in C$ such that the function $c(x, y) + \phi(x)$ of $x \in S^n$ is minimized at x_0. This proves c-convexity of ϕ.

Next we prove the gradient map $G_\phi(x)$ is well defined and an isomorphism. Since ϕ is differentiable everywhere, G_ϕ is well defined. Let $x = \exp_z(te)$, where $t \in [0, \pi]$ and $e \in T_z S^n$ with $|e| = 1$. Then the gradient map is explicitly given by $G_\phi(x) = \exp_z\left((t + f^{(1)}(t))e\right)$. If t moves from 0 to π, then $t + f^{(1)}(t)$ moves from 0 to π monotonically because $1 + f^{(2)}(t) > 0$ for almost all $t \in [0, \pi]$. Hence $G_\phi : S^n \to S^n$ is an isomorphism.

Lastly, ϕ is clearly twice differentiable whenever $x \neq z$ and $x \neq z'$. This completes the proof.

5 Discussion

We briefly discuss the Jacobian inequality for general manifolds.

In the proof of Theorem 1, we have used the closed property of cost-convex functions (Lemma 1), the Jacobian-ratio inequality (Lemma 10) and log-concavity of the Jacobian of the exponential map (Lemma 11). For any non-negatively cross-curved (or time-convex-sliding-mountain) manifold defined in Kim and McCann (2008), the former two lemmas are obtained in the same manner. However, Lemma 11 does not automatically follow from the non-negative cross-curvature condition.

The author does not know if any Riemannian manifold with non-negative cross-curvature satisfies the Jacobian inequality. At least, any product space of S^n and $\mathbb{R}^n$ satisfies the Jacobian inequality because the non-negative cross-curvature condition is preserved for products of manifolds (Kim and McCann (2008)) and the Jacobian determinant of the exponential map is also factorized into the Jacobian determinant on each space. This fact may enable us to describe dependency structures of multivariate directional data in statistics. We leave such an extension for future research.

Acknowledgements

This study was partially supported by the Global Center of Excellence "The research and training center for new development in mathematics" and by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Young Scientists (B), No. 19700258.

References

Akaike, H.: A new look at the statistical model identification. IEEE Trans. Automatic Control **19**(6), 716–723 (1974)

Brenier, Y.: Polar factorization and monotone rearrangement of vector-valued functions. Comm. Pure Appl. Math. **44**, 375–417 (1991)

Cordero-Erausquin, D., McCann, R.J., Schumuckenshläger, M.: A Riemannian interpolation inequality à la Borell, Brascamp and Lieb. Invent. Math. **146**, 219–257 (2001)

Delanoë, P., Loeper, G.: Gradient estimates for potentials of invertible gradient-mappings on the sphere. Calc. Var. **26**(3), 297–311 (2006)

Fisher, R.A.: Dispersion on a sphere. Proceedings of the Royal Society of London, Series A **217**, 295–305 (1953)

Kim, Y.H.: Counterexamples to continuity of optimal transportation on positively curved Riemannian manifolds (2007). Preprint. arXiv:0709.1653v1

Kim, Y.H., McCann, R.J.: Continuity, curvature, and the general covariance of optimal transportation (2007). Preprint. arXiv:0712.3077v1

Kim, Y.H., McCann, R.J.: Towards the smoothness of optimal maps on Riemmanian submersions and Riemannian products (of round spheres in particular) (2008). Preprint. arXiv:0806.0351v1

Loeper, G.: On the regularity of maps solutions of optimal transportation problems (2005). Preprint. arXiv:0504137v2

Ma, X.N., Trudinger, N.S., Wang, X.J.: Regularity of potential functions of the optimal transportation problem. Arch. Rational Mech. Anal. **177**, 151–183 (2005)

Mardia, K.V., Jupp, P.E.: Directional Statistics. John Wiley and Sons, Chichester (2000)

McCann, R.J.: Polar factorization of maps on Riemannian manifolds. Geom. Funct. Anal. **11**(3), 589–608 (2001)

Sei, T.: Parametric modeling based on the gradient maps of convex functions. Tech. rep., METR2006-51, Department of Mathematical Engineering, University of Tokyo (2006)

Villani, C.: Optimal Transport - Old and New. Springer, Berlin (2009)