MATHEMATICAL ENGINEERING TECHNICAL REPORTS

Scoring Rules for Statistical Models on Spheres

Yuya TAKASU, Keisuke YANO,
and Fumiyasu KOMAKI

METR 2017–09 March 2017

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

Scoring Rules for Statistical Models on Spheres

Yuya TAKASU [1], Keisuke YANO [1] and Fumiyasu KOMAKI [1,2]

[1]Department of Mathematical Informatics

Graduate School of Information Science and Technology

The University of Tokyo, Tokyo, JAPAN

[2]RIKEN Brain Science Institute, Wako-shi, JAPAN

{yuya_takasu,yano,komaki}@mist.i.u-tokyo.ac.jp

Abstract

Parameter estimation using the maximum likelihood method is often difficult since normalizing constants do not have explicit forms. We propose a new class of strictly proper scoring rules for statistical models on spheres that does not require the calculation of normalizing constants. The construction of the proposed class is based on divergence functions. We investigate orthogonally-invariant scoring rules in the proposed class. We show through numerical experiments that the proposed scoring rules work well.

1 Introduction

We consider statistical inference for parametric models on an n-dimensional unit sphere

$$\mathcal{X} := \left\{ x = (x^1, \dots, x^{n+1})^\top \in \mathbb{R}^{n+1} : (x^1)^2 + \cdots + (x^{n+1})^2 = 1 \right\}.$$

Let $\mathcal{M}$ be a parametric statistical model on $\mathcal{X}$. We assume that each element in $\mathcal{M}$ has a strictly positive and twice continuously differentiable probability density p with respect to the uniform measure μ.

Statistical inference on spheres has gathered much attention not only in directional statistics [13] but also in machine learning. For example, see [7] for context analysis, [14] for visual learning, [11] for genomic analysis, and [17] for morphometrics.

Using the maximum likelihood method to estimate parameters of statistical models on spheres is often difficult. Suppose that $\mathcal{M}$ is parametrized as $\{p(\cdot;\theta) = \tilde{p}(\cdot;\theta)/c(\theta) : \theta \in \Theta\}$ with $\Theta \subset \mathbb{R}^d$ and $d \in \mathbb{N}$. The normalizing constant $c(\theta)$ of $p(\cdot;\theta)$ often does not have an explicit form. A typical example of a distribution on $\mathcal{X}$ whose normalizing constant is difficult to represent explicitly is the Fisher–Bingham distribution [13]. To obtain maximum likelihood estimates for the Fisher–Bingham model, the saddle-point approximation [12] and the holonomic gradient method [15] have been proposed.

Instead of using the maximum likelihood method, we consider parameter estimation based on proper, 2-local, and homogeneous scoring rules. A scoring rule S is a loss function $S(x, Q) : \mathcal{X} \times \mathcal{M} \to \mathbb{R} \cup \{\infty\}$ that measures the quality of a distribution Q

as an estimate of the distribution of a random variable X on $\mathcal{X}$ when the realized value of X is x. It is said to be proper if the expected score $\int S(x,Q)\mathrm{d}P(x)$ is minimized at $Q = P$ for arbitrary $P \in \mathcal{M}$ and is said to be strictly proper if the minimizer is unique. If $\mathcal{M}$ is parametrized by θ, based on samples $x_1, x_2, \ldots, x_T$ and a proper scoring rule S, we estimate θ by

$$\hat{\theta}(x_1, \ldots, x_T) \in \underset{\theta\in\Theta}{\operatorname{argmin}} \sum_{t=1}^{T} S(x_t, Q_\theta).$$

A 2-local scoring rule S is a scoring rule represented by

$$S(x,Q) = s(x, q(x), \nabla\tilde{q}(x), \nabla^2\tilde{q}(x))$$

with $s : \mathcal{X} \times \mathbb{R}_+ \times \mathbb{R}^{n+1} \times \mathbb{R}^{(n+1)\times(n+1)} \to \mathbb{R} \cup \{\infty\}$ for all $x \in \mathcal{X}$ and all $Q \in \mathcal{M}$, where $q(x)$ is a probability density of $Q \in \mathcal{M}$ with respect to the uniform probability measure μ, $\tilde{q}$ is an extension of q to a function on $\mathbb{R}^{n+1} \setminus \{0\}$ such that $\tilde{q}(z) := q(z/\sqrt{z^\top z})$ for any $z \in \mathbb{R}^{n+1} \setminus \{0\}$, ∇ is the gradient operator on $\mathbb{R}^{n+1}$, and $\nabla^2\tilde{q} = \nabla(\nabla\tilde{q})^\top$. A 2-local scoring rule is said to be homogeneous if $s(x, q(x), \nabla\tilde{q}(x), \nabla^2\tilde{q}(x)) = s(x, \lambda q(x), \lambda\nabla\tilde{q}(x), \lambda\nabla^2\tilde{q}(x))$ for an arbitrary positive constant λ. To evaluate a 2-local and homogeneous scoring rule, we do not need the normalizing constant. This definition of a 2-local and homogeneous scoring rule is based on [9] and [16].

For construction of proper homogeneous scoring rules on the Euclidean space, Hyvärinen [10] proposed the Hyvärinen scoring rule, a strictly proper, 2-local and homogeneous scoring rule on the Euclidean space. Ehm and Gneiting [9] and Parry et al. [16] proposed a wide class of proper and homogeneous scoring rules on the Euclidean space.

In this paper, we introduce a useful class of strictly proper, 2-local and homogeneous scoring rules for parametric models on spheres. Focusing on the relationship between strictly proper scoring rules and divergence functions, we define divergence functions between probability distributions on $\mathcal{X}$ and construct the class including a scoring rule corresponding to the Hyvärinen scoring rule for models on the Euclidean space. Furthermore, we propose scoring rules that are invariant with respect to orthogonal transformations.

The rest of the paper is organized as follows. In Section 2, we prepare notations and the relationship between strictly proper scoring rules and divergence functions. In Section 3, we propose a class of strictly proper, 2-local and homogeneous scoring rules for parametric models on spheres. In Section 4, we investigate orthogonally-invariant scoring rules. In Section 5, we provide numerical experiments. In Section 6, we conclude the paper.

2 Preparation

2.1 Unit spheres as Riemannian manifolds

The metric tensor on a unit sphere $\mathcal{X}$ with respect to a local coordinate $(u^1, \ldots, u^n)$ is given by

$$g_{ab}(x) = \sum_{k=1}^{n+1} \frac{\partial x^k(u)}{\partial u^a} \frac{\partial x^k(u)}{\partial u^b}$$

for $a, b \in \{1, \ldots, n\}$ and for $x \in \mathcal{X}$, In this paper, we use a system of local coordinates $\{(U_\mathrm{u}, u_\mathrm{u}), (U_\mathrm{d}, u_\mathrm{d})\}$ defined by $U_\mathrm{u} = \mathcal{X} \setminus \{(0, \ldots, 0, -1)^\top\}$ and

$$u_\mathrm{u} = (u_\mathrm{u}^1, \ldots, u_\mathrm{u}^n)^\top = \left(\frac{x^1}{1 + x^{n+1}}, \ldots, \frac{x^n}{1 + x^{n+1}}\right)^\top,$$

and $U_\mathrm{d} = \mathcal{X} \setminus \{(0, \ldots, 0, 1)^\top\}$ and

$$u_\mathrm{d} = (u_\mathrm{d}^1, \ldots, u_\mathrm{d}^n)^\top = \left(\frac{x^1}{-1 + x^{n+1}}, \ldots, \frac{x^n}{-1 + x^{n+1}}\right)^\top,$$

respectively. We use the partition $\{H_\mathrm{u}, H_\mathrm{d}\}$ of $\mathcal{X}$ defined by $H_\mathrm{u} := \{x \in \mathcal{X} : x_{n+1} > 0\}$ and $H_\mathrm{d} := \{x \in \mathcal{X} : x_{n+1} \leq 0\}$.

Throughout the paper, we use the Einstein summation convention: if the same index appears in an upper position and in a lower position, a summation over the index is implied.

For a scalar function $h : \mathcal{X} \to \mathbb{R}$, the function $\tilde{h} : \mathbb{R}^{n+1} \setminus \{0\} \to \mathbb{R}$ denotes an extension of h such that $\tilde{h}(z) := h(z/\sqrt{z^\top z})$ for $z \in \mathbb{R}^{n+1} \setminus \{0\}$. Then, we have

$$\nabla \tilde{h}(x) = \frac{\partial h(x)}{\partial u^a} g^{ab} \frac{\partial x}{\partial u^b} \in \mathbb{R}^{n+1}, \tag{1}$$

where $g^{ab} = g^{ab}(x)$ is the (i, j)-component of the inverse matrix G^{-1} of the matrix $G = (g_{ab})$. We use this representation in the proof of Theorem 1.

2.2 Scoring rules and divergence functions

First, we give the definition of divergence functions.

Definition 1 (divergence function; see for example pp. 97–98 in [1])**.** A function $d : \mathcal{M} \times \mathcal{M} \to \mathbb{R} \cup \{\infty\}$ is said to be a divergence function if for $P, Q \in \mathcal{M}$, $d(P, Q) \geq 0$ with the equality if and only if $P = Q$.

For a scoring rule S, we define

$$d_S(P, Q) := \int S(x, Q) \mathrm{d}P(x) - \int S(x, P) \mathrm{d}P(x). \tag{2}$$

Lemma 1 (See for example [5])**.** If a scoring rule S is strictly proper, then d_S is a divergence function. If S is a scoring rule and d_S is a divergence function, then S is strictly proper.

We provide two examples of strictly proper scoring rules and the corresponding divergence functions.

Example 1 (The Bregman scoring rule; see for example [6, 16])**.** The Bregman scoring rule for a distribution Q on $\mathcal{X}$ is defined by

$$S(x, Q) = \phi'(q(x)) + \int (\phi(q(y)) - q(y)\phi'(q(y))) \mathrm{d}\mu(y),$$

where $\phi : \mathbb{R}_+ \to \mathbb{R}$ is strictly concave and differentiable and ϕ' denotes the derivative. Since

$$d_S(P, Q) = \int \{\phi\,(q(x)) - \phi\,(p(x)) + (p(x) - q(x))\,\phi'\,(q(x))\}\,\mathrm{d}\mu(x) \tag{3}$$

and ϕ is strictly concave, d_S is a divergence. Thus, by Lemma 1, the Bregman scoring rule is strictly proper. The function d_S is known as the separable Bregman divergence [4, 6, 8, 16]. However, it is neither 2-local nor homogeneous.

The following is an example of a scoring rule for parametric models on $\mathbb{R}^{n+1}$.

Example 2 (The Hyvärinen scoring rule; see [10])**.** The Hyvärinen scoring rule S for a distribution Q on $\mathbb{R}^{n+1}$ is defined by

$$S(x, Q) = \frac{1}{2}\|\nabla \log q(x)\|^2 + \Delta \log q(x),$$

where $\|\cdot\|$ is the Euclidean norm, ∇ is the gradient operator on the Euclidean space, Δ is the Laplacian on the Euclidean space, and q is a density with respect to the Lebesgue measure on $\mathbb{R}^{n+1}$. Since by the integration by parts, d_S is represented as

$$d_S(P, Q) = \frac{1}{2}\int p(x)\,\|\nabla \log q(x) - \nabla \log p(x)\|^2\,\mathrm{d}x, \tag{4}$$

d_S is a divergence. Thus, by Lemma 1, the Hyvärinen scoring rule is strictly proper. This divergence is known as the Hyvärinen divergence function [10]. The scoring rule is 2-local and homogeneous.

3 Proposed scoring rules on unit spheres

In this section, we introduce a useful class of strictly proper, 2-local, and homogeneous scoring rules for parametric models on an n-dimensional unit sphere $\mathcal{X}$.

Let f be a function $\mathcal{X} \times \mathbb{R}^{n+1} \to \mathbb{R}$ such that for each $x \in \mathcal{X}$, $z \mapsto f(x, z)$ is strictly concave and differentiable. We define

$$\begin{aligned} d_f(P, Q) = \int_{\mathcal{X}} p(x)\Big\{ & f\,(x, \nabla \log \tilde{q}(x)) - f\,(x, \nabla \log \tilde{p}(x)) \\ & - \left\langle \nabla \log \frac{\tilde{q}(x)}{\tilde{p}(x)}, (\nabla_2 f)\,(x, \nabla \log \tilde{q}(x)) \right\rangle \Big\}\mathrm{d}\mu(x) \end{aligned} \tag{5}$$

for $P, Q \in \mathcal{M}$, where $\langle \cdot, \cdot \rangle$ is the standard inner product in $\mathbb{R}^{n+1}$, $\tilde{p}$ and $\tilde{q}$ are the extensions of p and q, respectively, and

$$(\nabla_2 f)(x, z) := \left(\frac{\partial}{\partial z^1} f(x, z), \ldots, \frac{\partial}{\partial z^{n+1}} f(x, z) \right)^\top.$$

We show that the function d_f in (5) is a divergence function. Since for any $x \in \mathcal{X}$, $z \mapsto f(x, z)$ is strictly concave, we have $f(x, z_1) - f(x, z_2) > \langle z_1 - z_2, (\nabla_2 f)(x, z_1) \rangle$ for any $x \in \mathcal{X}$ and for any two distinct points $z_1, z_2 \in \mathbb{R}^{n+1}$; see p.70 in [3]. Thus, for all $P, Q \in \mathcal{M}$, $d_f(P,Q) \geq 0$ with the equality if and only if $P = Q$.

Let f be a function $\mathcal{X} \times \mathbb{R}^{n+1} \to \mathbb{R}$ such that for each $x \in \mathcal{M}$, $z \mapsto f(x,z)$ is twice continuously differentiable and such that for each $z_2 \in \mathbb{R}^{n+1}$, $z_1 \mapsto \nabla_2 \tilde{f}(z_1, z_2)$ is differentiable at any $z_1 \in \mathbb{R}^{n+1} \setminus \{0\}$. We define $S_f : \mathcal{X} \times \mathbb{R}^{n+1} \to \mathbb{R}$ by

$$\begin{aligned} S_f(x,Q) =& f\left(x, \nabla \log \tilde{q}(x)\right) - \left(\nabla \log \tilde{q}(x)\right)^\top \left(\nabla_2 f\right)\left(x, \nabla \log \tilde{q}(x)\right) \\ &- \left(\nabla_1 \cdot \nabla_2 \tilde{f}\right)\left(x, \nabla \log \tilde{q}(x)\right) \\ &- \operatorname{tr}\left(\left(\nabla_2^2 f\right)\left(x, \nabla \log \tilde{q}(x)\right) \nabla^2 \log \tilde{q}(x)\right) \\ &+ n x^\top \left(\nabla_2 f\right)\left(x, \nabla \log \tilde{q}(x)\right), \end{aligned} \tag{6}$$

where

$$(\nabla_1 \cdot \nabla_2 \tilde{f})(x,z) := \operatorname{tr}\{\nabla (\nabla_2 \tilde{f}(x,z))^\top\}.$$

Here tr is the trace of an $(n+1) \times (n+1)$ matrix.

The following theorem provides a class of strictly proper, 2-local, and homogeneous scoring rules on unit spheres.

Theorem 1. *Assume that $f : \mathcal{X} \times \mathbb{R}^{n+1} \to \mathbb{R}$ is a function such that for each $x \in \mathcal{X}$, $z \mapsto f(x,z)$ is strictly concave and twice continuously differentiable and such that for each $z_2 \in \mathbb{R}^{n+1}$, $z_1 \mapsto \nabla_2 \tilde{f}(z_1, z_2)$ is differentiable at every $z_1 \in \mathbb{R}^{n+1} \setminus \{0\}$. Then, the scoring rule S_f defined by (6) is strictly proper, 2-local, and homogeneous. The corresponding divergence function d_{S_f} defined by (2) and (6) is equal to d_f defined by* (5)*.*

Proof. By definition, S_f is 2-local and homogeneous. To prove that S_f is strictly proper, it suffices to show that $d_{S_f} = d_f$ since d_f is a divergence and Lemma 1 holds.

First, we show that $d_f = d_{S_f}$, by assuming the equality

$$\begin{aligned} &\int_{\mathcal{X}} p(x) \langle \nabla \log \tilde{p}(x), (\nabla_2 f)(x, \nabla \log \tilde{q}(x)) \rangle \mathrm{d}\mu(x) \\ &= - \sum_{\alpha \in \{\mathrm{u},\mathrm{d}\}} \int_{H_\alpha} p(x) \frac{\partial}{\partial u_\alpha^a} \left(g_\alpha^{ab} \left\langle \frac{\partial x}{\partial u_\alpha^b}, (\nabla_2 f)(x, \nabla \log \tilde{q}(x)) \right\rangle \sqrt{|G_\alpha|} \right) \\ &\quad \times \mathrm{d}u_\alpha^1 \wedge \cdots \wedge \mathrm{d}u_\alpha^n \end{aligned} \tag{7}$$

with $\{H_\mathrm{u}, H_\mathrm{d}\}$ and $\{u_\mathrm{u}, u_\mathrm{d}\}$ defined in Section 2, where for $\alpha \in \{\mathrm{u}, \mathrm{d}\}$, g_α is the metric tensor with respect to (U_α, u_α) and G_α is the matrix representation of g_α. The equality (7) will be proved later. From (5), we have

$$\begin{aligned} d_f(P,Q) = \int_{\mathcal{X}} p(x) \Big\{ & f\left(x, \nabla \log \tilde{q}(x)\right) - \langle \nabla \log \tilde{q}(x), \left(\nabla_2 f\right)\left(x, \nabla \log \tilde{q}(x)\right) \rangle \\ &+ \langle \nabla \log \tilde{p}(x), \left(\nabla_2 f\right)\left(x, \nabla \log \tilde{q}(x)\right) \rangle \Big\} \mathrm{d}\mu(x) - C(P), \end{aligned} \tag{8}$$

where $C(P)$ represents terms dependent only on P. From the assumption that equality (7) holds, we have

$$\begin{aligned} d_f(P,Q) = \sum_{\alpha\in\{\mathrm{u},\mathrm{d}\}} \int_{H_\alpha} p(x) \Big\{ & f\left(x, \nabla \log \tilde{q}(x)\right) - \langle \nabla \log \tilde{q}(x), \left(\nabla_2 f\right)\left(x, \nabla \log \tilde{q}(x)\right)\rangle \\ & - \frac{1}{\sqrt{|G_\alpha|}} \frac{\partial}{\partial u_\alpha^a} \left(g_\alpha^{ab} \left\langle \frac{\partial x}{\partial u_\alpha^b}, (\nabla_2 f)(x, \nabla \log \tilde{q}(x)) \right\rangle \sqrt{|G_\alpha|} \right) \Big\} \\ & \times \mathrm{d}u_\alpha^1 \wedge \ldots \wedge \mathrm{d}u_\alpha^n - C(P). \end{aligned} \tag{9}$$

By Appendix A.2, for $\alpha \in \{\mathrm{u}, \mathrm{d}\}$, for $x \in H_\alpha$,

$$\begin{aligned} S_f(x,Q) =& f\left(x, \nabla \log \tilde{q}(x)\right) - \left\langle \nabla \log \tilde{q}(x), \left(\nabla_2 f\right)\left(x, \nabla \log \tilde{q}(x)\right) \right\rangle \\ & - \frac{1}{\sqrt{|G_\alpha|}} \frac{\partial}{\partial u_\alpha^a} \left(g_\alpha^{ab} \left\langle \frac{\partial x}{\partial u_\alpha^b}, (\nabla_2 f)(x, \nabla \log \tilde{q}(x)) \right\rangle \sqrt{|G_\alpha|} \right). \end{aligned} \tag{10}$$

Combining (10) with (9) yields

$$d_f(P,Q) = \int_{\mathcal{X}} p(x) S_f(x,Q) \mathrm{d}\mu(x) - C(P).$$

Since $d_f(P,P) = 0$, we have

$$C(P) = \int_{\mathcal{X}} p(x) S_f(x,P) \mathrm{d}\mu(x).$$

Thus, we obtain $d_f = d_{S_f}$ under the assumption that equality (7) holds.

In the rest of the proof, we show that equality (7) holds. Let

$$\begin{aligned} \eta(x) := \sum_{a=1}^{n} & (-1)^{a-1} p(x) g^{ab}(x) \left\langle \frac{\partial x(u)}{\partial u^b}, (\nabla_2 f)(x, \nabla \log \tilde{q}(x)) \right\rangle \\ & \times \sqrt{|G(x)|} \mathrm{d}u^1(x) \wedge \cdots \wedge \mathrm{d}u^{a-1}(x) \wedge \mathrm{d}u^{a+1}(x) \wedge \cdots \wedge \mathrm{d}u^n(x), \end{aligned}$$

which is a differential $(n-1)$-form. As shown in Appendix A.3, η is independent of the choice of a coordinate u and is C^1. From Stoke's theorem (e.g.,[18]), we have $\int_{\mathcal{X}} \mathrm{d}\xi = 0$ for any differential $(n-1)$-form ξ on $\mathcal{X}$. Thus,

$$\begin{aligned} 0 &= \int_{\mathcal{X}} \mathrm{d}\eta \\ &= \sum_{\alpha\in\{\mathrm{u},\mathrm{d}\}} \int_{H_\alpha} \frac{\partial}{\partial u_\alpha^a} \left(p(x) g_\alpha^{ab} \left\langle \frac{\partial x}{\partial u_\alpha^b}, (\nabla_2 f)(x, \nabla \log \tilde{q}(x)) \right\rangle \sqrt{|G_\alpha|} \right) \\ &\qquad \times \mathrm{d}u_\alpha^1 \wedge \cdots \wedge \mathrm{d}u_\alpha^n \\ &= \sum_{\alpha\in\{\mathrm{u},\mathrm{d}\}} \int_{H_\alpha} \left(\frac{\partial}{\partial u_\alpha^a} p(x) \right) g_\alpha^{ab} \left\langle \frac{\partial x}{\partial u_\alpha^b}, \left(\nabla_2 f\right)\left(x, \nabla \log \tilde{q}(x)\right) \right\rangle \sqrt{|G_\alpha|} \\ &\qquad \times \mathrm{d}u_\alpha^1 \wedge \cdots \wedge \mathrm{d}u_\alpha^n \\ &\quad + \sum_{\alpha\in\{\mathrm{u},\mathrm{d}\}} \int_{H_\alpha} p(x) \frac{\partial}{\partial u_\alpha^a} \left(g_\alpha^{ab} \left\langle \frac{\partial x}{\partial u_\alpha^b}, (\nabla_2 f)(x, \nabla \log \tilde{q}(x)) \right\rangle \sqrt{|G_\alpha|} \right) \\ &\qquad \times \mathrm{d}u_\alpha^1 \wedge \cdots \wedge \mathrm{d}u_\alpha^n. \end{aligned} \tag{11}$$

Hence combining (1) with (11) yields

$$\begin{aligned}
&\int_{\mathcal{X}} \langle \nabla \log \tilde{p}(x), (\nabla_2 f)(x, \nabla \log \tilde{q}(x)) \rangle \mathrm{d}\mu(x) \\
&= \sum_{\alpha \in \{\mathrm{u},\mathrm{d}\}} \int_{H_\alpha} \left(\frac{\partial p(x)}{\partial u_\alpha^a} \right) g_\alpha^{ab} \left\langle \frac{\partial x}{\partial u_\alpha^b}, (\nabla_2 f)\left(x, \nabla \log \tilde{q}(x)\right) \right\rangle \sqrt{|G_\alpha|} \mathrm{d}u_\alpha^1 \wedge \cdots \wedge \mathrm{d}u_\alpha^n \\
&= - \sum_{\alpha \in \{\mathrm{u},\mathrm{d}\}} \int_{H_\alpha} p(x) \frac{\partial}{\partial u_\alpha^a} \left(g_\alpha^{ab} \left\langle \frac{\partial x}{\partial u_\alpha^b}, (\nabla_2 f)(x, \nabla \log \tilde{q}(x)) \right\rangle \sqrt{|G_\alpha|} \right) \mathrm{d}u_\alpha^1 \wedge \cdots \wedge \mathrm{d}u_\alpha^n \\
&= - \int_{\mathcal{X}} p(x) \frac{1}{\sqrt{|G|}} \frac{\partial}{\partial u_\alpha^a} \left(g_\alpha^{ab} \left\langle \frac{\partial x}{\partial u_\alpha^b}, (\nabla_2 f)(x, \nabla \log \tilde{q}(x)) \right\rangle \sqrt{|G_\alpha|} \right) \mathrm{d}\mu(x),
\end{aligned}$$

which shows that equality (7) holds. □

Example 3. Let $f_k(x, z) := -(||z||^2)^k$ for $k \geq 1$. The scoring rule S_{f_k} is

$$\begin{aligned}
S_{f_k}(x, Q) = (2k-1)\|\nabla \log \tilde{q}(x)\|^{2k} + 2k\|\nabla \log \tilde{q}(x)\|^{2(k-1)} \mathrm{tr}\left(\nabla^2 \log \tilde{q}(x)\right) \\
+ 4k(k-1)\|\nabla \log \tilde{q}(x)\|^{2(k-2)} \left(\nabla \log \tilde{q}(x)\right)^\top \nabla^2 \log \tilde{q}(x) \nabla \log \tilde{q}(x),
\end{aligned}$$

where $\| \cdot \|$ is the standard norm in $\mathbb{R}^{n+1}$. Since for all $x \in \mathcal{X}$, $f_k(x, \cdot)$ is strictly concave, S_{f_k} is strictly proper.

When $k = 1$, S_{f_1} is given by

$$S_{f_1}(s, Q) = \|\nabla \log \tilde{q}(x)\|^2 + 2\mathrm{tr}(\nabla^2 \log \tilde{q}(x)).$$

The corresponding divergence function $d_{S_{f_1}}$ in (5) is

$$d_{S_{f_1}}(P, Q) = \int p(x) \left\|\nabla \log \tilde{q}(x) - \nabla \log \tilde{p}(x)\right\|^2 \mathrm{d}\mu(x).$$

Here, S_{f_1} and $d_{S_{f_1}}$ for probability densities on the sphere correspond to the Hyvärinen scoring rule and the Hyvärinen divergence for probability densities on the Euclidean space, respectively.

4 Orthogonally-invariant scoring rules on unit spheres

In this section, we investigate orthogonally invariant scoring rules. We denote an orthogonal transformation with an orthogonal matrix V as $V(x) = Vx$.

In the following, suppose that a parametric model $\mathcal{M}$ on the sphere satisfies $Q \circ V \in \mathcal{M}$ for any $Q \in \mathcal{M}$ and any orthogonal transformation V, where $Q \circ V$ is the distribution given by $Q \circ V(A) = Q(V^{-1}(A))$ for any measurable set A. A scoring rule S on $\mathcal{X}$ is said to be orthogonally-invariant if

$$S(Vx, Q \circ V^{-1}) = S(x, Q).$$

The following lemma gives a sufficient condition for S_f in (6) to be orthogonally-invariant.

Lemma 2. The scoring rule S_f defined by (6) is orthogonally-invariant if f has the form

$$f(x, z) = g(\|z\|^2),$$

where $g : \mathbb{R}_+ \to \mathbb{R}$ is twice continuously differentiable and $\|\cdot\|$ is the standard norm in $\mathbb{R}^{n+1}$.

Proof. Let $H(x) := \nabla^2 \log \tilde{q}(x) \in \mathbb{R}^{(n+1)\times(n+1)}$ and let $\psi(x) := \nabla \log \tilde{q}(x) \in \mathbb{R}^{n+1}$.

Since $V^\top V = I_{n+1}$ with the $(n+1)\times(n+1)$ identity matrix I_{n+1} and since for $x \in \mathcal{X}$ and for $z \in \mathbb{R}^{n+1}$,

$$\begin{aligned} f(Vx, Vz) &= g(\|Vz\|^2) = g(\|z\|^2), \\ (\nabla_2 f)(Vx, Vz) &= 2g'(\|z\|^2)Vz, \\ (\nabla_2^2 f)(Vx, Vz) &= 2g'(\|z\|^2)I_n + 4g''(\|z\|^2)Vzz^\top V^\top, \end{aligned}$$

we obtain

$$\begin{aligned} &S_f(Vx, Q \circ V^{-1}) \\ =&f\left(Vx, V\psi(x)\right) - (\psi(x))^\top V^\top (\nabla_2 f)\left(Vx, V\psi(x)\right) \\ &\quad - \left(\nabla_1 \cdot \nabla_2 f\right)\left(Vx, V\psi(x)\right) - \mathrm{tr}\left(\left(\nabla_2^2 f\right)\left(Vx, V\psi(x)\right) H(x)\right) \\ &\quad + nx^\top V^\top (\nabla_2 f)\left(Vx, V\psi(x)\right) \\ =&g\left(\|\psi(x)\|^2\right) - 2g'\left(\|\psi(x)\|^2\right)\left\{\|\psi(x)\|^2 + \mathrm{tr}\left(H(x)\right)\right\} \\ &\quad - 4g''\left(\|\psi(x)\|^2\right)\left(\psi(x)\right)^\top H(x)\psi(x) \\ =&S_f(x, Q). \end{aligned}$$

□

Combining Lemma 2 with Theorem 1 yields the following theorem.

Theorem 2. *Suppose that a twice continuously differentiable function $g : \mathbb{R}_+ \to \mathbb{R}$ satisfies*

$$\lim_{w\to+0} g'(w)\sqrt{w} = 0, \tag{12}$$

$$\lim_{w\to+0} g''(w)w = 0, \tag{13}$$

$$g'(w) < 0 \text{ for } w \in \mathbb{R}_+, \tag{14}$$

and

$$g'(w) + 2g''(w)w < 0 \text{ for } w \in \mathbb{R}_+. \tag{15}$$

Then, the scoring rule S_f defined by (6) *with $f(x, z) = g(\|z\|^2)$ is strictly proper, 2-local, homogeneous, and orthogonally-invariant.*

The first assumption (12) and the second assumption (13) ensure that $f(x, z)$ in (5) and (6) is twice continuously differentiable with respect to z at $z = 0$. The third assumption (14) and the fourth assumption (15) ensure that for each $x \in \mathcal{X}$, $f(x, \cdot)$ is strictly concave.

5 Numerical experiments

In this section, we give several numerical experiments of parameter estimation using the proposed scoring rules. We consider the Fisher–Bingham distribution with density function

$$p(x; a, A) = \frac{1}{c(a, A)} \exp(a^\top x + x^\top A x),$$

where $a \in \mathbb{R}^{n+1}$ and $A \in \mathbb{R}^{(n+1)\times(n+1)}$ satisfying $A^\top = A$ and $\mathrm{tr}(A) = 0$.

Suppose that the dimension n of $\mathcal{X}$ is 3 and the true values of (a, A) are $a = (0, 0, 0, 0)^\top$ and $A = \mathrm{diag}(4, 2, -2, -4)$, where $\mathrm{diag}(d_1, \ldots, d_k)$ is the diagonal matrix of which the (i, i)-component is d_i.

Consider the estimation of A when a is known. We generate samples $x_1, \ldots, x_T$ and calculate the estimates $\hat{A}_g(x_1, \ldots, x_T)$ based on the scoring rule in Theorem 2 with $g : \mathbb{R}_+ \to \mathbb{R}$. We denote the scoring rule in Theorem 2 with g by S_g. We obtain $\hat{A}_g$ using the gradient descent method where the initial value is the zero matrix, and evaluate the squared error $\|\hat{A}_g - A^*\|_{\mathrm{F}}^2$ where $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm. We repeat the process above N times and obtain the average squared error.

We consider two classes of scoring rules. First, we consider $g_{1,k}(w) = -w^k$ with $k \geq 1/2$. Figure 1 shows the average squared error with respect to the sample size T when $k = 1$ and $N = 100$. Figures 2 and 3 show the average squared error with respect to k when $N = 1000$ and $T = 100$ and $T = 500$, respectively.

From Figure 1, we observe that $\hat{A}_{g_{1,1}}$ is consistent. From Figures 2 and 3, we see that average of $\|\hat{A}_{g_{1,k}} - A^*\|_{\mathrm{F}}^2$ is minimized at about $k = 1$.

Second, we consider $g_{2,k}(w) = -(1 + kw)\log(1 + kw)$ with $k > 0$. Figure 4 shows the average squared error with respect to the sample size T when $k = 1000$ and $N = 100$. Figure 5 and 6 show the average squared error with respect to k when $N = 1000$ and $T = 100$ and $T = 500$, respectively.

From Figure 4, we see that $\hat{A}_{g_{2,1000}}$ is consistent. From Figures 5 and 6, we see that the average of $\|\hat{A}_{g_{2,k}} - A^*\|_{\mathrm{F}}^2$ decreases as k gets larger.

Here, we compare the average squared errors of $\hat{A}_{g_{1,k}}$ and $\hat{A}_{g_{2,k}}$ to the expected squared error of the maximum likelihood estimator $\hat{A}_{\mathrm{MLE}}$ of A. Since we cannot calculate $\hat{A}_{\mathrm{MLE}}$ directly, we calculate the Fisher information $I(\theta)$ of

$$\theta = (a_{11}, a_{22}, \ldots, a_{nn}, a_{12}, a_{13}, \ldots, a_{1n+1}, a_{23}, \ldots, a_{nn+1})^\top$$

by the Monte Carlo method.

From Figure 3, we see that when $T = 500$, the minimal value of the average squared error of $\hat{A}_{g_{1,k}}$ is about 0.705: this is 113.3% of that of $\hat{A}_{\mathrm{MLE}}$. From Figure 6, we see that when $T = 500$, the minimal value of the average squared error of $\hat{A}_{g_{2,k}}$ is about 0.652: this is 104.8% of that of $\hat{A}_{\mathrm{MLE}}$.

These results show that the parameter estimation in the Fisher–Bingham distribution based on our scoring rules is comparable to the maximum likelihood estimation and show that the class $\{S_{g_{2,k}}\}$ work better than $\{S_{g_{1,k}}\}$ in this example.

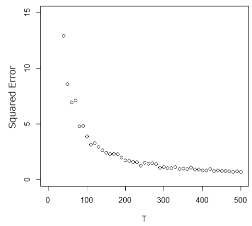

Figure 1. Average of $\|\hat{A}_{g_{1,1}} - A^*\|_{\mathrm{F}}^2$ with respect to T.

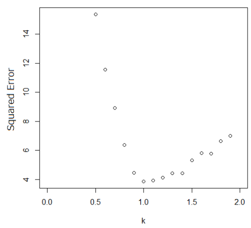

Figure 2. Average of $\|\hat{A}_{g_{1,k}} - A^*\|_{\mathrm{F}}^2$ with respect to k when $T = 100$.

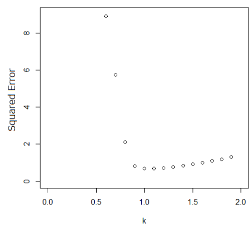

Figure 3. Average of $\|\hat{A}_{g_{1,k}} - A^*\|_{\mathrm{F}}^2$ with respect to k when $T = 500$.

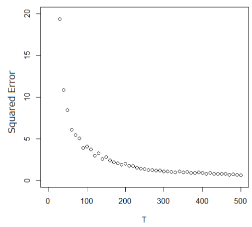

Figure 4. Average $\|\hat{A}_{g_{2,1000}} - A^*\|_{\mathrm{F}}^2$ with respect to T.

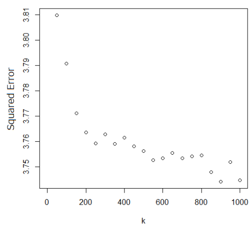

Figure 5. Average $\|\hat{A}_{g_{2,k}} - A^*\|_{\mathrm{F}}^2$ with respect to k when $T = 100$.

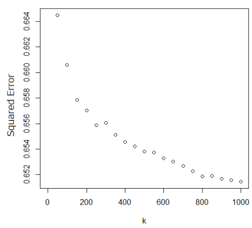

Figure 6. Average $\|\hat{A}_{g_{2,k}} - A^*\|_{\mathrm{F}}^2$ with respect to k when $T = 500$.

6 Conclusion

We have proposed a class of strictly proper scoring rules to estimate the parameters of statistical models on spheres. We have defined new divergence functions on probability distributions on spheres. To evaluate these scoring rules, we do not need normalizing constants because they are 2-local and homogeneous. Moreover, we have considered orthogonally-invariant scoring rules. The proposed scoring rules work well and the performance for parameter estimation is comparable to the maximum likelihood estimator with respect to the squared error throughout numerical experiments.

References

[1] Amari, S. (1985). Differential-Geometrical Methods in Statistics, Springer, New York.

[2] Bernardo, J. M. (1979). Expected information as expected utility. *The Annals of Statistics*, **7**, 686–690.

[3] Boyd, S. and Vandenberghe, L. (2004). Convex Optimization, Cambridge University Press, Cambridge.

[4] Bregman, L. M. (1967). The relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, **7**, 200–217.

[5] Dawid, A. P. (1998). Coherent measures of discrepancy, uncertainty and dependence, with applications to Bayesian predictive experimental design. Technical Report 139, Dept. Statistical Science, Univ. College London.

[6] Dawid, A. P. (2007). The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, **59**, 77–93.

[7] Dhillon, I. S. and Modha, D. S. (2001). Concept decompositions for large sparse text data using clustering. *Machine Learning*, **42**, 143–175.

[8] Eguchi. S. (2008). Information divergence geometry and the application to statistical machine learning. In *Information Theory and Statistical Learning* (F.Emmert-Streib and M. Dehmer,eds.), 309–332. Springer, New York.

[9] Ehm, W. and Gneiting, T. (2012). Local proper scoring rules of order two. *The Annals of Statistics*, **40**, 609–637.

[10] Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, **6**, 695–709.

[11] Janssen, P. J., Audit, B. and Ouzounis, C. A. (2001). Strain-specific genes of helicobacter pylori: distribution, function and dynamics. *Nucleic Acids Research*, **29**, 4395–4404.

[12] Kume, A. and Wood, A. T. A. (2005). Saddlepoint approximations for the Bingham and Fisher–Bingham normalising constants. *Biometrika*, **92**, 465–476.

[13] Mardia, K. and Jupp, P. (2000). *Directional Statistics*, 1st ed., Wiley, Chichester.

[14] Murase, H. and Nayar, S. K. (1995). Visual learning and recognition of 3D objects from appearance. *Journal of Computer Vision*, **14**, 5–24.

[15] Nakayama, H., Nishiyama, K., Noro, M., Ohara, K., Sei, T., Takayama, N. and Takemura, A. (2011). Holonomic gradient descent and its application to Fisher–Bingham integral. *Advances in Applied Mathematics*, **47**, 639–658.

[16] Parry, M., Dawid, A. P. and Lauritzen, S. (2012). Proper local scoring rules. *The Annals of Statistics*, **40**, 561–592.

[17] Slice, D. E. (2005). *Modern Morphometrics in Physical Anthropology*, Kluwer Academics, New York.

[18] Spivak, M. (1999). *A Comprehensive Introduction to Differential Geometry*, 3rd ed., Publish or Perish, Houston.

Appendix A

In this appendix, we provide the calculations used in Theorem 1.

A.1 Calculations over two local coordinates $(U_\mathrm{u}, u_\mathrm{u})$ and $(U_\mathrm{d}, u_\mathrm{d})$

We summarize calculations over $(U_\mathrm{u}, u_\mathrm{u})$ and $(U_\mathrm{d}, u_\mathrm{d})$ used in Appendices A.2 and A.3. Let $H(x) := \nabla^2 \log \tilde{q}(x) \in \mathbb{R}^{(n+1)\times(n+1)}$ and let $\psi(x) := \nabla \log \tilde{q}(x) \in \mathbb{R}^{n+1}$.

From the definition,

$$x^i = \begin{cases} \frac{2u_\mathrm{u}^i}{1+\sum_{j=1}^n (u_\mathrm{u}^j)^2}, & (i = 1, \dots, n), \\ -1 + \frac{2}{1+\sum_{j=1}^n (u_\mathrm{u}^j)^2}, & (i = n+1) \end{cases} \tag{16}$$

for all $x \in U_\mathrm{u}$. Therefore

$$\begin{aligned} \frac{\partial u_\mathrm{u}}{\partial x} &= \begin{pmatrix} \frac{\partial u_\mathrm{u}^1}{\partial x^1} & \cdots & \frac{\partial u_\mathrm{u}^1}{\partial x^{n+1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial u_\mathrm{u}^n}{\partial x^1} & \cdots & \frac{\partial u_\mathrm{u}^n}{\partial x^{n+1}} \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{1+x^{n+1}} I_n & -\frac{1}{1+x^{n+1}} u_\mathrm{u} \end{pmatrix}, \\ \frac{\partial x}{\partial u_\mathrm{u}} &= \begin{pmatrix} \frac{\partial x^1}{\partial u_\mathrm{u}^1} & \cdots & \frac{\partial x^1}{\partial u_\mathrm{u}^n} \\ \vdots & \ddots & \vdots \\ \frac{\partial x^{n+1}}{\partial u_\mathrm{u}^1} & \cdots & \frac{\partial x^{n+1}}{\partial u_\mathrm{u}^n} \end{pmatrix} \\ &= \begin{pmatrix} \frac{2}{1+\sum_{j=1}^n (u_\mathrm{u}^j)^2} I_n - \frac{4}{\{1+\sum_{j=1}^n (u_\mathrm{u}^j)^2\}^2} u_\mathrm{u} u_\mathrm{u}^\top \\ -\frac{4}{\{1+\sum_{j=1}^n (u_\mathrm{u}^j)^2\}^2} u_\mathrm{u}^\top \end{pmatrix} \\ &= \begin{pmatrix} (1+x^{n+1}) I_n - (1+x^{n+1})^2 u_\mathrm{u} u_\mathrm{u}^\top \\ -(1+x^{n+1})^2 u_\mathrm{u}^\top \end{pmatrix}. \end{aligned} \tag{17}$$

Hence the representation of the metric tensor on $(U_{\rm u}, u_{\rm u})$ denoted by $G_{\rm u}$ is

$$
\begin{aligned}
G_{\rm u} &= \left(\frac{\partial x}{\partial u_{\rm u}}\right)^\top \frac{\partial x}{\partial u_{\rm u}} \\
&= \left\{(1+x^{n+1})I_n - (1+x^{n+1})^2 u_{\rm u} u_{\rm u}^\top\right\}^2 + (1+x^{n+1})^4 u_{\rm u} u_{\rm u}^\top \\
&= (1+x^{n+1})^2 I_n + (1+x^{n+1})^3 \left\{-1 + x^{n+1} + (1+x^{n+1}) u_{\rm u}^\top u_{\rm u}\right\} u_{\rm u} u_{\rm u}^\top \\
&= (1+x^{n+1})^2 I_n, \qquad (18)
\end{aligned}
$$

and

$$|G_{\rm u}| = (1+s^{n+1})^{2n}. \qquad (19)$$

Here we use

$$
\begin{aligned}
u_{\rm u}^\top u_{\rm u} &= \sum_{j=1}^n (u_{\rm u}^j)^2 \\
&= \frac{1}{(1+x^{n+1})^2} \sum_{j=1}^n (x^j)^2 \\
&= \frac{1}{(1+x^{n+1})^2} (1 - (x^{n+1})^2) \\
&= \frac{1-x^{n+1}}{1+x^{n+1}}.
\end{aligned}
$$

Similarly,

$$
x^i = \begin{cases} -\frac{2u_{\rm d}^i}{1+\sum_{j=1}^n (u_{\rm d}^j)^2}, & (i = 1, \dots, n), \\ 1 - \frac{2}{1+\sum_{j=1}^n (u_{\rm d}^j)^2}, & (i = n+1) \end{cases} \qquad (20)
$$

for all $x \in (U_{\rm d})$. Therefore

$$
\begin{aligned}
\frac{\partial u_{\rm d}}{\partial x} &= \left(\tfrac{1}{-1+x^{n+1}} I_n \quad \tfrac{1}{1-x^{n+1}} u_{\rm d}\right), \\
\frac{\partial x}{\partial u_{\rm d}} &= \begin{pmatrix} (-1+x^{n+1}) I_n + (1-x^{n+1})^2 u_{\rm d} u_{\rm d}^\top \\ (1-x^{n+1})^2 u_{\rm d}^\top \end{pmatrix}. \qquad (21)
\end{aligned}
$$

Hence the representation of the metric tensor on $(U_{\rm d}, u_{\rm d})$ denoted by $G_{\rm d}$ is

$$G_{\rm d} = \left(\frac{\partial x}{\partial u_{\rm d}}\right)^\top \frac{\partial x}{\partial u_{\rm d}} = (1-x^{n+1})^2 I_n, \qquad (22)$$

$$|G_{\rm d}| = (1-x^{n+1})^{2n}, \qquad (23)$$

where we use

$$u_{\rm d}^\top u_{\rm d} = \frac{1+x^{n+1}}{1-x^{n+1}}.$$

A.2 Proof for equality (10)

We show that S_f is equal to the right-hand side in (10). To show this, it suffices to show that the third term in (10) is equal to

$$(\nabla_1 \cdot \nabla_2 f)(x, \psi(x)) + \operatorname{tr}\big((\nabla_2^2 f)(x, \psi(x)) H(x)\big) - n x^\top (\nabla_2 f)(x, \psi(x)).$$

First, for $\alpha \in \{\mathrm{u}, \mathrm{d}\}$, for $x \in U_\alpha$ the third term in (10) is expanded as

$$\begin{aligned}
&\frac{1}{\sqrt{|G_\alpha|}} \frac{\partial}{\partial u_\alpha^a} \left\{ g_\alpha^{ab} \left\langle \frac{\partial x}{\partial u_\alpha^b}, (\nabla_2 f)\left(x, \nabla \log \tilde{q}(x)\right) \right\rangle \sqrt{|G_\alpha|} \right\} \\
&\qquad = g_\alpha^{ab} \left\langle \frac{\partial x}{\partial u_\alpha^b}, \frac{\partial}{\partial u_\alpha^a} (\nabla_2 f)\left(x, \nabla \log \tilde{q}(x)\right) \right\rangle \\
&\qquad\quad + \frac{1}{\sqrt{|G_\alpha|}} \left\langle \frac{\partial}{\partial u_\alpha^a} \left(\sqrt{|G_\alpha|} g_\alpha^{ab} \frac{\partial x}{\partial u^b} \right), (\nabla_2 f)\left(x, \nabla \log \tilde{q}(x)\right) \right\rangle .
\end{aligned} \tag{24}$$

The first term in the above equality (24) is

$$\begin{aligned}
& g_\alpha^{ab} \left\langle \frac{\partial x}{\partial u_\alpha^b}, \frac{\partial}{\partial u_\alpha^a} (\nabla_2 f)\left(x, \nabla \log \tilde{q}(x)\right) \right\rangle \\
&= (\nabla_1 \cdot \nabla_2 f)\left(x, \nabla \log \tilde{q}(x)\right) + g_\alpha^{ab} \left\langle \frac{\partial x}{\partial u_\alpha^b}, (\nabla_2^2 f)\left(x, \nabla \log \tilde{q}(x)\right) \frac{\partial}{\partial u_\alpha^a} \nabla \log \tilde{q}(x) \right\rangle \\
&= (\nabla_1 \cdot \nabla_2 f)\left(x, \nabla \log \tilde{q}(x)\right) + \operatorname{tr}\left((\nabla_2^2 f)\left(x, \nabla \log \tilde{q}(x)\right) \nabla^2 \log \tilde{q}(x) \right).
\end{aligned}$$

Second, from (17), (18) and (19), for $x \in U_\mathrm{u}$, a part of the second term in (24) is calculated as

$$\begin{aligned}
& \frac{\partial}{\partial u_\mathrm{u}^a} \left(\sqrt{|G_\mathrm{u}|} g_\mathrm{u}^{ab} \frac{\partial x}{\partial u_\mathrm{u}^b} \right) \\
&\quad = \frac{\partial}{\partial u_\mathrm{u}} \begin{pmatrix} (1+x^{n+1})^{n-1} I_n - (1+x^{n+1})^n u_\mathrm{u} u_\mathrm{u}^\top & -(1+x^{n+1})^n u_\mathrm{u} \end{pmatrix} \\
&\quad = \frac{\partial}{\partial u_\mathrm{u}} \left\{ \left(\frac{2}{1+\sum_{j=1}^n (u_\mathrm{u}^j)^2} \right)^n \begin{pmatrix} \frac{1+\sum_{j=1}^n (u_\mathrm{u}^j)^2}{2} I_n - u_\mathrm{u} u_\mathrm{u}^\top & -u_\mathrm{u} \end{pmatrix} \right\} \\
&\quad = -n \left(\frac{2}{1+\sum_{j=1}^n (u_\mathrm{u}^j)^2} \right)^{n+1} \begin{pmatrix} \frac{1+\sum_{j=1}^n (u_\mathrm{u}^j)^2}{2} I_n - u_\mathrm{u} u_\mathrm{u}^\top \\ -u_\mathrm{u}^\top \end{pmatrix} u_\mathrm{u} \\
&\qquad\qquad + \left(\frac{2}{1+\sum_{j=1}^n (u_\mathrm{u}^j)^2} \right)^n \begin{pmatrix} u_\mathrm{u} - (n+1) u_\mathrm{u} \\ -n \end{pmatrix} \\
&\quad = \left(\frac{2}{1+\sum_{j=1}^n (u_\mathrm{u}^j)^2} \right)^n \begin{pmatrix} \{-n + n(1-x^{n+1}) - n\} u_\mathrm{u} \\ n(1-x^{n+1}) - n \end{pmatrix} \\
&\quad = (1+x^{n+1})^n \begin{pmatrix} -n(1+x^{n+1}) u_\mathrm{u} \\ -n x^{n+1} \end{pmatrix} \\
&\quad = -n \sqrt{|G_\mathrm{u}|} x.
\end{aligned} \tag{25}$$

From (21), (22) and (23), for $x \in U_\mathrm{d}$, a part of the second term of (24) is calculated as

$$
\begin{aligned}
&\frac{\partial}{\partial u_\mathrm{d}^a}\left(\sqrt{|G_\mathrm{d}|}g_\mathrm{d}^{ab}\frac{\partial x}{\partial u_\mathrm{d}^b}\right)\\
&= \frac{\partial}{\partial u_\mathrm{d}}\left(-(1-x^{n+1})^{n-1}I_n + (1-x^{n+1})^n u_\mathrm{d}u_\mathrm{d}^\top \quad (1-x^{n+1})^n u_\mathrm{d}\right)\\
&= \frac{\partial}{\partial u_\mathrm{d}}\left\{\left(\frac{2}{1+\sum_{j=1}^n (u_\mathrm{d}^j)^2}\right)^n \left(-\tfrac{1+\sum_{j=1}^n (u_\mathrm{d}^j)^2}{2} I_n + u_\mathrm{d}u_\mathrm{d}^\top \quad u_\mathrm{d}\right)\right\}\\
&= n\left(\frac{2}{1+\sum_{j=1}^n (u_\mathrm{d}^j)^2}\right)^{n+1}\begin{pmatrix}\tfrac{1+\sum_{j=1}^n (u_\mathrm{d}^j)^2}{2} I_n - u_\mathrm{d}u_\mathrm{d}^\top \\ -u_\mathrm{d}^\top\end{pmatrix} u_\mathrm{d}\\
&\qquad\qquad + \left(\frac{2}{1+\sum_{j=1}^n (u_\mathrm{d}^j)^2}\right)^n \begin{pmatrix}-u_\mathrm{d} + (n+1)u_\mathrm{d} \\ n\end{pmatrix}\\
&= \left(\frac{2}{1+\sum_{j=1}^n (u_\mathrm{d}^j)^2}\right)^n \begin{pmatrix}\{n - n(1+x^{n+1}) + n\}\, u_\mathrm{d} \\ -n(1+x^{n+1}) + n\end{pmatrix}\\
&= (1-x^{n+1})^n \begin{pmatrix}-n(1-x^{n+1})u_\mathrm{d} \\ -nx^{n+1}\end{pmatrix}\\
&= -n\sqrt{|G_\mathrm{d}|}x.
\end{aligned} \tag{26}
$$

From (25) and (26), we obtain

$$
\begin{aligned}
&\frac{1}{\sqrt{|G|}}\left\langle \frac{\partial}{\partial u^a}\left(\sqrt{|G|}g^{ab}\frac{\partial x}{\partial u^b}\right), (\nabla_2 f)(x, \nabla \log \tilde{q}(x))\right\rangle\\
&\qquad = -nx^\top (\nabla_2 f)(x, \nabla \log \tilde{q}(x)).
\end{aligned}
$$

Thus, the third term of (10) is equal to

$$
(\nabla_1 \cdot \nabla_2 f)(x, \psi(x)) + \mathrm{tr}\left((\nabla_2^2 f)(x, \psi(x))H(x)\right) - nx^\top (\nabla_2 f)(x, \psi(x)),
$$

which completes the proof.

A.3 Proofs about η in Theorem 2

We show that η is independent of coordinates and is of class C^1.

First, we show that η is independent of coordinates. We consider expressing η by another coordinate system $\tilde{u}$ that has the same orientation as u. We denote indices for u by $a, b, c, d \in \{1, \dots, n\}$ and denote indices for $\tilde{u}$ by $a', b', c', d' \in \{1, \dots, n\}$, respectively. The change of coordinates yields the following transformations:

$$
\begin{aligned}
&(-1)^{a-1}\mathrm{d}u^1 \wedge \cdots \wedge \mathrm{d}u^{a-1} \wedge \mathrm{d}u^{a+1} \wedge \cdots \wedge \mathrm{d}u^n\\
&= \sum_{c'=1}^n (-1)^{c'-1}\left|\frac{\partial u}{\partial \tilde{u}}\right| \frac{\partial \tilde{u}^{c'}}{\partial u^a}\mathrm{d}\tilde{u}^1 \wedge \cdots \wedge \mathrm{d}\tilde{u}^{c'-1} \wedge \mathrm{d}\tilde{u}^{c'+1} \wedge \cdots \wedge \mathrm{d}\tilde{u}^n,
\end{aligned} \tag{27}
$$

$$
g_{ab} = \tilde{g}_{a'b'}\frac{\partial \tilde{u}^{a'}}{\partial u^a}\frac{\partial \tilde{u}^{b'}}{\partial u^b}, \tag{28}
$$

$$g^{ab} = \tilde{g}^{a'b'} \frac{\partial u^a}{\partial \tilde{u}^{a'}} \frac{\partial u^b}{\partial \tilde{u}^{b'}}, \tag{29}$$

$$|G| = |\tilde{G}| \left| \frac{\partial \tilde{u}}{\partial u} \right|^2, \tag{30}$$

where the metric tensor of $\mathcal{X}$ defined through $\tilde{u}$ is denoted by $\tilde{g}$ and its matrix form is denoted by $\tilde{G}$. To see that equality (27) holds, note that

$$\begin{aligned}
&\mathrm{d}\tilde{u}^{c'} \wedge (-1)^{a-1} \mathrm{d}u^1 \wedge \cdots \wedge \mathrm{d}u^{a-1} \wedge \mathrm{d}u^{a+1} \wedge \cdots \wedge \mathrm{d}u^n \\
&= \mathrm{d}\tilde{u}^{c'} \wedge (-1)^{a-1} \left(\frac{\partial u^1}{\partial \tilde{u}^{b'_1}} \mathrm{d}\tilde{u}^{b'_1} \right) \wedge \cdots \wedge \left(\frac{\partial u^{a-1}}{\partial \tilde{u}^{b'_{a-1}}} \mathrm{d}\tilde{u}^{b'_{a-1}} \right) \\
&\qquad \wedge \left(\frac{\partial u^{a+1}}{\partial \tilde{u}^{b'_{a+1}}} \mathrm{d}\tilde{u}^{b'_{a+1}} \right) \wedge \cdots \wedge \left(\frac{\partial u^n}{\partial \tilde{u}^{b'_n}} \mathrm{d}\tilde{u}^{b'_n} \right) \\
&= \left(\frac{\partial u^1}{\partial \tilde{u}^{b'_1}} \mathrm{d}\tilde{u}^{b'_1} \right) \wedge \cdots \wedge \left(\frac{\partial u^{a-1}}{\partial \tilde{u}^{b'_{a-1}}} \mathrm{d}\tilde{u}^{b'_{a-1}} \right) \wedge \mathrm{d}\tilde{u}^{c'} \wedge \left(\frac{\partial u^{a+1}}{\partial \tilde{u}^{b'_{a+1}}} \mathrm{d}\tilde{u}^{b'_{a+1}} \right) \wedge \\
&\qquad \cdots \wedge \left(\frac{\partial u^n}{\partial \tilde{u}^{b'_n}} \mathrm{d}\tilde{u}^{b'_n} \right) \\
&= \sum_{\substack{\sigma:\text{permutation} \\ \sigma(k')=i}} \mathrm{sgn}(\sigma) \frac{\partial u^{\sigma(1)}}{\partial \tilde{u}^1} \cdots \frac{\partial u^{\sigma(k'-1)}}{\partial \tilde{u}^{k'-1}} \frac{\partial u^{\sigma(k'+1)}}{\partial \tilde{u}^{k'+1}} \cdots \frac{\partial u^{\sigma(n)}}{\partial \tilde{u}^n} \mathrm{d}\tilde{u}^1 \wedge \cdots \wedge \mathrm{d}\tilde{u}^n \\
&= \left| \frac{\partial u}{\partial \tilde{u}} \right| \frac{\partial \tilde{u}^{c'}}{\partial u^a} \mathrm{d}\tilde{u}^1 \wedge \cdots \wedge \mathrm{d}\tilde{u}^n \\
&= (-1)^{c'-1} \left| \frac{\partial u}{\partial \tilde{u}} \right| \frac{\partial \tilde{u}^{c'}}{\partial u^a} \mathrm{d}\tilde{u}^{c'} \wedge \mathrm{d}\tilde{u}^1 \wedge \cdots \wedge \mathrm{d}\tilde{u}^{c'-1} \wedge \mathrm{d}\tilde{u}^{c'+1} \wedge \cdots \wedge \mathrm{d}\tilde{u}^n.
\end{aligned}$$

From (27), (28), (29), and (30), we have

$$\begin{aligned}
&\sum_{a=1}^{n} (-1)^{a-1} g^{ab} \frac{\partial x}{\partial u^b} \sqrt{|G|} \mathrm{d}u^1 \wedge \cdots \wedge \mathrm{d}u^{a-1} \wedge \mathrm{d}u^{a+1} \wedge \cdots \wedge \mathrm{d}u^n \\
&\quad = \sum_{c'=1}^{n} (-1)^{c'-1} \left(\tilde{g}^{a'b'} \frac{\partial u^a}{\partial \tilde{u}^{a'}} \frac{\partial u^b}{\partial \tilde{u}^{b'}} \right) \left(\frac{\partial x}{\partial \tilde{u}^{d'}} \frac{\partial \tilde{u}^{d'}}{\partial u^b} \right) \left(\sqrt{|\tilde{G}|} \left| \frac{\partial \tilde{u}}{\partial u} \right| \right) \\
&\qquad \times \left| \frac{\partial u}{\partial \tilde{u}} \right| \frac{\partial \tilde{u}^{c'}}{\partial u^a} \mathrm{d}\tilde{u}^1 \wedge \cdots \wedge \mathrm{d}\tilde{u}^{c'-1} \wedge \mathrm{d}\tilde{u}^{c'+1} \wedge \cdots \wedge \mathrm{d}\tilde{u}^n \\
&\quad = \sum_{a'=1}^{n} (-1)^{a'-1} \tilde{g}^{a'b'} \frac{\partial x}{\partial \tilde{u}^{b'}} \sqrt{|\tilde{G}|} \mathrm{d}\tilde{u}^1 \wedge \cdots \wedge \mathrm{d}\tilde{u}^{a'-1} \wedge \mathrm{d}\tilde{u}^{a'+1} \wedge \cdots \wedge \mathrm{d}\tilde{u}^n.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\eta(x) = p(x) \sum_{a'=1}^{n} (-1)^{a'-1} \tilde{g}^{a'b'} &\left\langle \frac{\partial x(\tilde{u})}{\partial \tilde{u}^{b'}}, (\nabla_2 f)(x, \nabla \log \tilde{q}(x)) \right\rangle \\
&\times \sqrt{|\tilde{G}|} \mathrm{d}\tilde{u}^1 \wedge \cdots \wedge \mathrm{d}\tilde{u}^{a'-1} \wedge \mathrm{d}\tilde{u}^{a'+1} \wedge \cdots \wedge \mathrm{d}\tilde{u}^n,
\end{aligned}$$

which shows that η is independent of coordinates.

Next, we show that η is C^1. Since $U_\mathrm{u} \cup U_\mathrm{d} = \mathcal{X}$, we only need to prove that elements in η are of class C^1 over these two coordinates.

Consider $(U_\mathrm{u}, u_\mathrm{u})$. From (17), (18) and (19), the element of η with respect to $\mathrm{d}u_\mathrm{u}^1 \wedge \cdots \wedge \mathrm{d}u_\mathrm{u}^{a-1} \wedge \mathrm{d}u_\mathrm{u}^{a+1} \wedge \cdots \wedge \mathrm{d}u_\mathrm{u}^n$ is given as

$$\begin{aligned}
&(-1)^{a-1} p(x) g_\mathrm{u}^{ab} \left\langle \frac{\partial x(u_\mathrm{u})}{\partial u_\mathrm{u}^b}, (\nabla_2 f)(x, \nabla \log \tilde{q}(x)) \right\rangle \sqrt{|G_\mathrm{u}|} \\
&= (-1)^{a-1} p(x) \left(\frac{2}{1 + \sum_{b=1}^n \left(u_\mathrm{u}^b\right)^2} \right)^n \\
&\quad \times \left\langle \left(\frac{1 + \sum_{j=1}^n (u_\mathrm{u}^j)^2}{2} e_a - u_\mathrm{u}^a u_\mathrm{u}, -u_\mathrm{u}^a \right), (\nabla_2 f)(x, \nabla \log \tilde{q}(x)) \right\rangle
\end{aligned} \tag{31}$$

where e_a is a unit vector in $\mathbb{R}^{n+1}$ whose a-th element is 1 and the other elements are 0. To show that η is C^1, it suffices to show that each element in (31) is C^1. Since from (16), (17) and (18), x and $\nabla \log \tilde{q}(x)$ are given as

$$\begin{aligned}
x^i &= \begin{cases} \frac{2u_\mathrm{u}^i}{1+\sum_{j=1}^n (u_\mathrm{u}^j)^2}, & (i = 1, \ldots, n), \\ -1 + \frac{2}{1+\sum_{j=1}^n (u_\mathrm{u}^j)^2}, & (i = n+1), \end{cases} \\
\nabla \log \tilde{q}(x) &= \left(\frac{\partial x}{\partial u_\mathrm{u}} \right) G_\mathrm{u}^{-1} \frac{\partial \log q(x(u_\mathrm{u}))}{\partial u_\mathrm{u}} \\
&= \begin{pmatrix} \frac{1+\sum_{j=1}^n \left(u_\mathrm{u}^j\right)^2}{2} I_n - u_\mathrm{u} u_\mathrm{u}^\top \\ -u_\mathrm{u}^\top \end{pmatrix} \frac{\partial \log q(x(u_\mathrm{u}))}{\partial u_\mathrm{u}}
\end{aligned}$$

and since p and q are of class C^2, since $p(x) \neq 0$ and $q(x) \neq 0$ for all $x \in \mathcal{X}$, and since f is of class C^2, we conclude that η is of class C^1 over U_u.

Consider $(U_\mathrm{d}, u_\mathrm{d})$. From (21), (22) and (23), the element of η with respect to $\mathrm{d}u_\mathrm{d}^1 \wedge \cdots \wedge \mathrm{d}u_\mathrm{d}^{a-1} \wedge \mathrm{d}u_\mathrm{d}^{a+1} \wedge \cdots \wedge \mathrm{d}u_\mathrm{d}^n$ is given as

$$\begin{aligned}
&(-1)^{a-1} p(x) g_\mathrm{d}^{ab} \left\langle \frac{\partial x(u_\mathrm{d})}{\partial u_\mathrm{d}^b}, (\nabla_2 f)(x, \nabla \log \tilde{q}(x)) \right\rangle \sqrt{|G_\mathrm{d}|} \\
&= (-1)^{a-1} p(x) \left(\frac{2}{1 + \sum_{j=1}^n \left(u_\mathrm{d}^j\right)^2} \right)^n \\
&\quad \times \left\langle \left(\frac{1 + \sum_{j=1}^n (u_\mathrm{d}^j)^2}{2} e_a - u_\mathrm{d}^a u_\mathrm{d}, u_\mathrm{d}^a \right), (\nabla_2 f)(x, \nabla \log \tilde{q}(x)) \right\rangle .
\end{aligned} \tag{32}$$

Since from (20), (21) and (22), x and $\nabla \log \tilde{q}(x)$ are given as

$$\begin{aligned}
x^i &= \begin{cases} -\frac{2u_\mathrm{d}^i}{1+\sum_{j=1}^n (u_\mathrm{d}^j)^2}, & (i = 1, \ldots, n), \\ 1 - \frac{2}{1+\sum_{j=1}^n (u_\mathrm{d}^j)^2}, & (i = n+1), \end{cases} \\
\nabla \log \tilde{q}(x) &= \left(\frac{\partial x(u_\mathrm{d})}{\partial u_\mathrm{d}} \right) G_\mathrm{d}^{-1} \frac{\partial \log q(x(u_\mathrm{d}))}{\partial u_\mathrm{d}} \\
&= \begin{pmatrix} -\frac{1+\sum_{j=1}^n \left(u_\mathrm{d}^j\right)^2}{2} I_n + u_\mathrm{d} u_\mathrm{d}^\top \\ -u_\mathrm{d}^\top \end{pmatrix} \frac{\partial \log q(x(u_\mathrm{d}))}{\partial u_\mathrm{d}},
\end{aligned}$$

we conclude that η is of class C^1 over U_{d}. Thus, η is C^1.