Riemannian optimization on unit sphere with $p$-norm and its applications

Hiroyuki Sato

Abstract

This study deals with Riemannian optimization on the unit sphere in terms of $p$-norm with general $p > 1$. As a Riemannian submanifold of the Euclidean space, the geometry of the sphere with $p$-norm is investigated, and several geometric tools used for Riemannian optimization, such as retractions and vector transports, are proposed and analyzed. Applications to Riemannian optimization on the sphere with nonnegative constraints and $L_p$-regularization-related optimization are also discussed. As practical examples, the former includes nonnegative principal component analysis, and the latter is closely related to the Lasso regression and box-constrained problems. Numerical experiments verify that Riemannian optimization on the sphere with $p$-norm has substantial potential for such applications, and the proposed framework provides a theoretical basis for such optimization.

Keywords $p$-norm · Sphere · Riemannian optimization · Nonnegative PCA · Lasso regression · Box-constrained optimization

1 Introduction

In the Euclidean space $\mathbb{R}^n$, the $p$-norm of a vector $a \in \mathbb{R}^n$ whose $i$th element is $a_i \in \mathbb{R}$ is defined by

$$\|a\|_p := \sqrt[p]{\sum_{i=1}^{n} |a_i|^p},$$

1 Department of Applied Mathematics and Physics, Kyoto University, Yoshida-honmachi, Sakyo-ku, Kyoto 6068501, Japan

Hiroyuki Sato
hsato@i.kyoto-u.ac.jp
where \( p \geq 1 \) is a real value. When \( p = \infty \), the \( \infty \)-norm, or maximum norm, is defined by

\[
\|a\|_{\infty} := \max \{|a_1|, |a_2|, \ldots, |a_n|\}.
\]

In optimization and related fields, discussions are usually based on the 2-norm (Euclidean norm). The 1-norm is also important in, e.g., Lasso regression for sparse estimation [8]. Furthermore, for \( x \in \mathbb{R}^n \), the constraint \( \|x\|_{\infty} \leq c \) for some \( c \geq 0 \) is equivalent to the box constraint \(-c \leq x_i \leq c\) for all elements \( x_i \) of \( x \).

For \( p \geq 1 \) or \( p = \infty \), we define the unit sphere with \( p \)-norm in \( \mathbb{R}^n \) as

\[
S_p^{n-1} := \{x \in \mathbb{R}^n \mid \|x\|_p = 1\}.
\]

A particularly important and well-studied example is the case of \( p = 2 \), which reduces to the standard (hyper)sphere \( S_2^{n-1} = \{x \in \mathbb{R}^n \mid \|x\|_2 = 1\} \) in the sense of the Euclidean norm. In terms of optimization, as we discuss in Sect. 7, the case of \( p = 2p' \) can be used to implicitly impose the nonnegativity constraints on \( x \in S_{p'}^{n-1} \). A practical example of this is the case of \( p = 4 \) and \( p' = 2 \), which leads to a constrained optimization on the standard unit sphere \( S_2 \) with the constraint \( x \geq 0 \). Furthermore, the case of \( p = 1 \) is closely related to \( L_1 \) regularization in, e.g., Lasso [8], and the case of \( p = \infty \) is closely related to the box constraint.

This study addresses the geometry of \( S_p^{n-1} \) with \( p \in (1, \infty) \) and provides several mathematical tools required for Riemannian optimization, i.e., optimization on Riemannian manifolds, such as retractions and vector transports [1, 21]. A natural and practical retraction is defined through normalization in terms of \( p \)-norm, and we provide mathematical support for the validity of this retraction. Furthermore, we discuss projective and orthographic retractions on \( S_p^{n-1} \). Although it may be harder to use such retractions practically than the retraction based on normalization, their inverses are efficient and easy to implement. Thus, discussing them is meaningful. We also provide an explicit expression for the vector transport defined as the differentiated retraction associated with the retraction by normalization. Other contributions of this paper include applications of the sphere \( S_p^{n-1} \) to practical optimization problems related to, e.g., the nonnegative principal component analysis (PCA) and Lasso regression.

This paper is organized as follows. In Sect. 2, we introduce the notations used. We also review the differentiability and derivative of the \( p \)-norm, which are used throughout this paper. In Sect. 3, we prove that \( S_p^{n-1} \) is a Riemannian submanifold of \( \mathbb{R}^n \) and, as such, investigate its geometry. Section 4 provides a retraction on \( S_p^{n-1} \) based on normalization and its inverse. The respective formulas for the inverses of projective and orthographic retractions are also provided. In Sect. 5, we discuss a vector transport on \( S_p^{n-1} \) derived by differentiating a retraction. We also remark another vector transport based on the orthogonal projection. Section 6 is a reference to the geometric results in this paper. We present two types of applications of Riemannian optimization on \( S_p^{n-1} \) in Sect. 7. One is the application to Riemannian optimization problems on the sphere with the nonnegative constraint, which include nonnegative PCA as an important example.
The other is the application to $L_p$-regularization-related optimization problems, which include the Lasso regression and box-constrained problems. Section 8 concludes the paper.

2 Preliminaries

In this section, we provide preliminaries for the discussion in the later sections.

2.1 Notation

Throughout the paper, we use the following notation. The vector space of $n$-dimensional real column vectors is denoted by $\mathbb{R}^n$. We use the notation $\cdot^T$ to indicate transposition. The $n$-dimensional real vector whose $i$th element is $a_i \in \mathbb{R}$ is denoted by $[a_i] \in \mathbb{R}^n$, and we denote the $i$th element of $b \in \mathbb{R}^n$ by $b_i$ or $(b)_i$. A 2-dimensional column vector whose elements are $a_1$ and $a_2$ is denoted by $[a_1, a_2]^T$, while a closed interval $\{w \in \mathbb{R} \mid \alpha \leq w \leq \beta\}$ for $\alpha, \beta \in \mathbb{R}$ is denoted by $[\alpha, \beta]$. For $a = [a_i] \in \mathbb{R}^n$, we denote the element-wise power of $r \in \mathbb{R}$ by $a^r := [a_i^r] \in \mathbb{R}^n$ and the element-wise absolute value by $|a| := [|a_i|] \in \mathbb{R}^n$. Furthermore, the binary relation $\leq$ (resp. $\geq$) for vectors $a = [a_i], b = [b_i] \in \mathbb{R}^n$ means the element-wise relation $\leq$ (resp. $\geq$), i.e., $a \leq b$ (resp. $a \geq b$) is equivalent to $a_i \leq b_i$ (resp. $a_i \geq b_i$) for $i = 1, 2, \ldots, n$. In particular, $a \geq 0$ means that all elements of $a$ are nonnegative. We define the all-one vector as $\mathbf{1} := [1, 1, \ldots, 1]^T \in \mathbb{R}^n$. Then, the condition $\|x\|_p = 1$ is equivalent to $\|x\|_p = 1$ and rewritten as $\mathbf{1}^T|x|^p = 1$. The identity matrix of $n$th order is denoted by $I$. For $\mathbf{1} \in \mathbb{R}^n$ and $I \in \mathbb{R}^{n\times n}$, the size $n$ is determined by context.

We denote the sign function by $\text{sgn}$, i.e.,

$$\text{sgn}(w) := \begin{cases} 1 & \text{if } w > 0, \\ 0 & \text{if } w = 0, \\ -1 & \text{if } w < 0 \end{cases}$$

for $w \in \mathbb{R}$. Note that $\text{sgn}(w)|w| = w$ always holds. We also use the same notation for the element-wise application of $\text{sgn}$, i.e., for $a = [a_i] \in \mathbb{R}^n$, we define $\text{sgn}(a) := [\text{sgn}(a_i)] \in \mathbb{R}^n$.

The operator $\odot$ denotes the Hadamard product, which is the element-wise product, i.e., for $a = [a_i], b = [b_i] \in \mathbb{R}^n$, we define $a \odot b := [a_ib_i] \in \mathbb{R}^n$. We consider the Hadamard product only for column vectors in this paper. It is clear that the commutative law $a \odot b = b \odot a$ holds. Furthermore, for $c = [c_i] \in \mathbb{R}^n$, we have $a^T(b \odot c) = (a \odot b)^Tc$ because both sides are equal to $\sum_{i=1}^n a_ib_ic_i$. Using these facts, we can rewrite the condition $\|x\|_p = 1$ as $x^T(\text{sgn}(x) \odot |x|^{p-1}) = 1$ because we have

$$x^T(\text{sgn}(x) \odot |x|^{p-1}) = (\text{sgn}(x) \odot x)^T|x|^{p-1} = |x|^T|x|^{p-1} = \mathbf{1}^T|x|^p = \|x\|_p^p.$$
equip $\mathbb{R}^n$ with the standard inner product $\langle a, b \rangle := a^T b$ and the induced norm $\|a\| := \sqrt{\langle a, a \rangle} = \|a\|_2$, which coincides with the 2-norm, even when we discuss the sphere $S_p^{n-1}$ for general $p$. As discussed in Sect. 3, we regard $\mathbb{R}^n$ as a Riemannian manifold with the Riemannian metric induced by the standard inner product and consider $S_p^{n-1}$ for $p \in (1, \infty)$ as a Riemannian submanifold of $\mathbb{R}^n$.

For a manifold $\mathcal{M}$, we denote the tangent space of $\mathcal{M}$ at $x \in \mathcal{M}$ by $T_x \mathcal{M}$. Furthermore, when the manifold $\mathcal{M}$ is a Riemannian manifold with a Riemannian metric $\langle \cdot, \cdot \rangle$, each tangent space $T_x \mathcal{M}$ is endowed with the inner product $\langle \cdot, \cdot \rangle_x$ via the Riemannian metric $\langle \cdot, \cdot \rangle$, and the Riemannian gradient $\text{grad} f(x)$ of a $C^1$ function $f : \mathcal{M} \to \mathbb{R}$ at $x$ is defined as the unique tangent vector at $x$ satisfying $D f(x)[\xi] = \langle \text{grad} f(x), \xi \rangle_x$ for all $\xi \in T_x \mathcal{M}$, where $D f(x) : T_x \mathcal{M} \to T_{f(x)} \mathbb{R} \cong \mathbb{R}$ is the derivative of $f$ at $x \in \mathcal{M}$. For $\mathbb{R}^n$ as a Riemannian manifold with the Riemannian metric $\langle \xi, \eta \rangle_x := \xi^T \eta$ for any $x \in \mathbb{R}^n$ and $\xi, \eta \in T_x \mathbb{R}^n \cong \mathbb{R}^n$, the Riemannian gradient of a function $\tilde{f} : \mathbb{R}^n \to \mathbb{R}$ coincides with the standard Euclidean gradient $\nabla \tilde{f}$, i.e., $\nabla \tilde{f}(x) := [\partial \tilde{f}(x)/\partial x_i] \in T_x \mathbb{R}^n \cong \mathbb{R}^n$ for $x \in \mathbb{R}^n$. Furthermore, if $\mathcal{M}$ is an embedded submanifold of $\mathbb{R}^n$, then we simply say that $\mathcal{M}$ is a submanifold of $\mathbb{R}^n$.

### 2.2 Derivatives of $p$-norm functions

Here, we investigate the derivative or Euclidean gradient of the $p$-norm-related functions in $\mathbb{R}^n$. First, although the $p$-norm is defined for any $p \in [1, \infty]$, it is of class $C^1$ only for $p \in (1, \infty)$. In the remainder of this section, we assume $p \in (1, \infty)$. Then, it is easy to verify that

$$
\frac{d|w|^p}{dw} = p \text{sgn}(w)|w|^{p-1}
$$

for $w \in \mathbb{R}$. Regarding the $p$-norm of $x \in \mathbb{R}^n$, because $\|x\|_p^p = 1^T |x|^p$, its partial derivative with respect to the variable $x_i$ for $i \in \{1, 2, \ldots, n\}$ is

$$
\frac{\partial \|x\|_p^p}{\partial x_i} = \frac{\partial |x_i|^p}{\partial x_i} = p \text{sgn}(x_i)|x_i|^{p-1}.
$$

Therefore, the gradient of the function $x \mapsto \|x\|_p^p$ is equal to

$$
\nabla(x \mapsto \|x\|_p^p)(x) = [p \text{sgn}(x_i)|x_i|^{p-1}] = p \text{sgn}(x) \odot |x|^{p-1}. \quad (1)
$$

In the subsequent sections, we exploit the fact that the conditions $\|x\|_p = 1$ and $\|x\|_p = 1$—both of which characterize the unit sphere $S_p^{n-1}$—are equivalent to each other. Furthermore, $\|x\|_p^p$ usually seems to be easier to handle than $\|x\|_p$. For example,

---

1 We write $p \in [1, \infty]$ to indicate that $p \in [1, \infty)$ or $p = \infty$ holds.
the gradient of the $p$-norm function is computed as
\[
\nabla (x \mapsto \|x\|_p)(x) = \nabla \left( x \mapsto \left( \|x\|_p^p \right)^{\frac{1}{p}} \right)(x)
\]
\[
= \frac{1}{p} (\|x\|_p^{\frac{1}{p}-1} \cdot p \text{sgn}(x) \odot |x|^{p-1})
\]
\[
= \frac{\text{sgn}(x) \odot |x|^{p-1}}{\|x\|_p^{p-1}}.
\]
(2)

We prefer to use (1), which provides a simpler expression, rather than (2), unless (2) is essential in the discussion.

Note that $h(x) := \|x\|_p^p$ is not necessarily a $C^\infty$ function in $\mathbb{R}^n$. For example, consider the case $p = 3$ and $n = 2$, where $h(x) = |x_1|^3 + |x_2|^3$. Then, we have $\nabla h(x) = 3 \begin{bmatrix} |x_1| & |x_1| \\ |x_2| & |x_2| \end{bmatrix}$ and $\nabla^2 h(x) = 6 \begin{bmatrix} |x_1| & 0 \\ 0 & |x_2| \end{bmatrix}$. Hence, $h$ is of class $C^2$ in $\mathbb{R}^2$.

However, because $\frac{\partial^2 h(x)}{\partial x_1^2} = 6|x_1|$ (resp. $\frac{\partial^2 h(x)}{\partial x_2^2} = 6|x_2|$) is not partially differentiable with respect to $x_1$ (resp. $x_2$) at any $[0, x_2]^T \in \mathbb{R}^2$ (resp. $[x_1, 0]^T$), $h$ is not of class $C^3$ in $\mathbb{R}^2$. This causes nonsmoothness of $S^2_3$, which includes the points $[\pm 1, 0]^T$ and $[0, \pm 1]^T$, as a submanifold of $\mathbb{R}^2$. In the next section, we will prove that $S^{n-1}_p$ with $p \in (1, \infty)$ is still at least a $C^1$ submanifold of $\mathbb{R}^n$ (Theorem 1).

### 3 Geometry of $S^{n-1}_p$ and tools for Riemannian optimization

In this section, we discuss the geometry of the unit sphere with $p$-norm, i.e.,
\[
S^{n-1}_p = \{ x \in \mathbb{R}^n \mid \|x\|_p = 1 \},
\]
where $1 < p < \infty$. We use the following equivalent conditions interchangeably:
\[
\|x\|_p = 1 \iff \|x\|_p^p = 1 \iff 1^T |x|^p = 1 \iff x^T (\text{sgn}(x) \odot |x|^{p-1}) = 1.
\]

As expected, many properties of the standard sphere $S^{n-1}_2$ analogically hold for $S^{n-1}_p$ with any $p \in (1, \infty)$, especially even integer $p$, while some do not hold for $S^{n-1}_1$ or $S^{n-1}_\infty$.

#### 3.1 $S^{n-1}_p$ as a Riemannian submanifold of $\mathbb{R}^n$

First, we prove that $S^{n-1}_p$ is a submanifold of $\mathbb{R}^n$.

**Theorem 1** For $p \in (1, \infty)$, the unit sphere $S^{n-1}_p$ with $p$-norm is an $(n-1)$-dimensional $C^r$ embedded submanifold of $\mathbb{R}^n$, where $r = \infty$ if $p$ is an even
integer, \( r = p - 1 \) if \( p \) is an odd integer, and \( r = \lfloor p \rfloor \), which is the largest integer less than \( p \), if \( p \) is not an integer.\(^2\)

**Proof** We define \( h : \mathbb{R}^n \to \mathbb{R} \) as \( h(x) := \|x\|^p \). We can observe that \( h \) is a \( C^r \) function in \( \mathbb{R}^n \), where \( r \) is the integer in the statement of the theorem, as follows: If \( p \) is an even integer, \( h(x) = \sum_{i=1}^n |x_i|^p = \sum_{i=1}^n x_i^p \) is clearly a \( C^\infty \) function. If \( p \) is an odd integer, \( h(x) = \sum_{i=1}^n |x_i|^p \) is of class \( C^{p-1} \) because \( \partial^{p-1} h(x) / \partial x_i^{p-1} = (p!)|x_i| \) is continuous for any \( i \in \{1, 2, \ldots, n\} \). Similarly, if \( p \) is not an integer, \( h \) is of class \( C^{\lfloor p \rfloor} \) because we have

\[
\frac{\partial |p|!h}{\partial x_i^{\lfloor p \rfloor}}(x) = p(p-1) \cdots (p-\lfloor p \rfloor+1) \sgn(x_i)|x_i|^{p-\lfloor p \rfloor}
\]

\[
= \frac{\Gamma(p+1)}{\Gamma(p+1-\lfloor p \rfloor)} \sgn(x_i)|x_i|^{p-\lfloor p \rfloor},
\]

which is continuous because \( p - \lfloor p \rfloor > 0 \) in this case, where \( \Gamma(\cdot) \) is the gamma function. Therefore, \( h \) is of class \( C^r \) in every case.

Using the formula (1), the Jacobian matrix of \( h \) at \( x \in \mathbb{R}^n - \{0\} \), which is defined as \((Jh)_x := [\partial h(x)/\partial x_i] \in \mathbb{R}^{1 \times n}\), is computed as

\[
(Jh)_x = \nabla h(x)^T = p(\sgn(x) \odot |x|^{p-1})^T.
\]

For any \( x \in \mathbb{R}^n \) satisfying \( h(x) = 1 \), we have \((Jh)_x \neq 0 \) because such \( x \) is not 0. This implies that 1 is a regular value of \( h \). Therefore, it follows from the regular level set theorem [27, Theorem 9.9] that \( h^{-1}([-1]) = S^{n-1}_p \) is a \( C^r \) submanifold of \( \mathbb{R}^n \), whose dimension is \( n - \dim \mathbb{R} = n - 1 \). This completes the proof. \( \square \)

**Remark 1** Note that the integer \( r \) in Theorem 1 is greater than or equal to 1 in every case. Therefore, \( S^{n-1}_p \) with \( p \in (1, \infty) \) is always at least a \( C^1 \) submanifold of \( \mathbb{R}^n \). In contrast, if \( p = 1 \) or \( p = \infty \), the unit sphere \( S^{n-1}_p \) is not a \( C^1 \) embedded submanifold of \( \mathbb{R}^n \) because of their corners. Indeed, the above proof fails if \( p = 1 \) or \( p = \infty \) because \( x \mapsto \|x\|^p \) is not a \( C^1 \) function in such cases.

In what follows, we assume \( p \in (1, \infty) \) and define smoothness regarding \( S^{n-1}_p \) as \( C^r \) with \( r \ge 1 \) in Theorem 1.\(^3\) For example, we say that a function \( f \) on \( S^{n-1}_p \) is smooth if \( f \) is of class \( C^r \).

We endow the sphere \( S^{n-1}_p \) with the Riemannian metric as

\[
\langle \xi, \eta \rangle_x := \xi^T \eta, \ \xi, \ \eta \in T_x S^{n-1}_p, \ \ x \in S^{n-1}_p,
\]

which is induced from the Riemannian metric (the standard inner product)

\[
\langle a, b \rangle_x := a^T b, \ a, b \in T_x \mathbb{R}^n \simeq \mathbb{R}^n, \ x \in \mathbb{R}^n
\]

\(^2\) The statement can be rewritten as follows: for any positive integer \( k \), \( S^{n-1}_p \) is a \( C^{2k-1} \) submanifold of \( \mathbb{R}^n \) if \( 2k - 1 < p < 2k \), \( C^\infty \) submanifold if \( p = 2k \), and \( C^{2k} \) submanifold if \( 2k < p \le 2k + 1 \).

\(^3\) Since the tangent bundle \( T \mathcal{M} \) of a \( C^r \) manifold is a \( C^{r-1} \) manifold, we say that a map defined on \( T \mathcal{M} \) is smooth if it is of class \( C^{r-1} \).

\( \odot \) Springer
in the ambient space $\mathbb{R}^n$. Thus, $S_p^{n-1}$ is a Riemannian submanifold of $\mathbb{R}^n$.

### 3.2 Tangent space, normal space, and orthogonal projection

Defining $h(x) := \|x\|_p$, the tangent space $T_x S_p^{n-1}$ of $S_p^{n-1} = h^{-1}(\{1\})$ at $x$ is equal to the kernel of the linear map $Dh(x) : \mathbb{R}^n \cong T_x \mathbb{R}^n \to T_{h(x)} \mathbb{R} \cong \mathbb{R}$, i.e., $(Dh(x))^{-1}(\{0\})$. Here, it follows from (3) that the derivative $Dh(x)$ acts on $y \in \mathbb{R}^n$ as

$$Dh(x)[y] = (Jh)_x(y) = p(\text{sgn}(x) \odot |x|^{p-1})^T y.$$ 

Therefore, we have

$$T_x S_p^{n-1} = (Dh(x))^{-1}(\{0\}) = \{\xi \in \mathbb{R}^n \mid \xi^T (\text{sgn}(x) \odot |x|^{p-1}) = 0\}. \quad (4)$$

Since $S_p^{n-1}$ is a Riemannian submanifold of $\mathbb{R}^n$, we can define the normal space $N_x S_p^{n-1}$ of $S_p^{n-1}$ at a point $x$ as the orthogonal complement of $T_x S_p^{n-1} \subset T_x \mathbb{R}^n \cong \mathbb{R}^n$ in $\mathbb{R}^n$ with respect to the Riemannian metric in $\mathbb{R}^n$, i.e., the standard inner product. From the expression (4), we can observe that $T_x S_p^{n-1}$ is a hyperplane orthogonal to the vector $\text{sgn}(x) \odot |x|^{p-1} \in \mathbb{R}^n$. Hence, we have

$$N_x S_p^{n-1} := (T_x S_p^{n-1})^\perp = \{\alpha \text{sgn}(x) \odot |x|^{p-1} \mid \alpha \in \mathbb{R}\}. \quad (5)$$

For minimizing a smooth function $f : S_p^{n-1} \to \mathbb{R}$ on $S_p^{n-1}$, the Riemannian gradient of $f$ is important. Here, the Riemannian gradient $\text{grad } f(x)$ of $f$ at $x \in S_p^{n-1}$ can be obtained by orthogonally projecting $\nabla \tilde{f}(x) \in \mathbb{R}^n$ onto the tangent space $T_x S_p^{n-1}$ at $x$, where $\tilde{f}$ is a smooth extension of $f$ to the ambient space $\mathbb{R}^n$ and $\nabla \tilde{f}(x) := [\partial \tilde{f}(x)/\partial x_i] \in \mathbb{R}^n$ is the Euclidean gradient. That is, we have

$$\text{grad } f(x) = P_x(\nabla \tilde{f}(x)),$$

where $P_x$ is the orthogonal projection to the tangent space $T_x S_p^{n-1}$ at $x$. The projection $P_x : \mathbb{R}^n \to T_x S_p^{n-1}$ acts on any $d \in \mathbb{R}^n$ so that $d - P_x(d) \in N_x S_p^{n-1}$ holds. From (5), the normal vector $d - P_x(d)$ is written as $\alpha \text{sgn}(x) \odot |x|^{p-1}$ for some $\alpha \in \mathbb{R}$. Thus, we obtain the decomposition of $d$ as

$$d = P_x(d) + \alpha \text{sgn}(x) \odot |x|^{p-1}. \quad (6)$$

By noting the expression (4) and multiplying (6) by $(\text{sgn}(x) \odot |x|^{p-1})^T$ from the left, we obtain $\alpha = ((\text{sgn}(x) \odot |x|^{p-1})^T d)/\| |x|^{p-1} \|^2_2$, where we used the relation

$$(\text{sgn}(x) \odot |x|^{p-1})^T (\text{sgn}(x) \odot |x|^{p-1}) = ((\text{sgn}(x))^2)^T (|x|^{p-1})^2 = \| |x|^{p-1} \|^2_2 \neq 0.$$
Substituting the expression of $\alpha$ to (6), we obtain

$$P_x(d) = d - \frac{(\text{sgn}(x) \odot |x|^{p-1})^T d}{\|x|^{p-1}\|_2^2} \text{sgn}(x) \odot |x|^{p-1}$$

$$= \left( I - \frac{(\text{sgn}(x) \odot |x|^{p-1})(\text{sgn}(x) \odot |x|^{p-1})^T}{\|x|^{p-1}\|_2^2} \right) d.$$  

In other words, the linear map $P_x$ is represented as the matrix

$$P_x = I - \frac{(\text{sgn}(x) \odot |x|^{p-1})(\text{sgn}(x) \odot |x|^{p-1})^T}{\|x|^{p-1}\|_2^2}. \tag{7}$$

## 4 Retractions and their inverses

In an iterative Riemannian optimization algorithm on a Riemannian manifold $\mathcal{M}$, to compute the next point from the current point $x \in \mathcal{M}$ and search direction $\eta \in T_x \mathcal{M}$, a retraction on $\mathcal{M}$ is important [1, 3, 25]. A map $R : T \mathcal{M} \rightarrow \mathcal{M}$ is said to be a retraction on a $C^\infty$ manifold $\mathcal{M}$ if the restriction $R_x := R|_{T_x \mathcal{M}}$ of $R$ to $T_x \mathcal{M}$ for $x \in \mathcal{M}$ satisfies $R_x(0_x) = x$ and $DR_x(0_x) = \text{id}_{T_x \mathcal{M}}$, where $0_x$ is the zero vector in $T_x \mathcal{M}$ and $\text{id}_{T_x \mathcal{M}}$ is the identity map in $T_x \mathcal{M}$. Although retractions are usually discussed on $C^\infty$ manifolds, the manifold $S^{n-1}_p$ is a $C^r$ submanifold of $\mathbb{R}^n$, where $r$ is in Theorem 1 and may not be $\infty$. Therefore, analogous to the above definition of a retraction on a $C^\infty$ manifold [1], we define a retraction on $S^{n-1}_p$ as a $C^{r-1}$, which we say smooth, map on $T S^{n-1}_p$ satisfying the above properties.

Furthermore, the inverse of a retraction can be used in, e.g., the Riemannian conjugate gradient method [22, 30]. In the following, we discuss three types of retractions on $S^{n-1}_p$ and their respective inverses.

### 4.1 Retraction by normalization and its inverse

Intuitively, for any $x \in S^{n-1}_p$ and $\eta \in T_x S^{n-1}_p$, $x + \eta \in \mathbb{R}^n$ appears to be outside $S^{n-1}_p$ unless $\eta = 0$. This is actually true from the following proposition, based on which we can construct a retraction by normalizing $x + \eta$ (see Proposition 3). However, its proof for general $p > 1$ is not as easy as in the case of $p = 2$.

**Proposition 2** Assume that $p \in (1, \infty)$. For any $x \in S^{n-1}_p$ and $\eta \in T_x S^{n-1}_p$, if $\eta \neq 0$, then $\|x + \eta\|_p > 1$ holds.

**Proof** Note that the function $h(y) := \|y\|_p^p$ is not of class $C^2$ in the entire $\mathbb{R}^n$ when $1 < p < 2$. Therefore, we avoid using the Hessian matrix in the following discussion to address the general case.

We first show that $h$ is a strictly convex function in $\mathbb{R}^n$. For $y, z \in \mathbb{R}^n$ with $y \neq z$ and $\alpha \in (0, 1)$, Minkowski’s inequality (the triangle inequality for the $p$-norm) as well as monotonicity and convexity of the function $w \mapsto w^p$ on $\mathbb{R}_+ := \{w \in \mathbb{R} \mid w \geq 0\}$
yield that
\[
\|\alpha y + (1 - \alpha)z\|_p^p \leq (\alpha\|y\|_p + (1 - \alpha)\|z\|_p)^p \leq \alpha\|y\|_p^p + (1 - \alpha)\|z\|_p^p. \tag{8}
\]

We now assume that both equalities in (8) simultaneously hold. Then, the first equality implies that \(y = cz\) for some \(c \geq 0\) or \(z = 0\) from Minkowski’s inequality theory for \(p \in (1, \infty)\). Furthermore, from the second equality and the strict convexity of \(w \mapsto w^p\) on \(\mathbb{R}_+\), we have \(\|y\|_p = \|z\|_p\). If \(y = cz\) with \(c \geq 0\), then \(\|y\|_p = \|z\|_p\) implies \(c = 1\) or \(\|y\|_p = \|z\|_p = 0\). Otherwise, we have \(z = 0\); and \(\|y\|_p = \|z\|_p\) then means \(y = z = 0\). In any case, we have \(y = z\), which contradicts the assumption that \(y \neq z\). Therefore, both equalities in (8) do not hold at the same time, meaning
\[
\|\alpha y + (1 - \alpha)z\|_p^p < \alpha\|y\|_p^p + (1 - \alpha)\|z\|_p^p
\]
\[
= \alpha h(y) + (1 - \alpha)h(z).
\]
This proves that \(h\) is strictly convex.

By using the strict convexity of \(h\), we can show that \(\phi(t) := h(x + t\eta) = \|x + t\eta\|_p^p\) is a strictly convex function on \(\mathbb{R}\) for \(x \in S^n_{p-1}\) and \(\eta \in T_xS^n_{p-1}\) with \(\eta \neq 0\). Indeed, for any \(s, t \in \mathbb{R}^\ast\) with \(s \neq t\) and \(\alpha \in (0, 1)\), it follows from the strict convexity of \(h\) and the fact \(x + s\eta \neq x + t\eta\) that
\[
\phi(\alpha s + (1 - \alpha)t) = h(x + (\alpha s + (1 - \alpha)t)\eta)
\]
\[
= h(\alpha(x + s\eta) + (1 - \alpha)(x + t\eta))
\]
\[
< \alpha h(x + s\eta) + (1 - \alpha)h(x + t\eta)
\]
\[
= \alpha\phi(s) + (1 - \alpha)\phi(t).
\]
Subsequently, we show that \(t = 0\) is the unique minimizer of \(\phi\). Since \(\phi\) is strictly convex, it suffices to prove that \(\phi'(0) = 0\), which is shown as
\[
\phi'(0) = \nabla h(x)^T \eta = p(\text{sgn}(x) \otimes |x|^{p-1})^T \eta = 0
\]
from (1) and (4).

In conclusion, we obtain \(\|x + t\eta\|_p^p = \phi(t) > \phi(0) = \|x\|_p^p = 1\) for all \(t \neq 0\), where the case of \(t = 1\) implies that the desired inequality \(\|x + \eta\|_p > 1\) holds. \(\square\)

Considering Proposition 2, we propose a retraction \(R\) on \(S^n_{p-1}\) as
\[
R_x(\eta) := \frac{x + \eta}{\|x + \eta\|_p}, \quad \eta \in T_xS^n_{p-1}, \quad x \in S^n_{p-1}. \tag{9}
\]
This is simply the normalization (with respect to the \(p\)-norm) of \(x + \eta\), which is not on \(S^n_{p-1}\) when \(\eta \neq 0\). Note that the denominator in (9) is ensured to be nonzero from Proposition 2.
Proposition 3 Assume that \( p \in (1, \infty) \). The map \( R \) defined by (9) is a retraction on \( S_p^{n-1} \).

Proof It is clear that \( \| R_x(\eta) \|_p = 1 \) and \( R_x(0_x) = x \) hold for any \( x \in S_p^{n-1} \) and \( \eta \in T_x S_p^{n-1} \). To prove that \( DR_x(0_x) = \text{id}_{T_x S_p^{n-1}} \) holds, we use (2), i.e., the fact that the gradient of \( x \mapsto \| x \|_p \) is written as \( \| x \|_p^{1-p} \text{sgn}(x) \odot |x|^{p-1} \). Then, we can compute \( DR_x(0_x)[\eta] \) for \( \eta \in T_x S_p^{n-1} \) as

\[
DR_x(0_x)[\eta] = \frac{d}{dt} R_x(t\eta) \bigg|_{t=0} = \frac{\eta \| x + t\eta \|_p - (x + t\eta) \big( \| x + t\eta \|_p^{1-p} \text{sgn}(x + t\eta) \odot |x + t\eta|^{p-1} \big)^T \eta}{\| x + t\eta \|_p^2} \bigg|_{t=0} = \eta - x (\text{sgn}(x) \odot |x|^{p-1})^T \eta = \eta,
\]

where we used \( \| x \|_p = 1 \) and \((\text{sgn}(x) \odot |x|^{p-1})^T \eta = 0\) from (4).

To derive the inverse of \( R \), we fix \( x, y \in S_p^{n-1} \) and assume that \( \eta \in T_x S_p^{n-1} \) satisfies \( R_x(\eta) = y \). Then, \( \eta \) should satisfy \( x + \eta = \alpha y \) for some \( \alpha > 0 \). Multiplying the equality by \( (\text{sgn}(x) \odot |x|^{p-1})^T \) from the left and noting that \( x \in S_p^{n-1} \) and \( \eta \in T_x S_p^{n-1} \), we obtain \( \alpha = 1/(\text{sgn}(x) \odot |x|^{p-1})^T y \). Therefore, \( \eta \) should satisfy

\[
\eta = \alpha y - x = \frac{y}{(\text{sgn}(x) \odot |x|^{p-1})^T y} - x.\tag{11}\]

However, this is necessary but not sufficient for \( R_x(\eta) = y \). In fact, for certain \( x, y \in S_p^{n-1} \), there may not exist \( \eta \) such that \( R_x(\eta) = y \). The following proposition elaborates on this issue.

Specifically, when separating \( \mathbb{R}^n \) into two half spaces by the hyperplane \( T_x S_p^{n-1} \) (regarded as passing through the origin), \( x \) and \( y \) must be in the same half space.

Proposition 4 Assume that \( p \in (1, \infty) \). For any \( x \in S_p^{n-1} \), the inverse of \( R_x \) defined in (9) is given by

\[
R_x^{-1}(y) = \frac{y}{(\text{sgn}(x) \odot |x|^{p-1})^T y} - x, \quad y \in D_x,\tag{12}\]

where the domain \( D_x \) of \( R_x^{-1} \) is \( D_x = \{ y \in S_p^{n-1} \mid (\text{sgn}(x) \odot |x|^{p-1})^T y > 0 \} \).

Proof For \( y \) satisfying \((\text{sgn}(x) \odot |x|^{p-1})^T y = 0 \), the right-hand side of (12) is not defined. We assume that \((\text{sgn}(x) \odot |x|^{p-1})^T y \neq 0 \) and denote the right-hand side of (12) by \( \eta_{x,y} \), which is in \( T_x S_p^{n-1} \) because \((\text{sgn}(x) \odot |x|^{p-1})^T \eta_{x,y} = 1 - 1 = 0 \).
Then, we have

\[ R_x(\eta_{x,y}) = \frac{x + \eta_{x,y}}{\parallel x + \eta_{x,y} \parallel_p} \]

\[ = \frac{y}{\text{sgn}((\text{sgn}(x) \odot |x|^{p-1})^T y)} \]

\[ = \begin{cases} 
  y & \text{if } (\text{sgn}(x) \odot |x|^{p-1})^T y > 0 \\
  -y & \text{if } (\text{sgn}(x) \odot |x|^{p-1})^T y < 0
\end{cases} \]

Furthermore, if \( \eta \in T_x S_p^{n-1} \) satisfies \( R_x(\eta) = y \), then \( \eta \) should be equal to \( \eta_{x,y} \), as discussed in (11). Therefore, \( R_x(\eta) = y \) holds if and only if \( (\text{sgn}(x) \odot |x|^{p-1})^T y > 0 \) and \( \eta = \eta_{x,y} \). This completes the proof.

\[ \square \]

4.2 Inverse of projective retraction

Another natural retraction is the projective retraction [2]. The projective retraction \( R_{\text{proj}} \) on \( S_p^{n-1} \) is given by

\[ R_{\text{proj}}^x(\eta) = \arg \min_{y \in S_p^{n-1}} \parallel (x + \eta) - y \parallel_2, \quad \eta \in T_x S_p^{n-1}, \quad x \in S_p^{n-1}. \] (13)

**Remark 2** Note that the projection of a point onto any closed convex set in \( \mathbb{R}^n \) regarding the 2-norm uniquely exists [5, Section 8.1]. Therefore, because the unit ball \( B_p^n := \{ x \in \mathbb{R}^n \mid \parallel x \parallel_p \leq 1 \} \) with \( p \)-norm is obviously a closed convex set in \( \mathbb{R}^n \), vector \( y \in B_p^n \) that minimizes the distance \( \parallel (x + \eta) - y \parallel_2 \) uniquely exists for a given \( x \in S_p^{n-1} \) and \( \eta \in T_x S_p^{n-1} \). Since \( x + \eta \) is outside \( B_p^{n-1} \) unless \( \eta = 0 \) from Proposition 2, the right-hand side in (13) is equal to the uniquely existing projection of \( x + \eta \) onto \( B_p^n \) (clearly, we have \( R_{\text{proj}}^x(\eta) = x \) when \( \eta = 0 \)).

The vector \( R_{\text{proj}}^x(\eta) \) satisfies \( (x + \eta) - R_{\text{proj}}^x(\eta) \in N_{R_{\text{proj}}^x(\eta)} S_p^{n-1} \), which is implied by [2] or is a direct consequence of the Lagrange multiplier method. Therefore, there exists \( \alpha \in \mathbb{R} \) such that

\[ R_{\text{proj}}^x(\eta) = x + \eta - \alpha \text{sgn}(R_{\text{proj}}^x(\eta)) \odot |R_{\text{proj}}^x(\eta)|^{p-1}, \] (14)

where \( \alpha \) is determined such that \( R_{\text{proj}}^x(\eta) \in S_p^{n-1} \) holds, i.e.,

\[ \parallel x + \eta - \alpha \text{sgn}(R_{\text{proj}}^x(\eta)) \odot |R_{\text{proj}}^x(\eta)|^{p-1} \parallel_p = 1. \] (15)

However, it may be difficult to explicitly express \( R_{\text{proj}}^x(\eta) \) by solving (14) and (15).

**Remark 3** When \( p = 2 \), Eq. (14) is reduced to \( R_{\text{proj}}^x(\eta) = x + \eta - \alpha R_{\text{proj}}^x(\eta) \), i.e., we have \( (\alpha + 1)R_{\text{proj}}^x(\eta) = (x + \eta) \). Then, \( \parallel R_{\text{proj}}^x(\eta) \parallel_2 = 1 \) implies \( |\alpha + 1| = \parallel x + \eta \parallel_2 \).
Hence, we obtain \((x+\eta)/(\alpha+1) = \pm(x+\eta)/\|x+\eta\|_2\), among which \((x+\eta)/\|x+\eta\|_2\) is closer to \(x + \eta\). In summary, when \(p = 2\), we have \(R_{x}^{\text{proj}}(\eta) = (x + \eta)/\|x + \eta\|_2\), which is equal to the retraction by normalization in Sect. 4.1.

Although the above discussion implies that the projective retraction on \(S_p^{n-1}\) for general \(p \in (1, \infty)\) may not provide as successful a result as the retraction by normalization, the inverse of \(R_{x}^{\text{proj}}\) can be discussed more practically. For given \(x, y \in S_p^{n-1}\), if \(\eta \in T_x S_p^{n-1}\) satisfies \(R_{x}^{\text{proj}}(\eta) = y\), then \(x + \eta - y \in N_y S_p^{n-1}\) should hold. Therefore, there exists \(\alpha_{x,y} \in \mathbb{R}\) such that \(x + \eta - y = \alpha_{x,y} \text{sgn}(y) \odot |y|^{p-1}\). From \(x \in S_p^{n-1}\) and \(\eta \in T_x S_p^{n-1}\), we can obtain an explicit expression for \(\alpha_{x,y}\) as in Proposition 5.

Furthermore, it seems that \(\alpha_{x,y}\) should be nonnegative by analogy with the discussion of the case \(p = 2\) in Remark 3. We discuss these rigorously in the proof of the proposition using the Karush–Kuhn–Tucker (KKT) conditions.

**Proposition 5** Assume that \(p \in (1, \infty)\). For any \(x \in S_p^{n-1}\), the inverse of \(R_{x}^{\text{proj}}\) in (13) is given by
\[
(R_{x}^{\text{proj}})^{-1}(y) = y - x + \alpha_{x,y} \text{sgn}(y) \odot |y|^{p-1}
\]
\[
= \left( I - \frac{(\text{sgn}(y) \odot |y|^{p-1})(\text{sgn}(x) \odot |x|^{p-1})^T}{(\text{sgn}(y) \odot |y|^{p-1})^T(\text{sgn}(x) \odot |x|^{p-1})} \right) (y - x), \quad y \in D_x,
\]

(16)

where
\[
\alpha_{x,y} := \frac{1 - (\text{sgn}(x) \odot |x|^{p-1})^T y}{(\text{sgn}(x) \odot |x|^{p-1})^T(\text{sgn}(y) \odot |y|^{p-1})}
\]

(17)

and the domain of \((R_{x}^{\text{proj}})^{-1}\) is
\[
D_x = \{ y \in S_p^{n-1} \mid (\text{sgn}(x) \odot |x|^{p-1})^T(\text{sgn}(y) \odot |y|^{p-1}) \neq 0, \alpha_{x,y} \geq 0 \}. \quad (18)
\]

**Proof** The second equality in (16) directly follows from \((\text{sgn}(x) \odot |x|^{p-1})^T x = 1\).

We define \(\eta_{x,y} := y - x + \alpha_{x,y} \text{sgn}(y) \odot |y|^{p-1}\) with \(\alpha_{x,y}\) in (17). Then, what we need to prove is that \(R_{x}^{\text{proj}}(\eta) = y\) holds for \(\eta \in T_x S_p^{n-1}\) if and only if \(y\) belongs to the right-hand side of (18) and \(\eta = \eta_{x,y}\).

To see this in light of (13) and Remark 2, we must verify that \(z = y\) is the optimal solution to the following optimization problem with a fixed \(\eta \in T_x S_p^{n-1}\) if and only if \(\alpha_{x,y}\) in (17) is well-defined and nonnegative and \(\eta = \eta_{x,y}\):

\[
\begin{align*}
\text{minimize} \quad & \|x + \eta - z\|_2^2 \\
\text{subject to} \quad & \|z\|_p^p \leq 1, \quad z \in \mathbb{R}^n,
\end{align*}
\]

where the decision variable vector is \(z\). This is a convex optimization problem because both \(z \mapsto \|x + \eta - z\|_2^2\) and \(z \mapsto \|z\|_p^p - 1\) are convex. Furthermore, the problem
satisfies Slater’s condition [5, Section 5.2.3], i.e., it is strictly feasible (e.g., with \( z = 0 \)). Therefore, the condition that \( z = y \) is optimal for the optimization problem is equivalent to saying that there exists \( \lambda \in \mathbb{R} \) such that \( z = y \) and \( \lambda \) satisfy the KKT conditions for the problem, which are written as

\[
2(y - (x + \eta)) + \lambda p \operatorname{sgn}(y) \odot |y|^{p-1} = 0, \\
\|y\|_p^p \leq 1, \\
\lambda \geq 0, \\
\lambda(\|y\|_p^p - 1) = 0.
\]

Since \( \|y\|_p = 1 \), they are equivalent to

\[
\eta = y - x + \frac{p}{2} \lambda \operatorname{sgn}(y) \odot |y|^{p-1}, \\
\lambda \geq 0.
\]

(19) (20)

Noting that \( x \in S_p^{n-1} \) and \( \eta \in T_x S_p^{n-1} \), we multiply (19) by \( (\operatorname{sgn}(x) \odot |x|^{p-1})^T \) from the left to obtain

\[
2(1 - (\operatorname{sgn}(x) \odot |x|^{p-1})^T y) = p\lambda (\operatorname{sgn}(x) \odot |x|^{p-1})^T (\operatorname{sgn}(y) \odot |y|^{p-1}).
\]

If \( (\operatorname{sgn}(x) \odot |x|^{p-1})^T (\operatorname{sgn}(y) \odot |y|^{p-1}) = 0 \) holds, i.e., \( \operatorname{sgn}(y) \odot |y|^{p-1} \in T_x S_p^{n-1} \) holds, then \( 1 - (\operatorname{sgn}(x) \odot |x|^{p-1})^T y = 0 \) should hold, and \( \lambda \) can be any value. However, it then follows from (19) that \( y - x \in T_x S_p^{n-1} \), i.e., \( y = x + \xi \) for some \( \xi \in T_x S_p^{n-1} \). This, together with \( \|y\|_p = 1 \) and Proposition 2, implies that \( \xi = 0 \) and \( y = x \), contradicting the assumption \( (\operatorname{sgn}(x) \odot |x|^{p-1})^T (\operatorname{sgn}(y) \odot |y|^{p-1}) = 0 \). Hence, we have \( (\operatorname{sgn}(x) \odot |x|^{p-1})^T (\operatorname{sgn}(y) \odot |y|^{p-1}) \neq 0 \), and \( \lambda \) is written as

\[
\lambda = \frac{2}{p} \frac{1 - (\operatorname{sgn}(x) \odot |x|^{p-1})^T y}{(\operatorname{sgn}(x) \odot |x|^{p-1})^T (\operatorname{sgn}(y) \odot |y|^{p-1})} = \frac{2}{p} \alpha_{x,y}.
\]

Therefore, there exists \( \lambda \in \mathbb{R} \) such that \( z = y \) and \( \lambda \) satisfy the KKT conditions (19) and (20) if and only if \( \eta = y - x + \alpha_{x,y} \operatorname{sgn}(y) \odot |y|^{p-1} = \eta_{x,y} \) and \( \alpha_{x,y} \) is well-defined and nonnegative. This completes the proof. \( \square \)

### 4.3 Inverse of orthographic retraction

Other possibilities of retractions on \( S_p^{n-1} \) include the orthographic retraction. See [2] for a discussion of orthographic retractions on general Riemannian submanifolds.

For \( x \in S_p^{n-1} \) and \( \eta \in T_x S_p^{n-1} \), the orthographic retraction \( R_{\text{orth}} \) is defined to satisfy \( R_{\text{orth}}(\eta) = x + \eta + \xi \in S_p^{n-1} \) for some \( \xi \in N_x S_p^{n-1} \) with the smallest norm among all normal vectors in \( \{ \xi \in N_x S_p^{n-1} | x + \eta + \xi \in S_p^{n-1} \} \). Since we can express \( \xi \in N_x S_p^{n-1} \) as \( \xi = -\alpha \operatorname{sgn}(x) \odot |x|^{p-1} \) for some \( \alpha \in \mathbb{R} \), the relation \( \| R_{\text{orth}}(\eta) \|_p = 1 \)
yields the equation on $\alpha$ as
\[ \|x + \eta - \alpha \text{sgn}(x) \odot |x|^{p-1}\|_p = 1. \] (21)

**Remark 4** When $p = 2$, Eq. (21) is reduced to $(1-\alpha)^2 + \eta^T \eta = 1$, the smaller solution (with smaller absolute value) of which is given by $\alpha = 1 - \sqrt{1 - \eta^T \eta}$ if $\|\eta\|_2 \leq 1$. This gives the expression $R_x^{\text{orth}}(\eta) = \sqrt{1 - \eta^T \eta} x + \eta$, which is a well-known result.

For general $p \in (1, \infty)$, we have
\[ R_x^{\text{orth}}(\eta) = x + \eta - \alpha \text{sgn}(x) \odot |x|^{p-1}, \quad \eta \in T_x S_{p}^{n-1}, \quad x \in S_{p}^{n-1}, \]
where $\eta$ should be a tangent vector such that Eq. (21) has a solution, i.e., the norm of $\eta$ is sufficiently small (see Proposition 7 below for more detail), and $\alpha$ is the one with the smallest absolute value of the solutions. Furthermore, when such $\alpha$ exists, it should be nonnegative. Indeed, if we assume that there exists some $\alpha < 0$ satisfying (21), then for $\lambda := (1-\alpha\|x|^{p-1}\|_2^{-1})^{-1} \in (0, 1)$, we have $\|\lambda(x + \eta - \alpha \text{sgn}(x) \odot |x|^{p-1}\|_p = \lambda \|x + \eta - \alpha \text{sgn}(x) \odot |x|^{p-1}\| = \lambda < 1$. On the other hand, we have $\lambda(x + \eta - \alpha \text{sgn}(x) \odot |x|^{p-1}) = x + \xi$, where $\xi := \lambda(x + \eta - \alpha \text{sgn}(x) \odot |x|^{p-1}) - x \in T_x S_p^{n-1}$ holds because the definition of $\lambda$ implies $(\text{sgn}(x) \odot |x|^{p-1})^T \xi = \lambda(1-\alpha\|x|^{p-1}\|_2^{-1})^{-1} = 1$. Therefore, $\lambda(x + \eta - \alpha \text{sgn}(x) \odot |x|^{p-1})$ is the addition of $x$ and tangent vector $\xi \in T_x S_p^{n-1}$, and it follows from Proposition 2 that $\|\lambda(x + \eta - \alpha \text{sgn}(x) \odot |x|^{p-1}\|_p \geq 1$, leading to a contradiction. Thus, $\alpha \geq 0$ holds.

Unfortunately, as in the projective retraction in Sect. 4.2, it may be difficult to explicitly express such $\alpha$ for general $p$. However, the discussion on this retraction is still important because its inverse can be practically computed. Here, we assume that $\eta \in T_x S_{p}^{n-1}$ satisfies $R_x^{\text{orth}}(\eta) = y$ for given $x, y \in S_{p}^{n-1}$. Then, there exists $\alpha_{x,y} \geq 0$ such that $x + \eta - \alpha_{x,y} \text{sgn}(x) \odot |x|^{p-1} = y$. Since $\eta \in T_x S_{p}^{n-1}$, multiplying both sides by $(\text{sgn}(x) \odot |x|^{p-1})^T$ from the left yields
\[ \alpha_{x,y} = \frac{1 - (\text{sgn}(x) \odot |x|^{p-1})^T y}{(\text{sgn}(x) \odot |x|^{p-1})^T (\text{sgn}(x) \odot |x|^{p-1})} = \frac{1 - (\text{sgn}(x) \odot |x|^{p-1})^T y}{\| |x|^{p-1}\|_2^2}. \] (22)

Note that the denominator is nonzero because of $x \neq 0$. This observation, together with the discussion on when (22) is sufficient for $R_x^{\text{orth}}(\eta) = y$, leads to the following proposition.

**Proposition 6** Assume that $p \in (1, \infty)$. For any $x \in S_{p}^{n-1}$, the inverse of the retraction $R_x^{\text{orth}}$ is given by
\[ (R_x^{\text{orth}})^{-1}(y) = y - x + \alpha_{x,y} \text{sgn}(x) \odot |x|^{p-1}, \quad y \in D_x, \] (23)
where
\[ \alpha_{x,y} := \frac{1 - (\text{sgn}(x) \odot |x|^{p-1})^T y}{\| |x|^{p-1}\|_2^2}. \] (24)
and the domain of \((R^\text{orth}_x)^{-1}\) is

\[D_x = \{ y \in S_p^{n-1} \mid \alpha = \alpha_{x,y} \text{ is the solution to } \| y + (\alpha_{x,y} - \alpha) \text{ sgn}(x) \odot |x|^{p-1} \|= 1 \text{ with the smallest value} \}. \tag{25}\]

**Proof** For arbitrarily fixed \(x \in S_p^{n-1}\), let \(y\) be any point on \(S_p^{n-1}\) belonging to the right-hand side of (25) and let \(\eta_{x,y} := y - x + \alpha_{x,y} \text{ sgn}(x) \odot |x|^{p-1}\) be the right-hand side of (23) with \(\alpha_{x,y}\) in (24). We show that (23) holds. From the above discussion on (22), if \(\eta \in T_x S_p^{n-1}\) satisfies \(R^\text{orth}_x(\eta) = y\), then \(\eta = \eta_{x,y}\) should hold. To complete the proof of (23), we conversely show that \(R^\text{orth}_x(\eta) = y\) holds when \(y\) belongs to the right-hand side of (25) and \(\eta = \eta_{x,y}\) holds. Assume that \(y\) and \(\eta\) are such vectors, i.e., \(\alpha = \alpha_{x,y}\) is the solution to the equation \(\| y + (\alpha_{x,y} - \alpha) \text{ sgn}(x) \odot |x|^{p-1} \|= 1\) with the smallest value and \(\eta = \eta_{x,y}\). Then, from the definition of the orthographic retraction and the expression of \(\eta_{x,y}\), we have

\[R^\text{orth}_x(\eta) = R^\text{orth}_x(\eta_{x,y}) = x + \eta_{x,y} - \alpha \text{ sgn}(x) \odot |x|^{p-1},\]

where \(\alpha \geq 0\) is the solution to (21), i.e., \(\| y + (\alpha_{x,y} - \alpha) \text{ sgn}(x) \odot |x|^{p-1} \|= 1\), with the smallest value. Therefore, from the assumption, we have \(\alpha = \alpha_{x,y}\), and \(R^\text{orth}_x(\eta_{x,y}) = x + \eta_{x,y} - \alpha_{x,y} \text{ sgn}(x) \odot |x|^{p-1} = y\) holds. Thus, (23) is proved for \(y\) belonging to the right-hand side of (25).

Finally, to prove that the equality (25) for the domain of \((R^\text{orth}_x)^{-1}\) holds, we let \(y\) be any point on \(S_p^{n-1}\) that does not belong to the right-hand side of (25) and show that \(y\) is not in the domain of \((R^\text{orth}_x)^{-1}\), i.e., there does not exist \(\eta \in T_x S_p^{n-1}\) such that \(R^\text{orth}_x(\eta) = y\) holds. Since \(y\) does not belong to the right-hand side of (25), \(\alpha = \alpha_{x,y}\) is not the solution to the equation \(\| y + (\alpha_{x,y} - \alpha) \text{ sgn}(x) \odot |x|^{p-1} \|= 1\) with the smallest value. However, since \(\alpha = \alpha_{x,y}\) is obviously one of the solutions to the equation, this implies that there exists a smaller solution \(\alpha = \alpha_0 < \alpha_{x,y}\) to the equation. Therefore, for this \(\alpha_0\), we have

\[R^\text{orth}_x(\eta_{x,y}) = x + \eta_{x,y} - \alpha_0 \text{ sgn}(x) \odot |x|^{p-1} \neq x + \eta_{x,y} - \alpha_{x,y} \text{ sgn}(x) \odot |x|^{p-1} = y.\]

Because we already know that if \(R^\text{orth}_x(\eta) = y\) for some \(\eta \in T_x S_p^{n-1}\), then \(\eta\) should be \(\eta_{x,y}\), the fact \(R^\text{orth}_x(\eta_{x,y}) \neq y\) means that there does not exist \(\eta \in T_x S_p^{n-1}\) such that \(R^\text{orth}_x(\eta) = y\). We conclude that in this case \(y\) is not in the domain, \(D_x\), of \((R^\text{orth}_x)^{-1}\).

From the above discussion, the domain \(D_x\) is written as (25). This completes the proof.

To conclude this subsection, we discuss, for each \(x \in S_p^{n-1}\), the property of the subset of \(T_x S_p^{n-1}\) defined as \(^4\)

\[W_x S_p^{n-1} := \{ \eta \in T_x S_p^{n-1} \mid \text{There exists } \alpha \geq 0 \text{ such that (21) holds} \}. \tag{26}\]

\(^4\) As discussed before, if (21) holds for some \(\alpha \in \mathbb{R}\), then \(\alpha\) should be nonnegative.
If \( p = 2 \), then \( W_x S^{n-1}_2 \) is equal to the closed unit ball in the tangent space \( T_x S^{n-1}_p \) as Remark 4 implies, where we note that the norm in \( T_x S^{n-1}_p \) is induced by the 2-norm in \( \mathbb{R}^n \). In fact, for general \( p \in (1, \infty) \), the subset \( W_x S^{n-1}_p \subset T_x S^{n-1}_p \) is proved to be convex and compact. Thus, \( R_x^{\text{orth}}(\eta) \) is not defined for \( \eta \in T_x S^{n-1}_p \) whose norm is too large.

**Proposition 7** Assume that \( p \in (1, \infty) \). For any \( x \in S^{n-1}_p \), the subset \( W_x S^{n-1}_p \subset T_x S^{n-1}_p \) defined in (26) is convex and compact.

**Proof** To prove that \( W_x S^{n-1}_p \) is convex, we arbitrarily take \( \eta_1, \eta_2 \in W_x S^{n-1}_p \) and \( \lambda \in [0, 1] \) and show that \( (1 - \lambda) \eta_1 + \lambda \eta_2 \in W_x S^{n-1}_p \) holds. Then, from the definition of \( W_x S^{n-1}_p \), there exist \( \alpha_1, \alpha_2 \geq 0 \) satisfying \( \|x + \eta_i - \alpha_i \text{ sgn}(x) \cap |x|^{p-1}\|_p = 1 \) for \( i = 1, 2 \). For the convex combination \( (1 - \lambda) \eta_1 + \lambda \eta_2 \in T_x S^{n-1}_p \), it follows from the triangle inequality that

\[
\|x + (1 - \lambda) \eta_1 + \lambda \eta_2 - ((1 - \lambda) \alpha_1 + \lambda \alpha_2) \text{ sgn}(x) \cap |x|^{p-1}\|_p \\
\leq (1 - \lambda)\|x + \eta_1 - \alpha_1 \text{ sgn}(x) \cap |x|^{p-1}\|_p + \lambda\|x + \eta_2 - \alpha_2 \text{ sgn}(x) \cap |x|^{p-1}\|_p \\
= (1 - \lambda) + \lambda = 1.
\]

(27)

Noting that the function \( \varphi(\alpha) := \|x + (1 - \lambda) \eta_1 + \lambda \eta_2 - \alpha \text{ sgn}(x) \cap |x|^{p-1}\|_p \) is continuous, \( \lim_{\alpha \to \infty} \varphi(\alpha) = \infty \) from \( \text{sgn}(x) \cap |x|^{p-1} \neq 0 \), and \( \varphi((1 - \lambda) \alpha_1 + \lambda \alpha_2) \leq 1 \) from (27), we can conclude that there exists \( \alpha_0 \in [(1 - \lambda) \alpha_1 + \lambda \alpha_2, \infty) \) such that \( \varphi(\alpha_0) = 1 \). Therefore, \( (1 - \lambda) \eta_1 + \lambda \eta_2 \in W_x S^{n-1}_p \), indicating that \( W_x S^{n-1}_p \) is convex.

Subsequently, we prove that \( W_x S^{n-1}_p \) is compact. From the Heine–Borel theorem, it is sufficient to show that \( W_x S^{n-1}_p \) is closed and bounded in \( T_x S^{n-1}_p \). In the following, we use the equivalence between the \( p \)-norm and 2-norm in \( \mathbb{R}^n \), i.e., there exist constants \( c_p, C_p > 0 \) such that \( c_p \|w\|_2 \leq \|w\|_p \leq C_p \|w\|_2 \) holds for all \( w \in \mathbb{R}^n \).

To show that \( W_x S^{n-1}_p \) is closed, we define \( \Psi(\eta, \alpha) := \|x + \eta - \alpha \text{ sgn}(x) \cap |x|^{p-1}\|_p \) in \( T_x S^{n-1}_p \times \mathbb{R} \) and \( \psi(\eta) := \min_{\alpha \in \mathbb{R}} \Psi(\eta, \alpha) \) in \( T_x S^{n-1}_p \). The function \( \psi \) is well-defined, i.e., the minimum exists, since for any fixed \( \eta \), the function \( \psi(\eta, \alpha) \) is convex and continuous with respect to \( \alpha \), and we have \( \lim_{\alpha \to \pm \infty} \psi(\eta, \alpha) = \infty \). Furthermore, \( \psi \) thus defined is continuous. Indeed, for any \( \epsilon > 0 \), let \( \delta = \epsilon / C_p \). Then, for any \( \eta \in T_x S^{n-1}_p \) and \( \xi \in T_x S^{n-1}_p \) with \( \|\xi\|_x < \delta \), we have

\[
\begin{align*}
\left| \psi(\eta + \xi) - \psi(\eta) \right| &= \left| \min_{\alpha \in \mathbb{R}} \Psi(\eta + \xi, \alpha) - \min_{\alpha \in \mathbb{R}} \Psi(\eta, \alpha) \right| \\
&\leq \max_{\alpha \in \mathbb{R}} |\Psi(\eta + \xi, \alpha) - \Psi(\eta, \alpha)| \\
&= \max_{\alpha \in \mathbb{R}} \|x + (\eta + \xi) - \alpha \text{ sgn}(x) \cap |x|^{p-1}\|_p - \|x + \eta - \alpha \text{ sgn}(x) \cap |x|^{p-1}\|_p \\
&\leq \max_{\alpha \in \mathbb{R}} \|\xi\|_p \leq C_p \|\xi\|_2 = C_p \|\xi\|_x < C_p \delta = \epsilon.
\end{align*}
\]
where the second inequality follows from the triangle inequality. Then, noting that \( \Psi(\eta, 0) = \|x + \eta\| \geq 1 \), we can conclude that

\[
W_x S_p^{n-1} = \{ \eta \in T_x S_p^{n-1} \mid \text{There exists } \alpha \geq 0 \text{ such that } \Psi(\eta, \alpha) = 1 \}
= \{ \eta \in T_x S_p^{n-1} \mid \psi(\eta) \leq 1 \}
= \psi^{-1}(\{0, 1\})
\]

is closed since it is the preimage of the closed interval \([0, 1]\) under the continuous function \(\psi\).

Finally, we show that \(W_x S_p^{n-1}\) is bounded. For \(\eta \in T_x S_p^{n-1}\) and \(\alpha \geq 0\), noting that

\[
x^T(\text{sgn}(x) \odot |x|^{p-1}) = 1 \quad \text{and} \quad \eta^T(\text{sgn}(x) \odot |x|^{p-1}) = 0
\]
and putting

\[
d_x := \|\|x\|^{p-1}\|_2 > 0,
\]
we have

\[
\|x + \eta - \alpha \text{ sgn}(x) \odot |x|^{p-1}\|_2^2 = \|x + \eta\|_2^2 - 2\alpha + d_x^2 \alpha^2
\]
\[
= d_x^2(\alpha - d_x^{-2})^2 + \|x + \eta\|_2^2 - d_x^{-2}
\]
\[
\geq \|x + \eta\|_2^2 - d_x^{-2}
\]
\[
\geq (\|\eta\|_2 - \|x\|_2)^2 - d_x^{-2}
\]
\[
= (\|\eta\|_x - \|x\|_2)^2 - d_x^{-2},
\]
where the second inequality follows from the triangle inequality. Therefore, if \(\eta \in T_x S_p^{n-1}\) satisfies \(\|\eta\|_x > \|x\|_2 + \sqrt{c_p^{-2} + d_x^{-2}}\), then we have

\[
\|x + \eta - \alpha \text{ sgn}(x) \odot |x|^{p-1}\|_p \geq c_p \|x + \eta - \alpha \text{ sgn}(x) \odot |x|^{p-1}\|_2
\]
\[
\geq c_p \sqrt{(\|\eta\|_x - \|x\|_2)^2 - d_x^{-2}} > 1
\]
for any \(\alpha \geq 0\), which means that such \(\eta\) does not belong to \(W_x S_p^{n-1}\). Therefore, \(\eta \in W_x S_p^{n-1}\) implies that

\[
\|\eta\|_x \leq \|x\|_2 + \sqrt{c_p^{-2} + d_x^{-2}}.
\]

Hence, \(W_x S_p^{n-1}\) is bounded. This completes the proof. \(\square\)

### 4.4 Discussion on exponential retraction

On a general Riemannian manifold, another important retraction is the exponential retraction \(R := \text{Exp}\), where \(\text{Exp}\) is the exponential map \([1, 19]\). However, it may be difficult to use practically. Here, we discuss what the difficulties are in deriving an explicit form of the exponential map on \(S_p^{n-1}\). In the following discussion, we assume that \(p \geq 2\), which ensures that \(S_p^{n-1}\) is at least a \(C^2\) submanifold of \(\mathbb{R}^n\) from Theorem 1.
The exponential map Exp is defined as

\[
\text{Exp}_x(\eta) := \gamma_{x, \eta}(1), \quad \eta \in T_x S_p^{n-1}, \quad x \in S_p^{n-1},
\]

where \( \gamma_{x, \eta} \) is the geodesic on \( S_p^{n-1} \) emanating from \( x \) in the direction of \( \eta \). Specifically, the geodesic \( \gamma_{x, \eta} \) is defined to be a smooth curve on \( S_p^{n-1} \) satisfying \( \gamma_{x, \eta}(0) = x \) and \( \gamma_{x, \eta}'(0) = \eta \) with zero acceleration, i.e., \( \frac{d}{dt}\gamma_{x, \eta}'(t) = 0 \), where \( \frac{d}{dt} \) is the covariant derivative with respect to the Levi-Civita connection on \( S_p^{n-1} \). As a curve in \( \mathbb{R}^n \), let \( \gamma_{x, \eta}'(t) := \frac{d}{dt}\gamma_{x, \eta}(t) \) denote the second-order derivative of \( \gamma_{x, \eta} \). Then, the geodesic equation \( \frac{d}{dt}\gamma_{x, \eta}(t) = 0 \) is equivalent to \( \gamma_{x, \eta}(t) \in N_{\gamma_{x, \eta}} S_p^{n-1} \). For simplicity, we denote \( \gamma_{x, \eta}(t) \) by \( x(t) \). Then, \( x(t) \in S_p^{n-1} \), and \( 1_T |x(t)|^p = 1 \) holds. Differentiating both sides, we obtain

\[
(p - 1)(|x(t)|^{p-2} \circ \dot{x}(t)) T \dot{x}(t) + (\text{sgn}(x(t)) \circ |x(t)|^{p-1}) T \dot{x}(t) = 0. \tag{28}
\]

Since \( \dot{x}(t) \in N_{x(t)} S_p^{n-1} \) holds, it can be deduced that there exists \( \alpha(t) \in \mathbb{R} \) such that \( \dot{x}(t) = \alpha(t) \text{sgn}(x(t)) \circ |x(t)|^{p-1} \). Substituting this into (28), we obtain the expression \( \alpha(t) = -(p - 1)((|x(t)|^{p-2} T \dot{x}(t)^2)/||x(t)||^p \|x(t)||^{p-2}_2^2 \). Therefore, \( x(t) \) satisfies the geodesic equation

\[
\dot{x}(t) + \frac{(p - 1)(|x(t)|^{p-2} T \dot{x}(t)^2)}{||x(t)||^p \|x(t)||^{p-2}_2} \text{sgn}(x(t)) \circ |x(t)|^{p-1} = 0. \tag{29}
\]

Solving this equation for the case \( p \neq 2 \) may be difficult. Thus, this will be addressed in future work.

**Remark 5** When \( p = 2 \), Eq. (29) is reduced to \( \dot{x}(t) + (\dot{x}(t) T \dot{x}(t)) x(t) = 0 \), whose solution is \( x(t) = x \cos(\|\eta\| t) + (\eta/\|\eta\|) \sin(\|\eta\| t) \), where \( x(0) = x \) and \( \dot{x}(0) = \eta \), as shown in [1, Example 5.4.1].

## 5 Vector transports

In addition to a retraction, a vector transport is also an important geometric tool in Riemannian optimization methods, e.g., Riemannian conjugate gradient methods [1, 17, 18, 20, 23], Riemannian quasi-Newton methods [9, 10], and Riemannian stochastic optimization methods [24, 29]. Let \( \mathcal{M} \) be a Riemannian manifold and \( T \mathcal{M} \oplus T \mathcal{M} := \{ (\eta, \xi) \mid \eta, \xi \in T_x \mathcal{M}, \ x \in \mathcal{M} \} \) be the Whitney sum. A map \( T : T \mathcal{M} \oplus T \mathcal{M} \to \mathcal{M} : (\eta, \xi) \mapsto T_\eta(\xi) \) is called a vector transport on \( \mathcal{M} \) if there exists a retraction \( R \) on \( \mathcal{M} \) and the following conditions are satisfied for any \( x \in \mathcal{M} \):

1. \( T_\eta(\xi) \in T_{R_\eta(\eta)} \mathcal{M} \) for any \( \eta, \xi \in T_x \mathcal{M} \);
2. \( \mathcal{T}_\eta = \text{id}_{T_x \mathcal{M}} \);
3. For any \( \eta \in T_x \mathcal{M} \), the map \( T_\eta : T_x \mathcal{M} \to T_{R_\eta(\eta)} \mathcal{M} \) is linear.
5.1 Differentiated retraction

An important vector transport is the differentiated retraction $T^R$ \cite[Section 8.1.2]{1} associated with a retraction $R$ on $S_p^{n-1}$ defined by

$$T^R_{\eta}(\xi) := DR_x(\eta)[\xi], \quad \eta, \xi \in T_x S_p^{n-1}, \quad x \in S_p^{n-1}.$$ 

The differentiated retraction appears in the Riemannian (strong) Wolfe conditions \cite{21} and is thus used for line search in various algorithms.

Here, we derive the expression of $T^R$ with the retraction $R$ defined in (9). Noting (2), an analogous computation to (10) gives

$$T^R_{\eta}(\xi) = \frac{d}{dt} R_x(\eta + t\xi) \bigg|_{t=0} = \frac{\xi}{\|x + \eta\|_p^p} - \frac{(\text{sgn}(x + \eta) \odot |x + \eta|^{p-1})^T \xi}{\|x + \eta\|_p^{p+1}}(x + \eta).$$

5.2 Vector transport based on orthogonal projection

Since $S_p^{n-1}$ is a Riemannian submanifold of $\mathbb{R}^n$, another vector transport $T^P$ on $S_p^{n-1}$ is defined by the orthogonal projection \cite[Section 8.1.3]{1} as

$$T^P_{\eta}(\xi) := P_{R_x(\eta)}(\xi), \quad \eta, \xi \in T_x S_p^{n-1}, \quad x \in S_p^{n-1},$$

where the orthogonal projection $P$ is provided by (7). Specifically, if we use the retraction (9), we have

$$T^P_{\eta}(\xi) = \left(I - \frac{(\text{sgn}(R_x(\eta)) \odot |R_x(\eta)|^{p-1})(\text{sgn}(R_x(\eta)) \odot |R_x(\eta)|^{p-1})^T}{\|R_x(\eta)|^{p-1}\|_2^2}\right)\xi = \frac{(\text{sgn}(x + \eta) \odot |x + \eta|^{p-1})^T \xi}{\|x + \eta|^{p-1}\|_2^2} \text{sgn}(x + \eta) \odot |x + \eta|^{p-1}. $$

6 Summary of theoretical results

We investigated the geometry of $S_p^{n-1}$ and proposed several retractions and their inverses and vector transports. These results are summarized in Table 1.
where

\[ \xi, \eta \in \mathbb{R} \]

Vector transport by projection

\[ \xi = \frac{\eta (\eta) (\xi)}{\|\eta\|^2_2} \eta (\xi) \quad \text{for } \xi \in \mathbb{R}, \quad \eta (\xi) \quad \text{for } \eta \in \mathbb{R} \]

Applications

Table 1 Summary of theoretical results

| Sphere with \( p \)-norm \( S^n_p \) | \( S^n_p = \{ x \in \mathbb{R}^n \mid \| x \|_p = 1 \} \) |
|--------------------------------------|--------------------------------------------------|
| Riemannian metric on \( S^n_p \)    | \( \langle \xi, \eta \rangle_x = \xi^T \eta \) |
| Induced norm in \( T_x S^n_p \)     | \( \| \xi \|_x = \| \xi \|_2 = \sqrt{\xi^T \xi} \) |
| Tangent space at \( x \)            | \( T_x S^n_p = \{ \xi \in \mathbb{R}^n \mid \xi^T (\text{sgn}(x) \odot |x|^{p-1}) = 0 \} \) |
| Normal space at \( x \)             | \( N_x S^n_p = \{ \alpha \text{ sgn}(x) \odot |x|^{p-1} \mid \alpha \in \mathbb{R} \} \) |
| Orthogonal projection onto \( T_x S^n_p \) | \( P_x = I - \frac{(\text{sgn}(x) \odot |x|^{p-1}) (\text{sgn}(x) \odot |x|^{p-1})^T}{\| |x|^{p-1} \|^2_2} \) |

Retraction by normalization

\[ R_x(\eta) = \frac{x + \eta}{\|x + \eta\|_p} \]

Inverse of \( R_x \)

\[ R_x^{-1}(y) = \frac{y}{\langle \text{sgn}(x) \odot |x|^{p-1} \rangle^T y} - x, \]

where \( y \) satisfies \( \langle \text{sgn}(x) \odot |x|^{p-1} \rangle^T y > 0 \)

Inverse of projective retraction

\[ (R_x^{\text{proj}})^{-1}(y) = y - x + \alpha \text{ sgn}(y) \odot |y|^{p-1}, \]

where \( \alpha = \frac{1 - \langle \text{sgn}(x) \odot |x|^{p-1} \rangle^T y}{\|\langle \text{sgn}(x) \odot |x|^{p-1} \rangle \|_2^2} \)

and \( y \) is such that \( \alpha \) is well-defined and nonnegative

Inverse of orthographic retraction

\[ (R_x^{\text{orth}})^{-1}(y) = y - x + \alpha \text{ sgn}(x) \odot |x|^{p-1}, \]

where \( \alpha = \frac{1 - \langle \text{sgn}(x) \odot |x|^{p-1} \rangle^T y}{\|\langle \text{sgn}(x) \odot |x|^{p-1} \rangle \|_2^2} \)

and \( y \) is such that \( \alpha \geq 0 \) is the solution to the equation

\[ \| y + (\alpha - t) \text{ sgn}(x) \odot |x|^{p-1} \|_p^p = 1 \]

for \( t \) with the smallest value

Differentiated retraction of \( R \)

\[ T^{R}_{\eta}(\xi) = DR_x(\eta)[\xi] \]

\[ = \frac{\xi}{\|x + \eta\|_p} - \frac{\langle \text{sgn}(x + \eta) \odot |x + \eta|^{p-1} \rangle^T \xi}{\|x + \eta\|_p^{p+1}} \]

Vector transport by projection

\[ T^{P}_{\eta}(\xi) = P_{R_x(\eta)}(\xi) \]

\[ = \frac{\xi - \langle \text{sgn}(x + \eta) \odot |x + \eta|^{p-1} \rangle^T \xi}{\|x + \eta|^{p-1}\|_2^2} \text{ sgn}(x + \eta) \odot |x + \eta|^{p-1} \]

The sphere \( S^n_p \) is defined for \( p \in [1, \infty] \). However, the above results are for the case of \( p \in (1, \infty) \), where \( S^n_p \) is a \( C^1 \) submanifold of \( \mathbb{R}^n \). In addition, we assume \( x, y \in S^n_p \) and \( \xi, \eta \in T_x S^n_p \)

7 Applications

In this section, we discuss two types of applications of \( S^n_p \) for optimization.

7.1 Nonnegative constraints on spheres

In nonlinear optimization, we can introduce squared slack variables to handle nonnegative constraints [7]. Specifically, the constraint \( v \geq 0 \) for \( v \in \mathbb{R}^n \) is equivalent

\[ \square \text{ Springer} \]
to \( v = x^2 \) with \( x \in \mathbb{R}^n \). This idea can be used for optimization problems on the intersection of the sphere and nonnegative orthant.

### 7.1.1 Unconstrained and constrained optimization problems on spheres with different norms

For \( p' \geq 1 \) and \( p = 2p' \), \( S_p^{n-1} \) can be used to handle the variable on \( S_p^{n-1} \) with the nonnegative constraint. To see this, we consider the following problem:

\[
\begin{align*}
\text{minimize} & \quad g(v) \\
\text{subject to} & \quad v \geq 0, \ v \in S_p^{n-1},
\end{align*}
\]

where \( g: S_p^{n-1} \to \mathbb{R} \) is the objective function. This is a constrained Riemannian optimization problem on \( S_p^{n-1} \) with the constraint \( v \geq 0 \). Defining \( v := x^2 \geq 0 \) with \( x = [x_i] \in \mathbb{R}^n \), we can observe that the conditions \( v \in S_p^{n-1} \) and \( v \geq 0 \) on \( v \) are equivalent to \( \|x^2\|_{p'} = 1 \) on \( x \). Regarding the left-hand side, we have the relation \( \|x^2\|_{p'} = \sum_{i=1}^{n} |x_i|^2 \|p' = \sum_{i=1}^{n} |x_i|^{2p'} = \|x\|_{2p'}^{2p'} = \|x\|_{p}^{p} \). Therefore, \( \|x^2\|_{p'} = 1 \) is equivalent to \( \|x\|_{p} = 1 \), i.e., \( x \in S_p^{n-1} \). Hence, the aforementioned optimization problem is equivalent to the following problem:

\[
\begin{align*}
\text{minimize} & \quad -v^T A v \\
\text{subject to} & \quad v \geq 0, \ v \in S_2^{n-1},
\end{align*}
\]

which is an unconstrained Riemannian optimization problem on \( S_p^{n-1} \).

### 7.1.2 Application to nonnegative PCA

As a particular case of \( p' = 2 \) and \( p = 4 \), we can deduce from the above discussion that solving an optimization problem on \( S_2^{n-1} \) with the nonnegative constraint on the decision variable vector is equivalent to solving the corresponding optimization problem on \( S_4^{n-1} \) without constraint. An example where the nonnegative sphere can be used is graph learning [6]. Another example within this framework is the nonnegative PCA [28], which we focus on here.

In [14], the nonnegative PCA is formulated as follows:

\[
\begin{align*}
\text{minimize} & \quad -v^T A v \\
\text{subject to} & \quad v \geq 0, \ v \in S_2^{n-1},
\end{align*}
\]

where \( A \) corresponds to the variance–covariance matrix of the data to be analyzed. We assume that \( A \) is an \( n \times n \) symmetric positive definite matrix. The above problem is
equivalent to the following unconstrained problem on the sphere $S^{n-1}_4$ with 4-norm:

\[
\begin{align*}
\text{minimize} \quad & f(x) := -(x^2)^T A(x^2) \\
\text{subject to} \quad & x \in S^{n-1}_4.
\end{align*}
\] (31)

For an optimal solution $x^*$ to the latter problem, $v^* := x^2_*$ is an optimal solution to the former problem.

We can further show that, for any $v^*$ satisfying the first-order optimality conditions for Problem (30), $x^*$ that satisfies $v^* = x^2_*$ is a critical point of $f$ in Problem (31).

**Proposition 8** Let $v^*$ be an optimal solution to Problem (30) with an $n \times n$ symmetric positive definite matrix $A$. Then, $v^*$ satisfies

\[
v^* \geq 0, \quad v^T v^* = 1, \quad (I - v^* v^T) A v^* \leq 0.
\] (32)

Specifically, the $i$th element $(Av^*)_i$ of $Av^*$ satisfies $(Av^*)_i = (v^T A v^*) v_i$ if $(v^*)_i > 0$, and $(Av^*)_i \leq 0$ if $(v^*)_i = 0$. In particular, if $v^* > 0$, then $Av^* = (v^T A v^*) v^*$ holds, i.e., $v^T A v^*$ and $v^*$ are an eigenvalue and associated eigenvector of $A$, respectively.

**Proof** Problem (30) is equivalent to the following Euclidean optimization problem:

\[
\begin{align*}
\text{minimize} \quad & -v^T A v \\
\text{subject to} \quad & v \geq 0, \quad v^T v = 1, \quad v \in \mathbb{R}^n.
\end{align*}
\] (33)

Throughout this proof, among the $n$ constraints $v_1 \geq 0, v_2 \geq 0, \ldots, v_n \geq 0$, let $\mathcal{I}_0(v^*) := \{i_1, i_2, \ldots, i_m\} \subset \{1, 2, \ldots, n\}$ be the set of indices for the active inequality constraints at $v^*$ and $\mathcal{I}_+(v^*) := \{1, 2, \ldots, n\} - \mathcal{I}_0(v^*)$ be the complement of $\mathcal{I}_0(v^*)$ in $\{1, 2, \ldots, n\}$, i.e.,

\[
(v^*)_{i_1} = (v^*)_{i_2} = \cdots = (v^*)_{i_m} = 0,
\] (34)

and $(v^*)_i > 0$ for all $i \in \mathcal{I}_+(v^*)$. Then, letting $e_i \in \mathbb{R}^n$ be the vector whose $i$th element is 1 and the others are 0, the gradients of the $m$ functions defining the active inequality constraints are $e_{i_1}, e_{i_2}, \ldots, e_{i_m}$. Since $v^T v^* = 1, v^*$ is not 0; and $\mathcal{I}_+(v^*) \neq \emptyset$, i.e., there exists $i_0 \neq i_1, i_2, \ldots, i_m$ such that $(v^*)_{i_0} > 0$. Hence, the gradient of the equality constraint function $v^T v - 1$ at $v^*$, which is $2v^*$, and $e_{i_1}, e_{i_2}, \ldots, e_{i_m}$ are linearly independent. This means that the linear independent constraint qualification (LICQ) [15, Definition 12.4] holds at $v^*$, and the KKT conditions for (33) are necessary optimality conditions.
Writing the KKT conditions explicitly, there exist $\lambda \in \mathbb{R}^n$ and $\mu \in \mathbb{R}$ such that

$$-2Av_\ast - \lambda + 2\mu v_\ast = 0, \quad (35)$$
$$v_\ast \geq 0, \quad (36)$$
$$v_\ast^Tv_\ast = 1, \quad (37)$$
$$\lambda \geq 0, \quad (38)$$
$$\lambda \odot v_\ast = 0. \quad (39)$$

Under (36) and (38), Eq. (39) is equivalent to $\lambda^Tv_\ast = 0$. Using this and (37), and multiplying (35) by $v_\ast^T$ from the left, we obtain $\mu = v_\ast^TA v_\ast$. Therefore, (35) and (38) yield $\lambda = 2((v_\ast^T A v_\ast) I - A)v_\ast \geq 0$, which implies $(I - v_\ast v_\ast^T) A v_\ast \leq 0$. Thus, the conditions (32) are verified to hold.

Here, $I - v_\ast v_\ast^T$ is the orthogonal projection matrix to the orthogonal complement of the span of $v_\ast \geq 0$ with respect to the standard inner product in $\mathbb{R}^n$. Therefore, from (34), the intersection of the image of $I - v_\ast v_\ast^T$, $\text{Im}(I - v_\ast v_\ast^T)$, and the nonpositive orthant $\mathbb{R}_-^n := \{x \in \mathbb{R}^n \mid x \leq 0\}$ is

$$\text{Im}(I - v_\ast v_\ast^T) \cap \mathbb{R}_-^n = \{x = [x_i] \in \mathbb{R}_-^n \mid x_i = 0, \ i \in I_+(v_\ast)\}. \quad (40)$$

It follows from (32) and (40) that the $i$th element of $(I - v_\ast v_\ast^T) A v_\ast$ is

$$((I - v_\ast v_\ast^T) A v_\ast)_i \begin{cases} = 0 & \text{if } i \in I_+(v_\ast), \\ \leq 0 & \text{if } i \in I_0(v_\ast). \end{cases} \quad (41)$$

Rewriting this, we obtain $(A v_\ast)_i = (v_\ast^T A v_\ast)(v_\ast)_i$ if $i \in I_+(v_\ast)$, i.e., $(v_\ast)_i > 0$, and $(A v_\ast)_i \leq (v_\ast^T A v_\ast)(v_\ast)_i = 0$ if $i \in I_0(v_\ast)$, i.e., $(v_\ast)_i = 0$.

In particular, if $v_\ast > 0$, then $I_0(v_\ast) = \emptyset$, and $(A v_\ast)_i = (v_\ast^T A v_\ast)(v_\ast)_i$ holds for all $i \in \{1, 2, \ldots, n\}$, which is equivalent to $A v_\ast = (v_\ast^T A v_\ast)v_\ast$. This completes the proof. \hfill \Box

**Remark 6** Conversely, it can be readily checked that if $v_\ast \in \mathbb{R}^n$ satisfies the conditions (32), then $v_\ast, \lambda = 2((v_\ast^T A v_\ast) I - A)v_\ast$, and $\mu = v_\ast^T A v_\ast$ satisfy the KKT conditions (35)–(39). In summary, there exist $\lambda$ and $\mu$ such that the KKT conditions (35)–(39) are satisfied if and only if $v_\ast$ satisfies (32).

If $v_\ast$ is an optimal solution to Problem (30), $x_\ast$ satisfying $v_\ast = x_\ast^2$ is an optimal solution to Problem (31). Therefore, such $x_\ast$ satisfies $\nabla f(x_\ast) = 0$ on $S^n_{q-1}$. More generally, as the following proposition claims, if $v_\ast$ satisfies the first-order necessary optimality conditions (32) for Problem (30), then $x_\ast$ that satisfies $v_\ast = x_\ast^2$ is a critical point of $f$ in (31).
**Proposition 9** Consider Problem (31) with an $n \times n$ symmetric positive definite matrix $A$. The gradient of the objective function $f$ on $S^n_{n-1}$ satisfies

$$
\nabla f(x) = -4 \left( (Ax^2) \otimes x - \frac{(x^4)^T Ax^2}{\|x^3\|^2} x^3 \right) 
$$

(42)

for any $x \in S^n_{n-1}$. Furthermore, if $v_\ast \in S^n_{n-1}$ satisfies (32) and $x_\ast \in S^n_{n-1}$ satisfies $v_\ast = x_\ast^2$, then $\nabla f(x_\ast) = 0$.

**Proof** We first derive Eq. (42) for $\nabla f$. Let $\bar{f}(x) := -(x^2)^T A(x^2)$ in $\mathbb{R}^n$, which is a smooth extension of $f$ to $\mathbb{R}^n$. For any $d \in \mathbb{R}^n$, the directional derivative of $\bar{f}$ at $x$ in the direction of $d$ is computed as

$$
D \bar{f}(x)[d] = -4(x^2)^T A(x \odot d) = -4(Ax^2)^T (x \odot d) = -4((Ax^2) \odot x)^T d.
$$

Hence, we obtain $\nabla \bar{f}(x) = -4(Ax^2) \odot x$. The Riemannian gradient $\nabla f$ is then obtained by using the orthogonal projection (7) as

$$
\nabla f(x) = P_x(\nabla \bar{f}(x))
$$

$$
= -4 \left( I - \frac{\text{sgn}(x) \otimes |x|^3 (\text{sgn}(x) \otimes |x|^3)^T}{\|x^3\|^2} \right) ((Ax^2) \odot x)
$$

$$
= -4 \left( (Ax^2) \odot x - \frac{(x^4)^T Ax^2}{\|x^3\|^2} x^3 \right),
$$

where we used $\text{sgn}(x) \otimes |x|^3 = x \otimes |x|^2 = x^3$. Thus, (42) is proved.

Subsequently, we assume that $v_\ast \in S^n_{n-1}$ satisfies (32) and $x_\ast \in S^n_{n-1}$ satisfies $v_\ast = x_\ast^2$. As in the proof of Proposition 8, let $\mathcal{I}(v_\ast) := \{i_1, i_2, \ldots, i_m\}$ be the set of indices such that $(v_\ast)_{i_1} = (v_\ast)_{i_2} = \cdots = (v_\ast)_{i_m} = 0$ holds and $\mathcal{I}_+(v_\ast) := \{1, 2, \ldots, n\} - \mathcal{I}(v_\ast)$. Defining $\mu := v_\ast^T A v_\ast$, it follows from (41) that

$$
(v_\ast^2)^T A v_\ast = \sum_{i \in \mathcal{I}_+(v_\ast)} (v_\ast)^2_i (A v_\ast)_i = \sum_{i \in \mathcal{I}_+(v_\ast)} (v_\ast)^2_i \mu (v_\ast)_i = \mu \sum_{i \in \mathcal{I}_+(v_\ast)} (v_\ast)_i^3
$$

(43)

Here, from (41), we have $(I - v_\ast v_\ast^T) A v_\ast \odot v_\ast = 0$, which, together with $v_\ast \neq 0$ and (43), yields $(A v_\ast) \odot v_\ast = (v_\ast v_\ast^T A v_\ast) \odot v_\ast = \mu v_\ast = ((v_\ast^2)^T A v_\ast/\|v_\ast\|^3) v_\ast^2$. Substituting $v_\ast = x_\ast^2$, this is written as

$$
(A x_\ast^2) \odot x_\ast^2 = \frac{(x_\ast^4)^T A x_\ast^2}{\|x_\ast^2\|^3} x_\ast^4 = \frac{(x_\ast^4)^T A x_\ast^2}{\|x_\ast^2\|^2} x_\ast^4.
$$

(44)
In general, for any $a, b, c \in \mathbb{R}$, $ac^2 = bc^4$ is equivalent to $ac = bc^3$. Therefore, (44) is reduced to

$$(Ax^2) \odot x^* = \frac{(x^4)^T A x^2}{\|x^3\|^2} x^3,$$

which shows that grad $f(x^*) = 0$ in light of (42). This completes the proof. □

**Remark 7** The converse of the latter claim in Proposition 9 is not true, i.e., for a critical point $x^* \in S_n^{n-1}$ of $f$ in Problem (31), $v^* = x^2$ does not necessarily satisfy the first-order necessary optimality conditions (32) for Problem (30), as the following simple counterexample illustrates.

Let $n = 2$ and $A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$ in Problem (31), where $A$ is symmetric and positive definite. Then, $x^* = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \in S^1$ is a critical point of $f$ on $S^1$ since

$$\text{grad } f(x^*) = -4 \left( (Ax^2) \odot x^* - \frac{(x^4)^T A x^2}{\|x^3\|^2} x^3 \right) = -4 \left[ \begin{array}{c} 2 \\ 0 \end{array} \right] - \left[ \begin{array}{c} 2 \\ 0 \end{array} \right] = \left[ \begin{array}{c} 0 \\ 0 \end{array} \right].$$

For this $x^*$, let $v^* = x^2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \in S^1$. Then, we have

$$(I - v^* v^T) A v^* = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \not< \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

which means that $v^*$ violates (32).

In summary, we clarified the first-order necessary optimality conditions for constrained Problem (30) on $S^{n-1}_2$ ((32) in Proposition 8) and shown that $v^* \in S^{n-1}_2$ satisfying the conditions leads to a critical point of unconstrained Problem (31) on $S^{n-1}_4$ as $x^* \in S^{n-1}_4$ that satisfies $v^* = x^2$ (Proposition 9).

### 7.1.3 Numerical experiments for nonnegative PCA

The numerical experiments in this paper were carried out in double-precision floating-point arithmetic on a PC (Apple M1 Max, 64 GB RAM) equipped with MATLAB R2020b. Here, we demonstrate numerical experiments for the nonnegative PCA. To solve the constrained Problem (30) on $S^{n-1}_2$, we solve the unconstrained Problem (31) on $S^{n-1}_4$ to obtain $x^{\text{proposed}}_n \in S^{n-1}_4$. Then, we obtain $v^{\text{proposed}}_n := (x^{\text{proposed}}_n)^2$ as a solution to the original Problem (30) based on the proposed framework. For comparison, we also solve the constrained Euclidean optimization Problem (33), which is equivalent to Problem (30), using MATLAB’s `fmincon` function, to obtain $v^{\text{fmincon}}_n$.

We consider the three cases of $n = 10$, $n = 100$, and $n = 1000$. For each $n$, following [13, 14], we constructed an $n \times n$ symmetric positive definite matrix $A$ as $A =$
\( \sqrt{\text{SNR}} v_0 v_0^T + N \), where SNR is the signal-to-noise ratio, \( N \) is a random symmetric noise matrix, and \( v_0 \in S^{n-1}_2 \) is called a spiked signal in the spiked model. Here, SNR was set as 1.0, and the diagonal (resp. off-diagonal) elements of \( N \) were chosen from the Gaussian distribution \( \mathcal{N}(0, 2/n) \) (resp. \( \mathcal{N}(0, 1/n) \)), i.i.d. up to symmetry. Furthermore, the \( i \)th element of \( (v_0)_i = \begin{cases} 1/\sqrt{\# S} & \text{(if } i \in S) \\ 0 & \text{(otherwise)} \end{cases} \), where \( S \) is a subset of \{1, 2, \ldots, n\} whose elements are chosen from \{1, 2, \ldots, n\} uniformly at random with cardinality \( \# S = \lfloor \delta n \rfloor \) and \( \delta = 0.3 \), which determines sparsity of \( v_0 \). Implementing the orthogonal projection (7) and retraction (9) based on Manopt [4], we applied the Riemannian conjugate gradient method (conjugategradient) on \( S^{n-1}_4 \) to Problem (31) with \( n = 10, n = 100, \) and \( n = 1000 \). The initial point \( x_0 \) for solving Problem (31) was also randomly constructed, i.e., we constructed \( y_0 \in \mathbb{R}^n \) whose elements are independently chosen from the uniform distribution on the interval \([-1, 1]\) and then normalized it to obtain \( x_0 = y_0/\|y_0\|_4 \). We used \( v_0 := x_0^2 \) as the initial point for solving Problem (33) by \( \text{fmincon} \). The convergence criteria for the two methods followed the default settings of conjugategradient and \( \text{fmincon} \), i.e., both methods are determined to converge if the first-order optimality measure becomes less than \( 10^{-6} \).

For the case of \( n = 10 \), we observed that the difference of the two solutions \( v_{\text{proposed}}^{n=10} \) and \( v_{\text{fmincon}}^{n=10} \) was small as \( \|v_{\text{proposed}}^{n=10} - v_{\text{fmincon}}^{n=10}\|_2 = 5.311 \times 10^{-6} \). The CPU time was less than 0.1 seconds for each method. Both of the two methods solved the corresponding problem easily and yielded almost the same solution in this case.

For \( n = 100 \), a direct application of MATLAB’s \( \text{fmincon} \) function failed to solve Problem (30) because the number of function evaluations exceeded the default maximum number 3000. Therefore, we reset the maximum number of function evaluations in \( \text{fmincon} \) as 12000, which led to the successful termination of \( \text{fmincon} \). However, the resultant \( v_{\text{fmincon}}^{n=100} \) is still a little worse than \( v_{\text{proposed}}^{n=100} \) on several points. First, although both solutions satisfied the nonnegative constraint, \( v_{\text{fmincon}}^{n=100} \) slightly violated the equality constraint (norm constraint) as \( \|v_{\text{proposed}}^{n=100}\|_2 = 1 + 1.92 \times 10^{-12} \), which should be 1 in theory. Note that \( \|v_{\text{proposed}}^{n=100}\|_2 = 1 \) holds by construction. Second, the objective function values (the smaller, the better) were

\[
\frac{g(v_{\text{proposed}}^{n=100})}{g(v_{\text{fmincon}}^{n=100})} = -1.713928 < -1.713921 = g(v_{\text{fmincon}}^{n=100}).
\]

Third, sparsity of the solutions also differed. To measure sparsity of a vector \( c = [c_i] \in \mathbb{R}^n \), we used the Gini index [11], which is defined as

\[
G(c) := 1 - 2 \sum_{i=1}^{n} \frac{|c(i)|}{\|c\|_1} \left( \frac{n - i + 1/2}{n} \right),
\]

where \( c(1), c(2), \ldots, c(n) \) are obtained by reordering \( c_1, c_2, \ldots, c_n \) such that \( |c(1)| \leq |c(2)| \leq \cdots \leq |c(n)| \) holds. The Gini index is scale invariant, and a larger value of Gini...
Riemannian optimization on unit sphere with $p$-norm

index implies that a vector is sparser. For $v_{n=100}^{\text{proposed}}$ and $v_{n=100}^{\text{fmincon}}$, we observed

$$G(v_{n=100}^{\text{proposed}}) = 0.58620 > 0.58616 = G(v_{n=100}^{\text{fmincon}}).$$

Finally, the CPU times (resp. the number of iterations) for obtaining $v_{n=100}^{\text{proposed}}$ and $v_{n=100}^{\text{fmincon}}$ were 0.132 and 0.271 seconds (resp. 235 and 109), respectively. In these aspects, the proposed method performed slightly better than fmincon in this case.

Subsequently, for $n = 1000$, the fmincon function again failed to solve Problem (30) even when we reset the maximum number of function evaluations as 50000. With this setting, we obtained $v_{n=1000}^{\text{fmincon}}$, which satisfies nonnegativity constraint but violates the norm constraint as $\|v_{n=1000}^{\text{fmincon}}\|_2 = 1 + 3.87 \times 10^{-3}$, while $\|v_{n=1000}^{\text{proposed}}\|_2 = 1$ still holds for the proposed method by construction. The objective function values were

$$g(v_{n=1000}^{\text{proposed}}) = -1.659980 < -1.406485 = g(v_{n=1000}^{\text{fmincon}}).$$

Even if we normalize $v_{n=1000}^{\text{fmincon}}$ as $v_{n=1000}^{\text{fmincon}}/\|v_{n=1000}^{\text{fmincon}}\|_2 \in S^{n-1}_2$, we still have a large objective function value as $g(v_{n=1000}^{\text{fmincon}}/\|v_{n=1000}^{\text{fmincon}}\|_2) = -1.405398$, which is even larger than $g(v_{n=1000}^{\text{fmincon}})$. Furthermore, measuring sparsity by the Gini index (45) revealed

$$G(v_{n=1000}^{\text{proposed}}) = 0.61214 > 0.33833 = G(v_{n=1000}^{\text{fmincon}}),$$

which implies that the proposed method obtained a much sparser solution than fmincon. Finally, the CPU times (resp. the numbers of iterations) for obtaining $v_{n=1000}^{\text{proposed}}$ and $v_{n=1000}^{\text{fmincon}}$ were 1.490 and 28.714 seconds (resp. 328 and 49), respectively. Although conjugategradient took more iterations, the CPU time was much less than fmincon because time per iteration for conjugategradient was much shorter. We again note that fmincon did not give a feasible solution in this case and that the CPU time and number of iterations shown here regarding fmincon are only for reference.

Thus, we can conclude that the proposed framework efficiently solved the nonnegative PCA problem and yielded a much better solution, especially when the dimension is large. In addition, although a critical point $x \in S^{n-1}_2$ of $f$ does not necessarily give $v = x^2$ satisfying the optimality conditions (32) for Problem (30) as Remark 7, for all the above cases, $v = v_{n=10}^{\text{proposed}}$, $v_{n=100}^{\text{proposed}}$, $v_{n=1000}^{\text{proposed}}$ obtained by the proposed method satisfy (32) with sufficient precision. Specifically, $v \geq 0$ and $v^Tv = 1$ are satisfied by construction. Regarding the inequality $(I - vv^T)Au \leq 0$, the maximum value of the elements of $(I - vv^T)Au$ for $v = v_{n=10}^{\text{proposed}}$, $v_{n=100}^{\text{proposed}}$, $v_{n=1000}^{\text{proposed}}$ are $1.237 \times 10^{-6}$, $3.712 \times 10^{-7}$, and $4.044 \times 10^{-7}$, respectively, which can be regarded as almost 0 considering that we terminated conjugategradient when the gradient norm became less than $10^{-6}$.
7.2 $L_p$-regularization-related optimization

In certain applications, $L_p$ regularization is a frequently used technique, which considers an objective function as the weighted sum of the original objective function and the $p$-norm of the decision variable vector. In particular, $L_1$ regularization is used in Lasso for sparse estimation [8].

7.2.1 Relationship between regularized, constrained, and manifold optimization problems

We consider the following regularized optimization problem with $p \in [1, \infty]$:

$$
\text{minimize} \quad L(w) + \lambda \|w\|_p \\
\text{subject to} \quad w \in \mathbb{R}^n, 
$$

where $L: \mathbb{R}^n \to \mathbb{R}$ is a convex function, and $\lambda \geq 0$ is a predefined constant called a regularization parameter.

Intuitively, $L_p$ regularization is closely related to the constraint that the $p$-norm of the decision variable vector is not larger than a predefined nonnegative constant. Specifically, we consider the following constrained problem:

$$
\text{minimize} \quad L(w) \\
\text{subject to} \quad \|w\|_p \leq C, \ w \in \mathbb{R}^n, 
$$

where $C \geq 0$ is a constant. For $p = 1$, while Lasso is usually performed by solving the former Problem (46), the latter Problem (47) explains why Lasso tends to find a sparse solution [8]. For $p = 2$, the approaches of Problems (46) and (47) are called the Tikhonov [26] and Ivanov [12] regularizations, respectively, and they are essentially equivalent [16, Theorem 1]. This discussion can be generalized for general $p \in [1, \infty]$ as the following proposition. A complete proof is given in Appendix A.

**Proposition 10** Assume that $p \in [1, \infty]$, and let $L: \mathbb{R}^n \to \mathbb{R}$ be a convex function. If $w_*$ is an optimal solution to Problem (46) with a predefined constant $\lambda \geq 0$, then there exists $C \geq 0$ such that $w_*$ is an optimal solution to Problem (47) with $C$. Conversely, if $w_*$ is an optimal solution to Problem (47) with a predefined constant $C \geq 0$, then there exists $\lambda \geq 0$ such that $w_*$ is an optimal solution to Problem (46).

In the following, we assume $C > 0$ because the case $C = 0$ yields that $w = 0$ is the only feasible (and thus optimal) solution to Problem (47). From Proposition 10, we observe the importance of Problem (47) in dealing with Problem (46). Furthermore, Problem (47) is closely related to the following problem with an equality constraint:

$$
\text{minimize} \quad L(w) \\
\text{subject to} \quad \|w\|_p = C, \ w \in \mathbb{R}^n, 
$$
where $C$ is the constant in Problem (47). Indeed, if a minimum point $w_*$ of $L$ over the entire $\mathbb{R}^n$ lies in the ball $\{ w \in \mathbb{R}^n \mid \| w \|_p \leq C \}$, then $w_*$ is also an optimal solution to (47). Therefore, a possible approach to solving Problem (47) is to first minimize $L$ without any constraint to obtain $w_*$. If $\| w_* \|_p \leq C$, then Problem (47) is solved. Otherwise, there exists an optimal solution to Problem (47) that is on the boundary $\{ w \in \mathbb{R}^n \mid \| w \|_p = C \}$ of the ball since $L$ is convex in $\mathbb{R}^n$. In this case, it is sufficient to solve Problem (48) for obtaining an optimal solution to Problem (47). Furthermore, upon scaling $w \mapsto w/C$ and $L \mapsto L \circ CI$ and writing $w/C$ and $L \circ CI$ newly as $x$ and $f$, respectively, i.e., $f(x) := L(Cx) = L(w)$, Problem (48) essentially becomes equivalent to the following problem on the unit sphere $S^{n-1}_p$ with $p$-norm:

$$\text{minimize } f(x) \quad \text{subject to } x \in S^{n-1}_p. \quad (49)$$

As important cases, we consider $p = 1$ and $p = \infty$. However, these cases do not lie within the scope of the discussion in the previous sections. Therefore, we approximate $S^{n-1}_1$ and $S^{n-1}_\infty$ by $S^{n-1}_p$ with $p > 1$ being sufficiently close to 1 and $S^{n-1}_p$ with sufficiently large $p$, respectively. The following proposition evaluates the accuracy of such approximations.

**Proposition 11** Let $\varepsilon \in (0, 1)$ be a constant and assume that $n \geq 2$. If

$$1 < p \leq 1 + \frac{\log(1 + \varepsilon)}{\log n - \log(1 + \varepsilon)}, \quad (50)$$

then $x \in S^{n-1}_p$ satisfies $1 \leq \| x \|_1 \leq 1 + \varepsilon$. On the other hand, if

$$p \geq \frac{\log n}{-\log(1 - \varepsilon)}, \quad (51)$$

then $x \in S^{n-1}_p$ satisfies $1 - \varepsilon \leq \| x \|_\infty \leq 1$.

**Proof** First, we assume that $p$ satisfies (50). Since each element $x_i$ of $x \in S^{n-1}_p$ satisfies $|x_i| \leq 1$, we have $\| x \|_1 \geq \| x \|_p = 1$ from $p > 1$. Furthermore, since $w \mapsto w^p$ is convex on $\mathbb{R}_+$, we obtain the relation

$$1 = \| x \|_p^p = \sum_{i=1}^n |x_i|^p = n \sum_{i=1}^n \frac{1}{n} |x_i|^p \geq n \left( \sum_{i=1}^n \frac{1}{n} |x_i| \right)^p = \frac{1}{n^{p-1}} \| x \|_1^p.$$

---

5 The original Problem (47) has an advantage that it is a convex optimization problem; however, Problem (48) is not since the sphere $\{ w \in \mathbb{R}^n \mid \| w \|_p = C \}$ is not a convex set. Nevertheless, Problem (48) can be solved by unconstrained Riemannian optimization methods. Although the projection of an unconstrained optimal solution, i.e., $w^{\text{unconst}} \in \mathbb{R}^n$ that minimizes $L$ without constraint, onto $S^{n-1}_p$ is a naive feasible solution to (47), it may not be a good solution (see also the numerical experiments in Sects. 7.2.2 and 7.2.3). This is a motivation for exploring Problem (48) as an alternative to Problem (47).
Therefore, using (50), we can evaluate \( \|x\|_1 \) as
\[
\|x\|_1 \leq n^{1-1/p} \leq n^{\log n(1+\varepsilon)} = 1 + \varepsilon.
\]

Subsequently, we assume that \( p \) satisfies (51). Then, any \( x \in S_{p}^{n-1} \) satisfies \( \|x\|_\infty = \max_{i=1,2,...,n} |x_i| \leq 1 \). In addition, we have
\[
1 = \|x\|_p = \sum_{i=1}^{n} |x_i|^p \leq \sum_{i=1}^{n} \|x\|_\infty^p = n \|x\|_\infty^p.
\]
It follows from (51) that
\[
\|x\|_\infty \geq n^{-1/p} \geq n^{\log n(1-\varepsilon)} = 1 - \varepsilon.
\]
This completes the proof. \( \square \)

For example, consider the case of \( n = 10 \). To approximate \( S_1^{n-1} \) by \( S_p^{n-1} \) with guaranteeing that every \( x \in S_{p}^{n-1} \) satisfies \( \|x\|_1 \leq 1 + 10^{-4} =: 1 + \varepsilon \), from (50), it suffices to set \( p = 1 + \log(1 + \varepsilon) / (\log n - \log(1 + \varepsilon)) \approx 1 + 4.34 \times 10^{-5} \). To approximate \( S_{\infty}^{n-1} \) by \( S_p^{n-1} \) with guaranteeing that every \( x \in S_{p}^{n-1} \) satisfies \( \|x\|_\infty \geq 1 - 10^{-4} =: 1 - \varepsilon \), from (51), we can set \( p \geq \log n / (\log(1 - \varepsilon)) \), e.g., \( p = 23026 \) as an even number.

Furthermore, for \( p \in (1, \infty) \), (50) and (51) are equivalent to \( \varepsilon \geq n^{1-1/p} - 1 \) and \( \varepsilon \geq 1 - n^{-1/p} \), respectively. Since we assume \( 0 < \varepsilon < 1 \) in Proposition 11, these two inequalities make sense if the quantities \( n^{1-1/p} - 1 \) and \( 1 - n^{-1/p} \) in the right-hand sides are in the interval \((0, 1)\) for each case, which mean that \( 1 < p < \log_2 n / (\log_2 n - 1) \)\(^6\) and \( p > 1 \), respectively. Proposition 11 together with this observation yields the following result, which evaluates the accuracy of \( S_{p}^{n-1} \) with given \( p \) as approximation of \( S_1^{n-1} \) and \( S_{\infty}^{n-1} \).

**Corollary 12** Assume that \( n \geq 2 \). For \( p \in (1, \log_2 n / (\log_2 n - 1)) \), \( x \in S_{p}^{n-1} \) satisfies \( 1 \leq \|x\|_1 \leq n^{1-1/p} \). For \( p \in (1, \infty) \), \( x \in S_{p}^{n-1} \) satisfies \( n^{-1/p} \leq \|x\|_\infty \leq 1 \).

**7.2.2 Numerical experiment for Lasso regression**

Here, we consider the Lasso regression with simple artificial data. The data size is set as \( m = 100 \) and the number of variables as \( n = 13 \). We construct a data matrix \( X \in \mathbb{R}^{m \times n} \) with randomly generated elements, each of which is chosen from the standard Gaussian distribution, set \( w_s = [-5, -4, -3, -2, -1, 1, 2, 3, 4, 5, 0, 0, 0]^{T} \in \mathbb{R}^n \), and compute \( y = Xw_s + \epsilon \), where each element of \( \epsilon \in \mathbb{R}^n \) is randomly generated from a uniform distribution on the interval \([-1, 1]\). This means that, among \( n = 13 \)

\(^6\) When \( n = 2 \), we regard \( \log_2 n / (\log_2 n - 1) \) as \( \infty \). Indeed, \( 2^{1-1/p} - 1 \in (0, 1) \) holds for any \( p > 1 \).

\(^7\) When \( n = 2 \), we interpret this condition as \( p \in (1, \infty) \).
Table 2 Results obtained by solving the Lasso-related optimization problems for \( p = 1.000001 \) with several values of \( C \)

| \( w^{\text{nonreg}} \) | \( w^{\text{proposed}}_{C=1} \) | \( w^{\text{proposed}}_{C=10} \) | \( w^{\text{proposed}}_{C=22} \) | \( w^{\text{proposed}}_{C=50} \) | \( w^{\text{Lasso}}_{\lambda = 0.029} \) | \( w^{\text{Lasso}}_{\lambda = 0.746} \) |
|---|---|---|---|---|---|---|
| 1 | −5.055 | −0.162 | −1.645 | −3.702 | −8.040 | −4.989 | −3.727 |
| 2 | −3.904 | −0.133 | −1.411 | −3.132 | −5.762 | −3.881 | −3.212 |
| 3 | −3.022 | −0.137 | −1.314 | −2.564 | −4.521 | −3.023 | −2.712 |
| 4 | −2.039 | −0.057 | −0.644 | −1.336 | −3.657 | −1.986 | −1.192 |
| 5 | −1.036 | 0.000 | 0.000 | 0.000 | 0.000 | −2.872 | −0.993 | −0.002 |
| 6 | 0.967 | 0.006 | 0.000 | 0.370 | −0.979 | 0.939 | 0.224 |
| 7 | 1.972 | 0.174 | 1.124 | 1.751 | −0.608 | 1.959 | 1.831 |
| 8 | 3.028 | 0.001 | 0.549 | 1.974 | 5.474 | 2.976 | 1.890 |
| 9 | 4.036 | 0.054 | 1.019 | 2.776 | 6.346 | 3.981 | 2.791 |
| 10 | 5.060 | 0.270 | 2.217 | 4.173 | 7.002 | 5.037 | 4.314 |
| 11 | −0.008 | 0.006 | 0.073 | 0.000 | −2.290 | 0 | 0 |
| 12 | −0.032 | 0.000 | 0.003 | 0.000 | 0.979 | 0 | 0 |
| 13 | 0.052 | 0.000 | 0.000 | 0.000 | −1.471 | 0 | 0 |

The \( i \)th row shows the \( i \)th element of each solution

variables, the first 10 are essential and the last 3 have no effect on the data \( y \). We now estimate the coefficient parameter vector \( w_\ast \) without any information on it, i.e., by using only the observed data \( X \) and \( y \). An appropriate sparse estimation should yield a coefficient parameter vector whose last 3 elements are close to 0.

With the data \( X \) and \( y \), we consider Problem (48) with \( L(w) := \| Xw - y \|_2^2 \) and a constant \( C > 0 \), i.e., with the equality constraint \( \| w \|_p = C \). Note that we exclude the case when \( C = 0 \) since it yields the trivial solution \( w = 0 \). Let \( w^{\text{nonreg}} := (X^T X)^{-1} X^T y \) be the solution to the nonregularized (unconstrained) optimization problem of minimizing \( L \). If \( \| w^{\text{nonreg}} \|_p > C \), then a solution to Problem (48) is also a solution to (47), as discussed in Sect. 7.2.1. Solving Problem (48) is equivalent to minimizing \( f(x) := L(Cx) = \| Cx - y \|_2^2 \) with respect to \( x \in S_p^{n-1} \), i.e., solving Problem (49), and multiplying the resultant solution \( x_\ast \) by \( C \) to obtain the solution \( w_\ast = Cx_\ast \) to Problem (48).

The case of \( p = 1 \) corresponds to the Lasso regression. However, we can handle \( S_p^{n-1} \) with \( p > 1 \) using the Riemannian optimization techniques developed in the previous sections. Therefore, we adopt \( p = 1.000001 = 1 + 10^{-6} \) and expect that solving the problem on \( S_p^{n-1} \) yields a sparse solution. Corollary 12 guarantees that \( x \in S_p^{n-1} \) satisfies \( 1 \leq \| x \|_1 \leq n^{1-1/p} \approx 1 + 2.56 \times 10^{-6} \). Implementing the orthogonal projection (7) and retraction (9) based on Manopt, we applied the Riemannian conjugate gradient method for Problem (49) on \( S_p^{n-1} \).

In Table 2, as expected, \( w^{\text{nonreg}} \) is not sparse at all. Of course, if we scale \( w^{\text{nonreg}} \) as \( C w^{\text{nonreg}} / \| w^{\text{nonreg}} \|_p \), whose \( p \)-norm is \( C \), the scaled vector is not sparse either. We applied the Riemannian conjugate gradient method in the proposed framework with several values of \( C \) and obtained the solution \( w^{\text{proposed}}_C \) to Problem (48) for
each $C$. The results for $C = 1, 10, 22, 50$ are shown in the table. For small $C$ such as $C = 1, 10$, the resultant solutions are sparse but do not provide a good estimation. Indeed, for example, for the case $C = 10$, the 5th and 6th entries are almost zero while the 11th and 12th are nonzero. On the contrary, large $C$ does not contribute to sparse estimation at all. In fact, we have $\|w_{\text{nonreg}}\|_p = 30.21$. Therefore, for $C$ larger than this value, solving Problem (48) does not necessarily lead to a solution to Problem (47). Although finding the best value of $C$ is difficult, we observe that the case of $C = 22$ yields an appropriate solution in this experiment, which is a sparse solution with appropriate values. Indeed, the 11th, 12th, and 13th elements of $w_{C=22}^{\text{proposed}}$ are $2.37 \times 10^{-8}$, $1.12 \times 10^{-4}$, and $-6.60 \times 10^{-10}$, respectively.

For comparison, we also applied MATLAB’s lasso function, which successively increases the value of $\lambda$ and solves Problem (46) for each $\lambda$. For small $\lambda$’s, the corresponding solutions are dense, whereas the solution is 0 for a sufficiently large $\lambda$. We focus on the $\lambda$’s and corresponding solutions $w_{\lambda}^{\text{lasso}} \in \mathbb{R}^n$ such that only the last 3 elements of $w_{\lambda}^{\text{lasso}}$ are 0. The lasso function yielded several $\lambda$’s satisfying this condition. Among them, $w_{\lambda=0.029}^{\text{lasso}}$ and $w_{\lambda=0.746}^{\text{lasso}}$ correspond to the smallest and largest values of $\lambda$, respectively. We can observe from Table 2 that $w_{C=22}^{\text{proposed}}$ and $w_{\lambda=0.746}^{\text{lasso}}$ are close to each other. Furthermore, the values of Gini index (45) of them are $G(w_{C=22}^{\text{proposed}}) = 0.481$ and $G(w_{\lambda=0.746}^{\text{lasso}}) = 0.502$, showing that the two solutions are similar. As another measure of comparison between the two approaches, we use the signal-to-noise ratio, which is defined as

$$\text{SNR}(\hat{w}) = \frac{\text{Var}[X \hat{w}]}{\text{Var}[y - X \hat{w}]}$$

for the estimated coefficient vector $\hat{w} \in \mathbb{R}^n$. We have $\text{SNR}(w_{C=22}^{\text{proposed}}) = 9.53$ and $\text{SNR}(w_{\lambda=0.746}^{\text{lasso}}) = 9.75$, which are also close to each other.

Here, we note that conjugate gradient terminated for $p = 1.000001$ and $C = 22$ after it failed to find an appropriate step size. Thus, the gradient norm could not be sufficiently reduced to satisfy the convergence criterion. This is because $p = 1.000001$ is too close to 1. To understand this phenomenon in detail, we observed how the values of $p$ affect the resultant solutions. Here, we fixed $C = 22$ and solved Problem (49) by the conjugate gradient method to obtain a solution $w_p^{\text{proposed}}$ to Problem (48) for each of $p = 2, 1.1, 1.01, 1.001, 1.0001, 1.00001, 1.000001$. We observe from Table 3 that the gradient norm was not reduced enough in the cases of $p \leq 1.01$. Therefore, the CPU time and number of iterations for those cases are not important but are only for reference. Although the values of Gini index for the resultant solutions are similar to each other, the results obtained by smaller values of $p$ appear better because the 11th, 12th, and 13th elements of them are close to 0. Hence, for smaller values of $p$, we may not need to stick to reducing the gradient norm enough but just reduce the value of the objective function, which may lead to a better solution for the Lasso problem.
Table 3 Results obtained by solving the Lasso-related optimization problems for $C = 22$ with several values of $p$, showing the solutions and corresponding gradient norm of the objective function at $x = w/C$ (Grad.), value of Gini index (Gini), signal-to-noise ratio (SNR), time in seconds until termination of the conjugate gradient algorithm (Time), and number of iterations until termination (Iter.)

| $w_{\text{proposed}}$ | $w_{\text{proposed}}$ | $w_{\text{proposed}}$ | $w_{\text{proposed}}$ | $w_{\text{proposed}}$ | $w_{\text{proposed}}$ | $w_{\text{proposed}}$ |
|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|
| $p=2$                  | $p=1.1$                | $p=1.01$               | $p=1.001$              | $p=1.0001$             | $p=1.00001$            | $p=1.000001$           |
| 1                      | -11.910                | -4.499                 | -3.788                 | -3.712                 | -3.699                 | -3.703                 |
| 2                      | -5.269                 | -3.608                 | -3.204                 | -3.179                 | -3.128                 | -3.132                 |
| 3                      | -3.407                 | -2.820                 | -2.572                 | -2.536                 | -2.564                 | -2.564                 |
| 4                      | -6.187                 | -1.695                 | -1.327                 | -1.340                 | -1.336                 | -1.336                 |
| 5                      | -3.425                 | -0.694                 | -0.261                 | -0.217                 | -0.222                 | -0.221                 |
| 6                      | 1.183                  | 0.692                  | 0.409                  | 0.383                  | 0.373                  | 0.370                  |
| 7                      | 0.902                  | 1.924                  | 1.803                  | 1.812                  | 1.742                  | 1.751                  |
| 8                      | 6.511                  | 2.572                  | 2.013                  | 1.948                  | 1.975                  | 1.974                  |
| 9                      | 9.903                  | 3.519                  | 2.856                  | 2.802                  | 2.773                  | 2.776                  |
| 10                     | 9.537                  | 4.712                  | 4.230                  | 4.180                  | 4.169                  | 4.173                  |
| 11                     | -2.849                 | 0.021                  | 0.001                  | 0.000                  | 0.000                  | 0.000                  |
| 12                     | -1.534                 | 0.000                  | 0.000                  | 0.003                  | 0.000                  | 0.000                  |
| 13                     | 2.946                  | -0.000                 | 0.000                  | -0.020                 | -0.000                 | -0.000                 |
| Grad.                  | $4.07 \times 10^{-4}$  | $1.23 \times 10^{-3}$  | $> 1$                  | $> 1$                  | $> 1$                  | $> 1$                  |
| Gini                   | 0.378                  | 0.453                  | 0.479                  | 0.482                  | 0.480                  | 0.481                  |
| SNR                    | 3.823                  | 60.401                 | 11.080                 | 9.746                  | 9.489                  | 9.527                  |
| Time                   | 0.083                  | 0.126                  | 0.019                  | 0.028                  | 0.076                  | 0.028                  |
| Iter.                  | 45                     | 291                    | 34                     | 58                     | 173                    | 54                     |

The $i$th row shows the $i$th element of each solution.
7.2.3 Numerical experiment for box-constrained problem

Here, we consider the following box-constrained optimization problem:

\[
\begin{align*}
\text{minimize} & \quad L(w) \\
\text{subject to} & \quad l \leq w \leq u, \ w \in \mathbb{R}^n, \quad (52)
\end{align*}
\]

where \(l = [l_i], \ u = [u_i] \in \mathbb{R}^n\) are given constant vectors with \(l < u\). The constraint \(l \leq w \leq u\) means the box constraint \(l_i \leq w_i \leq u_i\) for \(i = 1, 2, \ldots, n\). Defining \(a := (u - l)/2 > 0\) and \(b := (l + u)/2\), this constraint is rewritten as \(-a \leq w - b \leq a\), which is equivalent to \(-1 \leq D^{-1}(w - b) \leq 1\), i.e., \(\|D^{-1}(w - b)\|_\infty \leq 1\), with \(D\) being the \(n \times n\) diagonal matrix with diagonal elements \(a_1, a_2, \ldots, a_n > 0\). Therefore, with the transformation \(x := D^{-1}(w - b) \in S^{n-1}_\infty\) and \(f(x) := L(a \odot x + b) = L(Dx + b) = L(w)\), solving Problem (52) is essentially equivalent to minimizing \(f\) in the unit ball \(B^n_\infty = \{x \in \mathbb{R}^n \mid \|x\|_\infty \leq 1\}\). Assume that \(L\) is convex in \(\mathbb{R}^n\); hence, \(f\) is also convex. Let \(x^{\text{unconst}} \in \mathbb{R}^n\) be a minimum point of \(f\) over \(\mathbb{R}^n\). If \(x^{\text{unconst}}\) is in \(B^n_\infty\), then it is a desired optimal solution. Otherwise, as discussed in Sect. 7.2.1, we only have to solve Problem (49) on the sphere \(S^{n-1}_p\) with \(p = \infty\). However, since \(p = \infty\) was excluded from the discussion in the previous sections, we instead need to consider a sufficiently large finite value \(p\) when solving the problem numerically.

We performed a numerical experiment for the following problem:

\[
\begin{align*}
\text{minimize} & \quad L(w) := \frac{1}{2} w^T A w + c^T w \\
\text{subject to} & \quad l \leq w \leq u, \ w \in \mathbb{R}^n, \quad (53)
\end{align*}
\]

where \(n = 10\), and the elements of the \(n \times n\) symmetric positive definite matrix \(A\) and vector \(c \in \mathbb{R}^n\) are randomly generated.

Specifically, we construct a matrix \(A_0 \in \mathbb{R}^{n \times n}\) each of whose element is chosen from the uniform distribution on the interval \([0, 2]\), and then compute \(A = A_0^T A_0\). Each element of \(c\) is chosen from the standard Gaussian distribution. We set \(l = [-1, -2, \ldots, -10]^T\) and \(u = [1, 2, \ldots, 10]^T\). Note that \(\nabla L(w) = A w + c\) and the minimum point of \(L\) over the entire \(\mathbb{R}^n\) is \(-A^{-1} c\). We checked that \(w^{\text{unconst}} := -A^{-1} c\) is not feasible for Problem (53) in this case. Therefore, as discussed above, if we minimize \(f(x) := L(a \odot x + b)\) with \(a := (u - l)/2\) and \(b := (l + u)/2\) on the sphere \(S^{n-1}_\infty\) to obtain \(w_\ast\), then \(w_\ast := a \odot x_\ast + b\) is an optimal solution to Problem (53). Here, we note that the Euclidean gradient of \(f\) is computed as \(\nabla f(x) = a \odot \nabla L(a \odot x + b) = a \odot (A(a \odot x + b) + c)\). We approximated \(S^{n-1}_\infty\) by \(S^{n-1}_p\) with several choices of \(p\), i.e., \(p = 10, 100, 1000, 10000, 50000\), and solved Problem (49) by the Riemannian conjugate gradient method based on Manopt. We denote the resultant approximate solution to the original Problem (53) by \(w_p^{\text{proposed}}\) for each \(p\) and the solution to Problem (53) obtained using MATLAB’s \texttt{fmincon}\ function by \(w^{\text{fmincon}}\). Here, the convergence criteria for the two methods followed the

\[8\] If \(l_i = u_i\) for some \(i\), then \(l_i\) is the only value that the corresponding \(w_i\) can take. By eliminating such a constant variable in advance, if necessary, we can assume \(l < u\) without loss of generality.
default settings of \texttt{conjugategradient} and \texttt{fmincon}, i.e., both methods are determined to converge if the first-order optimality measure becomes less than $10^{-6}$.

The results are shown in Table 4. Unfortunately, for each case of $p$ in the proposed method, \texttt{conjugategradient} failed to find an appropriate step size before convergence. However, the gradient norm was reduced to almost $10^{-6}$ in every case. As expected, the larger the value of $p$, the more accurate is the obtained solution. Note that, from Corollary 12, $x \in S_{p}^{n-1}$ satisfies $n^{-1/p} \leq \|x\|_{\infty} \leq 1$. For example, when $p = 50000$, we have $\|x\|_{\infty} \geq 1 - 4.61 \times 10^{-5}$.

Finally, we consider a naive feasible solution $\tilde{w}^{\text{unconst}}$ obtained by projecting $w^{\text{unconst}}$ onto the box, i.e., replacing the first and second elements of $w^{\text{unconst}}$, which are $-3.355$ and $4.331$ violating the constraints $-1 \leq w_{1} \leq 1$ and $-2 \leq w_{2} \leq 2$, with $-1$ and $2$, respectively. Specifically, we have

$$\tilde{w}^{\text{unconst}} = [-1, 2, -1.575, -0.383, \ldots, -1.885]^T,$$

which satisfies the box constraint $l \leq w \leq u$. However, the objective function value at $\tilde{w}^{\text{unconst}}$ is $L(\tilde{w}^{\text{unconst}}) = 6.65602 > 0$ and $\|\tilde{w}^{\text{unconst}} - w^{\text{fmincon}}\|_{2} = 4.97$. Therefore, such a naive feasible solution stemming from the unconstrained optimal solution is much worse than the other solutions discussed in this section.

\section{8 Concluding remarks}

This study investigated the geometry of the unit sphere defined via the $p$-norm as $S_{p}^{n-1} := \{x \in \mathbb{R}^{n} | \|x\|_{p} = 1\}$ with $p \in [1, \infty]$, especially $p \in (1, \infty)$, in detail. In particular, we derived formulas for retractions, their inverses, and vector transports, which can be used in Riemannian optimization algorithms. The results are summarized in Table 1 of Sect. 6.

Furthermore, we discussed two types of applications of optimization on $S_{p}^{n-1}$. The first was for optimization problems on the sphere with the nonnegative constraint, which include the nonnegative PCA problem. The second was for $L_p$-regularization-related optimization problems, which are closely related to the Lasso regression and box-constrained problems. To this end, we provided mathematical support for the applications and performed numerical experiments to verify the validity of the theory.

The applications addressed in this paper are examples of the proposed theory, and the corresponding numerical experiments are preliminary ones. Therefore, developing more efficient algorithms by combining the present theory and existing Riemannian optimization theory than state-of-the-art algorithms for specific problems, e.g., the nonnegative PCA and Lasso problems, are left for future work.

\textbf{Acknowledgements} The author would like to thank the editor and anonymous referee for the constructive comments that helped improve the paper.

\textbf{Funding} This work was supported by JSPS KAKENHI Grant Number JP20K14359.

\textbf{Data availability} The datasets generated during the current study are available from the corresponding author upon request.
Table 4 Results obtained by solving the box-constrained-related optimization problems with several values of $p$, showing the solutions and corresponding objective function value (Func.), gradient norm (Grad.), time in seconds until convergence (Time), number of iterations until convergence (Iter.), and distance to $w_{fmincon}$, i.e., $\| w - w_{fmincon}\|_2$ (Dist.).

|     | $w_{unconst}$ | $w_{proposed}$ | $w_{proposed}$ | $w_{proposed}$ | $w_{proposed}$ | $w_{proposed}$ | $w_{fmincon}$ |
|-----|---------------|----------------|----------------|----------------|----------------|----------------|---------------|
| 1   | -3.355        | -0.855         | -0.985         | -0.998         | -1.000         | -1.000         | -1.000        |
| 2   | 4.331         | 1.954          | 1.995          | 2.000          | 2.000          | 2.000          | 2.000         |
| 3   | -1.575        | -0.644         | -0.658         | -0.660         | -0.660         | -0.660         | -0.660        |
| 4   | -0.383        | -0.243         | -0.301         | -0.307         | -0.308         | -0.308         | -0.308        |
| 5   | -1.127        | -0.728         | -0.765         | -0.769         | -0.770         | -0.770         | -0.770        |
| 6   | 5.731         | 1.797          | 1.893          | 1.903          | 1.904          | 1.904          | 1.904         |
| 7   | -3.268        | -0.931         | -0.997         | -1.004         | -1.004         | -1.004         | -1.004        |
| 8   | -0.024        | 0.008          | 0.028          | 0.030          | 0.030          | 0.031          | 0.031         |
| 9   | 2.072         | 0.576          | 0.652          | 0.660          | 0.660          | 0.660          | 0.660         |
| 10  | -1.885        | -0.536         | -0.499         | -0.494         | -0.494         | -0.494         | -0.494        |

Func. – $-2.94983$ $-3.02746$ $-3.03539$ $-3.03618$ $-3.03625$ $-3.03627$

Grad. – $3.67 \times 10^{-6}$ $2.46 \times 10^{-6}$ $1.67 \times 10^{-6}$ $4.69 \times 10^{-6}$ $7.40 \times 10^{-6}$ –

Time – 0.120 0.109 0.123 0.109 0.206 0.038

Iter. – 273 248 279 253 500 23

Dist. $5.97$ $0.236$ $2.49 \times 10^{-2}$ $2.51 \times 10^{-3}$ $2.50 \times 10^{-4}$ $4.99 \times 10^{-5}$ 0

The $i$th row shows the $i$th element of each solution. For the solution by $f_{mincon}$, i.e., $w_{fmincon}$, instead of the gradient norm, the first-order optimality measure implemented in MATLAB was $5.80 \times 10^{-7}$.
Appendix A: Proof of Proposition 10

In this section, we provide a complete proof of Proposition 10 for self-containment.

Proof of Proposition 10 First, we fix \( \lambda \geq 0 \) and let \( w^* \) be an optimal solution to Problem (46). Then, for any \( w \in \mathbb{R}^n \), we have

\[
L(w^*) + \lambda \|w^*\|_p \leq L(w) + \lambda \|w\|_p. \tag{A1}
\]

We show that \( w^* \) is an optimal solution to Problem (47) with \( C := \|w^*\|_p \). For any feasible solution \( w \in \mathbb{R}^n \) to Problem (47), we have \( \|w\|_p \leq C = \|w^*\|_p \). Combining this and (A1), we have \( L(w^*) + \lambda \|w^*\|_p \leq L(w) + \lambda \|w^*\|_p \), which means \( L(w^*) \leq L(w) \). Furthermore, \( w^* \) is clearly a feasible solution to Problem (47) since \( \|w^*\|_p = C \). Therefore, \( w^* \) is an optimal solution to Problem (47).

Conversely, we fix \( C \geq 0 \) and let \( w^* \) be an optimal solution to Problem (47). Here, we additionally consider the Lagrange dual problem of (47):

\[
\text{maximize} \quad \inf_{w \in \mathbb{R}^n} (L(w) + \mu(\|w\|_p - C))
\]

\text{subject to} \quad \mu \geq 0, \quad \mu \in \mathbb{R}. \tag{A2}

If \( C > 0 \), then Slater’s condition for Problem (47), which is that there exists \( w \in \mathbb{R}^n \) with \( \|w\|_p < C \), clearly holds with \( w = 0 \). If \( C = 0 \), then the constraint \( \|w\|_p \leq C \) in Problem (47) is rewritten as the equality constraint \( w = 0 \), and Slater’s condition (which in this case is that a feasible solution exists) holds by taking \( w = 0 \). In each case, Slater’s condition for Problem (47) holds. Furthermore, Problem (47) is a convex optimization problem. Therefore, it follows from Slater’s theorem [5, Section 5.2.3] that strong duality holds. Hence, the optimal value \( L(w^*) \) of Problem (47) and the optimal value of the dual problem (A2) coincide. Letting \( \mu^* \geq 0 \) be an optimal solution to (A2), we have

\[
L(w^*) = \inf_{w \in \mathbb{R}^n} (L(w) + \mu^*(\|w\|_p - C)).
\]

Since \( \|w^*\|_p \leq C \) and \( \mu^* \geq 0 \), we obtain \( L(w^*) \leq L(w^*) + \mu^*(\|w^*\|_p - C) \leq L(w^*) \). Thus, \( L(w^*) = L(w^*) + \mu^*(\|w^*\|_p - C) \) holds, and \( w = w^* \) attains the minimum value of \( L(w) + \mu^*(\|w\|_p - C) \) over all \( w \in \mathbb{R}^n \). Since \( \mu^*C \) is a constant, \( w = w^* \) also attains the minimum value of \( L(w) + \mu^*\|w\|_p \) over \( \mathbb{R}^n \). This implies that \( w^* \) is an optimal solution to Problem (46) with \( \lambda = \mu^* \), thereby completing the proof. \( \square \)

Remark 8 Although we focus on the \( p \)-norm here, Proposition 10 can be further straightforwardly generalized to the case with a general norm in \( \mathbb{R}^n \). Indeed, in the proof of Proposition 10, we do not need any specific property of the \( p \)-norm but properties of a general norm.
References

1. Absil, P.-A., Mahony, R., Sepulchre, R.: Optimization Algorithms on Matrix Manifolds. Princeton University Press, Princeton (2008)
2. Absil, P.-A., Malick, J.: Projection-like retractions on matrix manifolds. SIAM J. Optim. 22(1), 135–158 (2012)
3. Adler, R.L., Dedieu, J.-P., Margulies, J.Y., Martens, M., Shub, M.: Newton’s method on Riemannian manifolds and a geometric model for the human spine. IMA J. Numer. Anal. 22(3), 359–390 (2002)
4. Bouna, N., Mishra, B., Absil, P.-A., Sepulchre, R.: Manopt, a Matlab toolbox for optimization on manifolds. J. Mach. Learn. Res. 15(1), 1455–1459 (2014)
5. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
6. Dong, S., Absil, P.-A., Gallivan, K.A.: Graph learning for regularized low-rank matrix completion. In: 23rd International Symposium on Mathematical Theory of Networks and Systems (MTNS 2018), pp. 460–467 (2018)
7. Fukuda, E.H., Fukushima, M.: A note on the squared slack variables technique for nonlinear optimization. J. Oper. Res. Soc. Jpn. 60(3), 262–270 (2017)
8. Hastie, T., Tibshirani, R., Wainwright, M.: Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press, Boca Raton (2015)
9. Huang, W., Gallivan, K.A., Absil, P.-A.: A Broyden class of quasi-Newton methods for Riemannian optimization. SIAM J. Optim. 25(3), 1660–1685 (2015)
10. Huang, W., Absil, P.-A., Gallivan, K.A.: A Riemannian BFGS method without differentiated retraction for nonconvex optimization problems. SIAM J. Optim. 28(1), 470–495 (2018)
11. Hurley, N., Rickard, S.: Comparing measures of sparsity. IEEE Trans. Inf. Theory 55(10), 4723–4741 (2009)
12. Ivanov, A.: The Theory of Approximate Methods and Their Applications to the Numerical Solution of Singular Integral Equations, vol. 2. Springer, Berlin (1976)
13. Khuzani, M.B., Li, N.: Stochastic primal-dual method on Riemannian manifolds of bounded sectional curvature. In: 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 133–140. IEEE (2017)
14. Liu, C., Boumal, N.: Simple algorithms for optimization on Riemannian manifolds with constraints. Appl. Math. Optim. 82(3), 949–981 (2020)
15. Nocedal, J., Wright, S.: Numerical Optimization, 2nd edn. Springer, Berlin (2006)
16. Oneto, L., Ridella, S., Anguita, D.: Tikhonov, Ivanov and Morozov regularization for support vector machine learning. Mach. Learn. 103(1), 103–136 (2016)
17. Ring, W., Wirth, B.: Optimization methods on Riemannian manifolds and their application to shape space. SIAM J. Optim. 22(2), 596–627 (2012)
18. Sakai, H., Iiduka, H.: Sufficient descent Riemannian conjugate gradient methods. J. Optim. Theory Appl. 190(1), 130–150 (2021)
19. Sakai, H., Iwai, T.: A new, globally convergent Riemannian conjugate gradient method with the weak Wolfe conditions. Comput. Optim. Appl. 64(1), 101–118 (2016)
20. Shub, M.: Some remarks on dynamical systems and numerical analysis. In: Dynamical Systems and Partial Differential Equations: Proceedings of VII ELAM, pp. 69–92 (1986)
21. Tu, L.W.: An Introduction to Manifolds. Springer, New York (2010)
22. Zass, R., Shashua, A.: Nonnegative sparse PCA. Adv. Neural Inf. Process. Syst. 19, 1561–1568 (2007)
23. Zhou, P., Yuan, X.-T., Yan, S., Feng, J.: Faster first-order methods for stochastic non-convex optimization on Riemannian manifolds. IEEE Trans. Pattern Anal. Mach. Intell. 43(2), 459–472 (2019)
30. Zhu, X., Sato, H.: Riemannian conjugate gradient methods with inverse retraction. Comput. Optim. Appl. 77(3), 779–810 (2020)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.