



# Manifold Optimization

Wen Perng

NTUEE

2024 October 1

Over the last course, we have developed the derivative of a function with respect to vectors and matrices with no constraints, i.e.:

$$\boldsymbol{x} \in \mathbb{R}^n \text{ and } X \in \mathbb{R}^{m \times n}.$$

And with the knowledge, we are able to extremize scalar functions of vector or matrix variables.

Over the last course, we have developed the derivative of a function with respect to vectors and matrices with no constraints, i.e.:

$$\boldsymbol{x} \in \mathbb{R}^n \text{ and } X \in \mathbb{R}^{m \times n}.$$

And with the knowledge, we are able to extremize scalar functions of vector or matrix variables.

But are rules of game still the same if, let's say, the matrix  $X$  is restricted to the set of **orthogonal matrices**? I.e. the set

$$\mathrm{O}(m) = \left\{ X \in \mathbb{R}^{m \times m} \mid X^T X = \mathbb{1} \right\}. \quad (1)$$

Obviously,

$$\arg \min_{X \in \mathbb{R}^{m \times m}} f(X) \text{ and } \arg \min_{X \in \mathrm{O}(m)} f(X)$$

will not coincide.

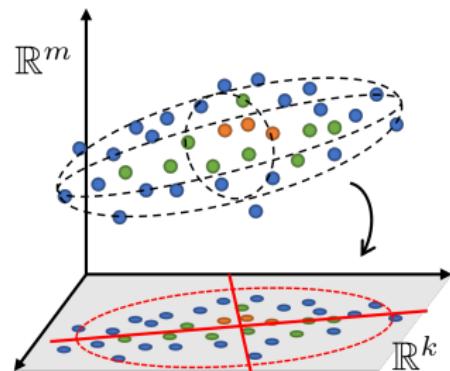
Or in **principal component analysis**:

given a data set

$X = [x_1, \dots, x_n] \in \mathbb{R}^{m \times n}$  ( $m < n$ ), we want to find a linear transformation

$$x_i \mapsto y_i = A^T x_i,$$

where  $A \in \mathbb{R}^{m \times k}$  with  $A^T A = \mathbb{1}_k$ , such that the new features  $y_i$ 's inherit the **maximum variance** of  $x_i$ 's.



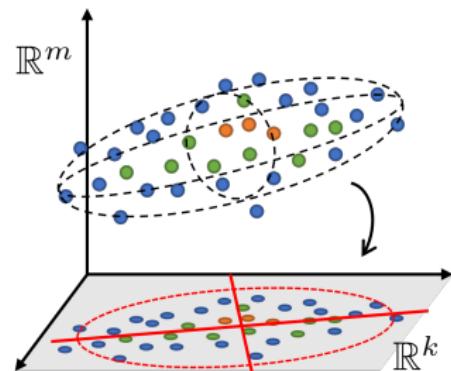
Or in **principal component analysis**:

given a data set

$X = [\mathbf{x}_1, \dots, \mathbf{x}_n] \in \mathbb{R}^{m \times n}$  ( $m < n$ ), we want to find a linear transformation

$$\mathbf{x}_i \mapsto \mathbf{y}_i = A^T \mathbf{x}_i,$$

where  $A \in \mathbb{R}^{m \times k}$  with  $A^T A = \mathbb{1}_k$ , such that the new features  $\mathbf{y}_i$ 's inherit the **maximum variance** of  $\mathbf{x}_i$ 's.

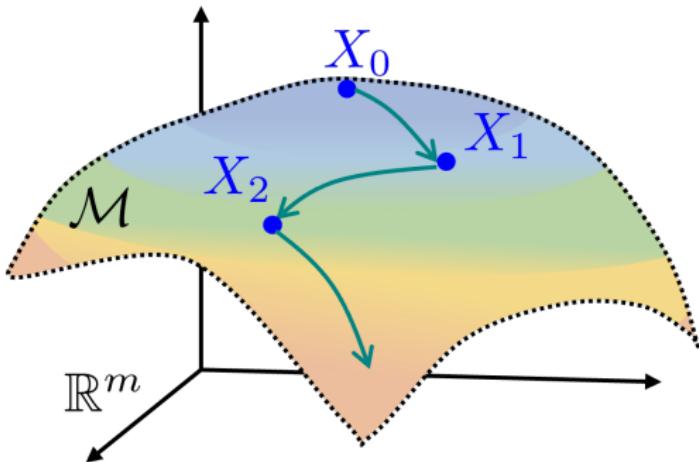


The corresponding optimization problem reads:

$$\max_A \text{Tr} \{(A^T X)(A^T X)^T\} \text{ s.t. } A^T A = \mathbb{1}_k. \quad (2)$$

Of course, constrained optimization can be handled using Lagrange multipliers, but we will not go down such path.

In general, we will view matrices as a surface in a high-dimensional space. And optimization will be done on this surface.



- 1 Recap: Matrix Differentiation**
- 2 Geometry of a Manifold**
- 3 Differentiation**
- 4 Manifold Optimization**

- 1 Recap: Matrix Differentiation**
  - Derivative ■ Inner Product ■ Gradient and Hessian
- 2 Geometry of a Manifold**
- 3 Differentiation**
- 4 Manifold Optimization**

# Derivative as a Linear Operator

Let  $x$  be a scalar in  $\mathbb{R}$  or a vector in  $\mathbb{R}^n$  or a matrix in  $\mathbb{R}^{m \times n}$ .

Given a function  $f(x)$ , if we set  $x = \gamma(t) = x_0 + th$ , we can Taylor expand it about  $t = 0$  into:

$$f \circ \gamma(t) = f \circ \gamma(0) + \frac{df \circ \gamma}{dt}(0)t + \frac{1}{2} \frac{d^2 f \circ \gamma}{dx^2} t^2 + \dots$$

## Derivative as a Linear Operator

Let  $x$  be a scalar in  $\mathbb{R}$  or a vector in  $\mathbb{R}^n$  or a matrix in  $\mathbb{R}^{m \times n}$ .

Given a function  $f(x)$ , if we set  $x = \gamma(t) = x_0 + th$ , we can Taylor expand it about  $t = 0$  into:

$$\begin{aligned} f \circ \gamma(t) &= f \circ \gamma(0) + \frac{df \circ \gamma}{dt}(0)t + \frac{1}{2} \frac{d^2f \circ \gamma}{dx^2} t^2 + \dots \\ &= f(x_0) + \frac{\partial f}{\partial x}(x_0)[h]t + \frac{1}{2} \frac{\partial^2 f}{\partial x^2}(x_0)[h]t^2 + \dots \end{aligned} \tag{3}$$

# Derivative as a Linear Operator

Let  $x$  be a scalar in  $\mathbb{R}$  or a vector in  $\mathbb{R}^n$  or a matrix in  $\mathbb{R}^{m \times n}$ .

Given a function  $f(x)$ , if we set  $x = \gamma(t) = x_0 + th$ , we can Taylor expand it about  $t = 0$  into:

$$\begin{aligned} f \circ \gamma(t) &= f \circ \gamma(0) + \frac{df \circ \gamma}{dt}(0)t + \frac{1}{2} \frac{d^2f \circ \gamma}{dx^2} t^2 + \dots \\ &= f(x_0) + \frac{\partial f}{\partial x}(x_0)[h]t + \frac{\partial^2 f}{\partial x^2}(x_0)[h]t^2 + \dots \end{aligned} \tag{3}$$

Thus, we have the first and second derivatives as:

$$\frac{\partial f}{\partial x}(x)[h] = \left. \frac{d}{dt} f(x + th) \right|_{t=0}, \tag{4}$$

$$\frac{\partial^2 f}{\partial x^2}(x)[h] = \left. \frac{d^2}{dt^2} f(x + th) \right|_{t=0}. \tag{5}$$

The second derivative can be further turned from a **quadratic form** into a **symmetric bilinear form**

The second derivative can be further turned from a **quadratic form** into a **symmetric bilinear form** by applying the **polarization identity**:

$$\frac{\partial^2 f}{\partial x^2}[h, k] = \frac{1}{4} \left( \frac{\partial^2 f}{\partial x^2}[h+k] - \frac{\partial^2 f}{\partial x^2}[h-k] \right), \quad (6)$$

with

$$\frac{\partial^2 f}{\partial x^2}[h] \equiv \frac{\partial^2 f}{\partial x^2}[h, h]. \quad (7)$$

The second derivative can be further turned from a **quadratic form** into a **symmetric bilinear form** by applying the **polarization identity**:

$$\frac{\partial^2 f}{\partial x^2}[h, k] = \frac{1}{4} \left( \frac{\partial^2 f}{\partial x^2}[h+k] - \frac{\partial^2 f}{\partial x^2}[h-k] \right), \quad (6)$$

with

$$\frac{\partial^2 f}{\partial x^2}[h] \equiv \frac{\partial^2 f}{\partial x^2}[h, h]. \quad (7)$$

An example is as follows: consider  $Q(\mathbf{x}) = \mathbf{x}^\top A \mathbf{x}$ ,

$$R(x, y) = \frac{1}{4} (Q(x+y) - Q(x-y)) = \mathbf{x}^\top A \mathbf{y} = \mathbf{y}^\top A \mathbf{x}.$$

"

$$R(y, x)$$

## Def. (Inner Product)

An inner product  $\langle \cdot, \cdot \rangle$  on a vector space  $V$  over the field  $\mathbb{R}$  is a map  $V \times V \rightarrow \mathbb{R}$  that satisfies: for  $x, y, z \in V$  and  $\lambda, \mu \in \mathbb{R}$

1. Symmetry:  $\langle x, y \rangle = \langle y, x \rangle$ ,
2. Linearity:  $\langle \lambda x + \mu y, z \rangle = \lambda \langle x, z \rangle + \mu \langle y, z \rangle$ ,
3. Positive-definiteness:  $\langle x, x \rangle > 0$  for all  $x \neq 0$

An example would be the usual dot product for column vectors  $\mathbf{x}$  and  $\mathbf{y} \in \mathbb{R}^n$ :

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y} = \sum_{i=1}^n x_i y_i. \quad (8)$$

A usual inner product used for matrices is the **Frobenius inner product** defined by

$$\langle X, Y \rangle = \text{Tr} \{ X^T Y \} = \sum_{i=1}^m \sum_{j=1}^n X_{ij} Y_{ij} \quad (X, Y \in \mathbb{R}^{m \times n}). \quad (9)$$

A usual inner product used for matrices is the **Frobenius inner product** defined by

$$\langle X, Y \rangle = \text{Tr} \{ X^T Y \} = \sum_{i=1}^m \sum_{j=1}^n X_{ij} Y_{ij} \quad (X, Y \in \mathbb{R}^{m \times n}). \quad (9)$$

But other inner products also exist:

- ▶ for  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ : given  $A \succ 0$  (positive definite),

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T A \mathbf{y}.$$

A usual inner product used for matrices is the **Frobenius inner product** defined by

$$\langle X, Y \rangle = \text{Tr} \{ X^T Y \} = \sum_{i=1}^m \sum_{j=1}^n X_{ij} Y_{ij} \quad (X, Y \in \mathbb{R}^{m \times n}). \quad (9)$$

But other inner products also exist:

- ▶ for  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ : given  $A \succ 0$  (positive definite),

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T A \mathbf{y}.$$

- ▶ given  $X^T X = \mathbb{1}_n = \cancel{XX^T}$ , for  $A, B \in \{ X\Omega \mid \Omega^T = -\Omega \}$ :

$X \in \mathbb{R}^{m \times n}$  ( $m > n$ )  $\langle A, B \rangle = \text{Tr} \left\{ A^T \left( \mathbb{1}_{\cancel{n}} - \frac{1}{2} XX^T \right) B \right\}.$

Our previous definition of the Gradient and Hessian:

### Def. (Euclidean Gradient)

Given an inner product  $\langle \cdot, \cdot \rangle$ , the **Euclidean gradient** of a function  $f(x)$  ( $x \in \mathbb{R}$ ,  $\mathbb{R}^n$  or  $\mathbb{R}^{m \times n}$ ) is  $\nabla f(x)$ , defined as

$$\frac{\partial f}{\partial x}(x)[h] = \langle \nabla f(x), h \rangle \quad (\forall h). \quad (10)$$

Our previous definition of the Gradient and Hessian:

### Def. (Euclidean Gradient)

Given an inner product  $\langle \cdot, \cdot \rangle$ , the **Euclidean gradient** of a function  $f(x)$  ( $x \in \mathbb{R}$ ,  $\mathbb{R}^n$  or  $\mathbb{R}^{m \times n}$ ) is  $\nabla f(x)$ , defined as

$$\frac{\partial f}{\partial x}(x)[h] = \langle \nabla f(x), h \rangle \quad (\forall h). \quad (10)$$

### Def. (Euclidean Hessian)

Given an inner product  $\langle \cdot, \cdot \rangle$ , the **Euclidean Hessian** of a function  $f(x)$  ( $x \in \mathbb{R}$ ,  $\mathbb{R}^n$ , or  $\mathbb{R}^{m \times n}$ ) is a linear map  $H_f(x)$ , defined as

$$\frac{\partial^2 f}{\partial x^2}(x)[h] = \langle H_f(x)[h], h \rangle \quad (\forall h). \quad (11)$$

- 1 Recap: Matrix Differentiation
- 2 Geometry of a Manifold
  - Matrix Manifolds ■ Tangent Space ■ Curves on Manifolds
- 3 Differentiation
- 4 Manifold Optimization

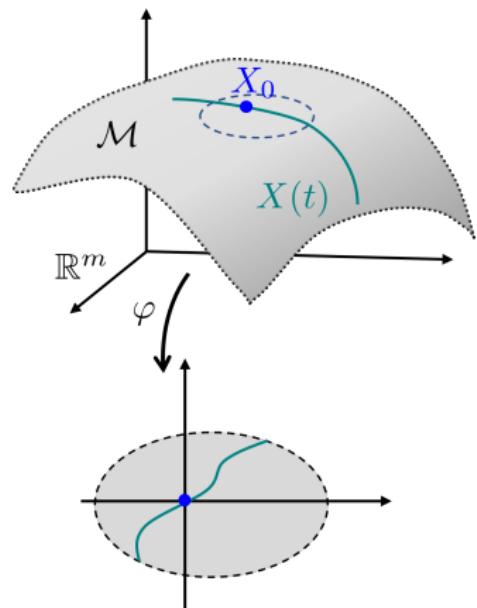
What is a manifold?

## Manifold

A manifold is a locally Euclidean, second countable, Hausdorff space.

A **differentiable manifold** is a manifold that is locally diffeomorphic to a vector space.

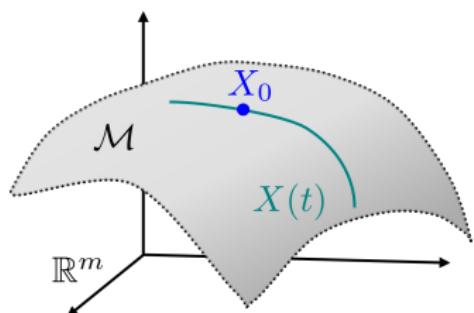
Jargons aside, we will be working with differentiable manifold  $\mathcal{M}$ , which are surfaces that can be parameterized locally using *charts*.



## Whitney Embedding Theorem

Any smooth real  $m$ -dimensional manifold can be smoothly embedded into the real  $2m$ -dimensional space  $\mathbb{R}^{2m}$ .

Simply put, the visualization on the right of **a surface in high dimensional Euclidean space** is valid for our purpose.



How do we define a manifold?

Thm. (Regular Value Theorem, pt.I)

Let  $f : \mathbb{R}^m \rightarrow \mathbb{R}^n$  with  $m > n$  be a continuously differentiable function. For any  $c \in \mathbb{R}^n$ , for which

$\partial f(x)/\partial x$  has **full rank**

for all  $x \in f^{-1}(c)$ , the set

$$\mathcal{M} = f^{-1}(c) \subset \mathbb{R}^m \qquad f(\vec{x}) = \vec{c}$$

is a  $(m - n)$ -dimensional manifold.

Note that a linear map from  $\mathbb{R}^m \rightarrow \mathbb{R}^n$  with  $m > n$  is full rank if and only if it is surjective.

e.g.  $f(\vec{x}) = \vec{x}^T \vec{x} = 1$

$$\hookrightarrow |\vec{x}|^2 = 1 \quad (\vec{x} \in \mathbb{R}^m, \text{ 曲面: } S^{m-1}, (m-1)\text{-sphere})$$

full rank?

$$\frac{\partial f}{\partial \vec{x}}[\vec{h}] = 2\vec{x}^T \vec{h} = k \in \mathbb{R} \rightarrow \vec{h} = \frac{k}{2} \vec{x}$$

e.g.  $O(m)$

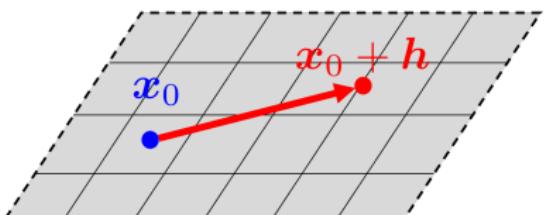
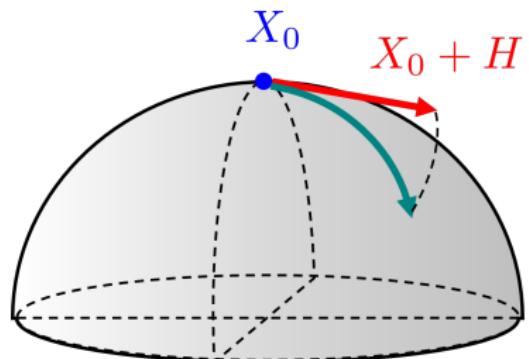
$$f(X) = X^T X - \mathbb{1} = \mathbb{O} \quad (X \in \mathbb{R}^{m \times m}) \rightsquigarrow \dim. = m^2 - \frac{m(m+1)}{2} = \frac{m(m-1)}{2}$$

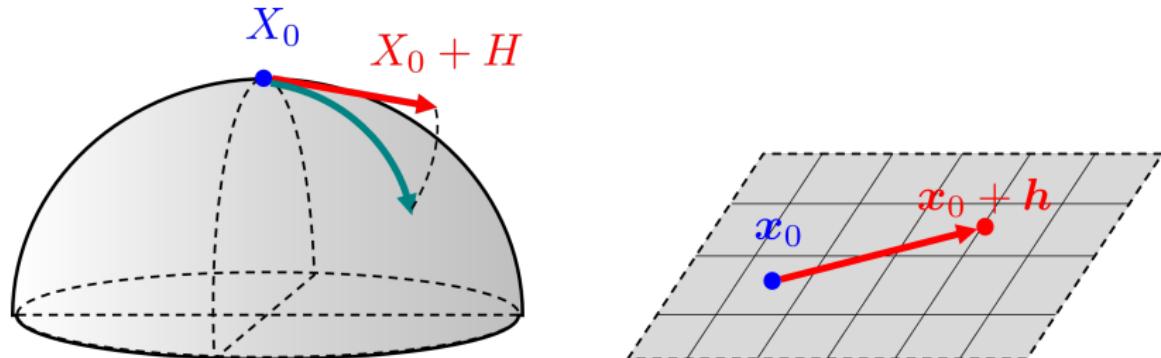
$$f: \mathbb{R}^{m \times m} \rightarrow \text{Sym}(m) = \{ X \in \mathbb{R}^{m \times m} \mid X^T = X \}$$

full rank?

$$\frac{\partial f}{\partial X}[H] = X^T H + H^T X = K \in \text{Sym}(m), H = \frac{1}{2} X K$$

# Manifolds vs. Vector Spaces





The differences to the geometry will change how we do optimization on manifolds. Ideas we need:

1. How do we differentiate on manifolds?
2. How do we move on manifolds?
3. How do we obtain *tangent vectors*?

Some examples of common matrix manifolds are given here:

### 1. Symmetric and Skew-Symmetric Matrices:

They form vector spaces:

$$\text{Sym}(m) := \{ X \in \mathbb{R}^{m \times m} \mid X^T = X \} \text{ and}$$

$$\text{Skew}(m) := \{ X \in \mathbb{R}^{m \times m} \mid X^T = -X \}.$$



$$\begin{aligned} & m + (m-1) + \dots + 1 \\ &= \frac{m(m+1)}{2} \end{aligned}$$



$$\begin{aligned} & (m-1) + \dots + 1 \\ &= \frac{m(m-1)}{2} \end{aligned}$$

Some examples of common matrix manifolds are given here:

### 1. Symmetric and Skew-Symmetric Matrices:

They form vector spaces:

$$\text{Sym}(m) := \{ X \in \mathbb{R}^{m \times m} \mid X^T = X \} \text{ and}$$

$$\text{Skew}(m) := \{ X \in \mathbb{R}^{m \times m} \mid X^T = -X \}.$$

Their dimensions are:

Some examples of common matrix manifolds are given here:

### 1. Symmetric and Skew-Symmetric Matrices:

They form vector spaces:

$$\text{Sym}(m) := \{ X \in \mathbb{R}^{m \times m} \mid X^T = X \} \text{ and}$$

$$\text{Skew}(m) := \{ X \in \mathbb{R}^{m \times m} \mid X^T = -X \}.$$

Their dimensions are:

$$\dim(\text{Sym}(m)) = \frac{m(m+1)}{2},$$

$$\dim(\text{Skew}(m)) = \frac{m(m-1)}{2}.$$

## 1. Symmetric and Skew-Symmetric Matrices:

They are mutually **orthogonal** with respect to the **Frobenius inner product**: for  $A \in \text{Sym}(m)$  and  $B \in \text{Skew}(m)$ ,

$$\langle A, B \rangle = \text{Tr} \{ A^T B \} = 0. \quad (12)$$

$$= \text{Tr} \{ B^T A \} = -\text{Tr} \{ A^T B \}$$

Hence, we have the following *orthogonal decomposition* of  $\mathbb{R}^{m \times m}$  with respect to the Frobenius inner product:

$$\mathbb{R}^{m \times m} = \text{Sym}(m) \overset{\perp}{\oplus} \text{Skew}(m). \quad (13)$$

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 1 & 2.5 \\ 2.5 & 4 \end{bmatrix} + \begin{bmatrix} 0 & -0.5 \\ 0.5 & 0 \end{bmatrix}$$

## 2. Orthogonal and Special Orthogonal Matrices:

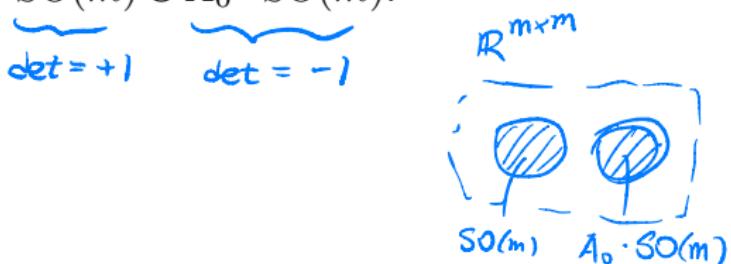
These matrices describe the  $m$ -dimensional transformation of *reflection* and *rotation*. They are no longer vector spaces.

$$\mathrm{O}(m) := \{ X \in \mathbb{R}^{m \times m} \mid X^T X = \mathbb{1} \} \text{ and}$$

$$\mathrm{SO}(m) := \{ X \in \mathbb{R}^{m \times m} \mid X^T X = \mathbb{1}, \det(X) = 1 \}.$$

Given matrix  $A_0 = \text{diag}(1, \dots, 1, -1)$ ,  $\det(A_0) = -1$ , we have

$$\mathrm{O}(m) = \underbrace{\mathrm{SO}(m)}_{\det = +1} \cup \underbrace{A_0 \cdot \mathrm{SO}(m)}_{\det = -1}.$$



## 2. Orthogonal and Special Orthogonal Matrices:

These matrices describe the  $m$ -dimensional transformation of *reflection* and *rotation*. They are no longer vector spaces.

$$\mathrm{O}(m) := \{ X \in \mathbb{R}^{m \times m} \mid X^T X = \mathbb{1} \} \text{ and}$$

$$\mathrm{SO}(m) := \{ X \in \mathbb{R}^{m \times m} \mid X^T X = \mathbb{1}, \det(X) = 1 \}.$$

Given matrix  $A_0 = \text{diag}(1, \dots, 1, -1)$ ,  $\det(A_0) = -1$ , we have

$$\mathrm{O}(m) = \mathrm{SO}(m) \cup A_0 \cdot \mathrm{SO}(m).$$

Hence, they have the same dimension:

$$\dim(\mathrm{O}(m)) = \dim(\mathrm{SO}(m)) = \frac{m(m-1)}{2}. \quad (14)$$

## 2. Orthogonal and Special Orthogonal Matrices:

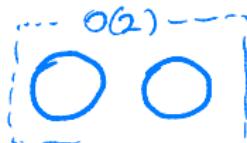
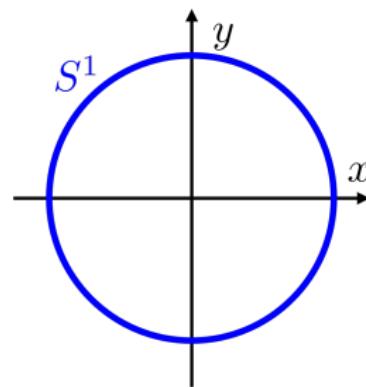
A special case would be the SO(2) and SO(3) matrices.

- ▶ SO(2): any 2D rotation can be parameterized by

$$\begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \quad (\theta \in (-\pi, \pi]).$$

Hence, SO(2) is a circle on a 2D plane, i.e. the 1-sphere  $S^1$ . It has dimension  $\frac{2 \cdot (2-1)}{2} = 1$ .

As a side note, O(2) will be two disconnected 1-spheres as a manifold.



## 2. Orthogonal and Special Orthogonal Matrices:

A special case would be the SO(2) and SO(3) matrices.

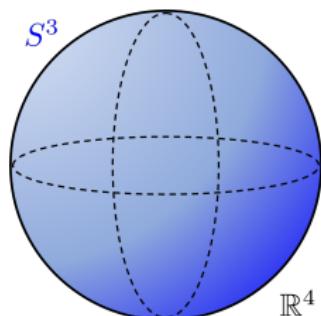
- ▶ SO(3): any 3D rotation can be parameterized by a *unit quaternion*

$$\begin{aligned} x &= x_0 + \mathbf{i}x_1 + \mathbf{j}x_2 + \mathbf{k}x_3 \\ &= \cos \frac{\theta}{2} + \sin \frac{\theta}{2} (\mathbf{i}n_1 + \mathbf{j}n_2 + \mathbf{k}n_3), \end{aligned} \tag{15}$$

with  $x_0^2 + x_1^2 + x_2^2 + x_3^2 = 1$ . The matrix representation will be

$$\begin{bmatrix} 2(x_0^2 + x_1^2) - 1 & 2(x_1x_2 - x_0x_3) & 2(x_1x_3 + x_0x_2) \\ 2(x_1x_2 + x_0x_3) & 2(x_0^2 + x_2^2) - 1 & 2(x_2x_3 - x_0x_1) \\ 2(x_1x_3 - x_0x_2) & 2(x_2x_3 + x_0x_1) & 2(x_0^2 + x_3^2) - 1 \end{bmatrix}.$$

Hence, SO(3) is a 3-sphere  $S^3$ .



### 3. Stiefel Manifold:

This is a generalization to orthogonal matrices.

$$\text{St}(m, k) = \left\{ X \in \mathbb{R}^{m \times k} \mid X^T X = \mathbb{1}_k \right\} \quad (m \geq k) \quad (16)$$

The dimension of the Stiefel manifold is

$$f(X) = X^T X - \mathbb{1}_k = \mathcal{O}$$

$$f: \mathbb{R}^{m \times k} \rightarrow \text{Sym}(k)$$

$$\dim(\text{St}(m, k)) = mk - \frac{k(k+1)}{2}$$

### 3. Stiefel Manifold:

This is a generalization to orthogonal matrices.

$$\text{St}(m, k) = \left\{ X \in \mathbb{R}^{m \times k} \mid X^T X = \mathbb{1}_k \right\} \quad (m \geq k) \quad (16)$$

The dimension of the Stiefel manifold is

$$\dim(\text{St}(m, k)) = mk - \frac{k(k+1)}{2}. \quad (17)$$

Recall that for principal component analysis, the manifold that we optimize on is the Stiefel manifold (pg.3).

Some more abstract manifolds exist, and the following discussions can also be applied to them, for example: Grassmannian and quotient manifolds.

$$\text{Gr}(n, k) = \left\{ \text{all } k\text{-dimensional subspaces} \subseteq \mathbb{R}^n \right\},$$

*equivalent relation*

$$\mathcal{M}/\sim = \{ [x] \mid x \in \mathcal{M} \}.$$

But we will not delve into them during this presentation.

*equivalent relation*

$$x \sim y \Leftrightarrow x = cy, c \in \mathbb{R} \setminus \{0\} \quad (x, y \neq 0)$$

$$x \sim x$$

$$x \sim y \Leftrightarrow y \sim x$$

*equivalent class*

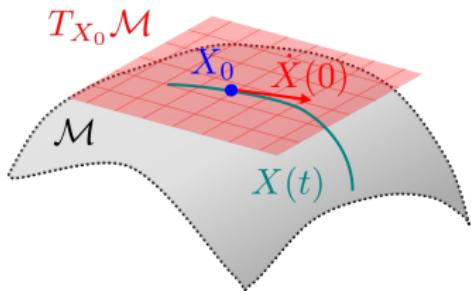
$$[x] = \{y \in \mathbb{R}^m \mid y \sim x\}$$

$$x \sim y, y \sim z \Rightarrow x \sim z$$

$$\mathbb{RP}^{m-1} = \{[x] \mid x \in \mathbb{R}^m\} = \mathbb{R}^m / \sim$$

# Tangent Vectors and Tangent Space

Consider a curve  $X(t)$  on our manifold embedded in a high dimensional  $\mathbb{R}^n$  that subjects to  $X(0) = X_0$ . If we take the time derivative of it at  $t = 0$ , we should obtain its velocity vector, or, a tangent vector at  $X_0$ .



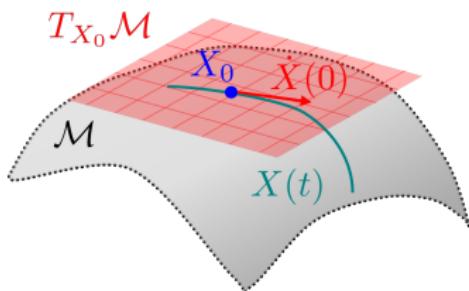
# Tangent Vectors and Tangent Space

Consider a curve  $X(t)$  on our manifold embedded in a high dimensional  $\mathbb{R}^n$  that subjects to  $X(0) = X_0$ . If we take the time derivative of it at  $t = 0$ , we should obtain its velocity vector, or, a tangent vector at  $X_0$ .

By collecting all the tangent vectors  $\dot{X}(0)$  of all possible such curves  $X(t)$ , we obtain the **tangent space** of the manifold  $\mathcal{M}$  at  $X_0$ :

$$T_{X_0}\mathcal{M}. \quad (18)$$

Note that the tangent space is a **vector space** (ref. proof via charts) with the same dimension as the manifold.



To find the tangent space, find the characterization to the tangent vectors. Some examples are as follows:

## 1. Skew( $m$ ):

$$T_X \text{Skew}(m)$$

To find the tangent space, find the characterization to the tangent vectors. Some examples are as follows:

1.  $\text{Skew}(m)$ :

$$T_X \text{Skew}(m) \xrightarrow{\sim} \text{Skew}(m). \quad (19)$$

$$X \in \text{Skew}(m)$$

To find the tangent space, find the characterization to the tangent vectors. Some examples are as follows:

1.  $\text{Skew}(m)$ :

$$T_X \text{Skew}(m) = \text{Skew}(m). \quad (19)$$

2.  $\text{SO}(m)$ :

To find the tangent space, find the characterization to the tangent vectors. Some examples are as follows:

1. **Skew( $m$ ):**

$$T_X \text{Skew}(m) = \text{Skew}(m). \quad (19)$$

2. **SO( $m$ ):**

Consider a curve  $X(t)$  subject to  $X(t)^\top X(t) = \mathbb{1}$ , then

$$\underbrace{\dot{X}(0)^\top}_{H} X_0 + X_0^\top \underbrace{\dot{X}(0)}_{H} = 0.$$

Hence,

$$T_X \text{SO}(m) := \left\{ H \in \mathbb{R}^{m \times m} \mid H^\top X + X^\top H = 0 \right\}$$

$H = X \Omega \quad \xrightarrow{\Omega^\top + \Omega = 0}$

To find the tangent space, find the characterization to the tangent vectors. Some examples are as follows:

1. **Skew( $m$ ):**

$$T_X \text{Skew}(m) = \text{Skew}(m). \quad (19)$$

2. **SO( $m$ ):**

Consider a curve  $X(t)$  subject to  $X(t)^\top X(t) = \mathbb{1}$ , then

$$\dot{X}(0)^\top X_0 + X_0^\top \dot{X}(0) = 0.$$

Hence,

$$\begin{aligned} T_X \text{SO}(m) &:= \left\{ H \in \mathbb{R}^{m \times m} \mid H^\top X + X^\top H = 0 \right\} \\ &= \left\{ X\Omega \mid \Omega \in \text{Skew}(m) \right\} \cong X \cdot \text{Skew}(m). \end{aligned} \quad (20)$$

### 3. $\text{St}(m, k)$ :

Consider a curve  $X(t)$  subject to  $X(t)^\top X(t) = \mathbb{1}_k$ , then

$$\dot{X}(0)^\top X_0 + X_0^\top \dot{X}(0) = 0.$$

3.  $\text{St}(m, k)$ :

Consider a curve  $X(t)$  subject to  $X(t)^\top X(t) = \mathbb{1}_k$ , then

$$\dot{X}(0)^\top X_0 + X_0^\top \dot{X}(0) = 0.$$

Hence,

$$T_X \text{St}(m, k) := \left\{ H \in \mathbb{R}^{m \times k} \mid H^\top X + X^\top H = 0 \right\}$$

### 3. $\text{St}(m, k)$ :

Consider a curve  $X(t)$  subject to  $X(t)^\top X(t) = \mathbb{1}_k$ , then

$$\dot{X}(0)^\top X_0 + X_0^\top \dot{X}(0) = 0.$$

Hence,

$$\begin{aligned} T_X \text{St}(m, k) &:= \left\{ H \in \mathbb{R}^{m \times k} \mid H^\top X + X^\top H = 0 \right\} \\ &= \left\{ X\Omega + X^\perp Z \mid \Omega \in \text{Skew}(k), Z \in \mathbb{R}^{(m-k) \times k} \right\}, \end{aligned} \tag{21}$$

where  $X^\perp \in \mathbb{R}^{m \times (m-k)}$  is the orthogonal complement to  $X$  such that their column vectors form a basis of  $\mathbb{R}^m$  and  $X^\top X^\perp = 0$ .

If we define a manifold using **regular value theorem**, then:

Thm. (Regular Value Theorem, pt.II)

Let our manifold  $\mathcal{M} = f^{-1}(C)$  for a constant (matrix)  $C$ . Then the tangent space at  $X \in \mathcal{M}$  is given by

$$T_X \mathcal{M} := \ker \left( \frac{\partial f}{\partial X}(X)[\cdot] \right). \quad (22)$$

kernel / null space 零空間

If we define a manifold using **regular value theorem**, then:

Thm. (Regular Value Theorem, pt.II)

Let our manifold  $\mathcal{M} = f^{-1}(C)$  for a constant (matrix)  $C$ . Then the tangent space at  $X \in \mathcal{M}$  is given by

$$T_X \mathcal{M} := \ker \left( \frac{\partial f}{\partial X}(X)[\cdot] \right). \quad (22)$$

Proof: consider a smooth curve  $\gamma$  on  $\mathcal{M} = f^{-1}(C)$  such that  $\gamma(0) = X \in \mathcal{M}$ . Then by our assumption,  $(f \circ \gamma)(t) = C$ ,

$$\frac{d}{dt}(f \circ \gamma)(0) = 0$$

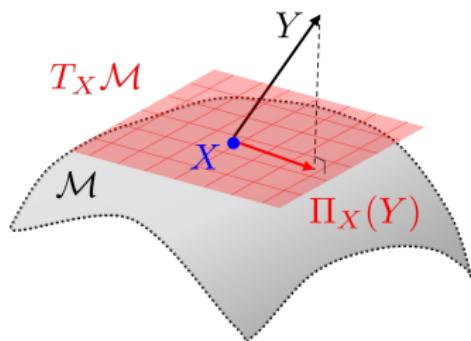
$$\frac{\partial f}{\partial X}[\dot{\gamma}(0)] = 0 \rightarrow \dot{\gamma}(0) \in \ker \left( \frac{\partial f}{\partial X}[\cdot] \right). \quad \blacksquare$$

Given any *vector* or *matrices*, we would like to *project* it onto the tangent space we just obtained.

### Def. (Projection)

A projection  $\Pi$  on a vector space  $V$  is an idempotent map. I.e.,

$$\Pi^2(X) = \Pi(X) \quad (\forall X \in V). \quad (23)$$



Given any *vector* or *matrices*, we would like to *project* it onto the tangent space we just obtained.

### Def. (Projection)

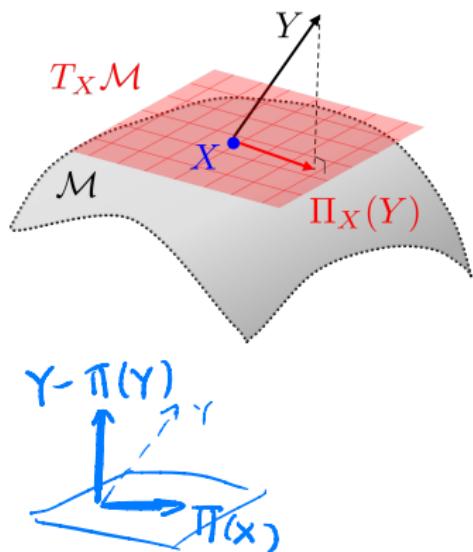
A projection  $\Pi$  on a vector space  $V$  is an idempotent map. I.e.,

$$\Pi^2(X) = \Pi(X) \quad (\forall X \in V). \quad (23)$$

### Def. (Orthogonal Projection)

Given an inner product  $\langle \cdot, \cdot \rangle$  on  $T_X\mathcal{M}$ , a projection is **orthogonal** if for any  $X, Y \in V$ ,

$$\langle \Pi(X), Y - \Pi(Y) \rangle = 0. \quad (24)$$



Given any *vector* or *matrices*, we would like to *project* it onto the tangent space we just obtained.

### Def. (Projection)

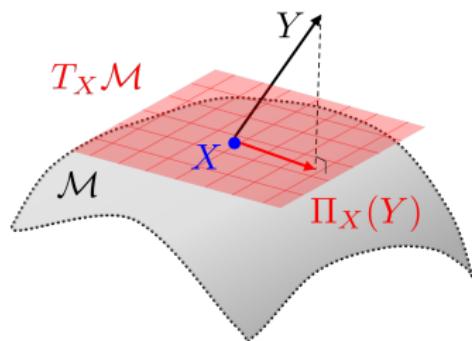
A projection  $\Pi$  on a vector space  $V$  is an idempotent map. I.e.,

$$\Pi^2(X) = \Pi(X) \quad (\forall X \in V). \quad (23)$$

### Def. (Orthogonal Projection)

Given an inner product  $\langle \cdot, \cdot \rangle$  on  $T_X\mathcal{M}$ , a projection is **orthogonal** if for any  $X, Y \in V$ ,

$$\langle \Pi(X), Y - \Pi(Y) \rangle = 0. \quad (24)$$



This is crucial for us to define the gradient and Hessian on the tangent space later.

Some examples of orthogonal projections onto the tangent space with respect to the Frobenius inner product are as follows:

### 1. Sym( $m$ ) and Skew( $m$ ):

Given a matrix  $X \in \mathbb{R}^{m \times m}$ , the respective orthogonal projections are

Some examples of orthogonal projections onto the tangent space with respect to the Frobenius inner product are as follows:

### 1. Sym( $m$ ) and Skew( $m$ ):

Given a matrix  $X \in \mathbb{R}^{m \times m}$ , the respective orthogonal projections are

$$\Pi_{\text{sym}}(X) = \text{sym}(\textcolor{red}{X}) := \frac{\textcolor{red}{X} + \textcolor{red}{X}^T}{2}, \quad (25)$$

$$\Pi_{\text{skew}}(X) = \text{skew}(\textcolor{red}{X}) := \frac{\textcolor{red}{X} - \textcolor{red}{X}^T}{2}. \quad (26)$$

Check the idempotence and orthogonality. Note that if different inner products are used, the orthogonal projection would be different.

$$\{\vec{x} \in \mathbb{R}^{n+1} \mid \vec{x}^T \vec{x} = 1\}$$

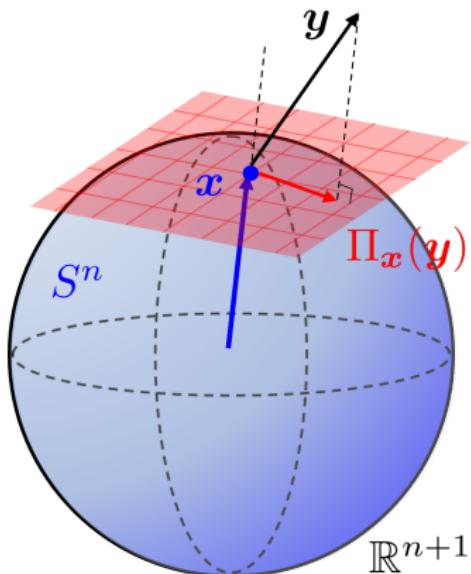
2.  $S^n$ :

The manifold of  $S^n$  is the set of unit vectors in  $\mathbb{R}^{n+1}$ .

Geometrically, we can obtain the orthogonal projection as

$$\Pi_x(y) = y - \langle x, y \rangle x, \quad (27)$$

where the inner product used is  
 $\langle x, y \rangle = x^T y$ .



However, for more obscure manifolds, is there a way for us to derive the orthogonal projection?

### Thm. (Regular Value Theorem, pt.III)

Let our manifold  $\mathcal{M} = f^{-1}(C)$  for a constant  $C$ . Then for  $X \in \mathcal{M}$ , we have the orthogonal projection of  $H$  onto  $T_X\mathcal{M}$  as

$$\Pi_X(\textcolor{red}{H}) = \left( \mathbb{1} - \left( \frac{\partial f}{\partial X} \right)^T \circ \left( \frac{\partial f}{\partial X} \circ \left( \frac{\partial f}{\partial X} \right)^T \right)^{-1} \circ \frac{\partial f}{\partial X} \right) [\textcolor{red}{H}], \quad (28)$$

However, for more obscure manifolds, is there a way for us to derive the orthogonal projection?

Thm. (Regular Value Theorem, pt. III)

Let our manifold  $\mathcal{M} = f^{-1}(C)$  for a constant  $C$ . Then for  $X \in \mathcal{M}$ , we have the orthogonal projection of  $H$  onto  $T_X\mathcal{M}$  as

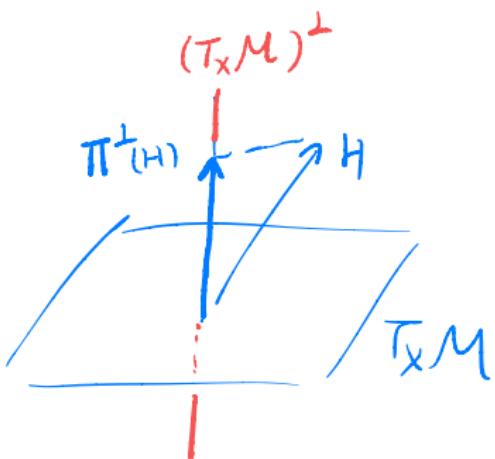
$$\Pi_X(\textcolor{red}{H}) = \left( \mathbb{1} - \left( \frac{\partial f}{\partial X} \right)^\top \circ \left( \frac{\partial f}{\partial X} \circ \left( \frac{\partial f}{\partial X} \right)^\top \right)^{-1} \circ \frac{\partial f}{\partial X} \right) [\textcolor{red}{H}], \quad (28)$$

where the **transposed** map is defined as

$$\begin{aligned} \langle x, y \rangle &= x^\top y \\ \langle x, Ay \rangle &= \langle Ax, y \rangle \end{aligned} \quad \left\langle \left( \frac{\partial f}{\partial X} \right)^\top [H], K \right\rangle = \left\langle H, \frac{\partial f}{\partial X}[K] \right\rangle. \quad (29)$$

The proof is as follows: consider any matrix  $H$ , and  $X \in \mathcal{M}$ . Let us denote the orthogonal complement projection  $\Pi_X^\perp$  as projection onto  $(T_X \mathcal{M})^\perp$ , then

$$\langle H - \Pi_X^\perp(H), (T_X \mathcal{M})^\perp \rangle = 0.$$



The proof is as follows: consider any matrix  $H$ , and  $X \in \mathcal{M}$ . Let us denote the orthogonal complement projection  $\Pi_X^\perp$  as projection onto  $(T_X \mathcal{M})^\perp$ , then

$$\langle H - \Pi_X^\perp(H), (T_X \mathcal{M})^\perp \rangle = 0.$$

Let us analyze each components. Since  $T_X \mathcal{M} = \ker(\underline{\partial f / \partial X})$  and  $\partial f / \partial X$  is full rank,

$$(T_X \mathcal{M})^\perp = \text{span}(\partial f / \partial X)^\top. \quad 0 = \begin{bmatrix} \textcircled{1} \\ \textcircled{2} \\ \vdots \\ \textcircled{n} \end{bmatrix} [\overset{\circ}{Y}]^\top \quad (30)$$

$\downarrow$   
 $T_X \mathcal{M}$   
 rowspace  $\perp T_X \mathcal{M}$

The proof is as follows: consider any matrix  $H$ , and  $X \in \mathcal{M}$ . Let us denote the orthogonal complement projection  $\Pi_X^\perp$  as projection onto  $(T_X \mathcal{M})^\perp$ , then

$$\left\langle H - \Pi_X^\perp(H), (T_X \mathcal{M})^\perp \right\rangle = 0.$$

Let us analyze each components. Since  $T_X \mathcal{M} = \ker(\partial f / \partial X)$  and  $\partial f / \partial X$  is full rank,

$$(T_X \mathcal{M})^\perp = \text{span}(\partial f / \partial X)^\top. \quad (30)$$

Hence, let us denote  $\Pi_X^\perp(H) = (\partial f / \partial X)^\top \hat{H}$ ,

The proof is as follows: consider any matrix  $H$ , and  $X \in \mathcal{M}$ . Let us denote the orthogonal complement projection  $\Pi_X^\perp$  as projection onto  $(T_X \mathcal{M})^\perp$ , then

$$\left\langle H - \Pi_X^\perp(H), \cancel{(T_X \mathcal{M})^\perp} \right\rangle = 0.$$

$(\frac{\partial f}{\partial X})^\top$

Let us analyze each components. Since  $T_X \mathcal{M} = \ker(\partial f / \partial X)$  and  $\partial f / \partial X$  is full rank,

$$(T_X \mathcal{M})^\perp = \text{span}(\partial f / \partial X)^\top. \quad (30)$$

$[000] \begin{bmatrix} * \\ * \\ * \end{bmatrix} = *0 + *0 + *0$

Hence, let us denote  $\Pi_X^\perp(H) = (\partial f / \partial X)^\top \hat{H}$ ,

$$\frac{\partial f}{\partial X} \left( H - \left( \frac{\partial f}{\partial X} \right)^\top \hat{H} \right) = 0 \Rightarrow \hat{H} = \left( \frac{\partial f}{\partial X} \left( \frac{\partial f}{\partial X} \right)^\top \right)^{-1} \frac{\partial f}{\partial X} [H]$$

$$\Pi_X(H) = (\mathbb{1} - \Pi_X^\perp)H \quad \blacksquare$$

$$\Pi_X(\textcolor{red}{H}) = \left( \mathbb{1} - \left( \frac{\partial f}{\partial X} \right)^T \circ \left( \frac{\partial f}{\partial X} \circ \left( \frac{\partial f}{\partial X} \right)^T \right)^{-1} \circ \frac{\partial f}{\partial X} \right) [\textcolor{red}{H}]$$

3. **St( $m, k$ ):** since  $f(X) = X^T X - \mathbb{1}_k$ ,

► For any  $K \in \mathbb{R}^{m \times k}$ ,  $\frac{\partial f}{\partial X}[K] = \textcolor{blue}{K^T X + X^T K}$

$$\Pi_X(\textcolor{red}{H}) = \left( \mathbb{1} - \left( \frac{\partial f}{\partial X} \right)^T \circ \left( \frac{\partial f}{\partial X} \circ \left( \frac{\partial f}{\partial X} \right)^T \right)^{-1} \circ \frac{\partial f}{\partial X} \right) [\textcolor{red}{H}]$$

3. **St( $m, k$ ):** since  $f(X) = X^T X - \mathbb{1}_k$ ,

- ▶ For any  $K \in \mathbb{R}^{m \times k}$ ,  $\frac{\partial f}{\partial X}[K] = K^T X + X^T K \in \text{Sym}(k)$ .
- ▶ For any  $H \in \text{Sym}(k)$  and  $K \in \mathbb{R}^{m \times k}$ ,

$$\left\langle H, \frac{\partial f}{\partial X}[K] \right\rangle_{\text{sym}} = \left\langle \left( \frac{\partial f}{\partial X} \right)^T [H], K \right\rangle_{\mathbb{R}^{m \times k}}$$

$$\text{Tr} \left\{ H^T (K^T X + X^T K) \right\} = \text{Tr} \left\{ K^T (XH + XH^T) \right\} = \left\langle \underline{\underline{XH}}, K \right\rangle$$

$$\left( \frac{\partial f}{\partial X} \right)^T [H]$$

$$\Pi_X(\textcolor{red}{H}) = \left( \mathbb{1} - \left( \frac{\partial f}{\partial X} \right)^\top \circ \left( \frac{\partial f}{\partial X} \circ \left( \frac{\partial f}{\partial X} \right)^\top \right)^{-1} \circ \frac{\partial f}{\partial X} \right) [\textcolor{red}{H}]$$

3. **St( $m, k$ ):** since  $f(X) = X^\top X - \mathbb{1}_k$ ,

► For any  $K \in \mathbb{R}^{m \times k}$ ,  $\frac{\partial f}{\partial X}[K] = K^\top X + X^\top K \in \text{Sym}(k)$ .

► For any  $H \in \text{Sym}(k)$  and  $K \in \mathbb{R}^{m \times k}$ ,

$$\left\langle H, \frac{\partial f}{\partial X}[K] \right\rangle_{\text{Sym}} = \text{Tr} \{ H^\top (K^\top X + X^\top K) \} = \langle 2XH, K \rangle_{\text{St}},$$

hence  $\left( \frac{\partial f}{\partial X} \right)^\top [H] = 2XH$ .

$$\Pi_X(\textcolor{red}{H}) = \left( \mathbb{1} - \left( \frac{\partial f}{\partial X} \right)^\top \circ \left( \frac{\partial f}{\partial X} \circ \left( \frac{\partial f}{\partial X} \right)^\top \right)^{-1} \circ \frac{\partial f}{\partial X} \right) [\textcolor{red}{H}]$$

3. **St( $m, k$ ):** since  $f(X) = X^\top X - \mathbb{1}_k$ ,

► For any  $K \in \mathbb{R}^{m \times k}$ ,  $\frac{\partial f}{\partial X}[K] = K^\top X + X^\top K \in \text{Sym}(k)$ .

► For any  $H \in \text{Sym}(k)$  and  $K \in \mathbb{R}^{m \times k}$ ,

$$\left\langle H, \frac{\partial f}{\partial X}[K] \right\rangle_{\text{Sym}} = \text{Tr} \{ H^\top (K^\top X + X^\top K) \} = \langle 2XH, K \rangle_{\text{St}},$$

hence  $\left( \frac{\partial f}{\partial X} \right)^\top [H] = 2XH$ .

►  $\frac{\partial f}{\partial X} \circ \left( \frac{\partial f}{\partial X} \right)^\top [H] = \frac{\partial f}{\partial X} [2XH] = 2H^\top X^\top X + 2X^\top XH = 4H$

$$\Pi_X(\textcolor{red}{H}) = \left( \mathbb{1} - \left( \frac{\partial f}{\partial X} \right)^\top \circ \left( \frac{\partial f}{\partial X} \circ \left( \frac{\partial f}{\partial X} \right)^\top \right)^{-1} \circ \frac{\partial f}{\partial X} \right) [\textcolor{red}{H}]$$

3. **St( $m, k$ ):** since  $f(X) = X^\top X - \mathbb{1}_k$ ,

► For any  $K \in \mathbb{R}^{m \times k}$ ,  $\frac{\partial f}{\partial X}[K] = K^\top X + X^\top K \in \text{Sym}(k)$ .

► For any  $H \in \text{Sym}(k)$  and  $K \in \mathbb{R}^{m \times k}$ ,

$$\left\langle H, \frac{\partial f}{\partial X}[K] \right\rangle_{\text{Sym}} = \text{Tr} \{ H^\top (K^\top X + X^\top K) \} = \langle 2XH, K \rangle_{\text{St}},$$

hence  $\left( \frac{\partial f}{\partial X} \right)^\top [H] = 2XH$ .

►  $\frac{\partial f}{\partial X} \circ \left( \frac{\partial f}{\partial X} \right)^\top [H] = \frac{\partial f}{\partial X}[2XH] = 4H$ .

► Finally,

$$\Pi_X(\textcolor{red}{H}) = \left( \mathbb{1} - 2X \cdot \frac{1}{4} \cdot \frac{\partial f}{\partial X} \right) [\textcolor{red}{H}]$$

$$\Pi_X(\textcolor{red}{H}) = \left( \mathbb{1} - \left( \frac{\partial f}{\partial X} \right)^\top \circ \left( \frac{\partial f}{\partial X} \circ \left( \frac{\partial f}{\partial X} \right)^\top \right)^{-1} \circ \frac{\partial f}{\partial X} \right) [\textcolor{red}{H}]$$

3. **St( $m, k$ ):** since  $f(X) = X^\top X - \mathbb{1}_k$ ,

► For any  $K \in \mathbb{R}^{m \times k}$ ,  $\frac{\partial f}{\partial X}[K] = K^\top X + X^\top K \in \text{Sym}(k)$ .

► For any  $H \in \text{Sym}(k)$  and  $K \in \mathbb{R}^{m \times k}$ ,

$$\left\langle H, \frac{\partial f}{\partial X}[K] \right\rangle_{\text{Sym}} = \text{Tr} \{ H^\top (K^\top X + X^\top K) \} = \langle 2XH, K \rangle_{\text{St}},$$

hence  $\left( \frac{\partial f}{\partial X} \right)^\top [H] = 2XH$ .

►  $\frac{\partial f}{\partial X} \circ \left( \frac{\partial f}{\partial X} \right)^\top [H] = \frac{\partial f}{\partial X}[2XH] = 4H$ .

► Finally,

$$\begin{aligned} \Pi_X(\textcolor{red}{H}) &= \left( \mathbb{1} - 2X \cdot (4)^{-1} \cdot \frac{\partial f}{\partial X} \right) [\textcolor{red}{H}] \\ &= \textcolor{red}{H} - \frac{X}{2} (\textcolor{red}{H}^\top X + X^\top \textcolor{red}{H}) = \textcolor{red}{H} - X \cdot \text{sym}(X^\top \textcolor{red}{H}). \end{aligned}$$

Thm. (Projection onto  $\text{St}(m, k)$ )

For  $X \in \text{St}(m, k)$ , the orthogonal projection from  $\mathbb{R}^{m \times k}$  to  $T_X \text{St}(m, k)$  is

$$\Pi_X(\textcolor{red}{H}) = \textcolor{red}{H} - X \cdot \text{sym}(X^T \textcolor{red}{H}). \quad (31)$$

Thm. (Projection onto  $\text{St}(m, k)$ )

For  $X \in \text{St}(m, k)$ , the orthogonal projection from  $\mathbb{R}^{m \times k}$  to  $T_X \text{St}(m, k)$  is

$$\Pi_X(\mathbf{H}) = \mathbf{H} - X \cdot \text{sym}(X^T \mathbf{H}). \quad (31)$$

Note that  $\text{SO}(m)$  is just a special case of the Stiefel manifold. Hence, they have the same formula for orthogonal projections. What's more, it can be further reduced to the form of

$$\begin{aligned}
\Pi_X(\mathbf{H}) &= \mathbf{H} - X \cdot \frac{1}{2} (X^T \mathbf{H} + \mathbf{H}^T X) \\
&= \mathbf{H} - \overbrace{\frac{1}{2} X^T \mathbf{H}}^{\in T_X \text{SO}(m)} - \frac{1}{2} X \mathbf{H}^T X \\
&= \frac{1}{2} (\mathbf{H} - X \mathbf{H}^T X) = X \cdot \frac{1}{2} (X^T \mathbf{H} - \mathbf{H}^T X) = X \cdot \text{skew}(X^T \mathbf{H})
\end{aligned}$$

Thm. (Projection onto  $\text{St}(m, k)$ )

For  $X \in \text{St}(m, k)$ , the orthogonal projection from  $\mathbb{R}^{m \times k}$  to  $T_X \text{St}(m, k)$  is

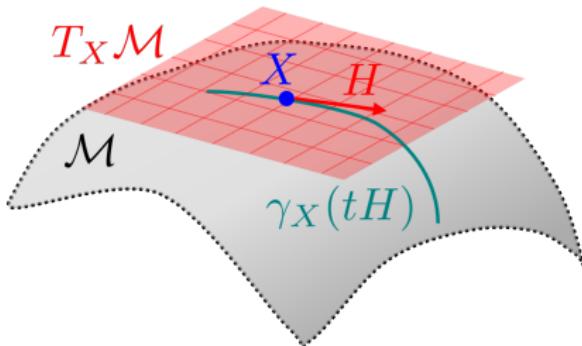
$$\Pi_X(\textcolor{red}{H}) = \textcolor{red}{H} - X \cdot \text{sym}(X^T \textcolor{red}{H}). \quad (31)$$

Note that  $\text{SO}(m)$  is just a special case of the Stiefel manifold. Hence, they have the same formula for orthogonal projections. What's more, it can be further reduced to the form of

$$\Pi_X(\textcolor{red}{H}) = X \cdot \text{skew}(X^T \textcolor{red}{H}) \in T_X \text{SO}(m). \quad (32)$$

For translation on the manifold, we need to introduce the idea of a geodesic. The geodesic is a *locally unique* curve that *minimizes* the distance between two points on the manifold.

To describe a geodesic on a given manifold  $\mathcal{M}$ , it has to satisfy the respective **geodesic equation**, a second order differential equation. The solution will be  $\gamma_X(tH)$ , a curve stemming from  $X \in \mathcal{M}$  with  $\dot{\gamma}_X(0) = H$  ( $H \in T_X\mathcal{M}$ ).



For example:

1.  $S^2$ :

The geodesic equation of it will be

$$\ddot{\gamma}(t) + \gamma(t)\dot{\gamma}(t)^T\dot{\gamma}(t) = 0, \quad \gamma(0) = \mathbf{x}, \quad \dot{\gamma}(0) = \mathbf{h} \quad (33)$$

under the usual Euclidean inner product.

For example:

1.  $S^2$ :

The geodesic equation of it will be

$$\ddot{\gamma}(t) + \gamma(t)\dot{\gamma}(t)^T\dot{\gamma}(t) = 0, \quad \gamma(0) = \mathbf{x}, \quad \dot{\gamma}(0) = \mathbf{h} \quad (33)$$

under the usual Euclidean inner product. The solution will be the *great circle* geodesic:


$$\gamma_{\mathbf{x}}(t\mathbf{h}) = \mathbf{x} \cos(t \|\mathbf{h}\|) + \frac{\mathbf{h}}{\|\mathbf{h}\|} \sin(t \|\mathbf{h}\|). \quad (34)$$

2.  $\text{SO}(m)$ :

The geodesic equation of it will be

$$\ddot{\Gamma}(t) + \Gamma(t)\dot{\Gamma}(t)^T\dot{\Gamma}(t) = \mathbf{0} \quad (35)$$

under the usual Frobenius inner product, with initial conditions

$$\Gamma(0) = X \in \text{SO}(m), \quad \dot{\Gamma}(0) = H \in X \cdot \text{Skew}(m). \quad (36)$$

## 2. $\text{SO}(m)$ :

The geodesic equation of it will be

$$\ddot{\Gamma}(t) + \Gamma(t)\dot{\Gamma}(t)^T\dot{\Gamma}(t) = \mathbf{0} \quad (35)$$

under the usual Frobenius inner product, with initial conditions

$$\Gamma(0) = X \in \text{SO}(m), \quad \dot{\Gamma}(0) = H \in X \cdot \text{Skew}(m). \quad (36)$$

The solution will be:

$$\Gamma_X(tH) = X \exp(\underbrace{tX^T H}_{\Omega \in \text{Skew}(m)}) = \exp(tHX^T)X. \quad (37)$$

$[\exp(\Omega)]^T = \exp(-\Omega)$   
 $\exp(\Omega) \in \text{SO}(m)$

This result has deep relation with Lie algebra, which we will not dig into in this presentation.

However, note that geodesic equations are really hard to solve.

### 3. $\text{St}(m, k)$ :

The geodesic equation of it will be

$$\ddot{\Gamma}(t) + \Gamma(t)\dot{\Gamma}(t)^T\dot{\Gamma}(t), \quad \Gamma(0) = X, \quad \dot{\Gamma}(0) = H \quad (38)$$

under the usual Frobenius inner product. The solution will be:

$$\Gamma_X(tH) = \exp(t(HX^T - XH^T)) \cdot X \cdot \exp(-tX^TH). \quad (39)$$

Note that the geodesic above coincides with that of  $\text{SO}(m)$  when  $m = k$ .

- 1 Recap: Matrix Differentiation
- 2 Geometry of a Manifold
- 3 Differentiation
  - Derivatives ■ Gradient and Hessian
- 4 Manifold Optimization

So how do we actually calculate the derivative on a manifold?  
Similar to the case of derivative over unconstrained matrices – we shall use the **directional derivative**. However, we need to **stay on the manifold**.

$$\frac{\partial f}{\partial x} [H] = \frac{d}{dt} f(x + tH) \Big|_0$$

So how do we actually calculate the derivative on a manifold? Similar to the case of derivative over unconstrained matrices – we shall use the **directional derivative**. However, we need to **stay on the manifold**. Hence, **geodesics** are used instead of addition.

### Def. (Differential Map)

Given a manifold  $\mathcal{M}$  with inner product  $\langle \cdot, \cdot \rangle$  and a function  $f$  defined on it. The differential map of  $f$  at  $X \in \mathcal{M}$  is defined to be

$$\frac{\partial f}{\partial X}[H] = \left. \frac{d}{dt} f(\gamma_X(tH)) \right|_{t=0} \quad (40)$$

where  $\gamma_X(tH)$  is the geodesic stemming from  $X$  in the direction of  $H \in T_X \mathcal{M}$ .

Given two manifolds  $\mathcal{M}$  and  $\mathcal{N}$ , the differential of the map  $f : \mathcal{M} \rightarrow \mathcal{N}$  has a fancy name termed the **pushforward**<sup>1</sup> besides our usual name of **differential map**.

### Def. (Pushforward)

The pushforward of  $f : \mathcal{M} \rightarrow \mathcal{N}$  at  $X \in \mathcal{M}$  is

$$\frac{\partial f}{\partial X} : T_X \mathcal{M} \rightarrow T_{f(X)} \mathcal{N}. \quad (41)$$



<sup>1</sup>More often, it is denoted as  $Df(X)$  instead of  $\partial f / \partial X$ .

Given two manifolds  $\mathcal{M}$  and  $\mathcal{N}$ , the differential of the map  $f : \mathcal{M} \rightarrow \mathcal{N}$  has a fancy name termed the **pushforward**<sup>1</sup> besides our usual name of **differential map**.

### Def. (Pushforward)

The pushforward of  $f : \mathcal{M} \rightarrow \mathcal{N}$  at  $X \in \mathcal{M}$  is

$$\frac{\partial f}{\partial X} : T_X \mathcal{M} \rightarrow T_{f(X)} \mathcal{N}. \quad (41)$$

Remember that the differential map is the mapping from the change in input to the change in output: consider a smooth curve  $\gamma(t) \in \mathcal{M}$  with  $\gamma(0) = X$ . then

$$\frac{\partial f}{\partial X} \left[ \underbrace{\dot{\gamma}(0)}_{\in T_X \mathcal{M}} \right] = \frac{d}{dt} (f \circ \gamma)(0) \in T_{f(X)} \mathcal{N}. \quad \blacksquare$$

---

<sup>1</sup>More often, it is denoted as  $Df(X)$  instead of  $\partial f / \partial X$ .

Let us make ourselves get used to the idea of pushforward: a mapping from tangent space to tangent space.

If  $X$  has a QR decomposition into  $X = QR$ , find  $\frac{\partial Q}{\partial X}[H]$  where  $H \in T_X \mathbb{R}^{m \times m} = \mathbb{R}^{m \times m}$ .

### Def. (QR Decomposition)

Any square matrix  $X \in \mathbb{R}^{m \times m}$  can be decomposed into a product of orthogonal matrix  $Q$  and upper triangular matrix  $R$ , i.e.

$$X = QR. \quad (42)$$

[Sol.] Since  $\mathbb{R}^{m \times m}$  is a vector space, the geodesic on it is just a straight line, i.e.  $\Gamma_X(tH) = X + tH$ . This returns to our usual unconstrained matrix differentiation.

$$X = QR$$

$$H = \frac{\partial Q}{\partial X}[H] \cdot R + Q \cdot \frac{\partial R}{\partial X}[H]$$

[Sol.] Since  $\mathbb{R}^{m \times m}$  is a vector space, the geodesic on it is just a straight line, i.e.  $\Gamma_X(tH) = X + tH$ . This returns to our usual unconstrained matrix differentiation.

$$H = \frac{\partial Q}{\partial X}[H] \cdot R + Q \cdot \frac{\partial R}{\partial X}[H]$$

$$Q^T H R^{-1} = Q^T \underbrace{\frac{\partial Q}{\partial X}[H]}_{T_Q O^{(m)} \text{ IIz}} + \underbrace{\frac{\partial R}{\partial X}[H] \cdot R^{-1}}_{T_R(\text{上三角}) \cong \text{上三角}} \in \text{上三角}$$

$Q \cdot \text{Skew}^{(m)}$

[Sol.] Since  $\mathbb{R}^{m \times m}$  is a vector space, the geodesic on it is just a straight line, i.e.  $\Gamma_X(tH) = X + tH$ . This returns to our usual unconstrained matrix differentiation.

$$H = \frac{\partial Q}{\partial X}[H] \cdot R + Q \cdot \frac{\partial R}{\partial X}[H]$$

$$Q^\top H R^{-1} = \underbrace{Q^\top \frac{\partial Q}{\partial X}[H]}_{\in \text{Skew}(m)} + \underbrace{\frac{\partial R}{\partial X}[H] \cdot R^{-1}}_{\text{upper triangular}}$$

$$\rightarrow \frac{\partial Q}{\partial X}[H] = Q \cdot \widetilde{\text{skew}}(Q^T H R^{-1}),$$

$$= \begin{bmatrix} 0 & 0 & -3 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

where  $\widetilde{\text{skew}}(\cdot)$  is the **unique** decomposition of a matrix into its skew symmetric part and upper-triangular part, the skew symmetric part is chosen.

Notes:

$$X(t) = \begin{bmatrix} t+1 & 0 \\ e^t & 2(t^2+1) \end{bmatrix} = \underbrace{\begin{bmatrix} (t+1)/\sqrt{(t+1)^2 + e^{2t}} & -e^t/\sqrt{(t+1)^2 + e^{2t}} \\ e^t/\sqrt{(t+1)^2 + e^{2t}} & (t+1)/\sqrt{(t+1)^2 + e^{2t}} \end{bmatrix}}_{Q(t)} \underbrace{\begin{bmatrix} \sqrt{(t+1)^2 + e^{2t}} & e^t \cdot 2(t^2+1)/\sqrt{(t+1)^2 + e^{2t}} \\ 0 & (t+1) \cdot 2(t^2+1)/\sqrt{(t+1)^2 + e^{2t}} \end{bmatrix}}_{R(t)}$$

$$X(0) = \begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \sqrt{2} & \sqrt{2} \\ 0 & \sqrt{2} \end{bmatrix}$$

$$\dot{Q}(0) = \frac{\partial Q}{\partial X} [\dot{X}(0)]$$

Notes:

$$X(t) = \begin{bmatrix} t+1 & 0 \\ e^t & 2(t^2+1) \end{bmatrix} = \underbrace{\begin{bmatrix} (t+1)/\sqrt{(t+1)^2 + e^{2t}} & -e^t/\sqrt{(t+1)^2 + e^{2t}} \\ e^t/\sqrt{(t+1)^2 + e^{2t}} & (t+1)/\sqrt{(t+1)^2 + e^{2t}} \end{bmatrix}}_{Q(t)} \underbrace{\begin{bmatrix} \sqrt{(t+1)^2 + e^{2t}} & e^t \cdot 2(t^2+1)/\sqrt{(t+1)^2 + e^{2t}} \\ 0 & (t+1) \cdot 2(t^2+1)/\sqrt{(t+1)^2 + e^{2t}} \end{bmatrix}}_{R(t)}$$

$$X(0) = \begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \sqrt{2} & \sqrt{2} \\ 0 & \sqrt{2} \end{bmatrix}$$

$$\begin{aligned} \dot{Q}(0) &= Q(0) \cdot \widetilde{\text{skew}}(Q^T(0) \dot{X}(0) R(0)) \\ &= \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \cdot \widetilde{\text{skew}} \left( \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} \end{bmatrix} \right) \\ &= \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \cdot \widetilde{\text{skew}} \left( \begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix} \right) = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \cdot 0 = 0 \end{aligned}$$

Notes:

$$X(t) = \begin{bmatrix} t+1 & 0 \\ e^t & 2(t^2+1) \end{bmatrix} = \underbrace{\begin{bmatrix} (t+1)/\sqrt{(t+1)^2 + e^{2t}} & -e^t/\sqrt{(t+1)^2 + e^{2t}} \\ e^t/\sqrt{(t+1)^2 + e^{2t}} & (t+1)/\sqrt{(t+1)^2 + e^{2t}} \end{bmatrix}}_{Q(t)} \underbrace{\begin{bmatrix} \sqrt{(t+1)^2 + e^{2t}} & e^t \cdot 2(t^2+1)/\sqrt{(t+1)^2 + e^{2t}} \\ 0 & (t+1) \cdot 2(t^2+1)/\sqrt{(t+1)^2 + e^{2t}} \end{bmatrix}}_{R(t)}$$

$$X(0) = \begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \sqrt{2} & \sqrt{2} \\ 0 & \sqrt{2} \end{bmatrix}$$

$$\dot{Q}(0) = Q(0) \cdot \widetilde{\text{skew}}(Q^T(0) \dot{X}(0) R'(0))$$

$$\begin{aligned} &= \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \cdot \widetilde{\text{skew}} \left( \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} \end{bmatrix} \right) \\ &= \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \cdot \widetilde{\text{skew}} \left( \begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix} \right) = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \cdot 0 = 0 \end{aligned}$$

$$\dot{Q}(t) \Big|_{t=0} = \begin{bmatrix} 1/\sqrt{(t+1)^2 + e^{2t}} - (t+1)(t+1 + e^{2t})/(\sqrt{(t+1)^2 + e^{2t}})^3 & -e^t/\sqrt{(t+1)^2 + e^{2t}} + e^t(t+1 + e^{2t})/(\sqrt{(t+1)^2 + e^{2t}})^3 \\ e^t/\sqrt{(t+1)^2 + e^{2t}} - e^t(t+1 + e^{2t})/(\sqrt{(t+1)^2 + e^{2t}})^3 & 1/\sqrt{(t+1)^2 + e^{2t}} - (t+1)(t+1 + e^{2t})/(\sqrt{(t+1)^2 + e^{2t}})^3 \end{bmatrix} \Big|_{t=0} = 0$$

## Thm. (Riesz Representation Theorem)

For a vector space  $V$  with inner product  $\langle \cdot, \cdot \rangle$  and a linear scalar function  $f$  defined on it, there exists a vector  $\mathbf{u} \in V$  such that for all  $\mathbf{v} \in V$ ,

$$f(\mathbf{v}) = \langle \mathbf{u}, \mathbf{v} \rangle. \quad (43)$$

This theorem states that any **linear function** can be represented as an **inner product**.

### Thm. (Riesz Representation Theorem)

For a vector space  $V$  with inner product  $\langle \cdot, \cdot \rangle$  and a linear scalar function  $f$  defined on it, there exists a vector  $\mathbf{u} \in V$  such that for all  $\mathbf{v} \in V$ ,

$$f(\mathbf{v}) = \langle \mathbf{u}, \mathbf{v} \rangle. \quad (43)$$

This theorem states that any **linear function** can be represented as an **inner product**.

We have the first derivative defined on the last page, it is a linear function. Since the domain (input) of the linear function is the  $T_X\mathcal{M}$ , a vector space, we can apply the theorem above.

## Def. (Riemannian Gradient)

Given an inner product  $\langle \cdot, \cdot \rangle$  on  $T_X\mathcal{M}$ , the **Riemannian gradient** of a function  $f(X)$  is  $\text{grad}f(X) \in T_X\mathcal{M}$ , defined as

$$\frac{\partial f}{\partial X}[H] = \langle \text{grad}f(X), H \rangle \quad (\forall H \in T_X\mathcal{M}). \quad (44)$$

## Def. (Riemannian Gradient)

Given an inner product  $\langle \cdot, \cdot \rangle$  on  $T_X\mathcal{M}$ , the **Riemannian gradient** of a function  $f(X)$  is  $\text{grad}f(X) \in T_X\mathcal{M}$ , defined as

$$\frac{\partial f}{\partial X}[H] = \langle \text{grad}f(X), H \rangle \quad (\forall H \in T_X\mathcal{M}). \quad (44)$$

e.g. Given  $Y \in \mathbb{R}^{m \times n}$  and  $Z \in \mathbb{R}^{k \times n}$ , compute the gradient of

$$f : \text{St}(m, k) \rightarrow \mathbb{R}, \quad f(X) := \|X^T Y - Z\|^2 \quad (45)$$

with respect to the Frobenius inner product.  $\|X\|^2 = \text{Tr}\{X^T X\}$

[Sol.] Recall that  $\Pi_X(H) = H - X \cdot \text{sym}(X^T H)$ ,

$$\frac{\partial}{\partial X} \left[ \text{Tr}\{(X^T Y - Z)(X^T Y - Z)^T\} \right]_H = \text{Tr}\{2H^T Y (X^T Y - Z)^T\}$$

## Def. (Riemannian Gradient)

Given an inner product  $\langle \cdot, \cdot \rangle$  on  $T_X \mathcal{M}$ , the **Riemannian gradient** of a function  $f(X) \in T_X \mathcal{M}$ , defined as

$$\frac{\partial f}{\partial X}[H] = \langle \text{grad}f(X), H \rangle \quad (\forall H \in T_X \mathcal{M}). \quad (44)$$

e.g. Given  $Y \in \mathbb{R}^{m \times n}$  and  $Z \in \mathbb{R}^{k \times n}$ , compute the gradient of

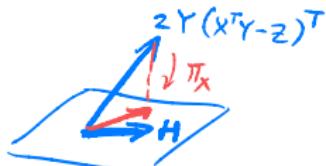
$$f : \text{St}(m, k) \rightarrow \mathbb{R}, \quad f(X) := \|X^T Y - Z\|^2 \quad (45)$$

with respect to the Frobenius inner product.

[Sol.] Recall that  $\Pi_X(\mathbf{H}) = \mathbf{H} - X \cdot \text{sym}(X^T \mathbf{H})$ ,

$$\frac{\partial f}{\partial X}[H] = \langle 2Y(X^T Y - Z)^T, H \rangle$$

$$\text{grad}f(X) = \Pi_X(2Y(X^T Y - Z)^T) = 2Y(X^T Y - Z)^T - X \cdot \text{sym}(X^T 2Y(X^T Y - Z)^T) \quad \blacksquare$$



Similarly, the second derivative is also defined via the directional derivative using geodesics:

### Def. (Second Derivative)

Given a manifold  $\mathcal{M}$  with inner product  $\langle \cdot, \cdot \rangle$  and a function  $f$  defined on it. The second derivative of  $f$  at  $X \in \mathcal{M}$  is defined to be

$$\frac{\partial^2 f}{\partial X^2}[H] = \left. \frac{d^2}{dt^2} f(\gamma_X(tH)) \right|_{t=0} \quad (46)$$

where  $\gamma_X(tH)$  is the geodesic stemming from  $X$  in the direction of  $H \in T_X \mathcal{M}$ . Further, using the **polarization identity**, we can obtain the bilinear form

$$\frac{\partial^2 f}{\partial X^2}[H, K] = \frac{1}{4} \left( \frac{\partial^2 f}{\partial X^2}[H + K] - \frac{\partial^2 f}{\partial X^2}[H - K] \right). \quad (47)$$

### Def. (Riemannian Hessian)

Given an inner product  $\langle \cdot, \cdot \rangle$  on  $T_X\mathcal{M}$ , the **Riemannian Hessian** of a function  $f(X)$  is a linear map  $\text{hess}f(X) : T_X\mathcal{M} \rightarrow T_X\mathcal{M}$ , defined as

$$\frac{\partial^2 f}{\partial X^2}[H, K] = \langle \text{hess}f(X)[H], K \rangle \quad (\forall H, K \in T_X\mathcal{M}). \quad (48)$$

$$\begin{aligned} & \langle v, H \rangle , H \in T_X\mathcal{M} \\ & \Downarrow \\ & \langle \pi_{T_X}(v), H \rangle \end{aligned}$$

## Example

Let  $A$  be a positive definite  $n \times n$  matrix and define

$$f : S^{n-1} \rightarrow \mathbb{R}, \quad \mathbf{x} \mapsto \frac{1}{2} \mathbf{x}^\top A \mathbf{x}.$$

Remember that the projection onto  $T_{\mathbf{x}} S^{n-1}$  is  $\Pi_{\mathbf{x}} = \mathbb{1}_n - \mathbf{x}\mathbf{x}^\top$ , please calculate the Riemannian gradient and Hessian to the function above.

[Sol.] The geodesic is  $\gamma_{\mathbf{x}}(t\mathbf{h}) = \mathbf{x} \cos(t \|\mathbf{h}\|) + \frac{\mathbf{h}}{\|\mathbf{h}\|} \sin(t \|\mathbf{h}\|)$ ,

$$\frac{\partial f}{\partial \mathbf{x}}[\mathbf{h}] = \frac{d}{dt} \left. \frac{1}{2} \mathbf{\gamma}^\top A \mathbf{\gamma} \right|_0 = \left. \mathbf{\gamma}^\top A \dot{\mathbf{\gamma}} \right|_0 = \mathbf{\gamma}^\top A \mathbf{h} = \langle A\mathbf{\gamma}, \mathbf{h} \rangle$$

## Example

Let  $A$  be a positive definite  $n \times n$  matrix and define

$$f : S^{n-1} \rightarrow \mathbb{R}, \quad \mathbf{x} \mapsto \frac{1}{2} \mathbf{x}^\top A \mathbf{x}.$$

Remember that the projection onto  $T_{\mathbf{x}} S^{n-1}$  is  $\Pi_{\mathbf{x}} = \mathbb{1}_n - \mathbf{x}\mathbf{x}^\top$ , please calculate the Riemannian gradient and Hessian to the function above.

[Sol.] The geodesic is  $\gamma_{\mathbf{x}}(t\mathbf{h}) = \mathbf{x} \cos(t \|\mathbf{h}\|) + \frac{\mathbf{h}}{\|\mathbf{h}\|} \sin(t \|\mathbf{h}\|)$ ,

$$\text{grad } f(\mathbf{x}) = (\mathbb{1}_n - \mathbf{x}\mathbf{x}^\top) A \mathbf{x}$$

$$\frac{d}{dt} \gamma^\top A \dot{\gamma} \Big|_0 = \dot{\gamma}^\top A \dot{\gamma} + \gamma^\top A \ddot{\gamma} \Big|_0$$

## Example

Let  $A$  be a positive definite  $n \times n$  matrix and define

$$f : S^{n-1} \rightarrow \mathbb{R}, \quad \mathbf{x} \mapsto \frac{1}{2} \mathbf{x}^\top A \mathbf{x}.$$

Remember that the projection onto  $T_{\mathbf{x}} S^{n-1}$  is  $\Pi_{\mathbf{x}} = \mathbb{1}_n - \mathbf{x}\mathbf{x}^\top$ , please calculate the Riemannian gradient and Hessian to the function above.

[Sol.] The geodesic is  $\gamma_{\mathbf{x}}(t\mathbf{h}) = \mathbf{x} \cos(t \|\mathbf{h}\|) + \frac{\mathbf{h}}{\|\mathbf{h}\|} \sin(t \|\mathbf{h}\|)$ ,

$$\text{grad } f(\mathbf{x}) = (\mathbb{1}_n - \mathbf{x}\mathbf{x}^\top) A \mathbf{x}$$

$$\frac{\partial^2 f}{\partial \mathbf{x}^2}[\mathbf{h}] = -(\mathbf{h}^\top \mathbf{h}) \mathbf{x}^\top A \mathbf{x} + \mathbf{h}^\top A \mathbf{h}$$

## Example

Let  $A$  be a positive definite  $n \times n$  matrix and define

$$f : S^{n-1} \rightarrow \mathbb{R}, \quad \mathbf{x} \mapsto \frac{1}{2} \mathbf{x}^\top A \mathbf{x}.$$

Remember that the projection onto  $T_{\mathbf{x}} S^{n-1}$  is  $\Pi_{\mathbf{x}} = \mathbb{1}_n - \mathbf{x}\mathbf{x}^\top$ , please calculate the Riemannian gradient and Hessian to the function above.

[Sol.] The geodesic is  $\gamma_{\mathbf{x}}(t\mathbf{h}) = \mathbf{x} \cos(t \|\mathbf{h}\|) + \frac{\mathbf{h}}{\|\mathbf{h}\|} \sin(t \|\mathbf{h}\|)$ ,

$$\text{grad } f(\mathbf{x}) = (\mathbb{1}_n - \mathbf{x}\mathbf{x}^\top) A \mathbf{x}$$

$$\frac{\partial^2 f}{\partial \mathbf{x}^2}[\mathbf{h}] = -(\mathbf{h}^\top \mathbf{h}) \mathbf{x}^\top A \mathbf{x} + \mathbf{h}^\top A \mathbf{h}$$

$$\frac{\partial^2 f}{\partial \mathbf{x}}[\mathbf{h}, \mathbf{k}] = \mathbf{k}^\top A \mathbf{h} - (\mathbf{k}^\top \mathbf{h}) \mathbf{x}^\top A \mathbf{x} = \langle A\mathbf{h} - \mathbf{x}^\top A \mathbf{x} \mathbf{h}, \mathbf{k} \rangle$$

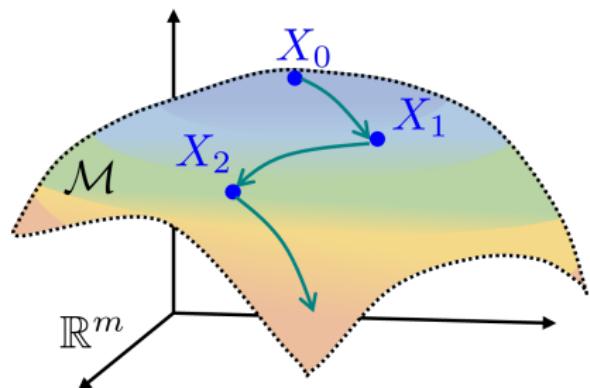
$$\text{hess } f(\mathbf{x})[\mathbf{h}] = (\mathbb{1}_n - \mathbf{x}\mathbf{x}^\top)(A - \mathbf{x}^\top A \mathbf{x} \mathbb{1}_n)\mathbf{h} \quad \blacksquare$$

- 1 Recap: Matrix Differentiation
- 2 Geometry of a Manifold
- 3 Differentiation
- 4 Manifold Optimization
  - Gradient Descent ■ Newton's Method

Now that we have a manifold  $\mathcal{M}$  and a scalar function  $f : \mathcal{M} \rightarrow \mathbb{R}$  defined on it. Let us find a way to optimize it!

What tools do we have?

- ▶ Manifold structure,
- ▶ Tangent space,
- ▶ Projection,
- ▶ Gradient,
- ▶ Hessian,
- ▶ Geodesics.



The scheme for gradient descent on a manifold is similar to that of its vector space counterpart. The pseudocode is as below<sup>2</sup>:

---

**Algorithm 1:** Gradient Descent on Manifold

---

**Input:** Function  $f(X)$  and initial condition  $X_0$ .

**Output:** Optimal  $X$  that minimizes  $f$  locally.

---

<sup>2</sup>The actual code implementation requires some extra care.

The scheme for gradient descent on a manifold is similar to that of its vector space counterpart. The pseudocode is as below<sup>2</sup>:

---

**Algorithm 1:** Gradient Descent on Manifold

---

**Input:** Function  $f(X)$  and initial condition  $X_0$ .

**Output:** Optimal  $X$  that minimizes  $f$  locally.

- 1 Set acceptable error;
  - 2 Initialize  $X = X_0$ ;
- 

<sup>2</sup>The actual code implementation requires some extra care.

The scheme for gradient descent on a manifold is similar to that of its vector space counterpart. The pseudocode is as below<sup>2</sup>:

---

**Algorithm 1:** Gradient Descent on Manifold

---

**Input:** Function  $f(X)$  and initial condition  $X_0$ .

**Output:** Optimal  $X$  that minimizes  $f$  locally.

- 1 Set acceptable error;
  - 2 Initialize  $X = X_0$ ;
  - 3 **while**  $\|X_{new} - X_{old}\| > error$  **do**
  - 4     Calculate gradient  $\text{grad}f(X)$ ;
  - 5     Set step size as  $\alpha$ ;
  - 6     Update  $X \leftarrow \gamma_X (-\alpha \cdot \text{grad}f(X))$ .
  - 7 **end**
- 

<sup>2</sup>The actual code implementation requires some extra care.

## Eigenvalue Decomposition

Let us consider minimizing the cost function

$$f(X) = \text{Tr} \{ X^T A X B \}, \quad \text{Brackett Fn.} \quad (49)$$

where  $X \in \text{SO}(n)$ ,  $A$  and  $B$  are non-degenerate<sup>3</sup> and symmetric.  
[sol.]

---

<sup>3</sup>A matrix is non-degenerate if no two eigenvalues of it are the same.

Let us consider minimizing the cost function

$$f(X) = \text{Tr} \{ X^T A X B \}, \quad (49)$$

where  $X \in \text{SO}(n)$ ,  $A$  and  $B$  are non-degenerate<sup>3</sup> and symmetric.

[sol.] The gradient is calculated to be

$$\begin{aligned} \frac{\partial f}{\partial X}[H] &= \text{Tr} \{ H^T A X B + X^T A H B \} = \text{Tr} \{ 2 H^T A X B \} \\ &= \langle H, 2 A X B \rangle \end{aligned}$$

$$\text{grad } f(X) = \Pi_X (2 A X B) = 2 X \cdot \text{skew}(X^T A X B)$$

---

<sup>3</sup>A matrix is non-degenerate if no two eigenvalues of it are the same.

Let us consider minimizing the cost function

$$f(X) = \text{Tr} \{ X^T A X B \}, \quad (49)$$

where  $X \in \text{SO}(n)$ ,  $A$  and  $B$  are non-degenerate<sup>3</sup> and symmetric.

[sol.] The gradient is calculated to be

$$\frac{\partial f}{\partial X}[H] = 2\text{Tr} \{ H^T A X B \} \rightarrow \text{grad } f(X) = 2X \cdot \text{skew}(X^T A X B)$$

Extremas occur at  $\text{grad } f(X) = 0$ ,

$$\begin{aligned} X^T A X B \text{ is symmetric} \\ \Leftrightarrow X^T A X \text{ & } B \text{ commutes} \\ (X^T A X B)^T = B X^T A X \end{aligned}$$

---

<sup>3</sup>A matrix is non-degenerate if no two eigenvalues of it are the same.

## Eigenvalue Decomposition

Let us consider minimizing the cost function

$$f(X) = \text{Tr} \{ X^T A X B \}, \quad (49)$$

where  $X \in \text{SO}(n)$ ,  $A$  and  $B$  are non-degenerate<sup>3</sup> and symmetric.

[sol.] The gradient is calculated to be

$$\frac{\partial f}{\partial X}[H] = \text{Tr} \{ H^T A X B \} \rightarrow \text{grad } f(X) = X \cdot \text{skew}(X^T A X B)$$

Extremas occur at  $\text{grad } f(X) = 0$ , i.e.  $X^T A X B$  is symmetric: if  $B$  is diagonal,

$$X^T A X \begin{bmatrix} \mu_1 & & \\ & \mu_2 & \\ & & \ddots \end{bmatrix} \Rightarrow X^T A X \text{ is also diagonal!}$$

---

<sup>3</sup>A matrix is non-degenerate if no two eigenvalues of it are the same.

## Eigenvalue Decomposition

Let us consider minimizing the cost function

$$f(X) = \text{Tr} \{ X^T A X B \}, \quad (49)$$

where  $X \in \text{SO}(n)$ ,  $A$  and  $B$  are non-degenerate<sup>3</sup> and symmetric.

[sol.] The gradient is calculated to be

$$\frac{\partial f}{\partial X}[H] = \text{Tr} \{ H^T A X B \} \rightarrow \text{grad } f(X) = X \cdot \text{skew}(X^T A X B)$$

Extremas occur at  $\text{grad } f(X) = 0$ , i.e.  $X^T A X B$  is symmetric: if  $B$  is diagonal,

$$(X^T A X)B = B(X^T A X) \rightarrow X^T A X \text{ is diagonal.}$$

Hence, by optimizing  $f$ , we are able to conduct eigenvalue decomposition on  $A$ .

---

<sup>3</sup>A matrix is non-degenerate if no two eigenvalues of it are the same.

We can further analyze which points are maxima and which are minima: *without loss of generality*, let us assume that  $B = \text{diag}(\mu_1, \mu_2, \dots, \mu_n)$  with  $\mu_i < \mu_j$  for  $i < j$ . At extrema, we also have

$$\mathcal{A} := X^T \textcolor{blue}{A} X = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n).$$

$$f(X) = \text{Tr}\{AB\} = \lambda_1 \mu_1 + \lambda_2 \mu_2 + \dots + \lambda_n \mu_n$$

排序不等式

$\mu_i$  increasing ,  $\begin{cases} \lambda_i \text{ increasing} \Rightarrow \text{maxima} \\ \lambda_i \text{ decreasing} \Rightarrow \text{minima} . \end{cases}$

We can further analyze which points are maxima and which are minima: *without loss of generality*, let us assume that  $B = \text{diag}(\mu_1, \mu_2, \dots, \mu_n)$  with  $\mu_i < \mu_j$  for  $i < j$ . At extrema, we also have

$$\mathcal{A} := X^T \mathbf{A} X = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n).$$

The Hessian can be calculated to be

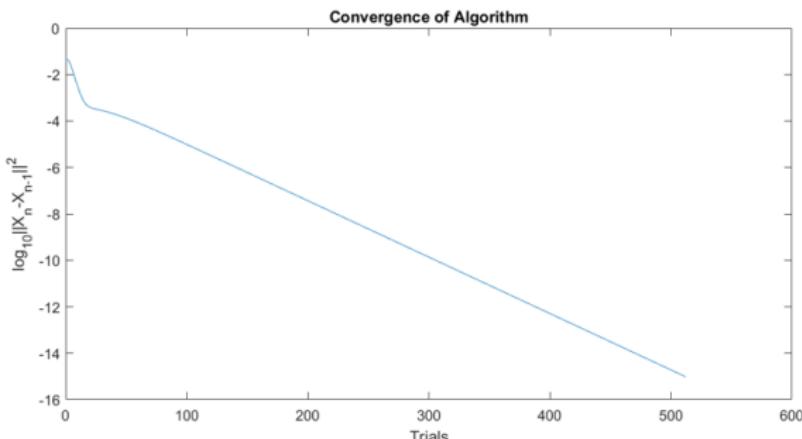
$$\text{hess } f(X)[X\Omega] = \underset{\mathbf{A}\in\text{Skew}(m)}{\cancel{X}} \text{skew} (\mathcal{A}\Omega B - \text{sym}(\mathcal{A}B)\Omega). \quad (50)$$

It can be calculated that for  $\lambda_i < \lambda_j$  ( $i < j$ ), the Hessian is negative definite; similarly for  $\lambda_i > \lambda_j$  ( $i < j$ ), the Hessian is positive definite; the other cases are all indefinite.

During each step of optimization, we can update  $X$  via

$$X \mapsto X \cdot \exp(-\alpha \cdot \text{skew}(X^T A X B)),$$

where  $\alpha$  is a suitable step size. If  $\alpha > 0$ , it converges to minima; if  $\alpha < 0$ , it converges to maxima. The code terminates when  $X$  converges.



Other optimization methods exist.

**Newton's method** is used to iteratively find the zero of a function. If we want to solve for the zero of  $g(x)$  ( $x \in \mathbb{R}$ ), then we can update  $x$  via the following iterative map:

$$x \mapsto x - \frac{g(x)}{g'(x)}. \quad (51)$$

Other optimization methods exist.

**Newton's method** is used to iteratively find the zero of a function. If we want to solve for the zero of  $g(x)$  ( $x \in \mathbb{R}$ ), then we can update  $x$  via the following iterative map:

$$x \mapsto x - \frac{g(x)}{g'(x)}. \quad (51)$$

To extremize a function  $f$  equals to finding the zero of its gradient. For Newton's method on vector spaces, we have the following iterative map:

$$\mathbf{x} \mapsto \mathbf{x} - (\text{hess } f(\mathbf{x}))^{-1} \text{grad } f(\mathbf{x}). \quad (52)$$

---

**Algorithm 2:** Newton's Method on Manifold

---

**Input:** Function  $f(X)$  and initial condition  $X_0$ .

**Output:** Optimal  $X$  that extremizes  $f$  locally.

- 1 Set acceptable error;
  - 2 Initialize  $X = X_0$ ;
  - 3 **while**  $\|X_{new} - X_{old}\| > error$  **do**
  - 4     Calculate Newton's step  $\Omega = -(\text{hess } f(X))^{-1} \text{grad } f(X)$ ;
  - 5     Update  $X \mapsto \gamma_X(\Omega)$ .
  - 6 **end**
-

---

**Algorithm 2:** Newton's Method on Manifold

---

**Input:** Function  $f(X)$  and initial condition  $X_0$ .

**Output:** Optimal  $X$  that extremizes  $f$  locally.

- 1 Set acceptable error;
  - 2 Initialize  $X = X_0$ ;
  - 3 **while**  $\|X_{new} - X_{old}\| > error$  **do**
  - 4     Calculate Newton's step  $\Omega = -(\text{hess } f(X))^{-1} \text{grad } f(X)$ ;
  - 5     Update  $X \mapsto \gamma_X(\Omega)$ .
  - 6 **end**
- 

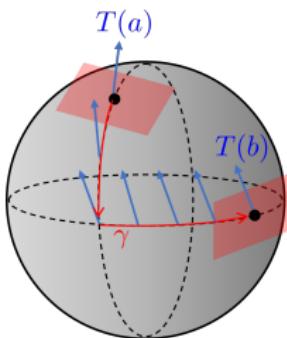
The Newton's method requires less iterations than gradient descent. However, it is prone to chaotic dynamic if the initial condition isn't chosen well enough.

Other methods of optimization requires more tools from geometry.  
For example, *retractions* can be used to replace geodesics in increasing calculation speed.

Other methods of optimization requires more tools from geometry.

For example, *retractions* can be used to replace geodesics in increasing calculation speed.

Or for one find an adaptive step size, Armijo condition and Wolfe condition can be used. The latter requires *parallel transport* for comparison of vectors from different tangent spaces.



Using geometry, we can generalize many previous ideas of optimization to more useful scenarios.

- [AMS08] P.-A. Absil, R. Mahony, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton: Princeton University Press, 2008. ISBN: 9781400830244. DOI: doi:10.1515/9781400830244. URL: <https://doi.org/10.1515/9781400830244>.
- [Tu17] Loring W. Tu. *Differential Geometry: Connections, Curvature, and Characteristic Classes*. Springer, 2017. ISBN: 9783319550824. DOI: doi:10.1007/9783319550848. URL: <https://link.springer.com/book/10.1007/978-3-319-55084-8>.
- [KSW24] Martin Kleinsteuber, Hao Shen, and Julian Wörmann. *Lecture Notes on Manifold Optimization for Representation Learning (MORL)*. July 2024.

Notes: