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.
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

$$\mathrm{O}(m) = \left\{ X \in \mathbb{R}^{m\times m} \,\middle|\, X^{\mathsf{T}}X = \mathbb{1} \right\}. \tag{1}$$

Obviously,

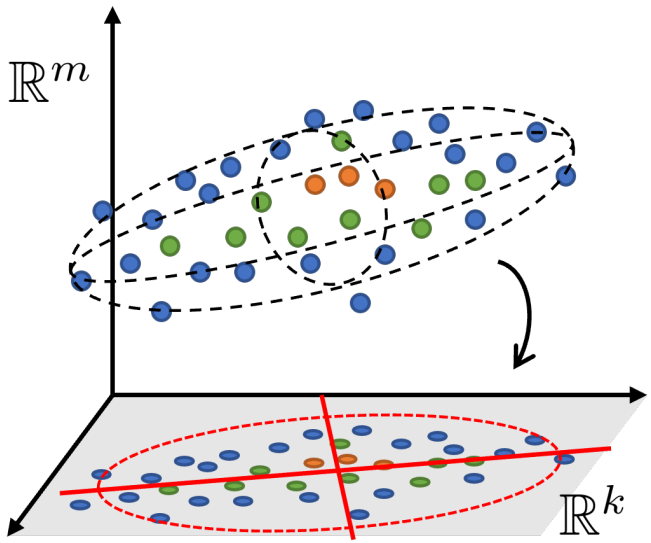
$$\arg\min_{X\in\mathbb{R}^{m\times m}} f(X) \text{ and } \arg\min_{X\in\mathrm{O}(m)} f(X)$$

will not coincide.

Or in **principal component analysis**:
given a data set
$X = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n] \in \mathbb{R}^{m\times n}$ $(m < n)$, we want to find a linear transformation

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

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



The corresponding optimization problem reads:

$$\max_{A} \operatorname{Tr}\left\{(A^{\mathsf{T}}X)(A^{\mathsf{T}}X)^{\mathsf{T}}\right\} \text{ s.t. } A^{\mathsf{T}}A = \mathbb{1}_k. \tag{2}$$

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.

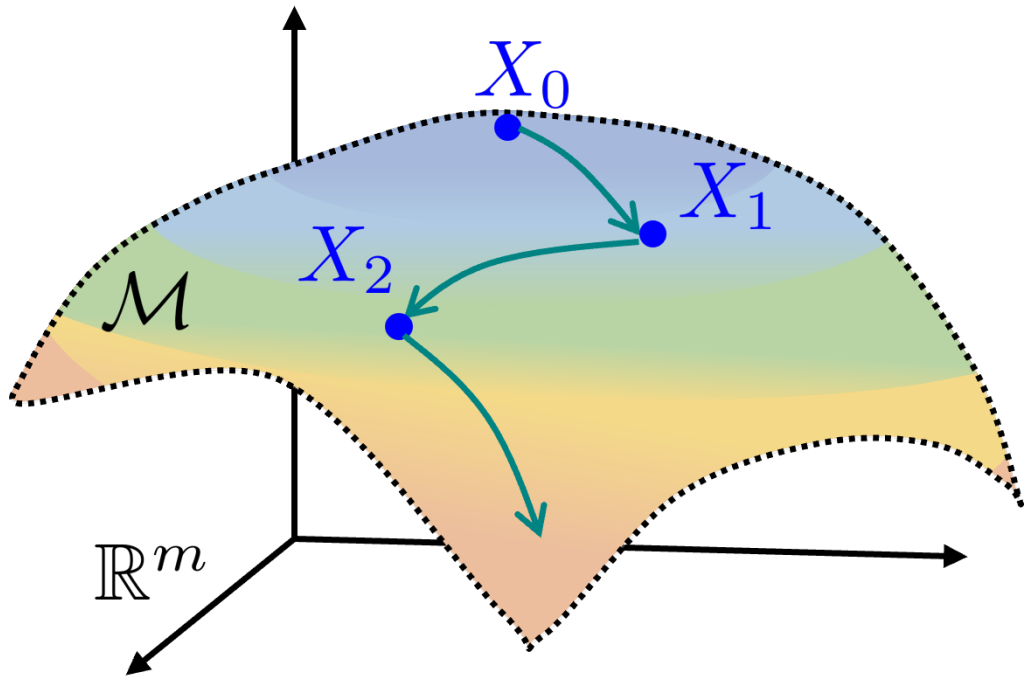


Table of Contents

Table of Contents

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:

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

Thus, we have the first and second derivatives as:

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

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

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

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

with

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

An example is as follows: consider $Q(\boldsymbol{x}) = \boldsymbol{x}^\mathsf{T} A \boldsymbol{x}$,

$$\frac{1}{4}\left(Q(\boldsymbol{x}+\boldsymbol{y}) - Q(\boldsymbol{x}-\boldsymbol{y})\right) = \boldsymbol{x}^\mathsf{T} A \boldsymbol{y} = \boldsymbol{y}^\mathsf{T} A \boldsymbol{x}.$$

Def. (Inner Product)

An inner product $\langle \cdot, \cdot \rangle$ on a vector space V over the field $\mathbb{R}$ is a map $V \times V \to \mathbb{R}$ 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 $\boldsymbol{x}$ and $\boldsymbol{y} \in \mathbb{R}^n$:

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^{\mathsf{T}} \boldsymbol{y} = \sum_{i=1}^{n} x_i y_i. \tag{8}$$

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

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

But other inner products also exist:

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

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

- given $X^{\mathsf{T}} X = \mathbb{1} = X X^{\mathsf{T}}$, for A, $B \in \left\{ X\Omega \,\middle|\, \Omega^{\mathsf{T}} = -\Omega \right\}$:

$$\langle A, B \rangle = \mathrm{Tr}\left\{ A^{\mathsf{T}} \left(\mathbb{1} - \frac{1}{2} X X^{\mathsf{T}} \right) B \right\}.$$

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

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

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 $\mathrm{H}_f(x)$, defined as

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

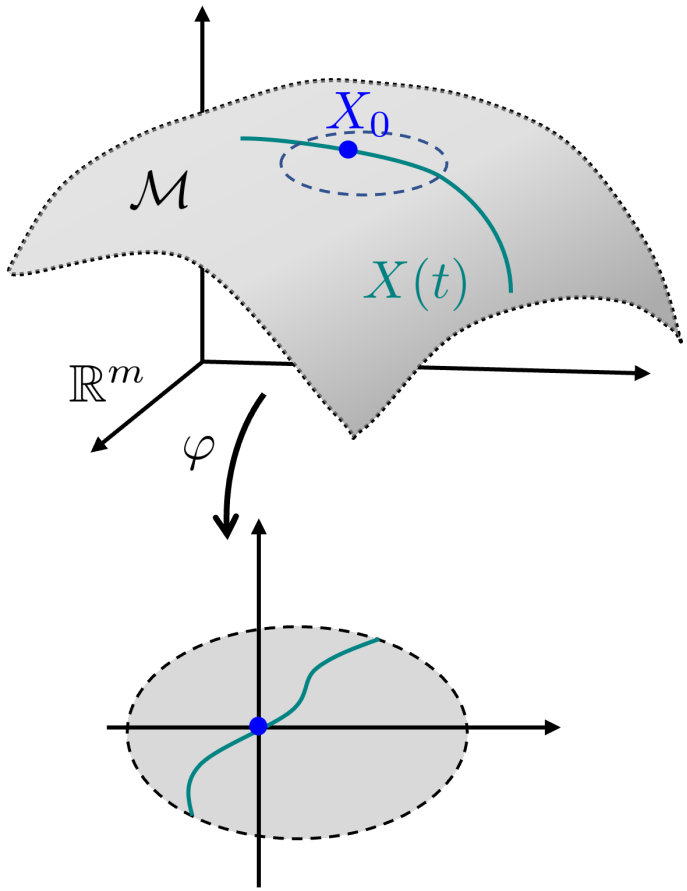
Table of Contents

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.

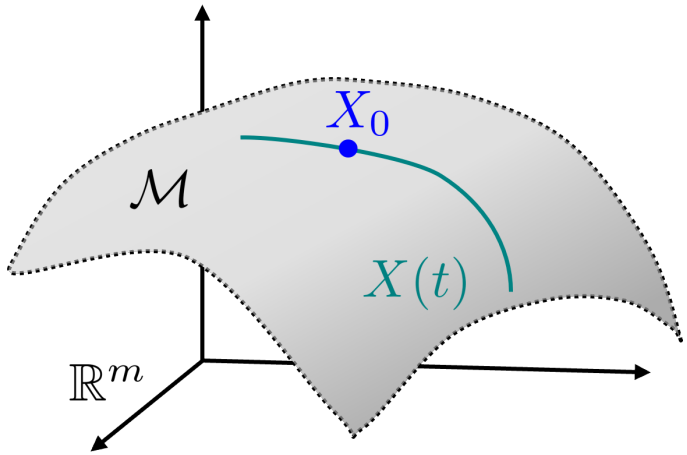
Jargons aside, we will be working with differentiable manifold $\mathcal{M}$, which are surfaces that can be parameterized locally using *charts*.



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

$$\partial \boldsymbol{f}(\boldsymbol{x})/\partial \boldsymbol{x} \text{ has full rank}$$

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

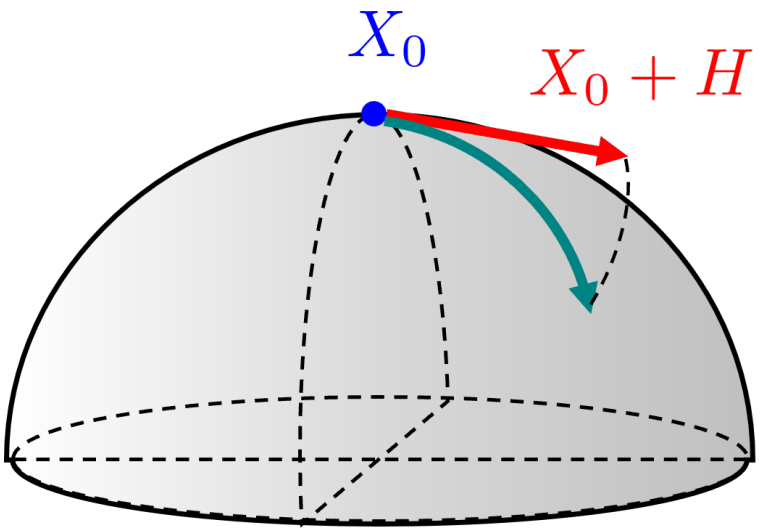
$$\mathcal{M} = \boldsymbol{f}^{-1}(\boldsymbol{c}) \subset \mathbb{R}^m$$

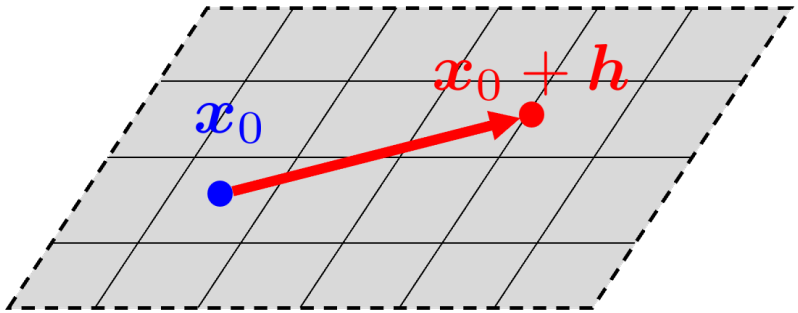
is a $(m-n)$-dimensional manifold.

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

e.g. $f(\boldsymbol{x}) = \boldsymbol{x}^\mathsf{T}\boldsymbol{x}$

e.g. $\mathrm{O}(m)$





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:

$$\begin{aligned}\mathrm{Sym}(m) &:= \left\{ X \in \mathbb{R}^{m\times m} \,\middle|\, X^{\mathsf{T}} = X \right\} \text{ and} \\ \mathrm{Skew}(m) &:= \left\{ X \in \mathbb{R}^{m\times m} \,\middle|\, X^{\mathsf{T}} = -X \right\}.\end{aligned}$$

   Their dimensions are:

$$\begin{aligned}\dim\left(\mathrm{Sym}(m)\right) &= \frac{m(m+1)}{2}, \\ \dim\left(\mathrm{Skew}(m)\right) &= \frac{m(m-1)}{2}.\end{aligned}$$

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

$$\langle A, B \rangle = \mathrm{Tr}\left\{A^{\mathsf{T}} B\right\} = 0. \tag{12}$$

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

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

2. **Orthogonal and Special Orthogonal Matrices**:
These matrices describe the m-dimensional transformation of *reflection* and *rotation*. They are no longer vector spaces.

$$\mathrm{O}(m) := \left\{ X \in \mathbb{R}^{m\times m} \,\middle|\, X^{\mathsf{T}}X = \mathbb{1} \right\} \text{ and}$$
$$\mathrm{SO}(m) := \left\{ X \in \mathbb{R}^{m\times m} \,\middle|\, X^{\mathsf{T}}X = \mathbb{1},\ \det(X) = 1 \right\}.$$

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

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

Hence, they have the same dimension:

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

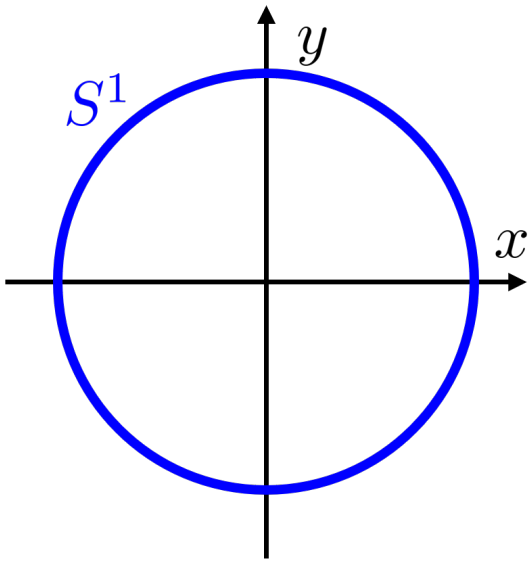
2. **Orthogonal and Special Orthogonal Matrices**:
   A special case would be the $\mathrm{SO}(2)$ and $\mathrm{SO}(3)$ matrices.

- $\mathrm{SO}(2)$: any 2D rotation can be parameterized by

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

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

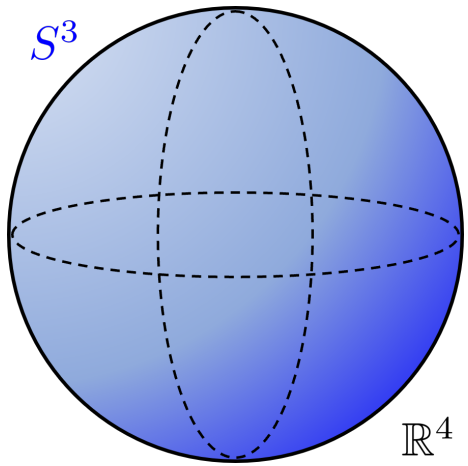
As a side note, $\mathrm{O}(2)$ will be two disconnected 1-spheres as a manifold.



2. **Orthogonal and Special Orthogonal Matrices**:
   A special case would be the $\mathrm{SO}(2)$ and $\mathrm{SO}(3)$ matrices.

- $\mathrm{SO}(3)$: 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, $\mathrm{SO}(3)$ is a 3-sphere S^3.

3. **Stiefel Manifold**:
   This is a generalization to orthogonal matrices.

$$\mathrm{St}(m,k) = \left\{ X \in \mathbb{R}^{m\times k} \,\middle|\, X^{\mathsf{T}}X = \mathbb{1}_k \right\} \qquad (m \geq k) \tag{16}$$

The dimension of the Stiefel manifold is

$$\dim\left(\mathrm{St}(m,k)\right) = mk - \frac{k(k+1)}{2}. \tag{17}$$

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.

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

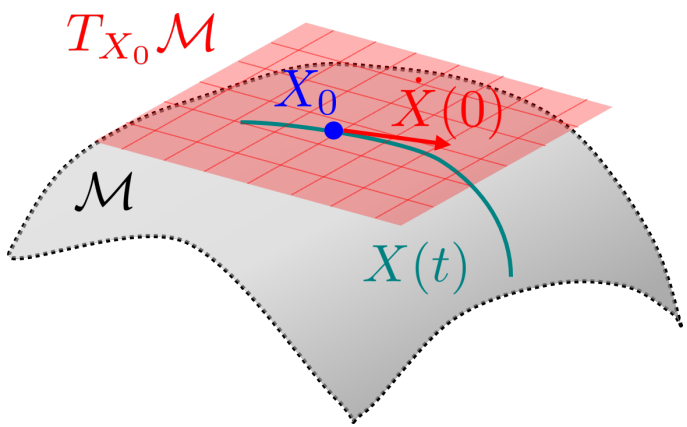
$$\mathcal{M}/\sim = \{\,[x] \,|\, x \in \mathcal{M}\,\}.$$

But we will not delve into them during this presentation.

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:

$$T_{X_0}\mathcal{M}. \tag{18}$$



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)**:
$$T_X\mathrm{Skew}(m) = \mathrm{Skew}(m). \tag{19}$$
2. **SO(m)**:
Consider a curve $X(t)$ subject to $X(t)^\mathsf{T}X(t) = \mathbb{1}$, then
$$\dot{X}(0)^\mathsf{T}X_0 + X_0^\mathsf{T}\dot{X}(0) = 0.$$
Hence,
$$\begin{aligned} T_X\mathrm{SO}(m) :&= \left\{ H \in \mathbb{R}^{m\times k} \,\middle|\, H^\mathsf{T}X + X^\mathsf{T}H = 0 \right\} \\ &= \{\, X\Omega \,|\, \Omega \in \mathrm{Skew}(k) \,\} \cong X \cdot \mathrm{Skew}(m). \end{aligned} \tag{20}$$

3. $\mathbf{St}(\boldsymbol{m}, \boldsymbol{k})$:
   Consider a curve $X(t)$ subject to $X(t)^\mathsf{T} X(t) = \mathbb{1}_k$, then

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

   Hence,

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

   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 $X^\mathsf{T} X^\perp = 0$.

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

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

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

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

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

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

Def. (Projection)

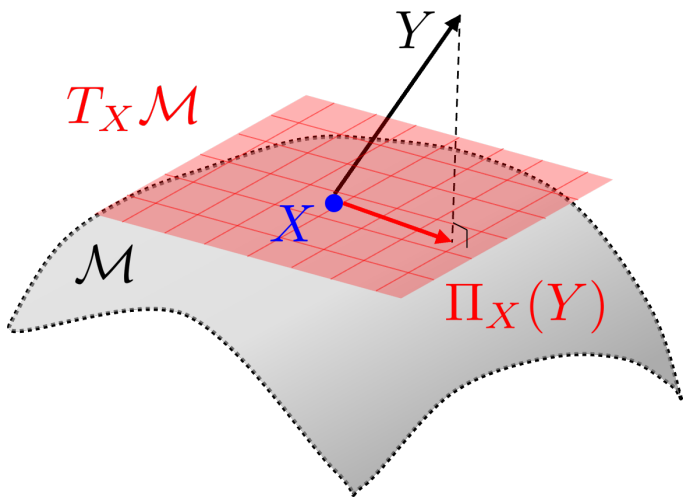
A projection Π on a vector space V is an idempotent map. I.e.,

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

Def. (Orthogonal Projection)

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

$$\langle \Pi(X), Y - \Pi(Y) \rangle = 0. \tag{24}$$



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. $\mathbf{Sym}(m)$ and $\mathbf{Skew}(m)$:
   Given a matrix $X \in \mathbb{R}^{m\times m}$, the respective orthogonal projections are

$$\Pi_{\mathrm{sym}}(X) = \mathrm{sym}(X) := \frac{X + X^{\mathsf{T}}}{2}, \tag{25}$$

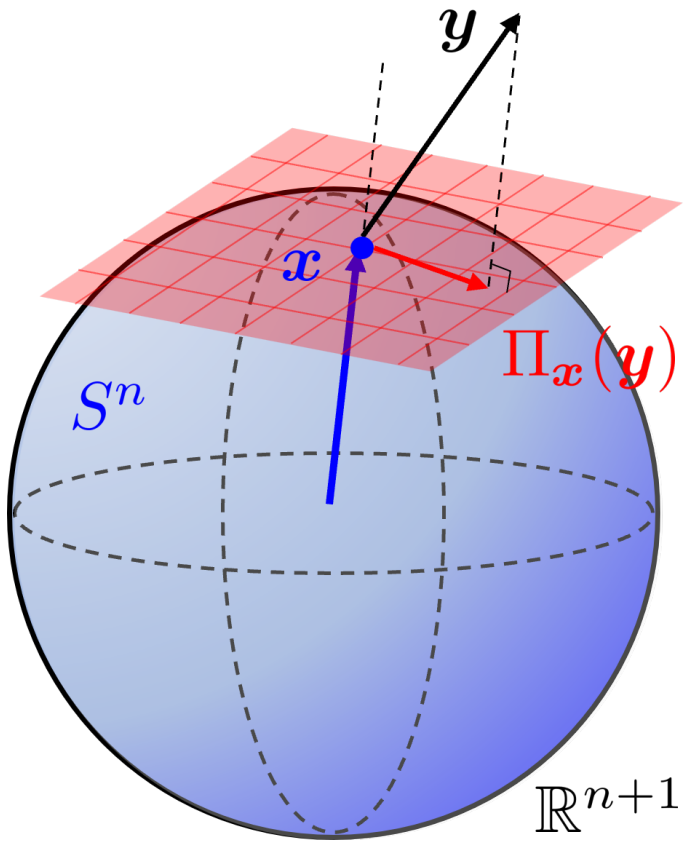
$$\Pi_{\mathrm{skew}}(X) = \mathrm{skew}(X) := \frac{X - X^{\mathsf{T}}}{2}. \tag{26}$$

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

2. $\boldsymbol{S^n}$:
   The manifold of S^n is the set of unit vectors in $\mathbb{R}^{n+1}$.
   Geometrically, we can obtain the orthogonal projection as

$$\Pi_{\boldsymbol{x}}(\boldsymbol{y}) = \boldsymbol{y} - \langle \boldsymbol{x}, \boldsymbol{y} \rangle \boldsymbol{x}, \qquad (27)$$

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



However, for more obscure manifolds, is there a way for us to *dervie* 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

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

where the **transposed** map is defined as

$$\left\langle \left(\frac{\partial f}{\partial X}\right)^{\mathsf{T}} [H], K \right\rangle = \left\langle H, \frac{\partial f}{\partial X}[K] \right\rangle. \tag{29}$$

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

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

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

$$(T_X\mathcal{M})^{\perp} = \operatorname{span}(\partial f/\partial X)^{\mathsf{T}}. \tag{30}$$

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

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

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

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

3. $\mathbf{St}(\boldsymbol{m}, \boldsymbol{k})$: since $f(X) = X^{\mathsf{T}}X - \mathbb{1}_k$,
   - For any $K \in \mathbb{R}^{m\times k}$, $\frac{\partial f}{\partial X}[K] = K^{\mathsf{T}}X + X^{\mathsf{T}}K \in \mathrm{Sym}(k)$.
   - For any $H \in \mathrm{Sym}(k)$ and $K \in \mathbb{R}^{m\times k}$,
     $\left\langle H, \frac{\partial f}{\partial X}[K] \right\rangle_{\mathrm{Sym}} = \mathrm{Tr}\left\{H^{\mathsf{T}}(K^{\mathsf{T}}X + X^{\mathsf{T}}K)\right\} = \langle 2XH, K\rangle_{\mathrm{St}}$,
     hence $\left(\frac{\partial f}{\partial X}\right)^{\mathsf{T}}[H] = 2XH$.
   - $\frac{\partial f}{\partial X} \circ \left(\frac{\partial f}{\partial X}\right)^{\mathsf{T}}[H] = \frac{\partial f}{\partial X}[2XH] = 4H$.
   - Finally,

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

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

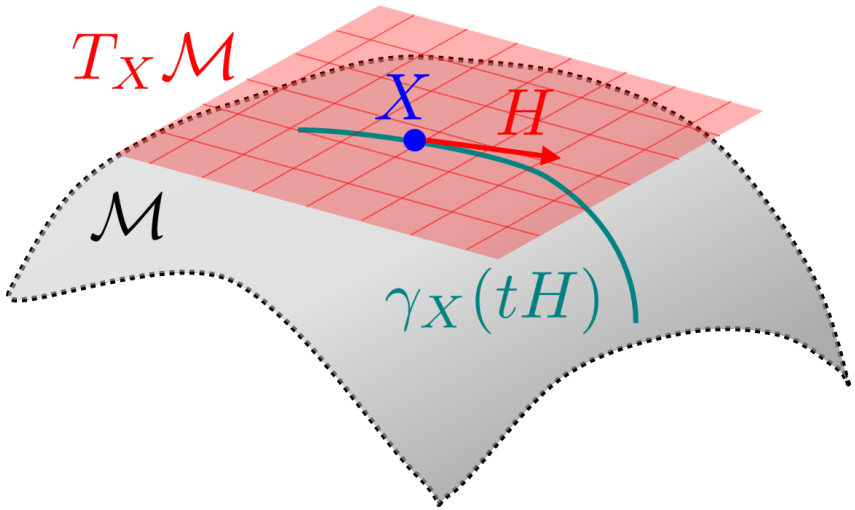
For $X \in \mathrm{St}(m,k)$, the orthogonal projection from $\mathbb{R}^{m\times k}$ to $T_X\mathrm{St}(m,k)$ is

$$\Pi_X(H) = H - X \cdot \mathrm{sym}\left(X^\mathsf{T} H\right). \tag{31}$$

Note that $\mathrm{SO}(m)$ 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

$$\Pi_X(H) = X \cdot \mathrm{skew}(X^\mathsf{T} H) \in T_X\mathrm{SO}(m). \tag{32}$$

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. $\boldsymbol{S^2}$:

   The geodesic equation of it will be

$$\ddot{\gamma}(t) + \gamma(t)\dot{\gamma}(t)^{\mathsf{T}}\dot{\gamma}(t) = 0,\ \gamma(0) = \boldsymbol{x},\ \dot{\gamma}(0) = \boldsymbol{h} \tag{33}$$

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

$$\gamma_{\boldsymbol{x}}(t\boldsymbol{h}) = \boldsymbol{x}\cos(t\,\|\boldsymbol{h}\|) + \frac{\boldsymbol{h}}{\|\boldsymbol{h}\|}\sin(t\,\|\boldsymbol{h}\|). \tag{34}$$

2. $\mathbf{SO}(m)$:
   The geodesic equation of it will be

$$\ddot{\Gamma}(t) + \Gamma(t)\dot{\Gamma}(t)^{\mathsf{T}}\dot{\Gamma}(t) \tag{35}$$

under the usual Frobenius inner product, with initial conditions

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

The solution will be:

$$\Gamma_X(tH) = X \exp(tX^{\mathsf{T}}H) = \exp(tHX^{\mathsf{T}})X. \tag{37}$$

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. $\mathbf{St}(\boldsymbol{m}, \boldsymbol{k})$:
   The geodesic equation of it will be

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

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

$$\Gamma_X(tH) = \exp\left(t(HX^{\mathsf{T}} - XH^{\mathsf{T}})\right) \cdot X \cdot \exp(-tX^{\mathsf{T}}H). \tag{39}$$

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

Table of Contents

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

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

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 $f : \mathcal{M} \to \mathcal{N}$ has a fancy name termed the **pushforward**[1] besides our usual name of **differential map**.

Def. (Pushforward)

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

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

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

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

[1] More often, it is denoted as $\mathrm{D}f(X)$ 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.

$$X = QR. \tag{42}$$

[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.

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

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

$$\to \frac{\partial Q}{\partial X}[H] = Q\cdot\widetilde{\mathrm{skew}}(Q^\mathsf{T}HR^{-1}),$$

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

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 $\boldsymbol{u} \in V$ such that for all $\boldsymbol{v} \in V$,

$$f(\boldsymbol{v}) = \langle \boldsymbol{u}, \boldsymbol{v} \rangle. \tag{43}$$

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 $\mathrm{grad} f(X) \in T_X\mathcal{M}$, defined as

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

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

$$f : \mathrm{St}(m,k) \to \mathbb{R}, \quad f(X) := \left\| X^\mathsf{T} Y - Z \right\|^2 \tag{45}$$

with respect to the Frobenius inner product.

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

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

$\mathrm{grad} f(X) = \Pi_X\left(2Y(X^\mathsf{T}Y - Z)^\mathsf{T}\right) = 2Y(X^\mathsf{T}Y - Z)^\mathsf{T} - X \cdot \mathrm{sym}\left(X^\mathsf{T} 2Y(X^\mathsf{T}Y - Z)^\mathsf{T}\right)$ ■

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

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

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

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

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 $\mathrm{hess} f(X) : T_X\mathcal{M} \to T_X\mathcal{M}$, defined as

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

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

$$f : S^{n-1} \to \mathbb{R}, \quad \boldsymbol{x} \mapsto \frac{1}{2}\boldsymbol{x}^{\mathsf{T}} A \boldsymbol{x}.$$

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

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

$$\operatorname{grad} f(\boldsymbol{x}) = (\mathbb{1}_n - \boldsymbol{x}\boldsymbol{x}^{\mathsf{T}})A\boldsymbol{x}$$

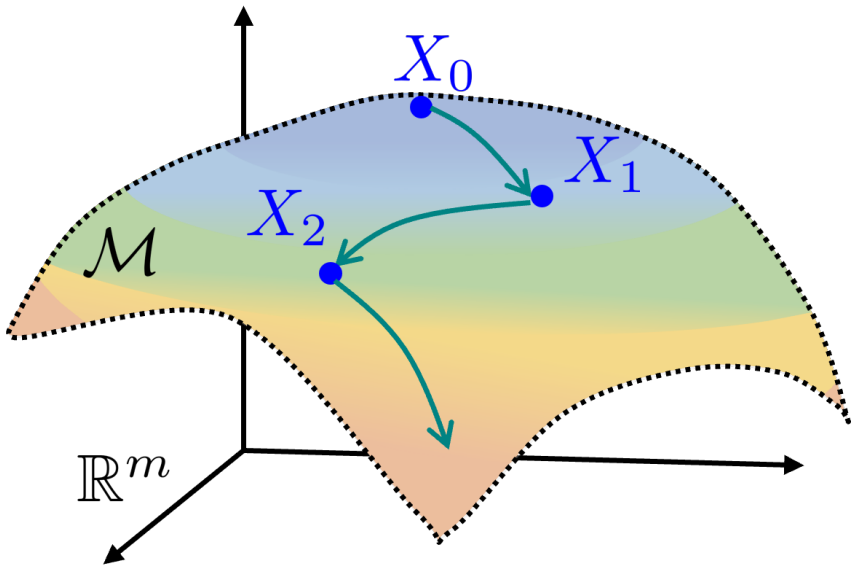
$$\begin{aligned}
\frac{\partial^2 f}{\partial \boldsymbol{x}^2}[\boldsymbol{h}] &= -(\boldsymbol{h}^{\mathsf{T}}\boldsymbol{h})\boldsymbol{x}^{\mathsf{T}} A \boldsymbol{x} \; + \boldsymbol{h}^{\mathsf{T}} A \boldsymbol{h} \\
\frac{\partial^2 f}{\partial \boldsymbol{x}}[\boldsymbol{h}, \boldsymbol{k}] &= \boldsymbol{k}^{\mathsf{T}} A \boldsymbol{h} - (\boldsymbol{k}^{\mathsf{T}}\boldsymbol{h})\boldsymbol{x}^{\mathsf{T}} A \boldsymbol{x} \\
\operatorname{hess} f(\boldsymbol{x})[\boldsymbol{h}] &= (\mathbb{1}_n - \boldsymbol{x}\boldsymbol{x}^{\mathsf{T}})(A - \boldsymbol{x}^{\mathsf{T}} A \boldsymbol{x}\mathbb{1}_n)\boldsymbol{h} \qquad \blacksquare
\end{aligned}$$

Table of Contents

Now that we have a manifold $\mathcal{M}$ and a scalar function $f : \mathcal{M} \to \mathbb{R}$ 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[2]:

Algorithm 1: Gradient Descent on Manifold

Input: Function $f(X)$ and initial condition X_0.
Output: Optimal X that minimizes f locally.

Set acceptable error;
Initialize $X = X_0$;
while $\|X_{new} - X_{old}\| > error$ **do**
    Calculate gradient $\mathrm{grad} f(X)$;
    Set step size as α;
    Update $X \mapsto \gamma_X\left(-\alpha \cdot \mathrm{grad} f(X)\right)$.
end

[2]The actual code implementation requires some extra care.

Let us consider minimizing the cost function

$$f(X) = \mathrm{Tr}\left\{X^{\mathsf{T}} A X B\right\}, \tag{49}$$

where $X \in \mathrm{SO}(n)$, A and B are non-degenerate[3] and symmetric.
[sol.] The gradient is calculated to be

$$\frac{\partial f}{\partial X}[H] = \mathrm{Tr}\left\{H^{\mathsf{T}} A X B\right\} \to \mathrm{grad} f(X) = X \cdot \mathrm{skew}(X^{\mathsf{T}} A X B)$$

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

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

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

[3] 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 $B = \mathrm{diag}(\mu_1, \mu_2, \cdots \mu_n)$ with $\mu_i < \mu_j$ for $i < j$. At extrema, we also have

$$\mathcal{A} := X^{\mathsf{T}} \mathcal{A} X = \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n).$$

The Hessian can be calculated to be

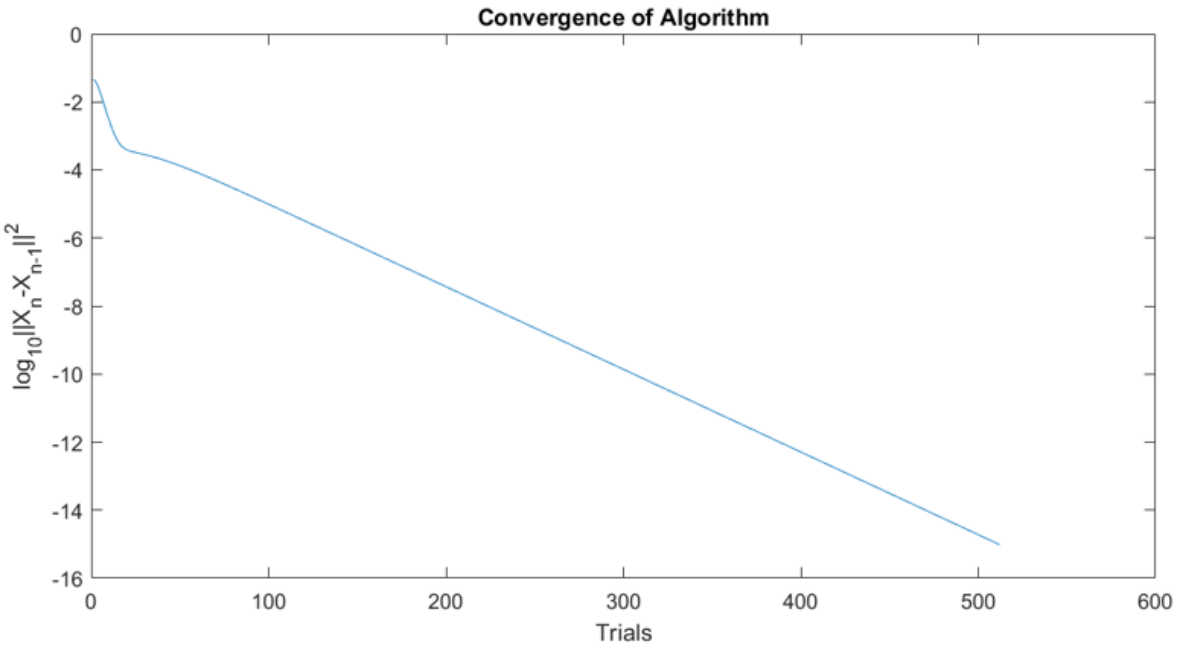
$$\mathrm{hess} f(X)[X\Omega] = \mathrm{skew}\left(\mathcal{A}\Omega B - \mathrm{sym}(\mathcal{A}B)\Omega\right). \tag{50}$$

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

$$X \mapsto X \cdot \exp\left(-\alpha \cdot \operatorname{skew}(X^{\mathsf{T}} A X B)\right),$$

where α 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:

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

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:

$$\boldsymbol{x} \mapsto \boldsymbol{x} - (\mathrm{hess} f(\boldsymbol{x}))^{-1} \mathrm{grad} f(\boldsymbol{x}). \tag{52}$$

Algorithm 2: Newton's Method on Manifold

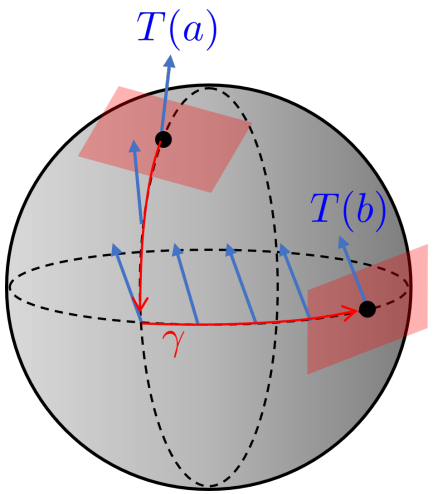
Input: Function $f(X)$ and initial condition X_0.

Output: Optimal X that extremizes f locally.

Set acceptable error;
Initialize $X = X_0$;
while $\|X_{new} - X_{old}\| > error$ **do**
    Calculate Newton's step $\Omega = -\left(\mathrm{hess} f(X)\right)^{-1} \mathrm{grad} f(X)$;
    Update $X \mapsto \gamma_X(\Omega)$.
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.
Or for one find an adpative 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: