# Introduction to Riemannian Optimization

Taskynov Anuar

Moscow State University

*taskynov.anuar@mail.ru*

April 5, 2019

# Overview

1. Manifolds
   - Definition
   - Tangent space
   - Riemannian Manifold
2. Riemannian Optimization
   - Riemannian gradient
   - Geodesic, exponential map
   - Retraction
   - Vector transport
3. Conclusion

# Manifold

## Definition (manifold)

**Manifold** $\mathcal{M}$ is a set which looks like Euclidean space around every point.

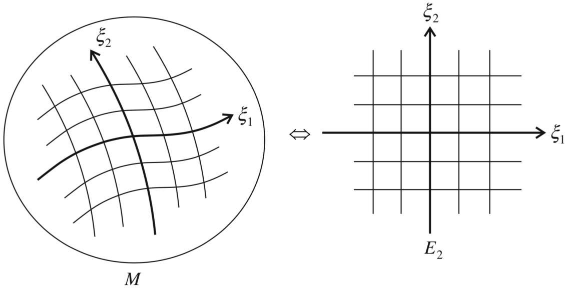


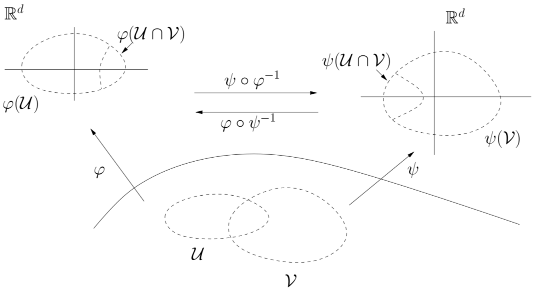

Figure: Manifold $\mathcal{M}$ and coordinate system $\xi$. $E_2$ is a two-dimensional Euclidean space

# Manifold

## Definition (manifold)

Formally $\mathcal{M}$ is a $d$-**dimensional manifold** if

- $\forall x \in \mathcal{M}$ , $\exists$ bijective function $\phi$: $\mathcal{U} \to \mathbb{R}^d$, where $\mathcal{U}$ — neighborhood at the point $x \in \mathcal{M}$;
- for neighborhoods $\mathcal{U}$ and $\mathcal{V}$ ($\mathcal{U} \cap \mathcal{V} \neq \emptyset$) the change of coordinates is smooth: $\phi \circ \psi^{-1}, \psi \circ \phi^{-1} \in C^\infty(\mathbb{R}^d)$;



# Manifold

- $\phi(x) \in \mathbb{R}^d$ is called the local (intrinsic) coordinates of point $x$.
- If $\mathcal{M} \subset \mathbb{R}^n$, then the point $x$ has global (extrinsic) coordinates $(\in \mathbb{R}^n)$.

Example:

- Circle, $S^1 = \{x = (x_1, x_2) \in \mathbb{R}^2 : x_1^2 + x_2^2 = 1\}$.
  $x \in \mathbb{R}^2$ — extrinsic coordinates.
  $t \in [0, 2\pi) \in \mathbb{R}$ — intrinsic coordinates.
  Mapping between coordinates: $\phi^{-1}(t) = (\cos t, \sin t)$.

# Manifold

Examples (subsets of finite Euclidean space):

- $\mathbb{R}^d$.
- (Real projective) $\mathbb{RP}^{n-1}$ is the set of all directions in $\mathbb{R}^n$.
- (Grassman) $\mathrm{Grass}(p, n)$ is the set, which parametrizes all $p$-dimensional linear subspaces of the $n$-dimensional vector space $\mathbb{R}^n$.
- (Stiefel) $\mathrm{St}(p, n) = \{X \in \mathbb{R}^{n\times p} : X^T X = I_p\}$.
- $r$-rank matrices $\mathcal{M}_r = \{X \in \mathbb{R}^{m\times p} : \mathrm{rank}(X) = r\}$.

Example (probability distribution):

- Gaussian distributions $p(x|\mu, \sigma^2) = \mathcal{N}(x|\mu, \sigma^2)$, $\phi(p) = (\mu, \sigma)$.

# Tangent space

**Definition (tangent space for $\mathcal{M} \subset \mathbb{R}^n$)**

Let $\gamma : (-a, a) \to \mathcal{M}$ is a smooth curve on a manifold, such that $\gamma(0) = x$. **The tangent space** at $x \in \mathcal{M}$, noted $T_x\mathcal{M}$, is the linear subspace ($\dim(T_x\mathcal{M}) = \dim(\mathcal{M})$) of $\mathbb{R}^n$ defined by:

$$T_x\mathcal{M} = \left\{ \xi \in \mathbb{R}^n : \xi = \gamma'(0) \right\}.$$

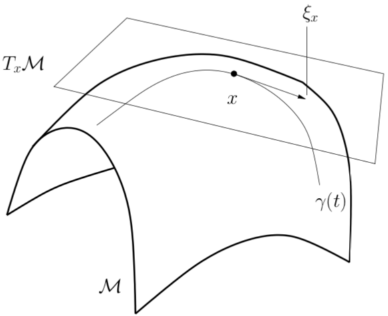



Figure: Tangent space $T_x\mathcal{M}$.

# Tangent space to a sphere $S^{n-1}$

Consider a sphere manifold $S^{n-1} = \left\{ x \in \mathbb{R}^n : x_1^2 + x_2^2 + \dots + x_n^2 = 1 \right\}$.
What is $T_{x_0} S^{n-1}$?
Let $x(t)$ is a curve on a sphere, $x(0) = x_0$. Since $x(t) \in S^{n-1}$ for all $t$, we have:

$$x(t)^T x(t) = 1.$$

Differentiating this equation with respect to $t = 0$:

$$x'(0)^T x_0 + x_0^T x'(0) = 0.$$

So we have:

$$T_{x_0} S^{n-1} = \left\{ z \in \mathbb{R}^n : z^T x_0 = 0 \right\}.$$

# Basis of the tangent space

Let $\phi : \mathcal{U} \to \mathbb{R}^d$, where $\mathcal{U}$ is a neighborhood at the point $x \in \mathcal{M} \subset \mathbb{R}^n$; $\hat{x} = \phi(x)$ — local coordinates.
Basis vectors $E_i$ defined as:

$$E_i = \lim_{\tau \to 0} \frac{\phi^{-1}(\hat{x} + \tau e_i) - \phi^{-1}(\hat{x})}{\tau} = \frac{\partial \phi^{-1}(\hat{x})}{\partial \hat{x}_i},$$

where $\frac{\partial \phi^{-1}(\hat{x})}{\partial \hat{x}_i} \in \mathbb{R}^n$ is a $i^{th}$ column of Jacobi matrix for $\phi^{-1}$ at the point $\hat{x}$.
Example for sphere, $S^2$:

$$\begin{pmatrix} \hat{x}_1 \\ \hat{x}_2 \end{pmatrix} \xrightarrow{\phi^{-1}} \begin{pmatrix} \cos \hat{x}_1 \\ \sin \hat{x}_1 \cos \hat{x}_2 \\ \sin \hat{x}_1 \sin \hat{x}_2 \end{pmatrix}$$

Jacobi matrix:

$$\frac{\partial \phi^{-1}}{\partial \hat{x}} = \begin{pmatrix} -\sin \hat{x}_1 & 0 \\ \cos \hat{x}_1 \cos \hat{x}_2 & -\sin \hat{x}_1 \sin \hat{x}_2 \\ \cos \hat{x}_1 \sin \hat{x}_2 & \sin \hat{x}_1 \cos \hat{x}_2 \end{pmatrix}$$

# Tangent Bundle

## Definition (tangent bundle)

**The tangent bundle**, noted $T\mathcal{M}$, is the set

$$T\mathcal{M} = \cup_{x \in \mathcal{M}} \left\{ (x, \xi) : \xi \in T_x\mathcal{M} \right\}$$

Example (circle):

$$TS^1 = \cup_{x \in S^1} \left\{ (x, \xi) : \xi \in T_x S^1 \simeq \mathbb{R} \right\} \simeq S^1 \times \mathbb{R}.$$

# Vector field on a manifold

**Definition (vector field on $\mathcal{M}$)**

A vector field $\mathbf{X} : \mathcal{M} \to T\mathcal{M}$,
$\mathbf{X}(x) = \xi$, $\xi \in T_x\mathcal{M}$.

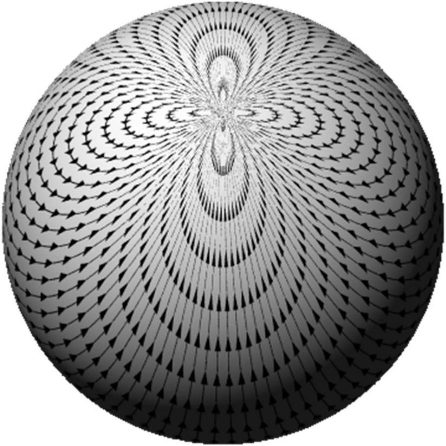


Figure: Vector field on a sphere.

# Riemannian manifold

## Definition (riemannian manifold)

A manifold whose tangent spaces are endowed with a smoothly varying inner product $g_x(\cdot,\cdot) = \langle \cdot, \cdot \rangle_x$ is called a **Riemannian manifold**.

Smoothly varying can be understood in the following sense: for all vector fields $\mathbf{X}, \mathbf{Y} \in \mathcal{X}(\mathcal{M})$, the function $x \to g_x(\mathbf{X}_x, \mathbf{Y}_x)$ is a smooth function from $\mathcal{M}$ to $\mathbb{R}$.

Inner product can be represented as:

$$g_x(\xi_x, \eta_x) \;=\; \left\{\xi_x \;=\; \sum_{i=1}^{d} (\hat{\xi}_x)_i E_i, \eta_x \;=\; \sum_{i=1}^{d} (\hat{\eta}_x)_i E_i\right\} \;=\; \hat{\xi}_x^T G_x \hat{\eta}_x, \quad (1)$$

where $G_x = \{\langle E_i, E_j \rangle\}_{i,j=1}^{d} \in \mathbb{R}^{d \times d}$ — symmetric, positive definite matrix, $\hat{\xi}_x, \hat{\eta}_x, \in \mathbb{R}^d$ is coordinate representation of tangent vectors.

# Riemannian gradient

Let $f : \mathbb{R}^n \to \mathbb{R}$ — differetiable function, $x, \xi \in \mathbb{R}^n$. Usual directional derivative:

$$Df(x)[\xi] = \lim_{\tau \to 0} \frac{f(x + \tau\xi) - f(x)}{\tau}.$$

If $f : \mathcal{M} \to \mathbb{R}$, $x \in \mathcal{M}$, $\xi \in T_x\mathcal{M}$:

$$Df(x)[\xi] = \left.\frac{df(\gamma(t))}{dt}\right|_{t=0},$$

where $\gamma$ is a differentiable curve on $\mathcal{M}$ satisfies $\gamma(0) = x$, $\gamma'(0) = \xi$.

**Definition (riemannian gradient)**

Given a smooth scalar field $f : \mathcal{M} \to \mathbb{R}$ on a Riemannian manifold, **the gradient** of $f$ at $x$, denoted by $\mathrm{grad} f(x)$, is defined as the unique element of $\mathcal{T}_x\mathcal{M}$ that satisfies:

$$\langle \mathrm{grad} f(x), \xi \rangle_x = Df(x)[\xi], \forall \xi \in T_x\mathcal{M}.$$

# Riemannian gradient

Let $f : \mathcal{M} \to \mathbb{R}$ $\text{grad} f : \mathcal{M} \to T\mathcal{M}$ is a vector field on $\mathcal{M}$.

$$\frac{\text{grad} f(x)}{\|\text{grad} f(x)\|_x} = \underset{\xi \in T_x\mathcal{M}, \|\xi\|_x = 1}{\arg\max} Df(x)[\xi].$$

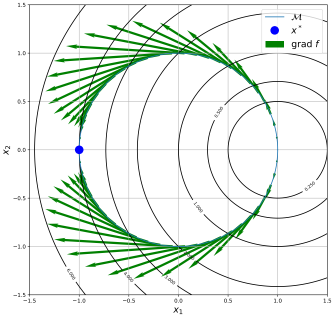



Figure: Function $f(x) = (x_1 - 1)^2 + x_2^2$ on a circle $S^1$.

# Riemannian Gradient

Coordinate expression: $\hat{\text{grad}} f(x) = G_x^{-1} \text{Grad} \hat{f}(\hat{x})$, where $\text{Grad} \hat{f}(\hat{x})$ is a vector of partial derivatives: $\text{Grad} \hat{f}(\hat{x}) = \left( \frac{\partial \hat{f}}{\partial \hat{x}_1} \dots \frac{\partial \hat{f}}{\partial \hat{x}_d} \right)^T$.

We need to inverse $G_x$.

Example:

Manifold: $S^2 = \left\{ x \in \mathbb{R}^2 : x_1^2 + x_2^2 + x_3^2 = 1 \right\}$.

Function $\phi^{-1}$: $\phi^{-1}(\hat{x}) = (\cos \hat{x}_1, \sin \hat{x}_1 \cos \hat{x}_2, \sin \hat{x}_1 \sin \hat{x}_2) = x$.

Let $f(x) : \mathbb{R}^3 \to \mathbb{R}$, $f(x) = (x_1 - 1)^2 + x_2^2 + x_3^2$, so $\hat{f}(\hat{x}) = f(\phi^{-1}(\hat{x}))$.

$\text{Grad} \hat{f}(\hat{x}) = \left( \frac{\partial \phi^{-1}}{\partial \hat{x}} \right)^T \frac{\partial f}{\partial \phi^{-1}} = \left( \frac{\partial \phi^{-1}}{\partial \hat{x}} \right)^T \nabla_x f(x)$.

Matrix $G_x = \left( \frac{\partial \phi^{-1}(\hat{x})}{\partial \hat{x}} \right)^T \frac{\partial \phi^{-1}(\hat{x})}{\partial \hat{x}} = \begin{pmatrix} \sin \hat{x}_1^2 & 0 \\ 0 & 1 \end{pmatrix}$.

Riemannian gradient: $\hat{\text{grad}} f(x) = G_x^{-1} \text{Grad} \hat{f}(\hat{x})$.

# Riemannian Gradient

Riemannian gradient for $f : \mathcal{M} \to \mathbb{R}$, where $\mathcal{M} \subset \mathbb{R}^n$:

$$\mathrm{grad} f(x) = \mathrm{Proj}_{T_x\mathcal{M}} \nabla \overline{f}(x),$$

where $\overline{f} : \mathbb{R}^n \to \mathbb{R}$ such that $f$ is a restriction of $\overline{f}$.
Only calculate projection to the tangent space.

Example:
Manifold: $S^2 = \left\{x \in \mathbb{R}^2 : x_1^2 + x_2^2 + x_3^2 = 1\right\}$.
Tangent space: $T_x S^2 = \left\{z \in \mathbb{R}^3 : z^T x = 0\right\}$.
Projection to $T_x S^2$: $\mathrm{Proj}_{T_x S^2}(y) = (I - xx^T)y$
Let $\overline{f}(x) : \mathbb{R}^3 \to \mathbb{R}$, $\overline{f}(x) = (x_1 - 1)^2 + x_2^2 + x_3^2$, $f : S^2 \to \mathbb{R}$ is a restriction of $\overline{f}$.
Riemannian gradient: $\mathrm{grad} f(x) = (I - xx^T)\nabla_x \overline{f}$

# Riemannian optimization

- Optimization problem: $f(x) \to \min_{x \in \mathcal{M}}$, where $\mathcal{M}$ is a riemannian manifold.
- How can you optimize this function?
- Usual gradient descent step: $x_{k+1} = x_k - \alpha_k \nabla f(x_k)$.
- For manifolds:
  - if you have intrinsic parametrization: $\hat{x}_{k+1} = \hat{x}_k - \alpha_k G_{x_k}^{-1} \text{Grad}\hat{f}(\hat{x})$.
  - if you have extrinsic parametrization: $x_{k+1} = x_k - \alpha_k \text{grad} f(x)$.

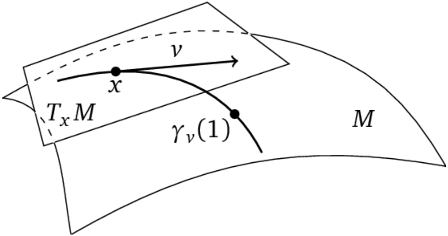



Figure: Mapping from tangent space $T_x\mathcal{M}$ to $\mathcal{M}$.

# Geodesic

- The length of a curve $\gamma : [0,1] \to \mathcal{M}$:

$$L(\gamma) = \int_0^1 \langle \gamma'(t), \gamma'(t) \rangle_{\gamma(t)} dt.$$

- Distance between two points $x, y \in \mathcal{M}$ on a Riemannian manifold:

$$\text{dist}(x,y) := \inf_{\gamma:\gamma(0)=x,\gamma(1)=y} \mathrm{L}(\gamma)$$

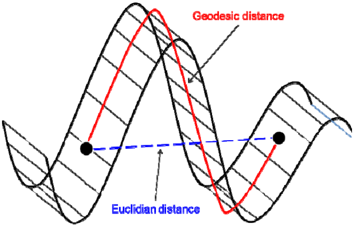



Figure: Geodesic distance

**Definition (geodesic)**

**Geodesic** is a curve $\gamma$ with minimal distance.

# Exponential map

**Definition (exponential map)**

Let $\mathcal{M} \subset \mathbb{R}^n$ — riemannian manifold and $x \in \mathcal{M}$. For every $\xi \in T_x\mathcal{M}$, there exists an open interval $(-a, a)$ and a unique geodesic $\gamma(t; x, \xi) : (-a, a) \to \mathcal{M}$ such that $\gamma(0) = x$ and $\gamma'(0) = \xi$. The mapping

$$\mathrm{Exp}_x : T_x\mathcal{M} \to \mathcal{M} : \xi \to \mathrm{Exp}_x(\xi) = \gamma(1; x, \xi)$$

is called **exponential map** at $x$. In particular, $\gamma(0; x, \xi) = x, \forall x \in \mathcal{M}$.

Optimization step: $x_{k+1} = \mathrm{Exp}_{x_k}(-\alpha_k \mathrm{grad} f(x_k))$.
Hard to compute! Because you should solve DE.

# Retraction

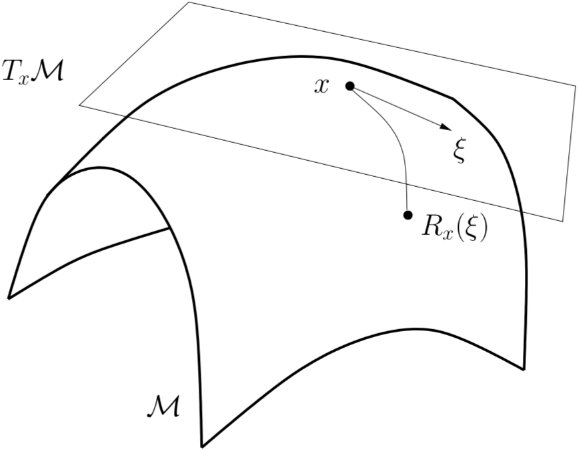



Figure: Retraction is an approximation of exponential map.

# Retraction

Exponential maps can be expensive to compute.

## Definition (Retraction)

A **retraction** on a manifold $\mathcal{M}$ is a smooth mapping $R : \mathcal{TM} \to \mathcal{M}$, $\mathcal{M} \subset \mathbb{R}^n$ with the following properties. Let $R_x$ denote the restriction of $R$ to $\mathcal{T}_x\mathcal{M}$.

- $R_x(0) = x$, $0 \in \mathcal{T}_x\mathcal{M}$.
- $\left.\frac{dR_x(t\xi)}{dt}\right|_{t=0} = \xi$, $\forall \xi \in T_x\mathcal{M}$.

We may do the following step: $x_{k+1} = R_{x_k}(\alpha_k \eta_k)$, $\eta_k \in T_{x_k}\mathcal{M}$. If $\eta_k = -\text{grad} f(x_k)$, then this step is one GD step.

# Retraction on $S^{n-1}$

Retraction $R_x(\xi) = \frac{x+\xi}{\|x+\xi\|}$, $\xi \in T_x S^{n-1}$, $x \in S^{n-1}$.
Let's check it!

- $\left(\frac{x+\xi}{\|x+\xi\|}\right)^T \frac{x+\xi}{\|x+\xi\|} = 1$, so $R_x(\xi) \in S^{n-1}$, $\forall (x, \xi) \in TS^{n-1}$.
- $R_x(0) = \frac{x}{\|x\|} = x$, because $x^T x = 1$.
- $$\frac{dR_x(t\xi)}{dt}\Big|_{t=0} = \left(\frac{x + t\xi}{\|x + t\xi\|}\right)'\Big|_{t=0} = \frac{\xi}{\|x\|} - \frac{x^T \xi x}{\|x\|^3} = \xi. \qquad (2)$$

# Retractions on $\mathrm{St}(p, n)$

Retractions on $\mathrm{St}(p, n)$:

- $R_X(\xi) = \mathrm{Proj}_{\mathrm{St}(p,n)}(X + \xi)$.
- $R_X(\xi) = (X + \xi)(I_p + \xi^T \xi)^{-1/2}$.
- $R_X(\xi) = \mathrm{QR}(X + \xi)$, where $\mathrm{QR}(\cdot)$ — return orthogonal matrix from QR decomposition.
- $R_X(\xi) = \mathrm{Cayley}\Big(-1/2(\xi X^T - X\xi^T)\Big)X$, where $\mathrm{Cayley}(A) = (I + A)^{-1}(I - A)$, $A \in \mathbb{R}^{n \times n}$ — skew-symmetric matrix.

# Gradient Descent with Momentum

Optimization step on $\mathbb{R}^n$:

$$\begin{cases} d_k = \beta d_{k-1} + \alpha_k \nabla f(x_k); \\ x_{k+1} = x_k - d_k. \end{cases}$$

Back to the manifold $\mathcal{M}$:

$$\begin{cases} d_k = \underbrace{\beta d_{k-1}}_{\in T_{x_{k-1}}\mathcal{M}} + \underbrace{\alpha_k \mathrm{grad} f(x_k)}_{\in T_{x_k}\mathcal{M}}; \\ x_{k+1} = R_{x_k}(-d_k); \end{cases}$$

# Vector Transport

## Definition (vector transport)

A **vector transport** on a manifold M is a smooth mapping:

$\mathrm{Transp} : \mathcal{TM} \times \mathcal{TM} \to \mathcal{TM}$,

satisfying the following properties for all $x \in \mathcal{M}$:

- $\exists R_x$, called the retraction associated with Transp: $\mathrm{Transp}_\eta(\xi) \in \mathcal{T}_{R_x(\eta)}\mathcal{M}$.
- $Transp_0(\xi) = \xi, \forall \xi \in \mathcal{T}_x\mathcal{M}$.
- $\mathrm{Transp}_\eta(a\xi + b\zeta) = a\mathrm{Transp}_\eta(\xi) + b\mathrm{Transp}_\eta(\zeta)$.

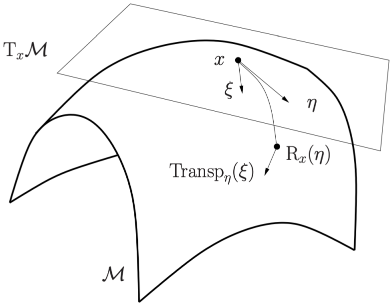



Figure: Vector transport

# Vector Transport

Typical vector transport:

$$\mathrm{Transp}_{\eta}(\xi) = \frac{d}{dt} R_x(\eta + t\xi)\Big|_{t=0}.$$

Example (sphere $S^{n-1}$):

- Retraction: $R_x(\xi) = \frac{x+\xi}{\|x+\xi\|}$.
- Vector transport:

$$\mathrm{Transp}_{\eta}(\xi) = \frac{1}{x+\eta}\Big(I - \frac{1}{\|x+\eta\|^2}(x+\eta)(x+\eta)^T\Big)\xi.$$

Another notation of vector transport: $\mathrm{Transp}_{T_{x \to R_x(\eta)}\mathcal{M}}(\xi)$

# Riemannian GD with Momentum

Optimization step:

$$\begin{cases} d_k = \mathrm{Transp}_{T_{x_{k-1} \to x_k}\mathcal{M}}(\beta d_{k-1}) + \alpha_k \mathrm{grad} f(x_k); \\ x_{k+1} = R_{x\,k}(-d_k) \end{cases}$$

# Conclusion

What do you need to optimize $f(x)$ on riemannian manifold $\mathcal{M}$?

- if you have intrinsic parametrization:
    - $\hat{\text{grad}} f(x) = G_{\hat{x}}^{-1} \nabla_{\hat{x}} f(\hat{x})$.
- if you have extrinsic parametrization:
    - Define the tangent space: $T_x\mathcal{M}$;
    - Riemannian gradient: $\text{grad} f(x) = \text{Proj}_{T_x\mathcal{M}} \nabla \overline{f}(x)$, for $\mathcal{M} \subset \mathbb{R}^n$;
    - Retraction operation: $R_x(\xi), \xi \in T_x\mathcal{M}$.
    - Vector transport operation: $\text{Transp}_{T_{x \to R_x(\eta)}\mathcal{M}}(\xi)$.