Optimization without retraction on the random generalized Stiefel manifold:

Landing algorithm for the stochastic CCA

Šimon Váry

SIAM LA24

Deptartment of Statistics
University of Oxford

Joint work with
Pierre Ablin (Apple Machine Learning Group, France)
P.-A. Absil (UCLouvain, Belgium)
Bin Gao (Chinese Academy of Sciences, China)

Outline

Optimization on the generalized Stiefel manifold

Optimization over the (generalized) Stiefel manifold

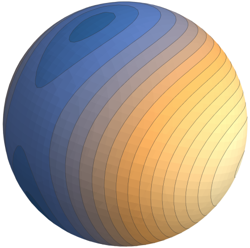
General form

$$\min_{X \in \mathbb{R}^{n \times p}} f(X)$$
$$\text{s.t. } X \in \mathrm{St}_B(p, n) := \{X \in \mathbb{R}^{n \times p} : X^\top B X = I_p\}$$

- $f : \mathbb{R}^{n \times p} \to \mathbb{R}$, continuously differentiable
- $B \in \mathbb{R}^{n \times n}$, positive definite
- $p(p+1)/2$ constraints: non-convex
- $\mathrm{St}_B(p, n)$, *(generalized) Stiefel manifold*

Challenges

- nonconvex constraints
- stochasticity
- preserving feasibility (large scale)
- parallel scalability

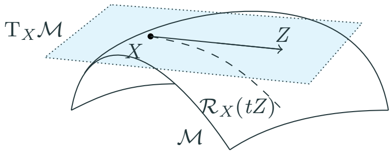


$f(x, y, z) = x^2 + 5y^2 - 3z^2 + 5x$

Riemannian optimization

♋ Riemannian gradient method

1. Choose search direction
   $Z^k = -\mathrm{grad} f(X^k)$
2. Perform a line search scheme and choose a suitable step size t_k
3. Retraction: $X^{k+1} = \mathcal{R}_{X^k}(t_k Z^k)$

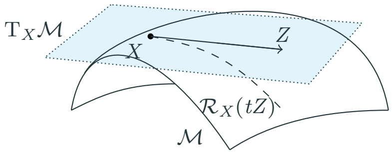


★ How to compute a Riemannian gradient for St_B?
  ☹ depends on chosen metric, but involves B^{-1} or $B^{-1/2}$ ...

★ How to construct a retraction map for St_B?
  ☹ Polar ($B^{-1/2}$), QR-based (Cholesky $LL^\top = X^\top BX$ and L^{-1})...

♋ Riemannian gradient method

1. Choose search direction
   $Z^k = -\mathrm{grad} f(X^k)$
2. Perform a line search scheme and choose a suitable step size t_k
3. Retraction: $X^{k+1} = \mathcal{R}_{X^k}(t_k Z^k)$



★ How to compute a Riemannian gradient for St_B?

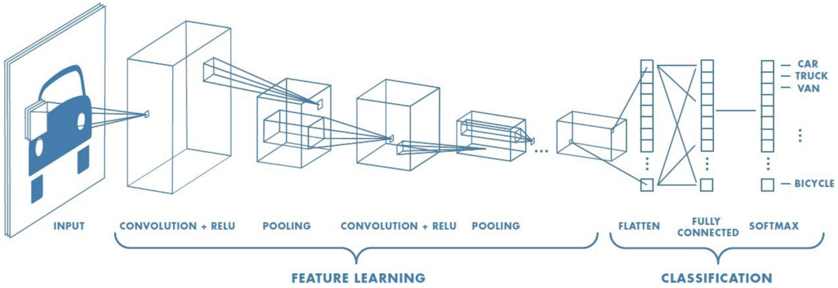
☹ depends on chosen metric, but involves B^{-1} or $B^{-1/2}$ ...

★ How to construct a retraction map for St_B?

☹ Polar ($B^{-1/2}$), QR-based (Cholesky $LL^\top = X^\top BX$ and L^{-1})...

⇝ **New challenges emerging from applications!**

Orthogonal weights in deep learning



Neural networks with Stiefel manifold [Bansal-Chen-Wang'18; Wang-Chen-Chakraborty-Yu'20]

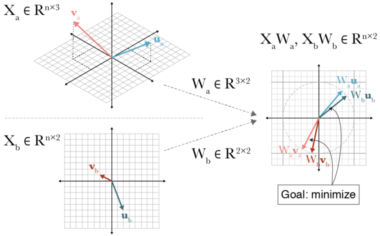
- random variable: ξ, resp. a dataset of N samples d_i:

$$\begin{aligned} \min_{X \in \mathbb{R}^{n\times p}} \quad & \mathbb{E}_\xi[f(X,\xi)] = \frac{1}{N}\sum_{i=1}^N f(X, d_i) \\ \text{s.t.} \quad & X \in \mathrm{St}(p,n) \end{aligned}$$

$\rightsquigarrow$ Cheap stochastic gradient

Similarity between neural network representations [Raghu et al.'17]

- datasets: $D_1 = (d_1^1, \dots, d_1^N)$, $D_2 = (d_2^1, \dots, d_2^N) \in \mathbb{R}^{n \times N}$
- the top-p most correlated principal components: $X, Y \in \mathbb{R}^{n \times p}$

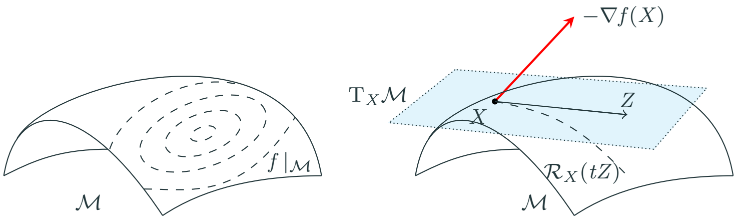


$$\min_{X, Y \in \mathbb{R}^{n \times p}} \quad \mathbb{E}_i \left[-\operatorname{tr}(X^\top d_1^i (d_2^i)^\top Y) \right]$$
$$\text{s.t.} \quad X^\top \mathbb{E}_i[d_1^i (d_1^i)^\top] X = I_p \text{ and } Y^\top \mathbb{E}_i[d_2^i (d_2^i)^\top] Y = I_p$$

$\rightsquigarrow$ **Random manifold**

- rank-deficient sample? *mini-batch*
- storage of B? $B = \begin{bmatrix} \mathbb{E}_i[d_1^i (d_1^i)^\top] & 0 \\ 0 & [d_2^i (d_2^i)^\top] Y = I_p \end{bmatrix}$

Can we still resort to geometric methods?

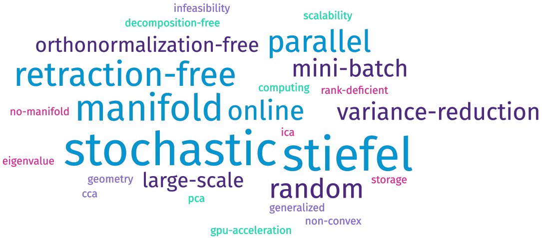


- choose search direction on the tangent space $Z = -\mathrm{grad} f(X)$
  - depends on the Riemannian metric $g(\cdot,\cdot)$, thus projection
- line search with a suitable step size t
- $X + tZ$?
  - retraction: $X^+ = \mathcal{R}_X(tZ)$

$\rightsquigarrow$ Intractable geometry with noisy samples

Landing field and landing flows

An algorithm to overcome the challenges



Desirable algorithm

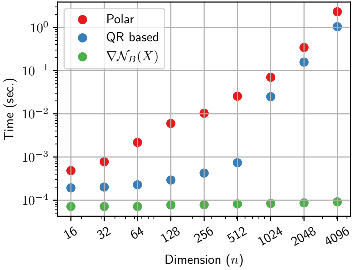
- retraction-free *orthonormalization-free*
- stochastic gradient *variance reduction*
- random manifold with noise *generalized manifold*
- mini-batch *rank-deficient covariance*
- online data *storage of manifold*
- GPU acceleration *parallel scalability*

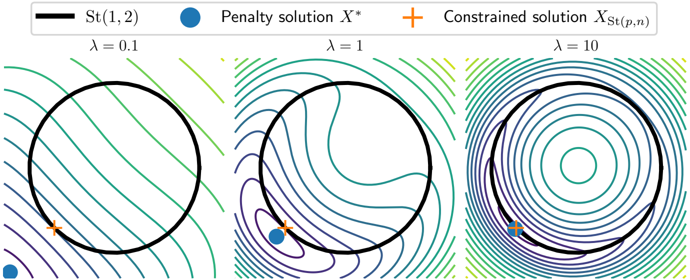
Penalty methods – inexact penalty

$$\min_{X\in\mathbb{R}^{n\times p}} \quad f(X) \qquad \text{s. t.} \quad X \in \mathrm{St}_B(p,n)$$

$$\mathcal{N}(X) = \frac{1}{4}\|X^\top BX - I_p\|_F^2$$

- **Quadratic penalty**: $f(X) + \omega\mathcal{N}(X)$ with $\nabla\mathcal{N}(X) = BX(X^\top BX - I_p)$ being cheap





- ω is small: minimizer is far from manifold
- ω is large: bad condition

Infeasible methods

Penalty $\to$ augmented Lagrangian *exact penalty*

- augmented Lagrangian function [Powell'69; Hestenes'69]

$$f(X) - \frac{1}{2}\langle \Lambda, X^\top BX - I_p\rangle + \omega \mathcal{N}(X)$$

- Fletcher's augmented Lagrangian [Fletcher'70]

$$f(X) - \frac{1}{2}\left\langle (BX)^\dagger[\nabla f(x)], X^\top BX - I_p\right\rangle + \omega \mathcal{N}(X)$$

- modified augmented Lagrangian function **(PLAM)**: [Gao-Liu-Yuan'19]

$$f(X) - \frac{1}{2}\langle \mathrm{sym}(\nabla f(X)^\top X), X^\top BX - I_p\rangle + \omega \mathcal{N}(X)$$

$\rightsquigarrow$ performance is sensitive to the penalty parameter:

$$\omega \geq \omega^* > 0$$

Landing field

Landing system *continuous-time*

$$\dot{X}(t) = -\Lambda\left(X\left(t\right)\right)$$

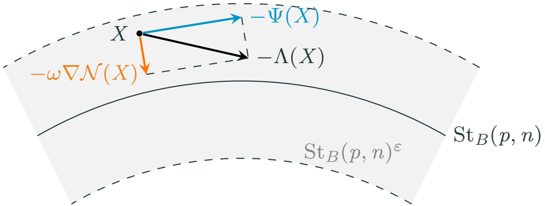
- *landing field*:

$$\Lambda(X) := \Psi(X) + \omega \nabla \mathcal{N}(X)$$

- *relative ascent direction*: $\Psi_B(X)$

$$\Psi_B(X) := 2\,\mathrm{skew}\left(\nabla f(X) X^\top B\right) BX$$

and $\nabla \mathcal{N}(X) = BX(X^\top BX - I_p)$.



Important points:

- $\langle \Psi_B(X), \nabla \mathcal{N}(X) \rangle = 0$, $\rightsquigarrow \omega > 0$
- For $X \in \mathrm{St}_B^\varepsilon$, can guarantee also: $X - \eta\Lambda(X) \in \mathrm{St}_B^\varepsilon$.

Safe step size to remain in the safe region

$$\mathrm{St}_B(p, n)^\varepsilon = \left\{ X \in \mathbb{R}^{n\times d} \,:\, \|X^\top BX - I_p\|_F^2 \leq \frac{\varepsilon^2}{4} \right\}$$

For $d = \|X^\top BX - I_p\|_F$ and $L_\mathcal{N} = \beta_1 + 4\kappa_B$

$$\eta \leq \eta(x) := \frac{\omega\|\nabla\mathcal{N}(X)\|^2 + \sqrt{\omega^2\|\nabla\mathcal{N}(X)\|^4 + L_\mathcal{N}\|\Lambda(X)\|^2(\varepsilon^2 - d^2)}}{L_\mathcal{N}\|\Lambda(X)\|^2},$$

the next iterate stays within the ε-region: $X^{k+1} \in \mathrm{St}_B(p, n)^\varepsilon$

Safe step size to remain in the safe region

$$\mathrm{St}_B(p,n)^\varepsilon = \left\{ X \in \mathbb{R}^{n\times d} \;:\; \|X^\top BX - I_p\|_F^2 \le \frac{\varepsilon^2}{4} \right\}$$

For $d = \|X^\top BX - I_p\|_F$ and $L_\mathcal{N} = \beta_1 + 4\kappa_B$

$$\eta \le \eta(x) := \frac{\omega\|\nabla\mathcal{N}(X)\|^2 + \sqrt{\omega^2\|\nabla\mathcal{N}(X)\|^4 + L_\mathcal{N}\|\Lambda(X)\|^2(\varepsilon^2 - d^2)}}{L_\mathcal{N}\|\Lambda(X)\|^2},$$

the next iterate stays within the ε-region: $X^{k+1} \in \mathrm{St}_B(p,n)^\varepsilon$

Lower bound for the safe step size

$$\eta(X) \ge \eta^* := \min\left\{ \frac{\varepsilon}{\sqrt{2L_\mathcal{N}}\,C_\Psi}, \frac{\omega\,\bar{C}_h^2\varepsilon^2}{L_\mathcal{N}(C_\Psi^2 + \omega^2 C_h\varepsilon^2)} \right\} \;,$$

for $\bar{C}_h = \sqrt{(1-\varepsilon)\kappa_B^{-1}}$, $C_h = \sqrt{(1+\varepsilon)\kappa_B}$, and $C_\Psi \ge \sup_{x\in\mathcal{M}^\varepsilon} \|\Psi(X)\|$.

Discrete-time convergence: global convergence

Merit function [Fletcher's Augmented Lagrangian] [Goyens et al. 2024]

$$\mathcal{L}(X) = f(X) - \frac{1}{2}\left\langle (BX)^{\dagger}[\nabla f(x)], X^\top BX - I_p \right\rangle + \beta\mathcal{N}(X)$$

for suitably chosen β.

Global convergence

For iterations from $X_0 \in \mathrm{St}_B^{\varepsilon}(p,n)$ with bounded $\eta \leq \min\left\{\frac{1}{\kappa_B^2 L_{\mathcal{L}}}, \eta^*\right\}$, and $\omega > 0$

$$\frac{1}{K}\sum_{k=1}^{K}\|\Psi_B(X_k)\|^2 \leq \frac{4(\mathcal{L}(X_0) - \mathcal{L}^*)}{\eta K} \quad \text{and} \quad \frac{1}{K}\sum_{k=1}^{K}\mathcal{N}(X_k) \leq \frac{2(\mathcal{L}(X_0) - \mathcal{L}^*)}{\eta\omega K},$$

where $\mathcal{L}^* = \min_{X\in\mathrm{St}_B^{\varepsilon}(p,n)} \mathcal{L}(X)$ and $L_{\mathcal{L}}$ is Lipschitz constant of $\mathcal{L}$.

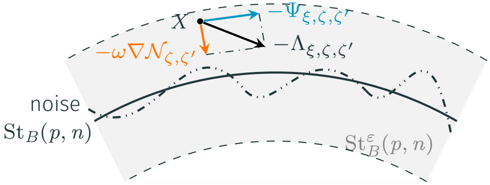
Worst-case complexity

$$\inf_{k\leq K} \|\Psi_B(X_k)\| = \mathcal{O}(1/\sqrt{K}) \qquad \text{and} \qquad \inf_{k\leq K} \|X_k^\top BX_k - I_p\|_F = \mathcal{O}(1/\sqrt{K})$$

Stochastic algorithms

Landing on random Stiefel manifolds

$$\min_{X \in \mathbb{R}^{n\times p}} \quad \mathbb{E}[f_\xi(X)]$$
$$\text{s. t.} \quad X \in \mathrm{St}_B(p,n) := \left\{X \in \mathbb{R}^{n\times p} | X^\top B X = I_p\right\} \text{ and } B = \mathbb{E}[B_\zeta]$$



Stochastic landing

$$X^{k+1} = X^k - \eta_k \Lambda_{\xi^k,\zeta^k,\zeta'^k}(X^k)$$

- $\Lambda_{\xi,\zeta,\zeta'}(X) = \Psi_{\xi,\zeta,\zeta'}(X) + \omega\nabla\mathcal{N}_{\zeta,\zeta'}(X)$
- $\Psi_{\xi,\zeta,\zeta'}(X) = 2\ \mathrm{skew}\left(\nabla f_\xi(X) X^\top B_\zeta\right) B_{\zeta'} X$
- $\nabla\mathcal{N}_{\zeta,\zeta'}(X) = 2B_\zeta X \left(X^\top B_{\zeta'} X - I_p\right)$ and $\mathcal{N}(X) = \frac{1}{4}\|X^\top B X - I_p\|_\mathrm{F}^2$
- Per-iteration complexity for r-batch size: $\mathcal{O}(npr)$ time and $\mathcal{O}(np)$ space

Landing stochastic gradient descent (Landing-SGD)

We have $\mathbb{E}[\Lambda_{\xi^k,\zeta^k,\zeta'^k}(X)] = \Lambda(X)$

$$X_{k+1} = X_k - \eta_k \Lambda_{\xi^k,\zeta^k,\zeta'^k}(X_k)$$

Decreasing step size

For $\eta_k = \eta_0 \times (1+k)^{-\frac{1}{2}}$ where $\eta_0 = 1/(\kappa_B^2 L_{\mathcal{L}})$ and assuming the segments $[X_k X_{k+1}] \in \mathrm{St}_B$

$$\inf_{k \leq K} \mathbb{E}[\|\Psi(X_k)\|^2] = \mathcal{O}\left(\frac{\log(K)}{\sqrt{K}}\right) \quad \text{and} \quad \inf_{k \leq K} \mathbb{E}[\mathcal{N}(X_k)] = \mathcal{O}\left(\frac{\log(K)}{\sqrt{K}}\right)$$

Sample complexity: $\mathcal{O}(\varepsilon^{-2})$ which matches the classic Riemannian SGD

Landing on general manifolds

$$\begin{array}{ll} \min_{x\in\mathbb{R}^d} & f(X) \\ \text{s.t.} & X \in \mathcal{M} := \left\{x \in \mathbb{R}^d : h(x) = 0\right\} \end{array}$$

General landing

$$x_{k+1} = x_k - \eta_k \Lambda(x_k)$$

$$\Lambda(x_k) = \Psi(x) + \omega \nabla \mathcal{N}(x)$$

$\mathcal{N}(X) = \frac{1}{2}\|h(x)\|^2$ $\quad \Big($stochastic $\Big[\Lambda(x^k) + \tilde{E}(x^k, \Xi^k)\Big]\Big)$

Relative ascent direction

A relative ascent direction $\Psi(x) : \mathbb{R}^d \to \mathbb{R}^d$, with a parameter $\rho > 0$ that may depend on ε satisfies:

1. (orthogonality) $\forall x \in \mathcal{M}^\varepsilon, \quad \forall v \in \operatorname{span}(\mathrm{D}h(x)^*) : \langle \Psi(x), v\rangle = 0$;
2. (gradient-related) $\forall x \in \mathcal{M}^\varepsilon$ we have that $\langle \Psi(x), \nabla f(x)\rangle \geq \rho\|\Psi(x)\|^2$;
3. (optimality) For $x \in \mathcal{M}$, we have that $\langle \Psi(x), \nabla f(x)\rangle = 0$ if and only if x is a critical point of f on $\mathcal{M}$

Numerical experiments

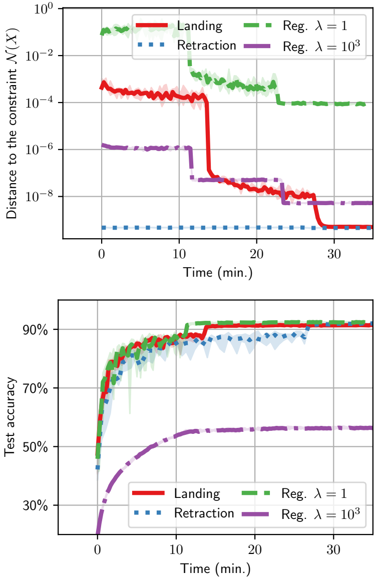
Numerical test on convolutional neural network with orthogonal kernels

Orthogonal CNN

$$\begin{array}{ll} \min\limits_{\theta} & \sum_i^N \ell(f_\Theta(x_i), y_i) \\ \text{s.t.} & \theta \in \Theta_{\text{orth}} : \theta_i \in \text{St}(p, n) \end{array}$$

- $f_\Theta(\cdot)$ is VGG16 convolutional neural network,
- Θ_{orth} includes 13 matrices of size $\approx 1\,000^2$,
- (x_i, y_i) samples from CIFAR-10,

with a batch size of 128 samples, fixed stepsize (decreasing every 50 epochs)

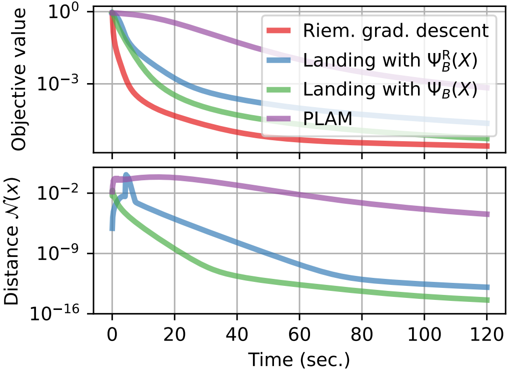


Numerical test on generalized eigenvalue problem

Generalized eigenvalue problem

$$\min_{X \in \mathbb{R}^{n\times p}} \ \mathrm{tr}(X^\top A X) \quad \text{s.t.} \quad X \in \mathrm{St}_B(p, n)$$

- condition number: $\kappa = 100$
- dimension: $n = 1000$ and $p = 500$
- $\lambda(A)_i \in [1/\kappa, 1]$
- $\lambda(B)_i \in [1/\kappa, 1]$.
- GPU acceleration: CUDA

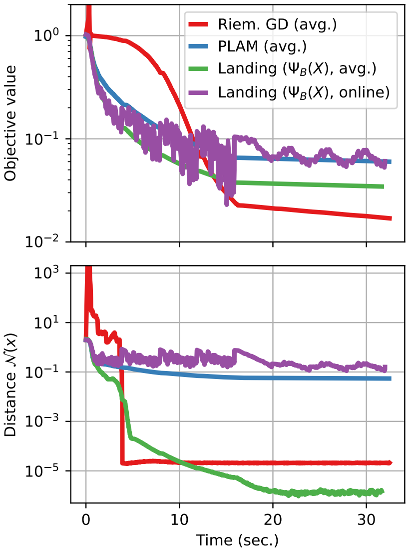


Numerical test on stochastic CCA

Stochastic CCA

$$\begin{aligned} \min_{X, Y \in \mathbb{R}^{n\times p}} \quad & \mathbb{E}_i \left[-\operatorname{tr}(X^\top d_1^i (d_2^i)^\top Y) \right] \\ \text{s.t.} \quad & X^\top \mathbb{E}_i[d_1^i (d_1^i)^\top] X = I_p \\ & Y^\top \mathbb{E}_i[d_2^i (d_2^i)^\top] Y = I_p \end{aligned}$$

- Benchmark test on MNIST (60 000 samples)
- dimension: $n = 28^2, p = 5$
- batch size: 512



Conclusion and perspectives

Take-home notes

- retraction-free algorithms
  decomposition-free; parallel scalability; BLAS operation
- stochastic gradient + noisy manifold
- generalized stiefel + general manifolds
- – higher-order landing flow
- – other manifolds
- – line-search?

References

✠ Pierre Ablin, P.-A. Absil, Bin Gao, Simon Vary

1. *Optimization flows landing on the Stiefel manifold*
   25th IFAC Symposium on Mathematical Theory of Networks and Systems (MTNS 2022), IFAC-PapersOnLine, 55-30 (2022), 25–30
2. *Infeasible deterministic, stochastic, and variance-reduction algorithms for optimization under orthogonality constraints.*
   arXiv:2303.16510, (2023)
3. *Optimization without retraction on the random generalized Stiefel manifold*, ICML, (2024)

Thanks for your attention!

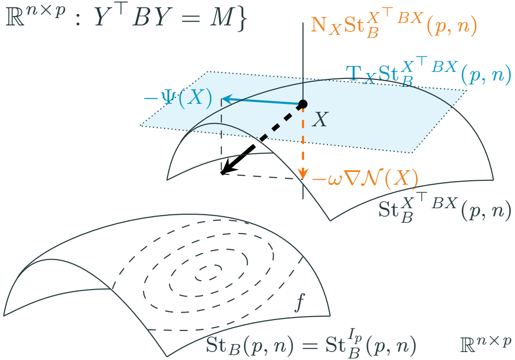
Email: simon.vary@uclouvain.be
Homepage: https://simonvary.github.io

Geometric interpretation of the landing

Geometry: $X \notin \mathrm{St}_B(p,n)$

$$\mathrm{St}_B^M(p,n) = \{Y \in \mathbb{R}^{n\times p} : Y^\top B Y = M\}$$

- diffeomorphism from $\mathrm{St}(p,n)$ to $\mathrm{St}_B^M(p,n)$:
  $\Phi_M : \mathbb{R}^{n\times p} \to \mathbb{R}^{n\times p} : X \mapsto Y = B^{-\frac{1}{2}} X M^{\frac{1}{2}}$
- tangent space:
  $\mathrm{T}_Y \mathrm{St}^M(p,n) = \{WBY : W \in \mathcal{S}_{\mathrm{skew}}^n\}$
- normal space:
  $\mathrm{N}_Y \mathrm{St}^M(p,n) = \{Y(Y^\top Y)^{-1} S : S \in \mathcal{S}_{\mathrm{sym}}^p\}$
- Riemannian gradient:
  $\mathrm{grad} f(X) = \mathrm{sk}(B^{-1}\nabla f(X) X^T) BX$



$$\Lambda(X) = \underbrace{\Psi(X)}_{\text{Relative desc. direction}} + \underbrace{\omega\nabla\mathcal{N}(X)}_{\text{normal direction}}$$