Optimization flows landing on the Stiefel manifold

continuous-time flows, deterministic and stochastic algorithms

Bin Gao

Academy of Mathematics and Systems Science
Chinese Academy of Sciences

Joint work with
Pierre Ablin (Apple, France)
P.-A. Absil (UCLouvain, Belgium)
Simon Vary (Oxford, UK)

Outline

Optimization on the Stiefel manifold

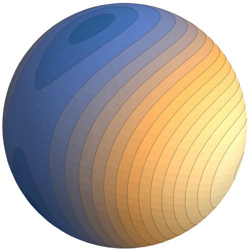
Optimization over the Stiefel manifold

General form

$$\min_{X \in \mathbb{R}^{n\times p}} \quad f(X)$$
$$\text{s.t.} \quad X^\top X = I_p \quad (p \ll n)$$

- $f : \mathbb{R}^{n\times p} \to \mathbb{R}$, continuously differentiable
- $p(p+1)/2$ constraints: nonconvex
- *Stiefel manifold*:

$$\mathrm{St}(p,n) := \{X \in \mathbb{R}^{n\times p} : X^\top X = I_p\}$$



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

Challenges

- nonconvex constraints
- NP-hard (special f)
- preserving feasibility (large scale)
- parallel scalability

♣ Optimization on matrix manifolds

- Steepest descent: [Helmke-Moore'94; Udriste'94]
- Conjugate gradient: [Edelman-Arias-Smith'98; Brace-Manton'06; Smith'94; Gallivan-Absil'10];
- Newton: [Smith'94; Edelman-Arias-Smith'98; Hu-Wen-Milzarek-Yuan'17; Zhao-Bai'22]
- Quasi-Newton: [Edelman-Arias-Smith'98; Brace-Manton'06; Gallivan-Absil'10; Huang-Gallivan-Absil'15]
- Trust region: [Absil-Baker-Gallivan'07]
- Geodesic search in canonical metric: [Abrudan-Eriksson-Koivunen'08]
- Cayley transformation: [Nishimori-Akaho'05]

♣ Searching in tangent space

- Projection-based method: [Manton'02; Absil-Mahony-Sepulchre'08]
- Constraint preserving update scheme: [Wen-Yin'12; Jiang-Dai '14]

♣ Other types of work

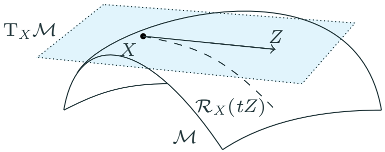
- Splitting and alternating: [Lai-Osher'14]
- Non-retraction based framework: [**G.**-Liu-Chen-Yuan'18; Wang-**G.**-Liu'21]
- Vector transport-free SVRG: [Liu-So-Wu'15; Jiang-Ma-So-Zhang'17]
- Constraint dissolving optimization: [Xiao-Liu-Toh'21-24]
- Gradient flows and PCG in DFT: [Dai-Zhou'14-'23]

Optimization algorithms on matrix manifolds [Absil-Mahony-Sepulchre'08]

An introduction to optimization on smooth manifolds [Boumal'23]

♋ 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: $X^{k+1} = \mathcal{R}_{X^k}(t_k Z^k)$



Retraction: For all $X \in \mathcal{M}$ *in general, it is globally defined*

1) $\mathcal{R}_X(0_X) = X$, where 0_X is the origin of $\mathrm{T}_X\mathcal{M}$;
2) $\frac{\mathrm{d}}{\mathrm{dt}}\mathcal{R}_X(tZ)|_{t=0} = Z$ for all $Z \in \mathrm{T}_X\mathcal{M}$

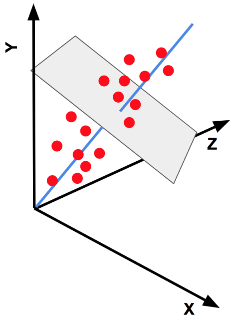
★ How to construct a retraction map for $\mathcal{M}$?
☹ Stiefel manifold: SVD, QR, Polar, Cayley...

$\rightsquigarrow$ **New challenges emerging from applications!**

Dimensionality reduction: $\mathbb{R}^m \longrightarrow \mathbb{R}^p$

[Pearson'01; Jolliffe'86; Oja'01; Zou'10-...]

- sample size: n
- feature space: $\mathbb{R}^m$
- observation data matrix: $A \in \mathbb{R}^{n\times m}$



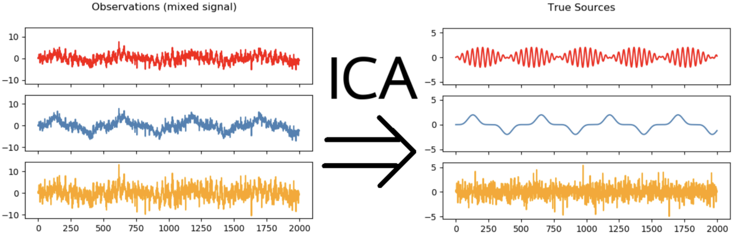
$$\begin{array}{cc} \min_{X\in\mathbb{R}^{n\times p}} & -\frac{1}{m}\operatorname{tr}\left(X^\top(A-\bar{A})(A-\bar{A})^\top X\right) \\ \text{s. t.} & X \in \operatorname{St}(p,n) \end{array}$$

where $\bar{A} = \frac{1}{m}\sum_{i=1}^{m} A_i \mathbf{1}^\top$

$\rightsquigarrow$ Large sample size n

- online PCA?
- GPU acceleration?

Independent Component Analysis (ICA)



Separation of a mixture of signals [Hyvarinen'99]

- data matrix: $A = [a_1, \ldots, a_N] \in \mathbb{R}^{N \times n}$
- scalar function: $\sigma(x) = \log(\cosh(x))$

$$\min_{X \in \mathbb{R}^{n \times n}} \quad \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{n} \sigma([AX]_{ij}) \qquad \text{s.t.} \quad X \in \mathrm{St}(n, n)$$

$\rightsquigarrow$ Average of N functions

- mini-batch?

Generalized eigenvalue problem

$$Ax = \lambda Bx$$

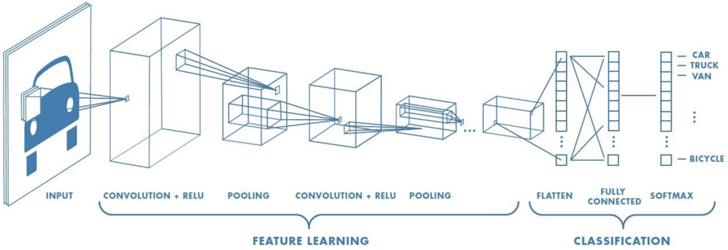
Rayleigh-Ritz trace minimization [Shustin-Avron'23]

- B: symmetric positive definite
- generalized Stiefel manifold: $\mathrm{St}_B(p, n) := \{X \in \mathbb{R}^{n\times p} : X^\top BX = I_p\}$

$$\begin{array}{ll} \min\limits_{X\in\mathbb{R}^{n\times p}} & \mathrm{tr}(X^\top AX) \\ \mathrm{s.\,t.} & X \in \mathrm{St}_B(p, n) \end{array} \quad \rightsquigarrow \quad \text{Generalized Stiefel manifold}$$

- matrix decomposition for B?

Orthogonal weights in deep learning



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

- random variable: ξ

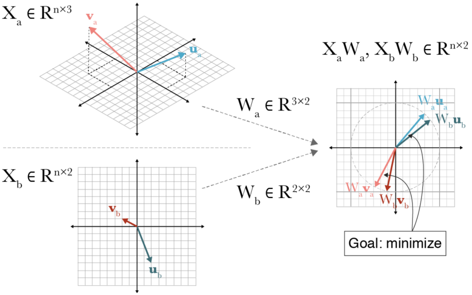
$$\begin{array}{ll} \min\limits_{X \in \mathbb{R}^{n\times p}} & \mathbb{E}_\xi[f(X,\xi)] \\ \text{s.t.} & X \in \mathrm{St}(p,n) \end{array} \quad \rightsquigarrow \quad \text{Stochastic gradient}$$

- variance reduction?

Canonical Correlation Analysis (CCA)

Measuring similarity between datasets [Raghu et al.'17]

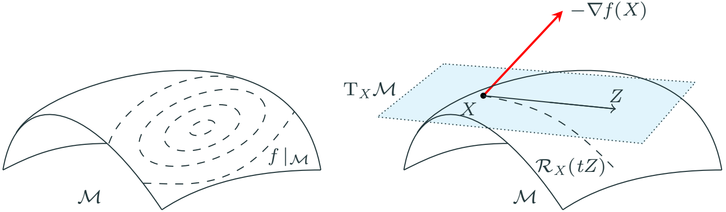
- sample size: N
- datasets: $D_1 = (d_1^1, \ldots, d_1^N)$, $D_2 = (d_2^1, \ldots, d_2^N) \in \mathbb{R}^{n \times N}$
- the top-p most correlated principal components: $X, Y \in \mathbb{R}^{n \times p}$



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

$\rightsquigarrow$ Random manifold

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

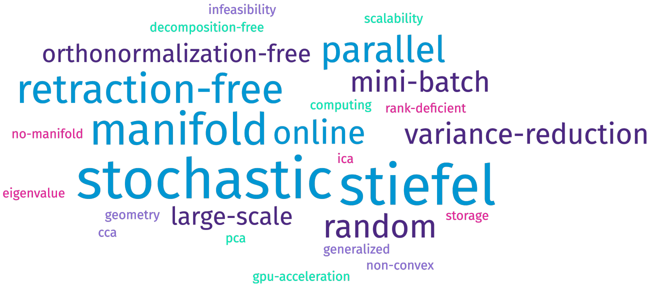


- choose search direction on the tangent space $Z = -\mathrm{grad} f(X)$
  - depends on the Riemannian metric $g(\cdot,\cdot)$, thus projection
- line search with a suitable step size t
- $X + tZ$?
  - retraction: $X^+ = \mathcal{R}_X(tZ)$

$\rightsquigarrow$ Intractable geometry with noisy

Landing field and landing flows

One-shot algorithm attempts to resolve all challenges



Desirable one-shot algorithm

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

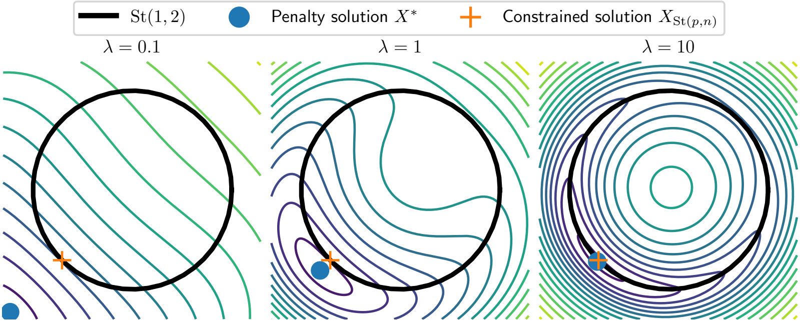
Penalty methods

Penalty *inexact penalty*

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

$$\mathcal{N}(X) = \frac{1}{4}\|X^\top X - I_p\|_{\mathrm{F}}^2$$

- **Quadratic penalty**: $f(X) + \lambda \mathcal{N}(X)$ [Xie-Xiong-Pu'17; Balestriero'18; Bansal-Chen-Wang'18]



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

Penalty $\rightarrow$ augmented Lagrangian *exact penalty*

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

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

- Fletcher's augmented Lagrangian [Fletcher'70]

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

- modified augmented Lagrangian function [G.-Liu-Yuan'19]

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

- constraint dissolving function [Xiao-Liu-Toh'23]

$$f\left(X\left(\frac{3}{2}I_p - \frac{1}{2}X^\top X\right)\right) + \lambda \mathcal{N}(X)$$

$\rightsquigarrow$ performance is sensitive to the penalty parameter $\lambda \geq \lambda^* > 0$

Landing system *continuous-time*

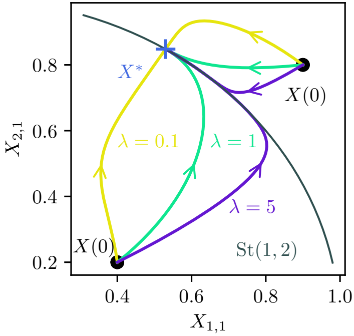
$$\dot{X}(t) = -\Lambda\left(X\left(t\right)\right)$$

- *landing field*:

$$\Lambda(X) := \psi(X)X + \lambda\,\nabla\mathcal{N}(X)$$

- *relative gradient*: $\psi(X)X$

$$\psi(X) := 2\,\text{skew}\left(\nabla f(X)X^\top\right)$$

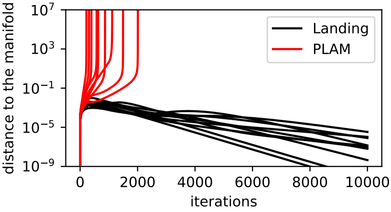


A cool feature

$$\langle \psi(X)X, \nabla\mathcal{N}(X)\rangle$$
$$= \left\langle \psi(X), X^\top(X^\top X - I)X\right\rangle$$
$$= 0$$

- always orthogonal $\rightsquigarrow \lambda > 0$
- PLAM: [G.-Liu-Yuan'19]

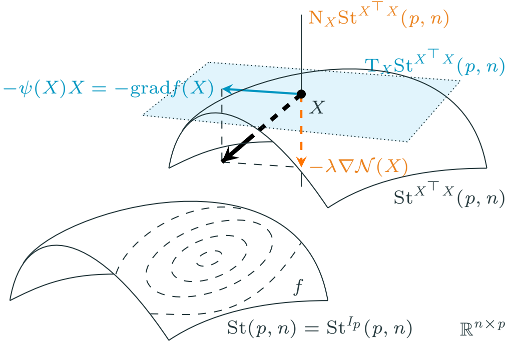
$$\nabla f(X) - X\,\text{sym}(\nabla f(X)^\top X) + \lambda\,\nabla\mathcal{N}(X)$$



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

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

- diffeomorphism from $\mathrm{St}(p,n)$ to $\mathrm{St}^M(p,n)$: $\Phi_M : \mathbb{R}^{n\times p} \to \mathbb{R}^{n\times p}$: $X \mapsto Y = XM^{\frac{1}{2}}$
- metric: $g_Y(\xi,\zeta) = \langle \xi, (I_n - \frac{1}{2}Y(Y^\top Y)^{-1}Y^\top)\zeta(Y^\top Y)^{-1}\rangle$.
- tangent space: $\mathrm{T}_Y\mathrm{St}^M(p,n) = \{WY : W \in \mathcal{S}^n_{\mathrm{skew}}\}$
- normal space: $\mathrm{N}_Y\mathrm{St}^M(p,n) = \{Y(Y^\top Y)^{-1}S : S \in \mathcal{S}^p_{\mathrm{sym}}\}$
- Riemannian gradient: $\mathrm{grad}f(X) = \psi(X)X$



$$\Lambda(X) = \underbrace{\psi(X)X}_{\text{Riemannian gradient}} + \underbrace{\lambda\nabla\mathcal{N}(X)}_{\text{normal vector}}$$

Continuous-time convergence: stability

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

- solutions (landing flow) exist and are unique:
  $\varphi_t(X_0)$ starting from $X_0 \in \mathbb{R}_*^{n\times p}$
- penalty is nonincreasing:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{N}(X(t)) = -\lambda\left\|\nabla\mathcal{N}(X(t))\right\|_{\mathrm{F}}^2 \leq 0$$

- convergence to the Stiefel manifold:

$$\lim_{t\to\infty}\mathcal{N}(\varphi_t(X_0)) = 0$$

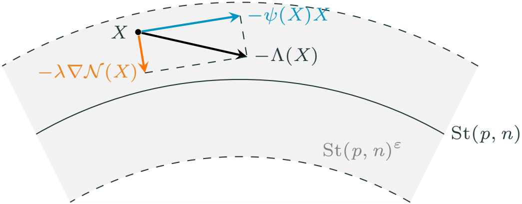
- convergence to the set of critical points:

$$X^* \in \{X^* \in \mathrm{St}(p,n) : \psi(X^*)X^* = 0\} \quad \text{if and only if} \quad \Lambda(X^*) = 0$$

- asymptotic stability: For all $X_0 \in \mathbb{R}_*^{n\times p}$, if X^* is a local minimum and isolated critical point of f relative to $\mathrm{St}(p,n)$, and if X^* is an ω-limit point of $\varphi_t(X_0)$, then $\lim_{t\to\infty}\varphi_t(X_0) = X^*$

Discrete-time convergence: safe step size

$$X_{k+1} = X_k - \eta_k \Lambda(X_k)$$
$$\Lambda(X_k) = \psi(X_k)X_k + \lambda \nabla \mathcal{N}(X_k)$$



Sage region and step size

$$\mathrm{St}(p,n)^\varepsilon = \{X \in \mathbb{R}^{n\times p} |\ \mathcal{N}(X) \leq \frac{1}{4}\varepsilon^2\}$$

Let $\mathcal{N}(X_k) = d^2 \leq \varepsilon^2$ and $g = \|\Lambda(X_k)\|_F$, then if

$$\eta_k \leq \eta(X_k) := \min\left\{\frac{\lambda d(1-d) + \sqrt{\lambda^2 d^2 (1-d)^2 + g^2(\varepsilon - d)}}{g^2}, \frac{1}{2\lambda}\right\},$$

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

Lower bound for step size

$$\eta(X_k) \geq \eta^* := \min\left\{\frac{\lambda(1-\varepsilon)\varepsilon}{a^2 + \lambda^2(1+\varepsilon)\varepsilon^2}, \sqrt{\frac{\varepsilon}{2a^2}}, \frac{1}{2\lambda}\right\}$$

where $a = \sup_{X\in \mathrm{St}^\varepsilon(p,n)} \|\psi(X)X\|_F$

Discrete-time convergence: global convergence

Merit function [G.-Liu-Yuan'19]

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

for suitably chosen $\mu > \frac{1}{2} \max_{X \in \mathrm{St}^\varepsilon(p,n)} \|\nabla f(X)\|_F$

Global convergence

For iterations from $X_0 \in \mathrm{St}^\varepsilon(p,n)$ with bounded $\eta \leq \min(\frac{1}{2L_g}, \frac{\mu}{4\lambda L_g \sqrt{1+\varepsilon}}, \eta^*)$

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

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

Worst-case complexity *$\mathcal{O}(\epsilon^{-2})$ iterations to ϵ-stationary point*

$$\inf_{k \leq K} \|\mathrm{grad} f(X_k)\| = \mathcal{O}(1/\sqrt{K}) \quad \text{and} \quad \inf_{k \leq K} \|X_k^\top X_k - I_p\|_F = \mathcal{O}(1/\sqrt{K})$$

Deterministic and stochastic algorithms

A family of landing algorithms

$$\begin{array}{ll} \min\limits_{X \in \mathbb{R}^{n \times p}} & \frac{1}{N} \sum_{i=1}^{N} f_i(X) \\ \text{s.t.} & X \in \mathrm{St}(p, n) \end{array}$$

Landing gradient descent: a prototype

$$X_{k+1} = X_k - \eta_k \Lambda(X_k)$$

- $\Lambda(X) = \frac{1}{N} \sum_{i=1}^{N} \Lambda_i(X)$
- $\Lambda_i(X) = \psi_i(X) + \lambda \nabla \mathcal{N}(X)$
- $\psi_i(X) = 2\,\mathrm{skew}\left(\nabla f_i(X) X^\top\right)$

Landing stochastic gradient descent (Landing-SGD)

Assume $\mathbb{E}_i[\Lambda_i(X)] = \Lambda(X)$

$$X_{k+1} = X_k - \eta_k \Lambda_{i_k}(X_k)$$

Decreasing step size

$\eta_k = \eta_0 \times (1+k)^{-\frac{1}{2}}$ and $\eta_0 = \min(\frac{1}{2L_g}, \frac{\nu}{4\lambda^2 L_g(1+\varepsilon)}, \eta^*)$

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

Constant step size

$\eta = \eta_0 \times (1+K)^{-\frac{1}{2}}$ and $\eta_0 = \min(\frac{1}{2L_g}, \frac{\nu}{4\lambda^2 L_g(1+\varepsilon)}, \eta^*)$

$$\inf_{k\leq K} \mathbb{E}[\|\mathrm{grad} f(X_k)\|^2] = \mathcal{O}\left(\frac{1}{\sqrt{K}}\right) \quad \text{and} \quad \inf_{k\leq K} \mathbb{E}[\|\mathcal{N}(X_k)\|^2] = \mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$$

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

Landing-SAGA: variance reduction

Assume $\mathbb{E}_i[\,\Lambda_k^{i_k}] = \Lambda(X)$

$$X_{k+1} = X_k - \eta\Lambda_k^{i_k}(X_k)$$

- batch size: m
- $\Lambda_k^{i_k}(X_k) =$
  $\mathrm{grad} f_{i_k}(X_k) - \mathrm{skew}(\Phi_k^{i_k} X_k^\top)X_k + \frac{1}{m}\sum_{j=1}^m \mathrm{skew}(\Phi_k^j X_k^\top)X_k + \lambda\nabla\mathcal{N}(X)$
- $\Phi_{k+1}^{i_k} = \nabla f_{i_k}(X_k)$ and $\Phi_{k+1}^j = \Phi_k^j$ for all $j \neq i_k$

Constant step size

Assume

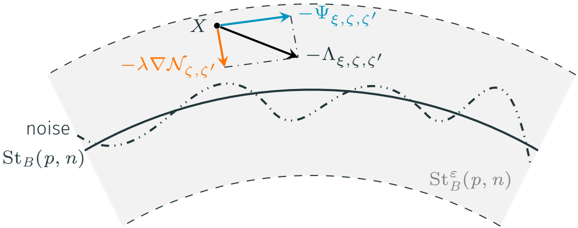
$$\eta \leq \min\left(\eta^*, \frac{\rho}{L_g}, \frac{1}{\sqrt{8N(1+\varepsilon)L_f}}, \left(\frac{\rho}{4N(4N+2)L_g L_f^2(1+\varepsilon)}\right)^{1/3}\right)$$

Then, we have

$$\inf_{k\leq K} \mathbb{E}[\|\mathrm{grad} f(X_k)\|^2] = O\left(\frac{1}{K}\right) \quad \text{and} \quad \inf_{k\leq K} \mathbb{E}[\|\mathcal{N}(X_k)\|^2] = O\left(\frac{1}{K}\right)$$

Sample complexity: $\mathcal{O}(N^{\frac{2}{3}}\varepsilon^{-1})$ which matches the Euclidean SAGA

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



Stochastic landing

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

- $\Lambda_{\xi,\zeta,\zeta'}(X) = \Psi_{\xi,\zeta,\zeta'}(X) + \lambda \nabla \mathcal{N}_{\zeta,\zeta'}(X)$
- $\Psi_{\xi,\zeta,\zeta'}(X) = 2 \,\mathrm{skew}\left(\nabla f_\xi(X) X^\top B_\zeta\right) B_{\zeta'} X$
- $\nabla \mathcal{N}_{\zeta,\zeta'}(X) = 2 B_{\zeta'} X \left(X^\top B_\zeta X - I_p\right)$ and $\mathcal{N}(X) = \frac{1}{4}\|X^\top B X - I_p\|_{\mathrm{F}}^2$

$$\min_{x\in\mathbb{R}^d} \quad f(X)$$
$$\text{s. t.} \quad X \in \mathcal{M} := \{x \in \mathbb{R}^d : h(x) = 0\}$$

General landing

$$x_{k+1} = x_k - \eta_k \Lambda(x_k)$$
$$\Lambda(x_k) = \Psi(x) + \lambda \nabla \mathcal{N}(x)$$

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

Relative descent direction

A relative descent 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 $\mathcal{M}$

Convergence of general landing

Sage region and step size

$$\mathcal{M}^{\varepsilon} = \left\{x \in \mathbb{R}^{d} \ : \ \|h(x)\| \leq \varepsilon\right\}$$

If

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

the next iterate stays within the ε-region: $x_{k+1} \in \mathcal{M}^{\varepsilon}$

Lower bound for step size

$$\eta(x) \geq \min\left\{\frac{\varepsilon}{\sqrt{2L_{\mathcal{N}}}C_{\Psi}}, \frac{\lambda\bar{C}_h^2\varepsilon^2}{L_{\mathcal{N}}(C_{\Psi}^2 + \lambda^2 C_h\varepsilon^2)}\right\}$$

Convergence

The landing iteration starting from $x_0 \in \mathcal{M}^{\varepsilon}$ satisfies

$$\frac{1}{K}\sum_{k=1}^{K}\|\Psi(x_k)\|^2 \leq 4\frac{\mathcal{L}(x^0) - \mathcal{L}^*}{\eta\rho K} \quad \text{and} \quad \frac{1}{K}\sum_{k=1}^{K}\|h(x_k)\|^2 \leq 4\frac{\mathcal{L}(x^0) - \mathcal{L}^*}{\eta\rho\lambda^2 K}$$

for a constant step size $\eta \leq \min\left\{\frac{\rho}{2L_{\mathcal{L}}}, \frac{\rho}{2L_{\mathcal{L}}C_h^2}, \frac{\varepsilon}{\sqrt{2L_{\mathcal{N}}}C_{\Psi}}, \frac{\lambda\bar{C}_h^2\varepsilon^2}{L_{\mathcal{N}}(C_{\Psi}^2+\lambda^2 C_h\varepsilon^2)}\right\}$

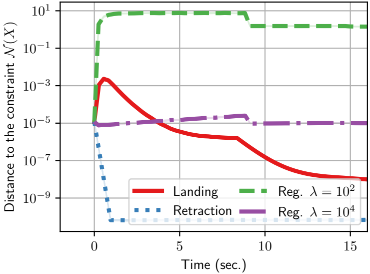
Numerical experiments

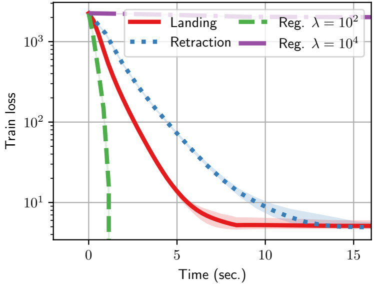
Numerical test on online PCA

Principal component analysis

$$\begin{array}{ll} \min\limits_{X \in \mathbb{R}^{n \times p}} & -\frac{1}{2} \|AX\|_F^2 \\ \text{s.t.} & X \in \mathrm{St}(p, n) \end{array}$$

- dimension: $n = 5000$
- sample size: $N = 15000$
- $A \in \mathbb{R}^{N \times n}$
- batch size: 128
- subspace dimension: $p = 500$

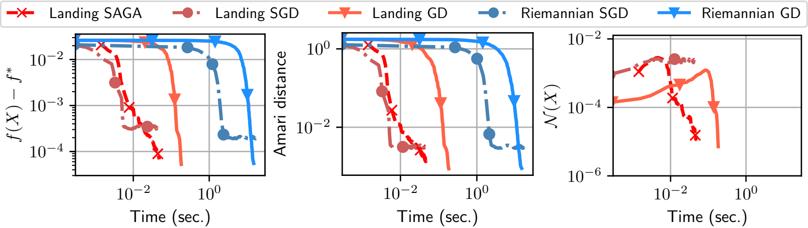




Independent component analysis

$$\begin{array}{ll} \min\limits_{X \in \mathbb{R}^{n\times n}} & \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{n}\sigma([AX]_{ij}) \\ \text{s.t.} & X \in \mathrm{St}(n,n) \end{array}$$

- dimension: $n = 10$
- sample size: $N = 10000$
- $A = SB^\top$ and $S \in \mathbb{R}^{N\times n}$

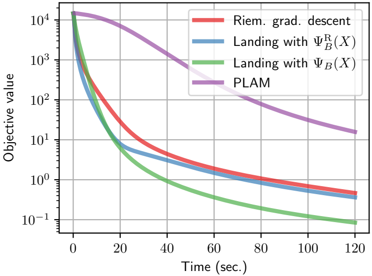


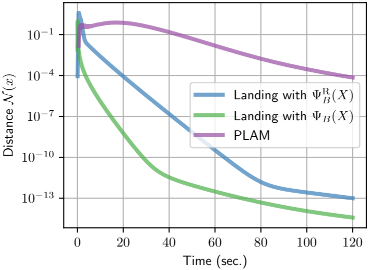
Numerical test on generalized eigenvalue problem

Generalized eigenvalue problem

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

- 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





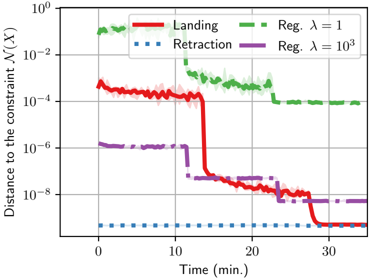
Numerical test on convolutional neural network with orthogonal kernels

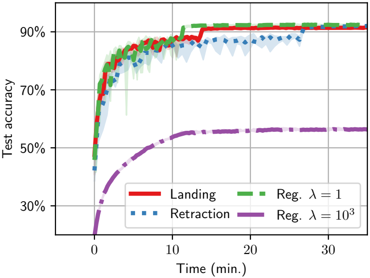
Orthogonal CNN

$$\begin{aligned} \min_{\theta} \quad & \sum_i^N \ell(f_\Theta(x_i), y_i) \\ \text{s.t.} \quad & \theta \in \Theta_{\text{orth}} : \theta_i \in \mathrm{St}(p, n) \end{aligned}$$

- $f_\Theta(\cdot)$ 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)



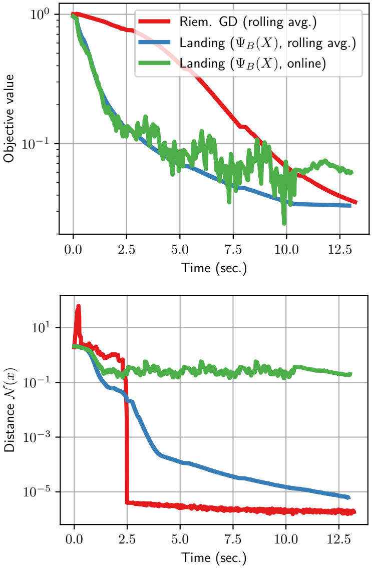


Numerical test on stochastic CCA

Stochastic CCA

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

- online
- dimension: $p = 5$
- 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

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.*
   Journal of Machine Learning Research, (2024), accepted.
3. *Optimization without retraction on the random generalized Stiefel manifold*
   ICML 2024

Thanks for your attention!

Email: gaobin@lsec.cc.ac.cn
Homepage: https://www.gaobin.cc
Group blog: https://www.gaobin.cc/popman

homepage

group blog