Low Rank Approximation
Lecture 8

Daniel Kressner
Chair for Numerical Algorithms and HPC
Institute of Mathematics, EPFL
daniel.kressner@epfl.ch

Manifold optimization

General setting: Aim at solving optimization problem

$$\min_{X \in \mathcal{M}_r} f(X),$$

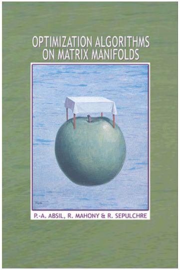
where $\mathcal{M}_r$ is a manifold of rank-r matrices or tensors.

Goal: Modify classical optimization algorithms (line search, Newton, quasi-Newton, ...) to produce iterates that stay on $\mathcal{M}_r$.

Advantages over ALS:

- No need to solve subproblems, at least for first-order methods;
- Relatively straightforward local convergence analysis.

Two valuable resources:



- Absil/Mahony/Sepulchre'2011: Optimization Algorithms on Matrix Manifolds. PUP, 2008. Available from `https://press.princeton.edu/absil`.
- Manopt, a Matlab toolbox for optimization on manifolds. Available from `https://manopt.org/`.

Manifolds

For *open* set $\mathcal{U} \subset \mathcal{M}$, chart is bijective function $\varphi : \mathcal{U} \to \mathbb{R}^d$.
Atlas of $\mathcal{M}$ into $\mathbb{R}^d$ is collection of charts $(\mathcal{U}_\alpha, \varphi_\alpha)$ such that:

- $\bigcup_\alpha \mathcal{U}_\alpha = \mathcal{M}$
- for any α, β with $\mathcal{U}_\alpha \cap \mathcal{U}_\beta \neq \{\emptyset\}$, change of coordinates

$$\varphi_\beta \circ \varphi_\alpha^{-1} : \mathbb{R}^d \to \mathbb{R}^d$$

is smooth (C^∞) on its domain $\varphi_\alpha(\mathcal{U}_\alpha \cap \mathcal{U}_\beta)$.

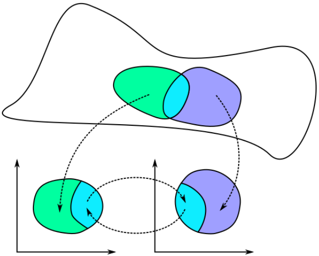


Illustration taken from Wikipedia.

Manifolds

In the following, we assume that atlas is maximal. Proper definition of smooth manifold $\mathcal{M}$ needs further properties (topology induced by maximal atlas is Hausdorff and second-countable). See [Lee'2003] and [Absil et al.'2008].

Properties of $\mathcal{M}$:

- finite-dimensional vector spaces are always manifolds;
- d = dimension of $\mathcal{M}$;
- $\mathcal{M}$ does not need to be connected (in the context of smooth optimization makes sense to consider connected manifolds only);
- function $f : \mathcal{M} \to \mathbb{R}$ differentiable at point $x \in \mathcal{M}$ if and only if

$$f \circ \varphi^{-1} : \varphi(\mathcal{U}) \subset \mathbb{R}^d \to \mathbb{R}$$

is differentiable at $\varphi(x)$ for some chart $(\mathcal{U}, \varphi)$ with $x \in \mathcal{U}$.

Manifolds: First examples

Lemma

Let $\mathcal{M}$ be a smooth manifold and $\mathcal{N} \subset \mathcal{M}$ an open subset. Then $\mathcal{N}$ is a smooth manifold (of equal dimension).

Proof: Given atlas for $\mathcal{M}$ obtain atlas for $\mathcal{N}$ by selecting charts $(\mathcal{U}, \varphi)$ with $\mathcal{U} \subset \mathcal{N}$.

Example: $\mathrm{GL}(n, \mathbb{R})$, the set of real invertible $n \times n$ matrices, is a smooth manifold.

EFY. Show that $\mathbb{R}_*^{m \times n}$, the set of real $m \times n$ matrices of full rank $\min\{m, n\}$, is a smooth manifold.

EFY. Show that the set of $n \times n$ symmetric positive definite matrices is a smooth manifold.

Two main classes of matrix manifolds:

- embedded submanifolds of $\mathbb{R}^{m \times n}$;
  Example: Stiefel manifold of orthonormal bases.
- quotient manifolds;
  Example: Grassmann manifold $\mathbb{R}_*^{m \times n}/\mathrm{GL}(n, \mathbb{R})$.

Will focus on embedded submanifolds (much easier to work with).

Immersions and submersion

Let $\mathcal{M}_1, \mathcal{M}_2$ be smooth manifolds and $F : \mathcal{M}_1 \to \mathcal{M}_2$. Let $x \in \mathcal{M}_1$ and $y = F(x) \in \mathcal{M}_2$. Choose charts φ_1, φ_2 around x, y. Then coordinate representation of F given by

$$\hat{F} := \varphi_2 \circ F \circ \varphi_1^{-1} : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}.$$

- F is called smooth if $\hat{F}$ is smooth.
- rank of F at $x \in \mathcal{M}_1$ defined as the rank of $D\hat{F}(\varphi(x_1))$ (Jacobian of $\hat{F}$ at $\varphi(x_1)$)
- F is called an immersion if its rank equals d_1 at every $x \in \mathcal{M}_1$.
- F is called a submersion if its rank equals d_2 at every $x \in \mathcal{M}_1$.

Embedded submanifolds

Subset $\mathcal{N} \subset \mathcal{M}$ is called an embedded submanifold of dimension k in $\mathcal{M}$ if for each point $p \in \mathcal{N}$ there is a chart $(\mathcal{U}, \varphi)$ in $\mathcal{M}$ such that all elements of $\mathcal{U} \cap \mathcal{N}$ are obtained by varying first k coordinates only. (See Chapter 5 of [Lee'2003] for more details.)

Theorem

Let $\mathcal{M}, \mathcal{N}$ be smooth manifolds and let $F : \mathcal{M} \to \mathcal{N}$ be a smooth map with constant rank ℓ. Then each level set

$$F^{-1}(y) := \{x \in \mathcal{M} : \; F(x) = y\}$$

is a closed embedded submanifold of codimension ℓ in $\mathcal{M}$.

Corollaries:

- If $F : \mathcal{M} \to \mathcal{N}$ is a submersion then each level is a closed embedded submanifold of codimension equal to the dimension of $\mathcal{N}$.
- In fact, by open submanifold lemma, only need to check full rank condition of submersion for points in the level set (replace $\mathcal{M}$ by the open set for which F has full rank).

The Stiefel manifold

For $m \geq n$, consider the set of all $m \times n$ matrices with orthonormal columns:

$$\mathrm{St}(m,n) := \{X \in \mathbb{R}^{m \times n} : X^T X = I_n\}.$$

Corollary

$\mathrm{St}(m,n)$ is an embedded submanifold of $\mathbb{R}^{m \times n}$.

Proof: Define $F : \mathbb{R}^{m \times n} \to \mathrm{symm}(n)$ as $F : X \mapsto X^T X$, where $\mathrm{symm}(n)$ denotes set of $n \times n$ symmetric matrices. At $X \in \mathrm{St}(m,n)$, consider Jacobian

$$DF(X) : H \mapsto X^T H + H^T X.$$

Given symmetric $Y \in \mathbb{R}^{n \times n}$, set $H = XY/2$. Then $DF(X)[H] = Y$; thus $DF(X)$ is surjective.

EFY. What is the dimension of the Stiefel manifold?

The manifold of rank-k matrices

Locality of definition of embedded submanifolds implies the following lemma (Lemma 5.5 in [Lee'2003]).

Lemma

Let $\mathcal{N}$ be subset of smooth manifold $\mathcal{M}$. Suppose every point $p \in \mathcal{N}$ has a neighborhood $\mathcal{U} \subset \mathcal{M}$ such that $\mathcal{U} \cap \mathcal{N}$ is an embedded submanifold of $\mathcal{U}$. Then $\mathcal{N}$ is an embedded submanifold of $\mathcal{M}$.

Theorem

Given $m \geq n$, the set

$$\mathcal{M}_k = \{A \in \mathbb{R}^{m \times n} : \mathrm{rank}(A) = k\}$$

is an embedded submanifold of $\mathbb{R}^{m \times n}$ for every $0 \leq k \leq n$.

The manifold of rank-k matrices

Choose arbitrary $A_0 \in \mathcal{M}_k$. After a suitable permutation, may assume w.l.o.g. that

$$A_0 = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \quad A_{11} \in \mathbb{R}^{k \times k} \text{ is invertible.}$$

This property remains true in an open neighborhood $U \subset \mathbb{R}^{m \times n}$ of A_0. Factorize $A \in U$ as

$$A = \begin{pmatrix} I & 0 \\ A_{21}A_{11}^{-1} & I \end{pmatrix} \begin{pmatrix} A_{11} & 0 \\ 0 & A_{22} - A_{21}A_{11}^{-1}A_{12} \end{pmatrix} \begin{pmatrix} I & A_{11}^{-1}A_{12} \\ 0 & I \end{pmatrix}.$$

Define $F : U \to \mathbb{R}^{(m-k)\times(n-k)}$ as $F : A \mapsto A_{22} - A_{21}A_{11}^{-1}A_{12}$. Then

$$F^{-1}(0) = U \cap \mathcal{M}_k.$$

The manifold of rank-k matrices

For arbitrary $Y \in \mathbb{R}^{(m-k)\times(n-k)}$, we obtain that

$$DF(A)\left[\begin{pmatrix} 0 & 0 \\ 0 & Y \end{pmatrix}\right] = Y.$$

Thus, F is a submersion. In turn, $\mathcal{U} \cap \mathcal{M}_k$ is an embedded submanifold of $\mathcal{U}$. By lemma, $\mathcal{M}_k$ is an embedded submanifold of $\mathbb{R}^{m\times n}$.

EFY. What is the dimension of $\mathcal{M}_k$?

EFY. Is $\mathcal{M}_k$ connected?

EFY. Prove that the set of symmetric rank-k matrices is an embedded submanifold of $\mathbb{R}^{n\times n}$. Is this manifold connected?

Tangent space

In the following, much of the discussion restricted to submanifolds $\mathcal{M}$ embedded in vector space V with inner product $\langle \cdot, \cdot \rangle$ and induced norm $\| \cdot \|$.

Given smooth curve $\gamma : \mathbb{R} \to \mathcal{M}$ with $x = \gamma(0)$, we call $\gamma'(0) \in V$ a tangent vector at x. The tangent space $T_x\mathcal{M} \subset V$ is the set of all tangent vectors at x.

Lemma

$T_x\mathcal{M}$ is a subspace of V.

Proof. If v_1, v_2 are tangent vectors then there are smooth curves γ_1, γ_2 such that $\gamma_1'(0) = v_1$, $\gamma_2'(0) = v_2$. To show that $\alpha v_1 + \beta v_2$ for $\alpha, \beta \in \mathbb{R}$ is again a tangent vector, consider chart $(\mathcal{U}, \varphi)$ around x such that $\varphi(x) = 0$. Define

$$\gamma(t) = \varphi^{-1}(\alpha\varphi(\gamma_1(t)) + \beta\varphi(\gamma_2(t)))$$

for t sufficiently close to 0. Then $\gamma(0) = x$ and $\gamma'(0) = \alpha v_1 + \beta v_2$.

EFY. Prove that the dimension of $T_x\mathcal{M}$ equals the dimension of $\mathcal{M}$ using a coordinate chart.

Tangent space

Application of definition to Stiefel manifold. Let

$$\gamma(t) = X + tY + \mathcal{O}(t^2)$$

be a smooth curve with $X \in \mathrm{St}(m, n)$. To ensure that $\gamma(t) \in \mathrm{St}(m, n)$, we require

$$I_n = \gamma(t)^T \gamma(t) = (X+tY)^T(X+tY)+\mathcal{O}(t^2) = I_n+t(X^TY+Y^TX)+\mathcal{O}(t^2).$$

Thus, $X^TY + Y^TX = 0$ characterizes tangent space:

$$\begin{aligned} T_x\mathrm{St}(m,n) &= \{Y \in \mathbb{R}^{m\times n} : X^TY = -Y^TX\} \\ &= \{XW + X_\perp W_\perp : W \in \mathbb{R}^{n\times n}, W = -W^T, W_\perp \in \mathbb{R}^{(m-n)\times n}\} \end{aligned}$$

where the columns of $X_\perp$ form basis of $\mathrm{span}(X)^\perp$

Tangent space

When $\mathcal{M}$ is defined (at least locally) as level set of constant rank function $F : V \to \mathbb{R}^N$, we have

$$T_x\mathcal{M} = \ker(DF(x)).$$

Proof. Let $v \in T_x\mathcal{M}$, that is, there is a curve $\gamma : \mathbb{R} \to \mathcal{M}$ such that $\gamma(0) = x$ and $\gamma'(0) = v$. Then, by chain rule,

$$DF(x)[v] = DF(x)[\gamma'(0)] = \frac{\partial}{\partial t} F(\gamma(t))\Big|_{t=0} = 0,$$

because F is constant on $\mathcal{M}$. Thus, $T_x\mathcal{M} \subset \ker(DF(x))$, which completes the proof by counting dimensions.

Tangent space of $\mathcal{M}_k$

Recall that $\mathcal{M}_k$ was obtained as level set of local submersion

$$F : A \mapsto A_{22} - A_{21} A_{11}^{-1} A_{12}.$$

Given $A \in \mathcal{M}_k$ consider SVD

$$A = \begin{pmatrix} U & U_\perp \end{pmatrix} \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} V & V_\perp \end{pmatrix}^T.$$

We have

$$DF \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix} [H] = H_{22}.$$

Thus, H is in the kernel if and only if $H_{22} = 0$. In terms of A this implies

$$\begin{aligned} T_A \mathcal{M}_k &= \ker(DF(A)) = \begin{pmatrix} U_k & U_\perp \end{pmatrix} \begin{pmatrix} \mathbb{R}^{k \times k} & \mathbb{R}^{k \times (n-k)} \\ \mathbb{R}^{(m-k) \times k} & 0 \end{pmatrix} \begin{pmatrix} V_k & V_\perp \end{pmatrix}^T \\ &= \{ UMV^T + U_p V^T + UV_p^T : M \in \mathbb{R}^{k \times k}, U_p^T U = V_p^T V = 0 \}. \end{aligned}$$

EFY. Compute the tangent space for the embedded submanifold of rank-k symmetric matrices.

Riemannian manifold and gradient

For submanifold $\mathcal{M}$ embedded in vector space V: Inner product $\langle \cdot, \cdot \rangle$ on V induces inner product on $T_x\mathcal{M}$. This turns $\mathcal{M}$ into a Riemannian manifold.[1]

The (Riemannian) gradient of smooth $f : \mathcal{M} \to \mathbb{R}$ at $x \in \mathcal{M}$ is defined as the unique element $\operatorname{grad} f(x) \in T_x\mathcal{M}$ that satisfies

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

EFY. Prove that the Riemannian gradient satisfies the steepest ascent property

$$\frac{\operatorname{grad} f(x)}{\|\operatorname{grad} f(x)\|_2} = \underset{\substack{\xi \in T_x\mathcal{M} \\ \|\xi\|=1}}{\arg\max}\; Df(x)[\xi].$$

[1] In general, for a Riemannian manifold one needs to have an inner product on $T_x\mathcal{M}$ that varies smoothly wrt x.

Riemannian gradient

For submanifold $\mathcal{M}$ embedded in vector space V: The (Euclidean) gradient of f in V admits the decomposition

$$\nabla f(x) = P_x \nabla f(x) + P_x^\perp \nabla f(x),$$

where $P_x, P_x^\perp$ are the orthogonal projections onto $T_x\mathcal{M}$, $T_x^\perp\mathcal{M}$. For every $\xi \in T_x\mathcal{M}$ we have

$$\begin{aligned} \langle P_x \nabla f(x), \xi \rangle &= \langle \nabla f(x) - P_x^\perp \nabla f(x), \xi \rangle \\ &= \langle \nabla f(x), \xi \rangle = Df(x)[\xi]. \end{aligned}$$

Hence,

$$\operatorname{grad} f(x) = P_x \nabla f(x).$$

The Riemannian gradient is the orthogonal projection of the Euclidean gradient onto the tangent space.

Riemannian gradient

Example: Given symmetric $n \times n$ matrix A, consider trace optimization problems

$$\min_{X \in \mathrm{St}(n,k)} \operatorname{trace}(X^T A X)$$

Study first-order perturbation

$$\begin{aligned} & \operatorname{trace}((X+H)^T A (X+H)) - \operatorname{trace}(X^T A X) \\ = \; & \operatorname{trace}(H^T A X) + \operatorname{trace}(X^T A H) + \mathcal{O}(\|H\|^2) \\ = \; & 2\langle H, AX \rangle + \mathcal{O}(\|H\|^2). \end{aligned}$$

$\rightsquigarrow$ Euclidean gradient at X given by $2AX$.

Note that $\operatorname{skew}(W) = (W - W^T)/2$ is orth projection on skew-symmetric matrices. Thus,

$$P_X(Z) = (I - XX^T)Z + X \cdot \operatorname{skew}(X^T Z).$$

$$\begin{aligned} \operatorname{grad} f(X) &= P_X(\nabla f(X)) = 2(I - XX^T)AX + 2X \cdot \operatorname{skew}(X^T A X) \\ &= 2(AX - X\,X^T AX). \end{aligned}$$

Riemannian gradient

Example: For $A \in \mathcal{M}_k$ consider SVD $A = U\Sigma V^T$ with $\Sigma \in \mathbb{R}^{k\times k}$. Define orthogonal projections onto span(U), span(V), and their complements:

$$P_U = UU^T,\ P_U^{\perp} = I - UU^T,\ P_V = VV^T,\ P_V^{\perp} = I - VV^T.$$

Recall that

$$T_A\mathcal{M}_k = \{UMV^T + U_pV^T + UV_p^T : M \in \mathbb{R}^{k\times k}, U_p^T U = V_p^T V = 0\}$$

The three terms of the sum are orthogonal to each other and can thus be considered separately $\rightsquigarrow$ Orthogonal projection onto $T_A\mathcal{M}_k$ given by

$$P_A(Z) = P_U Z P_V + P_U^{\perp} Z P_V + P_V^{\perp} Z P_U$$

EFY. Compute the Riemannian gradient of $f(A) = \|A - B\|_F^2$ on $\mathcal{M}_k$ for given $B \in \mathbb{R}^{m\times n}$.

Line search: Concepts from Euclidean case

$$\min_{x \in \mathbb{R}^N} f(x),$$

Line search is optimization algorithm of the form

$$x_{j+1} = x_j + \alpha_k \eta_j$$

with search direction η_j and step size $\alpha_j > 0$.

- *First-order* optimal choice of η_j: $\eta_j = -\nabla f(x_j) \rightsquigarrow$ gradient descent.
  Motivation for other choices: Faster local convergence (Newton-type methods), exact gradient computation too expensive, ...
  Gradient-related search directions: $\langle \eta_j, \nabla f(x_j) \rangle < \delta < 0$ for all j.

Line search: Concepts from Euclidean case

- Exact line search chooses

$$\alpha_j = \arg\min_{\alpha} f(x_j + \alpha_j \eta_j).$$

Only in exceptional cases simple optimization problem, e.g., admitting closed form solution.

EFY. Derive the closed form solution for exact line search applied to

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} x^T A x + b^T x$$

for symmetric positive $A \in \mathbb{R}^{n \times n}$ and $b \in \mathbb{R}^n$.

Alternative: Armijo rule. Let $\beta \in (0, 1)$ (typically $\beta = 1/2$) and $c \in (0, 1)$ (e.g., $c = 10^{-4}$) be fixed parameters. Determine largest $\alpha_j \in \{1, \beta, \beta^2, \beta^3, \ldots\}$ such that

$$f(x_j + \alpha_j \eta_j) - f(x_j) \leq c \alpha_j \nabla f(x_j)^T \eta_j$$

holds. (Such α_j always exists provided that η_j is descent direction, i.e., when $\langle \eta_j, \nabla f(x_j) \rangle < 0$.)

More details in [J. Nocedal and S. J. Wright. Numerical optimization. Second edition. Springer Series in Operations Research and Financial Engineering. Springer, 2006].

Line search: Extension to manifolds

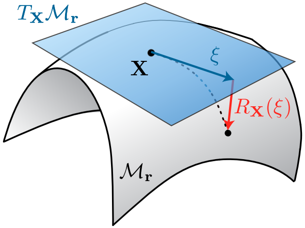
$$\min_{x \in \mathcal{M}} f(x)$$

Cannot use line search $x_{k+1} = x_j + \alpha_j \eta_j$, simply because addition is not well defined in $\mathcal{M}$.

Idea:

Search along smooth curve $\gamma(\alpha) \in \mathcal{M}$ with $\gamma(0) = x_j$ and $\gamma'(0) = \eta_j \in T_{x_j}\mathcal{M}$.

Step in direction $x_j + \alpha\eta_j \in x_j + T_{x_j}\mathcal{M}$ and go back to manifold via retraction:



Retraction

Tangent bundle $T\mathcal{M} := \bigcup_{x \in \mathcal{M}} T_x\mathcal{M}$

Definition

Mapping $R : T\mathcal{M} \to \mathcal{M}$ is called a retraction on $\mathcal{M}$ if for every $X_0 \in \mathcal{M}$ there exists neighborhood $\mathcal{U}$ around $(X_0, 0) \in T\mathcal{M}$ such that:

1. $\mathcal{U} \subset \mathrm{dom}(R)$ and $R|_{\mathcal{U}} : \mathcal{U} \to \mathcal{M}$ is smooth.
2. $R(x, 0) = x$ for all $(x, 0) \in \mathcal{U}$.
3. $DR(x, 0)[0, \xi] = \xi$ for all $(x, \xi) \in \mathcal{U}$.

Will write $R_x = R(x, \cdot) : T_x\mathcal{M} \to \mathcal{M}$ in the following.

Intuition behind definition:
Property 2 $=$ retraction does nothing to elements on manifold.
Property 3 $=$ retraction preserves direction of curves. Equivalent characterization: For every tangent vector $\xi \in T_x\mathcal{M}$, the curve $\gamma : \alpha \mapsto R_x(\alpha\xi)$ satisfies $\gamma'(0) = \xi$.

EFY. What is a retraction for the manifold of invertible $n \times n$ matrices (trick question)?

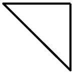
Retraction

Exponential maps are most natural choice of retraction from theoretical point of view but often too expensive/too cumbersome to compute.

In practice for matrix manifolds: Retractions are often built from matrix decompositions and metric projections.

Example $\mathrm{St}(n, k)$: Given $Y \in \mathbb{R}_*^{n \times k}$ (i.e., $\mathrm{rank}(Y) = k$), the economy-sized QR decomposition

$$Y = XR, \quad X^T X = I_k, \quad R =$$



is unique provided that diagonal elements of R are positive. This defines a diffeomorphism

$$\phi : \mathrm{St}(n, k) \times \mathrm{triu}_+(k) \to \mathbb{R}_*^{n \times k}, \quad \phi : (X, R) \mapsto XR,$$

where $\mathrm{triu}_+(k)$ denotes upper triangular matrix with positive diagonal elements. Applying ϕ^{-1} just means computing the QR decomposition. Note that

$$\dim \mathrm{St}(n, k) + \dim \mathrm{triu}_+(k) = \dim \mathbb{R}_*^{n \times k}.$$

Retraction

Abstract setting: Let $\mathcal{M}$ be embedded submanifold of vector space V and $\mathcal{N}$ smooth manifold such that

$$\dim(\mathcal{M}) + \dim(\mathcal{N}) = \dim(V).$$

Assume there is diffeomorphism

$$\phi : \mathcal{M} \times \mathcal{N} \to V_* : (x, y) \mapsto \phi(x, y)$$

for some open subset V_* of V. Moreover, assume $\exists$ neutral element $\mathrm{id} \in \mathcal{N}$ such that $\phi(x, \mathrm{id}) = x$ for all $x \in \mathcal{M}$.

Lemma

Under above assumptions,

$$R_x(\eta) := \pi_1(\phi^{-1}(x + \eta))$$

is a retraction on $\mathcal{M}$, where π_1 is projection onto first component: $\pi_1(x, y) = x$.

Retraction

Proof of lemma. Need to verify three properties of retraction.

Property 1: Immediately follows from assumptions that $R_x(\xi)$ is defined and smooth for all ξ in a neighborhood of $0 \in T_x\mathcal{M}$.

Property 2: $R_x(0) := \pi_1(\phi^{-1}(x)) = \pi_1(x, \mathrm{id}) = x$.

Property 3: Differentiating $x = \pi_1 \circ \phi^{-1}(\phi(x, \mathrm{id}))$ we obtain for any $\xi \in T_x\mathcal{M}$ that

$$\begin{aligned} \xi &= D(\pi_1 \circ \phi^{-1})[D\phi(x, \mathrm{id})[\xi, 0]] \\ &= D(\pi_1 \circ \phi^{-1})(x)[\xi] = DR_x(0)[\xi]. \end{aligned}$$

$\square$

Retraction

For $z \in V$ sufficiently close to $\mathcal{M}$, metric projection is well defined:

$$P_{\mathcal{M}}(z) := \arg\min_{x \in \mathcal{M}} \|z - x\|.$$

Corollary (Lewis/Malick'2008)

The map

$$R_x(\eta) := P_{\mathcal{M}}(x + \eta)$$

defines a retraction.

Examples for retractions based on metric projection:

- For $\mathrm{St}(n, k)$, polar factor $Y(Y^T Y)^{-1/2}$ of $Y \in \mathbb{R}_*^{n \times k}$ defines a retraction.
- For rank-k matrix manifold $\mathcal{M}_k$, best rank-k approximation $\mathcal{T}_k$ defines a retraction.
  There are other choices; see [Absil/Oseledets'2015: Low-rank retractions: a survey and new results].

EFY. For all examples discussed so far, develop algorithms that efficiently realize the retraction by exploiting the structure of $x + \eta$.

EFY. Find a retraction for the manifold of symmetric rank-k matrices.

Riemannian line search

$$x_{j+1} = R_{x_j}(\alpha_j \eta_j).$$

Assumption. Sequence $\{\eta_j\}$ is bounded and gradient related:

$$\limsup_{k\to\infty} \langle \operatorname{grad} f(x_j), \eta_j \rangle < 0.$$

Canonical choice: $\eta_j = -\operatorname{grad} f(x_j)$.

Extension of Armijo rule. Let $\beta \in (0,1)$ and $c \in (0,1)$ (e.g., $c = 10^{-4}$) be fixed parameters. Determine largest $\alpha_j \in \{1, \beta, \beta^2, \beta^3, \ldots\}$ such that

$$f(R_{x_j}(\alpha_j \eta_j)) - f(x_j) \leq c\alpha_j \langle \operatorname{grad} f(x_j), \eta_j \rangle \tag{1}$$

holds.

EFY. Show that the Armijo condition (1) can always be satisfied for sufficiently small α_j.

Riemannian line search

1: **for** j = 0,1,2,... **do**
2: Pick $\eta_j \in T_{x_j}\mathcal{M}$ such that sequence $\{\eta_j\}$ is gradient-related.
3: Choose $\alpha_j \in \{1, \beta, \beta^2, \beta^3, \ldots\}$ such that Armijo condition is satisfied.
4: Set $x_{j+1} = R_{x_j}(\alpha_j \eta_j)$.
5: **end for**

Convergence theory in Section 4.3 of [Absil'2008].

We call $x_* \in \mathcal{M}$ a critical point of f if $\operatorname{grad} f(x_*) = 0$.

Theorem

Every accumulation point of $\{x_j\}$ is a critical point of cost function f.

More can be said if manifold (or at least level set) is compact.

Corollary

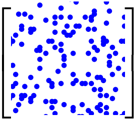
Assume that $\mathcal{L} = \{x \in \mathcal{M} : f(x) \leq f(x_0)\}$ is compact. Then $\lim_{j\to\infty} \|\operatorname{grad} f(x_j)\| \to 0$.

Note that $\mathcal{M}_k$ is *not* compact and it is not clear a priori whether $\mathcal{L}$ is compact..

Application to low-rank matrix and tensor completion

Matrix Completion

$$\mathrm{P}_\Omega A =$$



$$\overset{\text{recover?}}{\rightsquigarrow} \; A$$

$$\mathrm{P}_\Omega : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}, \quad \mathrm{P}_\Omega X = \begin{cases} X_{ij} & \text{if } (i,j) \in \Omega, \\ 0 & \text{else.} \end{cases}$$

Applications: image reconstruction, image inpainting, Netflix problem

Low-rank matrix completion:

$$\begin{aligned} \min_X \quad & \operatorname{rank}(X)\,, \qquad X \in \mathbb{R}^{m\times n} \\ \text{subject to} \quad & \mathrm{P}_\Omega X = \mathrm{P}_\Omega A \end{aligned}$$

Low-rank matrix completion: *($\rightsquigarrow$ NP-Hard)*

$$\begin{aligned} \min_X \quad & \operatorname{rank}(X)\,, \qquad X \in \mathbb{R}^{m\times n} \\ \text{subject to} \quad & \mathrm{P}_\Omega X = \mathrm{P}_\Omega A \end{aligned}$$

Nuclear norm relaxation: *($\rightsquigarrow$ convex, but expensive)*

$$\begin{aligned} \min_X \quad & \|X\|_* = \sum_i \sigma_i\,, \qquad X \in \mathbb{R}^{m\times n} \\ \text{subject to} \quad & \mathrm{P}_\Omega X = \mathrm{P}_\Omega A \end{aligned}$$

Robust low-rank completion: *(Assume rank is known)*

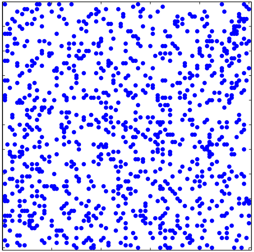
$$\begin{aligned} \min_X \quad & \frac{1}{2}\|\,\mathrm{P}_\Omega X - \mathrm{P}_\Omega A\|_F^2\,, \qquad X \in \mathbb{R}^{m\times n} \\ \text{subject to} \quad & \operatorname{rank}(X) = k \end{aligned}$$

Huge body of work! Overview: http://perception.csl.illinois.edu/matrix-rank/

Setting

$$\begin{array}{ll} \underset{X}{\text{minimize}} & f(X) := \frac{1}{2}\|P_\Omega(X - A)\|_F^2 \\ \text{subject to} & X \in \mathcal{M}_k := \left\{X \in \mathbb{R}^{m\times n} : \operatorname{rank}(X) = k\right\} \end{array}$$

$$\begin{aligned} P_\Omega : \quad & \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n} \\ & X_{ij} \mapsto \begin{cases} X_{ij} & \text{if } (i,j) \in \Omega, \\ 0 & \text{if } (i,j) \notin \Omega. \end{cases} \end{aligned}$$



Riemannian gradient given by:

$$\operatorname{grad} f(X) = P_{T_X\mathcal{M}_k}\big(P_\Omega(X - A)\big)$$

with orthogonal projection $P_{T_X\mathcal{M}_k} : \mathbb{R}^{m\times n} \to T_X\mathcal{M}_k$.

Geometric nonlinear CG for matrix completion

Input: Initial guess $X_0 \in \mathcal{M}_k$.

$\eta_0 \leftarrow -\mathrm{grad}\, f(X_0)$

$\alpha_0 \leftarrow \mathrm{argmin}_\alpha\, f(X_0 + \alpha\eta_0)$

$X_1 \leftarrow R_{X_0}(\alpha_0\eta_0)$

for $i = 1, 2, \ldots$ **do**

Compute gradient:

$\xi_i \leftarrow \mathrm{grad}\, f(X_i)$

Conjugate direction by PR+ updating rule:

$\eta_i \leftarrow -\xi_i + \beta_i \mathcal{T}_{X_{i-1} \to X_i} f(\eta_{i-1})$

Initial step size from linearized line search:

$\alpha_i \leftarrow \mathrm{argmin}_\alpha\, f(X_i + \alpha\eta_i)$

Armijo backtracking for sufficient decrease:

Find smallest integer $m \geq 0$ such that

$f(X_i) - f(R_{X_i}(2^{-m}\alpha_i\eta_i)) \geq -1 \cdot 10^{-4} \langle \xi_i, 2^{-m}\alpha_i\eta_i \rangle$

Obtain next iterate:

$X_{i+1} \leftarrow R_{X_i}(2^{-m}\alpha_i\eta_i)$

end for

Cost/iteration: $O((m+n)k^2 + |\Omega|k)$ ops.

Vector transport

Conjugate gradient method requires combination of gradients for subsequent iterates:

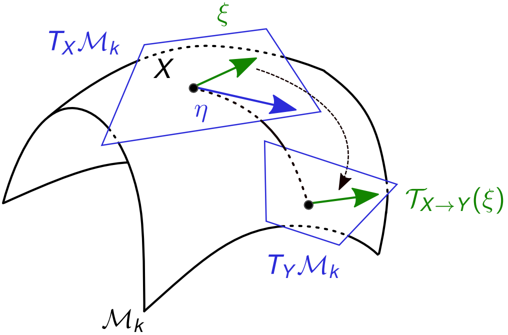
$$\operatorname{grad} f(X) \in T_{X}\mathcal{M}_k, \quad \operatorname{grad} f(Y) \in T_{Y}\mathcal{M}_k$$

$$\Rightarrow \quad \operatorname{grad} f(X) + \operatorname{grad} f(Y) \text{ ???}$$

Can be addressed by vector transport:

$$\mathcal{T}_{X \to Y} : T_X\mathcal{M}_k \to T_Y\mathcal{M}_k$$

$$\mathcal{T}_{X \to Y}(\xi) = P_{T_Y\mathcal{M}_k}(\xi).$$

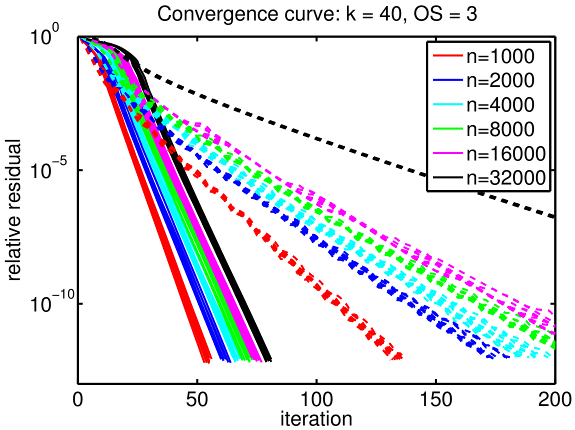


Can be implemented in $O((m+n)k^2)$ ops.

Numerical experiments

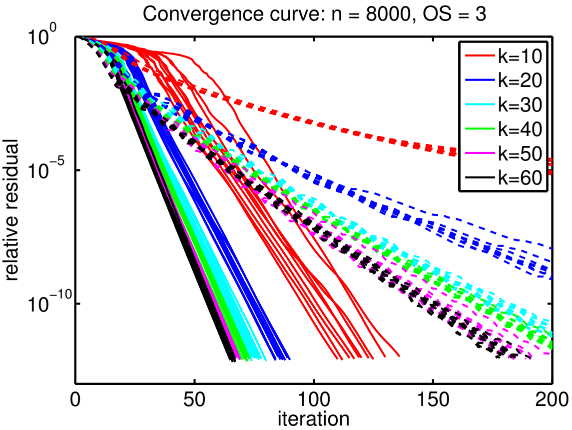
- Comparison to LMAFit [Wen/Yin/Zhang'2010]. `http://lmafit.blogs.rice.edu/`.
- Oversampling factor $\mathrm{OS} = |\Omega|/(k(2n-k))$.
- Purely academic example $A = A_L A_R^T$ with $A_L, A_R = \texttt{randn}$.

Influence of n



- Dashed lines: LMAFit. Solid lines: Nonlinear CG.
- time(1 iteration of Nonlinear CG)
  $\approx\ 2\times$ time(1 iteration of LMAFit)

Influence of rank



- Dashed lines: LMAFit. Solid lines: Nonlinear CG.
- time(1 iteration of Nonlinear CG)
  $\approx$ 2$\times$ time(1 iteration of LMAFit)

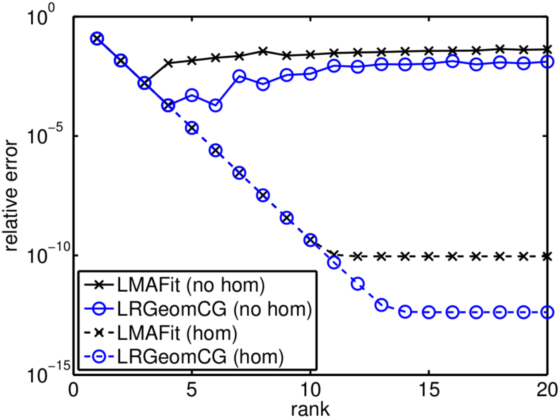
Numerical experiments

- Comparison to LMAFit [Wen/Yin/Zhang'2010]. `http://lmafit.blogs.rice.edu/`.
- Oversampling factor $\mathrm{OS} = |\Omega|/(k(2n-k)) = 8$.
- $8\,000 \times 8\,000$ matrix A is obtained from evaluating

$$f(x,y) = \frac{1}{1+|x-y|^2}$$

on $[0,1] \times [0,1]$.

Influence of rank

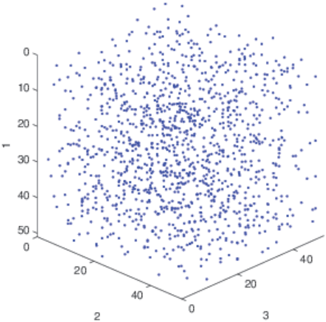


- Hom: Start with $k = 1$ and subsequently increase k, using previous result as initial guess.

Tensor Completion

Low-rank tensor completion:

$$\min_{\mathcal{X}} \quad \text{rank}(\mathcal{X})\,, \qquad \mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times \ldots \times n_d}$$
$$\text{subject to} \quad \mathrm{P}_\Omega\, \mathcal{X} = \mathrm{P}_\Omega\, \mathcal{A}$$



Applications:

- Completion of multidimensional data, e.g. hyperspectral images, CT Scans
- Compression of multivariate functions with singularities
- ...

Manifold of Tensors of fixed multilinear rank

$$\mathcal{M}_{\mathbf{k}} := \left\{\mathcal{X} \in \mathbb{R}^{n_1 \times \cdots \times n_d} : \text{rank}(\mathcal{X}) = \mathbf{k}\right\},$$

$$\dim(\mathcal{M}_{\mathbf{k}}) = \prod_{j=1}^{d} k_j + \sum_{i=1}^{d} \left(k_i n_i - \frac{k_i(k_i-1)}{2}\right).$$

- $\mathcal{M}_{\mathbf{k}}$ is a smooth manifold. Discussed for more general formats in [Holtz/Rohwedder/Schneider'2012], [Uschmajew/Vandereycken'2012]
- Riemannian with metric induced by standard inner product $\langle \mathcal{X}, \mathcal{Y} \rangle = \langle \mathcal{X}_{(1)}, \mathcal{Y}_{(1)} \rangle$ *(sum of element-wise product)*

Manifold structure used in

- dynamical low-rank approximation [Koch/Lubich'2010], [Arnold/Jahnke'2012], [Lubich/Rohwedder/Schneider/Vandereycken'2012], [Khoromskij/Oseledets/Schneider'2012], ...
- best multilinear approximation [Eldén/Savas'2009], [Ishteva/Absil/Van Huffel/De Lathauwer'2011], [Curtef/Dirr/Helmke'2012]

Gradients and Tangent Space $T_{\mathcal{X}}\mathcal{M}_{\mathbf{k}}$

Every ξ in the tangent space $T_{\mathcal{X}}\mathcal{M}_{\mathbf{k}}$ at $\mathcal{X} = \mathcal{C} \times_1 U \times_2 V \times_3 W$ can be written as:

$$\begin{aligned}\xi = \mathcal{S} &\times_1 U \ \times_2 V \ \times_3 W \\ + \mathcal{C} &\times_1 U_\perp \times_2 V \ \times_3 W \\ + \mathcal{C} &\times_1 U \ \times_2 V_\perp \times_3 W \\ + \mathcal{C} &\times_1 U \ \times_2 V \ \times_3 W_\perp\end{aligned}$$

for some $\mathcal{S} \in \mathbb{R}^{k_1 \times k_2 \times k_3}$, $U_\perp \in \mathbb{R}^{n_1 \times k_1}$ with $U_\perp^T U = 0, \ldots$
Again, we obtain the Riemannian gradient of the objective function

$$f(\mathcal{X}) := \frac{1}{2} \| \mathrm{P}_\Omega \mathcal{X} - \mathrm{P}_\Omega \mathcal{A} \|_F^2$$

by projecting the Euclidean gradient into the tangent space:

$$\operatorname{grad} f(\mathcal{X}) = \mathrm{P}_{T_{\mathcal{X}}\mathcal{M}_{\mathbf{k}}}(\mathrm{P}_\Omega \mathcal{X} - \mathrm{P}_\Omega \mathcal{A})$$

Retraction

Candidate for retraction: Metric projection

$$R_{\mathcal{X}}(\xi) = P_{\mathcal{X}}(\mathcal{X} + \xi) = \operatorname*{arg\,min}_{\mathcal{Z} \in \mathcal{M}_{\mathbf{k}}} \|\mathcal{X} + \xi - \mathcal{Z}\|.$$

No closed-form solution available

- Replaced by HOSVD truncation.
- Seems to work fine.
- HOSVD truncation is a retraction [K./Steinlechner/Vandereycken'13].

Geometric Nonlinear CG for Tensor Completion

Input: Initial guess $\mathcal{X}_0 \in \mathcal{M}_{\mathbf{k}}$.

$\eta_0 \leftarrow -\mathrm{grad}\, f(\mathcal{X}_0)$

$\alpha_0 \leftarrow \mathrm{argmin}_\alpha\, f(\mathcal{X}_0 + \alpha\eta_0)$

$\mathcal{X}_1 \leftarrow R_{\mathcal{X}_0}(\alpha_0\eta_0)$

for $i = 1, 2, \ldots$ **do**

Compute gradient:

$\xi_i \leftarrow \mathrm{grad}\, f(\mathcal{X}_i)$

Conjugate direction by PR+ updating rule:

$\eta_i \leftarrow -\xi_i + \beta_i \mathcal{T}_{\mathcal{X}_{i-1}\to\mathcal{X}_i} f(\eta_{i-1})$

Initial step size from linearized line search:

$\alpha_i \leftarrow \mathrm{argmin}_\alpha\, f(\mathcal{X}_i + \alpha\eta_i)$

Armijo backtracking for sufficient decrease:

Find smallest integer $m \geq 0$ such that

$f(\mathcal{X}_i) - f(R_{\mathcal{X}_i}(2^{-m}\alpha_i\eta_i)) \geq -1 \cdot 10^{-4}\langle \xi_i, 2^{-m}\alpha_i\eta_i\rangle$

Obtain next iterate:

$\mathcal{X}_{i+1} \leftarrow R_{\mathcal{X}_i}(2^{-m}\alpha_i\eta_i)$

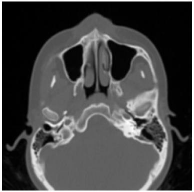
end for

Cost/iteration: $O(nk^d + |\Omega|k^{d-1})$ ops.

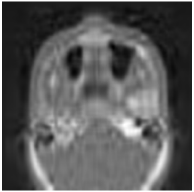
Reconstruction of CT Scan

199 $\times$ 199 $\times$ 150 tensor from scaled CT data set “INCISIX”,
(taken from OSIRIX MRI/CT data base
[www.osirix-viewer.com/datasets/])

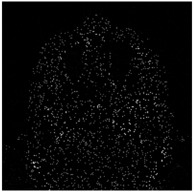
Slice of original tensor



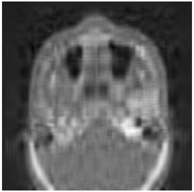
HOSVD approx. of rank 21



Sampled tensor (6.7%)



Low-rank completion of rank 21



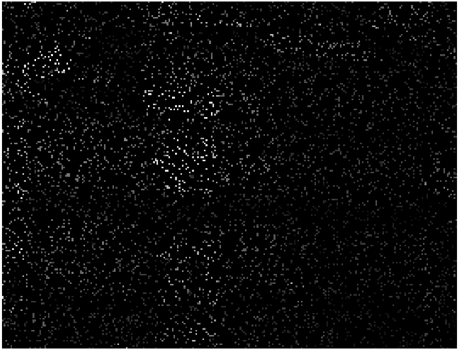
Compares very well with existing results w.r.t. low-rank recovery and speed, e.g., [Gandy/Recht/Yamada/’2011].

Hyperspectral Image

Set of photographs, (204 $\times$ 268 px) taken across a large range of wavelengths. 33 samples from ultraviolet to infrared [Image data: Foster et al.'2004]

Stacked into a tensor of size 204 $\times$ 268 $\times$ 33

10% of the Original Hyperspectral Imega Tensor, 16th Slice
Size of Tensor is [204, 268, 33]



Completed Tensor, 16th Slice
Final Rank is k = [50 50 6]

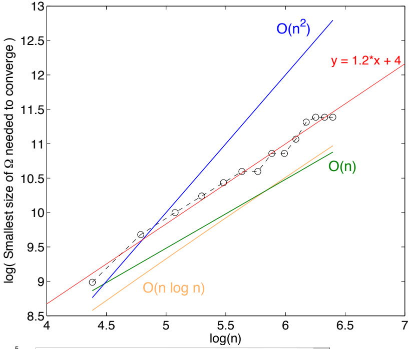
Here: Only 10% of entries known; [Signoretti et al.'2011] use 50%.

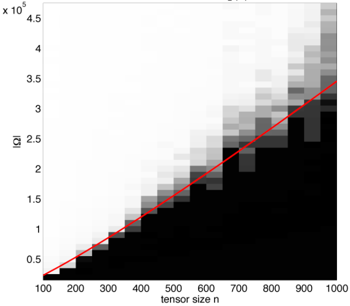
How many samples do we need?

Matrix case:
$O(n \cdot \log^{\beta} n)$ samples suffice!
[Candès/Tao'2009]
$\Rightarrow$ *Completion of tensor by applying matrix completion to matricization:* $O(n^2 \log(n))$. Gives upper bound!

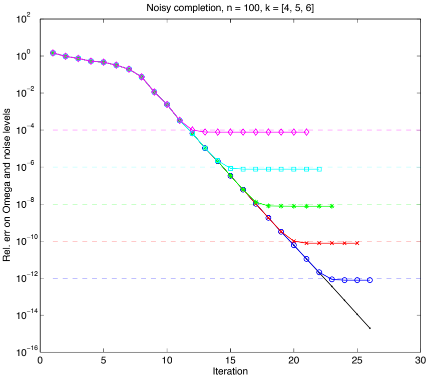
Tensor case:
Certainly: $|\Omega| \ll O(n^2)$
In all cases of convergence
$\rightsquigarrow$ exact reconstruction.

Conjecture: $|\Omega| = O(n \cdot \log^{\beta} n)$





Robustness of Convergence



- Random $100 \times 100 \times 100$ tensor of multilinear rank $(4, 5, 6)$ perturbed by white noise.
- Upon convergence $\rightsquigarrow$ reconstruction up to noise level.

Final remarks on Riemannian low-rank optimization

- Only discussed first-order methods. Fine for well-conditioned problems but slow convergence for ill-conditioned problems.
- Second-order methods (Newton-like) require Riemannian Hessian: painful and:
  - not of much help for well-conditioned problems (low-rank matrix completion).
  - linearized equations hard to solve efficiently for low-rank matrix and tensor manifolds.
- Low-rank matrices/tensors can also be viewed as products of quotient manifolds. Requires careful choice of metric to stay robust wrt small singular value σ_k [Ngo/Saad'2012], [Kasai/Mishra, ICML'2016].
- Lots of open problems concerning convergence analysis of low-rank Riemannian optimization!