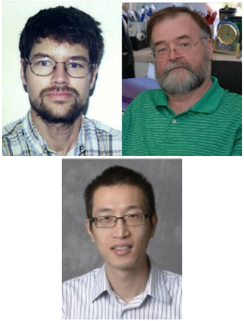
Riemannian Optimization and its Application to Phase Retrieval Problem

Wen Huang

Université catholique de Louvain

Joint work with:

- Pierre-Antoine Absil, Professor of Mathematical Engineering, *Université catholique de Louvain*
- Kyle A. Gallivan, Professor of Mathematics, *Florida State University*
- Xiangxiong Zhang, Assistant Professor, *Purdue University*

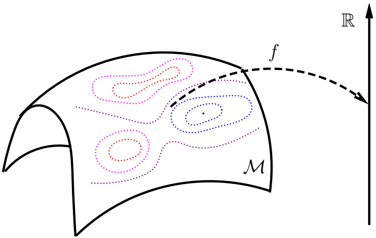


Riemannian Optimization

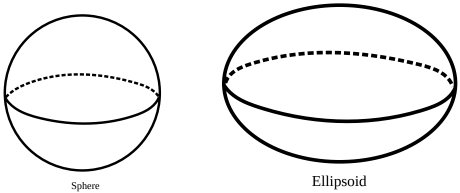
Problem: Given $f(x) : \mathcal{M} \to \mathbb{R}$, solve

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

where $\mathcal{M}$ is a Riemannian manifold.



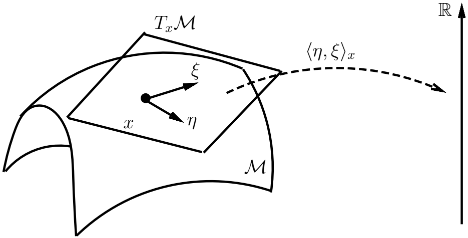
Examples of Manifolds



- Stiefel manifold: $\mathrm{St}(p, n) = \{X \in \mathbb{R}^{n \times p} | X^T X = I_p\}$
- Grassmann manifold: Set of all p-dimensional subspaces of $\mathbb{R}^n$
- Set of fixed rank m-by-n matrices
- And many more

Riemannian Manifolds

Roughly, a Riemannian manifold $\mathcal{M}$ is a smooth set with a smoothly-varying inner product on the tangent spaces.

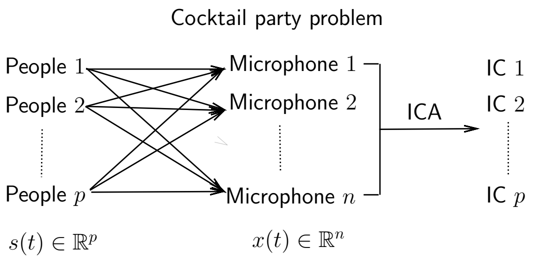


Applications

Three applications are used to demonstrate the importances of the Riemannian optimization:

- Independent component analysis [CS93]
- Matrix completion problem [Van12]
- Phase retrieval problem [CSV13, EM13]

Application: Independent Component Analysis



- Observed signal is $x(t) = As(t)$
- One approach:
  - Assumption: $E\{s(t)s(t+\tau)\}$ is diagonal for all τ
  - $C_\tau(x) := E\{x(t)x(x+\tau)^T\} = AE\{s(t)s(t+\tau)^T\}A^T$

Application: Independent Component Analysis

- Minimize joint diagonalization cost function on the Stiefel manifold [TI06]:

$$f : \mathrm{St}(p, n) \to \mathbb{R} : V \mapsto \sum_{i=1}^{N} \|V^T C_i V - \mathrm{diag}(V^T C_i V)\|_F^2.$$

- $C_1, \dots, C_N$ are covariance matrices and $\mathrm{St}(p, n) = \{X \in \mathbb{R}^{n \times p} | X^T X = I_p\}$.

Application: Matrix Completion Problem

Matrix completion problem

	Movie 1	Movie 2				Movie n
User 1		1			4	
User 2	3	5			4	
			5	1		
User m		2			5	3

Rate matrix M

- The matrix M is sparse
- The goal: complete the matrix M

Application: Matrix Completion Problem

$$
\begin{array}{ccc}
\text{movies} & \text{meta-user} & \text{meta-movie} \\
\begin{pmatrix} a_{11} & & & a_{14} \\ & & & a_{24} \\ & & a_{33} & \\ a_{41} & & & \\ & a_{52} & a_{53} & \end{pmatrix} = & \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \\ b_{41} & b_{42} \\ b_{51} & b_{52} \end{pmatrix} & \begin{pmatrix} c_{11} & c_{12} & c_{13} & c_{14} \\ c_{21} & c_{22} & c_{23} & c_{24} \end{pmatrix}
\end{array}
$$

- Minimize the cost function

$$f : \mathbb{R}_r^{m \times n} \to \mathbb{R} : X \mapsto f(X) = \|P_\Omega M - P_\Omega X\|_F^2.$$

- $\mathbb{R}_r^{m \times n}$ is the set of m-by-n matrices with rank r. It is known to be a Riemannian manifold.

Application: Phase Retrieval Problem

- The Phase Retrieval problem concerns recovering a signal given the modulus of its linear transform, e.g., the Fourier transform.
- It is important in many applications, e.g., X-ray crystallography imaging [Har93];
- A cost function in the PhaseLift [CSV13] framework is:

$$\min_{X \geq 0} \|b^2 - \mathrm{diag}(ZXZ^*)\|_2^2 + \kappa \mathrm{trace}\left(X\right),$$

  where b is the measurements, Z is the linear operator, and κ is a positive constant.
- The desired minimizer has rank one.

Application: Phase Retrieval Problem

- This motivates us to consider the optimization problem

$$\min_{X \geq 0} H(X) \tag{1}$$

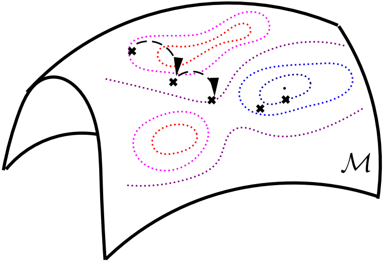
  and the desired minimizer has low rank.
- It is known that $\{X \in \mathbb{C}^{n \times n} | X \geq 0, \mathrm{rank}(X) \text{is fixed}\}$ is a manifold.
- Problem (1) can be solved by combining Riemannnian optimization with rank adaptive mechanism [JBAS10, ZHG+15]

More Applications

- Large-scale Generalized Symmetric Eigenvalue Problem and SVD
- Blind source separation on both Orthogonal group and Oblique manifold
- Low-rank approximate solution symmetric positive definite Lyapanov $AXM + MXA = C$
- Best low-rank approximation to a tensor
- Rotation synchronization
- Graph similarity and community detection
- Low rank approximation to role model problem

Comparison with Constrained Optimization

- All iterates on the manifold
- Convergence properties of unconstrained optimization algorithms
- No need to consider Lagrange multipliers or penalty functions
- Exploit the structure of the constrained set



Iterations on the Manifold

Consider the following generic update for an iterative Euclidean optimization algorithm:

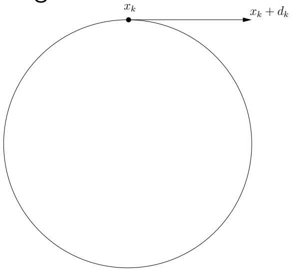
$$x_{k+1} = x_k + \Delta x_k = x_k + \alpha_k s_k \ .$$

This iteration is implemented in numerous ways, e.g.:

- Steepest descent: $x_{k+1} = x_k - \alpha_k \nabla f(x_k)$
- Newton's method: $x_{k+1} = x_k - \left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k)$
- Trust region method: Δx_k is set by optimizing a local model.

Riemannian Manifolds Provide

- Riemannian concepts describing directions and movement on the manifold
- Riemannian analogues for gradient and Hessian



Riemannian gradient and Riemannian Hessian

Definition

The Riemannian gradient of f at x is the unique tangent vector in $T_x M$ satisfying $\forall \eta \in T_x M$, the directional derivative

$$\mathrm{D}\, f(x)[\eta] = \langle \operatorname{grad} f(x), \eta \rangle$$

and $\operatorname{grad} f(x)$ is the direction of steepest ascent.

Definition

The Riemannian Hessian of f at x is a symmetric linear operator from $T_x M$ to $T_x M$ defined as

$$\operatorname{Hess} f(x) : T_x M \to T_x M : \eta \to \nabla_\eta \operatorname{grad} f,$$

where ∇ is the affine connection.

Retractions

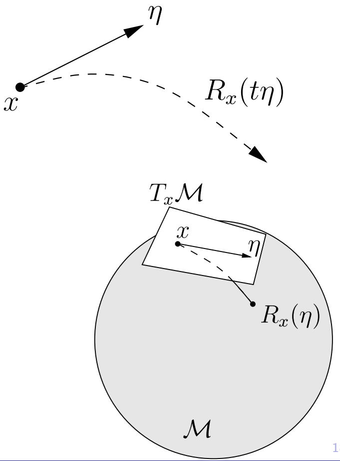
Euclidean	Riemannian
$x_{k+1} = x_k + \alpha_k d_k$	$x_{k+1} = R_{x_k}(\alpha_k \eta_k)$

Definition

A retraction is a mapping R from TM to M satisfying the following:

- R is continuously differentiable
- $R_x(0) = x$
- $\mathrm{D}\, R_x(0)[\eta] = \eta$

- maps tangent vectors back to the manifold
- defines curves in a direction

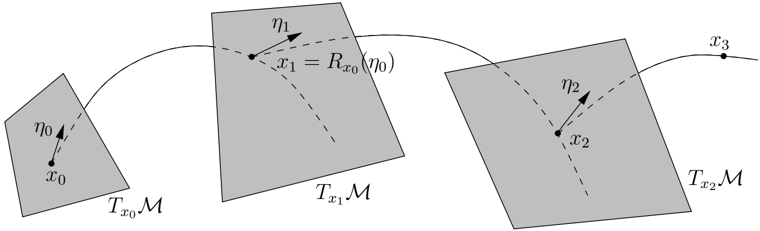


Generic Riemannian Optimization Algorithm

1. At iterate $x \in M$
2. Find $\eta \in T_x M$ which satisfies certain condition.
3. Choose new iterate $x_+ = R_x(\eta)$.
4. Goto step 1.

A suitable setting

This paradigm is sufficient for describing many optimization methods.



Categories of Riemannian optimization methods

Retraction-based: local information only

Line search-based: use local tangent vector and $R_x(t\eta)$ to define line

- Steepest decent
- Newton

Local model-based: series of flat space problems

- Riemannian trust region Newton (RTR)
- Riemannian adaptive cubic overestimation (RACO)

Categories of Riemannian optimization methods

Elements required for optimizing a cost function (M, g):

- an representation for points x on M, for tangent spaces T_xM, and for the inner products $g_x(\cdot, \cdot)$ on T_xM;
- choice of a retraction $R_x : T_xM \to M$;
- formulas for $f(x), \mathrm{grad}\ f(x)$ and $\mathrm{Hess}\ f(x)$ (or its action);
- Computational and storage efficiency;

Categories of Riemannian optimization methods

Retraction and transport-based: information from multiple tangent spaces

- Conjugate gradient: multiple tangent vectors
- Quasi-Newton e.g. Riemannian BFGS: transport operators between tangent spaces

Additional element required for optimizing a cost function (M, g):

- formulas for combining information from multiple tangent spaces.

Vector Transports

Vector Transport

- Vector transport: Transport a tangent vector from one tangent space to another
- $\mathcal{T}_{\eta_x}\xi_x$, denotes transport of ξ_x to tangent space of $R_x(\eta_x)$. R is a retraction associated with $\mathcal{T}$
- Isometric vector transport $\mathcal{T}_{\mathrm{S}}$ preserve the length of tangent vector

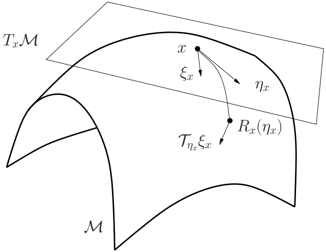


Figure: Vector transport.

Retraction/Transport-based Riemannian Optimization

Benefits

- Increased generality does not compromise the important theory
- Less expensive than or similar to previous approaches
- May provide theory to explain behavior of algorithms specifically developed for a particular application – or closely related ones

Possible Problems

- May be inefficient compared to algorithms that exploit application details

Some History of Optimization On Manifolds (I)

Luenberger (1973), *Introduction to linear and nonlinear programming*. Luenberger mentions the idea of performing line search along geodesics, "which we would use if it were computationally feasible (which it definitely is not)". Rosen (1961) essentially anticipated this but was not explicit in his Gradient Projection Algorithm.

Gabay (1982), *Minimizing a differentiable function over a differential manifold*. Steepest descent along geodesics; Newton's method along geodesics; Quasi-Newton methods along geodesics. On Riemannian submanifolds of $\mathbb{R}^n$.

Smith (1993-94), *Optimization techniques on Riemannian manifolds*. Levi-Civita connection ∇; Riemannian exponential mapping; parallel translation.

Some History of Optimization On Manifolds (II)

The "pragmatic era" begins:

Manton (2002), *Optimization algorithms exploiting unitary constraints* "The present paper breaks with tradition by not moving along geodesics". The geodesic update $\mathrm{Exp}_x \eta$ is replaced by a projective update $\pi(x + \eta)$, the *projection* of the point $x + \eta$ onto the manifold.

Adler, Dedieu, Shub, et al. (2002), *Newton's method on Riemannian manifolds and a geometric model for the human spine*. The exponential update is relaxed to the general notion of *retraction*. The geodesic can be replaced by any (smoothly prescribed) curve tangent to the search direction.

Absil, Mahony, Sepulchre (2007) Nonlinear conjugate gradient using retractions.

Some History of Optimization On Manifolds (III)

Theory, efficiency, and library design improve dramatically:

Absil, Baker, Gallivan (2004-07), Theory and implementations of Riemannian Trust Region method. Retraction-based approach. Matrix manifold problems, software repository

`http://www.math.fsu.edu/~cbaker/GenRTR`

Anasazi Eigenproblem package in Trilinos Library at Sandia National Laboratory

Absil, Gallivan, Qi (2007-10), Basic theory and implementations of Riemannian BFGS and Riemannian Adaptive Cubic Overestimation. Parallel translation and Exponential map theory, Retraction and vector transport empirical evidence.

Some History of Optimization On Manifolds (IV)

Ring and With (2012), combination of differentiated retraction and isometric vector transport for convergence analysis of RBFGS

Absil, Gallivan, Huang (2009-2015), Complete theory of Riemannian Quasi-Newton and related transport/retraction conditions, Riemannian SR1 with trust-region, RBFGS on partly smooth problems, A C++ library: `http://www.math.fsu.edu/~whuang2/ROPTLIB`

Sato, Iwai (2013-2015), Global convergence analysis using the differentiated retraction for Riemannian conjugate gradient methods

Many people Application interests start to increase noticeably

Current UCL/FSU Methods

- Riemannian Steepest Descent
- Riemannian Trust Region Newton: global, quadratic convergence
- Riemannian Broyden Family : global (convex), superlinear convergence
- Riemannian Trust Region SR1: global, $(d+1)-$superlinear convergence
- For large problems
  - Limited memory RTRSR1
  - Limited memory RBFGS
- Riemannian conjugate gradient (much more work to do on local analysis)
- A library is available at `www.math.fsu.edu/~whuang2/ROPTLIB`

Current/Future Work on Riemannian methods

- Manifold and inequality constraints
- Discretization of infinite dimensional manifolds and the convergence/accuracy of the approximate minimizers – specific to a problem and extracting general conclusions
- Partly smooth cost functions on Riemannian manifold

PhaseLift Framework

- A cost function in the PhaseLift [CSV13] framework is:

$$\min_{X \in \mathbb{C}^{n \times n}, X \geq 0} \|b^2 - \mathrm{diag}(ZXZ^*)\|_2^2 + \kappa \mathrm{trace}\,(X)\,;$$

- A desired minimizer has rank one;
- This motivates us to consider the optimization problem

$$\min_{X \in \mathbb{C}^{n \times n}, X \geq 0} H(X)$$

and the desired minimizer is low rank.

Optimization on Hermitian Positive Semidefinite Matrices

$$\min_{X \in \mathbb{C}^{n\times n}, X \geq 0} H(X)$$

- Suppose the rank of desired minimizer r^* is at most p.
- The domain $\{X \in \mathbb{C}^{n\times n} | X \geq 0\}$ can be replaced by $\mathcal{D}_p$, where $\mathcal{D}_p = \{X \in \mathbb{C}^{n\times n} | X \geq 0, \mathrm{rank}(X) \leq p\}$.
- An alternate cost function can be used:

$$F_p : \mathbb{C}^{n\times p} \to \mathbb{R} : Y_p \mapsto H(Y_p Y_p^*).$$

- Choosing $p > 1$ yields computational and theoretical benefits.
- This idea is not new and has been discussed in [BM03] and [JBAS10] for real positive semidefinite matrix constraints.

First Order Optimality Condition

Theorem

If $Y_p^* \in \mathbb{C}^{n \times p}$ is a rank deficient minimizer of F_p, then $Y_p Y_p^*$ is a stationary point of H.
In addition, if H is a convex cost function, $Y_p Y_p^*$ is a global minimizer of H.

- The real version of the optimality condition is given in [JBAS10].

Optimization Framework

- Equivalence: if $Y_p Y_p^* = \tilde{Y}_p \tilde{Y}_p^*$, then $F_p(Y_p) = F_p(\tilde{Y}_p)$;
- Quotient manifolds are used to remove the equivalence:
  - Equivalent class of $Y_r \in \mathbb{C}_*^{n \times r}$ is $[Y_r] = \{Y_r O_r | O_r \in \mathcal{O}_r\}$, where $1 \leq r \leq p$, $\mathbb{C}_*^{n \times r}$ denotes the n-by-r complex noncompact Stiefel manifold and $\mathcal{O}_r$ denote the r-by-r complex rotation group;
  - A fixed rank quotient manifold $\mathbb{C}_*^{n \times r} / \mathcal{O}_r = \{[Y_r] | Y_r \in \mathbb{C}_*^{n \times r}\}$, $1 \leq r \leq p$;
- Function on a fixed rank manifold is

$$f_r : \mathbb{C}_*^{n \times r} / \mathcal{O}_r \to \mathbb{R} : [Y_r] \mapsto F_r(Y_r) = H(Y_r Y_r^*);$$

- Optimize the cost function f_r and update r if necessary;
- A similar approach is used in [JBAS10] for real problems;

Update Rank Strategy

- Most of work is to choose a upper bound k for the rank and optimize over $\mathbb{C}^{n\times k}$ or $\mathbb{R}^{n\times k}$.
- Increasing rank by a constant [JBAS10, UV14]
  - Descent
  - Globally converge
- Dynamically search for a suitable rank [ZHG+15]
  - Not descent
  - Globally converge

Compare with a Convex Programming Solver

- Compare with convex programming
  - FISTA [BT09] in Matlab library TFOCS [BCG11];
  - X can be too large to be handled by the solver;
  - A low rank version of FISTA is used, denoted by LR-FISTA;
  - The approach is used in [CESV13, CSV13];
  - Works in practice but no theoretical results.

Comparisons

Table: $n_1 = n_2 = 64$; $n = n_1 n_2 = 4096$. k denotes the upper bound of the low-rank approximation in LR-FISTA. $\sharp$ represents the number of iterations reach the maximum. The relative mean-square error (RMSE) is $\min_{a:|a|=1} \|ax - x_*\|_2 / \|x_*\|_2$.

noiseless	Algorithm 1	LR-FISTA (k)				
		1	2	4	8	16
$iter$	124	1022	377	601	1554	$2000^\sharp$
nf	129	2212	804	1278	3360	4322
ng	124	1106	402	639	1680	2161
f_f	4.62_{-12}	8.18_{-12}	4.50_{-11}	4.64_{-12}	1.54_{-11}	1.27_{-9}
RMSE	6.34_{-6}	1.01_{-5}	1.74_{-5}	1.46_{-5}	1.10_{-4}	2.56_{-3}
t	2.12	1.27_2	5.25_1	9.35_1	3.48_2	6.86_2

- Algorithm 1 is faster and gives smaller RMSE.

The Gold Ball Data

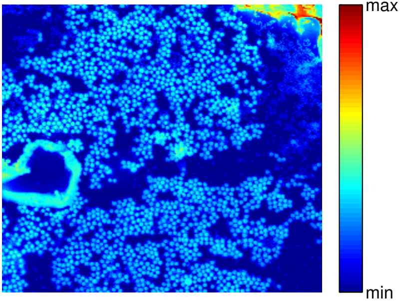


Figure: Image of the absolute value of the 256-by-256 complex-valued image. $n = 65536$. The pixel values correspond to the complex transmission coefficients of a collection of gold balls embedded in a medium.

Thank Stefano Marchesini at Lawrence Berkeley Notional Laboratory for providing the gold balls data set and granting permission to use it.

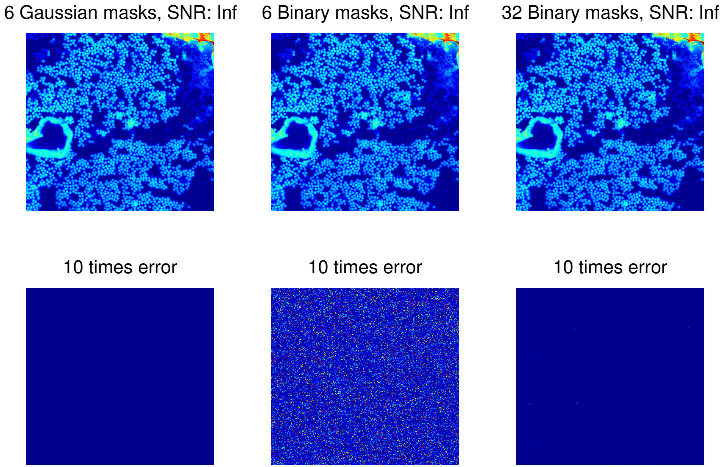
The Gold Ball Data

A set of binary masks contains a mask that is all 1 (which yields the original image) and several other masks comprising elements that are 0 or 1 with equal probability.

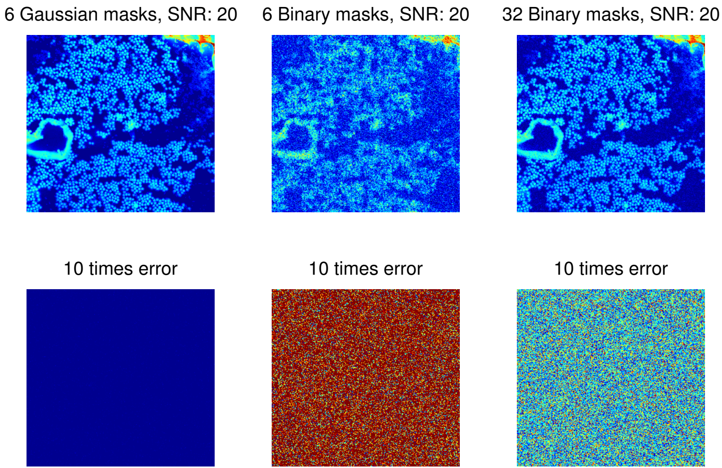
Table: RMSE and computational time (second) results with varying number and types of masks are shown in format RMSE/TIME. ♯ represents the computational time reaching 1 hour, i.e., 3.6_3 seconds.

	Algorithm 1			LR-FISTA		
SNR (dB)	20	40	inf	20	40	inf
6 Gaussian	$8.32_{-3}/4.30_1$	$8.32_{-5}/4.50_1$	$3.12_{-6}/4.19_1$	$8.32_{-3}/\sharp$	$3.12_{-4}/\sharp$	$3.12_{-4}/\sharp$
6 binary	$7.23_{-1}/7.90_2$	$1.29_{-1}/4.24_2$	$1.09_{-1}/4.42_2$	$8.24_{-1}/\sharp$	$4.98_{-1}/\sharp$	$4.98_{-1}/\sharp$
32 binary	$2.21_{-1}/6.84_2$	$3.02_{-3}/7.36_2$	$2.57_{-3}/6.54_2$	$6.07_{-1}/\sharp$	$5.82_{-1}/\sharp$	$5.78_{-1}/\sharp$

The Gold Ball Data



The Gold Ball Data



Summary

- Introduced the framework of Riemannian optimization and the state-of-the-art Riemannian algorithms
- Used applications to show the importance of Riemannian optimization
- Showed the performance of Riemannian optimization by using an optimization problem in the PhaseLift framework

Thanks!

References I

S. Becker, E. J. Cand, and M. Grant.
Templates for convex cone problems with applications to sparse signal recovery.
Mathematical Programming Computation, 3:165–218, 2011.

S. Burer and R. D. C. Monteiro.
A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization.
Mathematical Programming, 95(2):329–357, February 2003.
doi:10.1007/s10107-002-0352-8.

A. Beck and M. Teboulle.
A fast iterative shrinkage-thresholding algorithm for linear inverse problems.
SIAM Journal on Imaging Sciences, 2(1):183–202, January 2009.
doi:10.1137/080716542.

E. J. Candès, Y. C. Eldar, T. Strohmer, and V. Voroninski.
Phase retrieval via matrix completion.
SIAM Journal on Imaging Sciences, 6(1):199–225, 2013.
arXiv:1109.0573v2.

J. F. Cardoso and A. Souloumiac.
Blind beamforming for non-gaussian signals.
IEE Proceedings F Radar and Signal Processing, 140(6):362, 1993.

E. J. Candès, T. Strohmer, and V. Voroninski.
PhaseLift : Exact and stable signal recovery from magnitude measurements via convex programming.
Communications on Pure and Applied Mathematics, 66(8):1241–1274, 2013.

References II

Y. C. Eldar and S. Mendelson.
Phase retrieval: stability and recovery guarantees.
Applied and Computational Harmonic Analysis, 1:1–22, September 2013.
doi:10.1016/j.acha.2013.08.003.

R. W. Harrison.
Phase problem in crystallography.
Journal of the Optical Society of America A, 10(5):1046–1055, May 1993.
doi:10.1364/JOSAA.10.001046.

M. Journée, F. Bach, P.-A. Absil, and R. Sepulchre.
Low-rank optimization on the cone of positive semidefinite matrices.
SIAM Journal on Optimization, 20(5):2327–2351, 2010.

F. J. Theis and Y. Inouye.
On the use of joint diagonalization in blind signal processing.
2006 IEEE International Symposium on Circuits and Systems, (2):7–10, 2006.

A. Uschmajew and B. Vandereycken.
Line-search methods and rank increase on low-rank matrix varieties.
In *Proceedings of the 2014 International Symposium on Nonlinear Theory and its Applications (NOLTA2014)*, 2014.

B. Vandereycken.
Low-rank matrix completion by Riemannian optimization—extended version.
SIAM Journal on Optimization, 23(2):1214–1236, 2012.

References III

G. Zhou, W. Huang, K. A. Gallivan, P. Van Dooren, and P.-A. Absil.
Rank-constrained optimization: A Riemannian manifold approach.
In *Proceedings of The 23th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, number 1, 2015.