Geometry and Statistics in High-Dimensional Structured Optimization

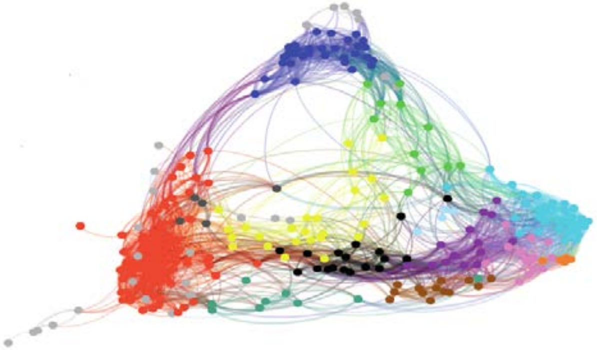
Yuanming Shi

ShanghaiTech University

Outline

- **Motivations**
  - Issues on computation, storage, nonconvexity,…
- **Two Vignettes:**
  - Structured Sparse Optimization
    - Geometry of Convex Statistical Optimization
    - Fast Convex Optimization Algorithms
  - Generalized Low-rank Optimization
    - Geometry of Nonconvex Statistical Optimization
    - Scalable Riemannian Optimization Algorithms
- **Concluding remarks**

Motivation: High-Dimensional Statistical Optimization

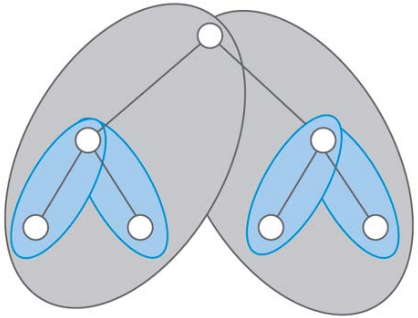


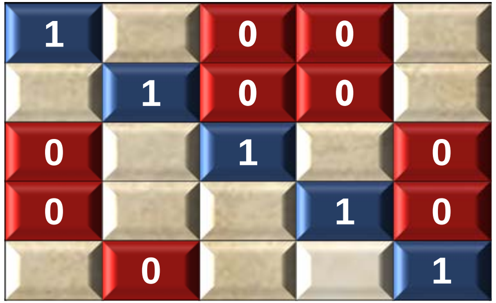
Motivations

- **The era of massive data sets**
  - Lead to new issues related to modeling, computing, and statistics.
- **Statistical issues**
  - Concentration of measure: high-dimensional probability
  - Importance of "low-dimensional" structures: sparsity and low-rankness
- **Algorithmic issues**
  - Excessively large problem dimension, parameter size
  - Polynomial-time algorithms often not fast enough
  - Non-convexity in general formulations

Issue A: Large-scale structured optimization

- Explosion in scale and complexity of the optimization problem for massive data set processing

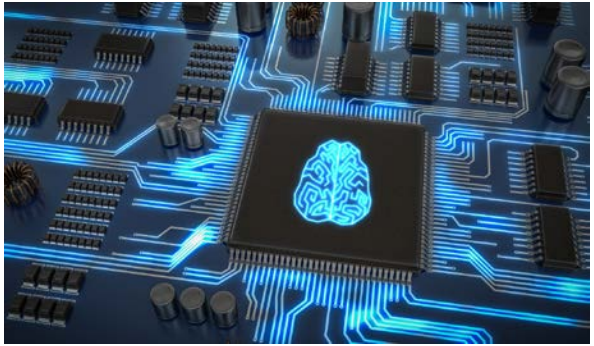




- **Questions:**
  - ➢ How to exploit the low-dimensional structures (e.g., sparsity and low-rankness) to assist efficient algorithms design?

Issue B: Computational vs. statistical efficiency

- Massive data sets require very fast algorithms but with rigorous guarantees: parallel computing and approximations are essential



- **Questions:**
  - When is there a gap between polynomial-time and exponential-time algorithms?
  - What are the trade-offs between computational and statistical efficiency?

Issue C: Scalable nonconvex optimization

- Nonconvex optimization may be super scary: saddle points, local optima

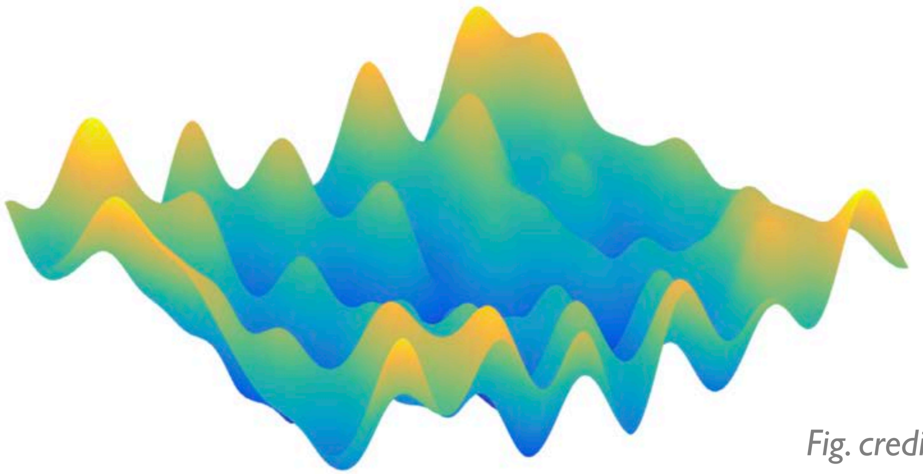
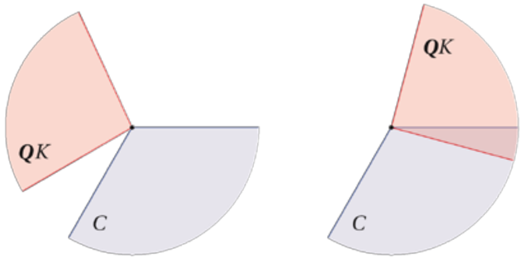


Fig. credit: Chen

- **Question:**
  - ➢ How to exploit the geometry of nonconvex programs to guarantee optimality and enable scalability in computation and storage?

Vignettes A: ***Structured Sparse Optimization***

1. Geometry of Convex Statistical Estimation
   1) Phase transitions of random convex programs
   2) Convex geometry, statistical dimension
2. Fast Convex Optimization Algorithms
   1) Homogeneous self-dual embedding
   2) Operator splitting, ADMM



High-dimensional sparse optimization

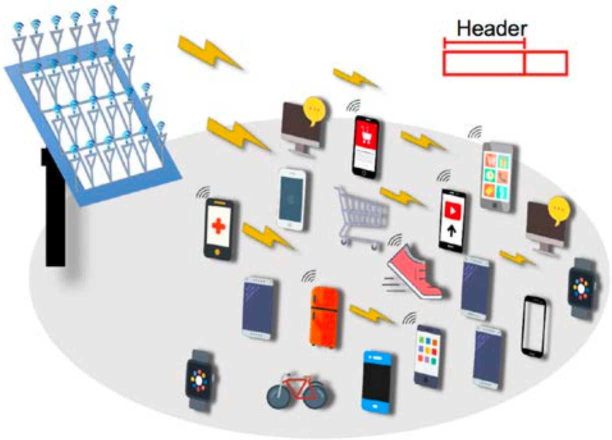
- Let $\boldsymbol{x}^\natural \in \mathbb{R}^d$ be an unknown structured sparse signal
  - Individual sparsity for compressed sensing
- Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function that reflects structure, e.g., ℓ_1-norm
- Let $\boldsymbol{A} \in \mathbb{R}^{m \times d}$ be a measurement operator
- **Observe** $\boldsymbol{z} = \boldsymbol{A}\boldsymbol{x}^\natural$
- Find estimate $\hat{\boldsymbol{x}}$ by solving **convex program**

$$\text{minimize} \quad f(\boldsymbol{x}) \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{z}$$

- **Hope:** $\hat{\boldsymbol{x}} = \boldsymbol{x}^\natural$

Application: High-dimensional IoT data analysis

- Machine-type communication (e.g., massive IoT devices) with sporadic traffic: massive device connectivity



Sporadic traffic: only a small fraction of potentially large number of devices are active for data acquisition (e.g., temperature measurement)

Application: High-dimensional IoT data analysis

- Cellular network with massive number of devices
  - Single-cell uplink with a BS with M antennas; Total N single-antenna devices, active devices (sporadic traffic) $\mathcal{S} \subset \{1, 2, \ldots, N\}$

$$\boldsymbol{y}(\ell) = \sum_{i \in \mathcal{S}} \boldsymbol{h}_i q_i(\ell), \ell = 1, \ldots, L$$

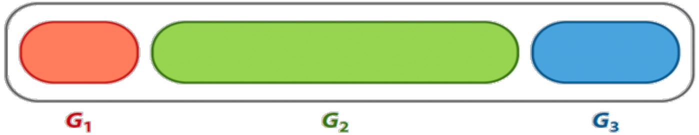
- Define diagonal activity matrix $\boldsymbol{A} \in \mathbb{R}^{N \times N}$ with $|\mathcal{S}|$ non-zero diagonals

$$\boldsymbol{Y} = \boldsymbol{QAH}$$

  - $\boldsymbol{Y} = [\boldsymbol{y}(1), \ldots, \boldsymbol{y}(L)]^T \in \mathbb{C}^{L \times M}$ denotes the received signal across M antennas
  - $\boldsymbol{H} = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_N]^T \in \mathbb{C}^{N \times M}$: channel matrix from all devices to the BS
  - $\boldsymbol{Q} = [\boldsymbol{q}(1), \ldots, \boldsymbol{q}(L)]^T \in \mathbb{C}^{L \times N}$: known transmit pilot matrix from devices

Group sparse estimation

- Let $\boldsymbol{\Theta}^\natural = \boldsymbol{A}\boldsymbol{H} \in \mathbb{C}^{N\times M}$ (unknown): group sparsity in rows $\boldsymbol{\theta}^{[i]}$ of matrix $\boldsymbol{\Theta}^\natural$

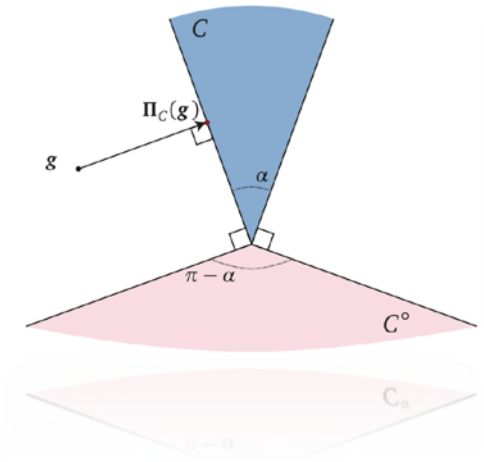


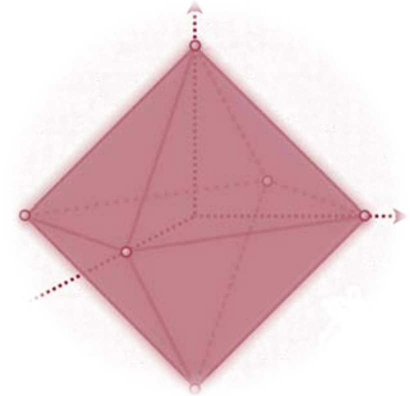
- Let $\boldsymbol{Q} \in \mathbb{C}^{L\times N}$ be a known measurement operator (pilot matrix)
- **Observe** $\boldsymbol{Y} = \boldsymbol{Q}\boldsymbol{\Theta}^\natural$
- Find estimate $\hat{\boldsymbol{\Theta}}$ by solving a **convex program**

$$\underset{\boldsymbol{\Theta}\in\mathbb{C}^{N\times M}}{\text{minimize}} \quad f(\boldsymbol{\Theta}) \quad \text{subject to} \quad \boldsymbol{Y} = \boldsymbol{Q}\boldsymbol{\Theta}$$

  - $f(\boldsymbol{\Theta}) = \sum_{i=1}^{N} \|\boldsymbol{\theta}^{[i]}\|_2$ is mixed ℓ_1/ℓ_2-norm to reflect group sparsity structure

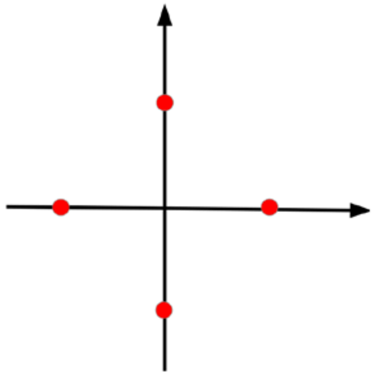
Geometry of Convex Statistical Optimization



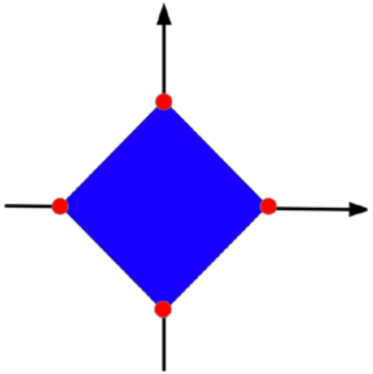


Geometric view: sparsity

- Sparse approximation via convex hull $\mathcal{D} := \mathrm{conv}\left(\{\pm \boldsymbol{e}_i | i \in [n]\}\right)$



$\Rightarrow$



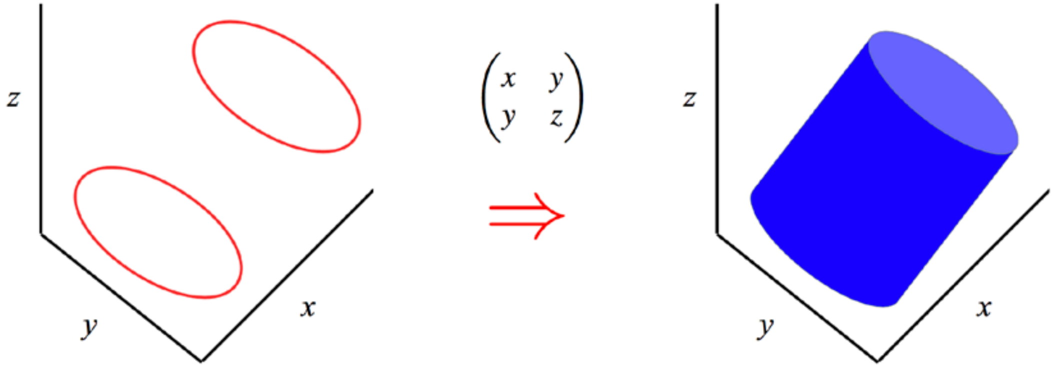
1-sparse vectors of
Euclidean norm 1

convex hull: ℓ_1-norm

$$\|\boldsymbol{z}\|_1 = \sum_{i=1}^n |z_i|$$

Geometric view: low-rank

- Low-rank approximation via convex hull



2x2 rank 1 symmetric matrices (normalized)

convex hull: nuclear norm

$$\|M\|_* = \sum_i \sigma_i(M)$$

Geometry of sparse optimization

- **Descent cone** of a function f at a point z is

$$\mathscr{D}(f, \boldsymbol{z}) := \{\boldsymbol{d} : f(\boldsymbol{z} + \epsilon \boldsymbol{d}) \le f(\boldsymbol{z}), \quad \text{for some } \epsilon > 0\}$$

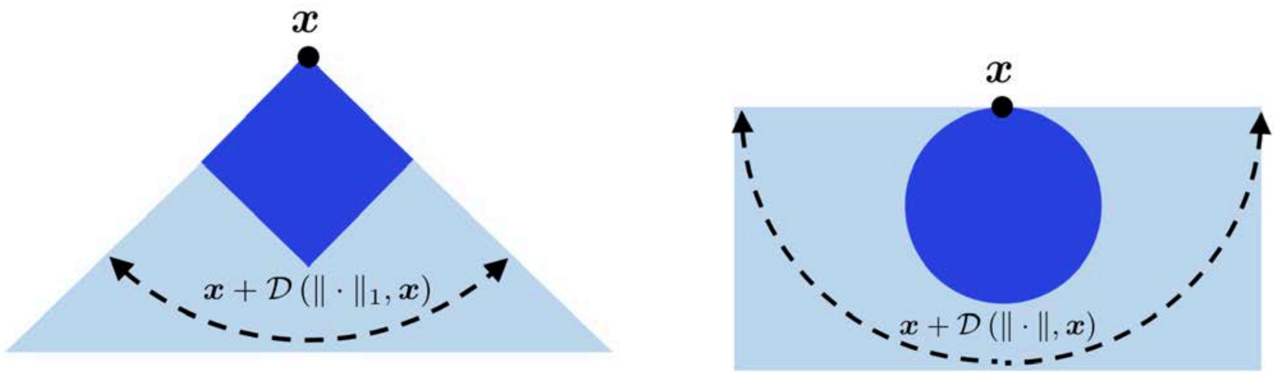


Fig. credit: Chen

References: Rockafellar 1970

Geometry of sparse optimization

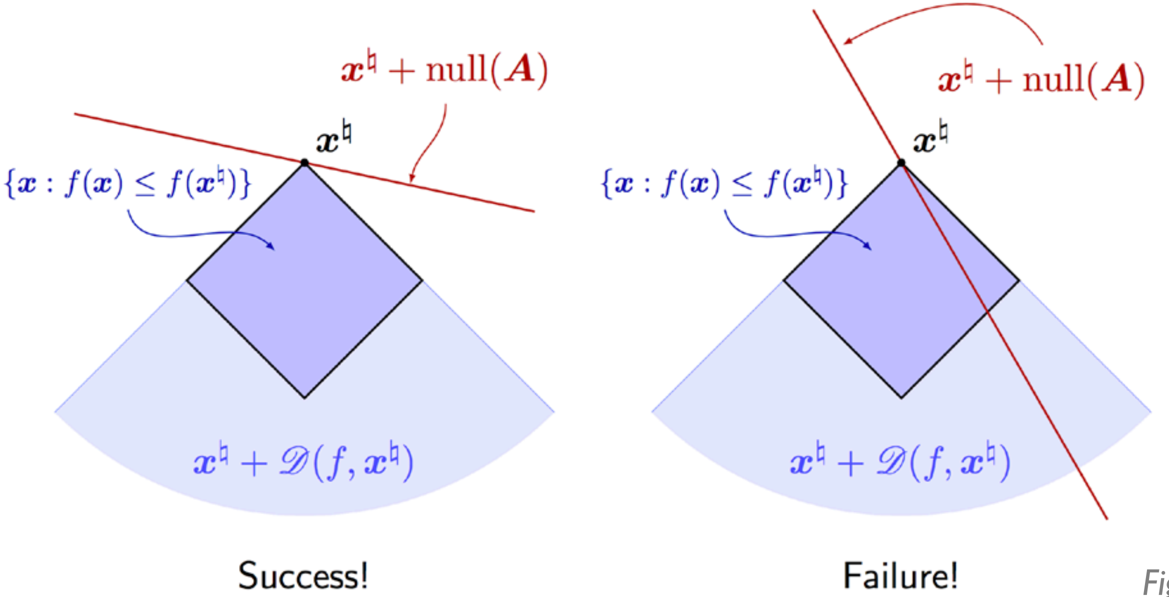


Fig. credit: Tropp

References: Candes–Romberg–Tao 2005, Rudelson–Vershynin 2006, Chandrasekaran et al. 2010, Amelunxen et al. 2013

Sparse optimization with random data

- **Assume**
  - The vector $\boldsymbol{x}^\natural \in \mathbb{R}^d$ is unknown
  - The observation $\boldsymbol{z} = \boldsymbol{A}\boldsymbol{x}^\natural$ where $\boldsymbol{A} \in \mathbb{R}^{m\times d}$ is standard normal
  - The vector $\hat{\boldsymbol{x}}$ solves

$$\text{minimize} \quad f(\boldsymbol{x}) \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{z}$$

- **Then**

$$m \gtrsim \delta(\mathscr{D}(f, \boldsymbol{x}^\natural)) \quad \Longrightarrow \quad \hat{\boldsymbol{x}} = \boldsymbol{x}^\natural, \quad \text{w.h.p.}$$
$$m \lesssim \delta(\mathscr{D}(f, \boldsymbol{x}^\natural)) \quad \Longrightarrow \quad \hat{\boldsymbol{x}} \neq \boldsymbol{x}^\natural, \quad \text{w.h.p.}$$

statistical dimension [Amelunxen-McCoy-Tropp'13]

Statistical dimension

- The statistical dimension of a closed, convex cone K is

$$\delta(K) := \mathbb{E}\left[\|\Pi_K(\boldsymbol{g})\|_2^2\right]$$

  - Π_K is the Euclidean projection onto K; $\boldsymbol{g}$ is a standard normal vector

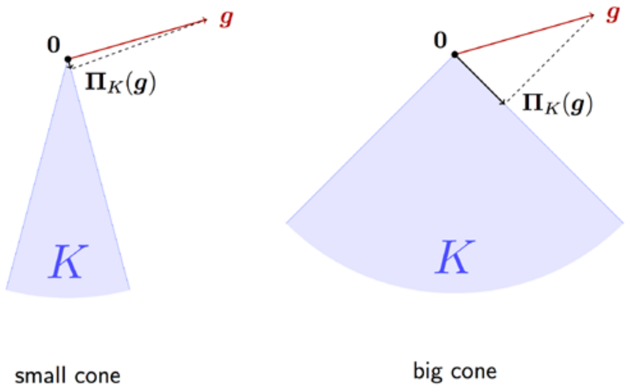


Fig. credit: Tropp

Cone	Notation	Statistical Dimension
j-dim subspace	L_j	j
Nonnegative orthant	$\mathbb{R}_+^d$	$\frac{1}{2}d$
Second-order cone	$\mathbb{L}^{d+1}$	$\frac{1}{2}(d+1)$
Real psd cone	$\mathbb{S}_+^d$	$\frac{1}{4}d(d-1)$

Examples for statistical dimension

- **Example I:** ℓ_1-minimization for compressed sensing
  - $\boldsymbol{x}^\natural \in \mathbb{R}^d$ with s non-zero entries

$$\delta\left(\mathscr{D}(\|\cdot\|_1, \boldsymbol{x}^\natural)\right) = \inf_{\tau \geq 0} \left\{ s(1+\tau^2) + (d-s)\sqrt{\frac{2}{\pi}} \int_\tau^\infty (z-\tau)^2 e^{-z^2} dz \right\}$$

- **Example II:** ℓ_1/ℓ_2-minimization for massive device connectivity
  - $\boldsymbol{X}^\natural \in \mathbb{R}^{N \times M}$ with s non-zero rows

$$\delta\left(\mathscr{D}(\|\cdot\|_{2,1}, \boldsymbol{X}^\natural)\right) = \inf_{\tau \geq 0} \left\{ s(M+\tau^2) + (N-s)\frac{2^{1-M/2}}{\Gamma(M/2)} \int_\tau^\infty (u-\tau)^2 u^{M-1} e^{-\frac{u^2}{2}} \mathrm{d}u \right\}$$

Numerical phase transition

- Compressed sensing with ℓ_1-minimization

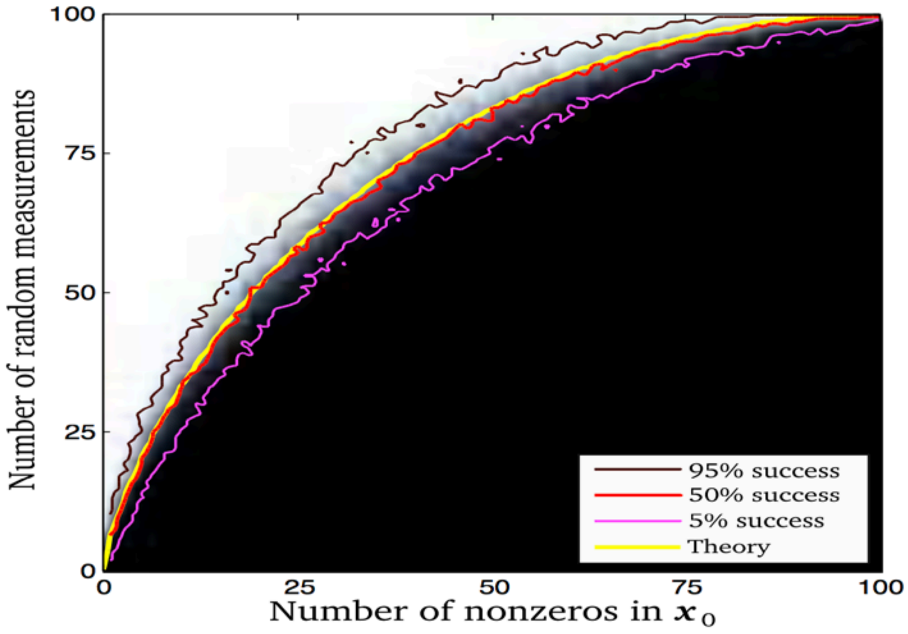
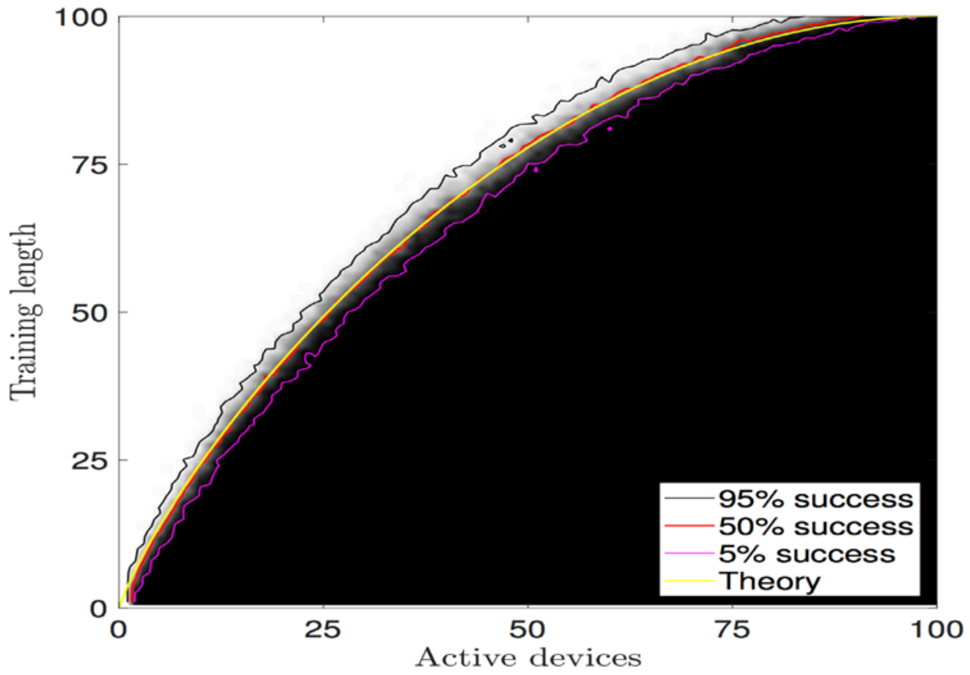


Fig. credit: Amelunxen-McCoy-Tropp'13

Numerical phase transition

- User activity detection via ℓ_1/ℓ_2-minimization



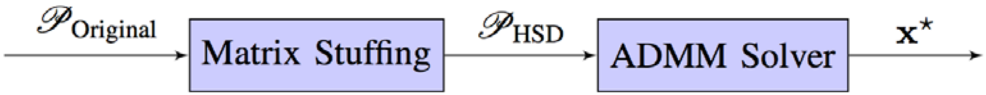
group-structured sparsity estimation

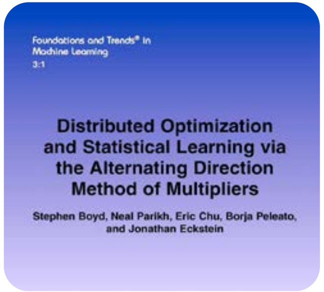
Summary of convex statistical optimization

- Theoretical foundations for sparse optimization
  - Convex relaxation: convex hull, convex analysis
  - Fundamental bounds for convex methods: convex geometry, high-dimensional statistics
- Computational limits for (convexified) sparse optimization
  - Custom methods (e.g., stochastic gradient descent): not generalizable for complicated problems
  - Generic methods (e.g., CVX): not scalable to large problem sizes

Can we design a unified framework for general large-scale convex programs?

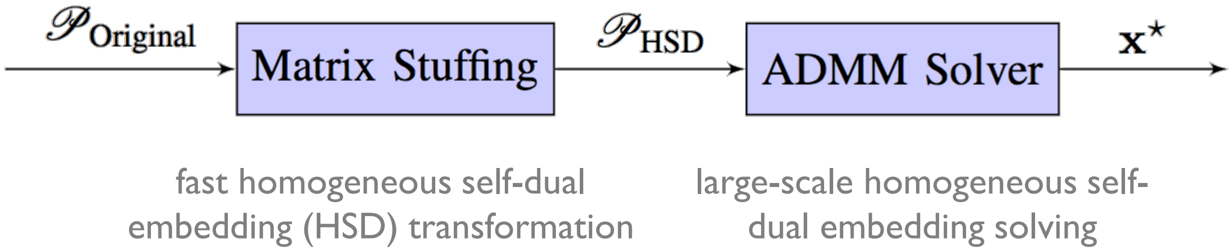
Fast Convex Optimization Algorithms





Large-scale convex optimization

- **Proposal:** Two-stage approach for large-scale convex optimization



  - Matrix stuffing: Fast homogeneous self-dual embedding (HSD) transformation
  - Operator splitting (ADMM): Large-scale homogeneous self-dual embedding

Smith form reformulation

- **Goal:** Transform the classical form to conic form

$$\begin{array}{ll} \underset{\boldsymbol{z}}{\text{minimize}} & f_0(\boldsymbol{z};\boldsymbol{\alpha}) \\ \text{subject to} & f_i(\boldsymbol{z};\boldsymbol{\alpha}) \le g_i(\boldsymbol{z};\boldsymbol{\alpha}), \\ & u_i(\boldsymbol{z};\boldsymbol{\alpha}) = v_i(\boldsymbol{z};\boldsymbol{\alpha}). \end{array} \quad \Rightarrow \quad \begin{array}{ll} \underset{\boldsymbol{\nu},\boldsymbol{\mu}}{\text{minimize}} & \mathbf{c}^T\boldsymbol{\nu} \\ \text{subject to} & \mathbf{A}\boldsymbol{\nu} + \boldsymbol{\mu} = \mathbf{b}, \\ & (\boldsymbol{\nu},\boldsymbol{\mu}) \in \mathbb{R}^n \times \mathcal{K}. \end{array}$$

- **Key idea:** Introduce a new variable for each subexpression in classical form [Smith '96]
  - The Smith form is ready for standard cone programming transformation

Example

- Coordinated beamforming problem family

$$\mathscr{P}_{\text{Original}} : \text{minimize} \quad \|\boldsymbol{v}\|_2^2$$
$$\text{subject to} \quad \|\boldsymbol{D}_l \boldsymbol{v}\|_2 \le \sqrt{P_l}, \forall l, \quad \text{Per-BS power constraint} \quad (1)$$
$$\|\boldsymbol{C}_k \boldsymbol{v} + \boldsymbol{g}_k\|_2 \le \beta_k \boldsymbol{r}_k^T \boldsymbol{v}, \forall k. \quad \text{QoS constraints} \quad (2)$$

- Smith form reformulation

$$\mathcal{G}_1(l) : \begin{cases} (y_0^l, \mathbf{y}_1^l) \in \mathcal{Q}^{KN_l+1} \\ y_0^l = \sqrt{P_l} \in \mathbb{R} \\ \mathbf{y}_1^l = \boldsymbol{D}_l \boldsymbol{v} \in \mathbb{R}^{KN_l} \end{cases}$$

Smith form for (1)

$$\mathcal{G}_2(k) : \begin{cases} (t_0^k, \mathbf{t}_1^k) \in \mathcal{Q}^{K+1} \\ t_0^k = \beta_k \mathbf{r}_k^T \mathbf{v} \in \mathbb{R} \\ \mathbf{t}_1^k = \mathbf{t}_2^k + \mathbf{t}_3^k \in \mathbb{R}^{K+1} \\ \mathbf{t}_2^k = \mathbf{C}_k \mathbf{v} \in \mathbb{R}^{K+1} \\ \mathbf{t}_3^k = \mathbf{g}_k \in \mathbb{R}^{K+1} \end{cases}$$

Smith form for (2)

The Smith form is readily to be reformulated as the standard cone program

Reference: Grant-Boyd'08

Optimality condition

- KKT conditions (necessary and sufficient, assuming strong duality)
  - Primal feasibility: $\mathbf{A}\boldsymbol{\nu}^\star + \boldsymbol{\mu}^\star - \mathbf{b} = \mathbf{0}$
  - Dual feasibility: $\mathbf{A}^T\boldsymbol{\eta}^\star - \boldsymbol{\lambda}^\star + \mathbf{c} = \mathbf{0}$
  - Complementary slackness: $\mathbf{c}^T\boldsymbol{\nu}^\star + \mathbf{b}^T\boldsymbol{\eta}^\star = 0$ zero duality gap
  - Feasibility: $(\boldsymbol{\nu}^\star, \boldsymbol{\mu}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\eta}^\star) \in \mathbb{R}^n \times \mathcal{K} \times \{0\}^n \times \mathcal{K}^*$

no solution if primal or dual problem infeasible/unbounded

Homogeneous self-dual (HSD) embedding

- **HSD embedding** of the primal-dual pair of transformed standard cone program (based on KKT conditions) [Ye et al. 94]

$$\boxed{\begin{aligned} \underset{\boldsymbol{\nu},\boldsymbol{\mu}}{\text{minimize}}\ & \mathbf{c}^T\boldsymbol{\nu} \\ \text{subject to}\ & \mathbf{A}\boldsymbol{\nu}+\boldsymbol{\mu}=\mathbf{b} \\ & (\boldsymbol{\nu},\boldsymbol{\mu})\in\mathbb{R}^n\times\mathcal{K}. \end{aligned}} + \boxed{\begin{aligned} \underset{\boldsymbol{\eta},\boldsymbol{\lambda}}{\text{maximize}}\ & -\mathbf{b}^T\boldsymbol{\eta} \\ \text{subject to}\ & -\mathbf{A}^T\boldsymbol{\eta}+\boldsymbol{\lambda}=\mathbf{c} \\ & (\boldsymbol{\lambda},\boldsymbol{\eta})\in\{0\}^n\times\mathcal{K}^* \end{aligned}} \Longrightarrow \boxed{\begin{aligned} \mathscr{F}_{\text{HSD}}:\ & \text{find } (\mathbf{x},\mathbf{y}) \\ \text{subject to}\ & \mathbf{y}=\mathbf{Q}\mathbf{x} \\ & \mathbf{x}\in\mathcal{C},\mathbf{y}\in\mathcal{C}^* \end{aligned}}$$

$$\underbrace{\begin{bmatrix}\boldsymbol{\lambda}\\ \boldsymbol{\mu}\\ \kappa\end{bmatrix}}_{\mathbf{y}} = \underbrace{\begin{bmatrix}\mathbf{0} & \mathbf{A}^T & \mathbf{c}\\ -\mathbf{A} & \mathbf{0} & \mathbf{b}\\ -\mathbf{c}^T & -\mathbf{b}^T & \mathbf{0}\end{bmatrix}}_{\mathbf{Q}} \underbrace{\begin{bmatrix}\boldsymbol{\nu}\\ \boldsymbol{\eta}\\ \tau\end{bmatrix}}_{\mathbf{x}}$$

finding a *nonzero* solution

- This feasibility problem is homogeneous and self-dual

Recovering solution or certificates

- Any HSD solution $(\boldsymbol{\nu}, \boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\eta}, \tau, \kappa)$ falls into one of three cases:
  - Case 1: $\tau > 0,\ \kappa = 0$, then $\hat{\boldsymbol{\nu}} = \boldsymbol{\nu}/\tau, \hat{\boldsymbol{\eta}} = \boldsymbol{\eta}/\tau, \hat{\boldsymbol{\mu}} = \boldsymbol{\mu}/\tau$ is a solution
  - Case 2: $\tau = 0,\ \kappa > 0$, implies $\mathbf{c}^T\boldsymbol{\nu} + \mathbf{b}^T\boldsymbol{\eta} < 0$
    - If $\mathbf{b}^T\boldsymbol{\eta} < 0$, then $\hat{\boldsymbol{\eta}} = \boldsymbol{\eta}/(-\mathbf{b}^T\boldsymbol{\eta})$ certifies primal infeasibility
    - If $\mathbf{c}^T\boldsymbol{\nu} < 0$, then $\hat{\boldsymbol{\nu}} = \boldsymbol{\nu}/(-\mathbf{c}^T\hat{\boldsymbol{\nu}})$ certifies dual infeasibility
  - Case 3: $\tau = \kappa = 0$, nothing can be said about original problem
- **HSD embedding:** 1) obviates need for phase I / phase II solves to handle infeasibility/unboundedness; 2) used in all interior-point cone solvers

Operator Splitting

$$\begin{aligned} \mathscr{F}_{\text{HSD}} : \text{find} \quad & (\mathbf{x}, \mathbf{y}) \\ \text{subject to} \quad & \mathbf{y} = \mathbf{Q}\mathbf{x} \\ & \mathbf{x} \in \mathcal{C}, \mathbf{y} \in \mathcal{C}^* \end{aligned}$$

Alternating direction method of multipliers

- **ADMM:** an operator splitting method solving convex problems in form

$$\mathscr{P}_{\text{ADMM}} : \text{minimize} \quad f(\mathbf{x}) + g(\boldsymbol{z}) \qquad \text{subject to} \;\; \mathbf{x} = \boldsymbol{z}$$

  - $f,\ g$ convex, not necessarily smooth, can take infinite values

- The basic ADMM algorithm [Boyd et al., FTML 11]

$$\begin{aligned}
\mathbf{x}^{[k+1]} &= \arg\min_{\mathbf{x}} \left(f(\mathbf{x}) + (\rho/2)\|\mathbf{x} - \boldsymbol{z}^{[k]} - \lambda^{[k]}\|_2^2 \right) \\
\boldsymbol{z}^{[k+1]} &= \arg\min_{\boldsymbol{z}} \left(g(\boldsymbol{z}) + (\rho/2)\|\mathbf{x}^{[k+1]} - \boldsymbol{z} - \lambda^{[k]}\|_2^2 \right) \\
\lambda^{[k+1]} &= \lambda^{[k]} - \mathbf{x}^{[k+1]} + \boldsymbol{z}^{[k+1]}
\end{aligned}$$

  - $\rho > 0$ is a step size; λ is the dual variable associated the constraint

Alternating direction method of multipliers

- **Convergence of ADMM:** Under benign conditions ADMM guarantees
  - $f(\mathbf{x}^k) + g(\boldsymbol{z}^k) \to p^\star$
  - $\lambda^k \to \lambda^\star$, an optimal dual variable
  - $\mathbf{x}^k - \boldsymbol{z}^k \to 0$
- Same as many other operator splitting methods for consensus problem, e.g., Douglas-Rachford method
- **Pros:** 1) with good robustness of method of multipliers; 2) can support decomposition

Operator splitting

- Transform HSD embedding $\mathscr{F}_{\mathrm{HSD}}$ in ADMM form: Apply the operating splitting method (ADMM)

$$\mathscr{P}_{\mathrm{ADMM}} : \underset{\mathbf{x},\tilde{\mathbf{x}},\mathbf{y},\tilde{\mathbf{y}}}{\text{minimize}} \quad I_{\mathcal{C}\times\mathcal{C}^*}(\mathbf{x},\mathbf{y}) + I_{\mathbf{Q}\tilde{\mathbf{x}}=\tilde{\mathbf{y}}}(\tilde{\mathbf{x}},\tilde{\mathbf{y}})$$
$$\text{subject to} \quad (\mathbf{x},\mathbf{y}) = (\tilde{\mathbf{x}},\tilde{\mathbf{y}})$$

- **Final algorithm**

$$\tilde{\mathbf{x}}^{[i+1]} = (\mathbf{I}+\mathbf{Q})^{-1}(\mathbf{x}^{[i]}+\mathbf{y}^{[i]})$$ subspace projection

$$\mathbf{x}^{[i+1]} = \Pi_{\mathcal{C}}(\tilde{\mathbf{x}}^{[i+1]}-\mathbf{y}^{[i]})$$ parallel cone projection

$$\mathbf{y}^{[i+1]} = \mathbf{y}^{[i]}-\tilde{\mathbf{x}}^{[i+1]}+\mathbf{x}^{[i+1]}$$ computationally trivial

Parallel cone projection

- **Proximal algorithms** for parallel cone projection [Parikn & Boyd, FTO 14]
  - Projection onto the second-order cone: $\mathcal{Q}^d = \{(z, \mathbf{x}) \in \mathbb{R} \times \mathbb{R}^{d-1} | \|\mathbf{x}\| \leq z\}$

$$\Pi_{\mathcal{C}}(\boldsymbol{\omega}, \tau) = \begin{cases} 0, \|\boldsymbol{\omega}\|_2 \leq -\tau \\ (\boldsymbol{\omega}, \tau), \|\boldsymbol{\omega}\|_2 \leq \tau \\ (1/2)(1 + \tau/\|\boldsymbol{\omega}\|_2)(\boldsymbol{\omega}, \|\boldsymbol{\omega}\|_2), \|\boldsymbol{\omega}\|_2 \geq |\tau|. \end{cases}$$

    - Closed-form, computationally scalable (we mainly focus on SOCP)
  - Projection onto positive semidefinite cone: $\mathbf{S}_+^n = \{\boldsymbol{M} \in \mathbb{R}^{n \times n} | \boldsymbol{M} = \boldsymbol{M}^T, \boldsymbol{M} \succeq \mathbf{0}\}$

$$\Pi_{\mathcal{C}}(\boldsymbol{V}) = \sum_{i=1}^n (\lambda_i)_+ \boldsymbol{u}_i \boldsymbol{u}_i^T$$

    - SVD is computationally expensive

Numerical results

- Power minimization coordinated beamforming problem (SOCP)

Network Size (L=K)		20	50	100	150
Interior-Point Solver	Solving Time [sec]	4.2835	**326.2513**	N/A	N/A
	Objective [W]	12.2488	6.5216	N/A	N/A
Operator Splitting	Solving Time [sec]	0.1009	**2.4821**	23.8088	81.0023
	Objective [W]	12.2523	6.5193	3.1296	2.0689

ADMM can speedup **130x** over the interior-point method

[Ref] Y. Shi, J. Zhang, B. O'Donoghue, and K. B. Letaief, "Large-scale convex optimization for dense wireless cooperative networks," IEEE Trans. Signal Process., vol. 63, no. 18, pp. 4729-4743, Sept. 2015. **(The 2016 IEEE Signal Processing Society Young Author Best Paper Award)**

Cone programs with random constraints

- Phase transitions in cone programming: independent standard normal entries in $\boldsymbol{c} \in \mathbb{R}^d$ and $\boldsymbol{A} \in \mathbb{R}^{m \times d}$

$$\underset{\boldsymbol{z} \in \mathbb{R}^d}{\text{minimize}} \quad \langle \boldsymbol{c}, \boldsymbol{z} \rangle \qquad \text{subject to} \quad \boldsymbol{A}\boldsymbol{z} = \boldsymbol{b} \quad \text{and} \quad \boldsymbol{z} \in \mathcal{C}$$

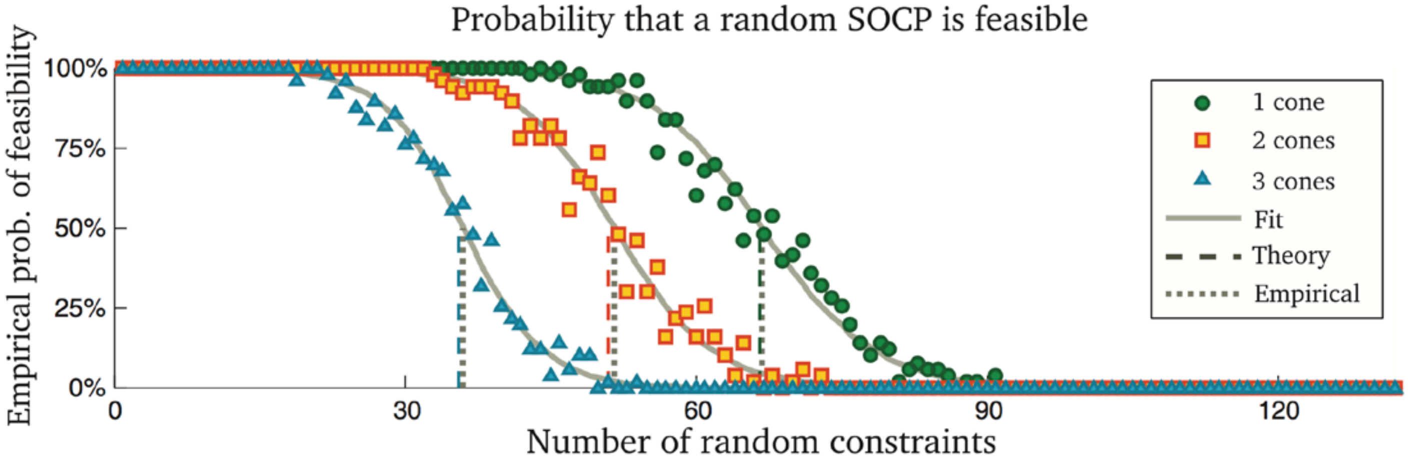
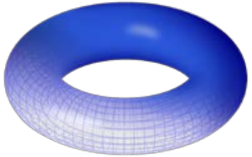
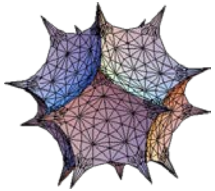


Fig. credit: Amelunxen-McCoy-Tropp'13

Vignette B: Generalized Low-Rank Optimization

1. Geometry of Nonconvex Statistical Estimation
2. Scalable Riemannian Optimization Algorithms

Optimization over Riemannian Manifolds (non-Euclidean geometry)

Generalized low-rank matrix optimization

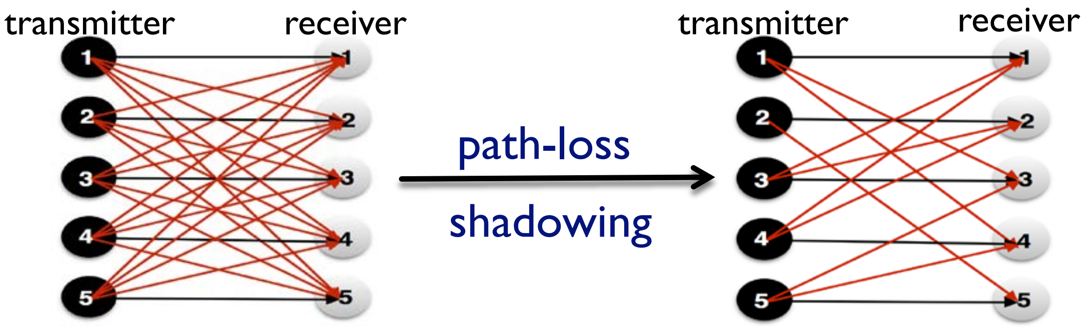
- Rank-constrained matrix optimization problem

$$\underset{\boldsymbol{M}\in\mathbb{R}^{n\times n}}{\text{minimize}} \quad f(\mathcal{A}(\boldsymbol{M})) \qquad \text{subject to} \quad \text{rank}(\boldsymbol{M}) = r$$

  - $\mathcal{A}:\mathbb{R}^{n\times n}\rightarrow\mathbb{R}^d$ is a real linear map on $n\times n$ matrices
  - $f:\mathbb{R}^d\rightarrow\mathbb{R}$ is convex and differentiable
  - A prevalent **model** in signal processing, statistics and machine learning (e.g., low-rank matrix completion)
- **Challenge I:** Reliably solve the low-rank matrix problem at scale
- **Challenge II:** Develop optimization algorithms with optimal storage $\Theta(rn)$

Application: Topological interference alignment

- **Blessings:** partial connectivity in dense wireless networks for massive data processing and transmission



Degrees of Freedom?

$$\mathsf{DoF} = \lim_{\mathsf{SNR}\to\infty} \frac{C(\mathsf{SNR})}{\log(\mathsf{SNR})}$$

- **Approach:** topological interference management (TIM) [Jafar, TIT 14]
  - ➢ **Maximize the achievable DoF:** only based on the network topology information **(no CSIT)**

Application: Topological interference alignment

- **Goal:** Deliver one data stream per user over N time slots
  - Transmitter i transmits $\mathbf{v}_i s_i$, receiver i receives

$$\mathbf{y}_i = \mathbf{v}_i h_{ii} s_i + \sum_{(i,j)\in\mathcal{S}, i\neq j} \mathbf{v}_j h_{ij} s_j + \mathbf{n}_i$$

$\mathcal{S}$: network connectivity pattern

  - Receiver decodes symbol s_i by projecting $\mathbf{y}_i$ into the space $\mathbf{u}_i \in \mathbb{C}^N$

$$\mathbf{u}_i^{\mathsf{H}}\mathbf{y}_i = \mathbf{u}_i^{\mathsf{H}}\mathbf{v}_i h_{ii} s_i + \sum_{(i,j)\in\mathcal{S}, i\neq j} \mathbf{u}_i^{\mathsf{H}}\mathbf{v}_j h_{ij} s_j + \mathbf{u}_i^{\mathsf{H}}\mathbf{n}_i$$

- **Topological interference alignment condition**

$$M_{ij} = \begin{cases} \mathbf{u}_i^{\mathsf{H}}\mathbf{v}_i = 1, & \forall i, \\ \mathbf{u}_i^{\mathsf{H}}\mathbf{v}_j = 0, & \forall i \neq j, (i,j)\in\mathcal{S}, \\ \star, & \text{otherwise.} \end{cases} \Longrightarrow$$

$$\mathbf{u}_i^{\mathsf{H}}\mathbf{y}_i = h_{ii} s_i + \mathbf{u}_i^{\mathsf{H}}\mathbf{n}_i$$

$$\text{DoF} = \frac{1}{\operatorname{rank}(\boldsymbol{M})} = \frac{1}{N}$$

Generalized low-rank model

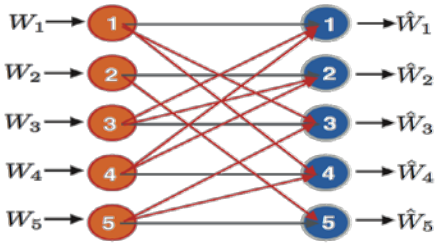
- Generalized low-rank optimization with network side information

$$\begin{aligned} \underset{\boldsymbol{M}\in\mathbb{C}^{K\times K}}{\text{minimize}} \quad & \text{rank}(\boldsymbol{M}) \\ \text{subject to} \quad & M_{ii} = 1, i = 1, \ldots, K \\ & M_{ij} = 0, \forall (i,j) \in \mathcal{S} \end{aligned}$$

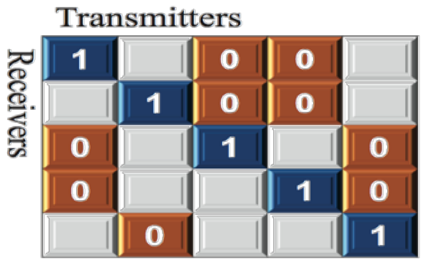
$\leftarrow$ topological interference alignment condition

➢ $\boldsymbol{M} = [\boldsymbol{u}_i^{\mathsf{H}} \boldsymbol{v}_j]$: precoding vectors and decoding vectors $\boldsymbol{u}_k, \boldsymbol{v}_k \in \mathbb{C}^N$

➢ $\text{rank}(\boldsymbol{M})$ equals the inverse of achievable degrees-of-freedom (DoF) $\frac{1}{N}$



(a) Topological interference alignment



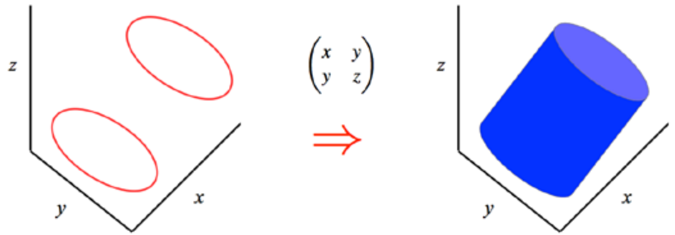
(b) Side information modeling matrix $\boldsymbol{M}$

side information $\mathcal{S}$

Nuclear norm fails

- **Convex relaxation fails:** always return the identity matrix!

$$\begin{aligned} \underset{\boldsymbol{M} \in \mathbb{C}^{K \times K}}{\text{minimize}} \quad & \|\boldsymbol{M}\|_* \\ \text{subject to} \quad & M_{ii} = 1, i = 1, \ldots, K \\ & M_{ij} = 0, \forall (i,j) \in \mathcal{S} \end{aligned}$$



  - **Fact:** $\mathrm{Trace}(\boldsymbol{M}) \leq \|\boldsymbol{M}\|_*$

- **Proposal:** Solve the nonconvex problems directly with rank adaptivity

$$\begin{aligned} \underset{\boldsymbol{M} \in \mathbb{C}^{K \times K}}{\text{minimize}} \quad & f(\boldsymbol{M}) := \|\mathcal{A}(\boldsymbol{M}) - \boldsymbol{z}\|_F^2 \\ \text{subject to} \quad & \mathrm{rank}(\boldsymbol{M}) = r \end{aligned}$$

manifold constraint

Riemannian manifold optimization problem

Recent advances in nonconvex optimization

- **2009–Present: Nonconvex heuristics**
  - Burer–Monteiro factorization idea + various nonlinear programming methods
  - Store low-rank matrix factors $\Theta(rn)$
- **Guaranteed solutions:** Global optimality with statistical assumptions
  - Matrix completion/recovery: [Sun-Luo'14], [Chen-Wainwright'15], [Ge-Lee-Ma'16],…
  - Phase retrieval: [Candes et al., 15], [Chen-Candes' 15], [Sun-Qu-Wright'16]
  - Community detection/phase synchronization [Bandeira-Boumal-Voroninski'16], [Montanari et al., 17],…

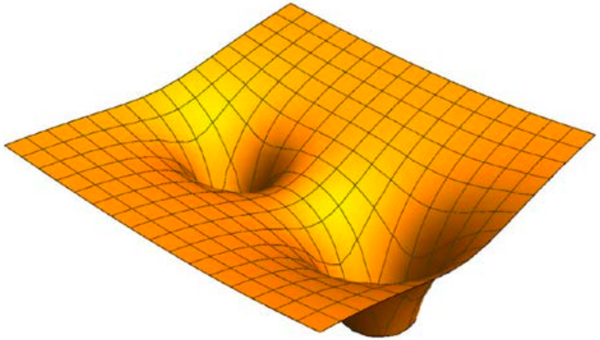
When are nonconvex optimization problems not scary?

Geometry of Nonconvex Statistical Optimization

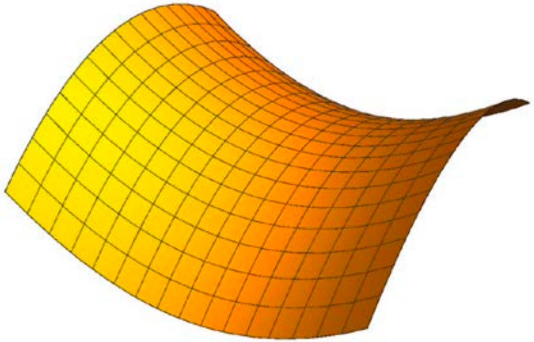
First-order stationary points

- Saddle points and local minima:

$$\lambda_{\min}(\nabla^2 f(\boldsymbol{z})) \begin{cases} > 0 & \text{local minimum} \\ = 0 & \text{local minimum or saddle point} \\ < 0 & \text{strict saddle point} \end{cases}$$



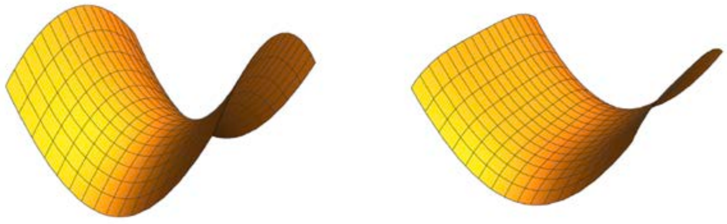
Local minima



Saddle points/local maxima

First-order stationary points

- **Applications:** PCA, matrix completion, dictionary learning etc.
  - **Local minima:** Either all local minima **are** global minima or all local minima **as good as** global minima
  - **Saddle points: Very poor** compared to global minima; **Several** such points



Strict saddle point Non-strict saddle point

- **Bottomline:** Local minima much more desirable than saddle points

Summary of nonconvex statistical optimization

- **Convex methods:**
  - Slow memory hogs
  - Convex relaxation fails sometimes, e.g., topological interference alignment
  - High computational complexity, e.g., eigenvalue decomposition
- **Nonconvex methods:** fast, lightweight
  - Under certain statistical models with benign global geometry: no spurious local optima

How to escape saddle points efficiently?

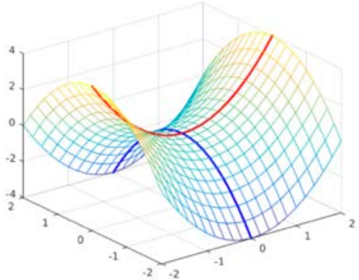
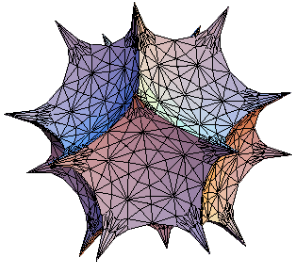


Fig credit: Sun, Qu & Wright

Riemannian Optimization Algorithms

Escape saddle pints via manifold optimization

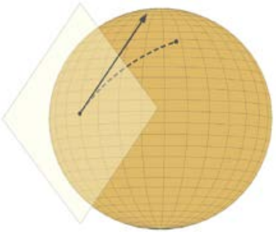


What is manifold optimization?

- Manifold (or manifold-constrained) optimization problem

$$\underset{\boldsymbol{M}\in\mathbb{C}^{m\times n}}{\text{minimize}} \quad f(\boldsymbol{M}) \quad \text{subject to} \quad \boldsymbol{M}\in\mathcal{M}$$

  - $f:\mathbb{R}^{m\times n}\to\mathbb{R}$ is a smooth function
  - $\mathcal{M}$ is a **Riemannian manifold:** spheres, orthonormal bases (Stiefel), rotations, positive definite matrices, ***fixed-rank matrices***, Euclidean distance matrices, semidefinite fixed-rank matrices, linear subspaces (Grassmann), phases, essential matrices, fixed-rank tensors, Euclidean spaces...



Escape saddle pints via manifold optimization

- Convergence guarantees for Riemannian **trust regions**
  - Global convergence to **second-order critical points**
  - Quadratic convergence rate locally
  - Reach ϵ-second order stationary point $\|\mathrm{grad} f(\boldsymbol{z})\| \leq \epsilon$ and $\nabla^2 f(\boldsymbol{z}) \succeq -\epsilon \boldsymbol{I}$ in $\mathcal{O}(1/\epsilon^3)$ iterations under Lipschitz assumptions [Cartis & Absil'16]

Escape strict saddle points via finding second-order stationary point

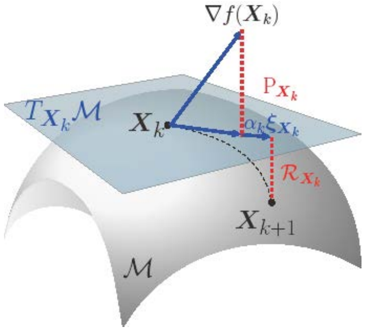
- **Other approaches:** Gradient descent by adding noise [Ge et al., 2015], [Jordan et al., 17] (slow convergence rate in general)

Recent applications of manifold optimization

- Matrix/tensor completion/recovery: [Vandereycken'13], [Boumal-Absil'15], [Kasai-Mishra'16],…
- Gaussian mixture models: [Hosseini-Sra'15], Dictionary learning: [Sun-Qu-Wright'17], Phase retrieval: [Sun-Qu-Wright'17],…
- Phase synchronization/community detection: [Boumal'16], [Bandeira-Boumal-Voroninski'16],…
- **Wireless transceivers design:** [Shi-Zhang-Letaief'16], [Yu-Shen-Zhang-K. B. Letaief'16], [Shi-Mishra-Chen'16],…

The power of manifold optimization paradigms

- Generalize Euclidean gradient (Hessian) to *Riemannian gradient (Hessian)*



$$\nabla_{\mathcal{M}} f(\mathbf{X}^{(k)}) = P_{\mathbf{X}^{(k)}}(\nabla f(\mathbf{X}^{(k)}))$$

Riemannian Gradient Euclidean Gradient

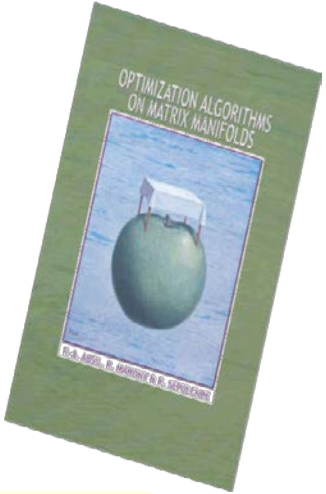
$$\mathbf{X}^{(k+1)} = \mathcal{R}_{\mathbf{X}^{(k)}}(-\alpha^{(k)} \nabla_{\mathcal{M}} f(\mathbf{X}^{(k)}))$$

Retraction Operator

- We need Riemannian geometry: 1) linearize search space $\mathcal{M}$ into a tangent space $T_{\mathbf{X}}\mathcal{M}$; 2) pick a metric on $T_{\mathbf{X}}\mathcal{M}$ to give intrinsic notions of gradient and Hessian

An excellent book

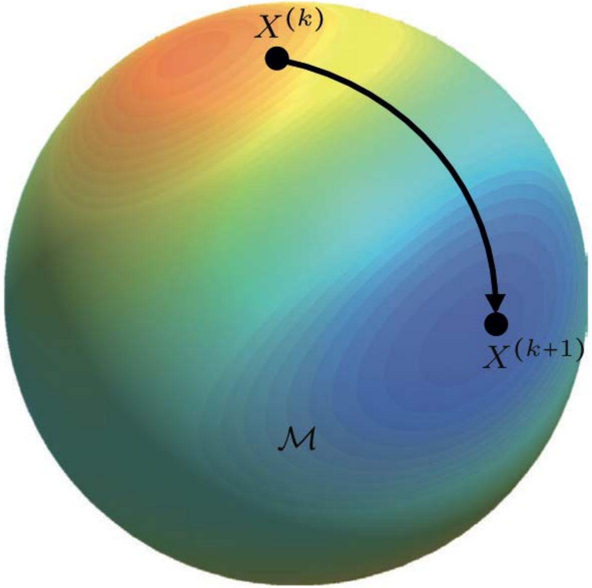
Optimization algorithms on matrix manifolds



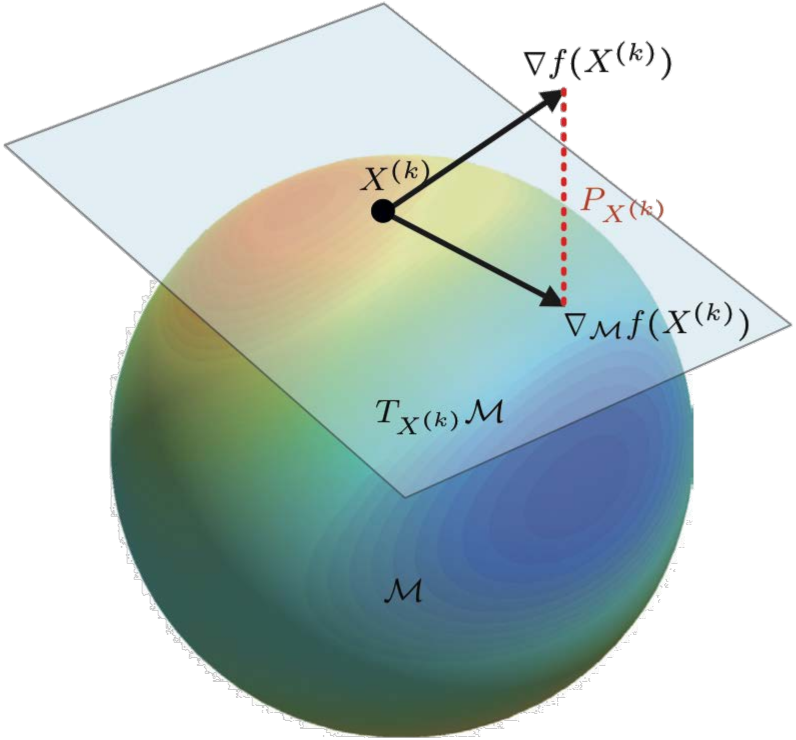
A Matlab toolbox

Taking A Close Look at ***Gradient Descent***

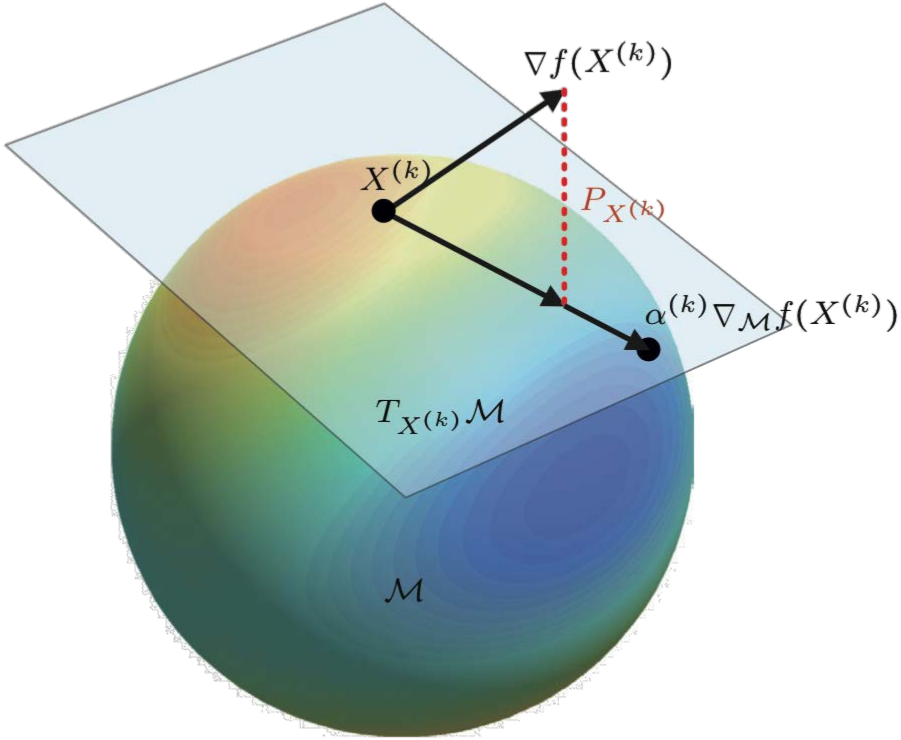
Optimization on the manifold: main idea



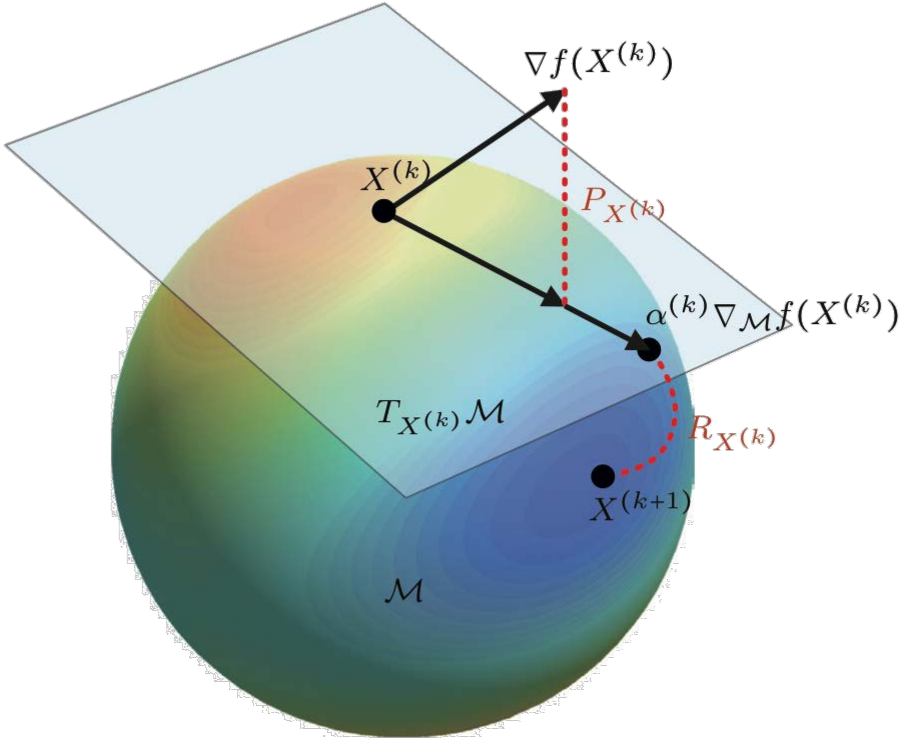
Optimization on the manifold: main idea



Optimization on the manifold: main idea



Optimization on the manifold: main idea



Example: Rayleigh quotient

- Optimization over (sphere) manifold $\mathbb{S}^{n-1} = \{x \in \mathbb{R}^n : x^T x = 1\}$

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \;\; f(x) = -x^T A x \quad \text{subject to} \;\; x^T x = 1$$

  - The cost function is smooth on $\mathbb{S}^{n-1}$, symmetric matrix $A \in \mathbb{R}^{n \times n}$

- Step 1: Compute the Euclidean gradient in $\mathbb{R}^n$

$$\nabla f(x) = -2Ax$$

- Step 2: Compute the Riemannian gradient on $\mathbb{S}^{n-1}$ via projecting $\nabla f(x)$ to the tangent space using the orthogonal projector $\mathrm{Proj}_x u = (I - xx^T)u$

$$\mathrm{grad} f(x) = \mathrm{Proj}_x \nabla f(x) = -2(I - xx^T)Ax$$

Example: Generalized low-rank optimization

- Generalized low-rank optimization for topological interference alignment via Riemannian optimization

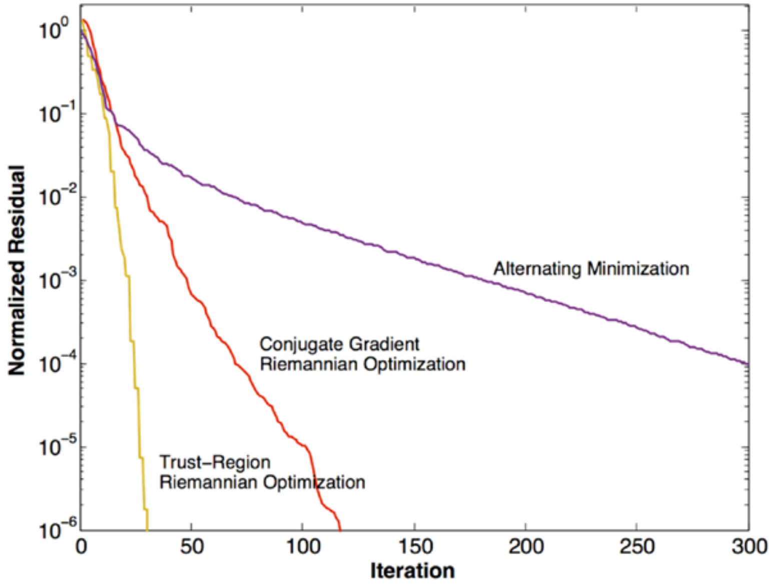
$$\underset{\boldsymbol{M}\in\mathbb{C}^{m\times n}}{\text{minimize}} \quad f(\boldsymbol{M}), \qquad \text{subject to} \quad \text{rank}(\boldsymbol{M}) = r$$

OPTIMIZATION-RELATED INGREDIENTS FOR PROBLEM $\mathscr{P}_r$

	$\mathscr{P}_r$: $\text{minimize}_{\mathbf{X}\in\mathcal{M}_r}\ f(\mathbf{X})$
Matrix representation of an element $\mathbf{X}\in\mathcal{M}_r$	$\mathbf{X} = (\mathbf{U}, \boldsymbol{\Sigma}, \mathbf{V})$
Computational space $\mathcal{M}_r$	$\text{St}(r, M)\times\text{GL}(r)\times\text{St}(r, M)$
Quotient space	$\text{St}(r, M)\times\text{GL}(r)\times\text{St}(r, M)/(\mathcal{O}(r)\times\mathcal{O}(r))$
Metric $g_{\mathbf{X}}(\xi_{\mathbf{X}}, \zeta_{\mathbf{X}})$ for $\xi_{\mathbf{X}}, \zeta_{\mathbf{X}}\in T_{\mathbf{X}}\mathcal{M}_r$	$g_{\mathbf{X}}(\xi_{\mathbf{X}}, \zeta_{\mathbf{X}}) = \langle\xi_U, \zeta_U\Sigma\Sigma^T\rangle + \langle\xi_\Sigma, \zeta_\Sigma\rangle + \langle\xi_V, \zeta_V\Sigma^T\Sigma\rangle$
Riemannian gradient $\text{grad}_{\mathbf{X}} f$	$\text{grad}_{\mathbf{X}} f = (\xi_U, \xi_\Sigma, \xi_V)$ (30)
Riemannian Hessian $\text{Hess}_{\mathbf{X}} f[\xi_{\mathbf{X}}]$	$\text{Hess}_{\mathbf{X}} f[\xi_{\mathbf{X}}] = \Pi_{\mathcal{H}_{\mathbf{X}}\mathcal{M}_r}(\nabla_{\xi_{\mathbf{X}}}\text{grad}_{\mathbf{X}} f)$ (40)
Retraction $\mathcal{R}_{\mathbf{X}}(\xi_{\mathbf{X}}) : \mathcal{H}_{\mathbf{X}}\mathcal{M}_r \to \mathcal{M}_r$	$(\text{uf}(\mathbf{U}+\xi_{\mathbf{X}}), \boldsymbol{\Sigma}+\xi_\Sigma, \text{uf}(\mathbf{V}+\xi_V))$

Convergence rates

- Optimize over fixed-rank matrices (quotient matrix manifold)

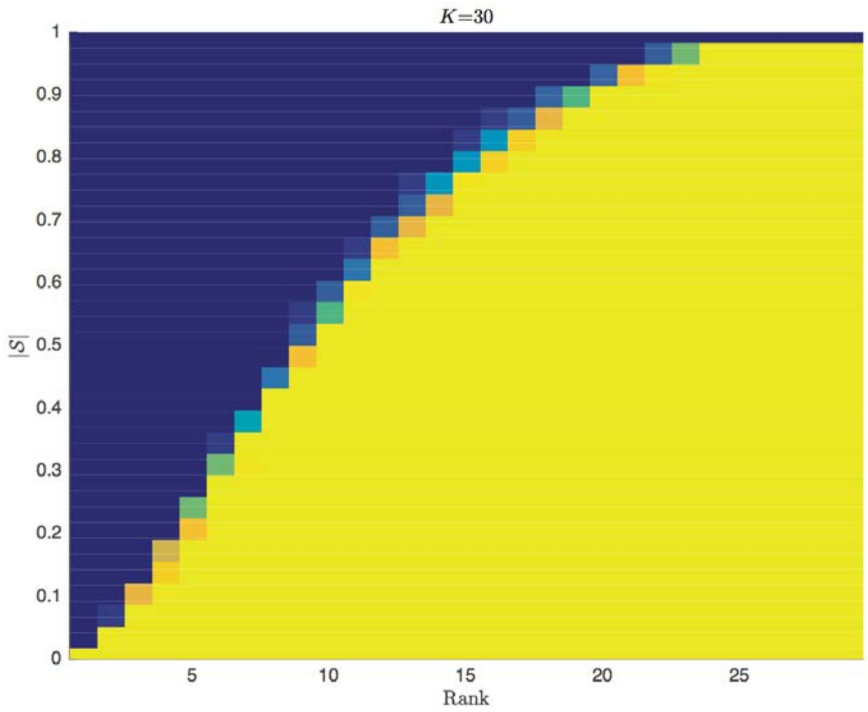


Riemannian algorithms:

1. Exploit the rank structure in a principled way
2. Develop second-order algorithms systematically
3. Scalable, SVD-free

[Ref] Y. Shi, J. Zhang, and K. B. Letaief, "Low-rank matrix completion for topological interference management by Riemannian pursuit," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, Jul. 2016.

Phase transitions for topological IA



$$\begin{array}{ll} \underset{\boldsymbol{M} \in \mathbb{C}^{K \times K}}{\operatorname{minimize}} & \operatorname{rank}(\boldsymbol{M}) \\ \text{subject to} & M_{ii} = 1, i = 1, \ldots, K \\ & M_{ij} = 0, \forall (i,j) \in \mathcal{S} \end{array}$$

The heat map indicates the empirical probability of success (blue=0%; yellow=100%)

Concluding remarks

- **Structured sparse optimization**
  - Convex geometry and analysis provide statistical optimality guarantees
  - Matrix stuffing for fast HSD embedding transformation
  - Operator splitting for solving large-scale HSD embedding
- **Future directions:**
  - Statistical analysis for more complicated problems, e.g., cone programs
  - Operator splitting for large-scale sparse SDP problems [Zheng-Fantuzzi-Papachristodoulou-Goulart-Wynn'17]
  - More applications: deep neural network compression via sparse optimization

Concluding remarks

- **Generalized low-rank optimization**
  - Nonconvex statistical optimization may not be that scary: no spurious local optima
  - Riemannian optimization is powerful: 1) Exploit the manifold geometry of fixed-rank matrices; 2) Escape saddle points
- **Future directions:**
  - Geometry of neural network loss surfaces via random matrix theory [Pennington-Bahri'17]: 1) Are all minima global? 2) What is the distribution of critical points?
  - More applications: blind deconvolution for IoT, big data analytics (e.g., ranking)

To learn more...

- **Web:** http://shiyuanming.github.io/
- **Papers:**
- Y. Shi, J. Zhang, and K. B. Letaief, "Group sparse beamforming for green Cloud-RAN," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, pp. 2809-2823, May 2014. (The 2016 Marconi Prize Paper Award)
- Y. Shi, J. Zhang, B. O'Donoghue, and K. B. Letaief, "Large-scale convex optimization for dense wireless cooperative networks," *IEEE Trans. Signal Process.*, vol. 63, no. 18, pp. 4729-4743, Sept. 2015. t. 2015. (The 2016 IEEE Signal Processing Society Young Author Best Paper Award)
- Y. Shi, J. Zhang, and K. B. Letaief, "Low-rank matrix completion for topological interference management by Riemannian pursuit," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 4703-4717, Jul. 2016.
- Y. Shi, J. Zhang, W. Chen, and K. B. Letaief, "Generalized sparse and low-rank optimization for ultra-dense networks," *IEEE Commun. Mag.*, to appear.