LUND UNIVERSITY

Low-Rank Inducing Norms with Optimality Interpretations

LU
2017
June 15

Christian Grussler

Pontus Giselsson, Anders Rantzer

Automatic Control, Lund University

LUND LINNAEUS

Problem

$$\begin{aligned} \underset{X \in \mathbb{R}^{m \times n}}{\text{minimize}} \quad & k(\|X\|) + h(X) \\ \text{subject to} \quad & \text{rank}(X) \leq r \end{aligned}$$

1. $k : \mathbb{R}_{\geq 0} \to \mathbb{R}$ is an increasing, convex, proper, closed function
2. $\|\cdot\|$ is a unitarily invariant norm
3. $h : \mathbb{R}^{m \times n} \to \mathbb{R}$ is a closed, proper, convex function

Vector-valued problems:

$$\begin{aligned} \underset{x \in \mathbb{R}^{n}}{\text{minimize}} \quad & k(\|\text{diag}(x)\|) + h(x) \\ \text{subject to} \quad & \underbrace{\text{rank}(\text{diag}(x))}_{\text{card}(x)} \leq r \end{aligned}$$

Example: Bilinear Regression [§]

Given $Y \in \mathbb{R}^{m\times n}$, $L \in \mathbb{R}^{k\times m}$, $R \in \mathbb{R}^{n\times k}$, $k \leq \min\{m,n\}$

$$\begin{aligned} \underset{X\in\mathbb{R}^{k\times k}}{\text{minimize}} \quad & \|Y - L^{\mathsf{T}} X R^{\mathsf{T}}\|_{\ell_2}^2 \\ \text{subject to} \quad & \text{rank}(X) \leq r \end{aligned}$$

where

- $X, Y \in \mathbb{R}^{m\times n}: \ \langle X, Y\rangle = \text{trace}(X^{\mathsf{T}}Y)$.
- $\|X\|_{\ell_2} = \sqrt{\langle X, X\rangle} = \sqrt{\sum_i \sigma_i^2(X)}$

[§] I.S. Dhillon '15

By assumption $\operatorname{rank}(\underbrace{L^{\mathsf{T}} X R^{\mathsf{T}}}_{=:M}) = \operatorname{rank}(X)$

$$\underset{M}{\text{minimize}} \quad \underbrace{\|M\|_{\ell_2}^2}_{k(\|M\|)} \underbrace{-2\langle Y, M\rangle + I_{\{M = L^{\mathsf{T}} X R^{\mathsf{T}}:\ X \in \mathbb{R}^{k\times k}\}}(M)}_{h(M)}$$

$$\text{subject to} \quad \operatorname{rank}(M) \leq r$$

Applications:

- Machine Learning: Principle Component Analysis, Multivariate Linear Regression, Data Compression, ...
- Control: Model Reduction, System Identification, ...

Explicit Solution:

$$\underset{\mathsf{rank}(X)\leq r}{\operatorname{argmin}} \ \|Y - L^{\mathsf{T}} X R^{\mathsf{T}}\|_{\ell_2}^2 = \{L^{\dagger} Y_r R^{\dagger} : Y_r \in \mathsf{svd}_r(Y)\}$$

$$\mathsf{svd}_r(Y) := \left\{ \sum_{i=1}^{r} \sigma_i(Y) u_i v_i^{\mathsf{T}} : Y = \sum_{i=1}^{q} \sigma_i(Y) u_i v_i^{\mathsf{T}} \text{ is SVD of } Y \right\}$$

with $\sigma_1(Y) \geq \cdots \geq \sigma_q(Y)$

Problem: Convex structural constraints?

$$\begin{aligned} \underset{X}{\text{minimize}} \quad & \|Y - L^{\mathsf{T}} X R^{\mathsf{T}}\|_{\ell_2}^2 + \tilde{h}(X) \\ \text{subject to} \quad & \operatorname{rank}(X) \leq r \end{aligned}$$

Examples:

- Nonnegative approximation: $\tilde{h}(X) = I_{\mathbb{R}^{k \times k}_{\geq 0}}(X)$.
- Hankel approximation: $\tilde{h}(X) = I_{\mathsf{Hankel}}(X)$.
- Feasibility problems: $Y = 0$ and $\tilde{h}(X) = I_{\mathcal{C}}(X)$.

Generally, no closed-form solutions are known!

Nuclear Norm Regularization

Standard approach today: Replace rank by nuclear norm [§]

$$\begin{aligned} \underset{X}{\text{minimize}} \quad & k(\|X\|) + h(X) \\ \text{subject to} \quad & \|X\|_{\ell_1} \leq \lambda \end{aligned}$$

- $\|X\|_{\ell_1} = \sum_i \sigma_i(X)$
- $\lambda \geq 0$ is fixed.

[§] Tibshirani, Chen, Donoho, Fazel, Boyd,...

Pros:

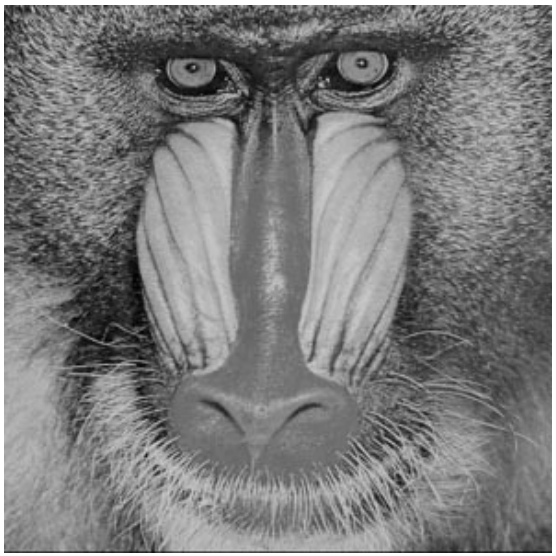
- Simple and generic heuristic $\implies$ No PhD needed!
- Probabilistic success guarantees [§]

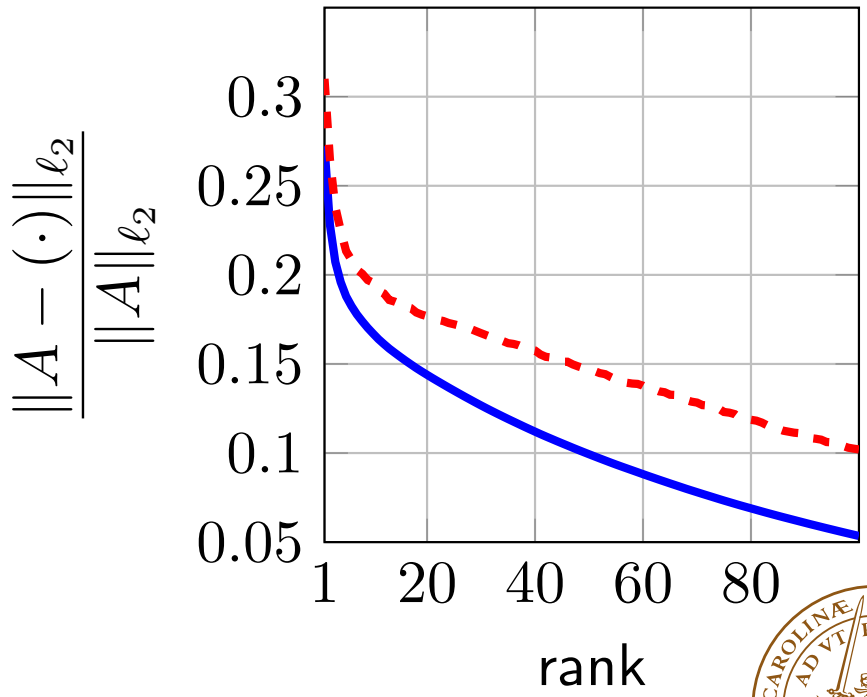
$$\begin{aligned} \underset{X}{\text{minimize}} \quad & \text{rank}(X) \\ \text{subject to} \quad & \mathcal{A}(X) = y \end{aligned} \implies \begin{aligned} \underset{X}{\text{minimize}} \quad & \|X\|_{\ell_1} \\ \text{subject to} \quad & \mathcal{A}(X) = y \end{aligned}$$

[§] Candès, Tao, Recht, Fazel, Parrilo, Chandrasekaran, ...

Baboon Approximation

$$\begin{aligned} \underset{X}{\text{minimize}} \quad & \|Y - X\|_{\ell_2}^2 + I_{\mathbb{R}_{\geq 0}^{m \times n}}(X) \\ \text{subject to} \quad & \text{rank}(X) \leq r \end{aligned}$$





$$\underset{X}{\text{minimize}} \quad k(\|X\|) + h(X) + \underbrace{\lambda \|X\|_{\ell_1}}_{\text{bias}}$$

Cons:

- Bias $\Longrightarrow$ May not solve the non-convex problem, e.g., Low-rank approximation
- No a posteriori check if the non-convex problem is solved
- Deterministic structure?
- Requires to sweep over a regularization parameter $\Rightarrow$ Cross-validation

Goal of this talk: Fix it for our problem class!

Modifications

Replace $\|\cdot\|_{\ell_1}$ with $\|\cdot\|_s$ [§]

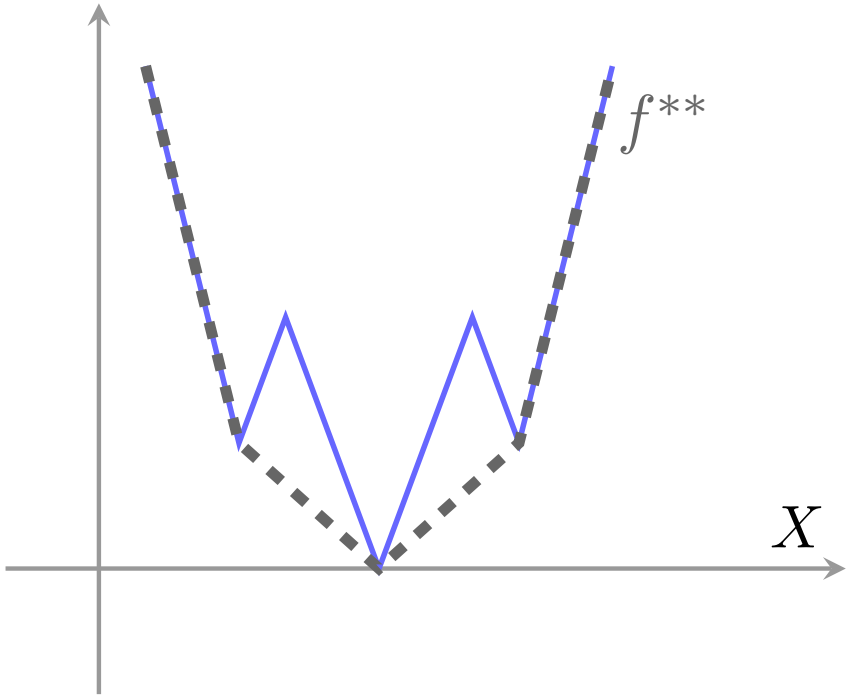
$$\underset{X}{\text{minimize}} \quad k(\|X\|) + h(X) + \underbrace{\lambda\|X\|_s}_{bias}.$$

Problem: Nothing really changed!

[§] Argyriou, Bach, Chandrasekaran, Eriksson, Mairal, Obozinski,...

Convex Envelope

$$\min_X \; f(X) \;=\; \min_X \; f^{**}(X)$$



$$f^{**}(X) = (f^*)^*(X)$$
$$f(X) \geq f^{**}(X)$$

Problem: $\left[k(\|\cdot\|) + I_{\mathsf{rank}(\cdot)\leq r} + h\right]^{**}$ unknown!

Old idea [§]

Replace $k(\|\cdot\|) + I_{\mathsf{rank}(\cdot)\leq r}(\cdot)$ with

$$\left[k(\|\cdot\|) + I_{\mathsf{rank}(\cdot)\leq r}\right]^{**}$$

Fact:

- $\left[k(\|\cdot\|) + I_{\mathsf{rank}(\cdot)\leq r}\right]^{**} = k\left(\left[\|\cdot\| + I_{\mathsf{rank}(\cdot)\leq r}\right]^{**}\right)$

[§] Lemaréchal 1973: $\min_x \sum_i f_i(x_i) \;\rightarrow\; \min_x \sum_i f_i^{**}(x_i)$

Low-Rank Inducing Norms

$$\|X\|_g := g(\sigma_1(X), \ldots, \sigma_{\min\{m,n\}}(X))$$

Example:

$$\|X\|_{\ell_2} \longrightarrow g(x) = \|x\|_{\ell_2}$$

$$\|X\|_{\ell_1} \longrightarrow g(x) = \|x\|_{\ell_1}$$

Dual norm

$$\|Y\|_{g^D} := \sup_{\|X\|_g \leq 1} \langle X, Y \rangle = g^D(\sigma_1(Y), \dots, \sigma_{\min\{m,n\}}(Y))$$

Examples:

$$\|Y\|_{\ell_2^D} = \|Y\|_{\ell_2}$$

$$\|Y\|_{\ell_1^D} = \|Y\|_{\ell_\infty} = \sigma_1(Y)$$

Truncated dual norms

$$\|Y\|_{g^D,r} := \sup_{\substack{\|X\|_g \leq 1 \\ \mathsf{rank}(X) \leq r}} \langle X, Y \rangle = \underbrace{g^D(\sigma_1(Y), \ldots, \sigma_r(Y))}_{= g^D(\sigma_1(Y), \ldots, \sigma_r(Y), 0 \ldots, 0)}$$

Examples:

$$\|Y\|_{\ell_2^D,r} = \sqrt{\sum_{i=1}^{r} \sigma_i^2(Y)}$$

$$\|Y\|_{\ell_1^D,r} = \|Y\|_{\ell_\infty}$$

Low-rank inducing norms [§]

$$\|X\|_{g,r*} := \sup_{\|Y\|_{g^D,r} \leq 1} \langle X, Y \rangle.$$

- If $\|\cdot\|_g$ SDP representable $\implies$ $\|\cdot\|_{g,r*}$ SDP repres.
- If $\text{prox}_{\|\cdot\|_g}$ computable

$$\implies \text{prox}_{\|\cdot\|_{g,r*}} \text{ computable}$$

$$\implies \text{prox}_{I_{\|\cdot\|_{g,r*} \leq t}}(\cdot, t) \text{ computable}$$

$$\implies k(\|\cdot\|_{g,r*}) = \min_t k(t) + I_{\|\cdot\|_{g,r*} \leq t}(\cdot, t)$$

Complexity for $g = \ell_2,\ \ell_\infty$: SVD + $\mathcal{O}(n \log n)$ ($n = \#$ SVs)

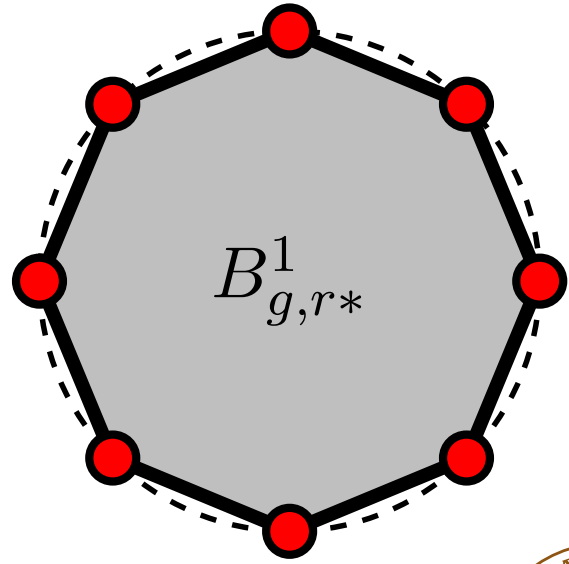
[§] Atomic norms, Overlapping norms, Support norms

Geometric Interpretation

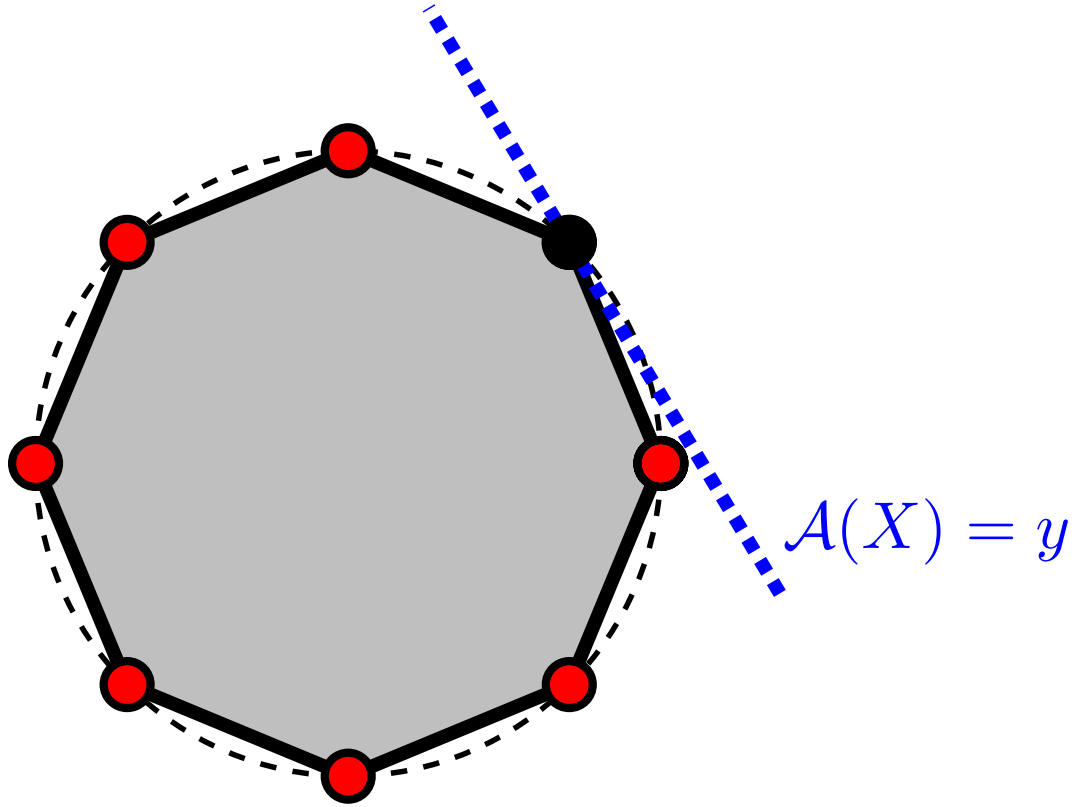
$$B^1_{g,r*} := \{X \in \mathbb{R}^{m\times n} : \|X\|_{g,r*} \leq 1\}$$

$$E_{g,r} := \{X \in \mathbb{R}^{m\times n} : \|X\|_g = 1,\ \mathsf{rank}(X) \leq r\}$$

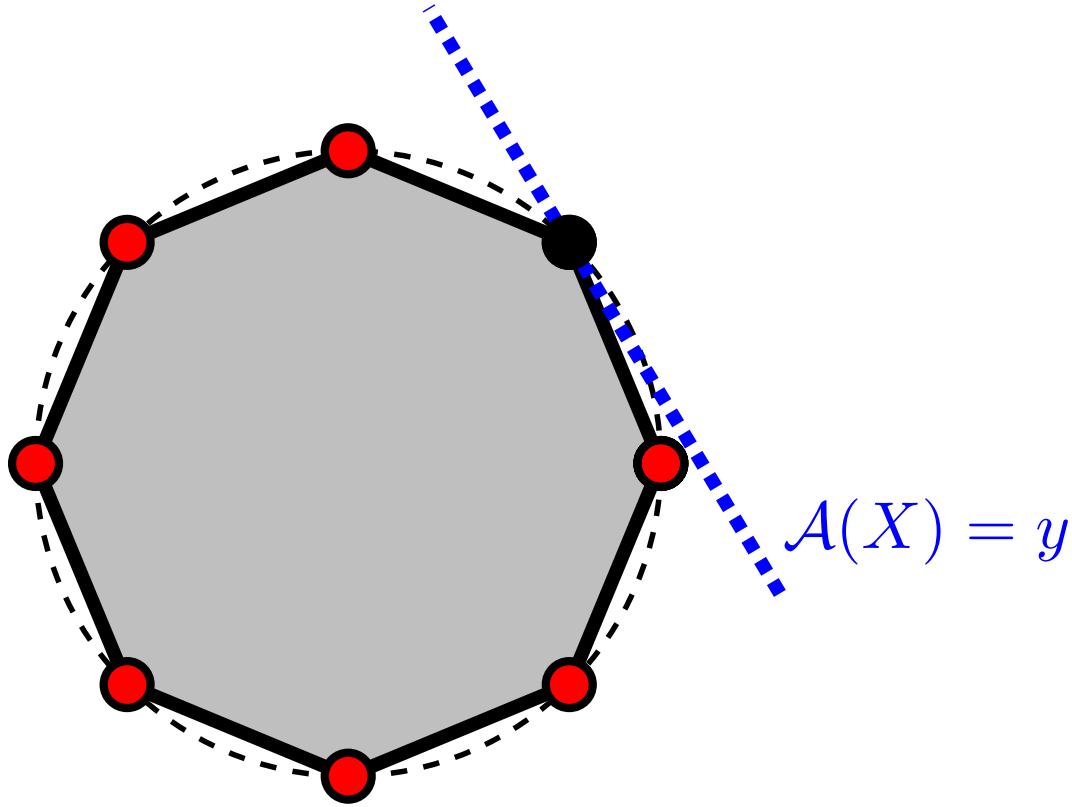
- $B^1_{g,r*} = \mathsf{conv}(E_{g,r})$
- $\|X\|_g \leq \|X\|_{g,r*}$
- $\|X\|_g = \|X\|_{g,r*},\ \mathsf{rank}(X) \leq r.$



$$\begin{aligned} \underset{X}{\text{minimize}} \quad & \|X\|_g \\ \text{subject to} \quad & \mathcal{A}(X) = y, \\ & \text{rank}(X) \leq r \end{aligned} \quad \Leftrightarrow \quad \begin{aligned} \underset{X}{\text{minimize}} \quad & \|X\|_{g,r*} \\ \text{subject to} \quad & \mathcal{A}(X) = y, \\ & \text{rank}(X) \leq r \end{aligned}$$



$$\begin{aligned} \underset{X}{\text{minimize}} \quad & \|X\|_g \\ \text{subject to} \quad & \mathcal{A}(X) = y, \\ & \text{rank}(X) \leq r \end{aligned} \quad \Leftrightarrow \quad \begin{aligned} \underset{X}{\text{minimize}} \quad & \|X\|_{g,r*} \\ \text{subject to} \quad & \mathcal{A}(X) = y, \end{aligned}$$



Best Convex Relaxation

$$\min_{\substack{X\in\mathbb{R}^{m\times n}\\ \mathsf{rank}(X)\leq r}} \left[k(\|X\|_g) + h(X)\right] \;\geq\; \min_{X\in\mathbb{R}^{m\times n}} \left[k(\|X\|_{g,r*}) + h(X)\right]$$

Best in the sense:

- $(k(\|\cdot\|_g) + I_{\mathsf{rank}(\cdot)\leq r}(\cdot) + h)^{**}$ unknown
- Simple **a posteriori test** for optimality
- Sweep over **discrete** r instead of λ

$$\Longrightarrow \textbf{ Cross-validation } \longleftrightarrow \text{ zero-duality gap}$$

Cost function replaced – NO BIAS!

Nuclear Norm

Standard interpretation:

$$\|\cdot\|_{\ell_1} = (\mathsf{rank}(\cdot) + I_{\|\cdot\|_{\ell_\infty} \leq 1})^{**}$$

Our interpretation # 1:

$$\|\cdot\|_{\ell_1} = (\|\cdot\|_{\ell_1} + I_{\mathsf{rank}(\cdot) \leq r})^{**}$$

Our interpretation # 2:

$$\|X\|_{\ell_1} = \|X\|_{g,1*} \geq \cdots \geq \|X\|_{g,r*} \geq \ldots \geq \|X\|_{g,q*} = \|X\|_g$$

$$\min_{\substack{X \in \mathbb{R}^{m \times n} \\ \mathsf{rank}(X) \leq 1}} [k(\|X\|_g) + h(X)] \geq \min_{X \in \mathbb{R}^{m \times n}} [k(\|X\|_{\ell_1}) + h(X)]$$

Some good news

- Zero-duality gap for bilinear regression

$$\begin{aligned} \underset{X \in \mathbb{R}^{k \times k}}{\text{minimize}} \quad & \|Y - L^{\mathsf{T}} X R^{\mathsf{T}}\|_{\ell_2}^2 \\ \text{subject to} \quad & \text{rank}(X) \leq r \end{aligned}$$

- Optimality interpretations, e.g., iterative re-weighting

$$\begin{aligned} \min_{\substack{X \in \mathbb{R}^{m \times n} \\ \text{rank}(X) \leq r}} \quad & [k(\|WX\|_g) + h(X)] \\ \geq \quad & \min_{X \in \mathbb{R}^{m \times n}} [k(\|WX\|_{g,r*}) + h(X)] \end{aligned}$$

- Extends to atomic sets

$$\min_{x \in \mathcal{A}} \left[k(G(x)) + h(x) \right] \;\geq\; \min_{x} \left[k(\|x\|_{\mathcal{A}_G}) + h(x) \right]$$

  - G is positively homogeneous
  - $\forall a \in \mathcal{A} \setminus \{0\} : G(a) > 0$
  - $\|x\|_{\mathcal{A}_G} = \inf\{t > 0 : t^{-1}x \in \mathsf{conv}(\mathcal{A}_G)\}$
  - $\mathcal{A}_G = \{a \in \mathsf{cone}(\mathcal{A}) : G(a) = 1\}$

Example:

$$\|\cdot\|_{\ell_2,r*} \;\longrightarrow\; G = \|\cdot\|_{\ell_2},\; \mathcal{A} = \{X : \mathsf{rank}(X) \leq r\}$$

Not bad news

$$X^\star \in \operatorname*{argmin}_X \left[k(\|X\|_{g,r*}) + h(X)\right]$$

$$Y^\star \in \operatorname*{argmin}_Y \left[k^+(\|Y\|_{g^D,r}) + h^*(Y)\right]$$

$k^+(y) := \sup_{x \geq 0}[xy - k(x)]$

- $\operatorname{rank}(X^\star) \leq r +$ uniqueness,

$$\sigma_r(Y^\star) \neq \sigma_{r+1}(Y^\star) \text{ or } \sigma_r(Y^\star) = 0$$

- $\operatorname{rank}(X^\star) \leq r + s$,

$$\sigma_r(Y^\star) = \cdots = \sigma_{r+s}(Y^\star) \neq \sigma_{r+s+1}(Y^\star)$$

Recovery Guarantees?

- Work in progress
- Why not using known tools? [§]

  Do not exploit **additional "knowledge"** provided by $\|\cdot\|_g$

[§] Chandrasekaran, Recht, Parrilo, Willsky '12

Example: Matrix Completion

Given partially known entries of a low-rank $Z \in \mathbb{R}^{m \times n}$, find the unknown entries.

Additional knowledge:

$$\begin{aligned} \underset{X}{\text{minimize}} \quad & \|X\|_g \\ \text{subject to} \quad & X_{ij} = Z_{ij}, \ (i,j) \in \mathcal{I} \\ & \text{rank}(X) \leq r \end{aligned}$$

- Small unknown entries: $k(\|\cdot\|) = \|\cdot\|_{\ell_2}$.

Example: Matrix Completion

$$H = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \in \mathbb{R}^{10\times 10}$$

$$Z := \sum_{i=1}^{5} \sigma_i(H) u_i u_i^{\mathsf{T}}$$

$$|Z_{ij} - H_{ij}| \leq \sigma_6(H) \implies \forall\ H_{ij} = 0 :\ |Z_{ij}| \leq \sigma_6(H)$$

$$Z = \begin{pmatrix} * & * & * & * & * & * & * & * & * & * \\ * & * & * & * & * & * & * & * & * & ? \\ * & * & * & * & * & * & * & * & ? & ? \\ * & * & * & * & * & * & * & ? & ? & ? \\ * & * & * & * & * & * & ? & ? & ? & ? \\ * & * & * & * & * & ? & ? & ? & ? & ? \\ * & * & * & * & ? & ? & ? & ? & ? & ? \\ * & * & * & ? & ? & ? & ? & ? & ? & ? \\ * & * & ? & ? & ? & ? & ? & ? & ? & ? \\ * & ? & ? & ? & ? & ? & ? & ? & ? & ? \end{pmatrix} \in \mathbb{R}^{10\times 10}$$

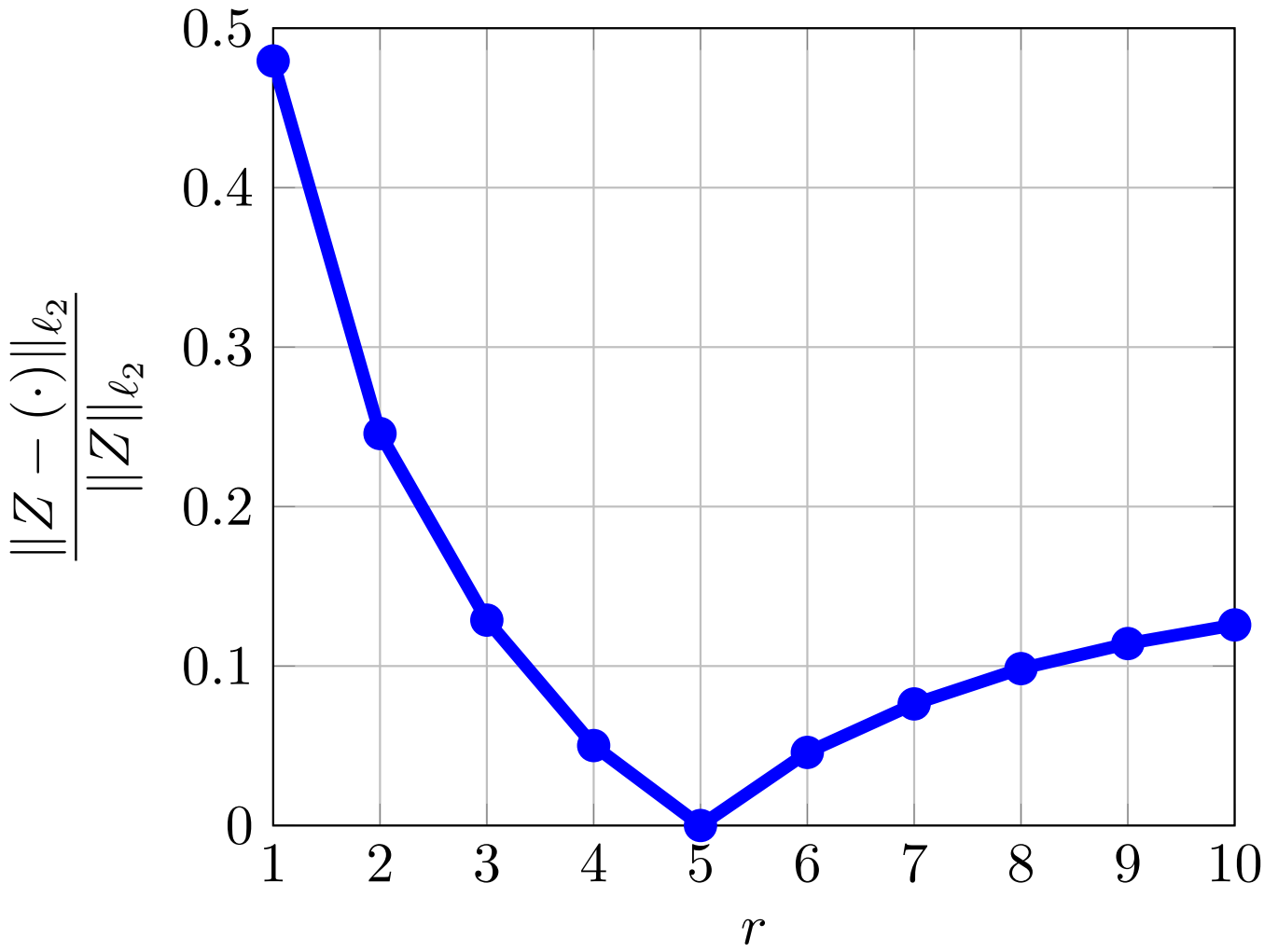
$$Z = \begin{pmatrix} * & * & * & * & * & * & * & * & * & * \\ * & * & * & * & * & * & * & * & * & ? \\ * & * & * & * & * & * & * & * & ? & ? \\ * & * & * & * & * & * & * & ? & ? & ? \\ * & * & * & * & * & * & ? & ? & ? & ? \\ * & * & * & * & * & ? & ? & ? & ? & ? \\ * & * & * & * & ? & ? & ? & ? & ? & ? \\ * & * & * & ? & ? & ? & ? & ? & ? & ? \\ * & * & ? & ? & ? & ? & ? & ? & ? & ? \\ * & ? & ? & ? & ? & ? & ? & ? & ? & ? \end{pmatrix} \in \mathbb{R}^{10\times 10}$$

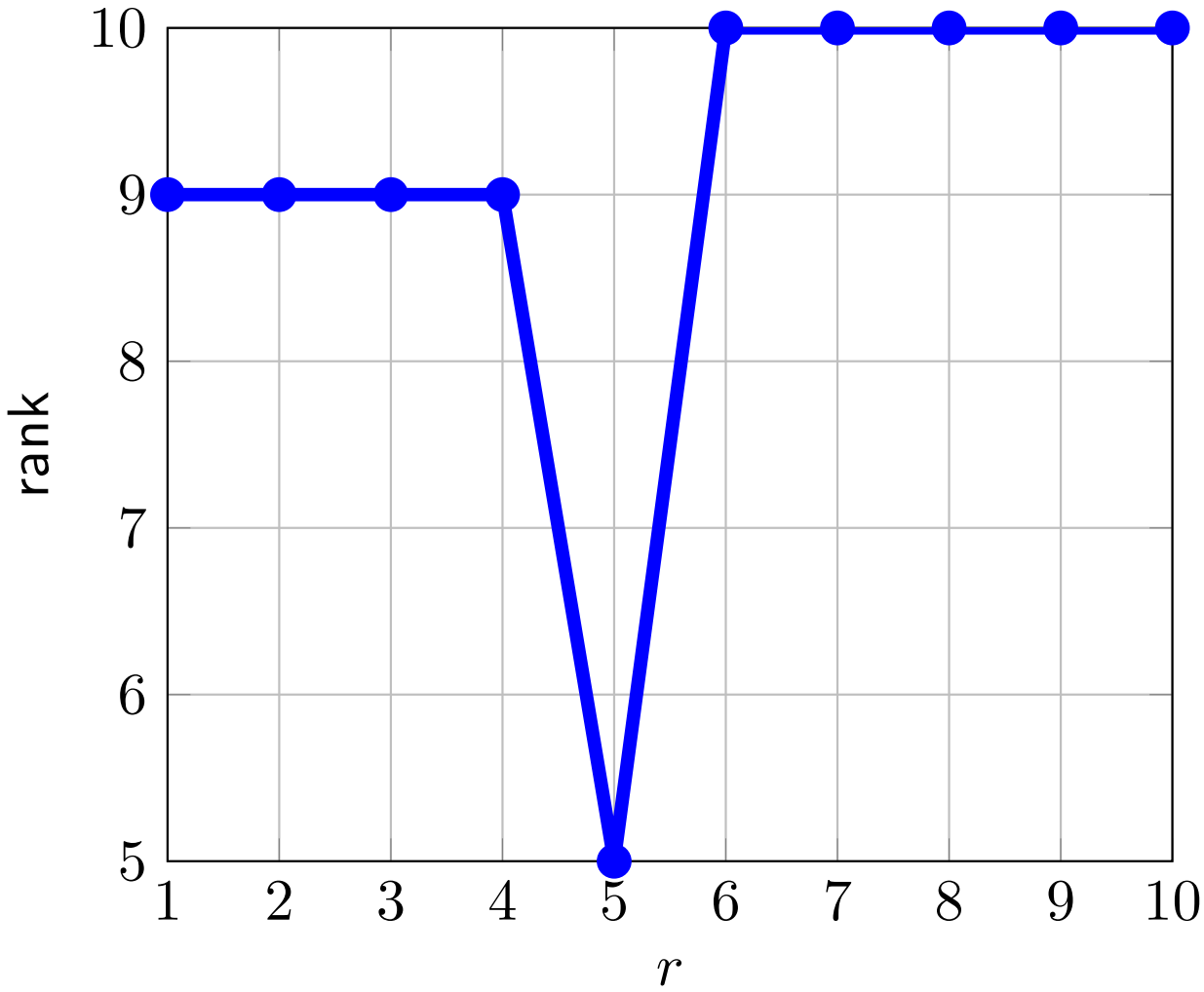
- **45 unknown entries (not randomly selected!)**
- Recovery guarantees with nuclear norm [§]:

$$3r(2n-r)+1 = \textbf{226} \text{ random Gaussian samples}$$

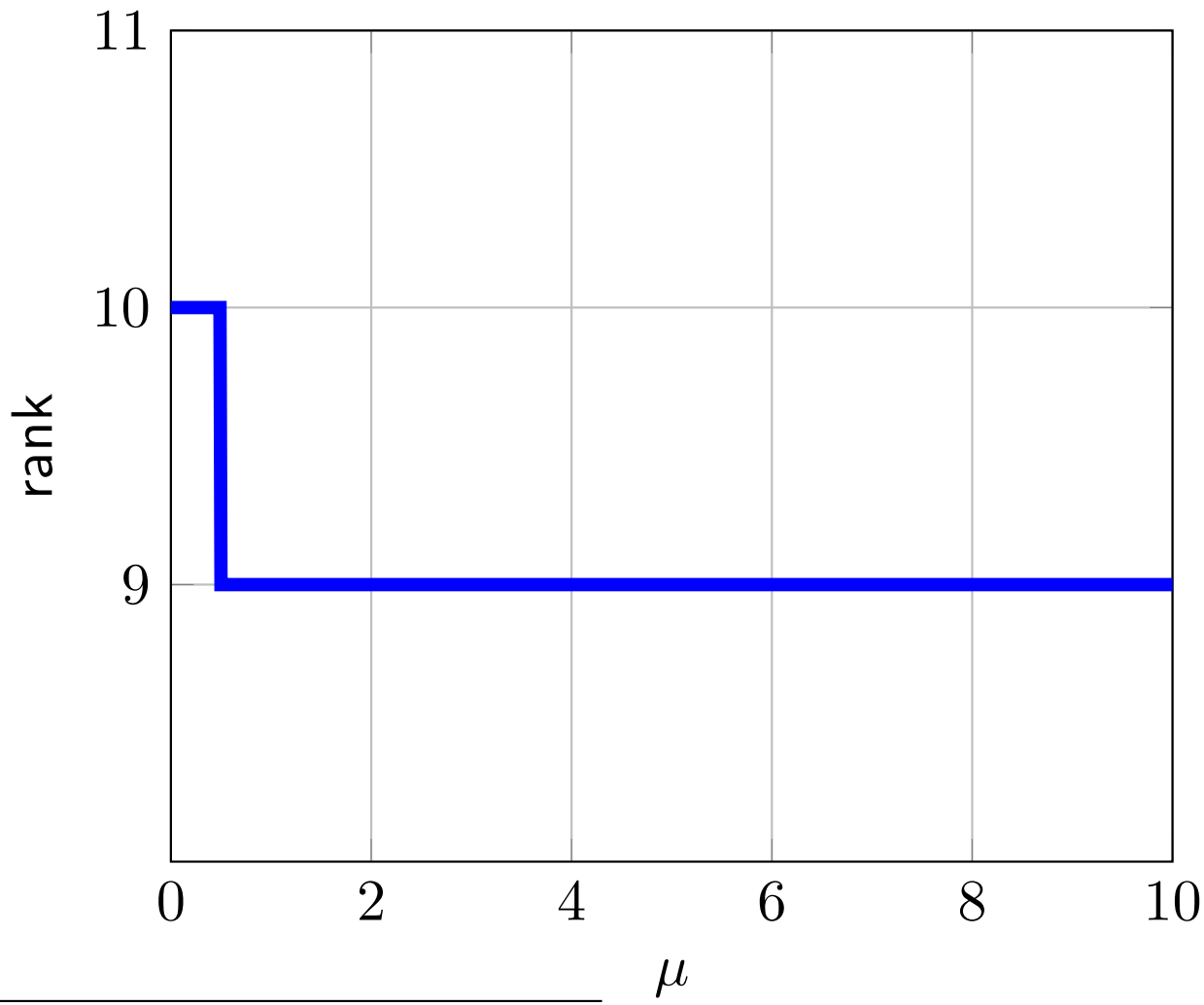
[§] Chandrasekaran, Recht, Parrilo, Willsky '12

$$\begin{aligned} \underset{X}{\text{minimize}} \quad & \|X\|_{\ell_2,r*} \\ \text{subject to} \quad & \forall (i,j) \in \mathcal{I} : X_{ij} = Z_{ij}. \end{aligned}$$




10
9
8
7
6
5
rank
1
2
3
4
5
6
7
8
9
10
r

$$\begin{aligned} \underset{X}{\text{minimize}} \quad & \frac{1}{2}\|X\|_{\ell_2}^2 + \mu\|X\|_{\ell_1} \\ \text{subject to} \quad & X_{ij} = Z_{ij}, \ (i,j) \in \mathcal{I} \end{aligned}$$



[§] Cai, Candès '10

Conclusion

- Simple a posterior test for optimality
- Prior information can/should be utilized

  $\Longrightarrow$ Model the non-convex problem
- Handles structured measurements
- Can be used to test performance of greedy methods

Most important: Replace – Don't add!

What I did not show you

- One can let r become real-valued through defining:

$$\|X\|_{g^D,r} = g^D(\sigma_1(X), \dots, \sigma_{\lfloor r \rfloor}(X), (r - \lfloor r \rfloor)\sigma_{\lceil r \rceil}(X)).$$

- Non-convex proximal splitting: $X_k = \text{prox}_{\gamma f_1}(Z_k)$

$$f_1 = k(\|\cdot\|_g) + I_{\text{rank}(\cdot) \leq r}$$

  - $\sigma_r(Y^\star) \neq \sigma_{r+1}(Y^\star)$: Local convergence to global minima
  - All $\sigma_r(Y^\star) \neq \sigma_{r+1}(Y^\star)$: All stationary points correspond to global minima ($\Longrightarrow$ Panos: global convergence)

Future Work

- Application to more control problems (Anders H., Mihailo)
- Can we learn a suitable norm? (Yong Sheng?)
- A priori deterministic and probabilistic guarantees (?)

Sources

- Low-Rank Inducing Norms with Optimality Interpretations.
- Low-Rank Optimization with Convex Constraints.
- PhD-thesis: Rank Reducing with Convex Constraints.
- The Use of the $r*$ Heuristic in Covariance Completion Problems.
- Local Convergence of Proximal Splitting Methods for Rank Constrained Problems

Questions?