A New Perspective on Low-Rank Optimization

Ryan Cory-Wright
MIT, Operations Research Center
Paper available at: ryancorywright.github.io
Questions? Email ryancw@mit.edu

Joint work with
Dimitris Bertsimas (MIT)
Jean Pauphilet (LBS)

Motivating Example: A Tale of Two Problems

Sparse Linear Regression:

$$\min_{\boldsymbol{w}\in\mathbb{R}^p} \quad \frac{1}{2n}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}\|_2^2+\frac{1}{2\gamma}\|\boldsymbol{w}\|_2^2+\mu\|\boldsymbol{w}\|_0$$

Routinely solved to optimality at scale, because convex relaxation well-studied.

Reduced Rank Regression:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{p\times n}} \quad \frac{1}{2m}\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\|_F^2+\frac{1}{2\gamma}\|\boldsymbol{\beta}\|_F^2+\mu\cdot\mathrm{Rank}(\boldsymbol{\beta})$$

Almost the same problem, but intractable, because no known relaxations give valid bounds.

We show: How to get strong & scalable relaxations.

Motivating Example: Linear Regression and Relaxations

Sparse Linear Regression: Fit interpretable model using small number of features :

$$\min_{\boldsymbol{w}\in\mathbb{R}^p} \quad \frac{1}{2n}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}\|_2^2+\frac{1}{2\gamma}\|\boldsymbol{w}\|_2^2+\mu\|\boldsymbol{w}\|_0$$

Perspective Relaxation:

$$\min_{\boldsymbol{w},\boldsymbol{\rho}\in\mathbb{R}^p,\boldsymbol{z}\in\{0,1\}^p} \quad \frac{1}{2n}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}\|_2^2+\frac{1}{2\gamma}\boldsymbol{e}^\top\boldsymbol{\rho}+\mu\cdot\boldsymbol{e}^\top\boldsymbol{z} \quad \text{s.t.} \quad z_i\rho_i\geq w_i^2 \quad \forall i\in[p].$$

Relaxation due to Frangioni and Gentile 2006, Günlük and Linderoth 2010: strong & scalable.

Allows exact solutions with $p=10^7$ features (Bertsimas and van Parys 2020, Hazimeh and Mazumder 2021)

Motivating Example: Rank Regression and Relaxations

Reduced Rank Regression: Fit interpretable model using small number of singular values

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{p\times n}} \quad \frac{1}{2m}\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\|_F^2+\frac{1}{2\gamma}\|\boldsymbol{\beta}\|_F^2+\mu\cdot\mathrm{Rank}(\boldsymbol{\beta})$$

Matrix Perspective Relaxation (new):

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{p\times n},\boldsymbol{W}\in\mathcal{S}_+^n,\boldsymbol{\theta}\in S_+^p} \quad \frac{1}{2m}\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\|_F^2+\frac{1}{2\gamma}\mathrm{tr}(\boldsymbol{\theta})+\mu\cdot\mathrm{tr}(\boldsymbol{W}) \quad \text{s.t.} \quad \boldsymbol{W}\preceq\mathbb{I},\begin{pmatrix}\boldsymbol{\theta} & \boldsymbol{\beta}\\ \boldsymbol{\beta}^\top & \boldsymbol{W}\end{pmatrix}\succeq\mathbf{0}.$$

In this talk, we derive this relaxation.

A generalization of perspective reformulation technique due to Frangioni and Gentile ’06.

Even Stronger Relaxations

Saddle-Point Sparsity Relaxation (Dong et al. '15): Also apply to diagonal "extracted" from $X^T X$

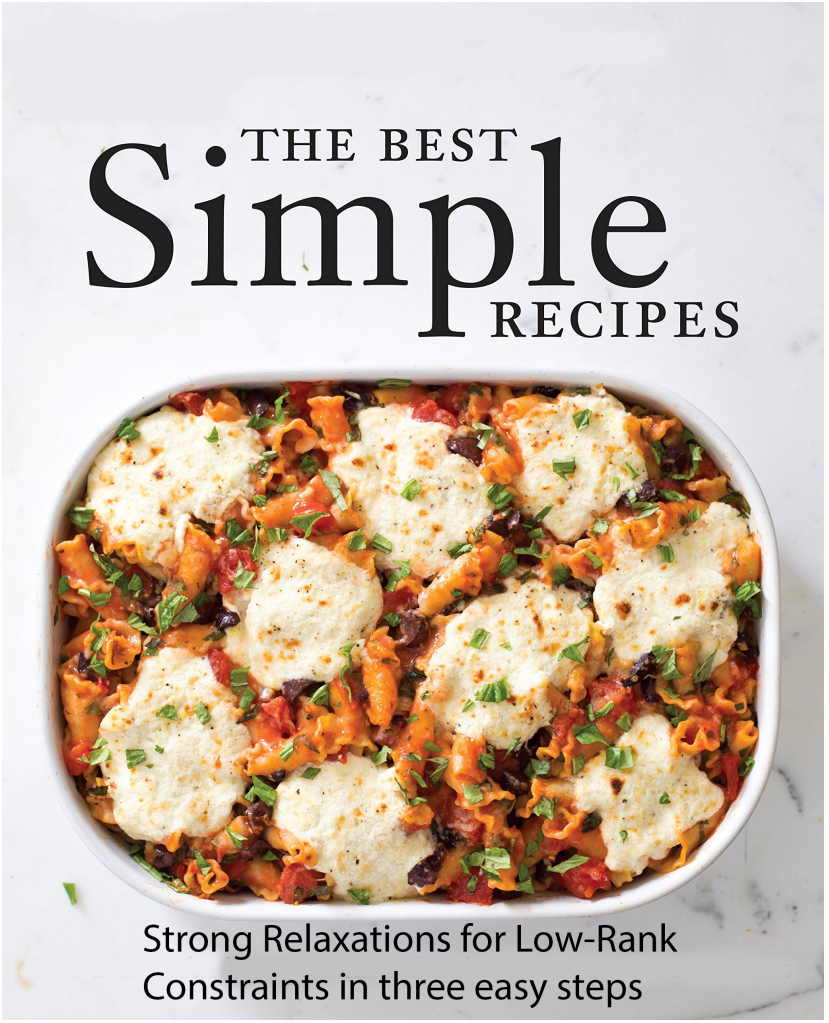
$$\begin{aligned} \min_{\boldsymbol{w}\in\mathbb{R}^p, \boldsymbol{z}\in[0,1]^p, \boldsymbol{W}\in S_+^p} \quad & \frac{1}{2n}\|\boldsymbol{y}\|_2^2 - \frac{1}{n}\langle \boldsymbol{y}, \boldsymbol{X}\boldsymbol{w}\rangle + \frac{1}{2}\langle \boldsymbol{W}, \frac{1}{\gamma}\mathbb{I} + \frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X}\rangle + \mu \boldsymbol{e}^\top \boldsymbol{z} \\ \text{s.t.} \quad & \boldsymbol{W} \succeq \boldsymbol{w}\boldsymbol{w}^\top, \ z_i W_{i,i} \geq w_i^2 \ \forall i \in [p]. \end{aligned}$$

Saddle-Point Rank Relaxation (new): Play same game in low-rank case

$$\begin{aligned} \min_{\boldsymbol{\theta}\in\mathcal{S}_+^n, \boldsymbol{\beta}\in\mathbb{R}^{p\times n}, \boldsymbol{B}\in\mathcal{S}_+^n, \boldsymbol{W}\in\mathcal{S}_+^n} \quad & \frac{1}{2m}\|\boldsymbol{Y}\|_F^2 - \frac{1}{m}\langle \boldsymbol{Y}, \boldsymbol{X}\boldsymbol{\beta}\rangle + \frac{1}{2}\langle \boldsymbol{B}, \frac{1}{\gamma}\mathbb{I} + \frac{1}{m}\boldsymbol{X}^\top \boldsymbol{X}\rangle + \mu \cdot \mathrm{tr}(\boldsymbol{W}) \\ \text{s.t.} \quad & \begin{pmatrix} \boldsymbol{B} & \boldsymbol{\beta} \\ \boldsymbol{\beta} & \boldsymbol{W} \end{pmatrix} \succeq \boldsymbol{0}, \boldsymbol{W} \preceq \mathbb{I}. \end{aligned}$$

Matrix Perspective Reformulation Technique: Recipe

1. Consider low-rank problem with spectral regularization
2. Formulate as mixed-projection optimization problem
3. Take matrix perspective of spectral regularizer.



Contributions

- Methodological: We propose a simple preprocessing technique which gives strong & scalable bounds for low-rank problems. Generalizes perspective reformulation technique from MIO.
- Theoretical: We invoke technique to explicitly characterize convex hulls of simple low-rank sets
- Algorithmic: We demonstrate technique's efficacy across diverse range of low-rank problems

Matrix Perspective Reformulation Technique I

Consider low-rank problem with spectral regularization

$$\min_{\boldsymbol{X}\in\mathcal{S}_+^n} \langle \boldsymbol{C}, \boldsymbol{X}\rangle + \Omega(\boldsymbol{X}) + \boxed{\mu \cdot \mathrm{Rank}(\boldsymbol{X})} \text{ s.t. } \langle \boldsymbol{A}_i, \boldsymbol{X}\rangle = b_i \ \forall i \in [m], \ \boldsymbol{X} \in \mathcal{K}, \ \boxed{\mathrm{Rank}(\boldsymbol{X}) \leq k},$$

Where:

- Rank ➡ ***X*** of minimal complexity,

Matrix Perspective Reformulation Technique I

Consider low-rank problem with spectral regularization

$$\min_{\boldsymbol{X}\in\mathcal{S}_+^n} \langle \boldsymbol{C}, \boldsymbol{X}\rangle + \Omega(\boldsymbol{X}) + \mu \cdot \mathrm{Rank}(\boldsymbol{X}) \text{ s.t. } \langle \boldsymbol{A}_i, \boldsymbol{X}\rangle = b_i \ \forall i \in [m],\ \boldsymbol{X} \in \mathcal{K},\ \mathrm{Rank}(\boldsymbol{X}) \leq k,$$

Where:

- $\Omega(\boldsymbol{X}) := \sum_{i=1}^{n} \omega\left(\lambda_i(\boldsymbol{X})\right)$ convex spectral function with $0 \in \omega$; ω closed convex.
- Example: $\Omega(X) = \|X\|_F^2$; like ridge regularization in regression.
- Example: $\Omega(X) = 0$ if $\|X\|_*$<= M, +∞ otherwise; like big-M penalty in regression.

Matrix Perspective Reformulation Technique II: Formulation

Low-rank problem

$$\min_{\boldsymbol{X}\in\mathcal{S}_+^n} \langle \boldsymbol{C}, \boldsymbol{X}\rangle + \Omega(\boldsymbol{X}) + \mu \cdot \mathrm{Rank}(\boldsymbol{X}) \text{ s.t. } \langle \boldsymbol{A}_i, \boldsymbol{X}\rangle = b_i \ \forall i \in [m],\ \boldsymbol{X} \in \mathcal{K},\ \mathrm{Rank}(\boldsymbol{X}) \leq k,$$

can be expressed as Mixed-Projection Optimization problem

$$\begin{aligned} \min_{\boldsymbol{Y}\in\mathcal{Y}_n^k} \min_{\boldsymbol{X}\in\mathcal{S}_+^n} \quad & \langle \boldsymbol{C}, \boldsymbol{X}\rangle + \mu \cdot \mathrm{tr}(\boldsymbol{Y}) + \mathrm{tr}(f(\boldsymbol{X})) \\ \text{s.t.} \quad & \langle \boldsymbol{A}_i, \boldsymbol{X}\rangle = b_i \quad \forall i \in [m],\ \boldsymbol{X} = \boldsymbol{Y}\boldsymbol{X},\ \boldsymbol{X} \in \mathcal{K} \end{aligned}$$

where $\boldsymbol{Y}$ is a projection matrix, f is a spectral function such that $\mathrm{tr}(\mathrm{f}(\mathrm{X})) = \Omega(X)$.

Matrix Perspective Reformulation Technique III: Reformulation

Mixed-Projection Conic Optimization problem

$$\begin{aligned} \min_{\boldsymbol{Y}\in\mathcal{Y}_n^k} \min_{\boldsymbol{X}\in\mathcal{S}_+^n} \quad & \langle \boldsymbol{C}, \boldsymbol{X}\rangle + \mu \cdot \mathrm{tr}(\boldsymbol{Y}) + \mathrm{tr}(f(\boldsymbol{X})) \\ \text{s.t.} \quad & \langle \boldsymbol{A}_i, \boldsymbol{X}\rangle = b_i \quad \forall i \in [m],\ \boldsymbol{X} = \boldsymbol{Y}\boldsymbol{X},\ \boldsymbol{X} \in \mathcal{K} \end{aligned}$$

Rewrite as equivalent problem which gives stronger relaxations:

$$\begin{aligned} \min_{\boldsymbol{Y}\in\mathcal{Y}_n^k} \min_{\boldsymbol{X}\in\mathcal{S}_+^n} \quad & \langle \boldsymbol{C}, \boldsymbol{X}\rangle + \mu \cdot \mathrm{tr}(\boldsymbol{Y}) + \mathrm{tr}(g_f(\boldsymbol{X}, \boldsymbol{Y})) + (n - \mathrm{tr}(\boldsymbol{Y}))\omega(0) \\ \text{s.t.} \quad & \langle \boldsymbol{A}_i, \boldsymbol{X}\rangle = b_i \quad \forall i \in [m],\ \boldsymbol{X} \in \mathcal{K}, \end{aligned}$$

where g_f is the matrix perspective of f (Effros, PNAS, 2009; Ebadian et al., PNAS, 2011)

$$g_f(\boldsymbol{X}, \boldsymbol{Y}) = \begin{cases} \boldsymbol{Y}^{\frac{1}{2}} f(\boldsymbol{Y}^{\dagger}\boldsymbol{X})\boldsymbol{Y}^{\frac{1}{2}} & \text{if } \boldsymbol{Y}^{\dagger}\boldsymbol{X} \in \mathcal{X}, \boldsymbol{Y} \succeq \boldsymbol{0}, \\ +\infty & \text{otherwise}, \end{cases}$$

Matrix Perspective Reformulation Technique III: Reformulation

Mixed-Projection Conic Optimization problem

$$\begin{aligned} \min_{\boldsymbol{Y}\in\mathcal{Y}_n^k} \min_{\boldsymbol{X}\in\mathcal{S}_+^n} \quad & \langle \boldsymbol{C}, \boldsymbol{X}\rangle + \mu \cdot \mathrm{tr}(\boldsymbol{Y}) + \mathrm{tr}(f(\boldsymbol{X})) \\ \text{s.t.} \quad & \langle \boldsymbol{A}_i, \boldsymbol{X}\rangle = b_i \quad \forall i \in [m],\ \boldsymbol{X} = \boldsymbol{Y}\boldsymbol{X},\ \boldsymbol{X} \in \mathcal{K} \end{aligned}$$

Rewrite as equivalent problem which gives stronger relaxations:

$$\begin{aligned} \min_{\boldsymbol{Y}\in\mathcal{Y}_n^k} \min_{\boldsymbol{X}\in\mathcal{S}_+^n} \quad & \langle \boldsymbol{C}, \boldsymbol{X}\rangle + \mu \cdot \mathrm{tr}(\boldsymbol{Y}) + \mathrm{tr}(g_f(\boldsymbol{X}, \boldsymbol{Y})) + (n - \mathrm{tr}(\boldsymbol{Y}))\omega(0) \\ \text{s.t.} \quad & \langle \boldsymbol{A}_i, \boldsymbol{X}\rangle = b_i \quad \forall i \in [m],\ \boldsymbol{X} \in \mathcal{K}, \end{aligned}$$

equivalent to applying perspective reformulation to eigenvalues of X, Y.

Can use relaxation to get feasible solutions!

Feasible Solutions via Greedily Rounding Semidefinite Relaxation

Take dual with respect to X, obtain saddle-point problem:

$$\min_{\boldsymbol{Y}\in \mathrm{Conv}(\mathcal{Y}_n^k)} \max_{\boldsymbol{\alpha},\boldsymbol{V}_{11},\boldsymbol{V}_{22}\in S^m} \lambda\cdot \mathrm{tr}(\boldsymbol{Y}) + h(\boldsymbol{\alpha}) - \Omega^\star\left(\boldsymbol{\alpha},\boldsymbol{Y},\boldsymbol{V}_{11},\boldsymbol{V}_{22}\right).$$

- Greedily rounding relaxation Y* using SVD gives projection matrix Y. How good is it?

THEOREM 7. *Let $\boldsymbol{Y}^\star$ denote a solution to the semidefinite relaxation (19), $\boldsymbol{Y}^\star = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ be a singular value decomposition of $\boldsymbol{Y}^\star$, $\mathcal{R}$ denote the indices of strictly fractional diagonal entries in $\boldsymbol{\Lambda}$, and $\boldsymbol{\alpha}^\star(\boldsymbol{Y})$ denote a best choice of $\boldsymbol{\alpha}$ for a given $\boldsymbol{Y}$. Suppose that for any $\boldsymbol{Y}\in\mathcal{Y}_k^n$, we have $\sigma_{\max}(\boldsymbol{\alpha}^\star(\boldsymbol{Y}))\le L$. Then, a greedy rounding of $\boldsymbol{Y}^\star$, i.e., $\boldsymbol{Y}_{greedy} = \boldsymbol{U}\boldsymbol{\Lambda}_{greedy}\boldsymbol{U}^\top$, where $\boldsymbol{\Lambda}_{greedy}$ is a diagonal matrix such that $\Lambda_{i,i}=1$ for the k highest diagonal coefficients in $\boldsymbol{\Lambda}^\star$, satisfies $0\le f(\boldsymbol{Y}_{greedy}) - f(\boldsymbol{Y}^\star)\le\epsilon$, where*

- $\epsilon = ML\min(|\mathcal{R}|, n-k)$ *for the spectral penalty.*
- $\epsilon = \frac{\gamma}{2}\min(|\mathcal{R}|, n-k)L^2$ *for the Frobenius penalty.*

Feasible Solutions via Greedily Rounding Semidefinite Relaxation

Take dual with respect to X, obtain saddle-point problem:

$$\min_{\boldsymbol{Y}\in \mathrm{Conv}(\mathcal{Y}_n^k)} \max_{\boldsymbol{\alpha},\boldsymbol{V}_{11},\boldsymbol{V}_{22}\in S^m} \lambda\cdot \mathrm{tr}(\boldsymbol{Y}) + h(\boldsymbol{\alpha}) - \Omega^\star\left(\boldsymbol{\alpha},\boldsymbol{Y},\boldsymbol{V}_{11},\boldsymbol{V}_{22}\right).$$

- Greedily rounding relaxation Y* using SVD gives projection matrix Y. How good is it?
- Answer: Near-exact in theory, often exact in practice.

Matrix Perspective Reformulation: Worked Example

Reduced Rank Regression: Fit interpretable model using small number of singular values

Step 1: Consider problem with spectral regularization:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{p\times n}} \quad \frac{1}{2m}\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\|_F^2 + \frac{1}{2\gamma}\|\boldsymbol{\beta}\|_F^2 + \mu\cdot\mathrm{Rank}(\boldsymbol{\beta})$$

Where $\Omega(X) = \frac{1}{2\gamma}\sum_{i=1}^{n}\lambda_i(\beta)^2 = \frac{1}{2\gamma}\|\beta\|_F^2$

Matrix Perspective Reformulation: Worked Example

Reduced Rank Regression: Fit interpretable model using small number of singular values

Step 2: Formulate as Mixed-Projection problem

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{p\times n},\boldsymbol{W}\in\mathcal{Y}_n^n} \quad \frac{1}{2m}\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\|_F^2+\frac{1}{2\gamma}\|\boldsymbol{\beta}\|_F^2+\mu\cdot\mathrm{tr}(\boldsymbol{W}),\boldsymbol{W}=\boldsymbol{\beta}\boldsymbol{W}$$

Where $\mathcal{Y}_n := \{\mathbf{P}\in S^n : \mathbf{P}^2=\mathbf{P}\}$ is set of $n\times n$ projection matrices

Matrix Perspective Reformulation: Worked Example

Reduced Rank Regression: Fit interpretable model using small number of singular values

Step 3: Reformulate by taking matrix perspective

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{p\times n},\boldsymbol{W}\in\mathcal{S}_+^n,\boldsymbol{\theta}\in S_+^p} \quad \frac{1}{2m}\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\|_F^2 + \boxed{\frac{1}{2\gamma}\mathrm{tr}(\boldsymbol{\theta})} + \mu\cdot\mathrm{tr}(\boldsymbol{W}) \quad \text{s.t.} \quad \boldsymbol{W}\preceq\mathbb{I}, \boxed{\begin{pmatrix}\boldsymbol{\theta} & \boldsymbol{\beta}\\ \boldsymbol{\beta}^\top & \boldsymbol{W}\end{pmatrix}\succeq\mathbf{0}}$$

Theoretical Contribution: Convex Hulls of Simple Low-Rank Sets

Bertsimas, Cory-Wright, and Pauphilet (21+): Theorem 2

Let T denote epigraph of spectral function under rank constraints:

$$\mathcal{T} = \left\{ \boldsymbol{X} \in \mathcal{S}_+^n : \mathrm{tr}(f(\boldsymbol{X})) + \mu \cdot \mathrm{Rank}(\boldsymbol{X}) \leq t, \mathrm{Rank}(\boldsymbol{X}) \leq k \right\}$$

$\omega(\cdot)$ scalar convex function such that $tr(f(X)) = \sum_{i=1}^{n} \omega(\lambda_i(X))$

Then, convex hull of T given by:

$$\mathcal{T}^c = \left\{ (\boldsymbol{X}, \boldsymbol{Y}) \in \mathcal{S}_+^n \times \mathrm{Conv}(\mathcal{Y}_n^k) : \mathrm{tr}(g_f(\boldsymbol{X}, \boldsymbol{Y})) + \mu \cdot \mathrm{tr}(\boldsymbol{Y}) + (n - \mathrm{tr}(\boldsymbol{Y}))\omega(0) \leq t \right\}$$

Where:

- g_f matrix perspective of f
- $\mathrm{Conv}(\mathcal{Y}_n^k) = \{ \boldsymbol{Y} \in S_+^n : \boldsymbol{Y} \preceq \mathbb{I}, \mathrm{tr}(\boldsymbol{Y}) \leq k \}$ is convex hull of rank-k projection matrices.

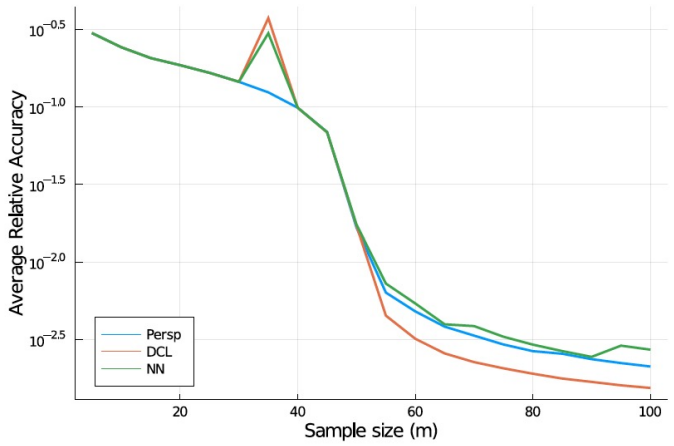
Matrix perspective reformulation gives convex hull of simple low-rank sets

Application I: Reduced Rank Regression

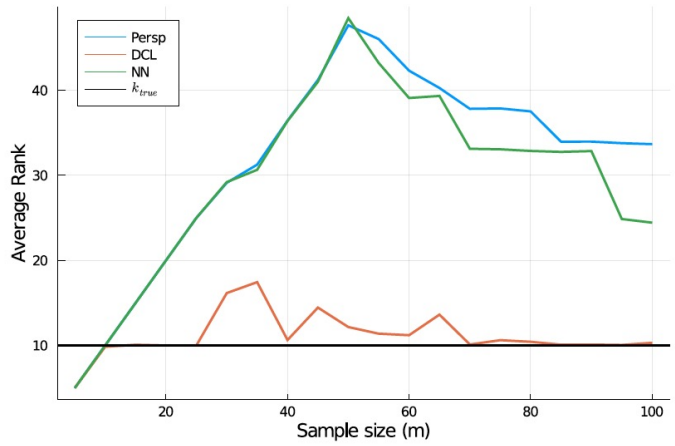
Example:

Recover rank-10 50 x m matrix:

- Vary m
- Measure MSE, rank from relaxations
- Compare against nuclear norm
- Matrix perspective dominates nuclear norm
- Saddle-point much more accurate and much faster than matrix perspective or NN, recovers true rank rapidly.
- Saddle-point w. Mosek solves for 300x300 matrices on Macbook Pro in minutes, nuclear norm takes hours for 150x150.
- Code available on GitHub: ryancorywright/MatrixPerspectiveSoftware



(a) Accuracy



(b) Rank

In practice, saddle-point relaxation *right* way to solve reduced rank regression.

Application II: Matrix Completion

Formulation:

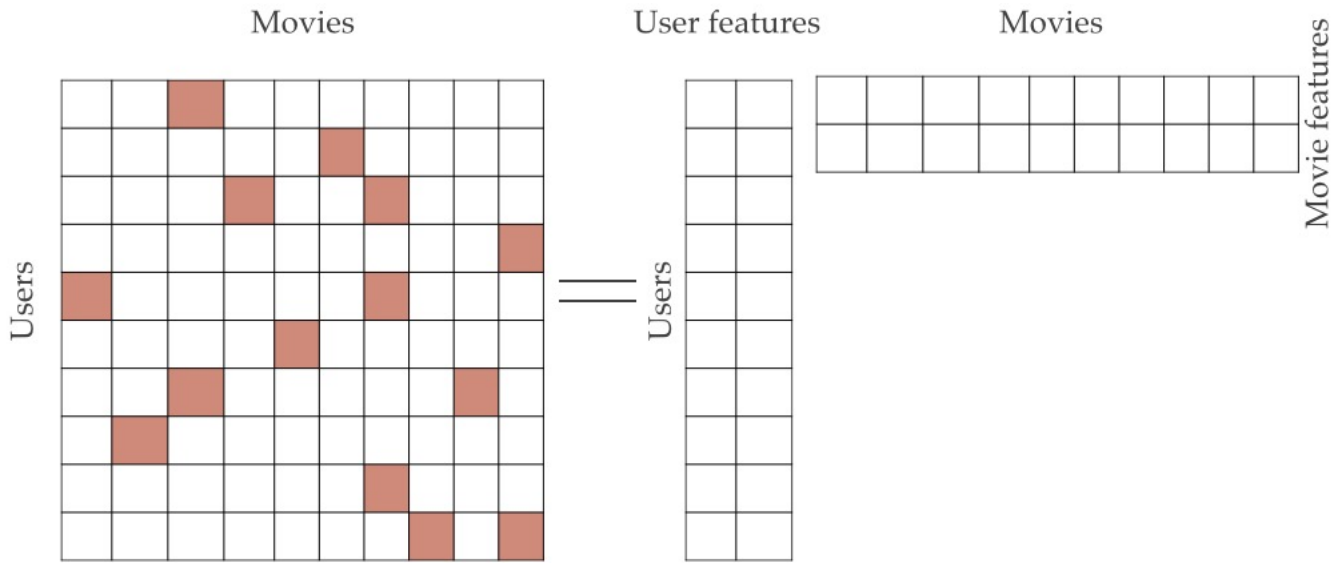
$$\min_{\boldsymbol{X}\in\mathbb{R}^{n\times p}} \quad \frac{1}{2}\sum_{(i,j)\in\mathcal{I}}(X_{i,j}-A_{i,j})^2 \quad \text{s.t.} \quad \text{Rank}(\boldsymbol{X})\leq k.$$

Decision variables/Problem data

$X_{i,j}$: Predicted rating movie *j* by user *i*
$A_{i,j}$: Reported rating movie *j* by user *i*

Movie Recommendation:

- Given user movie ratings, predict ratings for unseen movies.
- To make problem tractable, assume ratings depend on k factors (lead actor, lead actress, director, genre, year, ..)

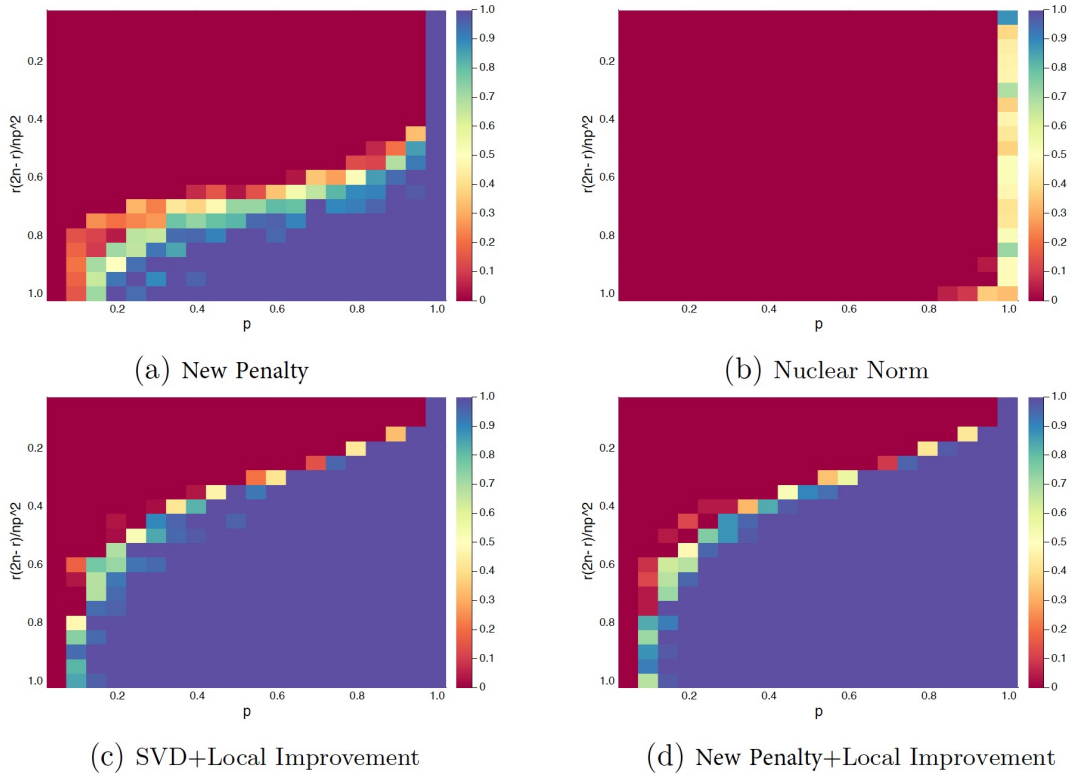


Application II: Matrix Completion

Example:

Recover low-rank 100x100 matrix:

- Vary rank, proportion entries sampled
- Measure % time recover matrix to 1% MSE (more purple=better)
- Nuclear norm *by far* worst approach
- New penalty better, new penalty with rounding much better
- Combining greedy rounding with local search: solutions within 1% of optimality in practice, 10% in theory.
- Code available on GitHub: ryancorywright/MixedProjectionSoftware



(a) New Penalty (b) Nuclear Norm

(c) SVD+Local Improvement (d) New Penalty+Local Improvement

In practice, new penalty is viable and *often* more accurate

Details in: *Mixed-Projection Conic Optimization: A New Paradigm for Modeling Rank Constraints*
D. Bertsimas, **R. Cory-Wright**, J. Pauphilet, Operations Research, 2021.
Winner, 2020 INFORMS George Nicholson Best Paper Competition

Technique Works for Many Other Applications!

Machine Learning	Decompose matrix into sparse plus low-rank matrix • See Nicholas Johnson's talk later in the session
Control Theory	Identify minimum complexity realization of a system
Algebraic Geometry	Minimum degree sum-of-squares decomposition of polynomial
Computer Vision	Decompose a matrix into a product of non-negative factors
Many, many more…	

Conclusion

Matrix perspective is natural generalization of perspective reformulation

- Exploit separability of eigenvalues to obtain "embarrassingly tight" formulation.
- Leads to relaxations which outperform state-of-the-art for central problems in OR/ML.
- Suggests this is a very general story, often useful to think about problems this way.

Two future directions:

1. Improve generality ➡ extend framework to tensors, non-negative polynomials.
2. Develop rounding ➡ find "best" rounding strategy for matrix perspective relax

Thank you for listening!
Lingering questions? Email ryancw@mit.edu