Low-rank Matrix Completion via Convex Optimization

Ben Recht
Center for the Mathematics of Information
Caltech

Recommender Systems

amazon.com

NETFLIX

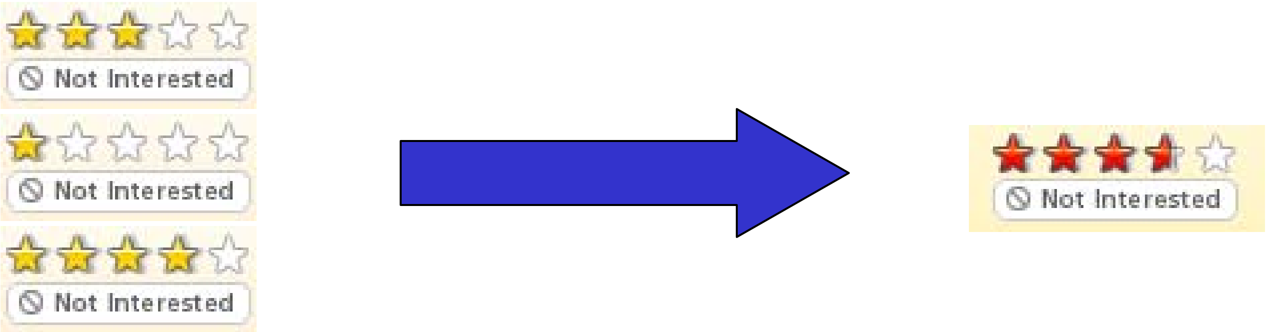
match.com

chemistry

eHarmony

Netflix Prize

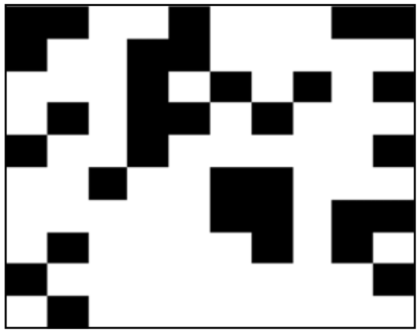
- One million big ones!
- Given 100 million ratings on a scale of 1 to 5, predict 3 million ratings to highest accuracy



- 17770 total movies x 480189 total users
- Over 8 billion total ratings
- How to fill in the blanks?

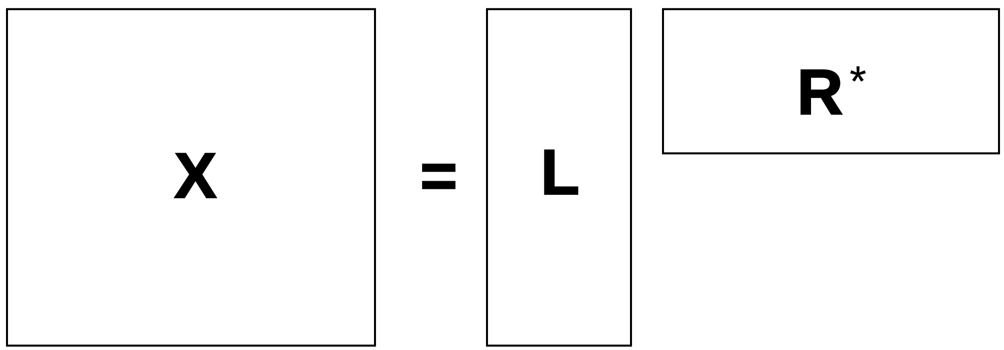
Abstract Setup: Matrix Completion

X =



X_{ij} known for black cells
X_{ij} unknown for white cells
Rows index movies
Columns index users

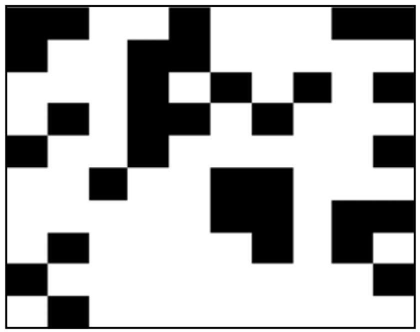
- How do you fill in the missing data?



k x n   k x r   r x n

kn entries   r(k+n) entries

Low-rank Matrix Completion

$\mathbf{X}$ = 

X_{ij} known for black cells
X_{ij} unknown for white cells

- How do you fill in the missing data?

$$\begin{array}{ll} \text{minimize} & \text{rank}(\mathbf{Z}) \\ \text{subject to} & Z_{ij} = X_{ij} \\ & \quad \forall\ (i,j) \in \Omega \end{array}$$

Recommender Systems

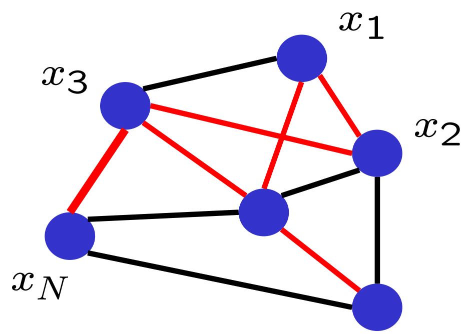
amazon.com

NETFLIX

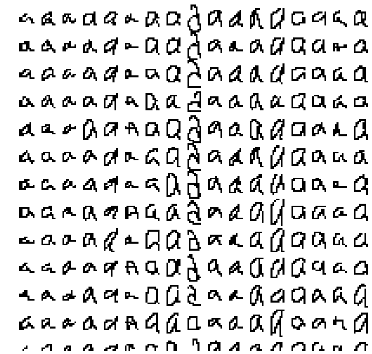
match.com

chemistry

Euclidean Embedding



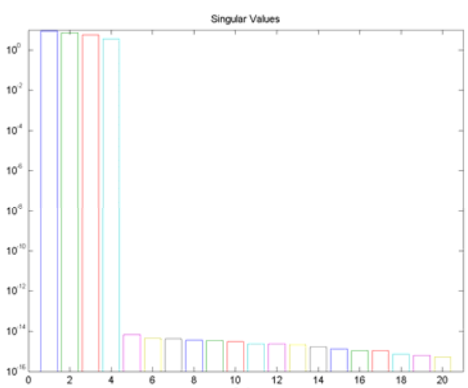
Multitask Learning



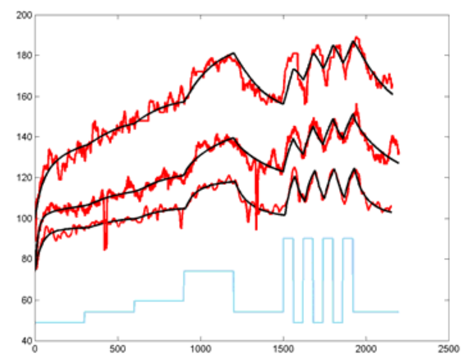
Rank of: Data Matrix

Gram Matrix

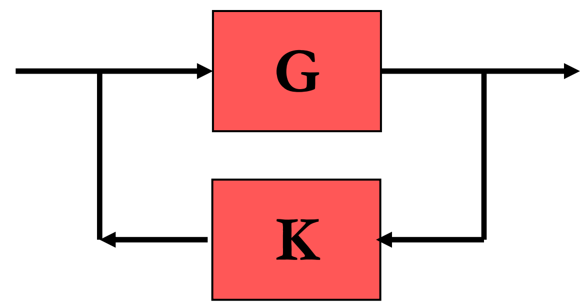
Matrix of Classifiers



Model Reduction



System Identification



Controller Design

Constraints involving the rank of the Hankel Operator, Matrix, or Singular Values

Affine Rank Minimization

- **PROBLEM:** Find the matrix of lowest rank that satisfies/approximates the underdetermined linear system

$$\mathcal{A}(\mathbf{X}) = \mathbf{b} \qquad \mathcal{A}: \mathbb{R}^{k \times n} \to \mathbb{R}^m$$

$$\begin{array}{ll} \text{minimize} & \text{rank}(\mathbf{X}) \\ \text{subject to} & \mathcal{A}(\mathbf{X}) = \mathbf{b} \end{array}$$

- **NP-HARD:**
  - Reduce to finding solutions to polynomial systems
  - Hard to approximate
  - Exact algorithms are awful

Proposed Heuristic

Affine Rank Minimization:

$$\begin{array}{ll} \text{minimize} & \text{rank}(\mathbf{X}) \\ \text{subject to} & \mathcal{A}(\mathbf{X}) = \mathbf{b} \end{array}$$

Convex Relaxation:

$$\begin{array}{ll} \text{minimize} & \|\mathbf{X}\|_* = \sum_{i=1}^{k} \sigma_i(\mathbf{X}) \\ \text{subject to} & \mathcal{A}(\mathbf{X}) = \mathbf{b} \end{array}$$

- Proposed by Fazel (2002).
- Nuclear norm is the “numerical rank” in numerical analysis
- The “trace heuristic” from controls if **X** is p.s.d.

Parsimonious Models

rank

$$x = \sum_{k=1}^{r} w_k \alpha_k$$

model   weights   atoms

- Search for best linear combination of fewest atoms
- “rank” = fewest atoms needed to describe the model

$$\|x\|_{\mathcal{A}} \equiv \inf_{(w,\alpha)} \sum_{k=1}^{r} |w_k|$$

- 2x2 matrices
- plotted in 3d

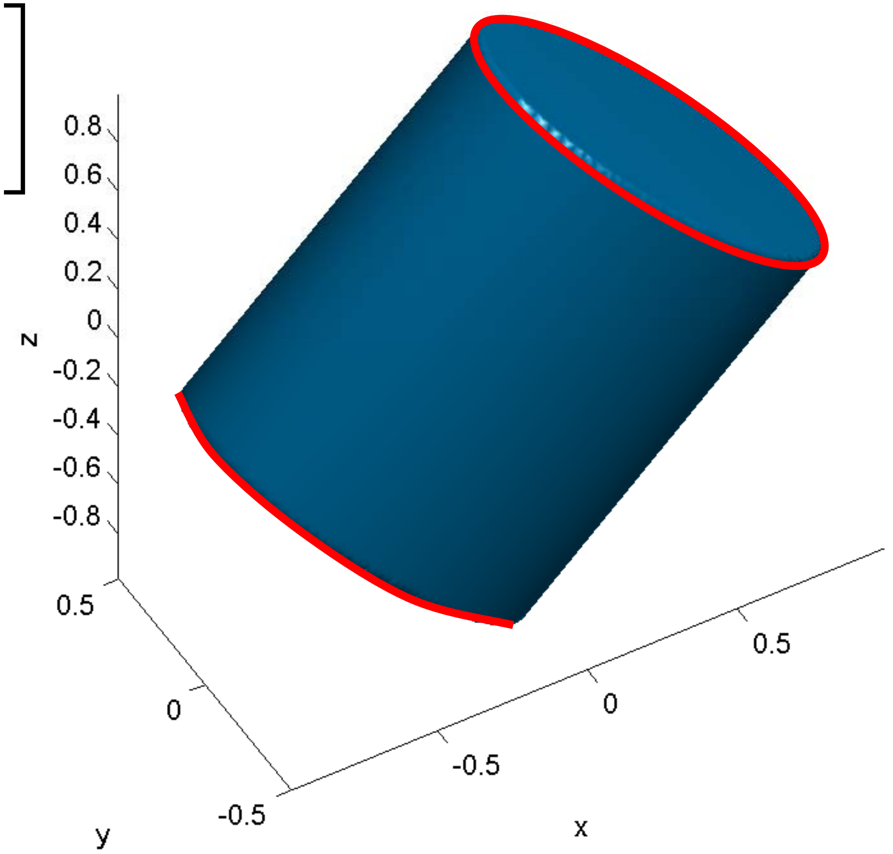
$$\begin{bmatrix} x & y \\ y & z \end{bmatrix}$$

rank 1

$x^2 + z^2 + 2y^2 = 1$

Convex hull:

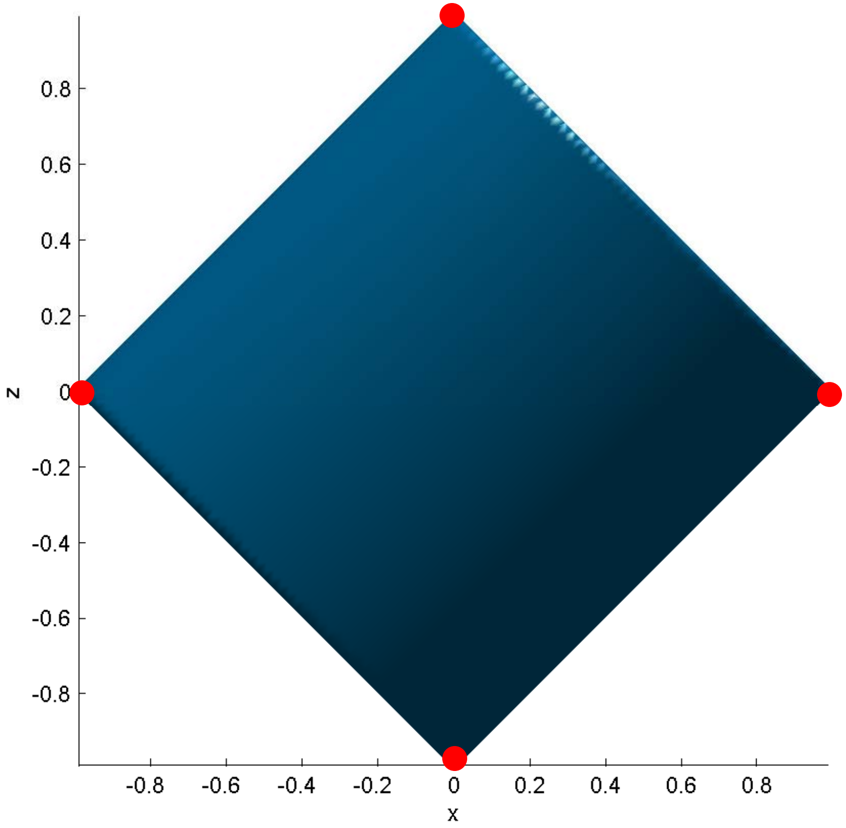
$$\{X \;:\; \|X\|_* \le 1\}$$



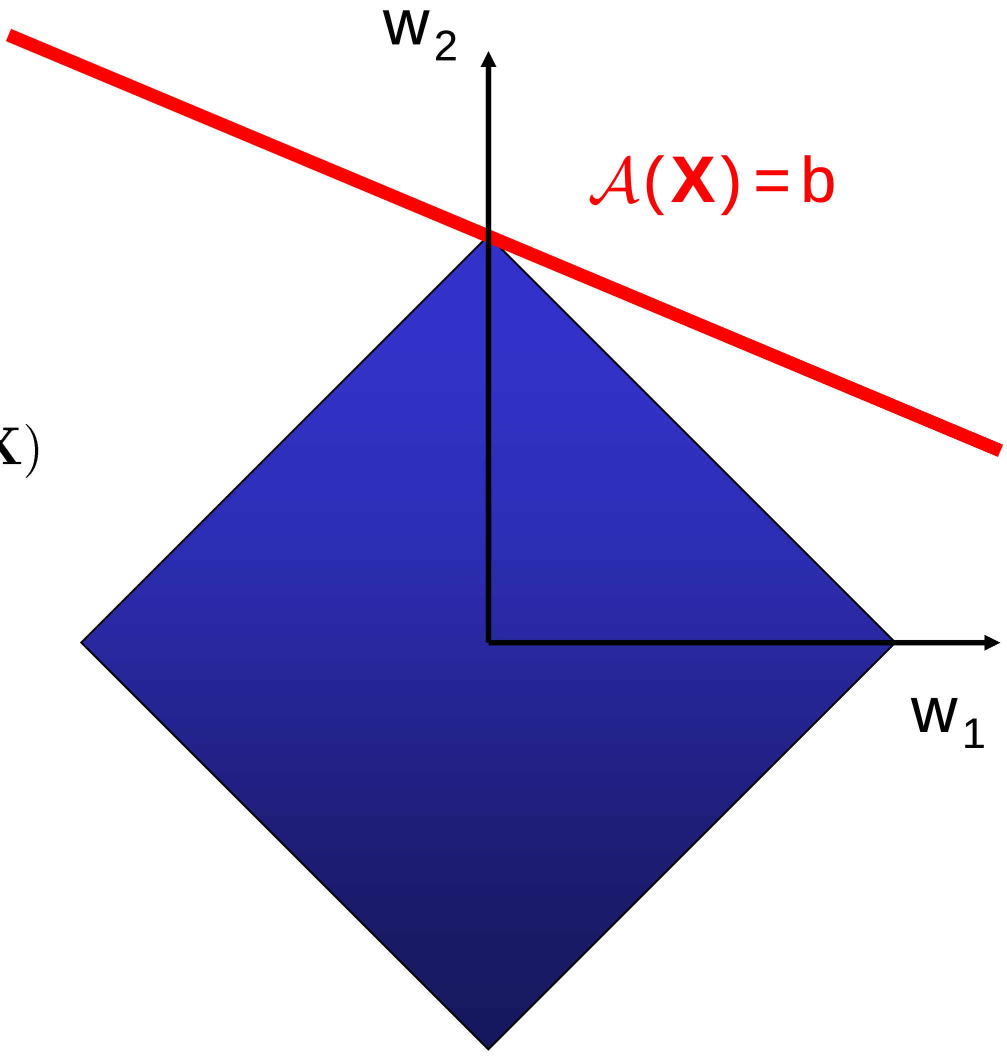
- 2x2 matrices
- plotted in 3d

$$\left\| \begin{bmatrix} x & 0 \\ 0 & z \end{bmatrix} \right\|_* \leq 1$$

- Projection onto x-z plane is l_1 ball



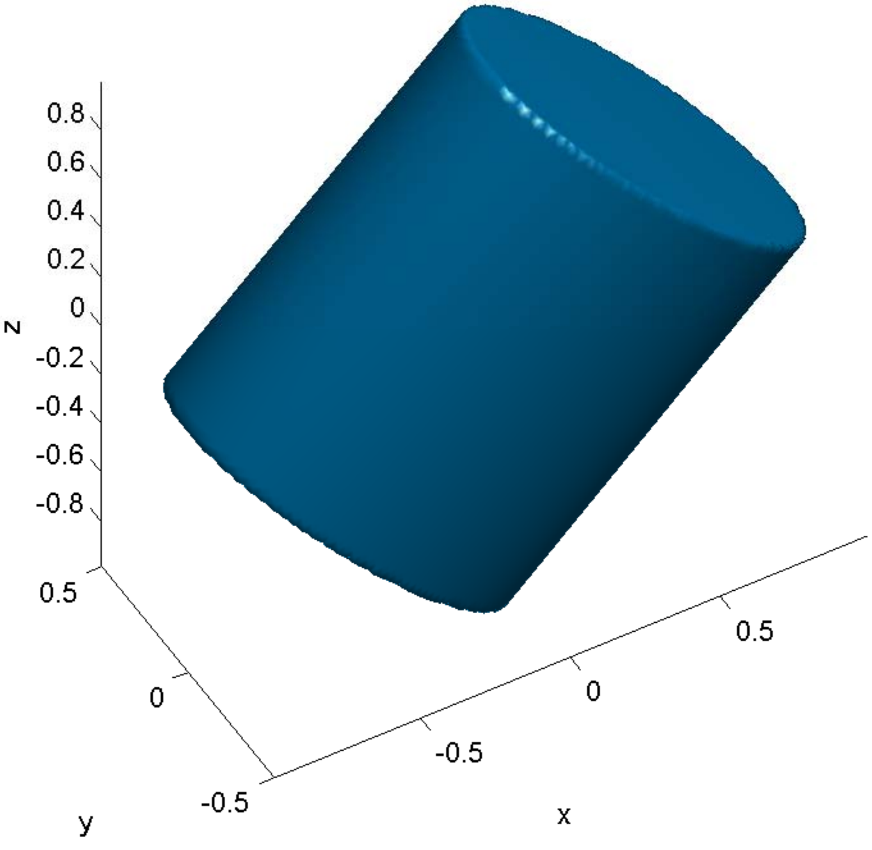
$$
\begin{array}{ll}
\text{minimize} & \|\mathbf{X}\|_* = \sum_{i=1}^{k} \sigma_i(\mathbf{X}) \\
\text{subject to} & \mathcal{A}(\mathbf{X}) = \mathbf{b}
\end{array}
$$



- 2x2 matrices
- plotted in 3d

$$\left\| \begin{bmatrix} x & y \\ y & z \end{bmatrix} \right\|_* \leq 1$$

- Not polyhedral...



So how do we compute it? And when does it work?

Equivalent Formulations

$$\begin{array}{ll}\text{minimize} & \|X\|_* \\ \text{subject to} & \mathcal{A}(X) = b\end{array} \quad \Leftrightarrow \quad \begin{array}{ll}\text{minimize} & \sum_{i=1}^k \sigma_i(X) \\ \text{subject to} & \mathcal{A}(X) = b\end{array}$$

- Semidefinite embedding:

$$X = U\Sigma V^*$$

$$\begin{bmatrix} W_1 & X \\ X^* & W_2 \end{bmatrix} = \begin{bmatrix} U \\ V \end{bmatrix} \Sigma \begin{bmatrix} U \\ V \end{bmatrix}^*$$

$$\begin{array}{ll}\text{minimize} & \frac{1}{2}(\mathrm{Tr}(W_1) + \mathrm{Tr}(W_2)) \\ \text{subject to} & \begin{bmatrix} W_1 & X \\ X^* & W_2 \end{bmatrix} \succeq 0 \\ & \mathcal{A}(X) = b\end{array}$$

- Low rank parametrization:

$$L = U\Sigma^{1/2}$$

$$R = V\Sigma^{1/2}$$

$$\begin{array}{ll}\text{minimize} & \frac{1}{2}(\|L\|_F^2 + \|R\|_F^2) \\ \text{subject to} & \mathcal{A}(LR^*) = b\end{array}$$

Computationally: Gradient Descent!

$$\mathcal{F}(\mathbf{L}, \mathbf{R}) = \sum_{i=1}^{k} \sum_{j=1}^{r} L_{ij}^2 + \sum_{i=1}^{n} \sum_{j=1}^{r} R_{ij}^2 + \lambda \left\| \mathcal{A}(\mathbf{L}\mathbf{R}^*) - \mathbf{b} \right\|^2$$

- "Method of multipliers"
- Schedule for λ controls the noise in the data
- Same global minimum as nuclear norm
- Dual certificate for the optimal solution

- When will this fail and when it might succeed?

First theory result

$$\mathcal{A}(\mathbf{X}) = \mathbf{b} \qquad \mathcal{A} : \mathbb{R}^{k \times n} \rightarrow \mathbb{R}^m$$

- If $m > c_0 r(k+n-r)\log(kn)$, the heuristic succeeds for most $\mathcal{A}$

Recht, Fazel, and Parrilo. 2007.

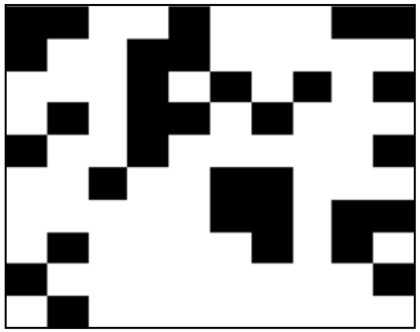
- Number of measurements c_0 $r(k+n-r)$ $\log(kn)$

constant **intrinsic dimension** **ambient dimension**

- **Approach:** Show that a random $\mathcal{A}$ is nearly an isometry on the manifold of low-rank matrices.
- Stable to noise in measurement vector **b** and returns as good an answer as a truncated SVD of the true **X**.

Low-rank Matrix Completion

X =



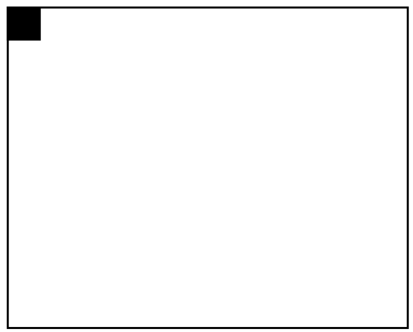
X_{ij} known for black cells
X_{ij} unknown for white cells

- How do you fill in the missing data?

$$\begin{aligned} &\text{minimize} && \text{rank}(\mathbf{Z}) \\ &\text{subject to} && Z_{ij} = X_{ij} \\ & && \quad \forall\ (i,j) \in \Omega \end{aligned}$$

Which matrices?

X =



$(= e_1 e_1^*)$

- Any subset of entries that misses the (1,1) component tells you nothing!

X =

$(= e_1 v^*)$

- Still need to see the entire first row
- Want each entry to provide nearly the same amount of information

Incoherence

- Let U be a subspace of $\mathbb{R}^n$ of dimension r and $\mathbf{P}_U$ be the orthogonal projection onto U. Then the *coherence* of U (with respect to the standard basis $\mathbf{e}_i$) is defined to be

$$\mu(U) \equiv \frac{n}{r} \max_{1 \leq i \leq n} \|\mathbf{P}_U \mathbf{e}_i\|^2.$$

- $\mu(U) \geq 1$
  - e.g., span of r columns of the Fourier transform
- $\mu(U) \leq n/r$
  - e.g., any subspace that contains a standard basis element
- $\mu(U) = O(1)$
  - sampled from the uniform distribution with $r > \log n$

Matrix Completion

- Suppose **X** is k x n ($k \leq$ n) has rank *r* and has row and column spaces with incoherence bounded above by μ. Then the nuclear norm heuristic recovers **X** from most subsets of entries Ω with cardinality at least

$$|\Omega| \geq C_1 \mu n^{5/4} \; r \; \log(n)$$

- If, in addition, $r \leq \mu^{-1} n^{1/5}$,

$$|\Omega| \geq C_2 \mu n^{6/5} \; r \; \log(n)$$

then entries suffice.

Candès and Recht. 2008

Proof Tools

- Convex Analysis
  - *KKT Conditions:* Find *dual* certificate proving minimum nuclear norm solution is the hidden low rank matrix
  - *Compressed Sensing:* Use *ansatz* for multiplier and bound its norm
- Probability on Banach Spaces
  - Moment bounds for norms of matrix valued random variables [Rudelson]
  - *Decoupling* [Bourgain-Tzafiri, de la Pena *et al*]: Indicators variables can be treated as independent
  - *Non-commutative Khintchine Inequality* [Lust-Piquard]: Tightly bound the operator norm in terms of the largest entry.

Mixture of hundreds of models, including nuclear norm

Gradient descent on low-rank nuclear norm parameterization

Netflix Prize

Leaderboard

Rank	Team Name	Best Score	% Improvement	Last Submit Time
--	No Grand Prize candidates yet	--	--	--
Grand Prize - RMSE <= 0.8563				
--	No Progress Prize candidates yet	--	--	--
Progress Prize - RMSE <= 0.8625				
1	When Gravity and Dinosaurs Unite	0.8675	8.82	2008-03-01 07:03:35
2	BellKor	0.8682	8.75	2008-02-28 23:40:45
3		0.8708	8.47	2008-02-06 14:12:44
...ze 2007 - RMSE = 0.8712 - Winning Team: KorBell				
4	KorBell	0.8712	8.43	2007-10-01 23:25:23
5	acmehill	0.8720	8.35	2008-03-02 05:08:12
6	Dan Tillberg	0.8727	8.27	2008-03-02 08:42:29
7	basho	0.8729	8.25	2007-11-24 14:27:00
8	Just a guy in a garage	0.8740	8.14	2008-02-06 12:16:40
9	BigChaos	0.8748	8.05	2008-03-01 17:26:06
10	Dinosaur Planet	0.8753	8.00	2007-10-04 04:56:45
⋮	⋮	⋮	⋮	⋮
50	amgl	0.8897	6.49	2007-12-23 18:44:03
51	Remco	0.8899	6.46	2007-04-04 06:16:56
52	mxlg	0.8900	6.45	2007-12-23 18:54:46
53	JustWithSVD	0.8900	6.45	2008-02-14 16:17:54
54		0.8900	6.45	2008-02-28 09:56:20
55		0.8901	6.44	2008-02-29 05:53:11
	Bozo_The_Clown	0.8902	6.43	2007-09-06 17:24:48

Parsimonious Modeling: A road map

- **Open Problems in rank minimization**: optimal bounds, noise performance, faster algorithms, more mining of connections with compressed sensing
- **Expanding the parsimony catalog**: dynamical systems, nonlinear models, tensors, completely positive matrices, Jordan Algebras, and beyond
- **Automatic parsimonious programming**: computational complexity of norms. algorithm and proof generation
- **Broad applied impact**: data mining time series in biology, medicine, social networks, and human computer interfaces

Acknowledgements

- See:

  http://www.ist.caltech.edu/~brecht/publications.html

  for all references

- Results developed in collaboration with Emmanuel Candès, John Doyle, Babak Hassibi, and Weiyu Xu at Caltech, Ali Rahimi at Intel Research, Harvey Cohen, Lawrence Recht, and John Whitin at Stanford, Maryam Fazel at U Washington, and Pablo Parrilo and the RealNose team at MIT.