Matrix Factorizations: A Tale of Two Norms

Nati Srebro
Toyota Technological Institute—Chicago

Maximum Margin Matrix Factorization
S, Jason Rennie, Tommi Jaakkola (MIT), NIPS 2004

Rank, Trace-Norm and Max-Norm
S, Adi Shraibman (Weizmann), COLT 2005

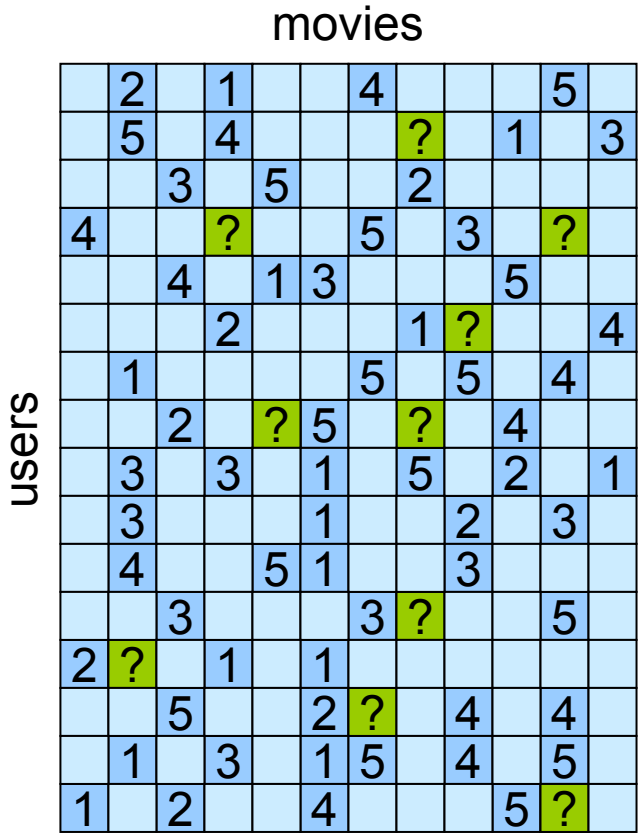
Collaborative Filtering in a Non-Uniform World: Learning with the Weighted Trace Norm
Ruslan Salakhutdinov (MIT), S, NIPS 2010

Practical Large-Scale Optimization for Max-Norm Regularization
Jason Lee (Stanford), Ben Recht (Wisconsin), Ruslan Salakhutdinov (MIT), S, Joel Tropp (CalTech), NIPS 2010

Concentration-Based Guarantees for Low-Rank Matrix Reconstruction
Rina Foygel (UChicago), S, 2011

Matrix Reconstruction (Collaborative Prediction)

- Observe Y_S, estimate other entries of Y



Matrix Reconstruction (Collaborative Prediction)

- Observe Y_S, estimate other entries of Y

$$\hat{X}(S) = \underset{\text{rank}(X) \leq k}{\arg\min} \; |X_S - Y_S|$$

$|X_S\text{-}Y_S|_1$ (abs err)
or
$|X_S\text{-}Y_S|_2$ (sq err)

- Convex alternative:

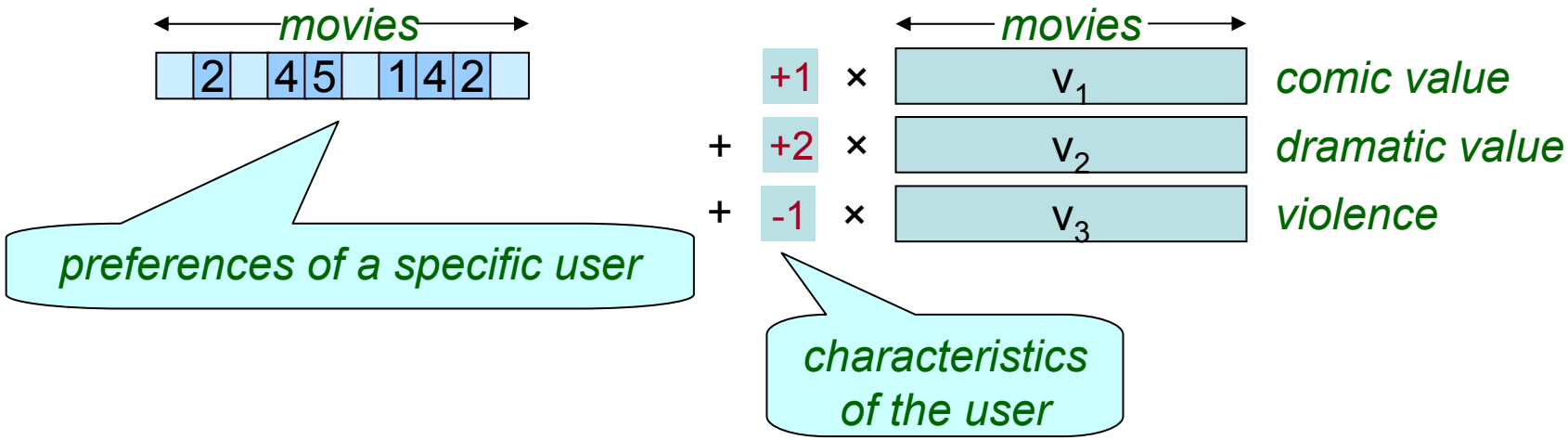
$$\hat{X}(S) = \underset{\|X\|_{\text{tr}} \leq k}{\arg\min} \; |X_S - Y_S|$$

$$\hat{X}(S) = \underset{\|X\|_{\max} \leq k}{\arg\min} \; |X_S - Y_S|$$

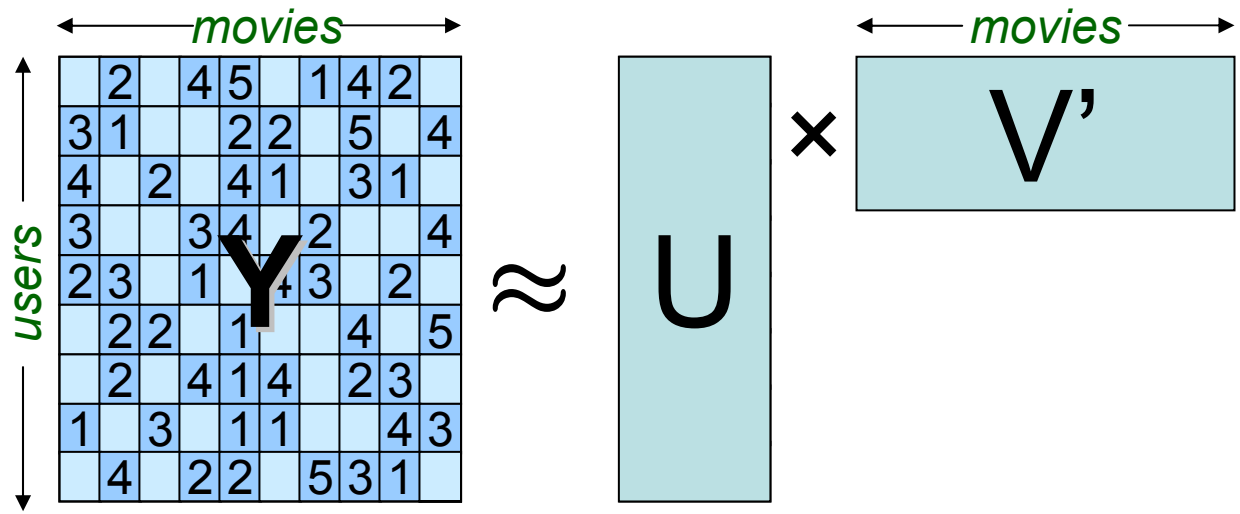
Outline

- Regularized Factorizations: $||X||_{tr}$ and $||X||_{max}$
- Optimizing with $||X||_{tr}$ and $||X||_{max}$
- Low-Rank Reconstruction Guarantees
- *Non-Uniform Sampling*
- Empirical Results

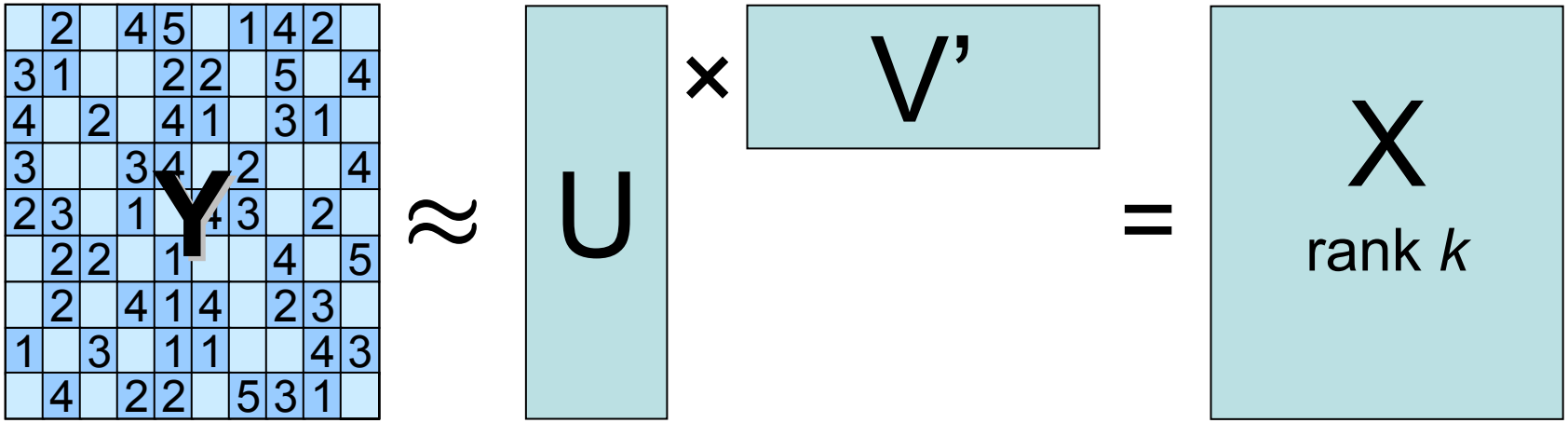
Linear Factor Model



Linear Factor Model



Unconstrained Matrix Factorization: Low Rank Approximation



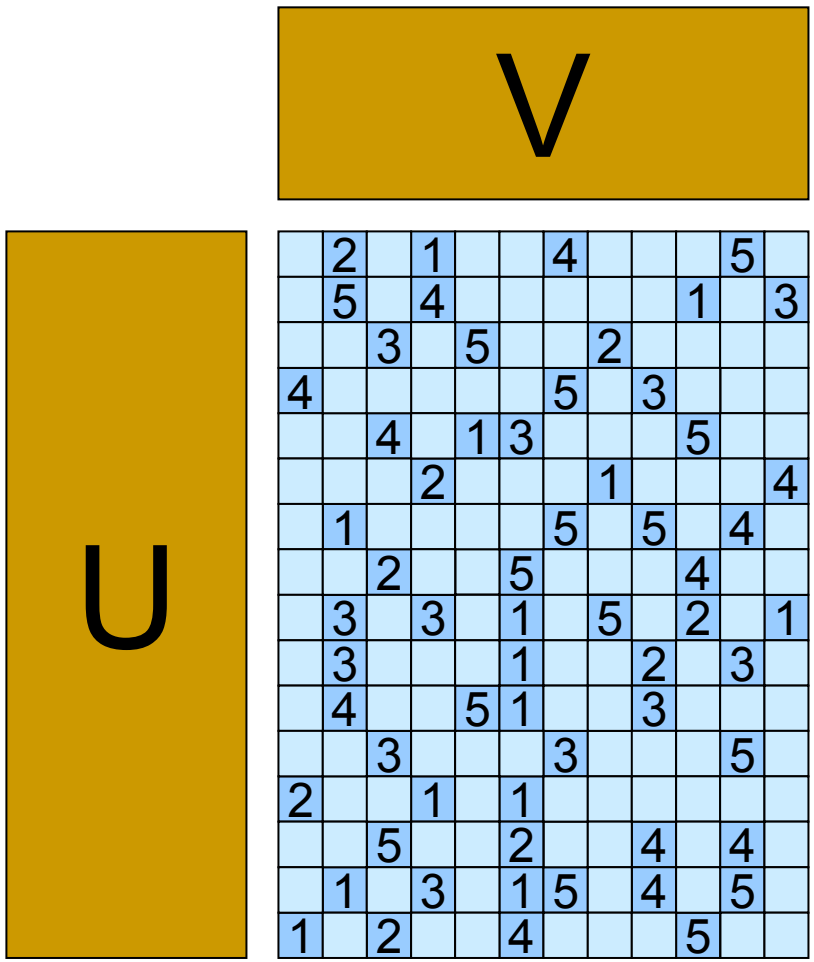
- Fitting low-rank model is hard
  - rank is non-convex
  - NP-hard to find low-rank matrix completion, or minimize error s.t. rank constraint
  - Only if Y is **fully observed**, and we want to find low-rank X minimizing **squared-error**, can use SVD

- Richer models allowing more factors?

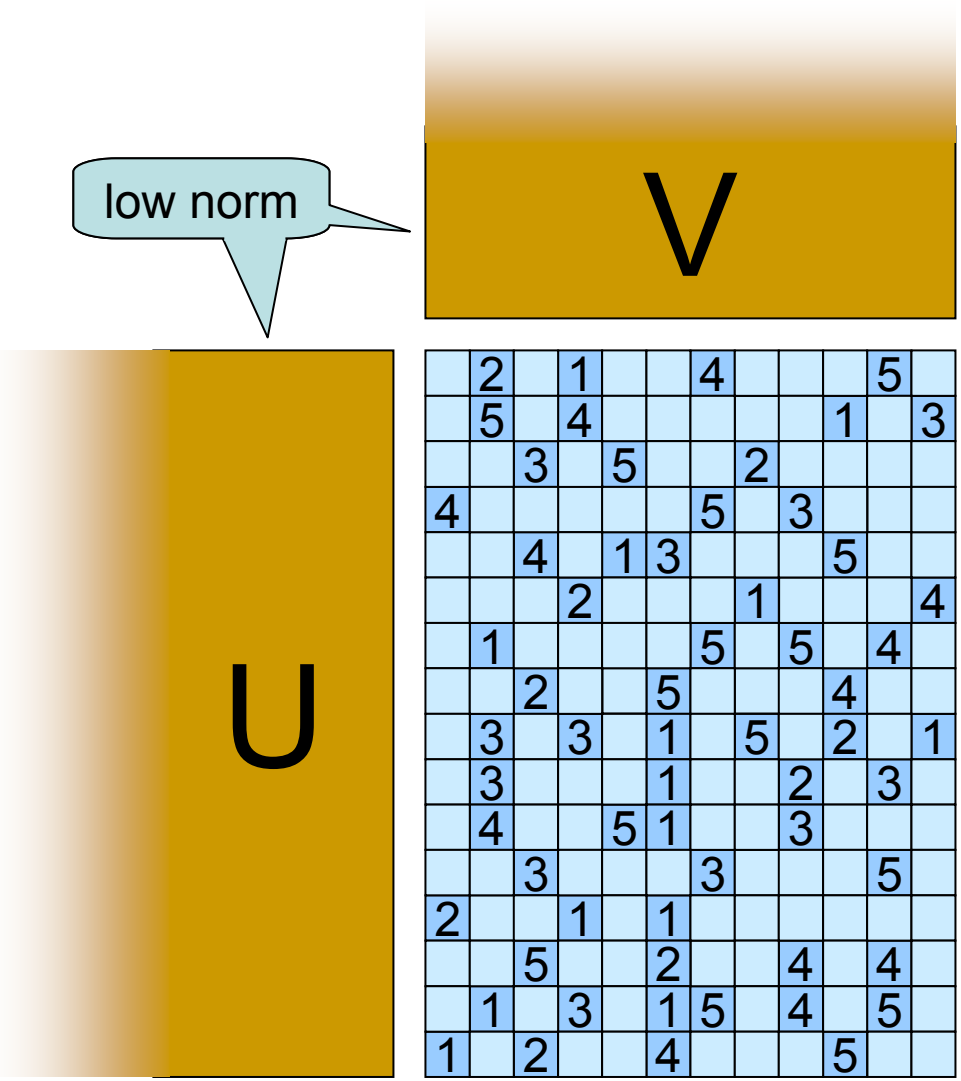
Matrix Factorization Models

Bound dimensionality of factorization:

$$\text{rank}(X) = \min_{X=UV'} \dim(U, V)$$
$$= |\text{sing values}|_0$$



Matrix Factorization Models



	2		1			4				5	
	5		4						1		3
		3		5			2				
4						5		3			
		4		1	3				5		
			2				1				4
	1					5		5		4	
		2			5				4		
	3		3		1		5		2		1
	3				1			2		3	
	4			5	1			3			
		3				3				5	
2			1		1						
		5			2			4		4	
	1		3		1	5		4		5	
1		2			4				5		

Bound dimensionality of factorization:

$$\text{rank}(X) = \min_{X=UV'} \dim(U,V)$$
$$= |\text{sing values}|_0$$

Bound avg norm of factorization:

$\|U\|_F^2 = \sum_i |U_i|^2$

$$\|X\|_{\text{tr}} = \min_{X=UV'} \|U\|_F \cdot \|V\|_F$$
$$= |\text{sing values}|_1$$

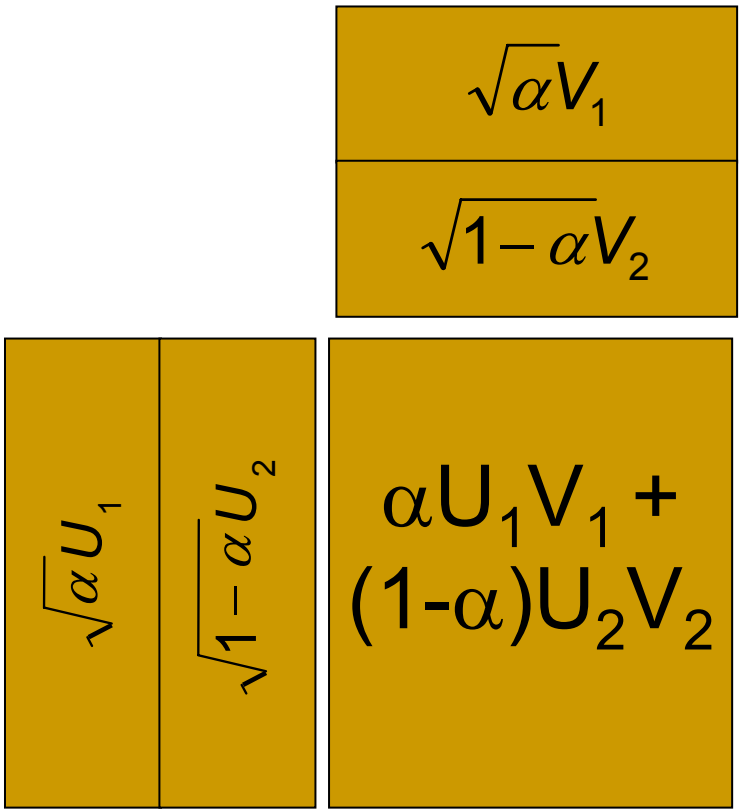
Bound norm of fact. uniformly:

$\|U\|_{2,\infty} = \max_i |U_i|$

$$\|X\|_{\max} = \min_{X=UV'} \|U\|_{2,\infty} \cdot \|V\|_{2,\infty}$$

aka γ_2:$\ell_1 \to \ell_\infty$ norm

Matrix Factorization Models: Convexity



Bound dimensionality of factorization:

$$\text{rank}(X) = \min_{X=UV'} \dim(U,V)$$
$$= |\text{sing values}|_0$$

Bound avg norm of factorization:

$\|U\|_F^2 = \sum_i |U_i|^2$

$$\|X\|_{\text{tr}} = \min_{X=UV'} \|U\|_F \cdot \|V\|_F$$
$$= |\text{sing values}|_1$$

Bound norm of fact. uniformly:

$\|U\|_{2,\infty} = \max_i |U_i|$

$$\|X\|_{\max} = \min_{X=UV'} \|U\|_{2,\infty} \cdot \|V\|_{2,\infty}$$

aka $\gamma_2: \ell_1 \to \ell_\infty$ norm

Trace-Norm and Max-Norm

Trace-Norm (On-Average Constraint):

Regularized factorization: $\|X\|_{\mathsf{tr}} = \min_{X=U'V} \sqrt{(\sum_i |U_i|^2)(\sum_j |V_j|^2)}$

Infinite comb. of factors: $\{X \,|\, \|X\|_{\mathsf{tr}} \leq 1\} = \mathsf{conv}\left(\left\{uv' \middle| |u| = |v| = 1\right\}\right)$

Lower bound on rank: $\mathsf{tc}(X) \doteq \frac{\|X\|_{\mathsf{tr}}^2}{nm} \leq \mathsf{rank}(X) \cdot \frac{\sum_{ij} |X_{ij}|^2}{nm}$

$\|X\|_{tr} \leq \mathsf{rank}^{1/2}(X)\, \|X\|_{Fro}$

Max-Norm (Uniform Constraint):

Regularized factorization: $\|X\|_{\mathsf{max}} = \min_{X=U'V} (\max_i |U_i|)(\max_j |V_j|)$

Infinite comb. of factors:

$$\mathsf{conv}\left(\left\{uv' \middle| u \in \pm 1^n, v \in \pm 1^m\right\}\right) \subseteq \{X \,|\, \|X\|_{\mathsf{max}} \leq 1\} \subseteq 2\mathsf{conv}\left(\left\{uv' \middle| u \in \pm 1^n, v \in \pm 1^m\right\}\right)$$

Grothendiek's Inequality

Lower bound on rank: $\mathsf{mc}(X) \doteq \|X\|_{\mathsf{max}}^2 \leq \mathsf{rank}(X) \cdot \max_{ij} |X_{ij}|^2$

Outline

- Regularized Factorizations: $||X||_{tr}$ and $||X||_{max}$
- → **Optimizing with $||X||_{tr}$ and $||X||_{max}$**
- Low-Rank Reconstruction Guarantees
- *Non-Uniform Sampling*
- Empirical Results

Optimization with the Trace-Norm

Covex

minimize loss(**X**) + $\lambda \cdot |\mathbf{X}|_{tr}$

$|\mathbf{X}|_{tr} = \sum$ (singular values of **X**) $= \min_{X=U'V} \frac{1}{2}(\sum_i |\mathbf{U}_i|^2 + \sum_j |\mathbf{V}_j|^2)$

$= \min_{A,B} \frac{1}{2}(tr(\mathbf{A}) + tr(\mathbf{B}))$

$\begin{pmatrix} \mathbf{A} & \mathbf{X} \\ \mathbf{X}' & \mathbf{B} \end{pmatrix}$ p.s.d.

[Fazel Hindi Boyd 2001]

SDP

minimize loss(**X**) + $\lambda/2$(tr(**A**)+tr(**B**))

$\begin{pmatrix} \mathbf{A} & \mathbf{X} \\ \mathbf{X}' & \mathbf{B} \end{pmatrix}$ p.s.d.

Optimization with the Trace-Norm

Not Covex!

But no local (non-global) minima!
follows from **[Burer Monteiro 03]**

minimize loss(**U'V**) + $\lambda/2(\sum_i |\mathbf{U}_i|^2 + \sum_j |\mathbf{V}_j|^2)$

X=UV

- Unconstrained opt on U,V
  [Rennie S 05]
- Stochastic Grad Desc on U,V
  [Salakhutdinov Minh 08]

Covex

$|\mathbf{X}|_{tr} = \sum$ (singular values of **X**) $= \min_{X=U'V} \frac{1}{2}(\sum_i |\mathbf{U}_i|^2 + \sum_j |\mathbf{V}_j|^2)$

minimize loss(**X**) + $\lambda \cdot |\mathbf{X}|_{tr}$

- Subgradient descent
- Smoothed Subgradient descent
- Singular Value Thresholding
  [Cai Candes Shen 08]
- . . .

All require SVD at each iteration

SDP

minimize loss(**X**) + $\lambda/2$(tr(**A**)+tr(**B**))

$$\begin{pmatrix} \mathbf{A} & \mathbf{X} \\ \mathbf{X}' & \mathbf{B} \end{pmatrix} \text{ p.s.d.}$$

- Can use standard SDP code

But only on small problems…

Optimization with the Max-Norm

Not Covex!

But no local (non-global) minima! (if U,V of high enough dim)
follows from **[Burer Monteiro 03]**

minimize loss(**U'V**) + $\lambda/2(\max_i |\mathbf{U}_i|^2 + \max_j |\mathbf{V}_j|^2)$
X=UV

Covex

$|\mathbf{X}|_{max} = \sum(\text{sing val of } \mathbf{X}) = \min_{X=U'V} \frac{1}{2}(\max_i|\mathbf{U}_i|^2 + \max_j|\mathbf{V}_j|^2)$
$= \min_{A,B} \frac{1}{2}(\max \text{diag}(\mathbf{A}) + \max \text{diag}(\mathbf{B}))$

minimize loss(**X**) + $\lambda \cdot |\mathbf{X}|_{max}$

$\begin{pmatrix} \mathbf{A} & \mathbf{X} \\ \mathbf{X}' & \mathbf{B} \end{pmatrix}$ p.s.d.

SDP

minimize loss(**X**) + $\lambda/2(\max \text{diag}(\mathbf{A}) + \max \text{diag}(\mathbf{B}))$

$\begin{pmatrix} \mathbf{A} & \mathbf{X} \\ \mathbf{X}' & \mathbf{B} \end{pmatrix}$ p.s.d.

Trace-Norm vs Max-Norm Optimization

	$\|X\|_{tr}$	$\|X\|_{max}$
SDP Represent able?	Yes	**Yes**
Calculating $\|X\|$	sum(SVD)	**Requires solving SDP (as hard as training)**
1st Order methods on X	Yes, SVD at each iteration (and also SVT)	**No**
Unconstrained Opt. of U,V (including Stoch Grad Descent)	Yes	**Yes** **[Lee Recht Salakhutdinov S Tropp NIPS'10]**

Trace-Norm vs Max-Norm: Empirical Matrix Completion Results

- 2004 experiments on 100x100 subsets of MovieLens and EachMovie using standard SDP code:

  MaxNorm better 90% of the time

- 2010 experiments on full NetFlix using stochastic gradient descent on U,V:

	RMSE	%improvement
NetFlix Cinematch:	0.9525	0 (baseline)
TraceNorm:	0.9235	3.04
MaxNorm:	0.9138	4.06
Winning team:	0.8553	10.2

Reconstruction Guarantees

Y=M+Z

- Using the **trace-norm**:

  #samples to get $\frac{1}{nm}\left|\hat{X}-Y\right|_1 \le \frac{1}{nm}|Z|_1+\epsilon$

  $\propto$ tc(M)·(n+m)log(n)/$\epsilon^2 \le$ rank(M)·(n+m)log(n)/ϵ^2

  tc(X) = $||X||^2_{tr}/nm \le$ rank(X) ($|X|^2_F/nm$)

  $|M_{ij}|^2$ = O(1) on average

- Using the **max-norm**:

  #samples to get $\frac{1}{nm}\left|\hat{X}-Y\right|_1 \le \frac{1}{nm}|Z|_1+\epsilon$

  $\propto$ mc(M)·(n+m)/$\epsilon^2 \le$ rank(M)·(n+m)/ϵ^2

  mc(X) = $||X||^2_{max} \le$ rank(X) $|X|^2_\infty$

  $|M_{ij}|^2$ = O(1) for all i,j

  $\sigma^2 = \frac{1}{nm}|Z|_2^2$

  #samples to get $\frac{1}{nm}\left|\hat{X}-Y\right|_2^2 \le \sigma^2+\epsilon$

  $\propto$ rank(M)·(n+m)·$\frac{\log^3 1/\epsilon}{\epsilon}\cdot\frac{\sigma^2+\epsilon}{\epsilon}$

Reconstruction Guarantees

Y=M+Z

- Using the **trace-norm**:

  #samples to get $\frac{1}{nm}\left|\hat{X}-Y\right|_1 \le \frac{1}{nm}|Z|_1+\epsilon$

  $\propto$ tc(M)·(n+m)log(n)/$\epsilon^2 \le$ rank(M)·(n+m)log(n)/ϵ^2

  $tc(X) = ||X||^2_{tr}/nm \le rank(X)\ (|X|^2_F/nm)$

  Only under uniform sampling.

  $|M_{ij}|^2$ = O(1) on average

- Using the **max-norm**:

  #samples to get $\left|\hat{X}-Y\right|_{L_1(\mathcal{D})} \le |Z|_{L_1(\mathcal{D})}+\epsilon$

  $\propto$ mc(M)·(n+m)/$\epsilon^2 \le$ rank(M)·(n+m)/ϵ^2

  $mc(X) = ||X||^2_{max} \le rank(X)\ |X|^2_\infty$

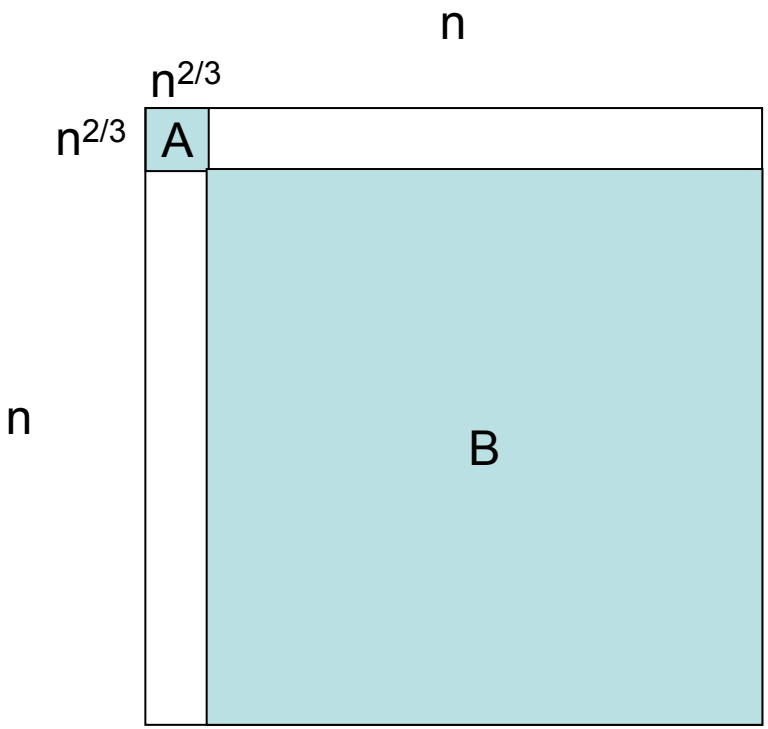
  $|M_{ij}|^2$ = O(1) for all i,j

  #samples to get $\left|\hat{X}-Y\right|^2_{L_2(\mathcal{D})} \le \sigma^2+\epsilon$

  $\propto$ rank(M)·(n+m)·$\frac{\log^3 1/\epsilon}{\epsilon}\cdot\frac{\sigma^2+\epsilon}{\epsilon}$

  Under arbitrary sampling!

The Trace-Norm with Non-Uniform Sampling



- Both A,B of rank 2
- Sampling:
  - uniform in A w.p. ½
  - uniform in B w.p. ½

- Regularizing with the rank or with the max-norm:
  sample complexity $\propto$ n
- Regularizing with the trace-norm:
  sample complexity $\propto n^{4/3}$!!!

[Salakhutdinov S NIPS'10]

Correcting for Non-Uniform Sampling: The Weighted Trace-Norm

[Salakhutdinov S NIPS'10]

- Instead of the Trace-Norm:

$$|X|_{tr} = \min_{X=UV} \tfrac{1}{2} \left(\sum_i |U_i|^2 + \sum_j |V_j|^2 \right)$$

- Regularize with the weighted Trace-Norm:

$$|X|_{tr(p)} = \min_{X=UV} \tfrac{1}{2} \left(\sum_i \mathbf{p(i)} |U_i|^2 + \sum_j \mathbf{p(j)} |V_j|^2 \right)$$

frequency of choosing row i

frequency of choosing column j

- On NetFlix:

	RMSE	%improvement
NetFlix Cinematch:	0.9525	0 (baseline)
TraceNorm:	0.9235	3.04
MaxNorm:	0.9138	4.06
Weighted TraceNorm:	**0.9105**	**4.41**
Winning team:	0.8553	10.20

Max-Norm and Trace-Norm Regularization

- Both lower bound on rank
- Both SDP representable, with practical large-scale opt method
- Reconstruction guarantees for both (better for max-norm)
- Good empirical performance (better for max-norm)
- **Better model then the rank: infinite factor model**
- Not only collaborative filtering!
  - Multi-task and multi-class learning
    [Argyriou et al 07] [Abernethy et al 08] [Amit et al 07]
  - Semi-supervised learning following eg **[Rie Zhang 05,07]**

Maximum Margin Matrix Factorization, S, Rennie, Jaakkola, NIPS 2004
Rank, Trace-Norm and Max-Norm S, Shraibman, COLT 2005
Collaborative Filtering in a Non-Uniform World: Learning with the Weighted Trace Norm
Salakhutdinov, S, NIPS 2010
Practical Large-Scale Optimization for Max-Norm Regularization
Lee, Recht, Salakhutdinov, S, Tropp, NIPS 2010
Concentration-Based Guarantees for Low-Rank Matrix Reconstruction, Foygel, S, 2011