Maximum **M**argin **M**atrix **F**actorization

Nati Srebro
University of Toronto

Jason Rennie
MIT

Tommi Jaakkola
MIT

Collaborative Prediction

Based on partially observed matrix:

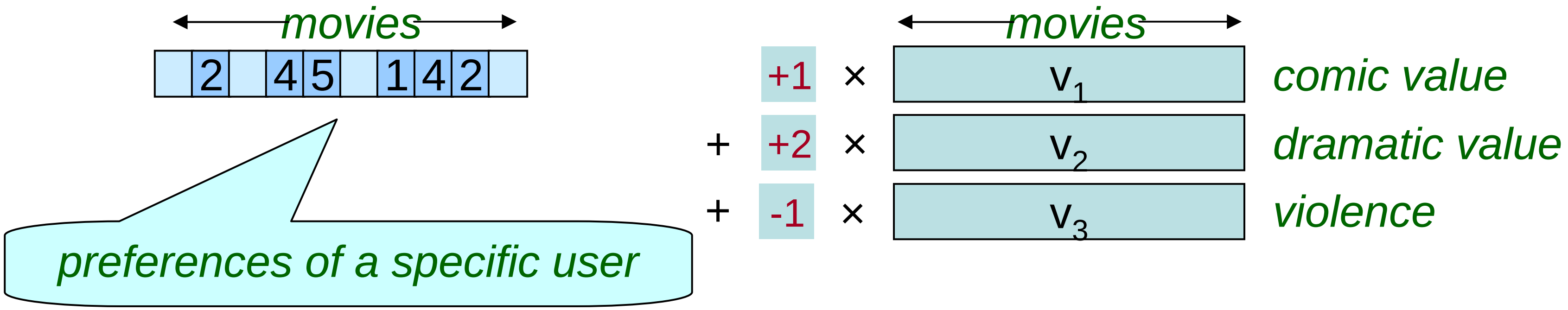
⇒ Predict unobserved entries "Will user *i* like movie *j*?"

movies

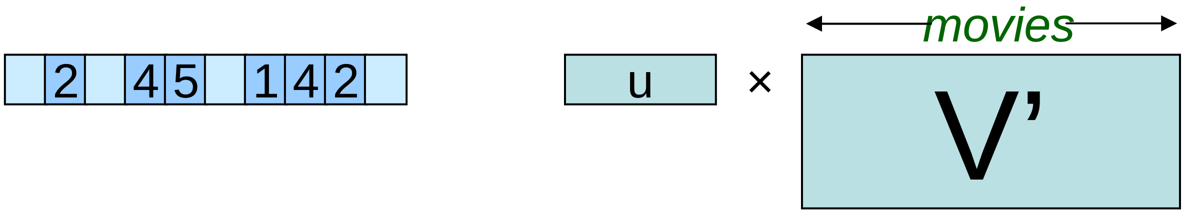
users

	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	?	

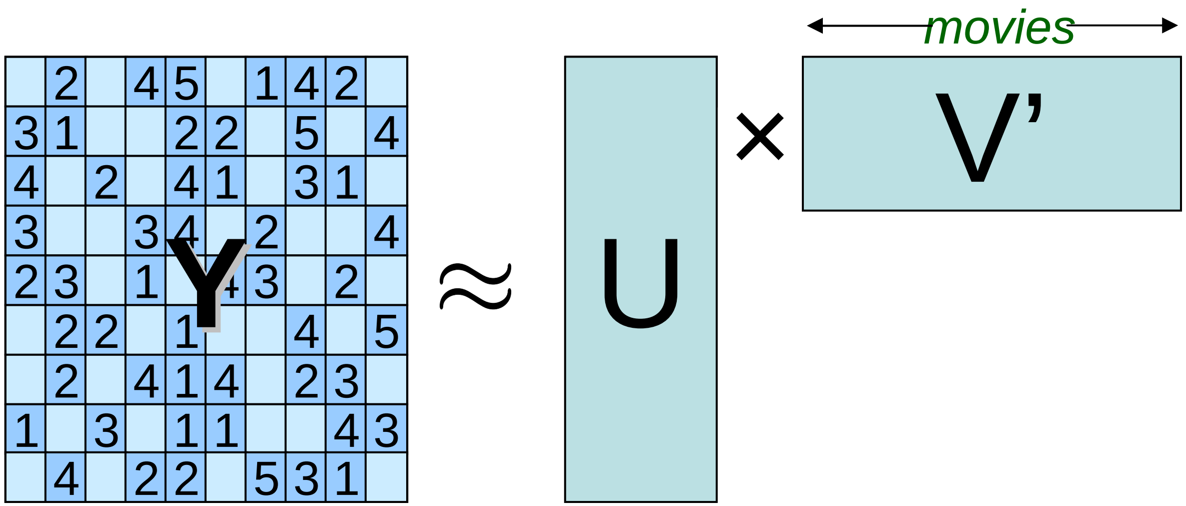
Linear Factor Model



Linear Factor Model

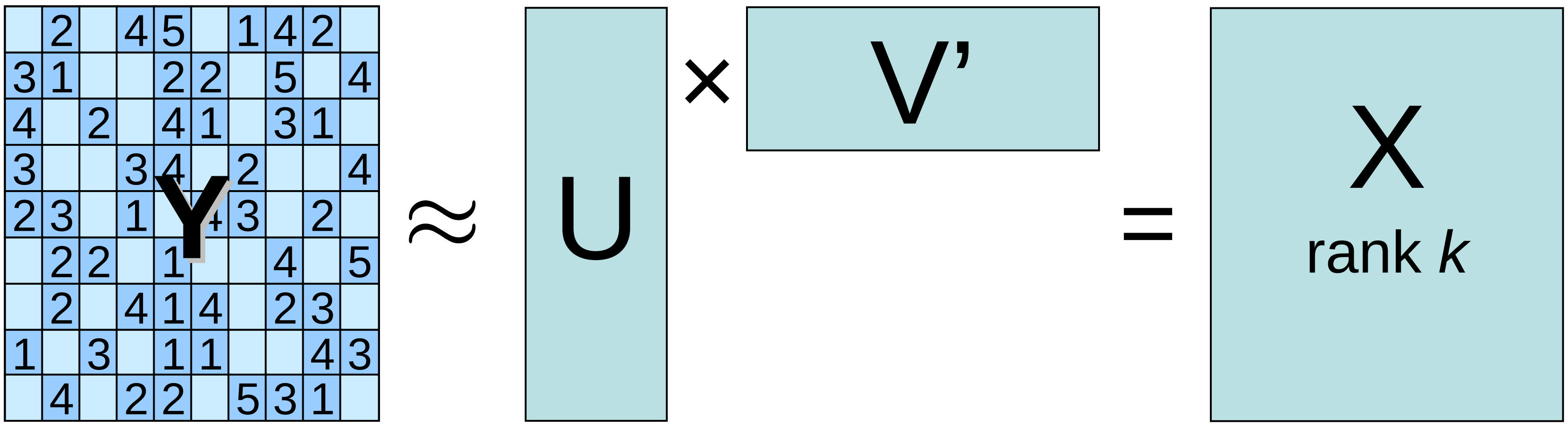


Linear Factor Model



Matrix Factorization

Unconstrained: Low Rank Approximation

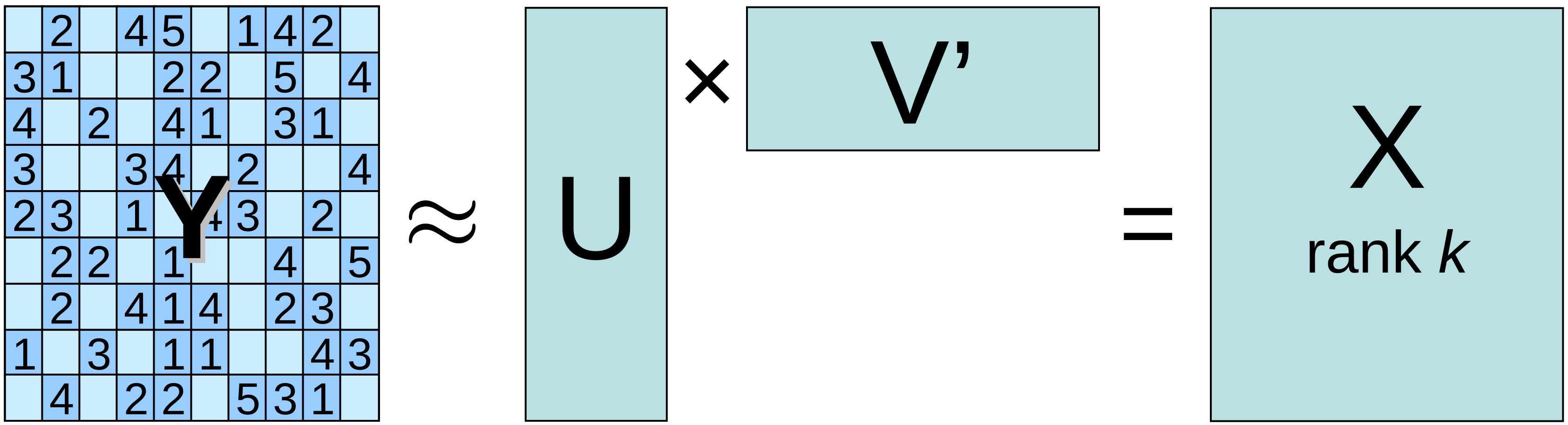


- Additive Gaussian noise: minimize $|Y\text{-}UV'|_{Fro}$
- General additive noise
- General conditional models
  - Multiplicative noise, Exponential-PCA **[Collins+01]**, Multinomial (pLSA **[Hofmann01]**), etc
- General loss functions
  - Hinge loss, loss functions appropriate for ratings, etc **[Gordon03]**

Unconstrained U,V,
fully observed Y
→ use SVD

non-convex,
no explicit solution

Matrix Factorization



- Non-Negativity **[LeeSeung99]**
- Stochasticity (convexity) **[LeeSeung97] [Hofmann01]**
- Sparsity
  - Clustering as an extreme (when rows of U sparse)

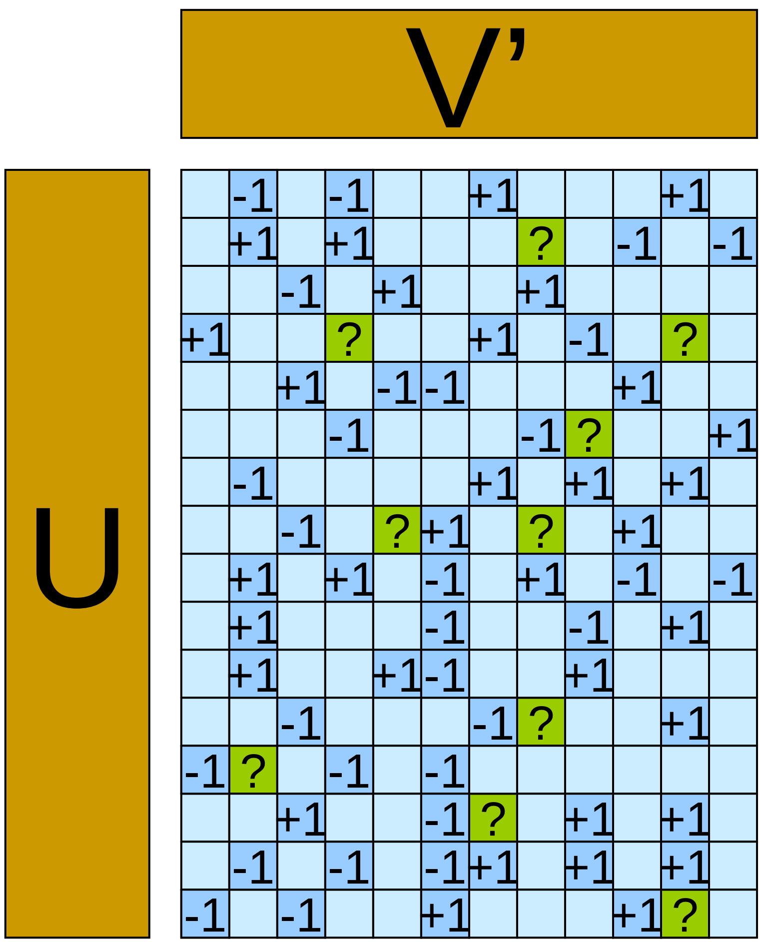
Overall number of factors still constrained

Non-convex optimization problems

Outline

- **M**aximum **M**argin **M**atrix **F**actorization
  - Unbounded number of factors
  - Convex!
- Learning MMMF: Semidefinite Programming
- Generalization Error Bounds

Collaborative Prediction with Matrix Factorization



Fit factorizable (low-rank) matrix **X=UV’** to observed entries.

minimize $\Sigma \text{loss}(\mathbf{X}_{ij};\mathbf{Y}_{ij})$

prediction

observation

Use matrix **X** to predict unobserved entries.

[Sarwar+00] [Azar+01] [Hofman04] [Marlin+04]

Collaborative Prediction with Matrix Factorization

linear classifiers

V'

feature vectors

			v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12
1.3	0.4	-1.5		-1		-1			+1				+1	
8.3	2.5	-4.8		+1		+1						-1		-1
0.7	-0.2	3.4			-1		+1			+1				
1.7	-5.2	1.6	+1						+1		-1			
-3.7	2.1	0.9			+1		-1	-1				+1		
4.3	-0.5	2.7				-1				-1				+1
4.7	0.2	6.4		-1					+1		+1		+1	
6.0	0.3	-5.8			-1			+1				+1		
-1.5	-3.7	0.4		+1		+1		-1		+1		-1		-1
-4.8	4.3	2.5		+1				-1			-1		+1	
3.4	4.7	-0.2		+1			+1	-1			+1			
1.6	6.0	-5.2			-1				-1				+1	
0.9	1.3	2.1	-1			-1		-1						
2.7	8.3	-0.5			+1			-1			+1		+1	
6.4	0.7	0.2		-1		-1		-1	+1		+1		+1	
-5.8	1.7	0.3	-1		-1			+1				+1		

When U is fixed,

each row is a linear classification problem:

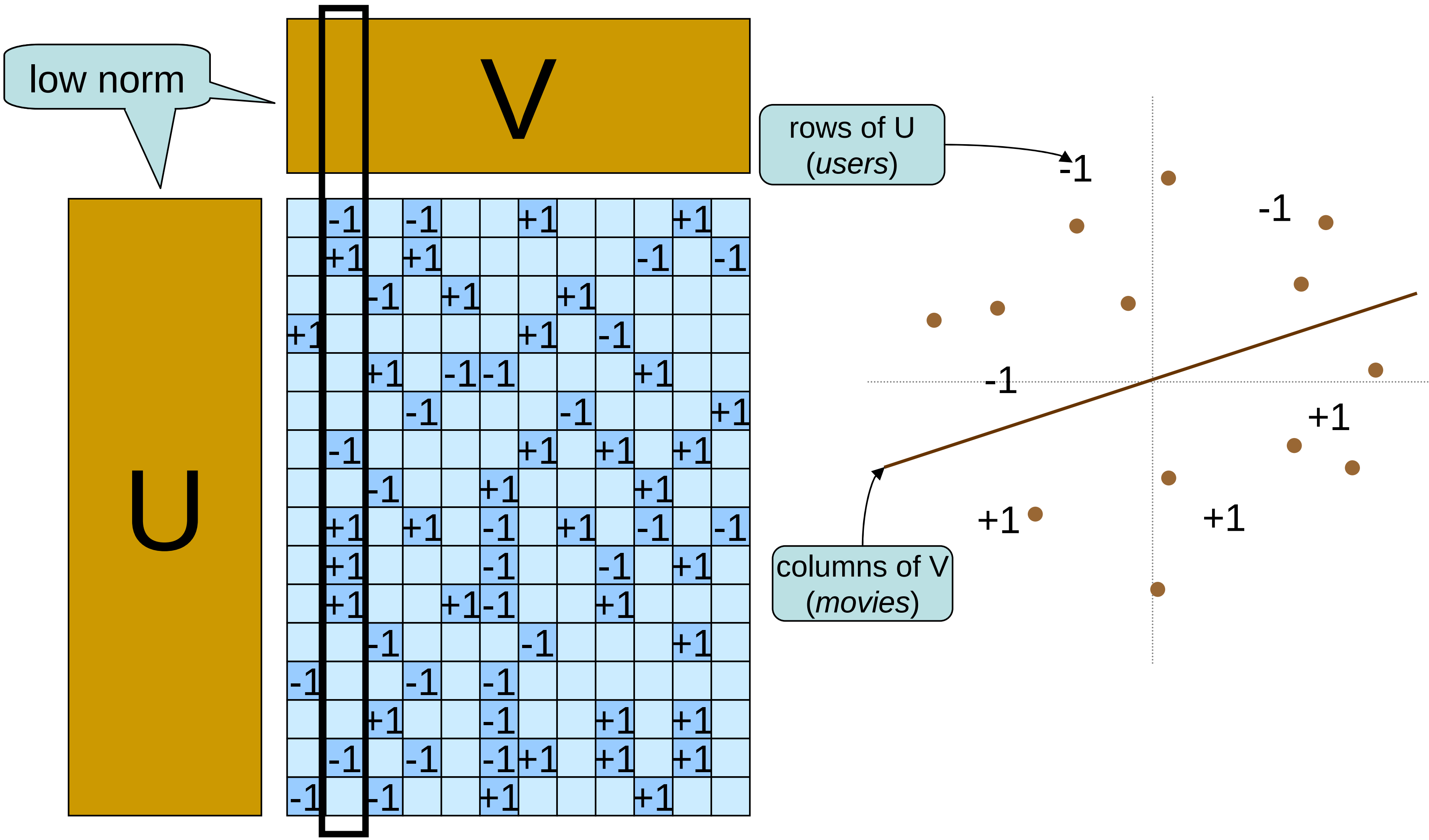
- rows of U are feature vectors
- columns of V are linear classifiers

Fitting U **and** V:

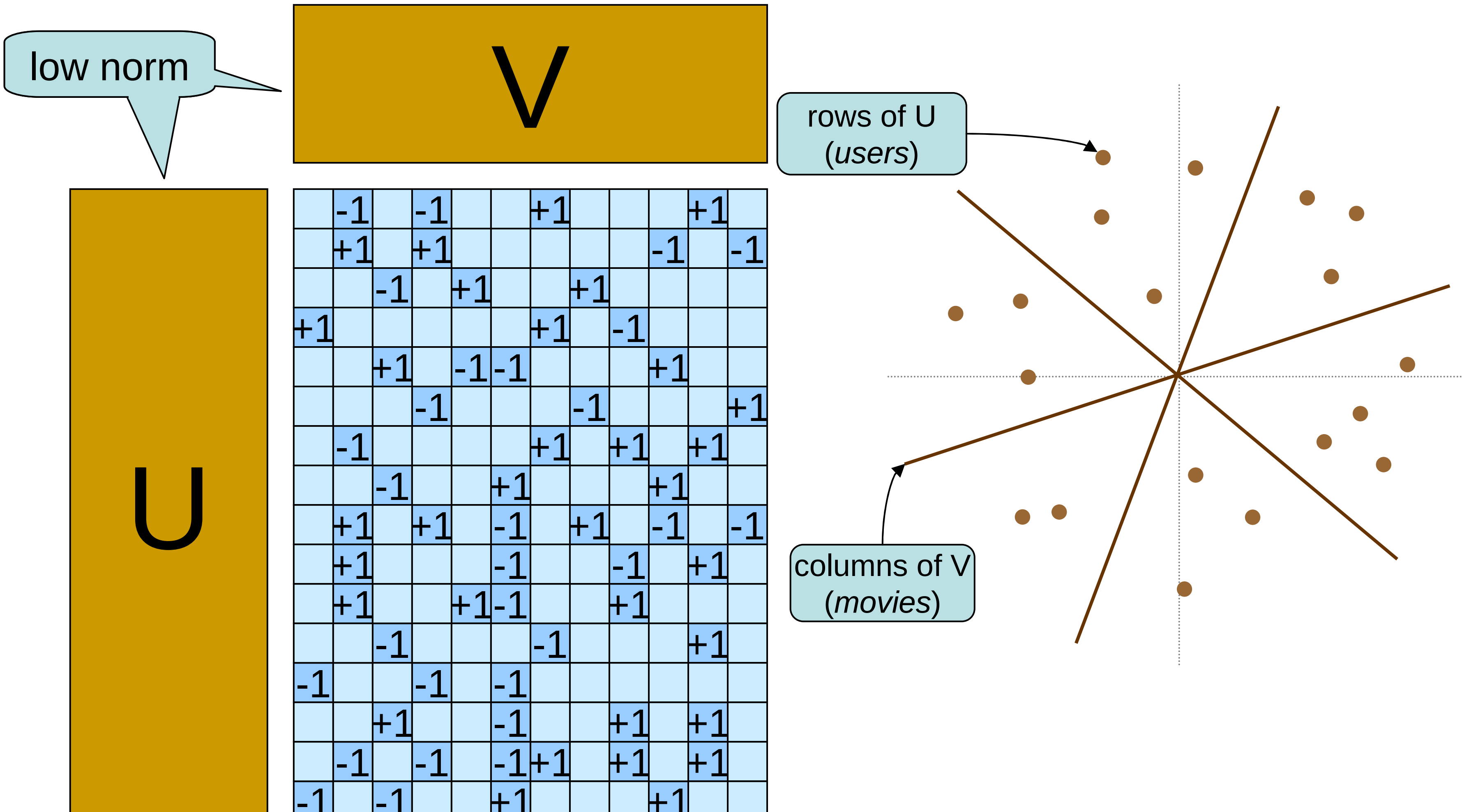
Learning features that work well across all classification problems.

Geometric Interpretation:

Co-embedding Points and Separating Hyperplanes

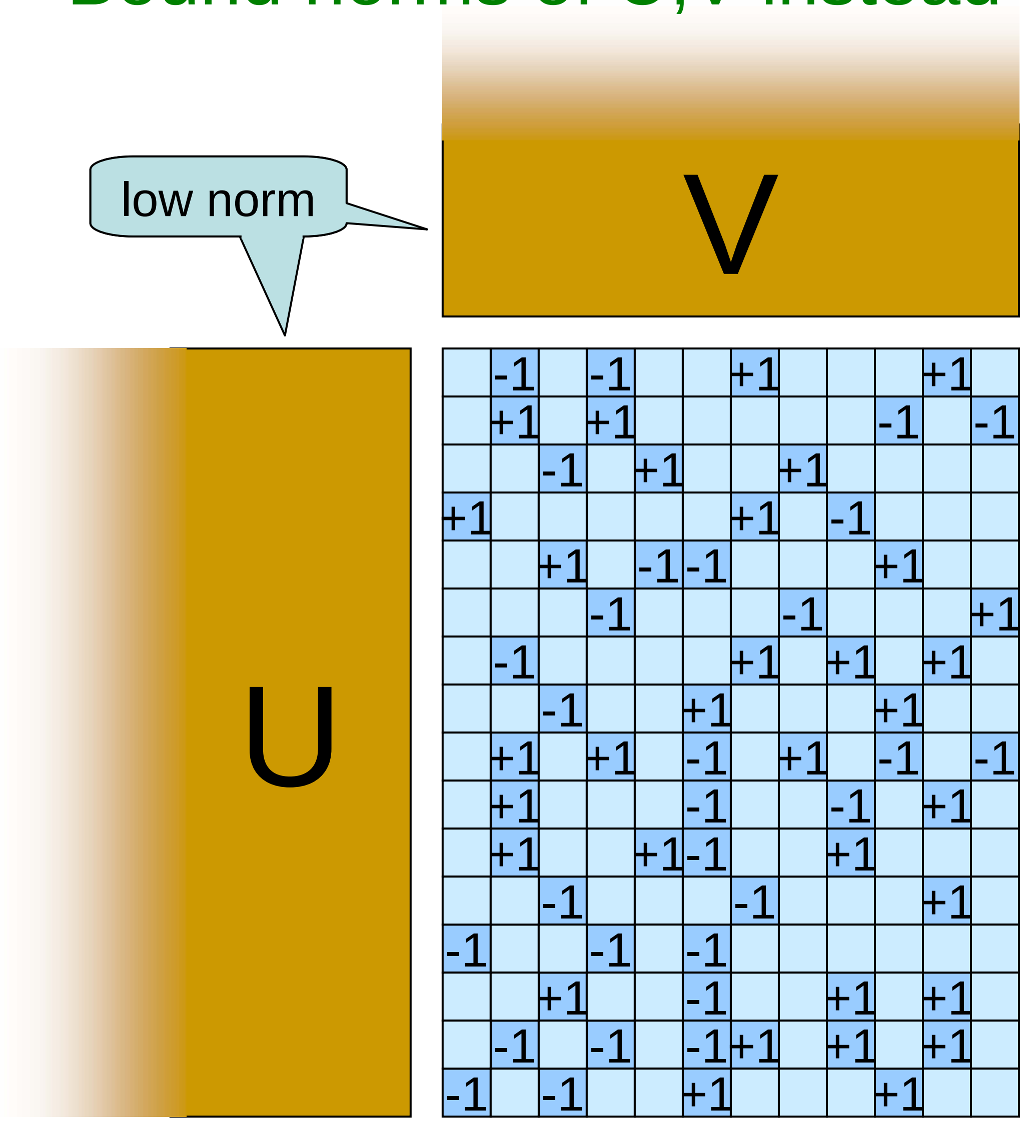


Geometric Interpretation:
Co-embedding Points and Separating Hyperplanes



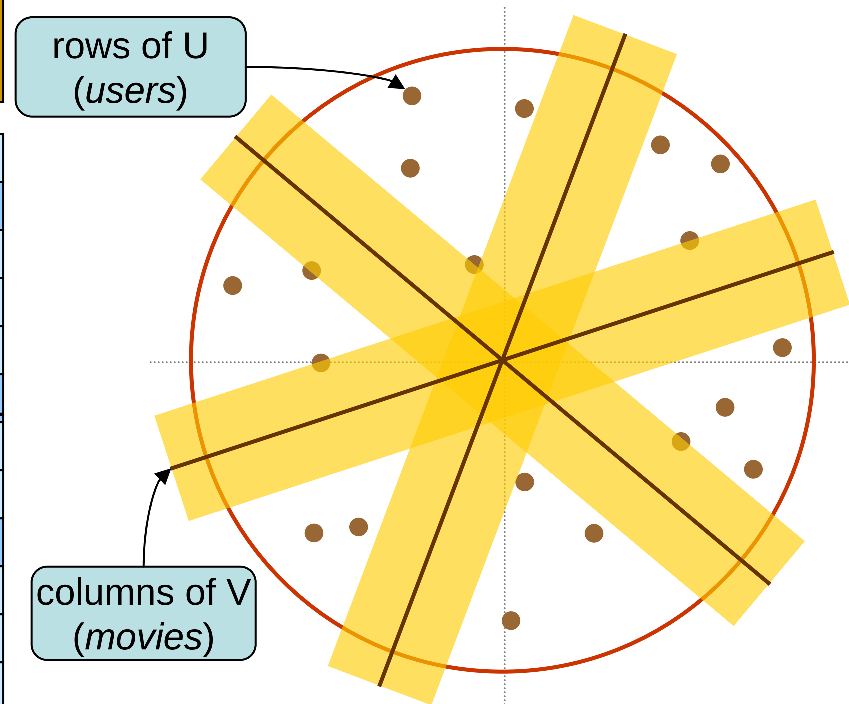
Max-Margin Matrix Factorization:

Bound norms of U,V instead of their dimensionality



bound norms uniformly:

$(\max_i |U_i|^2)\ (\max_j |V_j|^2) \leq 1$



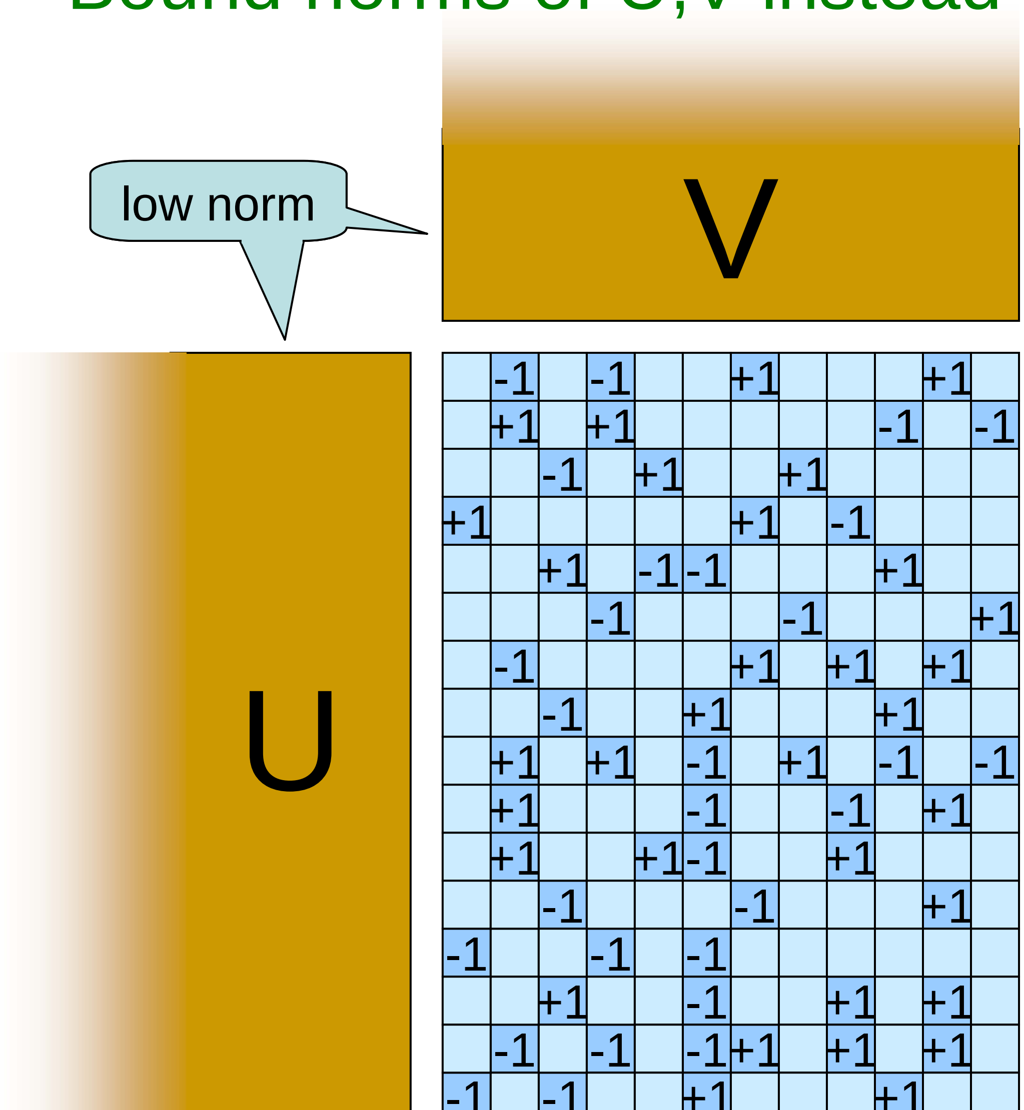
For observed $Y_{ij} \in \pm 1$:

$Y_{ij}\ X_{ij} \geq$ *Margin*

$\langle U_i, V_j \rangle$

Max-Margin Matrix Factorization:

Bound norms of U,V instead of their dimensionality



	-1		-1			+1				+1	
	+1		+1						-1		-1
		-1		+1			+1				
+1						+1		-1			
		+1		-1	-1				+1		
			-1				-1				+1
	-1					+1		+1		+1	
		-1			+1				+1		
	+1		+1		-1		+1		-1		-1
	+1				-1			-1		+1	
	+1			+1	-1			+1			
		-1				-1				+1	
-1			-1		-1						
		+1			-1			+1		+1	
	-1		-1		-1	+1		+1		+1	
-1		-1			+1				+1		

bound norms uniformly:

$(\max_i|U_i|^2)\ (\max_j|V_j|^2) \leq 1$

bound norms on average:

$(\sum_i |U_i|^2)\ (\sum_j |V_j|^2) \leq 1$

U is fixed:
each column of V is SVM

For observed $Y_{ij} \in \pm 1$:

$Y_{ij}\ X_{ij} \geq$ *Margin*

$\langle U_i, V_j \rangle$

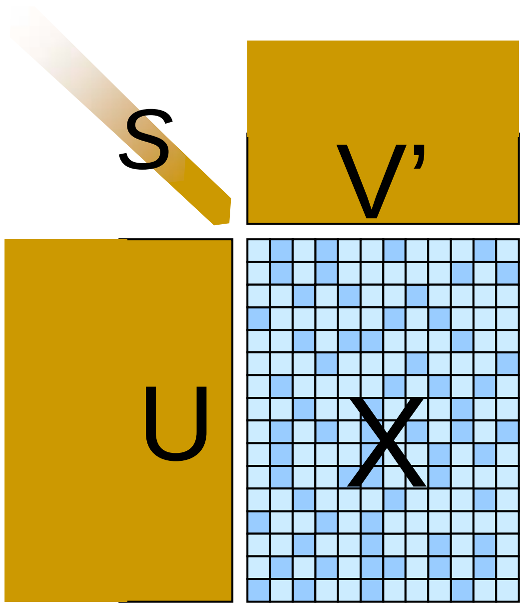
Finding Max-Margin Matrix Factorizations

maximize M

$\mathbf{Y}_{ij}\,\mathbf{X}_{ij} \geq M$

$\mathbf{X} = \mathbf{U}\mathbf{V}$

$$\underbrace{(\textstyle\sum_i |\mathbf{U}_i|^2)\,(\textstyle\sum_j |\mathbf{V}_j|^2)}_{|\mathbf{X}|_{tr} = \sum \text{(singular values of } \mathbf{X}\text{)}} \leq 1$$



Unlike rank(X) $\leq k$, this a convex constraint!

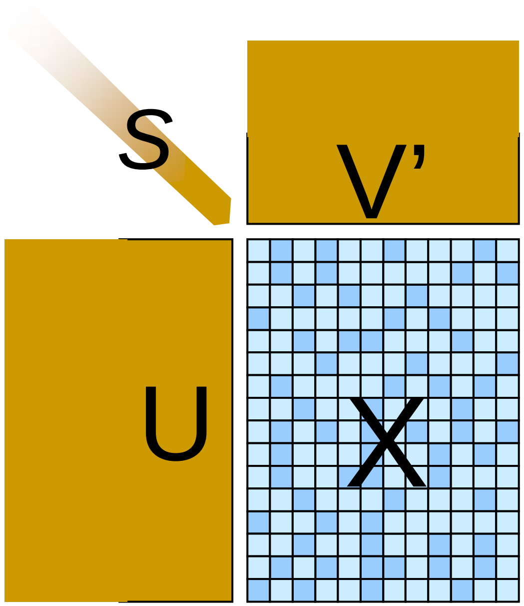
Finding Max-Margin Matrix Factorizations

maximize M

$\mathbf{Y}_{ij}\, \mathbf{X}_{ij} \geq M$

$\mathbf{X} = \mathbf{U}\mathbf{V}$

$$\underbrace{(\textstyle\sum_i |\mathbf{U}_i|^2)\ (\textstyle\sum_j |\mathbf{V}_j|^2)}_{|\mathbf{X}|_{tr} = \sum \text{(singular values of } \mathbf{X})} \leq 1$$

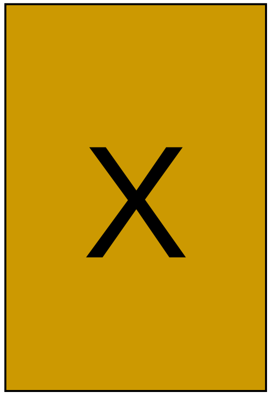


minimize $\mathrm{tr}(\mathbf{A}) + \mathrm{tr}(\mathbf{B})$

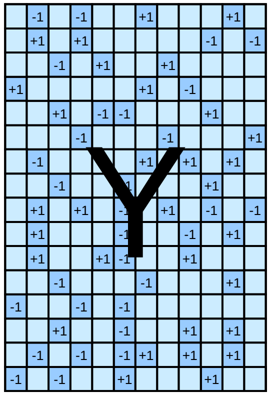
$\mathbf{Y}_{ij}\, \mathbf{X}_{ij} \geq 1$

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

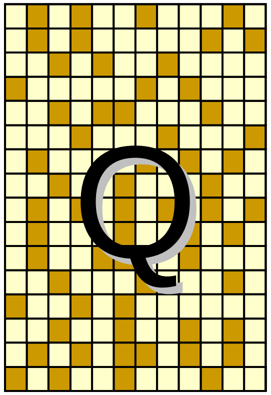
Finding Max-Margin Matrix Factorizations



dense primal



sparse observations
(constraints)



sparse dual

minimize $\mathrm{tr}(\mathbf{A})+\mathrm{tr}(\mathbf{B})$

$$\mathbf{Y}_{ij}\,\mathbf{X}_{ij} \geq 1$$

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

Dual variable $\mathbf{Q}_{ij}$ for each observed (i,j)

maximize $\sum \mathbf{Q}_{ij}$

$$0 \leq \mathbf{Q}_{ij}$$

$$||\mathbf{Q} \otimes \mathbf{Y}||_2 \leq 1$$

sparse elementwise product
(zero for unobserved entries)

Experimental Results on MovieLens Subset

		MMMF	K-median	Low Rank	per-user median rating
holdout set	1	**43.8**	41.5	42.5	41.3
	2	**44.7**	43.3	46.8	41.6
	3	**47.3**	43.9	45.7	43.2
	4	**45.0**	42.8	44.7	41.4
overall		**45.2**	42.9	44.2	41.9

% rating agreement (ratings are 1,2,3,4,5)

100 users $\times$ 100 movies, 7030 ratings

Generalization Error Bounds

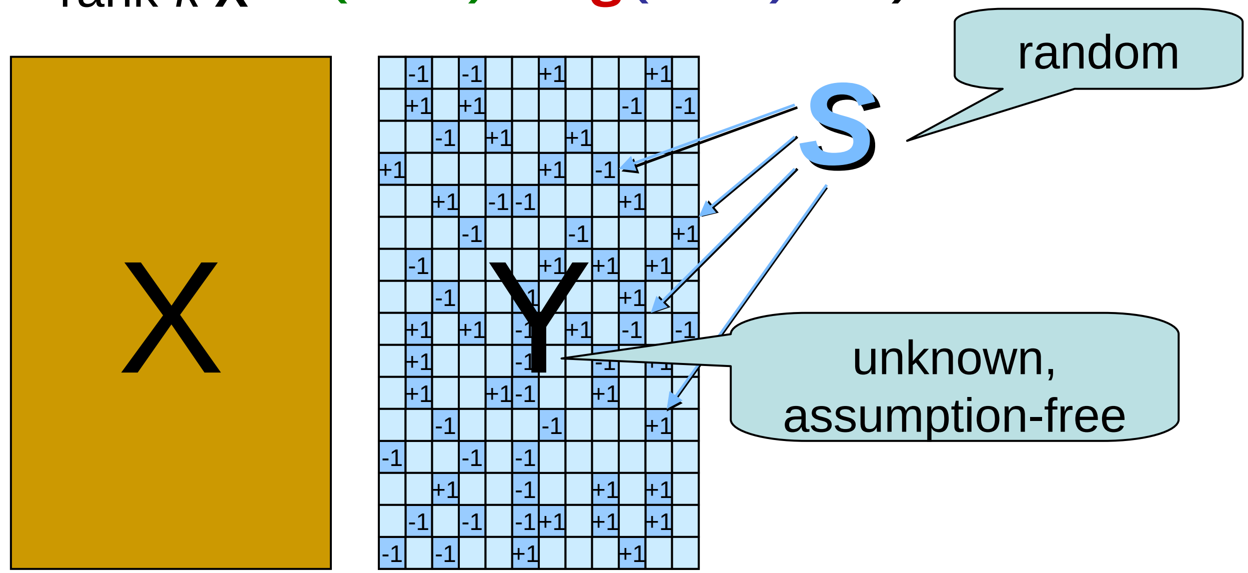
$D(X;Y) = \#_{ij}(X_{ij} \cdot Y_{ij} < 0)/nm$

generalization error

$D_S(X;Y) = \#_{ij \in S}(X_{ij} \cdot Y_{ij} < 1)/|S|$

empirical error

$$\forall_Y \Pr_S \left(\forall_{\text{rank-}k\ X}\ D(X;Y) < D_S(X;Y) + \varepsilon \right) > 1 - \delta$$



$(\sum |U_i|^2/n)(\sum |V_i|^2/m) \le R^2$: $\quad \varepsilon = K\sqrt[4]{\ln m}\sqrt{\frac{R^2(n+m)\log n + \log 1/\delta}{|S|}}$

rank(X) ≤ k: $\quad \varepsilon = \sqrt{\frac{k(n+m)\log\frac{8em}{k} + \log 1/\delta}{2|S|}}$

Maximum Margin Matrix Factorization as a Convex Combination of Classifiers

$$\{\, UV \mid (\textstyle\sum |U_i|^2)(\textstyle\sum |V_i|^2) \le 1 \,\}$$
$$= \text{convex-hull}(\, \underbrace{\{\, uv' \mid u \in R^n,\ v \in R^m\ |u|=|v|=1\}}_{\text{rank-one, unit-norm, matrices}} \,)$$

$$\text{conv}(\, \{\, uv' \mid u \in \{\pm 1\}^n,\ v \in \{\pm 1\}^m\} \,)$$
$$\subset \{\, UV \mid (\max |U_i|^2)(\max |V_j|^2) \le 1 \,\}$$
$$\subset 2\ \text{conv}(\, \underbrace{\{\, uv' \mid u \in \{\pm 1\}^n,\ v \in \{\pm 1\}^m\}}_{\text{rank-one sign matrices}} \,)$$

Maximum Margin Matrix Factorization

Unbounded number of factors

Learning is a convex problem!

- Correspondence with large margin linear classification
- Generalization error bounds
- Learning: Sparse SDP

poster
  - Making predictions using the dual solution
  - Other data types and loss functions
  - SDP for max-norm formulation
  - Slack

- Applicable in other domains where low-rank approximations are currently used
- Direct optimization of dual would enable large-scale applications

MATLAB code: http://www.cs.toronto.edu/~nati/mmmf