Fast **M**aximum **M**argin **M**atrix **F**actorization for Collaborative Prediction

Jason Rennie
MIT

Nati Srebro
Univ. of Toronto

Collaborative Prediction

Based on partially observed matrix:

$\Rightarrow$ Predict unobserved entries "Will user *i* like movie *j*?"

movies

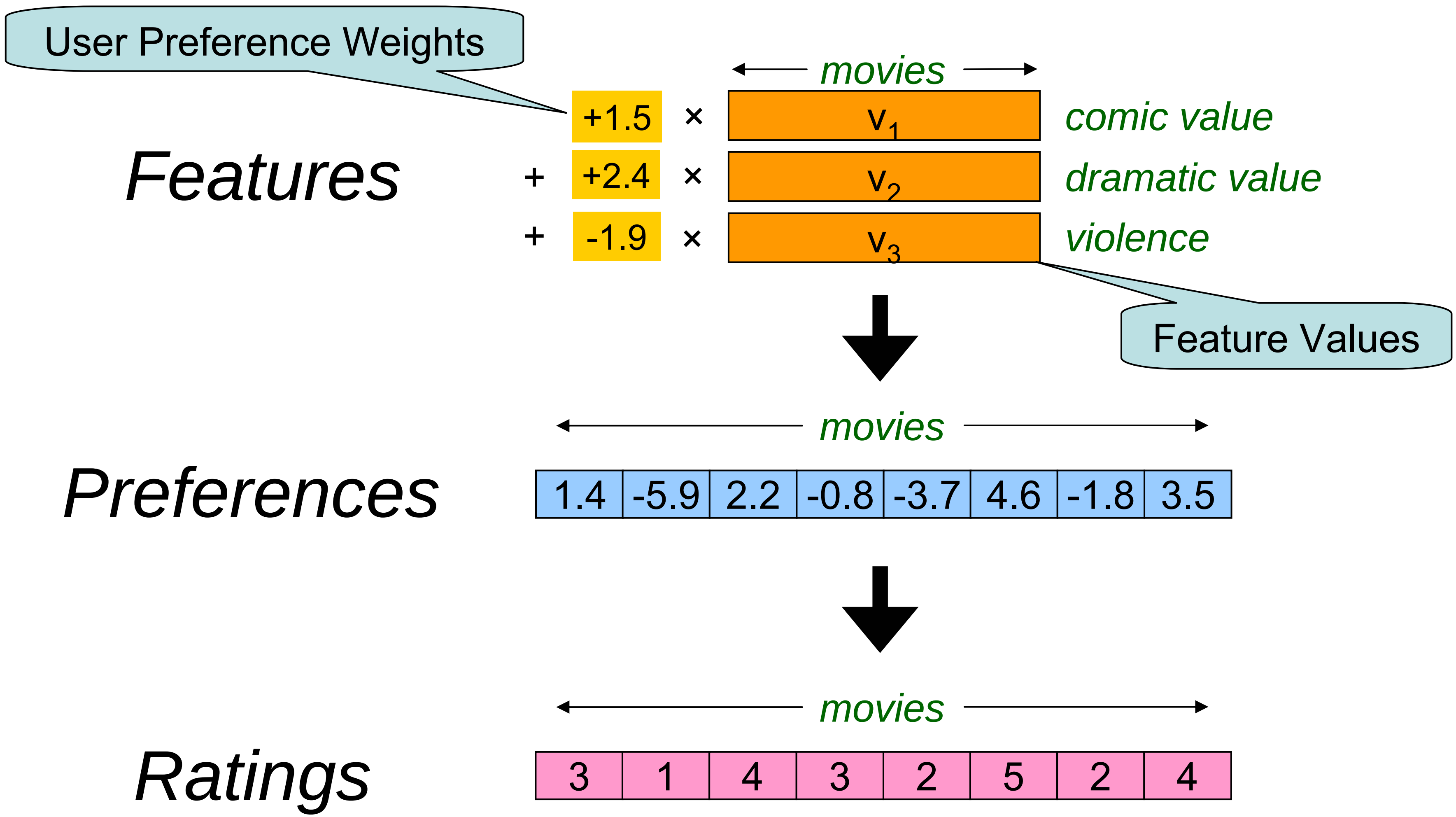
users

	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	?	

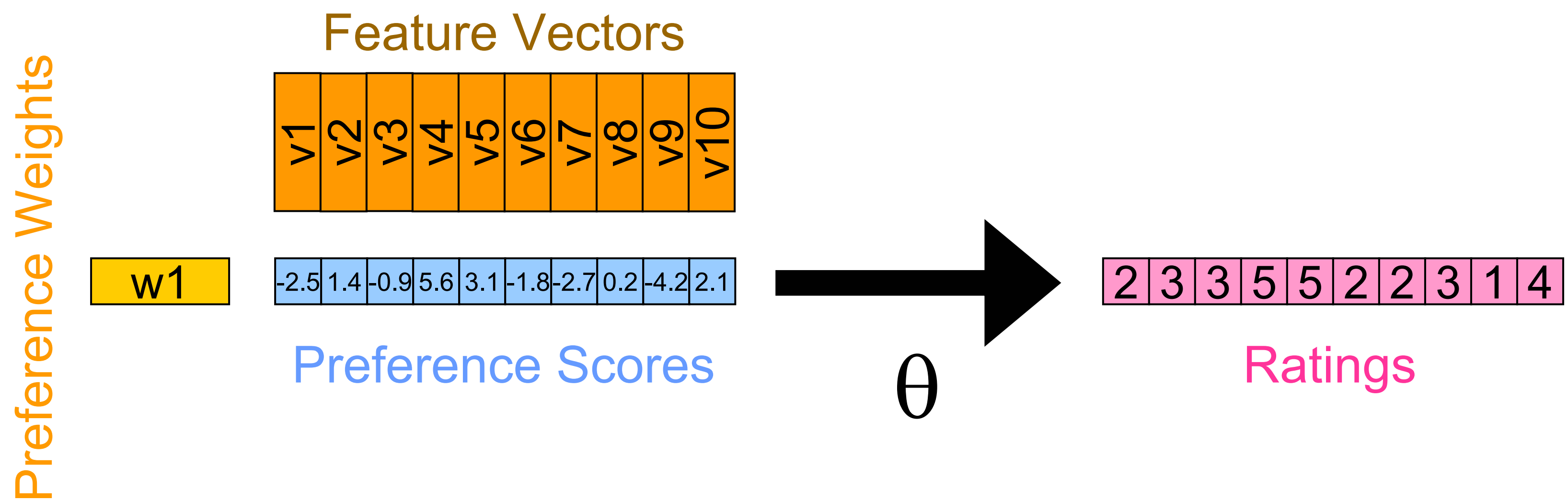
Problems to Address

- Underlying representation of preferences
  - Norm constrained matrix factorization (MMMF)
- Discrete, ordered labels
  - Threshold-based ordinal regression
- **Scaling-up MMMF to large problems**
  - **Factorized objective, gradient descent**
- *Ratings may not be missing at random*

Linear Factor Model



Ordinal Regression



Matrix Factorization

Feature Vectors

	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10
w1	-2.5	1.4	-0.9	5.6	3.1	-1.8	-2.7	0.2	-4.2	2.1
w2	0.2	-4.2	2.1	-2.5	1.4	-0.9	5.6	3.1	-1.8	-2.7
w3	3.1	-1.8	-2.7	0.2	-4.2	2.1	-2.5	1.4	-0.9	5.6
w4	1.4	-0.9	5.6	3.1	-1.8	-2.7	0.2	-4.2	2.1	-2.5
w5	-4.2	2.1	-2.5	1.4	-0.9	5.6	3.1	-1.8	-2.7	0.2
w6	-1.8	-2.7	0.2	-4.2	2.1	-2.5	1.4	-0.9	5.6	3.1
w7	-0.9	5.6	3.1	-1.8	-2.7	0.2	-4.2	2.1	-2.5	1.4
w8	2.1	-2.5	1.4	-0.9	5.6	3.1	-1.8	-2.7	0.2	-4.2
w9	-2.7	0.2	-4.2	2.1	-2.5	1.4	-0.9	5.6	3.1	-1.8
w10	5.6	3.1	-1.8	-2.7	0.2	-4.2	2.1	-2.5	1.4	-0.9
w11	-2.5	1.4	-0.9	5.6	3.1	-1.8	-2.7	0.2	-4.2	2.1
w12	0.2	-4.2	2.1	-2.5	1.4	-0.9	5.6	3.1	-1.8	-2.7
w13	3.1	-1.8	-2.7	0.2	-4.2	2.1	-2.5	1.4	-0.9	5.6
w14	1.4	-0.9	5.6	3.1	-1.8	-2.7	0.2	-4.2	2.1	-2.5
w15	-4.2	2.1	-2.5	1.4	-0.9	5.6	3.1	-1.8	-2.7	0.2

Preference Weights

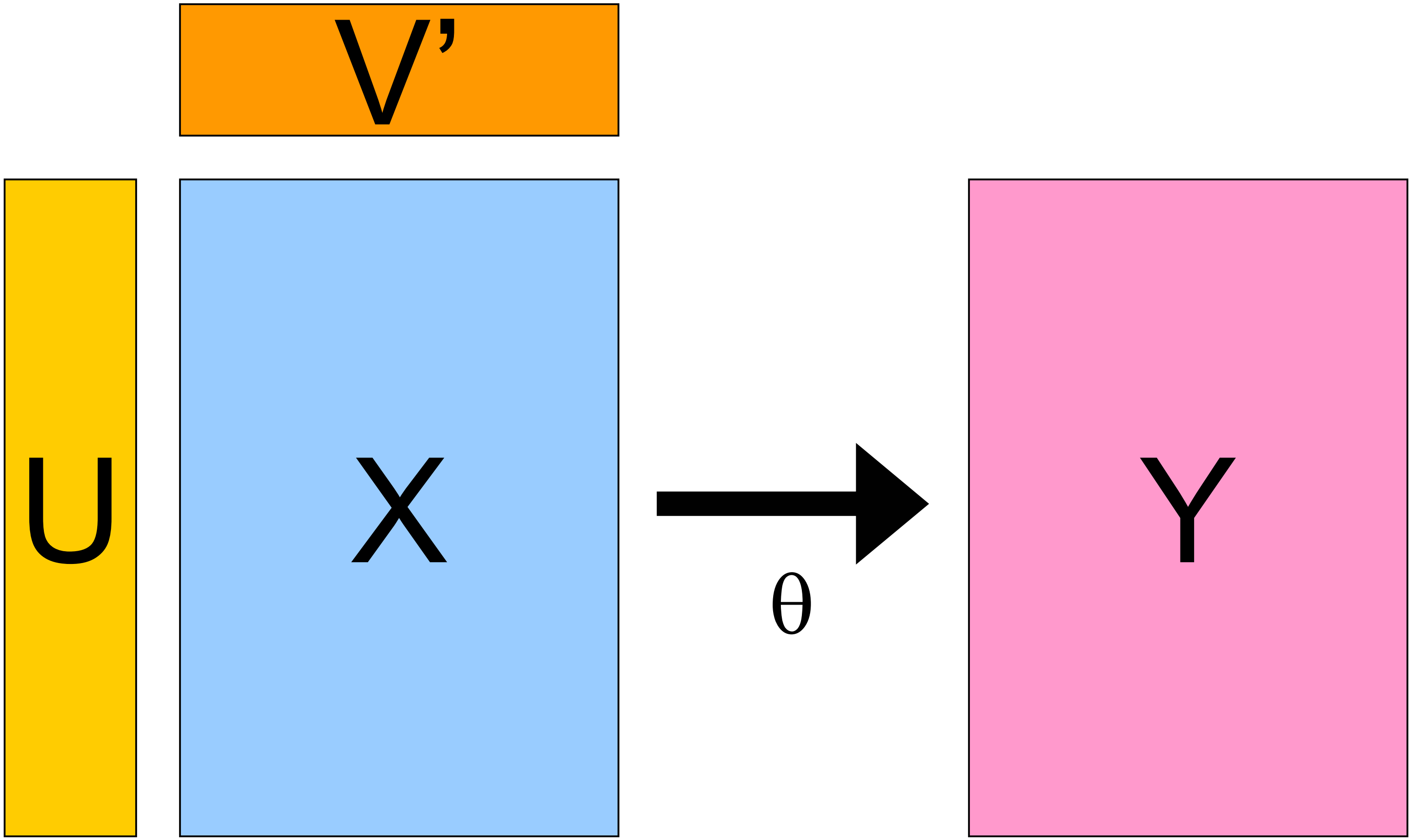
Preference Scores

$\xrightarrow{\theta}$

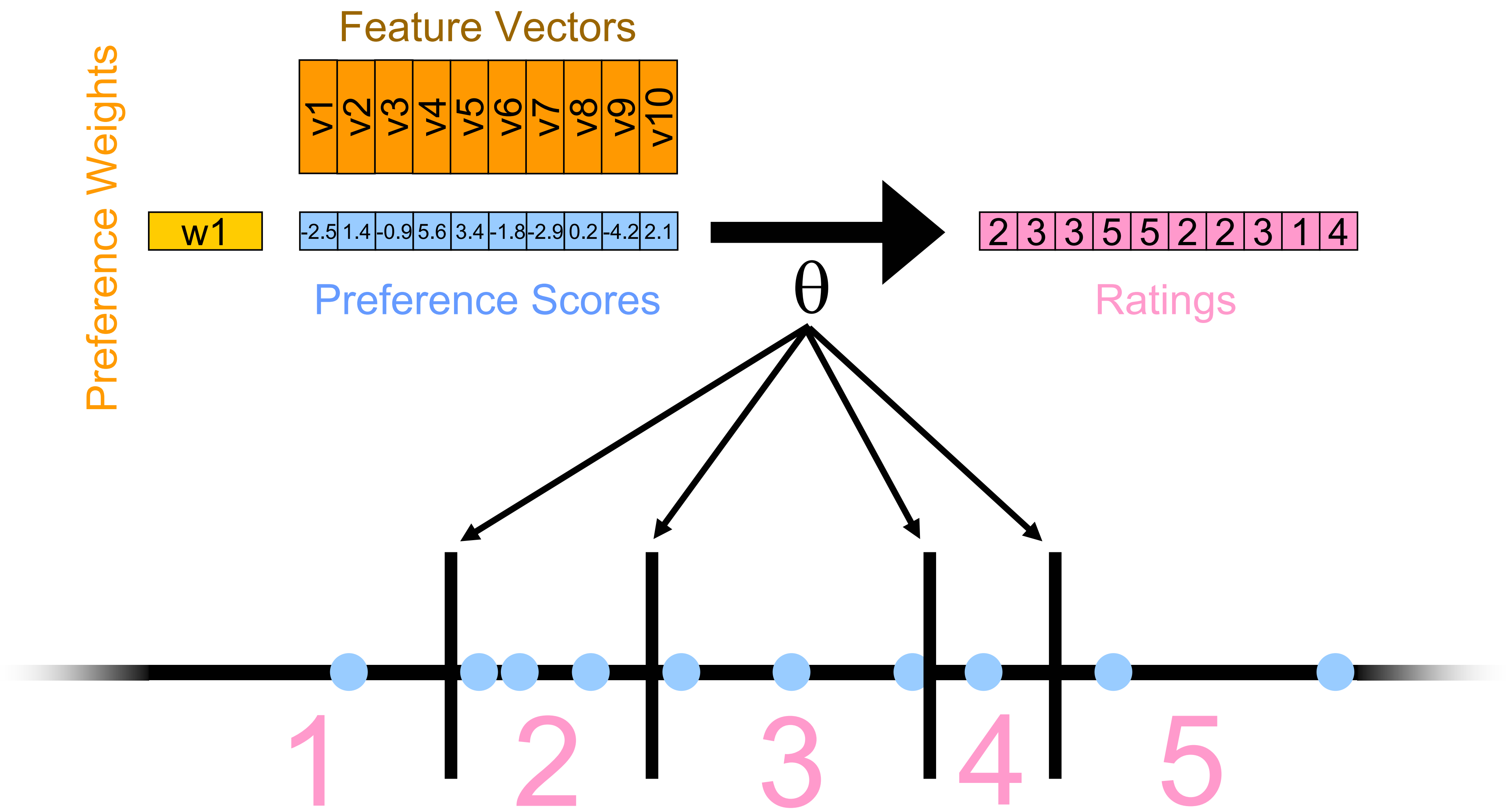
2	3	3	5	5	2	2	3	1	4
3	1	4	2	3	3	5	5	2	2
5	2	2	3	1	4	2	3	3	5
3	3	5	5	2	2	3	1	4	2
1	4	2	3	3	5	5	2	2	3
2	2	3	1	4	2	3	3	5	5
3	5	5	2	2	3	1	4	2	3
4	2	3	3	5	5	2	2	3	1
2	3	1	4	2	3	3	5	5	2
5	5	2	2	3	1	4	2	3	3
2	3	3	5	5	2	2	3	1	4
3	1	4	2	3	3	5	5	2	2
5	2	2	3	1	4	2	3	3	5
3	3	5	5	2	2	3	1	4	2
1	4	2	3	3	5	5	2	2	3

Ratings

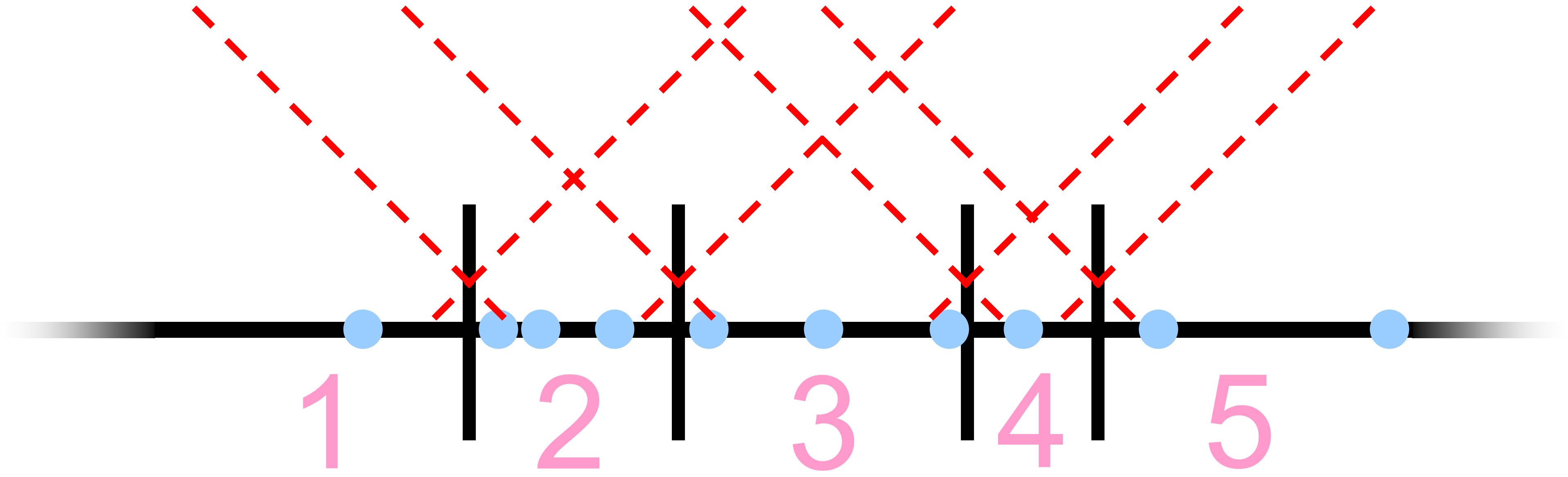
Matrix Factorization



Ordinal Regression



Max-Margin Ordinal Regression

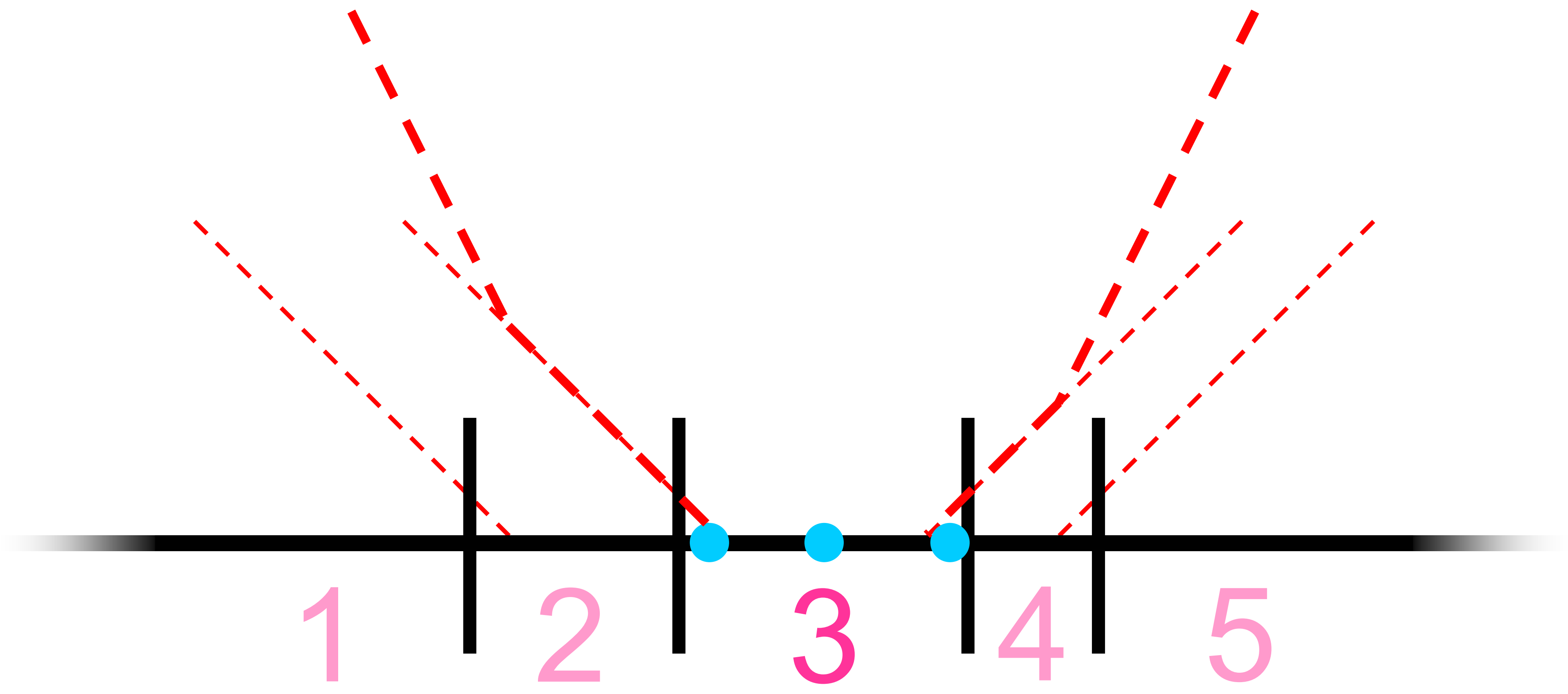


[Shashua & Levin, NIPS 2002]

Absolute Difference

- Shashua & Levin's loss bounds the misclassification error
- Ordinal Regression: we want to minimize the absolute difference between labels

All-Thresholds Loss



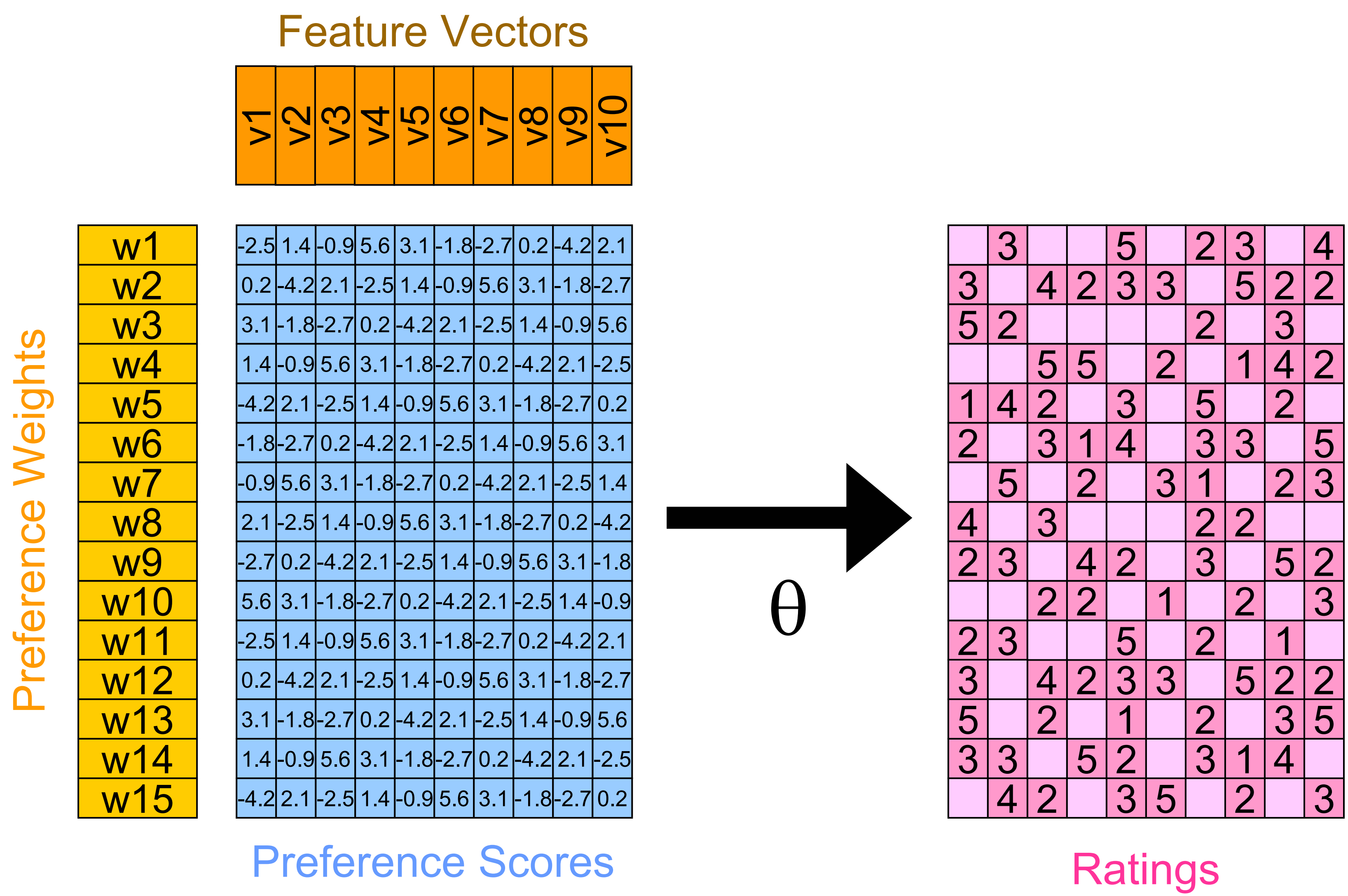
[Srebro et al., NIPS 2004] [Chu & Keerthi, ICML 2005]

All-Thresholds Loss

- Experiments comparing:
  - Least squares regression
  - Multi-class classification
  - Shashua & Levin’s Max-Margin OR
  - All-Thresholds OR
- All-Thresholds Ordinal Regression
  - Lowest misclassification error
  - Lowest absolute difference error

[Rennie & Srebro, IJCAI Wkshp 2005]

Learning Weights & Features



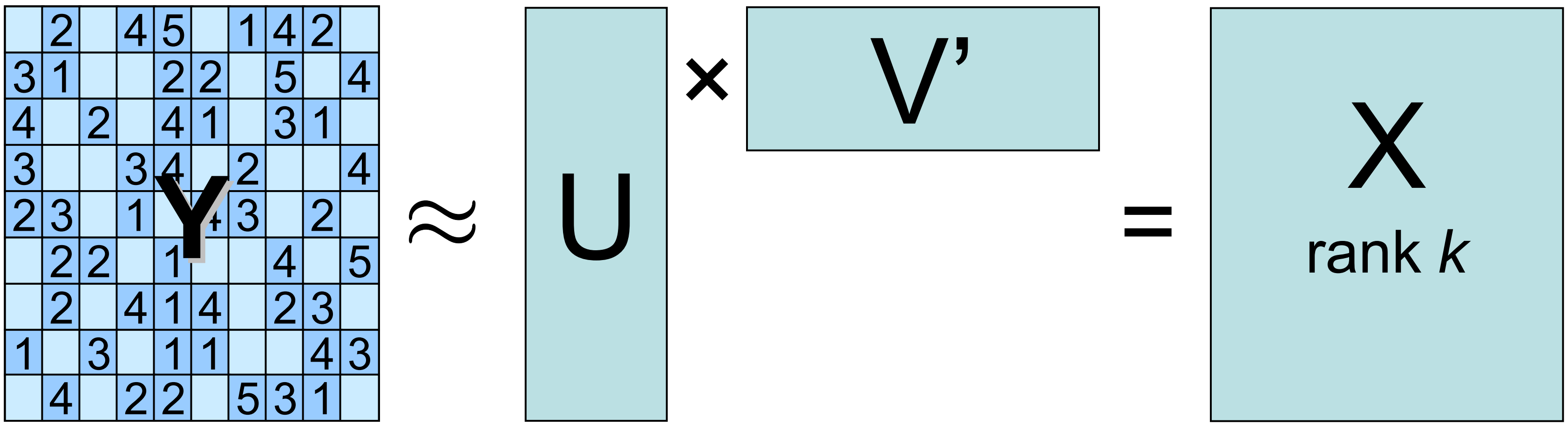
Preference Scores

	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10
w1	-2.5	1.4	-0.9	5.6	3.1	-1.8	-2.7	0.2	-4.2	2.1
w2	0.2	-4.2	2.1	-2.5	1.4	-0.9	5.6	3.1	-1.8	-2.7
w3	3.1	-1.8	-2.7	0.2	-4.2	2.1	-2.5	1.4	-0.9	5.6
w4	1.4	-0.9	5.6	3.1	-1.8	-2.7	0.2	-4.2	2.1	-2.5
w5	-4.2	2.1	-2.5	1.4	-0.9	5.6	3.1	-1.8	-2.7	0.2
w6	-1.8	-2.7	0.2	-4.2	2.1	-2.5	1.4	-0.9	5.6	3.1
w7	-0.9	5.6	3.1	-1.8	-2.7	0.2	-4.2	2.1	-2.5	1.4
w8	2.1	-2.5	1.4	-0.9	5.6	3.1	-1.8	-2.7	0.2	-4.2
w9	-2.7	0.2	-4.2	2.1	-2.5	1.4	-0.9	5.6	3.1	-1.8
w10	5.6	3.1	-1.8	-2.7	0.2	-4.2	2.1	-2.5	1.4	-0.9
w11	-2.5	1.4	-0.9	5.6	3.1	-1.8	-2.7	0.2	-4.2	2.1
w12	0.2	-4.2	2.1	-2.5	1.4	-0.9	5.6	3.1	-1.8	-2.7
w13	3.1	-1.8	-2.7	0.2	-4.2	2.1	-2.5	1.4	-0.9	5.6
w14	1.4	-0.9	5.6	3.1	-1.8	-2.7	0.2	-4.2	2.1	-2.5
w15	-4.2	2.1	-2.5	1.4	-0.9	5.6	3.1	-1.8	-2.7	0.2

Ratings

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

Low Rank Matrix Factorization

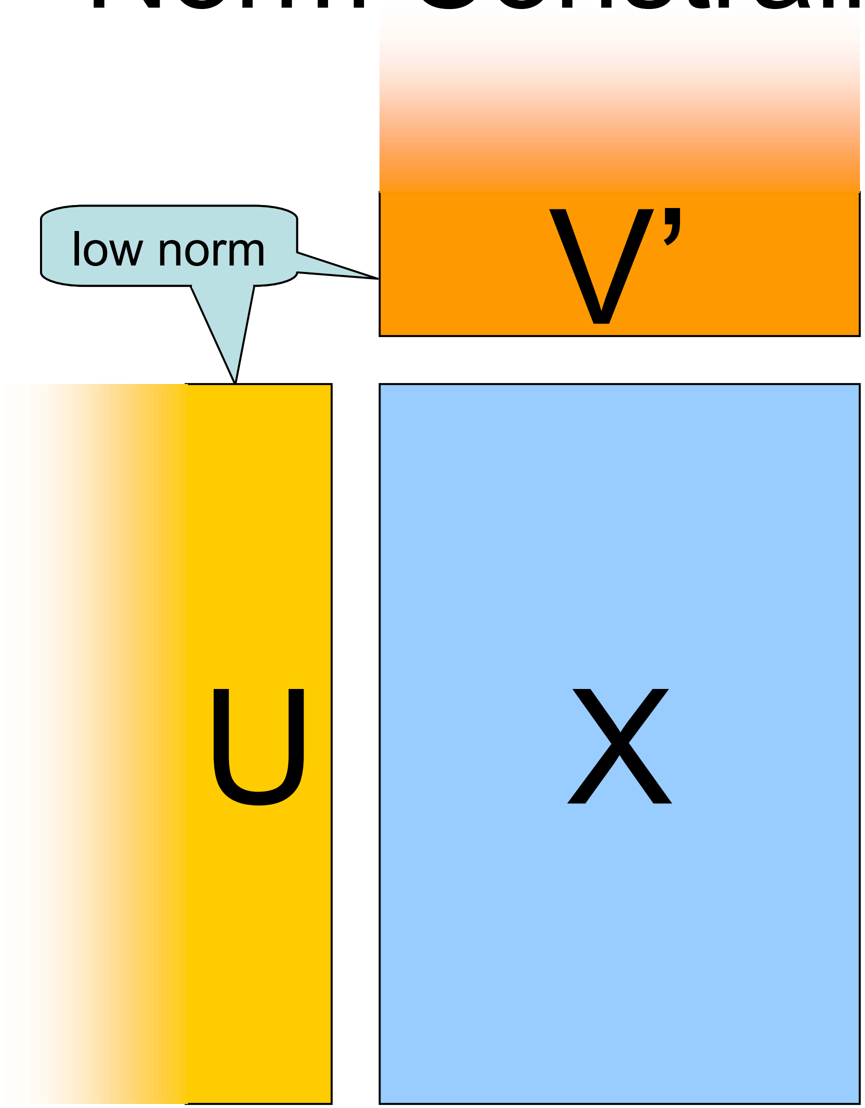


- **Sum-Squared Loss**
- **Fully Observed Y**
- **Classification Error Loss**
- **Partially Observed Y**

Use SVD to find Global Optimum

Non-convex No explicit soln.

Norm Constrained Factorization



$$\|X\|_{tr} = \min_{U,V} \left(\|U\|_{Fro}^2 + \|V\|_{Fro}^2\right)/2$$

$$\|U\|_{Fro}^2 = \sum_{i,j} U_{ij}^2$$

[Fazel et al., 2001]

MMMF Objective

Original Objective

All-Thresholds

$$\min_X \|X\|_{tr} + c \text{ loss}(X,Y)$$

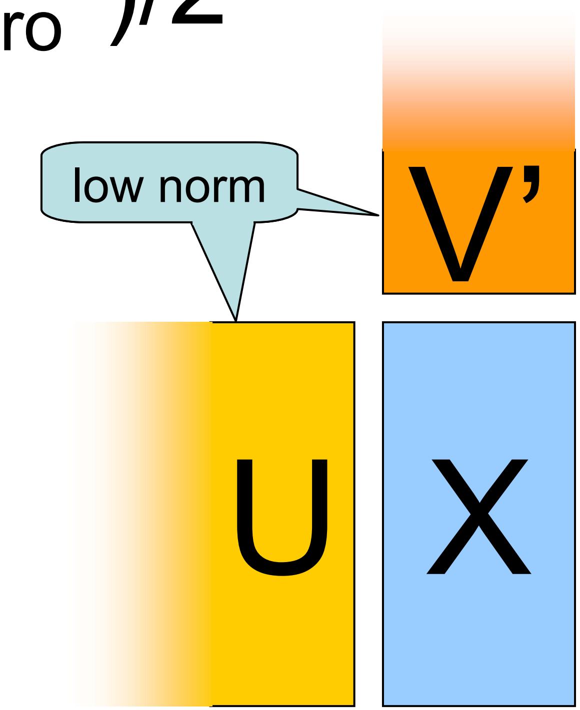
Factorized Objective

$$\min_{U,V} (\|U\|_{Fro}^2 + \|V\|_{Fro}^2)/2 + c \text{ loss}(UV',Y)$$

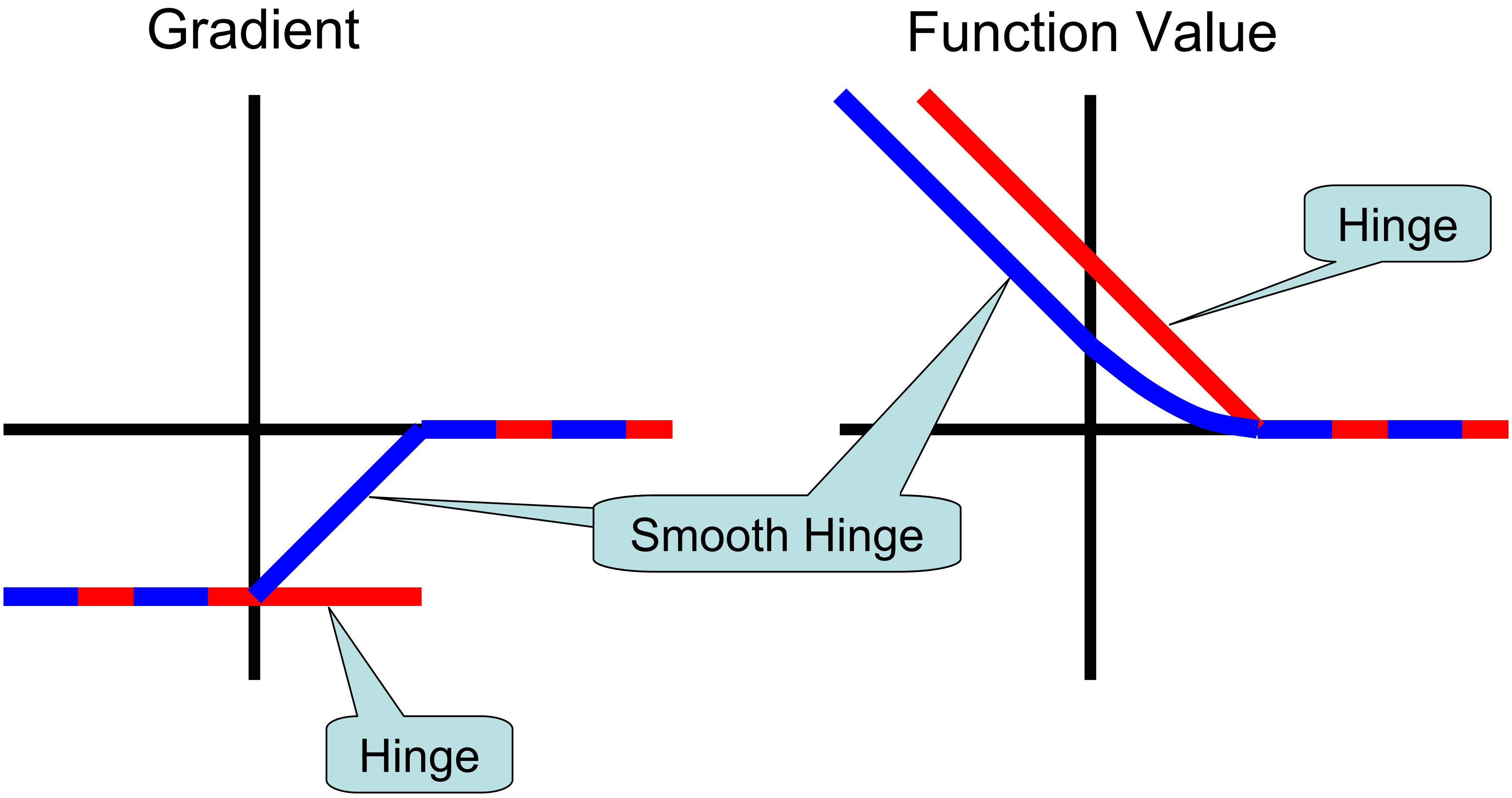
All-Thresholds

[Srebro et al., NIPS 2004]

$$\|U\|_{Fro}^2 = \sum_{i,j} U_{ij}^2$$



Smooth Hinge



Collaborative Prediction Results

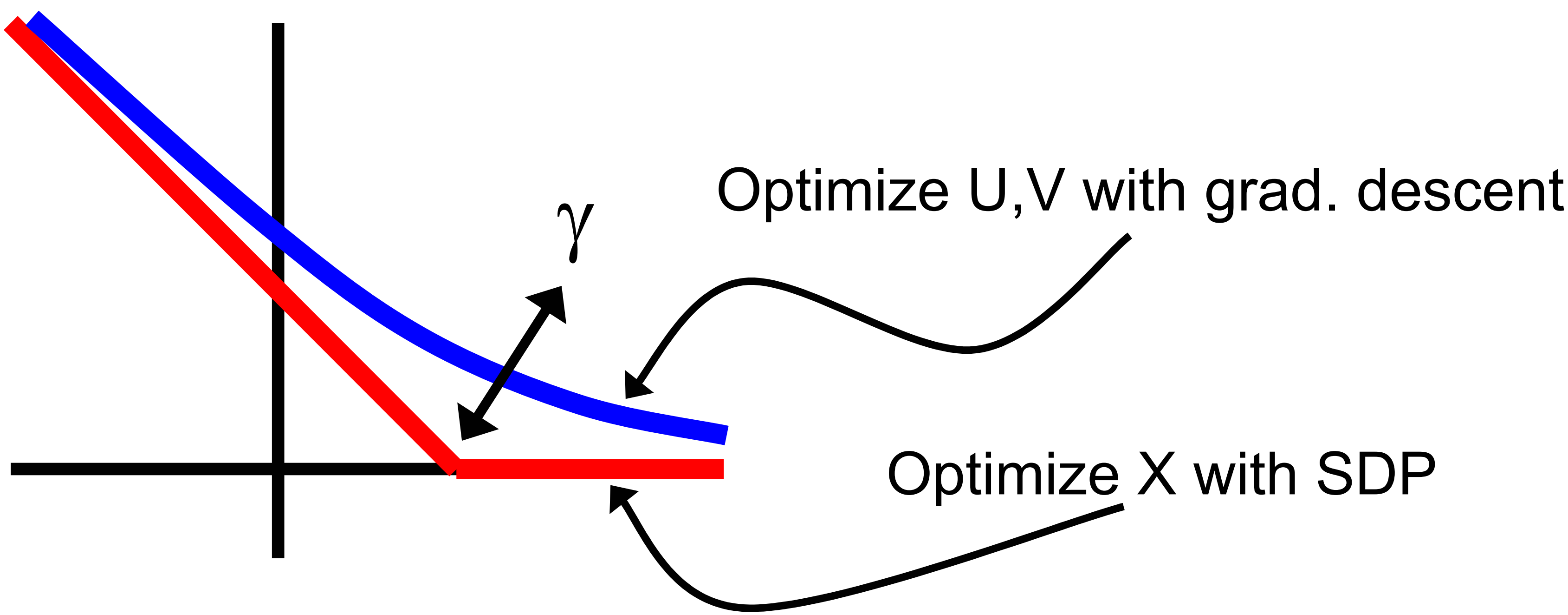
	EachMovie		MovieLens	
size, sparsity:	36656x1648, 96%		6040x3952, 96%	
Algorithm	Weak Error	Strong Error	Weak Error	Strong Error
URP	.4422	.4557	.4341	.4444
Attitude	.4520	.4550	.4320	.4375
MMMF	**.4397**	**.4341**	**.4156**	**.4203**

URP & Attitude Results: [Marlin, 2004]

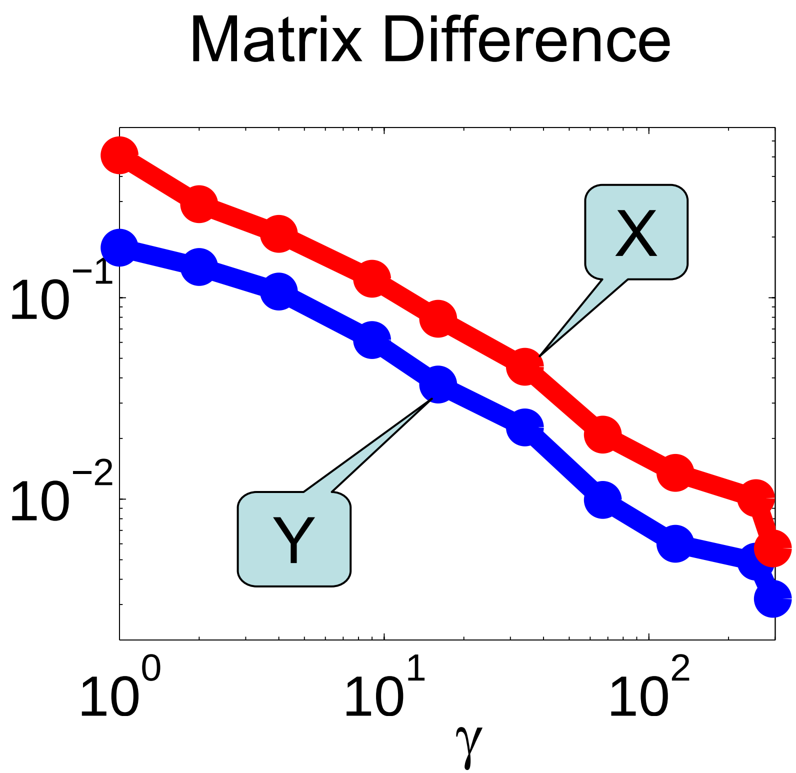
Local Minima?

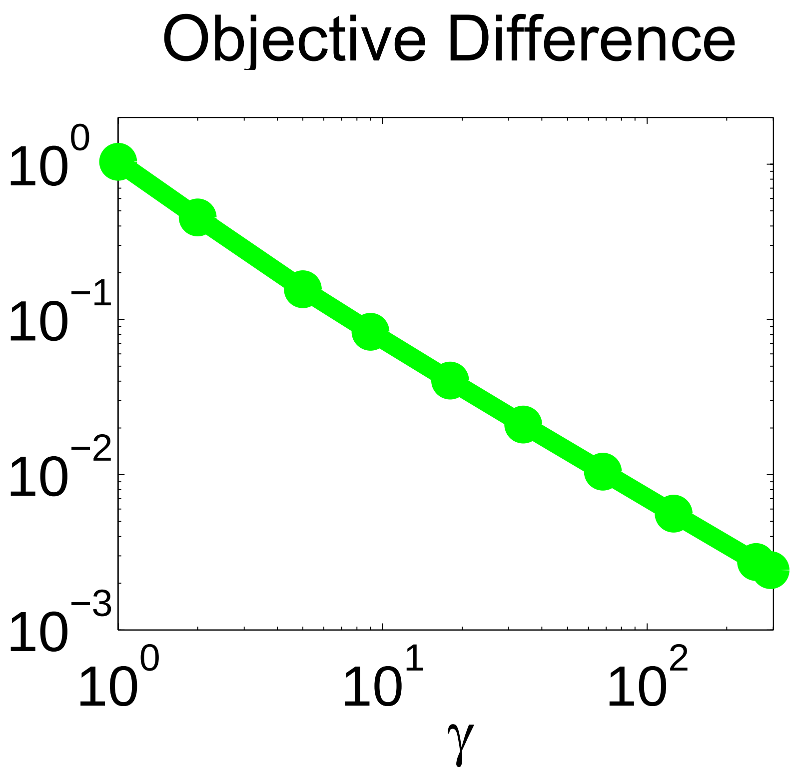
Factorized Objective

$$\min_{U,V} (||U||_{Fro}^2 + ||V||_{Fro}^2)/2 + c \; \mathrm{loss}(UV', Y)$$



Local Minima?





Data: 100 x 100 MovieLens, 65% sparse

Summary

- We scaled MMMF to large problems by optimizing the Factorized Objective
- Empirical tests indicate that local minima issues are rare or absent
- Results on large-scale data show substantial improvements over state-of-the-art

D'Aspremont & Srebro: large-scale SDP optimization methods. Train on 1.5 million binary labels in 20 hours.