ECS289: Scalable Machine Learning

Cho-Jui Hsieh
UC Davis

Oct 15, 2015

Outline

- Matrix Completion (Background)
- Alternating Least Squares (ALS)
- Stochastic Gradient method (SG)
- Coordinate Descent (CD)

Recommender Systems

Rating Matrix

Items

Users

Movie 1	Movie 2								Movie 10	Movie 11
1			5			3		5		2
	2		3			5		2	5	
				3	?	5		3		
2		5			3		4		2	
			5			5				1
	5			1				5		
1			1				2			4

Matrix Factorization Approach $A \approx WH^T$

H^T

-0.07	-0.11	-0.53	-0.46	-0.06	-0.05	-0.53	-0.07	-0.35	-0.19	-0.14
0.13	-0.42	0.45	0.17	-0.25	-0.17	-0.18	0.27	-0.59	0.05	0.14
-0.21	-0.43	-0.23	0.16	0.08	0.17	0.57	-0.39	-0.37	-0.08	-0.15

W

-8.72	0.03	-1.03
-7.56	-0.79	0.62
-4.07	-3.95	2.55
-3.52	3.73	-3.32
-7.78	2.34	2.33
-2.44	-5.29	-3.92
-1.78	1.90	-1.68

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

Matrix Factorization Approach $A \approx WH^T$

H^T

-0.07	-0.11	-0.53	-0.46	-0.06	-0.05	-0.53	-0.07	-0.35	-0.19	-0.14
0.13	-0.42	0.45	0.17	-0.25	-0.17	-0.18	0.27	-0.59	0.05	0.14
-0.21	-0.43	-0.23	0.16	0.08	0.17	0.57	-0.39	-0.37	-0.08	-0.15

W

-8.72	0.03	-1.03
-7.56	-0.79	0.62
-4.07	-3.95	2.55
-3.52	3.73	-3.32
-7.78	2.34	2.33
-2.44	-5.29	-3.92
-1.78	1.90	-1.68

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

Matrix Factorization Approach

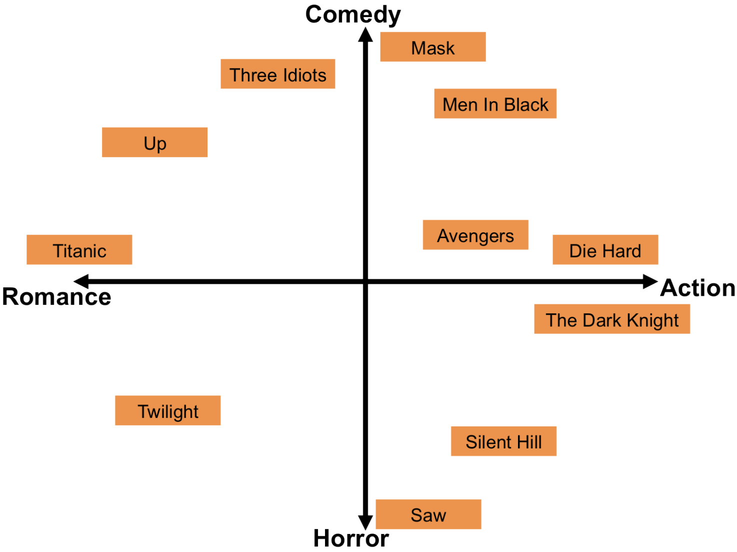
$$\min_{\substack{W \in \mathbb{R}^{m \times k} \\ H \in \mathbb{R}^{n \times k}}} \sum_{(i,j) \in \Omega} (A_{ij} - \boldsymbol{w}_i^T \boldsymbol{h}_j)^2 + \lambda \left(\|W\|_F^2 + \|H\|_F^2 \right),$$

- $\Omega = \{(i,j) \mid A_{ij} \text{ is observed}\}$
- Regularized terms to avoid over-fitting

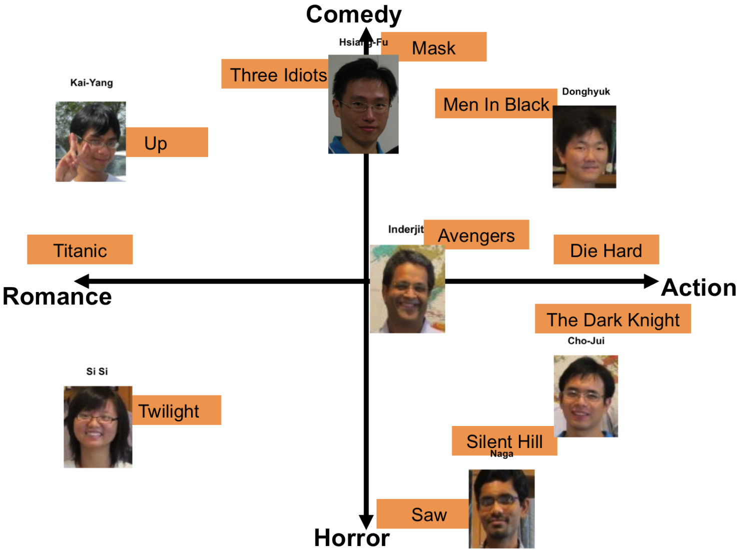
Matrix factorization maps users/items to latent feature space $\mathbb{R}^k$

- the i^{th} user $\Rightarrow$ i^{th} row of W, $\boldsymbol{w}_i^T$,
- the j^{th} item $\Rightarrow$ j^{th} row of H, $\boldsymbol{h}_j^T$.
- $\boldsymbol{w}_i^T \boldsymbol{h}_j$: measures the interaction between i^{th} user and j^{th} item.

Latent Feature Space



Latent Feature Space



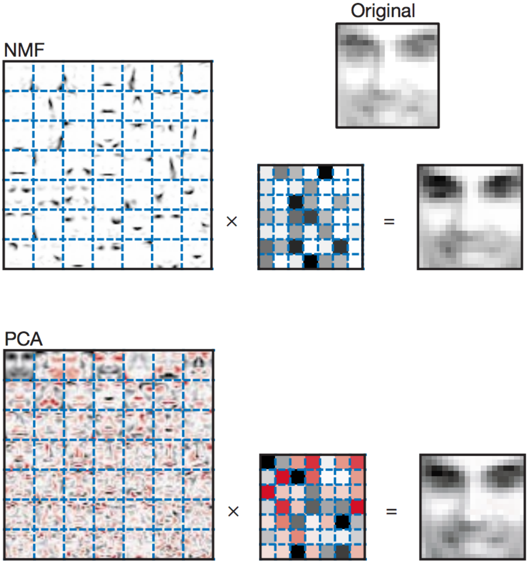
Other Factorizations

Nonnegative Matrix Factorization

$$\min_{W \geq 0, H \geq 0} \|A - WH^T\|_F^2 + \lambda \|W\|_F^2 + \lambda \|H\|_F^2$$

- Each entry is positive
- A is either fully or partially observed
- Goal: find nonnegative latent factors

NMF vs PCA

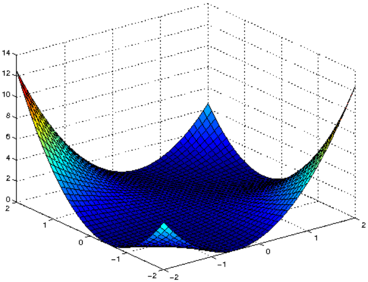


Optimization for Matrix Completion: Alternating Least Squares

Properties of the Objective Function

- Nonconvex problem (why?)
- Example: $f(x,y) = \frac{1}{2}(xy-1)^2$

  $\nabla f(0,0) = \mathbf{0}$, but clearly $[0,0]$ is not a global optimum



ALS: Alternating Least Squares

- Objective function:

$$\min_{W,H} \left\{ \frac{1}{2} \sum_{i,j \in \Omega} (A_{ij} - (WH^T)_{ij})^2 + \frac{\lambda}{2} \|W\|_F^2 + \frac{\lambda}{2} \|H\|_F^2 \right\} := f(W, H)$$

- Iteratively fix either H or W and optimize the other:

Input: partially observed matrix A, initial values of W, H
For $t = 1, 2, \ldots$
    Fix W and update H: $H \leftarrow \text{argmin}_H f(W, H)$
    Fix H and update W: $W \leftarrow \text{argmin}_W f(W, H)$

ALS: Alternating Least Squares

- Define: $\Omega_j := \{i \mid (i,j) \in \Omega\}$
- $\boldsymbol{w}_i$: the i-th row of W; $\boldsymbol{h}_j$: the j-th row of H
- The subproblem:

$$\begin{aligned}
\underset{H}{\operatorname{argmin}} & \frac{1}{2} \sum_{i,j\in\Omega} (A_{ij} - (WH^T)_{ij})^2 + \frac{\lambda}{2}\|H\|_F^2 \\
&= \sum_{j=1}^{n} \underbrace{\left(\frac{1}{2} \sum_{i\in\Omega_j} (A_{ij} - \boldsymbol{w}_i^T \boldsymbol{h}_j)^2 + \frac{\lambda}{2}\|\boldsymbol{h}_j\|^2 \right)}_{\text{ridge regression problem}}
\end{aligned}$$

ALS: Alternating Least Squares

- Define: $\Omega_j := \{i \mid (i,j) \in \Omega\}$
- $\boldsymbol{w}_i$: the i-th row of W; $\boldsymbol{h}_j$: the j-th row of H
- The subproblem:

$$\begin{aligned}
\operatorname*{argmin}_{H} \frac{1}{2} \sum_{i,j \in \Omega} (A_{ij} - (WH^T)_{ij})^2 + \frac{\lambda}{2}\|H\|_F^2 \\
= \sum_{j=1}^{n} \underbrace{\left(\frac{1}{2} \sum_{i \in \Omega_j} (A_{ij} - \boldsymbol{w}_i^T \boldsymbol{h}_j)^2 + \frac{\lambda}{2} \|\boldsymbol{h}_j\|^2 \right)}_{\text{ridge regression problem}}
\end{aligned}$$

- n ridge regression problems, each with k variables
  $\Rightarrow O(|\Omega| k^2 + nk^3)$
- Easy to parallelize (n independent ridge regression subproblems)

ALS: Alternating Least Squares

$$\begin{array}{cc} & \begin{pmatrix} \boxed{\quad H^T \quad} \end{pmatrix} \\ \begin{pmatrix} \boxed{\mathbf{w}_1^T} \\ \mathbf{w}_2^T \\ \mathbf{w}_3^T \end{pmatrix} & \begin{pmatrix} \boxed{A_{11} \;\; A_{12} \;\; A_{13}} \\ A_{21} \;\; A_{22} \;\; A_{23} \\ A_{31} \;\; A_{32} \;\; A_{33} \end{pmatrix} \end{array} \qquad \begin{array}{cc} & \begin{pmatrix} \boxed{\quad H^T \quad} \end{pmatrix} \\ \begin{pmatrix} \mathbf{w}_1^T \\ \boxed{\mathbf{w}_2^T} \\ \mathbf{w}_3^T \end{pmatrix} & \begin{pmatrix} A_{11} \;\; A_{12} \;\; A_{13} \\ \boxed{A_{21} \;\; A_{22} \;\; A_{23}} \\ A_{31} \;\; A_{32} \;\; A_{33} \end{pmatrix} \end{array}$$

Optimization for Matrix Completion: Stochastic Gradient Method

Stochastic Gradient Method

- n_i^W : number of nonzeroes in the i-th row of A
  n_j^H :number of nonzeroes in the j-th column of A
- Decompose the problem into Ω components:

$$f(W,H) = \frac{1}{2}\sum_{i,j\in\Omega}(A_{ij} - \boldsymbol{w}_i^T\boldsymbol{h}_j)^2 + \frac{\lambda}{2}\|W\|_F^2 + \frac{\lambda}{2}\|H\|_F^2$$

$$= \frac{1}{|\Omega|}\sum_{i,j\in\Omega}\underbrace{\left(\frac{|\Omega|}{2}(A_{ij} - \boldsymbol{w}_i^T\boldsymbol{h}_j)^2 + \frac{\lambda|\Omega|}{2n_i^W}\|\boldsymbol{w}_i\|^2 + \frac{\lambda|\Omega|}{2n_j^H}\|\boldsymbol{h}_j\|^2\right)}_{f_{i,j}(W,H)}$$

Stochastic Gradient Method

- n_i^W : number of nonzeroes in the i-th row of A
  n_j^H :number of nonzeroes in the j-th column of A
- Decompose the problem into Ω components:

$$f(W,H) = \frac{1}{2}\sum_{i,j\in\Omega}(A_{ij} - \boldsymbol{w}_i^T\boldsymbol{h}_j)^2 + \frac{\lambda}{2}\|W\|_F^2 + \frac{\lambda}{2}\|H\|_F^2$$

$$= \frac{1}{|\Omega|}\sum_{i,j\in\Omega}\underbrace{\left(\frac{|\Omega|}{2}(A_{ij} - \boldsymbol{w}_i^T\boldsymbol{h}_j)^2 + \frac{\lambda|\Omega|}{2n_i^W}\|\boldsymbol{w}_i\|^2 + \frac{\lambda|\Omega|}{2n_j^H}\|\boldsymbol{h}_j\|^2\right)}_{f_{i,j}(W,H)}$$

- The gradient of each component:

$$\nabla_{\boldsymbol{w}_i} f_{i,j}(W,H) = |\Omega|(\boldsymbol{w}_i^T\boldsymbol{h}_j - A_{ij})\boldsymbol{h}_j + \frac{\lambda|\Omega|}{n_i^W}\boldsymbol{w}_i$$

$$\nabla_{\boldsymbol{h}_j} f_{i,j}(W,H) = |\Omega|(\boldsymbol{w}_i^T\boldsymbol{h}_j - A_{ij})\boldsymbol{w}_i + \frac{\lambda|\Omega|}{n_j^H}\boldsymbol{h}_j$$

Stochastic Gradient Method

- SG algorithm:

  Input; partially observed matrix A, initial values of W, H

  For $t = 1, 2, \ldots$

  Randomly pick a pair $(i, j) \in \Omega$

  $\boldsymbol{w}_i \leftarrow (1 - \frac{\eta_t \lambda}{\boldsymbol{n}_i^W})\boldsymbol{w}_i - \eta_t(\boldsymbol{w}_i^T \boldsymbol{h}_j - A_{ij})\boldsymbol{h}_j$

  $\boldsymbol{h}_j \leftarrow (1 - \frac{\eta_t \lambda}{\boldsymbol{n}_j^H})\boldsymbol{h}_j - \eta_t(\boldsymbol{w}_i^T \boldsymbol{h}_j - A_{ij})\boldsymbol{w}_i$

Stochastic Gradient Method

- SG algorithm:

  Input; partially observed matrix A, initial values of W, H

  For $t = 1, 2, \ldots$

  Randomly pick a pair $(i, j) \in \Omega$

  $\boldsymbol{w}_i \leftarrow (1 - \frac{\eta_t \lambda}{n_i^W})\boldsymbol{w}_i - \eta_t(\boldsymbol{w}_i^T \boldsymbol{h}_j - A_{ij})\boldsymbol{h}_j$

  $\boldsymbol{h}_j \leftarrow (1 - \frac{\eta_t \lambda}{n_j^H})\boldsymbol{h}_j - \eta_t(\boldsymbol{w}_i^T \boldsymbol{h}_j - A_{ij})\boldsymbol{w}_i$

- Time complexity: $O(k)$ per iteration; $O(|\Omega|k)$ for one pass of all observed entries.

Stochastic Gradient Method

$$\begin{pmatrix} \mathbf{h}_1 & \mathbf{h}_2 & \mathbf{h}_3 \end{pmatrix} \qquad \qquad \begin{pmatrix} \mathbf{h}_1 & \mathbf{h}_2; & \mathbf{h}_3 \end{pmatrix}$$

$$\begin{pmatrix} \mathbf{w}_1^T \\ \mathbf{w}_2^T \\ \mathbf{w}_3^T \end{pmatrix} \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix} \qquad \qquad \begin{pmatrix} \mathbf{w}_1^T \\ \mathbf{w}_2^T \\ \mathbf{w}_3^T \end{pmatrix} \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix}$$

Optimization for Matrix Completion:

Distributed Stochastic Gradient Descent (DSGD)

How to parallelize SG?

- Two SG updates on (i_1, j_1) and (i_2, j_2) in the same time:
  - (i_1, j_1): Update $\boldsymbol{w}_{i_1}$ and $\boldsymbol{h}_{j_1}$
  - (i_2, j_2): Update $\boldsymbol{w}_{i_2}$ and $\boldsymbol{h}_{j_2}$
- Confliction happens when $i_1 = i_2$ or $j_1 = j_2$
- How to avoid confliction?

  Gemulla et al., "Large-Scale Matrix Factorization with Distributed Stochastic Gradient Descent". In KDD 2011.

DSGD: Distributed SGD [Gemulla et al, 2011]

$$\begin{array}{ccc} P_1 & P_2 & P_3 \end{array}$$

$$\begin{pmatrix} \boldsymbol{h}_1 & \boldsymbol{h}_2 & \boldsymbol{h}_3 \end{pmatrix}$$

$$\begin{pmatrix} \boldsymbol{w}_1^T \\ \boldsymbol{w}_2^T \\ \boldsymbol{w}_3^T \end{pmatrix} \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix}$$

DSGD: Distributed SGD

$$\begin{array}{ccc} P_1 & P_2 & P_3 \end{array}$$

$$\begin{pmatrix} \boldsymbol{h}_1 & \boldsymbol{h}_2 & \boldsymbol{h}_3 \end{pmatrix}$$

$$\begin{pmatrix} \boldsymbol{w}_1^T \\ \boldsymbol{w}_2^T \\ \boldsymbol{w}_3^T \end{pmatrix} \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix}$$

DSGD: Distributed SGD

$$\begin{array}{ccc} P_1 & P_2 & P_3 \end{array}$$

$$\begin{pmatrix} \boldsymbol{h}_1 & \boldsymbol{h}_2 & \boldsymbol{h}_3 \end{pmatrix}$$

$$\begin{pmatrix} \boldsymbol{w}_1^T \\ \boldsymbol{w}_2^T \\ \boldsymbol{w}_3^T \end{pmatrix} \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix}$$

Optimization for Matrix Completion: Coordinate Descent

Coordinate Descent

Update a variable at a time:

$$w_{it} \leftarrow \frac{\sum_{j \in \Omega_i} (A_{ij} - \boldsymbol{w}_i^T \boldsymbol{h}_j + w_{it} h_{jt}) h_{jt}}{\lambda + \sum_{j \in \Omega_i} h_{jt}^2}.$$

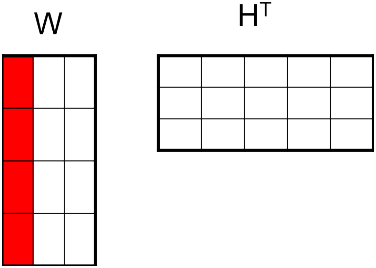
- Subproblem is just a univariate quadratic problem
- $\Omega_i = \{j : (i, j) \in \Omega\}$
- Can be done in $O(|\Omega_i|)$

Update Sequence:

- Item/user-wise update:
  - pick a user i or an item j
  - update the i-th row of W or the j-th column of H
- Feature-wise update:
  - pick a feature index $t \in \{1, \ldots, k\}$
  - update t-column of W and H alternatively

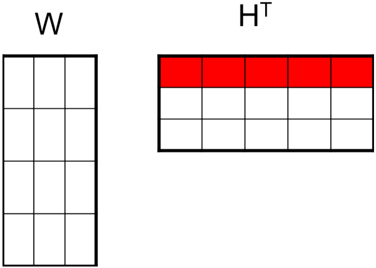
Feature-wise Update: CCD++

When $T = 2$



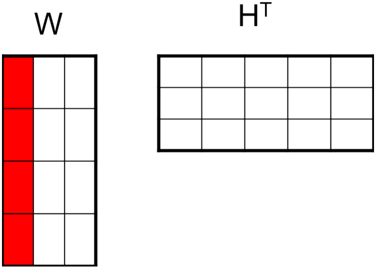
Feature-wise Update: CCD++

When $T = 2$



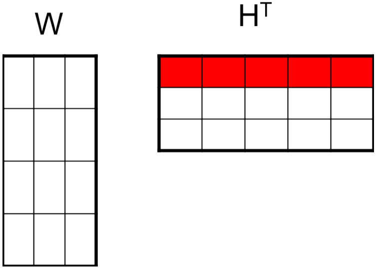
Feature-wise Update: CCD++

When $T = 2$



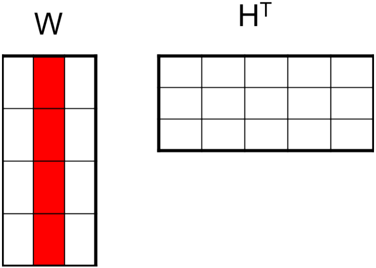
Feature-wise Update: CCD++

When $T = 2$



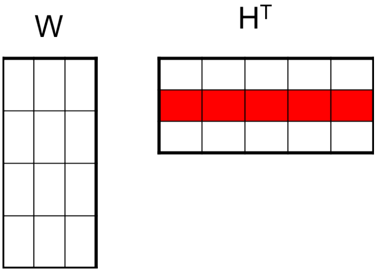
Feature-wise Update: CCD++

When $T = 2$



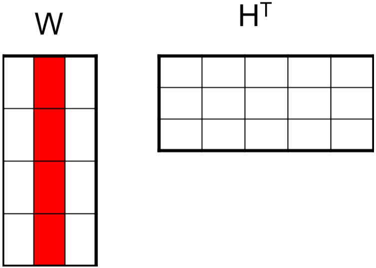
Feature-wise Update: CCD++

When $T = 2$



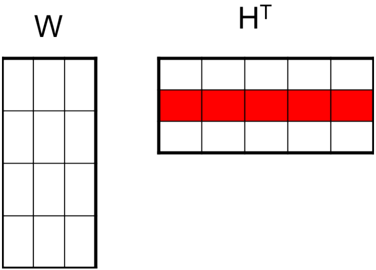
Feature-wise Update: CCD++

When $T = 2$



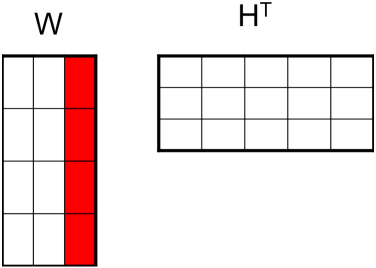
Feature-wise Update: CCD++

When $T = 2$



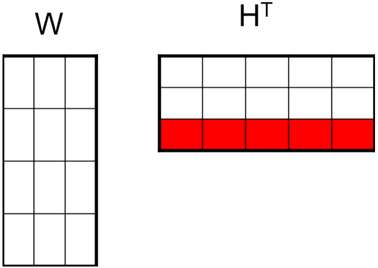
Feature-wise Update: CCD++

When $T = 2$



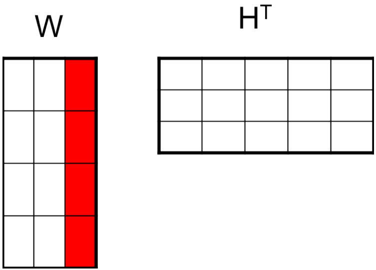
Feature-wise Update: CCD++

When $T = 2$



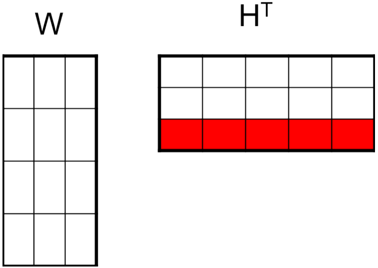
Feature-wise Update: CCD++

When $T = 2$



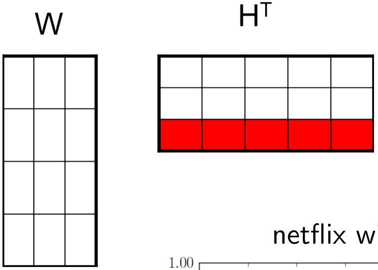
Feature-wise Update: CCD++

When $T = 2$

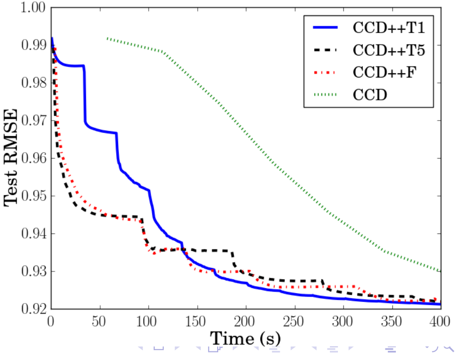


Feature-wise Update: CCD++

When $T = 2$



netflix with $k = 40$



- Cycle through k feature dimensions

Coming up

- Next class: other matrix completion topics

Questions?