Lecture 2 – Part 2
Research on the Netflix Problem (Mostly)

(January 16, 2015)

Mu Zhu
University of Waterloo

Content-boosted Matrix Factorization

- Zhu M (2014), "Making personalized recommendations in e-commerce", in *Statistics in Action: A Canadian Outlook*, J. F. Lawless, Ed., Chapman & Hall, pp. 259–268.

- Nguyen J, Zhu M (2013), "Content-boosted matrix factorization techniques for recommender systems", *Statistical Analysis and Data Mining* **6**, pp. 286–301.

- Forbes P, Zhu M (2011), "Content-boosted matrix factorization for recommender systems: Experiments with recipe recommendation", in *Proceedings of the 5th ACM Conference on Recommender Systems*, pp. 261–264.

Toy Example

	Moby-Dick (Melville)	*The Interpretation of Dreams* (Freud)	*The Origin of Species* (Darwin)	*Relativity* (Einstein)
Alice	90	~~70~~ ?	30	10
Bob	90	70	~~30~~ ?	10
Cathy	10	~~30~~ ?	70	90
David	10	30	~~70~~ ?	90

User-item ratings (matrix $\boldsymbol{R}$). A question mark (?) indicates that the corresponding rating is treated as if it were missing/unobserved and to be predicted.

Toy Example

	Moby-Dick	*Dreams*	*Species*	*Relativity*
Themes of conflict	$\surd$	$\surd$	$\surd$	$\times$
Elem. of moral phil.	$\surd$	$\surd$	$\times$	$\times$
Darkness of human nature	$\surd$	$\surd$	$\times$	$\times$
Other empirical evidence	$\times$	$\times$	$\surd$	$\surd$
Logical rigor	$\times$	$\times$	$\surd$	$\surd$
A grand new theory	$\times$	$\surd$	$\surd$	$\surd$

Content attributes (matrix $\boldsymbol{A}$). Sometimes, we may have content information for each item, such as these. Here, each $\boldsymbol{a}_{ij} \in \{0, 1\}$ is a binary indicator of whether item i has attribute j.

Boosting by Content

- intuition: content features may explain why some users prefer certain items to others
- toy example: any user's ratings of any two items are always $20 \times (4 - Z)$ points apart if the two items share Z content features in common ($Z = 0, 1, 2,$ or 3)
- reality: content information will never completely determine user preferences, but still more likely than not to be at least partially informative
- our work: two different ways to incorporate such content information into the matrix factorization approach

Two Ways to Boost by Content

- add extra penalty

$$\min_{\boldsymbol{p}_u, \boldsymbol{q}_i} \quad \text{(original objective)} - \lambda \left[\sum_{i=1}^{m} \sum_{i' \in \mathcal{S}_c(i)} \frac{\boldsymbol{q}_i^{\mathrm{T}} \boldsymbol{q}_{i'}}{|\mathcal{S}_c(i)|} \right]$$

  where

$$\mathcal{S}_c(i) \quad \equiv \quad \{i' < i : \quad \boldsymbol{a}_i^{\mathrm{T}} \boldsymbol{a}_{i'} \geq c\}$$

- enforce explicit constraint

$$\boldsymbol{q}_i = \boldsymbol{B}\boldsymbol{a}_i \quad \text{or} \quad \boldsymbol{q}_i(k) = \sum_j \boldsymbol{B}(k,j)\boldsymbol{a}_i(j)$$

(I) baseline MF; (II) extra penalty; (III) explicit constraint

		Moby-Dick	*Dreams*	*Species*	*Relativity*
(I)	Alice	89	**48**	30	10
	Bob	90	70	**51**	11
	Cathy	11	**52**	70	90
	David	10	30	**49**	89
(II)	Alice	89	**63**	30	10
	Bob	90	70	**37**	11
	Cathy	11	**37**	70	90
	David	10	30	**63**	89
(III)	Alice	90	**75**	30	10
	Bob	90	70	**25**	10
	Cathy	10	**25**	70	90
	David	10	30	**75**	90

using $K = 2$; $\lambda = 1$; $c = 3$

Expanded Alternating Optimization

Murdoch WJ, Zhu M (2014), "Expanded alternating optimization of nonconvex functions with applications to matrix factorization and penalized regression", arXiv:1412.4128.

Alternating Optimization (AO)

- coordinate descent $\in$ alternating optimization

$$\min_{\boldsymbol{z}\in\mathbb{R}^d} \quad f(\boldsymbol{z})$$

by partitioning $\boldsymbol{z}$ into blocks $\boldsymbol{z}_1, \boldsymbol{z}_2, ...$ and solving

$$\min_{\boldsymbol{z}_b} \quad f(\boldsymbol{z}_1, \boldsymbol{z}_2, ...)$$

successively over $b = 1, 2, ..., 1, 2, ...$

- attractive if each individual problem (over $\boldsymbol{z}_b$) is "easy"

Toy Example

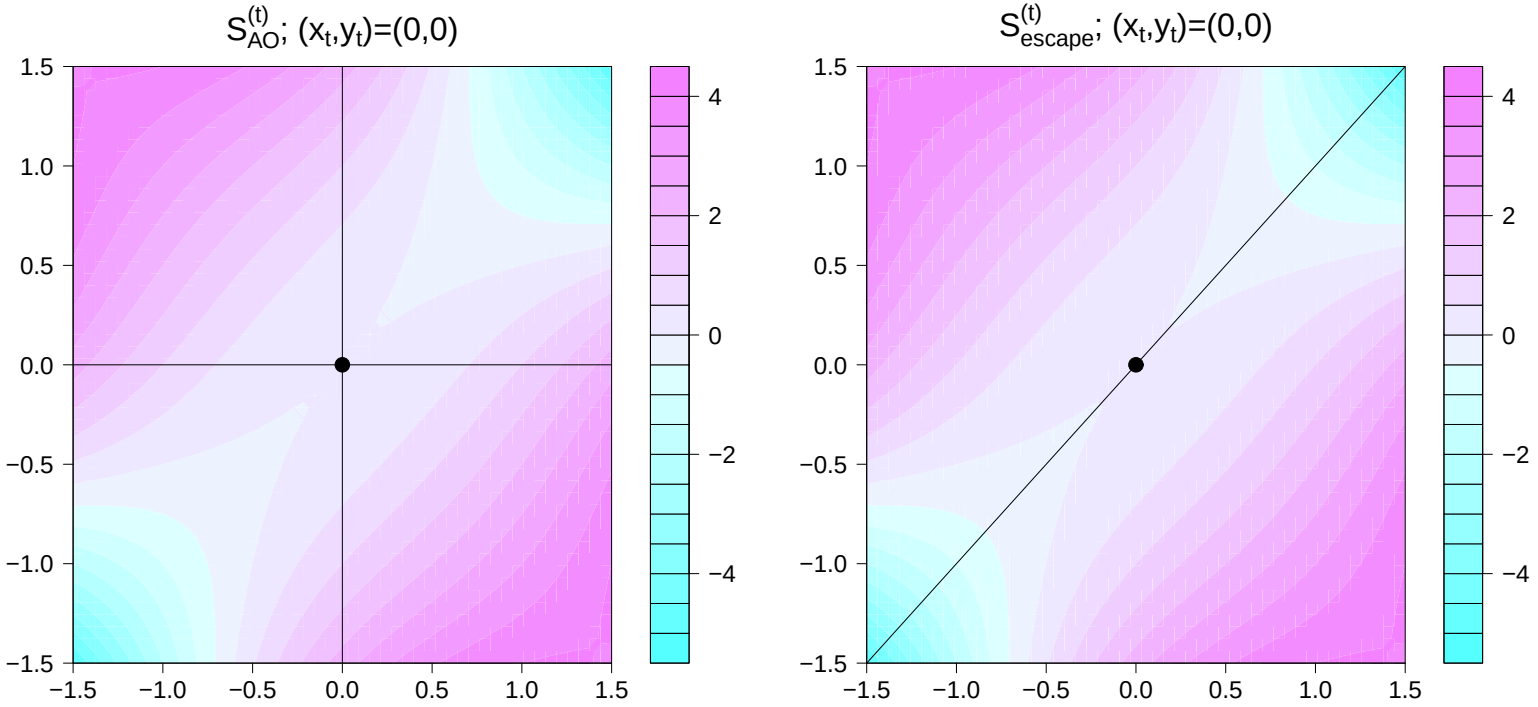
- minimize

$$f(x, y) = (x - y)^2 - x^2 y^2$$

  with coordinate descent

- suppose we reached $(x_t, y_t) = (0, 0)$ at iteration t
- fixing $x_t = 0$, $f(0, y) = y^2$ minimized at $y = 0$
- fixing $y_t = 0$, $f(x, 0) = x^2$ minimized at $x = 0$
- so, stuck at saddle point $(0, 0)$

Toy Example



Contours of $f(x,y) = (x-y)^2 - x^2y^2$

Toy Example

$$\mathcal{S}_{AO}^{(t)} = \{(x,y) : x = 0\} \cup \{(x,y) : y = 0\}$$

$$\mathcal{S}_{escape}^{(t)} = \{(x,y) : x = y\}$$

Scaling

A. Tayal, T. F. Coleman & Y. Li (2014), "Primal explicit max margin feature selection for nonlinear support vector machines", *Pattern Recognition* **47**, pp. 2153–2164.

- at each step,

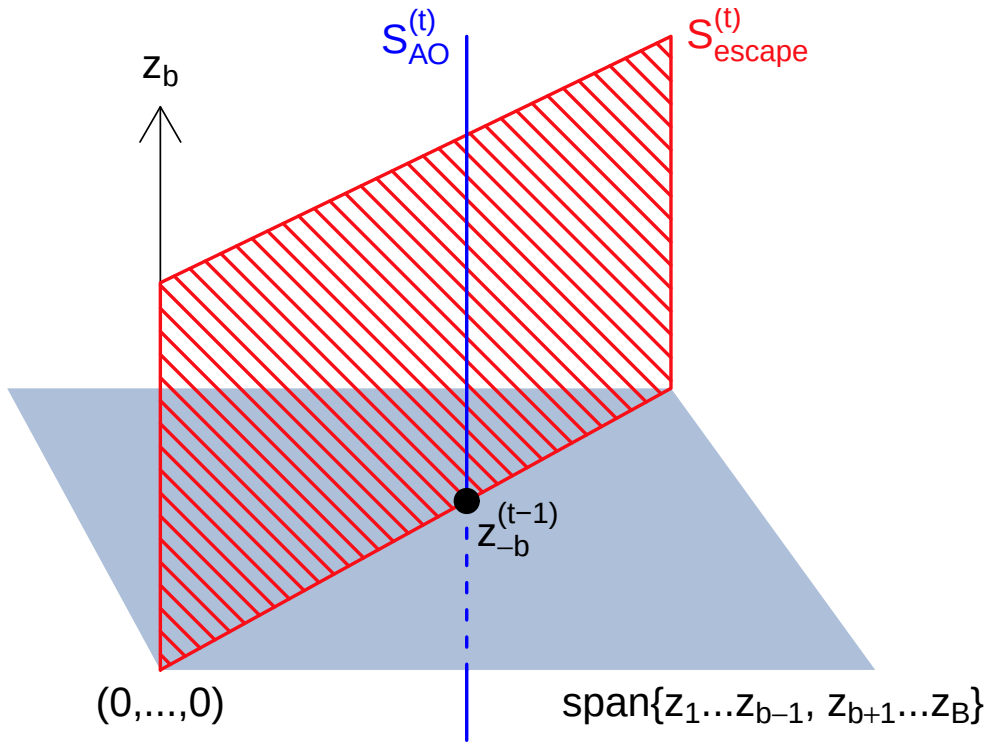
$$\min_{\boldsymbol{z}_b, v_b} \quad f\left(v_b \boldsymbol{z}_1, \quad \ldots, \quad v_b \boldsymbol{z}_{b-1}, \quad \boldsymbol{z}_b, \quad v_b \boldsymbol{z}_{b+1}, \quad \ldots, \quad v_b \boldsymbol{z}_B\right)$$

- let v_b^* denote the solution of v_b; all other components $\boldsymbol{z}_{-b}$ adjusted accordingly, i.e.,

$$\boldsymbol{z}_{-b} \quad \longleftarrow \quad v_b^* \boldsymbol{z}_{-b},$$

  before the next step

Scaling



$\underline{\mathcal{S}^{(t)}_{escape} \text{ larger than } \mathcal{S}^{(t)}_{AO}, \text{ but still not whole space}}$

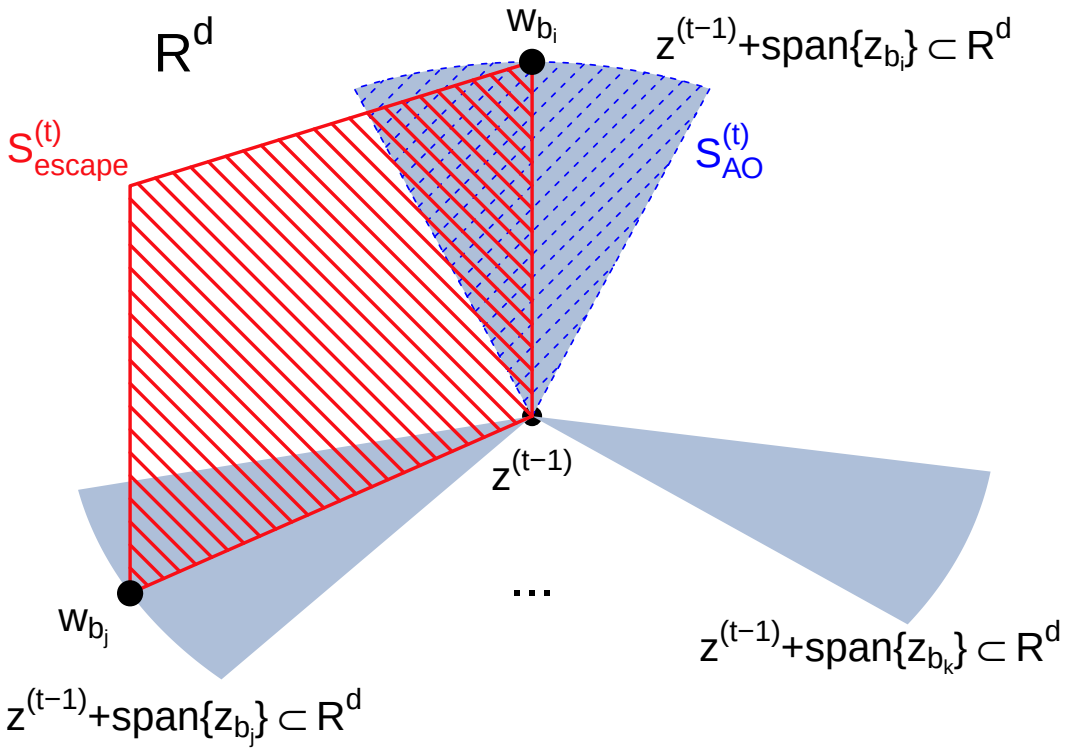
Restricted Joint Search (RJS)

$$\min_{\alpha_1,\ldots,\alpha_B} \quad f\left(\boldsymbol{z}_1 + \alpha_1 \boldsymbol{w}_1 I_1, \quad \boldsymbol{z}_2 + \alpha_2 \boldsymbol{w}_2 I_2, \quad \ldots, \quad \boldsymbol{z}_B + \alpha_B \boldsymbol{w}_B I_B\right)$$

$$\boldsymbol{w}_b \in \text{span}\{\boldsymbol{z}_b\}$$

$$I_b = \begin{cases} 1, & \text{if component } b \text{ chosen to "participate";} \\ 0, & \text{otherwise} \end{cases}$$

Restricted Joint Search (RJS)



$\underline{\mathcal{S}^{(t)}_{escape} \text{ never search by AO}}$

RJS for Matrix Factorization

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^n,\ \boldsymbol{\beta}\in\mathbb{R}^m} \quad \sum_{u,i\in\mathbb{T}} \left[r_{ui} - \left(\boldsymbol{p}_u + \alpha_u \boldsymbol{w}_p^u\right)^{\mathrm{T}} \left(\boldsymbol{q}_i + \beta_i \boldsymbol{w}_q^i\right)\right]^2 +$$

$$\lambda \left[\sum_u \|\boldsymbol{p}_u + \alpha_u \boldsymbol{w}_p^u\|^2 + \sum_i \|\boldsymbol{q}_i + \beta_i \boldsymbol{w}_q^i\|^2\right]$$

- search directions $\boldsymbol{w}_p^u, \boldsymbol{w}_q^i$ mostly zero
- nonzero ones selected ...
  - either randomly, e.g., $\boldsymbol{w}_p^u, \boldsymbol{w}_q^i \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I})$
  - or greedily (next slide)

Greedy Choices of $\boldsymbol{w}_p^u, \boldsymbol{w}_q^i$

- for $\boldsymbol{p}_u$ alone (fixing others), consider solving for α given $\boldsymbol{w}$:

$$\min_{\alpha} \quad L(\alpha) = \sum_{i \in \mathbb{T}_u} \left[r_{ui} - (\boldsymbol{p}_u + \alpha \boldsymbol{w})^{\mathrm{T}} \boldsymbol{q}_i \right]^2 + \lambda \| \boldsymbol{p}_u + \alpha \boldsymbol{w} \|^2$$

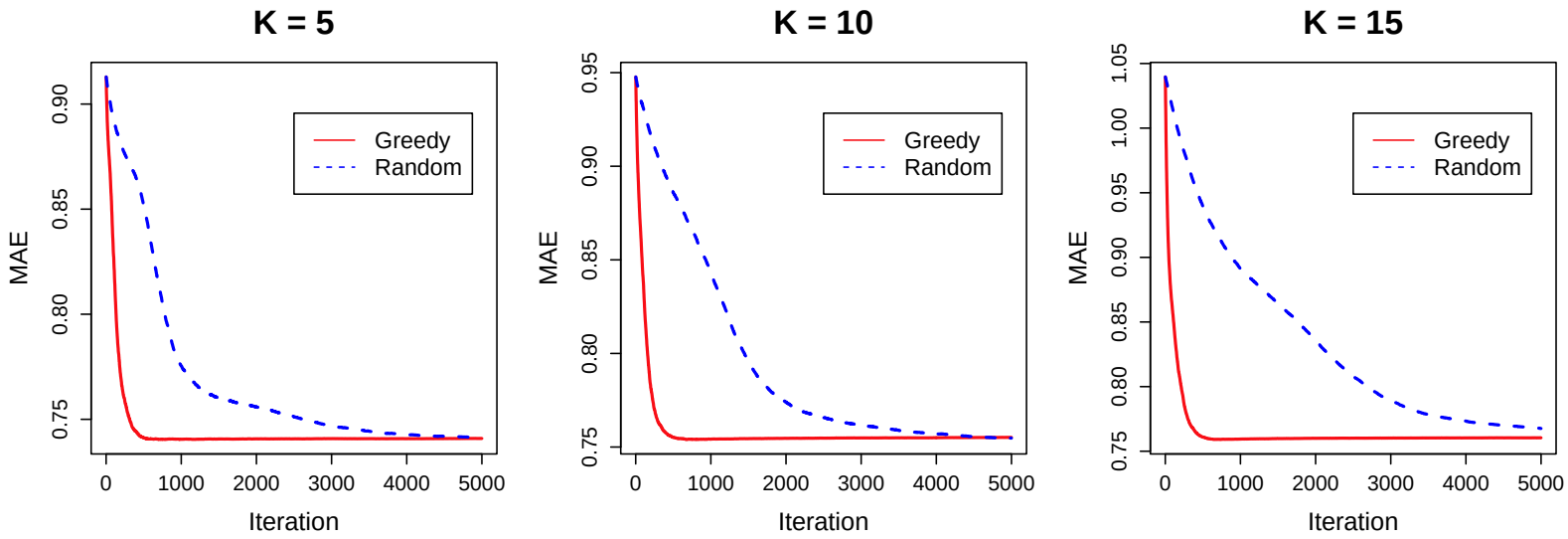
  where $\mathbb{T}_u = \{ i : r_{ui} \text{ is known} \}$

- optimal α must satisfy

$$\widehat{\alpha}(\boldsymbol{w}) = \frac{\sum_{i \in \mathbb{T}_u} \boldsymbol{w}^{\mathrm{T}} \boldsymbol{q}_i (r_{ui} - \boldsymbol{p}_u^{\mathrm{T}} \boldsymbol{q}_i) - \lambda \boldsymbol{w}^{\mathrm{T}} \boldsymbol{p}_u}{\sum_{i \in \mathbb{T}_u} (\boldsymbol{w}^{\mathrm{T}} \boldsymbol{q}_i)^2 + \lambda \| \boldsymbol{w} \|^2}$$

- can plug $\widehat{\alpha}(\boldsymbol{w})$ back into $L(\alpha)$ above and solve for "best" $\boldsymbol{w}$, e.g., with quasi-Newton

Example: Amazon Data



Mean absolute error (MAE) on the hold-out test set (average of 10 runs). Iterations start at convergence of baseline AO. Data source: McAuley & Leskovec (*RecSys* 2013). Original data contained $\sim$ 35.3M reviews from `amazon.com` (1995–2013). We took a dense subset consisting of users having rated ≥ 55 items & items with ≥ 24 ratings ($\sim$ 5.5M reviews).

Selective Scaling: Regression w/ MCP

- for given j, let

$$E_j \quad = \quad \{k \neq j \text{ such that } |\text{corr}(\boldsymbol{x}_k, \boldsymbol{x}_j)| > \rho_{min}\}$$

- solve

$$\min_{\beta_j, v} \quad \left\| \boldsymbol{y} - \beta_j \boldsymbol{x}_j - \sum_{\ell \in E_j^C \setminus \{j\}} \beta_\ell \boldsymbol{x}_\ell - \sum_{k \in E_j} (v\beta_k) \boldsymbol{x}_k \right\|^2 +$$

$$J(\beta_j) + \sum_{\ell \in E_j^C \setminus \{j\}} J(\beta_\ell) + \sum_{k \in E_j} J(v\beta_k)$$

- can solve explicitly — first order condition piecewise linear in v on $< \infty$ intervals, so just check if each interval crosses 0

Summary

- main message:
  - optimization of nonconvex functions is hard
- specific content:
  - content-boosted matrix factorization
  - expanded alternating optimization