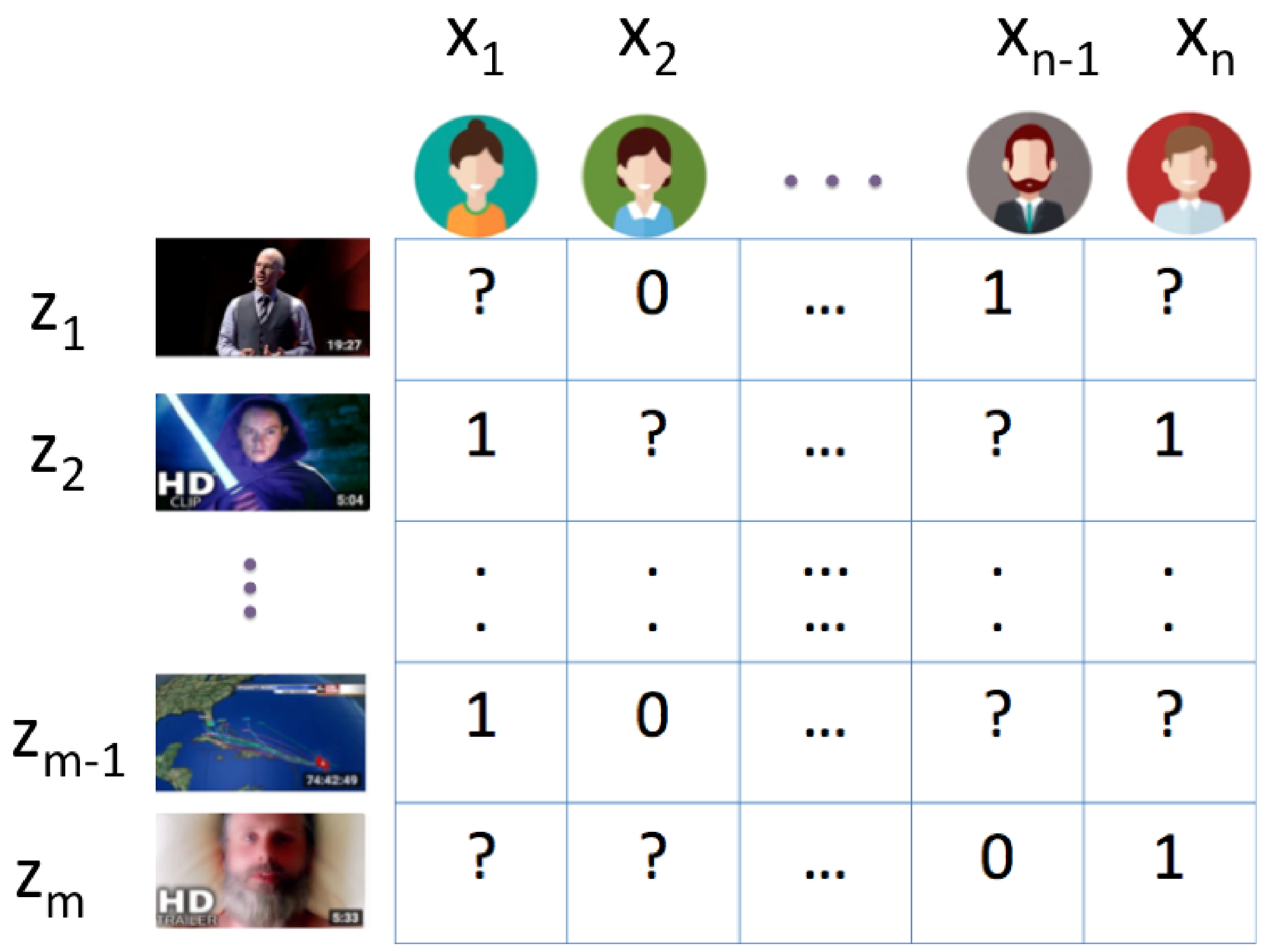
Provable Non-linear Inductive Matrix Completion

Kai Zhong[1], Zhao Song[2], Prateek Jain[3] and Inderjit S. Dhillon[4]

[1]Amazon [2]University of Washington [3]Microsoft [4]Amazon & University of Texas at Austin

Recommendation System with Side Information

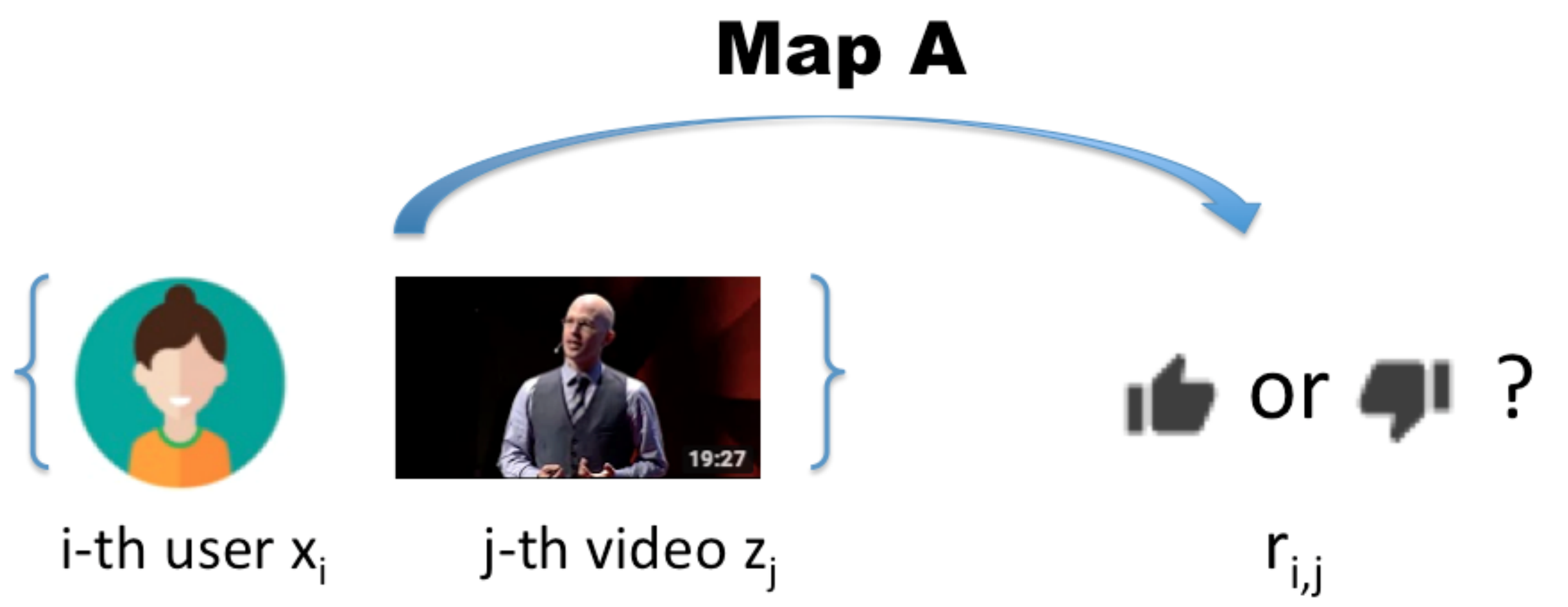
- Given *some observed* ratings (e.g. likes/dislikes) that users gave videos, the goal is to predict unobserved ratings.
- Additional side information:
  1. The i-th user has features $\boldsymbol{x}_i \in \mathbb{R}^{d_1}$: profile, subscriptions, etc..
  2. The j-th video has features $\boldsymbol{z}_j \in \mathbb{R}^{d_2}$: video features, description, etc...



	x_1	x_2	$\cdots$	x_{n-1}	x_n
z_1	?	0	...	1	?
z_2	1	?	...	?	1
$\vdots$	:	:	⋮	:	:
z_{m-1}	1	0	...	?	?
z_m	?	?	...	0	1

Learning Task

- Learning Task: Learn the mapping A from $(\boldsymbol{x}_i, \boldsymbol{z}_j)$ to $r_{i,j}$, where $r_{i,j}$ is the rating that the i-th user gives the j-th video,
$$r_{i,j} \approx A(\boldsymbol{x}_i, \boldsymbol{z}_j).$$



- Inductive: with side information, it is possible to predict ratings for new users/movies and provide personalized recommendation.

Inductive Matrix Completion

- Linear IMC: Ratings are in bi-linear relationship with $\boldsymbol{x}_i$, $\boldsymbol{z}_j$, i.e.,
$$r_{i,j} = \langle U^{*\top}\boldsymbol{x}_i, V^{*\top}\boldsymbol{z}_j\rangle,$$
where $U^* \in \mathbb{R}^{d_1\times k}$, $V^* \in \mathbb{R}^{d_2\times k}$ are the ground truth linear mappings.

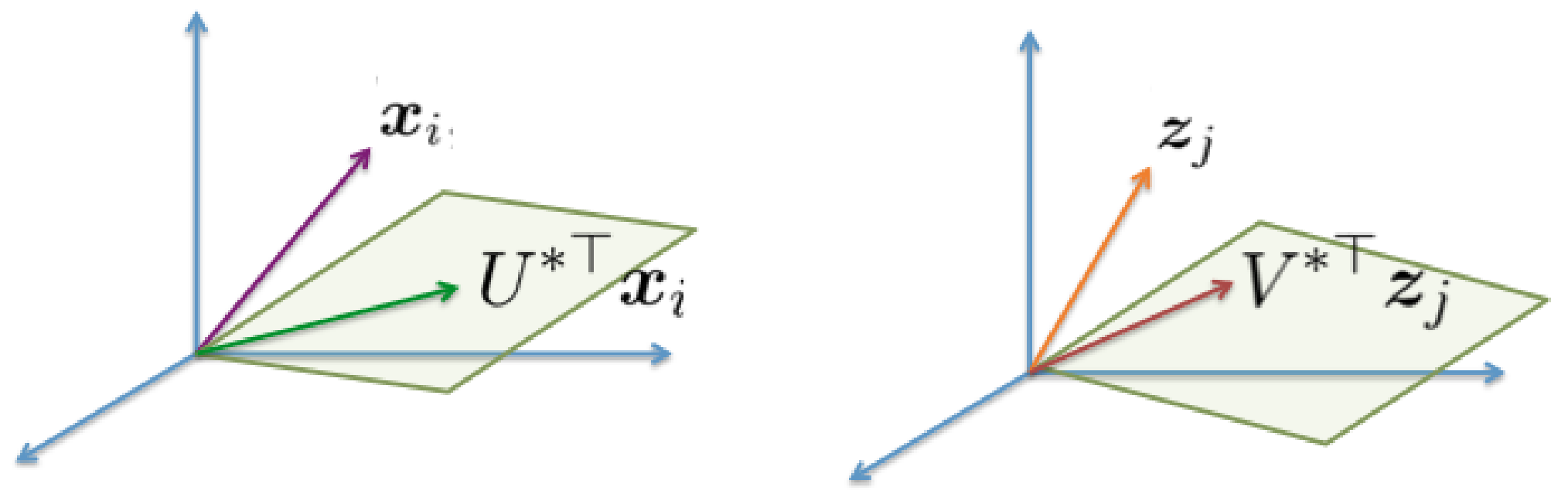


Figure: Linear low-dimensional embedding

- Cons: The model capacity is very limited.

Nonlinear Inductive Matrix Completion Based on One-layer NNs

- A nonlinear extension: (ϕ is a non-linear activation function)
$$r_{i,j} = \langle \phi(U^{*\top}\boldsymbol{x}_i), \phi(V^{*\top}\boldsymbol{z}_j)\rangle.$$

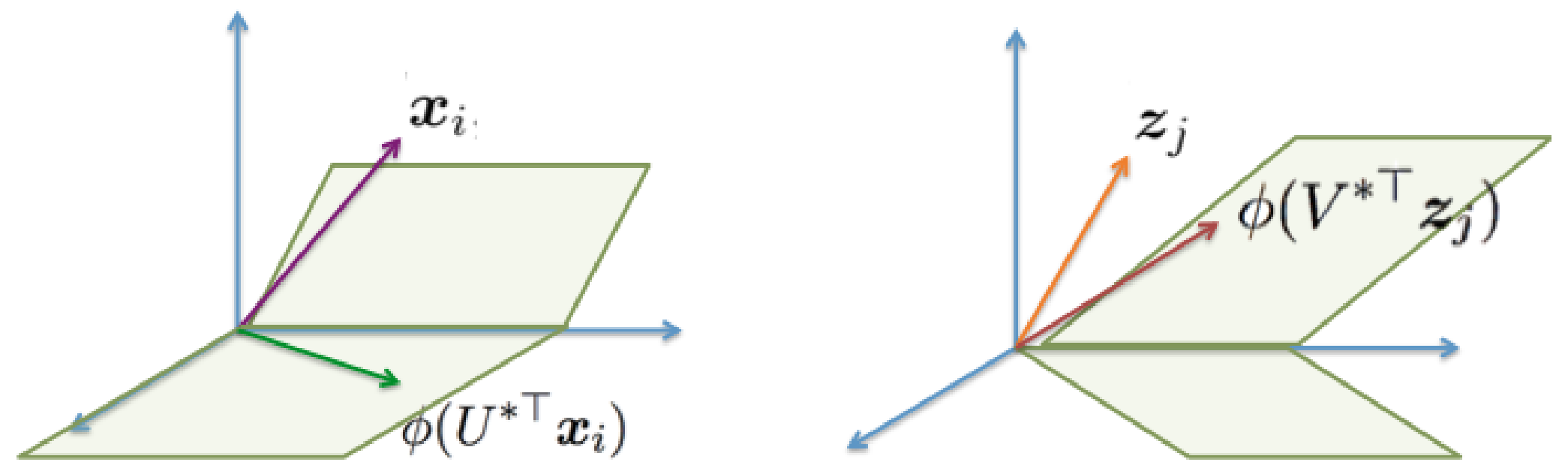
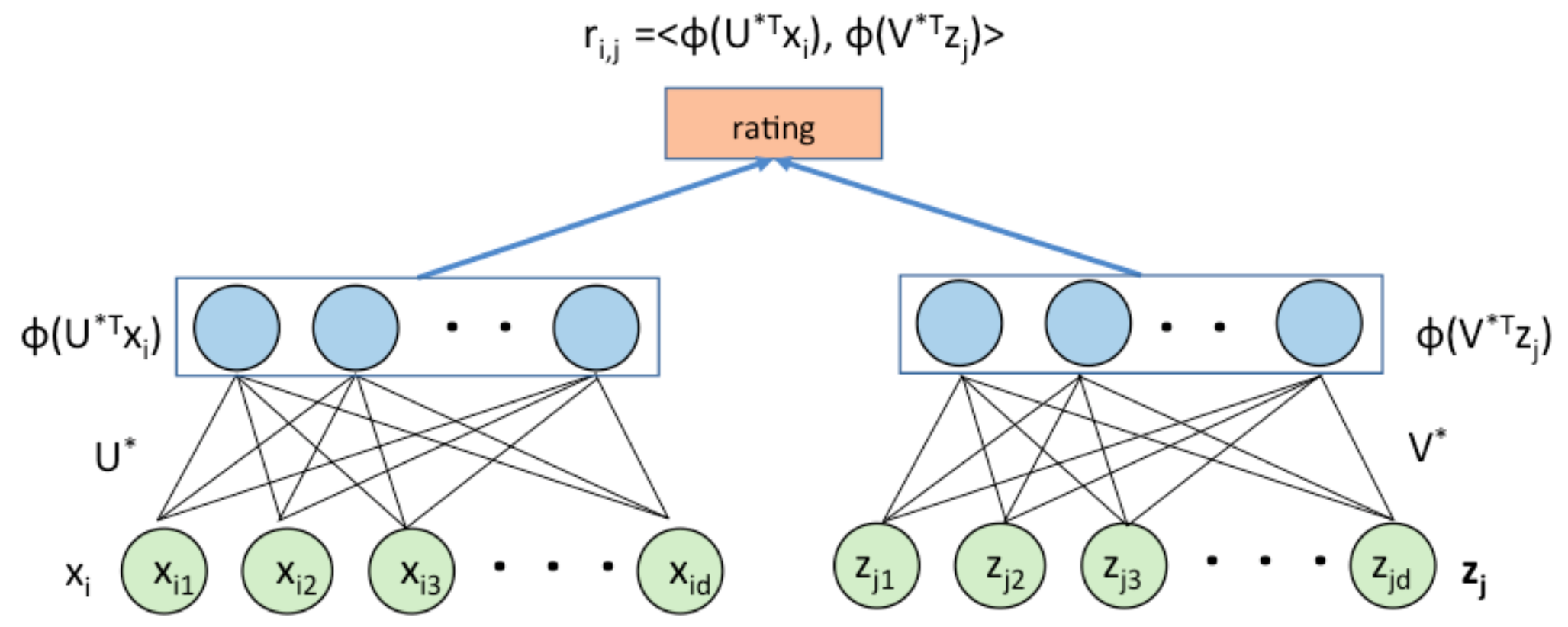

$\boldsymbol{x}_i$
$\phi(U^{*\top}\boldsymbol{x}_i)$
$\boldsymbol{z}_j$
$\phi(V^{*\top}\boldsymbol{z}_j)$

Figure: Non-linear embedding



- Optimization problem based on squared loss,
$$\min_{U\in\mathbb{R}^{d\times k}, V\in\mathbb{R}^{d\times k}} f_\Omega(U,V) = \frac{1}{2|\Omega|}\sum_{(i,j)\in\Omega}\left(\langle\phi(U^\top\boldsymbol{x}_i), \phi(V^\top\boldsymbol{z}_j)\rangle - r_{i,j}\right)^2,$$
where $\Omega \subset [n]\times[m]$ is the index set of observed user-video ratings.

Main Theoretical Challenges

1. Two sources of non-convexity
   a. non-linear activation function
   b. the inner product between two mapped features
2. Sampling error: the observed set Ω is a subset of $[m]\times[n]$.

Assumption and Definition

- Assumptions on data distribution:
$$\boldsymbol{x}_i \sim \mathcal{N}(0, I), \forall i \in [n],\ \text{i.i.d.}$$
$$\boldsymbol{z}_j \sim \mathcal{N}(0, I), \forall j \in [m],\ \text{i.i.d.}$$
$$\Omega = \{(i_l, j_l) \sim \text{uniform}([n]\times[m]),\ \forall l = 1, \cdots, |\Omega|,\ \text{i.i.d.}\}$$
- Assume the observations are from a planted model, i.e., there exists ground truth (U^*, V^*), such that
$$r_{i,j} = \langle U^{*\top}\boldsymbol{x}_i, V^{*\top}\boldsymbol{z}_j\rangle,$$
- Define $\lambda := \max\{\lambda(U^*), \lambda(V^*)\}$ and $\kappa := \max\{\kappa(U^*), \kappa(V^*)\}$, where $\lambda(U) = \sigma_1^k(U)/(\Pi_{i=1}^k\sigma_i(U))$, $\kappa(U) = \sigma_1(U)/\sigma_k(U)$, and $\sigma_i(U)$ denotes the i-th singular value of U with the ordering $\sigma_i \geq \sigma_{i+1}$.
- Define $d = \max\{d_1, d_2\}$.

Local Strong Convexity for Sigmoid and Tanh

Let the activation function ϕ be sigmoid or tanh. Then for any $t > 1$ and any given U, V, if
$$n \gtrsim t\lambda^4\kappa^2 d\log^2 d,$$
$$m \gtrsim t\lambda^4\kappa^2 d\log^2 d,$$
$$|\Omega| \gtrsim t\lambda^4\kappa^2 d\log^2 d,$$
$$\|U - U^*\| + \|V - V^*\| \lesssim 1/(\lambda^2\kappa),$$
then with probability at least $1 - d^{-t}$, the smallest eigenvalue of the Hessian of the objective $f_\Omega(U,V)$ is lower bounded by:
$$\lambda_{\min}(\nabla^2 f_\Omega(U,V)) \gtrsim 1/(\lambda^2\kappa).$$

Linear Convergence of Gradient Descent

Let $[U^c, V^c]$ be the parameters in the c-th iteration. Assuming $\|U^c - U^*\| + \|V^c - V^*\| \lesssim 1/(\lambda^2\kappa)$, then given a fresh sample set, Ω, that is independent of $[U^c, V^c]$, the next iterate using one step of gradient descent, i.e., $[U^{c+1}, V^{c+1}] = [U^c, V^c] - \eta\nabla f_\Omega(U^c, V^c)$, satisfies
$$\|U^{c+1} - U^*\|_F^2 + \|V^{c+1} - V^*\|_F^2 \leq (1 - M_l/M_u)(\|U^c - U^*\|_F^2 + \|V^c - V^*\|_F^2)$$
with probability $1 - d^{-t}$, where $\eta = \Theta(1/M_u)$ is the step size and $M_l \gtrsim 1/(\lambda^2\kappa)$ is the lower bound on the eigenvalues of the Hessian and $M_u \lesssim 1$ is the upper bound on the eigenvalues of the Hessian.

Main Theory for Overall Algorithm: Gradient Descent with Initialization by Tensor Method

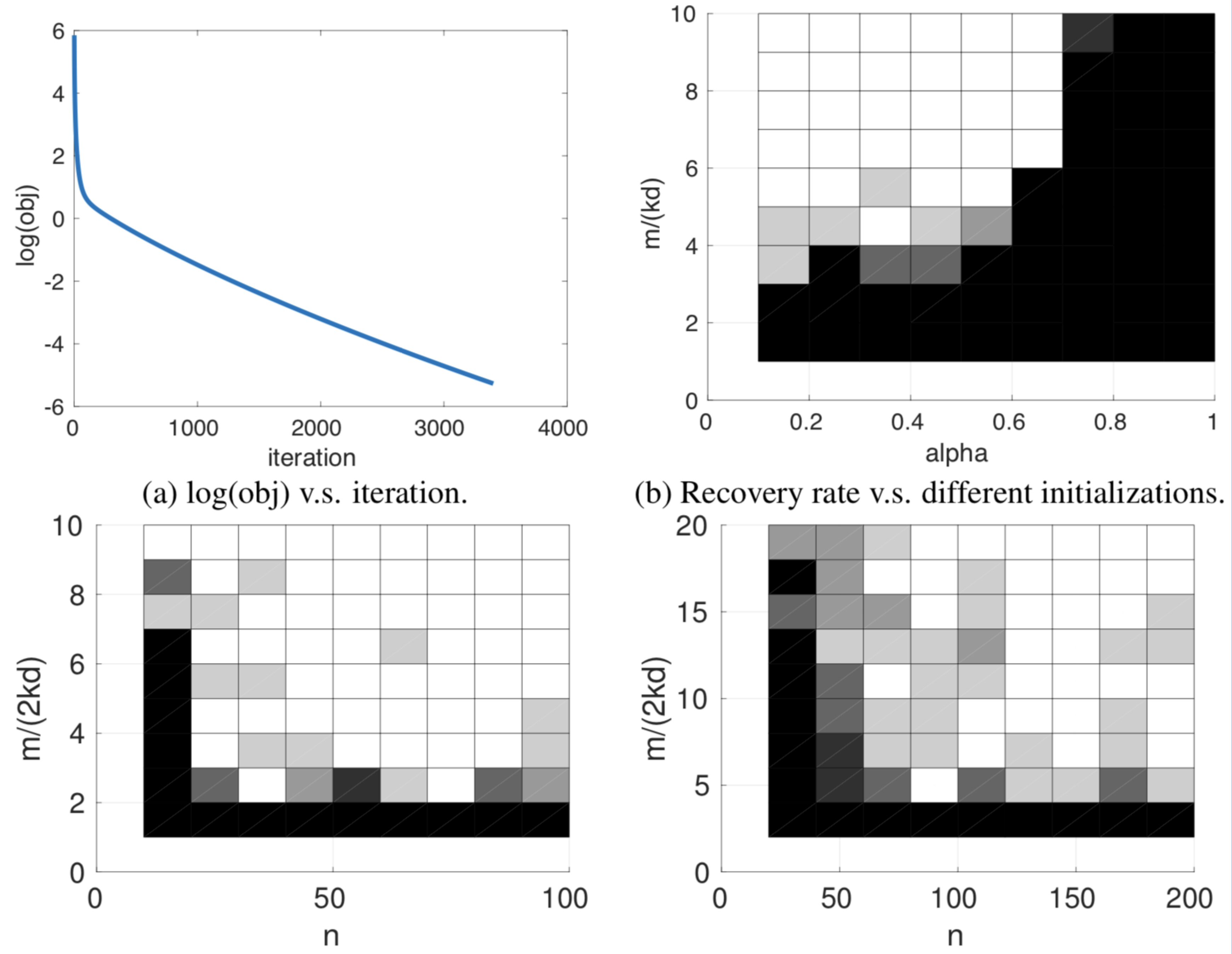
When $\phi(z)$ is sigmoid, applying gradient descent with tensor initialization can recover U^*, V^* w.h.p. to any precision $\epsilon > 0$ with both time complexity and sample complexity (refers to m, n and $|\Omega|$) $\text{poly}(d, \log(1/\epsilon), \kappa(U^*), \kappa(V^*))$,

Experimental Results: Vanilla (Linear) IMC v.s. Nonlinear IMC

Movie recommendation for new users on Movielens dataset

Dataset	#movies	#users	# ratings	# movie feat.	# user feat.	RMSE NIMC	RMSE IMC
ml-100k	1682	943	100,000	39	29	**1.034**	1.321
ml-1m	3883	6040	1,000,000	38	29	**1.021**	1.320

Simulation Results: Convergence Behavior & Sample Complexity



(a) log(obj) v.s. iteration. (b) Recovery rate v.s. different initializations.

(c) Recovery Rate v.s. (m, n) for Sigmoid (d) Recovery Rate v.s. (m, n) for ReLU

(a) shows the linear convergence of gradient descent. In (b,c,d), white blocks denote 100% recovery rate over 5 trials, while black means 100% failure. (b) shows the quality of the initialization, where smaller alpha means better initialization, affects the convergence to the ground truth. (c) and (d) show the recovery rate for sigmoid and ReLU activation functions respectively.

Mail: hsiangfu@, kaizhong@, isd@

Presented in AMLC 2019