Machine Learning

Lecture 16: Matrix Factorization

杨思蓓
SIST
Email: yangsb@shanghaitech.edu.cn

What Is Matrix Factorization?

$$X \in \mathcal{R}^{m \times n}$$

$$U \in \mathcal{R}^{m \times k}, \qquad V \in \mathcal{R}^{k \times n}$$

$$UV = X$$

1. $\Sigma \in \mathcal{R}^{k \times k}$, $\quad U\Sigma\Sigma^{-1}V = X$

$$U\Sigma V = X$$

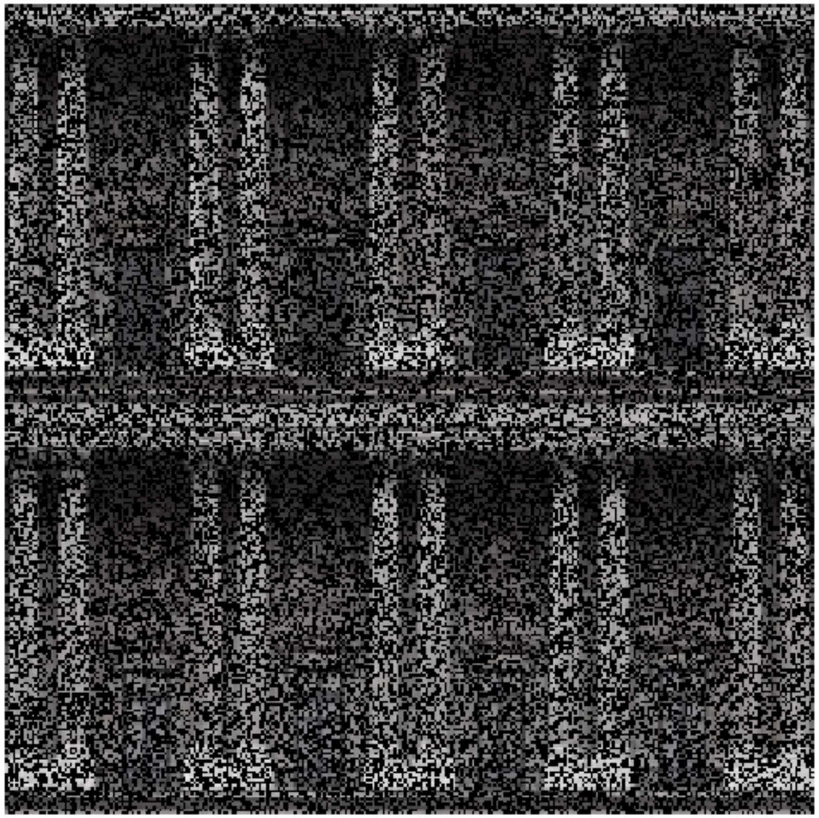
2. $UV = \tilde{X} \approx X$

$$\|X - UV\|_F^2$$

Why Matrix Factorization?

Application 1

Image Recovery



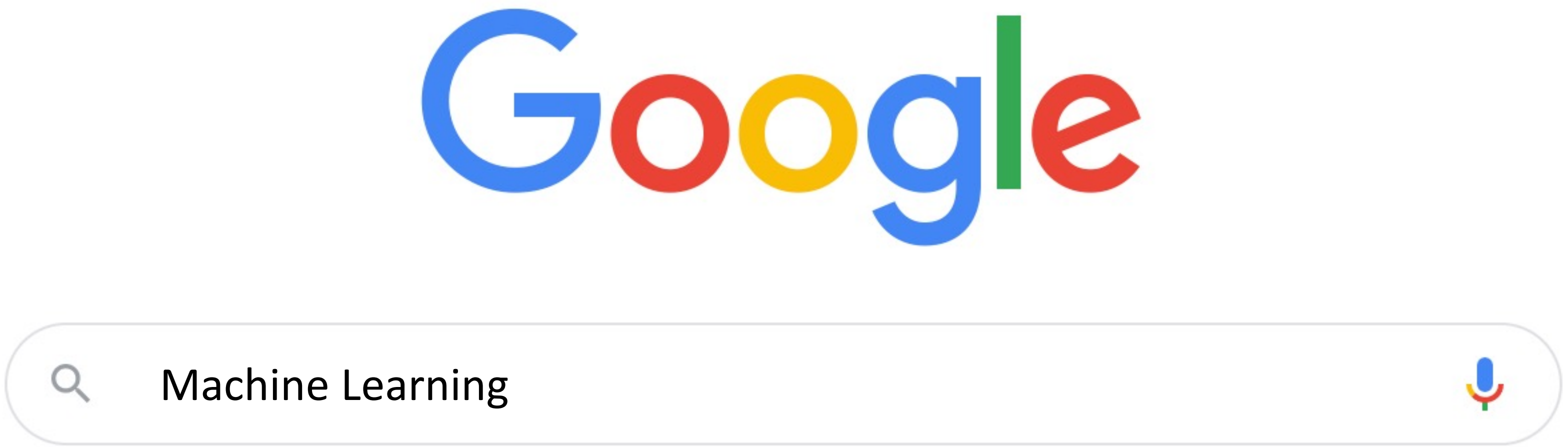
Application 2

Recommendation

	The Matrix	Star Wars	Roman Holiday	Titanic	Shrek	Madagascar
Alice	5	4	5	?	?	2
Bob	?	4	?	5	1	2
Tracy	5	5	4	5	2	?
Steven	4	?	?	5	?	2
John	4	5	2	?	2	?

Application 3

Search: Information Retrieval



Language Model Paradigm in IR

- Probabilistic relevance model
  - Random variables

$$R_d \in \{0,1\} \quad : \quad \text{relevance of document } d$$

$$q \subseteq \Sigma \quad : \quad \text{query, set of words}$$

  - Bayes' rule

probability of generating a query q to ask for relevant d

prior probability of relevance for document d (e.g. quality, popularity)

$$P(R_d = 1|q) = \frac{P(q|R_d = 1) \cdot P(R_d = 1)}{P(q)}$$

probability that document d is relevant for query q

J. Ponte and W.B. Croft, A Language Model Approach to Information Retrieval, ACM SIGIR, 1998.

Language Model Paradigm

$$P(R_d = 1|q) \propto \underbrace{P(q|R_d = 1)}_{2}\ \underbrace{P(R_d = 1)}_{1}$$

1

- First contribution: **prior probability of relevance**
  - simplest case: uniform (drops out for ranking)
  - **popularity**: document usage statistics (e.g. library circulation records, download or access statistics, hyperlink structure)

- Second contribution: **query likelihood**
  - query terms *q* are treated as a **sample** drawn from an (unknown) relevant document

2

Query Likelihood

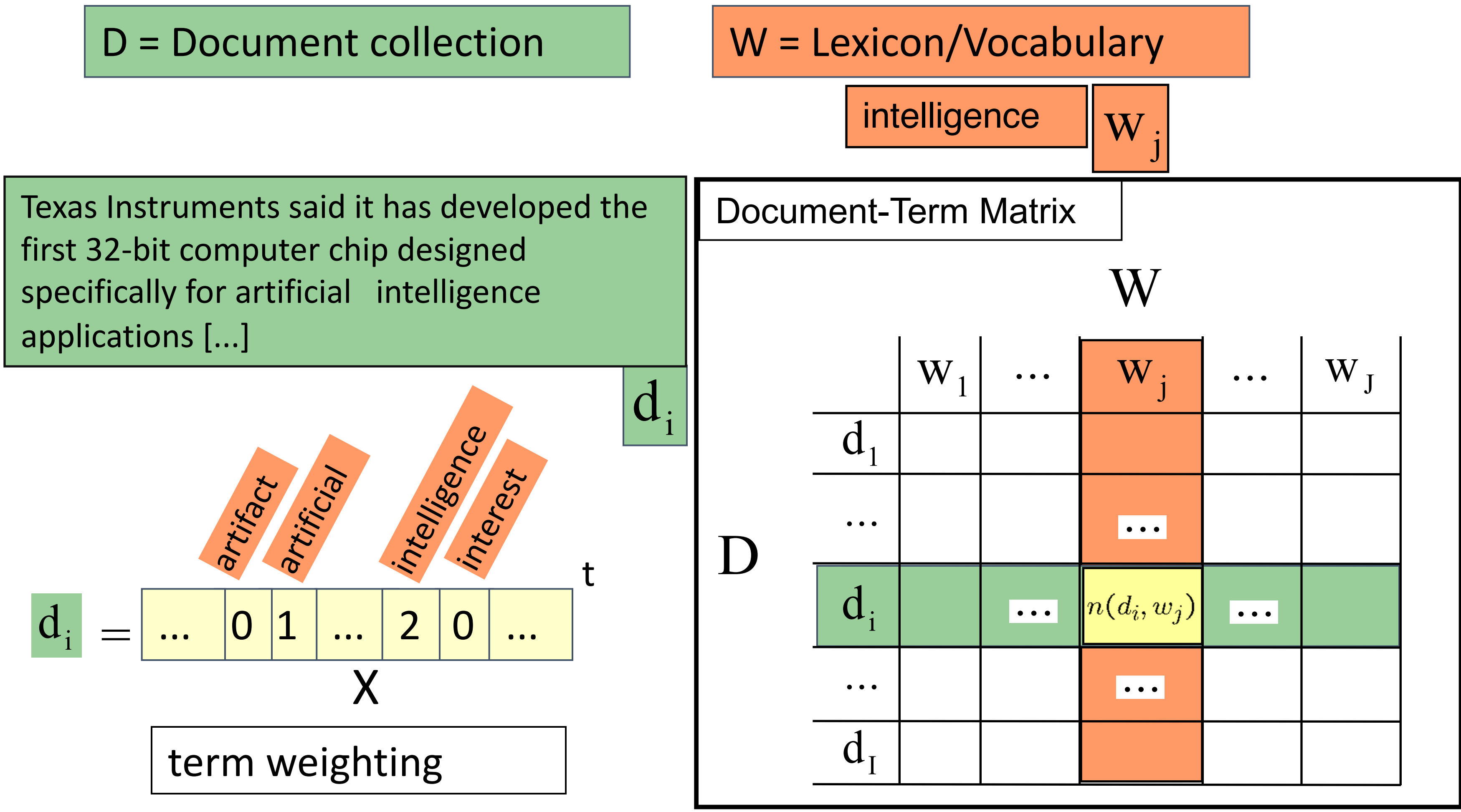
$$P(q|R_d = 1) \equiv P(q|d)$$

- $q = (w_1, \cdots, w_q)$
- Independent Assumption

$$P(q|d) = \Pi_{w \in q} P(w|d)$$

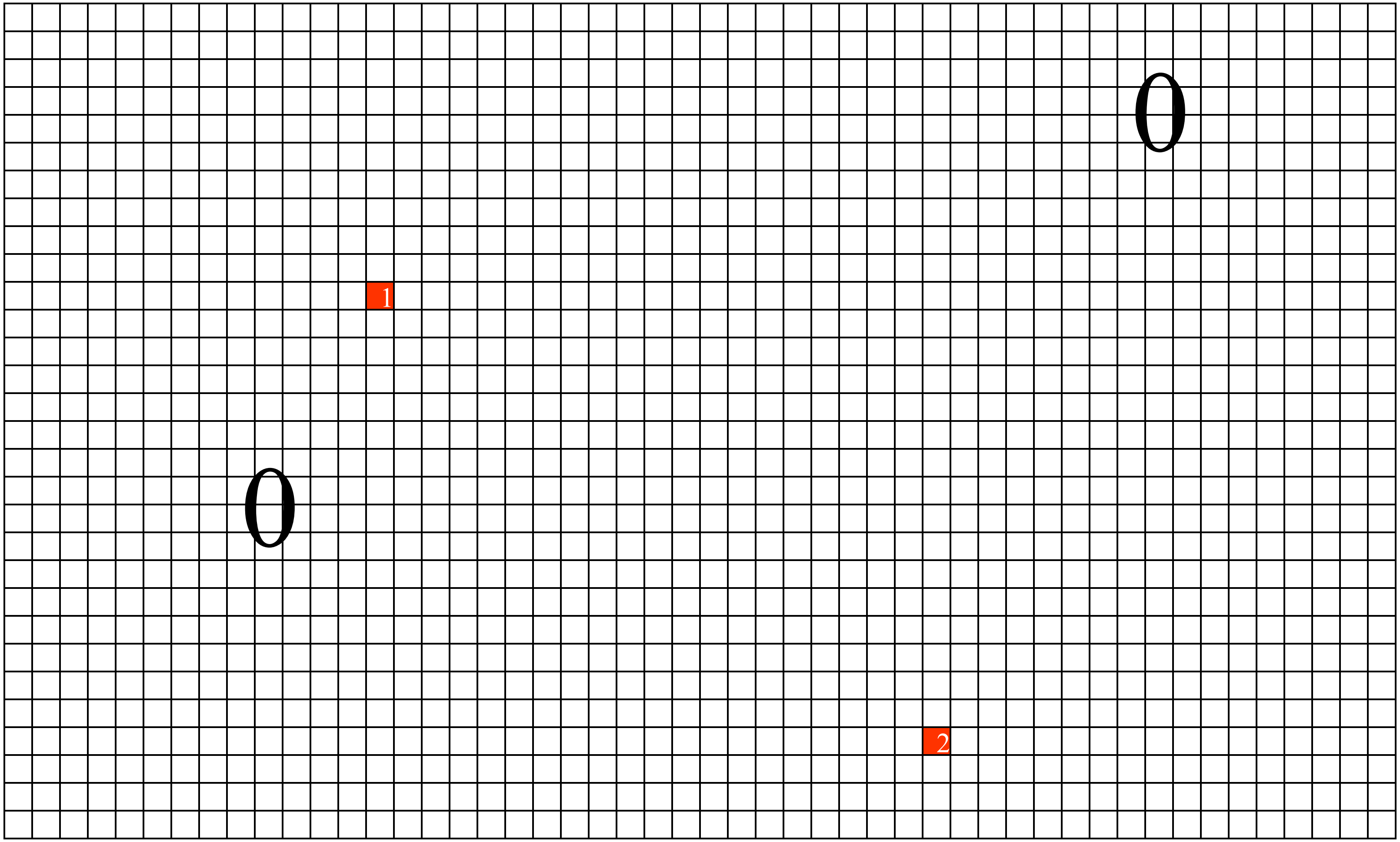
$$P(w|d)?$$

Document-Term Matrix



Document-Term Matrix

$\boldsymbol{X} =$



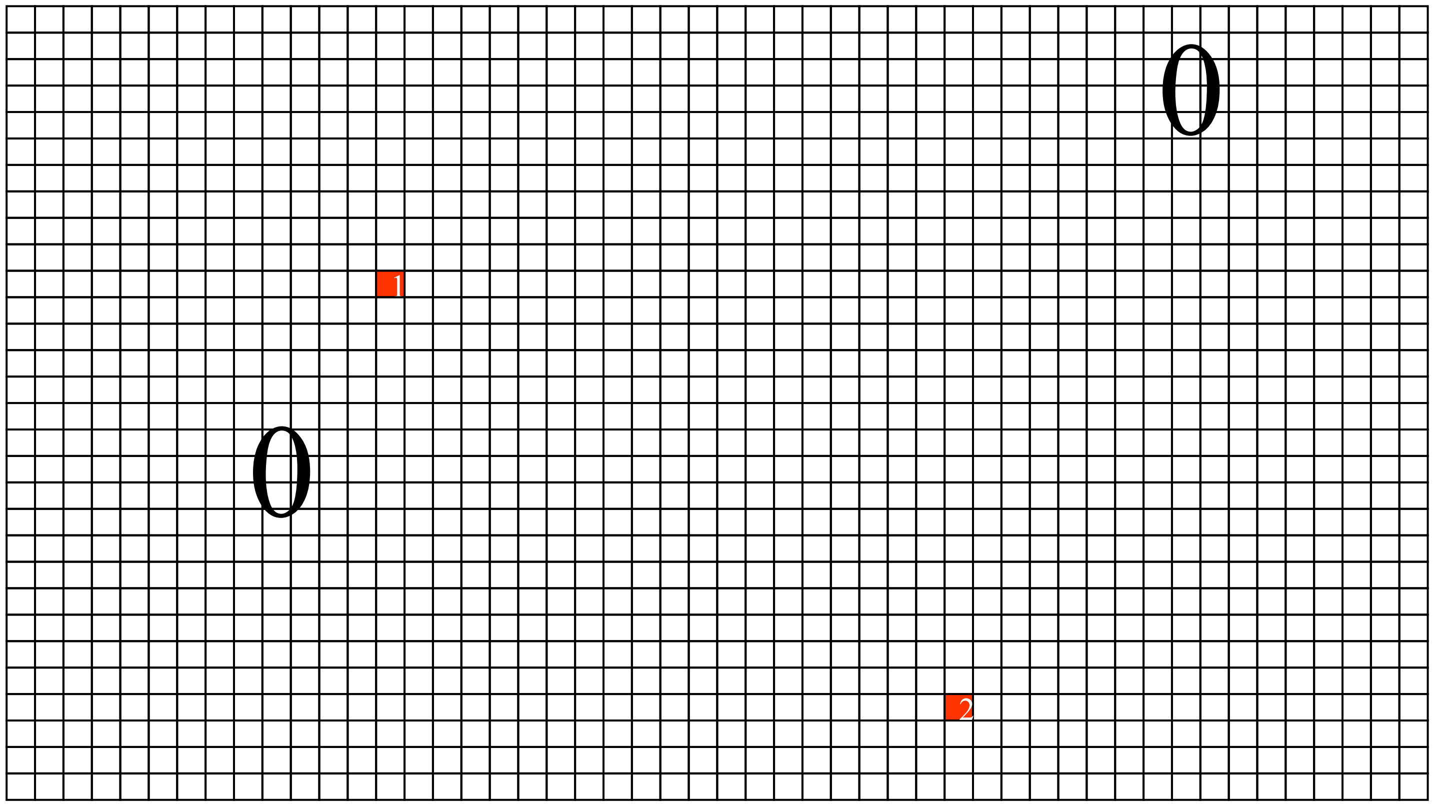
All The Three Applications

- Image Recovery
- Search
- Recommendation

- The Same Problem!

Incomplete Matrix

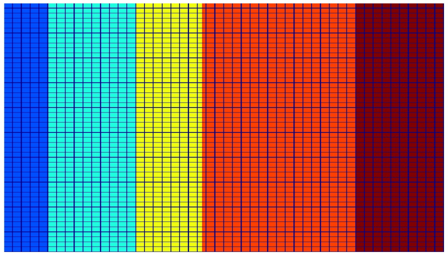
$X =$



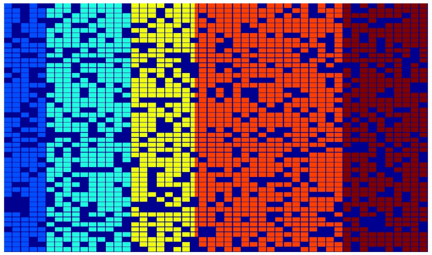
- Estimate the missing value
- Matrix completion

Matrix Completion

- If no assumption,
  - Mission impossible
- A reasonable assumption:
  - The matrix is of low rank



Low Rank Matrix



Incomplete Matrix

Why Matrix Factorization?

Matrix Factorization

$$X \in \mathcal{R}^{m \times n}$$

$$U \in \mathcal{R}^{m \times k}, \qquad V \in \mathcal{R}^{k \times n}$$

$$UV = \tilde{X} \approx X$$

- Low Rank Assumption
- k Hidden Factors

$$X = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots \boldsymbol{x}_n] \in \mathcal{R}^{m \times n}$$

Matrix Factorization

$$\begin{array}{cc} & n \\ m & \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ x_{13} & x_{23} & \cdots & x_{n3} \\ \vdots & \vdots & & \vdots \\ x_{1m} & x_{2m} & \cdots & x_{nm} \end{bmatrix} \end{array} \approx \begin{array}{cc} & k \\ m & \begin{bmatrix} u_{11} & \cdots & u_{k1} \\ u_{12} & \cdots & u_{k2} \\ u_{13} & \cdots & u_{k3} \\ \vdots & & \vdots \\ u_{1m} & \cdots & u_{km} \end{bmatrix} \end{array} \times \begin{array}{cc} n & \\ \begin{bmatrix} v_{11} & v_{21} & \cdots & v_{n1} \\ \vdots & \vdots & & \vdots \\ v_{1k} & v_{2k} & \cdots & v_{nk} \end{bmatrix} & k \end{array}$$

$$X \approx UV$$

$$\begin{bmatrix} \\ \mathbf{x}_i \\ \\ \end{bmatrix} \approx v_{i1} \cdot \begin{bmatrix} \\ \mathbf{u}_1 \\ \\ \end{bmatrix} + v_{i2} \cdot \begin{bmatrix} \\ \mathbf{u}_2 \\ \\ \end{bmatrix} + \cdots + v_{ik} \cdot \begin{bmatrix} \\ \mathbf{u}_k \\ \\ \end{bmatrix}$$

Algorithms

- Singular Value Decomposition
- Nonnegative Matrix Factorization
- Sparse Coding

Matrix Factorization

$$X \in \mathcal{R}^{m \times n}$$

$$U \in \mathcal{R}^{m \times k}, \qquad V \in \mathcal{R}^{k \times n}$$

$$UV = X$$

$$UV = \tilde{X} \approx X$$

$$\min_{rank(\tilde{X})=k} \left\| X - \tilde{X} \right\|_F^2$$

Singular Value Decomposition (SVD)

- For an arbitrary matrix $X \in \mathcal{R}^{m\times n}$ there exists a factorization as follows:

$$X = U\Sigma V$$

- where

$$U \in \mathcal{R}^{m\times m}, V \in \mathcal{R}^{n\times n}, UU^T = U^T U = I, VV^T = V^T V = I$$

$$\text{diagonal matrix } \Sigma \in \mathcal{R}^{m\times n}$$

- If $rank(X) = d$

$$U \in \mathcal{R}^{m\times d}, V \in \mathcal{R}^{d\times n}, U^T U = I, VV^T = I$$

$$\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \cdots \sigma_d) \;\; \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d > 0$$

SVD: Low-rank Approximation

- SVD can be used to compute optimal **low-rank approximations**.
- Approximation problem:

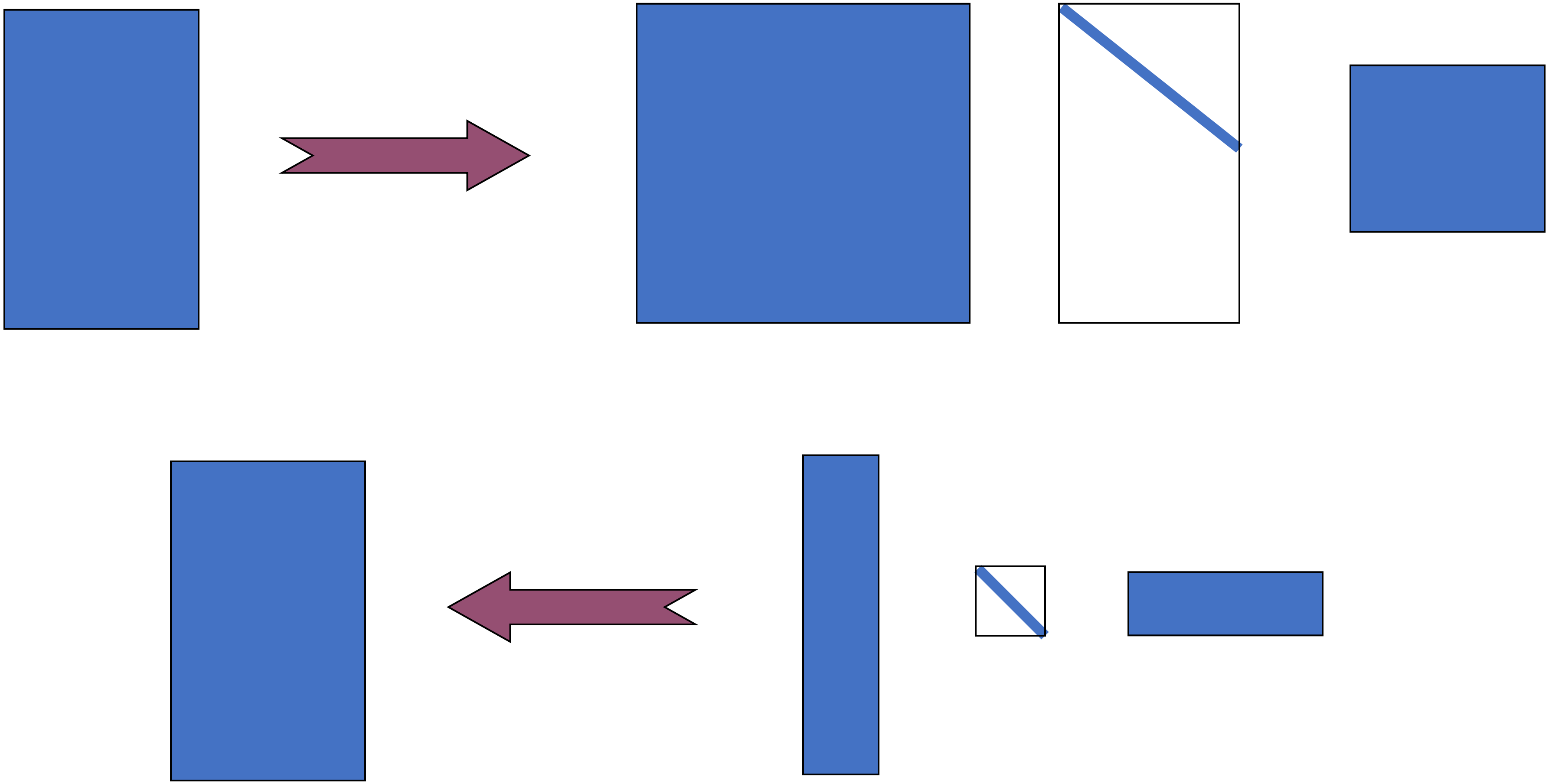
$$X^* = \underset{rank(\tilde{X})=k}{\operatorname{argmin}} \left\| X - \tilde{X} \right\|_F^2$$

- Solution via SVD

$$X^* = U\mathrm{diag}(\sigma_1, \cdots, \sigma_k, \underbrace{0, \cdots, 0}_{\text{set small singular values to zero}})V$$

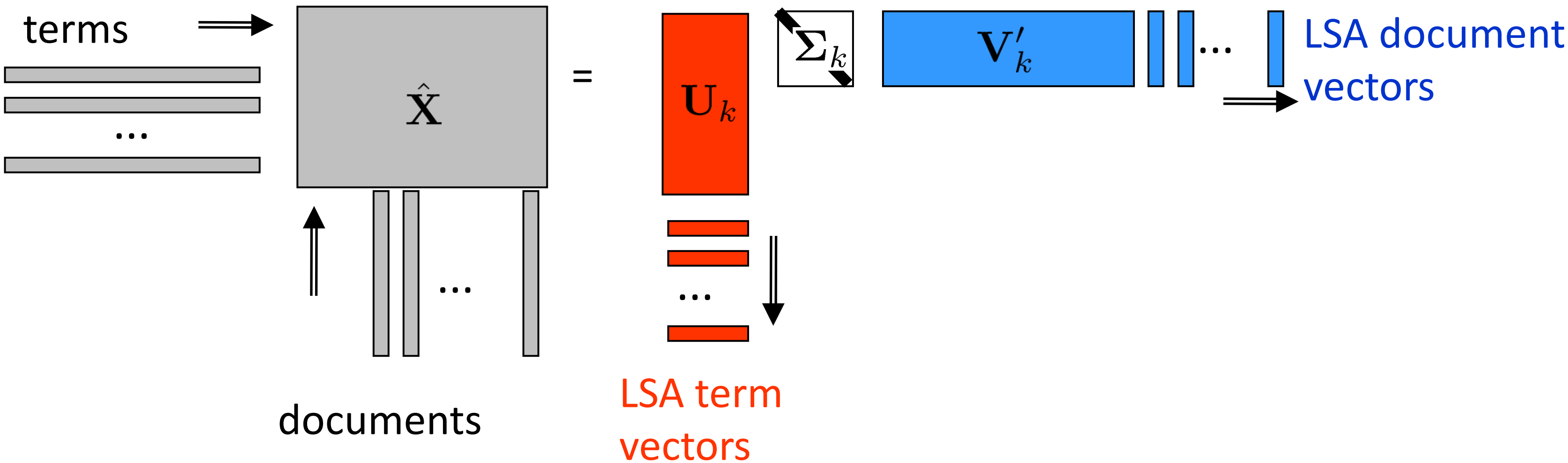
C. Eckart, G. Young, The approximation of a matrix by another of lower rank. Psychometrika, 1, 211-218, 1936.

Low rank approximation by SVD

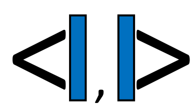


Latent Semantic Analysis (Indexing)

- The Latent Semantic Analysis via SVD can be summarized as follows:



- Document **similarity**
- $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle = \langle \Sigma_k \boldsymbol{v}_i, \Sigma_k \boldsymbol{v}_j \rangle$



Matrix Factorization: SVD

$$X \in \mathcal{R}^{m \times n}$$

$$U \in \mathcal{R}^{m \times k}, \qquad V \in \mathcal{R}^{k \times n}$$

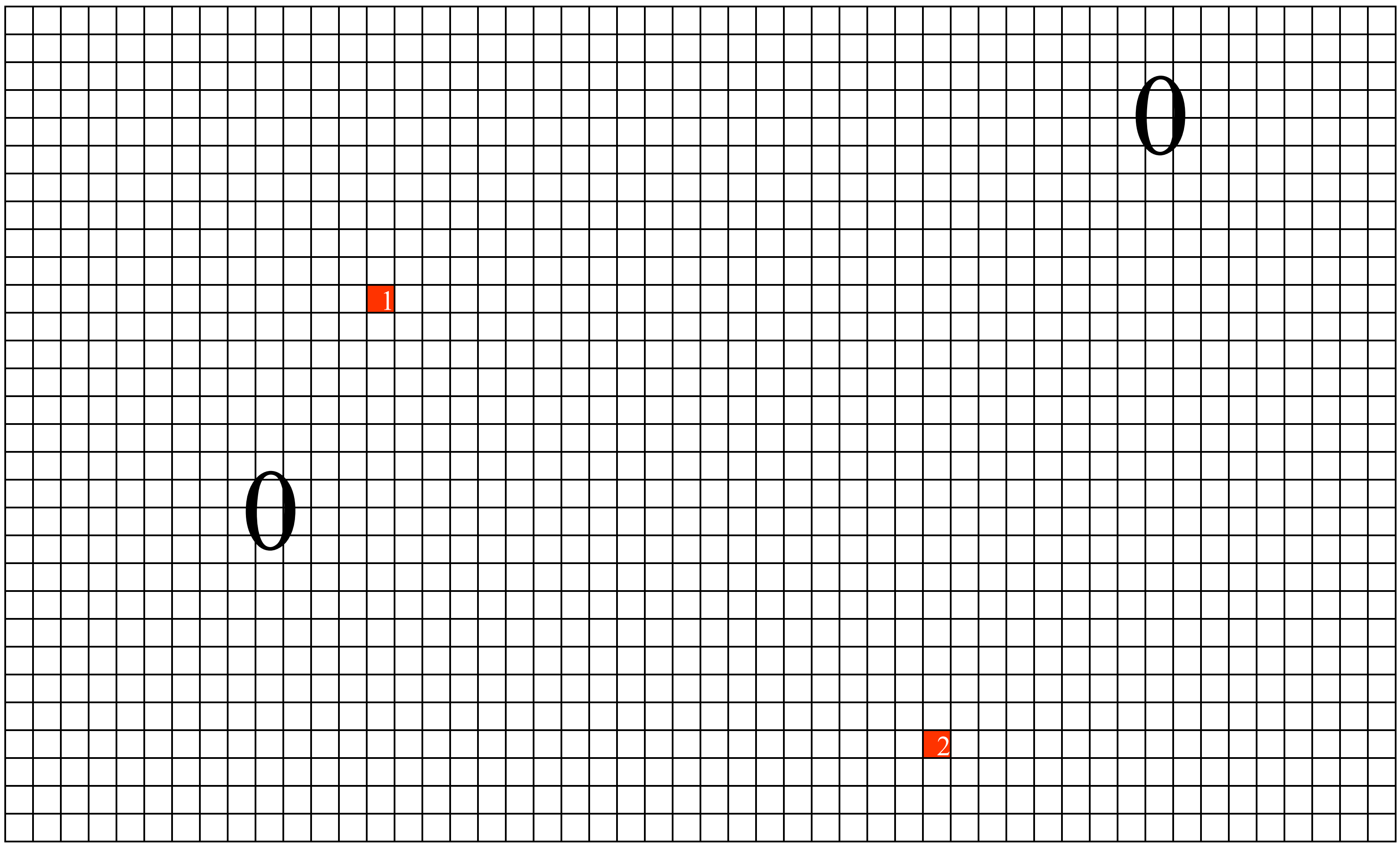
$$UV = \tilde{X} \approx X$$

$$\min_{rank(\tilde{X})=k} \left\| X - \tilde{X} \right\|_F^2$$

- Low Rank Assumption
- k Hidden Factors

Document-Term Matrix

$\boldsymbol{X} =$



Query Likelihood

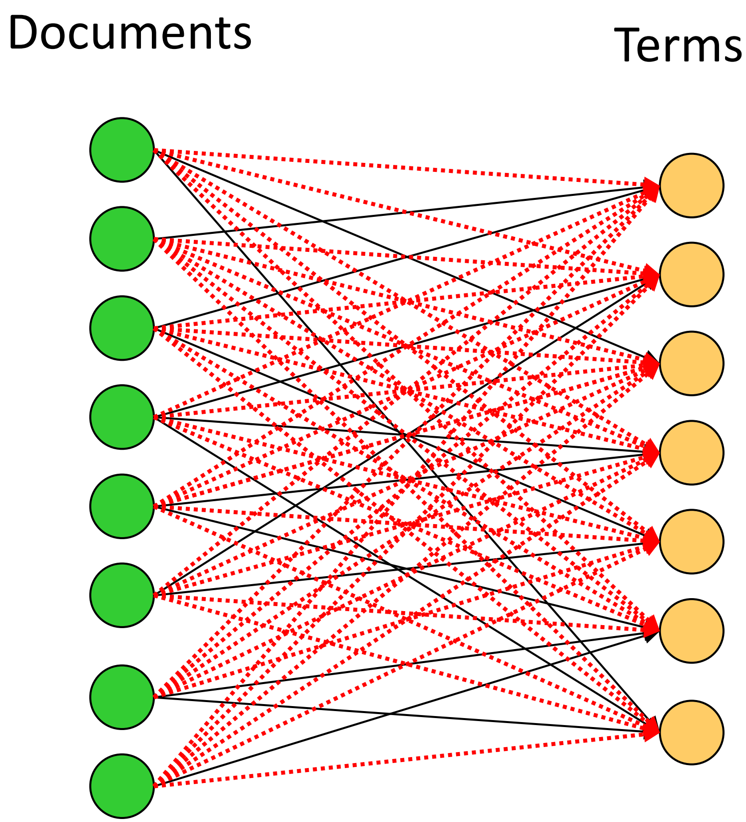
$$P(q|R_d = 1) \equiv P(q|d)$$

- $q = (w_1, \cdots, w_q)$
- Independent Assumption

$$P(q|d) = \Pi_{w \in q} P(w|d)$$

$$P(w|d)?$$

Naive Approach



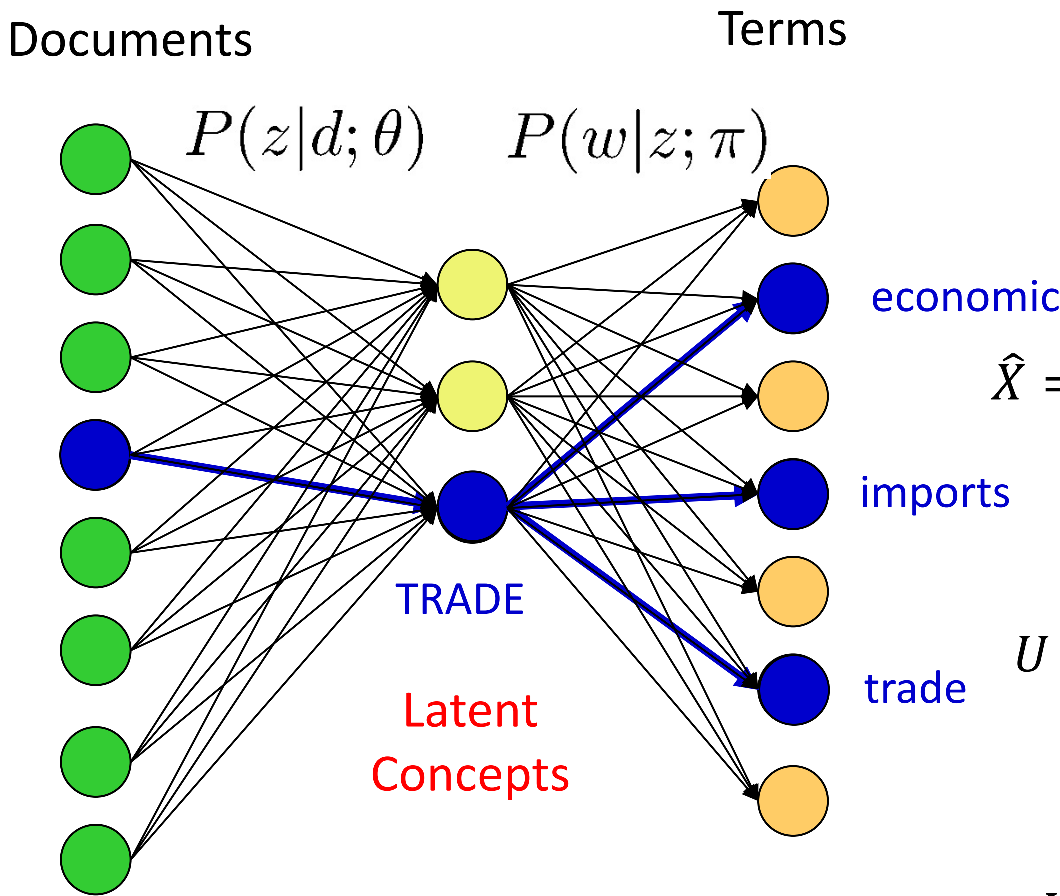
Maximum Likelihood Estimation

number of occurrences
of term *w* in document *d*

$$P(w|d) = \frac{n(d,w)}{\sum_{w'} n(d,w')}$$

$$X = \begin{bmatrix} P(w_1|d_1) & \cdots & P(w_1|d_n) \\ \vdots & \ddots & \vdots \\ P(w_m|d_1) & \cdots & P(w_m|d_n) \end{bmatrix}$$

Probabilistic Latent Semantic Analysis



$$\hat{P}(w|d) = \sum_z P(w|z)P(z|d)$$

$$\hat{X} = \begin{bmatrix} \hat{P}(w_1|d_1) & \cdots & \hat{P}(w_1|d_n) \\ \vdots & \ddots & \vdots \\ \hat{P}(w_m|d_1) & \cdots & \hat{P}(w_m|d_n) \end{bmatrix}$$

$$U = \begin{bmatrix} \hat{P}(w_1|z_1) & \cdots & \hat{P}(w_1|z_k) \\ \vdots & \ddots & \vdots \\ \hat{P}(w_m|z_1) & \cdots & \hat{P}(w_m|z_k) \end{bmatrix}$$

$$V = \begin{bmatrix} \hat{P}(z_1|d_1) & \cdots & \hat{P}(z_k|d_1) \\ \vdots & \ddots & \vdots \\ \hat{P}(z_1|d_n) & \cdots & \hat{P}(z_k|d_n) \end{bmatrix}$$

$$X \approx \hat{X} = UV^T$$

Matrix Factorization

$$\begin{array}{cc} & n \\ m & \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ x_{13} & x_{23} & \cdots & x_{n3} \\ \vdots & \vdots & & \vdots \\ x_{1m} & x_{2m} & \cdots & x_{nm} \end{bmatrix} \end{array} \approx \begin{array}{cc} & k \\ m & \begin{bmatrix} u_{11} & \cdots & u_{k1} \\ u_{12} & \cdots & u_{k2} \\ u_{13} & \cdots & u_{k3} \\ \vdots & & \vdots \\ u_{1m} & \cdots & u_{km} \end{bmatrix} \end{array} \times \begin{array}{cc} n & \\ \begin{bmatrix} v_{11} & v_{21} & \cdots & v_{n1} \\ \vdots & \vdots & & \vdots \\ v_{1k} & v_{2k} & \cdots & v_{nk} \end{bmatrix} & k \end{array}$$

$$X \approx UV$$

$$\begin{bmatrix} \\ \mathbf{x}_i \\ \\ \end{bmatrix} \approx v_{i1} \cdot \begin{bmatrix} \\ \mathbf{u}_1 \\ \\ \end{bmatrix} + v_{i2} \cdot \begin{bmatrix} \\ \mathbf{u}_2 \\ \\ \end{bmatrix} + \cdots + v_{ik} \cdot \begin{bmatrix} \\ \mathbf{u}_k \\ \\ \end{bmatrix}$$

Nonnegative Matrix Factorization

$$X \in \mathcal{R}^{m \times n}$$
$$U \in \mathcal{R}^{m \times k}, \qquad V \in \mathcal{R}^{k \times n}$$
$$UV = \tilde{X} \approx X$$
$$u_{ij} \geq 0, v_{ij} \geq 0$$

- Low rank assumption (k hidden factors)
- Nonnegative assumption

Non-negative Matrix Factorization

$$X \cong \hat{X} = UV^{T}, u_{ij} \geq 0, v_{ij} \geq 0$$

- Two cost functions
  - Euclidean distance

$$\left||A - B|\right|^2 = \Sigma_{ij}\left(A_{ij} - B_{ij}\right)^2$$

  - Divergence

$$D(A||B) = \Sigma_{ij}(A_{ij} \log \frac{A_{ij}}{B_{ij}} - A_{ij} + B_{ij})$$

Optimization Problems

- *Minimize* $||X - UV^T||^2$ *with respect to* U *and* V*, subject to the constraints* $U, V \geq 0$.

- *Minimize* $D(X||UV^T)$ *with respect to* U *and* V*, subject to the constraints* $U, V \geq 0$.

NMF Optimization (Euclidean Distance)

$$\min \left\| \left| X - UV^T \right| \right\|^2, s.t.\, u_{ij} \geq 0, v_{ij} \geq 0$$

$$J = \left\| \left| X - UV^T \right| \right\|^2 = \mathrm{tr}\left((X - UV^T)^T (X - UV^T) \right)$$

$$= \mathrm{tr}(X^T X - X^T U V^T - V U^T X + V U^T U V^T)$$

Γ, same size as U

Φ, same size as V

$$\mathcal{L} = \mathrm{tr}(X^T X) - 2\mathrm{tr}(X^T U V^T) + \mathrm{tr}(V U^T U V^T) + \mathrm{tr}(\Gamma U^T) + \mathrm{tr}(\Phi V^T)$$

$$\frac{\partial \mathcal{L}}{\partial U} = -2XV + 2UV^T V + \Gamma \qquad (UV^T V)_{ik} u_{ik} - (XV)_{ik} u_{ik} = 0$$

$$\boxed{u_{ik} \leftarrow \frac{(XV)_{ik}}{(UV^T V)_{ik}} u_{ik}}$$

$$\frac{\partial \mathcal{L}}{\partial V} = -2X^T U + 2VU^T U + \Phi \qquad (VU^T U)_{jk} v_{jk} - (X^T U)_{jk} v_{jk} = 0$$

$$\boxed{v_{jk} \leftarrow \frac{(X^T U)_{jk}}{(VU^T U)_{jk}} v_{jk}}$$

Multiplicative Update Rules

- *The Euclidean distance* $\left|\left|X - UV^T\right|\right|^2$ *is nonincreasing under the update rules*

$$u_{ik} \leftarrow \frac{(XV)_{ik}}{(UV^TV)_{ik}} u_{ik} \qquad v_{jk} \leftarrow \frac{(X^TU)_{jk}}{(VU^TU)_{jk}} v_{jk}$$

The Euclidean distance is invariant under these updates if and only if U and V are at a stationary point of the distance.

Matrix Factorization

$$X \in \mathcal{R}^{m \times n}$$
$$U \in \mathcal{R}^{m \times k}, \qquad V \in \mathcal{R}^{k \times n}$$
$$UV = \tilde{X} \approx X$$

- Low rank assumption (k hidden factors)
  - SVD
- Nonnegative assumption
  - NMF

Sparse Coding

$$X = \tilde{X} \approx U \cdot V^T$$

$$\begin{bmatrix} \\ \mathbf{x}_i \\ \\ \end{bmatrix} \approx v_{1i} \cdot \begin{bmatrix} \\ \mathbf{u}_1 \\ \\ \end{bmatrix} + v_{2i} \cdot \begin{bmatrix} \\ \mathbf{u}_2 \\ \\ \end{bmatrix} + \cdots + v_{ki} \cdot \begin{bmatrix} \\ \mathbf{u}_k \\ \\ \end{bmatrix}$$

$$\text{minimize}_{U,V} \; \left\| X - UV^T \right\|_F^2 + \lambda f(V)$$

subject to $\Sigma_i u_{i,k}^2 \leq c, \forall k = 1, \ldots, K.$

- Represent input vectors approximately as a weighted linear combination of a small number of "basis vectors."

Matrix Factorization: Summary

$$X \in \mathcal{R}^{m \times n}$$
$$U \in \mathcal{R}^{m \times k}, \qquad V \in \mathcal{R}^{k \times n}$$
$$UV = \tilde{X} \approx X$$

- Low rank assumption (k hidden factors)
  - SVD
- Nonnegative assumption
  - NMF
- Sparseness assumption
  - Sparse Coding