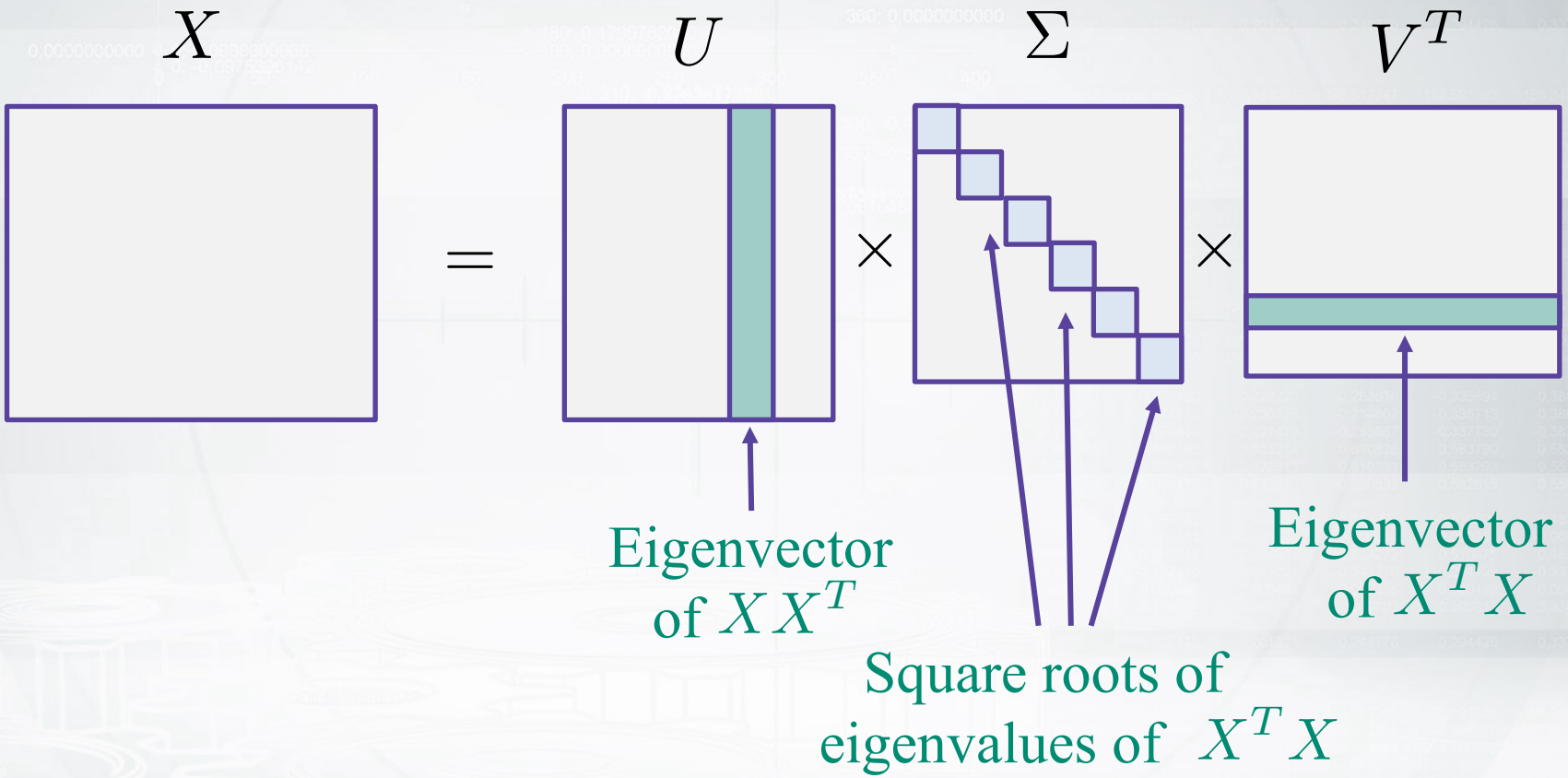
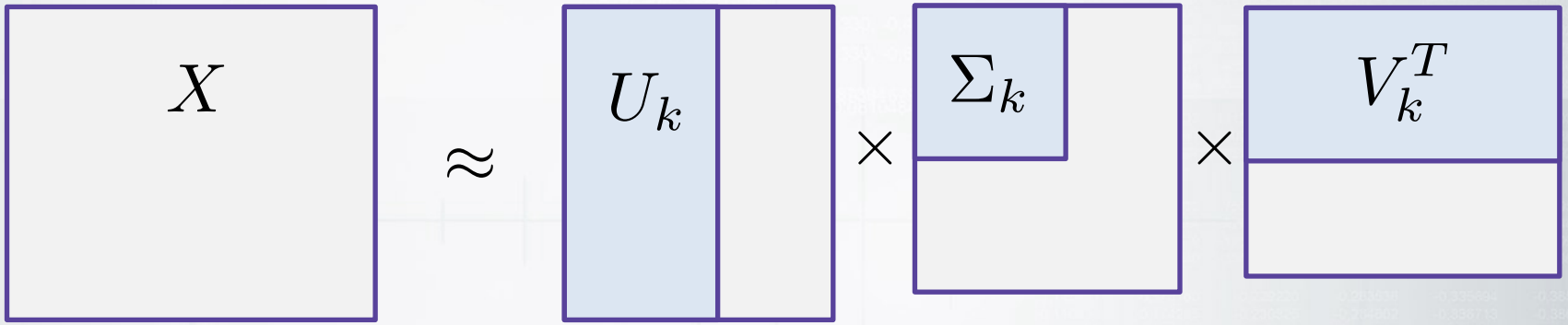
Explicit and implicit matrix factorization

Singular Value Decomposition (SVD)



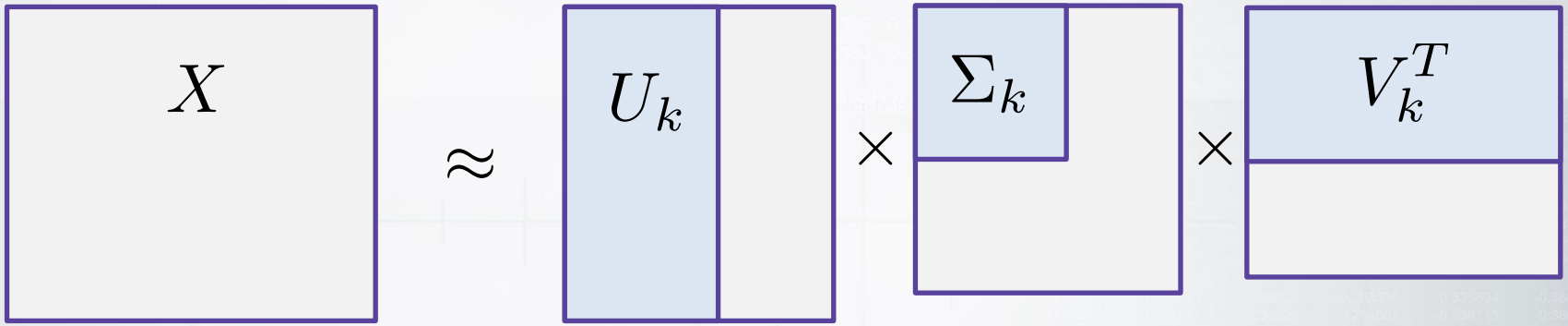
Truncated SVD

Keep only first k components: $\hat{X}_k = U_k \Sigma_k V_k^T$



Truncated SVD

Keep only first k components: $\hat{X}_k = U_k \Sigma_k V_k^T$

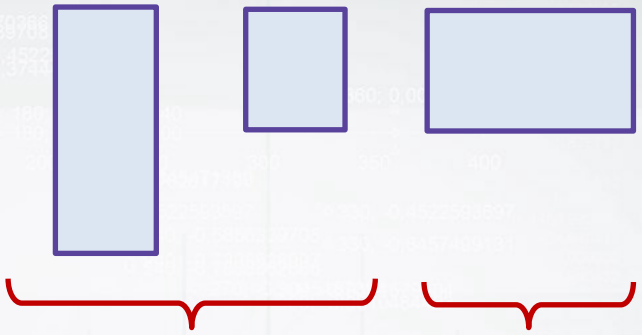


It's the best approximation of rank k in terms of Frobenius norm:

$$||X - \hat{X}||_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij} - \hat{x}_{ij})^2}$$

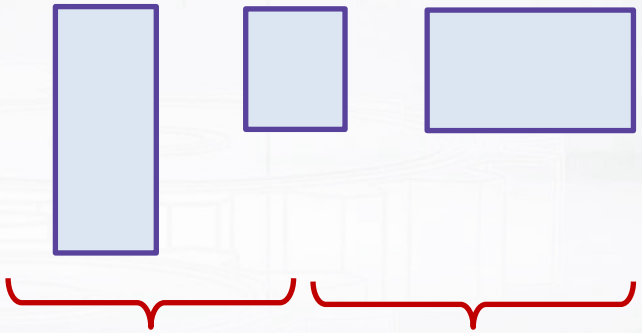
How do we use it?

Option 1:



$$\Phi = U_k \Sigma_k \quad \Theta = V_k^T$$

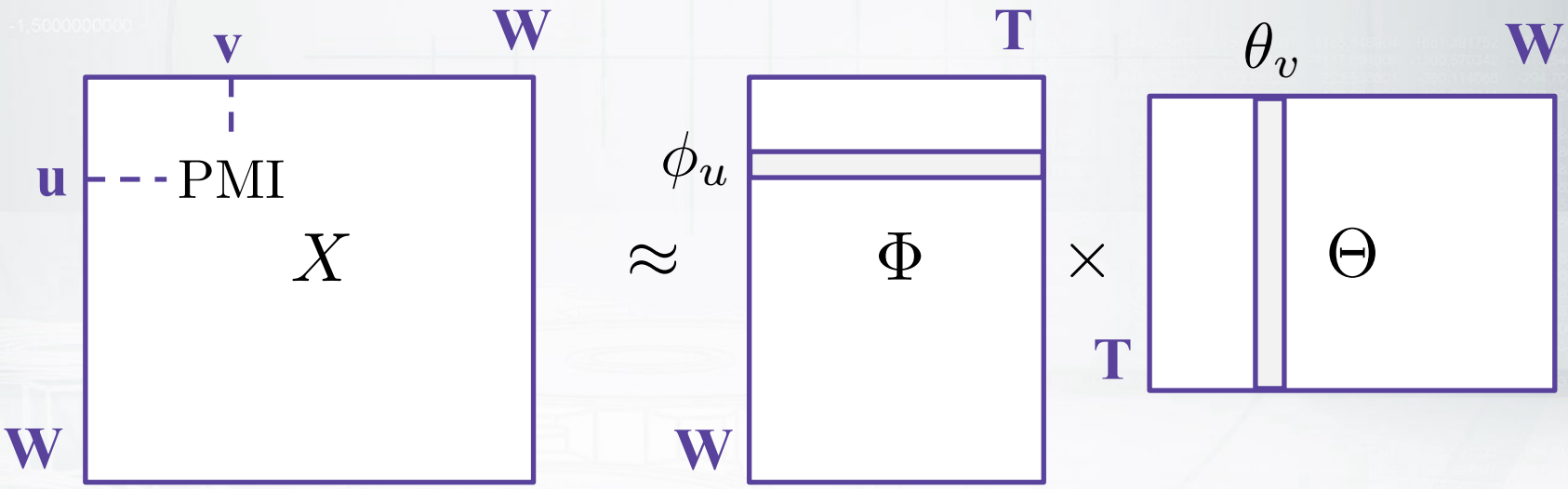
Option 2:



$$\Phi = U_k \sqrt{\Sigma_k} \quad \Theta = \sqrt{\Sigma_k} V_k^T$$

Vector Space Models of Semantics

- **Input:** word-word co-occurrences (counts, PMI, …)
- **Method:** dimensionality reduction (SVD, …)
- **Output:** similarity between vector representations of words

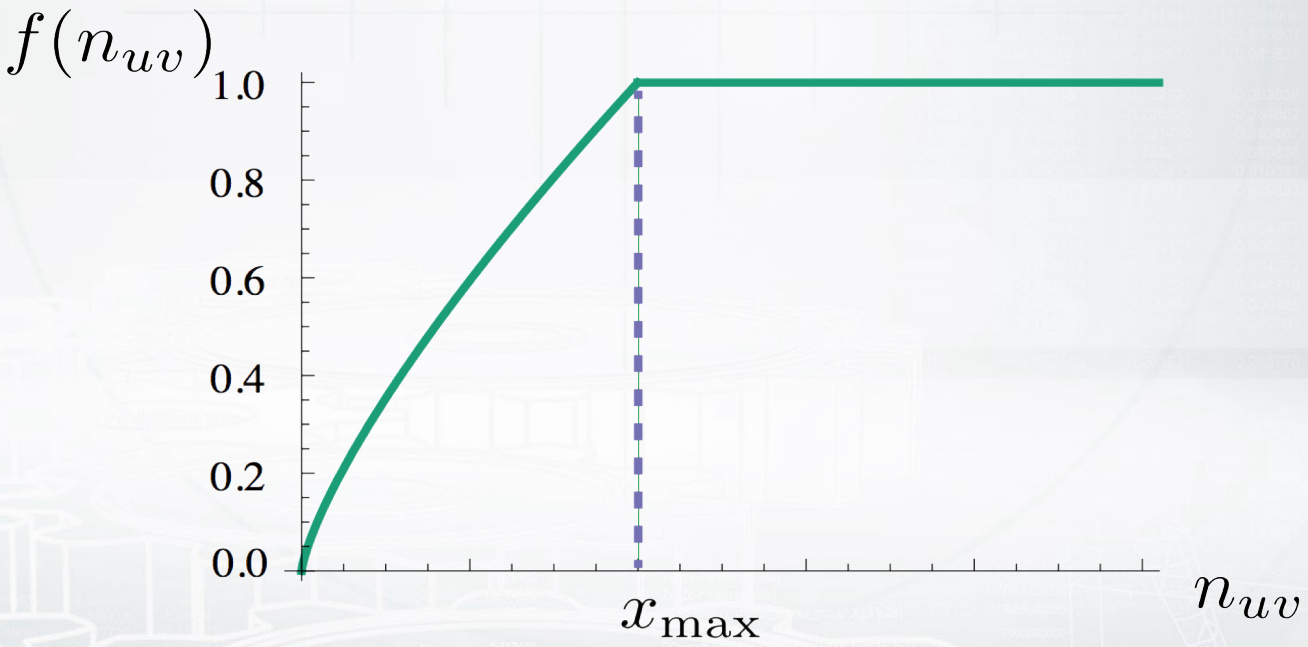


Turnay, P.D., Pantel, P.: from Frequency to Meaning: Vector Space Models of Semantics, 2010.

Weighted squared loss: GloVe

Fill X with $\log n_{uv}$ and try another objective:

$$\sum_{u \in W} \sum_{v \in W} f(n_{uv}) \left(\langle \phi_u, \theta_v \rangle + b_u + b'_v - \log n_{uv} \right)^2 \to \min_{\phi_u, \theta_v, b_u, b'_v}$$



Pennington et. al. GloVe: Global Vectors for Word Representation, 2014.

Word prediction: skip-gram model

Predict *context words* given a focus word:

$$p(w_{i-h}, \dots w_{i+h} | w_i) = \prod_{-h \le k \le h,\, k \neq 0} p(w_{i+k} | w_i)$$

Model each probability with a *softmax*:

$$p(u|v) = \frac{\exp \langle \phi_u, \theta_v \rangle}{\sum_{u' \in W} \exp \langle \phi_{u'}, \theta_v \rangle}$$

Still two matrices of parameters.

How do we train the model?

Log-likelihood maximization:

$$\mathcal{L} = \sum_{u \in W} \sum_{v \in W} n_{uv} \log p(u|v)$$

word co-occurrence

Method:

- SGD, online by word pairs in the corpus

Problem:

- *softmax* over vocabulary is slow!

Skip-gram Negative Sampling (SGNS)

Instead of predicting a word for another word, predict "yes" or "no" for word pairs:

$$\sum_{u \in W} \sum_{v \in W} n_{uv} \log \sigma \left(\langle \phi_u, \theta_v \rangle \right) +$$

$$k \, \mathbb{E}_{\bar{v}} \log \sigma \left(- \langle \phi_u, \theta_{\bar{v}} \rangle \right) \to \max_{\phi_u, \theta_v}$$

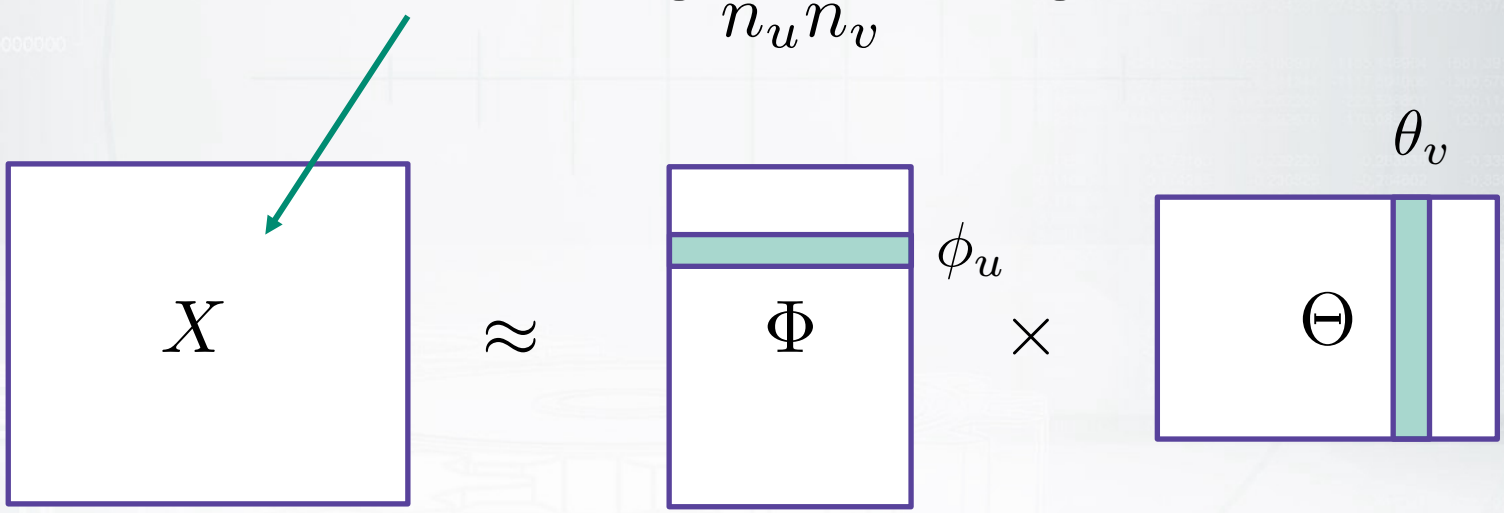
- Use positive examples from data: v co-occurred with u
- Sample negative examples: k random $\bar{v}$ from the vocabulary

Train with SGD to find two matrices of parameters (as usual).

SGNS as implicit matrix factorization

SGNS objective is maximized when $\langle \phi_u, \theta_v \rangle$ is equal to shifted Pointwise Mutual Information:

$$\text{sPMI} = \log \frac{n_{uv} n}{n_u n_v} - \log k$$



Levy and Goldberg. Neural Word Embedding as Implicit Matrix Factorization, 2014.