CSE 291

Algorithms for Non-negative Matrix Factorization

by D. D. Lee and H. S. Seung

Youngmin Cho

Abstract

- NMF review
- Cost functions and multiplicative algorithms
- Interpretation as gradient descent
- Proof of monotonic convergence

Question:

Given a non-negative matrix V,

find non-negative matrix factors W and H:

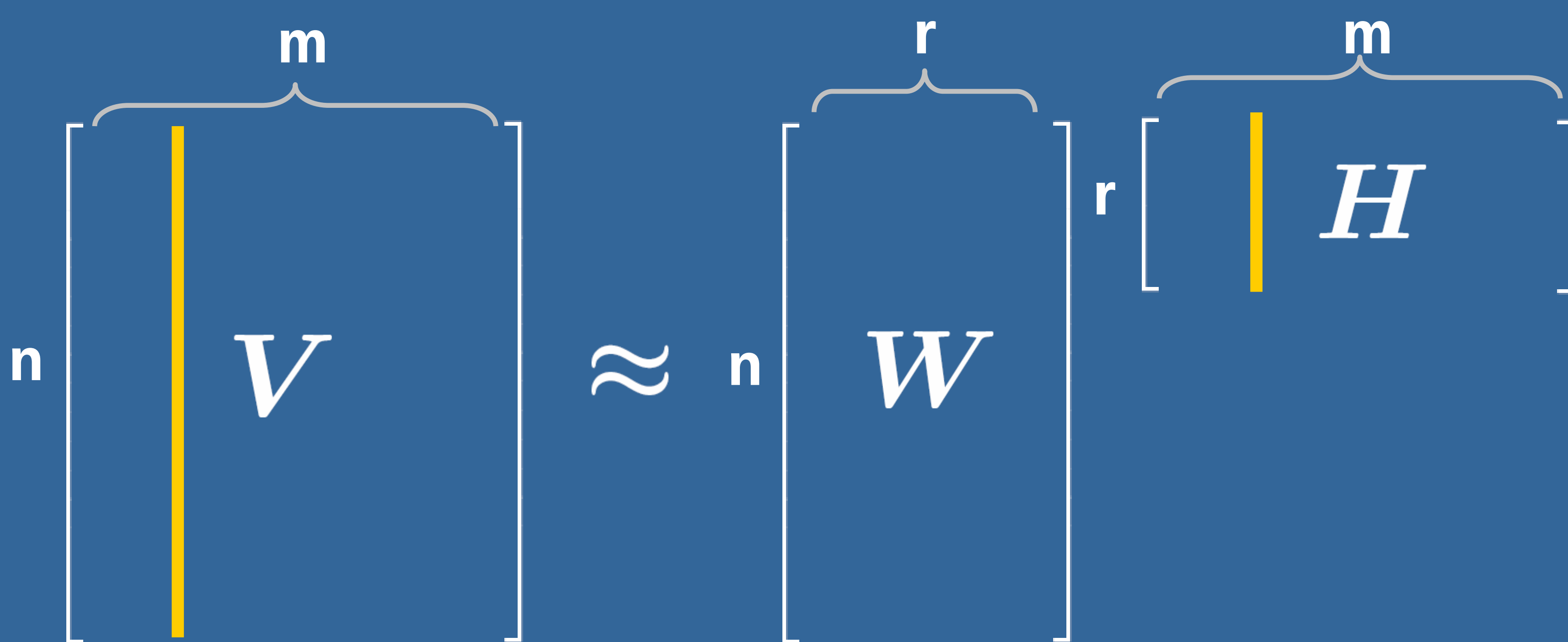
$$V \approx WH$$

Answer:

Non-negative Matrix Factorization (NMF)

NMF

$$V \approx WH \quad (1)$$



$$v \approx Wh$$

How to solve it

- Two simple and convergent algorithms

$$H_{a\mu} \leftarrow H_{a\mu} \frac{(W^T V)_{a\mu}}{(W^T W H)_{a\mu}}$$

$$W_{ia} \leftarrow W_{ia} \frac{(V H^T)_{ia}}{(W H H^T)_{ia}} \tag{4}$$

distance

$$H_{a\mu} \leftarrow H_{a\mu} \frac{\sum_i W_{ia} V_{i\mu}/(WH)_{i\mu}}{\sum_k W_{ka}}$$

$$W_{ia} \leftarrow W_{ia} \frac{\sum_\mu H_{a\mu} V_{i\mu}/(WH)_{i\mu}}{\sum_\nu H_{a\nu}} \tag{5}$$

divergence

Cost functions

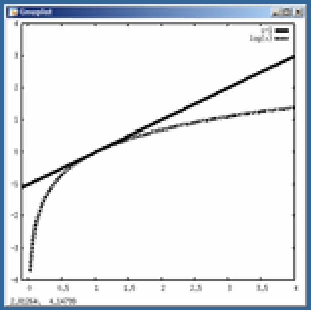
- Square of the Euclidean distance between A and B

$$\|A - B\|^2 = \sum_{ij} (A_{ij} - B_{ij})^2 \qquad (2)$$

- Generalized Kullback-Leibler divergence of A from B

$$D(A\|B) = \sum_{ij} \left(A_{ij} \log \frac{A_{ij}}{B_{ij}} - A_{ij} + B_{ij} \right) \qquad (3)$$

$$A_{ij} \log \frac{A_{ij}}{B_{ij}} - A_{ij} + B_{ij} = A_{ij} \left(\frac{B_{ij}}{A_{ij}} - 1 - \log \frac{B_{ij}}{A_{ij}} \right)$$



How to minimize it

- Minimize $\|V - WH\|^2$ or $D(V\|WH)$
- Convex in W only or H only
  (not convex in both variables)
- Goal - finding local minima
- Gradient descent?
  - Slow convergence
  - Sensitive to the step size
    (i.e., inconvenient for large apps)

Multiplicative update rules

distance

$$H_{a\mu} \leftarrow H_{a\mu} \frac{(W^T V)_{a\mu}}{(W^T W H)_{a\mu}}$$

$$W_{ia} \leftarrow W_{ia} \frac{(V H^T)_{ia}}{(W H H^T)_{ia}} \qquad (4)$$

divergence

$$H_{a\mu} \leftarrow H_{a\mu} \frac{\sum_i W_{ia} V_{i\mu} / (WH)_{i\mu}}{\sum_k W_{ka}}$$

$$W_{ia} \leftarrow W_{ia} \frac{\sum_\mu H_{a\mu} V_{i\mu} / (WH)_{i\mu}}{\sum_\nu H_{a\nu}} \qquad (5)$$

- Guaranteed to converge

Gradient descent?

- Gradient descent for $f(\vec{\theta})$

$$\vec{\theta} \leftarrow \vec{\theta} - \eta\left(\frac{\partial f}{\partial \vec{\theta}}\right)$$

- f = $\|V - WH\|^2$

$$H_{a\mu} \leftarrow H_{a\mu} + \eta_{a\mu}\left[(W^T V)_{a\mu} - (W^T W H)_{a\mu}\right] \quad (6)$$

- guaranteed to converge as long as η is sufficiently small

Multiplicative vs. Additive rules

$$H_{a\mu} \leftarrow H_{a\mu} + \eta_{a\mu}\left[(W^TV)_{a\mu} - (W^TWH)_{a\mu}\right] \quad (6)$$

$$\eta_{a\mu} = \frac{H_{a\mu}}{(W^TWH)_{a\mu}} \quad (7)$$

$$H_{a\mu} \leftarrow H_{a\mu}\frac{(W^TV)_{a\mu}}{(W^TWH)_{a\mu}} \quad (4)$$

- similar for W (distance), W and H (divergence)
- Is it convergent, even if η is not necessarily small enough?

Proof sketch for monotonic convergence

- Define an auxiliary function $G(h, h^t)$ for $F(h)$ (similar to EM)

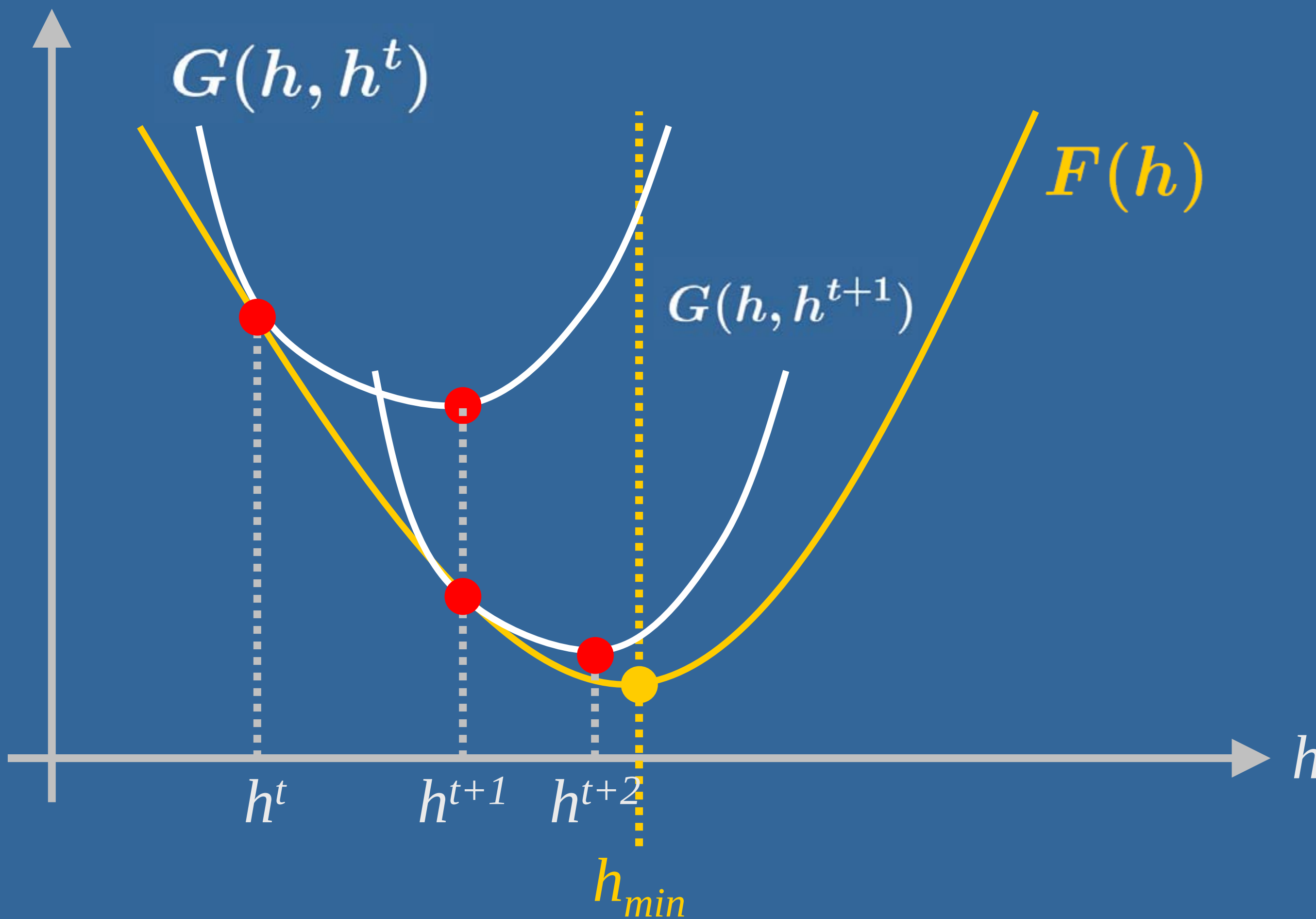
  $G(h, h') \geq F(h)$, $G(h, h) = F(h)$ (10)

- Find a local minimum of G by following repeatedly $h^{t+1} = \underset{h}{\operatorname{argmin}}\, G(h, h^t)$ (11)

- This sequence is converging to a local minimum of $F(h)$

  $(\because F(h^{t+1}) \leq G(h^{t+1}, h^t) \leq G(h^t, h^t) = F(h^t))$

Auxiliary function



$$h^{t+1} = \underset{h}{\operatorname{argmin}}\, G(h, h^t) \quad (11)$$

Updates for H of Euclidean distance

$$G(h, h^t) = F(h^t) + (h - h^t)^T \nabla F(h^t) + \frac{1}{2}(h - h^t)^T K(h^t)(h - h^t) \quad (14)$$

$$K_{ab}(h^t) = \delta_{ab}(W^T W h^t)_a / h^t_a \quad (13)$$

$$F(h) = \frac{1}{2}\sum_i (v_i - \sum_a W_{ia} h_a)^2 \quad (15)$$

- Proving steps

  : Show $G(h, h^t)$ is an auxiliary function for $F(h)$

  : Obtain the minimum of $G(h, h^t)$ by setting the gradient to zero

  : Check the equivalence between this updating rule and $H_{a\mu} \leftarrow H_{a\mu} \frac{(W^T V)_{a\mu}}{(W^T W H)_{a\mu}} \quad (4)$

Auxiliary function G(h, h^t) for F(h)

$$F(h) = \frac{1}{2}\sum_i (v_i - \sum_a W_{ia}h_a)^2 \quad (15)$$

$$F(h) = F(h^t) + (h - h^t)^T \nabla F(h^t) + \frac{1}{2}(h - h^t)^T (W^T W)(h - h^t) \quad (16)$$

$$G(h, h^t) = F(h^t) + (h - h^t)^T \nabla F(h^t) + \frac{1}{2}(h - h^t)^T K(h^t)(h - h^t) \quad (14)$$

$$0 \leq (h - h^t)^T \left[K(h^t) - W^T W\right](h - h^t) \quad (17)$$

$$M_{ab}(h^t) = h_a^t (K(h^t) - W^T W)_{ab} h_b^t \quad (18)$$

$$\nu^T M \nu = \sum_{ab} \nu_a M_{ab} \nu_b \quad = \sum_{ab} h_a^t (W^T W)_{ab} h_b^t \nu_a^2 - \nu_a h_a^t (W^T W)_{ab} h_b^t \nu_b \quad (20)$$

$$= \sum_{ab} (W^T W)_{ab} h_a^t h_b^t \left[\frac{1}{2}\nu_a^2 + \frac{1}{2}\nu_b^2 - \nu_a \nu_b\right] \quad (21)$$

$$= \frac{1}{2}\sum_{ab} (W^T W)_{ab} h_a^t h_b^t (\nu_a - \nu_b)^2 \quad (22)$$

Minimum of G(h, h^t) and update rules

$$h^{t+1} = h^t - K(h^t)^{-1} \nabla F(h^t) \qquad (24)$$

$$h_a^{t+1} = h_a^t \frac{(W^T v)_a}{(W^T W h^t)_a} \qquad (25)$$

$$H_{a\mu} \leftarrow H_{a\mu} \frac{(W^T V)_{a\mu}}{(W^T W H)_{a\mu}} \qquad (4)$$

- can be shown similarly for W of Euclidean distance

Yet another updates for H of divergence

$$G(h,h^t) = \sum_i (v_i \log v_i - v_i) + \sum_{ia} W_{ia} h_a - \sum_{ia} v_i \frac{W_{ia} h_a^t}{\sum_b W_{ib} h_b^t} \left(\log W_{ia} h_a - \log \frac{W_{ia} h_a^t}{\sum_b W_{ib} h_b^t} \right) \quad (26)$$

$$F(h) = \sum_i v_i \log \left(\frac{v_i}{\sum_a W_{ia} h_a} \right) - v_i + \sum_a W_{ia} h_a \quad (28)$$

- Proving steps

  : Show $G(h, h^t)$ is an auxiliary function for $F(h)$

  : Obtain the minimum of $G(h, h^t)$ by setting the gradient to zero

  : Check the equivalence between this updating rule and $H_{a\mu} \leftarrow H_{a\mu} \frac{\sum_i W_{ia} V_{i\mu}/(WH)_{i\mu}}{\sum_k W_{ka}} \quad (5)$

Proof of G(h, h^t) by convexity

$$G(h,h^t) = \sum_i (v_i \log v_i - v_i) + \sum_{ia} W_{ia}h_a - \sum_{ia} v_i \frac{W_{ia}h_a^t}{\sum_b W_{ib}h_b^t}\left(\log W_{ia}h_a - \log \frac{W_{ia}h_a^t}{\sum_b W_{ib}h_b^t}\right) \quad (26)$$

$$F(h) = \sum_i v_i \log\left(\frac{v_i}{\sum_a W_{ia}h_a}\right) - v_i + \sum_a W_{ia}h_a \quad (28)$$

$$-\log \sum\nolimits_a W_{ia}h_a = -\log \sum\nolimits_a \alpha_a \frac{W_{ia}h_a}{\alpha_a} \leq -\sum_a \alpha_a \log \frac{W_{ia}h_a}{\alpha_a} \quad (29)$$

$$\alpha_a = \frac{W_{ia}h_a^t}{\sum_b W_{ib}h_b^t} \quad (30)$$

$$-\log \sum\nolimits_a W_{ia}h_a \leq -\sum_a \frac{W_{ia}h_a^t}{\sum_b W_{ib}h_b^t}\left(\log W_{ia}h_a - \log \frac{W_{ia}h_a^t}{\sum_b W_{ib}h_b^t}\right) \quad (31)$$

$$h_a^{t+1} = \frac{h_a^t}{\sum_k W_{ka}} \sum_i \frac{v_i}{\sum_b W_{ib}h_b^t} W_{ia} \quad (33)$$

Recap: Proof sketch for monotonic convergence

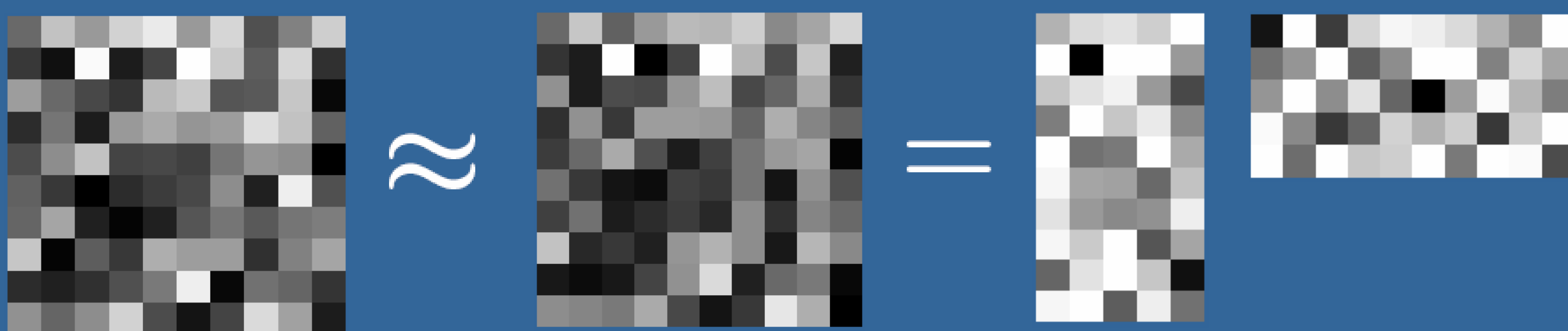
- Define an auxiliary function $G(h, h^t)$ for $F(h)$ (similar to EM)
- Find a local minimum of G by following repeatedly $h^{t+1} = \arg\min_h G(h, h^t) \quad (11)$

  : Obtain the minimum of $G(h, h^t)$ by setting the gradient to zero
- This sequence is converging to a local minimum of $F(h)$

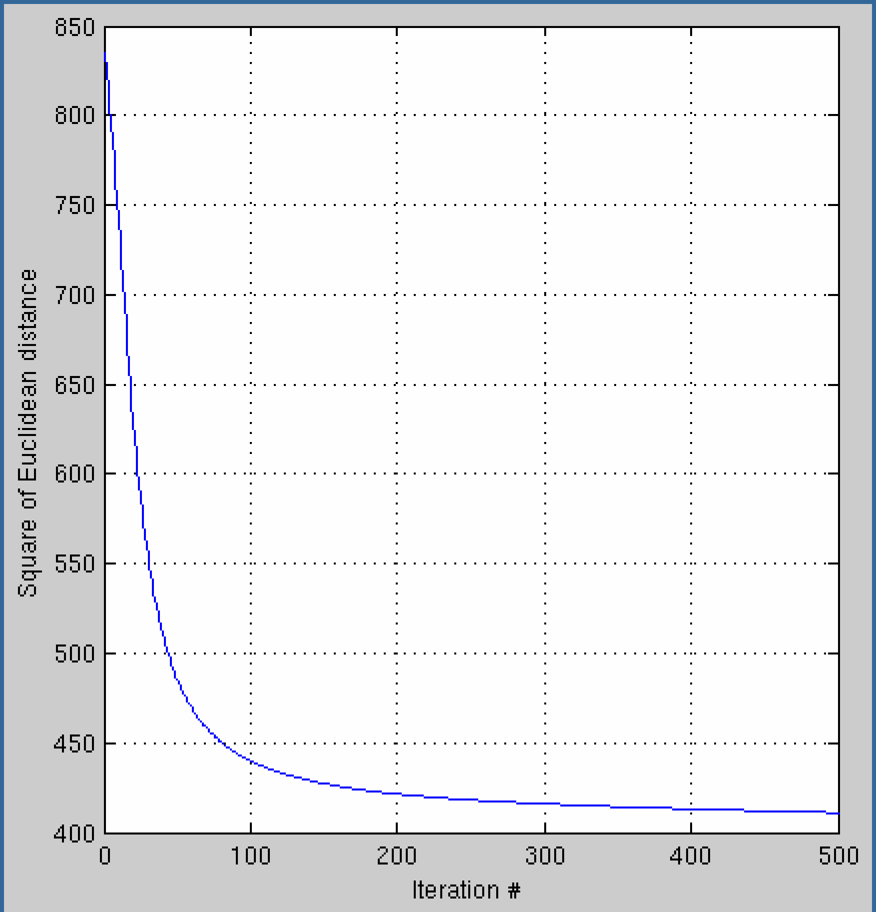
  : equivalent to the updating rules (4) and (5)

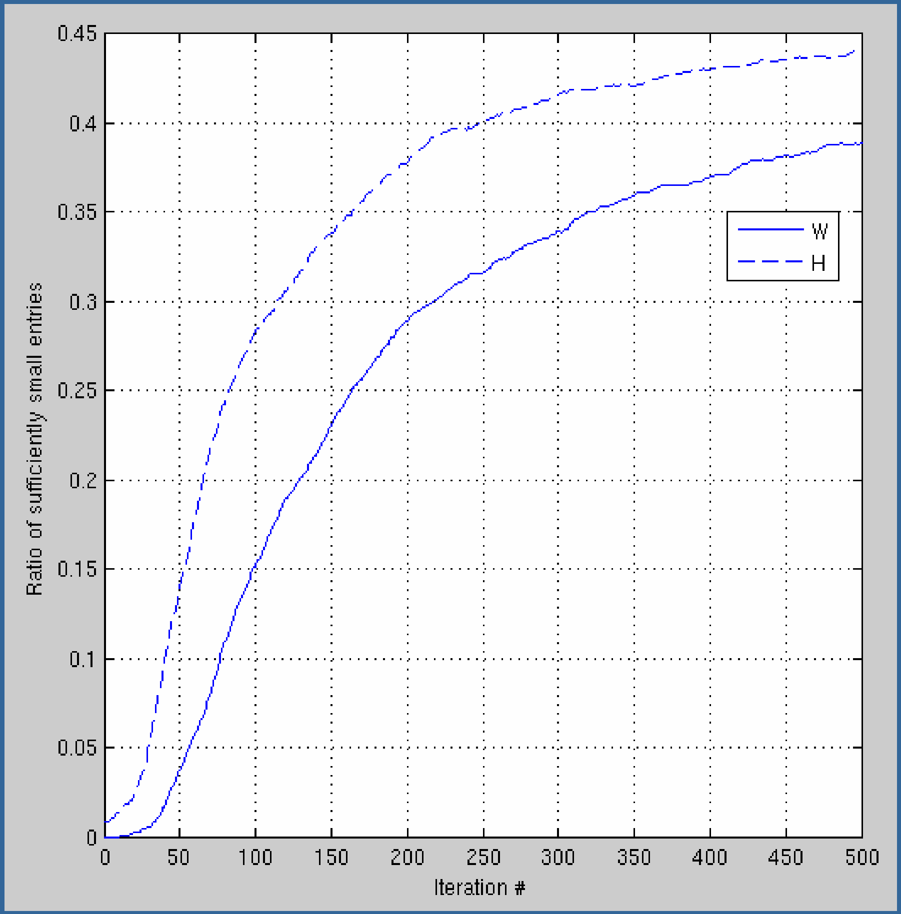
Example: Random matrix

$$V \approx WH$$



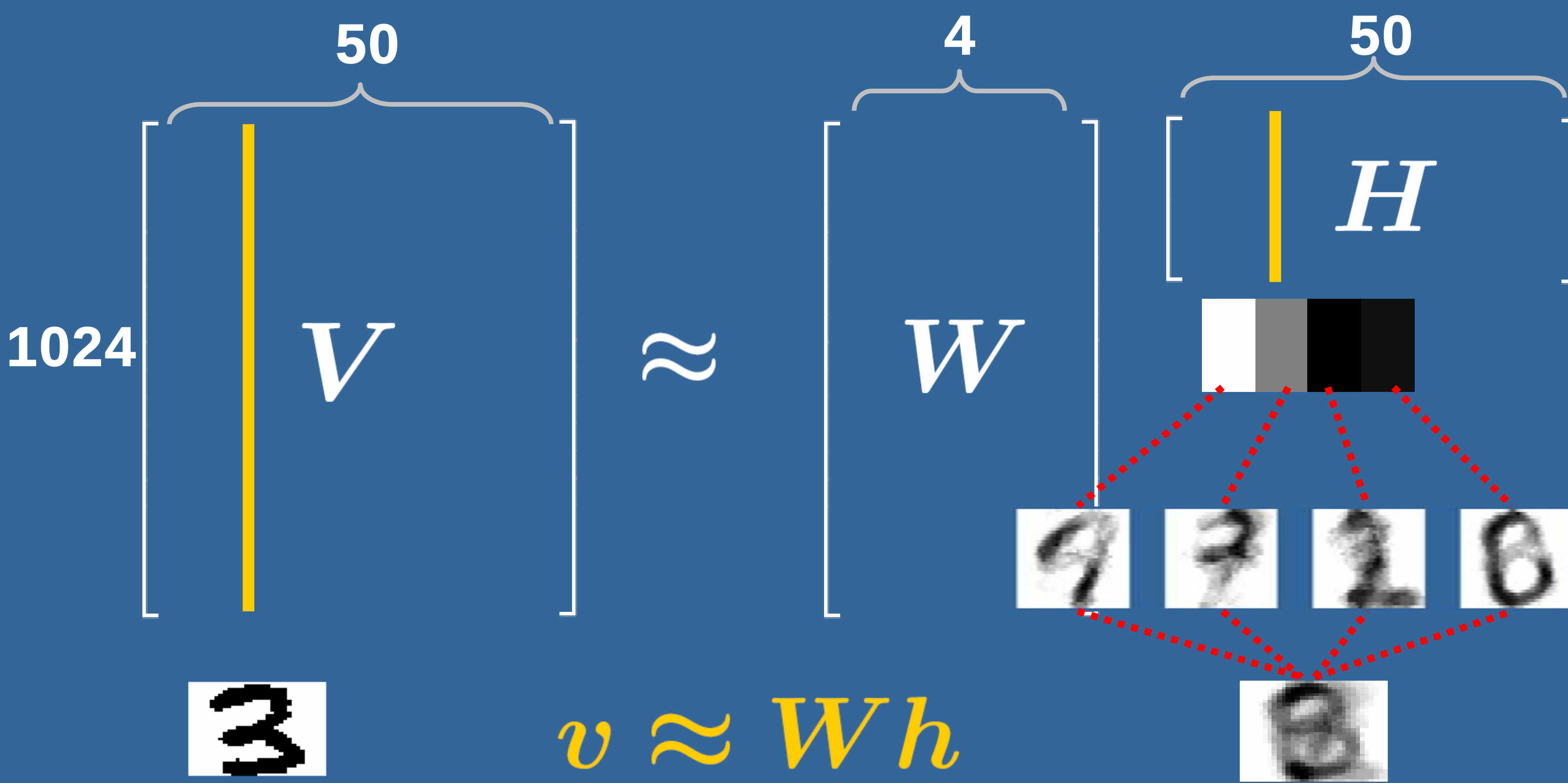
Convergence & Sparseness

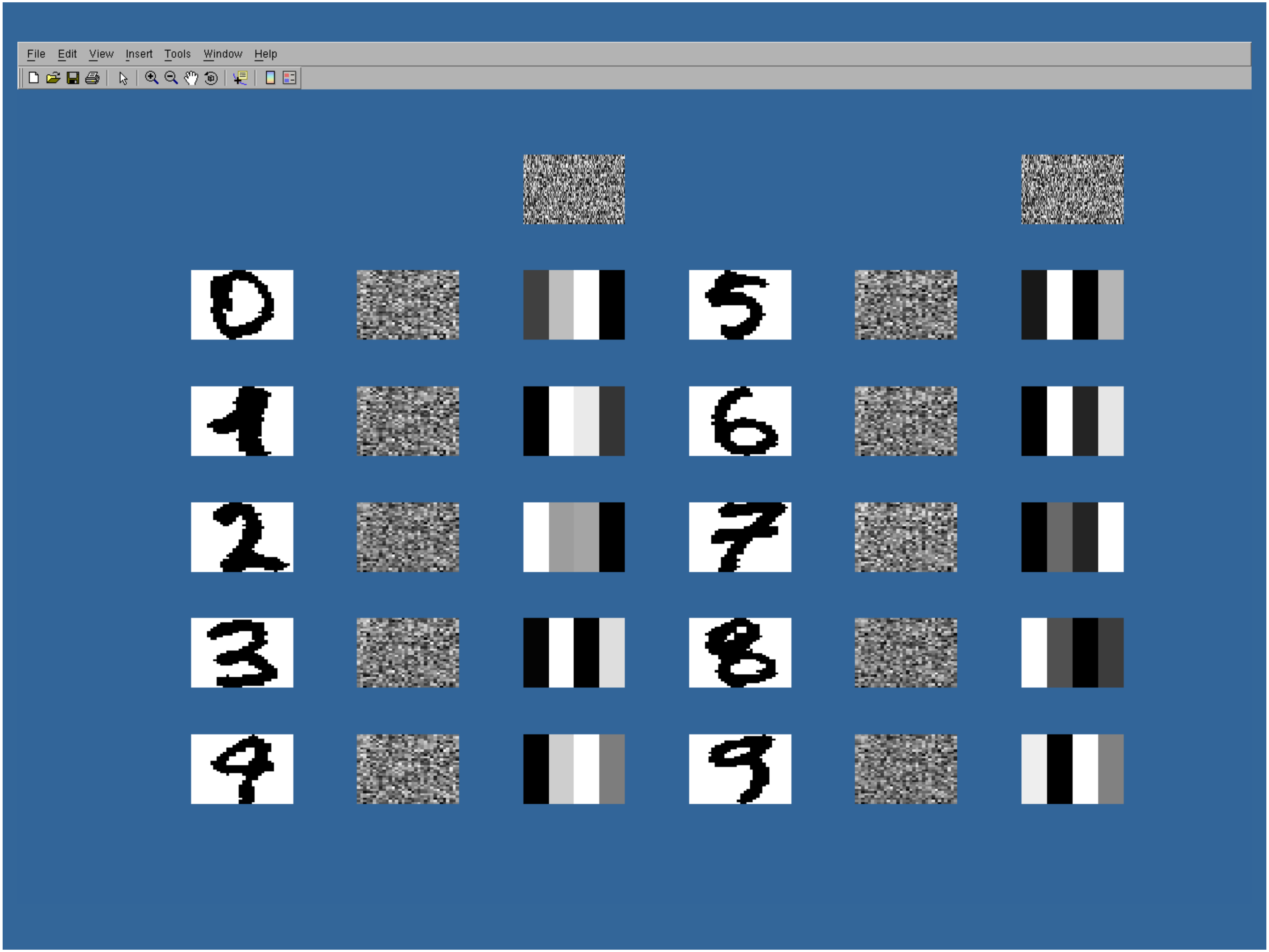




Example: Handwritten digits

$$V \approx WH \quad (1)$$




File Edit View Insert Tools Window Help

Summary

- NMF : $V \approx WH$
- Cost functions and multiplicative algorithms
  : Square of the Euclidean distance between A and B

$$H_{a\mu} \leftarrow H_{a\mu}\frac{(W^TV)_{a\mu}}{(W^TWH)_{a\mu}} \qquad W_{ia} \leftarrow W_{ia}\frac{(VH^T)_{ia}}{(WHH^T)_{ia}} \tag{4}$$

  : Generalized Kullback-Leibler divergence of A from B

$$H_{a\mu} \leftarrow H_{a\mu}\frac{\sum_i W_{ia}V_{i\mu}/(WH)_{i\mu}}{\sum_k W_{ka}} \qquad W_{ia} \leftarrow W_{ia}\frac{\sum_\mu H_{a\mu}V_{i\mu}/(WH)_{i\mu}}{\sum_\nu H_{a\nu}} \tag{5}$$

- Guaranteed monotonic convergence
- Interpretation as gradient descent