Solving Nonnegative Matrix Factorization using column-wise Block Coordinate Descent

HALS

Andersen Ang

Mathématique et recherche opérationnelle
UMONS, Belgium

manshun.ang@umons.ac.be Homepage: angms.science

First draft : August 2, 2017
Last update : May 10, 2020

(Block) Coordinate Descent

- Consider solving the minimization problem with vector variable $\mathbf{x} \in \mathbb{R}^n$:

$$\min_{\mathbf{x}} f(\mathbf{x}),$$

using iteratie algorithm.

- The "full update" is an update that act on the whole vector $\mathbf{x}$ in each iteration as :

$$\mathbf{x} \leftarrow \operatorname*{argmin}_{\mathbf{x}} f(\mathbf{x}).$$

- In Block coordinate descent : $\mathbf{x}$ is updated partially.
  For example : let $\mathbf{x}_{-i}$ be the vector $\mathbf{x}$ without the i-th component.
  Then BCD updates $\mathbf{x}$ in the following manner :
  - Update x_1 : $x_1 \leftarrow \arg\min_{x_1} f(x_1; \mathbf{x}_{-1})$
  - Update x_2 : $x_2 \leftarrow \arg\min_{x_2} f(x_2; \mathbf{x}_{-2})$
  - and so on for other i

The "coordinate" in BCD

- For problem with vector varaible $\mathbf{x}$, coordinate can be defined as
  1. element in the vector : x_i
  2. a subvector : a block of elements $x_{i_1}, x_{i_2}, \ldots, x_{i_n}$

- For problem with matrix varaible $\mathbf{X} \in \mathbb{R}^{m \times n}$, coordinate can be defined as
  1. element in the matrix X_{ij}
  2. rows or columns in the matrix
  3. submatrices of $\mathbf{X}$ the matrix

Coordinates in NMF

- Consider the Frobenius norm NMF minimization problem

$$\min_{\mathbf{W}\geq 0,\mathbf{H}\geq 0}\left\{f(\mathbf{W},\mathbf{H}) := \frac{1}{2}\|\mathbf{M}-\mathbf{W}\mathbf{H}\|_F^2\right\},$$

  where $\mathbf{M}\in\mathbb{R}^{m\times n}$, $\mathbf{W}\in\mathbb{R}^{m\times r}$ and $\mathbf{H}\in\mathbb{R}^{r\times n}$.

- Notation : let $\mathbf{w}_j$ be the j-th column of $\mathbf{W}$ and $\mathbf{h}^j$ be the j-th row of $\mathbf{H}$.

- By considering $\mathbf{w}_j, \mathbf{h}^j$ as coordinate, we now express f in terms of $\mathbf{w}_j$ and $\mathbf{h}^j$.

Dyadic expression of the NMF cost function

- In dyadic expression[1] we have

$$\mathbf{M} - \mathbf{WH} = \mathbf{M} - \sum_{i=1}^{r} \mathbf{w}_i \otimes \mathbf{h}^i.$$

  where $\mathbf{w}^{(i)} \otimes \mathbf{h}^{(i)}$ is the tensor product.

- Using the notation $\mathbf{w}_i\mathbf{h}^i$ to denote the tensor product, we have

$$\mathbf{M} - \mathbf{WH} = \mathbf{M} - \sum_{i=1}^{r} \mathbf{w}_i\mathbf{h}^i.$$

- Note that each dyad $\mathbf{w}_i\mathbf{h}^i$ is a rank-1 matrix, so $\sum_{i=1}^{r} \mathbf{w}_i\mathbf{h}^i$ is an rank-r matrix approximation of $\mathbf{M}$, and $\mathbf{M} - \sum_{i=1}^{r} \mathbf{w}_i\mathbf{h}^i$ is the residue between $\mathbf{M}$ and the rank-r approximation.

[1] See here for details.

Dyadic expression of the NMF cost function

- To express f in terms of $\mathbf{w}_j$ and $\mathbf{h}^j$, we have

$$\begin{aligned}
\mathbf{M} - \mathbf{WH} &= \mathbf{M} - \sum_{i=1}^{r} \mathbf{w}_i \mathbf{h}^i \\
&= \underbrace{\mathbf{M} - \sum_{i \neq j} \mathbf{w}_i \mathbf{h}^i}_{:=\mathbf{M}_j} - \mathbf{w}_j \mathbf{h}^j \\
&= \mathbf{M}_j - \mathbf{w}_j \mathbf{h}^j \\
\|\mathbf{M} - \mathbf{WH}\| &= \|\mathbf{M}_j - \mathbf{w}_j \mathbf{h}^j\|.
\end{aligned}$$

- Hence we have

$$f(\mathbf{w}_j, \mathbf{h}^j) = \frac{1}{2} \|\mathbf{M}_i - \mathbf{w}_j \mathbf{h}^j\|_F^2.$$

  Now we expand f in terms of $\mathbf{w}_j$ and $\mathbf{h}^j$.

Some algebra ... (1/2)

- Recall : by definition of F-norm we have $\|\mathbf{A}\|_F^2 = \mathrm{Tr}\,(\mathbf{A}^\top \mathbf{A})$, so

$$\begin{aligned}
& \|\mathbf{M}_j - \mathbf{w}_j \mathbf{h}^j\|_F^2 \\
= \; & \mathrm{Tr}\left((\mathbf{M}_j - \mathbf{w}_j \mathbf{h}^j)^\top (\mathbf{M}_j - \mathbf{w}_j \mathbf{h}^j)\right) \\
= \; & \mathrm{Tr}\left(\mathbf{M}_j^\top \mathbf{M}_i - \mathbf{M}_j^\top \mathbf{w}_j \mathbf{h}^j - (\mathbf{w}_j \mathbf{h}^j)^\top \mathbf{M}_i + (\mathbf{w}_j \mathbf{h}^j)^\top (\mathbf{w}_j \mathbf{h}^j)\right) \\
= \; & \mathrm{Tr}\left(\mathbf{M}_j^\top \mathbf{M}_j\right) - \mathrm{Tr}\left(\mathbf{M}_j^\top \mathbf{w}_j \mathbf{h}^j\right) - \mathrm{Tr}\left((\mathbf{w}_j \mathbf{h}^j)^\top \mathbf{M}_j\right) + \mathrm{Tr}\left((\mathbf{w}_j \mathbf{h}^j)^\top (\mathbf{w}_j \mathbf{h}^j)\right)\right) \\
= \; & \|\mathbf{M}_j\|_F^2 - \mathrm{Tr}\left(\mathbf{M}_j^\top \mathbf{w}_j \mathbf{h}^j\right) - \mathrm{Tr}\left((\mathbf{w}_j \mathbf{h}^j)^\top \mathbf{M}_j\right) + \mathrm{Tr}\left((\mathbf{w}_j \mathbf{h}^j)^\top (\mathbf{w}_j \mathbf{h}^j)\right)\right)
\end{aligned}$$

- As $\mathrm{Tr}\,(\mathbf{A}) = \mathrm{Tr}\,(\mathbf{A}^\top)$, we have

$$\|\mathbf{M}_j - \mathbf{w}_j \mathbf{h}^j\|_F^2 = \|\mathbf{M}_j\|_F^2 - 2\mathrm{Tr}\left(\mathbf{M}_j^\top \mathbf{w}_j \mathbf{h}^j\right) + \mathrm{Tr}\left((\mathbf{w}_j \mathbf{h}^j)^\top (\mathbf{w}_j \mathbf{h}^j)\right)\Big).$$

- For the last term, we have

$$(\mathbf{w}_j \mathbf{h}^j)^\top (\mathbf{w}_j \mathbf{h}^j) = {\mathbf{h}^j}^\top \mathbf{w}_j^\top \mathbf{w}_j \mathbf{h}^j.$$

  By associativity of product, we can perform multiplication in any order, so we first perform the product $\mathbf{w}_j^\top \mathbf{w}_j$, which gives a scalar, it can be moved out from the trace and gives

$$\|\mathbf{M}_j - \mathbf{w}_j \mathbf{h}^j\|_F^2 = \|\mathbf{M}_j\|_F^2 - 2\mathrm{Tr}\left(\mathbf{M}_j^\top \mathbf{w}_j \mathbf{h}^j\right) + \|\mathbf{w}_j\|_2^2 \mathrm{Tr}\left({\mathbf{h}^j}^\top \mathbf{h}^j\right).$$

Some algebra ... (2/2)

- **Lemma** For two column vectors $\mathbf{a}, \mathbf{b}$, we have $\mathrm{Tr}\left(\mathbf{a}\mathbf{b}^\top\right) = \mathbf{b}^\top\mathbf{a}$.
  Simple proof: $\mathrm{Tr}\left(\mathbf{a}\mathbf{b}^\top\right) = \sum_{i=1}^{n} a_i b_i = \mathbf{b}^\top\mathbf{a}$. □

- Using this equality, and note that $\mathbf{h}^j$ is a row vector, we have

$$\mathrm{Tr}\left(\mathbf{h}^{j\top}\mathbf{h}^j\right) = \|\mathbf{h}^j\|_2^2$$

  and, by treating $\mathbf{M}_j^\top\mathbf{w}_j$ as $\mathbf{a}$ and $\mathbf{h}^j$ as $\mathbf{b}^\top$, we have

$$\mathrm{Tr}\left(\mathbf{M}_j^\top\mathbf{w}_j\mathbf{h}^j\right) = \mathbf{h}^j\mathbf{M}_j^\top\mathbf{w}_j$$

  By $\langle\mathbf{A}\mathbf{x}, \mathbf{y}\rangle = \langle\mathbf{x}, \mathbf{A}^\top\mathbf{y}\rangle$

$$\mathbf{h}^j\mathbf{M}_j^\top\mathbf{w}_j = \langle\mathbf{M}_j\mathbf{h}^{j\top}, \mathbf{w}_j\rangle = \langle\mathbf{M}_j^\top\mathbf{w}_j, \mathbf{h}^{j\top}\rangle = \langle\mathbf{w}_j^\top\mathbf{M}_j, \mathbf{h}^j\rangle$$

- Finally, we have

$$\begin{aligned}\|\mathbf{M}_j - \mathbf{w}_j\mathbf{h}^j\|_F^2 &= \|\mathbf{M}_j\|_F^2 - 2\langle\mathbf{M}_j\mathbf{h}^{j\top}, \mathbf{w}_j\rangle + \|\mathbf{h}^j\|_2^2\|\mathbf{w}_j\|_2^2 \\ &= \|\mathbf{M}_j\|_F^2 - 2\langle\mathbf{w}_j^\top\mathbf{M}_j, \mathbf{h}^j\rangle + \|\mathbf{w}_j\|_2^2\|\mathbf{h}^j\|_2^2\end{aligned}$$

Quadratic form expressions

- We now have

$$f(\mathbf{w}_j) = \frac{1}{2}\|\mathbf{h}^j\|_2^2\|\mathbf{w}_j\|_2^2 - \langle \mathbf{M}_j\mathbf{h}^{j\top}, \mathbf{w}_j\rangle + \frac{1}{2}\|\mathbf{M}_j\|_F^2$$

$$f(\mathbf{h}^j) = \frac{1}{2}\|\mathbf{w}_j\|_2^2\|\mathbf{h}^j\|_2^2 - \langle \mathbf{w}_j^\top\mathbf{M}_j, \mathbf{h}^j\rangle + \frac{1}{2}\|\mathbf{M}_j\|_F^2$$

- The gradients and the corresponding Lipschitz constants are

$$\nabla_{\mathbf{w}} f(\mathbf{w}_j) = \|\mathbf{h}^j\|_2^2\mathbf{w}_j - \mathbf{M}_j\mathbf{h}^{j\top}, \quad L(\mathbf{w}_j) = \|\mathbf{h}^j\|_2^2$$

$$\nabla_{\mathbf{h}} f(\mathbf{h}^j) = \|\mathbf{w}_j\|_2^2\mathbf{h}^j - \mathbf{w}_j^\top\mathbf{M}_j, \quad L(\mathbf{h}^j) = \|\mathbf{w}_j\|_2^2$$

- The gradient update steps are

$$\mathbf{w}_j = \mathbf{w}_j - \frac{1}{L}\nabla_{\mathbf{w}} f(\mathbf{w}_j) = \mathbf{w}_j - \frac{\|\mathbf{h}^j\|_2^2\mathbf{w}_j - \mathbf{M}_j\mathbf{h}^{j\top}}{\|\mathbf{h}^j\|_2^2} = \frac{\mathbf{M}_j\mathbf{h}^{j\top}}{\|\mathbf{h}^j\|_2^2}$$

$$\mathbf{h}^j = \mathbf{h}^j - \frac{1}{L}\nabla_{\mathbf{h}} f(\mathbf{h}^j) = \mathbf{h}^j - \frac{\|\mathbf{w}_j\|_2^2\mathbf{h}^j - \mathbf{w}_j^\top\mathbf{M}_j}{\|\mathbf{w}_j\|_2^2} = \frac{\mathbf{w}_j^\top\mathbf{M}_j}{\|\mathbf{w}_j\|_2^2}$$

The update

- The update

$$\mathbf{w}_j = \frac{\mathbf{M}_j \mathbf{h}^{j\top}}{\|\mathbf{h}^j\|_2^2}, \qquad \mathbf{h}^j = \frac{\mathbf{w}_j^\top \mathbf{M}_j}{\|\mathbf{w}_j\|_2^2}$$

  are
  - the gradient update, and also the Newton's step on minimizing $f(\mathbf{w}_j)$ and $f(\mathbf{h}^j)$ (without constraint)
  - the exact solution to $f(\mathbf{w}_j)$ and $f(\mathbf{h}^j)$ (by takeing $\nabla f = 0$)

- With nonnegativity, we can just place the projection on top of the updates, that is

$$\mathbf{w}_j = \frac{[\mathbf{M}_j \mathbf{h}^{j\top}]_+}{\|\mathbf{h}^j\|_2^2}, \qquad \mathbf{h}^j = \frac{[\mathbf{w}_j^\top \mathbf{M}_j]_+}{\|\mathbf{w}_j\|_2^2}$$

  where $[\cdot]_+ = \max\{0, \cdot\}$.

- In fact, $\dfrac{[\mathbf{A}^\top \mathbf{b}]_+}{\|\mathbf{b}\|_2^2}$ is the optimal solution to the problem $\min_{\mathbf{x}\geq 0} \|\mathbf{b}\mathbf{x}^\top - \mathbf{A}\|_F^2$. See here for the proof.

Hierarchical Alternating Least Squares

This vector-wise exact BCD algorithm is called HALS in the literature.

Algorithm 1: Basic form of HALS

Input : $\mathbf{M} \in \mathbb{R}^{m \times n}$, r Initialize $\mathbf{W}_0, \mathbf{H}_0$;

for $k = 1, 2, \ldots$ **do**

1. Pick the index of the update component. e.g. using cyclic ordering

$$j = \mathrm{mod}(k-1, r) + 1$$

2. Pick $\mathbf{w}_j$ as the j-th column of $\mathbf{W}$ and $\mathbf{h}^j$ as j-th row of $\mathbf{H}$

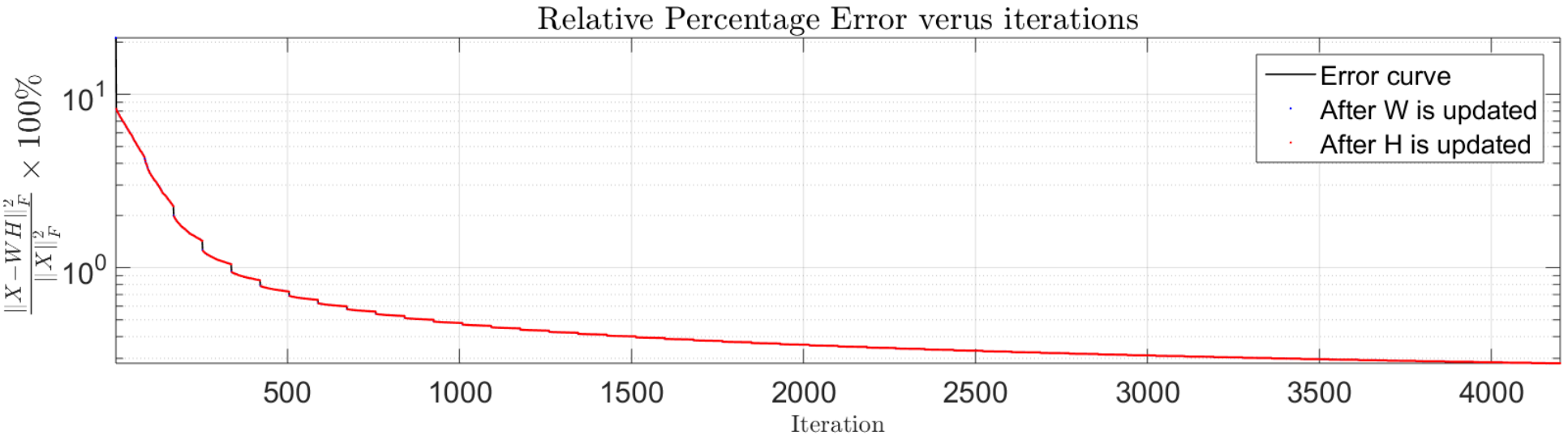
3. Compute $\mathbf{M}_j$, perform update : $\mathbf{w}_j = \dfrac{\left[\mathbf{M}_j \mathbf{h}^j\right]_+}{\|\mathbf{h}^j\|_2^2}$, $\mathbf{h}^j = \dfrac{\left[\mathbf{w}_j^\top \mathbf{M}_j\right]_+}{\|\mathbf{w}_j\|_2^2}$

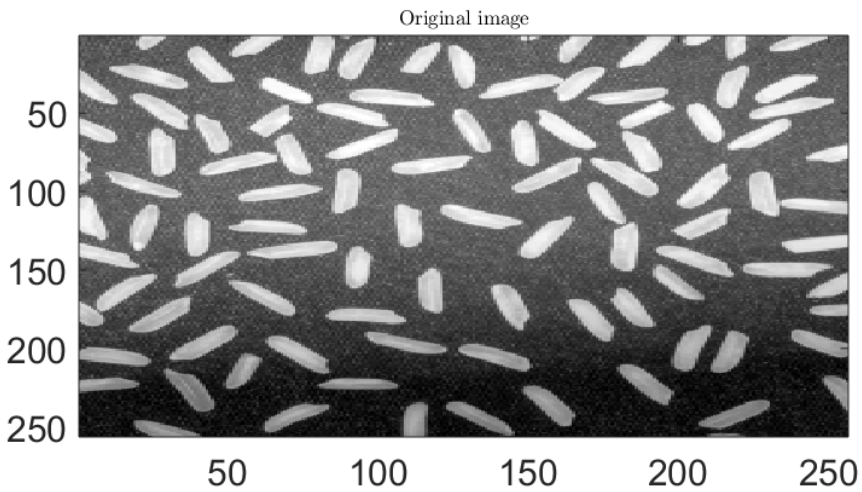
end

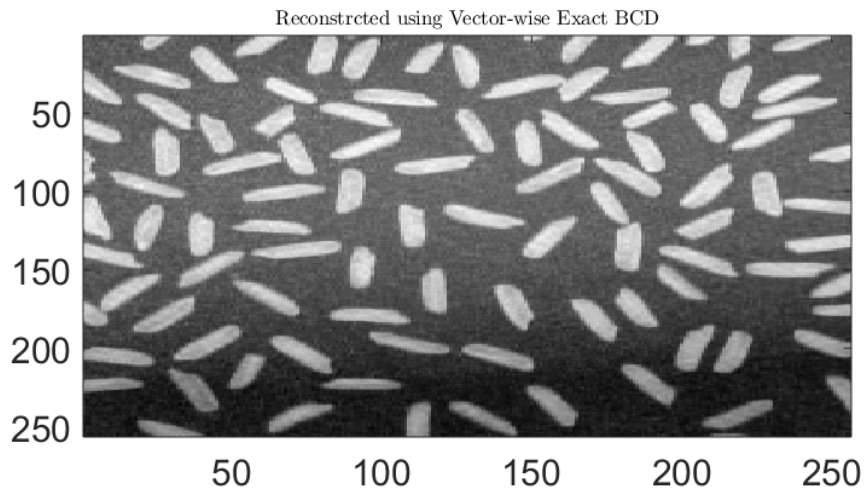
MATLAB code

There are various ways to accelerate HALS.

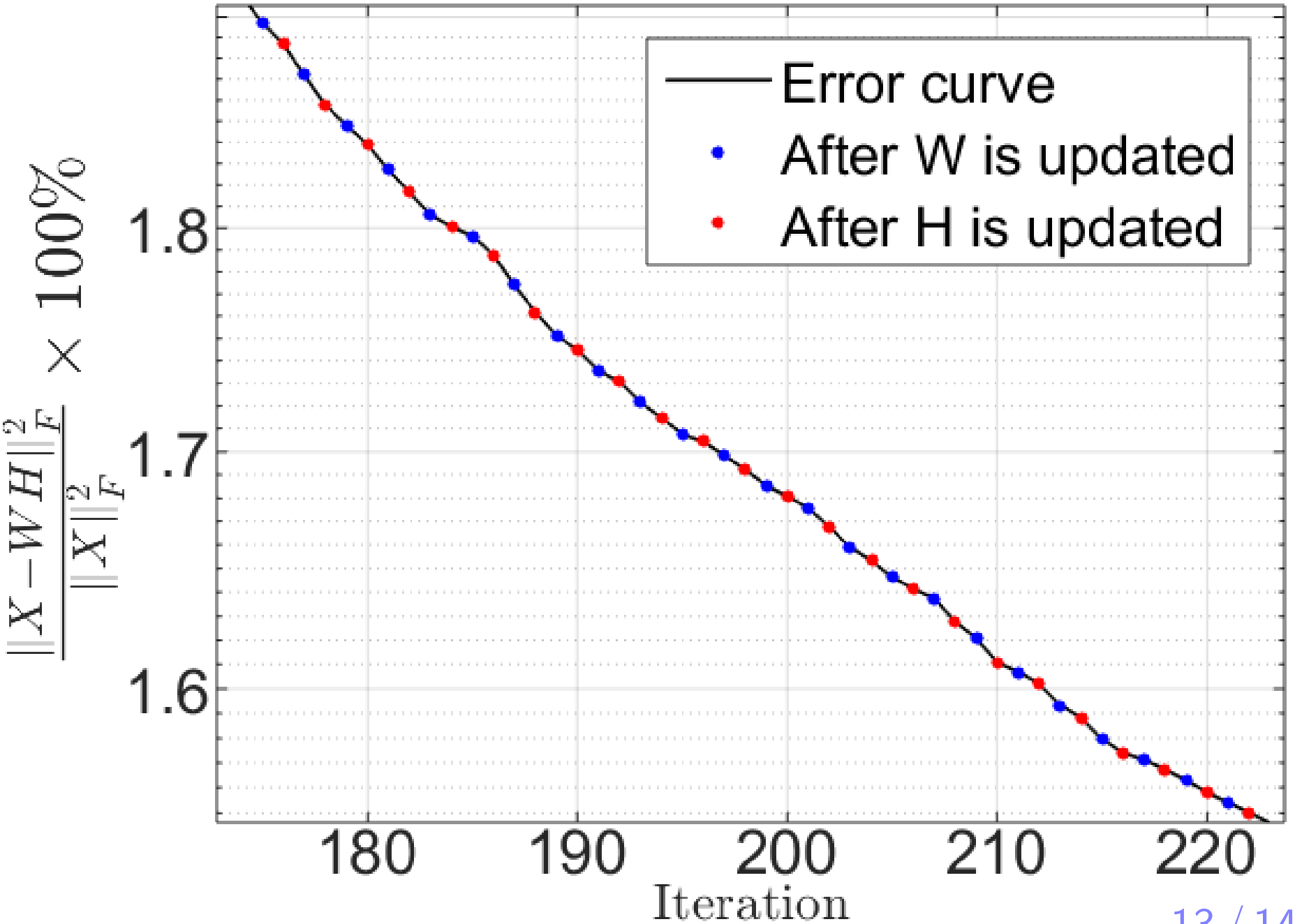
Example : MATLAB rice image $(m, n, r) = (256, 256, 84)$







Zoom in the error curve



Last page - summary

1. $f = \frac{1}{2}\|\mathbf{M} - \mathbf{WH}\|_F^2 = \frac{1}{2}\|\mathbf{M}_i - \mathbf{w}_j\mathbf{h}^j\|_F^2$

2. Vector-wise coorindate expression of f

3. The optimal exact update of coordinate component :

$$\mathbf{w}_j = \frac{\left[\mathbf{M}_j\mathbf{h}^j\right]_+}{\|\mathbf{h}^j\|_2^2}, \qquad \mathbf{h}^j = \frac{\left[\mathbf{w}_j^\top\mathbf{M}_j\right]_+}{\|\mathbf{w}_j\|_2^2}$$

4. Basic HALS algorithm

End of document