Detailed Derivation of the Update Rules for the Contrastive-NMF (C-NMF)

Giorgia Cantisani,1 Slim Essid,1 Gaël Richard,1
1 LTCI, Télécom Paris, Institut Polytechnique de Paris, France

The cost function of the Contrastive-NMF (C-NMF) is formulated as:

\[
C(W, H) = \underbrace{D_{KL}(X|WH)}_{\text{audio factorization}} + \mu \|H\|_1 + \beta \|W\|_1 - \delta (\|H_a S_a^T\|_F^2 - \|H_u S_a^T\|_F^2)
\]

\[
W, H, S_a \geq 0
\]

\[
\|h_k\|_2 = 1, \|s_k\|_2 = 1.
\]

(1)

where \(X \in \mathbb{R}^{M \times N}_+\) is the magnitude spectrogram of the mixture, the columns of \(W \in \mathbb{R}^{M \times K}_+\) are interpreted as non-negative audio spectral patterns expected to correspond to different sources and the rows of \(H \in \mathbb{R}^{K \times N}_+\) as their activations. \(M\) represents the number of frequency bins, \(N\) the number short-time Fourier transform frames and \(K\) the number of spectral patterns.

Considering the mixture \(x(t)\) given by the linear mixing of a target source \(s_a(t)\) and some interference sources \(s_u(t)\), we can define \(H_a \in \mathbb{R}^{K_a \times N}_+\) as the activations of the target and \(H_u \in \mathbb{R}^{(K-K_a) \times N}_+\) as the activations of the interference sources. \(S_a \in \mathbb{R}^{K_a \times N}_+\) are the activations of the target source reconstructed from the EEG response of a subject who was listening to \(x(t)\) and focusing on \(s_a(t)\). \(K_a\) are the number of spectral patterns used to describe the target source. The rows of \(H\) and \(S_a\) \((h_k, s_k)\) respectively are normalized in order to minimize the effect of a scale mismatch between the modalities.

Multiplicative Update Rule

To derive the multiplicative update rules of Eq.1, one can compute the gradient of the cost \(\nabla C(\theta)\), split it into is negative and positive parts and build the rules as following [Lee and Seung, 2001]:

\[
\theta \leftarrow \theta \odot \frac{\nabla \theta^T C(\theta)}{\nabla \theta^T C(\theta)}
\]

(2)

Since the variables are \(\theta = \{W, H\}\), the update rules will be:

\[
H \leftarrow H \odot \frac{\nabla H^T C(W, H)}{\nabla H^T C(W, H)}
\]

(3)

\[
W \leftarrow W \odot \frac{\nabla W^T C(W, H)}{\nabla W^T C(W, H)}
\]

(4)
**Update rule for W**

Since the cost function is completely separable, we can compute the gradient for the KL divergence and for the sparsity constraint separately.

**KL Divergence**

\[
\frac{\partial D_{KL}(X|WH)}{\partial w_{ij}} = \frac{\partial}{\partial w_{ij}} \sum_{m=1}^{M} \sum_{n=1}^{N} (x_{mn} \log \frac{x_{mn}}{WH_{mn}} - x_{mn} + WH_{mn}) = \\
= \sum_{m=1}^{M} \sum_{n=1}^{N} \frac{\partial}{\partial w_{ij}} (x_{mn} \log \frac{x_{mn}}{WH_{mn}}) + \sum_{m=1}^{M} \sum_{n=1}^{N} \frac{\partial}{\partial w_{ij}} (WH_{mn}) = \\
= \sum_{m=1}^{M} \sum_{n=1}^{N} x_{mn} \frac{\partial}{\partial w_{ij}} \log x_{mn} - \log WH_{mn} + \sum_{n=1}^{N} h_{jn} = \\
= \sum_{m=1}^{M} \sum_{n=1}^{N} x_{mn} \frac{\partial}{\partial w_{ij}} WH_{mn} + \sum_{n=1}^{N} h_{jn} = \\
= \sum_{n=1}^{N} \frac{-x_{in}}{WH_{in}} h_{jn} + \sum_{n=1}^{N} h_{jn} = \\
= [-(\Lambda^{-1} \otimes X)H^T + 1H^T]_{ij}
\]

(5)

**Sparsity**

\[
\frac{\partial \beta \|W\|_1}{\partial w_{ij}} = \frac{\partial}{\partial w_{ij}} \beta \sum_{m=1}^{M} \sum_{k=1}^{K} w_{mk} = \beta \frac{\partial}{\partial w_{ij}} w_{ij} = \beta
\]

(6)

**Update rule**

\[
W \leftarrow W \otimes \nabla_{W^{-C}(W,H)} = W \otimes \left( (\Lambda^{-1} \otimes X)H^T + 1H^T + \beta \right)
\]

(7)
**Update rule for** $H$

As for $W$, we can compute the gradient for the KL divergence, the sparsity constraint and for the margin term separately.

**KL divergence**

$$\frac{\partial \mathcal{D}_{KL}(X | WH)}{\partial h_{ij}} = \frac{\partial}{\partial h_{ij}} \sum_{m=1}^{M} \sum_{n=1}^{N} (x_{mn} \log \frac{x_{mn}}{WH_{mn}} - x_{mn} + WH_{mn}) =$$

$$= \sum_{m=1}^{M} \sum_{n=1}^{N} \frac{\partial}{\partial h_{ij}} (x_{mn} \log \frac{x_{mn}}{WH_{mn}}) + \sum_{m=1}^{M} \frac{\partial}{\partial h_{ij}} (WH_{mn}) =$$

$$= \sum_{m=1}^{M} \sum_{n=1}^{N} \frac{\partial}{\partial h_{ij}} (x_{mn} \log x_{mn} - \log WH_{mn}) + \sum_{m=1}^{M} \frac{\partial}{\partial h_{ij}} x_{mn} (\log WH_{mn}) + \sum_{m=1}^{M} w_{mi} =$$

$$= \sum_{m=1}^{M} \sum_{n=1}^{N} \frac{\partial}{\partial h_{ij}} x_{mn} (\log x_{mn} - \log WH_{mn}) + \sum_{m=1}^{M} \frac{\partial}{\partial h_{ij}} x_{mn} \log WH_{mn} + \sum_{m=1}^{M} w_{mi} =$$

$$= \sum_{m=1}^{M} \sum_{n=1}^{N} \frac{\partial}{\partial h_{ij}} x_{mn} \log x_{mn} - \sum_{m=1}^{M} \frac{\partial}{\partial h_{ij}} \log WH_{mn} + \sum_{m=1}^{M} \frac{\partial}{\partial h_{ij}} x_{mn} \log WH_{mn} + \sum_{m=1}^{M} w_{mi} =$$

$$= \sum_{m=1}^{M} \sum_{n=1}^{N} \frac{\partial}{\partial h_{ij}} x_{mn} \log x_{mn} - \sum_{m=1}^{M} \frac{\partial}{\partial h_{ij}} \log WH_{mn} + \sum_{m=1}^{M} \frac{\partial}{\partial h_{ij}} x_{mn} \log WH_{mn} + \sum_{m=1}^{M} w_{mi} =$$

$$= \left[-W^{T}(X \otimes \Lambda^{-1}) + W^{T}\Lambda\right]_{ij}$$

(8)

**Sparsity constrain**

$$\frac{\partial \mu \|H\|_{1}}{\partial h_{ij}} = \frac{\partial}{\partial h_{ij}} \mu \sum_{k=1}^{K} \sum_{n=1}^{N} h_{kn} = \mu \frac{\partial}{\partial h_{ij}} h_{ij} = \mu$$

(9)

**Contrast term**

Recall that the Frobenius norm can be rewritten as:

$$\|X\|_{F}^2 = \sum_{i=1}^{M} \sum_{j=1}^{N} x_{ij}^2 = \sqrt{\text{tr}(X^{T}X)}$$

(10)

Since $H_{a}S_{a}^{T}$ and $H_{u}S_{u}^{T}$ are square matrices, we have:

$$\|H_{a}S_{a}^{T}\|_{F}^2 = \text{tr}[(H_{a}S_{a}^{T})^{T}(H_{a}S_{a}^{T})] = \text{tr}[S_{a}H_{a}^{T}H_{a}S_{a}^{T}]$$

(11)

$$\|H_{u}S_{u}^{T}\|_{F}^2 = \text{tr}[(H_{u}S_{u}^{T})^{T}(H_{u}S_{u}^{T})] = \text{tr}[S_{u}H_{u}^{T}H_{u}S_{u}^{T}]$$

(12)

The gradient with respect to $H$, will be equal to the gradient computed with respect to $H_{a}$ for the first $K_{a}$ rows of $H$ and equal to the
gradient computed with respect to $H_u$ for the remaining rows:

$$
\nabla H(-\delta(\|H_u S_a^T\|^2_F - \|H_u S_a^T\|^2_F)) = \begin{cases} 
-\delta \nabla H_u (\|H_u S_a^T\|^2_F - \|H_u S_a^T\|^2_F), & \text{if } 1 < k < K_a \\
-\delta \nabla H_u (\|H_u S_a^T\|^2_F - \|H_u S_a^T\|^2_F), & \text{if } K_a + 1 < k < K
\end{cases}
$$

(13)

$$
\nabla H_u (\|H_u S_a^T\|^2_F - \|H_u S_a^T\|^2_F) = \nabla H_u \|H_u S_a^T\|^2_F = \nabla H_u \tr[S_a H_u^T H_u S_a^T] = \nabla H_u (H_u(S_a^T S_a) + H_u(S_a^T S_a)^T) = 2H_u S_a^T S_a
$$

(14)

Thus, we have:

$$
\nabla H(-\delta(\|H_u S_a^T\|^2_F - \|H_u S_a^T\|^2_F)) = \begin{cases} 
-2\delta H_u S_a^T S_a, & \text{if } 1 < k < K_a \\
2\delta H_u S_a^T S_a, & \text{if } K_a + 1 < k < K
\end{cases}
$$

(16)

**Update Rule**

$$
H \leftarrow H \otimes \frac{\nabla H \cdot C(W, H)}{\nabla H \cdot C(W, H) = H \otimes \frac{W^T (X \otimes \Lambda^{-1}) + \delta P}{W^T + \mu + \delta P^+}}
$$

(17)

where $P^-, P^+ \in \mathbb{R}^{K \times N}$ are auxiliary matrices defined as:

$$
P^- = \begin{cases} 
H_u S_a^T S_a, & \text{if } 1 < k < K_a \\
0, & \text{if } K_a + 1 < k < K
\end{cases}
$$

(18)

$$
P^+ = \begin{cases} 
0, & \text{if } 1 < k < K_a \\
H_u S_a^T S_a, & \text{if } K_a + 1 < k < K
\end{cases}
$$

(19)

**References**