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:

$$\begin{cases} C(\mathbf{W}, \mathbf{H}) = \underbrace{D_{KL}(\mathbf{X}|\mathbf{W}\mathbf{H})}_{\text{audio factorization}} + \underbrace{\mu\|\mathbf{H}\|_1 + \beta\|\mathbf{W}\|_1}_{\text{sparsity}} - \underbrace{\delta(\|\mathbf{H}_a\mathbf{S}_a^T\|_F^2 - \|\mathbf{H}_u\mathbf{S}_a^T\|_F^2)}_{\text{contrast}} \\ \mathbf{W}, \mathbf{H}, \mathbf{S}_a \geq 0 \\ \|\mathbf{h}_{k:}\|_2 = 1, \|\mathbf{s}_{k:}\|_2 = 1. \end{cases} \tag{1}$$

where $\mathbf{X} \in \mathbb{R}_+^{M\times N}$ is the magnitude spectrogram of the mixture, the columns of $\mathbf{W} \in \mathbb{R}_+^{M\times K}$ are interpreted as non-negative audio spectral patterns expected to correspond to different sources and the rows of $\mathbf{H} \in \mathbb{R}_+^{K\times N}$ 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 $\mathbf{H}_a \in \mathbb{R}_+^{K_a\times N}$ as the activations of the target and $\mathbf{H}_u \in \mathbb{R}_+^{(K-K_a)\times N}$ as the activations of the interference sources. $\mathbf{S}_a \in \mathbb{R}_+^{K_a\times N}$ 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 $\mathbf{H}$ and $\mathbf{S}_a$ ($\mathbf{h}_{k:}$ and $\mathbf{s}_{k:}$ 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 $\bigtriangledown C(\theta)$, split it into is negative and positive parts and build the rules as following [Lee and Seung, 2001]:

$$\theta \leftarrow \theta \otimes \frac{\bigtriangledown_{\theta^-} C(\theta)}{\bigtriangledown_{\theta^+} C(\theta)} \tag{2}$$

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

$$\mathbf{H} \leftarrow \mathbf{H} \otimes \frac{\bigtriangledown_{\mathbf{H}^-} C(\mathbf{W}, \mathbf{H})}{\bigtriangledown_{\mathbf{H}^+} C(\mathbf{W}, \mathbf{H})} \tag{3}$$

$$\mathbf{W} \leftarrow \mathbf{W} \otimes \frac{\bigtriangledown_{\mathbf{W}^-} C(\mathbf{W}, \mathbf{H})}{\bigtriangledown_{\mathbf{W}^+} C(\mathbf{W}, \mathbf{H})} \tag{4}$$

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

$$
\begin{aligned}
\frac{\partial D_{KL}(\mathbf{X}|\mathbf{WH})}{\partial w_{ij}} &= \frac{\partial}{\partial w_{ij}} \sum_{m=1}^{M} \sum_{n=1}^{N} (x_{mn} \log \frac{x_{mn}}{\mathbf{WH}|_{mn}} - x_{mn} + \mathbf{WH}|_{mn}) = \\
&= \sum_{m=1}^{M} \sum_{n=1}^{N} \frac{\partial}{\partial w_{ij}} (x_{mn} \log \frac{x_{mn}}{\mathbf{WH}|_{mn}}) + \sum_{m=1}^{M} \sum_{n=1}^{N} \frac{\partial}{\partial w_{ij}} (\mathbf{WH}|_{mn}) = \\
&= \sum_{m=1}^{M} \sum_{n=1}^{N} \frac{\partial}{\partial w_{ij}} x_{mn} (\log x_{mn} - log \mathbf{WH}|_{mn}) + \sum_{n=1}^{N} h_{jn} = \\
&= \sum_{m=1}^{M} \sum_{n=1}^{N} x_{mn} \frac{\partial}{\partial w_{ij}} (-\log \mathbf{WH}|_{mn}) + \sum_{n=1}^{N} h_{jn} = \\
&= \sum_{m=1}^{M} \sum_{n=1}^{N} \frac{-x_{mn}}{\mathbf{WH}|_{mn}} \frac{\partial}{\partial w_{ij}} (\mathbf{WH}|_{mn}) + \sum_{n=1}^{N} h_{jn} = \\
&= \sum_{n=1}^{N} \frac{-x_{in}}{\mathbf{WH}|_{in}} h_{jn} + \sum_{n=1}^{N} h_{jn} = \\
&= [-(\Lambda^{-1} \otimes \mathbf{X})\mathbf{H}^T + \mathbf{1}\mathbf{H}^T]|_{ij}
\end{aligned} \tag{5}
$$

- $D_{KL}(p, q) = p \log \frac{p}{q} - p + q$
- $\mathbf{WH}|_{mn} = \sum_k w_{mk} h_{kn}$
- derivative of the matrix product:

$$\frac{\partial}{\partial w_{ij}} \mathbf{WH}|_{mn} = \begin{cases} h_{jn} & \text{if } m = i \\ 0 & \text{if } m \neq i \end{cases}$$

- $\Lambda = \mathbf{WH}$

Sparsity

$$\frac{\partial \beta \|\mathbf{W}\|_1}{\partial w_{ij}} = \frac{\partial}{\partial w_{ij}} \beta \sum_{m=1}^{M} \sum_{k=1}^{K} w_{mk} = \beta \frac{\partial}{\partial w_{ij}} w_{ij} = \beta \tag{6}$$

Update rule

$$\boxed{\mathbf{W} \leftarrow \mathbf{W} \otimes \frac{\nabla_{\mathbf{W}^-} C(\mathbf{W}, \mathbf{H})}{\nabla_{\mathbf{W}^+} C(\mathbf{W}, \mathbf{H})} = \mathbf{W} \otimes \frac{(\Lambda^{-1} \otimes \mathbf{X})\mathbf{H}^T}{\mathbf{1}\mathbf{H}^T + \beta}} \tag{7}$$

Update rule for $\boldsymbol{H}$

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

KL divergence

$$
\begin{aligned}
\frac{\partial D_{KL}(\mathbf{X}|\mathbf{WH})}{\partial h_{ij}} &= \frac{\partial}{\partial h_{ij}} \sum_{m=1}^{M}\sum_{n=1}^{N}(x_{mn}\log\frac{x_{mn}}{\mathbf{WH}|_{mn}} - x_{mn} + \mathbf{WH}|_{mn}) = \\
&= \sum_{m=1}^{M}\sum_{n=1}^{N}\frac{\partial}{\partial h_{ij}}(x_{mn}\log\frac{x_{mn}}{\mathbf{WH}|_{mn}}) + \sum_{m=1}^{M}\sum_{n=1}^{N}\frac{\partial}{\partial h_{ij}}(\mathbf{WH}|_{mn}) = \\
&= \sum_{m=1}^{M}\sum_{n=1}^{N}\frac{\partial}{\partial h_{ij}}x_{mn}(\log x_{mn} - log\mathbf{WH}|_{mn}) + \sum_{m=1}^{M} w_{mi} = \\
&= \sum_{m=1}^{M}\sum_{n=1}^{N} x_{mn}\frac{\partial}{\partial h_{ij}}(-\log\mathbf{WH}|_{mn}) + \sum_{m=1}^{M} w_{mi} = \\
&= \sum_{m=1}^{M}\sum_{n=1}^{N}\frac{-x_{mn}}{\mathbf{WH}|_{mn}}\frac{\partial}{\partial h_{ij}}(\mathbf{WH}|_{mn}) + \sum_{m=1}^{M} w_{mi} = \\
&= \sum_{m=1}^{M}\frac{-x_{mj}}{\mathbf{WH}|_{mj}}w_{mi} + \sum_{m=1}^{M} w_{mi} = \\
&= [-\mathbf{W}^T(\mathbf{X}\otimes\Lambda^{-1}) + \mathbf{W}^T\mathbf{1}]|_{ij}
\end{aligned}
\tag{8}
$$

- $D_{KL}(p,q) = p\log\frac{p}{q} - p + q$
- $\mathbf{WH}|_{mn} = \sum_k w_{mk}h_{kn}$
- derivative of the matrix product:

$$\frac{\partial}{\partial h_{ij}}\mathbf{WH}|_{mn} = \begin{cases} w_{mi} & \text{if } n = j \\ 0 & \text{if } n \neq j \end{cases}$$

- $\Lambda = \mathbf{WH}$

Sparsity constrain

$$\frac{\partial \mu\|\mathbf{H}\|_1}{\partial h_{ij}} = \frac{\partial}{\partial h_{ij}}\mu\sum_{k=1}^{K}\sum_{n=1}^{N} h_{kn} = \mu\frac{\partial}{\partial h_{ij}}h_{ij} = \mu \tag{9}$$

Contrast term

Recall that the Frobenius norm can be rewritten as:

$$\|\mathbf{X}\|_F = \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N} x_{ij}^2} = \sqrt{tr(\mathbf{X}^T\mathbf{X})} \tag{10}$$

Since $\mathbf{H}_a\mathbf{S}_a^T$ and $\mathbf{H}_u\mathbf{S}_a^T$ are square matrices, we have:

$$\|\mathbf{H}_a\mathbf{S}_a^T\|_F^2 = tr[(\mathbf{H}_a\mathbf{S}_a^T)^T(\mathbf{H}_a\mathbf{S}_a^T)] = tr[\mathbf{S}_a\mathbf{H}_a^T\mathbf{H}_a\mathbf{S}_a^T] \tag{11}$$

$$\|\mathbf{H}_u\mathbf{S}_a^T\|_F^2 = tr[(\mathbf{H}_u\mathbf{S}_a^T)^T(\mathbf{H}_u\mathbf{S}_a^T)] = tr[\mathbf{S}_a\mathbf{H}_u^T\mathbf{H}_u\mathbf{S}_a^T] \tag{12}$$

The gradient with respect to $\mathbf{H}$, will be equal to the gradient computed with respect to $\mathbf{H}_a$ for the first K_a rows of $\mathbf{H}$ and equal to the gradient computed with respect to $\mathbf{H}_u$ for the remaining rows:

$$\nabla_{\mathbf{H}}(-\delta(\|\mathbf{H}_a\mathbf{S}_a^T\|_F^2 - \|\mathbf{H}_u\mathbf{S}_a^T\|_F^2)) = \begin{cases} -\delta \nabla_{\mathbf{H}_a} (\|\mathbf{H}_a\mathbf{S}_a^T\|_F^2 - \|\mathbf{H}_u\mathbf{S}_a^T\|_F^2), & \text{if } 1 < k < K_a \\ -\delta \nabla_{\mathbf{H}_u} (\|\mathbf{H}_a\mathbf{S}_a^T\|_F^2 - \|\mathbf{H}_u\mathbf{S}_a^T\|_F^2), & \text{if } K_a + 1 < k < K \end{cases} \tag{13}$$

$$\begin{aligned} \nabla_{\mathbf{H}_a}(\|\mathbf{H}_a\mathbf{S}_a^T\|_F^2 - \|\mathbf{H}_u\mathbf{S}_a^T\|_F^2)) = \nabla_{\mathbf{H}_a}\|\mathbf{H}_a\mathbf{S}_a^T\|_F^2 &= \nabla_{\mathbf{H}_a} tr[\mathbf{S}_a\mathbf{H}_a^T\mathbf{H}_a\mathbf{S}_a^T] = \\ &= \mathbf{H}_a(\mathbf{S}_a^T\mathbf{S}_a) + \mathbf{H}_a(\mathbf{S}_a^T\mathbf{S}_a)^T = \\ &= 2\mathbf{H}_a\mathbf{S}_a^T\mathbf{S}_a \end{aligned} \tag{14}$$

- $tr(\mathbf{ABC}) = tr(\mathbf{BCA}) = tr(\mathbf{CAB})$
- $\nabla_{\mathbf{X}} tr(\mathbf{XAX}^T) = \mathbf{X}(\mathbf{A}^T + \mathbf{A})$
- $(\mathbf{X}^T\mathbf{Y})^T = \mathbf{Y}^T\mathbf{X}$

$$\begin{aligned} \nabla_{\mathbf{H}_u}(\|\mathbf{H}_a\mathbf{S}_a^T\|_F^2 - \|\mathbf{H}_u\mathbf{S}_a^T\|_F^2)) = -\nabla_{\mathbf{H}_u}\|\mathbf{H}_u\mathbf{S}_a^T\|_F^2 &= -\nabla\, tr[\mathbf{S}_a\mathbf{H}_u^T\mathbf{H}_u\mathbf{S}_a^T] = \\ &= -(\mathbf{H}_u(\mathbf{S}_a^T\mathbf{S}_a) + \mathbf{H}_u(\mathbf{S}_a^T\mathbf{S}_a)^T) = \\ &= -2\mathbf{H}_u\mathbf{S}_a^T\mathbf{S}_a \end{aligned} \tag{15}$$

Thus, we have:

$$\nabla_{\mathbf{H}}(-\delta(\|\mathbf{H}_a\mathbf{S}_a^T\|_F^2 - \|\mathbf{H}_u\mathbf{S}_a^T\|_F^2)) = \begin{cases} -2\delta\mathbf{H}_a\mathbf{S}_a^T\mathbf{S}_a, & \text{if } 1 < k < K_a \\ +2\delta\mathbf{H}_u\mathbf{S}_a^T\mathbf{S}_a, & \text{if } K_a + 1 < k < K \end{cases} \tag{16}$$

Update Rule

$$\boxed{\mathbf{H} \leftarrow \mathbf{H} \otimes \frac{\nabla_{\mathbf{H}^-} C(\mathbf{W}, \mathbf{H})}{\nabla_{\mathbf{H}^+} C(\mathbf{W}, \mathbf{H})} = \mathbf{H} \otimes \frac{\mathbf{W}^T(\mathbf{X} \otimes \Lambda^{-1}) + \delta\mathbf{P}^-}{\mathbf{W}^T\mathbf{1} + \mu + \delta\mathbf{P}^+}} \tag{17}$$

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

$$\mathbf{P}^- = \begin{cases} \mathbf{H}_a\mathbf{S}_a^T\mathbf{S}_a, & \text{if } 1 < k < K_a \\ 0, & \text{if } K_a + 1 < k < K \end{cases} \tag{18}$$

$$\mathbf{P}^+ = \begin{cases} 0, & \text{if } 1 < k < K_a \\ \mathbf{H}_u\mathbf{S}_a^T\mathbf{S}_a, & \text{if } K_a + 1 < k < K \end{cases} \tag{19}$$

References

Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.