Block Nonnegative Matrix Factorization for Single Channel Source Separation

Minje Kim
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL 61801
minje@illinois.edu

1 Introduction

Nonnegative Matrix Factorization (NMF) [1, 2] has been widely used in audio research, e.g. automatic music transcription [3], musical source separation [4], and speech enhancement [5]. The key strategy for applying NMF to audio-related tasks is to find a lower rank representation of the Short Time Fourier Transformed (STFT) input signal and use the basis vectors as dictionaries. For example, in the single channel source separation, we assume that the dictionaries learned from different training sets of target sources are distinct so that their activations for reconstructing a mixture signal will have discriminant patterns per a source.

Although NMF provides an intuitive additive structure, i.e. weighted sum of basis vectors approximates the STFT matrix, that is suitable for audio analysis, its linear decomposition model can be limited in some sense, e.g. we do not learn a hierarchy of features from NMF. In this project I would like to elaborate the standard NMF model to have a deeper structure. At the same time, the unique additive reconstruction manner of NMF due to the nonnegativity constraint can be incorporated into the learning so that the features can have more parts-based representation.

To this end, I tried two different approaches: NMF with block sparsity and multi-layered auto-encoders. Although I got some source separation results that are comparable to the ones from the ordinary NMF with the newly derived nonnegative auto-encoders, there was no performance improvement for now. Therefore I would focus on the former approach in this report while setting aside the deep learning-based approach as an appendix with discussions on the future plan about it.

The proposed Block NMF (BNMF) is one of the recent attempts to preserve manifold of the audio signals that belong to a same source. In [6] an overcomplete dictionary model was proposed that fully makes use of the entire training samples (spectra) during the separation. Its behavior that respects manifold of the training data is because of the fact that it skips the NMF learning phase, which usually replace the training samples with convex cones (or hulls in the probabilistic versions). The convex cone representation is sometimes more useful than the original data as the cone can be defined a few basis vectors while it introduces unnecessary areas where no training data reside. Therefore, by skipping this step and learning sparse encoding of the entire training spectra as if they are the basis vectors, we can get the reconstructions of a source that apt to lie on the original manifold. However, it is burdensome to carry on the entire training data set for the source separation tasks that sometimes have to be done fast.

Another approach is to learn some important samples from the training data that preferably lie on the manifold instead of wrapping the data convexly[7]. If we apply the same sparse encoding aforementioned, or interpolate in-between those samples, we can effectively approximate the manifold of the data set with smaller number of data. Although this provides desired properties, it cannot efficiently represent unseen data points that are slightly off the interpolation line or the representative samples.

In this project I propose a new structured NMF where we can group the basis vectors into the ones that represent only local subset of data. A given data point is then reconstructed by using only one (or very small number) of group of bases using sample-wise block sparsity. On top of that I also add another regularization term that enforce each group as similar as possible.

After providing some introduction to existing methods, such as standard NMF in section 2.1, sparse encoding of overcomplete dictionary in section 2.2, and the manifold preserving quantization and interpolation in section 2.3, the theoretical background and derivation of the proposed method with toy examples are given in section 3. Section 4 is for crosstalk cancellation results. Appendix is for the deep learning part of the project that was set aside for further study.

2 Dictionary-based source separation using NMF

2.1 NMF with β-divergence [5][1]

NMF takes a nonnegative matrix $V \in \mathbb{R}_+^{M\times N}$ as input and tries to approximate it with a pair of factor matrices $W \in \mathbb{R}_+^{M\times R}$ and $H \in \mathbb{R}_+^{R\times N}$, where the set $\mathbb{R}_+$ stands for nonnegative real numbers, and R is for the number of latent components [1, 2]. A generalized way to measure the approximation error between the input V and the reconstruction $WH = \sum_{z=1}^{R} w_z h_z$ can be the β-divergence, which is defined by

$$\mathcal{D}_\beta(x|y) = \begin{cases} \frac{x^\beta + (\beta-1)y^\beta - \beta x y^{\beta-1}}{\beta(\beta-1)} & \beta \in \mathbb{R}\backslash\{0,1\} \\ x(\log x - \log y) + (y-x) & \beta = 1 \\ \frac{x}{y} - \log\frac{x}{y} - 1 & \beta = 0 \end{cases} \tag{1}$$

for any pair of elements x and y in the input and the reconstruction, respectively. Note that (1) reduces to Frobenius norm, unnormalized Kullback-Leibler divergence, and Itakura-Saito divergence [8] when β equals to 2, 1, and 0, respectively. Therefore, the objective function of NMF can be defined as follows:

$$\mathcal{J}_\beta = \mathcal{D}_\beta\Big(V\Big|WH\Big). \tag{2}$$

Using the fact that the derivative of the $\mathcal{D}_\beta(x|y)$ with respect to y is

$$\frac{\partial \mathcal{D}_\beta(x|y)}{\partial y} = y^{\beta-2}(y-x), \tag{3}$$

we can calculate the derivatives of the objective function (2) as follows:

$$\begin{aligned} \frac{\partial \mathcal{J}_\beta}{\partial W} &= \Big\{(WH)^{(\beta-2)} \odot (WH - V)\Big\}H^\top, \\ \frac{\partial \mathcal{J}_\beta}{\partial H} &= W^\top\Big\{(WH)^{(\beta-2)} \odot (WH - V)\Big\}, \end{aligned} \tag{4}$$

where $\odot$ is for Hadamard products and exponentiations are carried in the element-wise manner as well.

We can derive the multiplicative update rules of NMF by selecting the step size of the gradient descent method in such a way that it turns the update into a multiplicative form. An alternative view of this process is to simply choose the negative and positive terms of the derivative as the numerator and the denominator, respectively, which in turn produces following update rules:

$$\begin{aligned} W &\leftarrow W \odot \frac{\Big\{(WH)^{(\beta-2)} \odot V\Big\}H^\top}{(WH)^{(\beta-1)}H^\top}, \\ H &\leftarrow H \odot \frac{W^\top\Big\{(WH)^{(\beta-2)} \odot V\Big\}}{W^\top(WH)^{(\beta-1)}}. \end{aligned} \tag{5}$$

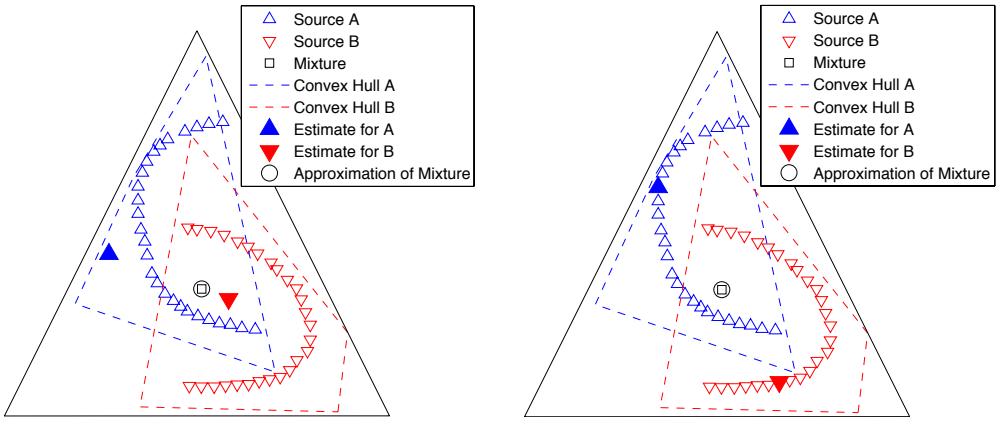
In the two sources case we learn basis vectors from clean training signals of the two sources separately. Suppose that they are $W^s \in \mathbb{R}_+^{M\times R^s}$ and $W^n \in \mathbb{R}_+^{M\times R^n}$, respectively. As for the unseen

[1] This clause is copied from the cited paper.

mixture STFT matrix $X = S^s + S^n$, we run another NMF, but with fixed bases $W = [W^s, W^n]$ this time. Therefore, encoding matrix H we get by using (5), but skipping the update for W, consists of two source groups, $H = [H^s; H^n]$, where ; stands for vertical matrix concatenation.

Now the separation is based on the grouped activation and bases. For the source S^s, for example, it can be recovered by using simple Wiener filter-like soft masking:

$$\hat{S}^s = \frac{W^s H^s}{WH} \odot X$$



(a) (b)

Figure 1: [7] Separation examples using (a) a plain topic model (b) sparse encoding

2.2 Sparse encoding of overcomplete dictionaries

Figure 1 (a) shows the separation results of Probabilistic Latent Semantic Indexing (PLSI) [9] that is equivalent to NMF with KL-divergence. After learning the red and blue convex hulls, it is usual to discard the training samples. Now the only way to decide whether the unseen data point belongs to source A or B is to see if either of the convex hull includes it. First, this is problematic since the hulls can overlap, so the filled red triangle can belong both sources. On top of that, if the unseen point is a mixture of the two unseen samples from the two sources, the infinite number of source reconstructions that can lie inside the hulls can be the solutions whether or not they are on the manifold.

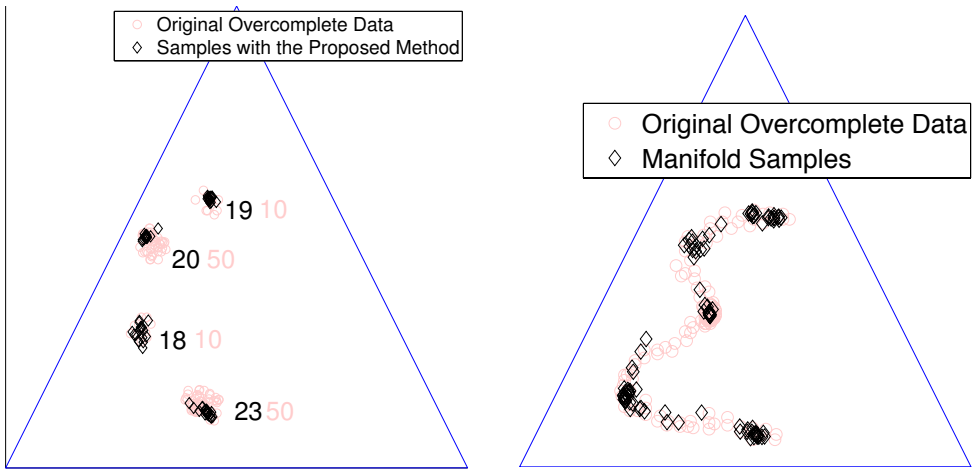
Figure 1 (b) is the sparse encoding of the overcomplete dictionaries. Now that we do not learn the hulls, but find a few active training samples, they can also reconstruct the mixture well and preserve the manifold at the same time.

To achieve this manifold preserving decomposition, elements in the overcomplete dictionary are sparsely encoded only small number of them are activated. The procedure is not so different from introducing sparsity over each column vector of H matrix except the fact that the method in [6] is based on PLSI.

Even if this method provides a way to respect the manifold of the training data, it is not very efficient to keep the entire training samples.

2.3 Manifold preserving quantization and interpolation

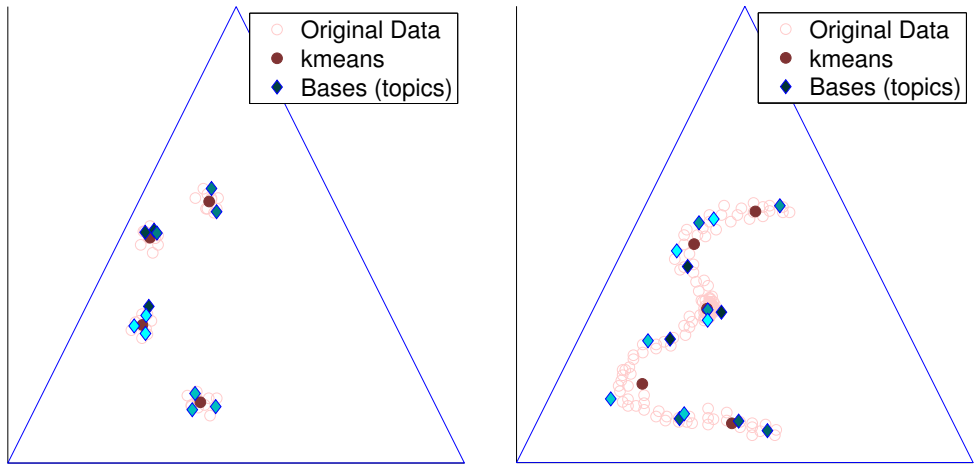
Figure 2 shows an alternative way to preserve the manifold. Instead of using the entire training samples or learning convex hulls. Figure 2 (a) is the case where the four clusters are not balanced (pink numbers are the number of samples in each cluster). The goal of this quantization is to learn four samples each of which represents a cluster. Black diamonds are four samples from 20 repeated experiments. In (b) we can also see that the five samples are efficiently represent five important areas of the manifold (once again, sampling was done 20 times).



(a) (b)

Figure 2: Learning manifold preserving samples from (a) four unbalanced clusters (b) epsilon shaped unbalanced dataset [7]

This manifold preserving quantization provides a multimodal structure that can theoretically deal with any nonlinear structure of the data. However, combined with interpolation, its expressive power is limited to the skeleton of the data manifold. What we actually want is a convex hull-based structure that is more flexible to the variance of unseen data, but also free from unwanted areas inside the hull.



(a) (b)

Figure 3: Learning manifold preserving local convex hulls from (a) four clusters (b) epsilon shaped unbalanced dataset. Diamonds with same colors are supposed to represent the same local structure (cluster).

3 Block NMF for manifold preserving hierarchical bases groups

The proposed method is a modified version of NMF to achieve locality preserving convex hulls. The basic concept that the method also tries to respect the local structure of the data is similar to the aforementioned manifold learning techniques, but now we can learn a few small convex hulls that correspond to clusters. I introduce the convex hull representation back, because it can construct small convex hulls for each underlying modality of data.

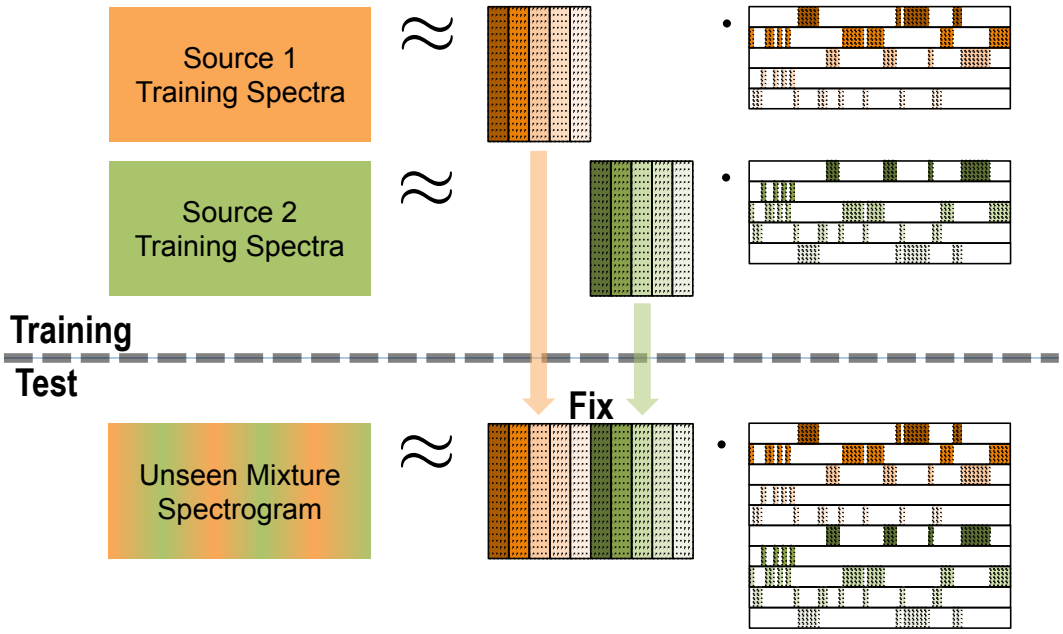


Figure 4: A block diagram for the full source separation procedure including source specific dictionary learning and its (block) sparse coding with the learned dictionaries.

Figure 3 presents some basis vectors that are learned from the proposed method. In (a) three bases (same colored diamonds) are allocated per a cluster. Because only one of the four sets are activated at a time, the original data points are exclusively approximated by those three-bases sets, not a big convex hull for all four clusters. In order for the bases within a set to be as close as possible to each other, I use K-means as the a prior for the bases. We can see the same effect of the proposed method in (b) as well. If we draw convex hulls each of which is defined by the same colored diamonds, we can see that those local convex hulls work as boundaries of the local subset of the data.

Figure 4 describes the source separation procedure using the proposed technique. For the separation task, we first have to learn basis vectors from each source (orange and green) that requires two distinct NMF runs in this example. The basis vectors for a source are grouped into the free defined number of sets, e.g. five bases per a group of five sets, and the H matrix is block-wise sparse. After learning those bases, we use them as a fixed dictionary for describing unseen mixtures while keeping the newly learned H matrix to be sparse as well.

The objective function of BNMF consists of three terms: the original reconstruction error, a penalty function for the block sparsity, and a prior for the bases concentration:

$$\mathcal{J} = \mathcal{D}(V|WH) + \lambda \sum_{t,g} \Omega(H_t^{(g)}) + \eta \sum_{g} \mathcal{D}(W^{(g)}|\mu^{(g)}). \tag{6}$$

The first term is the β-divergence in (1) from the original NMF algorithm. The function Ω on blocks g of each frame t is to give penalty to the solutions that are not sparse. For example, for a two dimensional variable, we can use a function that looks like Figure 5.

In particular, I used $\log / l_1$ penalty,

$$\Omega(H) = \sum_{t,g} \log(\epsilon + ||H_t^{(g)}||_1)$$

that was used in [10, 11], for its monotonicity and induced multiplicative updates.

The main difference between the universal speech model [11] and the proposed block NMF is that the former sets block sparsity on speakers. In other words, it selects relevant speech models in a global fashion, so the chosen one is always active while the proposed method selects the blocks

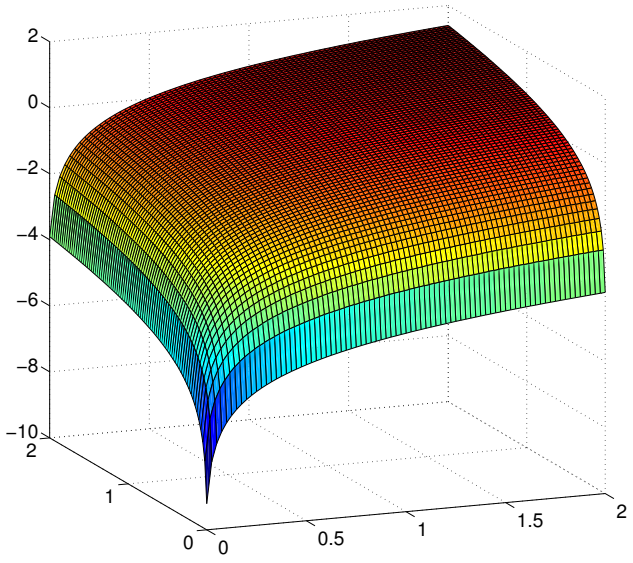


Figure 5: The penalty function used for the block sparsity. This function favors (has lower values) when one of the dimension is close to zero.

dynamically. Furthermore, the proposed model is not limited to correspond each bases group to a certain speaker.

For the data-driven way of learning those blocks, in the third term we can start from estimated clusters. For instance, we can make use of those manifold preserving samples [7] or simply Gaussian-based clustering. In this project, I used K-means algorithm to initialize those blocks. It provides a prior information about the basis vectors of a certain block that works like a Dirichlet prior in the simplex models, such as PLSI (red dots in Figure 3). For a given block g, the distance between bases vectors in the group and the corresponding mean $\mu^{(g)}$ is also minimized as an additional regularization.

We can derive multiplicative update rules for this new objective function as in the standard NMF case and the universal speech model, but on each frame of H as follows:

$$
\begin{aligned}
W^{(g)} &\leftarrow W^{(g)} \odot \frac{\Big\{(WH)^{(\beta-2)} \odot V\Big\} {H^{(g)}}^{\top} + \mu^{(g)}}{(WH)^{(\beta-1)} {H^{(g)}}^{\top}}, \\
H &\leftarrow H \odot \frac{W^{\top}\Big\{(WH)^{(\beta-2)} \odot V\Big\}}{W^{\top}(WH)^{(\beta-1)}} \\
H_t^{(g)} &\leftarrow H_t^{(g)} \odot \frac{1}{1 + \lambda/(\epsilon + ||H_t^{(g)}||_1)}
\end{aligned}
\tag{7}
$$

4 Experimental Results: Cross Talk Cancellation

For the experiment, I used two speakers from TIMIT corpus for training and testing. Nine sentences per a source speaker are transformed into magnitude STFT spectrograms and set aside to learn the block bases. Then, I picked up a sentence from each speaker and then mixed with 0dB SNR (the sound level of the two test signals are same).

An objective measurement that can assess the quality of the separated signals is proposed in [12]. From this, we can define three measures: Signal-to-Interference Ratio (SIR), Signal-to-Artifact Ra-

tion (SAR), and Signal-to-Distortion Ration (SDR). We want to get as high SIR as possible to maximize separation, but it can introduce artifacts that decrease SAR. SDR is an overall score.

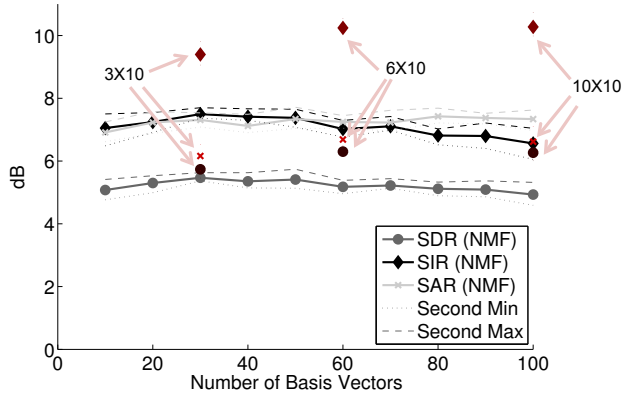


Figure 6: SDR, SIR, and SAR values from different choices of the number of bases and algorithms Red diamonds (SIR), crosses (SAR), and dots (SDR) are from the proposed method.

We can see in Figure 6 that the proposed method with enough number of bases (3, 6, and 10 bases per a cluster and ten clusters total) gave improved SIR by around 3dB. I believe that the improvement is because of the assumption that the proposed method can better preserve the data manifold as a combination of multiple sub-convex-hulls.

5 Conclusion and Future Works

In this project, I developed a manifold preserving NMF model where the components are grouped into several blocks that locally represents exclusive subsets of data. In this way we could get better-performing dictionaries of speech sources that could be used to produce improved source separation results.

Nothing stops it from evolving into deeper hierarchical model that further decomposes each block into subgroups on and on. On top of that, the initialization procedure with K-means clustering can be replaced with more sophisticated techniques. Comparisons to existing manifold preserving techniques should be performed as well.

Appendix: An Auto-encoder for NMF

Figure 7 depicts an auto-encoder that corresponds to NMF. This is not very different from a usual auto-encoder unless we further assume nonnegativity constraints for the parameters, W and H. Furthermore, the error between input and output can be defined by the β-divergence $\mathcal{D}_\beta(x|y)$ instead

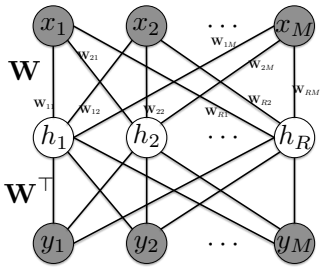


Figure 7: Auto-encoder for NMF.

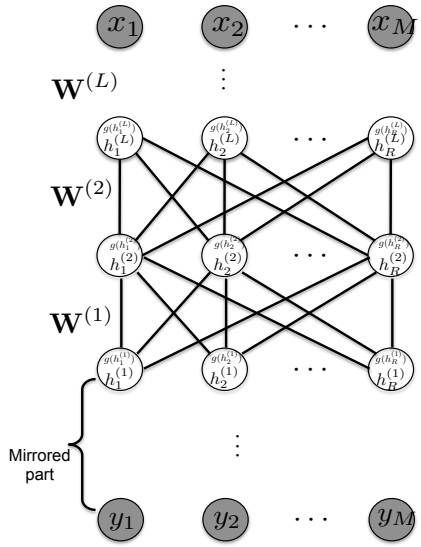


Figure 8: Auto-encoder for deep NMF.

of the mean squared error. The auto-encoder can be further extended to a multilayer network to make the best use of the deep structure as in Figure 8.

The relationships between nodes in adjacent layers are defined as follows:

$$z^{(l)} = g(h^{(l)})$$
$$h^{(l+1)} = w^{(l)} z^{(l)},$$

for some pre-defined nonlinearity $g(\cdot)$. Because the error between input and the output layer is the objective function of the stacked auto-encoders, and we can define the error in terms of β-divergence, the derived back-propagation algorithm is defined

$$\frac{\partial \mathcal{J}}{\partial W^{(L)}} = \frac{\partial \mathcal{J}}{\partial h^{(L)}} \frac{\partial h^{(L)}}{\partial W^{(L-1)}} = \frac{\partial \mathcal{D}(Y|X)}{\partial W^{(L-1)}}$$
$$\delta^{(L)} = \frac{\partial \mathcal{J}}{\partial z^{(L)}} = \frac{\partial \mathcal{D}(Y|X)}{\partial z^{(L)}}$$
$$\delta^{(l)} = \frac{\partial \mathcal{J}}{\partial z^{(l)}} = {W^{(l)}}^{\top} \delta^{(l+1)} g'(h^{(n+1)})$$
$$\frac{\partial \mathcal{J}}{\partial W^{(l)}} = \frac{\partial \mathcal{J}}{\partial h^{(l+1)}} \frac{\partial h^{(l+1)}}{\partial W^{(l)}} = \frac{\partial \mathcal{J}}{\partial h^{(l+1)}} {z^{(l)}}^{\top}$$
$$\frac{\partial \mathcal{J}}{\partial h^{(l)}} = \delta^{(l)} g'(h^{(n)}).$$

We can see that the $\delta^{(l)}$ at a given layer is a multiplication of current weights, previous $\delta^{(l+1)}$, and differentiation of the nonlinearity. However, the nonlinearity factor goes away in the multiplicative update rules.

I would like to study this model further, as for now it gives no better results than plain NMF case. One possible reason is that the training data that I used was not enough (around 1,000 frames) to train a network with hundreds of weights. It is also not clear yet how the multiplicative update rules in this network work instead of gradient descent with nonlinearity that enforces nonnegativity.

References

[1] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[2] ——, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 13. MIT Press, 2001.

[3] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, 2003, pp. 177–180.

[4] M. Kim, J. Yoo, K. Kang, and S. Choi, "Nonnegative matrix partial co-factorization for spectral and temporal drum source separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1192–1204, 2011.

[5] M. Kim and P. Smaragdis, "Single channel source separation using smooth nonnegative matrix factorization with markov random fields," in *Proceedings of the IEEE Workshop on Machine Learning for Signal Processing (MLSP)*, Southampton, UK, 2013.

[6] P. Smaragdis, M. Shashanka, and B. Raj, "A sparse non-parametric approach for single channel separation of known sounds," in *Advances in Neural Information Processing Systems (NIPS)*, Vancouver, BC, Canada, 2009.

[7] M. Kim and P. Smaragdis, "Manifold preserving hierarchical topic models for quantization and approximation," in *Proceedings of the International Conference on Machine Learning (ICML)*, Atlanta, Georgia, 2013.

[8] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[9] T. Hofmann, "Probablistic latent semantic indexing," in *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 1999.

[10] A. Lefèvre, F. Bach, and C. Févotte, "Itakura-saito non- negative matrix factorization with group sparsity," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011.

[11] D. L. Sun and G. J. Mysore, "Universal speech models for speaker independent single channel source separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada.

[12] E. Vincent, C. Fevotte, and R. Gribonval, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.