On the Performance of Manhattan Non-negative Matrix Factorization

Tongliang Liu and Dacheng Tao, Fellow, IEEE

Abstract—Extracting low-rank and sparse structures from matrices has been extensively studied in machine learning, compressed sensing, and conventional signal processing, and has been widely applied to recommendation systems, image reconstruction, visual analytics, and brain signal processing. Manhattan non-negative matrix factorization (MahNMF) is an extension of the conventional non-negative matrix factorization (NMF), which models the heavy-tailed Laplacian noise by minimizing the Manhattan distance between a non-negative matrix \( X \) and the product of two non-negative low-rank factor matrices. Fast algorithms have been developed to restore the low-rank and sparse structures of \( X \) in MahNMF. In this paper, we study the statistical performance of MahNMF in the frame of the statistical learning theory. We decompose the expected reconstruction error of MahNMF into the estimation error and the approximation error. The estimation error is bounded by the generalization error bounds of MahNMF, while the approximation error is analyzed using the asymptotic results of the minimum distortion of vector quantization. The generalization error bound is valuable for determining the size of the training sample needed to guarantee a desirable upper bound for the defect between the expected and empirical reconstruction errors. Statistical performance analysis shows how the reduced dimensionality affects the estimation and approximation errors. Our framework can also be used for analyzing the performance of NMF.

Index Terms—Manhattan distance, Non-negative matrix factorization, statistical analysis, estimation error, approximation error.

I. INTRODUCTION

The sheer volume of high-dimensional data generated from a wide variety of applications, such as data management [1], [2] and visual analytics [3], [4], present both a challenge and an opportunity for machine learning and algorithm development. Arguably, the high-dimensional data have low intrinsic dimensionality, particularly when sampled from a low-dimensional manifold, as shown by the success of manifold learning in many applications [5]–[7]. By stacking all data as column vectors of a matrix \( X \), the assumed low intrinsic dimensionality means that \( X \) is constructed by perturbing a low-rank matrix \( L \) with small noise \( N \), i.e., \( X = L + N \). Low-rank structure has been extensively studied, for instance in principal component analysis (PCA) [8], sparse wavelets [9] and non-negative matrix factorization (NMF) [10], to either exactly or approximately represent an arbitrary example in a dataset as a weighted sum of a few bases \( W \), i.e., \( L = WH \).

Since examples in a dataset can be associated with specific structures, e.g., face images under different lighting conditions, it is insufficient to assume that \( L \) is perturbed by small noise \( N \). This means that we need to model the random sparse structure \( S \), i.e., \( X = L + S + N \). If we simply ignore \( S \), the estimated rank of \( L \) will be significantly increased; for this reason, low-rank and sparse matrix decomposition has been extensively studied. Although the decomposition model itself is simple, it is more difficult to exactly or approximately recover \( L \) and \( S \), since the rank and cardinality constraints are not convex. Exact and unique decomposition does not always exist for an arbitrary matrix, and approximate decomposition results in a trade-off between low-rank and sparsity. Successful algorithms include rank-sparsity incoherence for matrix decomposition [11], robust PCA [12], GoDec [13], and Manhattan NMF (MahNMF) [14].

NMF, popularized by Lee and Seung [10], allows for the factorization of a non-negative data matrix into two low-rank non-negative matrices. Unlike other traditional dimensionality reduction methods, the found bases and the newly represented data matrices are required to be non-negative in NMF. The non-negativity allows only additive combinations and forces learning of parts-based representations [15]–[18]. NMF has been successfully applied as a dimensionality reduction approach in many fields, such as signal processing [19]–[21], data mining [22]–[24] and bioinformatics [25]–[27].

Numerous different algorithms have been developed for NMF applications, which can be classified into two groups depending on whether additional content information is used: the first group contains content-free algorithms [28], which attempt to find an optimization solution for NMF via numerical methods; the second contains content-based algorithms [29], in which the non-negative matrix is factorized with a constraint on \( W \) (or \( H \), or both). Due to the successful use of NMF in real-world applications, several popular computing environments, such as Matlab, R and Oracle Data Mining, have developed NMF packages.

MahNMF is an important extension of NMF that learns the low-rank and sparse structures simultaneously. In particular, MahNMF models the heavy-tailed Laplacian noise by minimizing the Manhattan distance between \( X \in \mathbb{R}^{m \times n}_+ \) and its low-rank approximation \( L = WH \), where \( W \in \mathbb{R}^{m \times r}_+ \) and \( H \in \mathbb{R}^{r \times n}_+ \); it is therefore effective when data are contaminated by outliers. It has been comprehensively and
empirically proven that MahNMF can robustly estimate the low-rank and sparse components of a non-negative matrix [14]. We also provide a thorough analysis to show how MahNMF is more robust to noise and outliers than NMF from an optimization viewpoint. Recently, learning with noisy examples has received a great deal of attention. The interested reader is referred to further examples [30]–[32].

Donoho and Stodden [33] discussed under what conditions NMF is unique (a sufficient condition for $X = WH$). The problem of when $X = WH$ has also been studied for a few special cases of NMF: Thomas [34] proved a sufficient and necessary condition for rank-$r$ NMF, and Kaykobad [35] presented a sufficient condition for symmetric NMF (or a non-negative symmetric matrix to be completely positive). We therefore know that in most cases, the solutions of MahNMF (and indeed NMF) returned by content-free algorithms are not unique and $X \neq WH$, and that even though solutions returned by content-based algorithms may be unique, $X$ is unlikely to be equal to $WH$. However, there are still two fundamental questions that have yet to be posed:

1) How much is $E\|X - W_nH_n\|$?
2) How much is $E\|X - W^*H^*\|$?

We use $W_n$ and $W^*$ to denote the learned and target bases of MahNMF, respectively, and $H_n$ and $H^*$ are the corresponding new representations. Maurer and Pontil [36] developed dimensionality-independent generalization error bounds for $k$-dimensional coding schemes, which can be used to analyze the dimensionality-independent generalization error bounds of NMF; this approach answers the first question for NMF. However, the MahNMF analysis is different, and therefore in this paper, we address these two questions specifically with respect to MahNMF.

Using statistic learning theory [37], [38], we first define the expected reconstruction error for MahNMF and decompose it into estimation and approximation errors. The estimation error, which is dependent on the used learning algorithm, is bounded by the generalization error bound of MahNMF. Using the Rademacher complexity [39] and covering number methods [40], we derive the generalization error bounds. The obtained bounds show, with high probability, that the expected reconstruction error of the learned bases $W_n$ is not more than $O(\sqrt{mn\ln n/n})$ worse than the empirical reconstruction error of the learned bases $W^*_n$; this answers the first question. The approximation error, which is dependent on the sample distribution, is analyzed by the asymptotic results of the minimum distortion of vector quantization [41]. In contrast to the estimation error, the approximation error is a non-increasing function of the reduced dimensionality $r$, and decreases in the order of $O(r^{-1/m})$ as $r$ approaches $m$. This answers the second question.

We note that our method based on directly bounding the covering number of the induced loss class can be used to derive tighter dimensionality-dependent generalization error bounds (than the dimensionality-independent bounds in [36]) for NMF, and that the method for deriving the approximation error bound can also be applied to the NMF using Euclidean distance loss.

The rest of this paper is organized as follows. In Section II, we present the setup of the problem, and provide the stochastic framework by decomposing the expected reconstruction error into estimation error and approximation error. Our main results about estimation error and approximation error are outlined in Section III and IV, respectively. The proofs of our main results are provided in Section V. Finally, Section VI concludes the paper.

II. PROBLEM SETUP

This section introduces notation that will be used throughout this paper, and then presents the normalized MahNMF. We also define reconstruction error for MahNMF and decompose the expected reconstruction error into estimation error and approximation error.

A. Notation

We denote $X = (x_1, \ldots, x_n) \in \mathbb{R}^{m \times n}$ as the data matrix consisting of $n$ independent and identically distributed observations drawn from a space $\mathcal{X}$ with a Borel measure $\rho$. We use capital letters to denote matrices, $A_i$ the $i$-th column of matrix $A$ and $A_{ij}$ the $i,j$-th entry of $A$. The $\ell_2$ (Euclidean) norm is denoted by $\|\cdot\|_2$, while the Manhattan distance between $A$ and $B$ is denoted by $\|A - B\|_M$ or $\|A - B\|_1$. The notation $O(\cdot)$ represents the big Omicron notation in Bachmann-Landau notation.

B. MahNMF

NMF minimizes the Euclidean distance loss and KLNMF$^1$ minimizes the Kullback-Leibler (KL) divergence loss as follows:

$$\min_{W,H} F_{\text{NMF}}(WH) = \|X - WH\|^2_2$$
$$= \sum_{i=1}^n \|x_i - WH_i\|^2_2, \quad (1)$$

subject to $W \in \mathbb{R}_+^{m \times r}, H \in \mathbb{R}_+^{r \times n}$ and

$$\min_{W,H} F_{\text{KLNMF}}(WH) = D(X\|WH)$$
$$= \sum_{i,j} \left( A_{ij} \ln \frac{A_{ij}}{B_{ij}} - A_{ij} + B_{ij} \right)$$

where $D(A\|B) = \sum_{ij} \left( A_{ij} \ln \frac{A_{ij}}{B_{ij}} - A_{ij} + B_{ij} \right)$ and the reduced dimensionality $r$ satisfies that $r < \min(m,n)$.

The Euclidean distance and KL divergence loss model the Gaussian noise and Poisson distribution, respectively. Assume that the residuals (or noise) $(x_i - WH_i), i = 1, \ldots, n$ are independently distributed with a Gaussian distribution. Problem (1) can be easily derived using the maximum likelihood approach.

$^1$For clarity, we denote NMF and KLNMF as the matrix factorization procedures minimizing Euclidean distance loss and KL divergence loss, respectively.
To derive problem (2), we assume that \( x_i \sim \text{Poisson}(WH_i) \), that is

\[
P(x_i|WH_i) = \frac{e^{-WH_i}(WH_i)^{x_i}}{x_i!}.
\]

Since we have assume that \( x_i, i = 1, \ldots, n \) are independently distributed, the Poisson likelihood of observing \( x_i, i = 1, \ldots, n \) given underlying parameters \( WH \) is given by

\[
P(X|WH) = \prod_{i=1}^{n} \frac{e^{-WH_i}(WH_i)^{x_i}}{x_i!}.
\]  
(3)

To maximize the likelihood with respect to \( WH \), take the log on both sides of Eq. (3), we have the following problem:

\[
\min_{W,H} \sum_{i=1}^{n} WH_i - x_i \log(WH_i),
\]

\[
\text{s.t. } W \in \mathbb{R}^{m \times r}, H \in \mathbb{R}^{r \times n},
\]

which is equal to the problem (2).

However, Gaussian and Poisson distributions are not heavy-tailed, as shown in Fig. 1. Considering that many practical contaminations, e.g., occlusions, are heavy-tailed, NMF and KLNMF therefore do not perform well in such practical scenarios.

Assume that the residuals \( (x_i - WH_i), i = 1, \ldots, n \) are independently sampled from a Laplace distribution, which is heavy-tailed as shown in Fig. 1. The objective function of MahNMF can be easily derived using the maximum likelihood approach. Thus, MahNMF employs the Manhattan distance between \( X \) and \( WH \) to model the heavy-tailed Laplacian noise:

\[
\min_{W,H} \sum_{i=1}^{n} ||x_i - WH_i||_1,
\]

\[
\text{s.t. } W \in \mathbb{R}^{m \times r}, H \in \mathbb{R}^{r \times n},
\]

where the reduced dimensionality \( r < \min(m, n) \).

Moreover, thanks to the Manhattan distance loss, MahNMF also effectively extracts the sparse structures, which makes MahNMF much more robust to outliers and capable to occlusions [14].

Besides the robustness advantage of MahNMF discussed from the viewpoint of noise distribution, we show the advantage of MahNMF from an optimization procedure viewpoint.

Let \( F(WH) \) be the objective function of a non-negative matrix factorization problem. Let \( f(t) = F(iWH), t \in \mathbb{R} \). We can verify that optimizing \( F(WH) \) is equal to finding a \( WH \) such that \( \delta f(1) = 0 \), where \( \delta f(t) \) denotes the subgradient of \( f(t) \).

We have

\[
\delta f_{\text{NMF}}(1) = \sum_{ij} 2(X - WH)_{ij}(-WH)_{ij}
\]  
(4)

and

\[
\delta f_{\text{KLNMF}}(1) = \sum_{ij} (WH - X)_{ij}
\]  
(5)

and

\[
\delta f_{\text{MahNMF}}(1) = \sum_{ij} \frac{2}{|X - WH|_{ij}} (X - WH)_{ij}(-WH)_{ij}.
\]  
(6)

We define \( \frac{2}{\| \cdot \|_{ij}} \in [-1,1] \) for MahNMF to satisfy the definition of subgradient.

We compare the robustness of MahNMF and NMF from an optimization viewpoint. Let

\[
c(X_{ij}, WH) = 2(X - WH)_{ij}(-WH)_{ij}
\]

be the contribution the \( j \)th entry of the \( i \)th training example made to the optimization procedure,

\[
w_{\text{MahNMF}}(X_{ij}, WH) = 1/|X - WH|_{ij}
\]

the weight to the contribution for the \( j \)th entry of the \( i \)th training example of MahNMF and

\[
w_{\text{NMF}}(X_{ij}, WH) = 1
\]

the weight to the contribution for the \( j \)th entry of the \( i \)th training example of NMF. Comparing equations (4) and (6), we notice that every entry has the same contribution but different weights. We further find that \( |X - WH|_{ij} \) represents the noise error added to the \( ij \)th entry of \( X \) or the very large error introduced by an outlier \( X_{ij} \) with regard to \( W \). We therefore interpret the optimization procedures as contribution-weighted optimization procedures regarding to the noise error of entries.

During the optimization procedure, a robust algorithm should assign a small weight to the contribution made by a large noise error entry, and a large weight to the contribution made by a small noise error entry. According to the contribution-weighted optimization procedure interpretation, MahNMF assigns a smaller weight to a contribution if the corresponding entry has a larger error while NMF provides the same weight to all entries. Thus, MahNMF is much more robust than NMF.

To compare the robustness of MahNMF and KLNMF, we set the contribution as

\[
c(X_{ij}, WH) = (WH - X)_{ij},
\]
the weight of MahNMF as
\[ w_{\text{MahNMF}}(X_{ij}, WH) = 2\left(1 - \frac{(X)_{ij}}{(WH)_{ij}}\right) \]
and the weight of KLNMF as
\[ w_{\text{KLNMF}}(X_{ij}, WH) = 1. \]
Since the KL divergence is asymmetric, we compare the robustness in two directions: \( X_{ij} > (WH)_{ij} \) and \( X_{ij} < (WH)_{ij} \). For each direction, it can be easily verified that when the noise error \(|X - WH|_{ij}\) is large, the weight function \( w_{\text{MahNMF}}(X_{ij}, WH) \) will be small while the weight function of KLNMF assigns the same weight to all entries. MahNMF is therefore much more robust than NMF. We empirically show the robustness and good performance of MahNMF in Appendix B.

For fixed bases \( W \), the representations \( H \) are determined by a convex problem. Therefore, it is sufficient to analyze the performance of MahNMF by studying the choice of bases \( W \).

Note that \( WH = WQ^{-1}QH \) for any scaling matrix \( Q \), and we can normalize \( \|W_i\|_1 \), \( i = 1, \ldots, r \) by choosing
\[ Q = \begin{pmatrix} \|W_1\|_1 & \|W_2\|_1 & \cdots & \|W_r\|_1 \end{pmatrix} \]
to limit the choice of bases \( W \) without changing the ability to represent the solutions to a MahNMF problem. Here, we call MahNMF with \( \ell_1 \)-normalized bases \( W \) the \( \ell_1 \)-normalized MahNMF, which has the following property:

**Lemma 1:** For \( \ell_1 \)-normalized MahNMF problems, if \( \|x\|_1 \leq R \), then \( \|h\|_1 \leq 2R \).

We defer the proof until Section V.

**C. Extensions of MahNMF**

In the previous subsection, we showed that MahNMF is more robust (to noise) than NMF and KLNMF. Then, it may interest many readers to exploit the extensibility of MahNMF for practical applications.

Guan et al. [14] provided a flexible framework for developing various algorithms of MahNMF for practical applications. They presented detailed optimization algorithms for MahNMF that are restricted by box-constraint [42], manifold regularization [43], group sparsity [44], elastic net [45] and symmetric bases [46]. MahNMF of cause can be extended to many other different scenarios; see, for example, online NMF learning [47], NMF that handles time varying data [48], and NMF that exploits the pairwise constraints that indicate the similarity/dissimilarity between two observations [49].

Note that the generalization error bounds of MahNMF derived by employing the complexity measures that measure the whole predefined loss class (such as the generalization error bounds in Theorems 1 and 2) can also be the upper bounds for the batch learning [50] extensions of MahNMF. Constraints will help produce a small active loss class, which is a subset of the predefined loss class. Then the complexity of the active loss class will be small. Taking the Rademacher complexity for an example, Bartlett and Mendelson have proven this property in Theorem 12 of [39]. So, if a constraint produces a small active loss class, the corresponding learning algorithm will share the generalization bound obtained by analyzing the complexity of the whole predefined loss class.

However, for different extensions of MahNMF, the estimation and approximation errors may vary. It is hard to analyze them in a uniform approach. We therefore, in the rest of the paper, analyze the performance of the empirical risk minimization (ERM) algorithm of MahNMF.

**D. Estimation Error and Approximation Error**

This subsection sets a framework for our results based on statistic learning theory. For any bases \( W \in \mathbb{R}^{m \times r} \), we define the reconstruction error of an example \( x \) as follows:
\[ f_W(x) = \min_{h \in \mathbb{R}^r} \|x - Wh\|_1. \]

The expected reconstruction error provides a criterion for choosing the bases \( W \), such that the expected reconstruction error is minimized.

**Definition 1 (Expected reconstruction error):** If the measurable space \( \mathcal{X} \) has a Borel measure \( \rho \) with distribution density function \( p(x) \), we define the expected reconstruction error of the bases \( W \) as
\[ R(W) = \int_{\mathcal{X}} f_W(x)dp(x) = \int_{\mathcal{X}} f_W(x)p(x)dx. \]

MahNMF attempts to search the bases \( W \) to make a small expected reconstruction error. However, the distribution density function \( p(x) \) is usually unknown, and \( R(W) \) cannot be directly minimized. The ERM algorithm provides a way to learn an approximation by minimizing the empirical reconstruction error.

**Definition 2 (Empirical reconstruction error):** If data \( X = (x_1, \ldots, x_n) \) are drawn independently and identically from \( \mathcal{X} \), we define the empirical reconstruction error of the bases \( W \) as
\[ R_n(W) = \frac{1}{n} \sum_{i=1}^{n} f_W(x_i). \]

Following statistic learning theory, we define the set of functions \( F_W \), indexed by \( \mathcal{W}_r = \mathbb{R}^{m \times r} \), as the loss class of MahNMF. For every \( f_W \in F_W \), \( f_W(x) \) measures the reconstruction error of \( x \) using the bases \( W \in \mathcal{W}_r \). Thus, choosing the bases \( W \) with a small empirical reconstruction error is equal to choosing a \( f_W \in F_W \), such that the empirical reconstruction error is minimized.

Our goal is to analyze the expected reconstruction error of the learned bases \( W_{n,r} \) of MahNMF, where
\[ W_{n,r} = \arg \min_{W \in \mathcal{W}_r} R_n(W). \]

We can decompose it into the estimation error and approximation error as follows
\[ R(W_{n,r}) = R(W_{n,r}) - R(W_r) + R(W_r). \]
We will study the estimation error, generalization error, and approximation error mostly depends on learning theory, the estimation error mostly depends on learning errors. It is well known that in the classical statistical learning theory, the expected reconstruction error, the estimation error, the generalization error, and the approximation error are shown by points and arrows.

where \( W_r = \arg \min_{W \in W_r} R(W) \).

In this paper, the estimation error and approximation error are bounded, respectively. The following inequalities provide a method to upper bound the estimation error:

\[
R(W_{n,r}) - R(W_r) \\
\leq |R(W_{n,r}) - R_n(W_{n,r})| \\
+ |R_n(W_{n,r}) - R_n(W_r)| + |R_n(W_r) - R(W_r)| \\
\leq 2 \sup_{W \in R_{n,r}^m} |R(W) - R_n(W)|. \tag{7}
\]

The second inequality holds because \( W_{n,r} \) is defined to be the global solution as \( W_{n,r} = \arg \min_W R_n(W) \). The defect in the last line is referred to as the generalization error.

The expected reconstruction error \( R(W_{n,r}) \) can be either bounded by bounding the estimation error \( R(W_{n,r}) - R(W_r) \) and approximation error \( R(W_r) \), or bounding the generalization error \( R(W_{n,r}) - R_n(W_{n,r}) \).

Figure 2 is a schematic diagram illustrating the loss class and errors. It is well known that in the classical statistical learning theory, the estimation error mostly depends on learning algorithms and the approximation error mostly depends on the choice of the loss class. However, for a MahNMF (and NMF) problem, the loss class is fixed because \( W_r = R_{m \times r} \) and the loss function \( f_{W_r} = \arg \min_{f_{W_r} \in F_{W_r}} R(W) \) is always in the loss class. Therefore, the approximation error for MahNMF (and NMF) depends only on the distribution of \( x \) (if \( r \) is fixed). We will study the estimation error, generalization error, and approximation error of MahNMF in the rest of the paper.

III. Generalization and Estimation Error Bounds

We bound the estimation error using the bound of the generalization error and inequalities (7). The generalization error is usually bounded by measuring the complexity of the loss class. The method based on Rademacher complexity is one of the most frequently used methods for deriving generalization error bounds, based on which the following theorem is obtained.

**Theorem 1:** For \( \ell_1 \)-normalized MahNMF problems, assume that \( \|x\|_1 \leq R \). For any \( W \in W_r \) and any \( \delta > 0 \), with probability at least \( 1 - \delta \), we have

\[
|R(W) - R_n(W)| \leq \sqrt{\frac{2\pi r m r R}{n}} + R\sqrt{\frac{\ln(2/\delta)}{2n}}.
\]

The Rademacher complexity usually leads to dimensionality-independent generalization bounds. However, our obtained bound (in Theorem 1) is dimensionality-dependent. One reason for this is that in the proof we used \( \|x\|_1 \leq \sqrt{m} \|x\|_2 \) to find a proper Gaussian process to upper bound the Rademacher complexity. We conjecture that the Rademacher complexity can be used to derive a dimensionality-independent generalization error bound for MahNMF by exploiting a more subtle proving approach.

The method of Rademacher complexity will introduce the worst-case dependency with regard to \( r \), which means the generalization bound in Theorem 1 is loose. Combining the Rademacher complexity and covering number to measure the complexity of the loss class may avoid the worse-case dependency (see, e.g., the proof method of Theorem 2 in [36]). The price to pay for this is that the obtained convergence rate is slower than that of Theorem 1.

To combine the measurements of Rademacher complexity and covering number, Maurer and Pontil [36] factored the bases \( W = US \) for NMF, where \( S \) is an \( r \times r \) matrix and \( U \) is an \( m \times r \) isometry, and measured the complexities induced by the \( r \times r \) matrices and isometries by exploiting the covering number and Rademacher complexity, respectively. However, for MahNMF problems, we cannot derive any tighter generalization bounds by using this method.

We introduce a method to directly bound the covering number of the loss function class induced by the reconstruction error. The obtained generalization bound will be much tighter.

**Theorem 2:** For \( \ell_1 \)-normalized MahNMF problems, assume that \( \|x\|_1 \leq R \). For any \( W \in W_r \), any \( \xi > 0 \) and any \( \delta > 0 \), and \( n \geq \frac{8m^2 R^2}{\xi^2} \), we have

\[
P \left\{ \|R(W) - R_n(W)\| \geq \xi \right\} \\
\leq 8 \left( \frac{16mR}{\xi} \right)^m \exp \left( -n\xi^2 \right),
\]

Let \( U \) be a bounded linear transformation from \( \mathbb{R}^r \) to \( \mathbb{R}^m \). If \( U \) is an isometry, it holds that \( \|Ux\| = \|x\| \) for all \( x \in \mathbb{R}^r \), where \( \|\cdot\| \) denotes a norm on the Euclidean space. It can be easily verified that the \( \ell_1 \)-normalized bases \( W \) compose an isometry with regard to the \( \ell_1 \) norm.
Upper bound of generalization

Theorem 1: For MahNMF problems, when the reduced dimensionality \( r \) approaches \( m \), we have

\[
R(W_{n,r}) - R(W_r) \leq 2R \sqrt{\frac{mr \ln(2mnR)}{2n}} + \frac{4}{n} + 2R \sqrt{\frac{\ln(2/\delta)}{2n}}.
\]

Theorem 4 shows that given \( m \) and \( r \), the estimation error, which is derived from the generalization bound, depends only on the learning algorithm of MahNMF, because some learning algorithms can further narrow the complexity of the loss class by using regularization terms.

IV. ASYMPTOTIC APPROXIMATION ERROR BOUND

The approximation errors \( R(W_r) \) are different for different distributions of \( X \). It is hard to derive a non-asymptotic uniform upper bound for the approximation error of different distributions. However, we can instead provide a tight asymptotic uniform upper bound for the approximation error, as follows:

\[
\text{Theorem 4: For MahNMF problems, when the reduced dimensionality } r \text{ approaches } m, \text{ we have }
R(W_r) \leq O(r^{-\frac{1}{m}}).
\]

The order of \( r \) is optimal.

A detailed proof is provided in the next section.

The asymptotic approximation error bound depends only on the reduced dimensionality \( r \). Our approximation error bound is somewhat weak (because \( m^{-1/m} \to 1, m \to \infty \)); however, it is the first to be derived for MahNMF (and NMF) and it provides insight into the problem, namely that when \( r \) and \( m \) are large, the approximation error will decrease with respect to the increase of \( r \) of order \( O(r^{-1/m}) \).

V. PROOFS

In this section, we present the proofs of the results in Sections II, III and IV. We begin by introducing the concentration inequalities, which play an important role in proving generalization error bounds.

A. Auxiliary Results

The following concentration inequality, well known as Hoeffding’s inequality [53], is widely used for deriving generalization error bounds.

\[
\text{Theorem 5: Let } X_1, \ldots, X_n \text{ be independent random variables with the range } [a_i, b_i] \text{ for } i = 1, \ldots, n. \text{ Let } S_n = \sum_{i=1}^{n} X_i. \text{ Then for any } \epsilon > 0, \text{ the following inequalities hold:}
\]

\[
\Pr\{S_n - ES_n \geq \epsilon\} \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^{n} (b_i - a_i)^2}\right),
\]

\[
\Pr\{ES_n - S_n \geq \epsilon\} \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^{n} (b_i - a_i)^2}\right).
\]

Another well-known concentration inequality is McDiarmid’s inequality [54] (also known as the bounded-difference inequality).

\[
\text{Theorem 6: Let } X = (x_1, \ldots, x_n) \text{ be an independent and identically distributed sample and } X^i \text{ a new sample with the}
\]

prove
Then, the following theorem holds:

\[ |f(X) - f(X')| \leq c_i, \forall i \in \{1, \ldots, n\}. \]

Then for any \( X \in \mathcal{X}^n \) and \( \epsilon > 0 \), the following inequalities hold:

\[
\Pr\{f(X) - Ef(X) \geq \epsilon\} \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^{n}(c_i)^2}\right),
\]

\[
\Pr\{Ef(X) - f(X) \geq \epsilon\} \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^{n}(c_i)^2}\right).
\]

Glivenko-Cantelli Theorem [40] is often used to analyze the uniform convergence of the empirical reconstruction error to the expected reconstruction error. A relatively small complexity of the loss class is essential to prove a Glivenko-Cantelli class, which provides consistent generalization error bounds. The Rademacher complexity and covering number are the most used complexity measures.

The Rademacher complexity and Gaussian complexity are data-dependent complexity measures. They are often used to derive dimensionality-independent generalization error bounds and defined as follows:

**Definition 3:** Let \( \sigma_1, \ldots, \sigma_n \) and \( \gamma_1, \ldots, \gamma_n \) be independent Rademacher variables and independent standard normal variables, respectively. Let \( x_1, \ldots, x_n \) be an independent and identically distributed sample and \( F \) a function class. The empirical Rademacher complexity and empirical Gaussian complexity are defined as:

\[
\mathcal{R}_n(F) = Ef_{\sigma} \sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(x_i)
\]

and

\[
\mathcal{G}_n(F) = Ef_{\gamma} \sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \gamma_i f(x_i),
\]

respectively. And the Rademacher complexity and Gaussian complexity are defined as:

\[
\mathcal{R}(F) = Ef \mathcal{R}_n(F)
\]

and

\[
\mathcal{G}(F) = Ef \mathcal{G}_n(F),
\]

respectively.

Using the symmetric distribution property of random variables, the following theorem holds:

**Theorem 7:** Let \( F \) be a real-valued function class on \( \mathcal{X} \) and \( X = (x_1, \ldots, x_n) \in \mathcal{X}^n \). Let

\[
\Phi(X) = \sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} (Ef(x) - f(x_i)).
\]

Then,

\[
Ef \Phi(X) \leq \mathcal{R}(F).
\]

**Proof.** Let \( X' = (x'_1, \ldots, x'_n) \in \mathcal{X}^n \) be another sample distributed independently of \( X \). We have

\[
E_{x} \Phi(X) = Ef_{\sigma} \sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} (Ef(x'_i) - f(x_i))
\]

\[
\leq Ef_{\sigma} \sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} (f(x'_i) - f(x_i))
\]

(Since \( f(x'_i) - f(x_i) \) has a symmetric distribution)

\[
= Ef_{\sigma} \sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(x'_i) - f(x_i)
\]

\[
= Ef_{\sigma} \sup_{f \in F} \frac{2}{n} \sum_{i=1}^{n} \sigma_i f(x_i)
\]

\[
= \mathcal{R}(F).
\]

This concludes the proof.

The following theorem [55] proven utilizing Theorem 7 and Hoeffding’s inequality plays an important role in proving generalization error bounds.

**Theorem 8:** Let \( F \) be a \([a, b]\)-valued function class on \( \mathcal{X} \) and \( X = (x_1, \ldots, x_n) \in \mathcal{X}^n \). For any \( \delta > 0 \), with probability at least \( 1 - \delta \), we have

\[
\sup_{f \in F} \left( Ef(x) - \frac{1}{n} \sum_{i=1}^{n} f(x_i) \right) \leq \mathcal{R}(F) + (b - a) \sqrt{\frac{\ln(1/\delta)}{2n}}.
\]

**Proof.** Let

\[
\Phi(X) = \sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} (Ef(x) - f(x_i)).
\]

We have

\[
|\Phi(X) - \Phi(X')| \leq \frac{1}{n} |f(x_i) - f(x'_i)| \leq \frac{b - a}{n}.
\]

Combining Theorem 7 and McDiarmid’s inequality, we have

\[
\Pr\{\Phi(X) - \mathcal{R}(F) \geq \epsilon\} \leq \Pr\{\Phi(X) - Ef(X) \geq \epsilon\} \leq \exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^{n} (b - a)^2}\right).
\]

Solving \( \delta = \exp\left(-\frac{2n^2\epsilon^2}{(b-a)^2}\right) \) for \( \epsilon \) gives the result.

The covering number [56] is also a data-dependent complexity measure of the complexity of a loss class.

**Definition 4:** Let \( B \) be a metric space with metric \( d \). Given observations \( X = (x_1, \ldots, x_n) \), and vectors \( f(X) = (f(x_1), \ldots, f(x_n)) \in B^n \), the covering number in \( p \)-norm, denoted as \( N_p(F, \epsilon, X) \), is the minimum number \( m \) of a collection of vectors \( v_1, \ldots, v_m \in B^n \), such that for \( f \in F, \forall v_j \):

\[
\|d(f(X), v_j)\|_p = \left[ \sum_{i=1}^{n} d(f(x_i), v_{ij}^p)^p \right]^{1/p} \leq n^{1/p} \epsilon,
\]

where \( v_{ij}^p \) is the \( i \)-th component of vector \( v_j \). We also define \( N_p(F, \epsilon, n) = \sup_X N_p(F, \epsilon, X) \).
The following well-known result is due to Pollard [57], whose proof combines the Hoeffding’s inequality and the covering number.

**Theorem 9:** Let $X_{1}^{2n} = \{x_1, \ldots, x_{2n}\}$ be $2n$ independent and identically distributed observations. For a function class $F$ with the range $[a, b]$, let $E f = \int f(x)d\rho(x)\) and $E_n f = \frac{1}{n} \sum_{i=1}^{n} f(x_i)$. For any $\xi > 0$ and any $n \geq \frac{8(b-a)^2}{\xi^2}$, we have

$$\Pr \left\{ \sup_{f \in F} |Ef - E_n f| \geq \xi \right\} \leq 8E N_1(F, \xi/8, X_{1}^{2n}) \exp \left( -\frac{n\xi^2}{32(b-a)^2} \right).$$

We provide an outline in Appendix A for the sake of self-containment of this paper.

**B. Proof of Lemma 1**

Let

$$h^* = \min_h \|x - Wh\|_1.$$  

We have

$$\|Wh^*\|_1 = \|x - Wh^*\|_1 \leq \|x - W0\|_1 = \|x\|_1.$$  

Then

$$\|Wh^*\|_1 \leq 2\|x\|_1 \leq 2R.$$  

We also have that

$$\|Wh^*\|_1 = \sum_{j=1}^{m} \|\langle Wh^* \rangle_j \| = \sum_{j=1}^{m} \|\sum_{k=1}^{r} W_{jk} h_k \|$$

$$= \sum_{k=1}^{m} \sum_{j=1}^{r} W_{jk} h_k = \sum_{k=1}^{m} \left( \sum_{j=1}^{r} W_{jk} \right) h_k$$

$$= \sum_{k=1}^{m} h_k = \|h^*\|_1.$$  

Thus,

$$\|h^*\|_1 \leq 2R.$$  

This concludes the proof of Lemma 1. 

**C. Proof of Theorem 1**

Because the reconstruction error

$$f_W(x) = \min_{h \in \mathbb{R}_+} \|x - Wh\|_1$$

has a minimum in its function, it is hard to bound the Rademacher complexity $\mathcal{R}(F_W)$ directly. We use the following two lemmas (see proofs in [58]) to bound $\mathcal{R}(F_W)$ by finding a proper Gaussian process which can be easily bounded to upper bound the Rademacher complexity $\mathcal{R}(F_W)$.

**Lemma 2 (Slepian’s Lemma):** A Gaussian process of $X$ is defined as $\Omega_X = \sum_{i=1}^{n} \gamma_i x_i$. Let $\Omega$ and $\Xi$ be mean zero, separable Gaussian processes indexed by a common set $S$, such that

$$E(\Omega_{s_1} - \Omega_{s_2})^2 \leq E(\Xi_{s_1} - \Xi_{s_2})^2, \forall s_1, s_2 \in S.$$  

Then

$$E \sup_{s \in S} \Omega_s \leq E \sup_{s \in S} \Xi_s.$$  

The Gaussian complexity is related to the Rademacher complexity by the following lemma.

**Lemma 3:**

$$\mathcal{R}(F) \leq \sqrt{\pi/2G(F)}.$$  

We are now going to find a proper Gaussian process to upper bound the Rademacher complexity $\mathcal{R}(F_W)$.

**Lemma 4:** For $\ell_1$-normalized MahNMF problems, assume that $\|x\|_1 \leq R$. Then,

$$\mathcal{R}(F_W) \leq \sqrt{2\pi \tau \ln R}/\sqrt{n}.$$  

**Proof.** Let

$$\Omega_W = \sum_{k=1}^{n} \gamma_k \min_h \|x_k - Wh\|_1$$

and

$$\Xi_W = 2\sqrt{\pi} R \sum_{k=1}^{n} \sum_{i=1}^{m} \gamma_{kij} \langle W^T e_j, e_i \rangle,$$

where $e_i, e_j$ are natural basis and $\gamma_k$ and $\gamma_{kij}$ are independent standard norm variables indexed by $k, i, j$.

We first prove that

$$E(\Omega_{W_1} - \Omega_{W_2})^2 \leq E(\Xi_{W_1} - \Xi_{W_2})^2.$$  

We have

$$E(\Omega_{W_1} - \Omega_{W_2})^2$$

$$= E \left( \sum_{k=1}^{n} \left( \gamma_k \min_{h \in \mathbb{R}_+} \|x_k - W_1 h\|_1 - \min_{h \in \mathbb{R}_+} \|x_k - W_2 h\|_1 \right) \right)^2$$

$$= \sum_{k=1}^{n} \left( \min_{h \in \mathbb{R}_+} \|x_k - W_1 h\|_1 - \min_{h \in \mathbb{R}_+} \|x_k - W_2 h\|_1 \right)^2$$

$$\leq \sum_{k=1}^{n} \left( \max_{h} \|x_k - W_1 h\|_1 - \|x_k - W_2 h\|_1 \right)^2$$

$$= \sum_{k=1}^{n} \left( \max_{h} \|W_1 h - W_2 h\|_1 \right)^2$$

$$= \sum_{k=1}^{n} \max_{h} \left( \sum_{i=1}^{r} h_i \|W_1 - W_2 e_i\|_1 \right)^2$$

(Using Cauchy-Schwarz inequality)

$$\leq \sum_{k=1}^{n} \sum_{h} h_i^2 \sum_{i=1}^{r} \sum_{i=1}^{r} \|W_1 - W_2 e_i\|_1^2$$

$$= \sum_{i=1}^{r} \sum_{i=1}^{r} \|W_1 - W_2 e_i\|_1^2.$$
Theorem 1 can be proven combing Theorem 8 and Lemma 4.

D. Proof of Theorem 2

The following result, which bounds the covering number of the loss class of MahNMF, plays a central role in proving Theorem 2.

**Lemma 5:** Let \( F_{W_i} \) be the loss function class of \( \ell_1 \)-normalized MahNMF. Then

\[
\ln N_1(F_{W_i}, \xi, 2n) \leq m \ln \left( \frac{2m}{\xi} \right).
\]

**Proof.** We will bound the covering number of \( F_W \) by bounding the covering number of the class \( W \). Cutting the subspace \( \{0, 1\}^m \subseteq \mathbb{R}^m \) into small \( m \)-dimensional regular solids with width \( \xi \), there are a total of

\[
\left[ \frac{1}{\xi} \right]^m \leq \left( \frac{1}{\xi} + 1 \right)^m \leq \left( \frac{2}{\xi} \right)^m
\]

such regular solids. If we pick out the centers of these regular solids and use them to make up \( W \), there are

\[
\left[ \frac{1}{\xi} \right]^m \leq \left( \frac{2}{\xi} \right)^m
\]

choices, denoted by \( S \), \(|S|\) is the upper bound of the \( \xi \)-cover of the class \( W \).

We will prove that for every \( W \), there exists a \( W' \in S \) such that \( |f_W - f_{W'}| \leq \xi' \), where \( \xi' = m\xi R \).

\[
|f_W - f_{W'}| = \left| \min_{h} \|x - Wh\|_{\ell_1} - \min_{h} \|x - Wh'\|_{\ell_1} \right|
\]

\[
\leq \max_{h} (\|x - Wh\|_{\ell_1} - \|x - Wh'\|_{\ell_1})
\]

\[
\leq \max_{h} (\|x - Wh\|_{\ell_1} - \|x - Wh'\|_{\ell_1})
\]

\[
\leq \max_{i} h_i (\|W - W'\|_{1})
\]

\[
\leq \max_{i} h_i m\xi / 2
\]

(Using Lemma 1)

\[
\leq m\xi R
\]

\[
= \xi'.
\]

The third inequality holds because \( |a| - |b| \leq |a - b| \).

Let the metric \( d \) be the absolute difference metric. According to Definition 4, for \( \forall f_W \in F_W \), there is a \( W' \in S \) such that

\[
\|d(f_W(x), f_{W'}(x))\|_{\ell_1}
\]

\[
= \left[ \sum_{i=1}^{2n} d(f_W(x_i), f_{W'}(x_i)) \right]
\]

\[
\leq 2n \xi'.
\]

Thus,

\[
N_1(F_W, \xi', 2n) \leq |S| \leq \left( \frac{2}{\xi} \right)^m = \left( \frac{2mR}{\xi'} \right)^m.
\]
Taking log on both sides, we have
\[ \ln N_1(F_W, \xi', 2n) \leq mr \ln \left( \frac{2mR}{\xi'} \right). \]

The first part of Theorem 2 can be proven combining Theorem 9 and Lemma 5.

To prove the second part. Let \( F_{W, \varepsilon} \) be a minimal \( \varepsilon \) cover of \( F_W \). It can be easily verified that
\[ \sup_{f_w \in F_W} |E_{f_w} - E_{f_w}| \leq 2\varepsilon + \sup_{f_w, \varepsilon \in F_{W, \varepsilon}} |E_{f_w, \varepsilon} - E_{f_{W, \varepsilon}}|. \]

Using Hoeffding’s inequality and the union bound, we have
\[ P \left\{ \sup_{f_w, \varepsilon \in F_{W, \varepsilon}} |E_{f_w} - E_{f_w, \varepsilon}| \geq \xi \right\} \leq 2|F_{W, \varepsilon}| \exp(-2nR^2\xi^2). \]

Let
\[ 2|F_{W, \varepsilon}| \exp(-2nR^2\xi^2) = 2 \left( \frac{2mR}{\varepsilon} \right)^m \exp(-2nR^2\xi^2) = \delta. \]

Then, with probability at least \( 1 - \delta \), we have
\[ \sup_{f_w \in F_{W, \varepsilon}} |E_{f_w} - E_{f_w}| \leq R \sqrt{\frac{m \ln(2mR/\varepsilon) + \ln(2/\delta)}{2n}}. \]

Thus, with probability at least \( 1 - \delta \), we get
\[ \sup_{f_w \in F_W} |E_{f_w} - E_{f_w}| \leq 2\varepsilon + R \sqrt{\frac{m \ln(2mR/\varepsilon) + \ln(2/\delta)}{2n}}. \]

Let \( \varepsilon = 1/n \), with probability at least \( 1 - \delta \), we have
\[ \sup_{f_w \in F_W} |E_{f_w} - E_{f_w}| \leq \frac{2}{n} + R \sqrt{\frac{m \ln(2mnR) + \ln(2/\delta)}{2n}}. \]

This concludes the proof of the second part of Theorem 2. ■

E. Proof of Theorem 4

According to Theorem 3 of Gruber [41], asymptotical results have been proven for the distortion of high-resolution vector quantization.

Let \( C_1, \ldots, C_k \) be \( k \) measurable sets which tile a measurable set \( C \subset \mathbb{R}^m \) and \( \{c_1, \ldots, c_k \} \subset \mathbb{R}^m \) be a codebook. To each signal \( x \in C \) assign the codevector \( c_i \) if \( x \in C_i \). Let \( p \) be the density of the source which generates the signal \( x \). Then, the distortion is defined as
\[ \int_{C_i} \min_{i=1, \ldots, k} \{ f(||c_i - x||) \} p(x) dx, \]
where \( f \) is a measure (loss function) for the quality of the quantizer and \( || \cdot || \) is a norm on \( \mathbb{R}^m \).

The minimum distortion is given by
\[ \inf_{C_1, \ldots, C_k} \left\{ \int_{C_i=1, \ldots, r} \min_{x \in \{c_i \}} \{ f(||c_i - x||) \} p(x) dx \right\}. \]

The following theorem describes the asymptotic result of the minimum distortion of vector quantization [41].

**Theorem 10:** Let \( f : [0, +\infty) \to [0, +\infty) \) satisfying \( f(0) = 0 \), \( f \) is continuous and strictly increasing and, for any given \( s > 1 \), the quotient \( f(st)/f(t) \) is decreasing and bounded above for \( t > 0 \). Then there are constants \( A, B > 0 \), depending only on \( f \) and \( \| \cdot \| \), such that the following statement holds: let \( C \subset \mathbb{R}^m \) be compact and measurable with \( |C| > 0 \) and \( p \) be the distribution density function of the source which generates \( x \). Then (8) is asymptotically equal to
\[ B \left( \int_X p(x) \frac{A^s}{m} f(r^{-\frac{1}{m}}) dx \right), \]
as \( r \to \infty \).

Now, we are ready to prove Theorem 4.

**Proof of Theorem 4.** MahNMF can be viewed as a coding scheme. Therefore, we can analyze MahNMF using Theorem 10. We relate the reconstruction error and the distortion of quantization as follows:

\[ R(W_r) = \min_{W \in \mathbb{R}^{m \times r}} \int_X f_W(x) dp = \min_{W \in \mathbb{R}^{m \times r}} \int_X \min_{h \in \mathbb{R}^r} \|x - Wh\|_1 p(x) dx \]
\[ \leq \min_{W \in \mathbb{R}^{m \times r}} \int_X \min_{x' \in \{W_1, \ldots, W_r \}} \|x - x'\|_1 p(x) dx \]
\[ = \int_X \min_{x' \in \{W_1, \ldots, W_r \}} \|x - x'\|_1 p(x) dx. \]

We can carefully choose \( \{W_1, \ldots, W_r \} \subset \mathbb{R}^m \) such that
\[ \int_X \min_{x' \in \{W_1, \ldots, W_r \}} \|x - x'\|_1 p(x) dx \]
is equal to the quantization distortion problem (8). Then, the minimum of \( R(W_r) \) is upper bounded by the minimum distortion.

Let \( f(x) = 1 \) and \( \| \cdot \|_1 = \| \cdot \|_1 \). It can be easily verified that \( f \) satisfies \( f : [0, +\infty) \to [0, +\infty) \), \( f(0) = 0 \), \( f \) is continuous and strictly increasing and, for any given \( s > 1 \), the quotient \( f(st)/f(t) \) is non-increasing and bounded above for \( t > 0 \). According to Theorem 10, the approximation error has an asymptotical bound as:
\[ R(W_r) \leq B \left( \int_X p(x) \frac{A^s}{m} f(r^{-\frac{1}{m}}) dx \right). \]

Since the reduced dimensionality \( r < \min(m, n) \), to use Theorem 10, we have to assume that the feature dimensionality \( m \) is sufficiently large. Thus, when \( m \) is sufficiently large and \( r \) approaches \( m \), it holds that \( R(W_r) \leq Br^{-\frac{1}{m}} \). When \( m \) is finite and \( r \) approaches \( m \), it trivially holds that \( R(W_r) = 0 \leq \)
by using the asymptotic results of the minimum distortion of vector quantization. Moreover, we proved that the bound is tight regarding to \( r \), and in doing so revealed a clear relationship between the approximation error and reduced dimensionality for MahNMF.

**APPENDIX A**

**THE OUTLINE OF THE PROOF OF THEOREM 9**

The outline of proof is of four steps.

1. **Symmetrization** (see a detailed proof in [59]). Let \( X' \) and \( X \) be independent and identically distributed sample sets drawn from \( X \), respectively. For any \( n \geq \frac{8(b-a)^2}{\xi^2} \), we have

\[
\Pr \left\{ \sup_{f \in F} \left| E_f - E_n f \right| \geq \xi \right\} \\
\leq 2 \Pr \left\{ \sup_{f \in F} \left| E'_n f - E_n f \right| \geq \xi / 2 \right\},
\]

where \( E'_n f = \frac{1}{n} \sum_{i=1}^{n} f(x'_i) \).

2. **Permutation.** Using the symmetric distribution property of \( f(x'_i) - f(x_i) \) and the union bound, it gives

\[
\Pr \left\{ \sup_{f \in F} \left| E'_n f - E_n f \right| \geq \xi / 2 \right\} \\
\leq 2 \Pr \left\{ \sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(x_i) \geq \xi / 4 \right\}.
\]

3. **Combination.** Setting \( G \) be an \( \xi / 8 \)-cover of \( F \) with respect to \( \ell_1 \) norm and applying the union bound, we have

\[
\Pr \left\{ \sup_{g \in \tilde{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(x_i) \geq \xi / 8 \right\} \\
\leq E_{N_1}(F, \xi / 8, X_1^{2n}) \Pr \left\{ \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(x_i) \geq \xi / 8 \right\}.
\]

4. **Concentration.** By the union bound of the Hoeffding’s inequality, we get

\[
\Pr \left\{ \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(x_i) \geq \xi / 8 \right\} \leq 2 \exp \left( -\frac{n\xi^2}{32(b-a)^2} \right).
\]

At last, combining the four steps, we have

\[
\Pr \left\{ \sup_{f \in F} \left| E_f - E_n f \right| \geq \xi \right\} \\
\leq 8 E_{N_1}(F, \xi / 8, X_1^{2n}) \exp \left( -\frac{n\xi^2}{32(b-a)^2} \right).
\]
APPENDIX B
THE ROBUSTNESS AND PERFORMANCE OF MAHNMF

In Section II-B we theoretically analyzed that MahNMF is more robust (to noise) than NMF and KLNMF. Guan et al. [14] had conducted extensive empirical experiments to validate this point. In this appendix, we borrow one reconstruction experiment on the PIE dataset [60] from [14] to show both the robustness and good performance of MahNMF.

The relative reconstruction error is defined as $\|X - X'\|^2 / \|X\|^2$, wherein $X$ and $X'$ denote the original image and reconstructed image, respectively. The smallest reconstruction error in each subline of Table I is shown in bold. The experiment settings are referred to [14]. Table I shows that MahNMF reconstructs the face images better in both the training and test sets when the training set is contaminated by occlusion, Laplace noise, salt and pepper noise and Gaussian noise. MahNMF therefore successfully handles the heavy-tailed noise and performs robustly in the presence of outliers. We do not repeat the other experiment results and suggest readers refer to [14] for detailed information.

ACKNOWLEDGMENT

The authors would like to thank Dr. Naiyang Guan for providing useful suggestions and codes of MahNMF.

REFERENCES
