Revisiting Maximal Response-Based Local Identification of Overcomplete Dictionaries

Zahra Shakeri and Waheed U. Bajwa
Dept. of Electrical and Computer Engineering, Rutgers University, Piscataway, New Jersey 08854
{zahra.shakeri, waheed.bajwa}@rutgers.edu

***Abstract*—This paper revisits the problem of recovery of an overcomplete dictionary in a local neighborhood from training samples using the so-called maximal response criterion (MRC). While it is known in the literature that MRC can be used for asymptotic exact recovery of a dictionary in a local neighborhood, those results do not allow for linear (in the ambient dimension) scaling of sparsity levels in signal representations. In this paper, a new proof technique is leveraged to establish that MRC can in fact handle linear sparsity (modulo a logarithmic factor) of signal representations. While the focus of this work is on asymptotic exact recovery, the same ideas can be used in a straightforward manner to strengthen the original MRC-based results involving noisy observations and finite number of training samples.**

I. Introduction

Dictionary learning is the problem of obtaining an overcomplete basis that results in sparse representations of signals. These sparse representations can then be used in a variety of applications, such as denoising [1], [2], classification [3], [4], and compressed sensing [5]. While initial focus in the literature has been on developing efficient algorithms for dictionary learning, it is important to also understand the performance of such algorithms theoretically.

Some recent works that focus on the theoretical aspects of dictionary learning algorithms and the required sample complexity for reliable recovery of the true dictionary include [6]–[16]. Among these works, [8], [12] focus on square dictionaries, while the rest study overcomplete dictionaries. In [6]–[11], global identification results are obtained while in [12]–[16] local identifiability is studied.

The focus of this paper is on a relatively-new maximization criterion proposed in [16] for dictionary learning called the *maximal response criterion* (MRC). Such a criterion not only leads to efficient computational algorithms for dictionary learning [17], but it is also shown in [16] that this new criterion results in provable local recovery of an $m \times p$ dictionary from training signals. Sample complexity results for dictionary learning under both noiseless and noisy settings are also provided in [16]. The common thread underlying these results is a decay constraint on sparse representations of the signals, which is a crucial element in the arguments used throughout [16]. Unfortunately, even in the best setting, the decay condition stated in [16] dictates that if the m-dimensional training signals have S-sparse representations in the dictionary then one must have $S = \mathcal{O}(\sqrt{m})$. Nonetheless, it is suggested in [16] that it may be possible to break this "square-root bottleneck" using different proof techniques (although no formal arguments are provided). Additionally, while the focus in [16] is on theoretical aspects of MRC, it is shown in [17] that efficient computational algorithms based on the MRC, collectively referred to as *iterative thresholding and K-means* (ITKM) algorithms, have strong (theoretical and experimental) convergence properties.

The work of the authors was supported under awards from the Army Research Office (W911NF-14-1-0295) and the National Science Foundation (CCF-1453073, CCF-1525276, and CCF-1218942).

In this paper, we revisit the MRC in [16] for dictionary learning and obtain an alternative decay condition on the coefficients of the sparse representations that is less restrictive than the one obtained in [16]. Specifically, the new decay condition allows us to break the square-root bottleneck in the sense that it can allow for asymptotic exact recovery of the true dictionary even if the sparse representations of the signals satisfy $S = \mathcal{O}(\frac{m}{\log p})$. Similar to [16], our focus here is on local analysis, i.e., we establish that there exists a neighborhood around the true dictionary in which only the true dictionary maximizes the objective function. Our new condition also results in a larger neighborhood compared to the one given in [16]. Our proofs rely on a new measure of dictionary coherence studied in [18], [19] as well as the *method of bounded differences* [20] and a complex variant of *Azuma's inequality* [21]. We conclude by noting that our proof techniques can be used in a straightforward manner to also strengthen the results reported in [16] for dictionary learning in both noisy and finite sample settings.

Notational Convention: Bold upper-case, bold lower-case, and lower-case letters denote matrices, vectors, and scalars, respectively. The k-th column of a matrix $\mathbf{X}$ is denoted by $\mathbf{x}_k$, $\mathbf{X}_{\mathcal{I}}$ is the matrix consisting of columns of $\mathbf{X}$ with indices $\mathcal{I}$, v_i denotes the i-th element of a vector $\mathbf{v}$, and $\mathbf{e}_j$ denotes the j-th column of the identity matrix. Furthermore, $\mathbf{v}_1 \odot \mathbf{v}_2$ denotes the pointwise product of $\mathbf{v}_1$ and $\mathbf{v}_2$. We write $[K]$ for $\{1, \ldots, K\}$. For two matrices $\mathbf{A}$ and $\mathbf{B}$ of same dimensions $m \times p$, we define their distance to be $d(\mathbf{A}, \mathbf{B}) = \max_{i \in [p]} \|\mathbf{a}_i - \mathbf{b}_i\|_2$. For any matrix $\mathbf{X} \in \mathbb{R}^{m \times p}$ consisting of unit-norm columns, we define its *worst-case coherence* as $\mu = \max_{\substack{i,j \in [p] \\ i \neq j}} |\langle \mathbf{x}_i, \mathbf{x}_j \rangle|$ and its *average coherence* as $\nu = \frac{1}{p-1} \max_{i \in [p]} \big| \sum_{\substack{j \in [p] \\ j \neq i}} \langle \mathbf{x}_i, \mathbf{x}_j \rangle \big|$, where $\langle \cdot, \cdot \rangle$ denotes the inner product. Finally, we use the notation $f(\varepsilon) = \mathcal{O}(g(\varepsilon))$ if $\lim_{\varepsilon \to 0} f(\varepsilon)/g(\varepsilon) = c < \infty$ for some constant c.

II. System Model

In dictionary learning, we assume that an observation $\mathbf{y} \in \mathbb{R}^m$ is generated according to

$$\mathbf{y} = \mathbf{D}\mathbf{x} + \mathbf{n}, \tag{1}$$

where $\mathbf{D} \in \mathbb{R}^{m\times p}$ is a fixed dictionary, $\mathbf{x} \in \mathbb{R}^p$ is the signal coefficient vector, and $\mathbf{n} \in \mathbb{R}^m$ is the underlying noise vector. Given a signal matrix $\mathbf{Y}$ consisting of observations $\mathbf{y}_k$, $k \in [N]$, the goal is to find a representative dictionary, $\mathbf{D}^*$, and a coefficient matrix $\mathbf{X}^*$ consisting of signal coefficient vectors $\mathbf{x}_k^*$, $k \in [N]$, such that the representation error is minimized. In other words,

$$(\mathbf{D}^*, \mathbf{X}^*) = \underset{\mathbf{D}'\in\mathcal{D}, \mathbf{X}'\in\mathcal{X}}{\arg\min} \|\mathbf{Y} - \mathbf{D}'\mathbf{X}'\|_F^2. \tag{2}$$

Here, the dictionary class $\mathcal{D}$ is defined as

$$\mathcal{D} \triangleq \{\mathbf{D}' \in \mathbb{R}^{m\times p}, \|\mathbf{d}_j'\|_2 = 1, \forall j \in [p], \operatorname{rank}(\mathbf{D}') = m \le p\},$$

while we assume the coefficient vectors are sparse, i.e.,

$$\mathcal{X} \triangleq \{\mathbf{X}' \in \mathbb{R}^{p\times N}, \|\mathbf{x}_j'\|_0 \le S, \forall j \in [N]\}, \tag{3}$$

where S denotes the sparsity of the coefficient vectors and it is assumed that $S \ll m$.

Similar to [16], we solve (2) for $\mathbf{D}$ using the MRC:

$$\mathbf{D}^* = \max_{\mathbf{D}\in\mathcal{D}} \sum_{k\in[N]} \max_{|\mathcal{I}|=S} \|\mathbf{D}_{\mathcal{I}}^T\mathbf{y}_k\|_1. \tag{4}$$

Note that (4) is maximizing the ℓ_1 norm of the S largest responses (inner products of dictionary columns and observations), and can be interpreted as a generalization of the K-means objective function [16]. Our focus in this work is on the asymptotic version of (4), which can be stated as

$$\mathbf{D}^* = \max_{\mathbf{D}\in\mathcal{D}} \mathbb{E}_{\mathbf{y}} \left\{ \max_{|\mathcal{I}|=S} \|\mathbf{D}_{\mathcal{I}}^T\mathbf{y}\|_1 \right\}. \tag{5}$$

In [16], local identifiability results are obtained using the MRC for dictionaries generated from randomly sparse signal coefficients in the presence of noise. To describe that signal coefficient model, we consider a sequence $\mathbf{c} \in \mathbb{R}^p$ satisfying

$$c_1 \ge c_2 \ge \cdots \ge c_p \ge 0, \ \|\mathbf{c}\|_2 = 1. \tag{6}$$

We construct the signal coefficient vectors using the relation

$$\mathbf{x} = \mathbf{q} \odot \mathbf{P}\mathbf{c}, \tag{7}$$

where $\mathbf{P} \in \mathbb{R}^{p\times p}$ is a random permutation matrix, $\mathbf{P} = [\mathbf{e}_{\pi(1)}, \mathbf{e}_{\pi(2)}, \ldots, \mathbf{e}_{\pi(p)}]^T$ for random permutation $(\pi_1, \ldots, \pi_p)$, and $\mathbf{q} \in \mathbb{R}^p$ is a sign vector with elements taking values ± 1 randomly. In this case, given $\mathbf{c}$, the coefficient vector $\mathbf{x}$ takes a particular value with probability $\frac{1}{2^p p!}$. Note that while we do not require $\mathbf{x}$ to be sparse, an additional constraint on the decay of the elements of $\mathbf{c}$ will be imposed to prove identifiablity results for the underlying dictionary.

III. Asymptotic Identifiability Results

In this section, we prove a variant of Proposition 6 in [16]. While this result is for the most basic setting where noise is not present, the proof technique can be used in all theorems in [16] to improve the results stated in there.

Theorem 1. *Consider observations generated via* (1) *with noise variance $\sigma = 0$. Let $\mathbf{D} \in \mathcal{D}$ be a dictionary with worst-case coherence μ and average coherence ν, and let $\mathbf{x}$ be the signal coefficient vector generated according to* (7)*. If $\nu \le \mu\sqrt{\frac{\log p}{p}}$ and $\mathbf{c}$ satisfies*

$$c_S > c_{S+1} + 26\mu\sqrt{\log p}, \tag{8}$$

then there is a local maximum of (5) *at $\mathbf{D}$ with high probability. Moreover, for any perturbation of the true dictionary, $\widetilde{\mathbf{D}} = (\widetilde{\mathbf{d}}_1, \ldots, \widetilde{\mathbf{d}}_p)$ with $d(\mathbf{D}, \widetilde{\mathbf{D}}) \le \varepsilon$, we have $\mathbb{E}_{\mathbf{y}}\{\max_{|\mathcal{I}|=S} \|\widetilde{\mathbf{D}}_{\mathcal{I}}^T\mathbf{y}\|_1^2\} < \mathbb{E}_{\mathbf{y}}\{\max_{|\mathcal{I}|=S} \|\mathbf{D}_{\mathcal{I}}^T\mathbf{y}\|_1^2\}$ with high probability as soon as*

$$\varepsilon \le \frac{c_S - c_{S+1} - 26\mu\sqrt{\log p}}{1 + 3\sqrt{\log\left(\frac{25p^2S\sqrt{B}}{(c_S - c_{S+1} - 26\mu\sqrt{\log p})(\sum_{i\in[S]} c_i)}\right)}}, \tag{9}$$

where $\sqrt{B}$ is the largest singular value of $\mathbf{D}$.

Outline of Proof: The main steps for the proof of the theorem follow from the steps taken in [16]. Specifically, we show that for a fixed permutation, the maximal response is obtained by $\mathbf{D}_{\mathcal{I}_s}$, where $\mathcal{I}_s$ denotes the indices of the coefficient vector elements corresponding to $\{c_i\}_{i\in[S]}$. The biggest difference between our proof and that in [16] is that we introduce the decaying condition in (8), which is less restrictive than the decaying condition in [16] for the decay of elements of $\mathbf{c}$. The rest of the proof is similar to the proof of Proposition 6 in [16]. For ε-perturbations of the original dictionary, i.e., $d(\mathbf{D}, \widetilde{\mathbf{D}}) \le \varepsilon$, we show that for small perturbations of the original dictionary and most sign sequences, the maximal response is obtained by $\widetilde{\mathbf{D}}_{\mathcal{I}_s}$. Then, comparing $\mathbb{E}_{\mathbf{y}}\{\max_{|\mathcal{I}|=S} \|\widetilde{\mathbf{D}}_{\mathcal{I}}^T\mathbf{y}\|_1^2\}$ with $\mathbb{E}_{\mathbf{y}}\{\max_{|\mathcal{I}|=S} \|\mathbf{D}_{\mathcal{I}}^T\mathbf{y}\|_1^2\}$, it is shown that (9) ensures that $\mathbf{D}$ maximizes (5) with high probability.

The technical proof of Theorem 1 relies on the following lemma, whose proof is provided in the appendix.

Lemma 1. *Consider observations generated according to* (1) *with noise variance $\sigma = 0$, where the dictionary $\mathbf{D} \in \mathcal{D}$ has worst-case coherence μ and average coherence ν, and let $\mathbf{x}$ be generated according to* (7)*. Then, for any $i \in [p]$, any $\varepsilon > 0$, and any p satisfying $\sqrt{p} \le \varepsilon/\nu$, we have*

$$\mathbb{P}\left\{\left|\sum_{\substack{j\in[p]\\ j\ne i}} x_j\langle\mathbf{d}_i, \mathbf{d}_j\rangle\right| > \varepsilon\right\} \le 4\exp\left(-\frac{(\varepsilon - \nu\sqrt{p})^2}{144\mu^2}\right). \tag{10}$$

Proof of Theorem 1. The objective function in (5) can be restated as

$$\begin{aligned}\mathbb{E}_{\mathbf{y}}\left\{\max_{|\mathcal{I}|=S} \|\mathbf{D}_{\mathcal{I}}^T\mathbf{y}\|_1\right\} &= \mathbb{E}_{\pi}\mathbb{E}_{\mathbf{q}}\left\{\max_{|\mathcal{I}|=S} \|\mathbf{D}_{\mathcal{I}}^T\mathbf{D}\mathbf{x}\|_1\right\} \\ &= \mathbb{E}_{\pi}\mathbb{E}_{\mathbf{q}}\left\{\max_{|\mathcal{I}|=S} \sum_{i\in\mathcal{I}} |\langle\mathbf{d}_i, \mathbf{D}\mathbf{x}\rangle|\right\}.\end{aligned} \tag{11}$$

We now show that the maximum of (11) is obtained via $\mathcal{I} = \mathcal{I}_s$, where $\mathcal{I}_s = \pi^{-1}(\{1, 2, \ldots, S\})$. Selecting $\varepsilon = 13\mu\sqrt{\log p}$, as long as the condition $\nu \leq \mu\sqrt{\frac{\log p}{p}}$ is satisfied, we have $\varepsilon - \nu\sqrt{p} \geq 0$ and $\exp\left(-\frac{(\varepsilon-\nu\sqrt{p})^2}{144\mu^2}\right) \leq p^{-1}$. Therefore, with high probability, for any $i \in \mathcal{I}_s$, we have

$$|\langle \mathbf{d}_i, \mathbf{Dx}\rangle| = \left| q_i c_{\pi(i)} + \sum_{\substack{i\in[p]\\ j\neq i}} q_j c_{\pi(j)} \langle \mathbf{d}_i, \mathbf{d}_j\rangle \right|$$
$$\overset{(a)}{\geq} c_S - \left| \sum_{\substack{i\in[p]\\ j\neq i}} q_j c_{\pi(j)} \langle \mathbf{d}_i, \mathbf{d}_j\rangle \right| \overset{(b)}{\geq} c_S - 13\mu\sqrt{\log p}, \quad (12)$$

where (a) follows from the triangle inequality and (b) follows from substituting $\varepsilon = 13\mu\sqrt{\log p}$ in (10). Similarly, for all $i \notin \mathcal{I}_s$, we have

$$\begin{aligned} |\langle \mathbf{d}_i, \mathbf{Dx}\rangle| &= \left| q_i c_i + \sum_{\substack{i\in[p]\\ j\neq i}} q_j c_{\pi(j)} \langle \mathbf{d}_i, \mathbf{d}_j\rangle \right| \\ &\leq c_{S+1} + \left| \sum_{\substack{i\in[p]\\ j\neq i}} q_j c_{\pi(j)} \langle \mathbf{d}_i, \mathbf{d}_j\rangle \right| \\ &\leq c_{S+1} + 13\mu\sqrt{\log p}, \end{aligned} \quad (13)$$

with high probability. Thus, (8) ensures the maximum of the objective function is attained at $\mathcal{I}_s$. The rest of the proof is the same as the proof of Proposition 6 in [16], in which wherever $\mu\|\mathbf{c}\|_1$ appears, it can be replaced by $13\mu\sqrt{\log p}$. □

IV. Discussion and Conclusion

A natural question to ask about Theorem 1 is whether there exist dictionaries that satisfy the $\nu \leq \mu\sqrt{\frac{\log p}{p}}$ condition. In this regard, note that this condition is implied by conditions $\frac{p}{\log p} \leq m$ and $\nu \leq \frac{\mu}{\sqrt{m}}$ and according to [18], there exist dictionaries, such as Gaussian matrices, that satisfy $\nu \leq \frac{\mu}{\sqrt{m}}$.

Next, to analyze our result and compare it to the analogous result in [16], we study the basic setting where $\mathbf{c}$ is S-sparse and $\{c_i\}_{i=1}^S = \frac{1}{\sqrt{S}}$, resulting in $\|\mathbf{c}\|_1 = \sqrt{S}$. According to the decay condition in [16], $c_S > c_{S+1} + 2\mu\|\mathbf{c}\|_1$, which guarantees recovery of the true dictionary as long as $S < \frac{1}{2\mu}$. From the Welch bound [22], this at best translates to sparsity levels of order $\mathcal{O}(\sqrt{m})$. With the new decay condition $c_S > c_{S+1} + 26\mu\sqrt{\log p}$, we can recover the true dictionary even if the sparsity levels are of order $\mathcal{O}(\frac{m}{\log p})$. Thus, our analysis is able to overcome the fundamental limitations of [16], where regardless of the dictionary, there is a square-root bottleneck for S.

We conclude by noting that although we have only studied the noiseless asymptotic case, the decay condition for the coefficient vector can also be used in noisy and finite sample settings. And while we have not discussed the computational aspects of MRC-based dictionary learning algorithms, such analysis and discussion has been carried out in [17] that shows that the ITKM algorithms originating from the MRC have low complexity and strong convergence properties.

Appendix

Lemma 2 (The Complex Azuma Inequality [18])**.** *Assuming the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and let $\widetilde{M}_1, \ldots, \widetilde{M}_n$ be a complex-valued martingale difference sequence on $(\Omega, \mathcal{F}, \mathbb{P})$ with $|\widetilde{M}_i| \leq c_i$ for $i \in [n]$. Then for any $t > 0$,*

$$\mathbb{P}\left\{ \left| \sum_{i\in[n]} \widetilde{M}_i \right| \geq t \right\} \leq 4\exp\left(-\frac{t^2}{4\sum_{i\in[n]} c_i^2} \right). \quad (14)$$

Proof of Lemma 1. The proof follows similar steps as Lemma 3 in [18]. The measurement vector $\mathbf{y}$ can be stated as

$$\mathbf{y} = \mathbf{Dx} = \mathbf{D}(\mathbf{q} \odot \mathbf{Pc}) = \mathbf{D}_\Pi(\mathbf{q}_\Pi \odot \mathbf{c}), \quad (15)$$

where $\Pi = \{\pi(i)\}_{i=1}^p$, $\mathbf{D}_\Pi$ is the column-wise permuted version of $\mathbf{D}$, and $\mathbf{q}_\Pi$ is the permuted version of $\mathbf{q}$. We introduce the method of bounded differences (MOBD) [20] that uses Azuma's inequality for bounded martingale difference sequences (BMDS). For a fixed index i, conditioned on the event $\mathcal{A}_{i'} = \{\pi(i) = i'\}$ and the sign vector $\mathbf{q}$, writing the coefficient vector elements as $x_j = q_j c_{\pi(j)}, j \in [p]$, we get

$$\begin{aligned} &\mathbb{P}\left\{ \left| \sum_{\substack{j\in[p]\\ j\neq i'}} q_j c_{\pi(j)} \langle \mathbf{d}_{\pi(i)}, \mathbf{d}_j\rangle \right| > \varepsilon \middle| \mathcal{A}_{i'}, \mathbf{q} \right\} \\ &\quad = \mathbb{P}\left\{ \left| \sum_{\substack{j\in[p]\\ j\neq i}} q_{\pi(j)} c_j \langle \mathbf{d}_{i'}, \mathbf{d}_{\pi(j)}\rangle \right| > \varepsilon \middle| \mathcal{A}_{i'}, \mathbf{q} \right\}. \end{aligned} \quad (16)$$

To obtain an upper bound for (16), we define a random $(p-1)$-tuple $\Pi^{-i} = \{\pi(k)\}_{k=1}^p, k \neq i$ and construct a Doob Martingale $(M_0, M_1, \ldots, M_{p-1})$:

$$M_0 = \mathbb{E}\left\{ \sum_{\substack{j\in[p]\\ j\neq i}} q_{\pi(j)} c_j \langle \mathbf{d}_{i'}, \mathbf{d}_{\pi(j)}\rangle \middle| \mathcal{A}_{i'}, \mathbf{q} \right\}, \text{and}$$
$$M_\ell = \mathbb{E}\left\{ \sum_{\substack{j\in[p]\\ j\neq i}} q_{\pi(j)} c_j \langle \mathbf{d}_{i'}, \mathbf{d}_{\pi(j)}\rangle \middle| \{\pi_k^{-i}\}_{k=1}^\ell, \mathcal{A}_{i'},\ \mathbf{q} \right\},$$

for $\ell \in [p-1]$, where $\{\pi_k^{-i}\}_{k=1}^\ell$ denotes the first ℓ elements of $\prod^{-i}$. Similar to [18], we can bound $|M_0|$ by

$$\begin{aligned} |M_0| &= \left| \mathbb{E}\left\{ \sum_{\substack{j\in[p]\\ j\neq i}} q_{\pi(j)} c_j \langle \mathbf{d}_{i'}, \mathbf{d}_{\pi(j)}\rangle \middle| \mathcal{A}_{i'}, \mathbf{q} \right\} \right| \\ &\leq \sum_{\substack{j\in[p]\\ j\neq i}} \left| q_{\pi(j)} c_j \mathbb{E}\left\{ \langle \mathbf{d}_{i'}, \mathbf{d}_{\pi(j)}\rangle \middle| \mathcal{A}_{i'}, \mathbf{q} \right\} \right| \\ &\leq \sum_{\substack{j\in[p]\\ j\neq i}} c_j \left| \sum_{\substack{q\in[p]\\ q\neq i'}} \frac{\langle \mathbf{d}_{i'}, \mathbf{d}_q\rangle}{p-1} \right| \leq \nu\|\mathbf{c}\|_1 \leq \nu\sqrt{p}. \end{aligned} \quad (17)$$

In order to utilize Azuma's Inequality, we have to construct a BMDS from $(M_0, \ldots, M_{p-1})$. Defining $\widetilde{M}_\ell = M_\ell - M_{\ell-1}$ for $\ell \in [p-1]$, it is necessary to find an upper bound on $|\widetilde{M}_\ell|$. According to [23], we have $|\widetilde{M}_\ell| \leq \sup_{r,s}[M_\ell(r) - M_\ell(s)]$ where for $\ell \in [p-1]$, $M_\ell(r)$ is defined as $M_\ell(r) \triangleq \mathbb{E}\left\{ \sum_{j\in[p]\, j\neq i} q_{\pi(j)} c_j \langle \mathbf{d}_{i'}, \mathbf{d}_{\pi(j)}\rangle \middle| \{\pi_k^{-i}\}_{k=1}^{\ell-1}, \pi_\ell^{-i} = r, \mathcal{A}_{i'}, \mathbf{q} \right\}$.

To find an upper bound for $|M_\ell(r) - M_\ell(s)|$, we have

$$\begin{aligned}
&|M_\ell(r) - M_\ell(s)| \\
&= \Big| \sum_{\substack{j\in[p]\\ j\neq i}} q_{\pi(j)} c_j \big(\mathbb{E}\big[\langle \mathbf{d}_{i'}, \mathbf{d}_{\pi(j)}\rangle \big| \{\pi_k^{-i}\}_{k=1}^{\ell-1}, \pi_\ell^{-i} = r, \mathcal{A}_{i'}, \mathbf{q}\big] \\
&\quad - \mathbb{E}\big\{ \langle \mathbf{d}_{i'}, \mathbf{d}_{\pi(j)}\rangle \big| \{\pi_k^{-i}\}_{k=1}^{\ell-1}, \pi_\ell^{-i} = s, \mathcal{A}_{i'}, \mathbf{q}\big\}\big) \Big| \\
&\le \sum_{\substack{j\in[p]\\ j\neq i}} c_j \Big| \mathbb{E}\left\{ \langle \mathbf{d}_{i'}, \mathbf{d}_{\pi(j)}\rangle \big| \{\pi_k^{-i}\}_{k=1}^{\ell-1}, \pi_\ell^{-i} = r, \mathcal{A}_{i'}, \mathbf{q} \right\} \\
&\quad - \mathbb{E}\left\{ \langle \mathbf{d}_{i'}, \mathbf{d}_{\pi(j)}\rangle \big| \{\pi_k^{-i}\}_{k=1}^{\ell-1}, \pi_\ell^{-i} = s, \mathcal{A}_{i'}, \mathbf{q}\right\} \Big| \\
&= \sum_{\substack{j\le \ell+1\\ j\neq i}} c_j |d_{\ell,j}| + \sum_{\substack{j> \ell+1\\ j\neq i}} c_j |d_{\ell,j}|, \qquad (18)
\end{aligned}$$

where $d_{l,j} \triangleq \mathbb{E}\left\{ \langle \mathbf{d}_{i'}, \mathbf{d}_{\pi(j)}\rangle \big| \{\pi_k^{-i}\}_{k=1}^{\ell-1}, \pi_\ell^{-i} = r, \mathcal{A}_{i'}, \mathbf{q}\right\} - \mathbb{E}\left\{ \langle \mathbf{d}_{i'}, \mathbf{d}_{\pi(j)}\rangle \big| \{\pi_k^{-i}\}_{k=1}^{\ell-1}, \pi_\ell^{-i} = s, \mathcal{A}_{i'}, \mathbf{q}\right\}$.

We consider various cases to upper bound (18). For the case where $\ell \notin [p-3]$, Π is deterministic. In this case, if $i \le \ell$,

$$\sum_{\substack{j\in[\ell+1]\\ j\neq i}} c_j |d_{\ell,j}| = c_{\ell+1} |\langle \mathbf{d}_{i'}, \mathbf{d}_r\rangle - \langle \mathbf{d}_{i'}, \mathbf{d}_s\rangle| \le 2\mu c_{\ell+1}. \qquad (19)$$

Similarly, if $i > \ell$, $\sum_{\substack{j\in[\ell+1]\\ j\neq i}} c_j |d_{\ell,j}| \le 2\mu c_\ell$. If $\ell \in [p-3]$, for any $j > \ell+1, j \neq i$, $\pi(j)$ has a uniform distribution over $[p] - \{\{\pi_k^{-i}\}_{k=1}^{\ell-1}, \pi_\ell^{-i} = r, \mathcal{A}_{i'}\}$ and $[p] - \{\{\pi_k^{-i}\}_{k=1}^{\ell-1}, \pi_\ell^{-i} = s, \mathcal{A}_{i'}\}$, conditioned on $\{\{\pi_k^{-i}\}_{k=1}^{\ell-1}, \pi_\ell^{-i} = r, \mathcal{A}_{i'}\}$ and $\{\{\pi_k^{-i}\}_{k=1}^{\ell-1}, \pi_\ell^{-i} = s, \mathcal{A}_{i'}\}$, respectively and we have

$$|d_{\ell,j}| = \frac{1}{p-\ell-1} |\langle \mathbf{d}_{i'}, \mathbf{d}_r\rangle - \langle \mathbf{d}_{i'}, \mathbf{d}_s\rangle| \le \frac{2\mu}{p-\ell-1}. \qquad (20)$$

If $\ell \in [p-3]$, for any $j \le \ell+1$, we study three cases for i. If $i < \ell$, $\sum_{\substack{j\in[\ell+1]\\ j\neq i}} c_j |d_{\ell,j}| \le 2\mu c_{\ell+1}$, if $i = \ell$, $\sum_{\substack{j\in[\ell+1]\\ j\neq i}} c_j |d_{\ell,j}| \le 2\mu c_\ell$ and if $i > \ell+1$, $\sum_{\substack{j\in[\ell+1]\\ j\neq i}} c_j |d_{\ell,j}| \le 2\mu(c_\ell + \frac{c_{\ell+1}}{p-\ell-1})$. Denoting $d_\ell \triangleq \sum_{\substack{j\in[p]\\ j\neq i}} c_j |d_{\ell,j}|$, we have $\sup_{r,s}[M_\ell(r) - M_\ell(s)] \le 2\mu d_\ell$, where

$$d_\ell = \begin{cases} c_\ell + c_{\ell+1} + \dfrac{1}{p-\ell-1} \sum_{j=\ell+2}^{p} c_j, & \ell \in [p-3], \\ c_\ell & \ell \notin [p-3]. \end{cases}$$

To use the complex Azuma inequality, it is necessary to upper bound $\sum_{\ell\in[p-1]} d_\ell^2$:

$$\begin{aligned}
&\sum_{\ell\in[p-1]} d_\ell^2 \\
&= \sum_{\ell\in[p-3]} \big(c_\ell + c_{\ell+1} + \frac{1}{p-\ell-1} \sum_{j=\ell+2}^{p} c_j\big)^2 + \sum_{\ell=p-2}^{p-1} c_\ell^2 \\
&= \sum_{\ell\in[p-3]} \Big(c_\ell^2 + c_{\ell+1}^2 + 2c_\ell c_{\ell+1} + \frac{2(c_\ell + c_{\ell+1})}{p-\ell-1} \sum_{j=\ell+2}^{p} c_j \\
&\quad + \big(\frac{1}{p-\ell-1} \sum_{j=\ell+2}^{p} c_j\big)^2 \Big) + c_{p-2}^2 + c_{p-1}^2. \qquad (21)
\end{aligned}$$

Since $\mathbf{c}$ is non-negative and non-increasing, $2c_\ell c_{\ell+1} \le 2c_\ell$ and we can write

$$\sum_{\ell=1}^{p-3} \big(c_\ell^2 + c_{\ell+1}^2 + 2c_\ell c_{\ell+1}\big) \le 4\|\mathbf{c}\|_2^2 - c_{p-2}^2 - c_{p-1}^2. \qquad (22)$$

Denoting $\|\mathbf{c}\|_1^{-n} \triangleq \|\mathbf{c}\|_1 - \sum_{i\in[n]} c_i$, which has $p-n$ elements, we have $\|\mathbf{c}\|_1^{-n} \le (p-n)c_{n+1}$. Therefore,

$$\begin{aligned}
\sum_{\ell\in[p-3]} \frac{2(c_\ell + c_{\ell+1})}{p-\ell-1} \sum_{j=\ell+2}^{p} c_j &\le \sum_{\ell\in[p-3]} \frac{4c_\ell \|\mathbf{c}\|_1^{-(\ell+1)}}{p-\ell-1} \\
&\le \sum_{\ell\in[p-3]} \frac{4c_\ell(p-\ell-1)c_{\ell+2}}{p-\ell-1} = \sum_{\ell\in[p-3]} 4c_\ell c_{\ell+2} \le 4\|\mathbf{c}\|_2^2. \qquad (23)
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
\sum_{\ell\in[p-3]} \left(\frac{1}{p-\ell-1} \sum_{j=\ell+2}^{p} c_j\right)^2 &= \sum_{\ell\in[p-3]} \left(\frac{\|\mathbf{c}\|_1^{-(\ell+1)}}{p-\ell-1}\right)^2 \\
&\le \sum_{\ell\in[p-3]} c_{\ell+2}^2 \le \|\mathbf{c}\|_2^2. \qquad (24)
\end{aligned}$$

Adding the upper bounds in (22), (23), and (24) for (21) results in $\sum_{\ell\in[p-1]} d_\ell^2 \le 9\|\mathbf{c}\|_2^2$. We have established that $(\widetilde{M}_1, \ldots, \widetilde{M}_{p-1})$ is a BDMS with $|\widetilde{M}_\ell| \le 2\mu d_\ell$ for $\ell \in [p-1]$. We therefore have

$$\begin{aligned}
&\mathbb{P}\left\{ \Big| \sum_{\substack{j\in[p]\\ j\neq i}} q_{\pi(j)} c_j \langle \mathbf{d}_{i'}, \mathbf{d}_{\pi(j)}\rangle \Big| > \varepsilon \big| \mathcal{A}_{i'}, \mathbf{q} \right\} \\
&\overset{(a)}{\le} \mathbb{P}\left\{ |M_{p-1} - M_0| > \varepsilon\|\mathbf{c}\|_2 - \nu\sqrt{p}\|\mathbf{c}\|_2 \big| \mathcal{A}_{i'}, \mathbf{q}\right\} \\
&= \mathbb{P}\left\{ \Big| \sum_{i\in[p-1]} \widetilde{M}_i \Big| > \varepsilon\|\mathbf{c}\|_2 - \nu\sqrt{p}\|\mathbf{c}\|_2 \big| \mathcal{A}_{i'}, \mathbf{q} \right\} \\
&\overset{(b)}{\le} 4\exp\left(-\frac{(\varepsilon - \nu\sqrt{p})^2 \|\mathbf{c}\|_2^2}{16\mu^2 \sum_{\ell=1}^{p-1} d_\ell} \right) \\
&\le 4\exp\left(-\frac{(\varepsilon - \nu\sqrt{p})^2}{144\mu^2} \right), \qquad (25)
\end{aligned}$$

where (a) follows from (17) and (b) follows from the complex Azuma inequality for BDMS in Lemma 2. Taking the union bound over all events $\mathcal{A}_{i'}$ and sign sequences, we have

$$\begin{aligned}
&\mathbb{P}\left\{ \Big| \sum_{\substack{j\in[p]\\ j\neq i}} q_{\pi(j)} c_j \langle \mathbf{d}_{i'}, \mathbf{d}_{\pi(j)}\rangle \Big| > \varepsilon \right\} \\
&\le \sum_{j\in[p]} \sum_{i'\in[p]} \mathbb{P}\left\{ \Big| \sum_{\substack{j\in[p]\\ j\neq i}} q_{\pi(j)} c_j \langle \mathbf{d}_{i'}, \mathbf{d}_{\pi(j)}\rangle \Big| > \varepsilon\|\mathbf{c}\|_2 \big| \mathcal{A}_{i'}, \mathbf{q} \right\} \\
&\qquad \times \mathbb{P}(\mathcal{A}_{i'}) \mathbb{P}(q_j) \\
&\le 4\exp\left(-\frac{(\varepsilon - \nu\sqrt{p})^2}{144\mu^2} \right), \qquad (26)
\end{aligned}$$

where i' can be replaced with any $i \in [p], i \neq i'$ and the inequality holds for all i. □

REFERENCES

[1] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Transactions on Information Theory*, vol. 52, no. 1, pp. 6–18, 2006.

[2] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image Processing*, vol. 15, no. 12, pp. 3736–3745, 2006.

[3] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: transfer learning from unlabeled data," in *Proceedings of 24th international conference on Machine learning*. ACM, 2007, pp. 759–766.

[4] J. Mairal, J. Ponce, G. Sapiro, A. Zisserman, and F. R. Bach, "Supervised dictionary learning," in *Advances in neural information processing systems*, 2009, pp. 1033–1040.

[5] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[6] P. Georgiev, F. Theis, and A. Cichocki, "Sparse component analysis and blind source separation of underdetermined mixtures," *IEEE Transactions on Neural Networks*, vol. 16, no. 4, pp. 992–996, 2005.

[7] M. Aharon, M. Elad, and A. M. Bruckstein, "On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them," *Linear algebra and its applications*, vol. 416, no. 1, pp. 48–67, 2006.

[8] D. A. Spielman, H. Wang, and J. Wright, "Exact recovery of sparsely-used dictionaries," *Journal of Machine Learning Research*, vol. 23, pp. 37.1–37.18, 2012.

[9] A. Agarwal, A. Anandkumar, P. Jain, and P. Netrapalli, "Learning sparsely used overcomplete dictionaries via alternating minimization," *arXiv preprint arXiv:1310.7991*, 2013.

[10] A. Agarwal, A. Anandkumar, and P. Netrapalli, "Exact recovery of sparsely used overcomplete dictionaries," *arXiv preprint arXiv:1309.1952v1*, 2013.

[11] S. Arora, R. Ge, and A. Moitra, "New algorithms for learning incoherent and overcomplete dictionaries," in *Proceedings of 27th Conference on Learning Theory*, 2014, pp. 779–806.

[12] R. Gribonval and K. Schnass, "Dictionary identification-sparse matrix-factorisation via ℓ_1-minimisation," *IEEE Transactions on Information Theory*, vol. 56, no. 7, pp. 3523–3539, 2010.

[13] Q. Geng, H. Wang, and J. Wright, "On the local correctness of ℓ_1 minimization for dictionary learning," *arXiv preprint arXiv: 1101: 5672*, 2011.

[14] K. Schnass, "On the identifiability of overcomplete dictionaries via the minimisation principle underlying K-SVD," *Applied and Computational Harmonic Analysis*, vol. 37, no. 3, pp. 464–491, 2014.

[15] R. Gribonval, R. Jenatton, and F. Bach, "Sparse and spurious: dictionary learning with noise and outliers," *IEEE Transactions on Information Theory*, vol. 61, no. 11, pp. 6298–6319, 2015.

[16] K. Schnass, "Local identification of overcomplete dictionaries," *Journal of Machine Learning Research*, vol. 16, pp. 1211–1242, 2015.

[17] ——, "Convergence radius and sample complexity of ITKM algorithms for dictionary learning," *arXiv preprint arXiv:1503.07027*, 2015.

[18] W. U. Bajwa, R. Calderbank, and S. Jafarpour, "Why Gabor frames? Two fundamental measures of coherence and their role in model selection," *Journal of Communications and Networks*, vol. 12, no. 4, pp. 289–307, 2010.

[19] W. U. Bajwa, R. Calderbank, and D. G. Mixon, "Two are better than one: Fundamental parameters of frame coherence," *Appl. Comput. Harmon. Anal.*, vol. 33, no. 1, pp. 58–78, Jul. 2012.

[20] C. McDiarmid, "On the method of bounded differences," *Surveys in combinatorics*, vol. 141, no. 1, pp. 148–188, 1989.

[21] K. Azuma, "Weighted sums of certain dependent random variables," *Tohoku Math. J.*, vol. 19, no. 3, pp. 357–367, 1967.

[22] L. R. Welch, "Lower bounds on the maximum cross correlation of signals," *IEEE Transactions on Information Theory*, vol. 20, no. 3, pp. 397–399, 1974.

[23] R. Motwani and P. Raghavan, *Randomized algorithms*. Chapman & Hall/CRC, 2010.