DEEP CNN SPARSE CODING ANALYSIS: TOWARDS AVERAGE CASE

Michael Murray[†§], *Jared Tanner*[†§],

[†]University of Oxford, [§]Alan Turing Institute

ABSTRACT

Deep convolutional sparse coding (D-CSC) is a framework reminiscent of deep convolutional neural nets (DCNN), but by omitting the learning of the dictionaries one can more transparently analyse the role of the activation function and its ability to recover activation paths through the layers. Papyan, Romano, and Elad conducted an analysis of such an architecture [1], showed the relationship with DCNNs, and proved conditions under which a D-CSC is guaranteed to recover activation paths. A technical innovation of their work highlights that one can view the efficacy of the ReLU nonlinear activation function of a DCNN through the new variant of the tensor's sparsity, referred to as stripe-sparsity, and by which they can prove that the density of activations can be proportional to the ambient dimension of the data. We extend their uniform guarantees to a slightly modified model and prove that with high probability the desired activation is typically possible to recover for a greater density of activations per layer. Our extension follows from incorporating the prior work on one step thresholding by Schnass and Vandergheynst [2] into the appropriately modified architecture of [1].

Index Terms— Deep Learning, Sparse Coding, Deep Convolutional Sparse Coding, Rademacher Concentration.

1. INTRODUCTION

It has been posited that the enforcement of sparsity through ReLU nonlinear activation functions may be one of the key ingredients behind deep convolutional neural nets (DCNNs) effectiveness at developing composite representations at depth [3]. By omitting the learning of the dictionaries in DCNNs, one can more transparently analyse the role of the activation function and its ability to recover activation paths through the layers. Papyan, Romano, and Elad proposed a deep convolutional sparse coding (D-CSC) model with this aim and used techniques from sparse coding to prove conditions where a DCNN is guaranteed to activate the desired nodes in a given layer, [1]. They propose the model

$$\begin{aligned}
\hat{\mathbf{X}}^{(0)} &= \mathbf{A}^{(1)}\mathbf{D}^{(1)}\hat{\mathbf{X}}^{(1)} + \mathbf{V}^{(0)}\\
\hat{\mathbf{X}}^{(1)} &= \mathbf{A}^{(2)}\mathbf{D}^{(2)}\hat{\mathbf{X}}^{(2)} + \mathbf{V}^{(1)}\\
&\dots\\
\hat{\mathbf{X}}^{(L-1)} &= \mathbf{A}^{(L)}\mathbf{D}^{(L)}\hat{\mathbf{X}}^{(L)} + \mathbf{V}^{(L-1)}
\end{aligned} \tag{1}$$

but with $\mathbf{D}^{(l)}$ being the identity matrix and the other constituents:

- $\hat{\mathbf{X}}^{(0)} \in \mathbb{R}^{M\times d}$ is matrix containing the observed data, with each of the d columns a data point of dimension M. To be clear then: $\hat{\mathbf{X}}^{(0)} = [\hat{\mathbf{x}}_1^{(0)}\ \hat{\mathbf{x}}_2^{(0)}\ ...\ \hat{\mathbf{x}}_d^{(0)}]$ where throughout lower case vectors typically denote columns of their respective matrices,
- $\hat{\mathbf{X}}^{(l)} \in \mathbb{R}^{n_l M\times d}$ is a matrix containing the representation of the observed data $\hat{\mathbf{X}}^{(0)}$ at layer l,
- $\mathbf{A}^{(l)} \in \mathbb{R}^{(n_{l-1})M\times n_l M}$ is the transpose of the weight matrix mapping between layers $l-1$ and l. This matrix has a convolutional structure (described in [1] and [4]), is circular and banded, and is created by shifting a local dictionary $\mathbf{A}^{(l)}_{Local} \in \mathbb{R}^{m_l\times n_l}$. For $l \geq 2$ there is a stride between each spatially shifted $\mathbf{A}^{(l)}$ we denote s_l, in [1] $s_l = n_{l-1}$. The columns of $\mathbf{A}^{(l)}$ have unit ℓ^2 norm.
- $\mathbf{V}^{(l)} \in \mathbb{R}^{n_l M\times d}$ is the error matrix $\mathbf{V}^{(l)} = \hat{\mathbf{X}}^{(l)} - \mathbf{X}^{(l)}$.

In [1] Papyan, Romano, and Elad consider the $\hat{\mathbf{X}}^{(l)}$ obtained by applying a feed forward algorithm on Model (1) and prove that this model admits a solution under certain sparsity constraints. To achieve this they extend the traditional sparsity counting measure $\|\mathbf{x}\|_0$ which counts the number of nonzeros in x to the localized sparsity models:

- $\|\mathbf{x}\|_{0,\infty}^{P^{(l)}} = \max_i \|P_i^{(l)}\mathbf{x}\|_0$ where $P_i^{(l)}$, the patch operator, takes m_l consecutive elements of $\mathbf{x}$ starting at x_i
- $\|\mathbf{x}\|_{0,\infty}^{Q^{(l)}} = \max_i \|Q_i^{(l)}\mathbf{x}\|_0$ where $Q_j^{(l)}$, the stripe operator, takes $\lfloor((2(m_l/s_l)-1)n_l)\rfloor$ consecutive elements of $\mathbf{x}$ starting at x_i
- $\|\mathbf{X}^{(l)}\|_{0,\infty}^{Q^{(l)}} = \max_j \|\mathbf{x}_j\|_{0,\infty}^{Q^{(l)}}$

This publication is based on work supported by the Alan Turing Institute under the EPSRC grant EP/N510129/1. The authors thank David L. Donoho, Vardan Papyan for motivating this work and Jeremias Sulam for stimulating discussions.

Their analysis relies heavily on the notion of the coherence of a dictionary,

$$\mu(\mathbf{A}) = \max_{i \neq j} |\mathbf{a}_i^* \mathbf{a}_j| \tag{2}$$

where $\mathbf{a}_i$ is the i^{th} column of $\mathbf{A}$. The main technical innovations in [1] include how bounds from traditional sparse approximation propagate through multiple layers, and the aforementioned sparsity models in order to ameliorate the limited lower bound on (2) which is often observed for matrices with the convolutional structure as in the case of $\mathbf{A}^{(l)}$. In particular, the main results in [1] most relevant to our extensions lets $\mathbf{X}^{(l)}$ be the sparse matrices in Model (1) with $\mathbf{V}^{(l)} = 0$, i.e., the data is generated without noise and the activation functions in the presence of noise are computed recursively from the data matrix $\hat{\mathbf{X}}^{(0)}$ as

$$\hat{\mathbf{X}}^{(l)} = Proj_{\|\cdot\|_{0,\infty}^{Q^{(l)}} \leq S_l} \left((\mathbf{A}^{(l)})^T \hat{\mathbf{X}}^{(l-1)} \right) \tag{3}$$

for $l = 1, 2, \cdots L$ where $Proj_{\|\cdot\|_{0,\infty}^{Q^{(l)}} \leq S_l}(\cdot)$ projects to the largest entries within the sparsity model. In particular, they prove[1] that if the noise free data and activation functions satisfy a sparsity bound $\|\mathbf{X}^{(l)}\|_{0,\infty}^{Q^{(l)}} \leq S_l$ and some approximation error power is bounded by η_l, then if

$$S_l < \frac{\mu(\mathbf{A}^{(l)})^{-1}}{|X_{max}^{(l)}|} \left(\frac{1}{2} |X_{min}^{(l)}| - \eta_l \right) + \frac{1}{2} \tag{4}$$

where $|X_{min}^{(l)}|$ and $|X_{max}^{(l)}|$ are the smallest and largest non-zeros in $\mathbf{X}^{(l)}$ respectively, then the location of non-zeros in $\hat{\mathbf{X}}^{(l)}$ exactly coincide with the location of non-zeros in $\mathbf{X}^{(l)}$.

2. MAIN RESULT: AVERAGE CASE PERFORMANCE

Notable in the sparsity bound (4) is the presence of $\mu(\mathbf{A}^{(l)})^{-1}$ which is the factor that allows for nontrivial S_l. Bounds of the form (4) are prevalent in the theory of sparse approximation, see for instance [5, Chapter 5], where it is known [6] that for a generic matrix $\mathbf{B} \in \mathbb{R}^{m \times \gamma m}$ that $\mu(\mathbf{B}) > m^{-1/2}\sqrt{1 - \gamma^{-1}}$ which is colloquially referred to as the square-root bottleneck in that $\mu^{-1} \sim m^{1/2}$. In many applications e.g, imaging m_i is typically not more than 7^2 and n_i is typically about $2m_i$. For Model (1) there is also the additional challenge due to $\mathbf{A}^{(l)}$ being convolutional in structure, this can result in a large mutual coherence if the stride is between shifted versions of the local dictionaries $\mathbf{A}_{Local}^{(l)}$ is insufficient.

It is well known from the work of Schnass and Vandergheynst [2] in the single layer context that, if one introduces a randomized sign pattern, the Rademacher concentration inequality can be used to derive bounds showing that the recovery of the correct activation locations is *typically* possible with the sparsity bounds relaxed to depend on $\mu(\mathbf{A})^{-2}$. Our main result is to combine the techniques used in [2] to the multi-layer setting of [1]. In order to do so we adapt Model (1) to include randomized sign patterns on the masks of the network, or equivalently activations at the next layer. That is the inclusion of:

- $\mathbf{D}^{(l)} \in \mathbb{R}^{n_l M \times n_l M}$ a random, diagonal matrix whose diagonal entries are independent Rademacher random variables. This matrix multiplies the weight matrix mapping between the $l - 1$th and lth layers, applying a random sign pattern to the columns of $\mathbf{A}^{(l)}$.

Under this adaption we are able to provide Theorem 1.

Theorem 1. *Let $\hat{\boldsymbol{X}}^{(0)}$ be a date matrix consistent with Model* (1) *with $\|\boldsymbol{V}^{(l)}\|_{2,\infty}^{P^{(l)}} \leq \zeta_l$, $\|\boldsymbol{X}^{(l)}\|_{0,\infty}^{Q^{(l)}} \leq S_l$ for all $l = 0, \dots, L-1$. Further assume that $\boldsymbol{D}^{(l)}$ is a random diagonal matrix with independent Rademacher random variables on the diagonal entries drawn independent of the dictionaries $\boldsymbol{A}^{(l)}$. Then let $\hat{\boldsymbol{X}}^{(l)}$ be computed as in* (3) *and denote as Z_L the event that the location of the non-zeros in $\boldsymbol{X}^{(l)}$ and $\hat{\boldsymbol{X}}^{(l)}$ exactly coincide for $l = 0, 1, \dots, L$; then the probability this event doesn't hold, $\bar{Z}_L$, is at most*

$$P(\bar{Z}_L) \leq 2dM \sum_{l=1}^{L} n_l \exp\left(- \frac{|X_{min}^{(l)}|^2}{8\left(|X_{max}^{(l)}|^2 \mu_l^2 S_l + \zeta_{l-1}^2 \right)} \right) \tag{5}$$

Furthermore when Z_L does occur then $\forall j$,

$$\|\hat{\boldsymbol{x}}_j^{(l)} - \boldsymbol{x}_j^{(l)}\|_{2,\infty}^{P^{(l)}} \leq \zeta_l \tag{6}$$

where

$$\zeta_l = \sqrt{\|\hat{\boldsymbol{X}}^{(l)}\|_{0,\infty}^{P^{(l)}}} \left(\mu_l (S_l - 1) |X_{max}^{(l)}| + \zeta_{l-1} \right). \tag{7}$$

Note that in the above formulation we have adopted the shortened notation $\mu_l = \mu(\mathbf{A}^{(l)})$. A key implication of Theorem 1 is that the derived probability bound scales proportional to $\mu(\mathbf{A}^{(l)})^{-2}$ rather than $\mu(\mathbf{A}^{(l)})^{-1}$ across a given layer. To be precise for a given representation $\hat{\mathbf{x}}^{(l-1)}$ then for an arbitrary $\delta \in [0, 1]$ then $P(\bar{W}_l) \leq \delta$ if the following inequality on the sparsity is satisfied:

$$S_l \leq \left(\frac{|x_{min}^{(l)}|^2}{8|x_{max}^{(l)}|^2 \ln\left(\frac{2Mn_l}{\delta}\right)} - \frac{\zeta_{l-1}^2}{|x_{max}^{(l)}|^2} \right) \mu_l^{-2} \tag{8}$$

[1] The analysis of Papyan, Romano, and Elad in [1] are wide-ranging, including showing conditions under which the solution to Model (1) is unique, considering various thresholding operators such as soft and hard thresholding, and more advanced algorithms to compute $\hat{\mathbf{X}}^{(l)}$ from $\mathbf{A}^{(l)}$ and $\hat{\mathbf{X}}^{(l-1)}$; due to space constraints we do not speak to the majority of their contributions.

3. PROOF OF MAIN RESULT

We develop the proof of Theorem 1 by first considering the failure to recover the sparse representation of a single vector across a single layer and then a single vector across multiple layers. Consider then the recovery of the support for a single vector across a single layer, the signal model for this we can write as:

$$\hat{\mathbf{x}}^{(l-1)} = \mathbf{A}\mathbf{D}\hat{\mathbf{x}}^{(l)} + \mathbf{v}^{(l-1)} \tag{9}$$

where we have omitted the superscript layer notation on the matrices $\mathbf{A}$ and $\mathbf{D}$ for typographical clarity. Inclusion of the diagonal Rademacher matrix $\mathbf{D}$ allows us to derive probabilistic bounds using the following concentration of measure inequality:

Theorem 2 (Rademacher concentration [7])**.** *Let α be an arbitrary real vector and ε a random vector whose elements are independent random variables pulled from a Rademacher distribution $\{-1, 1\}$. Then for all $t > 0$*

$$P\left(|\sum_i \varepsilon_i \alpha_i| > t\right) \le 2\exp\left(-\frac{t^2}{2\|\alpha\|_2^2}\right). \tag{10}$$

built upon the single layer vector case given in Lemma 3.

Lemma 3. *Suppose we have a data point $\hat{\boldsymbol{x}}^{(l-1)}$, and assume it was generated under signal model* (9)*. Let $\hat{\boldsymbol{x}}^{(l)}$ be given by*

$$\hat{\boldsymbol{x}}^{(l)} = Proj_{\|\cdot\|_{0,\infty}^{Q^{(l)}} \le S_l}\left(\boldsymbol{D}^T\boldsymbol{A}^T\hat{\boldsymbol{x}}^{(l-1)}\right) \tag{11}$$

and let $\boldsymbol{x}^{(l)}$ be a solution of (9) *with $\|\boldsymbol{x}^{(l)}\|_{0,\infty}^{Q^{(l)}} \le S_l$ and $\boldsymbol{v}^{(l-1)} = 0$. Denote as $\bar{W}_\Lambda$ the event that the location, Λ, of the nonzeros in $\boldsymbol{x}^{(l)}$ and $\hat{\boldsymbol{x}}^{(l)}$ differ, then*

$$P(\bar{W}_\Lambda) \le 2nM\exp\left(-\frac{|x_{min}^{(l)}|^2}{8\left(|x_{max}^{(l)}|^2\mu_l^2 S_l + \zeta_{l-1}^2\right)}\right), \tag{12}$$

furthermore in the setting where the location of the nonzeros in $\boldsymbol{x}^{(l)}$ and $\hat{\boldsymbol{x}}^{(l)}$ are the same, then

$$\|\hat{\boldsymbol{x}}^{(l)} - \boldsymbol{x}^{(l)}\|_{2,\infty}^{P^{(l)}} \le \zeta_l$$

where

$$\zeta_l = \sqrt{\|\hat{\boldsymbol{x}}^{(l)}\|_{0,\infty}^{P^{(l)}}}\left(\mu_l(S_l - 1)|x_{max}^{(l)}| + \zeta_{l-1}\right).$$

Proof. Lemma 3 extends bounds in [2] to include additive noise and the stripe sparsity model $\|\cdot\|_{0,\infty}^{Q^{(l)}} \le S_l$ present for the convolutional matrices $\mathbf{A}^{(l)}$ considered here.

For $\bar{W}_\Lambda$ to occur requires the condition that $\exists$ some $i \in \Lambda$ and some $k \in \bar{\Lambda}$ such that $|\langle \varepsilon_i \mathbf{a}_i, \hat{\mathbf{x}}^{(l-1)}\rangle| < |\langle \varepsilon_k \mathbf{a}_k, \hat{\mathbf{x}}^{(l-1)}\rangle|$. This condition is equivalent to requiring $\min_{i\in\Lambda}|\langle \varepsilon_i \mathbf{a}_i, \hat{\mathbf{x}}^{(l-1)}\rangle| < \max_{k\notin\Lambda}|\langle \varepsilon_k \mathbf{a}_k, \hat{\mathbf{x}}^{(l-1)}\rangle|$ for which:

$$P(\bar{W}_\Lambda) = P(\min_i |\langle \varepsilon_i \mathbf{a}_i, \hat{\mathbf{x}}^{(l-1)}\rangle| < \max_k |\langle \varepsilon_k \mathbf{a}_k, \hat{\mathbf{x}}^{(l-1)}\rangle|)$$

Introducing an arbitrary real valued scalar threshold $p > 0$, we form the following inequality

$$\begin{aligned} P(\bar{W}_\Lambda) \le & P(\min_i |\langle \varepsilon_i \mathbf{a}_i, \hat{\mathbf{x}}^{(l-1)}\rangle| < p) \\ & + P(\max_k |\langle \varepsilon_k \mathbf{a}_k, \hat{\mathbf{x}}^{(l-1)}\rangle| > p) \end{aligned}$$

for which we provide bounds on each of the terms on the right using the Rademacher concentration of Theorem 2. Denoting $\bar{W}'_\Lambda$ as the event $\max_{k\notin\Lambda}|\langle \varepsilon_k \mathbf{a}_k, \hat{\mathbf{x}}^{(l-1)}\rangle| > p$ then:

$$\begin{aligned} P(\bar{W}'_\Lambda) &\le \sum_{k\in\bar{\Lambda}} P\left(|\langle \varepsilon_k \mathbf{a}_k, \hat{\mathbf{x}}^{(l-1)}\rangle| > p\right) \\ &= \sum_{k\in\bar{\Lambda}} P\left(|\sum_{j\in\Lambda} \varepsilon'_j x_j^{(l)}\langle \mathbf{a}_k, \mathbf{a}_j\rangle + \varepsilon_k\langle \mathbf{a}_k, \mathbf{v}\rangle| > p\right) \\ &\le 2\sum_{k\in\bar{\Lambda}} \exp\left(\frac{-p^2}{2\left(\sum_{j\in\Lambda\cap\Gamma}|x_j^{(l)}|^2|\langle \mathbf{a}_k, \mathbf{a}_j\rangle|^2 + \zeta_{l-1}^2\right)}\right) \\ &\le 2(mN - |\Lambda|)\exp\left(\frac{-p^2}{2\left(|x_{max}^{(l)}|^2 S_l\mu_l^2 + \zeta_{l-1}^2\right)}\right) \end{aligned} \tag{13}$$

Here the first line and inequality arises from $\max_{k\notin\Lambda}\{|\langle \varepsilon_k \mathbf{a}_k, \hat{\mathbf{x}}^{(l-1)}\rangle| > p\} \in \cup_{k\notin\Lambda}|\langle \varepsilon_k \mathbf{a}_k, \hat{\mathbf{x}}^{(l-1)}\rangle| > p$ and then bounding using the disjoint union. The second is a simple expansion of the inner product, where we define a new Rademacher random variable $\varepsilon'_j = \varepsilon_j\varepsilon_k$ (note that the set of random variables $\{\{\varepsilon'_i\} \cup \varepsilon_k\}$ is also mutually independent). Moving from the second to the third we use Theorem 2 (note that the set Γ refers to the indices of columns of $\mathbf{A}^{(l)}$ which have a nonzero inner product with the column $\mathbf{a}_k$), and the final line follows from $|\Lambda \cap \Gamma| \le S_l$.

Denoting $\bar{W}''_\Lambda$ as the event $\min_i\{|\langle \varepsilon_i \mathbf{a}_i, \hat{\mathbf{x}}^{(l-1)}\rangle| < p\}$, which we bound from below as

$$|\langle \varepsilon_i \mathbf{a}_i, \hat{\mathbf{x}}^{(l-1)}\rangle| \ge |x_i| - \left|\sum_{j\in\Lambda, j\ne i} \varepsilon'_j x_j\langle \mathbf{a}_i, \mathbf{a}_j\rangle + \varepsilon_i\langle \mathbf{a}_i, \mathbf{v}\rangle\right|,$$

and noting that we want to minimise this term then:

$$\begin{aligned} &P(\bar{W}''_\Lambda) \le \sum_{i\in\Lambda} P(\left|\sum_{j\in\Lambda, j\ne i} \varepsilon'_j x_j\langle \mathbf{a}_i, \mathbf{a}_j\rangle + \varepsilon_i\zeta_{l-1}\right| > |x_{min}^{(l)}| - p) \\ &\le 2|\Lambda|\exp\left(\frac{-(|x_{min}^{(l)}| - p)^2}{2\left(|x_{max}^{(l)}|^2 S_l\mu_l^2 + \zeta_{l-1}^2\right)}\right) \end{aligned}$$

which follows as in (13). Since p can be arbitrary, and since it is equidistant between the expectations of the two distributions of $|\langle \varepsilon_i \mathbf{a}_i, y \rangle|$ and $|\langle \varepsilon_k \mathbf{a}_k, y \rangle|$, choose $p = |x_{min}|/2$. As a result, combining the bounds on $P(\bar{W}'_\Lambda)$ and $P(\bar{W}''_\Lambda)$ we obtain the desired bound on $P(\bar{W}_\Lambda)$.

□

With the single layer, single vector case proven in Lemma 3, we now proceed to investigate the single vector, multilayer case.

Lemma 4. *Suppose we have a column vector $\hat{\boldsymbol{x}}^{(0)}$ taken from a matrix generated under date Model* (1)*. If Y_L denotes the event that the thresholding operation exactly recovers the support Λ of this vector at all layers up to layer L then:*

$$P(\bar{Y}_L) \leq 2M \sum_{l=1}^{L} n_l \exp\left(-\frac{|x_{min}^{(l)}|^2}{8\left(|x_{max}^{(l)}|^2 \mu_l^2 S_l + \zeta_{l-1}^2\right)} \right) \tag{14}$$

Proof. This result can be proved easily via induction. For the sake of convenience let:

$$\gamma^{(l)} = n_l \exp\left(-\frac{|x_{min}^{(l)}|^2}{8\left(|x_{max}^{(l)}|^2 \mu_l^2 S_l + \zeta_{l-1}^2\right)} \right)$$

Note that the bound on the error at each layer is conditioned on the correct recovery of the support at the previous layer. Letting $\bar{W}_\Lambda^{(l)}$ be the event that support at the lth layer is not correctly recovered, then

$$\begin{aligned}
&P(\bar{W}_\Lambda^{(1)}) \leq \gamma^{(1)} \\
&P(\bar{W}_\Lambda^{(2)} | \bar{W}_\Lambda^{(1)}) \leq \gamma^{(2)} \\
&\dots \\
&P(\bar{W}_\Lambda^{(l)} | \cap_i^{(l-1)} \bar{W}_\Lambda^{(i)}) \leq \gamma^{(l)}
\end{aligned}$$

The result for $L = 1$ is trivial so we will proceed with the proof by induction by considering the base case $L = 2$. Applying to $\bar{Y}_2$ De Morgan's theorem:

$$\begin{aligned}
P(\bar{Y}_2) &= P(\bar{W}_\Lambda^{(1)} \cup \bar{W}_\Lambda^{(2)}) \\
&= P(\bar{W}_\Lambda^{(1)}) + P(\bar{W}_\Lambda^{(2)} \cup W_\Lambda^{(1)}) \\
&= P(\bar{W}_\Lambda^{(1)}) + P(\bar{W}_\Lambda^{(2)} | W_\Lambda^{(1)}) P(W_\Lambda^{(1)}) \\
&\leq \gamma^{(1)} + \gamma^{(2)} P(W_\Lambda^{(1)}) \\
&\leq \gamma^{(1)} + \gamma^{(2)}
\end{aligned}$$

Hence our theorem is correct for $l = 1$ and $l = 2$. Now assume that the result holds true for the kth layer, i.e.:

$$P(\bar{Y}_k) \leq \sum_{l=1}^{k} \gamma^{(l)}$$

Consider then $\bar{Y}_{k+1}$:

$$\begin{aligned}
P(\bar{Y}_{k+1}) &= P(\bar{W}_\Lambda^{(k+1)} \cup \bar{Y}_k) \\
&= P(\bar{Y}_k) + P(\bar{W}_\Lambda^{(2)} \cup Y_k) \\
&= P(\bar{Y}_k) + P(\bar{W}_\Lambda^{(k+1)} | Y_k) P(Y_k) \\
&\leq \sum_{l=1}^{k} \gamma^{(l)} + \gamma^{(k+1)} P(Y_k) \\
&\leq \sum_{l=1}^{k} \gamma^{(l)} + \gamma^{(k+1)} = \sum_{l=1}^{k+1} \gamma^{(l)}
\end{aligned}$$

This proves the $k + 1$th case, and given our base and $k + 1$th cases hold then all others must follow. □

Theorem 1 follows from Lemma 4 as follows. Suppose that $Z_j^{(L)}$ is the event that supports $\{\Lambda_j^{(l)}\}_{l=1}^{L}$ of the representations $\{\mathbf{x}_j^{(l)}\}_{l=1}^{L}$ of the jth are recovered from $\{\hat{\mathbf{x}}_j^{(l)}\}_{l=1}^{L}$. Define

$$Z_L = \bigcap_{j=1}^{d} Z_j^{(L)}$$

and applying De Morgan's rule, then $\bar{Z}^{(L)} = \bigcup_{j=1}^{d} \bar{Z}_j^{(L)}$ and as a result:

$$\begin{aligned}
P(\bar{Z}_L) &\leq \sum_{j=1}^{d} P(\bar{Z}_j^{(L)}) \\
&= 2M \sum_{j=1}^{d} \sum_{l=1}^{L} \gamma_j^{(l)} \leq 2Md \sum_{l=1}^{L} \gamma_{max}
\end{aligned}$$

which gives the claimed probability bound in Theorem 1. Bounding the error in Theorem 1 under the assumption that the support Λ is recovered follows exactly as in Theorem 8 of [1].

4. CONCLUSION

Over recent years there has been growing number of researchers working to better understand deep learning, to highlight just a few contributions [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23]. Our contribution in this paper is Theorem 1, which extends the prior uniform bounds in [1] to high probability bounds, with the proportionality to the dictionary coherence improving from μ^{-1} to μ^{-2} respectively. Assuming the weight matrices are suitably conditioned, this indicates that the forward pass algorithm is likely to recover the latent representations in Model 1 for a more complex (in terms of the number of non-zeros) family of signals than previously thought. In summary, this suggests that if optimal sparse coding is indeed an important factor explaining the success of certain CNN architectures, then explicitly encouraging weight matrices with low coherence during training would improve the CNN's performance.

5. REFERENCES

[1] V. Papyan, Y. Romano, and M. Elad, "Convolutional Neural Networks Analyzed via Convolutional Sparse Coding," *ArXiv e-prints*, July 2016.

[2] K. Schnass and P. Vandergheynst, "Average performance analysis for thresholding," *IEEE Signal Processing Letters*, vol. 14, no. 11, pp. 828–831, Nov 2007.

[3] Xavier Glorot, Antoine Bordes, and Yoshua Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, Geoffrey Gordon, David Dunson, and Miroslav Dudk, Eds., Fort Lauderdale, FL, USA, 11–13 Apr 2011, vol. 15 of *Proceedings of Machine Learning Research*, pp. 315–323, PMLR.

[4] V. Papyan, J. Sulam, and M. Elad, "Working Locally Thinking Globally - Part I: Theoretical Guarantees for Convolutional Sparse Coding," *ArXiv e-prints*, July 2016.

[5] R. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*, Applied and Numerical Harmonic Analysis. Birkhäuser Basel, 2013.

[6] L R. Welch, "Lower bounds on the maximum cross correlation of signals (corresp.)," vol. 20, pp. 397 – 399, 06 1974.

[7] M. Ledoux and M. Talagrand, *Probability in Banach Spaces*, Classics in Mathematics. Springer-Verlag Berlin Heidelberg, 2011.

[8] A. C. Gilbert, Y. Zhang, K. Lee, Y. Zhang, and H. Lee, "Towards Understanding the Invertibility of Convolutional Neural Networks," *ArXiv e-prints*, May 2017.

[9] Sanjeev Arora, Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli, "A compressed sensing view of unsupervised text embeddings, bag-of-n-grams, and LSTMs," in *International Conference on Learning Representations*, 2018.

[10] S. Mallat, "Understanding deep convolutional networks," *Philosophical Transactions of the Royal Society of London Series A*, vol. 374, pp. 20150203, Apr. 2016.

[11] Jeffrey Pennington and Yasaman Bahri, "Geometry of neural network loss surfaces via random matrix theory," 2017.

[12] S. S. Schoenholz, J. Gilmer, S. Ganguli, and J. Sohl-Dickstein, "Deep Information Propagation," *ArXiv e-prints*, Nov. 2016.

[13] Jeffrey Pennington, Samuel S. Schoenholz, and Surya Ganguli, "Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice," *CoRR*, vol. abs/1711.04735, 2017.

[14] Tomaso A. Poggio and Qianli Liao, "Theory ii: Landscape of the empirical risk in deep learning," *CoRR*, vol. abs/1703.09833, 2017.

[15] Chiyuan Zhang, Qianli Liao, Alexander Rakhlin, Karthik Sridharan, Brando Miranda, Noah Golowich, and Tomaso Poggio, "Theory of deep learning iii: Generalization properties of sgd," 04/2017 2017.

[16] Kenji Kawaguchi, "Deep learning without poor local minima," in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds., pp. 586–594. Curran Associates, Inc., 2016.

[17] Benjamin D. Haeffele and René Vidal, "Global optimality in tensor factorization, deep learning, and beyond," *CoRR*, vol. abs/1506.07540, 2015.

[18] C. D. Freeman and J. Bruna, "Topology and Geometry of Half-Rectified Network Optimization," *ArXiv e-prints*, Nov. 2016.

[19] H. N. Mhaskar and T. Poggio, "Deep vs. shallow networks: An approximation theory perspective," *Analysis and Applications*, vol. 14, no. 06, pp. 829–848, 2016.

[20] Thomas Wiatowski and Helmut Bölcskei, "A mathematical theory of deep convolutional neural networks for feature extraction," *CoRR*, vol. abs/1512.06293, 2015.

[21] Ankit B Patel, Minh Tan Nguyen, and Richard Baraniuk, "A probabilistic framework for deep learning," in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds., pp. 2558–2566. Curran Associates, Inc., 2016.

[22] Julien Mairal, "End-to-end kernel learning with supervised convolutional kernel networks," in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds., pp. 1399–1407. Curran Associates, Inc., 2016.

[23] Jeffrey Pennington and Pratik Worah, "Nonlinear random matrix theory for deep learning," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., pp. 2637–2646. Curran Associates, Inc., 2017.