PROJECTING ONTO THE MULTI-LAYER CONVOLUTIONAL SPARSE CODING MODEL

Jeremias Sulam[‡], Vardan Papyan[‡], Yaniv Romano[†] and Michael Elad[‡]

[‡]Computer Science Department,
[†]Electrical Engineering Department,
Technion – Israel Institute of Technology

ABSTRACT

The recently proposed Multi-Layer Convolutional Sparse Coding (ML-CSC) model, consisting of a cascade of convolutional sparse layers, provides a new interpretation of Convolutional Neural Networks (CNNs). Under this framework, the forward pass in a CNN is equivalent to an algorithm that estimates nested sparse representation vectors from a given input signal. Despite having served as a pivotal connection between CNNs and sparse modeling, it is still unclear how to develop pursuit algorithms that serve this model exactly. In this work, we propose a new pursuit formulation by adopting a projection approach. We provide new and improved bounds on the stability of the resulting convolutional sparse representations, and we propose a multi-layer projection algorithm to retrieve them. We demonstrate this algorithm numerically, showing that it is superior to the Layered Basis Pursuit alternative in retrieving the representations of signals belonging to the ML-CSC model.

Index Terms— Convolutional Sparse Coding, Multi-layer Pursuit, Stability Guarantees, Convolutional Neural Networks.

1. INTRODUCTION

Sparse representation modeling assumes that natural signals can be (well) described as a linear combination of only a few building blocks or components, commonly known as atoms [1]. This model has been extensively studied, and a plethora of works have provided different methods to carry out the pursuit of such decompositions and train the model from real data [2]. Neural networks, on the other hand, is a classification algorithm whose origins can be traced to almost half a century ago [3, 4]. Recently, a convolutional variant – convolutional neural networks (CNN) – together other small modifications have enabled the development of state-of-the-art machine learning methods for a wide variety of problems and with impressive performance [5].

Most works on this new research field, termed deep learning, have been motivated mostly by intuition and with empirical justifications. Recently, some research groups have started to provide a more theoretical understanding of CNN, borrowing ideas from harmonic analysis [6], assuming Gaussian weights [7] and employing tensor factorization [8], among other approaches. A precise connection between sparse modeling and CNNs was recently presented in [11], and its contribution is centered in defining the Multi-Layer Convolutional Sparse Coding (ML-CSC) model. When deploying this model to real signals, compromises were made in way that each layer is only *approximately* explained by the following one. With this relaxation in the pursuit of the convolutional representations, the main observation of this work is that the inference stage of CNNs – nothing but the forward-pass – can be interpreted as a very crude pursuit algorithm seeking for unique sparse representations. The work in [11] further proposed an improved pursuit for approximating the sparse representations of the network, termed Layered Basis Pursuit. Nonetheless, neither this nor the forward pass serve the ML-CSC model exactly, as they do not provide signals that comply with the model assumptions. In addition, the theoretical guarantees accompanying these layered approaches suffer from bounds that become looser with the network's depth.

In this work we focus on defining a pursuit problem for signals belonging to this model. In particular, given proper convolutional dictionaries, we study the question of how to project signals onto the ML-CSC model. This problem is fundamentally different to the one studied in [11], as its solution should satisfy the model constraints exactly. This in turn enables the development of tighter recovery guarantees than current ones. We further derive a simple algorithm to carry out this projection in practice. The resulting multi-layer pursuit is compared experimentally with the Layered Basis Pursuit algorithm [11], showing that it provides better recovery of the multi-layer representations of signals in this model.

2. SINGLE AND MULTI-LAYER CONVOLUTIONAL SPARSE CODING

The Convolutional Sparse Coding (CSC) model assumes a signal $\mathbf{x} \in \mathbb{R}^N$ admits a decomposition as $\mathbf{D}_1\boldsymbol{\gamma}_1$, where $\boldsymbol{\gamma}_1 \in \mathbb{R}^{Nm_1}$ is sparse and $\mathbf{D}_1 \in \mathbb{R}^{N\times Nm_1}$ has a convolutional structure. This dictionary consists of m_1 local n_1-dimensional filters at every possible location (Figure 1 top). As a result, each j^{th} *patch* $\mathbf{P}_j\mathbf{x} \in \mathbb{R}^{n_1}$ from the

signal $\mathbf{x}$ can be expressed in terms of a shift-invariant local model corresponding to a *stripe* from the global sparse vector, $\mathbf{S}_j\boldsymbol{\gamma}_1 \in \mathbb{R}^{(2n_1-1)m_1}$. In this context, the sparsity of the representation is better captured through the $\ell_{0,\infty}$ pseudo-norm [12]. This measure, as opposed to the traditional ℓ_0, provides a notion of local sparsity and it is defined by the maximal number of non-zeros in a stripe from $\boldsymbol{\gamma}$. Formally,

$$\|\boldsymbol{\gamma}\|_{0,\infty}^{s} = \max_i \|\mathbf{S}_i\boldsymbol{\gamma}\|_0. \tag{1}$$

The Multi-Layer Convolutional Sparse Coding (ML-CSC) model is a natural extension of the CSC described above, as it assumes that a sparse representation $\boldsymbol{\gamma}_1$ also allows for a decomposition in terms of CSC model. Formally, given a set of convolutional dictionaries $\{\mathbf{D}_i\}_{i=1}^{L}$ of appropriate dimensions, a signal $\mathbf{x} \in \mathbb{R}^N$ admits a representation in terms of the ML-CSC model if

$$\begin{aligned}
\mathbf{x} &= \mathbf{D}_1\boldsymbol{\gamma}_1, \quad \|\boldsymbol{\gamma}_1\|_{0,\infty}^{s} \le \lambda_1,\\
\boldsymbol{\gamma}_1 &= \mathbf{D}_2\boldsymbol{\gamma}_2, \quad \|\boldsymbol{\gamma}_2\|_{0,\infty}^{s} \le \lambda_2,\\
&\vdots\\
\boldsymbol{\gamma}_{L-1} &= \mathbf{D}_L\boldsymbol{\gamma}_L, \quad \|\boldsymbol{\gamma}_L\|_{0,\infty}^{s} \le \lambda_L.
\end{aligned}$$

We denote as $\mathcal{M}_{\boldsymbol{\lambda}}$ the set of signals satisfying the ML-CSC model assumptions with parameters given by $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_L]$, and we will employ the notation $\mathbf{x}(\boldsymbol{\gamma}_i) \in \mathcal{M}_{\boldsymbol{\lambda}}$ to emphasize that $\mathbf{x}$ allows for a decomposition in terms of the ML-CSC representations $\{\boldsymbol{\gamma}_i\}_{i=1}^{L}$. Note that $\mathbf{x}(\boldsymbol{\gamma}_i) \in \mathcal{M}_{\boldsymbol{\lambda}}$ can also be expressed as $\mathbf{x} = \mathbf{D}_1\mathbf{D}_2\ldots\mathbf{D}_L\boldsymbol{\gamma}_L$. We will further denote $\mathbf{D}^{(i)}$ as the *effective* dictionary for the i^{th} level, i.e., $\mathbf{D}^{(i)} = \mathbf{D}_1\mathbf{D}_2\ldots\mathbf{D}_i$. Therefore, for any layer, $\mathbf{x} = \mathbf{D}^{(i)}\boldsymbol{\gamma}_i$.

2.1. Pursuit in the noisy setting

Real signals often contain noise or deviate from the above idealistic model assumption, preventing us from enforcing the above model exactly. Consider the measurements $\mathbf{y} = \mathbf{x}(\boldsymbol{\gamma}_i) + \mathbf{v}$, with $\mathbf{x}(\boldsymbol{\gamma}_i) \in \mathcal{M}_{\boldsymbol{\lambda}}$ and $\mathbf{v}$ a nuisance vector of bounded energy, $\|\mathbf{v}\|_2 \le \mathcal{E}_0$. In this setting, the pursuit problem becomes searching for sparse convolutional representations that provide an approximation to $\mathbf{y}$. In its most general form, this pursuit is represented by the Deep Coding Problem (DCP$_{\boldsymbol{\lambda}}^{\boldsymbol{\mathcal{E}}}$), as introduced in [11] and given by

$$\begin{aligned}
(\text{DCP}_{\boldsymbol{\lambda}}^{\boldsymbol{\mathcal{E}}}): \quad & \text{find } \{\boldsymbol{\gamma}_i\}_{i=1}^{L} \text{ s.t.} \\
& \|\mathbf{y} - \mathbf{D}_1\boldsymbol{\gamma}_1\|_2 \le \mathcal{E}_0, \qquad \|\boldsymbol{\gamma}_1\|_{0,\infty}^{s} \le \lambda_1\\
& \|\boldsymbol{\gamma}_1 - \mathbf{D}_2\boldsymbol{\gamma}_2\|_2 \le \mathcal{E}_1, \qquad \|\boldsymbol{\gamma}_2\|_{0,\infty}^{s} \le \lambda_2\\
& \qquad\qquad \vdots \qquad\qquad\qquad\quad \vdots\\
& \|\boldsymbol{\gamma}_{L-1} - \mathbf{D}_L\boldsymbol{\gamma}_L\|_2 \le \mathcal{E}_{L-1}, \quad \|\boldsymbol{\gamma}_L\|_{0,\infty}^{s} \le \lambda_L,
\end{aligned} \tag{2}$$

where the scalars λ_i and $\mathcal{E}_i$ are the i^{th} entries of $\boldsymbol{\lambda}$ and $\boldsymbol{\mathcal{E}}$, respectively. The solution to this problem was shown to be stable in terms of a bound on the ℓ_2-distance between the estimated representations $\hat{\boldsymbol{\gamma}}_i$ and the true ones $\boldsymbol{\gamma}_i$. These results depend on the characterization of the dictionaries through their mutual coherence, $\mu(\mathbf{D})$, which measures the maximal normalized correlation between atoms in the dictionary. Formally, assuming the atoms are normalized as $\|\mathbf{d}_i\|_2 = 1\ \forall i$, this measure is defined as

$$\mu(\mathbf{D}) = \max_{i \ne j} |\mathbf{d}_i^T\mathbf{d}_j|. \tag{3}$$

Relying on this measure, Theorem 5 in [11] shows that given a signal $\mathbf{x}(\boldsymbol{\gamma}_i) \in \mathcal{P}_{\mathcal{M}_{\boldsymbol{\lambda}}}$ contaminated with noise of known energy $\mathcal{E}_0^2$, if the representations satisfy the sparsity constraint

$$\|\boldsymbol{\gamma}_i\|_{0,\infty}^{s} < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{D}_i)}\right), \tag{4}$$

then the solution to the DCP$_{\boldsymbol{\lambda}}^{\boldsymbol{\mathcal{E}}}$ given by $\{\hat{\boldsymbol{\gamma}}_i\}_{i=1}^{L}$ satisfies[1]

$$\|\boldsymbol{\gamma}_i - \hat{\boldsymbol{\gamma}}_i\|_2^2 \le 4{\mathcal{E}_0}^2 \prod_{j=1}^{i} \frac{4^{i-1}}{1 - (2\|\boldsymbol{\gamma}_j\|_{0,\infty}^{s} - 1)\mu(\mathbf{D}_j)}. \tag{5}$$

Interestingly, several layer-wise algorithms, including the forward pass of CNN, provide approximations to the solution of this problem [11]. Despite its significance, we note two drawbacks of this problem. First, these bounds increase with the number of layers or the depth of the network, which is a direct consequence of the layer-wise relaxation in the above pursuit. On the other hand, given the underlying signal $\mathbf{x}(\boldsymbol{\gamma}_i) \in \mathcal{M}_{\boldsymbol{\lambda}}$,

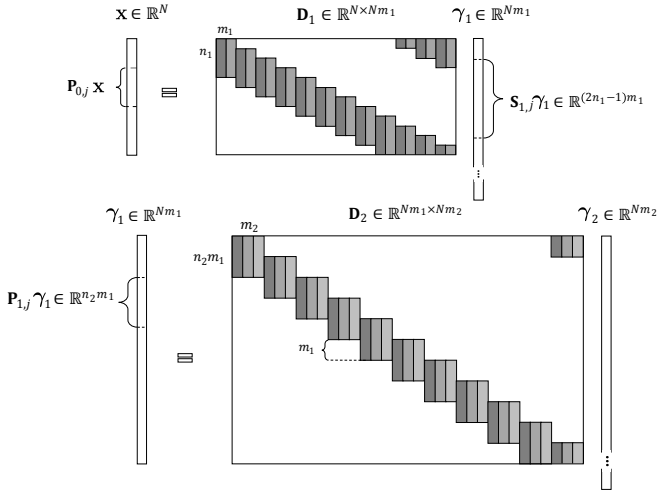

Fig. 1: The CSC model (top), and its ML-CSC extension by imposing a similar model on $\boldsymbol{\gamma}_1$ (bottom). From a local perspective, a patch from the signal, $\mathbf{P}_{0,j}\mathbf{x}$ has a corresponding sparse stripe given by $\mathbf{S}_{1,j}\boldsymbol{\gamma}_1$. An analogous decomposition can be stated for a patch from the signal $\boldsymbol{\gamma}_1$, represented by $\mathbf{P}_{1,j}\boldsymbol{\gamma}_1$.

[1] In the particular instance of the DCP$_{\boldsymbol{\lambda}}^{\boldsymbol{\mathcal{E}}}$ where $\mathcal{E}_i = 0$ for $1 \le i \le L-1$, the above bound can be made tighter by a factor of 4^{i-1} while preserving the same form.

this problem searches estimates $\{\hat{\gamma}_i\}_{i=1}^L$ that *approximately* explain each layer. However, because $\|\hat{\gamma}_{i-1} - \mathbf{D}_i\hat{\gamma}_i\|_2 > 0$, this problem *does not* provide a signal that satisfies the ML-CSC model assumptions.

3. A PROJECTION ALTERNATIVE

We now focus on the problem of finding an estimate $\hat{\mathbf{x}}$ that, unlike the previous approach, would have an expression in terms of a ML-CSC decomposition. In other words, we are interested in projecting the measurements $\mathbf{y}$ onto the set $\mathcal{M}_{\boldsymbol{\lambda}}$. We formalize this problem as

$$(\mathcal{P}_{\mathcal{M}_{\boldsymbol{\lambda}}}): \quad \min_{\{\boldsymbol{\gamma}_i\}_{i=1}^L} \quad \|\mathbf{y} - \mathbf{x}(\boldsymbol{\gamma}_i)\|_2 \quad \text{s.t.} \quad \mathbf{x}(\boldsymbol{\gamma}_i) \in \mathcal{M}_{\boldsymbol{\lambda}}. \tag{6}$$

Note that the solution to this problem, $\{\hat{\gamma}_i\}$, is required to belong to the set $\mathcal{M}_{\boldsymbol{\lambda}}$, implying that $\boldsymbol{\gamma}_{i-1} = \mathbf{D}_i\boldsymbol{\gamma}_i \ \forall i$. A solution to the DCP$_{\boldsymbol{\lambda}}^{\boldsymbol{\mathcal{E}}}$ problem, on the other hand, would provide estimates $\hat{\boldsymbol{\gamma}}_i$ that explain each layer approximately, but such that $\boldsymbol{\gamma}_{i-1} \neq \mathbf{D}_i\boldsymbol{\gamma}_i \ \forall i$.

3.1. Stability of the projection $\mathcal{P}_{\mathcal{M}_{\boldsymbol{\lambda}}}$

Clearly, both problems (2) and (6) provide estimates $\boldsymbol{\gamma}_i$ for the measurement $\mathbf{y} = \mathbf{x}(\boldsymbol{\gamma}_i) + \mathbf{v}$. In light of the stability result of the DCP$_{\boldsymbol{\lambda}}^{\boldsymbol{\mathcal{E}}}$ problem, how close will the solution of the $\mathcal{P}_{\mathcal{M}_{\boldsymbol{\lambda}}}$ problem be from the underlying representations? We now provide such a stability guarantee.

Theorem 1. *Stability of the solution to the $\mathcal{P}_{\mathcal{M}_{\boldsymbol{\lambda}}}$ problem: Suppose $\mathbf{x}(\boldsymbol{\gamma}_i) \in \mathcal{M}_{\boldsymbol{\lambda}}$ is observed through $\mathbf{y} = \mathbf{x} + \mathbf{v}$, where $\mathbf{v}$ is a bounded noise vector, $\|\mathbf{v}\|_2 \leq \mathcal{E}_0$, and $\|\boldsymbol{\gamma}_i\|_{0,\infty}^s = \lambda_i < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{D}^{(i)})}\right)$, for $1 \leq i \leq L$. Consider the set $\{\hat{\gamma}_i\}_{i=1}^L$ to be the solution of the $\mathcal{P}_{\mathcal{M}_{\boldsymbol{\lambda}}}$ problem. Then,*

$$\|\boldsymbol{\gamma}_i - \hat{\boldsymbol{\gamma}}_i\|_2^2 \leq \frac{4\mathcal{E}_0^2}{1 - (2\|\boldsymbol{\gamma}_i\|_{0,\infty}^s - 1)\mu(\mathbf{D}^{(i)})}. \tag{7}$$

Interestingly, this theorem provides a bound on the distance to the true representation $\boldsymbol{\gamma}_i$ which is not cumulative over the layers, allowing for generally tighter results than those in (5). In other words, the bound does not increase with the network's depth. This is achieved by relying on the mutual coherence of the effective dictionary for that layer, $\mathbf{D}^{(i)}$, as opposed to the mutual coherence each individual dictionary. Note that this is a potentially useful characterization, as the effective dictionary is expected to become less correlated as one considers more global atoms. Lastly, while the conditions imposed on the sparse vectors $\boldsymbol{\gamma}_i$ might seem prohibitive, one should remember that this results from a worst case analysis. Moreover, one can effectively construct analytic nested convolutional dictionaries with small coherence measures, as shown in [11].

We now prove the above result.

Algorithm 1: ML-CSC Projection Algorithm

```
Init: x* = 0 ;
for k = 1 : λ_L do
    γ̂_L ← Pursuit(y, D^(L), k) ;
    for j = L : −1 : 1 do
        γ̂_{j−1} ← D_j γ̂_j;
    if ‖γ̂_i‖^s_{0,∞} > λ_i for any 1 ≤ i < L then
        break;
    else
        x* ← D^(i) γ̂_i;
return x*
```

Proof. Denote the solution to the $\mathcal{P}_{\mathcal{M}_{\boldsymbol{\lambda}}}$ problem by $\hat{\mathbf{x}}$; i.e., $\hat{\mathbf{x}} = \mathbf{D}^{(i)}\hat{\boldsymbol{\gamma}}_i$. Given that the original signal $\mathbf{x}$ satisfies $\|\mathbf{y} - \mathbf{x}\|_2 \leq \mathcal{E}_0$, the solution to the $\mathcal{P}_{\mathcal{M}_{\boldsymbol{\lambda}}}$ problem, $\hat{\mathbf{x}}$ must satisfy

$$\|\mathbf{y} - \hat{\mathbf{x}}\|_2 \leq \|\mathbf{y} - \mathbf{x}\|_2 \leq \mathcal{E}_0, \tag{8}$$

as this is the signal which provides the shortest ℓ_2 (data-fidelity) distance from $\mathbf{y}$. Note that because $\hat{\mathbf{x}}(\boldsymbol{\gamma}_i) \in \mathcal{M}_{\boldsymbol{\lambda}}$, we can have that $\hat{\mathbf{x}} = \mathbf{D}^{(i)}\hat{\boldsymbol{\gamma}}_i$, $\forall\ 1 \leq i \leq L$. Recalling Lemma 1 in [13], the product $\mathbf{D}_1\mathbf{D}_2\ldots\mathbf{D}_i$ is a convolutional dictionary. In addition, we have required that $\|\hat{\boldsymbol{\gamma}}_i\|_{0,\infty}^s \leq \lambda_i < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{D}^{(i)})}\right)$. Therefore, from the same arguments presented in [12], it follows that

$$\|\boldsymbol{\gamma}_i - \hat{\boldsymbol{\gamma}}_i\|_2^2 \leq \frac{4\mathcal{E}_0^2}{1 - (2\|\boldsymbol{\gamma}_i\|_{0,\infty}^s - 1)\mu(\mathbf{D}^{(i)})}. \tag{9}$$

□

3.2. ML-CSC Pursuit

The above result provides a stability bound for the solution to the $\mathcal{P}_{\mathcal{M}_{\boldsymbol{\lambda}}}$ problem, but it does not specify how to solve it practically. In the context of the DCP$_{\boldsymbol{\lambda}}^{\boldsymbol{\mathcal{E}}}$, one can approximate its solution in a layer-wise manner, solving for the sparse representations $\hat{\boldsymbol{\gamma}}_i$ progressively from $i = 1, \ldots, L$. Surprisingly, the Forward Pass of a CNN *is* one such algorithm, and it provides an approximate solution of this problem [11]. A better alternative was also proposed in that work, where each representation $\hat{\boldsymbol{\gamma}}_i$ is sparse coded in a Basis Pursuit formulation given the previous representation $\hat{\boldsymbol{\gamma}}_{i-1}$ and dictionary $\mathbf{D}_i$. This algorithm is called Layered Basis Pursuit (BP).

Moving to the variation proposed in this work, how can one solve the $\mathcal{P}_{\mathcal{M}_{\boldsymbol{\lambda}}}$ problem in practice? Note that employing a similar layer-wise approach is fruitless: after obtaining a necessarily distorted estimate $\hat{\boldsymbol{\gamma}}_1$ one cannot proceed with equalities for the next layers, as $\boldsymbol{\gamma}_1$ does not necessarily have a perfectly sparse representation with respect to $\mathbf{D}_2$. Herein we present a solution based on global sparse coding strategy that propagates the obtained solution, at every iteration, to all the convolutional layers, as detailed in Algorithm 1.

This projection algorithm progressively recovers sparse representations to provide a projection for any given signal $\mathbf{y}$.

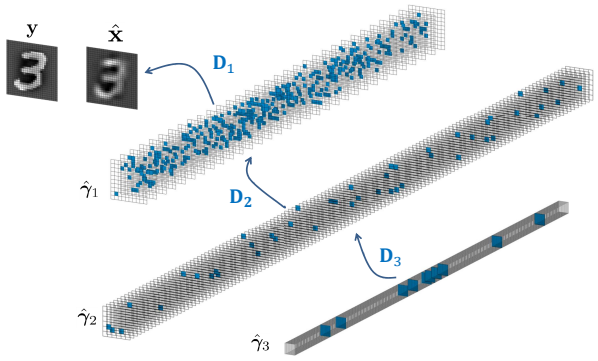

Fig. 2: Decompositions of an MNIST digit in terms of its sparse features $\boldsymbol{\gamma}_i$ and convolutional dictionaries $\mathbf{D}_i$.

The solution is initialized with the zero vector, and then a Pursuit algorithm is applied with a progressively larger $\ell_{0,\infty}$ constraint on the deepest representation, from 1 to λ_L. As shown in [12], several sparse coding methods can be employed (and proven) to solve this CSC problem. At each step, given the estimated $\hat{\boldsymbol{\gamma}}_L$, the intermediate features and their $\ell_{0,\infty}$ norms are computed. If all sparsity constraints are satisfied, then the algorithm proceeds. If, on the other hand, any of the constraints is violated, the previously computed $\mathbf{x}^*$ is reported as the solution.

4. NUMERICAL SIMULATIONS

In this section we demonstrate the ML-CSC Projection Algorithm and compare it with the Layered approach from [11] for the problem of (multi-layer) sparse recovery. We employ the ML-CSC dictionary from [13], trained on a corpus of real digits (MNIST) and consisting of 3 convolutional layers[2]. Importantly, this model allows for projecting digit images onto the ML-CSC model, as illustrated in Figure 2.

To study the recovery of sparse vectors, we take 500 digits from the MNIST dataset and project them on the trained model, essentially running Algorithm 1 and obtaining the representations $\boldsymbol{\gamma}_i$. We then create the noisy measurements as $\mathbf{y} = \mathbf{D}^{(i)}\boldsymbol{\gamma}_i + \mathbf{v}$, where $\mathbf{v}$ is Gaussian noise with $\sigma = 0.02$, providing nothing but noisy digits. In order to evaluate our projection approach, we run Algorithm 1 employing the Subspace Pursuit algorithm [14] for the sparse coding step, with the oracle target cardinality[3] k. Recall that once the deepest representations $\hat{\boldsymbol{\gamma}}_L$ have been obtained, the inner ones are simply computed as $\hat{\boldsymbol{\gamma}}_{i-1} = \mathbf{D}_i\hat{\boldsymbol{\gamma}}_i$. In the layered approach from [11], on the other hand, the pursuit of the representations progresses sequentially: first running a pursuit for $\hat{\boldsymbol{\gamma}}_1$, then employing this estimate to run another pursuit for $\hat{\boldsymbol{\gamma}}_2$, etc. In the same spirit, we employ Subspace Pursuit layer by layer, employing the oracle cardinality of the representation at each stage. The results are presented in Figure 3: at the top

[2]We refer the reader to [13] for more details on this model and its training.

[3]In order to run Algorithm 1, one also needs to define the constants λ_i. For simplicity, we set these to be $\lambda_{i-1} = \|\mathbf{D}_i\|_0\lambda_i$, and $\lambda_L = k$.

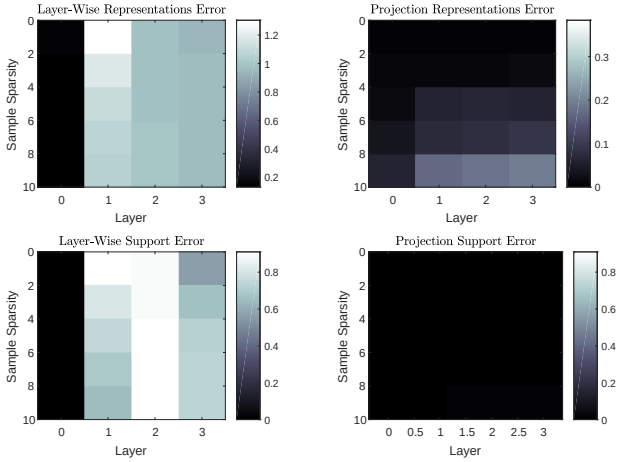

Fig. 3: Top: normalized ℓ_2 error between the estimated and the true representations. Bottom: normalized intersection between the estimated and the true support.

we depict the relative ℓ_2 error of the recovered representations ($\|\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i\|_2/\|\boldsymbol{\gamma}_i\|_2$) and, at the bottom, the normalized intersection of the supports [2], both as a function of the sample cardinality k and the layer depth.

One can see that the ML-CSC Projection Algorithm manages to retrieve the representations $\hat{\boldsymbol{\gamma}}_i$ more accurately than the layered pursuit, as evidenced by the ℓ_2 error and the error in the estimated support. The chief reason behind the difficulty of the layered approach is that the overall success relies on the correct recovery of the first layer representations, $\hat{\boldsymbol{\gamma}}_1$. If these are not properly estimated (as evidenced by the bottom-left graph), there is little hope for the recovery of the deeper ones. The projection alternative, on the other hand, relies on the estimation of the deepest $\hat{\boldsymbol{\gamma}}_L$, which are very sparse. Once these are estimated, the remaining ones are simply computed by propagating them to the shallower layers.

5. CONCLUSION

We have revisited the ML-CSC model and formalized the pursuit of the nested sparse representations in terms of a projection problem. The solution to this problem was shown to be stable, providing bounds that do not scale with the network's depth and that are generally tighter than previous results. We further proposed a simple algorithm to implement this in practice, and demonstrated it for the problem of sparse recovery. Many open questions arise from the ideas presented in this paper. For instance, is the result of the ML-CSC Pursuit algorithm stable? Is this algorithm optimal? how can one guarantee that all $\hat{\boldsymbol{\gamma}}_{i-1}$ would be sparse in a more general case? and, of course, how do we train the filters from real data in an efficient way? All these constitute directions of our current work.

6. ACKNOWLEDGMENTS

The research leading to these results has received funding in part from the European Research Council under EUs 7th Framework Program, ERC under Grant 320649.

7. REFERENCES

[1] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From Sparse Solutions of Systems of Equations to Sparse Modeling of Signals and Images," *SIAM Review.*, vol. 51, no. 1, pp. 34–81, Feb. 2009.

[2] Michael Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*, Springer Publishing Company, Incorporated, 1st edition, 2010.

[3] Yann LeCun, Bernhard E Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne E Hubbard, and Lawrence D Jackel, "Handwritten digit recognition with a back-propagation network," in *Advances in neural information processing systems*, 1990, pp. 396–404.

[4] David E Rumelhart, Geoffrey E Hinton, Ronald J Williams, et al., "Learning representations by back-propagating errors," *Cognitive modeling*, vol. 5, no. 3, pp. 1, 1988.

[5] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[6] Joan Bruna and Stéphane Mallat, "Invariant scattering convolution networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1872–1886, 2013.

[7] Raja Giryes, Guillermo Sapiro, and Alex M Bronstein, "Deep neural networks with random gaussian weights: A universal classification strategy," *CoRR, abs/1504.08291*, 2015.

[8] Nadav Cohen, Or Sharir, and Amnon Shashua, "On the expressive power of deep learning: A tensor analysis," in *29th Annual Conference on Learning Theory*, Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, Eds., Columbia University, New York, New York, USA, 23–26 Jun 2016, vol. 49 of *Proceedings of Machine Learning Research*, pp. 698–728, PMLR.

[9] Karol Gregor and Yann LeCun, "Learning fast approximations of sparse coding," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 399–406.

[10] Bo Xin, Yizhou Wang, Wen Gao, David Wipf, and Baoyuan Wang, "Maximal sparsity with deep networks?," in *Advances in Neural Information Processing Systems*, 2016, pp. 4340–4348.

[11] Vardan Papyan, Yaniv Romano, and Michael Elad, "Convolutional neural networks analyzed via convolutional sparse coding," *To appear in JMLR. arXiv preprint arXiv:1607.08194*, 2016.

[12] Vardan Papyan, Jeremias Sulam, and Michael Elad, "Working locally thinking globally: Theoretical guarantees for convolutional sparse coding," *To appear in IEEE Transactions of signal processing. Preprint arXiv:1607.02009*, 2017.

[13] Jeremias Sulam, Vardan Papyan, Yaniv Romano, and Michael Elad, "Multi-layer convolutional sparse modeling: Pursuit and dictionary learning," *arXiv preprint arXiv:1708.08705*, 2017.

[14] Wei Dai and Olgica Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," *IEEE Transactions on Information Theory*, vol. 55, no. 5, pp. 2230–2249, 2009.