Dense and Sparse Coding: Theory and Architectures

Anonymous Authors

Abstract

The sparse representation model has been successfully utilized in a number of signal and image processing tasks; however, recent research has highlighted its limitations in certain deep-learning architectures. This paper proposes a novel dense and sparse coding model that considers the problem of recovering a dense vector $\mathbf{x}$ and a sparse vector $\mathbf{u}$ given linear measurements of the form $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}$. Our first theoretical result proposes a new natural geometric condition based on the minimal angle between subspaces corresponding to the measurement matrices $\mathbf{A}$ and $\mathbf{B}$ to establish the uniqueness of solutions to the linear system. The second analysis shows that, under mild assumptions and sufficient linear measurements, a convex program recovers the dense and sparse components with high probability. The standard RIPless analysis cannot be directly applied to this setup. Our proof is a non-trivial adaptation of techniques from anisotropic compressive sensing theory and is based on an analysis of a matrix derived from the measurement matrices $\mathbf{A}$ and $\mathbf{B}$. We begin by demonstrating the effectiveness of the proposed model on simulated data. Then, to address its use in a dictionary learning setting, we propose a dense and sparse auto-encoder (DenSaE) that is tailored to it. We demonstrate that a) DenSaE denoises natural images better than architectures derived from the sparse coding model ($\mathbf{Bu}$), b) training the biases in the latter amounts to implicitly learning the $\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}$ model, and c) $\mathbf{A}$ and $\mathbf{B}$ capture low- and high-frequency contents, respectively.

1 Introduction

Given a data set, it is now well accepted that learning a dictionary in which each example admits a sparse representation is tremendously useful in a number of tasks (Aharon et al., 2006; Mairal et al., 2011). This problem, known as sparse coding (Olshausen and Field, 1997) or dictionary learning (Garcia-Cardona and Wohlberg, 2018), has been the subject of significant investigation in recent years in the signal processing community. Convolutional sparse coding (CSC) refers to the case when the dictionary comprises translations of filters. In the deep-learning literature, early work suggests that the hidden layers of successful deep neural networks with ReLU activation functions (Zeiler et al., 2010; Glorot et al., 2011) produce sparse representations of their inputs. Since then, there has been a growing body of work that develops the connection between sparse coding and deep ReLU networks (Gregor and Lecun, 2010; Papyan et al., 2017; Sulam et al., 2018; Tolooshams et al., 2019), further highlighting the importance of sparsity in modern data analysis.

Recent work has highlighted some limitations of the convolutional sparse coding model (Simon and Elad, 2019) and its multi-layer and deep generalizations (Sulam et al., 2019; Zazo et al., 2019). In (Simon and Elad, 2019), the authors argue that the sparsity levels that CSC allows can only accommodate very sparse vectors, making it unsuitable to capture all features of signals such as natural images, which exhibit both edges and smooth, texture-like components. To mitigate this, the authors propose to compute the minimum mean-squared error solution under the CSC model, which is a dense vector that can capture a richer set of features than a sparse one. Starting with a sparse code, multi-layer sparse coding (Sulam et al., 2018) employs a sequence of transformations to generate outputs that are sparse, except for the last, which yields the signal of interest. One limitation

Preprint. Under review.

of this model is that sparsity decreases after each transformation, which puts limits on how deep this model can go (Sulam et al., 2019).

Related work: Given the measurements $\mathbf{y}$, the problem of recovering $\mathbf{x}$ and $\mathbf{u}$ is similar in flavor to sparse recovery in the union of dictionaries (Donoho and Huo, 2001; Elad and Bruckstein, 2002; Donoho and Elad, 2003; Soltani and Hegde, 2017; Studer et al., 2011; Studer and Baraniuk, 2014). Most results in this literature take the form of an uncertainty principle that relates the sum of the sparsity of $\mathbf{x}$ and $\mathbf{u}$ to the mutual coherence between $\mathbf{A}$ and $\mathbf{B}$, and which guarantees that the representation is unique and identifiable by ℓ_1 minimization.

To address the aforementioned limitations of classical sparse coding, we propose a dense and sparse coding model that represents a signal as the sum of two components: one that admits a dense representation $\mathbf{x}$ in a dictionary $\mathbf{A}$, and another whose representation $\mathbf{u}$ is sparse in a second dictionary $\mathbf{B}$. In (Zazo et al., 2019), the authors argue that the multi-layer extension of this model can, in principle, have arbitrary depth. To our knowledge, the dense and sparse coding model has not been yet fully analyzed. Our contributions are

Conditions for identifiability and recovery by convex optimization: We begin by deriving theoretical conditions, expressed via the minimum principal angle between the column space of $\mathbf{A}$ and the span of s columns in $\mathbf{B}$, under which the dense and sparse representation is unique. The proposed condition is geometric and provides an interpretable framework to guide a suitable choice of measurement matrices. Then, we propose a convex program for recovery that minimizes $||\mathbf{A}\mathbf{x}||_2^2 + ||\mathbf{u}||_1$, subject to linear constraints. To our knowledge, the analysis of this program is novel and in sharp contrast to classical settings in sparse approximation, in which the objective consists of a single sparsifying norm, rather than the combination of different norms. Robust PCA (Candès et al., 2011), which decomposes a matrix as the sum of low-rank and sparse matrices, uses the combination of the ℓ_1 and nuclear norms, giving it a flavor similar to our problem. Our analysis uses some techniques from RIPless compressed sensing (Kueng and Gross, 2014). One key challenge for our setup is accounting for the two operators $\mathbf{A}$ and $\mathbf{B}$. The properties of a matrix derived from both are crucial for the success of the convex program. Our analysis provides a template for programs that optimize the combination of norms.

Phase-transition curves: We demonstrate through simulations that the convex program can successfully solve the dense and sparse coding problem. We give plots of the probability of successful recovery as a function of the sparsity of $\mathbf{u}$ and the number of measurements. The smaller the size of $\mathbf{A}$ compared to $\mathbf{B}$, the easier the recovery. Understandably, for a given sparsity level, recovery of both $\mathbf{u}$ and $\mathbf{x}$ requires more measurements than the classical setting where $\mathbf{x} = \mathbf{0}$.

An application to image denoising: We formulate a *dense* and sparse dictionary learning problem and use deep unfolding (Hershey et al., 2014; Monga et al., 2019) to design a neural network, the dense and sparse auto-encoder (DenSaE), that solves the proposed problem. With supervised training of DenSaE, we demonstrate its superiority over architectures derived from the sparse coding model, and show that, when training the biases of the latter, these implicitly learn the dense and sparse model. We also show that $\mathbf{A}$ and $\mathbf{B}$ respectively span low- and high-frequency subspaces.

Notation: Lowercase and uppercase boldface letters denote column vectors and matrices, respectively. Given a vector $\mathbf{x} \in \mathbb{R}^n$ and a support set $S \subset \{1, ..., n\}$, $\mathbf{x}_S$ denotes the restriction of $\mathbf{x}$ to indices in S. For a matrix $\mathbf{A} \in \mathbb{R}^{m \times p}$, $\mathbf{A}_S$ is a submatrix of size $m \times |S|$ with column indices in S. The column space of a matrix $\mathbf{A}$ (the span of the columns of $\mathbf{A}$) is designated by $\mathrm{Col}(\mathbf{A})$, its null space by $\mathrm{Ker}(\mathbf{A})$. We denote the Euclidean, ℓ_1 and ℓ_∞ norms of a vector, respectively as $||\mathbf{x}||_2$, $||\mathbf{x}||_1$, and $||\mathbf{x}||_\infty$. The operator and infinity norm of a matrix $\mathbf{A}$ are respectively denoted as $||\mathbf{A}||$ and $||\mathbf{A}||_\infty$. The sign function, applied componentwise to a vector $\mathbf{x}$, is denoted by $\mathrm{sgn}(\mathbf{x})$. The indicator function is denoted by $\mathbb{1}$. The column vector $\mathbf{e}_i$ denotes the vector of zeros except a 1 at the i-th location. The orthogonal complement of a subspace $\boldsymbol{W}$ denoted by $\mathbf{W}^\perp$. The operator $\mathcal{P}_\mathbf{W}$ denotes the orthogonal projection operator onto the subspace $\mathbf{W}$.

Organization: **Section 2** discusses the theoretical analysis of the dense and sparse coding problem. Numerical and denoising experiments appear in **Section 3**. We conclude in **Section 4**.

2 Theoretical Analysis

The dense and sparse coding problem studies the solutions of the linear system $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}$. Given matrices $\mathbf{A} \in \mathbb{R}^{m\times p}$ and $\mathbf{B} \in \mathbb{R}^{m\times n}$ and a vector $\mathbf{y} \in \mathbb{R}^m$, the goal is to provide conditions under which there is a unique solution $(\mathbf{x}^*, \mathbf{u}^*)$, where $\mathbf{u}^*$ is s-sparse, and an algorithm for recovering it.

2.1 Uniqueness results for the feasibility problem

In this subsection, we study the uniqueness of solutions to the linear system accounting for the different structures the measurement matrices $\mathbf{A}$ and $\mathbf{B}$ can have. For more details of all the different cases we consider, we refer the reader to **Appendix A**. The main result of this first part of the analysis is Theorem 3 which, under a natural geometric condition based on the minimum principal angle between the column space of $\mathbf{A}$ and the span of s columns in $\mathbf{B}$, establishes a uniqueness result for the dense and sparse coding problem. Since the vector $\mathbf{u}$ in the proposed model is sparse, we consider the classical setting of an overcomplete measurement matrix $\mathbf{B}$ with $n \gg m$. The next theorem provides a uniqueness result assuming a certain direct sum representation of the space $\mathbb{R}^m$.

Theorem 1 *Assume that there exists at least one solution to $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}$, namely the pair $(\mathbf{x}^*, \mathbf{u}^*)$. Let S, with $|S| = s$, denote the support of $\mathbf{u}^*$. If $\mathbf{A}$ and $\mathbf{B}_S$ have full column rank and $\mathbb{R}^m = \mathrm{Col}(\mathbf{A}) \oplus \mathrm{Col}(\mathbf{B}_S)$, the only unique solution to the linear system, with the condition that any feasible s-sparse vector $\mathbf{u}$ is supported on S, is $(\mathbf{x}^*, \mathbf{u}^*)$.*

Proof 1 *Let $(\mathbf{x}, \mathbf{u})$, with $\mathbf{u}$ supported on S, be another solution pair. It follows that $[\mathbf{A} \quad \mathbf{B}_S]\begin{bmatrix}\mathbf{x}-\mathbf{x}^*\\ \mathbf{u}_S-\mathbf{u}_S^*\end{bmatrix} = \mathbf{0}$. Noting that the matrix $[\mathbf{A} \quad \mathbf{B}_S]$ has full column rank, the homogeneous problem admits the trivial solution implying that $\mathbf{x}-\mathbf{x}^* = \mathbf{0}$ and $\mathbf{u}_S - \mathbf{u}_S^* = \mathbf{0}$. Therefore, $(\mathbf{x}^*, \mathbf{u}^*)$ is the only unique solution.*

The uniqueness result in the above theorem hinges on the representation of the space $\mathbb{R}^m$ as the direct sum of the subspaces $\mathrm{Col}(\mathbf{A})$ and $\mathrm{Col}(\mathbf{B}_S)$. We use the definition of the minimal principal angle between two subspaces, and its formulation in terms of singular values (Björck and Golub, 1973), to derive an explicit geometric condition for the uniqueness analysis of the linear system.

Definition 2 *Let $\mathbf{U} \in \mathbb{R}^{m\times r}$ and $\mathbf{V} \in \mathbb{R}^{m\times q}$ be matrices whose columns are the orthonormal basis of $\mathrm{Col}(\mathbf{A})$ and $\mathrm{Col}(\mathbf{B})$ respectively. The minimum principal angle between the subspaces $\mathrm{Col}(\mathbf{A})$ and $\mathrm{Col}(\mathbf{B})$ is defined as follows*

$$\mu(\mathbf{U}, \mathbf{V}) = \max_{i,j} \langle \mathbf{u}_i, \mathbf{v}_j \rangle, \tag{1}$$

where $\mathbf{u}_i$ denotes the i-th column of $\mathbf{U}$ and $\mathbf{v}_j$ denotes the j-th column of $\mathbf{V}$. The minimum angle $\mu(\mathbf{U}, \mathbf{V})$ is also equal to the largest singular value of $\mathbf{U}^T\mathbf{V}$, $\mu(\mathbf{U}, \mathbf{V}) = \sigma_1(\mathbf{U}^T\mathbf{V})$.

Theorem 3 *Assume that there exists at least one solution to $\mathbf{y} = \mathbf{A}\mathbf{x}+\mathbf{B}\mathbf{u}$, namely the pair $(\mathbf{x}^*, \mathbf{u}^*)$. Let S, with $|S| = s$, denote the support of $\mathbf{u}^*$. Assume that $\mathbf{A}$ and $\mathbf{B}_S$ have full column rank . Let $\mathbf{U} \in \mathbb{R}^{m\times r}$ and $\mathbf{V} \in \mathbb{R}^{m\times q}$ be matrices whose columns are the orthonormal bases of $\mathrm{Col}(\mathbf{A})$ and $\mathrm{Col}(\mathbf{B}_S)$ respectively. If $\mu(\mathbf{U}, \mathbf{V}) = \sigma_1(\mathbf{U}^T\mathbf{V}) < 1$, the only unique solution to the linear system, with the condition that any feasible s-sparse vector $\mathbf{u}$ is supported on S, is $(\mathbf{x}^*, \mathbf{u}^*)$.*

Proof 2 *Consider any candidate solution pair $(\mathbf{x}^* + \mathbf{x}, \mathbf{u}^* + \mathbf{u})$. We will prove uniqueness by showing that $\mathbf{A}\mathbf{x} + \mathbf{B}_S\mathbf{u}_S = 0$ if and only if $\mathbf{x} = \mathbf{0}$ and $\mathbf{u}_S = \mathbf{0}$. Using the orthonormal basis set $\mathbf{U}$ and $\mathbf{V}$, $\mathbf{A}\mathbf{x} + \mathbf{B}_S\mathbf{u}_S$ can be represented as follows: $\mathbf{A}\mathbf{x} + \mathbf{B}_S\mathbf{u}_S = [\mathbf{U} \quad \mathbf{V}]\begin{bmatrix}\mathbf{U}^T\mathbf{A}\mathbf{x}\\ \mathbf{V}^T\mathbf{B}_S\mathbf{u}_S\end{bmatrix}$. For simplicity of notation, let $\mathbf{K}$ denote the block matrix: $\mathbf{K} = [\mathbf{U} \quad \mathbf{V}]$. If we can show that the columns of $\mathbf{K}$ are linearly independent, it follows that $\mathbf{A}\mathbf{x} + \mathbf{B}_S\mathbf{u}_S = \mathbf{0}$ if and only if $\mathbf{A}\mathbf{x} = \mathbf{0}$ and $\mathbf{B}_S\mathbf{u}_S = \mathbf{0}$. We now consider the matrix $\mathbf{K}^T\mathbf{K}$ which has the following representation*

$$\mathbf{K}^T\mathbf{K} = \begin{bmatrix}[\mathbf{I}]_{r\times r} & [\mathbf{U}^T\mathbf{V}]_{r\times q}\\ [\mathbf{V}^T\mathbf{U}]_{q\times r} & [\mathbf{I}]_{q\times q}\end{bmatrix} = \begin{bmatrix}[\mathbf{I}]_{r\times r} & [\mathbf{0}]_{r\times q}\\ [\mathbf{0}]_{q\times r} & [\mathbf{I}]_{q\times q}\end{bmatrix} + \begin{bmatrix}[\mathbf{0}]_{r\times r} & [\mathbf{U}^T\mathbf{V}]_{r\times q}\\ [\mathbf{V}^T\mathbf{U}]_{q\times r} & [\mathbf{0}]_{q\times q}\end{bmatrix}.$$

With the singular value decomposition of $\mathbf{U}^T\mathbf{V}$ being $\mathbf{U}^T\mathbf{V} = \mathbf{Q}\mathbf{\Sigma}\mathbf{R}^T$, the last matrix in the above representation has the following equivalent form $\begin{bmatrix}\mathbf{0} & \mathbf{U}^T\mathbf{V}\\ \mathbf{V}^T\mathbf{U} & \mathbf{0}\end{bmatrix} = \begin{bmatrix}\mathbf{Q} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{bmatrix}\begin{bmatrix}\mathbf{0} & \mathbf{\Sigma}\\ \mathbf{\Sigma} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{Q} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{bmatrix}^T$.

It now follows that $\begin{bmatrix} \mathbf{0} & \mathbf{U}^T\mathbf{V} \\ \mathbf{V}^T\mathbf{U} & \mathbf{0} \end{bmatrix}$ *is* similar *to the matrix* $\begin{bmatrix} \mathbf{0} & \mathbf{\Sigma} \\ \mathbf{\Sigma} & \mathbf{0} \end{bmatrix}$. *Hence, the nonzero eigenvalues of* $\mathbf{K}^T\mathbf{K}$ *are* $1 \pm \sigma_i$, $1 \leq i \leq \min(p, q)$, *with* σ_i *denoting the* i*-th largest singular value of* $\mathbf{U}^T\mathbf{V}$. *Using the assumption* $\sigma_1 < 1$ *results the bound* $\lambda_{\min}\left(\mathbf{K}^T\mathbf{K}\right) > 0$. *It follows that the columns of* $\mathbf{K}$ *are linearly independent, and hence* $\mathbf{A}\mathbf{x} = \mathbf{0}$ *and* $\mathbf{B}_S\mathbf{u}_S = \mathbf{0}$. *Since* $\mathbf{A}$ *and* $\mathbf{B}_S$ *are assumed to have full column rank,* $\mathbf{x} = \mathbf{0}$ *and* $\mathbf{u}_S = \mathbf{0}$. *This concludes the proof.*

A restrictive assumption of the above theorem is that the support of the sought-after s-sparse solution $\mathbf{u}^*$ is known. We can remove this assumption by considering $\mathrm{Col}(\mathbf{A})$ and $\mathrm{Col}(\mathbf{B}_T)$ where T is an arbitrary subset of $\{1, 2, ..., n\}$ with $|T| = s$. More precisely, we state the following corollary whose proof is similar to the proof of Theorem 3.

Corollary 4 *Assume that there exists at least one solution to* $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}$, *namely the pair* $(\mathbf{x}^*, \mathbf{u}^*)$. *Let* S, *with* $|S| = s$, *denote the support of* $\mathbf{u}^*$ *and* T *be an arbitrary subset of* $\{1, 2, ..., n\}$ *with* $|T| = s$. *Assume that* $\mathbf{A}$ *and* $\mathbf{B}_T$ *have full column rank (for any choice of* T*) . Let* $\mathbf{U} \in \mathbb{R}^{m\times p}$ *and* $\mathbf{V} \in \mathbb{R}^{m\times q}$ *be matrices whose columns are the orthonormal bases of* $\mathrm{Col}(\mathbf{A})$ *and* $\mathrm{Col}(\mathbf{B}_S)$ *respectively. If* $\mu(\mathbf{U}, \mathbf{V}) = \sigma_1(\mathbf{U}^T\mathbf{V}) < 1$, *holds for all choices of* T, *the only unique solution to the linear system is* $(\mathbf{x}^*, \mathbf{u}^*)$ *with the condition that any feasible* $\mathbf{u}$ *is* s*-sparse.*

The mutual coherence plays a key role for the success of sparse approximation in the union of bases (Donoho and Elad, 2003). While the definitions of the minimum principle angle and the mutual coherence condition look similar, they are markedly different: the former relates *subspaces*, while the latter relates pairs of vectors. An open question is to determine the class of matrices for which the minimum angle bound holds. In the next subsection, we propose a convex program to recover the dense and sparse vectors. Theorem 7 establishes uniqueness and complexity results for the proposed optimization program.

2.2 Dense and sparse recovery via convex optimization

Given that the dense and sparse coding problem seeks a dense vector $\mathbf{x}^*$ and a sparse solution $\mathbf{u}^*$, we propose the following convex optimization program

$$\min_{\mathbf{x},\mathbf{u}} \; ||\mathbf{A}\mathbf{x}||_2^2 + ||\mathbf{u}||_1 \quad \text{s.t.} \quad \mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} = \mathbf{A}\mathbf{x}^* + \mathbf{B}\mathbf{u}^*. \tag{2}$$

In this section, we show that, under certain conditions, the above minimization problem admits a unique solution. Our proof is a non-trivial adaptation of the existing analysis in (Kueng and Gross, 2014) for the anistropic compressive sensing problem. This analysis is based on a single measurement matrix and can not be directly applied to our scenario. Let $\mathbf{a}_1, ..., \mathbf{a}_m$ be a sequence of zero-mean i.i.d random vectors drawn from some distribution F on $\mathbb{R}^p$ and let $\mathbf{b}_1, ..., \mathbf{b}_m$ be a sequence of zero-mean i.i.d random vectors drawn from some distribution G on $\mathbb{R}^n$. We can eliminate the dense component in the linear constraint by projecting the vector $\mathbf{y}$ onto the orthogonal complement of $\mathrm{Col}(\mathbf{A})$ to obtain $\mathcal{P}_{\mathrm{Col}(\mathbf{A})^\perp}(\mathbf{y}) = \mathcal{P}_{\mathrm{Col}(\mathbf{A})^\perp}(\mathbf{B}\mathbf{u})$. With this, the matrix $\mathcal{P}_{\mathrm{Col}(\mathbf{A})^\perp}(\mathbf{B})$ is central in the analysis to follow. We denote the i-th measurement vector, corresponding to a row of this matrix, with $\mathbf{c}_i = [\mathcal{P}_{\mathrm{Col}(A)^\perp}(\mathbf{B})]^T\mathbf{e}_i$ and $\mathbf{C} = \frac{1}{\sqrt{m}}\sum_{i=1}^m \mathbf{e}_i\mathbf{c}_i^T$. Let $\mathbf{\Sigma}$ denote the covariance matrix $\mathbf{\Sigma} = E[\mathbf{c}_i\mathbf{c}_i^T]^{\frac{1}{2}}$. Further technical discussion on the matrix $\mathbf{C}$ is deferred to **Appendix B**. We use the measurement matrix $\mathbf{C}$ introduced above and adapt the anisotropic compressive sensing theory in (Kueng and Gross, 2014) to analyze uniqueness of the proposed program. Below, we give brief background to this theory highlighting important assumptions and results following the notation closely therein.

Anisotropic compressive sensing: Given a sequence of zero-mean i.i.d random vectors $\mathbf{d}_1, ..., \mathbf{d}_m$ drawn from some distribution F on $\mathbb{R}^n$, the anisotropic compressive sensing problem studies the following optimization program

$$\min_{\mathbf{u}} \; ||\mathbf{u}||_1 \quad \text{s.t.} \quad \mathbf{y} = \mathbf{D}\mathbf{u} = \mathbf{D}\mathbf{u}^*, \tag{3}$$

where $\mathbf{D} = \frac{1}{\sqrt{m}}\sum_{i=1}^m \mathbf{e}_i\mathbf{d}_i^T$ and $\mathbf{u}^*$ is the sought-out sparse solution. The analysis makes three important assumptions.

Completeness: The covariance matrix $\mathbf{\Sigma}$ is invertible with condition number denoted by κ.

Incoherence: The incoherence parameter is the smallest number ν such that

$$\max_{1\leq i\leq n} |\langle \mathbf{d}\,, \mathbf{e}_i\rangle|^2 \leq \nu, \quad \max_{1\leq i\leq n} |\langle \mathbf{d}\,, E[\mathbf{c}\mathbf{c}^*]^{-1}\mathbf{e}_i|^2 \leq \nu, \tag{4}$$

hold almost surely.

Conditioning of the covariance matrix: We start with the following definition of the s-sparse condition number restated from (Kueng and Gross, 2014).

Definition 5 *(Kueng and Gross, 2014) The largest and smallest s-sparse eigenvalue of a matrix $\mathbf{X}$ are given by*

$$\lambda_{\max}(s, \mathbf{X}) := \max_{\mathbf{v}, ||\mathbf{v}||_0\leq s} \frac{||\mathbf{X}\mathbf{v}||_2}{||\mathbf{v}||_2}, \quad \lambda_{\min}(s, \mathbf{X}) := \min_{\mathbf{v}, ||\mathbf{v}||_0\leq s} \frac{||\mathbf{X}\mathbf{v}||_2}{||\mathbf{v}||_2}. \tag{5}$$

The s-sparse condition number of $\mathbf{X}$ is cond$(s, \mathbf{X}) = \dfrac{\lambda_{\max}(s, \mathbf{X})}{\lambda_{\min}(s, \mathbf{X})}$.

Given these assumptions, the main result in (Kueng and Gross, 2014) reads

Theorem 6 *(Kueng and Gross, 2014) With $\kappa_s = \max\{$cond$(s, \mathbf{\Sigma})$, cond$(s, \mathbf{\Sigma}^{-1})\}$ let $\mathbf{u} \in \mathbb{C}^n$ be an s-sparse vector and let $\omega \geq 1$. If the number of measurements fulfills $m \geq C\kappa_s\, \nu\, \omega^2\, s \log n$, then the solution $\mathbf{u}$ of the convex program* (3) *is unique and equal to $\mathbf{u}^*$ with probability at least $1 - e^{-\omega}$.*

The proof of Theorem 6 is based on the dual certificate approach. The idea is to first propose a dual certificate vector $\mathbf{v}$ with sufficient conditions that ensure uniqueness of the minimization problem. It then remains to construct the dual certificate satisfying the conditions. We seek a similar result for the uniqueness of the convex program corresponding to the dense and sparse coding model. However, the standard analysis can not be directly applied since it only considers a single measurement matrix. This requires us to analyze the matrix $\mathbf{C}$ introduced earlier. The anisotropic compressive sensing analysis in (Kueng and Gross, 2014) assumes the following conditions on the dual certificate $\mathbf{v}$

$$||\mathbf{v}_S - \operatorname{sgn}(\mathbf{u}^*_S)||_2 \leq \tfrac{1}{4} \;\text{ and }\; ||\mathbf{v}_{S^\perp}||_\infty \leq \tfrac{1}{4}. \tag{6}$$

The following condition follows from the assumptions in Theorem 6

$$||\mathbf{\Delta}_S||_2 \leq 2||\mathbf{\Delta}_{S^\perp}||_2, \tag{7}$$

where $\mathbf{\Delta} \in \text{Ker}(\mathbf{D})$. The conditions (6) and (7) will be used in the proof of our main result. The main part of the technical analysis in (Kueng and Gross, 2014) is using the assumptions in Theorem 6 and showing that the above conditions (6) and (7) hold with high probability.

Main result: Using the the background discussed above, we assume completeness, incoherence, and conditioning of the covariance matrix $\mathbf{\Sigma}$. Our main result is stated below.

Theorem 7 *Assume that there exists at least one solution to $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}$, namely the pair $(\mathbf{x}^*, \mathbf{u}^*)$. Let $\omega \geq 1$ and define $\kappa_s = \max\{$cond$(s, \mathbf{\Sigma})$, cond$(s, \mathbf{\Sigma}^{-1})\}$. Assume the two conditions*

$$||\mathbf{B}_S^T\mathbf{A}|| \leq \frac{1}{32||\mathbf{x}^*||_2}, \quad ||\mathbf{B}_{S^\perp}^T\mathbf{A}||_\infty \leq \frac{1}{32||\mathbf{x}^*||_\infty}. \tag{8}$$

If the number of measurements fulfills $m \geq C\kappa_s\, \nu\, \omega^2\, s \log n$, then the solution of the convex program (2) *is unique and equal to $(\mathbf{x}^*, \mathbf{u}^*)$ with probability at least $1 - e^{-\omega}$.*

Proof sketch 1 *Consider a feasible solution pair $(\mathbf{x}^* + \mathbf{\Delta_1}, \mathbf{u}^* + \mathbf{\Delta_2})$ and let the function $f(\mathbf{x}, \mathbf{u})$ denote the objective in the optimization program. The idea of the proof is to show that any feasible solution is not minimal in the objective value, $f(\mathbf{x}^* + \mathbf{\Delta_1}, \mathbf{u}^* + \mathbf{\Delta_2}) > f(\mathbf{x}, \mathbf{u})$. Using duality of the ℓ_1 norm and characterization of the subgradient $\mathbf{\Lambda}$ of the ℓ_1 norm, we first show that $f(\mathbf{x}^* + \mathbf{\Delta_1}, \mathbf{u}^* + \mathbf{\Delta_2}) > f(\mathbf{x}^*, \mathbf{u}^*) + \langle \operatorname{sgn}(u^*) + \mathbf{\Lambda} - \mathbf{v} - 2\mathbf{B}^T\mathbf{A}\mathbf{x}^*\,, \mathbf{\Delta_2}\rangle$ where $\mathbf{v} \in \text{Col}(\mathbf{C}^T)$, with $\mathbf{C} = \mathbf{B} - \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{B}$ denoting the dual certificate. It then remains to show that the term $\langle \operatorname{sgn}(\mathbf{u}^*) + \mathbf{\Lambda} - \mathbf{v} - 2\mathbf{B}^T\mathbf{A}\mathbf{x}^*\,, \mathbf{\Delta_2}\rangle$ is positive . To show this, we further analyze this term and make use of the assumptions of the theorem, the dual certificate conditions* (6)*, and the deviation inequality in* (7) *to arrive at the desired result. For complete proof, see **Appendix B**.*

Complexity compared to ℓ_1 minimization: The sample complexity of solving the convex program corresponding to the dense and sparse coding problem is larger than that of ℓ_1 minimization for the compressive sensing problem. Essentially, the constants κ_s and ν in our analysis are expected to scale with $p + n$, in contrast to the compressive sensing analysis where they scale with n.

3 Experiments

3.1 Phase transition curves

We generate *phase transition curves* and present how the success rate of the recovery, using the proposed model, changes under different scenarios. To generate the data, we fix the number of columns of $\mathbf{B}$ to be $n = 100$. Then, we vary the sampling ratio $\sigma = \frac{m}{n+p} \in [0.05, 0.95]$ and the sparsity ratio $\rho = \frac{s}{m}$ in the same range. The sensing matrix in our model is $[\mathbf{A} \quad \mathbf{B}]$, hence the apparent difference in the definition of σ compared to "traditional" compressive sensing. In the case where we revert to the compressive sensing scenario ($p = 0$), the ratios coincide.

We generate random matrices $\mathbf{A} \in \mathbb{R}^{m \times p}$ and $\mathbf{B} \in \mathbb{R}^{m \times n}$ whose columns have expected unit norm. The vector $\mathbf{u} \in \mathbb{R}^n$ has s randomly chosen indices, whose entries are drawn according to standard normal distribution, and $\mathbf{x} \in \mathbb{R}^p$ is generated as follows: we generate a random vector $\boldsymbol{\gamma} \in \mathbb{R}^m$, and then construct $\mathbf{x} = \mathbf{A}^T\boldsymbol{\gamma}$. The construction ensures that $\mathbf{x}$ does not belong in the null space of $\mathbf{A}$, and hence ignores trivial solutions with respect to this dense component. We normalize both $\mathbf{x}$ and $\mathbf{u}$ to have unit norm, and generate the measurement vector $\mathbf{y} \in \mathbb{R}^m$ as $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}$. We solve the convex optimization problem in (2) to obtain the numerical solution pair $(\hat{\mathbf{x}}, \hat{\mathbf{u}})$ using `CVXPY`, and register a successful recovery if both $\frac{\|\hat{\mathbf{x}}-\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \leq \epsilon$ and $\frac{\|\hat{\mathbf{u}}-\mathbf{u}\|_2}{\|\mathbf{u}\|_2} \leq \epsilon$, with $\epsilon = 10^{-3}$. For each choice of σ and ρ we average 100 independent runs to estimate the success rate.

Figure 1 shows the phase transition curves for $p \in \{0.1m, 0.5m\}$ to highlight different ratios between p and n. We observe that increasing p leads to a deterioration in performance. This is expected, as this creates a greater *overlap* on the spaces spanned by $\mathbf{A}$ and $\mathbf{B}$. We can view our model as explicitly modeling the noise of the system. In such a case, the number of columns of $\mathbf{A}$ explicitly encodes the complexity of the noise model: as p increases, so does the span of the noise space.

Extending the signal processing interpretation, note that we model the noise signal $\mathbf{x}$ as a dense vector, which can be seen as encoding smooth areas of the signal that correspond to *low-frequency* components. On the contrary, the signal $\mathbf{u}$ has, by construction, a sparse structure, containing *high-frequency* information, an interpretation that will be further validated in the next subsection. Further numerical experiments comparing the dense and sparse coding model to the conventional compressive sensing problem can be found in **Appendix C**.

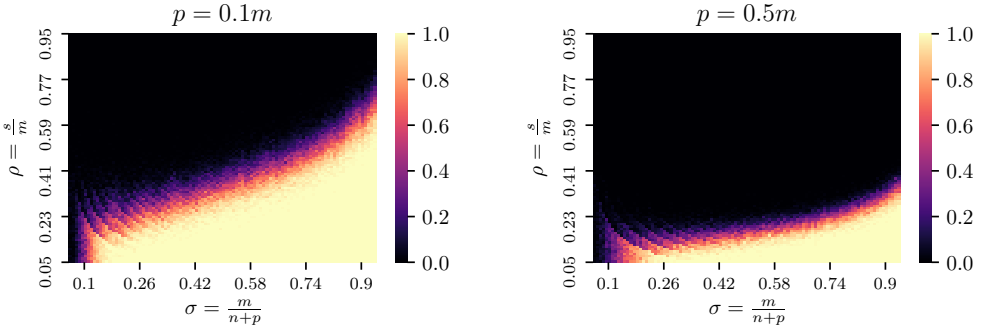


Figure 1: Phase transition curves for $p = 0.1m$ (left) and $p = 0.5m$ (right).

3.2 Natural image denoising

We formulate the dense and sparse dictionary learning problem as follows

$$\min_{\mathbf{A},\mathbf{B},\{\mathbf{x}^j\}_{j=1}^J,\{\mathbf{u}^j\}_{j=1}^J} \sum_{j=1}^{J} \frac{1}{2}\|\mathbf{y}^j - \mathbf{A}\mathbf{x}^j - \mathbf{B}\mathbf{u}^j\|_2^2 + \frac{1}{2\lambda_x}\|\mathbf{A}\mathbf{x}^j\|_2^2 + \lambda_u\|\mathbf{u}^j\|_1, \tag{9}$$

where J is the number of images, λ_x controls the smoothness of $\mathbf{A}\mathbf{x}^j$ (parameter of a Gaussian prior on $\mathbf{A}\mathbf{x}^j$), and λ_u (associated with a Laplace prior) controls the degree of sparsity. We minimize this bi-convex objective function by constructing an auto-encoder architecture, which we term the dense and sparse auto-encoder (DenSaE).

The encoder architecture is a recurrent network that performs dense and sparse coding: it maps $\mathbf{y}^j$ into a dense representation $\mathbf{x}_T^j$ and a sparse one $\mathbf{u}_T^j$ by unfolding (Gregor and Lecun, 2010; Simon and Elad, 2019; Tolooshams et al., 2019) T iterations of the following proximal gradient algorithm

$$\begin{aligned}\mathbf{x}_t^j &= \mathbf{x}_{t-1}^j + \alpha_x(\mathbf{A}^{\mathrm{T}}(\mathbf{y}^j - (1+\frac{1}{\lambda_x})\mathbf{A}\mathbf{x}_{t-1}^j - \mathbf{B}\mathbf{u}_{t-1}^j)),\\ \mathbf{u}_t^j &= \mathcal{S}_b\big(\mathbf{u}_{t-1}^j + \alpha_u\mathbf{B}^{\mathrm{T}}(\mathbf{y}^j - \mathbf{A}\mathbf{x}_{t-1}^j - \mathbf{B}\mathbf{u}_{t-1}^j)\big),\end{aligned} \tag{10}$$

where α_x and α_u are step sizes, and $\mathcal{S}_b$, with $b = \alpha_u\lambda_u$, is the Shrinkage operator for general sparse coding (Tolooshams et al., 2019), and ReLU for non-negative sparse coding

$$\text{Shrinkage}_b(z) = \text{ReLU}_b(z) - \text{ReLU}_b(-z), \qquad \text{ReLU}_b(z) = (z-b)\cdot\mathbb{1}_{z\geq b}. \tag{11}$$

The decoder uses the dense and sparse estimates $\mathbf{x}_T^j$ and $\mathbf{u}_T^j$ to reconstruct the image $\hat{\mathbf{y}}^j = \mathbf{A}\mathbf{x}_T^j + \mathbf{B}\mathbf{u}_T^j$. The dictionaries $\mathbf{A}$ and $\mathbf{B}$ can now be learned through back-propagation by minimizing the reconstruction loss $\mathcal{L}_{\mathbf{A},\mathbf{B}} = \frac{1}{J}\sum_{j=1}^{J}\frac{1}{2}\|\mathbf{y}^j - \hat{\mathbf{y}}^j\|_2^2$. The DenSaE architecture is shown in **Appendix D**. We examined the following questions

(a) How does the denoising performance change as the number of filters in $\mathbf{A}$ vs. $\mathbf{B}$ varies?
(b) What is the performance of DenSaE compared to networks, such as CSCNet (Simon and Elad, 2019), that perform sparse coding ($\mathbf{y} = \mathbf{B}\mathbf{u}$)?
(c) What characteristics of the images does the model capture?
(d) How do sparse coding networks, such as CSCNet (Simon and Elad, 2019), behave as the regularization parameter (i.e., bias) is trained?

We address the case when $\mathbf{A}$ and $\mathbf{B}$ are strided convolutional matrices, with little or no overlap between translations of the filters. Convolutional models with strides equal to the filter size are equivalent to patch-based models with dense dictionary matrices (Pfister and Bresler, 2019), namely the model we analyzed in **Section 2**. We evaluate the model in the presence of Gaussian noise with standard deviation of $\tau = \{15, 25, 50, 75\}$, and follow an approach similar to (Simon and Elad, 2019) when reconstructing an input image.

We trained DenSaE for image denoising in a supervised manner using 432 images from the Berkeley Segmentation Dataset (BSD432) and tested it on images from BSD68 (Martin et al., 2001). We used a non-informative prior on $\mathbf{A}\mathbf{x}$ (i.e., $\lambda_x \to \infty$). The network contains a total of 64 filters of size 7×7 with strides of 5. We varied the ratio of number of filters in $\mathbf{A}$ and $\mathbf{B}$ as the overall number of filters was kept constant. As baselines, we trained two variants, $\text{CSCNet}_{\text{hyp}}^{\text{tied}}$ and $\text{CSCNet}_{\text{LS}}^{\text{tied}}$, of CSCNet (Simon and Elad, 2019). Both networks are auto-encoders. Within each network, we tied the weights so that they are interpretable as dictionaries learned for the sparse coding generative model. In $\text{CSCNet}_{\text{hyp}}^{\text{tied}}$, the bias is a shared hyper-parameter. In $\text{CSCNet}_{\text{LS}}^{\text{tied}}$, we learn a different bias for each filter by minimizing the reconstruction loss. In both cases, the bias controls the sparsity of the feature maps. Further details of the network architectures and training parameters are summarized in **Appendix D**.

Table 1: DenSaE's denoising performance on test BSD68 as the ratio of filters in $\mathbf{A}$ and $\mathbf{B}$ changes.

τ	$\frac{1\mathbf{A}}{63\mathbf{B}}$	$\frac{4\mathbf{A}}{60\mathbf{B}}$	$\frac{8\mathbf{A}}{56\mathbf{B}}$	$\frac{16\mathbf{A}}{48\mathbf{B}}$	$\frac{32\mathbf{A}}{32\mathbf{B}}$
15	**30.21**	30.18	30.18	30.14	29.89
25	**27.70**	**27.70**	27.65	27.56	27.26
50	**24.81**	**24.81**	24.43	24.44	23.68
75	23.31	**23.33**	23.09	22.09	20.09

Ratio of number of filters in $\mathbf{A}$ and $\mathbf{B}$: Table 1 shows that, for a range of noise levels, the smaller the number of filters associated with $\mathbf{A}$, the better DenSaE can denoise images. We hypothesize that this is a direct consequence of our findings from **Section 2** that the smaller the number of columns of $\mathbf{A}$, the easier the recovery $\mathbf{x}$ and $\mathbf{u}$. Indeed, it would appear that the success of the dictionary learning step relies on the success of the recovery step, similar to classical dictionary learning.

Dense and sparse coding vs. sparse coding: Table 2 shows that DenSaE (best network from Table 1) denoises images better than $\text{CSCNet}_{\text{hyp}}^{\text{tied}}$, suggesting that the dense and sparse coding model models images better than sparse coding. We discuss the performance of $\text{CSCNet}_{\text{LS}}^{\text{tied}}$ below.

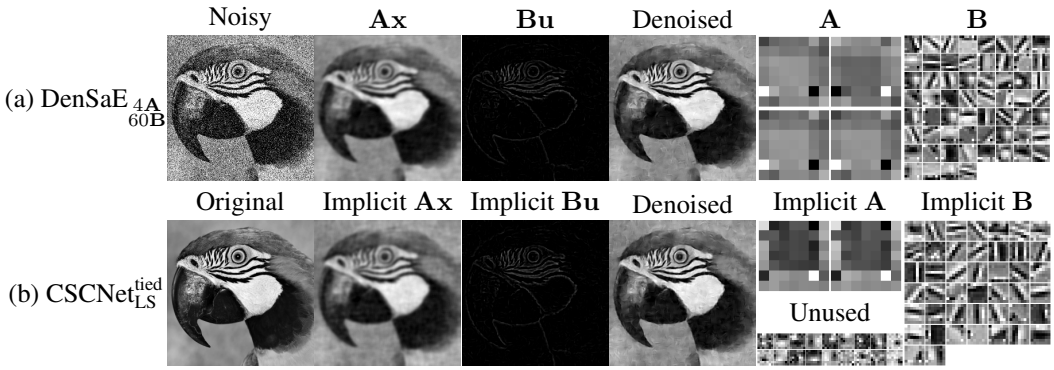


Figure 2: Visualization of a test image for $\tau = 50$. (a) DenSaE $(4\mathbf{A}, 60\mathbf{B})$, (b) CSCNet$^{\text{tied}}_{\text{LS}}$.

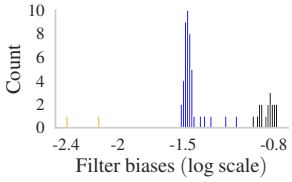
Dictionary characteristics: Figure 2(a) shows the decomposition of a noisy test image ($\tau = 50$) by DenSaE, into its $\mathbf{Ax}$ and $\mathbf{Bu}$ components. The figure demonstrates that $\mathbf{Ax}$ captures low-frequency content, *despite the use of a non-informative prior*, while $\mathbf{Bu}$ captures high-frequency details (edges). This is corroborated by the smoothness of the filters associated with $\mathbf{A}$, and the Gabor-like nature of those associated with $\mathbf{B}$ (Mehrotra et al., 1992). Because the filters are much smaller than the images, locally, frequencies around zero (DC) dominate low frequencies. Thus, it is not surprising that the $\mathbf{A}$ filters look constant. This is, likely, another reason why DenSaE denoise images well with very few low-frequency filters (Table 1). We observed similar performance when we tuned λ_x. We found that, as λ_x decreases, $\mathbf{Ax}$ captures a lower range of frequencies, and $\mathbf{Bu}$ a broader range.

Table 2: Performance of DenSaE on test BSD68 against CSCNet.

τ	DenSaE	CSCNet$^{\text{tied}}_{\text{hyp}}$	CSCNet$^{\text{tied}}_{\text{LS}}$
15	30.21	30.12	**30.34**
25	27.70	27.51	**27.75**
50	**24.81**	24.54	**24.81**
75	**23.33**	22.83	23.32

CSCNet implicitly learns $\mathbf{Ax} + \mathbf{Bu}$ model: Figure 3 shows that CSCNet$^{\text{tied}}_{\text{LS}}$ comprises three groups of filters: one with small bias values, one with intermediate ones, and a third with large values. We found that the feature maps associated with the large bias values are all zero, i.e., they do not contribute to the representation. We also found that the majority of feature maps are associated with intermediate bias values, and are sparse, in contrast to the small number of feature maps associated with small bias values, which are dense. These observations suggest that *auto-encoder architectures implementing the sparse coding model ($\mathbf{y} = \mathbf{Bu}$), when learning the biases by minimizing reconstruction error, implicitly perform two functions.* First, they automatically select the optimal number of filters. Second, they automatically partition the filters into two groups: one that yields a dense representation of the input, and another that yields a sparse one. In other words, the architectures trained in this fashion *implicitly learn the dense and sparse coding model* ($\mathbf{y} = \mathbf{Ax} + \mathbf{Bu}$). That's why DenSaE and CSCNet$^{\text{tied}}_{\text{LS}}$ perform similarly. Figure 2(b) shows the filters associated with each of the three groups of bias values described above.

Figure 3: Biases from CSCNet$^{\text{tied}}_{\text{LS}}$ ($\tau = 50$).



4 Conclusions

This paper proposed a novel dense and sparse coding model for a flexible representation of a signal as $\mathbf{y} = \mathbf{Ax} + \mathbf{Bu}$. Our first result gives a verifiable condition that guarantees uniqueness of the model. Our second result uses tools from RIPless compressed sensing to shows that, with sufficiently many linear measurements, a convex program with ℓ_1 and ℓ_2 regularizations can recover the components $\mathbf{x}$ and $\mathbf{u}$ uniquely with high probability. Numerical experiments on synthetic data confirm our observations. Finally, we proposed a dense and sparse auto-encoder, DenSaE, tailored to the $\mathbf{Ax} + \mathbf{Bu}$ model. We showed that DenSaE is superior to networks implementing the sparse coding model, and shed light on the implicit behavior of these networks when their biases are trained. We also found that DenSaE naturally decomposes signals into low- and high-frequency components.

Broader Impact

Two criticisms of modern, deep neural network architectures are their lack of interpretability, and the fact that training them can exert a tremendous footprint on the planet. Starting from a generative model, deep unfolding/algorithm unrolling refers to the process of converting inference algorithms that arise from the model into a deep neural network architecture. In contrast to a standard deep network, whose weights at different layers are distinct by default, the layers of architectures derived by deep unrolling share weights, which they inherit from the generative model. In other words, the generative model constrains the weights, as well as the activation functions of the deeply unrolled network. Consequently, a deeply-unrolled network, unlike a conventional deep neural network, is interpretable. In addition, because it has significantly fewer parameters, its carbon footprint is much smaller.

We introduced a novel generative model, the dense and sparse coding model, that is a generalization of the classical sparse coding model. Deep unrolling, in the context of the sparse coding model, gives rise to ReLU networks, which are interpretable and have significantly fewer parameters than conventional ReLU networks. For theoretical reasons, multi-layer extensions of sparse coding give rise to architectures with limited depth. In contrast, the dense and sparse coding model we propose can, in principle, give rise to architectures with arbitrary depth, *without* increasing the number of parameters of the associated networks. Thus, deep unrolling of dense and sparse coding is a principled way towards the design of interpretable, very deep neural networks that exert a low carbon footprint on the planet.

References

M. Aharon, M. Elad, and A. Bruckstein, "k-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on signal processing*, vol. 54, no. 11, pp. 4311–22, 2006.

J. Mairal, F. Bach, and J. Ponce, "Task-driven dictionary learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 4, pp. 791–804, 2011.

B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by v1?" *Vision research*, vol. 37, no. 23, pp. 3311–3325, 1997.

C. Garcia-Cardona and B. Wohlberg, "Convolutional dictionary learning: A comparative review and new algorithms," *IEEE Transactions on Computational Imaging*, vol. 4, no. 3, pp. 366–81, Sep. 2018.

M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks," in *Proc. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2010, pp. 2528–35.

X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. 14th International Conference on Artificial Intelligence and Statistics*, 2011, pp. 315–23.

K. Gregor and Y. Lecun, "Learning fast approximations of sparse coding," in *Proc. International Conference on Machine Learning (ICML)*, 2010, pp. 399–406.

V. Papyan, Y. Romano, and M. Elad, "Convolutional neural networks analyzed via convolutional sparse coding," *Journal of Machine Learning Research*, vol. 18, pp. 1–52, 2017.

J. Sulam, V. Papyan, Y. Romano, and M. Elad, "Multilayer convolutional sparse modeling: Pursuit and dictionary learning," *Signal Processing, IEEE Transactions on*, vol. 66, no. 15, pp. 4090–104, 2018.

B. Tolooshams, S. Dey, and D. Ba, "Deep residual auto-encoders for expectation maximization-based dictionary learning," *arXiv preprint arXiv:1904.08827*, 2019.

D. Simon and M. Elad, "Rethinking the csc model for natural images," in *Proc. Advances in Neural Information Processing Systems*, 2019, pp. 2271–2281.

J. Sulam, A. Aberdam, A. Beck, and M. Elad, "On multi-layer basis pursuit, efficient algorithms and convolutional neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2019.

J. Zazo, B. Tolooshams, and D. Ba, "Convolutional dictionary learning in hierarchical networks," in *Proc. 2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. IEEE, 2019, pp. 131–135.

D. L. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE transactions on information theory*, vol. 47, no. 7, pp. 2845–2862, 2001.

M. Elad and A. M. Bruckstein, "A generalized uncertainty principle and sparse representation in pairs of bases," *IEEE Transactions on Information Theory*, vol. 48, no. 9, pp. 2558–2567, 2002.

D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization," *Proc. the National Academy of Sciences*, vol. 100, no. 5, pp. 2197–2202, 2003.

M. Soltani and C. Hegde, "Fast algorithms for demixing sparse signals from nonlinear observations," *IEEE Transactions on Signal Processing*, vol. 65, no. 16, pp. 4209–4222, 2017.

C. Studer, P. Kuppinger, G. Pope, and H. Bolcskei, "Recovery of sparsely corrupted signals," *IEEE Transactions on Information Theory*, vol. 58, no. 5, pp. 3115–3130, 2011.

C. Studer and R. G. Baraniuk, "Stable restoration and separation of approximately sparse signals," *Applied and Computational Harmonic Analysis*, vol. 37, no. 1, pp. 12–35, 2014.

E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM (JACM)*, vol. 58, no. 3, pp. 1–37, 2011.

R. Kueng and D. Gross, "Ripless compressed sensing from anisotropic measurements," *Linear Algebra and its Applications*, vol. 441, pp. 110–123, 2014.

J. R. Hershey, J. L. Roux, and F. Weninger, "Deep unfolding: Model-based inspiration of novel deep architectures," *arXiv preprint arXiv:1409.2574*, 2014.

V. Monga, Y. Li, and Y. C. Eldar, "Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing," *arXiv preprint arXiv:1912.10557*, 2019.

A. Björck and G. H. Golub, "Numerical methods for computing angles between linear subspaces," *Mathematics of computation*, vol. 27, no. 123, pp. 579–594, 1973.

L. Pfister and Y. Bresler, "Learning filter bank sparsifying transforms," *IEEE Transactions on Signal Processing*, vol. 67, no. 2, pp. 504–519, Jan 2019.

D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. 8th Int'l Conf. Computer Vision*, vol. 2, July 2001, pp. 416–423.

R. Mehrotra, K. Namuduri, and N. Ranganathan, "Gabor filter-based edge detection," *Pattern Recognition*, vol. 25, no. 12, pp. 1479–94, 1992.