DICTIONARY LEARNING UNDER GENERATIVE COEFFICIENT PRIORS, WITH APPLICATIONS TO COMPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

There is a rich literature on recovering data from limited measurements under the assumption of sparsity in some basis, whether known (compressed sensing) or unknown (dictionary learning). In particular, classical dictionary learning assumes the given dataset is well-described by sparse combinations of an unknown basis set. However, this assumption is of limited validity on real-world data. Recent work spanning theory and computational science has sought to replace the canonical sparsity assumption with more complex data priors, demonstrating how to incorporate pretrained generative models into frameworks such as compressed sensing and phase retrieval. Typically, the dimensionality of the input space of the generative model is much smaller than that of the output space, paralleling the "low description complexity," or compressibility, of sparse vectors. In this paper, we study dictionary learning under this kind of known generative prior on the coefficients, which may capture non-trivial low-dimensional structure in the coefficients. This is a distributional learning approach to compression, in which we learn a suitable dictionary given access to a small dataset of training instances and a specified generative model for the coefficients. Equivalently, it may be viewed as transfer learning for generative models, in which we learn a new linear layer (the dictionary) to fine-tune a pretrained generative model (the coefficient prior) on a new dataset. We give, to our knowledge, the first provable algorithm for recovering the unknown dictionary given a suitable initialization. Finally, we compare our approach to traditional dictionary learning algorithms on synthetic compression and denoising tasks, demonstrating empirically the advantages of incorporating finer-grained structure than sparsity.

1 INTRODUCTION

Sparsity is ubiquitous throughout data science as a tractable formalization of the structure of naturally occurring signals. Early work in signal processing involved defining specific bases, such as Daubechies wavelets, of which natural images are well-approximated by sparse linear combinations, as well as other domain-specific bases such as curvelets and ridgelets (Daubechies, 1988; Candès & Demanet, 2002; Candès, 2003). These innovations alone revolutionized signal processing, yielding such recognizable technologies as JPEG image compression. While such basis transform methods for compression required a linear number of measurements, i.e. using all n pixels of an image, a revolution came with the landmark works of Donoho and Candes, Romberg, and Tao, establishing the field of compressed sensing (Donoho, 2006; Candes et al., 2006). In particular, they demonstrated that if an n-dimensional signal is k-sparse in a known basis that satisfies the restricted isometry property (e.g. from i.i.d. Gaussian or Fourier matrices), its sparse coefficients can be efficiently recovered using only $O(k \log n)$ linear measurements. Letting $A \in \mathbb{R}^{m \times n}$ with $m << n$ denote the measurement matrix and $x \in \mathbb{R}^n$ the unknown signal, one measures $y = Ax \in \mathbb{R}^m$; the signal x is recovered by minimizing $||x||_1$ subject to the m measurements. This innovation has inspired many offshoots and applications, notably providing a substantial speedup to MRI (Lustig et al., 2007).

In the previous settings, the linear measurements were either randomized, as in the case of compressed sensing, or deterministic, as in the case of wavelet bases, but they were always fully known. In contrast, dictionary learning — alternately known as sparse coding — posits that the signal is

sparse in an *unknown* basis, which may be specific to the task or dataset at hand. Thus, one must learn the matrix A, whose columns are sometimes referred to as atoms, simultaneously to learning the sparse coefficients x. This would be impossible given only a single measurement vector y, but can be done given multiple measurements of the form $y^j = Ax^j$. Dictionary learning has proven useful in a variety of applications, including denoising (Elad & Aharon, 2006), facial recognition (Xu et al., 2017), feature extraction (Ceylan, 2018), and medical image processing (Zhao et al., 2021; Li et al., 2012). In this work, we are particularly interested in applications to compression and denoising, which we emphasize are applications that make sense even when the dictionary is not uniquely defined. Practical algorithms for solving dictionary learning include the Method of Optimal Directions (Engan et al., 1999) and K-SVD (Aharon et al., 2006). Under certain randomized generative models on the sparse coefficients x, alternating minimization algorithms similar to MOD have been shown to run in polynomial time and provide provable guarantees on the output solution to the sparse coding problem (Arora et al., 2014; Agarwal et al., 2014; Arora et al., 2015).

However, it is important to note that sparsity is only an *approximation* to the structure of real-world signals. Images are not truly sparse in a particular dictionary; sparsity simply provides a tractable model that has been useful in applications. In fact, recent empirical results demonstrate that sparsity is too coarse an assumption in practice, as deep neural networks can provide a far better approximation to real-world signals. State-of-the-art audio and image processing — including denoising, featurization, classification, inpainting, and beyond – is now almost universally achieved by deep learning methods. For example, the top accuracies of $> 99.8\%$ on the Labeled Faces in the Wild Database (Huang et al., 2007), once a testbed for dictionary learning and hand-featurization methods, are now attained entirely by deep nets trained using outside data. The advent of generative adversarial networks (GANs), variational autoencoders (VAEs), and other such schemes have led to generative priors capable of generating photorealistic, yet fake, images (Karras et al., 2017), which far surpass the reconstruction quality that results from simply assuming sparsity in a particular basis. Correspondingly, there has been a renaissance in solving inverse problems via learned generative priors. Such works have spanned both scientific and vision domains, including phase retrieval (Hand et al., 2018), image inpainting (Yeh et al., 2017), super-resolution (Ledig et al., 2016), and denoising (Lempitsky et al., 2018). Such works demonstrate that the shared representation space of deep generative models is more powerful on real-world data than sparsity.

From a theoretical perspective, Bora et al. (2017) introduced a corresponding *theory* of compressed sensing (a particular inverse problem) under generative models, in which the sparse signal x to be recovered is instead constrained only to lie in the range of a generative model, $x = G(z)$. They show that simple empirical risk minimization suffices to recover x with sample complexity m on the order of $k = \dim(z)$, comparably to the canonical k-sparse case.

Regularizing reconstruction problems via generative priors, instead of traditional sparsity priors, has thus been a productive endeavor at the interface of signal processing and machine learning. In this paper, we naturally extend this line of work to dictionary learning by considering generative priors for the coefficients $x = G(z)$, where z lives in a small, k-dimensional input space and parallels the low description complexity of k-sparsity. Here, the naturally occurring data is thus modeled by $A^*G(z)$ and not $G(z)$ itself (a departure from the interpretation of Bora et al. (2017)). Particular constructions for G can capture k-sparsity (Kamath et al., 2019), recovering ordinary dictionary learning, but introducing a generative model for coefficients allows us to model strictly more complex statistical relationships among the features as well.

To understand this concretely, consider the following thought experiment. Recall that ordinary dictionary learning provides a model for how structure is embedded in commonly observed signals, via sparse combinations of a common pool of atoms. However, the dependencies between these individual features is important for providing a finer-grained model of structure. For example, an atom representing the leg of an item of furniture is useful on its own, but in practice we expect it to co-occur only in certain combinations with other atoms, such as the seat of a chair or the surface of a table. One can iterate this reasoning further, e.g. chairs themselves may only appear together with tables or desks. If the data y is made up of such atomic building blocks, which enjoy complex hierarchical structure in their co-occurrence patterns, then modeling $y = AG(z)$ is natural. This is not just hypothetical: there is precedent across multiple domains for incorporating more complex structures than sparsity in dictionary learning. For example, multi-layer convolutional sparse coding assumes that the dictionary coefficients are themselves representable as a sparse combination of separate dictionary elements, and so on, thus forming a hierarchical linear model of composed

dictionaries (Papyan et al., 2017; Sulam et al., 2018). In addition, Li et al. (2012) proposes a model of "group sparsity" for medical image denoising, in which certain groups of atoms are more likely to co-occur together, according to predefined overlapping subsets. As we show in Section 6, this structure can be captured by a simple one-layer generative model. In fact, the ability of deep nets to capture hierarchical structure is thought to be a key driver of their success, particularly when learning common structure that generalizes across label classes. Thus, incorporation of a generative coefficient prior $G(z)$ generalizes sparsity, group sparsity, and multi-layer dictionary learning, while also drawing on the myriad successes of deep learning in capturing complex structure in data.

Crucially, we assume throughout that this coefficient prior G (e.g. the group identities) is known. (If $G(z)$ itself were unknown as well, this problem would reduce to the quite general one of learning a complete nonlinear generative model for a dataset from scratch. This is itself an open problem for each type of generative model, see e.g. Kodali et al. (2018) .) Access to such a G can occur through transfer learning: one might have a large dataset D', for which an unsupervised method can be used to learn a generative model $M(z)$, as well as a smaller but related dataset D. If $M(z) = F(G(z))$ is a deep net, one can remove the final few layers to obtain a reasonable $G(z)$ for use as a coefficient prior with D. We explore this setting in the experiments in Section 6.

1.1 Our contributions

In this paper, we formulate a new version of dictionary learning that incorporates non-trivial dependencies in the coefficients via a generative prior. We give an intuitive alternating minimization algorithm (Section 4) for recovering the unknown dictionary, with provable convergence guarantees under variants of standard assumptions in the literature (Section 3). The strongest of these assumptions is a decoding optimization oracle, which is also needed by previous work on compressed sensing under generative priors (Bora et al., 2017). However, to validate this assumption, we extend the tools of Hand & Voroninski (2018) to our more challenging case, where A^* is unknown. In particular, we show that the optimization landscape is favorable (in the sense that there is always a direction of descent) at every iteration of our algorithm in a particular idealized setting, without noise and when the generative model is an expansive Gaussian ReLU network (Section 5). Finally, we adapt our algorithm into a PyTorch implementation and demonstrate that it outperforms classic dictionary learning algorithms on denoising tasks (Section 6). We obtain more accurate reconstructions (using only the k free parameters of the generative model, as compared to k nonzero coefficients) from noisy data, validating our choice of both model and algorithm. More generally, we hope the flexible hybrid of dictionary learning and deep generative priors proposed here will open the door to the application of dictionary learning techniques to a far broader scope of applications, thereby bringing the deep learning revolution to dictionary learning.

Notation Given a matrix M, let $(M)_i$ denote the i^{th} column of M. For a vector u, let u_i denote its i^{th} entry. In contrast, we will use superscripts to denote sample indices, i.e. $y^j = A^*G(z^j) + \eta^j$ is the j^{th} vector sample. Let $B_p(u, r) = \{x : ||x - u||_p \leq r\}$. Similarly, let $B_p(M, r)$ denote all matrices whose i^{th} columns lie in $B_p((M)_i, r)$. Let Supp(μ) denote the support of a probability distribution μ. We say that $f(x) = O(g(x))$ if $\lim_{x\to\infty} \frac{f(x)}{g(x)} \leq C$ for some constant C, and $f(x) = \omega(g(x))$ if $\lim_{x\to\infty} \frac{f(x)}{g(x)}$ diverges to infinity. Let $u^{\odot n}$ denote a vector's n^{th} element-wise power, i.e. $(u^{\odot n})_i = (u_i)^n$. Let $\|M\|$ or $\|M\|_2$ denote the spectral norm of the matrix M.

2 Related work

Dictionary learning. Dictionary learning was first studied in the neuroscience community in the seminal work of Olshausen & Field (1997). Efficient algorithms were proposed in the following decade, including the Method of Optimal Directions (Engan et al., 1999) and K-SVD (Aharon et al., 2006). Since then it has found important applications in many domains, e.g. medical imaging (see (Zhao et al., 2021) and references therein). Supervised variants of dictionary learning have been proposed (Mairal et al., 2009b), and provable algorithms under different distributional, dimension, and sparsity assumptions were given (Chatterji & Bartlett, 2017; Spielman et al., 2012). Closest to the alternating minimization algorithm we propose is the work of Arora et al. (2015), who first gave strong convergence guarantees for a neurally plausible instantiation of alternating minimization and

whose convergence analysis loosely inspired ours. Hong et al. (2018) studied a union-of-subspaces model, and Li et al. (2012) proposed a related group sparsity structure, both of which are in the same spirit of generalized structural dependencies among coefficients. The compositional coefficient model of multi-layer convolutional sparse coding is also similar to our work: the signal model $y = A_1^* A_2^* \dots A_\ell^* z$ provides a non-trivial compositional model for the dictionary coefficients, and the forward pass of a convolutional neural network is shown to essentially decode the current convolutional dictionaries (Papyan et al., 2017; Sulam et al., 2018). However, their coefficient prior is restricted to linear models $A_2^* \dots A_\ell^*$, whereas we consider an arbitrary Lipschitz non-linear $G(z)$. Moreover, these works do not assume any of the composed dictionaries are known, whereas we are more motivated by transfer learning and assume access to G.

Compressed sensing with additional structure. Several preceding works studied compressed sensing under more exotic models than generic sparsity, such as graph-structured sparsity (Hegde et al., 2015), but Bora et al. (2017) initiated the study of compressed sensing under arbitrary generative models. They showed that a suitable optimization oracle for minimizing reconstruction error could recover a good estimate of the unknown non-sparse coefficients, with the required number of measurements scaling with the input dimension of the generative model and the number of layers in the network. A related paper by Hand and Voroninski then gave an efficient implementation of the optimization oracle in an interesting class of models (Hand & Voroninski, 2018), and later used very similar tools to analyze phase retrieval under a generative prior (Hand et al., 2018).

Generative priors for inverse problems. On the practical side, recent works have harnessed both trained and untrained generative priors in inverse problems such as phase retrieval, image inpainting, and denoising. Pretrained generative models are obtained from existing datasets (e.g. that of Tramel et al. (2016)), while untrained generative models are generally more expressive and confer an inductive bias towards natural data by virtue of their architecture alone (e.g. the deep image prior (Lempitsky et al., 2018), and the deep decoder (Heckel & Hand, 2019; Bostan et al., 2020; Lawrence et al., 2021)). In contrast to this paper, the forward model is fully specified in all of these works.

Learning distributional transformations. From a distributional perspective, our problem fits into the broad class of problems of trying to learn a linear transformation A of a known distribution p_x (given by the push-forward of p_z through G). Arora et al. (2012) gives a moment-based algorithm for recovering such an unknown linear transformation A. However, this work only applies when the original distribution p_x satisfies stringent coordinate-wise independence conditions, whereas the main utility of our model is in allowing structured dependencies across coordinates.

3 Problem Formulation

Let $A^* \in \mathbb{R}^{m \times n}$ denote the unknown dictionary, which we assume has unit-normed columns, and let G be a generative model with $G : \mathbb{R}^k \to \mathbb{R}^n$. Measurements are of the form $y^j = A^* G(z^j) + \eta^j$, where each z^j is i.i.d. according to a fixed distribution p_z and η is entry-wise i.i.d. zero-mean noise with $||\eta|| \leq N$. Given vector samples $\{y^1, \dots, y^m\}$ and access to G, our task is to recover a good approximation of A^*. Note that in our complete problem setup, including the S-REC and optimization oracle defined below, a good estimate of A^* enables good approximations of each z^j (see Theorem A.2), so this is a reasonable metric. We make the following additional assumptions:

Assumption 1 (Bounded-norm outputs.) The norms of the outputs of the generative model G are upper-bounded by a constant R: $||G(z)||_2 \leq R \, \forall z \in \text{Supp}(p_z)$. A simple way of achieving this is to assume G is L-Lipschitz, and that $\text{Supp}(p_z) \subseteq B_2(0, \frac{R}{L})$ for some values R and L. Bora et al. (2017) similarly require a Lipschitz network, although the easier nature of the problem allows for a weaker dependence on $\log L$ and for $||z||_2$ to be bounded only with high probability.

Assumption 2 (Set-restricted eigenvalue condition (S-REC).) The S-REC was defined in Bora et al. (2017) as an analogue of the restricted isometry property from classical compressed sensing. It ensures that one can detect, based on the measurements y_i, whether the corresponding generative model outputs were similar. Without such a condition, one could not hope to recover the $G(z_i)$ based on only $A^* G(z_i)$, and thus could not take advantage of G. Since our problem generalizes that of Bora et al. (2017), we must also assume the S-REC: $||A^*(G(z^1) - G(z^2)|| \geq \gamma ||G(z^1) - G(z^2)|| - \delta \, \forall z^1, z^2 \in S$. In particular, we set $S = B_2(0, \frac{R}{L})$, because we have assumed $\text{Supp}(p_z) \subseteq B_2(0, \frac{R}{L})$.

Assumption 3 (Initialization.) We require access to an initial guess A^0 of A^* whose columns are Δ-close in ℓ_2 norm to those of A^*, i.e. $||(A^0)_i - (A^*)_i|| \leq \Delta \; \forall i$. We only need that $\Delta < 1$, although our ultimate guarantees have some dependence on Δ. An initialization assumption is common in the dictionary learning literature, as the problem geometry becomes more favorable (convex) close to the true solution; for instance, Arora et al. (2015) requires $\Delta = \frac{1}{\log n}$, and Schnass (2015) tolerates comparatively higher sparsities (and thus richer signals, much like how G also provides a richer signal model) at the expense of giving only a local analysis. In practice, random initializations of A^0 still performed well in our experiments; see Section 6. However, note that A^* is not necessarily identifiable, even up to permutation. This is dependent on the inherent symmetries of $G(z)$; for example, if $G(z)$ is distributed as $BG(z)$ for a matrix B, then one can never hope to distinguish between A^* and A^*B. Thus, this assumption theoretically breaks any symmetries by starting iterations sufficiently close to *one* viable A^*, i.e. such that $y \sim A^*G(z)$.

Assumption 4 (Optimization oracle.) Given any y and any A that is as close to A^* as the initialization, $||(A)_i - (A^*)_i|| \leq \Delta \; \forall i$, we can compute z in $B_2(0, \frac{R}{L})$ (i.e. over the domain of G) s.t. $||y - AG(z)|| \leq \min_{||\tilde{z}|| \leq \frac{R}{L}} ||y - AG(\tilde{z})|| + \theta$. Note that when $A = A^*$, this is exactly the optimization oracle required in Bora et al. (2017). In Theorem 5.3, we validate the existence of the optimization oracle in certain cases by showing that $||y - AG(\tilde{z})||$ always has a descent direction.

Assumption 5 (Outer product estimate.) Let $C^* = \mathbb{E}_{z^* \sim p_z}[G(z^*)G(z^*)^T]$. We require that C^* is invertible, and that we have access to an estimate C^{-1} such that $||C^{-1} - (C^*)^{-1}||_2 \leq \nu$. Despite the choice of notation, C^{-1} does not need to satisfy any properties (e.g. invertibility) besides the stated one. A simple estimation procedure for obtaining C^{-1} is simply to invert an estimate for C^*.

Assumption 6 (Interplay of parameters.) Finally, our theoretical analysis requires the following technical condition governing the previously introduced parameters: $\left(\nu||C^*|| + 10n||C^{-1}||\frac{R^2}{\gamma}\right) \leq \frac{1}{4}$. This is a technical condition that simplifies the analysis, but can likely be weakened with additional effort, as suggested by the success of our experiments under diverse generative models.

4 Our Algorithm

Algorithm 1: Basic alternating minimization

Inputs: Initialization $A^0 \in B_2(A^*, \Delta)$ and TP i.i.d. samples $\tilde{y}^0, \ldots, \tilde{y}^{TP-1}$, $s = 1, \ldots, T$;
for $s = 0, \ldots, T-1$ **do**
  Consider P fresh samples $(y^0, \ldots, y^{P-1}) = (\tilde{y}^{sP}, \ldots, \tilde{y}^{(s+1)P-1})$
  Decode: Using the optimization oracle, compute $z^0, \ldots, z^{P-1} \in B_2(0, \frac{R}{L})$ such that

$$||y^j - A^sG(z^j)|| \leq \min_{||z*|| \leq \frac{R}{L}} ||y^j - A^sG(z^*)|| + \theta \quad \forall j = 0, \ldots, P-1$$

  Update: $A^{s+1} = \text{Proj}_{\mathcal{B}} A^s - \eta \hat{g}^s$, where $\hat{g}^s = \frac{1}{P}\sum_{p=0}^{P-1}(A^sG(z^p) - y^p)G(z^p)^TC^{-1}$ is a finite-sample estimate for $g^s = \mathbb{E}_{z^*,\eta}[(A^sG(z) - y)G(z)^T] \cdot C^{-1}$ (in which z is a function of the random variables z^* and η and the constant A^s), and $\mathcal{B} = B_2(A^0, \frac{\Delta}{2})$.
end
Return: A^T, an estimate for A^*

To recover A^* given samples $y_1, \ldots, y_m$, one would ideally like to minimize reconstruction over the latent variables $z_1, \ldots, z_m$ and A^* at once by solving $\min_{A^*, z_1, \ldots, z_m} \sum_{i=1}^m ||A^*G(z_i) - y_i||_2^2$. However, the reconstruction objective is non-convex, even for ordinary sparse coding (where G outputs sparse vectors). As is common in MOD algorithms for dictionary learning, we adopt an alternating minimization procedure which alternates between "decoding" a good choice for each z_i based on the current estimate A of A^*, and "updating" the estimate A to better match these z_i. As in Arora et al. (2015), we take a single step in the direction of the gradient of A at each update step, and draw new i.i.d. measurement samples at each iteration for ease of analysis.

The algorithm departs from precedent in pre-conditioning by C^{-1}, the estimate of $\mathbb{E}[G(z^*)G(z^*)]^{-1}$, which we can motivate as follows. Intuitively, observe that $\mathbb{E}_{z^*,\eta}[A^sG(z) - A^*G(z^*)G(z^*)^T] \approx (A^s - A^*)\mathbb{E}[G(z^*)G(z^*)^T]$, when z is close enough to z^*. Furthermore,

$-(A^s - A^*)$ is exactly the direction from the current iterate, A^s, to the correct dictionary, A^*. Thus, the role of C^{-1} is to approximately invert the $\mathbb{E}[G(z^*)G(z^*)^T]$ term, resulting in update steps which make progress by stepping towards the correct answer, A^*.

5 THEORETICAL RESULTS

Under the assumptions of Section 3, we can bound the convergence of Algorithm 1 to the correct A^*. First, for a cleaner statement, Theorem 5.1 provides our result in an idealized noiseless setting:

Theorem 5.1 (Geometric convergence under best circumstances) *Let Assumptions 1-6 of Section 3 hold with $N = \theta = \delta = 0$ and $||C^{-1}|| = 1$. This corresponds to a noiseless setting, with a perfect optimization oracle, zero-valued S-REC offset parameter δ, and unit-normed estimate C^{-1}. Then Algorithm 1 with $\eta = \frac{2}{25}$ yields a solution A with $||(A)_i - (A^*)_i||_2 \leq \epsilon$ in $O\Big(\log\big(\frac{\Delta^2}{\epsilon}\big)\Big)$ iterations, with $O(n^{poly(\log(R),\log(\Delta),\log(1/\epsilon))})$ fresh samples required per iteration.*

Remark 1 *Note that, in contrast to traditional dictionary learning guarantees, there is no explicit dependence in the theorem on k, the underlying dimensionality of the coefficients. This is because it implicitly plays a role in determining the S-REC constants γ and δ; see Bora et al. (2017).*

More generally, we show the following rate of convergence, which is geometric with a bias:

Theorem 5.2 (General convergence rate) *Let Assumptions 1-6 hold, let $\zeta = \theta + N$, $\eta = \frac{2}{25}$, and assume $P = n^{2L}\log^2 n$ fresh samples per iteration. Then the iterates of Algorithm 1 satisfy:*

$$\mathbb{E}[||A_i^t - A_i^*||^2] \leq \Big(\frac{24}{25}\Big)^t ||A_i^0 - A_i^*||^2 + \Big(\frac{400n^2||C^{-1}||^2(\zeta+\delta)^2R^2}{\gamma^2} + \frac{4(n\Delta R + \zeta)R||C^{-1}||}{n^L}\Big)$$

The proof of these theorems is outlined in Subsection 5.1 and given in full detail in Subsection A.2.

Separately, we show a secondary result validating that the optimization oracle is reasonable under particular conditions, borrowing the machinery of Hand & Voroninski (2018). We provide definitions so that this paper is self-contained, but refer to (Hand & Voroninski, 2018) for details.

Suppose that G is an L-layer ReLU net, $G(z) = \text{ReLU}(W_L \dots \text{ReLU}(W_2\text{ReLU}(W_1 z)) \dots)$, where each W_i satisfies the *Weight Distribution Condition (WDC)* with constant ϵ as defined by Hand & Voroninski (2018) (their Definition 2). The full terms of the WDC are technical and we include the complete definition in Subsection A.3, but the WDC essentially requires that the neuron weights at each layer are Gaussian-like in their distribution; it is satisfied with high probability (dependent on ϵ) when $n \geq k \log k$ and each entry of W is i.i.d. $N(0, \frac{1}{n})$. There is also a condition jointly on A^* and G:

Range Restricted Isometry Condition (RRIC). A^* and G satisfy the RRIC with constant ϵ if

$$\frac{\Big|\Big\langle A^*(G(z_1) - G(z_2)), A^*(G(z_3) - G(z_4))\Big\rangle - \Big\langle G(z_1) - G(z_2), G(z_3) - G(z_4)\Big\rangle\Big|}{||G(z_1) - G(z_2)||_2||G(z_3) - G(z_4)||_2} \leq \epsilon$$

This property is similar to the S-REC, as both ask that A^* act approximately like an isometry on outputs of G. One does not strictly imply the other, but the RRIC with $z_1 = z_3$ and $z_2 = z_4$ implies the S-REC with $\delta = 0$ and $\gamma = \sqrt{1-\epsilon}$. We only require the RRIC to be satisfied for A^* and G, because we later show that this induces the RRIC with a slightly worse constant for A close to A^*.

Theorem 5.3 *Let the WDC and RRIC be satisfied with respect to some constant ϵ for A^* and G. Define $c = \epsilon + 2||A^* - A|| \cdot ||A^*|| + ||A^* - A||^2$ and let A be any matrix such that $||A - A^*|| \leq \frac{c}{||A||}$. This is easily achieved if e.g. $||A^*|| \geq \frac{1}{2}||A||$. We require that $K_1 L^8 \epsilon^{1/4} \leq 1$ and $K_1 L^8 c^{1/4} \leq 1$ for an absolute constant K_1. Finally, assume there is no measurement noise: we observe $y^i = A^*G(z^i)$. Let $D_v f(x) = \lim_{t\to 0^+} \frac{f(x+tv)-f(x)}{t||v||_2}$. Then when optimizing the objective function $\frac{1}{2}||AG(z) - A^*G(z^*)||^2$ with respect to z, there is always a descent direction $-v_{z,z^*}$ outside of a*

small ball around z^* and its negative multiple:

$$D_{-v_{z,z^*}} f(z) < -K_3 \frac{\sqrt{\epsilon} L^3}{2^L} \max(||z||, ||z^*||) \ \forall x \notin B(z^*, R_1) \cup B(-\rho_L z^*, R_2) \cup \{0\}$$

Here, $R_1 = K_2 L^3 \epsilon^{1/4} ||z^*||$ and $R_2 = K_2 L^{13} \epsilon^{1/4} ||z^*||$. Moreover, 0 is a local maximum: $D_y f(0) < -\frac{1}{8\pi 2^L} ||z^*|| \ \forall y \neq 0$. As in Hand & Voroninski (2018), K_1, K_2, and K_3 are constants, and $\rho_L > 0$ converges to 1 as the number of network layers L goes to infinity.

Thus when A is sufficiently close to A^*, the optimization objective of the decoding step always admits a descent direction, outside of small neighborhoods around the correct z^* and its negative scalar multiple. A first order method such as Algorithm 1 of Hand et al. (2018) is then guaranteed to converge to a good solution. This validates the optimization oracle in a certain setting. The proof of Theorem 5.3 is in Subsection A.3 and outlined in Subsection 5.2.

5.1 Proof of convergence (outline)

The core of the proof of Theorem 5.2 is to show that the infinite-sample gradient at every update step, g^s, has sufficiently large inner product (column-wise) with the direction of the correct dictionary, $A^s - A^*$, in the following sense. Given an iterative algorithm with steps of the form $v^{s+1} = v^s - \eta g^s$ and desired solution v^*, we say the vector g^s is **$(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\epsilon_s})$-correlated** with $v^s - v^*$ if $\langle g, v^s - v^* \rangle \geq \alpha ||z^s - z^*||^2 + \beta ||g^s||^2 - \epsilon_s$. Moreover, a random vector $\hat{g}^s$ is **$(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\epsilon_s})$-correlated-w.h.p.** if this holds with probability $1 - n^{-\omega(1)}$. In our case, we will think of v^s as a column of iterate A^s, v^* as the corresponding column of A^*, and g^s as it is already defined in Algorithm 1. Given such a property, we make progress at every iteration, and v^s converges geometrically to v^* (see e.g. Theorem 6 of (Arora et al., 2015)). To prove that correlation holds for $\hat{g}^s$, we formalize several intuitive facts that hold at the beginning of each update step. For ease of notation, let z refer to one of the z^j computed in the decode step, z^* to the latent variable generating measurement y^j, and A to A^s. First, we can bound the reconstruction error that was computed in the decode step (all proofs in Subsection A.2):

Lemma 5.4 (Closeness between y and $AG(z)$.) *Let $A \in B_2(A^*, \Delta)$, $\Sigma = A - A^*$, and $y = A^* G(z^*) + \eta$ for some unknown z^*. We can apply the optimization oracle to y with our current guess A to give an estimate z. The resultant decoding objective value is then bounded as follows:*

$$||y - AG(z)|| \leq ||A^* - A|| \cdot R + \theta + ||\eta|| = ||\Sigma|| R + \theta + ||\eta||$$

Combining Theorem 5.4 with the S-REC, we then bound $||G(z) - G(z^*)||$ in terms of $||A - A^*||$:

Lemma 5.5 (Closeness between $G(z)$ and $G(z^*)$.) *Let A be an arbitrary matrix and $\Sigma = A - A^*$. Let z be the output of the optimization oracle corresponding to A with any fixed measurement generated as $y = A^* G(z^*) + \eta$. Then we have:*

$$||G(z) - G(z^*)|| \leq \frac{2||\Sigma|| R + \theta + \delta + ||\eta||}{\gamma}$$

Theorem 5.5 implies that, given a good estimate A of A^*, the optimization oracle can find a good estimate z of z^*. Finally, by assumption, we have that $\mathbb{E}[G(z^*) G(z^*)^T]^{-1} \approx C^{-1}$. Putting all of these facts, together, one can obtain as desired that:

$$\begin{aligned} g^s &= \mathbb{E}_{z^*,\eta}[(A^s G(z) - y) G(z)^T] \cdot C^{-1} = \mathbb{E}_{z^*,\eta}[(A^s G(z) - A^* G(z^*) - \eta) G(z)^T] \cdot C^{-1} \\ &\approx \mathbb{E}_{z^*}[(A^s G(z^*) - A^* G(z^*)) G(z^*)^T] \cdot C^{-1} \approx A^s - A^* \end{aligned}$$

In practice, we don't have access to g^s exactly; instead, we approximate the expectation with P fresh samples, $\hat{g}^s = \frac{1}{P} \sum_{p=0}^{P-1} (AG(z^p) - y^p) G(z^p)^T C^{-1}$. If $\hat{g}^s$ is close enough to g^s, however, then $\hat{g}^s$ will remain correlated-w.h.p. with the correct direction $A^s - A^*$:

Theorem 5.6 ($\hat{g}^s$ is correlated-w.h.p. with $A^s - A^*$) *Draw $P = n^{2L} \log^2 n$ fresh samples at every iteration, and assume initialization $A^0 \in B_2(A^*, \Delta)$. Then $\hat{g}_i^s$ is $(\frac{1}{4}, \frac{1}{25}, 100 n^2 ||C^{-1}||^2 \frac{(\theta + \delta + ||\eta||)^2 R^2}{\gamma^2} + \frac{(n\Delta R + \theta + N) R ||C^{-1}||}{n^L})$-correlated-w.h.p. with $A_i^s - A_i^*$.*

Finally, Theorem 40 of Arora et al. (2015) converts correlation-w.h.p. into a statement of biased geometric convergence to the optimum. Applying this theorem with our Theorem 5.6 yields our biased geometric convergence result, Theorem 5.2. For all technical details, see Subsection A.2.

(a) Dependencies: $m = 10$, $n = 9$.

(b) Group sparsity: $m = 784$, $n = 300$.

Figure 1: Error bars show one standard deviation in log-scale over 4 trials, with $n_B = 5$ and $B = 30$.

5.2 Proof of descent directions at decoding step (outline)

The original work of Hand & Voroninski (2018) proves the existence of descent directions for the decoding step when $A = A^*$, which implicitly assumes A^* is known. In contrast, we seek to *learn* A^*, and at each iteration minimize a modified loss function $f = ||AG(z) - A^*G(z^*)||^2$ where $A \approx A^*$. However, one could hope that the optimization landscape of f closely resembles that of $\bar{f}(z) = ||AG(z) - AG(z^*)||^2$ when A is a good approximation of A^*, and this is indeed how we prove Theorem 5.3. To reason about the descent directions of $||AG(z) - AG(z^*)||^2$, we must prove that A itself also satisfies the RRIC. A calculation (Theorem A.9) shows that if A^* satisfies the RRIC with respect to G with constant ϵ, then any A sufficiently close to A^* satisfies the RRIC with constant $c\epsilon + 2||A^* - A|| \cdot ||A^*|| + ||A^* - A||^2$. Then, the rest of Hand and Voroninski's original proof goes through with the added error from $||\nabla f - \nabla \bar{f}||$; see Subsection A.3 for details.

6 Experiments

We compare dictionary learning with and without a generative prior on a variety of experiments using PyTorch, and demonstrate that inclusion of accurate, finer-grained coefficient priors than sparsity can broadly improve reconstruction quality. In all experiments, we begin with a specified generative prior. In the first set of experiments, we fix a ground truth A^* and obtain synthetic measurements as $y \sim A^*G(z^*) + \eta$, where $z^* \sim N(0, I_{k\times k})$ and $\eta \sim N(0, \sigma^2 I_{m\times m})$ with $\sigma = 0.1$. The second set of experiments are instead transfer learning problems: given a generative model pre-trained on a large dataset, we remove the final layer to obtain a coefficient prior G for a related, but much smaller dataset. (Thus there is no ground truth A^*.) This will demonstrate the practical value of using a generative coefficient prior, even when it is not explicitly used in data generation.

Both the ground-truth A^*, when applicable, and the entries of A^0, are initialized uniformly at random between 0 and 1. As baselines, we use the mini-batched implementation of MOD from scikit-learn (Mairal et al., 2009a), and an open-source implementation of K-SVD (Rubinstein et al., 2008). As implied by the theory, we consider an independent batch of samples at each iteration of the minimization; in subsequent error plots, all methods are evaluated after each successive batch of samples (on the x-axes). However, it is helpful to make a few practical modifications to Algorithm 1. We use the Adam optimizer for all optimizations (Kingma & Ba, 2015), and obtain good performance without explicitly pre-conditioning by C^{-1} or projecting onto $B_2(A^0, \frac{\Delta}{2})$. We then employ two main variants. The first variant, "altmin", most closely follows Algorithm 1. To perform the optimization in the decoding step, we use a learning rate of 10^{-2} with 1000 iterations of the Adam optimizer. In the update step, we find it helpful to run 1000 iterations of the Adam optimizer with learning rate 10^{-3} instead of taking a single gradient step. In the second variant, "autodiff", we optimize over A and z_i at once with respect to the nonconvex reconstruction objective in Section 4 for each batch of samples. Our algorithmic guarantees do not apply in this setting, but it turns out to be an efficient and practical heuristic. Further experiments and details are included in the Appendix.

We do not directly compare the dictionaries learned from each method to a "ground-truth" dictionary, even when one is available, because two models can learn different dictionaries that both represent the data well. In the absence of identifiability, how can one quantify whether a "good" dictionary was learned? Drawing on both the widespread utility of dictionary learning for denoising and the shared notion of "low description complexity" between latent sparse vectors and our latent inputs z to G, we take a compression perspective on evaluating the learned dictionary. Given two competing models, we assess which model can most accurately reconstruct data from noisy samples given a *fixed bit complexity* for z (or the sparse coefficient vector for baselines), enforced by round-

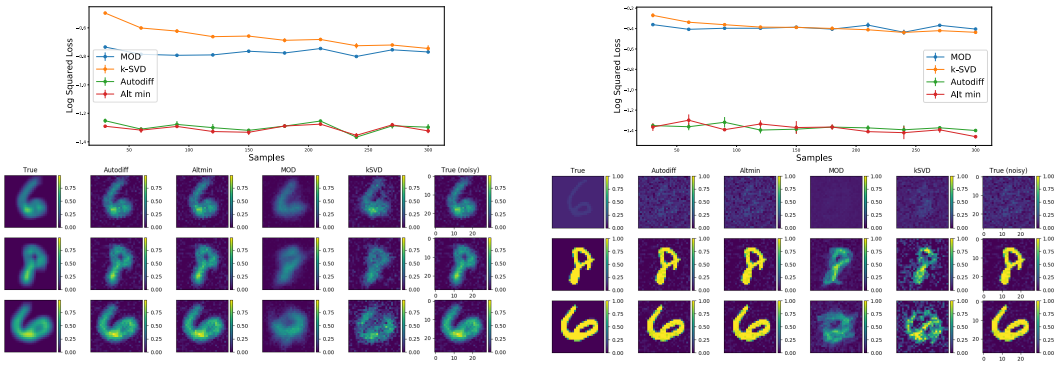

Figure 2: In error plots (first row), error bars show one standard deviation in log-scale over 4 trials, each with 2 batches of test samples and $B = 30$. In reconstruction plots (bottom row), each row is a different sample and each column a different method (in order: noiseless, autodiff, altmin, MOD, kSVD, noisy). The left and right columns are the blurred and recolored datasets, respectively.

ing to 4-digit precision. As such, after learning a dictionary A, we draw a batch of B i.i.d. samples $y^i = A^*G(z^i) + \eta^i$ and compute loss $\frac{1}{B}\sum_{i=0}^{B-1} \frac{||A^*G(z^i) - A^*x^i||^2}{||A^*G(z^i)||^2}$ over n_B test batches, where x^i is either the minimizer of the loss in the range of G or the best k-sparse approximation for baselines.

In all cases, we find that autodiff and altmin strongly outperform sparsity-based dictionary learning, validating that generative priors greatly improve the reconstruction ability of a learned dictionary.

Generative prior capturing dependencies First, in Figure 1a we capture the intuition that G may produce non-sparse outputs, which still contain meaningful correlations among the indices. To this end, we set $G(z) = [z;\ z^{\odot 2};\ \sin(z)]$ with $k = 3$. It is important not to merely concatenate linear functions of z: any linear generative model $G(z) = Mz$ collapses to $y = A^*G(z) = (A^*M)z$ for noiseless data. Such a model would reduce to solving a linear system.

Generative prior capturing group sparsity. Next, in Figure 1b we set $G(z) = \text{ReLU}(M(z^{\odot 2}))$, where $M \in \{0,1\}^{n\times k}$ is entrywise i.i.d. Bern(0.6). We interpret this setting as an approximation of "group sparsity", in which the k free parameters correspond to k groups of motifs, $G_i = \{j\ s.t.\ M_{ji} = 1\}$, such that elements of G_i always appear at the same intensity.

Generative prior derived from an MNIST VAE. Finally, in Figure 2 we begin with a pretrained VAE decoder on MNIST (see Subsection A.1 for details). The decoder architecture is $S(W_2\text{ReLU}(W_1 z))$, where $S(\cdot)$ is the sigmoid function and W_1 and W_2 are matrices. G is obtained by slicing off the last layer of the decoder: $G(z) = \text{ReLU}(W_1 z)$. We then transform a subset of MNIST digits via blurring, colorshift, or flipping. This creates a *transfer learning* task between the original, large and new, small datasets. Given samples y' from a transformed dataset, we thus apply the "transferred" knowledge of G to denoising, where $y = y' + \eta$ and $\eta \sim N(0, 0.0025I)$.

Remark 2 (Unsupervised baselines) *There are no comparisons to unsupervised deep learning baselines, such as retraining the original VAE from scratch, because the sample complexity at which our method succeeds is too low for these methods. Our code includes the option to verify this.*

7 Conclusion

In this work, we propose a new model for dictionary learning that can capture complex co-occurrence relationships between dictionary elements. We prove theoretically that an intuitive alternating minimization algorithm will recover the unknown dictionary at a geometric rate in the noiseless case, and simulated experiments validate the incorporation of generative priors and demonstrate the practical benefits of such a model. Several interesting directions for future work remain. Although we treat $G(z)$ as a coefficient model, problems such as blind deconvolution apply when $G(z)$ is an image prior, and A^* a convolution matrix. One future direction is to modify the algorithm to preserve such known structure in A^*. We also expect our model to prove useful on real datasets, and it would be valuable to test it in non-simulated distributional learning settings.

REFERENCES

Alekh Agarwal, Animashree Anandkumar, Prateek Jain, Praneeth Netrapalli, and Rashish Tandon. Learning sparsely used overcomplete dictionaries. In *Conference on Learning Theory*, pp. 123–137. PMLR, 2014.

M. Aharon, M. Elad, and A. Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. on Signal Processing*, pp. 4311–4322, 2006.

Sanjeev Arora, Rong Ge, Ankur Moitra, and Sushant Sachdeva. Provable ica with unknown gaussian noise, with implications for gaussian mixtures and autoencoders. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper/2012/file/09c6c3783b4a70054da74f2538ed47c6-Paper.pdf.

Sanjeev Arora, Rong Ge, and Ankur Moitra. New algorithms for learning incoherent and overcomplete dictionaries. In *Conference on Learning Theory*, pp. 779–806. PMLR, 2014.

Sanjeev Arora, Rong Ge, Tengyu Ma, and Ankur Moitra. Simple, efficient, and neural algorithms for sparse coding. *Journal of Machine Learning Research*, 40(2015), January 2015. ISSN 1532-4435.

Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G. Dimakis. Compressed sensing using generative models. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 537–546. PMLR, 06–11 Aug 2017. URL http://proceedings.mlr.press/v70/bora17a.html.

Emrah Bostan, Reinhard Heckel, Michael Chen, Michael Kellman, and Laura Waller. Deep phase decoder: self-calibrating phase microscopy with an untrained deep neural network. *Optica*, 7 (6):559, June 2020. ISSN 2334-2536. doi: 10.1364/OPTICA.389314. URL https://www.osapublishing.org/abstract.cfm?URI=optica-7-6-559.

Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8), 2006. ISSN 0010-3640.

Emmanuel Candès and Laurent Demanet. Curvelets and fourier integral operators, 2002.

Emmanuel J. Candès. Ridgelets: Estimating with ridge functions, 2003.

Rahime Ceylan. The effect of feature extraction based on dictionary learning on ecg signal classification. *International Journal of Intelligent Systems and Applications in Engineering*, 6(1):40–46, Mar. 2018. doi: 10.18201/ijisae.2018637929. URL https://www.ijisae.org/IJISAE/article/view/647.

Niladri Chatterji and Peter L Bartlett. Alternating minimization for dictionary learning with random initialization. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3210ddbeaa16948a702b6049b8d9a202-Paper.pdf.

Ingrid Daubechies. Orthonormal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics*, 41(7):909–996, 1988. doi: https://doi.org/10.1002/cpa.3160410705. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.3160410705.

David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

Michael Elad and Michal Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 15(12), Dec 2006.

K. Engan, S. Aase, and J. Hakon-Husoy. method of optimal directions for frame design. In *ICASSP*, pp. 2443–2446, 1999.

Paul Hand and Vladislav Voroninski. Global guarantees for enforcing deep generative priors by empirical risk. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet (eds.), *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 970–978. PMLR, 06–09 Jul 2018. URL http://proceedings.mlr.press/v75/hand18a.html.

Paul Hand, Oscar Leong, and Vlad Voroninski. Phase retrieval under a generative prior. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/1bc2029a8851ad344a8d503930dfd7f7-Paper.pdf.

Reinhard Heckel and Paul Hand. Deep Decoder: Concise Image Representations from Untrained Non-convolutional Networks. *International Conference on Learning Representations*, 2019. URL http://arxiv.org/abs/1810.03982.

Chinmay Hegde, Piotr Indyk, and Ludwig Schmidt. A nearly-linear time framework for graph-structured sparsity. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 928–937, Lille, France, 07–09 Jul 2015. PMLR. URL http://proceedings.mlr.press/v37/hegde15.html.

David Hong, Robert P. Malinas, Jeffrey A. Fessler, and Laura Balzano. Learning dictionary-based unions of subspaces for image denoising. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pp. 1597–1601, 2018. doi: 10.23919/EUSIPCO.2018.8553117.

Gary B. Huang, Marwan Mattar, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, 2007.

Akshay Kamath, Sushrut Karmalkar, and Eric Price. Lower bounds for compressed sensing with generative models. *CoRR*, abs/1912.02938, 2019. URL http://arxiv.org/abs/1912.02938.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. pp. 4311–4322, 2017. URL https://arxiv.org/abs/1710.10196.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Naveen Kodali, James Hays, Jacob D. Abernethy, and Zsolt Kira. On convergence and stability of gans. *arXiv: Artificial Intelligence*, 2018.

Jonas Moritz Kohler and Aurelien Lucchi. Sub-sampled cubic regularization for non-convex optimization. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1895–1904. PMLR, 06–11 Aug 2017. URL http://proceedings.mlr.press/v70/kohler17a.html.

Hannah Lawrence, David A. Barmherzig, Henry Li, Michael Eickenberg, and Marylou Gabrié. Phase retrieval with holography and untrained priors: Tackling the challenges of low-photon nanoscale imaging. In *Mathematical and Scientific Machine Learning*, 2021.

Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. *CoRR*, abs/1609.04802, 2016. URL http://dblp.uni-trier.de/db/journals/corr/corr1609.html#LedigTHCATTWS16.

Victor Lempitsky, Andrea Vedaldi, and Dmitry Ulyanov. Deep Image Prior. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9446–9454. IEEE, jun 2018. ISBN 978-1-5386-6420-9. doi: 10.1109/CVPR.2018.00984. URL https://box.skoltech.ru/index.php/s/ib52BOoV58ztuPM{#}pdfviewerhttps://ieeexplore.ieee.org/document/8579082/.

Shutao Li, Haitao Yin, and Leyuan Fang. Group-sparse representation with dictionary learning for medical image denoising and fusion. *IEEE Transactions on Biomedical Engineering*, 59(12): 3450–3459, 2012. doi: 10.1109/TBME.2012.2217493.

Michael Lustig, David Donoho, and John M. Pauly. Sparse mri: The application of compressed sensing for rapid mr imaging. *Magnetic Resonance in Medicine*, 58(6):1182–1195, 2007. doi: https://doi.org/10.1002/mrm.21391. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.21391.

Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 689–696, New York, NY, USA, 2009a. Association for Computing Machinery. ISBN 9781605585161. doi: 10.1145/1553374.1553463. URL https://doi.org/10.1145/1553374.1553463.

Julien Mairal, Jean Ponce, Guillermo Sapiro, Andrew Zisserman, and Francis Bach. Supervised dictionary learning. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou (eds.), *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2009b. URL https://proceedings.neurips.cc/paper/2008/file/c0f168ce8900fa56e57789e2a2f2c9d0-Paper.pdf.

Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: a strategy employed by v1. *Vision Research*, 37:3311–3325, 1997.

Vardan Papyan, Yaniv Romano, and Michael Elad. Convolutional neural networks analyzed via convolutional sparse coding. *J. Mach. Learn. Res.*, 18(1):2887–2938, January 2017. ISSN 1532-4435.

Ron Rubinstein, Michael Zibulevsky, and Michael Elad. Efficient implementation of the k-svd algorithm using batch orthogonal matching pursuit, 2008.

Karin Schnass. Local identification of overcomplete dictionaries. *J. Mach. Learn. Res.*, 16:1211–1242, 2015.

Daniel A. Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. In Shie Mannor, Nathan Srebro, and Robert C. Williamson (eds.), *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pp. 37.1–37.18, Edinburgh, Scotland, 25–27 Jun 2012. JMLR Workshop and Conference Proceedings. URL http://proceedings.mlr.press/v23/spielman12.html.

Jeremias Sulam, Vardan Papyan, Yaniv Romano, and Michael Elad. Multilayer convolutional sparse modeling: Pursuit and dictionary learning. *IEEE Transactions on Signal Processing*, 66(15): 4090–4104, 2018. doi: 10.1109/TSP.2018.2846226.

Eric W. Tramel, Andre Manoel, Francesco Caltagirone, Marylou Gabrié, and Florent Krzakala. Inferring sparsity: Compressed sensing using generalized restricted Boltzmann machines. In *2016 IEEE Information Theory Workshop (ITW)*, pp. 265–269. IEEE, sep 2016. ISBN 978-1-5090-1090-5. doi: 10.1109/ITW.2016.7606837. URL http://ieeexplore.ieee.org/document/7606837/.

Yong Xu, Zhengming Li, Jian Yang, and David Zhang. A survey of dictionary learning algorithms for face recognition. *IEEE Access*, 5:8502–8514, 2017. doi: 10.1109/ACCESS.2017.2695239.

Raymond A. Yeh, Chen Chen, Teck Yian Lim, Alexander G. Schwing, Mark Hasegawa-Johnson, and Minh N. Do. Semantic image inpainting with deep generative models. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6882–6890, 2017. doi: 10.1109/CVPR.2017.728.

R. Zhao, H. Li, and X. Liu. A survey of dictionary learning in medical image analysis and its application for glaucoma diagnosis. *Arch Computat Methods Eng*, 28:463–471, 2021.