Journal of Machine Learning Research 23 (2022) 1-23 Submitted 12/20; Revised 4/22; Published 5/22

Recovery and Generalization in Over-Realized Dictionary Learning

Jeremias Sulam JSULAM1@JHU.EDU
Department of Biomedical Engineering &
Mathematical Institute for Data Science
Johns Hopkins University
Baltimore, MD 21205, USA

Chong You CYOU@BERKELEY.EDU
Department of Electrical Engineering & Computer Sciences
University of California
Berkeley, CA 94720-1776, USA

Zhihui Zhu ZHIHUI.ZHU@DU.EDU
Department of Electrical & Computer Engineering
University of Denver
Denver, CO 80210, USA

Editor: David Wipf

Abstract

In over two decades of research, the field of dictionary learning has gathered a large collection of successful applications, and theoretical guarantees for model recovery are known only whenever optimization is carried out in the *same* model class as that of the underlying dictionary. This work characterizes the surprising phenomenon that dictionary recovery can be facilitated by searching over the space of larger *over-realized* models. This observation is general and independent of the specific dictionary learning algorithm used. We thoroughly demonstrate this observation in practice and provide an analysis of this phenomenon by tying recovery measures to generalization bounds. In particular, we show that *model recovery* can be upper-bounded by the empirical risk, a model-dependent quantity and the generalization gap, reflecting our empirical findings. We further show that an efficient and provably correct distillation approach can be employed to recover the correct atoms from the over-realized model. As a result, our meta-algorithm provides dictionary estimates with consistently better recovery of the ground-truth model.

Keywords: Dictionary learning, model recovery, sparse models, over-realization, over-parameterization

1. Introduction

Latent variable models have been very successful for a variety of unsupervised learning problems, from regularizing inverse problems of different kinds to enabling clustering, classification, or other down-stream supervised learning problems (Bengio et al., 2013). We focus on sparse representation models, which posit that data $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$ admits a sparse decomposition in terms of a redundant dictionary $\mathbf{D} \in \mathcal{D} \subset \mathbb{R}^{d \times p}$, where $p > d$ and $\mathcal{D}$ is an appropriate constraint set. In other words, $\mathbf{x} = \mathbf{D}\boldsymbol{\gamma}$, where the number of nonzero entries is small: $\|\boldsymbol{\gamma}\|_0 \leq k \ll d$. These models are most useful when the model $\mathbf{D}$ is learned from a collection of samples $\{\mathbf{x}_i\}_{i=1}^n$, thus allowing for greater sparsity or representation power. This task goes by the name of *dictionary learning*, and many algorithms have been proposed over the last two decades to (most often approximately) solve this problem (Aharon et al., 2006a; Mairal et al., 2010; Engan et al., 1999; Olshausen and Field, 1997; Arora et al., 2015).

A central problem in dictionary learning is that of model recovery. More precisely, assuming that the training samples follow such a generative model, $\mathbf{x}_i = \mathbf{D}\boldsymbol{\gamma}_i$, and one has access to a learning algorithm that provides an estimate $\widehat{\mathbf{D}}$, how close will the obtained model be from the true generating dictionary? There exists by now a rich literature on these questions. Some of these results are concerned with providing recovery guarantees for popular and practical dictionary learning methods, such as the K-SVD (Aharon et al., 2006b; Schnass, 2014) or simpler online learning algorithms (Olshausen and Field, 1997; Arora et al., 2015). Others instead propose new algorithms with recovery guarantees, most often in an alternating minimization manner (Agarwal et al., 2016, 2014; Arora et al., 2014a,b; Arora and Risteski, 2017), while other results study local identifiability (Geng and Wright, 2014; Gribonval et al., 2015a) or fundamental limits and min-max optimal bounds (Shakeri et al., 2018; Jung et al., 2016). Naturally, these guarantees depend on the minimum number of training samples, n, as well as on the parameters of the model: d, p and k, the particular distribution of the non-zero values, and possibly the amount of noise contamination in the observations.

Though dictionary learning algorithms vary, by and large they share the following common scheme: given the constraint set $\mathcal{D}_p$ of the ground-truth model, typically $\mathcal{D}_p = \{\mathbf{D} \in \mathbb{R}^{d\times p} : \|\mathbf{D}_i\|_2 = 1,\ \forall i \in \{1,\dots,p\}\}$, and given a collection of n samples from this model, one searches for an estimate $\widehat{\mathbf{D}} \in \mathcal{D}_p$ by means of some optimization approach. The first question we pose in this work is the following: *Why should one limit to the set $\mathcal{D}_p$ instead of searching over a larger class of models?* Somewhat surprisingly, we will show that dictionary recovery can be consistently improved if one allows the learning algorithm to search for models $\widehat{\mathbf{D}} \in \mathcal{D}_{p'} \subset \mathbb{R}^{d\times p'}$, where $p' > p$. In other words, we will search for a larger set of atoms than those that are strictly necessary to sparsely represent the training data—an *over-realized* model.

While it is certainly natural that a larger model of $p' > p$ atoms can approximate the training samples better than one with p atoms, it is not immediately obvious that this might lead to a better overall dictionary recovery. After all, how can one evaluate model recovery if the estimate and ground-truth models belong to different spaces? To this end, we propose a new dissimilarity metric and show that it can be upper-bounded by a function of the empirical risk (i.e. training error) and the generalization gap, both of which are computable. This result links recovery guarantees to generalization bounds, allowing us to characterize the behaviour observed in our experiments, and leading to a uniform upper bound to the recovery error.

Even if one can improve recovery with a larger model, one might be interested in obtaining a dictionary of the original size, i.e. only with p columns. We therefore study a second driving question: *given a trained model $\widehat{\mathbf{D}} \in \mathcal{D}_{p'}$, can one distill from it an estimate $\widetilde{\mathbf{D}} \in \mathcal{D}_p$ and, in doing so, improve the recovery of the true dictionary?* We will answer this question in the affirmative, providing a provably correct algorithm under incoherence assumptions. As a result, we will provide a meta-algorithm for dictionary learning via over-realized models that improves model recovery over conventional (non over-realized) approaches, across a variety of model parameters and learning algorithms.

The study of over-realized models in unsupervised learning has received some—but limited—attention in the past. The work by Dasgupta and Schulman (2007) showed more than a decade ago that the recovery of k clusters by k-means (Lloyd, 1982) can be improved by a two-step process, whereby in the first round one uses more random guesses as initialization (more precisely, $\mathcal{O}(k \log k)$); see also the recent analysis (Qian et al., 2021; Hong et al., 2022). The recent inspiring work by Buhai et al. (2020) is the first to show empirical benefits of over-realized models in representation learning settings. More precisely, the authors demonstrate that over-realization can lead to higher log-likelihood and improved recovery in three different latent variable models and show that this phenomenon is robust, in the sense that it persists across different training algorithms and parameter settings. Buhai et al. (2020) carry out a large empirical study for noisy-OR networks, dictionary learning, and probabilistic context-free grammar models, demonstrating that in all cases the ground-truth model can be better recovered by first searching over a larger model class, followed by an ad-hoc pruning of the latent components. Alas, no analysis is provided in this work.

In the neural networks community, a new and growing body of work has shown that a large number of parameters is key to obtaining good empirical performance (Zhang et al., 2021), bringing forth a surge of interests for providing theoretical support (Goldt et al., 2019; Tian, 2019; Mei and Montanari, 2019; Belkin et al., 2019; Yang et al., 2020). This *over-parameterization* regime refers to models having a larger number of parameters than training samples. In contrast, in this work we study and analyze how *over-realization* (having more parameters than that of the underlying generative model) improves recovery in dictionary learning.

Summary of contributions: In this work we center our study of over-realization in the specific problem of dictionary learning. We provide a notion of dissimilarity that allows for the quantification of the recovery error of a ground-truth dictionary through a larger one—that is, one with more atoms. We do this via a key Lemma linking the recovery error by a measure of the expected risk (see Lemma 3.1), which can in turn be bounded by the empirical risk employing standard generalization bounds (see Theorem 3.2). We then present a distillation procedure that provably recovers correct components (those that are close to the ground-truth p atoms) under incoherence and sparsity assumptions (see Theorem 4.1). Throughout the presentation of these results, we numerically illustrate the benefits obtained through over-realized dictionaries, as well as via our distillation algorithm.

Overview: We first introduce our notation and provide the necessary background in Section 2. We then address the recovery problem in the over-realized case in Section 3, providing examples and presenting our main theoretical result. Section 4 tackles the question of the distillation of larger models, and provides a provably correct algorithm as well as extensive empirical evidence. We finally delineate final remarks and conclude in Section 5.

2. Preliminaries

We consider data $\mathbf{x} \in \mathbb{R}^d$, and a redundant dictionary $\mathbf{D}_0 \in \mathcal{D}_p$, $p > d$. We consider the following generative model for $\mathbf{x}$ throughout this work, providing a sampling distribution $\mathbb{P}$: a sparse representation $\boldsymbol{\gamma} \in \mathbb{R}^p$ is sampled from a set of k-sparse vectors by (i) sampling its support S uniformly from the set of all possible $\binom{p}{k}$ supports of cardinality k, and (ii) sampling its non-zero values *i.i.d.* from a distribution with mean zero and unit variance (for simplicity). Samples are then obtained as $\mathbf{x} = \mathbf{D}_0\boldsymbol{\gamma}$. Given $\mathbf{x}$ and $\mathbf{D}_0$, the problem of retrieving the representation $\boldsymbol{\gamma}$ is termed *sparse coding*, and it involves solving a problem of the form

$$\min_{\boldsymbol{\gamma}} \frac{1}{2}\|\mathbf{x} - \mathbf{D}_0\boldsymbol{\gamma}\|_2^2 + g(\boldsymbol{\gamma}), \tag{1}$$

where $g(\boldsymbol{\gamma})$ is a sparsity-promoting function that regularizes the ill-posed recovery problem. Typical choices for g are the non-convex and non-smooth ℓ_0 pseudo-norm, or its convex relaxation, the ℓ_1 norm. Alternatively, g may denote an indicator function over a constraint set, such as

$$g_k(\boldsymbol{\gamma}) = \begin{cases} 0 & \text{if } \|\boldsymbol{\gamma}\|_0 \leq k, \\ +\infty & \text{otherwise.} \end{cases} \tag{2}$$

In either case, numerous pursuit algorithms exist that allow for the provable recovery of $\boldsymbol{\gamma}$ under assumptions like restricted isometry property (Candes and Tao, 2005) or incoherence (Tropp, 2004; Donoho and Elad, 2003). These exact recovery guarantees are naturally extended to approximate recovery in the case of noisy measurements. When $g(\boldsymbol{\gamma}) = \|\boldsymbol{\gamma}\|_1$, the problem is termed Basis Pursuit DeNoising or Lasso (Tibshirani, 1996) (and Basis Pursuit when an ℓ_1 ball is used as a constraint set). Alternatively, one may employ greedy algorithms such as the popular Orthogonal Matching Pursuit (OMP) (Pati et al., 1993), which approximates the solution to the ℓ_0-constrained problem.

When the dictionary is not known, the dictionary learning problem attempts to recover an estimate as close as possible to the ground-truth model given a set of n training samples $\mathbf{x}_i$ from it.

The quality of a dictionary in approximating a sample $\mathbf{x}$ is measured by the function value of the cost above, namely

$$f_{\mathbf{x}}(\mathbf{D}) \coloneqq \inf_{\boldsymbol{\gamma}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{x}-\mathbf{D}\boldsymbol{\gamma}\|_2^2 + g(\boldsymbol{\gamma}). \tag{3}$$

In this way, the dictionary learning problem minimizes this loss over the n samples, and can be written as

$$\min_{\mathbf{D}\in\mathcal{D}_p} \frac{1}{n}\sum_{i=1}^{n} f_{\mathbf{x}_i}(\mathbf{D}). \tag{4}$$

The resulting optimization problem is non-convex and hard to analyze in general (Tillmann, 2014), but this has not prevented the development of many—and very successful—algorithms. One such methods is the Online Dictionary Learning (ODL) from Mairal et al. (2010), which minimizes Equation (4) in an online manner. In a nutshell, given a current estimate for the dictionary, this algorithm iterates between drawing a sample (or a mini-batch thereof) at random, then employing a pursuit algorithm to minimize Equation (1), and finally updating the dictionary so as to minimize a surrogate of the cost in Equation (4). The approach is general in that it can accommodate different pursuit algorithms for different penalty functions $g(\boldsymbol{\gamma})$, and it scales well to large data sets. The very popular K-SVD (Aharon et al., 2006a), on the other hand, is a batch-learning approach that alternates between sparse coding (typically with OMP) and dictionary update, which is characteristically carried out column-by-column by performing rank-1 approximations to atom-wise residuals.

2.1 Recovery

A central question in this setting is that of model recovery, which studies how far the recovered estimate $\widehat{\mathbf{D}} \in \mathcal{D}_p$ is from the ground-truth dictionary, $\mathbf{D}_0 \in \mathcal{D}_p$. To formalize this question one needs an appropriate measure of dissimilarity between matrices. The problem in Equation (4) is permutation (and sign) invariant: the columns of the dictionary can be arbitrarily permuted (or multiplied by -1) without modifying the cost $f_{\mathbf{x}}(\mathbf{D})$. Thus, different measures of recovery have been used in previous works accounting for such invariance, such as (Arora et al., 2015)

$$\min_{P\in\Pi} \|\mathbf{D}_0 - \widehat{\mathbf{D}}P\|_F^2, \tag{5}$$

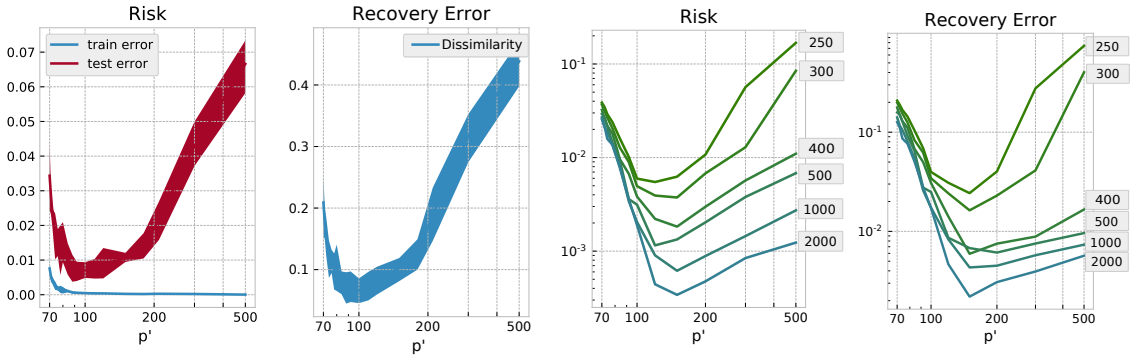
where Π is the set of signed permutation matrices, i.e. orthogonal matrices that contain only $\{0, \pm 1\}$. Several works have addressed these questions of recovery over the last decade. Some of these show local linear convergence to the global optimum (i.e. the true model) via alternating minimization employing ℓ_1 penalty functions (Agarwal et al., 2014, 2016) or to an ϵ-close optimum via ℓ_0 constraints (Arora et al., 2015). In the simpler case of orthonormal dictionaries the optimization landscape is better understood (Zhai et al., 2020), as in the case of learning only one atom (Sun et al., 2015; Qu et al., 2019). In these settings, these non-convex problems have a benign geometry structure that allows for provable algorithms. On the other hand, Jung et al. (2016) develops minimax risk bounds for dictionary recovery, and Shakeri et al. (2018) studies these as a function of their tensor structure. All of these results, however, analyze the conventional setting whereby the constraint sets of the ground-truth dictionary and the one enforced during optimization are the same.

2.2 Generalization gap

From a statistical learning standpoint, the dictionary learning problem consists in finding a model $\widehat{\mathbf{D}} \in \mathcal{D}_p$ that minimizes the above function in expectation over the population, i.e.,

$$\widehat{\mathbf{D}} \in \operatorname*{argmin}_{\mathbf{D}\in\mathcal{D}_p} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}} \left[f_{\mathbf{x}}(\mathbf{D})\right]. \tag{6}$$

Since one does not typically have access to the underlying distribution, the empirical risk minimization algorithm (ERM) minimizes the empirical estimate of the above risk, which is precisely the problem



(a) (b) (c) (d)

Figure 1: (a) and (b): Risk of the estimated dictionary and dissimilarity with the ground truth model, as defined in Equation (8), trained with 300 samples. (c) and (d): Risk (test error) and recovery error for different size of the training data (as indicated by the numbers next to each line). The dictionary size p' refers to that of the estimated matrix, whereas the original one remains fixed containing $p = 70$ atoms.

in Equation (4). In this context, a central question is given by the generalization gap, which quantifies the extent to which the empirical error, $\mathcal{R}_S(\mathbf{D}) = \frac{1}{n}\sum_{i=1}^{n} f_{\mathbf{x}_i}(\mathbf{D})$, differs from its expectation in Equation (6), termed generalization error, or risk. Uniform bounds have recently been developed for these models (Maurer and Pontil, 2010; Vainsencher et al., 2011; Seibert, 2019). More specifically, the work by Gribonval et al. (2015b) shows that, with overwhelming probability over the draw of the samples, this gap is uniformly bounded,

$$\sup_{\mathbf{D}\in\mathcal{D}_p} \left| \mathcal{R}_S(\mathbf{D}) - \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}}[f_{\mathbf{x}}(\mathbf{D})] \right| \leq \eta_n, \tag{7}$$

where η_n, depends on the model capacity, the number of samples, as well as the data distribution and properties of the penalty function g. Slightly more specifically, η_n is $\mathcal{O}(\sqrt{(dp)\log n/n})$, where (dp) is the number of parameters in the dictionary with p atoms. This type of bounds are very useful, since they provide an upper bound to the expected (real) risk given the empirical risk, and they reflect the natural trade-off between the model size (number of atoms, p) and the number of training samples, n. The bound above holds not just for norms and *norm-like* regularization functions (like the ℓ_1 norm) but also for indicator sets as g_k in Equation (2). We will keep our derivations maximally general by simply referring to η_n, and we refer the reader to (Gribonval et al., 2015b) for further details on the involved constants.

3. Searching for over-realized dictionaries

In this work we focus on the over-realized setting, in which the minimization in Equation (4) is done over a class of dictionaries $\mathcal{D}_{p'}$, with $p' > p$, i.e. larger than the original model. One might wonder as to the need for this change. After all, there exists indeed a global minimum ($\mathbf{D}_0$) with p atoms that achieves both zero training and testing errors. Nonetheless, the problem in Equation (4) is non-convex, and practical alternating minimization and local-search algorithms may converge to only local minima. In many settings, however, non-convex optimization problems have been shown to become easier in over-parametrized settings, in the sense that spurious local minima decrease and local algorithms are more likely to converge to the global minimum (Safran and Shamir, 2018; Buhai et al., 2020). As a result, it may be possible to obtain better dictionaries by searching over a larger

class of matrices. As we will show, this is not only true in terms of their risk, but also with respect to their dissimilarity to the true generating model.

We first require a dissimilarity measure between dictionaries of potentially different sizes.[1] We will use the following definition for a *dissimilarity* between a dictionary $\mathbf{D}_0 \in \mathcal{D}_p$ and an estimate $\widehat{\mathbf{D}} \in \mathcal{D}_{p'}$:

$$d(\mathbf{D}_0, \widehat{\mathbf{D}}) \coloneqq \frac{1}{p} \sum_{i=1}^{p} \min_{j \in [p']} \min_{c \in \{-1,1\}} \|\mathbf{D}_i^0 - c\, \widehat{\mathbf{D}}_j\|_2^2. \tag{8}$$

Note that this quantity is zero if and only if there exists a match for each of the atoms in $\mathbf{D}_0$ in the estimated $\widehat{\mathbf{D}}$, irrespective the size p'. Moreover, this expression provides a generalization of the commonly used distance measure in Equation (5).[2] Lastly, note that this definition does not require the minimizer over $j \in [p']$ to be unique. On the other hand, Equation (8) does allow for an atom in $\widehat{\mathbf{D}}$ to be chosen as the closest neighbor for two different atoms in $\mathbf{D}_0$. This, however, would only occur in cases where $\mathbf{D}_0$ is very coherent.

We now explore the first question posed above, namely: can one obtain an estimate with better generalization error *and* lower recovery error by searching in a hypothesis class bigger than that of the original dictionary? As a motivating example, we construct the following experimental setting. Data is sampled as described in the previous section from a ground-truth dictionary (with normalized Gaussian atoms) of size 50×70, from representations with cardinality $k = 3$. We construct 300 such samples for training, leaving 1000 to estimate the population statistics. As a learning algorithm, we employ ODL (Mairal et al., 2010) for 2000 iterations, which are more than sufficient for convergence.[3] We employ OMP for the sparse coding step.

In Figure 1a we depict the risk, or error, on both training and test sets, as a function of the number of atoms in the estimated dictionary $\widehat{\mathbf{D}}$, from 70 (the size of the ground-truth model) to 500. We repeat the experiment 20 times, and present the mean together with the 25% and 75% percentiles. Interestingly, both train and testing errors, shown in Figure 1a, *improve* with increasing dictionary size $p' > p$ within some range. More surprisingly, the dissimilarity to the estimate to the ground truth $\mathbf{D}_0$ also improves as one searches for bigger dictionaries. Note that because of our definition of dissimilarity in Equation (8), a small dissimilarity implies a close recovery of the true atoms, irrespective of the "extra" ones. At the same time, this behaviour is tightly related to that of model capacity and over-fitting: while increased dictionary size allows for better recovery, the finite training size eventually becomes insufficient to train the larger model and the generalization error increases (while perfectly fitting the training data). This is verified in Figure 1c and Figure 1d, seeing that the generalization error—*and dictionary recovery*—is precisely controlled by the size of the training set. In this figure, only the means of the 20 realizations are depicted for the sake of clarity. For completeness, in Appendix C.1 we present analogous results to those in Figures 1a and 1b, but reporting the best run out of 30 random initializations.

3.1 Recovery guarantees via generalization bounds

While the behaviour observed in Figure 1a and Figure 1c is well understood in the statistical learning literature, this is still surprising in light of the fact that *there exist* a ground truth model with just p atoms that achieves zero risk. Moreover, how this relates to improved recovery of the ground-truth dictionary in over-realized settings—as shown in Figure 1b and Figure 1d—is, to the best of our knowledge, unknown. Learning bounds and recovery guarantees for dictionary learning have so far

1. We will use $\mathbf{D}_i^0$ to denote the i^{th} column, or atom, from $\mathbf{D}_0$.
2. Note that our definition in Equation (8) generalizes that in Equation (5) by allowing the set of permutation matrices to become column-selection (non-square) ones.
3. Available at `spams-devel.gforge.inria.fr/`. Note that ODL can accommodate different formulations and algorithms for the sparse coding step (and not just an ℓ_1 minimization), which will enable us to explore different experimental settings.

remained mostly separated. We will now precisely connect the model recovery error with its expected risk, providing a theoretical characterization for this phenomenon.

Let $f_{\mathbf{x}}^{[s]}(\widehat{\mathbf{D}}) = \inf_{\boldsymbol{\gamma}:\|\boldsymbol{\gamma}\|_0 \leq s} \frac{1}{2}\|\mathbf{x} - \widehat{\mathbf{D}}\boldsymbol{\gamma}\|_2^2$ denote the loss measured with s non-zero coefficients. We will denote the mutual coherence of a dictionary by $\mu(\mathbf{D}) = \max_{i \neq j} |\langle \mathbf{D}_i, \mathbf{D}_j \rangle|$ (recall that columns are normalized). Furthermore, for a given atom in the estimate dictionary, $\widehat{\mathbf{D}}_j$, consider its closest atom in the ground truth dictionary, $\mathbf{D}^0_{i_{(j)}}$, where $i_{(j)} = \operatorname{argmin}_{i \in [p]} \min_{c \in \{-1,1\}} \|\mathbf{D}_i^0 - c\widehat{\mathbf{D}}_j\|_2$. We will also need a cross-dictionary coherence, defined as

$$\nu(\widehat{\mathbf{D}}, \mathbf{D}_0) = \max_j \max_{k \neq i_{(j)}} \left| \langle \widehat{\mathbf{D}}_j, \mathbf{D}_k^0 \rangle \right|.$$

In words, $\nu(\widehat{\mathbf{D}}, \mathbf{D}_0)$ quantifies the coherence between $\widehat{\mathbf{D}}$ and $\mathbf{D}_0$ after excluding the closest neighbor of each atom.[4] While this expression might seem somewhat convoluted, this simply reduces to the traditional mutual coherence of the dictionary, $\mu(\mathbf{D}_0)$, in the case that $\widehat{\mathbf{D}} = \mathbf{D}_0$. With these definitions, we have the following central Lemma.

Lemma 3.1. *For a ground-truth dictionary $\mathbf{D}_0 \in \mathcal{D}_p$ generating samples $\mathbf{x}_i = \mathbf{D}_0\boldsymbol{\gamma}_i$, where $\boldsymbol{\gamma}_i$ are k-sparse with non-zero entries sampled i.i.d. from a zero mean and unit variance distribution, and for any dictionary $\widehat{\mathbf{D}} \in \mathcal{D}_{p'}$, we have that*

$$\frac{2}{k} \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathbb{P}} [f_{\mathbf{x}}^{[k]}(\widehat{\mathbf{D}})] \;\leq\; d(\mathbf{D}_0, \widehat{\mathbf{D}}) \;\leq\; \frac{4}{k} \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathbb{P}} [f_{\mathbf{x}}^{[1]}(\widehat{\mathbf{D}})] - \frac{4}{k}\zeta_k(k-1). \tag{9}$$

where $\zeta_k = \max\left\{0 \;,\; 1 - (k-2)\mu(\mathbf{D}_0) - 2\nu(\widehat{\mathbf{D}}, \mathbf{D}_0)^2\right\}$.

Note that this result links the dissimilarity, $d(\mathbf{D}_0, \widehat{\mathbf{D}})$, with the expected risk, as measured by $f_{\mathbf{x}}^{[1]}(\widehat{\mathbf{D}})$ and $f_{\mathbf{x}}^{[k]}(\widehat{\mathbf{D}})$. We will comment on further implications of this shortly, but first we present our main result as a consequence of Lemma 3.1, which is of practical relevance. Employing the generalization bound from Equation (7), we can bound the dictionary dissimilarity by informative quantities, as presented in the following main result.

Theorem 3.2. *For a ground-truth dictionary $\mathbf{D}_0 \in \mathcal{D}_p$ generating samples $\mathbf{x}_i$ with sparsity of k, and for any estimate $\widehat{\mathbf{D}} \in \mathcal{D}_{p'}$, with overwhelming probability, we have that*

$$\frac{k}{4} d(\mathbf{D}_0, \widehat{\mathbf{D}}) \leq \frac{1}{n} \sum_i^n f_{\mathbf{x}_i}^{[1]}(\widehat{\mathbf{D}}) - \zeta_k(k-1) + \mathcal{O}\left(\sqrt{\frac{dp' \log(n)}{n}}\right). \tag{10}$$

First, this result shows that the dissimilarity to the true model can be upper-bounded by the empirical risk up to the generalization gap and a model-dependent quantity. This reflects an important implicit trade-off: dictionary recovery can be decreased by increasing the model capacity (dictionary size) as long as the generalization gap is kept small by increasing the sample size appropriately. This is precisely the behaviour observed in Figure 1d above. Second, the term $\zeta_k(k-1)$ appearing in both results above accounts for the fact that the upper bound is constructed via $f_{\mathbf{x}_i}^{[1]}$, as opposed to $f_{\mathbf{x}_i}^{[k]}$. Indeed, note that this term vanishes when $k = 1$. When $k > 1$, the empirical estimate of $f_{\mathbf{x}_i}^{[1]}$ will necessarily be greater than zero. It is in these cases where the term $\zeta_k(k-1)$ provides a non-trivial tighter bound, as long as $k \leq 2 + 1/\mu(\mathbf{D}_0) - 2\nu(\widehat{\mathbf{D}}, \mathbf{D}_0)^2/\mu(\mathbf{D}_0)$, which are mild conditions. Moreover, whenever $k > 1$ and $\mu(\mathbf{D}_0) \approx \nu(\widehat{\mathbf{D}}, \mathbf{D}_0) \approx 0$, then $\zeta_k \approx 1$. This represents the best case scenario as ζ_k will thus decrease the upper bound the most. As the dictionaries become more coherent, the extent to which ζ_k improves the bound decreases. However, this result holds for any sparsity level of k.

4. Note that this quantity is still lower than 1 even if several atoms in $\widehat{\mathbf{D}}$ are close to the same atom in $\mathbf{D}_0$.

It is natural to inquire how large the gap between $f_{\mathbf{x}_i}^{[k]}(\widehat{\mathbf{D}})$ and $f_{\mathbf{x}_i}^{[1]}(\widehat{\mathbf{D}})$ can be. A lower bound on the difference between the expected values of these quantities can be readily derived from our Lemma 3.1. From this, one can obtain that $2\mathbb{E}\left[f_{\mathbf{x}}^{[1]}(\widehat{\mathbf{D}})\right] - \mathbb{E}\left[f_{\mathbf{x}}^{[k]}(\widehat{\mathbf{D}})\right] \geq \zeta_k(k-1)$. Based on the discussion above, the lower bound between these terms is controlled by the dictionary coherence of $\mathbf{D}_0$, as well as the cross-dictionary coherence. An upper bound can also be derived, and we show in Appendix A.1 that under mild conditions,

$$f_{\mathbf{x}}^{[1]}(\widehat{\mathbf{D}}) - f_{\mathbf{x}}^{[k]}(\widehat{\mathbf{D}}) \lesssim (k-1)\|\mathbf{x}\|_2\|\hat{\boldsymbol{\gamma}}\|_\infty, \tag{11}$$

where $\hat{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\gamma}} f_{\mathbf{x}}^{[k]}(\widehat{\mathbf{D}})$. This is tight when $k = 1$.

While we defer the proof of Lemma 3.1 to Appendix A, let us provide a brief proof sketch. The upper and lower bound for $d(\mathbf{D}_0, \widehat{\mathbf{D}})$ are obtained independently, though with similar techniques. For the upper bound, we make the observation that the risk $\mathbb{E}[f_{\mathbf{x}}^{[1]}(\widehat{\mathbf{D}})]$ can be expressed analytically in closed form, and can be further decomposed in three terms. Relying on the fact that the non-zero entries are drawn i.i.d. with mean zero and unit variance, the expectation one of these vanishes. Another term can be lower-bounded by $\zeta_k(k-1)$, while the remaining term can be lower bounded by a quantity that is proportional to the dictionary dissimilarity $d(\mathbf{D}_0, \widehat{\mathbf{D}})$. The lower bound, on the other hand, is obtained by constructing an analytical (and potentially sub-optimal) solution for the sparse coding problem represented by $f_{\mathbf{x}}^{[k]}(\widehat{\mathbf{D}})$ relying on the atoms that are closest to $\mathbf{D}_0$, thus upper bounding this risk. A series of algebraic manipulations and the final evaluation of the expectation provide the final upper bound on $\mathbb{E}[f_{\mathbf{x}}^{[k]}(\widehat{\mathbf{D}})]$ as a function of the dissimilarity $d(\mathbf{D}_0, \widehat{\mathbf{D}})$.

As we see, these results provide an answer in support of learning larger dictionaries, not only to minimize the expected risk but also to obtain estimates with small dissimilarity to the ground-truth model. The reader should note that this result does not explain *why* this improvement is achieved by common dictionary learning methods. In other words, our analysis does not address the question of when obtaining good minimizers for the empirical risk is possible. Instead, we have shown that *when* those good minimizers are obtained, good recovery is possible if the generalization gap is controlled appropriately. Moreover, a different question also remains: can one *distill* the estimated over-realized $\widehat{\mathbf{D}}$ to recover the best p atoms that are the closest to the real model? This is the question we address in the next section.

4. Distilling the over-realized model

In this section, we will first show that the recovered atoms in the over-realized dictionary exhibit two distinct behaviors: any recovered atom is either (very) close to a true atom in $\mathbf{D}_0$, or is significantly far apart from all atoms in $\mathbf{D}_0$. We will also show that this clustering behaviour correlates with the atom usage in the estimated model. From this observation, we will then derive a provably correct pruning strategy based on the atom's usage frequency. This distillation approach will recover an estimate, $\widetilde{\mathbf{D}} \in \mathcal{D}_p$, of the original size with a lower recovery error than the traditional (non over-realized) learning approach.

As before, given 500 training samples created as the linear combination of $k = 3$ atoms from a ground-truth dictionary $\mathbf{D}_0$ with 70 atoms in 50 dimensions, we train an over-realized dictionary $\widehat{\mathbf{D}}$ with 90 atoms using ODL (with OMP for sparse coding). We then measure, per estimated atom $\widehat{\mathbf{D}}_j$, the similarity to its closest neighbor in the ground-truth $\mathbf{D}_0$ (computed as $-\log\|\widehat{\mathbf{D}}_j - \mathbf{D}_{i_{(j)}}^0\|_2^2$). We plot these similarities as a function of the atom's *usage*: the relative number of times it is used by the training samples upon completion of training. The results are depicted in Figure 2a, and two observations are worth noting: the recovered atoms *either* have a high similarity with those in the ground-truth dictionary *or* are markedly distinct, with a clear separation between groups. This is similar to the observation made in (Buhai et al., 2020) in the context of noisy-or networks and approximate sparse coding. Second, there exists a strong correlation between the

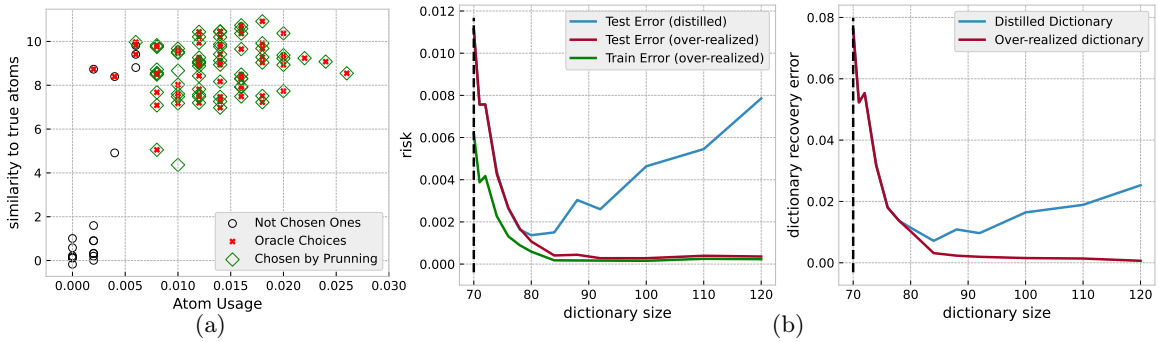


Figure 2: (a) Atoms in the over-realized $\widehat{\mathbf{D}}$: their similarity to their closest atom in the ground-truth dictionary $\mathbf{D}_0$ and its usage frequency. (b) Risk and dissimilarity to the ground truth model by the over-realized dictionary (i.e. with $p' > p$) and by the distilled version, of the same size as the original model ($p' = p$).

Algorithm 1: Meta-algorithm for over-realized dictionary learning

Data: Set of n training samples, $\mathbf{X} \in \mathbb{R}^{d \times n}$; dictionary size p, and sparse coding parameters (cardinality or penalty) λ;

Initialization: Choose $p' > p$ and dictionary learning method `LearnD`(p,λ);

Dictionary Learning: $\widehat{\mathbf{D}} \leftarrow$`LearnD`($p'$,$\lambda$) ;

Distillation: $\widetilde{\mathbf{D}} \leftarrow$ {top p-used atoms in $\widehat{\mathbf{D}}$};

Result: Estimated dictionary $\widetilde{\mathbf{D}}$

former measure—which cannot be computed in practice, without the original model—and the number of times an estimated atom is used by the training samples—which can.

Following this observation, we then propose the following simple meta-algorithm, summarized in Algorithm 1: after learning an over-realized dictionary, we keep the p most frequently used atoms by the training samples. Other works have suggested similar approaches that prune the over-realized model to a subset of components and then continue the optimization with these as better initializations (Dasgupta and Schulman, 2007). This is not needed in our setting, however, likely due to the significantly more accurate coding step. Our proposed distillation approach is also similar to the heuristic proposed by Buhai et al. (2020), which simply discards all atoms that were not used after a step of sparse coding. In light of this, ours can be thought of a refinement of the same idea. Figure 2b illustrates the same experiment as that in Figure 1a and Figure 1b, though now with the statistics provided by our distillation strategy. While clearly the distillation procedure introduces some errors, it still provides a considerable advantage over the traditional approach (i.e. training with the original size p) by significantly diminishing the recovery error. This is further explained by the details in Figure 2a, comparing the atoms chosen by this distillation procedure and the *oracle* choices—those atoms that are the closest to the ground-truth dictionary. As can be seen, most atoms selected by this strategy coincide with the oracle ones.

4.1 Theoretical guarantees for distillation

We now strengthen our argument for our distillation strategy. In the following result, we show that if the atom usage of the over-realized estimate $\widehat{\mathbf{D}}$ is measured via OMP (with $k = 1$), and $\widehat{\mathbf{D}}$ contains at least p atoms that are ϵ-close to the real ones (plus others that are not), then OMP is guaranteed to select the correct (i.e. closest) ones, thus retaining them in the pruning stage.

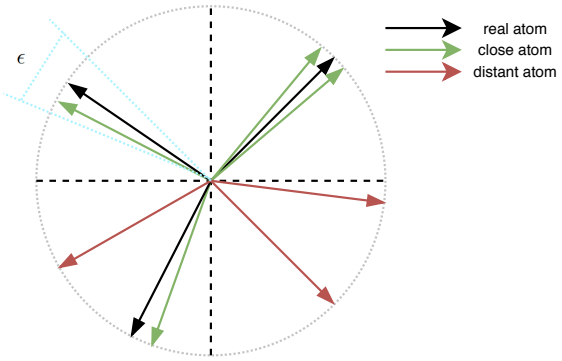


Figure 3: Illustration of true and estimated atoms.

Let $\mathbf{D}_0 \in \mathbb{R}^{d\times p}$ and consider, without loss of generality, that $\widehat{\mathbf{D}} = [\widehat{\mathbf{D}}_0, \mathbf{A}] \in \mathbb{R}^{d\times p'}$, $p' > p$, with $\widehat{\mathbf{D}}_0 \in \mathbb{R}^{d\times m}$, with $p \leq m \leq p'$, such that $d(\widehat{\mathbf{D}}_i^0, \mathbf{D}_0) \leq \epsilon$ for all $i \in \{1, \dots, m\}$, and $d(\mathbf{A}_j, \mathbf{D}_0) > \epsilon$ for all $j \in \{1, \dots, p' - m\}$. In other words, $\widehat{\mathbf{D}}_0$ contains all those m atoms that are ϵ-close to those in $\mathbf{D}_0$, while $\mathbf{A}$ contains those that are further away. Additionally, we require that each atom in $\mathbf{D}_0$ has at least one ϵ-neighbor in $\widehat{\mathbf{D}}_0$; i.e. $d(\mathbf{D}_i^0, \widehat{\mathbf{D}}_0) \leq \epsilon$ for all $i \in \{1, \dots, p\}$. We allow $m \geq p$ since the over-realized estimate $\widehat{\mathbf{D}}$ may naturally contain several atoms that close to a real one. Also suppose that both $\mathbf{D}_0$ and $\widehat{\mathbf{D}}$ are column-wise normalized for simplicity. These assumptions, which reflect the behavior depicted in Figure 2a, are illustrated in Figure 3 below. Lastly, let us denote by $\mu(\mathbf{D}_0, \mathbf{A}) = \max_{i,j} \left|\langle \mathbf{D}_i^0, \mathbf{A}_j \rangle\right|$ the mutual coherence between $\mathbf{D}_0$ and $\mathbf{A}$.

With these definitions, we have the following result, which we prove in Appendix B.

Theorem 4.1. *Let* $\mathbf{x}$ *be a* k*-sparse signal under* $\mathbf{D}_0$*, i.e., there exists* $\boldsymbol{\gamma} \in \mathbb{R}^p$ *with* $\|\boldsymbol{\gamma}\|_0 \leq k$ *such that* $\mathbf{x} = \mathbf{D}_0\boldsymbol{\gamma}$*, and let* $\widehat{\mathbf{D}}$ *be defined as above. Then,* $\operatorname{argmax}_i |\mathbf{x}^T \widehat{\mathbf{D}}_i| \in [m]$ *as long as*

$$k \leq \frac{1 - \frac{\epsilon}{2} + \sqrt{\epsilon} + \mu(\mathbf{D}_0)}{\mu(\mathbf{D}_0) + \sqrt{\epsilon} + \mu(\mathbf{D}_0, \mathbf{A})}.$$

Note that, on one hand, if the dissimilarity $\epsilon = 0$ and we replace $\mu(\mathbf{D}_0, \mathbf{A})$ with $\mu(\mathbf{D}_0)$, our condition can be compared to the traditional incoherence condition for OMP that requires $k < \frac{1}{2}(1 + \frac{1}{\mu(\mathbf{D}_0)})$. As shown by the results in Figure 2a, we indeed observe that the similarity in the un-related atoms to those in $\mathbf{D}_0$ is quite low, i.e. $\mu(\mathbf{D}_0, \mathbf{A})$ is very small. Then, in this case (with $\epsilon = 0$) our condition is milder than the one for OMP, leading to relaxed and improved guarantees. This is natural, since we must only select atoms in $\widehat{\mathbf{D}}$ that belong to $\widehat{\mathbf{D}}_0$—as opposed to demanding the recovery of *the* correct atoms within it. On the other hand, $\mathbf{A}$ itself is allowed to be very coherent, as our condition only requires $\mu(\mathbf{D}_0, \mathbf{A})$ to be small. Lastly, the result above is more general in that we allow for $\epsilon > 0$, which better reflects the empirical behavior depicted in Figure 2a.

4.2 Generalization to different model parameters and algorithms

Thus far we have employed the same experimental setting (dimension, dictionary size and sparsity) for all the above examples for simplicity. However, the reported findings are general and hold for a variety of parameters and algorithms. We now demonstrate this in Figure 4 where we report the risk and dictionary error for the estimates produced by learning a dictionary (with ODL+OMP) in the traditional setting (i.e., $\widehat{\mathbf{D}} \in \mathcal{D}_p$) and that produced by searching over a larger set (i.e., $\widehat{\mathbf{D}} \in \mathcal{D}_{p'}$, with $p' > p$) followed by our distillation strategy. In this way, all reported measures are computed on estimates of the same size as the original model. Note that an important improvement in risk, but most importantly in dictionary recovery, is observed across a wide range of parameters. Moreover, the

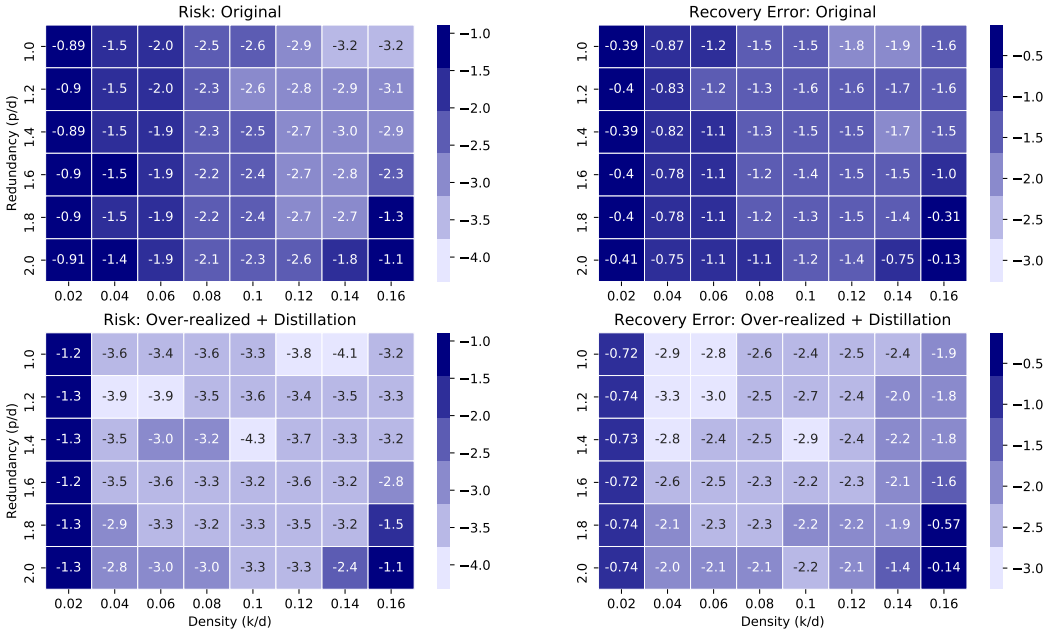
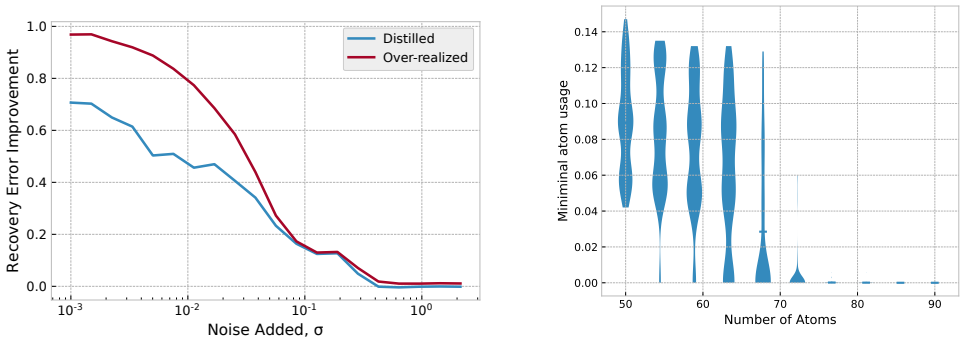


Figure 4: Risk and dictionary recovery error ($\log_{10}$ thereof, lower is better) of the estimate provided by *traditional* dictionary learning (i.e. $\widehat{\mathbf{D}} \in \mathcal{D}_p$) and that resulting from the proposed over-realized approach (i.e. $\widehat{\mathbf{D}} \in \mathcal{D}_{p'}$) followed by distillation to the original size, over a number of parameters (sparsity, dimension, and redundancy).

phenomenon is general across different model parameters and across different learning algorithms and regularization functions g. In Appendix C.2 we show that similar behaviour (albeit less pronounced) can be obtained by employing: (i) the ODL method from (Mairal et al., 2010) with an ℓ_1 regularizer, i.e. employing Lasso for sparse coding, and (ii) the batch algorithm K-SVD (Aharon et al., 2006a).

On a different note, we have considered the noiseless setting throughout; i.e. each sample $\mathbf{x}$ can be exactly expressed as $\mathbf{x} = \mathbf{D}_0\boldsymbol{\gamma}$. In more realistic cases, samples contain measurement noise, or model deviations, which can be modelled by assuming that $\mathbf{x} = \mathbf{D}_0\boldsymbol{\gamma} + \mathbf{v}$, where $\mathbf{v}$ is a nuisance vector. While the thorough study of this setting is out of the scope of this work, we will now show empirically that the benefit of over-realization is robust to noise contamination. To this end, and in a similar manner to the above experiments, we contaminate the samples with noise $\mathbf{v}$ sampled from a Gaussian distribution with covariance $\sigma^2\mathbf{I}$. We then measure the risk and recovery error achieved by traditional dictionary learning (i.e., $p' = p$), by the over-realization approach and by our proposed distillation procedure. These quantities are reported in Figure 5a as relative (normalized) improvement over the traditional setting.[5] As one can see, the benefits of searching over larger model deteriorates smoothly with increasing noise, both for the general over-realized model as well as for our practical distillation approach.

5. More precisely, the quantity measured is $\left(d(\mathbf{D}_0, \widehat{\mathbf{D}}_p) - d(\mathbf{D}_0, \widehat{\mathbf{D}}_{p'})\right) / d(\mathbf{D}_0, \widehat{\mathbf{D}}_p)$ where $\widehat{\mathbf{D}}_p$ is the estimate found by traditional dictionary learning ($p = p'$) and $\widehat{\mathbf{D}}_{p'}$ denotes the estimate found in the over-realized setting, where $p' > p$. The improvement for the distilled version of the estimate is computed analogously.



(a) (b)

Figure 5: (a) Improvement in dictionary recovery relative to that of traditional dictionary learning for the over-realized case and our distillation procedure, as a function of added Gaussian noise. (b) Minimal number of atom usage in the estimated dictionary as a function of the model size, for different dimensions and sparsity levels, and a ground-truth model size of 70.

5. Final Remarks

We have shown that learning over-realized dictionaries can be beneficial not just to provide lower training and population risk, but to also improve the recovery of the underlying model. Our characterization of this phenomenon relies on the connection between the recovery error and the expected risk, thus providing an upper-bound to the former in terms of the empirical risk and a generalization gap. Moreover, we showed that an estimate of the original size can be distilled from the larger model, consistently improving recovery error across different model parameters and algorithms.

At the same time, several questions remain unanswered. It is still unclear what determines the optimal degree of over-realization. Importantly, a complete understanding of the reasons behind the benefits of over-realization is still missing, and is likely to involve an optimization perspective. A natural hypothesis is that the improvement might be due to having a larger number of "initial guesses", since a bigger model will provide a larger covering of the space at initialization. As a result, certain initial atoms will be more likely to fall *close* to some of the ground-truth atoms. This does not seem to be the sole responsible factor, however, as repeating the training process with only those atoms found by distillation (from their initialization) deteriorates performance.

On the other hand, we have noted that the optimization problem presents a phase transition of sorts as the model size grows. We demonstrate this by measuring the minimal number of times any atom in the estimated dictionary is used, for increasing number of atoms, and we compute this statistic for different models dimensions and sparsity levels. One can see in Figure 5b that as soon as the number of atoms exceeds the ground truth size (70 in this case), this statistic drastically drops, reflecting the two-type behaviour illustrated in Section 4. On the one hand, this can in fact provide a practical way of determining the (unknown) size of the ground-truth model in practice, which might be worth in its own right. On the other, we believe this might reflects a fundamental change in the optimization landscape. In the $p' > p$ setting, the learning problem might become more amenable to practical optimization algorithms, thus finding a better solution. Further research in this direction will enable to characterize the reported results better, and might extend the application of these ideas to other unsupervised machine learning models.

Acknowledgements

This work was supported by NSF grants CCF 2007649 and CCF 2008460. The authors thank Qing Qu and Jacopo Teneggi for insightful comments at different stages of this work.

Appendix A. Recovery Guarantees

Lemma 3.1. *For a ground-truth dictionary $\mathbf{D}_0 \in \mathbb{R}^{d\times p}$ generating samples $\mathbf{x}_i = \mathbf{D}_0\boldsymbol{\gamma}_i$, where γ_i are k-sparse with non-zeros sampled iid from a zero mean and unit variance distribution, and for any estimate $\widehat{\mathbf{D}} \in \mathcal{D}_{p'}$, with overwhelming probability, we have that*

$$\frac{2}{k}\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}}[f_{\mathbf{x}}^{[k]}(\widehat{\mathbf{D}})] \le d(\mathbf{D}_0,\widehat{\mathbf{D}}) \le \frac{4}{k}\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}}[f_{\mathbf{x}}^{[1]}(\widehat{\mathbf{D}})] - \frac{2}{k}\zeta_k(k-1). \tag{12}$$

where $\zeta_k = \max\Big\{0\ ,\ 1-(k-2)\mu(\mathbf{D}) - 2\nu(\widehat{\mathbf{D}},\mathbf{D}_0)^2\Big\}$.

Proof Recall that $\mathbf{x}$ is sampled from distribution $\mathbb{P}$ by first sampling its support S from a uniform distribution of all possible supports with k elements, followed by sampling the non-zeros of its representation given the support. These non-zero entries are sampled i.i.d. from a distribution with mean zero and variance of 1. The sample is finally constructed as $\mathbf{x} = \mathbf{D}\boldsymbol{\gamma}$, with the ground truth dictionary $\mathbf{D}$.

Upper bound Let us first show the upper bound. Let $S = \mathrm{supp}(\boldsymbol{\gamma})$. Then,

$$f_{\mathbf{x}}^{[1]}(\widehat{\mathbf{D}}) = \inf_{\boldsymbol{\alpha}:\|\boldsymbol{\alpha}\|_0=1}\frac{1}{2}\|\mathbf{x}-\widehat{\mathbf{D}}\boldsymbol{\alpha}\|_2^2 \tag{13}$$

$$= \min_j \min_{\alpha_j}\frac{1}{2}\|\mathbf{D}_S\boldsymbol{\gamma}_S - \widehat{\mathbf{D}}_j\alpha_j\|_2^2 \tag{14}$$

$$= \frac{1}{2}\|\mathbf{D}_S\boldsymbol{\gamma}_S - \widehat{\mathbf{D}}_{j^*}\left(\widehat{\mathbf{D}}_{j^*}^T\mathbf{D}_S\boldsymbol{\gamma}_S\right)\|_2^2, \tag{15}$$

where the last inequality follows by solving for the optimal $\alpha_j^* = \widehat{\mathbf{D}}_j^T\mathbf{x}$, and j^* denotes the optimal choice of the atom index, given by (recall atoms are normalized)

$$j^* = \arg\min_j \|\mathbf{x}-\widehat{\mathbf{D}}_j\alpha_j^*\|_2^2 = \arg\max_j \big|\langle \mathbf{D}_S\boldsymbol{\gamma}_S, \widehat{\mathbf{D}}_j\rangle\big|. \tag{16}$$

See (Elad, 2010, Section 3.1) for a more detailed derivation. Let us denote by $\mathbf{D}_i$ the closest atom to $\widehat{\mathbf{D}}_{j^*}$ in S; i.e. $i = \arg\min_{k\in S}\min_{c\in\{+1,-1\}}\|\mathbf{D}_k - c\widehat{\mathbf{D}}_{j^*}\|_2$. Then, expand the expression above as follows

$$2f_{\mathbf{x}}^{[1]}(\widehat{\mathbf{D}}) = \|\left(\mathbf{D}_i\gamma_i + \mathbf{D}_{S\setminus i}\boldsymbol{\gamma}_{S\setminus i}\right) - \widehat{\mathbf{D}}_{j^*}\widehat{\mathbf{D}}_{j^*}^T\left(\mathbf{D}_i\gamma_i + \mathbf{D}_{S\setminus i}\boldsymbol{\gamma}_{S\setminus i}\right)\|_2^2 \tag{17}$$

$$= \|(\mathbf{D}_i - \widehat{\mathbf{D}}_{j^*}\widehat{\mathbf{D}}_{j^*}^T\mathbf{D}_i)\gamma_i + (\mathbf{I}-\widehat{\mathbf{D}}_{j^*}\widehat{\mathbf{D}}_{j^*}^T)\mathbf{D}_{S\setminus i}\boldsymbol{\gamma}_{S\setminus i}\|_2^2 \tag{18}$$

$$= \|(\mathbf{D}_i - \widehat{\mathbf{D}}_{j^*}\widehat{\mathbf{D}}_{j^*}^T\mathbf{D}_i)\gamma_i\|_2^2 + \|(\mathbf{I}-\widehat{\mathbf{D}}_{j^*}\widehat{\mathbf{D}}_{j^*}^T)\mathbf{D}_{S\setminus i}\boldsymbol{\gamma}_{S\setminus i}\|_2^2 + \dots \tag{19}$$

$$\cdots + 2\big\langle\ (\mathbf{D}_i - \widehat{\mathbf{D}}_{j^*}\widehat{\mathbf{D}}_{j^*}^T\mathbf{D}_i)\gamma_i\ ,\ (\mathbf{I}-\widehat{\mathbf{D}}_{j^*}\widehat{\mathbf{D}}_{j^*}^T)\mathbf{D}_{S\setminus i}\boldsymbol{\gamma}_{S\setminus i}\ \big\rangle \tag{20}$$

$$= A_i + B_i + C_i. \tag{21}$$

Let us now analyze $\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}}[2f_{\mathbf{x}}^{[1]}(\widehat{\mathbf{D}})] = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}}[A_i] + \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}}[B_i] + \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}}[C_i]$.

Consider first

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}}[A_i] = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}}[\|(\mathbf{D}_i - \widehat{\mathbf{D}}_{j^*}\widehat{\mathbf{D}}_{j^*}^T\mathbf{D}_i)\gamma_i\|_2^2] \tag{22}$$

$$= \mathop{\mathbb{E}}_{S}\left[\mathop{\mathbb{E}}_{\boldsymbol{\gamma}_S}[\|\mathbf{D}_i - \widehat{\mathbf{D}}_{j^*}(\widehat{\mathbf{D}}_{j^*}^T\mathbf{D}_i)\|_2^2\gamma_i^2 \mid S]\right] \tag{23}$$

$$= \mathop{\mathbb{E}}_{S}\left[\|\mathbf{D}_i - \widehat{\mathbf{D}}_{j^*}(\widehat{\mathbf{D}}_{j^*}^T\mathbf{D}_i)\|_2^2\right] \tag{24}$$

$$= \frac{k}{p}\sum_{i=1}^{p}\|\mathbf{D}_i - \rho_i\widehat{\mathbf{D}}_{j^*}\|_2^2, \tag{25}$$

where we used the fact that $\mathbb{E}[\gamma_i^2] = 1$ and we defined $\rho_i \coloneqq \widehat{\mathbf{D}}_{j^*}^T \mathbf{D}_i$.

Looking at the third term,

$$\frac{1}{2} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}} [C_i] = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}} \left[\left\langle \ (\mathbf{D}_i - \widehat{\mathbf{D}}_{j^*}\widehat{\mathbf{D}}_{j^*}^T \mathbf{D}_i)\gamma_i \ , \ (\mathbf{I} - \widehat{\mathbf{D}}_{j^*}\widehat{\mathbf{D}}_{j^*}^T)\mathbf{D}_{S\setminus i}\boldsymbol{\gamma}_{S\setminus i} \ \right\rangle \right] \tag{26}$$

$$= \mathop{\mathbb{E}}_{S} \left[\mathop{\mathbb{E}}_{\boldsymbol{\gamma}_S} \left[\left\langle \ (\mathbf{D}_i - \rho_i \widehat{\mathbf{D}}_{j^*})\gamma_i \ , \ (\mathbf{I} - \widehat{\mathbf{D}}_{j^*}\widehat{\mathbf{D}}_{j^*}^T)\mathbf{D}_{S\setminus i}\boldsymbol{\gamma}_{S\setminus i} \ \right\rangle \mid S \right] \right] \tag{27}$$

$$= \mathop{\mathbb{E}}_{S} \left[\sum_{q\in S\setminus i} \mathop{\mathbb{E}}_{\boldsymbol{\gamma}_S} \left[\left\langle (\mathbf{D}_i - \rho_i \widehat{\mathbf{D}}_{j^*})\gamma_i, (\mathbf{I} - \widehat{\mathbf{D}}_{j^*}\widehat{\mathbf{D}}_{j^*}^T)\mathbf{D}_q\gamma_q \right\rangle \mid S \right] \right] \tag{28}$$

$$= \mathop{\mathbb{E}}_{S} \left[\sum_{q\in S\setminus i} \mathop{\mathbb{E}}_{\boldsymbol{\gamma}_S} \left[\gamma_i\gamma_q \langle (\mathbf{D}_i - \rho_i \widehat{\mathbf{D}}_{j^*}), (\mathbf{I} - \widehat{\mathbf{D}}_{j^*}\widehat{\mathbf{D}}_{j^*}^T)\mathbf{D}_q \rangle \mid S \right] \right] \tag{29}$$

$$= 0 \tag{30}$$

because $\mathbb{E}[\gamma_i\gamma_q] = \mathbb{E}[\gamma_i]\mathbb{E}[\gamma_q] = 0$, since the variables are independent and of zero mean. Thus, so far we have that

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}} [f_{\mathbf{x}}^{[1]}(\widehat{\mathbf{D}})] = \frac{k}{2p} \sum_{i=1}^{p} \|\mathbf{D}_i - \rho_i \widehat{\mathbf{D}}_{j^*}\|_2^2 + \frac{1}{2} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}} [B_i]. \tag{31}$$

First, note that $\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}} [B_i] > 0$. Consider a tighter lower bound as follows

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}} [B_i] = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}} \ \|(\mathbf{I} - \widehat{\mathbf{D}}_{j^*}\widehat{\mathbf{D}}_{j^*}^T)\mathbf{D}_{S\setminus i}\boldsymbol{\gamma}_{S\setminus i}\|_2^2 \tag{32}$$

$$= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}} \ \left[\|\mathbf{D}_{S\setminus i}\boldsymbol{\gamma}_{S\setminus i}\|_2^2 + \|\widehat{\mathbf{D}}_{j^*}\widehat{\mathbf{D}}_{j^*}^T\mathbf{D}_{S\setminus i}\boldsymbol{\gamma}_{S\setminus i}\|_2^2 - 2\langle \mathbf{D}_{S\setminus i}\boldsymbol{\gamma}_{S\setminus i}, \widehat{\mathbf{D}}_{j^*}\widehat{\mathbf{D}}_{j^*}^T\mathbf{D}_{S\setminus i}\boldsymbol{\gamma}_{S\setminus i} \rangle \right] \tag{33}$$

$$= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}} \ \|\mathbf{D}_{S\setminus i}\boldsymbol{\gamma}_{S\setminus i}\|_2^2 + \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}} \|\widehat{\mathbf{D}}_{j^*}\widehat{\mathbf{D}}_{j^*}^T\mathbf{D}_{S\setminus i}\boldsymbol{\gamma}_{S\setminus i}\|_2^2 - 2 \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}} (\widehat{\mathbf{D}}_{j^*}^T\mathbf{D}_{S\setminus i}\boldsymbol{\gamma}_{S\setminus i})^2 \tag{34}$$

$$\geq \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}} \ \|\mathbf{D}_{S\setminus i}\boldsymbol{\gamma}_{S\setminus i}\|_2^2 - 2 \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}} (\widehat{\mathbf{D}}_{j^*}^T\mathbf{D}_{S\setminus i}\boldsymbol{\gamma}_{S\setminus i})^2 \tag{35}$$

$$= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}} \ \|\mathbf{D}_{S\setminus i}\boldsymbol{\gamma}_{S\setminus i}\|_2^2 - 2 \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}} \left(\sum_{k\in S\setminus i} \widehat{\mathbf{D}}_{j^*}^T\mathbf{D}_k\gamma_k \right)^2 \tag{36}$$

$$\geq \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}} \ \|\mathbf{D}_{S\setminus i}\boldsymbol{\gamma}_{S\setminus i}\|_2^2 - 2 \max_{k\in S\setminus i} \left| \widehat{\mathbf{D}}_{j^*}^T\mathbf{D}_k \right|^2 \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}} \Big(\sum_{k\in S\setminus i} \gamma_k \Big)^2 \tag{37}$$

$$\geq \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}} \ \|\mathbf{D}_{S\setminus i}\boldsymbol{\gamma}_{S\setminus i}\|_2^2 - 2 \max_{k\in [p]\setminus i} \left| \widehat{\mathbf{D}}_{j^*}^T\mathbf{D}_k \right|^2 \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}} \Big(\sum_{k\in S\setminus i} \gamma_k \Big)^2 \tag{38}$$

$$\geq \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}} \ \|\mathbf{D}_{S\setminus i}\boldsymbol{\gamma}_{S\setminus i}\|_2^2 - 2\nu^2(k-1) \tag{39}$$

where we used the fact that $\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}} \Big(\sum_{k\in S\setminus i} \gamma_k \Big)^2 = k-1$ since the variables are independent and have variance of 1. Additionally, we defined $\nu = \max_j \max_{k\in[p]\setminus i^*} \left| \widehat{\mathbf{D}}_j^T\mathbf{D}_k \right|$, with $i^* = \arg\max_{k\in[p]} \left| \widehat{\mathbf{D}}_j^T\mathbf{D}_k \right|$. In other words, i^* denotes the nearest neighbor in $\mathbf{D}$ for every $\widehat{\mathbf{D}}_j$. Continuing from above,

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}} [B_i] \geq \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}} \ \|\mathbf{D}_{S\setminus i}\boldsymbol{\gamma}_{S\setminus i}\|_2^2 - 2\nu^2(k-1) \tag{40}$$

$$\geq \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}} \ (1 - \delta_{k-1})\|\boldsymbol{\gamma}_{S\setminus i}\|_2^2 - 2\nu^2(k-1) \tag{41}$$

$$\geq (1 - (k-2)\mu(\mathbf{D}))(k-1) - 2\nu^2(k-1) \tag{42}$$

$$= \max\{ \big[1 - (k-2)\mu(\mathbf{D}) - 2\nu^2 \big](k-1), 0 \} \tag{43}$$

where δ_{k-1} is the $(k-1)$-RIP constant of $\mathbf{D}$, and we then used the bound with the mutual coherence $\delta_k \leq (k-1)\mu(\mathbf{D})$. In the last line, we added the condition that $\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}}[B_i] \geq 0$.

Thus, defining $\zeta_k \coloneqq \max\left\{0, 1-(k-2)\mu(\mathbf{D})-2\nu^2\right\}$, we can write

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}}[f_{\mathbf{x}}^{[1]}(\widehat{\mathbf{D}})] \geq \frac{k}{2p}\sum_{i=1}^{p}\|\mathbf{D}_i - \rho_i\widehat{\mathbf{D}}_{j^*}\|_2^2 + \frac{1}{2}\big[1-(k-2)\mu(\mathbf{D})-2\nu^2\big](k-1) \tag{44}$$

$$\geq \frac{k}{2p}\sum_{i=1}^{p}\|\mathbf{D}_i - \rho_i\widehat{\mathbf{D}}_{j^*}\|_2^2 + \frac{1}{2}\zeta_k(k-1). \tag{45}$$

Finally, recalling the definition of ρ_i (and that the atoms have unit norm) note that

$$\|\mathbf{D}_i - (\mathbf{D}_i^T\widehat{\mathbf{D}}_{j^*})\widehat{\mathbf{D}}_{j^*}\|_2^2 \geq \frac{1}{2}\min(\|\mathbf{D}_i - \widehat{\mathbf{D}}_{j^*}\|_2^2, \|\mathbf{D}_i + \widehat{\mathbf{D}}_{j^*}\|_2^2) = \frac{1}{2}d(\mathbf{D}_i, \hat{\mathbf{D}}_{j^*})$$

Recall that $\mathbf{D}_i$ is the closest atom to $\widehat{\mathbf{D}}_{j^*}$ out of those in the support S, and their distance might be equal or larger to the closest atom in $\widehat{\mathbf{D}}$ to $\mathbf{D}_i$; i.e.

$$d(\mathbf{D}_i, \hat{\mathbf{D}}_{j^*}) \geq \min_j d(\mathbf{D}_i, \hat{\mathbf{D}}_j).$$

Thus,

$$\frac{1}{p}\sum_{i=1}^{p}\min_j d(\mathbf{D}_i, \hat{\mathbf{D}}_j) = d(\mathbf{D}, \widehat{\mathbf{D}}) \leq \frac{4}{k}\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}}[f_{\mathbf{x}}^{[1]}(\widehat{\mathbf{D}})] - \frac{2}{k}\zeta_k(k-1). \tag{46}$$

Lower bound Let us now focus on the lower bound for $d(\mathbf{D}, \widehat{\mathbf{D}})$. For any S, let $\hat{\mathbf{D}}_{\hat{S}}$ contain the atoms from $\hat{\mathbf{D}}$ that are closest to the ones in $\mathbf{D}_S$, i.e.,

$$d(\mathbf{D}_{S(i)}, \hat{\mathbf{D}}_{\hat{S}(i)}) = d(\mathbf{D}_{S(i)}, \hat{\mathbf{D}}),\ \forall i \leq k.$$

Then,

$$\begin{aligned} f_{\mathbf{x}}^{[k]}(\widehat{\mathbf{D}}) &= \inf_{\boldsymbol{\alpha}:\|\boldsymbol{\alpha}\|_0=k}\frac{1}{2}\|\mathbf{x}-\widehat{\mathbf{D}}\boldsymbol{\alpha}\|_2^2 \\ &\leq \min_{\boldsymbol{\alpha}_{\hat{S}}}\frac{1}{2}\|\mathbf{D}_S\boldsymbol{\gamma}_S - \widehat{\mathbf{D}}_{\hat{S}}\boldsymbol{\alpha}_{\hat{S}}\|_2^2 \\ &= \frac{1}{2}\|\mathbf{D}_S\boldsymbol{\gamma}_S - \widehat{\mathbf{D}}_{\hat{S}}(\widehat{\mathbf{D}}_{\hat{S}}^T\widehat{\mathbf{D}}_{\hat{S}})^{-1}\widehat{\mathbf{D}}_{\hat{S}}^T\mathbf{D}_S\boldsymbol{\gamma}_S\|_2^2, \end{aligned}$$

which implies

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\gamma}_S}[f_{\mathbf{x}}^{[k]}(\widehat{\mathbf{D}})] &= \frac{1}{2}\mathbb{E}_{\boldsymbol{\gamma}_S}[\|\mathbf{D}_S\boldsymbol{\gamma}_S - \widehat{\mathbf{D}}_{\hat{S}}(\widehat{\mathbf{D}}_{\hat{S}}^T\widehat{\mathbf{D}}_{\hat{S}})^{-1}\widehat{\mathbf{D}}_{\hat{S}}^T\mathbf{D}_S\boldsymbol{\gamma}_S\|_2^2] \\ &= \frac{1}{2}\sum_{i=1}^{k}\|\mathbf{D}_{S(i)} - \widehat{\mathbf{D}}_{\hat{S}}(\widehat{\mathbf{D}}_{\hat{S}}^T\widehat{\mathbf{D}}_{\hat{S}})^{-1}\widehat{\mathbf{D}}_{\hat{S}}^T\mathbf{D}_{S(i)}\|_2^2 \\ &\leq \frac{1}{2}\sum_{i=1}^{k}\|\mathbf{D}_{S(i)} - \hat{\mathbf{D}}_{\hat{S}(i)}\hat{\mathbf{D}}_{\hat{S}(i)}^T\mathbf{D}_{S(i)}\|_2^2 \\ &\leq \frac{1}{2}\sum_{i=1}^{k}d(\mathbf{D}_{S(i)}, \hat{\mathbf{D}}_{\hat{S}(i)}) = \frac{1}{2}\sum_{i=1}^{k}d(\mathbf{D}_{S(i)}, \hat{\mathbf{D}}), \end{aligned}$$

where the first line utilizes the fact that each entry of $\boldsymbol{\gamma}_S$ is i.i.d. with variance 1, and the third line follows because $\|\mathbf{D}_{S(i)} - \widehat{\mathbf{D}}_{\hat{S}}(\widehat{\mathbf{D}}_{\hat{S}}^T\widehat{\mathbf{D}}_{\hat{S}})^{-1}\widehat{\mathbf{D}}_{\hat{S}}^T\mathbf{D}_{S(i)}\|_2^2$ is the projection residual of $\mathbf{D}_{S(i)}$ onto the subspace spanned by $\widehat{\mathbf{D}}_{\hat{S}}$, which is smaller than the one onto a particular column of $\widehat{\mathbf{D}}_{\hat{S}}$. The last line follows because

$$\|\mathbf{a} - \mathbf{a}\mathbf{a}^T\mathbf{b}\|^2 = \|\mathbf{a}\|^2 - (\mathbf{a}^T\mathbf{b})^2 \leq \min\{\|\mathbf{a}\|^2 - 2(\mathbf{a}^T\mathbf{b}) + \|\mathbf{b}\|^2, \|\mathbf{a}\|^2 + 2(\mathbf{a}^T\mathbf{b}) + \|\mathbf{b}\|^2\} = d(\mathbf{a}, \mathbf{b})$$

for any unit norm vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$. Thus, finally,

$$\begin{aligned}
\mathbb{E}[f_{\mathbf{x}}^{[k]}(\widehat{\mathbf{D}})] &= \mathbb{E}_S\left[\mathbb{E}_{\boldsymbol{\gamma}_S}[f_{\mathbf{x}}^{[k]}(\widehat{\mathbf{D}}) \mid S]\right] \\
&\leq \frac{1}{2}\mathbb{E}_S\left[\sum_{i=1}^{k} d(\mathbf{D}_{S(i)}, \hat{\mathbf{D}})\right] \\
&= \frac{1}{2}\frac{\binom{p-1}{k-1}}{\binom{p}{k}}\sum_{i=1}^{p} d(\mathbf{D}_i, \hat{\mathbf{D}}) = \frac{1}{2}\frac{\frac{(p-1)!}{(k-1)!(p-k)!}}{\frac{(p)!}{(k)!(p-k)!}}\sum_{i=1}^{p} d(\mathbf{D}_i, \hat{\mathbf{D}}) = \frac{1}{2}\frac{k}{p}\sum_{i=1}^{p} d(\mathbf{D}_i, \hat{\mathbf{D}}) \\
&\leq \frac{k}{2} d(\mathbf{D}, \hat{\mathbf{D}}).
\end{aligned}$$

■

A.1 Differences between $f_{\mathbf{x}}^{[1]}$ and $f_{\mathbf{x}}^{[k]}$

Our main result provides an upper bound on the recovery of a ground truth dictionary $\mathbf{D}$ based on the measure $f_{\mathbf{x}}^{[1]}$, as opposed to the risk $f_{\mathbf{x}}^{[k]}$. A lower bound on the expectation of their difference can readily be provided by Lemma 3.1, from which one can obtain $\mathbb{E}\left[f_{\mathbf{x}}^{[1]} - f_{\mathbf{x}}^{[k]}\right] \geq \frac{\zeta_k}{2}(k-1)$.

An upper bound between these measures can be derived as follows. Recall that

$$f_{\mathbf{x}}^{[s]}(\widehat{\mathbf{D}}) = \inf_{\boldsymbol{\gamma}:\|\boldsymbol{\gamma}\|_0 \leq s} \frac{1}{2}\|\mathbf{x} - \widehat{\mathbf{D}}\boldsymbol{\gamma}\|_2^2.$$

Then,

$$\begin{aligned}
f_{\mathbf{x}}^{[1]}(\widehat{\mathbf{D}}) &= \min_j \min_{\alpha_j} \frac{1}{2}\|\mathbf{x} - \widehat{\mathbf{D}}_j\alpha_j\|_2^2 && (47)\\
&= \frac{1}{2}\|\mathbf{x} - \widehat{\mathbf{D}}_{j^*}(\widehat{\mathbf{D}}_{j^*}^T\mathbf{D}_S\boldsymbol{\gamma}_S)\|_2^2 && (48)\\
&= \frac{1}{2}\|\mathbf{x} - \widehat{\mathbf{D}}_{j^*}\hat{\alpha}_{j^*}\|_2^2 && (49)
\end{aligned}$$

from the same steps as in the proof of Lemma 3.1. On the other hand, there exists a support $\hat{S}$ so that

$$\begin{aligned}
f_{\mathbf{x}}^{[k]}(\widehat{\mathbf{D}}) &= \frac{1}{2}\|\mathbf{x} - \sum_{i\in\hat{S}} \widehat{\mathbf{D}}_i\hat{\gamma}_i\|_2^2 && (50)\\
&= \frac{1}{2}\|\mathbf{x} - \widehat{\mathbf{D}}_{\hat{S}}\hat{\gamma}_{\hat{S}}\|_2^2. && (51)
\end{aligned}$$

We now assume that the selected atom for $f_{\mathbf{x}}^{[1]}$ is included among the ones selected for $f_{\mathbf{x}}^{[k]}$, i.e. $j^* \in \hat{S}$. This holds by making stronger generative assumption (e.g. by requiring k and $d(\mathbf{D}, \widehat{\mathbf{D}})$ to be

small), or by employing specific algorithms to compute $f_{\mathbf{x}}^{[k]}$, such as Matching Pursuit and other greedy pursuit variations.

Then,

$$2\left(f_{\mathbf{x}}^{[1]}(\widehat{\mathbf{D}}) - f_{\mathbf{x}}^{[k]}(\widehat{\mathbf{D}})\right) = \|\mathbf{x} - \widehat{\mathbf{D}}_{j^*}\hat{\alpha}_{j^*}\|_2^2 - \|\mathbf{x} - \widehat{\mathbf{D}}_{\hat{S}}\hat{\gamma}_{\hat{S}}\|_2^2 \tag{52}$$

$$= \|\widehat{\mathbf{D}}_{j^*}\hat{\alpha}_{j^*}\|_2^2 - \|\widehat{\mathbf{D}}_{\hat{S}}\hat{\gamma}_{\hat{S}}\|_2^2 - 2\mathbf{x}^T\left(\widehat{\mathbf{D}}_{j^*}\hat{\alpha}_{j^*} - \widehat{\mathbf{D}}_{\hat{S}}\hat{\gamma}_{\hat{S}}\right) \tag{53}$$

$$\leq \|\widehat{\mathbf{D}}_{j^*}\hat{\alpha}_{j^*}\|_2^2 - \|\widehat{\mathbf{D}}_{\hat{S}}\hat{\gamma}_{\hat{S}}\|_2^2 + 2\|\mathbf{x}\|_2\|\widehat{\mathbf{D}}_{j^*}\hat{\alpha}_{j^*} - \widehat{\mathbf{D}}_{\hat{S}}\hat{\gamma}_{\hat{S}}\|_2. \tag{54}$$

We will now show that $\|\widehat{\mathbf{D}}_{j^*}\hat{\alpha}_{j^*}\|_2^2 \leq \|\widehat{\mathbf{D}}_{\hat{S}}\hat{\gamma}_{\hat{S}}\|_2^2$. To see this, note first that $\widehat{\mathbf{D}}_{\hat{S}}\widehat{\mathbf{D}}_{\hat{S}}^+ \succcurlyeq 0$. Letting $\widehat{\mathbf{D}}_{\hat{S}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, we have $\widehat{\mathbf{D}}_{\hat{S}}\widehat{\mathbf{D}}_{\hat{S}}^+ = \mathbf{U}\mathbf{U}^T$. Since $j^* \in \hat{S}$, $\widehat{\mathbf{D}}_{j^*} \in \text{span}(\widehat{\mathbf{D}}_{\hat{S}})$, which means that we can write $\widehat{\mathbf{D}}_{j^*} = \mathbf{U}\mathbf{b}$, for some $\mathbf{b} \in \mathbb{R}^{|\hat{S}|}$. Therefore,

$$\widehat{\mathbf{D}}_{\hat{S}}\widehat{\mathbf{D}}_{\hat{S}}^+ - \widehat{\mathbf{D}}_{j^*}\widehat{\mathbf{D}}_{j^*}^T = \mathbf{U}\mathbf{U}^T - \mathbf{U}\mathbf{b}\mathbf{b}^T\mathbf{U}^T = \mathbf{U}\left(\mathbf{I} - \mathbf{b}\mathbf{b}^T\right)\mathbf{U}^T \succcurlyeq 0,$$

because $\|\mathbf{b}\|_2 = \|\widehat{\mathbf{D}}_{j^*}\|_2 = 1$. As a result, and recalling that $\hat{\alpha}_j = \widehat{\mathbf{D}}_{j^*}^T\mathbf{x}$ and $\hat{\boldsymbol{\gamma}}_{\hat{S}} = \widehat{\mathbf{D}}_{\hat{S}}^+\mathbf{x}$, we have that

$$\mathbf{x}^T\left(\widehat{\mathbf{D}}_{\hat{S}}\widehat{\mathbf{D}}_{\hat{S}}^+ - \widehat{\mathbf{D}}_{j^*}\widehat{\mathbf{D}}_{j^*}^T\right)\mathbf{x} \geq 0 \tag{55}$$

$$\mathbf{x}^T\widehat{\mathbf{D}}_{\hat{S}}\widehat{\mathbf{D}}_{\hat{S}}^+\mathbf{x} \geq \mathbf{x}^T\widehat{\mathbf{D}}_{j^*}\widehat{\mathbf{D}}_{j^*}^T\mathbf{x} \tag{56}$$

$$\|\widehat{\mathbf{D}}_{\hat{S}}\hat{\boldsymbol{\gamma}}_{\hat{S}}\|_2^2 = \|\widehat{\mathbf{D}}_{\hat{S}}\mathbf{D}_{\hat{S}}^+\mathbf{x}\|_2^2 \geq \|\widehat{\mathbf{D}}_{j^*}\widehat{\mathbf{D}}_{j^*}^T\mathbf{x}\|_2^2 = \|\widehat{\mathbf{D}}_{j^*}\hat{\alpha}_j\|_2^2. \tag{57}$$

Thus, resuming from Equation (54) and defining $\Delta_{j^*} = \hat{\alpha}_{j^*} - \hat{\boldsymbol{\gamma}}_{j^*}$, we have

$$f_{\mathbf{x}}^{[1]}(\widehat{\mathbf{D}}) - f_{\mathbf{x}}^{[k]}(\widehat{\mathbf{D}}) \leq \|\mathbf{x}\|_2\|\widehat{\mathbf{D}}_{j^*}\hat{\alpha}_{j^*} - \widehat{\mathbf{D}}_{\hat{S}}\hat{\gamma}_{\hat{S}}\|_2 \tag{58}$$

$$\leq \|\mathbf{x}\|_2\left(|\Delta_{j^*}| + \|\widehat{\mathbf{D}}_{\hat{S}\setminus j^*}\hat{\gamma}_{\hat{S}\setminus j^*}\|_2\right) \tag{59}$$

$$\leq \|\mathbf{x}\|_2\left(|\Delta_{j^*}| + (k-1)\|\hat{\gamma}_{\hat{S}}\|_\infty\right). \tag{60}$$

Note that $\Delta_{j^*} = 0$ whenever $k = 1$, and this quantity is bounded by $2\|\hat{\gamma}_{\hat{S}}\|_\infty$ whenever $k > 1$. Thus, $|\Delta_{j^*}| \leq 2(k-1)\|\hat{\gamma}_{\hat{S}}\|_\infty$. Putting everything together,

$$f_{\mathbf{x}}^{[1]}(\widehat{\mathbf{D}}) - f_{\mathbf{x}}^{[k]}(\widehat{\mathbf{D}}) \leq 3(k-1)\|\mathbf{x}\|_2\|\hat{\gamma}\|_\infty. \tag{61}$$

Note that this is tight when $k = 1$.

As a last remark, one could provide a bound for the maximum value of the coefficients $\hat{\gamma}_{\hat{S}}$ for any $\hat{S}$ by noting that

$$\|\hat{\gamma}_{\hat{S}}\|_\infty^2 \leq \|\hat{\gamma}_{\hat{S}}\|_2^2 = \|\widehat{\mathbf{D}}_{\hat{S}}^+\mathbf{x}\|_2^2 \leq \|\widehat{\mathbf{D}}_{\hat{S}}^+\|_2^2\|\mathbf{x}\|_2^2 \leq \frac{1}{(1-\mu(\widehat{\mathbf{D}})(k-1))^2}\|\mathbf{x}\|_2^2.$$

See Ben-Haim et al. (2010) for a proof for the last step.

Appendix B. Pruning Guarantees

Let $\mathbf{D}_0 \in \mathbb{R}^{d\times p}$ and consider, without loss of generality, that $\widehat{\mathbf{D}} = [\widehat{\mathbf{D}}_0, \mathbf{A}] \in \mathbb{R}^{d\times p'}$ with $\hat{\mathbf{D}}_0 \in \mathbb{R}^{d\times m}$, with $m \leq p'$, such that $d(\hat{\mathbf{D}}_i^0, \mathbf{D}_0) \leq \epsilon$ for all $i \in \{1,\ldots,m\}$, and $d(\mathbf{A}_j, \mathbf{D}_0) > \epsilon$ for all $j \in \{1,\ldots,p'-m\}$. In other words, $\hat{\mathbf{D}}_0$ contains all those m atoms that are ϵ-close to those in $\mathbf{D}_0$, while $\mathbf{A}$ contains those that are further away. Additionally, we require that each atom in $\mathbf{D}_0$ has at least one ϵ-neighbor in $\hat{\mathbf{D}}_0$; i.e. $d(\mathbf{D}_i^0, \hat{\mathbf{D}}_0) \leq \epsilon$ for all $i \in \{1,\ldots,p\}$. We allow $m \geq p$

since the over-realized estimate $\hat{\mathbf{D}}$ may naturally contain several atoms that are close to a real one. Also suppose that both $\mathbf{D}_0$ and $\hat{\mathbf{D}}$ are column-wise normalized for simplicity. Let us denote by $\mu(\mathbf{D}_0, \mathbf{A}) = \max_{i,j} \left|\langle \mathbf{D}_i^0, \mathbf{A}_j \rangle\right|$ the mutual coherence between $\mathbf{D}_0$ and $\mathbf{A}$. With these definitions, we have the following result:

Theorem 4.1. *Let* $\mathbf{x}$ *be a* k*-sparse signal under* $\mathbf{D}_0$*, i.e., there exists* $\boldsymbol{\gamma} \in \mathbb{R}^p$ *with* $\|\boldsymbol{\gamma}\|_0 \leq k$ *such that* $\mathbf{x} = \mathbf{D}_0\boldsymbol{\gamma}$*. Then,* $\text{argmax}_k \left|\mathbf{x}^T \hat{\mathbf{d}}_i\right| \in [m]$ *as long as*

$$k \leq \frac{1 - \frac{\epsilon}{2} + \sqrt{\epsilon} + \mu(\mathbf{D}_0)}{\mu(\mathbf{D}_0) + \sqrt{\epsilon} + \mu(\mathbf{D}_0, \mathbf{A})}. \tag{62}$$

Proof Without of loss generality, we assume that the entries of $\boldsymbol{\gamma}$ are placed in the decreasing order of the values $|\gamma_i|$. Recall that we require each atom in $\mathbf{D}_0$ has at least one ϵ-neighbor in $\hat{\mathbf{D}}_0$; i.e. $d(\mathbf{D}_i^0, \hat{\mathbf{D}}_0) \leq \epsilon$ for all $i \in \{1, \ldots, p\}$. For simplicity, we assume $d(\mathbf{D}_i^0, \hat{\mathbf{D}}_i) = \|\mathbf{D}_i^0 - \hat{\mathbf{D}}_i\|^2 \leq \epsilon$ for all $i \in \{1, \ldots, p\}$, i.e., the i-th column of $\hat{\mathbf{D}}_0$ (or $\hat{\mathbf{D}}$) is ϵ-close to the i-th atom of $\mathbf{D}_0$.

To show the atom that has the largest correlation with $\mathbf{x}$ must be within the first m columns of $\widehat{\mathbf{D}}$, we need to find $i \in [m]$ such that

$$\left|\mathbf{x}^\top \widehat{\mathbf{D}}_i^0\right| > \left|\mathbf{x}^\top \mathbf{A}_\ell\right|, \ \forall \ell. \tag{63}$$

Towards that goal, we choose $i = 1$ (as $|\gamma_1|$ is the largest sparse coefficient) to get

$$\begin{aligned} \left|\mathbf{x}^\top \widehat{\mathbf{D}}_1^0\right| = \left|\sum_{i=1}^k \gamma_i (\mathbf{D}_i^0)^\top \widehat{\mathbf{D}}_1^0\right| &\geq (1 - \frac{\epsilon}{2}) \left|\gamma_1\right| - (\mu(\mathbf{D}_0) + \sqrt{\epsilon}) \sum_{i=2}^k \left|\gamma_i\right| \\ &\geq \left((1 - \frac{\epsilon}{2}) - (k-1)(\mu(\mathbf{D}_0) + \sqrt{\epsilon})\right) \left|\gamma_1\right|, \end{aligned} \tag{64}$$

where the first inequality follows because

$$(\mathbf{D}_1^0)^\top \widehat{\mathbf{D}}_1^0 = 1 - \frac{1}{2}\|\mathbf{D}_1^0 - \widehat{\mathbf{D}}_1^0\|_2^2 \geq 1 - \frac{\epsilon}{2}$$

and

$$(\mathbf{D}_i^0)^\top \widehat{\mathbf{D}}_1^0 = (\mathbf{D}_i^0)^\top \mathbf{D}_1^0 + (\mathbf{D}_i^0)^\top (\widehat{\mathbf{D}}_1^0 - \mathbf{D}_1^0) \leq \mu(\mathbf{D}_0) + \|\widehat{\mathbf{D}}_1^0 - \mathbf{D}_1^0\|_2 \leq \mu(\mathbf{D}_0) + \sqrt{\epsilon}$$

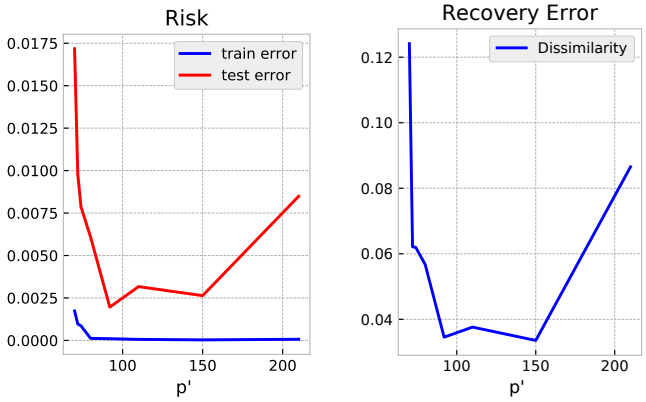
for all $2 \leq i \leq p$. On the other hand, we have

$$\left|\mathbf{x}^\top \mathbf{A}_\ell\right| = \left|\sum_{i=1}^k \gamma_i (\mathbf{D}_i^0)^\top \mathbf{A}_\ell\right| \leq \mu(\mathbf{D}_0, \mathbf{A}) \sum_{i=1}^k \left|\gamma_i\right| \leq k\mu(\mathbf{D}_0, \mathbf{A}) \left|\gamma_1\right|, \ \forall \ell.$$

which together with Equation (64) and Equation (62) gives Equation (63), implying that the element chosen by the first step of OMP must correspond to the one that is close to the correct dictionary, $\mathbf{D}_0$. ■

Appendix C. Further Numerical Results

C.1 Descriptive results on over-realized dictionary learning



(a) (b)

Figure 6: Same experimental setting as in Figure 1, but reporting the best run for each p'.

C.2 Results on recovery and risk improvements based on over-realized dictionary learning followed by distillation

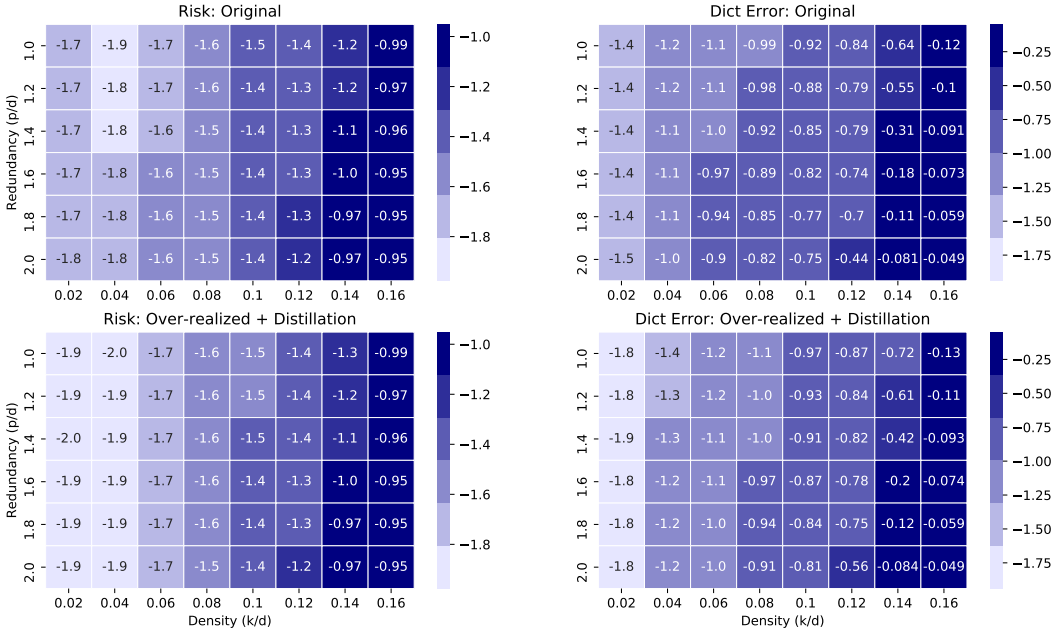


Figure 7: Risk and dictionary recovery error ($\log_{10}$ thereof, lower is better) of the estimate provided by *traditional* dictionary learning (i.e. $\widehat{\mathbf{D}} \in \mathcal{D}_p$) and that resulting from the proposed over-realized approach (i.e. $\widehat{\mathbf{D}} \in \mathcal{D}_{p'}$) followed by distillation to the original size, over a number of parameters (sparsity, dimension, and redundancy). Algorithm: ODL+Lasso.

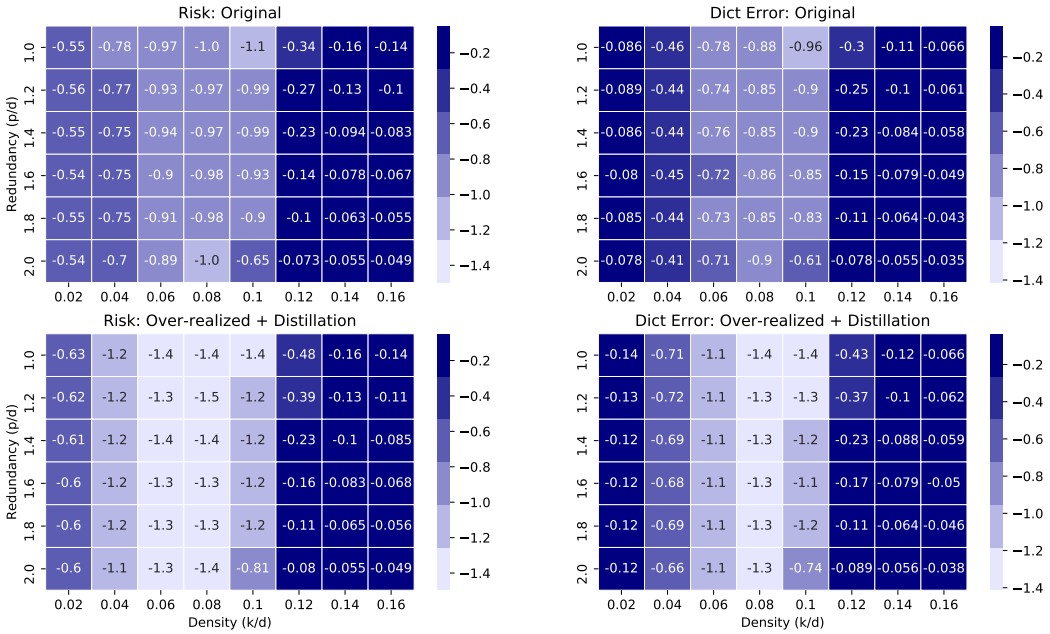


Figure 8: Risk and dictionary recovery error ($\log_{10}$ thereof, lower is better) of the estimate provided by *traditional* dictionary learning (i.e. $\widehat{\mathbf{D}} \in \mathcal{D}_p$) and that resulting from the proposed over-realized approach (i.e. $\widehat{\mathbf{D}} \in \mathcal{D}_{p'}$) followed by distillation to the original size, over a number of parameters (sparsity, dimension, and redundancy). Algorithm: K-SVD.

References

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

Alekh Agarwal, Animashree Anandkumar, Prateek Jain, and Praneeth Netrapalli. Learning sparsely used overcomplete dictionaries via alternating minimization. *SIAM Journal on Optimization*, 26 (4):2775–2799, 2016.

M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006a.

Michal Aharon, Michael Elad, and Alfred M Bruckstein. On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them. *Linear algebra and its applications*, 416(1): 48–67, 2006b.

Sanjeev Arora and Andrej Risteski. Provable benefits of representation learning. *arXiv preprint arXiv:1706.04601*, 2017.

Sanjeev Arora, Aditya Bhaskara, Rong Ge, and Tengyu Ma. More algorithms for provable dictionary learning. *arXiv preprint arXiv:1401.0579*, 2014a.

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

Sanjeev Arora, Rong Ge, Tengyu Ma, and Ankur Moitra. Simple, efficient, and neural algorithms for sparse coding. 2015.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Zvika Ben-Haim, Yonina C Eldar, and Michael Elad. Coherence-based performance guarantees for estimating a sparse vector under random noise. *IEEE Transactions on Signal Processing*, 58(10): 5030–5043, 2010.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Rares-Darius Buhai, Yoni Halpern, Yoon Kim, Andrej Risteski, and David Sontag. Empirical study of the benefits of overparameterization in learning latent variable models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1211–1219. PMLR, 2020.

Emmanuel J Candes and Terence Tao. Decoding by linear programming. *IEEE transactions on information theory*, 51(12):4203–4215, 2005.

Sanjoy Dasgupta and Leonard Schulman. A probabilistic analysis of em for mixtures of separated, spherical gaussians. *Journal of Machine Learning Research*, 8(Feb):203–226, 2007.

David L Donoho and Michael Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, 2003.

Michael Elad. *Sparse and redundant representations: from theory to applications in signal and image processing*. Springer Science & Business Media, 2010.

Kjersti Engan, Sven Ole Aase, and J Hakon Husoy. Method of optimal directions for frame design. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*, volume 5, pages 2443–2446. IEEE, 1999.

Quan Geng and John Wright. On the local correctness of ℓ_1-minimization for dictionary learning. In *2014 IEEE International Symposium on Information Theory*, pages 3180–3184. IEEE, 2014.

Sebastian Goldt, Madhu Advani, Andrew M Saxe, Florent Krzakala, and Lenka Zdeborová. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. In *Advances in Neural Information Processing Systems*, pages 6979–6989, 2019.

Rémi Gribonval, Rodolphe Jenatton, and Francis Bach. Sparse and spurious: dictionary learning with noise and outliers. *IEEE Transactions on Information Theory*, 61(11):6298–6319, 2015a.

Rémi Gribonval, Rodolphe Jenatton, Francis Bach, Martin Kleinsteuber, and Matthias Seibert. Sample complexity of dictionary learning and other matrix factorizations. *IEEE Transactions on Information Theory*, 61(6):3469–3486, 2015b.

Jiazhen Hong, Wei Qian, Yudong Chen, and Yuqian Zhang. A geometric approach to k-means. *arXiv preprint arXiv:2201.04822*, 2022.

Alexander Jung, Yonina C Eldar, and Norbert Görtz. On the minimax risk of dictionary learning. *IEEE Transactions on Information Theory*, 62(3):1501–1515, 2016.

Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2): 129–137, 1982.

Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11(Jan):19–60, 2010.

Andreas Maurer and Massimiliano Pontil. k-dimensional coding schemes in hilbert spaces. *IEEE Transactions on Information Theory*, 56(11):5839–5846, 2010.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.

Yagyensh Chandra Pati, Ramin Rezaiifar, and Perinkulam Sambamurthy Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on*, pages 40–44. IEEE, 1993.

Wei Qian, Yuqian Zhang, and Yudong Chen. Structures of spurious local minima in k-means. *IEEE Transactions on Information Theory*, 68(1):395–422, 2021.

Qing Qu, Yuexiang Zhai, Xiao Li, Yuqian Zhang, and Zhihui Zhu. Geometric analysis of nonconvex optimization landscapes for overcomplete learning. In *International Conference on Learning Representations*, 2019.

Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer relu neural networks. In *International Conference on Machine Learning*, pages 4433–4441, 2018.

Karin Schnass. On the identifiability of overcomplete dictionaries via the minimisation principle underlying k-svd. *Applied and Computational Harmonic Analysis*, 37(3):464–491, 2014.

Matthias Seibert. *Sample Complexity of Representation Learning for Sparse and Related Data Models*. PhD thesis, Technische Universität München, 2019.

Zahra Shakeri, Waheed U Bajwa, and Anand D Sarwate. Minimax lower bounds on dictionary learning for tensor data. *IEEE Transactions on Information Theory*, 64(4):2706–2726, 2018.

Ju Sun, Qing Qu, and John Wright. When are nonconvex problems not scary? *arXiv preprint arXiv:1510.06096*, 2015.

Yuandong Tian. Over-parameterization as a catalyst for better generalization of deep relu network. *arXiv preprint arXiv:1909.13458*, 2019.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Andreas M Tillmann. On the computational intractability of exact and approximate dictionary learning. *IEEE Signal Processing Letters*, 22(1):45–49, 2014.

Joel A Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004.

Daniel Vainsencher, Shie Mannor, and Alfred M Bruckstein. The sample complexity of dictionary learning. *Journal of Machine Learning Research*, 12(Nov):3259–3281, 2011.

Zitong Yang, Yaodong Yu, Chong You, Jacob Steinhardt, and Yi Ma. Rethinking bias-variance trade-off for generalization of neural networks. In *International Conference on Machine Learning*, pages 10767–10777. PMLR, 2020.

Yuexiang Zhai, Zitong Yang, Zhenyu Liao, John Wright, and Yi Ma. Complete dictionary learning via l4-norm maximization over the orthogonal group. *J. Mach. Learn. Res.*, 21(165):1–68, 2020.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.