On the Minimax Risk of Dictionary Learning

Alexander Jung, Yonina C. Eldar, *Fellow, IEEE*, and Norbert Görtz, *Senior Member, IEEE*

Abstract—We consider the problem of learning a dictionary matrix from a number of observed signals, which are assumed to be generated via a linear model with a common underlying dictionary. In particular, we derive lower bounds on the minimum achievable worst case mean squared error (MSE), regardless of computational complexity of the dictionary learning (DL) schemes. By casting DL as a classical (or frequentist) estimation problem, the lower bounds on the worst case MSE are derived following an established information-theoretic approach to minimax estimation. The main contribution of this paper is the adaption of these information-theoretic tools to the DL problem in order to derive lower bounds on the worst case MSE of any DL algorithm. We derive three different lower bounds applying to different generative models for the observed signals. The first bound only requires the existence of a covariance matrix of the (unknown) underlying coefficient vector. By specializing this bound to the case of sparse coefficient distributions and assuming the true dictionary satisfies the restricted isometry property, we obtain a lower bound on the worst case MSE of DL methods in terms of the signal-to-noise ratio (SNR). The third bound applies to a more restrictive subclass of coefficient distributions by requiring the non-zero coefficients to be Gaussian. Although the applicability of this bound is the most limited, it is the tightest of the three bounds in the low SNR regime. A particular use of our lower bounds is the derivation of necessary conditions on the required number of observations (sample size), such that DL is feasible, i.e., accurate DL schemes might exist. By comparing these necessary conditions with sufficient conditions on the sample size such that a particular DL technique is successful, we are able to characterize the regimes, where those algorithms are optimal in terms of required sample size.

Index Terms—Compressed sensing, dictionary learning, minimax risk, Fano inequality.

I. Introduction

ACCORDING to [1], the worldwide internet traffic in 2016 will exceed the Zettabyte threshold.[1] In view of the pervasive massive datasets generated at an ever increasing speed [2], [3], it is mandatory to be able to extract relevant information out of the observed data. A recent approach to this challenge, which has proven extremely useful for a wide

A. Jung is with the Department of Computer Science, Aalto University, Espoo 02150, Finland (e-mail: alexander.jung@aalto.fi).

Y. C. Eldar is with the Department of EE, Technion–Israel Institute of Technology, Haifa 32000, Israel (e-mail: yonina@ee.technion.ac.il).

N. Görtz is with the Institute of Telecommunications, Vienna University of Technology, Vienna 1040, Austria (e-mail: norbert.goertz@nt.tuwien.ac.at).

[1]One Zettabyte equals 10^{21} bytes.

range of applications, is *sparsity* and the related theory of *compressed sensing* (CS) [4]–[7]. In our context, sparsity means that the observed signals can be represented by a linear combination of a small number of prototype functions or atoms. In many applications the set of atoms is pre-specified and stored in a dictionary matrix. However, in some settings it might be necessary or beneficial to adaptively determine a dictionary based on the observations [8]–[10]. The task of adaptively determining the underlying dictionary matrix is referred to as *dictionary learning* (DL). DL has been considered for a wide range of applications, such as image processing [11]–[15], blind source separation [16], sparse principal component analysis [17], and more.

We consider observing N signals $\mathbf{y}_k \in \mathbb{R}^m$ generated via a fixed (but unknown) underlying dictionary $\mathbf{D} \in \mathbb{R}^{m \times p}$ which we would like to estimate. More precisely, the observations $\mathbf{y}_k$ are modeled as noisy linear combinations

$$\mathbf{y}_k = \mathbf{D}\mathbf{x}_k + \mathbf{n}_k, \tag{1}$$

where $\mathbf{n}_k$ is assumed to be zero-mean with i.i.d. components of variance σ^2. To formalize the estimation problem underlying DL, we assume the coefficient vectors $\mathbf{x}_k$ to be zero-mean random vectors with finite covariance matrix $\mathbf{\Sigma}_x$. Our first main result, i.e., Theorem 1 applies to a very wide class of coefficient distributions since it only requires a finite covariance matrix $\mathbf{\Sigma}_x$. In particular, it holds for non-sparse random coefficient vectors. However, the main focus of the paper (Corollary 2 and Theorem 3) will be on distributions for which the coefficient vector $\mathbf{x}_k$ is strictly s-sparse with probability one. In this work, we analyze the difficulty inherent to the problem of estimating the true dictionary $\mathbf{D} \in \mathbb{R}^{m \times p}$, which is deterministic but unknown, from the measurements $\mathbf{y}_k$, generated according to the linear model (1).

If we stack the observations $\mathbf{y}_k$, for $k = 1, \ldots, N$, columnwise into the data matrix $\mathbf{Y} \in \mathbb{R}^{m \times N}$, one can cast DL as a matrix factorization problem [18]. Given the data matrix $\mathbf{Y}$, we aim to find a dictionary matrix $\mathbf{D} \in \mathbb{R}^{m \times p}$ such that

$$\mathbf{Y} = \mathbf{D}\mathbf{X} + \mathbf{N} \tag{2}$$

where the column sparse matrix $\mathbf{X} \in \mathbb{R}^{p \times N}$ contains in its kth column the sparse expansion coefficients $\mathbf{x}_k$ of the signal $\mathbf{y}_k$. The noise matrix $\mathbf{N} = (\mathbf{n}_1, \ldots, \mathbf{n}_N) \in \mathbb{R}^{m \times N}$ accounts for small modelling and measurement errors.

A plethora of DL methods have been proposed and analyzed in the literature (e.g., [8], [19]–[27]). In a Bayesian setting, i.e., modeling the dictionary as random with a known prior distribution, the authors of [24], [25], and [28] devise a variant of the *approximate message passing* scheme [29] to the DL problem. The authors of [20]–[23] and [30] model

the dictionary as non-random and estimate the dictionary by solving the (non-convex) optimization problem

$$\min_{\mathbf{D}\in\mathcal{D},\mathbf{X}\in\mathbb{R}^{p\times N}} \|\mathbf{Y}-\mathbf{DX}\|_F^2 + \lambda\|\mathbf{X}\|_1, \qquad (3)$$

where $\|\mathbf{X}\|_1 \triangleq \sum_{k,l} |\mathbf{X}_{k,l}|$ and $\mathcal{D} \subseteq \mathbb{R}^{m\times p}$ denotes a constraint set, e.g., requiring the columns of the learned dictionary to have unit norm. The term $\lambda\|\mathbf{X}\|_1$ (with sufficiently large λ) in the objective (3) enforces the columns of the coefficient matrix $\mathbf{X}$ to be (approximately) sparse.

Assuming the true dictionary $\mathbf{D} \in \mathbb{R}^{m\times p}$ deterministic but unknown (with known size p) and the observations $\mathbf{y}_k$ i.i.d. according to the model (1), the works in [20]–[22] provide upper bounds on the distance between the generating dictionary $\mathbf{D}$ and the closest local minimum of (3). For the square (i.e., $p = m$) and noiseless ($\mathbf{N} = \mathbf{0}$) setting, [22] showed that $N = \mathcal{O}(p\log(p))$ observations suffice to guarantee that the dictionary is a local minimum of (3). Using the same setting (square dictionary and noiseless measurements), [26] proved the scaling $N = \mathcal{O}(p\log(p))$, for arbitrary sparsity levels, to be sufficient such that the dictionary matrix can be recovered perfectly from the measurements $\mathbf{y}_k$.[2] Our analysis, in contrast, takes measurement noise into account and yields lower bounds on the required sample size in terms of SNR. While the results on the square-dictionary and noiseless case are theoretically important, their practical relevance is limited. Considering the practically more common case of an overcomplete ($p > m$) dictionary $\mathbf{D}$ and noisy measurements ($\mathbf{N} \neq \mathbf{0}$), the authors of [21] show that a sample size of $N = \mathcal{O}(p^3 m)$ i.i.d. measurements $\mathbf{y}_k$ suffices for the existence of a local minimum of the cost function in (3) which is close to the true dictionary $\mathbf{D}$.

In contrast to approaches based on solving (3), a recent line of work [8], [26], [27] presents DL methods based on (graph-)clustering techniques. In particular, the set of observed samples $\mathbf{y}_k$ is clustered such that the elements within each cluster share a single generating column $\mathbf{d}_j$ of the underlying dictionary. The authors of [27] show that a sample size $N = \mathcal{O}(p^2 \log p)$ suffices for their clustering-based method to accurately recover the true underlying dictionary. However, this result applies only for sufficiently incoherent dictionaries $\mathbf{D}$ and for the case of vanishing sparsity rate, i.e., $s/p \to 0$. The scaling of the required sample size with the square of the number p of dictionary columns (neglecting logarithmic terms) is also predicted by our bounds. What sets our work apart from [27] is that we state our results in a non-asymptotic setting, i.e., our bounds can be evaluated for any given number p of dictionary atoms, dimension m of observed signals and nominal sparsity level s of the coefficient vector $\mathbf{x}_k$ in (1).

Although numerous DL schemes have been proposed and analyzed, existing results typically yield sufficient conditions (e.g., on the sample size N) such that DL is feasible. In contrast, necessary conditions which apply to any DL algorithm (irrespective of computational complexity) are far more limited. We are only aware of a single fundamental result that applies to a Bernoulli-Gauss prior for the coefficient vectors $\mathbf{x}_k$

in (1): This result, also known as the "coupon collector phenomenon" [26], states that in order to have every column $\mathbf{d}_j$ of the dictionary contributing in at least one observed signal $\mathbf{y}_k$ (i.e., the corresponding entry $x_{k,j}$ of the coefficient vector in (1) is non-zero) the sample size has to scale linearly with $(1/\theta)\log p$ where θ denotes the probability $\mathrm{P}\{x_{k,j} \neq 0\}$. For the choice $\theta = s/p$, which yields s-sparse coefficient vectors with high probability, this requirement effectively becomes $N \geq c_1(p/s)\log p$, with some absolute constant c_1.

In this paper we contribute to the understanding of necessary conditions or fundamental recovery thresholds for DL, by deriving lower bounds on the minimax risk for the DL problem. We define the risk incurred by a DL scheme as the mean squared error (MSE) using the Frobenius norm of the deviation from the true underlying dictionary. Since the minimax risk is defined as the minimum achievable worst case MSE, our lower bounds apply to the worst case MSE of any algorithm, regardless of its computational complexity. This paper seems to contain the first analysis that targets directly the fundamental limits on the achievable MSE of any DL method.

For the derivation of the lower bounds, we apply an established information-theoretic approach (cf. Section II) to minimax estimation, which is based on reducing a specific multiple hypothesis problem to minimax estimation of the dictionary matrix. Although this information-theoretic framework has been successfully applied to several other (sparse) minimax estimation problems [31]–[35], the adaptation of this method to DL seems to be new. The lower bounds on the minimax risk give insight into the dependencies of the achievable worst case MSE on the model parameters, i.e., the sparsity s, the dictionary size p, the dimension m of the observed signal and the SNR. Our lower bounds on the minimax risk have direct implications on the required sample size of accurate DL schemes. In particular our analysis reveals that, for a sufficiently incoherent underlying dictionary, the minimax risk of DL is lower bounded by $c_1 p^2/(\mathrm{SNR}N)$, where c_1 is some absolute constant. Thus, for a vanishing minimax risk it is necessary for the sample size N to scale linearly with the square of the number p of dictionary columns and inversely with the SNR. Finally, by comparing our lower bounds (on minimax risk and sample size) with the performance guarantees of existing learning algorithms, we can test if these methods perform close to optimal.

Some recent work on the sample complexity of DL [36], [37] presents upper bounds on the sample size such that the (expected) performance of an ideal learning method is close to its empirical performance observed when applied to the samples. While the authors of [36] measure the quality of the estimate $\widehat{\mathbf{D}}$ via the residual error obtained when sparsely approximating the observed vectors $\mathbf{y}_k$, we use a different risk measure based on the squared Frobenius norm of the deviation from the true underlying dictionary. Clearly, these two risk measures are related. Indeed, if the Frobenius norm $\|\widehat{\mathbf{D}} - \mathbf{D}\|_F$ is small, we can also expect that any sparse linear combination $\mathbf{Dx}$ using the dictionary $\mathbf{D}$ can also be well represented by a sparse linear combination $\widehat{\mathbf{D}}\mathbf{x}'$ using $\widehat{\mathbf{D}}$. Our results are somewhat complementary to the upper bounds in [36] in that they yield

---

[2]With high probability and up to scaling and permutations of the dictionary columns.

lower bounds on the required sample size such that there may exist accurate learning schemes (regardless of computational complexity).

The remainder of this paper is organized as follows: We introduce the minimax risk of DL and the information-theoretic method for lower bounding it in Section II. Lower bounds on the minimax risk for DL are presented in Section III. We also put our bounds into perspective by comparing their implications to the available performance guarantees of some DL schemes. Detailed proofs of the main results are contained in Section IV.

Throughout the paper, we use the following notation: Given a natural number $k \in \mathbb{N}$, we define the set $[k] \triangleq \{1, \ldots, k\}$. For a matrix $\mathbf{A} \in \mathbb{R}^{m \times p}$, we denote its Frobenius norm and its spectral norm by $\|\mathbf{A}\|_{\mathrm{F}} \triangleq \sqrt{\mathrm{Tr}\{\mathbf{A}\mathbf{A}^T\}}$ and $\|\mathbf{A}\|_2$, respectively. The open (Frobenius-norm) ball of radius $r > 0$ and center $\mathbf{D} \in \mathbb{R}^{m \times p}$ is denoted $\mathcal{B}(\mathbf{D}, r) \triangleq \{\mathbf{D}' \in \mathbb{R}^{m \times p} : \|\mathbf{D} - \mathbf{D}'\|_{\mathrm{F}} < r\}$. For a square matrix $\mathbf{A}$, the vector containing the elements along the diagonal of $\mathbf{A}$ is denoted $\mathrm{diag}\{\mathbf{A}\}$. Analogously, given a vector $\mathbf{a}$, we denote by $\mathrm{diag}\{\mathbf{a}\}$ the diagonal matrix whose diagonal is obtained from $\mathbf{a}$. The kth column of the identity matrix is denoted $\mathbf{e}_k$. For a matrix $\mathbf{X} \in \mathbb{R}^{p \times N}$, we denote by $\mathrm{supp}(\mathbf{X})$ the N-tuple $\big(\mathrm{supp}(\mathbf{x}_1), \ldots, \mathrm{supp}(\mathbf{x}_N)\big)$ of subsets given by concatenating the supports $\mathrm{supp}(\mathbf{x}_k)$ of the columns $\mathbf{x}_k$ of the matrix $\mathbf{X}$. The complementary Kronecker delta is denoted $\bar{\delta}_{l,l'}$, i.e., $\bar{\delta}_{l,l'} = 0$ if $l = l'$ and equal to one otherwise. We denote by $\mathbf{0}$ the vector or matrix with all entries equal to 0. The determinant of a square matrix $\mathbf{C}$ is denoted $|\mathbf{C}|$. The identity matrix is written as $\mathbf{I}$ or $\mathbf{I}_d$ when the dimension $d \times d$ is not clear from the context. Given a positive semidefinite (psd) matrix $\mathbf{C}$, we write its smallest eigenvalue as $\lambda_{\min}(\mathbf{C})$. The natural and binary logarithm of a number b are $\log(b)$ and $\log_2(b)$, respectively. For two sequences $g(N)$ and $f(N)$, indexed by the natural number N, we write $g = \mathcal{O}(f)$ and $g = \Theta(f)$ if, respectively, $g(N) \leq C' f(N)$ and $g(N) \geq C'' f(N)$ for some constants $C', C'' > 0$. If $g(N)/f(N) \to 0$, then we write $g = o(f)$. We denote by $\mathrm{E}_{\mathbf{X}} f(\mathbf{X})$ the expectation of the function $f(\mathbf{X})$ of the random vector (or matrix) $\mathbf{X}$.

II. Problem Formulation

A. Basic Setup

For our analysis we assume the observations $\mathbf{y}_k$ are i.i.d. realizations according to the random linear model

$$\mathbf{y} = \mathbf{D}\mathbf{x} + \mathbf{n}. \tag{4}$$

Thus, the vectors $\mathbf{y}_k$, $\mathbf{x}_k$ and $\mathbf{n}_k$, for $k = 1, \ldots, N$, in (1) are i.i.d. realizations of the random vectors $\mathbf{y}$, $\mathbf{x}$ and $\mathbf{n}$ in (4). Here, the matrix $\mathbf{D} \in \mathbb{R}^{m \times p}$, with $p \geq m$, represents the deterministic but unknown underlying dictionary, whose columns are the building blocks of the observed signals $\mathbf{y}_k$. The vector $\mathbf{x}$ represents zero mean random expansion coefficients, whose distribution is assumed to be known. Our analysis applies to a wide class of distributions. In fact, we only require the existence of the covariance matrix

$$\Sigma_x \triangleq \mathrm{E}_{\mathbf{x}}\{\mathbf{x}\mathbf{x}^T\}. \tag{5}$$

The effect of modeling and measurement errors are captured by the noise vector $\mathbf{n}$, which is assumed independent of $\mathbf{x}$ and is white Gaussian noise (AWGN) with zero mean and known variance σ^2. When combined with a sparsity enhancing prior on $\mathbf{x}$, the linear model (4) reduces to the sparse linear model (SLM) [38], which is the workhorse of CS [6], [39], [40]. However, while the works on the SLM typically assume the dictionary $\mathbf{D}$ in (4) perfectly known, we consider the situation where $\mathbf{D}$ is unknown.

In what follows, we assume the columns of the dictionary $\mathbf{D}$ to be normalized, i.e.,

$$\mathbf{D} \in \mathcal{D} \triangleq \{\mathbf{B} \in \mathbb{R}^{m \times p} | \mathbf{e}_k^T \mathbf{B}^T \mathbf{B} \mathbf{e}_k = 1, \text{ for all } k \in [p]\}. \tag{6}$$

The set $\mathcal{D}$ is known as the *oblique manifold* [21], [41], [42]. For fixed problem dimensions p, m and s, requiring (6) effectively amounts to identifying SNR with the quantity $\|\Sigma_x\|_2/\sigma^2$. Our analysis is local in the sense that we consider the true dictionary $\mathbf{D}$ to belong to a small neighborhood, i.e.,

$$\begin{aligned} \mathbf{D} \in \mathcal{X}(\mathbf{D}_0, r) &\triangleq \mathcal{B}(\mathbf{D}_0, r) \cap \mathcal{D} \\ &= \{\mathbf{D}' \in \mathcal{D} : \|\mathbf{D}' - \mathbf{D}_0\|_{\mathrm{F}} < r\} \end{aligned} \tag{7}$$

with a fixed and known "reference dictionary" $\mathbf{D}_0 \in \mathcal{D}$ and known radius[3] $r \leq 2\sqrt{p}$. This local analysis avoids ambiguity issues that are intrinsic to DL (cf. [26]). However, the lower bounds on the minimax risk derived under the locality constraint (7) trivially also apply to the global DL problem, i.e., where we only require (6).

B. The Minimax Risk

We will investigate fundamental limits on the accuracy achievable by any DL scheme, irrespective of its computational complexity. By a DL scheme, we mean an estimator $\widehat{\mathbf{D}}(\cdot)$ which maps the observation $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$ to an estimate $\widehat{\mathbf{D}}(\mathbf{Y})$ of the true underlying dictionary $\mathbf{D}$. The accuracy of a given learning method will be measured via the MSE $\mathrm{E}_{\mathbf{Y}}\{\|\widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D}\|_{\mathrm{F}}^2\}$, which is the expected squared distance of the estimate $\widehat{\mathbf{D}}(\mathbf{Y})$ from the true dictionary, measured in Frobenius norm. Note that the MSE of a given learning algorithm $\widehat{\mathbf{D}}(\mathbf{Y})$ depends on the true underlying dictionary $\mathbf{D}$, which is fixed but unknown. Therefore, the MSE cannot be minimized uniformly for all $\mathbf{D}$ [43]. However, for a given estimator $\widehat{\mathbf{D}}(\cdot)$, a reasonable performance measure is the worst case MSE $\sup_{\mathbf{D} \in \mathcal{X}(\mathbf{D}_0, r)} \mathrm{E}_{\mathbf{Y}}\{\|\widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D}\|_{\mathrm{F}}^2\}$ [44]. The optimum estimator under this criterion has smallest worst case MSE among all possible estimators. This smallest worst case MSE (referred to as minimax risk) is an intrinsic property of the estimation problem and does not depend on a specific estimator.

In general, the minimax risk ε^* depends on the sample size N, the dimension m of the observed signals, the number p of dictionary elements, the sparsity degree s and the noise variance σ^2. Moreover, the minimax risk also depends on the distribution of the coefficient vector $\mathbf{x}$ in (4). However, in what follows we assume this distribution to be fixed and known.

---

[3]Considering only values not exceeding $2\sqrt{p}$ for the radius r in (7) is reasonable since for any radius $r > 2\sqrt{p}$ we would obtain $\mathcal{X}(\mathbf{D}_0, r) = \mathcal{D}$ yielding the global DL problem.

For the sake of light notation, we will not make these dependencies explicit. Concretely, the minimax risk ε^* is defined as

$$\varepsilon^* \triangleq \inf_{\widehat{\mathbf{D}}} \sup_{\mathbf{D} \in \mathcal{X}(\mathbf{D}_0, r)} \mathrm{E}_{\mathbf{Y}}\{\|\widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D}\|_{\mathrm{F}}^2\}. \qquad (8)$$

We highlight that the minimax risk is defined here for a fixed and known distribution of the coefficient vector $\mathbf{x}_k$ in (1). In what follows, we derive three different lower bounds on the minimax risk by considering different types of coefficient distributions.

Note that while, at first sight, the locality assumption (7) may suggest that our analysis yields weaker results than the general setting, the opposite is actually true. Indeed, our lower bounds on the minimax risk predict that even under the additional a-priori knowledge that the true dictionary belongs to the (small) neighborhood of a known reference dictionary $\mathbf{D}_0$, the minimax risk is lower bounded by a strictly positive number which, for a sufficiently large sample size, does not depend on the size of the neighborhood at all. Also, from the definition (8) it is obvious that any lower bound on the minimax risk ε^* under the locality constraint (7) is simultaneously a lower bound on the minimax risk for global DL, which is obtained from (8) by replacing the constraint $\mathbf{D} \in \mathcal{X}(\mathbf{D}_0, r)$ in the inner maximization with the constraint $\mathbf{D} \in \mathcal{D}$.

The minimax problem (8) typically cannot be solved in closed-form. Instead of trying to exactly solve (8) and determine ε^*, we will derive lower bounds on ε^* by adapting an established information-theoretic methodology (cf., e.g., [31], [33], [45]) to the DL problem. Having a lower bound on the minimax risk ε^* allows to asses the performance of a given DL scheme. In particular, if the worst case MSE of a given method is close to the lower bound, then there is no point in searching for alternative algorithms with substantially better performance (in terms of worst case MSE). We emphasize that our bounds apply to any DL scheme, regardless of its computational complexity. In particular, these bounds apply to DL methods which do not exploit knowledge of the sparse coefficient distribution nor of the noise variance.

C. Information-Theoretic Lower Bounds on the Minimax Risk

A principled approach [31], [33], [45] to lower bounding the minimax risk ε^* is based on reducing a specific multiple hypothesis testing problem to minimax estimation of the dictionary $\mathbf{D}$. More precisely, if there exists an estimator with small worst case MSE, then this estimator can be used to solve a hypothesis testing problem. However, using Fano's inequality, there is a fundamental limit on the error probability for the hypothesis testing problem. This limit induces a lower bound on the worst case MSE of any estimator, i.e., on the minimax risk. Let us now outline the details of this approach.

First, within this approach one assumes that the true dictionary $\mathbf{D}$ in (4) is taken uniformly at random (u.a.r.) from a finite subset $\mathcal{D}_0 \triangleq \{\mathbf{D}_l\}_{l \in [L]} \subseteq \mathcal{X}(\mathbf{D}_0, r)$ for some $L \in \mathbb{N}$ (cf. Fig. 1). This subset $\mathcal{D}_0$ is constructed such that (i) any two distinct dictionaries $\mathbf{D}_l, \mathbf{D}_{l'} \in \mathcal{D}_0$ are separated by at least $\sqrt{8\varepsilon}$, i.e., $\|\mathbf{D}_l - \mathbf{D}_{l'}\|_{\mathrm{F}} \geq \sqrt{8\varepsilon}$ and (ii) it is hard to detect the true dictionary $\mathbf{D}$, drawn u.a.r. out of $\mathcal{D}_0$, based on observing
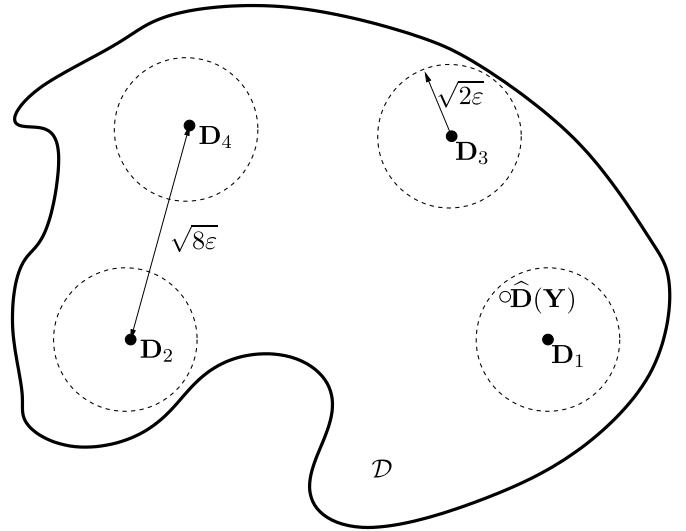

Fig. 1. A finite ensemble $\mathcal{D}_0 = \{\mathbf{D}_l\}_{l \in [L]}$, containing $L = 4$ dictionaries $\mathbf{D}_l \in \mathcal{D}$, which is used for deriving a lower bound $\varepsilon^* \geq \varepsilon$ on the minimax risk ε^* (cf. (8)). For the true dictionary $\mathbf{D} = \mathbf{D}_1$, we also depict a typical realization of an estimator $\widehat{\mathbf{D}}$ achieving the minimax risk.

$\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$. The existence of such a set $\mathcal{D}_0$ yields a relation between the sample size N and the remaining model parameters, i.e., m, p, s, σ which has to be satisfied such that at least one estimator with minimax-risk not exceeding ε may exist.

In order to find a lower bound $\varepsilon^* \geq \varepsilon$ on the minimax risk ε^* (cf. (8)), we hypothesize the existence of an estimator $\widehat{\mathbf{D}}(\mathbf{Y})$ achieving the minimax risk in (8). Then, the minimum distance detector

$$\underset{\mathbf{D}' \in \mathcal{D}_0}{\arg\min} \|\widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D}'\|_{\mathrm{F}} \qquad (9)$$

recovers the correct dictionary $\mathbf{D} \in \mathcal{D}_0$ if $\widehat{\mathbf{D}}(\mathbf{Y})$ belongs to the open ball $\mathcal{B}(\mathbf{D}, \sqrt{2\varepsilon})$ (indicated by the dashed circles in Fig. 1) centered at $\mathbf{D}$ and with radius $\sqrt{2\varepsilon}$. The information-theoretic method [31], [32], [45] of lower bounding the minimax risk ε^* consists then in relating, via Fano's inequality [46, Ch. 2], the error probability $\mathrm{P}\{\widehat{\mathbf{D}}(\mathbf{Y}) \notin \mathcal{B}(\mathbf{D}, \sqrt{2\varepsilon})\}$ to the mutual information (MI) between the observation $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$ and the dictionary $\mathbf{D}$ in (4), which is assumed to be drawn u.a.r. out of $\mathcal{D}_0$.

Thus, within this approach, the DL estimation problem is interpreted as a communication problem, as illustrated in Fig. 2. The source selects the true dictionary $\mathbf{D} = \mathbf{D}_l$ by drawing u.a.r. an element $\mathbf{D}_l$ from the set $\mathcal{D}_0$. This element $\mathbf{D}_l$ then generates the "channel output" $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$ via the model (4) for N channel uses. The observation model (4) acts as a channel model, relating the input $\mathbf{D} = \mathbf{D}_l$ to the output $\mathbf{Y}$. A crucial step in the information-theoretic method is the analysis of the MI defined by [46]

$$I(\mathbf{Y}; l) \triangleq \mathrm{E}_{\mathbf{Y}, l}\left\{ \log \frac{p(\mathbf{Y}, l)}{p(\mathbf{Y}) p(l)} \right\},$$

where $p(\mathbf{Y}, l)$, $p(\mathbf{Y})$ and $p(l)$ denote the joint and marginal distributions, respectively, of the channel output $\mathbf{Y}$ and the random index l. A key challenge in applying this method to DL is that the model (4) does not correspond to a simple
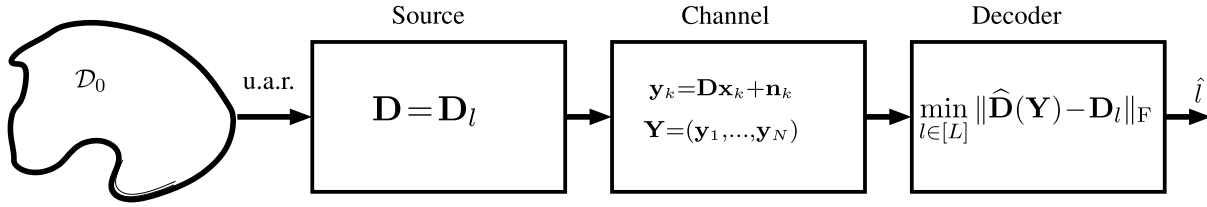
Fig. 2. Information-theoretic method for lower bounding the minimax risk.

AWGN channel, for which the MI between output and input can be characterized easily. Indeed, the model (4) corresponds to a fading channel with the vector $\mathbf{x}$ representing fading coefficients. As is known from the analysis of non-coherent channel capacity, characterizing the MI between output and input for fading channels is much more involved than for AWGN channels [47]. In particular, we require a tight upper bound on the MI $I(\mathbf{Y}; l)$ between the output $\mathbf{Y}$ and a random index l which selects the input $\mathbf{D} = \mathbf{D}_l$ u.a.r. from a finite set $\mathcal{D}_0 \subseteq \mathcal{X}(\mathbf{D}_0, r)$. Upper bounding $I(\mathbf{Y}; l)$ typically involves the analysis of the Kullback-Leibler (KL) divergence between the distributions of $\mathbf{Y}$ induced by different dictionaries $\mathbf{D} = \mathbf{D}_l$, $l \in [L]$.

Unfortunately, an exact characterization of the KL divergence between Gaussian mixture models is in general not possible and one has to resort to approximations or bounds [48]. A main conceptual contribution of this work is a strategy to avoid evaluating KL divergences between Gaussian mixture models. Instead, similar to the approach of [32], we assume that, in addition to the observation $\mathbf{Y}$, we also have access to some side information $\mathbf{T}(\mathbf{X})$, which depends only on the coefficient vector $\mathbf{x}_k$, for $k \in [N]$, stored column-wise in the matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$. Clearly, any lower bound on the minimax risk for the situation with the additional side information $\mathbf{T}(\mathbf{X})$ is trivially also a lower bound for the case of no side information, since the optimal learning scheme for the latter situation may simply ignore the side information. As we show rigorously in Appendix, we have the upper bound $I(\mathbf{Y}; l) \le I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X}))$, where $I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X}))$ is the conditional mutual information, given the side information $\mathbf{T}(\mathbf{X})$, between the observed data matrix $\mathbf{Y}$ and the random index l. Thus, in order to control the MI $I(\mathbf{Y}; l)$ it is sufficient to control the conditional MI $I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X}))$, which turns out to be a much easier task. We will use two specific choices for $\mathbf{T}(\mathbf{X})$: $\mathbf{T}(\mathbf{X}) = \mathbf{X}$ and $\mathbf{T}(\mathbf{X}) = \text{supp}(\mathbf{X})$. The choice $\mathbf{T}(\mathbf{X}) = \mathbf{X}$ yields tighter bounds for the case of high SNR, while using $\mathbf{T}(\mathbf{X}) = \text{supp}(\mathbf{X})$ results in more accurate bounds in the low SNR regime. In Section IV we show that the problem of upper bounding $I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X}))$ is tractable for both choices.

III. LOWER BOUNDS ON THE MINIMAX RISK

We now state our main results, i.e., lower bounds on the minimax risk of DL. The first bound applies to any distribution of the coefficient vector $\mathbf{x}$, requiring only the existence of the covariance matrix Σ_x. Two further, more specialized, lower bounds assume sparse coefficient vectors and require the underlying dictionary $\mathbf{D}$ in (1) to satisfy a restricted isometry property (RIP) [49].

A. General Coefficients

In this section, we consider the DL problem based on the model (4) with a zero-mean random coefficient vector $\mathbf{x}$. We make no further assumptions on the statistics of $\mathbf{x}$ except that the covariance matrix Σ_x exists. For this setup, the side information $\mathbf{T}(\mathbf{X})$ (cf. our discussion at the end of Section II-C) for the derivation of lower bounds on the minimax risk will be chosen as the coefficients, i.e., $\mathbf{T}(\mathbf{X}) = \mathbf{X}$. Our first main result is the following lower bound on the minimax risk for the DL problem.

Theorem 1: Consider a DL problem based on N i.i.d. observations following the model (4) *and with true dictionary satisfying* (7) *for some $r \le 2\sqrt{p}$. If*

$$p(m - 1) \ge 50, \tag{10}$$

then the minimax risk ε^* is lower bounded as

$$\varepsilon^* \ge \frac{1}{320} \min\left\{r^2, \frac{(p(m-1)/10 - 1)\sigma^2}{N\|\Sigma_x\|_2}\right\}. \tag{11}$$

The first bound in (11), i.e., $\varepsilon^* \ge r^2/320$,[4] complies (up to fixed constants) with the worst case MSE of a dumb estimator $\widehat{\mathbf{D}}$ which ignores the observation $\mathbf{Y}$ and always delivers a fixed dictionary $\mathbf{D}_1 \in \mathcal{X}(\mathbf{D}_0, r)$. Since the true dictionary $\mathbf{D}$ also belongs to the neighborhood $\mathcal{X}(\mathbf{D}_0, r)$, the MSE of this estimator is upper bounded by

$$\|\widehat{\mathbf{D}} - \mathbf{D}\|_F^2 = \|\mathbf{D}_1 - \mathbf{D}\|_F^2$$
$$\le \left(\|\mathbf{D}_1 - \mathbf{D}_0\|_F + \|\mathbf{D}_0 - \mathbf{D}\|_F\right)^2 \overset{(7)}{\le} 4r^2.$$

The second bound in (11) (ignoring constants) is essentially the minimax risk ε' of a simple signal in noise problem

$$\mathbf{z} = \mathbf{s} + \mathbf{n} \tag{12}$$

with AWGN noise, $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \frac{\sigma^2}{\|\Sigma_x\|_2}\mathbf{I}_{p(m-1)})$ and an unknown non-random signal $\mathbf{s}$ of dimension $p(m - 1)$, which is also the dimension of the oblique manifold $\mathcal{D}$ [42]. A standard result in classical estimation theory is that, given the observation of N i.i.d. realizations $\mathbf{z}_k$ of the vector $\mathbf{z}$ in (12), the minimax risk

$$\varepsilon' \triangleq \inf_{\widehat{\mathbf{s}}(\cdot)} \sup_{\mathbf{s} \in \mathbb{R}^{p(m-1)}} \mathrm{E}\left\{\|\widehat{\mathbf{s}}(\mathbf{z}_1, \ldots, \mathbf{z}_N) - \mathbf{s}\|_2^2\right\}$$

of estimating $\mathbf{s} \in \mathbb{R}^{p(m-1)}$ is [44, pp. 403, Exercise 5.8]

$$\varepsilon' = \frac{\sigma^2}{N\|\Sigma_x\|_2} p(m - 1). \tag{13}$$

---

[4]The constant $1/320$ is an artifact of our proof technique and might be improved by a more pedantic analysis.

For fixed ratio $\|\Sigma_x\|_2/\sigma^2$, the bound (11) predicts that $N = \Theta(pm)$ samples are required for accurate DL. Remarkably, this scaling matches the scaling of the sample size found in [36] to be sufficient for successful DL. Note, however, that the analysis [36] is based on the sparse representation error of a dictionary, whereas we target the Frobenius norm of the deviation from the true underlying dictionary.

B. Sparse Coefficients

In this section we focus on a particular subclass of probability distributions for the zero mean coefficient vector $\mathbf{x}$ in (4). More specifically, the random support $\mathrm{supp}(\mathbf{x})$ of the coefficient vector $\mathbf{x}$ is assumed to be distributed uniformly over the set $\Xi \triangleq \{\mathcal{S} \subseteq [p] : |\mathcal{S}| = s\}$, i.e.,

$$P(\mathrm{supp}(\mathbf{x}) = \mathcal{S}) = \frac{1}{|\Xi|} = \frac{1}{\binom{p}{s}}, \text{ for any } \mathcal{S} \in \Xi. \quad (14)$$

We also assume that, conditioned on the support $\mathcal{S} = \mathrm{supp}(\mathbf{x})$, the non-zero entries of $\mathbf{x}$ are i.i.d. with variance σ_a^2, i.e.,

$$E_{\mathbf{x}}\{\mathbf{x}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}^T|\mathcal{S}\} = \sigma_a^2 \mathbf{I}_s. \quad (15)$$

The sparse coefficient support model (14) is useful for performing sparse coding of the observed samples $\mathbf{y}_k$. Indeed, once we have learned the dictionary $\mathbf{D}$, we can estimate for each observed sample $\mathbf{y}_k$, using a standard CS recovery method, the sparse coefficient vector $\mathbf{x}_k$. Sparse source coding is then accomplished by using the sparse coefficient vector to represent the signal $\mathbf{y}_k$. For sparse source coding to be robust against noise, one has to require the underlying dictionary $\mathbf{D}$ to be well conditioned for sparse signals. While there are various ways of quantifying the conditioning of a dictionary, e.g., based on the dictionary coherence [22], [27], we will focus here on the restricted isometry property (RIP) [33], [49], [50]. A dictionary $\mathbf{D}$ is said to satisfy the RIP of order s with constant δ_s if

$$(1 - \delta_s)\|\mathbf{z}\|^2 \leq \|\mathbf{D}\mathbf{z}\|^2 \leq (1 + \delta_s)\|\mathbf{z}\|^2,$$
$$\text{for any } \mathbf{z} \in \mathbb{R}^p \text{ such that } \|\mathbf{z}\|_0 \leq s. \quad (16)$$

While it is intractable to determine if a given matrix $\mathbf{D}$ satisfies the RIP, it turns out that there are random constructions, i.e., the matrix $\mathbf{D}$ is obtained by drawing it from a certain probability distribution, which yield matrices having the RIP with high probability [49]. Moreover, the RIP is stable in the sense that if a matrix $\mathbf{D}$ satisfies the RIP, then so will the perturbed matrix $\mathbf{D} + \Delta$ for a sufficiently small perturbation $\|\Delta\|_F$ (we make use of this fact when proving Theorem 4 below).

Let us formally define the signal-to-noise ratio (SNR) for the observation model (4) as

$$\mathrm{SNR} \triangleq E_{\mathbf{x}}\{\|\mathbf{D}\mathbf{x}\|_2^2\}/E_{\mathbf{n}}\{\|\mathbf{n}\|_2^2\}. \quad (17)$$

Note that the SNR depends on the unknown underlying dictionary $\mathbf{D}$. However, if $\mathbf{D}$ satisfies the RIP (16) with constant δ_s, then

$$\frac{(1 - \delta_s)s\sigma_a^2}{m\sigma^2} \leq \mathrm{SNR} \leq \frac{(1 + \delta_s)s\sigma_a^2}{m\sigma^2}. \quad (18)$$

These bounds depend on $\mathbf{D}$ only via the RIP constant δ_s. For a small constant δ_s, (18) justifies the approximation $\mathrm{SNR} \approx \frac{s\sigma_a^2}{m\sigma^2}$.

As can be verified easily, any random coefficient vector $\mathbf{x}$ conforming with (14) and (15) possesses a finite covariance matrix, given explicitly by

$$\Sigma_x = (s/p)\sigma_a^2 \mathbf{I}_p. \quad (19)$$

Therefore we can invoke Theorem 1, which, combined with (18) and (19), yields the following corollary.

Corollary 2: Consider a DL problem based on N i.i.d. observations according to the model (4) *and with true dictionary satisfying* (7) *for some $r \leq 2\sqrt{p}$. Furthermore, the random coefficient vector $\mathbf{x}$ in* (4) *conforms with* (14) *and* (15). *If the dictionary $\mathbf{D}$ satisfies the RIP* (16) *with RIP-constant $\delta_s \leq 1/2$ and moreover*

$$p(m - 1) \geq 50, \quad (20)$$

then the minimax risk ε^* is lower bounded as

$$\varepsilon^* \geq \frac{1}{320} \min \left\{ r^2, \frac{2p(p(m-1)/10 - 1)}{\mathrm{SNR}Nm} \right\}. \quad (21)$$

For sufficiently large sample size N the second bound in (21) will be in force, leading to a scaling of the minimax risk as $\varepsilon^* = \Theta(p^2/(N\mathrm{SNR}))$. In particular, this bound suggests a decay of the worst case MSE via $1/N$. This agrees with empirical results in [21], indicating that the MSE of popular DL methods typically decay with $1/N$. Moreover, the dependence on the sample size via $1/N$ is theoretically sound, since averaging the outcomes of a learning scheme over N independent observations reduces the estimator variance by $1/N$. Note that, as long as the first bound in (21) is not in force, the overall lower bound (21) scales with $1/\mathrm{SNR}$, which agrees with the basic behavior of the upper bound derived in [21] on the distance of the closest local minimum of (3) to the true dictionary $\mathbf{D}$.

If we consider a fixed SNR (cf. (17)), then our lower bound predicts that for a vanishing minimax risk ε^* the sample size N has to scale as $N = \Theta(p^2)$. This scaling is considerably smaller than the sample size requirement $N = \mathcal{O}(p^3 m)$, which [21] proved to be sufficient in the noisy and over-complete setting, such that minimizing (3) yields an accurate estimate of the true dictionary $\mathbf{D}$. However, for vanishing sparsity rate ($s/p \to 0$), the scaling $N = \Theta(p^2)$ matches the required sample size of the algorithms put forward in [8] and [27], certifying that, for extremely sparse signals, they perform close to the information-theoretic optimum for fixed SNR.

We next derive an alternative lower bound on the minimax risk for DL based on the sparse coefficient model (14) and (15) by additionally assuming the non-zero coefficients to be Gaussian. In particular, let us denote by $\mathbf{P}$ a random matrix which is drawn u.a.r. from the set of all permutation matrices of size $p \times p$. Furthermore, we denote by $\mathbf{z} \in \mathbb{R}^s$ a multivariate normal random vector with zero mean and covariance matrix $\Sigma_z = \sigma_a^2 \mathbf{I}_s$. Based on the matrix $\mathbf{P}$ and vector $\mathbf{z}$, we generate the coefficient vector $\mathbf{x}$ as

$$\mathbf{x} = \mathbf{P}\big(\mathbf{z}^T, \mathbf{0}_{1 \times (p-s)}\big)^T \text{ with } \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \sigma_a^2 \mathbf{I}_s). \quad (22)$$

Theorem 3 below presents a lower bound on the minimax risk for the low SNR regime where SNR $\leq$ $(1/(9\sqrt{80}))m/(2s)$.

Theorem 3: Consider a DL problem based on the model (4) *such that* (7) *holds with some* $r \leq 2\sqrt{p}$ *and the underlying dictionary* **D** *satisfies the RIP of order s with constant* $\delta_s \leq 1/2$ *(cf.* (16)*). We assume the coefficients* **x** *in* (4) *to be distributed according to* (22) *with* SNR $\leq (1/(9\sqrt{80}))m/(2s)$. *If*

$$p(m-1) \geq 50, \qquad (23)$$

then the minimax risk ε^* *is lower bounded as*

$$\varepsilon^* \geq \frac{1}{12960} \min \left\{ r^2/s, \frac{(p(m-1)/10 - 1)p}{\text{SNR}^2 N m^2} \right\}. \qquad (24)$$

The main difference between the bounds (21) and (24) is their dependence on the SNR (cf. (17)). While the bound (21), which applies to arbitrary coefficient statistics and does not exploit the sparse structure of the model (22), depends on the SNR via $1/\text{SNR}$, the bound (24) shows a dependence via $1/\text{SNR}^2$. Thus, in the low SNR regime where SNR $\ll 1$, the bound (24) tends to be tighter, i.e. higher, than (21).

We now show that the dependence of the bound (24) on the SNR via $1/\text{SNR}^2$ agrees with the basic behavior of the constrained Cramér–Rao bound (CCRB) [51]. Indeed, if we assume for simplicity that $p = s = 1$ and the true dictionary (which is now a vector) is $\mathbf{d} = \mathbf{e}_1$, then the CCRB gives [51, Th. 1]

$$\text{E}_\mathbf{Y}\{(\widehat{\mathbf{d}}(\mathbf{Y}) - \mathbf{d})(\widehat{\mathbf{d}}(\mathbf{Y}) - \mathbf{d})^T\} \succeq \frac{1}{\text{SNR}^2 m^2 N}(\mathbf{I} - \mathbf{e}_1 \mathbf{e}_1^T) \qquad (25)$$

for any unbiased estimate $\widehat{\mathbf{d}}(\mathbf{Y})$, for which $\text{E}_\mathbf{Y}\{\widehat{\mathbf{d}}(\mathbf{Y})\} = \mathbf{d}$.[5] Thus, in this simplified setting, the dependence of the minimax bound (24) on the SNR is also reflected by the CCRB.

We emphasize that the bound in Theorem 3 is derived by exploiting the (conditional) Gaussianity of the non-zero entries in the coefficient vector. By contrast, the bounds in Theorem 1 and Corollary 2 do not require the non-zero entries to be Gaussian.

C. A Partial Converse

Given the lower bounds on the minimax risk presented in Sections III-A and III-B it is natural to ask wether these are sharp, i.e., there exist DL schemes whose worst case MSE comes close to the lower bounds. To this end, we consider a simple instance of the DL problem and analyze the MSE of a very basic DL method. As it turns out, in certain regimes, the worst case MSE of this simple DL approach essentially matches the lower bound (21). The particular DL method that we will analyze is described by Algorithm 1.

---

[5]Using the notation of [51], we obtained (25) from [51, Th. 1] by using the matrix $\mathbf{U} = (\mathbf{e}_2, \ldots, \mathbf{e}_m)$ which forms an orthonormal basis for the null space of the gradient mapping $\mathbf{F}(\mathbf{d}) = \frac{\partial f(\mathbf{d})}{\partial \mathbf{d}}$ with the constraint function $f(\mathbf{d}) = \|\mathbf{d}\|_2^2 - 1$. Moreover, for evaluating [51, Th. 1] we used the formula $J_{k,l} = (1/2)\text{Tr}\left\{\mathbf{C}^{-1}(\mathbf{d})\frac{\partial \mathbf{C}(\mathbf{d})}{\partial d_k}\mathbf{C}^{-1}(\mathbf{d})\frac{\partial \mathbf{C}(\mathbf{d})}{\partial d_l}\right\}$ [52] for the elements of the Fisher information matrix, which applies to a Gaussian observation with zero mean and covariance matrix $\mathbf{C}(\mathbf{d})$.

Algorithm 1 A Very Simple DL Method.

Input: data matrix $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N) \in \mathbb{R}^{m \times N}$
Output: learned dictionary $\widehat{\mathbf{D}}(\mathbf{Y}) \in \mathbb{R}^{m \times p}$

1) *Compute an estimate* $\widehat{\mathbf{X}}$ *of the coefficient matrix* $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ *by element-wise thresholding, i.e.,*

$$\widehat{\mathbf{X}} = (\widehat{\mathbf{x}}_1, \ldots, \widehat{\mathbf{x}}_N),$$

$$\text{with } \hat{x}_{k,l} = \begin{cases} 1, & \text{if } y_{k,l} > 0.5 \\ 0, & \text{if } |y_{k,l}| \leq 0.5 \\ -1, & \text{if } y_{k,l} < 0.5. \end{cases} \qquad (26)$$

2) *For each column-index* $j \in [p]$, *define*

$$\widetilde{\mathbf{d}}_j \triangleq \frac{p}{Ns} \sum_{k \in [N]} \hat{x}_{k,j} \mathbf{y}_k. \qquad (27)$$

3) *Output*

$$\widehat{\mathbf{D}}(\mathbf{Y}) \triangleq (\widehat{\mathbf{d}}_1, \ldots, \widehat{\mathbf{d}}_p), \text{ with } \widehat{\mathbf{d}}_l = \mathbf{Q}\widetilde{\mathbf{d}}_l. \qquad (28)$$

Here, $\mathbf{Q}\mathbf{d} \triangleq \text{argmin}_{\mathbf{d}' \in \overline{\mathcal{B}}(1)} \|\mathbf{d}' - \mathbf{d}\|_2$ *denotes the projection of the vector* $\mathbf{d} \in \mathbb{R}^m$ *onto the closed unit ball* $\overline{\mathcal{B}}(1) \triangleq \{\mathbf{d}' \in \mathbb{R}^m : \|\mathbf{d}'\|_2 \leq 1\}$.

Note that the learned dictionary $\widehat{\mathbf{D}}(\mathbf{Y})$ obtained by Algorithm 1 might not have unit-norm columns so that it might not belong to the oblique manifold $\mathcal{D}$. While this is somewhat counter-intuitive, as the true dictionary **D** belongs to $\mathcal{D}$, this fact is not relevant for the derivation of upper bounds on the MSE incurred by $\widehat{\mathbf{D}}(\mathbf{Y})$.

We have the following characterzation of the performance of Algorithm 1.

Theorem 4: Consider a DL problem based on N i.i.d. observations according to the model (4) *with true dictionary satisfying* (7) *with* $\mathbf{D}_0 = \mathbf{I}$ *and some* $r \leq 2\sqrt{p}$. *Furthermore, the random coefficient vector* **x** *in* (4) *conforms with* (14) *and* (15)*. Moreover, the non-zero entries of* **x** *have magnitude equal to one, i.e.,* $\mathbf{x} \in \{-1, 0, 1\}^p$. *If* $r\sqrt{s} \leq 1/10$ *and* $\sigma \leq 0.4$, *then the MSE of Algorithm 1 satisfies*

$$\text{E}_\mathbf{Y}\{\|\widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D}\|_\text{F}^2\} \leq 4(p^2/N)[(1-r)^2/\text{SNR} + 1]$$
$$+ 2p \exp(-pN 0.4^2/(2\sigma^2)), \qquad (29)$$

for any $\mathbf{D} \in \mathcal{X}(\mathbf{D}_0, r)$.

Admittedly, the scope of Theorem 4 is limited. In particular, it applies only to the regime of extremely sparse signals where $s \leq 1/(100r^2)$. However, the point of Theorem 4 is that there exists a non trivial regime where the lower bound (21) is essentially tight, i.e., gives an accurate characterization of the minimax risk. The proof of Theorem 4 is provided at the end of Section IV.

According to Theorem 4, in the low-SNR regime, i.e., where SNR $= o(1)$, and for sufficiently small noise variance, such that $\sigma \leq 0.4$ and

$$p \exp(-pN 0.4^2/(2\sigma^2)) = o((p^2/N)(1-r)^2/\text{SNR}), \qquad (30)$$

the MSE of the DL scheme given by Algorithm 1 scales as

$$
E_{\mathbf{Y}}\{\|\widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D}\|_{\mathrm{F}}^2\} = \mathcal{O}\left(\frac{p^2(1-r)^2}{N\mathrm{SNR}}\right). \tag{31}
$$

We highlight that the scaling of the upper bound (31) essentially matches the scaling of the lower bound (21), certifying that the bound of Corollary 2 is tight in certain regimes.

IV. PROOF OF THE MAIN RESULTS

Before stating the detailed proofs of Theorem 1 and Theorem 3, we present the key ideas and the main ingredients used for their proofs. At their core, the proofs of Theorem 1 and Theorem 3 are based on the construction of a finite set $\mathcal{D}_0 \triangleq \{\mathbf{D}_1, \ldots, \mathbf{D}_L\} \subseteq \mathcal{D}$ (cf. (6)) of L distinct dictionaries having the following desiderata:

- For any two dictionaries $\mathbf{D}_l, \mathbf{D}_{l'} \in \mathcal{D}_0$,

$$
\|\mathbf{D}_l - \mathbf{D}_{l'}\|_{\mathrm{F}}^2 \geq \bar{\delta}_{l,l'} 8\varepsilon. \tag{32}
$$

- If the true dictionary in (4) is chosen as $\mathbf{D} = \mathbf{D}_l \in \mathcal{D}_0$, where l is selected u.a.r. from $[L]$, then the conditional MI between $\mathbf{Y}$ and l, given the side information $\mathbf{T}(\mathbf{X})$,[6] is bounded as

$$
I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X})) \leq \eta \tag{33}
$$

for some small η.

To verify the existence of such a set $\mathcal{D}_0$, we rely on the following result:

Lemma 5: For $P \in \mathbb{N}$ such that

$$
\log(P)/d < (1 - 2/10)^2/4, \tag{34}
$$

there exists a set $\mathcal{P} \triangleq \{\mathbf{b}_l\}_{l \in [P]}$ of P distinct binary vectors $\mathbf{b}_l \in \{-1, 1\}^d$ satisfying

$$
\|\mathbf{b}_l - \mathbf{b}_{l'}\|_0 \geq d/10, \quad \text{for any } l, \ l' \in [P] \text{ with } l \neq l'. \tag{35}
$$

Proof: We construct the set $\mathcal{P}$ sequentially by drawing i.i.d. realizations $\mathbf{b}_l$ from a standard Bernoulli vector $\mathbf{b} \in \{-1, 1\}^d$. Consider two different indices $l, l' \in [P]$. Define the vector $\tilde{\mathbf{b}} \triangleq \mathbf{b}_l \odot \mathbf{b}_{l'}$ by element-wise multiplication and observe that

$$
\|\mathbf{b}_l - \mathbf{b}_{l'}\|_0 = (1/2)\left(p - \sum_{r \in [d]} \tilde{b}_r\right). \tag{36}
$$

Each one of the three vectors $\mathbf{b}_l, \mathbf{b}_{l'}, \tilde{\mathbf{b}} \in \{-1, 1\}^d$ contains zero-mean i.i.d. Bernoulli variables. Therefore,

$$
P\{\|\mathbf{b}_l - \mathbf{b}_{l'}\|_0 \leq d/10\} \overset{(36)}{=} P\{(d - \sum_{r \in [d]} \tilde{b}_r) \leq d/5\}
$$
$$
= P\{\sum_{r \in [d]} \tilde{b}_r \geq (4/5)d\}. \tag{37}
$$

According to Hoeffding's inequality [49, Th. 7.20] (see Lemma 9 in Appendix),

$$
P\{\sum_{r \in [d]} \tilde{b}_r \geq (1 - 2/10)d\} \leq \exp(-d(1 - 2/10)^2/2). \tag{38}
$$

[6]Particular choices for $\mathbf{T}(\mathbf{X})$ are discussed at the end of Section II-C.

Taking a union bound over all $\binom{P}{2}$ pairs $l, l' \in [P]$, we have from (37) and (38) that the probability of P i.i.d. draws $\{\mathbf{b}_l\}_{l \in [P]}$ violating (35) is upper bounded by

$$
P_1 \leq \exp(-d(1 - 2/10)^2/2 + 2\log P), \tag{39}
$$

which is strictly lower than 1 if (34) is valid. Thus, there must exist at least one set $\mathcal{P} = \{\mathbf{b}_l\}_{l \in [P]}$ of cardinality P whose elements satisfy (35). $\square$

The following result gives a sufficient condition on the cardinality L and threshold η such that there exists at least one subset $\mathcal{D}_0 \subseteq \mathcal{D}$ of L distinct dictionaries satisfying (32) and (33).

Lemma 6: Consider a DL problem based on the generative model (4) such that (7) holds with some $r \leq 2\sqrt{p}$. If $(m-1)p \geq 50$, then there exists a set $\mathcal{D}_0 \subseteq \mathcal{D}$ of cardinality $L \geq 2^{(m-1)p/5}$ such that (32) and (33) are satisfied (for the side information $\mathbf{T}(\mathbf{X}) = \mathbf{X}$) with

$$
\eta = 320N\|\Sigma_x\|_2\varepsilon/\sigma^2 \tag{40}
$$

and

$$
\varepsilon \leq r^2/320. \tag{41}
$$

Proof: According to Lemma 5, for $(m-1)p \geq 50$, there is a set of L matrices $\mathbf{D}_{1,l} \in (1/\sqrt{4(m-1)p})\{-1, 1\}^{(m-1)\times p}$, $l \in [L]$ with $L \geq 2^{(m-1)p/5}$, such that

$$
\|\mathbf{D}_{1,l} - \mathbf{D}_{1,l'}\|_{\mathrm{F}}^2 \geq 1/40 \text{ for } l \neq l'. \tag{42}
$$

Since the matrices $\mathbf{D}_{1,l} \in \mathbb{R}^{(m-1)\times p}$, for $l \in [L]$, have entries with values in $(1/\sqrt{4(m-1)p})\{-1, 1\}$ their columns all have norm equal to $1/\sqrt{4p}$.

Based on the matrices $\mathbf{D}_{1,l} \in \mathbb{R}^{(m-1)\times p}$, we now construct a modified set of matrices $\mathbf{D}_{2,l} \in \mathbb{R}^{m\times p}$, $l \in [L]$. Let $\mathbf{U}_j$ denote an arbitrary $m \times m$ unitary matrix satisfying

$$
\mathbf{d}_{0,j} = \mathbf{U}_j\mathbf{e}_1. \tag{43}
$$

Here, $\mathbf{d}_{0,j}$ denotes the jth column of $\mathbf{D}_0 \in \mathbb{R}^{m\times p}$. Then, we define the matrix $\mathbf{D}_{2,l}$ column-wise, by constructing its jth column $\mathbf{d}_{2,l,j}$ as

$$
\mathbf{d}_{2,l,j} = \mathbf{U}_j\begin{pmatrix} 0 \\ \mathbf{d}_{1,l,j} \end{pmatrix}, \tag{44}
$$

where $\mathbf{d}_{1,l,j}$ is the jth column of the matrix $\mathbf{D}_{1,l}$. Note that, for any $l \in [L]$, the jth column $\mathbf{d}_{2,l,j}$ of $\mathbf{D}_{2,l}$ is orthogonal to the column $\mathbf{d}_{0,j}$ and has norm equal to $1/\sqrt{4p}$, i.e.,

$$
\mathrm{diag}\{\mathbf{D}_0^T\mathbf{D}_{2,l}\} = \mathbf{0}, \text{ and}
$$
$$
\mathrm{diag}\{\mathbf{D}_{2,l}^T\mathbf{D}_{2,l}\} = \frac{1}{4p}\mathbf{1} \text{ for any } l \in [L]. \tag{45}
$$

Moreover, for two distinct indices $l, l' \in [L]$, we have

$$
\|\mathbf{D}_{2,l} - \mathbf{D}_{2,l'}\|_{\mathrm{F}}^2 \overset{(44)}{=} \|\mathbf{D}_{1,l} - \mathbf{D}_{1,l'}\|_{\mathrm{F}}^2 \overset{(42)}{\geq} 1/40. \tag{46}
$$

Consider the matrices $\mathbf{D}_l$,

$$
\mathbf{D}_l = \sqrt{1 - \varepsilon'/(4p)}\mathbf{D}_0 + \sqrt{\varepsilon'}\mathbf{D}_{2,l}, \tag{47}
$$

where $l \in [L]$ and

$$
\varepsilon' \triangleq 320\varepsilon. \tag{48}
$$

The construction (47) is feasible, since (41) guarantees $\varepsilon' \le r^2 \le 4p$. We now verify that the matrices $\mathbf{D}_l$, for $l \in [L]$, belong to $\mathcal{X}(\mathbf{D}_0, r)$ and moreover are such that (32) and (33) are satisfied, with η given in (40).

$\mathbf{D}_l$ *belongs to* $\mathcal{X}(\mathbf{D}_0, r)$: Consider the jth columns $\mathbf{d}_{l,j}$, $\mathbf{d}_{0,j}$ and $\mathbf{d}_{2,l,j}$ of $\mathbf{D}_l$, $\mathbf{D}_0$ and $\mathbf{D}_{2,l}$, respectively. Then

$$\|\mathbf{d}_{l,j}\|_2^2 \overset{(45),(47)}{=} (1 - \varepsilon'/(4p))\|\mathbf{d}_{0,j}\|_2^2 + \varepsilon'\|\mathbf{d}_{2,l,j}\|_2^2$$
$$\overset{(45)}{=} (1 - \varepsilon'/(4p)) + (\varepsilon'/(4p)) = 1. \qquad (49)$$

Thus, the columns of any $\mathbf{D}_l$, for $l \in [L]$, have unit norm. Moreover,

$$\|\mathbf{D}_l - \mathbf{D}_0\|_F^2 \overset{(47)}{=} \|(1 - \sqrt{1 - \varepsilon'/(4p)})\mathbf{D}_0 - \sqrt{\varepsilon'}\mathbf{D}_{2,l}\|_F^2$$
$$\overset{(45)}{=} (1 - \sqrt{1 - \varepsilon'/(4p)})^2\|\mathbf{D}_0\|_F^2 + \varepsilon'\|\mathbf{D}_{2,l}\|_F^2$$
$$\overset{(45)}{=} (1 - \sqrt{1 - \varepsilon'/(4p)})^2\|\mathbf{D}_0\|_F^2 + \varepsilon'/4$$
$$\overset{\varepsilon'/(4p)\le 1, \mathbf{D}_0 \in \mathcal{D}}{\le} (\varepsilon'/(4p))^2 p + \varepsilon'/4$$
$$\overset{\varepsilon'/(4p)\le 1}{\le} (1/2)\varepsilon'$$
$$\overset{(41)}{\le} r^2.$$

Lower Bounding $\|\mathbf{D}_l - \mathbf{D}_{l'}\|_F^2$: The squared distance between two different matrices $\mathbf{D}_l$ and $\mathbf{D}_{l'}$ is obtained as

$$\|\mathbf{D}_l - \mathbf{D}_{l'}\|_F^2 \overset{(47)}{=} \varepsilon'\|\mathbf{D}_{2,l} - \mathbf{D}_{2,l'}\|_F^2$$
$$\overset{(46)}{\ge} \varepsilon'/40. \qquad (50)$$

Thus, we have verified

$$\|\mathbf{D}_l - \mathbf{D}_{l'}\|_F^2 \ge \varepsilon'/40 \overset{(48)}{=} 8\varepsilon, \qquad (51)$$

for any two different $l, l' \in [L]$.

Upper Bounding $I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X}))$: We now upper bound the conditional MI $I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X}))$, conditioned on the side information $\mathbf{T}(\mathbf{X}) = \mathbf{X}$, between the observation $\mathbf{Y}$ and the index l of the true dictionary $\mathbf{D} = \mathbf{D}_l \in \mathcal{D}_0$ in (4). Here, the random index l is taken u.a.r. from the set $[L]$. First, note that the dictionaries $\mathbf{D}_l$ given by (47), satisfy

$$\|\mathbf{D}_l - \mathbf{D}_{l'}\|_F^2 \overset{(47)}{=} \varepsilon'\|\mathbf{D}_{2,l} - \mathbf{D}_{2,l'}\|_F^2$$
$$\le \varepsilon'(\|\mathbf{D}_{2,l}\|_F + \|\mathbf{D}_{2,l'}\|_F)^2$$
$$= 4\varepsilon'\|\mathbf{D}_{2,l}\|_F^2$$
$$\overset{(45),(48)}{=} 320\varepsilon. \qquad (52)$$

According to our observation model (4), conditioned on the coefficients $\mathbf{x}_k$, the observations $\mathbf{y}_k$ follow a multivariate Gaussian distribution with covariance matrix $\sigma^2\mathbf{I}$ and mean vector $\mathbf{D}\mathbf{x}_k$. Therefore, we can employ a standard argument based on the convexity of the Kullback-Leibler (KL) divergence (see, e.g., [32]) to upper bound $I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X}))$ as

$$I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X})) \le \frac{1}{L^2} \sum_{l,l' \in [L]} \mathbb{E}_{\mathbf{X}}\{D(f_{\mathbf{D}_l}(\mathbf{Y}|\mathbf{X})\|f_{\mathbf{D}_{l'}}(\mathbf{Y}|\mathbf{X}))\}, \quad (53)$$

where $D(f_{\mathbf{D}_l}(\mathbf{Y}|\mathbf{X})\|f_{\mathbf{D}_{l'}}(\mathbf{Y}|\mathbf{X}))$ denotes the KL divergence between the conditional probability density functions

(given the coefficients $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)$) of the observations $\mathbf{Y}$ for the true dictionary being either $\mathbf{D}_l$ or $\mathbf{D}_{l'}$. Since, given the coefficients $\mathbf{X}$, the observations $\mathbf{y}_k$ are independent multivariate Gaussian random vectors with mean $\mathbf{D}\mathbf{x}_k$ and common covariance matrix $\sigma^2\mathbf{I}_m$, we can apply the formula [53, eq. (3)] for the KL divergence to obtain

$$D(f_{\mathbf{D}_l}(\mathbf{Y}|\mathbf{X})\|f_{\mathbf{D}_{l'}}(\mathbf{Y}|\mathbf{X})) = \sum_{k \in [N]} \frac{1}{2\sigma^2}\|(\mathbf{D}_l - \mathbf{D}_{l'})\mathbf{x}_k\|^2$$
$$= \sum_{k \in [N]} \frac{1}{2\sigma^2} \text{Tr}\{(\mathbf{D}_l - \mathbf{D}_{l'})^T(\mathbf{D}_l - \mathbf{D}_{l'})\mathbf{x}_k\mathbf{x}_k^T\}. \qquad (54)$$

Inserting (54) into (53) and using (52) as well as

$$\text{Tr}\{\mathbf{A}^T\mathbf{A}\Sigma_x\} \le \|\Sigma_x\|_2\|\mathbf{A}\|_F^2,$$

yields

$$I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X})) \le \frac{320N\|\Sigma_x\|_2\varepsilon}{\sigma^2} \qquad (55)$$

completing the proof. □

For the proof of Theorem 3 we will need a variation of Lemma 6, which is based on using the side information $\mathbf{T}(\mathbf{X}) = \text{supp}(\mathbf{X})$ instead of $\mathbf{X}$ itself.

Lemma 7: Consider a DL problem based on the generative model (4) *such that* (7) *holds with some* $r \le 2\sqrt{p}$. *The random sparse coefficients* $\mathbf{x}$ *are distributed according to* (22) *with* SNR $\le (1/(9\sqrt{80}))m/(2s)$. *We assume that the reference dictionary* $\mathbf{D}_0$ *satisfies the RIP of order* s *with constant* $\delta_s \le 1/2$.

If $(m-1)p \ge 50$ *then there exists a set* $\mathcal{D}_0 \subseteq \mathcal{D}$ *of cardinality* $L = 2^{(m-1)p/5}$ *such that* (32) *and* (33) *are satisfied, for the side information* $\mathbf{T}(\mathbf{X}) = \text{supp}(\mathbf{X})$, *with*

$$\eta = 12960N\text{SNR}^2m^2\varepsilon/p, \qquad (56)$$

and

$$\varepsilon \le r^2/(320s). \qquad (57)$$

Proof: We use the same ensemble $\mathcal{D}_0$ (cf. (47)) as in the proof of Lemma 6 (note that condition (57) implies (41) since $s \ge 1$). Thus, we already verified in the proof of Lemma 6 that $\mathcal{D}_0 \subseteq \mathcal{X}(\mathbf{D}_0, r)$ and (32) is satisfied.

Upper bounding $I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X}))$: We upper bound the conditional MI $I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X}))$, conditioned on the side information $\mathbf{T}(\mathbf{X}) = \text{supp}(\mathbf{X})$, between the observation $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_N)$ and the index l of the true dictionary $\mathbf{D} = \mathbf{D}_l \in \mathcal{D}_0$ in (4). Here, the random index l is taken u.a.r. from the set $[L]$ and the conditioning is w.r.t. the random supports $\text{supp}(\mathbf{X}) = (\text{supp}(\mathbf{x}_1), \dots, \text{supp}(\mathbf{x}_N))$ of the coefficient vectors $\mathbf{x}_k$, being i.i.d. realizations of the sparse vector $\mathbf{x}$ given by (22).

Let us introduce the shorthand $\mathcal{S}_k \triangleq \text{supp}(\mathbf{x}_k)$. Note that, conditioned on $\mathcal{S}_k$, the columns of the matrix $\mathbf{Y}$, i.e., the observed samples $\mathbf{y}_k$ are independent multivariate Gaussian random vectors with zero mean and covariance matrix

$$\Sigma_k = \sigma_a^2\mathbf{D}_{\mathcal{S}_k}\mathbf{D}_{\mathcal{S}_k}^T + \sigma^2\mathbf{I}. \qquad (58)$$

Thus, according to [31, eq. (18)], we can use the following bound on the conditional MI

$$I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X}))$$
$$\leq \mathrm{E}_{\mathbf{T}(\mathbf{X})}\Bigg\{ \sum_{\substack{k\in[N]\\l,l'\in[L]}} \frac{1}{L^2}\mathrm{Tr}\big\{\big[\Sigma_{k,l}^{-1}-\Sigma_{k,l'}^{-1}\big]\big[\Sigma_{k,l'}-\Sigma_{k,l}\big]\big\}\Bigg\} \quad (59)$$

with

$$\Sigma_{k,l} \triangleq \sigma_a^2 \mathbf{D}_{l,\mathcal{S}_k}\mathbf{D}_{l,\mathcal{S}_k}^T + \sigma^2\mathbf{I}. \quad (60)$$

Here, $\mathrm{E}_{\mathbf{T}(\mathbf{X})}\{\,\cdot\,\}$ denotes expectation with respect to the side information $\mathbf{T}(\mathbf{X}) = (\mathcal{S}_1,\ldots,\mathcal{S}_N)$ which is distributed uniformly over the N-fold product $\Xi \times \ldots \times \Xi$ (cf. (14)). Since any of the matrices $\Sigma_{k,l}$ is made up of the common component $\sigma^2\mathbf{I}$ and the individual component $\sigma_a^2\mathbf{D}_{l,\mathcal{S}_k}\mathbf{D}_{l,\mathcal{S}_k}^T$, which has rank not larger than s, for any two $l,l' \in [L]$, the difference $\Sigma_{k,l} - \Sigma_{k,l'}$ satisfies

$$\mathrm{rank}\,\big\{\Sigma_{k,l}-\Sigma_{k,l'}\big\} \leq 2s. \quad (61)$$

Therefore, using $\mathrm{Tr}\{\mathbf{A}\} \leq \mathrm{rank}\{\mathbf{A}\}\|\mathbf{A}\|_2$ and (61), we can rewrite (59) as

$$I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X}))$$
$$\leq 2s\mathrm{E}_{\mathbf{T}(\mathbf{X})}\Bigg\{ \sum_{\substack{k\in[N]\\l,l'\in[L]}} \frac{1}{L^2}\big\|\Sigma_{k,l}^{-1}-\Sigma_{k,l'}^{-1}\big\|_2\big\|\Sigma_{k,l'}-\Sigma_{k,l}\big\|_2 \Bigg\}. \quad (62)$$

In what follows, we will first upper bound the spectral norm $\|\Sigma_{k,l'} - \Sigma_{k,l}\|_2$ and subsequently, using a perturbation result [54] for matrix inversion, upper bound the spectral norm $\|\Sigma_{k,l}^{-1} - \Sigma_{k,l'}^{-1}\|_2$. Inserting these two bounds into (62) will then yield the final upper bound on $I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X}))$.

Due to the construction (47),

$$\Sigma_{k,l}-\Sigma_{k,l'} \overset{(60)}{=} \sigma_a^2\big(\mathbf{D}_{l,\mathcal{S}_k}\mathbf{D}_{l,\mathcal{S}_k}^T - \mathbf{D}_{l',\mathcal{S}_k}\mathbf{D}_{l',\mathcal{S}_k}^T\big)$$
$$\overset{(47)}{=} \sigma_a^2\sqrt{1-\varepsilon'/4p}\sqrt{\varepsilon'}\,\big(\overline{\mathbf{D}_{0,\mathcal{S}_k}\mathbf{D}_{2,l,\mathcal{S}_k}^T} - \overline{\mathbf{D}_{0,\mathcal{S}_k}\mathbf{D}_{2,l',\mathcal{S}_k}^T}\big)$$
$$+ \sigma_a^2\varepsilon'\big(\mathbf{D}_{2,l,\mathcal{S}_k}\mathbf{D}_{2,l,\mathcal{S}_k}^T - \mathbf{D}_{2,l',\mathcal{S}_k}\mathbf{D}_{2,l',\mathcal{S}_k}^T\big) \quad (63)$$

with the shorthand $\overline{\mathbf{X}} \triangleq \mathbf{X} + \mathbf{X}^T$. In what follows, we need

$$\begin{aligned}\|\mathbf{D}_{0,\mathcal{S}}\|_2 &\leq \sqrt{3/2}, \\ \|\mathbf{D}_{2,l,\mathcal{S}}\|_2 &\leq \sqrt{s/(4p)}, \text{ and} \\ \|\Sigma_{k,l}^{-1}\|_2 &\leq 1/\sigma^2, \end{aligned} \quad (64)$$

for any $l \in [L]$ and any subset $\mathcal{S} \subset [p]$ with $|\mathcal{S}| \leq s$. The first bound in (64) follows from the assumed RIP (with constant $\delta_s \leq 1/2$) of the reference dictionary $\mathbf{D}_0$. The second bound in (64) is valid because the matrices $\mathbf{D}_{2,l}$ have columns with norm equal to $1/\sqrt{4p}$ (cf. (45)). For the verification of the last bound in (64) we note that, according to (60), $\lambda_{\min}(\Sigma_{k,l}) \geq \sigma^2$. Therefore,

$$\begin{aligned}\|\Sigma_{k,l}-\Sigma_{k,l'}\|_2 &\overset{(63),(64)}{\leq} \sqrt{6\varepsilon' s/(4p)}\sigma_a^2\sqrt{1-\varepsilon'/(4p)} \\ &\qquad + 2\sigma_a^2\varepsilon' s/(4p) \\ &\overset{(57)}{\leq} 4.5\sigma_a^2\sqrt{\varepsilon' s/(4p)}. \end{aligned} \quad (65)$$

Since the true dictionary $\mathbf{D}$ is assumed to satisfy the RIP with constant $\delta_s \leq 1/2$, the low SNR condition $\mathrm{SNR} \leq m/(2s)$ implies via (18),

$$\frac{\sigma_a^2}{\sigma^2} \leq \frac{1}{9\sqrt{80}}. \quad (66)$$

Since

$$\begin{aligned}\big\|\Sigma_{k,l}^{-1}\big(\Sigma_{k,l}-\Sigma_{k,l'}\big)\big\|_2 &\leq \big\|\Sigma_{k,l}^{-1}\big\|_2\big\|\Sigma_{k,l}-\Sigma_{k,l'}\big\|_2 \\ &\overset{(64),(65)}{\leq} 4.5(\sigma_a/\sigma)^2\sqrt{\varepsilon' s/p} \\ &\overset{(66),(57)}{\leq} 1/2, \end{aligned} \quad (67)$$

we can invoke [54, Th. 2.3.4.] yielding

$$\begin{aligned}\big\|\Sigma_{k,l}^{-1}-\Sigma_{k,l'}^{-1}\big\|_2 &\leq 2\big\|\Sigma_{k,l}^{-1}\big\|_2^2\big\|\Sigma_{k,l}-\Sigma_{k,l'}\big\|_2 \\ &\overset{(64)}{\leq} 2\sigma^{-4}\big\|\Sigma_{k,l'}-\Sigma_{k,l}\big\|_2. \end{aligned} \quad (68)$$

Inserting (65) and (68) into (62) yields the bound

$$\begin{aligned}I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X})) &\leq 4Ns\sigma^{-4}(1/L^2)\sum_{l,l\in[L]}\big\|\Sigma_{k,l'}-\Sigma_{k,l}\big\|_2^2 \\ &\overset{(65)}{\leq} 4\cdot4.5^2Ns^2(\sigma_a/\sigma)^4\varepsilon'/(4p) \\ &\overset{\varepsilon'=320\varepsilon}{\leq} 6480Ns^2(\sigma_a/\sigma)^4\varepsilon/p \\ &\overset{\delta_s\leq1/2,(18)}{\leq} 12960N\mathrm{SNR}^2m^2\varepsilon/p, \end{aligned} \quad (69)$$

completing the proof. $\qquad\qquad\square$

The next result relates the cardinality L of a subset $\mathcal{D}_0 = \{\mathbf{D}_1,\ldots,\mathbf{D}_L\} \subseteq \mathcal{D}$ to the conditional MI $I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X}))$ between the observation $\mathbf{Y} = (\mathbf{y}_1,\ldots,\mathbf{y}_N)$, with $\mathbf{y}_k$ i.i.d. according to (4), and a random index l selecting the true dictionary $\mathbf{D}$ in (4) u.a.r. from $\mathcal{D}_0$.

Lemma 8: Consider the DL problem (4) *with minimax risk ε^* (cf.* (8)*), which is assumed to be upper bounded by a positive number ε, i.e., $\varepsilon^* \leq \varepsilon$. Assume there exisits a finite set $\mathcal{D}_0 = \{\mathbf{D}_1,\ldots,\mathbf{D}_L\} \subseteq \mathcal{D}$ consisting of L distinct dictionaries $\mathbf{D}_l \in \mathbb{R}^{m\times p}$ such that*

$$\|\mathbf{D}_l - \mathbf{D}_{l'}\|_F^2 \geq 8\bar{\delta}_{l,l'}\varepsilon. \quad (70)$$

Then, for any function $\mathbf{T}(\mathbf{X})$ of the true coefficients $\mathbf{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_N)$,

$$I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X})) \geq (1/2)\log_2(L) - 1. \quad (71)$$

Proof: Our proof idea closely follows that of [33, Th. 1]. Consider a minimax estimator $\widehat{\mathbf{D}}(\mathbf{Y})$, whose worst case MSE is equal to ε^*, i.e.,

$$\sup_{\mathbf{D}\in\mathcal{D}} \mathrm{E}_{\mathbf{Y}}\big\{\|\widehat{\mathbf{D}}(\mathbf{Y})-\mathbf{D}\|_F^2\big\} = \varepsilon^*, \quad (72)$$

and, in turn since $\mathcal{D}_0 \subseteq \mathcal{D}$,

$$\sup_{\mathbf{D}\in\mathcal{D}_0} \mathrm{E}_{\mathbf{Y}}\big\{\|\widehat{\mathbf{D}}(\mathbf{Y})-\mathbf{D}\|_F^2\big\} \leq \varepsilon^*. \quad (73)$$

Based on the estimator $\widehat{\mathbf{D}}(\mathbf{Y})$, we define a detector $\hat{l}(\mathbf{Y})$ for the index of true underlying dictionary $\mathbf{D}_l \in \mathcal{D}_0$ via

$$\hat{l}(\mathbf{Y}) \triangleq \operatorname*{argmin}_{l'\in[L]} \|\mathbf{D}_{l'}-\widehat{\mathbf{D}}(\mathbf{Y})\|_F^2. \quad (74)$$

In case of ties, i.e., when there are multiple indices l' such that $\mathbf{D}_{l'}$ achieves the minimum in (74), we randomly select one of the minimizing indices as the estimate $\hat{l}(\mathbf{Y})$. Let us now assume that the index l is selected u.a.r. from $[L]$ and bound the probability P_e of a detection error, i.e., $P_e \triangleq \mathrm{P}\{\hat{l}(\mathbf{Y}) \neq l\}$. Note that if

$$\|\widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D}_l\|_{\mathrm{F}}^2 < 2\varepsilon \tag{75}$$

then for any wrong index $l' \in [L] \setminus \{l\}$,

$$
\begin{aligned}
\|\widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D}_{l'}\|_{\mathrm{F}} &= \|\widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D}_{l'} + \mathbf{D}_l - \mathbf{D}_l\|_{\mathrm{F}} \\
&\geq \|\mathbf{D}_l - \mathbf{D}_{l'}\|_{\mathrm{F}} - \|\widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D}_l\|_F \\
&\overset{(70),(75)}{\geq} (\sqrt{8} - \sqrt{2})\sqrt{\varepsilon} \\
&= \sqrt{2\varepsilon} \\
&\overset{(75)}{>} \|\widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D}_l\|_{\mathrm{F}}.
\end{aligned}
\tag{76}
$$

Thus, the condition (75) guarantees that the detector $\hat{l}(\mathbf{Y})$ in (74) delivers the correct index l. Therefore, in turn, a detection error can only occur if $\|\widehat{\mathbf{D}} - \mathbf{D}_l\|_{\mathrm{F}}^2 \geq 2\varepsilon$ implying that

$$
\begin{aligned}
P_e &\leq \mathrm{P}\{\|\widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D}_l\|_{\mathrm{F}}^2 \geq 2\varepsilon\} \\
&\overset{(a)}{\leq} \frac{1}{2\varepsilon}\mathrm{E}_{\mathbf{Y}}\{\|\widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D}_l\|_{\mathrm{F}}^2\} \\
&\overset{(73)}{\leq} \frac{\varepsilon^*}{2\varepsilon} \\
&\overset{\varepsilon^* \leq \varepsilon}{\leq} 1/2,
\end{aligned}
\tag{77}
$$

where (a) is due to the Markov inequality [55]. However, according to Lemma 10 (see Appendix), we also have

$$I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X})) \geq \log_2(L) - P_e\log_2(L) - 1, \tag{78}$$

and, in turn, since $P_e \leq 1/2$ by (77),

$$I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X})) \geq (1/2)\log_2(L) - 1,$$

completing the proof. $\square$

Finally, we simply have to put the pieces together to obtain Theorem 1 and Theorem 3.

Proof of Theorem 1: According to Lemma 6, if $(m-1)p \geq 50$ and for any $\varepsilon < r^2/320$ (this condition is implied by the first bound in (11)), then there exists a set $\mathcal{D}_0 \subseteq \mathcal{X}(\mathbf{D}_0, r)$ of cardinality $L = 2^{(m-1)p/5}$ satisfying (32) and (33) with $\eta = 320N\|\Sigma_x\|_2\varepsilon/\sigma^2$. Applying Lemma 8 to the set $\mathcal{D}_0$ yields,

$$320N\|\Sigma_x\|_2\varepsilon/\sigma^2 \geq I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X})) \geq (1/2)\log_2(L) - 1 \tag{79}$$

implying

$$
\begin{aligned}
\varepsilon &\geq \frac{\sigma^2}{320N\|\Sigma_x\|_2}((1/2)\log_2(L) - 1) \\
&\geq \frac{\sigma^2}{320N\|\Sigma_x\|_2}((m-1)p/10 - 1).
\end{aligned}
\tag{80}
$$

Proof of Theorem 3: According to Lemma 7, if $(m-1)p \geq 50$ and for any $\varepsilon < r^2/(320s)$ (this condition is implied by the first bound in (24)), then there

exists a set $\mathcal{D}_0 \subseteq \mathcal{X}(\mathbf{D}_0, r)$ of cardinality $L = 2^{(m-1)p/5}$ satisfying (32) and (33) with $\eta = 12960Nm^2\mathrm{SNR}^2\varepsilon/p$. Applying Lemma 8 to the set $\mathcal{D}_0$ yields,

$$
\begin{aligned}
12960Nm^2\mathrm{SNR}^2\varepsilon/p &\geq I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X})) \\
&\geq (1/2)\log_2(L) - 1
\end{aligned}
\tag{81}
$$

so that

$$
\begin{aligned}
\varepsilon &\geq \frac{\mathrm{SNR}^{-2}p}{12960Nm^2}((1/2)\log_2(L) - 1) \\
&\geq \frac{\mathrm{SNR}^{-2}p}{12960Nm^2}((m-1)p/10 - 1).
\end{aligned}
\tag{82}
$$

Proof of Theorem 4: First note that an arbitrary dictionary $\mathbf{D} \in \mathcal{X}(\mathbf{D}_0 = \mathbf{I}, r)$ can be written as

$$\mathbf{D} = \mathbf{I} + \Delta, \quad \text{with}\|\Delta\|_{\mathrm{F}} \leq r. \tag{83}$$

Any matrix $\mathbf{D}$ of the form (83) satisfies the RIP with constant δ_s such that

$$(1 - r)^2 \leq 1 - \delta_s \leq 1 + \delta_s \leq (1 + r)^2. \tag{84}$$

Moreover, since we assume the coefficient vectors $\mathbf{x}_k$ in (1) to be discrete-valued $\mathbf{x}_k \in \{-1, 0, 1\}^p$ and complying with (14),

$$\mathrm{E}_{\mathbf{x}_k}\{\mathbf{x}_{k,t}^2\} = s/p, \tag{85}$$

and

$$\|\mathbf{x}_k\|_2^2 = s. \tag{86}$$

For (86), we used the fact that the non-zero entries of $\mathbf{x}_k$ all have the same magnitude equal to one. Combining (86) with (84), we obtain the following bound on the SNR:

$$
\begin{aligned}
\mathrm{SNR} &= \mathrm{E}_{\mathbf{x}}\{\|\mathbf{D}\mathbf{x}\|_2^2\}/\mathrm{E}_{\mathbf{n}}\{\|\mathbf{n}\|_2^2\} \\
&\overset{(16)}{\geq} (1 - \delta_s)s/(m\sigma^2) \\
&\overset{(84)}{\geq} (1 - r)^2 s/(m\sigma^2).
\end{aligned}
\tag{87}
$$

In order to derive an upper bound on the MSE of the DL scheme given by Algorithm 1, we first split the MSE of $\widehat{\mathbf{D}}(\mathbf{Y}) = (\widehat{\mathbf{d}}_1(\mathbf{Y}), \ldots, \widehat{\mathbf{d}}_p(\mathbf{Y}))$ into a sum of the MSE for the individual columns of the dictionary, i.e.,

$$\mathrm{E}_{\mathbf{Y}}\{\|\widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D}\|_{\mathrm{F}}^2\} = \sum_{l \in [p]} \mathrm{E}_{\mathbf{Y}}\{\|\widehat{\mathbf{d}}_l(\mathbf{Y}) - \mathbf{d}_l\|_2^2\}. \tag{88}$$

Thus, we may analyze the column-wise MSE $\mathrm{E}_{\mathbf{Y}}\{\|\widehat{\mathbf{d}}_l(\mathbf{Y}) - \mathbf{d}_l\|_2^2\}$ separately for each index $l \in [p]$. Note that, by construction

$$\|\widehat{\mathbf{d}}_l(\mathbf{Y}) - \mathbf{d}_l\|_2^2 \leq 2, \tag{89}$$

since the columns of $\widehat{\mathbf{D}}(\mathbf{Y})$ and $\mathbf{D}$ have norm at most one.

We analyze the MSE of Algorithm 1 by conditioning on a specific event $\mathcal{C}$, defined as

$$\mathcal{C} \triangleq \bigcap_{\substack{k \in [N] \\ l \in [p]}} \{|n_{k,l}| < 0.4\}. \tag{90}$$

Assuming $r\sqrt{s} \leq 1/10$, the occurrence of $\mathcal{C}$ implies the estimated coefficient matrix $\widehat{\mathbf{X}} = (\widehat{\mathbf{x}}_1, \ldots, \widehat{\mathbf{x}}_N)$ (cf. (26)) to coincide with the true coefficients $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$, i.e.,

$$P\{\mathbf{X} = \widehat{\mathbf{X}}|\mathcal{C}\} = 1. \tag{91}$$

Indeed, if $r\sqrt{s} \leq 1/10$ and $|n_{k,l}| < 0.4$ for every $k \in [N]$ and $l \in [p]$, then $y_{k,j} > 0.5$ if $x_{k,j} = 1$ (implying $\hat{x}_{k,j} = 1$), and $y_{k,j} < -0.5$ if $x_{k,j} = -1$ (implying $\hat{x}_{k,j} = -1$) as well as $|y_{k,j}| \leq 0.5$ if $x_{k,j} = 0$ (implying $\hat{x}_{k,j} = 0$). Characterising the probability of the event $\mathcal{C}$ is straightforward, since the noise entries $n_{k,l}$ are assumed i.i.d. Gaussian variables with zero mean and variance σ^2. In particular, the tail bound [49, Proposition 7.5]) together with a union bound over all entries of the coefficient matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N) \in \mathbb{R}^{p \times N}$, yields

$$P\{\mathcal{C}^c\} \leq \exp(-pN0.4^2/(2\sigma^2)). \tag{92}$$

As a next step we upper bound the MSE using the law of total expectation:

$$
\begin{aligned}
E_{\mathbf{Y}}\{\|\widehat{\mathbf{d}}_l(\mathbf{Y}) - \mathbf{d}_l\|_2^2\} &= E_{\mathbf{Y},\mathbf{N}}\{\|\widehat{\mathbf{d}}_l(\mathbf{Y}) - \mathbf{d}_l\|_2^2|\mathcal{C}\}P(\mathcal{C}) \\
&\quad + E_{\mathbf{Y},\mathbf{N}}\{\|\widehat{\mathbf{d}}_l(\mathbf{Y}) - \mathbf{d}_l\|_2^2|\mathcal{C}^c\}P(\mathcal{C}^c) \\
&\overset{(89)}{\leq} E_{\mathbf{Y},\mathbf{N}}\{\|\widehat{\mathbf{d}}_l(\mathbf{Y}) - \mathbf{d}_l\|_2^2|\mathcal{C}\}P(\mathcal{C}) + 2P(\mathcal{C}^c) \\
&\leq E_{\mathbf{Y},\mathbf{N}}\{\|\widehat{\mathbf{d}}_l(\mathbf{Y}) - \mathbf{d}_l\|_2^2|\mathcal{C}\} \\
&\quad + 2\exp(-pN0.4^2/(2\sigma^2)). \tag{93}
\end{aligned}
$$

The conditional MSE $E\{\|\widehat{\mathbf{d}}_l(\mathbf{Y}) - \mathbf{d}_l\|_2^2|\mathcal{C}\}$ can then be bounded by

$$
\begin{aligned}
&E_{\mathbf{Y},\mathbf{N}}\{\|\widehat{\mathbf{d}}_l(\mathbf{Y}) - \mathbf{d}_l\|_2^2|\mathcal{C}\} \\
&= E_{\mathbf{Y},\mathbf{N}}\{\|\mathbf{Q}\widetilde{\mathbf{d}}_l(\mathbf{Y}) - \mathbf{d}_l\|_2^2|\mathcal{C}\} \\
&\overset{(a)}{\leq} E_{\mathbf{Y},\mathbf{N}}\{\|\widetilde{\mathbf{d}}_l(\mathbf{Y}) - \mathbf{d}_l\|_2^2|\mathcal{C}\} \\
&= E_{\mathbf{Y},\mathbf{N}}\{\|\frac{p}{Ns}\sum_{k\in[N]}\hat{x}_{k,l}\mathbf{y}_k - \mathbf{d}_l\|_2^2|\mathcal{C}\} \\
&\overset{(1)}{=} E_{\mathbf{Y},\mathbf{X},\mathbf{N}}\{\|\frac{p}{Ns}\sum_{k\in\mathcal{C}_l}\hat{x}_{k,l}(\mathbf{D}\mathbf{x}_k + \mathbf{n}_k) - \mathbf{d}_l\|_2^2|\mathcal{C}\} \\
&\overset{(b)}{=} E_{\mathbf{X},\mathbf{N}}\{\|\frac{p}{Ns}\sum_{k\in\mathcal{C}_l}x_{k,l}(\mathbf{D}\mathbf{x}_k + \mathbf{n}_k) - \mathbf{d}_l\|_2^2|\mathcal{C}\} \tag{94}
\end{aligned}
$$

where steps (a) and (b) are valid because of $\mathbf{d}_l \in \overline{\mathcal{B}}(1)$ (cf. (7), (6)) and $P(x_{k,l} = \hat{x}_{k,l}|\mathcal{C}) = 1$ (cf. (91)), respectively. Applying the inequality $\|\mathbf{y} + \mathbf{z}\|_2^2 \leq 2(\|\mathbf{y}\|_2^2 + \|\mathbf{z}\|_2^2)$ to (94) yields

$$
\begin{aligned}
&E_{\mathbf{Y},\mathbf{N}}\{\|\widehat{\mathbf{d}}_l(\mathbf{Y}) - \mathbf{d}_l\|_2^2|\mathcal{C}\} \\
&\leq 2E_{\mathbf{X},\mathbf{N}}\{\|\frac{p}{Ns}\sum_{k\in[N]}x_{k,l}\mathbf{n}_k\|_2^2|\mathcal{C}\} \\
&\quad + 2E_{\mathbf{X},\mathbf{N}}\{\|\mathbf{d}_l - \frac{p}{Ns}\sum_{\substack{k\in[N]\\t\in[p]}}x_{k,l}x_{k,t}\mathbf{d}_t\|_2^2|\mathcal{C}\}. \tag{95}
\end{aligned}
$$

Our strategy will be to separately bound the two expectations in (95) from above.

In order to upper bound $E_{\mathbf{X},\mathbf{N}}\{\|\frac{p}{Ns}\sum_{k\in[N]}x_{k,l}\mathbf{n}_k\|_2^2|\mathcal{C}\}$, we note that the conditional distribution $f(n_{k,t}|\mathcal{C})$ of $n_{k,t}$, given the event $\mathcal{C}$, is given by

$$f(n_{k,t}|\mathcal{C}) = \frac{\mathcal{I}_{[-0.4,0.4]}(n_{k,t}) \cdot e^{-\frac{n_{k,t}^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}(\mathcal{Q}(-0.4/\sigma) - \mathcal{Q}(0.4/\sigma))}, \tag{96}$$

where $\mathcal{I}_{[-0.4,0.4]}(\cdot)$ is the indicator function for the interval $[-0.4, 0.4]$ and $\mathcal{Q}(x) \triangleq \int_{z=x}^{\infty}(1/\sqrt{2\pi})\exp(-(1/2)z^2)dz$ denotes the tail probability of the standard normal distribution. In particular, the conditional variance $\sigma_{n_{k,t}}^2$ can be bounded as

$$\sigma_{n_{k,t}}^2 \leq \sigma^2/\underbrace{(\mathcal{Q}(-0.4/\sigma) - \mathcal{Q}(0.4/\sigma))}_{\triangleq v}. \tag{97}$$

Since, conditioned on $\mathcal{C}$, the variables $\hat{x}_{k,l}$ and $n_{k,t}$ are independent, we obtain

$$
\begin{aligned}
&E_{\mathbf{X},\mathbf{N}}\{\|\frac{p}{Ns}\sum_{k\in[N]}x_{k,l}\mathbf{n}_k\|_2^2|\mathcal{C}\} \\
&= \left(\frac{p}{Ns}\right)^2\sum_{\substack{k\in[N]\\t\in[m]}}E_{\mathbf{X},\mathbf{N}}\{x_{k,l}^2|\mathcal{C}\}E_{\mathbf{X},\mathbf{N}}\{n_{k,l}^2|\mathcal{C}\} \\
&\overset{(97)}{\leq} \left(\frac{p}{Ns}\right)^2 NE_{\mathbf{X},\mathbf{N}}\{x_{k,l}^2|\mathcal{C}\}m\sigma^2/v \\
&\overset{(a)}{=} \left(\frac{p}{Ns}\right)^2 NE_{\mathbf{X}}\{x_{k,l}^2\}m\sigma^2/v \\
&\overset{(85)}{=} \left(\frac{p}{Ns}\right)^2 N(s/p)m\sigma^2/v \\
&\overset{(87)}{\geq} (p/N)(1-r)^2/(v\text{SNR}), \tag{98}
\end{aligned}
$$

where step (a) is due to the fact that $x_{k,l}^2$ is independent of the event $\mathcal{C}$.

As to the second expectation in (95), we first observe that

$$
\begin{aligned}
&E_{\mathbf{X},\mathbf{N}}\{\|\mathbf{d}_l - \frac{p}{Ns}\sum_{\substack{k\in[N]\\t\in[p]}}x_{k,l}x_{k,t}\mathbf{d}_t\|_2^2|\mathcal{C}\} \\
&= E_{\mathbf{X}}\{\|\mathbf{d}_l - \frac{p}{Ns}\sum_{\substack{k\in[N]\\t\in[p]}}x_{k,l}x_{k,t}\mathbf{d}_t\|_2^2\} \tag{99}
\end{aligned}
$$

since the coefficients $x_{k,t}$ are independent of the event $\mathcal{C}$. Next, we expand the squared norm und apply the relations

$$
E_{\mathbf{X}}\{x_{k,l}x_{k,t}x_{k',l}x_{k,t'}\} =
\begin{cases}
(s/p)^2, & \text{for } k' = k, \text{ and } t = t' \neq l \\
(s/p)^2, & \text{for } k' \neq k, \text{ and } t = t' = l \\
(s/p), & \text{for } k' = k, \text{ and } t = t' = l \\
0 & \text{else.}
\end{cases} \tag{100}
$$

A somewhat lengthy calculation reveals that

$$E_{\mathbf{X}}\{\|\mathbf{d}_l - \frac{p}{Ns}\sum_{\substack{k\in[N]\\t\in[p]}}x_{k,l}x_{k,t}\mathbf{d}_t\|_2^2\} = \frac{p + p/s - 2}{N}$$

$$\leq 2p/N. \tag{101}$$

Inserting (101) into (99) yields

$$\mathrm{E}_{\mathbf{X},\mathbf{N}}\Big\{\|\mathbf{d}_l - \frac{p}{Ns}\sum_{\substack{k\in[N]\\t\in[p]}} x_{k,l}x_{k,t}\mathbf{d}_t\|_2^2\big|\mathcal{C}\Big\} \le 2p/N. \quad (102)$$

Combining (102) and (98) with (95) and inserting into (93), we finally obtain

$$\mathrm{E}_{\mathbf{Y}}\{\|\widehat{\mathbf{d}}_l(\mathbf{Y}) - \mathbf{d}_l\|_2^2\} \le 2\frac{p}{N}\left[\frac{(1-r)^2}{\nu\mathrm{SNR}} + 2\right]$$
$$+ 2\exp(-pN0.4^2/(2\sigma^2)), \quad (103)$$

and in turn, by summing over all column indices $l \in [p]$ (cf. (88)),

$$\mathrm{E}_{\mathbf{Y}}\{\|\widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D}\|_{\mathrm{F}}^2\} \le 2\frac{p^2}{N}\left[\frac{(1-r)^2}{\nu\mathrm{SNR}} + 2\right]$$
$$+ 2p\exp(-pN0.4^2/(2\sigma^2)). \quad (104)$$

The upper bound (29) follows by noting that $\nu = \mathcal{Q}(-0.4/\sigma) - \mathcal{Q}(0.4/\sigma) \ge 1/2$ for $\sigma \le 0.4$.

V. CONCLUSION

By adapting an established information-theoretic approach to minimax estimation, we derived lower bounds on the minimax risk of DL using certain random coefficient models for representing the observations as linear combinations of the columns of an underlying dictionary matrix. These lower bounds on the optimum achievable performance, quantified in terms of worst case MSE, seem to be the first results of their kind for DL. Our first bound applies to a wide range of coefficient distributions, and only requires the existence of the covariance matrix of the coefficient vector. We then specialized this bound to a sparse coefficient model with normally distributed non-zero coefficients. Exploiting the specific structure induced by the sparse coefficient model, we derived another lower bound which tends to be tighter in the low SNR regime. Our bounds apply to the practically relevant case of overcomplete dictionaries and noisy measurements. An analysis of a simple DL scheme for the low SNR regime, reveals that our lower bounds are tight, as they are attained by the worst case MSE of a particular DL algorithm. Moreover, for fixed SNR and vanishing sparsity rate, the necessary scaling $N = \Theta(p^2)$ of the sample size N implied by our lower bound matches the sufficient condition (upper bound) on the sample size such that the learning methods proposed in [8] and [27] are successful. Hence, in certain regimes, the DL techniques suggested in [8] and [27] are essentially optimal in terms of sample size requirements.

APPENDIX
TECHNICALITIES

The derivation of our main results makes use of Hoeffding's inequality, a powerful result in large deviation analysis, which we state here for the readers convenience.

Lemma 9 [49, Th. 7.20]: Let x_r, $r \in [k]$, be a sequence of i.i.d. zero mean, bounded random variables, i.e., $|x_r| \le a$ for some constant a. Then,

$$\mathrm{P}\Big\{\sum_{r\in[k]} x_r \ge t\Big\} \le \exp\Big(-\frac{t^2}{2ka^2}\Big). \quad (105)$$

The next result is a specialisation of Fano's inequality of information theory to our setting.

Lemma 10: Consider the DL problem based on observing the data matrix $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$ with columns being i.i.d. realizations of the vector $\mathbf{y}$ in (4). We stack the corresponding realizations $\mathbf{x}_k$ of the coefficient vector $\mathbf{x}$ into the matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$. The true dictionary in (4) is obtained by selecting u.a.r., and statistically independent of the random coefficients $\mathbf{x}_k$, an element of the set $\mathcal{D}_0 = \{\mathbf{D}_1, \ldots, \mathbf{D}_L\}$, i.e., $\mathbf{D} = \mathbf{D}_l$ where the index $l \in [L]$ is drawn u.a.r. from $[L]$. Let $\mathbf{T}(\mathbf{X})$ denote an arbitrary function of the coefficients. Then, the error probability $\mathrm{P}\{\hat{l}(\mathbf{Y}) \ne l\}$ of any detector $\hat{l}(\mathbf{Y})$ which is based on observing $\mathbf{Y}$ is lower bounded as

$$\mathrm{P}\{\hat{l}(\mathbf{Y}) \ne l\} \ge 1 - \frac{I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X})) + 1}{\log_2(L)}. \quad (106)$$

Here, $I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X}))$ denotes the conditional MI between $\mathbf{Y}$ and l given the side information $\mathbf{T}(\mathbf{X})$.

Proof: According to Fano's inequality [46, p. 38],

$$\mathrm{P}\{\hat{l}(\mathbf{Y}) \ne l\} \ge \frac{H(l|\mathbf{Y}) - 1}{\log_2(L)}. \quad (107)$$

Combining this with the identity [46, p. 21]

$$I(l; \mathbf{Y}) = H(l) - H(l|\mathbf{Y}), \quad (108)$$

and the fact that $H(l) = \log_2(L)$, since l is distributed uniformly over $[L]$, yields

$$\mathrm{P}\{\hat{l}(\mathbf{Y}) \ne l\} \ge 1 - \frac{I(l; \mathbf{Y}) + 1}{\log_2(L)}. \quad (109)$$

By the chain rule of MI [46, Ch. 2]

$$I(\mathbf{Y}; l) = I(\mathbf{Y}, \mathbf{T}(\mathbf{X}); l) - I(l; \mathbf{T}(\mathbf{X})|\mathbf{Y})$$
$$= I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X})) + \underbrace{I(l; \mathbf{T}(\mathbf{X}))}_{=0} - I(l; \mathbf{T}(\mathbf{X})|\mathbf{Y})$$
$$= I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X})) - I(l; \mathbf{T}(\mathbf{X})|\mathbf{Y}). \quad (110)$$

Here, we used $I(l; \mathbf{T}(\mathbf{X})) = 0$, since the coefficients $\mathbf{X}$ and the index l are independent. Since $I(l; \mathbf{T}(\mathbf{X})|\mathbf{Y}) \ge 0$ [46, Ch. 2], we have from (110) that $I(\mathbf{Y}; l) \le I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X}))$. Thus,

$$\mathrm{P}\{\hat{l}(\mathbf{Y}) \ne l\} \overset{(109),(110)}{\ge} 1 - \frac{I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X})) + 1}{\log_2(L)}, \quad (111)$$

completing the proof. $\square$

REFERENCES

[1] "The zettabyte era—Trends and analysis," (white paper) CISCO, San Jose, CA, USA, Tech. Rep., May 2015.

[2] *The Data Deluge*, Economist, New York, NY, USA, Feb. 2010.

[3] *A Special Report on Managing Information: Data, Data Everywhere*, Economist, New York, NY, USA, Feb. 2010.

[4] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[5] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[6] Y. C. Eldar and G. Kutyniok, *Compressed Sensing: Theory and Applications*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[7] Y. C. Eldar, *Sampling Theory: Beyond Bandlimited Systems*. Cambridge, U.K.: Cambridge Univ. Press, 2015.

[8] S. Arora, R. Ge, and A. Moitra. (2013). "New algorithms for learning incoherent and overcomplete dictionaries." [Online]. Available: http://arxiv.org/abs/1308.6273

[9] M. Protter and M. Elad, "Image sequence denoising via sparse and redundant representations," *IEEE Trans. Image Process.*, vol. 18, no. 1, pp. 27–35, Jan. 2009.

[10] G. Peyré, "Sparse modeling of textures," *J. Math. Imag. Vis.*, vol. 34, no. 1, pp. 17–31, May 2009.

[11] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?" *Vis. Res.*, vol. 37, no. 23, pp. 3311–3325, Dec. 1997.

[12] I. Tosic and P. Frossard, "Dictionary learning for stereo image representation," *IEEE Trans. Image Process.*, vol. 20, no. 4, pp. 921–934, Apr. 2011.

[13] M. Turkan and C. Guillemot, "Online dictionaries for image prediction," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2011, pp. 293–296.

[14] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Non-local sparse models for image restoration," in *Proc. 12th IEEE Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 2272–2279.

[15] J. Mairal, M. Elad, and G. Sapiro, "Sparse representation for color image restoration," *IEEE Trans. Image Process.*, vol. 17, no. 1, pp. 53–69, Jan. 2008.

[16] M. Zibulevsky and B. Pearlmutter, "Blind source separation by sparse decomposition in a signal dictionary," *Neural Comput.*, vol. 13, no. 4, pp. 863–882, Apr. 2001.

[17] R. Jenatton, G. Obozinski, and F. Bach. (Sep. 2009). "Structured sparse principal component analysis." [Online]. Available: http://arxiv.org/abs/0909.1440

[18] F. Bach, J. Mairal, and J. Ponce, "Convex sparse matrix factorizations," Tech. Rep., 2008.

[19] K. Schnass, "On the identifiability of overcomplete dictionaries via the minimisation principle underlying K-SVD," *Appl. Comput. Harmon. Anal.*, vol. 37, no. 3, pp. 464–491, Nov. 2014.

[20] M. Aharon, M. Elad, and A. M. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[21] R. Jenatton, R. Gribonval, and F. Bach. (Oct. 2012). "Local stability and robustness of sparse dictionary learning in the presence of noise." [Research Report]. Available: http://arxiv.org/abs/1210.0685

[22] R. Gribonval and K. Schnass, "Dictionary identification—Sparse matrix-factorization via ℓ_1-minimization," *IEEE Trans. Inf. Theory*, vol. 56, no. 7, pp. 3523–3539, Jul. 2010.

[23] M. Yaghoobi, T. Blumensath, and M. E. Davies, "Dictionary learning for sparse approximations with the majorization method," *IEEE Trans. Signal Process.*, vol. 57, no. 6, pp. 2178–2191, Jun. 2009.

[24] J. T. Parker, P. Schniter, and V. Cevher, "Bilinear generalized approximate message passing—Part I: Derivation," *IEEE Trans. Inf. Theory*, vol. 62, no. 22, pp. 5839–5853, Nov. 2014.

[25] J. T. Parker, P. Schniter, and V. Cevher, "Bilinear generalized approximate message passing—Part II: Applications," *IEEE Trans. Inf. Theory*, vol. 62, no. 22, pp. 5854–5867, Nov. 2014.

[26] D. A. Spielman, H. Wang, and J. Wright, "Exact recovery of sparsely-used dictionaries," in *Proc. Conf. Learn. Theory*, Jun. 2012, pp. 37.1–37.18.

[27] A. Agarwal, A. Anandkumar, P. Jain, P. Netrapalli, and R. Tandon, "Learning sparsely used overcomplete dictionaries via alternating minimization," *J. Mach. Learn. Res.*, vol. 35, pp. 1–15, Jul. 2014.

[28] F. Krzakala, M. Mezard, and L. Zdeborova, "Phase diagram and approximate message passing for blind calibration and dictionary learning," in *Proc. IEEE ISIT*, Jul. 2013, pp. 659–663.

[29] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proc. Nat. Academy Sci. USA*, vol. 106, no. 45, pp. 18914–18919, 2009.

[30] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proc. 26th Annu. Int. Conf. Mach. Learn. (ICML)*, Montreal, QC, Canada, 2009, pp. 689–696.

[31] W. Wang, M. J. Wainwright, and K. Ramchandran, "Information-theoretic bounds on model selection for Gaussian Markov random fields," in *Proc. IEEE ISIT*, Austin, TX, USA, Jun. 2010, pp. 1373–1377.

[32] M. J. Wainwright, "Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting," *IEEE Trans. Inf. Theory*, vol. 55, no. 12, pp. 5728–5741, Dec. 2009.

[33] E. J. Candès and M. A. Davenport, "How well can we estimate a sparse vector?" *Appl. Comput. Harmon. Anal.*, vol. 34, no. 2, pp. 317–323, Mar. 2013.

[34] N. P. Santhanam and M. J. Wainwright, "Information-theoretic limits of selecting binary graphical models in high dimensions," *IEEE Trans. Inf. Theory*, vol. 58, no. 7, pp. 4117–4134, Jul. 2012.

[35] T. T. Cai and H. H. Zhou, "Optimal rates of convergence for sparse covariance matrix estimation," *Ann. Stat.*, vol. 40, no. 5, pp. 2389–2420, 2012.

[36] D. Vainsencher, S. Mannor, and A. M. Bruckstein, "The sample complexity of dictionary learning," *J. Mach. Learn. Res.*, vol. 12, pp. 3259–3281, Feb. 2011.

[37] R. Gribonval, R. Jenatton, F. Bach, M. Kleinsteuber, and M. Seibert, "Sample complexity of dictionary learning and other matrix factorizations," *IEEE Trans. Inf. Theory*, vol. 61, no. 6, pp. 3469–3486, Jun. 2015.

[38] A. Jung, S. Schmutzhard, F. Hlawatsch, Y. C. Eldar, and Z. Ben-Haim, "Minimum variance estimation of a sparse vector within the linear Gaussian model: An RKHS approach," *IEEE Trans. Inf. Theory*, vol. 60, no. 10, pp. 6555–6575, Oct. 2014.

[39] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, Mar. 2008.

[40] R. G. Baraniuk, "Compressive sensing [lecture notes]," *IEEE Signal Process. Mag.*, vol. 24, no. 4, pp. 118–121, Jul. 2007.

[41] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*. Princeton, NJ, USA: Princeton Univ. Press, 2008.

[42] P.-A. Absil and K. A. Gallivan, "Joint diagonalization on the oblique manifold for independent component analysis," in *Proc. IEEE ICASSP*, Toulouse, France, May 2006, pp. 1–5.

[43] Y. C. Eldar, "Rethinking biased estimation: Improving maximum likelihood and the Cramér–Rao bound," *Found. Trends Signal Process.*, vol. 1, no. 4, pp. 305–449, Jan. 2007,

[44] E. L. Lehmann and G. Casella, *Theory of Point Estimation*, 2nd ed. New York, NY, USA: Springer, 1998.

[45] B. Yu, "Assouad, Fano, and Le Cam," in *Festschrift for Lucien Le Cam*, D. Pollard, E. Torgersen, and G. L. Yang, Eds. New York, NY, USA: Springer, 1997, pp. 423–435.

[46] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.

[47] A. Lapidoth and S. M. Moser, "Capacity bounds via duality with applications to multiple-antenna systems on flat-fading channels," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2426–2467, Oct. 2003.

[48] J. R. Hershey and P. A. Olsen, "Approximating the Kullback Leibler divergence between Gaussian mixture models," in *Proc. IEEE ICASSP*, Honolulu, HI, Apr. 2007, pp. 317–320.

[49] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. New York, NY, USA: Springer, 2012.

[50] E. J. Candès, "The restricted isometry property and its implications for compressed sensing," *Comp. Rendus Math.*, vol. 346, nos. 9–10, pp. 589–592, May 2008.

[51] P. Stoica and B. C. Ng, "On the Cramér–Rao bound under parametric constraints," *IEEE Signal Process. Lett.*, vol. 5, no. 7, pp. 177–179, Jul. 1998.

[52] K. V. Mardia and R. J. Marshall, "Maximum likelihood estimation of models for residual covariance in spatial regression," *Biometrika*, vol. 71, no. 1, pp. 135–146, Apr. 1984.

[53] J.-L. Durrieu, J.-P. Thiran, and F. Kelly, "Lower and upper bounds for approximation of the Kullback–Leibler divergence between Gaussian mixture models," in *Proc. IEEE ICASSP*, Kyoto, Japan, Mar. 2012, pp. 4833–4836.

[54] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. Baltimore, MD, USA: The Johns Hopkins Univ. Press, 1996.

[55] P. Billingsley, *Probability and Measure*, 3rd ed. New York, NY, USA: Wiley, 1995.

Alexander Jung received the Diplom-Ingenieur and Dr. techn. degrees in electrical engineering/signal processing from Vienna University of Technology, Vienna, Austria, in 2008, and 2011, respectively.

He is currently assistant professor for machine learning within the Department of Computer Science at Aalto University. His research interests are in statistical signal processing and machine learning for big data with emphasis on sparse models as well as trade-offs between accuracy and computational/communication complexity of learning algorithms. He received several national and international awards including a Best Student Paper award at the worlds largest signal processing conference, the IEEE ICASSP 2011.

Yonina C. Eldar (S'98–M'02–SM'07–F'12) received the B.Sc. degree in Physics in 1995 and the B.Sc. degree in Electrical Engineering in 1996 both from Tel-Aviv University (TAU), Tel-Aviv, Israel, and the Ph.D. degree in Electrical Engineering and Computer Science in 2002 from the Massachusetts Institute of Technology (MIT), Cambridge.

From January 2002 to July 2002 she was a Postdoctoral Fellow at the Digital Signal Processing Group at MIT. She is currently a Professor in the Department of Electrical Engineering at the Technion - Israel Institute of Technology, Haifa, Israel, where she holds the Edwards Chair in Engineering.

She is also a Research Affiliate with the Research Laboratory of Electronics at MIT and was a Visiting Professor at Stanford University, Stanford, CA.

Her research interests are in the broad areas of statistical signal processing, sampling theory and compressed sensing, optimization methods, and their applications to biology and optics.

Dr. Eldar has received numerous awards for excellence in research and teaching, including the IEEE Signal Processing Society Technical Achievement Award (2013), the IEEE/AESS Fred Nathanson Memorial Radar Award (2014), and the IEEE Kiyo Tomiyasu Award (2016).

She was a Horev Fellow of the Leaders in Science and Technology program at the Technion and an Alon Fellow. She received the Michael Bruno Memorial Award from the Rothschild Foundation, the Weizmann Prize for Exact Sciences, the Wolf Foundation Krill Prize for Excellence in Scientific Research, the Henry Taub Prize for Excellence in Research (twice), the Hershel Rich Innovation Award (three times), the Award for Women with Distinguished Contributions, the Andre and Bella Meyer Lectureship, the Career Development Chair at the Technion, the Muriel & David Jacknow Award for Excellence in Teaching, and the Technionís Award for Excellence in Teaching (two times). She received several best paper awards and best demo awards together with her research students and colleagues including the SIAM outstanding Paper Prize and the IET Circuits, Devices and Systems Premium Award, and was selected as one of the 50 most influential women in Israel.

She is a member of the Young Israel Academy of Science and Humanities and the Israel Committee for Higher Education, and an IEEE Fellow.

She is the Editor in Chief of *Foundations and Trends in Signal Processing*, a member of the IEEE Sensor Array and Multichannel Technical Committee and serves on several other IEEE committees. In the past, she was a Signal Processing Society Distinguished Lecturer, member of the IEEE Signal Processing Theory and Methods and Bio Imaging Signal Processing technical committees, and served as an associate editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the *EURASIP Journal of Signal Processing*, the *SIAM Journal on Matrix Analysis and Applications*, and the *SIAM Journal on Imaging Sciences*. She was Co-Chair and Technical Co-Chair of several international conferences and workshops.

She is author of the book "*Sampling Theory: Beyond Bandlimited Systems*" and co-author of the books "*Compressed Sensing*" and "*Convex Optimization Methods in Signal Processing and Communications*," all published by Cambridge University Press.

Norbert Görtz (S'97–M'99–SM'03) received the Dipl.-Ing. degree in Electrical Engineering from the Ruhr-University Bochum, Germany, in 1993, the Dr.-Ing. degree from Christian-Albrechts-University Kiel, Germany, in 1999, for research work in coded speech transmission over noisy channels, and the Habilitation degree from Munich University of Technology, Germany, in 2004, with a thesis on joint source-channel coding. From 1999 until 2000 he was a Post-doctoral Researcher with the Institute of Network and Systems Theory, Christian-Albrechts-University Kiel. After that he became a Post-doctoral Researcher with the Institute for Communications Engineering, Munich University of Technology, until 2004. After a three-month research visit of the IT Department at Lund University, Sweden, and a temporary C4 professorship at the University of Kassel, Germany, he went to Scotland in October 2004, where he became a Lecturer and a Senior Lecturer with the Institute for Digital Communications, School of Engineering at the University of Edinburgh. Since September 2008 he has been a Full Professor of multimedia signal processing with the Institute of Telecommunications, Vienna University of Technology, Austria. His research interests include source and channel coding, in particular LDPC channel codes, resource allocation and scheduling under delay constraints, cross-layer design and, more recently, compressed sensing and iterative recovery with applications in signal processing and communications.