Dictionary Learning for the Almost-Linear Sparsity Regime

Alexei Novikov NOVIKOV@PSU.EDU and **Stephen White** SEW347@PSU.EDU
Department of Mathematics, Penn State University, University Park, PA 16802 *

Editors: Shipra Agrawal and Francesco Orabona

Abstract

Dictionary learning, the problem of recovering a sparsely used matrix $\mathbf{D} \in \mathbb{R}^{M\times K}$ and N independent $K\times 1$ s-sparse vectors $\mathbf{X} \in \mathbb{R}^{K\times N}$ from samples of the form $\mathbf{Y} = \mathbf{DX}$, is of increasing importance to applications in signal processing and data science. Early papers on provable dictionary learning identified that one can detect whether two samples $\mathbf{y}_i, \mathbf{y}_j$ share a common dictionary element by testing if their inner product (correlation) exceeds a certain threshold: $|\langle \mathbf{y}_i, \mathbf{y}_j\rangle| > \tau$. These correlation-based methods work well when sparsity is small, but suffer from declining performance when sparsity grows faster than $\sqrt{M}$; as a result, such methods were abandoned in the search for dictionary learning algorithms when sparsity is nearly linear in M.

In this paper, we revisit correlation-based dictionary learning. Instead of seeking to recover individual dictionary atoms, we employ a spectral method to recover the subspace spanned by the dictionary atoms in the support of each sample. This approach circumvents the primary challenge encountered by previous correlation methods, namely that when sharing information between two samples it is difficult to tell *which* dictionary element the two samples share. We prove that under a suitable random model the resulting algorithm recovers dictionaries in polynomial time for sparsity linear in M up to log factors. Our results improve on the best known methods by achieving a decaying error bound in dimension M; the best previously known results for the overcomplete ($K > M$) setting achieve polynomial time in the linear regime only for constant error bounds. Numerical simulations confirm our results.

Keywords: Compressed Sensing, Dictionary Learning, Sparsity, Sparse Coding

1. Introduction

The problem of finding sparse representations for large datasets is of tremendous importance in data science and machine learning applications. Sparse representations have obvious advantages for data storage and processing, while offering insight into a dataset's intrinsic structure. This *sparse recovery* problem can often be formulated as that of how to recover a sparse vector $\mathbf{x} \in \mathbb{R}^K$ from a dense sample of the form $\mathbf{y} = \mathbf{Dx}$ for some known sparsely-used matrix $\mathbf{D}$ called the "dictionary." Depending on applications, this dictionary can be known from physics or hand-designed, such as the wavelet bases used in image processing (e.g., Vetterli and Kovačevic, 1995). Beyond numerous further applications in signal and image processing (see Elad (2010) for a summary of developments), sparse representations have been fruitfully applied in areas including computational neuroscience (Olshausen and Field, 1996b,a, 1997) and machine learning (Argyriou et al., 2006; Ranzato et al., 2007).

Yet as the tasks of understanding and compressing data become increasingly central to the needs of modern technology, there is a corresponding interest in methods which learn from data not only the underlying sparse codes but also the dictionary itself. This is the *dictionary learning problem*.

* Both authors were partially supported by grants NSF DMS-1813943 and AFOSR FA9550-20-1-0026.

Specifically, we aim to recover a sparsely used matrix $\mathbf{D} \in \mathbb{R}^{M \times K}$ from N measurements of the form $\mathbf{y} = \mathbf{D}\mathbf{x}$, where $\mathbf{x} \in \mathbb{R}^K$ is sufficiently sparse (that is, $\mathbf{x}$ has significantly fewer than K nonzero entries). Written in matrix form, we seek to recover $\mathbf{D}$ from a matrix $\mathbf{Y} = \mathbf{D}\mathbf{X} \in \mathbb{R}^{M \times N}$ with the prior knowledge that rows of $\mathbf{X}$ are sparse. In most applications, practitioners are interested in recovering *overcomplete* dictionaries which satisfy $K > M$, as these allow for greater flexibility in basis selection and for sparser representation (e.g. Chen et al., 1998; Donoho et al., 2006).

1.1. Prior Work

Dictionary learning is typically formulated as a nonconvex optimization problem of finding the dictionary $\mathbf{D}$ and matrix with sparse columns $\mathbf{X}$ such that $\mathbf{X}$ is as sparse as possible:

$$\text{Find } \mathbf{D}, \mathbf{X} \text{ minimizing } \|\mathbf{X}\|_0 \text{ subject to } \mathbf{Y} = \mathbf{D}\mathbf{X}.$$

As a nonconvex optimization, finding solutions to this problem is computationally challenging. The most popular heuristic for solving the dictionary learning problem is *alternating minimization*. Alternating minimization algorithms rely on the fact that when $\mathbf{D}$ or $\mathbf{X}$ is known, the other can be solved using known methods, most frequently based on ℓ_1 relaxations that make the problem convex (Candés et al., 2006). Alternating minimization techniques thus alternate between a "sparse coding" step in which a guess for $\mathbf{D}$ is fixed and the algorithm solves for $\mathbf{X}$, and a "dictionary update" step in which $\mathbf{X}$ is fixed and the dictionary is updated. This process is repeated until a convergence criterion is met. These algorithms often lack theoretical guarantees, though some recent work has found conditions under which particular alternating minimzation algorithms converge to a global minimum in the dictionary learning setting (Chatterji and Bartlett, 2017; Agarwal et al., 2016).

In this paper, we are interested in dictionary learning algorithms with provable guarantees. Initial theoretical study of provable dictionary learning focused on the case when s is no greater than $\sqrt{M}$; this is a well-known recovery boundary even when the dictionary $\mathbf{D}$ is known. Spielman et al. (2012) developed an algorithm that accurately recovers the dictionary in this sparsity regime, but their algorithm does not generalize to recovery of overcomplete dictionaries. Arora et al. (2013) and Agarwal et al. (2017) then independently introduced similar correlation and clustering methods, which enjoy similar theoretical guarantees for the $s \sim \sqrt{M}$ regime.

Candés et al. (2006) showed that when the dictionary $\mathbf{D}$ is known and satisfies certain properties such as the restricted isometry property (Candes and Tao, 2005), it is possible to recover $\mathbf{x}_i$ from $\mathbf{y}_i = \mathbf{D}\mathbf{x}_i$ when $\mathbf{x}_i$ has linearly-many nonzeros in M. Accordingly, there was tremendous interest in determining whether recovery in this scaling regime remained possible when $\mathbf{D}$ is unknown. In (Arora et al., 2014), the authors develop provable methods for recovering dictionaries with sparsity $s = \mathcal{O}(M)$ up to logarithmic factors, but their method requires quasipolynomial running time. In a pair of papers, Sun, Qu, and Wright develop a polynomial-time method which can provably recover invertible dictionaries with $s = \mathcal{O}(M)$ (Sun et al., 2017a,b). However, this algorithm depends intimately on properties of orthogonal and invertible matrices and thus is limited to the case of complete dictionaries ($M = K$); moreover, their theoretical guarantees demand a high sample complexity of $K \gg M^8$.

More recently, Zhai et al. (2020b) introduced a method based on ℓ^4 norm optimization. Despite many nice properties of this approach further elaborated by Zhai et al. (2020a), it remains limited to the complete case and the authors prove only the accuracy of a global optimum with no guarantee of convergence in arbitrary dimension.

For the overcomplete setting, Barak et al. (2015) developed a tensor decomposition method based on the sum-of-squares hierarchy that can recover overcomplete dictionaries with sparsity up to $s = \mathcal{O}\left(M^{1-\delta}\right)$ for any $\delta > 0$ in polynomial time, but this time tends to super-polynomial as $\delta \to 0$ and requires a constant error as $M \to \infty$. This and related methods have generally enjoyed the best theoretical guarantees for efficient dictionary learning in the overcomplete linear sparsity regime due to their impressive generality, especially after their runtime was improved to polynomial time by Ma et al. (2016) provided that the target error remains constant. Yet the requirement of constant error is strict—with these methods, even in sublinear sparsity regimes such as $s \sim M^{0.99}$, an inverse logarithmic decay in error requires superpolynomial time.

1.2. Intuition and Our Contribution

The correlation-based clustering method of Arora et al. (2014) offers an appealing intuition in the $s \sim \sqrt{M}$ regime: that pairs of elements which "look similar" in the sense of being highly correlated are likely to share support. If the dictionary $\mathbf{D}$ is incoherent (that is, $|\langle \mathbf{d}_{k_1}, \mathbf{d}_{k_2} \rangle| \leq c/\sqrt{M}$ for $k_1 \neq k_2$) and the coefficients $\mathbf{x}_i, \mathbf{x}_j$ are symmetric, then if $s \ll \sqrt{M}$ then as $M \to \infty$, the *correlation*[1] $\langle \mathbf{y}_i, \mathbf{y}_j \rangle$ will concentrate near zero if $\mathbf{y}_i$ and $\mathbf{y}_j$ share no support, but concentrate around ± 1 if they do. As a result, thresholding $|\langle \mathbf{y}_i, \mathbf{y}_j \rangle|$ becomes a reliable indicator of whether $\mathbf{y}_i$ and $\mathbf{y}_j$ share a common dictionary element in their support. By constructing a graph with an edge between i and j whenever $|\langle \mathbf{y}_i, \mathbf{y}_j \rangle| \geq \tau$ for some threshold τ (say 1/2), one can determine which groups of $\mathbf{y}_i$'s share *the same* common support element by applying overlapping clustering methods. One can then recover the individual dictionary vectors by a spectral method on each of the resulting clusters.

Yet correlation-based clustering cannot be performed with accuracy once sparsity exceeds $\sqrt{M}$, as above this threshold correlation no longer reliably indicates whether two samples share a common dictionary element. As a result, methods based on the correlation $\langle \mathbf{y}_i, \mathbf{y}_j \rangle$ have not been widely employed in subsequent attempts to solve dictionary learning in the linear ($s \sim M$) sparsity regime, with practitioners instead turning to more technical machinery such as the sum-of-squares hierarchy (Barak et al., 2015) or Riemannian trust-regions (Sun et al., 2017a,b).

In our present work, we revisit correlation-based dictionary learning methods. By adopting a different approach which sidesteps the key challenges of previous correlation thresholding methods, we are able to apply these methods successfully even in the linear sparsity regime. In this paper, we introduce the **Spectral Subspace Dictionary Learning** (SSDL) algorithm for solving the overcomplete dictionary learning problem in the linear sparsity regime up to logarithmic factors. We show that for a suitable probabilistic model, the algorithm runs in polynomial time and results in an error which decays in M. In other words, our algorithm actively performs better in high dimensions compared to alternatives: the previous best known methods for the overcomplete linear regime, that is, those of (Ma et al., 2016), require super-polynomial (quasi-polynomial) runtime to achieve errors with even an inverse-logarithmic decay in M. This holds true even in our proposed "almost-linear" regime where sparsity differs from linear in dimension by only a logarithmic factor.

Our method is a natural adaptation of the correlation-based approach of Arora et al. (2013) to the linear sparsity regime. Instead of immediately attempting to recover dictionary elements, we first pursue an intermediate step of recovering *spanning subspaces*, the subspaces $\mathcal{S}_i$ spanned by the supporting dictionary elements of each sample $\mathbf{y}_i$. Once these subspaces are recovered, the individual dictionary elements can be recovered through pairwise comparison of subspaces to find

1. We note this is a slight abuse of notation as the vectors $\mathbf{y}_i, \mathbf{y}_j$ may not be unit vectors.

their intersection. This can be interpreted as reversing the order of the algorithm of Arora et al. (2013): whereas those authors proceeded by first detecting support information about $\mathbf{X}$ then using this to extract geometric information about $\mathbf{D}$, we propose first to recover information about the geometry of $\mathbf{D}$ in the form of spanning subspaces, then to use this subspace information to find shared support among columns of $\mathbf{X}$.

The primary advantage of this approach is that the subspace recovery step effectively recovers geometric information from all of $\mathbf{y}_i$'s supporting dictionary elements at once. In particular, if $\mathbf{y}_i$ and $\mathbf{y}_j$ are highly correlated, we no longer need to concern ourselves with which particular element they share in their support. Accordingly, this approach does not require the cumbersome clustering method used in previous correlation-based methods.

To recover these subspaces, we employ a spectral method based on extracting the eigenvectors of a modified covariance matrix of the samples $\mathbf{Y}$. Specifically, for each j we examine the *correlation-weighted covariance matrix* $\widehat{\Sigma}_j$, defined as the sample covariance of the correlation-weighted samples $\langle \mathbf{y}_i, \mathbf{y}_j \rangle \mathbf{y}_i$. By design, these reweighted samples will have greater variance in the directions of the support elements of $\mathbf{y}_j$, meaning $\widehat{\Sigma}_j$ will have a rank-s "spike" in the directions spanned by support elements of $\mathbf{y}_j$, while a random-matrix assumption on $\mathbf{D}$ guarantees that, with high probability, there will be no comparable spikes in other directions. As a result, the s leading eigenvectors of $\widehat{\Sigma}_j$ (that is, the s eigenvectors corresponding to the s largest eigenvalues) will reliably span a subspace close to that spanned by the support elements of $\mathbf{y}_j$. We provide theoretical guarantees that this method accurately recovers spanning subspaces with sparsity $s \sim M \log^{-(6+\eta)}(M)$ for any $\eta > 0$, which allows for recovery of the individual dictionary elements by a further subspace intersection process (see Algorithm 2).

The rigorous proof of this result follows three broad steps. First, using concentration inequalities from high-dimensional probability, we prove that with high probability, the dictionary $\mathbf{D}$ and samples satisfy certain geometric properties needed for recovery. These properties are all aspects of the fact that a random dictionary in high dimensions will have fairly uniform behavior with no spikes in any particular direction. From this, for fixed j we can prove that the expectation of the correlation-weighted covariance $\widehat{\Sigma}_j$ essentially consists of a multiple of the unweighted covariance plus a rank-s spike in the direction of the spanning subspace of $\mathbf{y}_i$. We conclude by proving that the sample version of the correlation-weighted covariance converges to this expectation, giving a matrix whose top s eigenvectors approximately span the spanning subspace of $\mathbf{y}_i$. The result follows for all j by a union bound.

The resulting algorithm is conceptually simple, easy to implement, and its iterative nature makes it highly parallelizable. Moreover, proving its performance guarantees requires only standard techniques from high-dimensional probability. We emphasize that, unlike many algorithms for dictionary learning with theoretical guarantees, SSDL requires no initialization; accordingly, it is an ideal candidate for use as an initializer for a subsequent refinement by an iterative method.

The sample complexity required for SSDL to recover subspaces accurately depends on the particular sparsity regime. In the most challenging linear-sparsity regime that is our main focus, up to $\log$ factors subspace recovery requires a sample complexity of at most M^4, but in the "easier" regime $s \ll \sqrt{M}$, the required sample complexity eases to $N \sim M$ (see Theorem 4.2). In the linear-sparsity case, the bottleneck is caused by approximation of a covariance matrix in Frobenius norm, which is known to require a factor of M additional samples than does estimation in the operator norm. We believe that in future work this step can be replaced by an approach requiring

approximation only in the ℓ_2 operator norm, in which case the sample complexity would be lowered to M^3 in the linear-sparsity case.

1.3. Structure of Paper

In section 2, we technically specify the problem to be solved and introduce our notations, parameter scaling, and probabilistic model. In section 3, we motivate and detail the SSDL algorithm. Section 4 contains an overview of our main theoretical results, while section 5 sketches their proof, though we defer detailed proofs of most technical lemmas to the appendix. Lastly, section 6 contains the results of numerical experiments validating our results.

2. Parameter Scaling, Data Model, and Conventions

We begin by stating the sparse dictionary learning problem explicitly:

Definition 1 (Sparse Dictionary Learning) *Let* $\mathbf{D} = \begin{pmatrix} \mathbf{d}_1 & \mathbf{d}_2 & \dots & \mathbf{d}_K \end{pmatrix}$ *be an (unknown)* $M \times K$ *matrix with unit vector columns, called the* dictionary. *Let* $\mathbf{x}$ *be an* s*-sparse random vector, and define the random vector* $\mathbf{y} = \mathbf{D}\mathbf{x}$*. The sparse dictionary learning problem is:*

Given $\mathbf{Y} = \mathbf{D}\mathbf{X}$ where $\mathbf{X}$ is a $K \times N$ matrix with columns $\{\mathbf{x}_i\}_{i=1}^N$ i.i.d. copies of $\mathbf{x}$, recover $\mathbf{D}$.

It is clear from the definition that $\mathbf{D}$ can only be recovered up to sign and permutation. Accordingly, we employ the following definition for comparing two dictionaries, due to Arora et al. (2013): we say that two dictionaries are *column-wise* ε*-close* if their columns are close in Euclidean norm after an appropriate permutation and change of sign. In detail:

Definition 2 (Column-wise ε-close (Arora et al., 2013)) *Two dictionaries* $\mathbf{D} = \begin{pmatrix} \mathbf{d}_1 & \mathbf{d}_2 & \dots & \mathbf{d}_K \end{pmatrix}$ *and* $\mathbf{D}' = \begin{pmatrix} \mathbf{d}'_1 & \mathbf{d}'_2 & \dots & \mathbf{d}'_{K'} \end{pmatrix}$ *are column-wise* ε*-close if they have the same dimensions* $M \times K$ *and there exists a permutation* π *of* $\{1, \dots, K\}$ *and a* K*-element sequence* $\theta_k \in \{-1, 1\}$ *such that for all* $k = 1, \dots, K$*:*

$$\|\mathbf{d}_k - \theta_k \mathbf{d}'_{\pi(k)}\|_2 \leq \varepsilon$$

2.1. Parameter Scaling

We denote the following parameters and their scaling:

Definition 3 (Parameters and scaling) *We define* M *to be the dimension of the samples* $\mathbf{y}_j$*,* s *the sparsity level,* K *the dictionary size,* N *the number of samples,* J *the number of recovered subspaces, and* ℓ *the maximum intersection size. We assume the following parameter scaling (all parameters are assumed to grow at most polynomially in* M*):*

- $0 < \gamma < \eta$ *constant in* M
- $s = M \log^{-(4+\eta)}(M)$
- $K = M \log^{2+\gamma}(M)$
- $N \gg \max\left\{\frac{s^{10}\log^{12} M}{M^6}, \frac{K^2 s^4 \log^{10} M}{M^3}\right\} = \max\{M^4 \log^{-32-10\eta}(M), M^3 \log^{(-12+2\gamma-4\eta)}(M)\}$

- $J = K \log^3 K$
- $\ell = \left\lceil \frac{\log(2K)}{\log(K/s)} \right\rceil$, *the smallest integer such that* $(s/K)^\ell \leq \frac{1}{2K}$.

Throughout the text, we will often encounter the terms s/M and K/M. One should think of s/M as a "slightly less than 1" term decaying slowly, and of K/M as a "slightly more than 1" term that grows slowly.

2.2. Data Model

We begin by defining the following distributions for our dictionary $\mathbf{D}$ and sparsity pattern $\mathbf{X}$:

Definition 4 ($\mathcal{U}$ distribution) *A random matrix $\mathbf{D} \in \mathbb{R}^{M \times K}$ follows the $\mathcal{U}$ distribution if its columns $\{\mathbf{d}_k\}_{k=1}^K$ are K independent and uniformly distributed unit vectors in $\mathbb{R}^M$.*

Definition 5 ($\mathcal{X}(W)$ distribution) *Let W be a symmetric random variable satisfying $|W| \in [c, C]$ almost surely for $0 < c \leq C$. A random vector $\mathbf{X} \in \mathbb{R}^{K \times N}$ follows a $\mathcal{X}(W)$ distribution if:*

- *The supports $\Omega_i = \mathbf{supp}\,(\mathbf{x}_i)$ of each column $\mathbf{x}_i$ of $\mathbf{X}$ are independent, uniformly random s-element subsets of $\{1, \ldots, K\}$.*
- *Nonzero entries of $\mathbf{X}$ are i.i.d. copies of W.*

This definition implies columns of $\mathbf{X}$ are independent when distributed according to a $\mathcal{X}(W)$ distribution. In our theoretical results, we assume that $\mathbf{D} \sim \mathcal{U}$ and $\mathbf{X} \sim \mathcal{X}(W)$. As the extension to bounded symmetric random variables is trivial, we assume $W = \pm 1$ with equal probability. This choice of particular distribution for $\mathbf{X}$ is made for theoretical convenience and significantly simplifies the analysis, but we expect our results to hold with minimal modifications for the more commonly used Bernoulli-Gaussian model used by Spielman et al. (2012) and others.

Our result differs from many other provable results on dictionary learning in that we assume the dictionary $\mathbf{D}$ to be a random matrix. The specific geometric properties required to recover $\mathbf{D}$, which are reliably satisfied by a $\mathcal{U}$-distributed random matrix, are outlined in Definition 10. Our treatment is similar to that of the well-known restricted isometry property (Candes and Tao, 2005), a deterministic property often assumed in recovery guarantees for compressed sensing (that is, when the dictionary is known); indeed, we believe that many of the geometric properties we will prove individually may be consequences of the stronger RIP property. The restricted isometry property is known to hold for many types of random matrices, but no families of deterministic matrices for which it holds are yet known (e.g., Bandeira et al., 2013). Accordingly, such recovery results implicitly make a random-matrix assumption for $\mathbf{D}$, as we do here explicitly.

2.3. Notation and conventions

Vectors are represented by boldface lowercase letters, while matrices will be written as boldface uppercase letters. Roman letters (both upper- and lowercase) will be used for both scalars and random variables depending on context. We will use the notation $|\mathcal{A}|$ for the number of elements in a finite set $\mathcal{A}$, and $\mathcal{A}^c$ for its complement.

We use two matrix norms at different points in the text. The standard l_2 operator norm will be denoted by $\| \bullet \|_2$ while the Frobenius norm will be denoted $\| \bullet \|_F$. Vector norms always refer to

the standard l_2 (Euclidean) norm, and will be denoted $\| \bullet \|_2$. We will use the notation $a \ll b$, where both a and b are scalars depending on M, to mean $\lim_{M\to\infty} |a|/|b| = 0$, where the norm in question may depend on context.

The index-free notation $\mathbf{y} = \mathbf{D}\mathbf{x}$ will refer to a generic independent copy drawn from the sampling distribution, used for index-independent properties of this distribution such as expectation, while we reserve the indexed notation $\mathbf{y}_i$ to refer to a particular random vector in the sample $\mathbf{Y}$. Given a sample $\mathbf{y}_i$, its *support*, denoted Ω_i, is defined as the set of indices of the dictionary vectors in its construction with nonzero coefficients:

$$\Omega_i := \mathbf{supp}\left(\mathbf{x}_i\right) = \{k \in \{1, \ldots, K\} : x_{ik} \neq 0\}$$

The "support vectors" of $\mathbf{y}_i$ refer to the dictionary elements indexed by Ω_i, the set $\{\mathbf{d}_k\}_{k\in\Omega_i}$. We use the notation $\mathcal{A} - \mathcal{B}$ for the relative complement of set $\mathcal{B}$ in set $\mathcal{A}$: $\mathcal{A} - \mathcal{B} = \{x : x \in \mathcal{A}, x \notin \mathcal{B}\}$. We denote the dimension of a vector subspace $\mathcal{S}$ with the shorthand $\dim(\mathcal{S})$. We reserve δ for the Dirac delta function: $\delta_{k\in\Omega_i}$ equals one for $k \in \Omega_i$ and zero otherwise.

Throughout this text, "with high probability" means that an event occurs with probability converging to 1 faster than any polynomial in M; often these will be bounds with the approximate form $M^{-\log M}$:

Definition 6 (High Probability) *A sequence of events ω_M is said to occur* with high probability *in M provided that for any constant $\alpha > 0$,*

$$\lim_{M\to\infty} M^{\alpha}(1 - \mathbb{P}\left(\omega_M\right)) = 0.$$

We will frequently make use of the fact that under this definition, the union of polynomially-many events occurring with high probability also occurs with high probability. As our results are asymptotic in nature, we implicitly assume without statement that, where necessary, M is sufficiently large for our results to hold. Lastly, constants c and C are used to represent "some sufficiently large constant" and may change between lines.

3. Algorithm

In this section, we outline the key elements of the spectral subspace dictionary learning algorithm (SSDL). SSDL consists of two main steps: *subspace recovery*, wherein we aim to recover the subspaces spanned by the support vectors of each sample $\mathbf{y}_i$, and *subspace intersection*, which combines the information from subsets of the recovered subspaces to recover individual dictionary elements.

3.1. History and Motivation

The key concept underlying **SSDL** and its proof is the idea that for a dictionary with approximately orthogonal columns (typically $|\langle \mathbf{d}_k, \mathbf{d}_m \rangle| \leq C/\sqrt{M}$ for $k \neq m$), given two different samples $\mathbf{y}_i$ and $\mathbf{y}_j$, the absolute inner product $|\langle \mathbf{y}_i, \mathbf{y}_j \rangle|$ should be larger when they share an element in their support. This idea was used by Arora et al. (2013) for the case $s \ll \sqrt{M}$. Indeed, since dictionary vectors are unit vectors, we have:

$$\langle \mathbf{y}_i, \mathbf{y}_j \rangle = \sum_{k\in\Omega_i} \sum_{m\in\Omega_j} x_{ik}x_{jm} \langle \mathbf{d}_k, \mathbf{d}_m \rangle = \sum_{k\in\Omega_i\cap\Omega_j} x_{ik}x_{jk} + \sum_{k\in\Omega_i-\Omega_j} \sum_{m\in\Omega_j-\Omega_i} x_{ik}x_{jm} \langle \mathbf{d}_k, \mathbf{d}_m \rangle$$

For a random dictionary, the inner product $\langle \mathbf{d}_k, \mathbf{d}_m \rangle$ will be of order approximately $1/\sqrt{M}$ with high probability. Thus if nonzero coefficients x_{ik} are bounded below, each term in the sum over the intersection $\Omega_i \cap \Omega_j$ is much larger than the terms in the second sum. In many cases, particularly when $s \ll \sqrt{M}$, the intersection $\Omega_i \cap \Omega_j$ will contain at most one element, in which case the first sum is either 0 or ± 1. On the other hand, the second term will be a sum of approximately s^2 random variables of magnitude $1/\sqrt{M}$. In particular, if nonzero entries of $\mathbf{x}_i$ are sub-Gaussian and have mean zero, the Hanson-Wright inequality (Hanson and Wright, 1971) guarantees that with high probability (up to possible log factors):

$$\left| \sum_{k \in \Omega_i - \Omega_j} \sum_{m \in \Omega_j - \Omega_i} x_{ik} x_{jm} \langle \mathbf{d}_k, \mathbf{d}_m \rangle \right| \leq \frac{\sqrt{|\Omega_i - \Omega_j|}\sqrt{|\Omega_j - \Omega_i|}}{\sqrt{M}} \approx \frac{s}{\sqrt{M}}$$

From this one can consider the following heuristic normal approximation:

$$\langle \mathbf{y}_i, \mathbf{y}_j \rangle \sim \begin{cases} N(0, C^2 s^2/M), & |\Omega_i \cap \Omega_j| = 0 \\ N(\pm 1, C^2 s^2/M), & |\Omega_i \cap \Omega_j| = 1 \end{cases} \tag{1}$$

When $s \ll \sqrt{M}$, the absolute inner product $|\langle \mathbf{y}_i, \mathbf{y}_j \rangle|$ behaves close to an indicator function for whether $\mathbf{y}_i$ and $\mathbf{y}_j$ share support (the case $|\Omega_i \cap \Omega_j| \geq 2$ occurs with negligible probability for $s \ll \sqrt{M}$ and $K \geq M$).

Arora et al. (2013) used these inner products to construct a graph where i and j shared an edge if $|\langle \mathbf{y}_i, \mathbf{y}_j \rangle|$ exceeded some threshold τ. Overlapping clustering methods were then used to recover overlapping communities $\mathcal{C}_k$ each corresponding to a dictionary element, after which the dictionary elements could be recovered by averaging or by taking the top eigenvalue of the community covariance $\frac{1}{N}\sum_{i \in \mathcal{C}_k} \mathbf{y}_i \mathbf{y}_i^T$.

However, when $s \gg \sqrt{M}$, this approach breaks down, as the variance of the terms in 1 dominates the mean term, meaning $|\langle \mathbf{y}_i, \mathbf{y}_j \rangle|$ can no longer be used as a reliable indicator of shared support. Indeed, in this case for any fixed threshold τ, $\mathbb{P}(\langle \mathbf{y}_i, \mathbf{y}_j \rangle \geq \tau)$ will tend to 1, making reliable community detection nearly impossible. This was perceived as a fundamental roadblock to generalizing the correlation-based techniques of Arora et al. (2013) beyond the $s \sim \sqrt{M}$ barrier; accordingly, subsequent research on dictionary learning with theoretical guarantees used entirely different techniques.

In this work, our primary observation is that even in the $s \gg \sqrt{M}$ regime, correlations still contain sufficient information on shared support for the dictionary to be recovered. Even though, in this regime, $\langle \mathbf{y}_i, \mathbf{y}_j \rangle$ is dominated by terms originating from non-shared support elements, there is still a small bias in favor of shared support: $|\langle \mathbf{y}_i, \mathbf{y}_j \rangle|$ will, on average, be larger when $|\Omega_i \cap \Omega_j| > 0$.

As already noted, in this regime the correlations are not strong enough to directly infer the sparsity pattern as in (Arora et al., 2013). Therefore in this work we propose an intermediate step: before attempting to recover the dictionary elements, for each sample $\mathbf{y}_i$ we recover its *spanning subspace* $\mathcal{S}_i$:

Definition 7 (Spanning Subspace) *Given a sample $\mathbf{y}_i = \mathbf{D}\mathbf{x}_i$, the* spanning subspace *of sample i is the subspace $\mathcal{S}_i$ defined as $\mathcal{S}_i = \mathbf{span}\{\mathbf{d}_k : k \in \mathbf{supp}(\mathbf{x}_i)\}$.*

First recovering the spanning subspaces obviates any need to perform community detection on an unreliable connection graph, which was the immediate point of failure for the correlation-based method of Arora et al. (2013) in the $s \gg \sqrt{M}$ setting.

3.2. Subspace Recovery

At a high level, given a sample $\mathbf{y}_j$ the subspace recovery step is a spectral method that constructs a matrix which, with high probability, will have lead s eigenvectors spanning a subspace close to the true spanning subspace $\mathcal{S}_j$. To estimate $\mathcal{S}_j$, we consider a statistic based on the classical estimator for the covariance of $\mathbf{y}$, the *sample covariance matrix* $\widehat{\Sigma}$:

$$\widehat{\Sigma} = \frac{1}{N}\mathbf{Y}\mathbf{Y}^T = \frac{1}{N}\sum_{i=1}^{N}\mathbf{y}_i\mathbf{y}_i^T$$

As long as $E\mathbf{y} = 0$, it is easy to see that $\widehat{\Sigma}$ is an unbiased estimator (that is, $E\widehat{\Sigma} = E\mathbf{y}\mathbf{y}^T$); by the law of large numbers, then, $\widehat{\Sigma} \to E\mathbf{y}\mathbf{y}^T$ in N almost surely (later we will use quantitative versions of this result; see, for instance, Vershynin (2018), theorems 4.7.1 and 5.6.1).

To find the subspace $\mathcal{S}_j$, though, we need an estimator which is biased towards those directions spanned by the support elements of $\mathbf{y}_j$. Our goal is to weight the sample covariance matrix in such a way that a sample $\mathbf{y}_i$ is given more weight the larger the shared support between $\mathbf{y}_i$ and $\mathbf{y}_j$. At first, we employed a method based on the thresholding scheme of Arora et al. (2013):

$$\widehat{\Sigma}_j^{\tau} := \frac{1}{N}\sum_{i\neq j}\mathbb{1}_{\{|\langle \mathbf{y}_j,\mathbf{y}_i\rangle|\geq\tau\}}\mathbf{y}_i\mathbf{y}_i^T$$

where τ was a fixed threshold parameter. Our idea was that, although the bias would be small, $\mathbb{P}\left(\langle \mathbf{y}_j,\mathbf{y}_i\rangle \mid \geq \tau\right)$ would nonetheless be greater when $\mathbf{y}_i$ shares support with $\mathbf{y}_j$ even when $s \gg \sqrt{M}$. This worked well in numerical simulations, but proved intractable for theoretical work.

Noting that the above is the sample covariance matrix for the random vector $\mathbb{1}_{\{|\langle \mathbf{y}_j,\mathbf{y}\rangle|\geq\tau\}}\mathbf{y}$, we replaced this nonlinear thresholding function $\mathbb{1}_{\{|\langle \mathbf{y}_j,\mathbf{y}\rangle|\geq\tau\}}$ with the quadratic weight $\langle \mathbf{y}_j,\mathbf{y}\rangle^2$. This change allowed for much cleaner computations by the linearity of expectations. Accordingly, we now introduce the key statistic of the subspace recovery step, the *correlation-weighted covariance* Σ_j:

$$\Sigma_j := E[\langle \mathbf{y}_j,\mathbf{y}\rangle^2\,\mathbf{y}\mathbf{y}^T]$$

and its sample version $\widehat{\Sigma}_j$:

$$\widehat{\Sigma}_j := \frac{1}{N}\sum_{i\neq j}^{N}\langle \mathbf{y}_j,\mathbf{y}_i\rangle^2\,\mathbf{y}_i\mathbf{y}_i^T$$

We point out that Σ_j and $\widehat{\Sigma}_j$ are the covariance and the sample covariance estimator, respectively, for the random vector $\langle \mathbf{y}_j,\mathbf{y}\rangle\,\mathbf{y}$. Not only is this a more theoretically tractable object, but it also resulted in a an immediate improvement in the accuracy of our empirical simulations. As before, the idea is that samples $\mathbf{y}_i$ which share support elements with $\mathbf{y}_j$ will have a higher correlation and therefore the covariance will be "stretched" in favor of the directions spanned by the support elements of $\mathbf{y}_j$.

A major challenge is that when $s \gg \sqrt{M}$ this bias remains small, with the result that $\widehat{\Sigma}_k$ will be close to the unweighted covariance matrix of $\mathbf{y}$, with only a small perturbation in the directions $\mathcal{S}_j$. However, this covariance can be accurately estimated by the sample covariance $\widehat{\Sigma} = \frac{1}{N}\mathbf{Y}\mathbf{Y}^T$. With this estimate in hand, we remove the $\mathbf{D}\mathbf{D}^T$ component by "covariance projection:" taking the orthogonal complement of $\widehat{\Sigma}_j$ (in the Frobenius sense) with respect to the unweighted sample

covariance matrix $\widehat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{y}_i \mathbf{y}_i^T = \frac{1}{N}\mathbf{Y}\mathbf{Y}^T$, with the goal of leaving only the bias component. Thus for our spectral method, we ultimately look at the span of the eigenvectors corresponding to the s largest eigenvalues of the matrix

$$\widehat{\Sigma}_j^{\text{proj}} := \widehat{\Sigma}_j - \frac{\left\langle \widehat{\Sigma}_j, \widehat{\Sigma} \right\rangle_F}{\|\widehat{\Sigma}\|_F^2} \widehat{\Sigma}$$

As sample covariance matrices, $\widehat{\Sigma}_0$ and $\widehat{\Sigma}$ can be made arbitrarily close to their expectations with sufficiently large sample size, so this statistic will have spectral properties close to the bias matrix $\sum_{k\in\Omega_j} \mathbf{d}_k \mathbf{d}_k^T$. We detail the precise process in Algorithm 1.

Algorithm 1: SSR: Single Subspace Recovery
Input: index j, $M \times N$ data matrix $\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1 & \mathbf{y}_2 & \dots & \mathbf{y}_N \end{pmatrix}$, est. covariance matrix $\widehat{\Sigma} = \frac{1}{N}\mathbf{Y}\mathbf{Y}^T$
Output: s-dimensional subspace $\widehat{\mathcal{S}}_j$
Correlation-Weighted Covariance: Compute $\widehat{\Sigma}_j = \frac{1}{N}\sum_{i=1}^{N} \langle \mathbf{y}_j, \mathbf{y}_i \rangle^2 \mathbf{y}_i \mathbf{y}_i^T$
Covariance Projection: Compute $\widehat{\Sigma}_j^{\text{proj}} = \widehat{\Sigma}_j - \mathbf{proj}_{\widehat{\Sigma}}(\widehat{\Sigma}_j)$
Spectral Recovery: Compute the leading s eigenvectors of $\widehat{\Sigma}_j$ and set $\widehat{\mathcal{S}}_j$ equal to their span.
return $\widehat{\mathcal{S}}_j$

Naturally, the subspace $\widehat{\mathcal{S}}_j$ recovered by Algorithm 1 will only approximately match the true subspace $\mathcal{S}_j$. For this reason we introduce the following metric on subspaces of the same dimension:

Definition 8 (Subspace Distance) *For two s-dimensional subspaces $\mathcal{S}_1$, $\mathcal{S}_2$ of $\mathbb{R}^M$, let $\mathbf{E}_1$ be an orthonormal basis of $\mathcal{S}_1$ and let $\mathbf{F}_2$ be an orthonormal basis of $\mathcal{S}_2^\perp$, the orthogonal complement of $\mathcal{S}_2$. We define the* subspace distance $\mathcal{D}$ *between $\mathcal{S}_1$ and $\mathcal{S}_2$ as:*

$$\mathcal{D}(\mathcal{S}_1, \mathcal{S}_2) = \sup_{\mathbf{z}\in\mathcal{S}, \|\mathbf{z}\|_2=1} \|\mathbf{z} - \mathbf{proj}_{\mathcal{S}_2}(\mathbf{z})\|_2 = \sup_{\mathbf{z}\in\mathcal{S}, \|\mathbf{z}\|_2=1} \|\mathbf{proj}_{\mathcal{S}_2^\perp}(\mathbf{z})\|_2 = \left\|\mathbf{F}_2\mathbf{F}_2^T\mathbf{E}_1\right\|_2 = \|\mathbf{F}_2^T\mathbf{E}_1\|_2$$

In section 4.1, we demonstrate that the recovered $\widehat{\mathcal{S}}_j$ is close to the true $\mathcal{S}_j$ for each j simultaneously with high probability. We also show in Theorem 4.4 that with high probability, up to logarithmic factors only the first $\mathcal{O}(K)$ subspaces are required to find every dictionary vector via subspace intersection.

3.3. Subspace Intersection

Since the eigenvectors returned by this spectral method are orthonormal, they do not correspond to dictionary elements directly, but instead form a basis for a subspace $\widehat{\mathcal{S}}_i$ close to the true subspace $\mathcal{S}_i$. Thus an additional *subspace intersection step* is needed to recover actual dictionary elements from the estimated subspaces.

To motivate our subspace intersection algorithm, we note that if the subspaces $\mathcal{S}_i$ were known exactly, there is a particularly simple algorithm to find a dictionary element when subspaces $\mathcal{S}_i$ are known exactly. Since $\mathcal{S}_i = \mathbf{span}\{\mathbf{d}_k\}_{k\in\Omega_i}$, $\mathcal{S}_i \cap \mathcal{S}_j = \mathbf{span}\{\mathbf{d}_k\}_{k\in\Omega_i\cap\Omega_j}$ (almost surely). It follows that if $\dim(\mathcal{S}_i \cap \mathcal{S}_j) = 1$ exactly, then $\mathcal{S}_i \cap \mathcal{S}_j = \mathbf{span}\{\mathbf{d}_k\}$ where k is the unique element in $\Omega_i \cap \Omega_j$.

Letting $\mathbf{F}_i$ and $\mathbf{F}_j$ be orthonormal basis matrices for $\mathcal{S}_i$ and $\mathcal{S}_j$, respectively, we can write an element in $\mathcal{S}_i$ as $\mathbf{F}_i\mathbf{v}$ for $\mathbf{v} \in \mathbb{R}^s$. Since the matrix for projection onto subspace $\mathcal{S}_j$ is $\mathbf{F}_j\mathbf{F}_j^T$, it follows that $\dim(\mathcal{S}_i \cap \mathcal{S}_j) = \dim(\ker(\mathbf{F}_i - \mathbf{F}_j\mathbf{F}_j^T\mathbf{F}_i))$; denote this matrix $\mathbf{P}_{ij}$. Then any $\mathbf{v}$ in the kernel of $\mathbf{P}_{ij}$ corresponds to a vector $\mathbf{F}_i\mathbf{v}$ in $\mathcal{S}_i \cap \mathcal{S}_j$. Then if $\dim(\ker(\mathbf{P}_{ij})) = 1$, we can easily recover a basis for the intersection $\mathcal{S}_i \cap \mathcal{S}_j$.

As long as s^2/K is small, $|\mathcal{S}_i \cap \mathcal{S}_j|$ will typically have either 0 or 1 element, so an intersection between two subspaces will rarely have dimension above one. When $s^2/K \gg 1$—typically the case in our setting—we will instead need to perform subspace intersection on more than two subspaces. It is easy to see that:

$$E|\Omega_1 \cap \Omega_2 \cap \ldots \Omega_\ell| = \frac{s^{\ell+1}}{K^\ell}$$

Therefore, for $\ell \geq \frac{\log s}{\log K/s}$, we have $E|\Omega_0 \cap \Omega_2 \cap \ldots \Omega_\ell] \leq 1$ (this bound is made precise in Theorem 4.4 with a slightly larger ℓ).

In the case when we only know approximate subspaces $\widehat{\mathcal{S}}_i$, $\mathbf{P}_{ij}$ will almost surely have trivial kernel, so we relax the condition $\dim(\ker(\mathbf{P}_{ij})) = 1$ to the condition that, given some small threshold τ, $\mathbf{P}_{ij}$ has exactly one singular value $\sigma \leq \tau$. This works under the assumption that dictionary elements are nearly orthogonal, which holds when columns of $\mathbf{D}$ are i.i.d. random vectors for a broad class of random vectors (Vershynin, 2018, e.g.,). In practice, τ should not need to be very small; $\tau = 1/2$ was adequate in numerical experiments. We thus define the *approximate subspace intersection* of two subspaces $\mathcal{S}_i, \mathcal{S}_j$ as follows:

Definition 9 (Approximate Subspace Intersection) *Let $\mathcal{S}_i$ and $\mathcal{S}_j$ be subspaces of $\mathbb{R}^M$ with respective orthonormal basis matrices $\mathbf{F}_i, \mathbf{F}_j$. Denote by $\mathbf{P}_{ij} = (\mathbf{I} - \mathbf{F}_j\mathbf{F}_j^T)\mathbf{F}_i$ the projection matrix of $\mathcal{S}_i$ onto $\mathcal{S}_j^\perp$. The* approximate subspace intersection *of $\mathcal{S}_i$ onto $\mathcal{S}_j$ with threshold τ is the subspace $\mathcal{A}_\tau(\mathcal{S}_i, \mathcal{S}_j)$ of $\mathbb{R}^M$ defined as the span of all right singular vectors of $\mathbf{P}_{ij}$ corresponding to sufficiently small singular values:*

$$\mathcal{A}_\tau(\mathcal{S}_i, \mathcal{S}_j) = \mathbf{span}\{\mathbf{v} : \mathbf{v} \textit{ is a right singular vector of } \mathbf{P}_{ij} \textit{ corresponding to a singular value } \sigma_\mathbf{v} \leq \tau\}$$

with the convention that $\mathbf{span}(\emptyset) = \{\mathbf{0}\}$.

We note that this definition ensures $\mathcal{A}_\tau(\mathcal{S}_i, \mathcal{S}_j)$ is a subspace of $\mathcal{S}_i$.

In Algorithm 2, we detail the approximate subspace intersection algorithm for a fixed list of ℓ subspaces $\mathcal{S}_{i_1}, \ldots, \mathcal{S}_{i_\ell}$. In Theorem 4.4, we show that with high probability that to recover all K dictionary elements it suffices to consider only the non-overlapping intersections $\bigcap_{p=1}^{\ell} \mathcal{S}_{\ell(j-1)+p}$ for $j = 1, \ldots, K\log^3 K)/\ell$. To recover an entire dictionary, then, we first employ subspace recovery to learn the subspaces $\{\widehat{\mathcal{S}}_1, \ldots, \widehat{\mathcal{S}}_J\}$ for the first $J = K\log^3 K$ samples, then take the intersection of each consecutive set of ℓ subspaces. (Duplicates, those estimated dictionary elements which are close to one another based on absolute inner product, can be handled by rejecting duplicates or averaging them together.)

3.4. Time Complexity

To compute the correlation-weighted covariance $\widehat{\Sigma}_j$ for a single j, the single subspace recovery algorithm adds N matrices of the form $\langle \mathbf{y}_j, \mathbf{y}_i \rangle^2 \mathbf{y}_i\mathbf{y}_i^T$, each of which can be computed in $\mathcal{O}\left(M^2\right)$

Algorithm 2: SSI, Approximate Subspace ℓ-fold Intersection
Input: List of subspaces $\mathcal{S}_1, \ldots, \mathcal{S}_\ell$, threshold $\tau < 1$
Output: Estimated dictionary element, or **False** if no element is found.
$\mathcal{S} \leftarrow \mathcal{S}_1$
for $i \in \{2, \ldots, \ell\}$ **do**
  $\mathcal{A} = \mathcal{A}_\tau(\mathcal{S}, \widehat{\mathcal{S}}_i)$
  if $\dim(\mathcal{A}) = 0$ **then**
    return False
  else if $\dim(\mathcal{A}) = 1$ **then**
    Set $\hat{\mathbf{d}}$ a basis of $\mathcal{A}$
    return $\hat{\mathbf{d}}$
  else if $\dim(\mathcal{A}) \geq 2$ **then**
    $\mathcal{S} = \mathcal{A}$
  return False
end

time, meaning $\widehat{\Sigma}_j$ can be computed in time $\mathcal{O}\left(NM^2\right)$. Finding the top s eigenvalues and eigenvectors then takes an additional $\mathcal{O}\left(sM^2\right)$ operations, meaning the entire subspace recovery step for a single sample $\mathbf{y}_i$ can be completed in $\mathcal{O}\left(NM^2\right)$ time. As only the first $K \log^3 K$ subspaces are required to find every dictionary vector via subspace intersection, computing all subspaces will take time $\mathcal{O}\left(NKM^2\right)$ (again, up to log factors).

The runtime of the each subspace intersection step is dominated by matrix multiplication and finding eigenvectors, which both have order $\mathcal{O}\left(M^3\right)$. Under the assumption that the support of each sample is uniformly distributed among s-element subsets of $\{1, \ldots, K\}$, we show in Theorem 4.4 that with high probability, one only needs to check fewer than $\mathcal{O}\left(K \log^3 K\right)$ intersections in order to recover each dictionary element. Accordingly, with high probability the subspace intersection step to take $\mathcal{O}\left(\ell KM^3\right)$ time, so we conclude that the entire SSDL process takes $\mathcal{O}\left(NKM^2 + KM^3\right)$ with high probability, up to log factors.

4. Main Results

We are now ready to present our main result, which states that most dictionaries $\mathbf{D} \sim \mathcal{U}$ can be recovered with sparsity linear in M up to a logarithmic factor:

Theorem 4.1 *Fix parameters $0 < \gamma < \eta$ and set $s = M \log^{-(4+\eta)}(M)$, $K = M \log^{2+\gamma}(M)$, $\ell = \left\lceil \frac{\log(2K)}{\log(K/s)} \right\rceil$, and $N \gg \max\left\{ \frac{s^{10} \log^{12} M}{M^6}, \frac{K^2 s^4 \log^{10} M}{M^3} \right\}$ as in 3. Suppose that $\mathbf{Y} = \mathbf{DX}$ with $\mathbf{D} \sim \mathcal{U}$ and $\mathbf{X} \sim \mathcal{X}(W)$. Then for M large enough and*

$$\varepsilon = \frac{\sqrt{M}}{\sqrt{K}} + \frac{Ks \log^2 M}{M^2} + \frac{Ks^2 \log^4 M}{M^{3/2}\sqrt{N}} + \frac{s^5 \log^5 M}{M^3 \sqrt{N}}$$

sufficiently small, **SSDL** *recovers a dictionary $\widehat{\mathbf{D}}$ that is column-wise $(C\ell\varepsilon)$-close to $\mathbf{D}$ with high probability: $\widehat{\mathbf{D}}$ and $\mathbf{D}$ are the same size and there exists a sequence of signs θ, and a permutation π of $\{1, \ldots, K\}$ such that for all $k = 1, \ldots, K$:*

$$\|\mathbf{d}_k - \theta_k \hat{\mathbf{d}}_{\pi(k)}\|_2 \leq C\ell\varepsilon \leq C\varepsilon \log M$$

for an absolute constant C.

Under the scaling in 3, $\varepsilon \ll 1/\log M$ and therefore the recovered dictionary converges to the true dictionary column-wise. We note that this theorem can be adapted to the case where $s = M^{1-2\eta}$ and $K = M^{1+\gamma}$ for some $0 < \gamma < \eta$; we restrict our proof to the case that s and K are linear up to logarithmic factors, as this is the most challenging regime for the dictionary learning problem. That said, our later results will suggest that the sample complexity required by SSDL is lower in these easier sparsity regimes.

4.1. Guarantees for Subspace Recovery

Theorem 4.1 is based on separate guarantees for the subspace recovery and subspace intersection steps, which we review individually, beginning with an overview of our guarantees for the subspace recovery step. We have the following result, which states that the subspaces recovered by SSDL are close to the true spanning subspaces:

Theorem 4.2 (Subspace Recovery) *Fix parameters $0 < \gamma < \eta$ and set $s = M\log^{-(4+\eta)}(M)$, $K = M\log^{2+\gamma}(M)$, and $N \gg \max\left\{\frac{s^{10}\log^{12} M}{M^6}, \frac{K^2 s^4 \log^{10} M}{M^3}\right\}$ as in 3. Suppose that $\mathbf{Y} = \mathbf{DX}$ with $\mathbf{D} \sim \mathcal{U}$ and $\mathbf{X} \sim \mathcal{X}(W)$. Let $\mathcal{S}_j$ be the spanning subspace of sample $\mathbf{y}_j$, while $\widehat{\mathcal{S}}_j$ is the subspace recovered by Algorithm 1. As long as*

$$\varepsilon = \frac{\sqrt{M}}{\sqrt{K}} + \frac{Ks\log^2 M}{M^2} + \frac{Ks^2\log^4 M}{M^{3/2}\sqrt{N}} + \frac{s^5\log^5 M}{M^3\sqrt{N}},$$

is sufficiently small, then with high probability

$$\mathcal{D}\left(\mathcal{S}_j, \widehat{\mathcal{S}}_j\right) \le C\varepsilon.$$

for all $j \in \{1,\dots,N\}$. It follows that as $M \to \infty$, $\widehat{\mathcal{S}}_j \to \mathcal{S}_j$ for all j with high probability.

Details of the proof follow in Section 5, but the key ingredients are these: by a union bound, it suffices to prove that the desired bound holds with high probability for a single j. Using concentration of measure results, we show that $\widehat{\Sigma}_j$ converges to its expectation, which will be close to a rank-s matrix with eigenvectors spanning $\mathcal{S}_j$. We then apply Weyl's Theorem and the Davis-Kahan Theorem on continuity of eigenvalues and invariant subspaces, respectively, of symmetric matrices under perturbation (Weyl, 1912; Davis and Kahan, 1970), allowing us to bound the distance between the recovered subspace $\widehat{\mathcal{S}}_j$ and true subspace $\mathcal{S}_j$.

4.1.1. SAMPLE COMPLEXITY

Since the sample covariance matrix is an unbiased estimator of the true covariance matrix, the law of large numbers will guarantee that for fixed dictionary $\mathbf{D}$, $\widehat{\Sigma}_j \to E[\widehat{\Sigma}_j|\mathbf{D}]$ almost surely with enough samples N. We claim that the empirically observed matrix $\widehat{\Sigma}_0$ will have the same spectral properties as its expectation with high probability as long as N is larger than $\max\left\{s^{10}\log^{10} M/M^6, K^2 s^4 \log^8 M/M^3\right\}$. The additional two factors of $\log M$ in 3 reflect the fact that up to $\ell \le C\log M$ intersections will be taken during the subspace intersection step, allowing for the error to potentially magnify by a factor of $\log M$ by the triangle inequality. Overall, this translates to a worst-case sample complexity of order M^4 up to log factors. See Lemma 5.5 for details.

4.2. Subspace Intersection

In this section, we present guarantees stating that that with high probability, the subspace intersection step rejects groups of subspaces which do not contain a unique dictionary element in their intersection. If they do intersect, then subspace intersection returns a vector close to the true vector. Specifically:

Theorem 4.3 (Subspace Intersection) *Fix parameters $0 < \gamma < \eta$ and set $s = M \log^{-(4+\eta)}(M)$, $K = M \log^{2+\gamma}(M)$, $\ell = \left\lceil \frac{\log(2K)}{\log(K/s)} \right\rceil$, and $N \gg \max\left\{ \frac{s^{10}\log^{12} M}{M^6}, \frac{K^2 s^4 \log^{10} M}{M^3} \right\}$ as in 3. Let $\mathcal{J}$ be a collection of at most polynomially-many ℓ-element subsets of $\{1, \dots, N\}$. With high probability as $M \to \infty$, the following holds for every $\mathcal{I} \in \mathcal{J}$:*

If $\bigcap_{i \in \mathcal{I}} \mathcal{S}_i = \mathbf{span}(\mathbf{d}_k)$, then $\hat{\mathbf{d}}$ will be returned by Algorithm 2 with $\tau = 1/2$ and will satisfy

$$\min_{t \in \{-1,1\}} \{\|\hat{\mathbf{d}} - t\mathbf{d}_k\|_2\} \leq C\ell\varepsilon \leq C\varepsilon \log M$$

*for ε as in Theorem 4.2. Moreover, if $\mathbf{dim}\left(\bigcap_{i \in \mathcal{I}} \mathcal{S}_i\right) \neq 1$, the algorithm returns **False**.*

This result follows from the subspace recovery bound from Theorem 4.2 and the fact that two random s-dimensional subspaces in $\mathbb{R}^M$ will not be closer than any constant with high probability (Lemma 5.10).

To complete the accuracy guarantees of Theorem 4.1, we conclude by showing that it is possible to isolate each dictionary element by looking at only a polynomial number of ℓ-element intersections. This ensures we can recover every column in the dictionary in polynomial time with high probability. Specifically, for $\ell = \left\lceil \frac{\log 2K}{\log K/s} \right\rceil$ and $J = K \log^3 K$, we claim that with high probability, choosing $\mathcal{J}$ in Theorem 4.3 to be the first J/ℓ disjoint ℓ-element subsets is sufficient to recover every dictionary element at least once.

Theorem 4.4 (Polynomially-Many Intersections Suffice) *Let ℓ be the smallest integer such that $(s/K)^\ell < 1/2K$, and assume Ω_i, $i = 1, \dots, K$ are uniformly distributed among s-element subsets of $\{1, \dots, K\}$. For positive integer J and $j \in \{1, \dots, J/\ell\}$, define the non-intersecting ℓ-fold intersections Ω_j^ℓ as:*

$$\Omega_j^\ell = \bigcap_{p=1}^{\ell} \Omega_{(j-1)\ell + p}$$

Then as long as $J \geq K \log^3 K$, with high probability for every $k \in \{1, \dots, K\}$ there exists a $j \in \{1, \dots, J/\ell\}$ such that $\Omega_j^\ell = \{k\}$.

Theorem 4.4 can be proven by fixing k and noting that since the sets Ω_j^ℓ do not overlap, they will be independent, then calculating the probability that none of them contain k as a unique element of intersection; a union bound completes the proof. Details are in the appendix.

Theorem 4.4 guarantees that subspace intersection will recover every dictionary element at least once while requiring only $K \log^3 K$ intersections. Theorem 4.3 ensures subspace intersection (Algorithm 2) will correctly detect overlapping support with high probability, and will recover the associated dictionary element up to error ε by Theorem 4.2. Taken together, these results complete the proof of Theorem 4.1.

5. Proofs of Theoretical Guarantees

In this section, we provide sketches of the proof for the results outlined in the previous section.

5.1. Proofs for Subspace Recovery

The most involved of our theoretical guarantees is Theorem 4.2, which states that the subspaces recovered by Algorithm 1 are close to the true spanning subspaces. It will suffice to prove the result for a fixed index j, as the result will then follow by a union bound over $\{1, \dots, N\}$. In the proof, we use $j = 0$ to indicate this fixed index; although this is a minor abuse of notation as j ranges from 1 to N, this notation emphasizes the distinction between the sample $\mathbf{y}_0$ and the others, while the change from N to $N + 1$ samples is negligible.

We begin with a broad overview of our approach and the steps involved:

1. **The "Good Event."** We introduce the geometric properties of $\mathbf{D}$ required for proving our result, which will occur with high probability.
2. **Expectation computation.** We compute the expectation $E[\widehat{\Sigma}_0|\mathbf{D}]$ of the correlation-weighted covariance.
3. **Bounds on expectation error.** We separate the computed expectations into "signal" and "noise" terms and compute bounds on the noise terms. These bounds depend only on the parameters M, s, and K and not on the number of samples N.
4. **Bound on estimation error.** We bound the probability that the sample $\mathbf{Y}$ produces a correlation-weighted covariance that is far from its expectation. These bounds are controlled by the number of samples N.
5. **Subspace Comparison.** We convert the stated bounds on the correlation-weighted covariance to bounds on their s-leading subspaces.

5.1.1. THE GOOD EVENT

To prove our results, we define a "good event" $\mathcal{G}_0$ for fixed $\mathbf{x}_0$, which occurs with high probability. $\mathcal{G}_0$ describes sufficient geometric properties of $\mathbf{D}$ for the dictionary to be recovered successfully. This allows us to prove separately that $\mathcal{G}_0$ occurs with high probability, after which we can treat these properties as deterministic while proving our main results.

Many of the following facts can be inferred heuristically using the standard approximation that in high dimensions, uniformly distributed unit vectors are nearly distributed as $N(0, \frac{1}{M}\mathbf{I})$ random vectors. Specifically, since the $N(0, \frac{1}{M}I)$ distribution is rotationally invariant, we can assume the existence of a collection $\mathbf{w}_1, \dots, \mathbf{w}_K$ of independent $N(0, \mathbf{I})$ random vectors such that $\mathbf{d}_k = \frac{\mathbf{w}_k}{\|\mathbf{w}_k\|_2}$ for all $k = 1, \dots, K$. It was shown by Stam (1982) that the $N(0, \frac{1}{M}I)$ and uniform distributions converge in total variation; we will only use weaker properties that can be derived from standard concentration of measure results.

We now define $\mathcal{G}_0$ rigorously:

Definition 10 (Good Event for $\mathbf{x}_0$) *The* good event for $\mathbf{x}_0$*, denoted $\mathcal{G}_0$, is the event that the following conditions hold simultaneously. We use c and C for small and large positive constants respectively; constants vary between list items.*

$\mathcal{G}_0.1$ $\left\|\mathbf{D}\mathbf{D}^T - \frac{K}{M}\mathbf{I}\right\|_2 \leq \frac{C\sqrt{K}}{\sqrt{M}}$

$\mathcal{G}_0.2$ $\frac{cK}{\sqrt{M}} \leq \left\|\mathbf{D}\mathbf{D}^T\right\|_F \leq CK/\sqrt{M}$.

$\mathcal{G}_0.3$ *All eigenvalues of the matrix* $\sum_{k\in\Omega_0}\mathbf{d}_k\mathbf{d}_k^T$ *lie in an interval* (c, C) *for constants* $0 < c < C$.

$\mathcal{G}_0.4$ $\sup_{k\neq m}|\langle\mathbf{d}_k, \mathbf{d}_m\rangle| \leq \frac{C\log M}{\sqrt{M}}$.

$\mathcal{G}_0.5$ *Defining* $\tilde{\mathbf{y}}_0^k = \mathbf{y}_0 - \delta_{k\in\Omega_0}\mathbf{d}_k$, *we have* $\sup_k|\langle\tilde{\mathbf{y}}_0^k, \mathbf{d}_k\rangle| \leq \frac{C\sqrt{s}\log M}{\sqrt{M}}$ *for all* k.

$\mathcal{G}_0.6$ *Conditional on* $\mathbf{D}$ *and* $\mathbf{y}_0$, $\|\langle\mathbf{y}_0, \mathbf{y}\rangle\mathbf{y}\|_2 \leq \frac{Cs^{3/2}\log M}{\sqrt{M}}$ *with high probability.*

We refer to the event that the good events for all $i \in \{0, \ldots, N\}$ hold simultaneously simply as the "good event" $\mathcal{G} = \cap_{i=0}^N\mathcal{G}_i$. We have the following lemma, with proof in Appendix A.1:

Lemma 5.1 *The good event* $\mathcal{G}$ *holds with high probability.*

As N is at-most polynomial in M, to prove this lemma it suffices to prove that $\mathcal{G}_0$ holds with high probability.

5.1.2. EXPECTATION COMPUTATION

We begin with the following result on the expectations of the correlation-weighted covariance:

Lemma 5.2 (Expectation Computation) *Let* $\mathbf{v}_0 = \mathbf{D}\mathbf{D}^T\mathbf{y}_0$ *and* $\tilde{\mathbf{y}}_0^k = \mathbf{y}_0 - \delta_{k\in\Omega_0}\mathbf{d}_k$. *We have*

$$E[\langle\mathbf{y}_0, \mathbf{y}\rangle^2\mathbf{y}\mathbf{y}^T|\mathbf{D}] = \frac{s}{K}\left[\frac{s-1}{K-1}\mathbf{v}_0\mathbf{v}_0^T + \left(1 - \frac{2(s-1)}{K-1}\right)\sum_{k\in\Omega_0}\mathbf{d}_k\mathbf{d}_k^T\right] +$$
$$+ \frac{s}{K}\left[\left(1 - \frac{2(s-1)}{K-1}\right)\left(2\sum_{k\in\Omega_0}^K\left\langle\tilde{\mathbf{y}}_0^k, \mathbf{d}_k\right\rangle\mathbf{d}_k\mathbf{d}_k^T + \sum_{k=1}\left\langle\tilde{\mathbf{y}}_0^k, \mathbf{d}_k\right\rangle^2\mathbf{d}_k\mathbf{d}_k^T\right) + \left(\frac{s-1}{K-1}\sum_{k=1}^K\langle\mathbf{y}_0, \mathbf{d}_k\rangle^2\right)\mathbf{D}\mathbf{D}^T\right]. \tag{2}$$

Here the first row consists of signal terms, which form a matrix with s-leading eigenvalues approximately spanning the subspace $\mathcal{S}_0$. By contrast, the second row consists of nuisance terms which will ultimately not affect this subspace information: they either have small magnitude (the third and fourth terms) or they will be removed by covariance projection (the fifth term). The proof of this lemma is a computation and is deferred to Appendix A.2.

5.1.3. BOUNDING EXPECTATION ERROR

In this section we bound the difference between our desired biased covariance matrix and the actual expectation of $\widehat{\Sigma}_0$, resulting in error bounds intrinsic to the dimension M. Specifically, we prove the following lemma:

Lemma 5.3 (Expectation Error Bound) *With high probability,*

$$\left\|\frac{K}{s}E[\widehat{\Sigma}_0|\mathbf{D}] - \left(\frac{s-1}{K-1}\mathbf{v}_0\mathbf{v}_0^T + \sum_{k\in\Omega_0}\mathbf{d}_k\mathbf{d}_k^T + \left(\frac{s-1}{K-1}\sum_{k=1}^K\langle\mathbf{y}_0, \mathbf{d}_k\rangle^2\right)\mathbf{D}\mathbf{D}^T\right)\right\|_2 \leq \frac{CKs\log^2 M}{M^2} \tag{3}$$

and

$$\left\| \frac{K}{s} \left(E[\widehat{\Sigma}_0 | \mathbf{D}] - \mathbf{proj}_{\mathbf{D}\mathbf{D}^T}(E[\widehat{\Sigma}_0 | \mathbf{D}]) \right) - \left(\frac{s-1}{K-1} \mathbf{v}_0 \mathbf{v}_0^T + \sum_{k \in \Omega_0} \mathbf{d}_k \mathbf{d}_k^T \right) \right\|_2 \leq \frac{CKs \log^2 M}{M^2} \quad (4)$$

where $\mathbf{v}_0 = \mathbf{D}\mathbf{D}^T \mathbf{y}_0$.

In other words, up to scaling, the expectation of $\widehat{\Sigma}_0$ after covariance projection is approximately equal to the low-rank matrix $\frac{s-1}{K-1} \mathbf{v}_0 \mathbf{v}_0^T + \sum_{k \in \Omega_0} \mathbf{d}_k \mathbf{d}_k^T$. We begin by bounding the nuisance terms from 2:

$$2 \sum_{k \in \Omega_0} \left\langle \tilde{\mathbf{y}}_0^k, \mathbf{d}_k \right\rangle \mathbf{d}_k \mathbf{d}_k^T + \sum_{k=1}^{K} \left\langle \tilde{\mathbf{y}}_0^k, \mathbf{d}_k \right\rangle^2 \mathbf{d}_k \mathbf{d}_k^T.$$

By $\mathcal{G}_0.3$ and $\mathcal{G}_0.5$, the first term is bounded by $C\sqrt{s} \log M / \sqrt{M}$; for the second term, we have from $\mathcal{G}_0.1$, $\mathcal{G}_0.5$, and the triangle inequality that

$$\left\| \sum_{k=1}^{K} \left\langle \tilde{\mathbf{y}}_0^k, \mathbf{d}_k \right\rangle^2 \mathbf{d}_k \mathbf{d}_k^T \right\|_2 \leq \sup_{k \in K} \left\langle \tilde{\mathbf{y}}_0^k, \mathbf{d}_k \right\rangle^2 \|\mathbf{D}\mathbf{D}^T\|_2 \leq \frac{CKs \log^2 M}{M^2}.$$

This proves line 3 from Lemma 5.3.

We now incorporate the covariance projection step. We need to show that Frobenius projection of $E[\widehat{\Sigma}_0]$ onto $E\mathbf{y}\mathbf{y}^T = \frac{s}{K}\mathbf{D}\mathbf{D}^T$ removes the $\mathbf{D}\mathbf{D}^T$ term in equation 2 while contributing only a negligible factor elsewhere. Since projection is scale-invariant, it suffices to prove this for projection onto $\mathbf{D}\mathbf{D}^T$ in place of $E\mathbf{y}\mathbf{y}^T$. We prove the following lemma:

Lemma 5.4 (Projection Error Bound) *On* $\mathcal{G}_0$,

$$\left\| \frac{K}{s} \mathbf{proj}_{\mathbf{D}\mathbf{D}^T}(E[\widehat{\Sigma}_0 | \mathbf{D}]) - \left(\frac{s}{K} \sum_{k=1}^{K} \langle \mathbf{y}_0, \mathbf{d}_k \rangle^2 \right) \mathbf{D}\mathbf{D}^T \right\|_2 \leq \frac{Cs \log^2 M}{M} + \frac{CKs^2}{M^3}.$$

This lemma follows from $\mathcal{G}$, the triangle inequality, and the definition of Frobenius projection. A detailed proof can be found in Appendix A.3. This lemma implies that the error in Lemma 5.3, Equation 4 is dominated by the error from Equation 3, confirming Lemma 5.3.

5.2. Bounding Estimation Error

We now bound the resulting error from observing only the finite sample $\mathbf{Y}$ consisting of N random vectors, which will result in the following lemma:

Lemma 5.5 (Estimation Error Bound) *Recall that* $\widehat{\Sigma}_0^{proj} = \widehat{\Sigma}_0 - \mathbf{proj}_{\widehat{\Sigma}}(\widehat{\Sigma}_0)$. *Then with high probability*

$$\left\| \frac{K}{s} \widehat{\Sigma}_0^{proj} - \left(\frac{s-1}{K-1} \mathbf{v}_0 \mathbf{v}_0^T + \sum_{k \in \Omega_0} \mathbf{d}_k \mathbf{d}_k^T \right) \right\|_2 \leq \frac{CKs \log^2 M}{M^2} + \frac{CKs^2 \log^4 M}{M^{3/2}\sqrt{N}} + \frac{Cs^5 \log^5 M}{M^3 \sqrt{N}}.$$

In particular, estimation error will be small so long as $N \gg \max\left\{ \frac{s^{10} \log^{10} M}{M^6}, \frac{K^2 s^4 \log^8 M}{M^3} \right\}$.

We will employ the following theorem (Vershynin (2018), theorem 5.6.1) on covariance estimation, versions of which are well-known in the literature:

Theorem 5.1 (Vershynin (2018), Theorem 5.6.1, General Covariance Estimation (Tail Bound)) *Let* $\mathbf{z}$ *be a random vector in* $\mathbb{R}^M$*. Assume that, for some* $\kappa \geq 1$*,*

$$\|\mathbf{z}\|_2 \leq \kappa \sqrt{E[\|\mathbf{z}\|_2^2]}$$

almost surely. Then, for every positive integer N*,* $\{\mathbf{z}_i\}_{i=1}^N$ *i.i.d. copies of* $\mathbf{z}$*, and* $t \geq 0$*, we have:*

$$\left\| E\mathbf{z}\mathbf{z}^T - \frac{1}{N}\sum_{i=1}^N \mathbf{z}_i \mathbf{z}_i^T \right\|_2 \leq C\|E\mathbf{z}\mathbf{z}^T\|_2 \left(\sqrt{\frac{\kappa^2 M(\log M + t)}{N}} + \frac{\kappa^2 M(\log M + t)}{N} \right)$$

with probability at least $1 - 2\exp(-t)$*.*

We aim to apply this theorem to the correlation-weighted random vectors $\mathbf{z} = \langle \mathbf{y}_0, \mathbf{y} \rangle \mathbf{y}$ given $\mathbf{D}$, for which we will derive the following bounds on the expectation of $\|\mathbf{z}\|_2$:

Lemma 5.6 *Suppose that* $\mathcal{G}_0$ *holds. Then the following bounds hold:*

$$E\| \langle \mathbf{y}_0, \mathbf{y} \rangle \mathbf{y} | \mathbf{D}\|_2^2 \leq \frac{Cs^3}{M}$$

$$\|E \langle \mathbf{y}_0, \mathbf{y} \rangle^2 \mathbf{y}\mathbf{y}^T | \mathbf{D}\|_2 = \|E[\widehat{\Sigma}_0 | \mathbf{D}]\|_2 \leq \frac{Cs^3}{M^2}.$$

$$\|E \langle \mathbf{y}_0, \mathbf{y} \rangle^2 \mathbf{y}\mathbf{y}^T | \mathbf{D}\|_F = \|E[\widehat{\Sigma}_0 | \mathbf{D}]\|_F \leq \frac{Cs^3}{M^{3/2}}.$$

The result follows from $\mathcal{G}$ and similar computations to 5.2; details are in Appendix A.4.

With the use of a truncation trick, we can use these bounds to prove the following bound on the deviation of $\widehat{\Sigma}_0 = \frac{1}{N}\sum_{i=1}^N \langle \mathbf{y}_0, \mathbf{y}_i \rangle \mathbf{y}_i \mathbf{y}_i^T$ from its expectation:

Lemma 5.7 (Correlation-Weighted Covariance Estimation Bound) *Suppose that* $\mathcal{G}_0$ *holds. Then with high probability:*

$$\|\widehat{\Sigma}_0 - E[\widehat{\Sigma}_0 | \mathbf{D}]\|_2 = \left\| \frac{1}{N}\sum_{i=1}^N \langle \mathbf{y}_0, \mathbf{y}_i \rangle^2 \mathbf{y}_i \mathbf{y}_i^T - E[\widehat{\Sigma}_0 | \mathbf{D}] \right\|_2 \leq \frac{Cs^3 \log^2 M}{M^{3/2}\sqrt{N}}$$

This lemma follows by applying Theorem 5.1 to the truncated random vector $\mathbf{z} = \mathbb{1}_\omega \langle \mathbf{y}_0, \mathbf{y} \rangle \mathbf{y}$ where ω is the event that $\| \langle \mathbf{y}_0, \mathbf{y} \rangle \mathbf{y}\|_2 \leq Cs^{3/2} \log M / \sqrt{M}$; details are in the appendix. We immediately get the following bound for the Frobenius norm as a corollary:

Corollary 11 *Suppose that* $\mathcal{G}_0$ *holds. Then with high probability,*

$$\|\widehat{\Sigma}_0 - E[\widehat{\Sigma}_0 | \mathbf{D}]\|_F \leq \frac{Cs^3 \log^2 M}{M\sqrt{N}}$$

We can derive an analogous bound for the unweighted sample covariance:

Corollary 12 *Recall that* $\widehat{\Sigma} = \frac{1}{N}\mathbf{Y}\mathbf{Y}^T = \frac{1}{N}\sum_{i=1}^N \mathbf{y}_i\mathbf{y}_i^T$ *with* $E[\widehat{\Sigma}|\mathbf{D}] = E[\mathbf{y}\mathbf{y}^T|\mathbf{D}] = \frac{s}{K}\mathbf{D}\mathbf{D}^T$. *Then with high probability:*

$$\left\|\frac{1}{N}\mathbf{Y}\mathbf{Y}^T - E[\mathbf{y}\mathbf{y}^T|\mathbf{D}]\right\|_2 = \left\|\widehat{\Sigma} - \frac{s}{K}\mathbf{D}\mathbf{D}^T\right\|_2 \leq \frac{Cs\log M}{\sqrt{M}\sqrt{N}}$$

and

$$\left\|\frac{1}{N}\mathbf{Y}\mathbf{Y}^T - E[\mathbf{y}\mathbf{y}^T|\mathbf{D}]\right\|_F = \left\|\widehat{\Sigma} - \frac{s}{K}\mathbf{D}\mathbf{D}^T\right\|_F \leq \frac{Cs\log M}{\sqrt{N}}$$

This can be proven with the same technique used to prove Lemma 5.7, so a detailed proof is omitted.

Having proven the above bounds, proving convergence of $\widehat{\Sigma}_0^{\text{proj}}$ amounts to an extended computation with the triangle inequality, which can be found in Appendix A.6. Combined with Lemma 5.3, these computations complete the proof of Lemma 5.5.

5.3. Bounding Subspace Error

It remains to demonstrate that the first s eigenvectors of $\widehat{\Sigma}_0^{\text{proj}}$ span a subspace close to $\mathcal{S}_0 = \mathbf{span}\{\mathbf{d}_k\}_{k\in\Omega_0}$. It is easy to see that this holds in an asymptotic sense: by $\mathcal{G}_0.1$, we know $\mathbf{D}\mathbf{D}^T\mathbf{y}_0/\|\mathbf{D}\mathbf{D}^T\mathbf{y}_0\|_2 \to \mathbf{y}_0/\|\mathbf{y}_0\|_2$, while $\sum_{k\in\Omega_0}\mathbf{d}_k\mathbf{d}_k^T$ will be close to an identity matrix on the subspace $\mathcal{S}_0$. However, acquiring quantitative bounds on the subspace distance is more challenging and requires some technical machinery.

We begin by introducing the following notation: given a matrix $\mathbf{A} \in \mathbb{R}^{M\times M}$, let $\lambda_1(\mathbf{A}), \ldots, \lambda_M(\mathbf{A})$ be the eigenvalues of $\mathbf{A}$ in descending order. Similarly, for $m \in \{1, \ldots, M\}$ let $\mathcal{S}_m(\mathbf{A})$ denote the subspace spanned by the eigenvectors of $\mathbf{A}$ corresponding to $\lambda_1(\mathbf{A}), \ldots, \lambda_m(\mathbf{A})$.

We will prove the following result:

Theorem 5.2 *Let* $\mathcal{S}_0 = \mathbf{span}\{\mathbf{d}_k\}_{k\in\Omega_0}$. *As long as*

$$\tilde{\varepsilon} = \frac{Ks\log^2 M}{M^2} + \frac{Ks^2\log^4 M}{M^{3/2}\sqrt{N}} + \frac{s^5\log^5 M}{M^3\sqrt{N}},$$

is sufficiently small, then with high probability

$$\mathcal{D}\left(\mathcal{S}_0, \mathcal{S}_s\left(\widehat{\Sigma}_0^{proj}\right)\right) \leq \frac{C\sqrt{M}}{\sqrt{K}} + C\tilde{\varepsilon}.$$

We can recognize the $\tilde{\varepsilon}$ in this bound as the bound from Lemma 5.5, while the $\sqrt{M}/\sqrt{K}$ term arises from the deviation of $\mathbf{v}_0$ from a multiple of $\mathbf{y}_0$. To prove the theorem, we can write the result of Lemma 5.5 as

$$\widehat{\Sigma}_0^{\text{proj}} = \frac{s}{K}\left(\mathbf{v}_0\mathbf{v}_0^T + \sum_{k\in\Omega_0}\mathbf{d}_k\mathbf{d}_k^T + \mathcal{E}\right)$$

where $\mathcal{E}$ is a symmetric matrix with norm bounded by $C\tilde{\varepsilon}$ by Lemma 5.5. We will subsequently ignore the outer factor of s/K as this has no effect on the matrix's invariant subspaces.

We claim that the spectral properties of $\frac{s-1}{K-1}\mathbf{v}_0\mathbf{v}_0^T + \sum_{k\in\Omega_0}\mathbf{d}_k\mathbf{d}_k^T$ are essentially the same as those of $\mathbf{y}_0\mathbf{y}_0^T + \sum_{k\in\Omega_0}\mathbf{d}_k\mathbf{d}_k^T$, which clearly has s-leading subspace $\mathcal{S}_0$. If this holds, it then follows from a routine application of the Davis-Kahan theorem on $\mathcal{E}$ that $\widehat{\Sigma}_0^{\text{proj}}$ have lead s eigenvectors approximately spanning $\mathcal{S}_0$ with an additional error proportional to $\tilde{\varepsilon}$ as in the theorem. Therefore, proving the theorem reduces to proving the following lemma:

Lemma 5.8 *Let*

$$\mathbf{B} = \frac{s-1}{K-1}\mathbf{v}_0\mathbf{v}_0^T + \sum_{k\in\Omega_0}\mathbf{d}_k\mathbf{d}_k^T$$

Then for all $i \leq s$, $\lambda_i(\mathbf{B}) > c$ and for all $i > s$, $\lambda_i(\mathbf{B}) \leq C\sqrt{M}/\sqrt{K}$. Moreover, recalling that $\mathcal{S}_s(\mathbf{B})$ is the subspace spanned by the leading s eigenvectors of $\mathbf{B}$, we have

$$\mathcal{D}(\mathcal{S}_0, \mathcal{S}_s(\mathbf{B})) \leq \frac{C\sqrt{M}}{\sqrt{K}}.$$

The key ingredients in our proof are the following variants of Weyl's and the Davis-Kahan Theorems:

Theorem 5.3 (Weyl (1912)) *Let $\mathbf{A}$ and $\mathcal{E}$ be symmetric $M \times M$ matrices, and let $\lambda_i(\cdot)$ represent the i-th eigenvalue in descending order. Then for all $i = 1, \ldots, M$:*

$$\lambda_i(\mathbf{A}) + \lambda_M(\mathcal{E}) \leq \lambda_i(\mathbf{A}+\mathcal{E}) \leq \lambda_i(\mathbf{A}) + \lambda_1(\mathcal{E}).$$

It follows that

$$|\lambda_i(\mathbf{A}) - \lambda_i(\mathbf{A}+\mathcal{E})| \leq \|\mathcal{E}\|_2.$$

Theorem 5.4 (Davis and Kahan (1970)) *Let $A = E_0A_0E_0^T + E_1A_1E_1^T$ and $A+\mathcal{E} = F_0\Lambda_0F_0^T + F_1\Lambda_1F_1^T$ be symmetric matrices with $[E_0, E_1]$ and $[F_0, F_1]$ orthogonal. If the eigenvalues of A_0 are contained in an interval (a, b), and the eigenvalues of Λ_1 are excluded from the interval $(a-\delta, b+\delta)$ for some $\delta > 0$, then*

$$\|F_1^TE_0\|_2 \leq \frac{\|F_1^T\mathcal{E}E_0\|_2}{\delta} \leq \frac{\|\mathcal{E}\|_2}{\delta}$$

for any unitarily invariant matrix norm $\|\cdot\|$.

Put in the language of subspace distances, this theorem immediately gives the following corollary:

Corollary 13 *Let $A = E_0A_0E_0^T + E_1A_1E_1^T$, $A+\mathcal{E} = F_0\Lambda_0F_0^T + F_1\Lambda_1F_1^T$, and δ be as in Theorem 5.4. If $\mathcal{S}_A$ and $\mathcal{S}_{A+\mathcal{E}}$ are the subspaces spanned by columns of E_0 and F_0, respectively, then*

$$\mathcal{D}(\mathcal{S}_A, \mathcal{S}_{A+\mathcal{E}}) \leq \frac{\|\mathcal{E}\|_2}{\delta}.$$

We now apply Weyl's theorem and the Davis-Kahan theorem to our particular situation via the following lemmas. The constants we derive are likely suboptimal, but adequate for our purposes. We have the following lemma:

Lemma 5.9 *Let $\mathbf{A}$ be a rank-s symmetric positive-semidefinite matrix with nonzero eigenvalues satisfying:*

$$0 < \alpha \leq \lambda_1(\mathbf{A}), \ldots, \lambda_s(\mathbf{A}) \leq \beta$$

for some constants $\alpha < \beta$. Let $\mathbf{v} \in \mathcal{S}$ and $\mathbf{u} \in \mathcal{S}^\perp$ be vectors such that $\|\mathbf{v}\| = \|\mathbf{u}\|$ and $\|\mathbf{v}+\varepsilon\mathbf{u}\|_2 = 1$. For any $\varepsilon \in (0, \alpha/54)$ and $Z > \max\{2\beta, 1\}$, define the matrix $\mathbf{B}$ as:

$$\mathbf{B} = Z(\mathbf{v}+\varepsilon\mathbf{u})(\mathbf{v}+\varepsilon\mathbf{u})^T + \mathbf{A}$$

Then for all $i > s$, $\lambda_i(\mathbf{B}) \leq 24\varepsilon$ and

$$\mathcal{D}(\mathcal{S}(\mathbf{A}), \mathcal{S}(\mathbf{B})) \leq \frac{78\beta\varepsilon}{\alpha^2}.$$

This lemma is mainly a restatement of Weyl's Theorem and the Davis-Kahan Theorem to a specific situation; a proof is in Appendix A.7. We apply this lemma to prove Lemma 5.8:

Proof We now apply Lemma 5.9 to $\mathbf{B} = \frac{s-1}{K-1}\mathbf{v}_0\mathbf{v}_0^T + \sum_{k\in\Omega_0}\mathbf{d}_k\mathbf{d}_k^T$. Since $\sum_{k\in\Omega_0}\mathbf{d}_k\mathbf{d}_k^T$ is symmetric, rank s, and symmetric positive-semidefinite with eigenvalues in $[c, C]$ by $\mathcal{G}_0.3$, it is a valid choice for $\mathbf{A}$; it remains to designate $\mathbf{v}$ and $\mathbf{u}$. To this end, we write:

$$\mathbf{v}_0 = \mathbf{z}_0 + \mathbf{u}_0$$

where $\mathbf{z}_0$ is the component of $\mathbf{v}$ lying in $\mathcal{S}_0$ and $\mathbf{u}_0$ is the component in its orthogonal complement. Since $\mathbf{v}_0 = \mathbf{D}\mathbf{D}^T\mathbf{y}_0$, we have

$$\mathbf{v}_0 = \frac{K}{M}\mathbf{y}_0 + \left(\mathbf{D}\mathbf{D}^T - \frac{K}{M}\mathbf{I}\right)\mathbf{y}_0.$$

Thus, since $\mathbf{y}_0 \in \mathcal{S}_0$, applying $\mathcal{G}_0.1$ yields:

$$\|\mathbf{v}_0\|_2 \geq \|\mathbf{z}_0\|_2 \geq \frac{cK}{M}\|\mathbf{y}_0\|_2 \geq \frac{cK\sqrt{s}}{M}$$

and moreover that $\|\mathbf{u}_0\|_2 \leq C\sqrt{Ks}/\sqrt{M}$. We can thus write

$$\mathbf{v}_0 = \|\mathbf{v}_0\|_2\left(\frac{\mathbf{z}_0}{\|\mathbf{v}_0\|} + \frac{\mathbf{u}_0}{\|\mathbf{v}_0\|}\right) = \|\mathbf{v}_0\|_2\left(\frac{\mathbf{z}_0}{\|\mathbf{v}_0\|} + \left(\frac{\|\mathbf{u}_0\|_2}{\|\mathbf{z}\|_2}\right)\frac{\|\mathbf{z}_0\|\mathbf{u}_0}{\|\mathbf{v}_0\|_2\|\mathbf{u}_0\|}\right).$$

We can now apply Lemma 5.9 with $Z = \frac{s-1}{K-1}\|\mathbf{v}_0\|_2^2$, $\mathbf{v} = \mathbf{z}_0/\|\mathbf{v}_0\|_2$, $\mathbf{u} = \frac{\|\mathbf{z}\|_2\mathbf{u}_0}{\|\mathbf{v}_0\|_2\|\mathbf{u}_0\|_2}$, and $\varepsilon = \frac{\|\mathbf{u}_0\|_2}{\|\mathbf{z}_0\|_2}$, which tells us that

$$\mathcal{D}(\mathcal{S}_0, \mathcal{S}_s(\mathbf{B})) \leq \frac{C\|\mathbf{u}_0\|_2}{\|\mathbf{z}_0\|_2} \leq \frac{C\sqrt{M}}{\sqrt{K}}$$

and that all eigenvalues past the s-th are bounded by $C\sqrt{M}/\sqrt{K}$, as desired. ■

This completes the proof of Theorem 4.2.

5.4. Guarantees for subspace intersection

We now prove Theorem 4.4, which states that the intersection step with close enough estimated subspaces $\widehat{\mathcal{S}}_1, \ldots, \widehat{\mathcal{S}}_\ell$ accurately approximates the intersection of true subspaces $\mathcal{S}_1, \ldots, \mathcal{S}_\ell$. The result follows from the following lemma:

Lemma 5.10 *Suppose Ω_1, Ω_2 are at-most-s-element subsets of $\{1, 2, \ldots, K\}$ such that $\Omega_1 \cap \Omega_2 = \emptyset$. Let $\mathcal{S}_1 = \mathbf{span}\{\mathbf{d}_k\}_{k\in\Omega_1}$ and $\mathcal{S}_j = \mathbf{span}\{\mathbf{d}_m\}_{m\in\Omega_2}$. Then with high probability,*

$$\mathcal{D}(\mathcal{S}_1, \mathcal{S}_2) \geq 1 - \frac{Cs}{M}$$

The proof of this lemma can be found in Appendix A.8, and is derived from the fact that the vectors in a random $\mathcal{U}$-distributed dictionary are nearly orthogonal.

It follows from Lemma 5.10 that for large enough M, two subspaces $\mathcal{S}_i$ and $\mathcal{S}_j$ contain vectors closer than a constant threshold if and only if they share support ($|\Omega_i \cap \Omega_j| \geq 1$). Since this result

holds with high probability, this holds for all pairs i, j simultaneously. Since this holds for pairwise intersections, the analogous result automatically holds for ℓ-wise intersections as well. Theorem 4.3 now follows almost immediately: by Theorem 4.2, $|\mathcal{D}(\widehat{\mathcal{S}}_i, \widehat{\mathcal{S}}_j) - \mathcal{D}(\mathcal{S}_i, \mathcal{S}_j)| \leq C\varepsilon$, meaning $\mathcal{D}(\widehat{\mathcal{S}}_i, \widehat{\mathcal{S}}_j)$ well approximates $\mathcal{D}(\mathcal{S}_i, \mathcal{S}_j)$. Therefore by Lemma 5.10 above, $\mathcal{D}(\widehat{\mathcal{S}}_i, \widehat{\mathcal{S}}_j)$ will be small if and only if samples $\mathbf{y}_i$ and $\mathbf{y}_j$ share support. Theorem 4.2 then provides the quantitative bound of order ε for a single intersection; since there are at most ℓ intersections, the triangle inequality bounds the total error at $C\ell\varepsilon \leq C\varepsilon \log M$. This completes the proof of Theorem 4.3.

6. Numerical Simulations

All code used in these simulations, including an open-source Python implementation of SSDL, is publicly available at https://github.com/sew347/spectral_dict_learn. In this section, we supplement our theoretical results with numerical simulations. We consider two metrics. The first is convergence in angle: how close is $\left|\left\langle \hat{\mathbf{d}}, \mathbf{d} \right\rangle\right|$ to 1. This measure can always be converted to the error in L_2 norm by the identity $\|\hat{\mathbf{d}} - \mathbf{d}\|_2^2 = 2 - 2\left\langle \hat{\mathbf{d}}, \mathbf{d} \right\rangle$ (up to possible differences in sign). The second main metric is the proportion of *false recoveries*: a *false recovery* occurs when subspace intersection either returns a vector when a dictionary element does not exist in the intersection of true subspaces, or when subspace intersection fails to return a vector when one is in the intersection of true subspaces.

To test our theoretical hypotheses, in Figure 1, we show the maximum sparsity that can be tolerated by SDL while remaining above a certain accuracy threshold. Specifically, this figure shows the highest sparsity for which our test returns average angular accuracy above 0.95 with false recovery proportion below 0.08. Tests were run on the first 50 subspaces over 5 dictionaries in each dimension. To ensure a nontrivial number of overlaps, the support of each of the first 50 samples was seeded so that $1 \in \Omega_1, \Omega_2, 2 \in \Omega_3, \Omega_4, \ldots, 25 \in \Omega_{49}, \Omega_{50}$. In this test $K = 2M$, and N was pegged to $N = 30000$ for $M = 500$ then allowed to grow as different powers of M. The results confirm the findings of Theorem 4.1: sparsity growth is linear when $N \sim M^4$, but slower than linear when $N \sim M^3$ or $N \sim M^2$.

7. Conclusion

We introduced SSDL as an efficient method for recovering dictionaries from high-dimensional samples even in the linear sparsity regime. In this regime, SSDL achieves decaying errors in M in polynomial time, improving on the best-known provable alternatives. Reproducible numerical simulations validate these results.

Our initial research on SSDL suggests several avenues for future research. First, we suspect it is possible to reduce the sample complexity from approximately M^4 to closer to M^3 by replacing the covariance projection step with a more sophisticated method for controlling the dominant term in 2. This is because the worst term in the error in Theorem 4.2 comes from estimating the covariance matrix in the Frobenius rather than in the operator norm, which is known to have worse sample complexity ($N \sim M^2$) than estimation in the operator norm ($N \sim M$). We are also interested in investigating to what degree the uniformity assumption in the generation of support sets Ω_j can be relaxed, to allow for more general sampling distributions. As our method involves computation of a

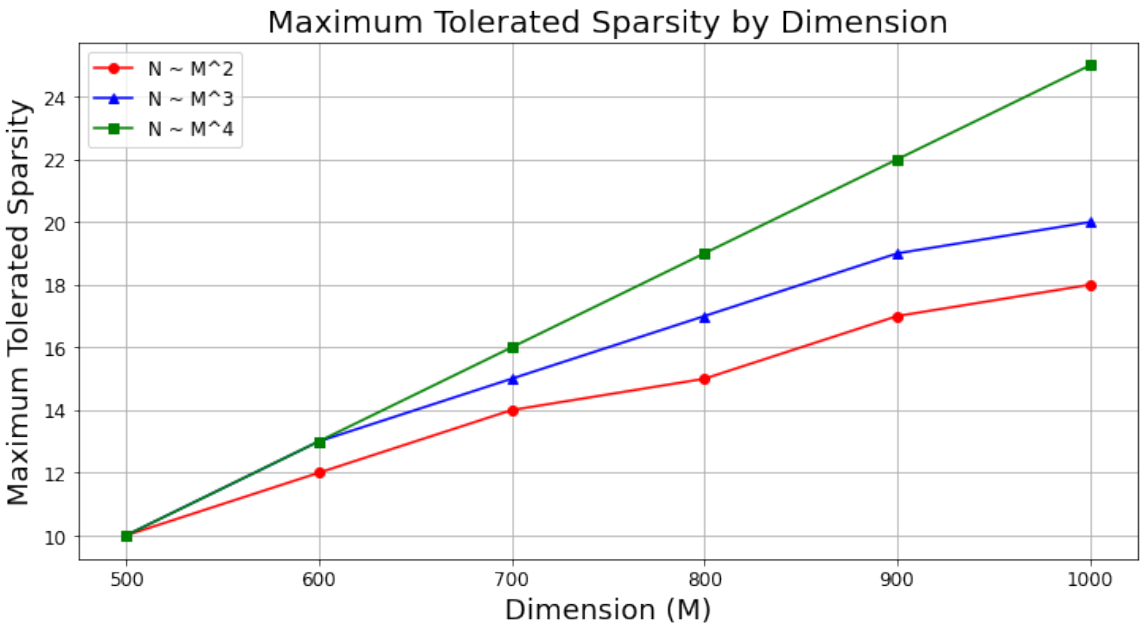

Figure 1: **Maximum tolerated sparsity.** As predicted by Theorem 4.1, $N \sim M^4$ allows for linear sparsity growth in M, while fewer samples result in sublinear growth.

fourth-order statistic, it bears some similarity to the recently introduced ℓ_4-based dictionary learning methods; future work will seek to study these connections in greater detail.

References

Alekh Agarwal, Animashree Anandkumar, Prateek Jain, and Praneeth Netrapalli. Learning sparsely used overcomplete dictionaries via alternating minimization. *SIAM Journal on Optimization*, 26(4):2775–2799, 2016. doi: 10.1137/140979861. URL https://doi.org/10.1137/140979861.

Alekh Agarwal, Animashree Anandkumar, and Praneeth Netrapalli. A clustering approach to learning sparsely used overcomplete dictionaries. *IEEE Transactions on Information Theory*, 63(1):575–592, 2017. doi: 10.1109/TIT.2016.2614684.

Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Multi-task feature learning. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006. URL https://proceedings.neurips.cc/paper/2006/file/0afa92fc0f8a9cf051bf2961b06ac56b-Paper.pdf.

Sanjeev Arora, Rong Ge, and Ankur Moitra. New algorithms for learning incoherent and overcomplete dictionaries. *Journal of Machine Learning Research*, 35, 08 2013.

Sanjeev Arora, Aditya Bhaskara, Rong Ge, and Tengyu Ma. More algorithms for provable dictionary learning, 2014. URL https://arxiv.org/abs/1401.0579.

Afonso S. Bandeira, Matthew Fickus, Dustin G. Mixon, and Percy Wong. The road to deterministic matrices with the restricted isometry property. *Journal of Fourier Analysis and Applications*, 19(6):1123–1149, 2013. doi: 10.1007/s00041-013-9293-2. URL https://doi.org/10.1007/s00041-013-9293-2.

Boaz Barak, Jonathan A. Kelner, and David Steurer. Dictionary learning and tensor decomposition via the sum-of-squares method. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, STOC '15, page 143–151, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450335362. doi: 10.1145/2746539.2746605. URL https://doi.org/10.1145/2746539.2746605.

E.J. Candes and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005. doi: 10.1109/TIT.2005.858979.

Emmanuel Candés, Justin Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59, 08 2006. doi: 10.1002/cpa.20124.

Niladri S. Chatterji and Peter L. Bartlett. Alternating minimization for dictionary learning: Local convergence guarantees. arXiv, 2017. doi: 10.48550/ARXIV.1711.03634. URL https://arxiv.org/abs/1711.03634.

Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998. doi: 10.1137/S1064827596304010. URL https://doi.org/10.1137/S1064827596304010.

Chandler Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970. ISSN 00361429. URL http://www.jstor.org/stable/2949580.

D.L. Donoho, M. Elad, and V.N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory*, 52(1):6–18, 2006. doi: 10.1109/TIT.2005.860430.

Devdatt P. Dubhashi and Desh Ranjan. Balls and bins: A study in negative dependence. *BRICS Report Series*, 3(25), Jan. 1996. doi: 10.7146/brics.v3i25.20006. URL https://tidsskrift.dk/brics/article/view/20006.

Michael Elad. *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Springer Publishing Company, Incorporated, 1st edition, 2010. ISBN 144197010X.

D. L. Hanson and F. T. Wright. A Bound on Tail Probabilities for Quadratic Forms in Independent Random Variables. *The Annals of Mathematical Statistics*, 42(3):1079 – 1083, 1971. doi: 10.1214/aoms/1177693335. URL https://doi.org/10.1214/aoms/1177693335.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302 – 1338, 2000. doi: 10.1214/aos/1015957395. URL https://doi.org/10.1214/aos/1015957395.

Tengyu Ma, Jonathan Shi, and David Steurer. Polynomial-time tensor decompositions with sum-of-squares, 2016. URL https://arxiv.org/abs/1610.01980.

B A Olshausen and D J Field. Natural image statistics and efficient coding. *Network: Computation in Neural Systems*, 7(2):333–339, 1996a. doi: 10.1088/0954-898X_7_2_014. URL https://doi.org/10.1088/0954-898X_7_2_014. PMID: 16754394.

Bruno A. Olshausen and David J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996b. doi: 10.1038/381607a0. URL https://doi.org/10.1038/381607a0.

Bruno A. Olshausen and David J. Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research*, 37(23):3311–3325, 1997. ISSN 0042-6989. doi: https://doi.org/10.1016/S0042-6989(97)00169-7. URL https://www.sciencedirect.com/science/article/pii/S0042698997001697.

Marc' aurelio Ranzato, Y-lan Boureau, and Yann Cun. Sparse feature learning for deep belief networks. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper/2007/file/c60d060b946d6dd6145dcbad5c4ccf6f-Paper.pdf.

Daniel Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. *IJCAI International Joint Conference on Artificial Intelligence*, 23, 06 2012.

A. J. Stam. Limit theorems for uniform distributions on spheres in high-dimensional euclidean spaces. *Journal of Applied Probability*, 19(1):221–228, 1982. ISSN 00219002. URL http://www.jstor.org/stable/3213932.

Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere i: Overview and the geometric picture. *IEEE Transactions on Information Theory*, 63(2):853–884, 2017a. doi: 10.1109/TIT.2016.2632162.

Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere II: Recovery by riemannian trust-region method. *IEEE Transactions on Information Theory*, 63(2):885–914, feb 2017b. doi: 10.1109/tit.2016.2632149. URL https://doi.org/10.1109%2Ftit.2016.2632149.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

Martin Vetterli and Jelena Kovačevic. *Wavelets and Subband Coding*. Prentice-Hall, Inc., USA, 1995. ISBN 0130970808.

Hermann Weyl. Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). *Mathematische Annalen*, 71(4):441–479, 1912. doi: 10.1007/BF01456804. URL https://doi.org/10.1007/BF01456804.

Yuexiang Zhai, Hermish Mehta, Zhengyuan Zhou, and Yi Ma. Understanding l4-based dictionary learning: Interpretation, stability, and robustness. In *International Conference on Learning Representations*, 2020a. URL https://openreview.net/forum?id=SJeY-1BKDS.

Yuexiang Zhai, Zitong Yang, Zhenyu Liao, John Wright, and Yi Ma. Complete dictionary learning via l4-norm maximization over the orthogonal group. *Journal of Machine Learning Research*, 21 (165):1–68, 2020b. URL http://jmlr.org/papers/v21/19-755.html.

Appendix A. Technical Proofs

In this section, we include proofs of technical lemmas which we omitted from the main text.

A.1. Proof of Lemma 5.1

We begin with the proof of our main probability result, that $\mathcal{G}$ occurs with high probability.

Lemma 5.1 *The good event $\mathcal{G}$ holds with high probability.*

As noted in the main text, it suffices by a union bound to prove the result for $\mathcal{G}_0$. We prove each item in Definition 10 as a separate lemma. In many of these we will make use of the relationship $\mathbf{d}_k = \mathbf{w}_k/\|\mathbf{w}_k\|_2$ where $\mathbf{w}_k$ are i.i.d. random vectors with independent $N(0,1)$ entries. Expressed in matrix form, we have $\mathbf{D} = \mathbf{WN}$ where $\mathbf{W}$ is a matrix with columns $\mathbf{w}_k$ and $\mathbf{N}$ is a diagonal matrix with k-th diagonal entry $1/\|\mathbf{w}_k\|_2$. We will make use of the following lemma:

Lemma A.1 *Let $\mathbf{w}$ be an $N(0,\mathbf{I})$ random vector. Then with high probability, $\sqrt{M} - \log M \leq \|\mathbf{w}\|_2 \leq \sqrt{M} + \log M$.*

Proof By definition, entries of $\mathbf{w}$, denoted w_m for $m = 1, \ldots, M$, are i.i.d. normal random variables with variance 1. It follows that $\|\mathbf{w}\|_2^2 = \sum_{m=1}^M w_m^2$ will be distributed as a chi-squared random variable with M degrees of freedom. Such a variable obeys the following concentration inequalities (Laurent and Massart, 2000):

$$\begin{aligned} \mathbb{P}\left(\|\mathbf{w}\|_2^2 \geq M + \sqrt{Mt} + 2t\right) &\leq e^{-t} \\ \mathbb{P}\left(\|\mathbf{w}\|_2^2 \leq M - \sqrt{Mt}\right) &\leq e^{-t} \end{aligned} \tag{5}$$

which for $t \leq M$ yields the symmetric bound:

$$\mathbb{P}\left(\left|\|\mathbf{w}\|_2^2 - M\right| \geq 3\sqrt{Mt}\right) \leq e^{-t} \tag{6}$$

To convert this to a bound on $\|\mathbf{w}\|_2$, we use the fact that for any $z, \delta \geq 0$, $|z-1| \geq \delta$ implies $|z^2 - 1| \geq \max\{\delta, \delta^2\}$. Accordingly,

$$\mathbb{P}\left(\left|\frac{1}{\sqrt{M}}\|\mathbf{w}\|_2 - 1\right| \geq \delta\right) \leq \mathbb{P}\left(\left|\frac{1}{M}\|\mathbf{w}\|_2^2 - 1\right| \geq \max\{\delta, \delta^2\}\right)$$

Setting $\delta = \log M/\sqrt{M}$ and applying 6 with $t = (\log^2 M)/9 \leq M$, we have:

$$\mathbb{P}\left(|\|\mathbf{w}\|_2 - \sqrt{M}| \leq \log M\right) \leq \mathbb{P}\left(|\|\mathbf{w}\|_2^2 - M| \leq \sqrt{M}\log M\right) \leq e^{-(\log^2 M)/9}$$

which proves the result. ■

Lemma A.2 ($\mathcal{G}_0$.1) $\left\|\mathbf{D}\mathbf{D}^T - \frac{K}{M}\mathbf{I}\right\|_2 \leq \frac{C\sqrt{K}}{\sqrt{M}}$.

Proof We use the normal identity $\mathbf{D} = \mathbf{W}\mathbf{N}$ to write:

$$\left\|\mathbf{D}\mathbf{D}^T - \frac{K}{M}\mathbf{I}\right\|_2 = \left\|\mathbf{W}\mathbf{N}^2\mathbf{W}^T - \frac{K}{M}\mathbf{I}\right\|_2.$$

By the triangle inequality,

$$\left\|\mathbf{W}\mathbf{N}^2\mathbf{W}^T - \frac{K}{M}\mathbf{I}\right\|_2 \leq \left\|\frac{1}{M}\mathbf{W}\mathbf{W}^T - \frac{K}{M}\mathbf{I}\right\|_2 + \left\|\mathbf{W}\mathbf{N}^2\mathbf{W}^T - \frac{1}{M}\mathbf{W}\mathbf{W}^T\right\|_2 \tag{7}$$

By Lemma A.1, with high probability $M - \sqrt{M}\log M \leq \|\mathbf{w}_k\|_2^2 \leq M + \sqrt{M}\log M$ for all k. Therefore:

$$\left\|\mathbf{W}\mathbf{N}^2\mathbf{W}^T - \frac{1}{M}\mathbf{W}\mathbf{W}^T\right\|_2 \leq \|\mathbf{W}\|_2^2\left\|\mathbf{N}^2 - \frac{1}{M}\mathbf{I}\right\|_2 \leq \frac{C\|\mathbf{W}\|_2^2}{M^{3/2}}$$

We now bound $\left\|\mathbf{W}\mathbf{W}^T - K\mathbf{I}\right\|_2$. We employ an ε-net argument (e.g. Vershynin, 2018). We let $\mathcal{M}$ be a $1/4$-net on the unit sphere; that is, $\mathcal{M}$ is a finite set such that every point on the unit sphere is at most Euclidean distance $1/4$ from a point in $\mathcal{M}$. It is a fact of ε-nets that for $\varepsilon = 1/4$, we can choose $\mathcal{M}$ such that $|\mathcal{M}| \leq 9^M$. Moreover:

$$\left\|\mathbf{W}\mathbf{W}^T - K\mathbf{I}\right\|_2 = \sup_{\|\mathbf{z}\|=1} \mathbf{z}^T\left(\mathbf{W}\mathbf{W}^T - K\mathbf{I}\right)\mathbf{z} \leq 2\max_{\mathbf{z}\in\mathcal{M}} \mathbf{z}^T\left(\mathbf{W}\mathbf{W}^T - K\mathbf{I}\right)\mathbf{z}$$

We now fix $\mathbf{z} \in \mathcal{M}$. We have:

$$\mathbf{z}^T\left(\mathbf{W}\mathbf{W}^T - K\mathbf{I}\right)\mathbf{z} = \sum_{k=1}^{K}\langle\mathbf{z}, \mathbf{w}_k\rangle^2 - K.$$

By rotational invariance, $\langle\mathbf{z}, \mathbf{w}_k\rangle \sim N(0,1)$, so we can concentrate this sum using the chi-squared concentration inequalities 5. Choosing $t = CM$, we have:

$$\left|\left(\sum_{k=1}^{K}\langle\mathbf{z}, \mathbf{w}_k\rangle^2\right) - K\right| \leq \sqrt{CKM} + 2CM$$

with probability at least $1 - e^{-CM}$. Unfixing $\mathcal{M}$ by a union bound, this holds for all $\mathbf{z} \in \mathcal{M}$ with probability at least $1 - 9^M e^{-CM}$, which represents high probability for sufficiently large C. Thus with high probability, changing C if necessary,

$$\left\|\mathbf{W}\mathbf{W}^T - K\mathbf{I}\right\|_2 \leq C\sqrt{KM}$$

Plugging back into 7, noting that the above implies $\|\mathbf{W}\|_2^2 \leq CK$, we conclude:

$$\left\|\mathbf{D}\mathbf{D}^T - \frac{K}{M}\mathbf{I}\right\|_2 \leq \frac{C\sqrt{K}}{\sqrt{M}} + \frac{CK}{M^{3/2}}$$

with high probability. Since $K/M^{3/2} \ll \sqrt{K}/\sqrt{M}$, this completes the proof. ■

Lemma A.3 ($\mathcal{G}_0$.2) $\frac{cK}{\sqrt{M}} \leq \left\|\mathbf{D}\mathbf{D}^T\right\|_F \leq CK/\sqrt{M}$.

Proof This follows immediately from $\mathcal{G}_0$.1 and the inequality $\|\mathbf{D}\mathbf{D}^T\|_F \leq \sqrt{M}\|\mathbf{D}\mathbf{D}^T\|_2$. ■

Lemma A.4 ($\mathcal{G}_0$.3) *All eigenvalues of the matrix $\sum_{k\in\Omega_0} \mathbf{d}_k\mathbf{d}_k^T$ lie in an interval (c, C) for constants $0 < c < C$.*

Proof We can write $\sum_{k\in\Omega_0} \mathbf{d}_k\mathbf{d}_k^T = \mathbf{D}_{\Omega_0}\mathbf{D}_{\Omega_0}^T$ where $\mathbf{D}_{\Omega_0}$ is the submatrix of $\mathbf{D}$ with columns indexed by Ω_0. Since the nonzero eigenvalues of $\mathbf{D}_{\Omega_0}\mathbf{D}_{\Omega_0}^T$ are the same as those of $\mathbf{D}_{\Omega_0}^T\mathbf{D}_{\Omega_0}$, we may prove the bound for the latter matrix. By Lemma A.1, it suffices to prove the result for the matrix $\frac{1}{M}\mathbf{W}_{\Omega_0}^T\mathbf{W}_{\Omega_0}$.

As in the proof of $\mathcal{G}_0$.1, we employ an ε-net argument. Let $\mathcal{M}$ be a $1/4$-net of the unit sphere in $\mathbb{R}^s$, which can be chosen with at most 9^s elements. For fixed $\mathbf{z} \in \mathcal{M}$, we have:

$$\mathbf{z}^T\left(\frac{1}{M}\mathbf{W}_{\Omega_0}^T\mathbf{W}_{\Omega_0} - \mathbf{I}\right)\mathbf{z}$$

Since $\mathbf{W}$ has i.i.d. $N(0,1)$ entries, $\mathbf{z}^T\mathbf{W}_{\Omega_0}^T\mathbf{W}_{\Omega_0}\mathbf{z}$ will be distributed as a chi-squared random variable with M degrees of freedom. Accordingly, by 5, we have that

$$|\mathbf{z}^T\mathbf{W}_{\Omega_0}^T\mathbf{W}_{\Omega_0}\mathbf{z} - M| \leq \sqrt{CMs} + 2Cs$$

with probability at least $1 - e^{-Cs}$. Unfixing $\mathbf{z}$ by a union bound, this holds for all $\mathbf{z} \in \mathcal{M}$ with probability at least $1 - 9^s e^{-Cs}$, which is high probability for large enough C. We conclude that with high probability,

$$\left\|\frac{1}{M}\mathbf{W}_{\Omega_0}^T\mathbf{W}_{\Omega_0} - \mathbf{I}\right\|_2 \leq \frac{C\sqrt{s}}{\sqrt{M}}$$

which implies the result. ■

Lemma A.5 ($\mathcal{G}_0$.4) $\sup_{k\neq m} |\langle \mathbf{d}_k, \mathbf{d}_m\rangle| \leq \frac{C\log M}{\sqrt{M}}$.

Proof Since $k \neq m$, $\mathbf{d}_k$ and $\mathbf{d}_m$ are independent. Then by rotational invariance, we can treat $\mathbf{d}_m$ as a fixed vector, say the first coordinate vector $\mathbf{e}_1$: $\langle \mathbf{d}_k, \mathbf{d}_m\rangle \sim \langle \mathbf{d}_k, \mathbf{e}_1\rangle$. Writing $\mathbf{d}_k = \mathbf{w}_k/\|\mathbf{w}_k\|_2$ for $\mathbf{w}_k \sim N(0, \mathbf{I})$, we have $\langle \mathbf{d}_k, \mathbf{e}_1\rangle = \frac{1}{\|\mathbf{w}\|_2}\langle \mathbf{w}_k, \mathbf{e}_1\rangle \sim \frac{1}{\|\mathbf{w}\|_2} \times N(0,1)$. It is known (see (Vershynin, 2018), proposition 2.1.2) that a normally distributed random variable $Z_\sigma \sim N(0, \sigma^2)$ obeys the concentration inequality:

$$\mathbb{P}(Z_\sigma \geq t) \leq \frac{\sigma}{\sqrt{2\pi}t}\exp\left(\frac{-t^2}{2\sigma^2}\right) \tag{8}$$

Applying this to $\langle \mathbf{w}_k, \mathbf{e}_1\rangle$, we have $|\langle \mathbf{w}_k, \mathbf{e}_1\rangle| \leq C\log M$ with high probability. By Lemma A.1, we know that $\|\mathbf{w}_k\|_2 \geq c/\sqrt{M}$ with high probability, so we conclude the result. ■

Lemma A.6 ($\mathcal{G}_0$.5) *Defining $\tilde{\mathbf{y}}_0^k = \mathbf{y}_0 - \delta_{k\in\Omega_0}\mathbf{d}_k$, we have $\sup_k |\langle \tilde{\mathbf{y}}_0^k, \mathbf{d}_k\rangle| \leq \frac{C\sqrt{s}\log M}{\sqrt{M}}$ for all k.*

Proof By definition of $\mathbf{w}_k$, we know that:

$$\left|\left\langle \tilde{\mathbf{y}}_0^k, \mathbf{d}_k \right\rangle\right| = \left|\left\langle \tilde{\mathbf{y}}_0^k, \frac{\mathbf{w}_k}{\|\mathbf{w}_k\|_2} \right\rangle\right| = \frac{\|\tilde{\mathbf{y}}_0^k\|_2}{\|\mathbf{w}_k\|_2} \left|\left\langle \frac{\tilde{\mathbf{y}}_0^k}{\|\tilde{\mathbf{y}}_0^k\|_2}, \mathbf{w}_k \right\rangle\right|$$

We know from Lemma A.1 that $\|\mathbf{w}_k\|_2 > 1/\sqrt{2M}$ for all k. By $\mathcal{G}_0.3$, $\|\tilde{\mathbf{y}}_0^k\|_2 \leq C\sqrt{s}$. Thus,

$$|\left\langle \tilde{\mathbf{y}}_0^k, \mathbf{d}_k \right\rangle| = \frac{\|\tilde{\mathbf{y}}_0^k\|_2}{\|\mathbf{w}_k\|_2} \left|\left\langle \frac{\tilde{\mathbf{y}}_0^k}{\|\tilde{\mathbf{y}}_0^k\|_2}, \mathbf{w}_k \right\rangle\right| \leq C\sqrt{s} \left|\left\langle \frac{\tilde{\mathbf{y}}_0^k}{\|\tilde{\mathbf{y}}_0^k\|_2}, \mathbf{w}_k \right\rangle\right|$$

By definition, $\mathbf{w}_k$ and $\tilde{\mathbf{y}}_0^k$ are independent. Therefore by rotational invariance of the normal distribution, for each k, $\left\langle \frac{\tilde{\mathbf{y}}_0^k}{\|\tilde{\mathbf{y}}_0^k\|_2}, \mathbf{w}_k \right\rangle \sim N(0,1)$. Applying the normal concentration inequality 8 to $\left\langle \frac{\tilde{\mathbf{y}}_0^k}{\|\tilde{\mathbf{y}}_0^k\|_2}, \mathbf{w}_k \right\rangle$, we have $\left|\left\langle \frac{\tilde{\mathbf{y}}_0^k}{\|\tilde{\mathbf{y}}_0^k\|_2}, \mathbf{w}_k \right\rangle\right| \leq C \log M$ with high probability; it follows that $|\left\langle \tilde{\mathbf{y}}_0^k, \mathbf{d}_k \right\rangle| \leq \frac{C\sqrt{s}\log M}{\sqrt{M}}$ with high probability. The result then follows from a union bound over k. ■

Lemma A.7 ($\mathcal{G}_0.6$) *Conditional on* $\mathbf{D}$ *and* $\mathbf{y}_0$, $\|\langle \mathbf{y}_0, \mathbf{y} \rangle \mathbf{y}\|_2 \leq \frac{Cs^{3/2}\log M}{\sqrt{M}}$ *with high probability.*

Proof By the same logic used to prove $\mathcal{G}_0.3$, we have $\|\mathbf{y}_i\|_2 \leq C\sqrt{s}$ with high probability. Therefore:

$$\|\langle \mathbf{y}_0, \mathbf{y}_i \rangle \mathbf{y}_i\|_2 \leq |\langle \mathbf{y}_0, \mathbf{y}_i \rangle| \|\mathbf{y}_i\|_2 \leq C|\langle \mathbf{y}_0, \mathbf{y}_i \rangle|\sqrt{s}$$

We now control the term $|\langle \mathbf{y}_0, \mathbf{y}_i \rangle|$:

$$\langle \mathbf{y}_0, \mathbf{y}_i \rangle = \sum_{k \in \Omega_i} x_{ik} \langle \mathbf{y}_0, \mathbf{d}_k \rangle = \sum_{k \in \Omega_0 - \Omega_i} x_{ik} + \sum_{k \in \Omega_i} x_{ik} \left\langle \tilde{\mathbf{y}}_0^k, \mathbf{d}_k \right\rangle$$

The magnitude of the first term will depend on the size of the intersection $\Omega_0 \cap \Omega_i$. It is easy to see that $E|\Omega_0 \cap \Omega_i| = s^2/K$. Therefore, using a Chernoff bound[2] (e.g., Vershynin, 2018, theorem 2.3.1) we see that $|\Omega_0 \cap \Omega_i| \leq 2s^2 \log M/K$ with high probability. Then by $\mathcal{G}_0.5$ and Hoeffding's inequality, with high probability we have:

$$|\langle \mathbf{y}_0, \mathbf{y}_i \rangle| \leq C \log M \left(\frac{s \log M}{\sqrt{K}} + \frac{s}{\sqrt{M}} \right) \leq \frac{Cs \log M}{\sqrt{M}}$$

completing the proof. ■

A.2. Proof of Lemma 5.2

We now prove Lemma 5.2:

2. Strictly speaking, as elements in Ω_i are not chosen independently, a Chernoff bound cannot be applied directly. However, the negative correlation of sampling with replacement guarantees better concentration properties than if elements of Ω_i were chosen independently with probability s/K; see Dubhashi and Ranjan (1996) for details.

Lemma 5.2 (Expectation Computation) *Let* $\mathbf{v}_0 = \mathbf{D}\mathbf{D}^T\mathbf{y}_0$ *and* $\tilde{\mathbf{y}}_0^k = \mathbf{y}_0 - \delta_{k\in\Omega_0}\mathbf{d}_k$. *We have*

$$E[\langle \mathbf{y}_0, \mathbf{y}\rangle^2 \mathbf{y}\mathbf{y}^T|\mathbf{D}] = \frac{s}{K}\left[\frac{s-1}{K-1}\mathbf{v}_0\mathbf{v}_0^T + \left(1-\frac{2(s-1)}{K-1}\right)\sum_{k\in\Omega_0}\mathbf{d}_k\mathbf{d}_k^T\right]+$$
$$+\frac{s}{K}\left[\left(1-\frac{2(s-1)}{K-1}\right)\left(2\sum_{k\in\Omega_0}^{K}\left\langle\tilde{\mathbf{y}}_0^k,\mathbf{d}_k\right\rangle\mathbf{d}_k\mathbf{d}_k^T+\sum_{k=1}\left\langle\tilde{\mathbf{y}}_0^k,\mathbf{d}_k\right\rangle^2\mathbf{d}_k\mathbf{d}_k^T\right)+\left(\frac{s-1}{K-1}\sum_{k=1}^{K}\langle\mathbf{y}_0,\mathbf{d}_k\rangle^2\right)\mathbf{D}\mathbf{D}^T\right]. \tag{2}$$

Proof As all expectations in this lemma are conditional on $\mathbf{D}$ and $\mathbf{y}_0$, we will not write this explicitly in the proof. We can expand:

$$E\langle\mathbf{y}_0,\mathbf{y}\rangle^2\mathbf{y}\mathbf{y}^T = E\sum_{k_1\in\Omega}\sum_{k_2\in\Omega}\sum_{k_3\in\Omega}\sum_{k_4\in\Omega}x_{k_1}x_{k_2}x_{k_3}x_{k_4}\langle\mathbf{y}_0,\mathbf{d}_{k_1}\rangle\langle\mathbf{y}_0,\mathbf{d}_{k_2}\rangle\mathbf{d}_{k_3}\mathbf{d}_{k_4}^T$$

Since $Ex_{ik} = 0$ for all i,k, terms in the above expectation will be nonzero only when the indices are paired. Accordingly:

$$E\langle\mathbf{y}_0,\mathbf{y}\rangle^2\mathbf{y}\mathbf{y}^T = E\sum_{k\in\Omega}\sum_{m\in\Omega}x_k^2x_m^2\langle\mathbf{y}_0,\mathbf{d}_k\rangle^2\mathbf{d}_m\mathbf{d}_m^T + E\sum_{k\in\Omega}\sum_{m\in\Omega,m\neq k}x_k^2x_m^2\langle\mathbf{y}_0,\mathbf{d}_k\rangle\langle\mathbf{y}_0,\mathbf{d}_m\rangle\mathbf{d}_k\mathbf{d}_m^T$$

$$= E\sum_{k=1}^{K}\sum_{m=1}^{K}\mathbb{1}_{\{k,m\}\subseteq\Omega}\langle\mathbf{y}_0,\mathbf{d}_k\rangle^2\mathbf{d}_m\mathbf{d}_m^T + E\sum_{k=1}^{K}\sum_{m\neq k}^{K}\mathbb{1}_{\{k,m\}\subseteq\Omega}\langle\mathbf{y}_0,\mathbf{d}_k\rangle\langle\mathbf{y}_0,\mathbf{d}_m\rangle\mathbf{d}_k\mathbf{d}_m^T$$

$$= \sum_{k=1}^{K}\sum_{m=1}^{K}\mathbb{P}(\{k,m\}\subseteq\Omega)\langle\mathbf{y}_0,\mathbf{d}_k\rangle^2\mathbf{d}_m\mathbf{d}_m^T + \sum_{k=1}^{K}\sum_{m\neq k}^{K}\mathbb{P}(\{k,m\}\subseteq\Omega)\langle\mathbf{y}_0,\mathbf{d}_k\rangle\langle\mathbf{y}_0,\mathbf{d}_m\rangle\mathbf{d}_k\mathbf{d}_m^T$$

$$= \frac{s}{K}\sum_{k=1}^{K}\langle\mathbf{y}_0,\mathbf{d}_k\rangle^2\mathbf{d}_k\mathbf{d}_k^T + \frac{s(s-1)}{K(K-1)}\sum_{k=1}^{K}\sum_{m\neq k}^{K}\langle\mathbf{y}_0,\mathbf{d}_k\rangle^2\mathbf{d}_m\mathbf{d}_m^T + \frac{s(s-1)}{K(K-1)}\sum_{k=1}^{K}\sum_{m\neq k}^{K}\langle\mathbf{y}_0,\mathbf{d}_k\rangle\langle\mathbf{y}_0,\mathbf{d}_m\rangle\mathbf{d}_k\mathbf{d}_m^T$$

Next, we complete the square by transferring part of the first term into the other two:

$$E\langle\mathbf{y}_0,\mathbf{y}\rangle^2\mathbf{y}\mathbf{y}^T = \left(\frac{s}{K}-\frac{2s(s-1)}{K(K-1)}\right)\left(\sum_{k=1}^{K}\langle\mathbf{y}_0,\mathbf{d}_k\rangle^2\mathbf{d}_k\mathbf{d}_k^T\right)+$$
$$+\frac{s(s-1)}{K(K-1)}\left(\sum_{k=1}^{K}\langle\mathbf{y}_0,\mathbf{d}_k\rangle^2\right)\left(\sum_{k=1}^{K}\mathbf{d}_k\mathbf{d}_k^T\right)+\frac{s(s-1)}{K(K-1)}\left(\sum_{k=1}^{K}\langle\mathbf{y}_0,\mathbf{d}_k\rangle\mathbf{d}_k\right)\left(\sum_{k=1}^{K}\langle\mathbf{y}_0,\mathbf{d}_k\rangle\mathbf{d}_k\right)^T \tag{9}$$

Lastly, we note that for $k\in\Omega_0$, $\langle\mathbf{y}_0,\mathbf{d}_k\rangle = 1+\langle\mathbf{y}_0-\mathbf{d}_k,\mathbf{d}_k\rangle = 1+\left\langle\tilde{\mathbf{y}}_0^k,\mathbf{d}_k\right\rangle$, while for $k>s$, $\tilde{\mathbf{y}}_0^k = \mathbf{y}_0$. Accordingly we can substitute:

$$\sum_{k=1}^{K}\langle\mathbf{y}_0,\mathbf{d}_k\rangle^2\mathbf{d}_k\mathbf{d}_k^T = \sum_{k\in\Omega_0}\mathbf{d}_k\mathbf{d}_k^T + 2\sum_{k\in\Omega_0}\left\langle\tilde{\mathbf{y}}_0^k,\mathbf{d}_k\right\rangle\mathbf{d}_k\mathbf{d}_k^T + \sum_{k=1}^{K}\left\langle\tilde{\mathbf{y}}_0^k,\mathbf{d}_k\right\rangle^2\mathbf{d}_k\mathbf{d}_k^T$$

The result follows by making these substitutions in 9, factoring out s/K, and noting that $\sum_{k=1}^{K}\mathbf{d}_k\mathbf{d}_k^T = \mathbf{D}\mathbf{D}^T$ and $\mathbf{v}_0 = \mathbf{D}\mathbf{D}^T\mathbf{y}_0 = \sum_{k=1}^{K}\langle\mathbf{y}_0,\mathbf{d}_k\rangle\mathbf{d}_k$ by definition. ■

A.3. Proof of Lemma 5.4

We now provide a detailed proof of Lemma 5.4:

Lemma 5.4 (Projection Error Bound) *On $\mathcal{G}_0$,*

$$\left\| \frac{K}{s}\mathbf{proj}_{\mathbf{D}\mathbf{D}^T}(E[\widehat{\Sigma}_0|\mathbf{D}]) - \left(\frac{s}{K}\sum_{k=1}^{K}\langle \mathbf{y}_0, \mathbf{d}_k\rangle^2\right)\mathbf{D}\mathbf{D}^T \right\|_2 \leq \frac{Cs\log^2 M}{M} + \frac{CKs^2}{M^3}.$$

The proof will employ the following easily proven lemma, which states that Frobenius projection with $\mathbf{D}$ will not increase the 2-norm of a matrix by more than a constant factor.

Lemma A.8 *Supposes $\mathcal{G}_0$ holds. Then for any matrix $\mathbf{A} \in \mathbb{R}^{M\times M}$,*

$$\left\| \frac{\langle \mathbf{A}, \mathbf{D}\mathbf{D}^T\rangle_F}{\|\mathbf{D}\mathbf{D}^T\|_F^2}\mathbf{D}\mathbf{D}^T \right\|_2 \leq C\|\mathbf{A}\|_2.$$

Proof By the Cauchy-Schwarz inequality, $\mathcal{G}_0.1$, and $\mathcal{G}_0.2$

$$\left\| \frac{\langle \mathbf{A}, \mathbf{D}\mathbf{D}^T\rangle_F}{\|\mathbf{D}\mathbf{D}^T\|_F^2}\mathbf{D}\mathbf{D}^T \right\|_2 \leq \frac{\|\mathbf{A}\|_F\|\mathbf{D}\mathbf{D}^T\|_2}{\|\mathbf{D}\mathbf{D}^T\|_F} \leq \frac{C\|\mathbf{A}\|_F}{\sqrt{M}}.$$

The result follows from the fact that $\|\mathbf{A}\|_F \leq \sqrt{M}\|\mathbf{A}\|_2$. ■

We now prove Lemma 5.4:

Proof [Proof of Lemma 5.4] From 2, we have:

$$\frac{K}{s}E[\langle \mathbf{y}_0, \mathbf{y}\rangle^2 \mathbf{y}\mathbf{y}^T|\mathbf{D}] = \frac{s-1}{K-1}\mathbf{v}_0\mathbf{v}_0^T + \left(1 - \frac{2(s-1)}{K-1}\right)\sum_{k\in\Omega_0}\mathbf{d}_k\mathbf{d}_k^T +$$
$$+\left(1-\frac{2(s-1)}{K-1}\right)\left(2\sum_{k\in\Omega_0}\langle \tilde{\mathbf{y}}_0^k, \mathbf{d}_k\rangle \mathbf{d}_k\mathbf{d}_k^T + \sum_{k=1}\langle \tilde{\mathbf{y}}_0^k, \mathbf{d}_k\rangle^2 \mathbf{d}_k\mathbf{d}_k^T\right) + \left(\frac{s-1}{K-1}\sum_{k=1}^{K}\langle \mathbf{y}_0, \mathbf{d}_k\rangle^2\right)\mathbf{D}\mathbf{D}^T \tag{10}$$

The fifth term will be removed entirely by projection. Since the third and fourth terms are known to be small from the main text, Lemma A.8 indicates we only need to bound the projection of the first and second terms.

We first bound inner products $\left\langle \sum_{k\in\Omega_0}\mathbf{d}_k\mathbf{d}_k^T, \mathbf{D}\mathbf{D}^T\right\rangle_F$ and $\left\langle \mathbf{v}_0\mathbf{v}_0^T, \mathbf{D}\mathbf{D}^T\right\rangle_F$. We consider each term individually, beginning with the first. We write:

$$\left\langle \sum_{k\in\Omega_0}\mathbf{d}_k\mathbf{d}_k^T, \mathbf{D}\mathbf{D}^T\right\rangle_F = \sum_{k\in\Omega_0}\sum_{m=1}^{K}\langle \mathbf{d}_k, \mathbf{d}_m\rangle^2 = \sum_{k\in\Omega_0}1 + \sum_{k\in\Omega_0}\sum_{m\neq k}^{K}\langle \mathbf{d}_k, \mathbf{d}_m\rangle^2$$

which by the triangle inequality and $\mathcal{G}_0.4$ will be bounded by

$$\left|\left\langle \sum_{k\in\Omega_0}\mathbf{d}_k\mathbf{d}_k^T, \mathbf{D}\mathbf{D}^T\right\rangle_F\right| \leq s + \frac{CKs\log^2 M}{M} \tag{11}$$

Lastly, we bound $\left\langle \frac{s-1}{K-1}\mathbf{v}_0\mathbf{v}_0^T, \mathbf{D}\mathbf{D}^T \right\rangle_F$. We have:

$$\left| \left\langle \frac{s-1}{K-1}\mathbf{v}_0\mathbf{v}_0^T, \mathbf{D}\mathbf{D}^T \right\rangle_F \right| \leq \frac{s}{K} \left| \sum_{k=1}^K \left\langle \mathbf{v}_0\mathbf{v}_0^T, \mathbf{d}_k\mathbf{d}_k^T \right\rangle_F \right| = \frac{s}{K} \sum_{k=1}^K \langle \mathbf{v}_0, \mathbf{d}_k \rangle^2 .$$

Noting that $\mathbf{v}_0 = \mathbf{D}\mathbf{D}^T\mathbf{y}_0$, we know that

$$\frac{s}{K} \sum_{k=1}^K \langle \mathbf{v}_0, \mathbf{d}_k \rangle^2 = \frac{s}{K} \mathbf{v}_0^T (\mathbf{D}\mathbf{D}^T)\mathbf{v}_0 = \frac{s}{K} \mathbf{y}_0^T (\mathbf{D}\mathbf{D}^T)^3 \mathbf{y}_0 \leq \frac{s}{K} \|\mathbf{D}\mathbf{D}^T\|_2^3 \|\mathbf{y}_0\|_2^2 \leq \frac{CK^2 s^2}{M^3} \quad (12)$$

by $\mathcal{G}_0.1$ and $\mathcal{G}_0.3$.

Further, by $\mathcal{G}_0.1$ and $\mathcal{G}_0.2$, we have:

$$\left\| \frac{\mathbf{D}\mathbf{D}^T}{\|\mathbf{D}\mathbf{D}^T\|_F^2} \right\|_2 \leq \frac{C}{K}$$

We combine this with 11 and 12 to conclude that

$$\left\| \frac{\left\langle \sum_{k\in\Omega_0} \mathbf{d}_k\mathbf{d}_k^T, \mathbf{D}\mathbf{D}^T \right\rangle_F + \left\langle \frac{s-1}{K-1}\mathbf{v}_0\mathbf{v}_0^T, \mathbf{D}\mathbf{D}^T \right\rangle_F}{\|\mathbf{D}\mathbf{D}^T\|_F^2} \mathbf{D}\mathbf{D}^T \right\|_2 \leq \frac{Cs\log^2 M}{M} + \frac{CKs^2}{M^3}$$

as desired. ■

A.4. Proof of Lemma 5.6

We now prove Lemma 5.6:

Lemma 5.6 *Suppose that $\mathcal{G}_0$ holds. Then the following bounds hold:*

$$E\| \langle \mathbf{y}_0, \mathbf{y} \rangle \mathbf{y} | \mathbf{D} \|_2^2 \leq \frac{Cs^3}{M}$$

$$\|E \langle \mathbf{y}_0, \mathbf{y} \rangle^2 \mathbf{y}\mathbf{y}^T | \mathbf{D}\|_2 = \|E[\widehat{\Sigma}_0 | \mathbf{D}]\|_2 \leq \frac{Cs^3}{M^2}.$$

$$\|E \langle \mathbf{y}_0, \mathbf{y} \rangle^2 \mathbf{y}\mathbf{y}^T | \mathbf{D}\|_F = \|E[\widehat{\Sigma}_0 | \mathbf{D}]\|_F \leq \frac{Cs^3}{M^{3/2}}.$$

Proof We repeat the computations of Lemma 5.2, replacing outer products with inner products. This yields:

$$E[\| \langle \mathbf{y}_0, \mathbf{y} \rangle \mathbf{y}\|_2^2 | \mathbf{D}] = \left(\frac{s}{K} - \frac{2s(s-1)}{K(K-1)} \right) \sum_{k=1}^K \langle \mathbf{y}_0, \mathbf{d}_k \rangle^2 + \frac{s(s-1)}{K(K-1)} \left(K \sum_{k=1}^K \langle \mathbf{y}_0, \mathbf{d}_k \rangle^2 + \left\| \sum_{k=1}^K \langle \mathbf{y}_0, \mathbf{d}_k \rangle \mathbf{d}_k \right\|_2^2 \right)$$

$$= \left(\frac{s(s-1)}{(K-1)} + \frac{s}{K} - \frac{2s(s-1)}{K(K-1)} \right) \sum_{k=1}^K \langle \mathbf{y}_0, \mathbf{d}_k \rangle^2 + \frac{s(s-1)}{K(K-1)} \left\| \mathbf{D}\mathbf{D}^T \mathbf{y}_0 \right\|_2^2 .$$

We know by $\mathcal{G}_0.1$ and $\mathcal{G}_0.3$ that

$$\sum_{k=1}^{K} \langle \mathbf{y}_0, \mathbf{d}_k \rangle^2 = \mathbf{y}_0^T \mathbf{D}\mathbf{D}^T \mathbf{y}_0 \leq \frac{CKs}{M}$$

and that

$$\left\| \mathbf{D}\mathbf{D}^T \mathbf{y}_0 \right\|_2^2 = \mathbf{y}_0^T (\mathbf{D}\mathbf{D}^T)^2 \mathbf{y}_0 \leq \frac{CK^2 s}{M^2}.$$

We conclude that

$$E[\| \langle \mathbf{y}_0, \mathbf{y} \rangle \mathbf{y} \|_2^2 | \mathbf{D}] \leq \frac{Cs^3}{M}.$$

We proceed to bound the covariance $E \langle \mathbf{y}_0, \mathbf{y} \rangle^2 \mathbf{y}\mathbf{y}^T = E[\widehat{\Sigma}_0 | \mathbf{D}]$. From Lemma 5.3, we can infer that $E \langle \mathbf{y}_0, \mathbf{y} \rangle^2 \mathbf{y}\mathbf{y}^T$ satisfies:

$$E[\widehat{\Sigma}_0 | \mathbf{D}] = \frac{s}{K} \left(\frac{s-1}{K-1} \mathbf{v}_0 \mathbf{v}_0^T + \sum_{k \in \Omega_0} \mathbf{d}_k \mathbf{d}_k^T + \left(\frac{s}{K} \sum_{k=1}^{K} \langle \mathbf{y}_0, \mathbf{d}_k \rangle^2 \right) \mathbf{D}\mathbf{D}^T + \mathcal{E} \right)$$

where $\mathcal{E}$ has negligible norm, as does $\sum_{k \in \Omega_0} \mathbf{d}_k \mathbf{d}_k^T$ by $\mathcal{G}_0.3$. By $\mathcal{G}_0.1$ and $\mathcal{G}_0.3$, we know that

$$\frac{s}{K} \left\| \frac{s-1}{K-1} \mathbf{v}_0 \mathbf{v}_0^T \right\|_2 \leq \frac{s^2}{K^2} \|\mathbf{v}_0\|_2^2 = \frac{s^2}{K^2} \times \mathbf{y}_0^T (\mathbf{D}\mathbf{D}^T)^2 \mathbf{y}_0 \leq \frac{s^2}{K^2} \times \frac{CK^2 s}{M^2} = \frac{Cs^3}{M^2}$$

and likewise that

$$\left(\frac{s^2}{K^2} \sum_{k=1}^{K} \langle \mathbf{y}_0, \mathbf{d}_k \rangle^2 \right) \mathbf{D}\mathbf{D}^T = \frac{s^2}{K^2} \times \mathbf{y}_0^T \mathbf{D}\mathbf{D}^T \mathbf{y}_0 \times \|\mathbf{D}\mathbf{D}\|_2 \leq \frac{s^2}{K^2} \times \frac{CK^2 s}{M^2} = \frac{Cs^3}{M^2}.$$

This proves the bound on $\|E[\widehat{\Sigma}_0 | \mathbf{D}]\|_2$; the bound on the Frobenius norm follows immediately from the fact that $\|E[\widehat{\Sigma}_0 | \mathbf{D}]\|_F \leq \sqrt{M} \|E[\widehat{\Sigma}_0 | \mathbf{D}]\|_2$. ■

A.5. Proof of Lemma 5.7

In this section, we detail the proof of Lemma 5.7:

Lemma 5.7 (Correlation-Weighted Covariance Estimation Bound) *Suppose that $\mathcal{G}_0$ holds. Then with high probability:*

$$\|\widehat{\Sigma}_0 - E[\widehat{\Sigma}_0 | \mathbf{D}]\|_2 = \left\| \frac{1}{N} \sum_{i=1}^{N} \langle \mathbf{y}_0, \mathbf{y}_i \rangle^2 \mathbf{y}_i \mathbf{y}_i^T - E[\widehat{\Sigma}_0 | \mathbf{D}] \right\|_2 \leq \frac{Cs^3 \log^2 M}{M^{3/2} \sqrt{N}}$$

Proof In this lemma, the expectations conditioned on $\mathbf{D}$ are implied and we do not write them explicitly. We are interested in estimating the true covariance matrix of the random vector $\langle \mathbf{y}_0, \mathbf{y} \rangle \mathbf{y}$ by its sample covariance $\frac{1}{N} \sum_{i=1}^{N} \langle \mathbf{y}_0, \mathbf{y}_i \rangle^2 \mathbf{y}_i \mathbf{y}_i^T$. By $\mathcal{G}_0.6$, we know that with high probability,

$$\| \langle \mathbf{y}_0, \mathbf{y}_i \rangle \mathbf{y}_i \|_2 \leq C \log M \sqrt{E[\| \langle \mathbf{y}_0, \mathbf{y} \rangle \mathbf{y} \|_2^2]}$$

We would like to apply Theorem 5.1 with $\kappa = C \log M$, but since this is only with high probability, not almost surely, we cannot apply it directly.

Instead, let ω be the event that $\| \langle \mathbf{y}_0, \mathbf{y} \rangle \mathbf{y} \|_2^2 \leq C s^{3/2} \log M / \sqrt{M}$ and consider the truncated random vector $\mathbf{z} = \mathbb{1}_\omega \langle \mathbf{y}_0, \mathbf{y} \rangle \mathbf{y}$. By $\mathcal{G}_0.6$, ω occurs with high probability, and therefore $\mathbf{z} = \mathbf{y}$ with high probability. Since $\mathbf{y}_0$ and $\mathbf{y}$ are each sum of s unit vectors, it is easy to see that $\| \langle \mathbf{y}_0, \mathbf{y} \rangle \mathbf{y}_0 \|_2^2 \leq s^6$. Therefore:

$$E \| \langle \mathbf{y}_0, \mathbf{y} \rangle \mathbf{y} \|_2^2 - (1 - \mathbb{P}(\omega)) s^6 \leq E \|\mathbf{z}\|_2^2 \leq E \| \langle \mathbf{y}_0, \mathbf{y} \rangle \mathbf{y} \|_2^2$$

Since ω occurs with high probability, $E\|\mathbf{z}\|_2^2 = E \| \langle \mathbf{y}_0, \mathbf{y} \rangle \mathbf{y} \|_2^2$ up to a correction vanishing faster than any polynomial in M. Therefore we may safely substitute $E\|\mathbf{z}\|_2^2 = E \| \langle \mathbf{y}_0, \mathbf{y} \rangle \mathbf{y} \|_2^2$ in the following bounds. With this substitution, we have that almost surely, $\|\mathbf{z}\|_2 \leq C \log M \sqrt{E\|\mathbf{z}\|_2^2}$.

We now consider the sample $\{\mathbf{y}_i\}_{i=1}^N$ consisting of N i.i.d. copies of $\mathbf{y}$, and define $\mathbf{z}_i = \mathbb{1}_{\omega_i} \langle \mathbf{y}_0, \mathbf{y}_i \rangle \mathbf{y}_i$ where ω_i is the event that $\| \langle \mathbf{y}_0, \mathbf{y}_i \rangle \mathbf{y}_i \|_2^2 \leq C s^{3/2} \log M / \sqrt{M}$. The $\mathbf{z}_i$ are i.i.d. copies of $\mathbf{z}$, so we can apply Theorem 5.1 to the sample covariance of the random vectors $\mathbf{z}_i = \mathbb{1}_{\omega_i} \langle \mathbf{y}_0, \mathbf{y} \rangle \mathbf{y}$ with $\kappa = C \log M$, yielding:

$$\left\| \frac{1}{N} \sum_{i=1}^N \mathbf{z}_i \mathbf{z}_i^T - E \mathbf{z}_i \mathbf{z}_i^T \right\|_2 \leq \sqrt{\frac{CM(\log^3 M + t \log^2 M)}{N}} \times \|E\widehat{\Sigma}_0\|_2$$

with probability at least $1 - \exp(-t)$. Choosing $t = \log^2 M$ and applying Lemma 5.6, we have with high probability:

$$\left\| \frac{1}{N} \sum_{i=1}^N \mathbf{z}_i \mathbf{z}_i^T - E \mathbf{z}_i \mathbf{z}_i^T \right\|_2 \leq \sqrt{\frac{CM \log^4 M}{N}} \times \frac{s^3 \log^4 M}{M^2} \leq \frac{C s^3 \log^6 M}{M^{3/2} \sqrt{N}}$$

Yet we have already noted that with high probability, $\mathbf{z}_i = \langle \mathbf{y}_0, \mathbf{y}_i \rangle \mathbf{y}_i$ for all i. Therefore, with high probability,

$$\left\| \frac{1}{N} \sum_{i=1}^N \langle \mathbf{y}_0, \mathbf{y}_i \rangle^2 \mathbf{y}_i \mathbf{y}_i^T - E \mathbf{z}_i \mathbf{z}_i^T \right\|_2 \leq \frac{C s^3 \log^2 M}{M^{3/2} \sqrt{N}} \tag{13}$$

Since we have already concluded that $\|E\widehat{\Sigma}_0 - E\mathbf{z}_i \mathbf{z}_i^T\|_2$ is small, we conclude the result. ∎

A.6. Proof of Lemma 5.5

In this section, we complete the proof of Lemma 5.5.

Lemma 5.5 (Estimation Error Bound) *Recall that* $\widehat{\Sigma}_0^{proj} = \widehat{\Sigma}_0 - \mathbf{proj}_{\widehat{\Sigma}}(\widehat{\Sigma}_0)$. *Then with high probability*

$$\left\| \frac{K}{s} \widehat{\Sigma}_0^{proj} - \left(\frac{s-1}{K-1} \mathbf{v}_0 \mathbf{v}_0^T + \sum_{k \in \Omega_0} \mathbf{d}_k \mathbf{d}_k^T \right) \right\|_2 \leq \frac{CKs \log^2 M}{M^2} + \frac{CKs^2 \log^4 M}{M^{3/2} \sqrt{N}} + \frac{C s^5 \log^5 M}{M^3 \sqrt{N}}.$$

In particular, estimation error will be small so long as $N \gg \max \left\{ \frac{s^{10} \log^{10} M}{M^6}, \frac{K^2 s^4 \log^8 M}{M^3} \right\}$.

Proof All expectations in the following are implicitly conditioned on $\mathbf{D}$. We aim to bound the difference between the result of covariance projection with expected versus sample covariance matrices:

$$\|\widehat{\Sigma}_0 - \mathbf{proj}_{\widehat{\Sigma}}(\widehat{\Sigma}_0) - (E\widehat{\Sigma}_0 - \mathbf{proj}_{\mathbf{DD}^T}(E\widehat{\Sigma}_0))\|_2.$$

By the triangle inequality this is bounded by:

$$\|\widehat{\Sigma}_0 - E\widehat{\Sigma}_0\|_2 + \|\mathbf{proj}_{\mathbf{DD}^T}(E\widehat{\Sigma}_0) - \mathbf{proj}_{\mathbf{DD}^T}(\widehat{\Sigma}_0)\|_2 + \|\mathbf{proj}_{\mathbf{DD}^T}(\widehat{\Sigma}_0) - \mathbf{proj}_{\widehat{\Sigma}}(\widehat{\Sigma}_0)\|_2 \quad (14)$$

From Lemma 5.7, we know that $\|\widehat{\Sigma}_0 - \Sigma\|_2 \leq \frac{Cs^3 \log^2 M}{M^{3/2}\sqrt{N}}$. We now expand the second term in 14:

$$\|\mathbf{proj}_{\mathbf{DD}^T}(E\widehat{\Sigma}_0) - \mathbf{proj}_{\mathbf{DD}^T}(\widehat{\Sigma}_0)\|_2 = \|\mathbf{proj}_{\mathbf{DD}^T}(E\widehat{\Sigma}_0 - \widehat{\Sigma}_0)\|_2 =$$

$$= \frac{\left|\left\langle \mathbf{DD}^T, E\widehat{\Sigma}_0 - \widehat{\Sigma}_0 \right\rangle_F\right|}{\|\mathbf{DD}^T\|_F^2}\|\mathbf{DD}^T\|_2 \leq \frac{\|E\widehat{\Sigma}_0 - \widehat{\Sigma}_0\|_F}{\|\mathbf{DD}^T\|_F}\|\mathbf{DD}^T\|_2$$

by the Cauchy-Schwarz inequality. By Lemma 5.7, $\mathcal{G}_0.1$, and $\mathcal{G}_0.2$:

$$\frac{\|E\widehat{\Sigma}_0 - \widehat{\Sigma}_0\|_F}{\|\mathbf{DD}^T\|_F}\|\mathbf{DD}^T\|_2 \leq \frac{Cs^3 \log^2 M}{M\sqrt{N}} \times \frac{C\sqrt{M}}{K} \times \frac{CK}{M} \leq \frac{Cs^3 \log^2 M}{M^{3/2}\sqrt{N}}$$

We now turn to the final term in 14, which can be controlled as follows:

$$\|\mathbf{proj}_{\mathbf{DD}^T}(\widehat{\Sigma}_0) - \mathbf{proj}_{\widehat{\Sigma}}(\widehat{\Sigma}_0)\|_2 = \left\| \frac{\left\langle \widehat{\Sigma}_0, \mathbf{DD}^T \right\rangle_F}{\|\mathbf{DD}^T\|_F^2} - \frac{\left\langle \widehat{\Sigma}_0, \widehat{\Sigma} \right\rangle_F}{\|\widehat{\Sigma}\|_F^2} \right\|_2 \|\widehat{\Sigma}_0\|_2$$

$$= \left| \left\langle \widehat{\Sigma}_0, \Sigma_{\mathbf{D}} - \frac{\|\mathbf{DD}^T\|_F^2}{\|\widehat{\Sigma}\|_F^2}\widehat{\Sigma} \right\rangle_F \right| \times \frac{\|\widehat{\Sigma}_0\|_2}{\|\mathbf{DD}^T\|_F^2} \leq \left\| \mathbf{DD}^T - \frac{\|\mathbf{DD}^T\|_F^2}{\|\widehat{\Sigma}\|_F^2}\widehat{\Sigma} \right\|_2 \times \frac{\|\widehat{\Sigma}_0\|_F \|\widehat{\Sigma}_0\|_2}{\|\mathbf{DD}^T\|_F^2} \quad (15)$$

By Lemma 5.7, we can substitute $\|E\widehat{\Sigma}_0\|_2$ and $\|E\widehat{\Sigma}_0\|_F$ for $\|\widehat{\Sigma}_0\|_2$ and $\|\widehat{\Sigma}_0\|_F$ respectively, up to a constant factor. We know from Lemma 5.6, $\|E\widehat{\Sigma}_0\|_F \leq \frac{Cs^3 \log^4 M}{M^{3/2}}$ while $\|E\widehat{\Sigma}_0\|_2 \leq \frac{Cs^3 \log^4 M}{M^2}$. Then by $\mathcal{G}_0.2$:

$$\frac{\|\widehat{\Sigma}_0\|_F \|\widehat{\Sigma}_0\|_2}{\|\mathbf{DD}^T\|_F^2} \leq \frac{Cs^3 \log^2 M}{M^{3/2}} \times \frac{Cs^3 \log^2 M}{M^2} \times \frac{CM}{K^2} \leq \frac{Cs^6 \log^4 M}{K^2 M^{5/2}}$$

To bound the first term in equation 15, we use the triangle inequality, which yields:

$$\left\| \mathbf{DD}^T - \frac{\|\mathbf{DD}^T\|_F^2}{\|\widehat{\Sigma}\|_F^2}\widehat{\Sigma} \right\|_2 \leq \frac{K}{s}\left(\left\| \frac{s}{K}\mathbf{DD}^T - \widehat{\Sigma} \right\|_2 + \left| \frac{\left\| \frac{s}{K}\mathbf{DD}^T \right\|_F^2}{\|\widehat{\Sigma}\|_F^2} - 1 \right| \|\widehat{\Sigma}\|_2 \right) \quad (16)$$

By corollary 12 and $\mathcal{G}_0.1$, we have $\|\widehat{\Sigma}\|_2 \leq \frac{s}{K} \times \frac{CK}{M} = \frac{Cs}{M}$. By the same corollary and $\mathcal{G}_0.2$, we have $\|\widehat{\Sigma}\|_F^2 \geq \frac{s}{K} \times \frac{cK^2}{M} = \frac{cKs}{M}$. Thus, noting that $|\|a\| - \|b\|| \leq \|a - b\|$ for any norm, we have:

$$\left| \frac{\left\| \frac{s}{K}\mathbf{DD}^T \right\|_F^2}{\|\widehat{\Sigma}\|_F^2} - 1 \right| = \frac{\left| \left\| \frac{s}{K}\mathbf{DD}^T \right\|_F^2 - \|\widehat{\Sigma}\|_F^2 \right|}{\|\widehat{\Sigma}\|_F^2} \leq \frac{\|\widehat{\Sigma} - \frac{s}{K}\mathbf{DD}^T\|_F}{\|\widehat{\Sigma}\|_F^2} \leq \frac{Cs \log M}{\sqrt{N}} \times \frac{CM}{Ks} \leq \frac{CM \log M}{K\sqrt{N}}$$

where in the second to last inequality we again used corollary 12. Combining the pieces in 16, we have:

$$\left\|\mathbf{D}\mathbf{D}^T - \frac{\|\mathbf{D}\mathbf{D}^T\|_F^2}{\|\widehat{\Sigma}\|_F^2}\widehat{\Sigma}\right\|_2 \leq \frac{K}{s}\left(\frac{Cs\log M}{\sqrt{M}\sqrt{N}} + \frac{CM\log M}{K\sqrt{N}} \times \frac{Cs}{M}\right) \leq \frac{CK\log M}{\sqrt{M}\sqrt{N}}$$

Plugging back into 15, we have:

$$\|\mathbf{proj}_{\mathbf{D}\mathbf{D}^T}(E\widehat{\Sigma}_0) - \mathbf{proj}_{\mathbf{D}\mathbf{D}^T}(\widehat{\Sigma}_0)\|_2 \leq \frac{CK\log M}{\sqrt{M}\sqrt{N}} \times \frac{Cs^6\log^4 M}{K^2M^{5/2}} \leq \frac{Cs^6\log^5 M}{KM^3\sqrt{N}}$$

Combining our estimates for the three terms, we have:

$$\|\widehat{\Sigma}_0 - \mathbf{proj}_{\widehat{\Sigma}}(\widehat{\Sigma}_0) - (E\widehat{\Sigma}_0 - \mathbf{proj}_{\mathbf{D}\mathbf{D}^T}(E\widehat{\Sigma}_0))\|_2 \leq \frac{Cs^3\log^4 M}{M^{3/2}\sqrt{N}} + \frac{Cs^6\log^5 M}{KM^3\sqrt{N}}$$

Factoring out a factor of $\frac{s}{K}$ and plugging in to Lemma 5.4 completes the proof. ■

A.7. Proof of Lemma 5.9

In this section, we prove Lemma 5.9:

Lemma 5.9 *Let* $\mathbf{A}$ *be a rank-*s *symmetric positive-semidefinite matrix with nonzero eigenvalues satisfying:*

$$0 < \alpha \leq \lambda_1(\mathbf{A}), \ldots, \lambda_s(\mathbf{A}) \leq \beta$$

for some constants $\alpha < \beta$*. Let* $\mathbf{v} \in \mathcal{S}$ *and* $\mathbf{u} \in \mathcal{S}^\perp$ *be vectors such that* $\|\mathbf{v}\| = \|\mathbf{u}\|$ *and* $\|\mathbf{v}+\varepsilon\mathbf{u}\|_2 = 1$*. For any* $\varepsilon \in (0, \alpha/54)$ *and* $Z > \max\{2\beta, 1\}$*, define the matrix* $\mathbf{B}$ *as:*

$$\mathbf{B} = Z(\mathbf{v}+\varepsilon\mathbf{u})(\mathbf{v}+\varepsilon\mathbf{u})^T + \mathbf{A}$$

Then for all $i > s$*,* $\lambda_i(\mathbf{B}) \leq 24\varepsilon$ *and*

$$\mathcal{D}(\mathcal{S}(\mathbf{A}), \mathcal{S}(\mathbf{B})) \leq \frac{78\beta\varepsilon}{\alpha^2}.$$

We will use the following lemma:

Lemma A.9 *Let* $\mathbf{v}$ *and* $\mathbf{u}$ *be orthogonal unit vectors, and let* $\varepsilon \in (0,1)$*. Then there exist matrices* $\mathbf{B}_1, \mathbf{B}_2$ *with orthonormal columns spanning the orthogonal complements of* $\mathbf{v}$ *and* $\mathbf{v}+\varepsilon\mathbf{u}$*, respectively, such that*

$$\|\mathbf{B}_1 - \mathbf{B}_2\|_2 \leq 4\varepsilon$$

Proof Without loss of generality, we can assume that $\mathbf{v} = \mathbf{e}_1$ and $\mathbf{u} = \mathbf{e}_2$. We can choose $\mathbf{B}_1$ and $\mathbf{B}_2$ to be identically equal on the orthogonal complement of the span of $\mathbf{e}_1, \mathbf{e}_2$. Accordingly, it suffices to prove the result for two-dimensional matrices. Let

$$\mathbf{B}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{B}_2 = \frac{1}{\sqrt{1+\varepsilon^2}}\begin{pmatrix} 1 & -\varepsilon \\ \varepsilon & 1 \end{pmatrix}.$$

The result follows from taking the difference $\mathbf{B}_1 - \mathbf{B}_2$ and applying the inequality $\|\mathbf{B}_1 - \mathbf{B}_2\|_2 \leq \|\mathbf{B}_1 - \mathbf{B}_2\|_F$. ■

We now use this lemma to prove 5.9:

Proof [Proof of Lemma 5.9] Since $\mathbf{B}$ has rank at most $s+1$, we may restrict consideration to the $s+1$-dimensional subspace given by the direct sum of $\mathcal{S}$ with the span of $\mathbf{u}$. We note that Weyl's theorem ensures that the lead eigenvalue λ of $\mathbf{B}$ satisfies $Z \leq \lambda \leq Z + \beta$.

We first show that the lead eigenspaces of $\mathbf{B}$ differs from that of $Z(\mathbf{v}+\varepsilon\mathbf{u})(\mathbf{v}+\varepsilon\mathbf{u})^T$ by at most ε/Z. We first prove this result in the case that $\mathbf{A} = \mathbf{I}_s$, the first s columns of an identity matrix. Consider the lead eigenvector $\mathbf{B}$ corresponding to lead eigenvalue λ. There exists $\mathbf{w}$ orthogonal to $\mathbf{v} + \varepsilon\mathbf{u}$ such that

$$\mathbf{B}(\mathbf{v} + \varepsilon\mathbf{u} + \mathbf{w}) = Z(\mathbf{v} + \varepsilon\mathbf{u}) + \mathbf{v} + \mathbf{I}_s\mathbf{w} = \lambda(\mathbf{v} + \varepsilon\mathbf{u} + \mathbf{w})$$

If $\mathbf{w} = 0$, the result holds, so we assume $\|\mathbf{w}\|_2 > 0$. Taking inner products of this equation with $\mathbf{w}$, we have:

$$\langle \mathbf{v}, \mathbf{w}\rangle + \langle \mathbf{I}_s\mathbf{w}, \mathbf{w}\rangle = \lambda\|\mathbf{w}\|_2^2$$

Since $0 \leq \langle \mathbf{I}_s\mathbf{w}, \mathbf{w}\rangle \leq \|\mathbf{w}\|_2^2$, we have further that

$$\langle \mathbf{v}, \mathbf{w}\rangle \geq (\lambda - 1)\|\mathbf{w}\|_2^2.$$

Therefore, since $\langle \mathbf{v} + \varepsilon\mathbf{u}, \mathbf{w}\rangle = 0$, we have $\langle \mathbf{v}, \mathbf{w}\rangle = -\varepsilon\langle \mathbf{u}, \mathbf{w}\rangle$, so by the Cauchy-Schwarz inequality,

$$\varepsilon\|\mathbf{u}\|_2\|\mathbf{w}\|_{\geq}(\lambda - 1)\|\mathbf{w}\|_2^2 \implies \varepsilon\|\mathbf{u}\|_2 \geq (\lambda - 1)\|\mathbf{w}\|_2$$

where division by $\|\mathbf{w}\|_2$ is permissible as we assumed $\mathbf{w} \neq 0$. Since $\lambda \geq Z$, we conclude that $\|\mathbf{w}\|_2 \leq \varepsilon/(Z-1) \leq 2\varepsilon/Z$.

We now extend this to the s-dimensional subspace $\mathcal{S}$ by studying the complement of the eigenvectors. Let $\mathbf{E}$ be a matrix with columns forming and orthogonal basis of the orthogonal complement of $\mathbf{v} + \varepsilon\mathbf{u}$, and let $\mathbf{E}_\mathbf{w}$ be the same for $\mathbf{v} + \varepsilon\mathbf{u} + \mathbf{w}$. By Lemma A.9, we can choose these such that $\|\mathbf{E}_1 - \mathbf{E}_2\|_2 \leq 8\varepsilon/Z$. We consider the matrix:

$$\mathbf{E}_\mathbf{w}^T\mathbf{B}\mathbf{E}_\mathbf{w} = \mathbf{E}^T\mathbf{B}\mathbf{E} + (\mathbf{E}_w - \mathbf{E})^T\mathbf{B}\mathbf{E} + \mathbf{E}^T\mathbf{B}(\mathbf{E}_\mathbf{w} - \mathbf{E}) + (\mathbf{E}_\mathbf{w} - \mathbf{E})^T\mathbf{B}(\mathbf{E}_\mathbf{w} - \mathbf{E}).$$

Since $\|\mathbf{E} - \mathbf{E}_\mathbf{w}\|_2 \leq 8\varepsilon/Z$, we have that

$$\|(\mathbf{E}_w - \mathbf{E})^T\mathbf{B}\mathbf{E} + \mathbf{E}^T\mathbf{B}(\mathbf{E}_\mathbf{w} - \mathbf{E}) + (\mathbf{E}_\mathbf{w} - \mathbf{E})^T\mathbf{B}(\mathbf{E}_\mathbf{w} - \mathbf{E})\|_2 \leq 27\varepsilon.$$

as each term is bounded individually by 9ε. Since $\mathbf{E}$ is a basis of the complement of $\mathbf{v} + \varepsilon\mathbf{u}$, $\mathbf{E}^T\mathbf{B}\mathbf{E} = \mathbf{E}^T\mathbf{A}\mathbf{E}$, which will have eigenvectors spanning the orthogonal complement of $\mathcal{S}_s(\mathbf{A})$ with respect to $\mathbf{v} + \varepsilon\mathbf{u}$. Since the above matrices are symmetric, Weyl's Theorem tells us that any eigenvalues of $\mathbf{E}^T\mathbf{B}\mathbf{E}$ beyond the first (s-1) will be bounded by 24ε. Since $\varepsilon < \alpha/54$, the first $s-1$ eigenvalues of $\mathbf{E}^T\mathbf{A}\mathbf{E}$ are separated from zero by at least $\alpha/2$; thus we can apply the Davis-Kahan theorem to conclude that

$$\mathcal{D}(\mathcal{S}_{s-1}(\mathbf{E}^T\mathbf{A}\mathbf{E}), \mathcal{S}_{s-1}(\mathbf{E}_\mathbf{w}^T\mathbf{B}\mathbf{E}_\mathbf{w}) \leq 54\varepsilon/\alpha.$$

We now repeat this application of Lemma A.9 with bases for the spans of $\mathbf{v}$ and $\mathbf{v}+\varepsilon\mathbf{u}$. We now use $\mathbf{E}_\mathbf{v}$ as a basis for the orthogonal complement of $\mathbf{v}$, chosen according the lemma, with $\mathbf{E}$ representing a different basis if necessary. Since

$$\mathbf{E}^T\mathbf{A}\mathbf{E} = \mathbf{E}_\mathbf{v}^T\mathbf{A}\mathbf{E}_\mathbf{v} (\mathbf{E}-\mathbf{E}_\mathbf{v})^T\mathbf{B}\mathbf{E} + \mathbf{E}^T\mathbf{B}(\mathbf{E}-\mathbf{E}_\mathbf{v}) + (\mathbf{E}-\mathbf{E}_\mathbf{v})^T\mathbf{B}(\mathbf{E}-\mathbf{E}_\mathbf{v}),$$

applying the lemma gives us that $\mathbf{E}^T\mathbf{A}\mathbf{E} \approx \mathbf{E}_\mathbf{v}^T\mathbf{A}\mathbf{E}_\mathbf{v}$ up to a correction with norm at most 12ε. By Weyl's and the Davis-Kahan theorem, then,

$$\mathcal{D}(\mathcal{S}_{s-1}(\mathbf{E}^T\mathbf{A}\mathbf{E}), \mathcal{S}_{s-1}(\mathbf{E}_\mathbf{v}^T\mathbf{A}\mathbf{E}_\mathbf{v}) \le 24\varepsilon/\alpha.$$

We conclude by the triangle inequality that the subspace spanned by eigenvectors 2 through $s-1$ of $\mathbf{B}$ is distance at most $78\varepsilon/\alpha$ from the orthogonal complement of $\mathbf{v}$ in $\mathcal{S}_s(\mathbf{A})$. Since we already concluded that the leading eigenvector of $\mathbf{B}$ is close to $\mathbf{v}$, the result for $\mathbf{A}=\mathbf{I}_s$ follows.

It remains to extend this result to all rank s positive-semidefinite matrices $\mathbf{A}$. We consider the matrix $\mathcal{A}$ which equals $\mathbf{A}$ on $\mathcal{S}_s(\mathbf{A})$ and is the identity on its complement. This matrix will be positive-definite and therefore has a square-root-inverse $\mathcal{A}^{-1/2}$ with norm at most $1/\sqrt{\alpha}$. By construction, then, $\mathcal{A}^{-1/2}\mathbf{A}\mathcal{A}^{-1/2} = \mathbf{I}_s$ as before. We have:

$$\mathbf{B} = \mathcal{A}^{1/2}\left(Z\mathcal{A}^{-1/2}(\mathbf{v}+\varepsilon\mathbf{u})(\mathbf{v}+\varepsilon\mathbf{u})\mathcal{A}^{-1/2} + \mathcal{I}_s\right)\mathcal{A}^{1/2}$$

Since these matrices remain symmetric, we can then apply the previous result to the matrix inside the parentheses with $\mathbf{v} \to \frac{\|\mathbf{v}\|_2}{\|\mathcal{A}^{-1/2}\mathbf{v}\|_2}\mathcal{A}^{-1/2}\mathbf{v}$ and $Z \to Z\|\mathcal{A}^{-1/2}\mathbf{v}\|^2 \le Z/\alpha$, giving an subspace error of at most $\frac{78\varepsilon}{\alpha^2}$. Applying the copies of $\mathcal{A}^{1/2}$ outside the parentheses can magnify this by at most a further factor of β, so we conclude that

$$\mathcal{D}(\mathcal{S}_s(\mathbf{B}), \mathcal{S}_s(\mathbf{A})) \le \frac{78\beta\varepsilon}{\alpha^2}$$

for generic $\mathbf{A}$ as desired. ■

A.8. Proof of Lemma 5.10

We now prove Lemma 5.10:

Lemma 5.10 *Suppose Ω_1, Ω_2 are at-most-s-element subsets of $\{1, 2, \dots, K\}$ such that $\Omega_1 \cap \Omega_2 = \emptyset$. Let $\mathcal{S}_1 = \mathbf{span}\{\mathbf{d}_k\}_{k\in\Omega_1}$ and $\mathcal{S}_j = \mathbf{span}\{\mathbf{d}_m\}_{m\in\Omega_2}$. Then with high probability,*

$$\mathcal{D}(\mathcal{S}_1, \mathcal{S}_2) \ge 1 - \frac{Cs}{M}$$

Proof Without loss of generality, we may assume $|\Omega_1| = |\Omega_2| = s$. Let $\mathbf{P}_1, \mathbf{P}_2$ represent the projection matrices onto $\mathcal{S}_1$ and $\mathcal{S}_2$ respectively. By rotation invariance, we may assume that $\mathcal{S}_2$ is fixed while $\mathcal{S}_1$ remains random.

We apply an ε-net argument over the unit sphere of $\mathcal{S}_2$: let $\mathcal{M}_2$ be a 1/4-net of the unit sphere in $\mathcal{S}_2$, chosen to have cardinality at most 9^s. Now let $\mathbf{z} \in \mathcal{M}_2$. Again applying rotation invariance, we may assume that $\mathcal{S}_1$ is fixed while $\mathbf{z}$ is random. It can be shown that

$$\mathbb{P}\left(\|\mathbf{P}_1\mathbf{z}\|_2 > \frac{C\sqrt{s}}{\sqrt{M}}\right) \le 10^{-s}$$

(say) for large enough C; see, for example, (Vershynin, 2018, Lemma 5.3.2) for details. Unfixing $\mathbf{z} \in \mathcal{M}_2$ by a union bound, it follows that

$$\mathbb{P}\left(\sup_{\mathbf{z}\in\mathcal{S}_2} \frac{\|\mathbf{P}_1\mathbf{z}\|_2}{\|\mathbf{z}\|_2} > \frac{2C\sqrt{s}}{\sqrt{M}}\right) \leq \mathbb{P}\left(\max_{\mathbf{z}\in\mathcal{M}_2} \|\mathbf{P}_1\mathbf{z}\|_2 > \frac{C\sqrt{s}}{\sqrt{M}}\right) \leq (9/10)^s$$

This implies that, with high probability, for every unit vector $\mathbf{z} \in \mathcal{S}$, $|\langle \mathbf{P}_1\mathbf{z}, \mathbf{z}\rangle| \leq C\sqrt{s}/\sqrt{M}$. Accordingly, by the Pythagorean theorem,

$$\mathcal{D}(\mathcal{S}_1, \mathcal{S}_2) \geq \sqrt{1 - Cs/M}.$$

substituting a new constant C if necessary. Applying the fact that $\sqrt{1-x} \geq 1 - x/4$ for x near zero, we conclude the result. ■

A.9. Proof of Theorem 4.4

Lastly, we prove Theorem 4.4:

Theorem 4.4 (Polynomially-Many Intersections Suffice) *Let ℓ be the smallest integer such that $(s/K)^\ell < 1/2K$, and assume Ω_i, $i = 1, \ldots, K$ are uniformly distributed among s-element subsets of $\{1, \ldots, K\}$. For positive integer J and $j \in \{1, \ldots, J/\ell\}$, define the non-intersecting ℓ-fold intersections Ω_j^ℓ as:*

$$\Omega_j^\ell = \bigcap_{p=1}^{\ell} \Omega_{(j-1)\ell+p}$$

Then as long as $J \geq K \log^3 K$, with high probability for every $k \in \{1, \ldots, K\}$ there exists a $j \in \{1, \ldots, J/\ell\}$ such that $\Omega_j^\ell = \{k\}$.

Proof We begin by fixing k and then computing the probability that k is the unique element of intersection for a fixed Ω_j^ℓ. We know that $\mathbb{P}\left(k \in \Omega_j^\ell\right) = (s/K)^\ell$, while the probability that another element is in the intersection is bounded by $K(s/K)^\ell$, so we have:

$$\mathbb{P}\left(\bigcap_{j\in\mathcal{I}_i} \Omega_i = \{k\}\right) \geq \left(\frac{s}{K}\right)^\ell \left(1 - K\left(\frac{s}{K}\right)^\ell\right) \geq \frac{1}{2}\left(\frac{s}{K}\right)^\ell$$

We now unfix the set $\mathcal{I}$ as follows. We divide $\{1, 2, \ldots, N\}$ into disjoint ℓ-element subsets $\mathcal{I}_i$. We define the random variables H_i to be indicators for the events $\left\{\bigcap_{j\in\mathcal{I}_i} \Omega_i = \{k\}\right\}$. Since the sets $\mathcal{I}_i$ do not overlap, these are J/ℓ independent Bernoulli random variables with success probability at least $(s/K)^\ell/2$. Since $(s/K)^\ell < 1/2K$, it follows that:

$$\mathbb{P}\left(\sum_{i=1}^{J/\ell} H_i = 0\right) \leq \left(1 - \frac{1}{2}\left(\frac{s}{K}\right)^\ell\right)^{J/\ell} \leq \left(1 - \frac{1}{4K}\right)^{J/\ell} \leq \exp\left(-\frac{J}{4K\ell}\right)$$

Since $J \geq K \log^3 K$ and $\ell = \left\lceil \frac{\log(2K)}{\log(K/s)} \right\rceil$, we conclude the above bound tends to zero faster than any polynomial, which still holds when unfixing k by a union bound. This completes the proof. ■