Fast Algorithms for Demixing Sparse Signals from Nonlinear Observations

Mohammadreza Soltani and Chinmay Hegde
Electrical and Computer Engineering Department
Iowa State University*

Abstract

We study the problem of *demixing* a pair of sparse signals from nonlinear observations of their superposition. Mathematically, we consider a nonlinear signal observation model, $y_i = g(a_i^T x) + e_i,\ i = 1, \ldots, m$, where $x = \Phi w + \Psi z$ denotes the superposition signal, Φ and Ψ are orthonormal bases in $\mathbb{R}^n$, and $w, z \in \mathbb{R}^n$ are sparse coefficient vectors of the constituent signals. Further, we assume that the observations are corrupted by a subgaussian additive noise. Within this model, g represents a nonlinear *link* function, and $a_i \in \mathbb{R}^n$ is the i-th row of the measurement matrix, $A \in \mathbb{R}^{m \times n}$. Problems of this nature arise in several applications ranging from astronomy, computer vision, and machine learning.

In this paper, we make some concrete algorithmic progress for the above demixing problem. Specifically, we consider two scenarios: (i) the case when the demixing procedure has no knowledge of the link function, and (ii) the case when the demixing algorithm has perfect knowledge of the link function. In both cases, we provide fast algorithms for recovery of the constituents w and z from the observations. Moreover, we support these algorithms with a rigorous theoretical analysis, and derive (nearly) tight upper bounds on the sample complexity of the proposed algorithms for achieving stable recovery of the component signals. Our analysis also shows that the running time of our algorithms is essentially as good as the best possible.

We also provide a range of numerical simulations to illustrate the performance of the proposed algorithms on both real and synthetic signals and images. Our simulations show the superior performance of our algorithms compared to existing methods for demixing signals and images based on convex optimization. In particular, our proposed methods yield demonstrably better sample complexities as well as improved running times, thereby enabling their applicability to large-scale problems.

1 Introduction

1.1 Setup

In numerous signal processing applications, the problem of *demixing* is of special interest. In simple terms, demixing involves disentangling two (or more) constituent signals from observations of their linear superposition. Formally, consider a discrete-time signal $x \in \mathbb{R}^n$ that can be expressed as the superposition of two signals:

$$x = \Phi w + \Psi z\,,$$

where Φ and Ψ are orthonormal bases of $\mathbb{R}^n$, and $w, z \in \mathbb{R}^n$ are the corresponding basis coefficients. The goal of signal demixing, in this context, is to reliably recover the constituent signals (equivalently, their basis representations w and z) from the superposition signal x.

*This work was supported in part by the National Science Foundation under the grant CCF-1566281. Parts of this work also appear in an Iowa State University technical report [1] and a conference paper to be presented in the 2016 Asilomar Conference in November 2016 [2].

Demixing suffers from a fundamental *identifiability* issue since the number of unknowns ($2n$) is greater than the number of observations (n). This is easy to see: suppose for simplicity that $\Phi = \Psi = I_n$, the canonical basis of $\mathbb{R}^n$, and therefore, $x = w + z$. Now, suppose that both w and z have only one nonzero entry in the first coordinate. Then, there is an infinite number of w and z that are consistent with the observations x, and any hope of recovering the true components is lost. Therefore, for the demixing problem to have an identifiable solution, one inevitably has to assume some type of *incoherence* between the constituent signals (or more specifically, between the corresponding bases Φ and Ψ) [3, 4]. Such an incoherence assumption certifies that the components are sufficiently "distinct" and that the recovery problem is well-posed. Please see Section 3 for a formal definition of incoherence.

However, even if we assume that the signal components are sufficiently incoherent, demixing poses additional challenges under stringent observation models. Suppose, now, that we only have access to undersampled linear measurements of the signal, i.e., we record:

$$y = Ax\,, \tag{1.1}$$

where $A \in \mathbb{R}^{m\times n}$ denotes the measurement operator and where $m < n$. In this scenario, the demixing problem is further confounded by the fact that A possesses a nontrivial null space. In this case, it might seem impossible to recover the components x and z since A possesses a nontrivial null space. Once again, this problem is highly ill-posed and further structural assumptions on the constituent signals are necessary. Under-determined problems of this kind have recently received significant attention in signal processing, machine learning, and high-dimensional statistics. In particular, the emergent field of *compressive sensing* [5–7] shows that it is indeed possible to exactly reconstruct the underlying signals under certain assumptions on x, provided the measurement operator is designed carefully. This intuition has enabled the design of a wide range of efficient architectures for signal acquisition and processing [8, 9].

In this paper, we address an *even* more challenging question in the demixing context. Mathematically, we consider a *noisy, nonlinear* signal observation model, formulated as follows:

$$y_i = g(\langle a_i, \Phi w + \Psi z\rangle) + e_i,\ i = 1, \dots, m\,. \tag{1.2}$$

Here, as before, the superposition signal is modeled as $x = \Phi w + \Psi z$. Each observation is generated by the composition of a linear functional of the signal $\langle a_i, x\rangle$, with a (scalar) nonlinear function g. Here, g is sometimes called a *link* or *transfer* function, and a_i denotes the i^{th} row of a linear measurement matrix $A \in \mathbb{R}^{m\times n}$. For full generality, in (1.2) we assume that each observation y_i is corrupted by additive noise; the noiseless case is realized by setting $e_i = 0$. We will exclusively consider the "measurement-poor" regime where the number of observations m is much smaller than the ambient dimension n.

For all the reasons detailed above, the problem of recovering the coefficient vectors w and z from the measurements y seems daunting. Therefore, we make some structural assumptions. Particularly, we assume that w and z are *s-sparse* (i.e., they contain no more than s nonzero entries). Further, we will assume perfect knowledge of the bases Φ and Ψ, and the measurement matrix A. The noise vector $e \in \mathbb{R}^m$ is assumed to be stochastic, zero mean, and bounded. Under these assumptions, we will see that it is indeed possible to stably recover the coefficient vectors, with a number of observations that is proportional to the sparsity level s, as opposed to the ambient dimension n.

The nonlinear link function g plays a crucial role in our algorithm development and analysis. In signal processing applications, such nonlinearities may arise due to imperfections caused during a measurement process, or inherent limitations of the measurement system, or due to quantization or calibration errors. We discuss such practical implications more in detail below. On an abstract level, we consider two distinct scenarios. In the first scenario, the link function may be non-smooth, non-invertible, or even unknown to the recovery procedure. This is the more challenging case, but we will show that recovery of the components is possible even without knowledge of g. In the second scenario; the link function is a known, smooth, and strictly monotonic function. This is the somewhat simpler case, and we will see that this leads to significant improvements in recovery performance both in terms of theory and practice.

1.2 Our Contributions

In this paper, we make some concrete algorithmic progress in the demixing problem under nonlinear observations. In particular, we study the following scenarios depending on certain additional assumptions made on (1.2):

1. **Unknown** g. We first consider the (arguably, more general) scenario where the nonlinear link function g may be non-smooth, non-invertible, or even unknown. In this setting, we do not explicitly model the additive noise term in (1.2). For such settings, we introduce a novel demixing algorithm that is non-iterative, does not require explicit knowledge of the link function g, and produces an estimate of the signal components. We call this algorithm OneShot to emphasize its non-iterative nature. It is assumed that OneShot possess oracle knowledge of the measurement matrix A, and orthonormal bases Φ and Ψ.

   We supplement our proposed algorithm with a rigorous theoretical analysis and derive upper bounds on the *sample complexity* of demixing with nonlinear observations. In particular, we prove that the sample complexity of OneShot to achieve an estimation error κ is given by $m = \mathcal{O}(\frac{1}{\kappa^2} s \log \frac{n}{s})$ provided that the entries of the measurement matrix are i.i.d. standard normal random variables.

2. **Known** g. Next, we consider the case where the nonlinear link function g is known, smooth, and monotonic. In this setting, the additive noise term in (1.2) is assumed to be bounded either absolutely, or with high probability. For such (arguably, milder) settings, we provide an iterative algorithm for demixing of the constituent signals in (1.2) given the nonlinear observations y. We call this algorithm Demixing with Hard Thresholding, or DHT for short. In addition to knowledge of g, we assume that DHT possesses oracle knowledge of A, Φ, and Ψ.

   Within this scenario, we also analyze two special sub-cases:

   Case 2a: Isotropic measurements. We assume that the measurement vectors a_i are independent, isotropic random vectors that are incoherent with the bases Φ and Ψ. This assumption is more general than the i.i.d. standard normal assumption on the measurement matrix made in the first scenario, and is applicable to a wider range of measurement models. For this case, we show that the sample complexity of DHT is upper-bounded by $m = \mathcal{O}(s \text{ polylog } n)$, independent of the estimation error κ.

   Case 2b: Subgaussian measurements. we assume that the rows of the matrix A are independent subgaussian isotropic random vectors. This is also a generalization of the i.i.d. standard normal assumption made above, but more restrictive than Case 2a. In this setting, we obtain somewhat better sample complexity. More precisely, we show that the sample complexity of DHT is $m = \mathcal{O}(s \log \frac{n}{s})$ for sample complexity, matching the best known sample complexity bounds for recovering a superposition of s-sparse signals from *linear* observations [10, 11].

In both the above cases, the underlying assumption is that the bases Φ and Ψ are sufficiently incoherent, and that the sparsity level s is small relative to the ambient dimension n. In this regime, we show that DHT exhibits a *linear* rate of convergence, and therefore the computational complexity of DHT is only a logarithmic factor higher than OneShot. Table 1 provides a summary of the above contributions for the specific case where Φ is the identity (canonical) basis and Ψ is the discrete cosine transform (DCT) basis, and places them in the context of the existing literature on some nonlinear recovery methods [12–14]. We stress that these previous works do not explicitly consider the demixing problem, but in principle the algorithms of [12–14] can be extended to the demixing setting as well.

1.3 Techniques

At a high level, our recovery algorithms are based on the now-classical method of *greedy* iterative thresholding. In both methods, the idea is to first form a proxy of the signal components, followed by hard thresholding to promote sparsity of the final estimates of the coefficient vectors w and z. The key distinguishing factor from existing methods is that the greedy thresholding procedures used to estimate w and z are *deliberately myopic*, in the sense that each thresholding step operates *as if the other component did not exist at all.*

Table 1: *Summary of our contributions, and comparison with existing methods for the concrete case where Φ is the identity and Ψ is the DCT basis. Here, s denotes the sparsity level of the components, n denotes the ambient dimension, m denotes the number of samples, and κ denotes estimation error.*

Algorithms	Sample complexity	Running time	Measurements	Link function
LASSO [12]	$\mathcal{O}(\frac{s}{\kappa^2}\log\frac{n}{s})$	poly(n)	Gaussian	unknown
OneShot	$\mathcal{O}(\frac{s}{\kappa^2}\log\frac{n}{s})$	$\mathcal{O}(mn)$	Gaussian	unknown
DHT	$\mathcal{O}(s \text{ polylog } n)$	$\mathcal{O}(mn\log\frac{1}{\kappa})$	Isotropic rows	known
DHT	$\mathcal{O}(s\log\frac{n}{s})$	$\mathcal{O}(mn\log\frac{1}{\kappa})$	Subgaussian	known

Despite this apparent shortcoming, we are still able to derive bounds on recovery performance when the signal components are sufficiently incoherent.

Our first algorithm, OneShot, is based on the recent, pioneering approach of [12], which describes a simple (but effective) method to estimate a high-dimensional signal from unknown nonlinear observations. Our first main contribution of this paper is to extend this idea to the nonlinear demixing problem, and to precisely characterize the role of incoherence in the recovery process. Indeed, a variation of the approach of [12] (described in Section 5) can be used to solve the nonlinear demixing problem as stated above, with a similar two-step method of first forming a proxy, and then performing a convex estimation procedure (such as the LASSO [15]) to produce the final signal estimates. However, as we show below in our analysis and experiments, OneShot offers superior performance to this approach. The analysis of OneShot is based on a geometric argument, and leverages the *Gaussian mean width* for the set of sparse vectors, which is a statistical measure of complexity of a set of points in a given space.

While OneShot is simple and effective, one can potentially do much better if the link function g were available at the time of recovery. Our second algorithm, DHT, leverages precisely this intuition. First, we formulate our nonlinear demixing problem in terms of an optimization problem with respect to a specially-defined loss function that depends on the nonlinearity g. Next, for solving the proposed optimization problem, we propose an iterative method to solve the optimization problem, up to an additive approximation factor. Each iteration with DHT involves a proxy calculation formed by computing the gradient of the loss function, followed by (myopic) projection onto the constraint sets. Again, somewhat interestingly, this method can be shown to be *linearly convergent*, and therefore only incurs a small (logarithmic) overhead in terms of running time. The analysis of DHT is based on bounding certain parameters of the loss function known as the restricted strong convexity (RSC) and restricted strong smoothness (RSS) constants.[1]

Finally, we provide a wide range of simulations to verify empirically our claims both on synthetic and real data. We first compare the performance of OneShot with the convex optimization method of [12] for nonlinear demixing via a series of phase transition diagrams. Our simulation results show that OneShot outperforms this convex method significantly in both demixing efficiency as well as running time, and consequently makes it an attractive choice in large-scale problems. However, as discussed below, the absence of knowledge of the link function induces an inevitable scale ambiguity in the final estimation[2]. For situations where we know the link function precisely, our simulation results show that DHT offers much better statistical performance compared to OneShot, and is even able to recover the scale of the signal components explicitly. We also provide simulation results on real-world natural images and astronomical data to demonstrate robustness of our approaches.

[1] Quantifying algorithm performance by bounding RSC and RSC constants of a given loss function are quite widespread in the machine learning literature [**?**, 16–18], but have not studied in the context of signal demixing.

[2] Indeed, following the discussion in [12], any demixing algorithm that does not leverage knowledge of g is susceptible to such a scale ambiguity.

1.4 Organization

The rest of this paper is organized as follows. Section 2 describes several potential applications of our proposed approach, and relationship with prior art. Section 3 introduces some key notions that are used throughout the paper. Section 4 contains our proposed algorithms, accompanied by analysis of their performance; complete proofs are deferred to Section 6. Section 5 lists the results of a series of numerical experiments on both synthetic and real data, and Section 7 provides concluding remarks.

2 Applications and Related Work

Demixing problems of various flavors have been long studied in research areas spanning signal processing, statistics, and physics, and we only present a small subset of relevant related work. In particular, demixing methods have been the focus of significant research over the fifteen years, dating back at least to [19]. The work of Elad et al. [3] and Bobin et al. [20] posed the demixing problem as an instance of *morphological components analysis* (MCA), and formalized the observation model (1.1). Specifically, these approaches posed the recovery problem in terms of a convex optimization procedure, such as the LASSO [15]. The work of Pope et al. [21] analyzed somewhat more general conditions under which stable demixing could be achieved.

More recently, the work of [22] showed a curious phase transition behavior in the performance of the convex optimization methods. Specifically, they demonstrated a sharp statistical characterization of the achievable and non-achievable parameters for which successful demixing of the signal components can be achieved. Moreover, they extended the demixing problem to a large variety of signal structures beyond sparsity via the use of general *atomic norms* in place of the ℓ_1-norm in the above optimization. See [23] for an in-depth discussion of atomic norms, their statistical and geometric properties, and their applications to demixing.

Approaches for (linear) demixing has also considered a variety of signal models beyond sparsity. The robust PCA problem [24–26] involves the separation of low-rank and sparse matrices from their sum. This idea has been used in several applications ranging from video surveillance to sensor network monitoring. In machine learning applications, the separation of low-rank and sparse matrices has been used for latent variable model selection [27] as well as the robust alignment of multiple occluded images [28]. Another type of signal model is the *low-dimensional manifold* model. In [10, 11], the authors proposed a greedy iterative method for demixing signals, arising from a mixture of known low-dimensional manifolds by iterative projections onto the component manifolds.

The problem of signal demixing from linear measurements belongs to a class of linear inverse problems that underpin *compressive sensing* [5, 6]; see [7] for an excellent introduction. There, the overarching goal is to recover signals from (possibly randomized) linear measurements of the form (1.1). More recently, it has been shown that compressive sensing techniques can also be extended to inverse problems where the available observations are manifestly *nonlinear*. For instance, in 1-bit compressive sensing [29, 30] the linear measurements of a given signal are quantized in the extreme fashion such that the measurements are binary (± 1) and only comprise the sign of the linear observation. Therefore, the amplitude of the signal is completely discarded by the quantization operator. Another class of such nonlinear recovery techniques can be applied to the classical signal processing problem of phase retrieval [31] which is somewhat more challenging than 1-bit compressive sensing. In this problem, the phase information of the signal measurements may be irrecovably lost and we have only access to the amplitude information of the signal [31]. Therefore, the recovery task here is to retrieve the phase information of the signal from random observations. Other related works include approaches for recovering low-rank matrices from nonlinear observations [32, 33]. We mention in passing that inverse problems involving nonlinear observations have also long been studied in the statistical learning theory literature; see [34–37] for recent work in this area. Analogous to our scenarios above, these works consider both known as well as unknown link functions; these two classes of approaches are respectively dubbed as Generalized Linear Models (GLM) learning methods and Single Index Model (SIM) learning methods.

For our algorithmic development, we build upon a recent line of efficient, iterative methods for signal estimation in high dimensions [12, 16–18, 38, 39]. The basic idea is to pose the recovery as a (non-convex) optimization problem in which an objective function is minimized over the set of s-sparse vectors. Essentially, these algorithms are based on well-known iterative thresholding methods proposed in the context of sparse recovery and compressive sensing [40, 41]. The analysis of these methods heavily depends on the assumption that the objective function satisfies certain (restricted) regularity conditions; see Sections 3 and 6 for details. Crucially, we adopt the approach of [42], which introduces the concept of the restricted strong convexity (RSC) and restricted strong smoothness (RSS) constants of a loss function. Bounding these constants in terms of problem parameters n and s, as well as the level of incoherence in the components, enables explicit characterization of both sample complexity and convergence rates.

3 Preliminaries

In this section, we introduce some notation and key definitions. Throughout this paper, $\|.\|_p$ denotes the ℓ_p-norm of a vector in $\mathbb{R}^n$, and $\|A\|$ denotes the spectral norm of the matrix $A \in \mathbb{R}^{m\times n}$. Let Φ and Ψ be orthonormal bases of $\mathbb{R}^n$. Define the set of sparse vectors in the bases Φ and Ψ as follows:

$$K_1 = \{\Phi a \mid \|a\|_0 \leq s_1\},$$
$$K_2 = \{\Psi a \mid \|a\|_0 \leq s_2\},$$

and define $K = \{a \mid \|a\|_0 \leq s\}$.

In order to bound the sample complexity of our proposed algorithms, we will need some concepts from high-dimensional geometry. First, we define a statistical measure of complexity of a set of signals, following [12].

Definition 3.1. *(Local gaussian mean width.) For a given set $K \in \mathbb{R}^n$, the local gaussian mean width (or simply, local mean width) is defined as follows $\forall\ t > 0$:*

$$W_t(K) = \mathbb{E} \sup_{x,y\in K, \|x-y\|_2\leq t} \langle g, x - y\rangle.$$

where $g \sim \mathcal{N}(0, I_{n\times n})$.

Next, we define the notion of a *polar norm* with respect to a given subset Q of the signal space:

Definition 3.2. *(Polar norm.) For a given $x \in \mathbb{R}^n$ and a subset of $Q \in \mathbb{R}^n$, the polar norm with respect to Q is defined as follows:*

$$\|x\|_{Q^\circ} = \sup_{u\in Q} \langle x, u\rangle.$$

Furthermore, for a given subset of $Q \in \mathbb{R}^n$, we define $Q_t = (Q - Q) \cap tB_2^n$. Since Q_t is a symmetric set, one can show that the polar norm with respect to Q_t defines a semi-norm. Next, we use the following standard notions from random matrix theory [43]:

Definition 3.3. *(Subgaussian random variable.) A random variable X is called subgaussian if it satisfies the following:*

$$\mathbb{E} \exp\left(\frac{cX^2}{\|X\|_{\psi_2}^2}\right) \leq 2,$$

where $c > 0$ is an absolute constant and $\|X\|_{\psi_2}$ denotes the ψ_2-norm which is defined as follows:

$$\|X\|_{\psi_2} = \sup_{p\geq 1} \frac{1}{\sqrt{p}} (\mathbb{E}|X|^p)^{\frac{1}{p}}.$$

Definition 3.4. *(Isotropic random vectors.) A random vector-valued variable $v \in \mathbb{R}^n$ is said to be isotropic if $\mathbb{E}vv^T = I_{n\times n}$.*

In order to analyze the computational aspects of our proposed algorithms (in particular, DHT), we will need the following definition from [**?**]:

Definition 3.5. *A loss function f satisfies Restricted Strong Convexity/Smoothness (RSC/RSS) if:*

$$m_{4s} \leq \|\nabla^2_\xi f(t)\| \leq M_{4s},$$

where $\xi = \text{supp}(t_1) \cup \text{supp}(t_2)$, for all $\|t_i\|_0 \leq 2s$ and $i = 1, 2$. Also, m_{4s} and M_{4s} are (respectively) called the RSC and RSS constants.

As discussed earlier, the underlying assumption in all demixing problems of the form (3.4) is that the constituent bases are sufficiently *incoherent* as per the following definition:

Definition 3.6. *(ε-incoherence.) The orthonormal bases Φ and Ψ are said to be ε-incoherent if:*

$$\varepsilon = \sup_{\substack{\|u\|_0\leq s,\ \|v\|_0\leq s \\ \|u\|_2=1,\ \|v\|_2=1}} |\langle \Phi u, \Psi v\rangle|. \tag{3.1}$$

The parameter ε is related to the so-called mutual coherence parameter of a matrix. Indeed, if we consider the (overcomplete) dictionary $\Gamma = [\Phi\, \Psi]$, then the mutual coherence of Γ is given by $\gamma = \max_{i\neq j} |(\Gamma^T\Gamma)_{ij}|$. Moreover, one can show that $\varepsilon \leq s\gamma$ [7].

We now formally establish our *signal model*. Consider a signal $x \in \mathbb{R}^n$ that is the superposition of a pair of sparse vectors in different bases, i.e.,

$$x = \Phi w + \Psi z\,, \tag{3.2}$$

where $\Phi, \Psi \in \mathbb{R}^{n\times n}$ are orthonormal bases, and $w, z \in \mathbb{R}^n$ such that $\|w\|_0 \leq s$, and $\|z\|_0 \leq s$. We define the following quantities:

$$\bar{x} = \frac{\Phi\bar{w} + \Psi\bar{z}}{\|\Phi\bar{w} + \Psi\bar{z}\|_2} = \alpha(\Phi\bar{w} + \Psi\bar{z}), \tag{3.3}$$

where $\alpha = \frac{1}{\|\Phi\bar{w}+\Psi\bar{z}\|_2}$, $\bar{w} = \frac{w}{\|w\|_2}$, $\bar{z} = \frac{z}{\|z\|_2}$. Also, define the coefficient vector, $t = [w\ z]^T \in \mathbb{R}^{2n}$. as the vector obtaining by stacking the individual coefficient vectors w and z of the component signals.

We now state our *measurement model*. Consider the nonlinear observation model:

$$y_i = g(a_i^T x) + e_i,\ i = 1 \dots m, \tag{3.4}$$

where $x \in \mathbb{R}^n$ is the superposition signal given in (3.2), and $g : \mathbb{R} \mapsto \mathbb{R}$ represents a nonlinear link function. We denote $\Theta(x) = \int_{-\infty}^{x} g(u)du$ as the integral of g. As mentioned above, depending on the knowledge of the link function g, we consider two scenarios:

1. In the first scenario, the nonlinear link function may be non-smooth, non-invertible, or even unknown. In this setting, we assume the noiseless observation model, i.e., $y = g(Ax)$. In addition, we assume that the measurement matrix is populated by i.i.d. unit normal random variables.

2. In this setup, g represents a known nonlinear, differentiable, and strictly monotonic function. Further, in this scenario, we assume that the observation y_i is corrupted by a subgaussian additive noise with $\|e_i\|_{\psi_2} \leq \tau$ for $i = 1, \dots, m$. We also assume that the additive noise has zero mean and independent from a_i, i.e., $\mathbb{E}\,(e_i) = 0$ for $i = 1, \dots, m$. In addition, we assume that the measurement matrix consists of either (2a) isotropic random vectors that are incoherent with Φ and Ψ, or (2b) populated with subgaussian random variables.

We highlight some additional clarifications for the second case. In particular, we make the following :

Assumption 3.7. *There exist nonnegative* $l_1, l_2 > 0$ *(resp., nonpositive parameters* $l_1, l_2 < 0$*) such that* $0 < l_1 \leq g'(x) \leq l_2$ *(resp.* $l_1 \leq g'(x) \leq l_2 < 0$*).*

In words, the derivative of the link function is strictly bounded either within a positive interval or within a negative interval. In this paper, we focus on the case when $0 < l_1 \leq g'(x) \leq l_2$. The analysis of the complementary case is similar.

The lower bound on $g'(x)$ guarantees that the function g is a monotonic function, i.e., if $x_1 < x_2$ then $g(x_1) < g(x_2)$. Moreover, the upper bound on $g'(x)$ guarantees that the function g is *Lipschitz* with constant l_2. Such assumptions are common in the nonlinear recovery literature [**?**, 39].[3]

In Case 2a, the vectors a_i (i.e., the rows of A) are independent isotropic random vectors. For this case, in addition to incoherence between the component bases, we also need to define a measure of *cross*-coherence between the measurement matrix A and the dictionary Γ. The following notion of cross-coherence was introduced in the early literature of compressive sensing [44]:

Definition 3.8. *(Cross-coherence.) The cross-coherence parameter between the measurement matrix* A *and the dictionary* $\Gamma = [\Phi\ \Psi]$ *is defined as follows:*

$$\vartheta = \max_{i,j} \frac{a_i^T \Gamma_j}{\|a_i\|_2}, \tag{3.5}$$

where a_i *and* Γ_j *denote the* i^{th} *row of the measurement matrix* A *and the* j^{th} *column of the dictionary* Γ*.*

The cross-coherence assumption implies that $\left\|a_i^T \Gamma_\xi\right\|_\infty \leq \vartheta$ for $i = 1, \ldots, m$, where Γ_ξ denotes the restriction of the columns of the dictionary to set $\xi \subseteq [2n]$, with $\|\xi\|_0 \leq 4s$ such that $2s$ columns are selected from each basis Φ and Ψ.

4 Algorithms and Theoretical Results

Having defined the above quantities, we now present our main results. As per the previous section, we study two distinct scenarios:

4.1 When the link function g is unknown

Recall that we wish to recover components w and z given the nonlinear measurements y and the matrix A. Here and below, for simplicity we assume that the sparsity levels s_1 and s_2, specifying the sets K_1 and K_2, are equal, i.e., $s_1 = s_2 = s$. The algorithm (and analysis) effortlessly extends to the case of unequal sparsity levels. Our proposed algorithm, that we call OneShot, is described in pseudocode form below as Algorithm 1.

The mechanism of OneShot is simple, and *deliberately* myopic. At a high level, OneShot first constructs a *linear estimator* of the target superposition signal, denoted by $\widehat{x}_{\text{lin}} = \frac{1}{m} A^T y$. Then, it performs independent projections of $\widehat{x}_{\text{lin}}$ onto the constraint sets K_1 and K_2. Finally, it combines these two projections to obtain the final estimate of the target superposition signal.

In the above description of OneShot, we have used the following *projection* operators:

$$\widehat{w} = \mathcal{P}_s(\Phi^* \widehat{x}_{\text{lin}}), \quad \hat{z} = \mathcal{P}_s(\Psi^* \widehat{x}_{\text{lin}}).$$

Here, $\mathcal{P}_s$ denotes the projection onto the set of (canonical) s-sparse signals K and can be implemented by hard thresholding, i.e., any procedure that retains the s largest coefficients of a vector (in terms of absolute

[3] Using the monotonicity property of g that arises from Assumption 3.7, one might be tempted to simply apply the inverse of the link function on the measurements y_i in (3.4) convert the nonlinear demixing problem to the more amenable case of linear demixing, and then use any algorithm (e.g., [11]) for recovery of the constituent signals. However, this naïve way could result in a large error in the estimation of the components, particularly in the presence of the noise e_i in (3.4). This issue has been also considered in [39] for generic nonlinear recovery both from a theoretical as well as empirical standpoint.

Algorithm 1 ONESHOT

Inputs: Basis matrices Φ and Ψ, measurement matrix A, measurements y, sparsity level s.
Outputs: Estimates $\widehat{x} = \Phi\widehat{w} + \Psi\widehat{z}$, $\widehat{w} \in K_1$, $\widehat{z} \in K_2$

$\widehat{x}_{\text{lin}} \leftarrow \frac{1}{m}A^T y$	{form linear estimator}
$b_1 \leftarrow \Phi^* \widehat{x}_{\text{lin}}$	{forming first proxy}
$\widehat{w} \leftarrow \mathcal{P}_s(b_1)$	{sparse projection}
$b_2 \leftarrow \Psi^* \widehat{x}_{\text{lin}}$	{forming second proxy}
$\widehat{z} \leftarrow \mathcal{P}_s(b_2)$	{sparse projection}
$\widehat{x} \leftarrow \Phi\widehat{w} + \Psi\widehat{z}$	{Estimating $\widehat{x}$}

value) and sets the others to zero[4]. Ties between coefficients are broken arbitrarily. Observe that ONESHOT is *not* an iterative algorithm, and this in fact enables us to achieve a fast running time.

We now provide a rigorous performance analysis of ONESHOT. Our proofs follow the geometric approach provided in [12], specialized to the demixing problem. In particular, we derive an upper bound on the estimation error of the *component* signals w and z, modulo scaling factors. In our proofs, we use the following result from [12], restated here for completeness.

Lemma 4.1. *(Quality of linear estimator). Given the model in Equation* (3.2)*, the linear estimator,* $\widehat{x}_{lin}$*, is an unbiased estimator of* $\bar{x}$ *(defined in* (3.3)*) up to constants. That is,* $\mathbb{E}(\widehat{x}_{lin}) = \mu\bar{x}$ *and:* $\mathbb{E}\|\widehat{x}_{lin} - \mu\bar{x}\|_2^2 = \frac{1}{m}[\sigma^2 + \eta^2(n-1)]$*, where* $\mu = \mathbb{E}(y_1\langle a_1, \bar{x}\rangle)$*,* $\sigma^2 = Var(y_1\langle a_1, \bar{x}\rangle)$*,* $\eta^2 = \mathbb{E}(y_1^2)$*.*

We now state our first main theoretical result, with the full proof provided below in Section 6.

Theorem 4.2. *(Performance of* ONESHOT*) Let* $y \in \mathbb{R}^m$ *be given the set of nonlinear measurements. Let* $A \in \mathbb{R}^{m\times n}$ *be a random matrix with i.i.d. standard normal entries. Also, let* $\Phi, \Psi \in \mathbb{R}^{n\times n}$ *are bases with incoherence* ε*, as defined in Def. 3.6. If we use* ONESHOT *to recover* w *and* z *(up to a scaling) described in equations* (3.2) *and* (3.3)*, then the estimation error of the constituent signal,* w *(similarly,* z*) satisfies the following upper bound* $\forall t > 0$*:*

$$\mathbb{E}\|\widehat{w} - \mu\alpha\bar{w}\| \le t + \frac{2\sqrt{2}\sigma}{\sqrt{m}}\left(\frac{1+\varepsilon}{\sqrt{1-\varepsilon}}\right) + 2\sqrt{2}\mu\left(\frac{\varepsilon}{\sqrt{1-\varepsilon}}\right) + \frac{2\eta}{t\sqrt{m}}W_t(K). \tag{4.1}$$

The authors of [12, 45] provide upper bounds on the local mean width $W_t(K)$ of the set of s-sparse vectors. In particular, for any $t > 0$ they show that $W_t(K) \le Ct\sqrt{s\log(2n/s)}$ for some absolute constant C. By plugging in this bound and letting $t \to 0$, we can combine components $\widehat{w}$ and $\widehat{z}$ which gives the following:

Corollary 4.3. *With the same assumptions as Theorem 4.1, the error of nonlinear estimation incurred by the final output* $\widehat{x}$ *satisfies the upper bound:*

$$\mathbb{E}\|\widehat{x} - \mu\bar{x}\| \le \frac{4\sqrt{2}\sigma}{\sqrt{m}}\left(\frac{1+\varepsilon}{\sqrt{1-\varepsilon}}\right) + 4\sqrt{2}\mu\left(\frac{\varepsilon}{\sqrt{1-\varepsilon}}\right) + \frac{C\eta}{\sqrt{m}}\sqrt{s\log\frac{2n}{s}}. \tag{4.2}$$

Corollary 4.4. *(Example quantitative result). The constants* σ, η, μ *depend on the nature of the nonlinear function* f*, and are often rather mild. For example, if* $f(x) = \text{sign}(x)$*, then we may substitute*

$$\mu = \sqrt{\frac{2}{\pi}} \approx 0.8, \qquad \sigma^2 = 1 - \frac{2}{\pi} \approx 0.6, \qquad \eta^2 = 1,$$

in the above statement. Hence, the bound in (4.2) *becomes:*

$$\mathbb{E}\|\widehat{x} - \mu\bar{x}\| \le \frac{4}{\sqrt{m}}\left(\frac{1+\varepsilon}{\sqrt{1-\varepsilon}}\right) + 4.53\left(\frac{\varepsilon}{\sqrt{1-\varepsilon}}\right) + \frac{C}{\sqrt{m}}\sqrt{s\log\frac{2n}{s}}\,. \tag{4.3}$$

[4]The typical way is to sort the coefficients by magnitude and retain the s largest entries, but other methods such as randomized selection can also be used.

Proof. Using Lemma 4.1, $\mu = \mathbb{E}(y_i \langle a_i, \bar{x} \rangle)$ where $y_i = \text{sign}(\langle a_i, x \rangle)$. Since $a_i \sim \mathcal{N}(0, I)$ and $\bar{x}$ has unit norm, $\langle a_i, \bar{x} \rangle \sim \mathcal{N}(0, 1)$. Thus, $\mu = \mathbb{E}|g| = \sqrt{\frac{2}{\pi}}$ where $g \sim \mathcal{N}(0, I)$. Moreover, we can write $\sigma^2 = \mathbb{E}(|g|^2) - \mu^2 = 1 - \frac{2}{\pi}$. Here, we have used the fact that $|g|^2$ obeys the χ_1^2 distribution with mean 1. Finally, $\eta^2 = \mathbb{E}(y_1^2) = 1$. $\square$

In contrast with demixing algorithms for traditional (linear) observation models, our estimated signal $\widehat{x}$ outputting from OneShot can differ from the true signal x by a scale factor. Next, suppose we fix $\kappa > 0$ as a small constant, and suppose that the incoherence parameter $\varepsilon = c\kappa$ for some constant c, and that the number of measurements scales as:

$$m = \mathcal{O}\left(\frac{s}{\kappa^2} \log \frac{n}{s}\right). \tag{4.4}$$

Then, the (expected) estimation error $\|\widehat{x} - \mu\bar{x}\| \leq O(\kappa)$. In other words, the *sample complexity* of OneShot is given by $m = \mathcal{O}(\frac{1}{\kappa^2} s \log(n/s))$, which resembles results for the linear observation case [11, 12][5].

We observe that the estimation error in (4.2) is upper-bounded by $\mathcal{O}(\varepsilon)$. This is meaningful only when $\varepsilon \ll 1$, or when $s\gamma \ll 1$. Per the Welch Bound [7], the mutual coherence γ satisfies $\gamma \geq 1/\sqrt{n}$. Therefore, Theorem 4.2 provides non-trivial results only when $s = o(\sqrt{n})$. This is consistent with the *square-root bottleneck* that is often observed in demixing problems; see [46] for detailed discussions.

The above theorem obtains a bound on the expected value of the estimation error. We can derive a similar upper bound that holds with high probability. In this theorem, we assume that the *measurements* y_i for $i = 1, 2, \ldots, m$ have a *sub-gaussian* distribution (according to Def. 3.3). We obtain the following result, with full proof deferred to Section 6.

Theorem 4.5. *(High-probability version of Thm. 4.2.) Let $y \in \mathcal{R}^m$ be a set of measurements with a sub-gaussian distribution. Assume that $A \in \mathbb{R}^{m \times n}$ is a random matrix with i.i.d standard normal entries. Also, assume that $\Phi, \Psi \in \mathbb{R}^{n \times n}$ are two bases with incoherence ε as in Definition 3.6. Let $0 \leq s \leq \sqrt{m}$. If we use* OneShot *to recover w and z (up to a scaling) described in* (3.2) *and* (3.3)*, then the estimation error of the output of* OneShot *satisfies the following:*

$$\|\widehat{x} - \mu\bar{x}\| \leq \frac{2\sqrt{2}\eta s}{\sqrt{m}} \left(\frac{1+\varepsilon}{\sqrt{1-\varepsilon}}\right) + 4\sqrt{2}\mu \left(\frac{\varepsilon}{\sqrt{1-\varepsilon}}\right) + \frac{C\eta}{\sqrt{m}} \sqrt{s \log \frac{2n}{s}} + 4\frac{\eta s}{\sqrt{m}}, \tag{4.5}$$

with probability at least $1 - 4\exp(-\frac{cs^2\eta^4}{\|y_1\|_{\psi_2}^4})$ where $C, c > 0$ are absolute constants. The coefficients μ, σ, and η are given in Lemma 4.1. Here, $\|y_1\|_{\psi_2}$ denotes the ψ_2-norm of the first measurement y_1 (Definition 3.3).

In Theorem 4.5, we stated the tail probability bound of the estimation error for the superposition signal, x. Similar to Theorem 4.2, we can derive a completely analogous tail probability bound in terms of the constituent signals w and z.

4.2 When the link function g is known

The advantages of OneShot is that it enables fast demixing, and can handle even unknown, non-differentiable link functions. But its primary weakness is that the sparse components are recovered only up to an arbitrary scale factor. This can lead to high estimation errors in practice, and this can be unsatisfactory in applications. Moreover, even for reliable recovery up to a scale factor, its sample complexity is inversely dependent on the estimation error. To solve these problems, we propose a different, iterative algorithm for recovering the signal components. Here, the main difference is that the algorithm is assumed to possess (perfect) knowledge of the nonlinear link function, g.

[5]Here, we use the term "sample-complexity" as the number of measurements required by a given algorithm to achieve an estimation error κ. However, we must mention that algorithms for the linear observation model are able to achieve stronger sample complexity bounds that are independent of κ.

Recall that we define $\Gamma = [\Phi\ \Psi]$ and $t = [w; z]^T$. First, we formulate our demixing problem as the minimization of a special loss function $F(t)$:

$$\begin{aligned} \min_{t\in\mathbb{R}^{2n}} \quad & F(t) = \frac{1}{m}\sum_{i=1}^{m} \Theta(a_i^T\Gamma t) - y_i a_i^T \Gamma t \\ \text{s. t.} \quad & \|t\|_0 \leq 2s. \end{aligned} \tag{4.6}$$

Observe that the loss function $F(t)$ is *not* the typical squared-error function commonly encountered in statistics and signal processing applications. In contrast, it heavily depends on the nonlinear link function g (via its integral Θ). Instead, such loss functions are usually used in GLM and SIM estimation in the statistics literature [**?**]. In fact, the objective function in (4.6) can be considered as the *sample* version of the problem:

$$\min_{t\in\mathbb{R}^{2n}} \ \mathbb{E}(\Theta(a^T\Gamma t) - y a^T\Gamma t),$$

where a, y and Γ satisfies the model (3.4). It is not hard to show that the solution of this problem satisfies $\mathbb{E}(y_i|a_i) = g(a_i^T\Gamma t)$. We note that the gradient of the loss function can be calculated in closed form:

$$\begin{aligned} \nabla F(t) &= \frac{1}{m}\sum_{i=1}^{m} \Gamma^T a_i g(a_i^T\Gamma t) - y_i \Gamma^T a_i, \\ &= \frac{1}{m}\Gamma^T A^T (g(A\Gamma t) - y). \end{aligned} \tag{4.7}$$

We now propose an *iterative* algorithm for solving (4.6) that we call it DEMIXING WITH HARD THRESHOLDING (DHT). The method is detailed in Algorithm 2. At a high level, DHT iteratively refines its estimates of the constituent signals w, z (and the superposition signal x). At any given iteration, it constructs the gradient using (4.7). Next, it updates the current estimate according to the gradient update being determined in Algorithm 2. Then, it performs hard thresholding using the operator $\mathcal{P}_{2s}$ to obtain the new estimate of the components w and z. This procedure is repeated until a stopping criterion is met. See Section 5 for the choice of stopping criterion and other details. We mention that the initialization step in Algorithm 2 is arbitrary and can be implemented (for example) by running ONESHOT and obtaining initial points $\left(x^0, w^0, z^0\right)$. We use this initialization in our simulation results.

Implicitly, we have again assumed that both component vectors w and z are s-sparse; however, as above we mention that Algorithm 2 and the corresponding analysis easily extend to differing levels of sparsity in the two components. In Algorithm 2, $\mathcal{P}_{2s}$ denotes the projection of vector $\tilde{t}^k \in \mathbb{R}^{2n}$ on the set of $2s$ sparse vectors, again implemented via hard thresholding.

We now provide our second main theoretical result, supporting the convergence analysis of DHT. In particular, we derive an upper bound on the estimation error of the constituent vector t (and therefore, the component signals w, z). The proofs of Theorems 4.6, 4.7 and 4.8 are deferred to section 6.

Theorem 4.6. *(Performance of* DHT*) Consider the measurement model* (3.4) *with all the assumptions mentioned for the second scenario in Section 3. Suppose that the corresponding objective function F satisfies the RSS/RSC properties with constants M_{6s} and m_{6s} on the set J with $\|J\|_0 \leq 6s$ such that $1 \leq \frac{M_{6s}}{m_{6s}} \leq \frac{2}{\sqrt{3}}$. Choose a step size parameter η' with $\frac{0.5}{M_{6s}} < \eta' < \frac{1.5}{m_{6s}}$. Then,* DHT *outputs a sequence of estimates (w^k, z^k) such that the estimation error of the constituent vector satisfies the following upper bound (in expectation) for any $k \geq 1$:*

$$\|t^{k+1} - t^*\|_2 \leq (2q)^k \|t^0 - t^*\|_2 + C\tau\sqrt{\frac{s}{m}}, \tag{4.8}$$

where $q = 2\sqrt{1 + \eta'^2 M_J^2 - 2\eta' m_J}$ and $C > 0$ is a constant that depends on the step size η' and the convergence rate q.

Algorithm 2 Demixing with Hard Thresholding (DHT)

Inputs: Bases Φ and Ψ, measurement matrix A, link function g, measurements y, sparsity level s, step size η'.
Outputs: Estimates $\widehat{x} = \Phi\widehat{w} + \Psi\widehat{z}$, $\widehat{w}$, $\widehat{z}$
Initialization:
$\left(x^0, w^0, z^0\right) \leftarrow$ ARBITRARY INITIALIZATION
$k \leftarrow 0$
while $k \leq N$ **do**
$t^k \leftarrow [w^k; z^k]$ {forming constituent vector}
$t_1^k \leftarrow \frac{1}{m}\Phi^T A^T(g(Ax^k) - y)$
$t_2^k \leftarrow \frac{1}{m}\Psi^T A^T(g(Ax^k) - y)$
$\nabla F^k \leftarrow [t_1^k; t_2^k]$ {forming gradient}
$\tilde{t}^k = t^k - \eta'\nabla F^k$ {gradient update}
$[w^k; z^k] \leftarrow \mathcal{P}_{2s}\left(\tilde{t}^k\right)$ {sparse projection}
$x^k \leftarrow \Phi w^k + \Psi z^k$ {estimating $\widehat{x}$}
$k \leftarrow k + 1$
end while
Return: $(\widehat{w}, \widehat{z}) \leftarrow \left(w^N, z^N\right)$

Equation (4.8) indicates that Algorithm 2 (DHT) enjoys a linear rate of convergence. In particular, for the noiseless case $\tau = 0$, this implies that Alg. 2 returns a solution with accuracy κ after $N = \mathcal{O}(\log \frac{\|t^0 - t\|_2}{\kappa})$ iterations. The proof of Theorem 4.6 leverages the fact that the objective function $F(t)$ in (4.6) satisfies the RSC/RSS conditions specified in Definition 3.5. Please refer to Section 6 for a more detailed discussion. Moreover, we observe that in contrast with ONESHOT, DHT can recover the components w and z without any ambiguity in scaling factor, as depicted in the bound (4.8). We also verify this observation empirically in our simulation results in Section 5.

Echoing our discussion in Section 3, we consider two different models for the measurement matrix A and derive upper bounds on the sample complexity of DHT corresponding to each case. First, we present the sample complexity of Alg. 2 when the measurements are chosen to be isotropic random vectors, corresponding to Case (2a) described in the introduction:

Theorem 4.7. *(Sample complexity when the rows of A are isotropic.) Suppose that the rows of A are independent isotropic random vectors. In order to achieve the requisite RSS/RSC properties of Theorem 4.6, the number of samples needs to scale as:*

$$m = \mathcal{O}(s \log n \log^2 s \log(s \log n)),$$

provided that the bases Φ and Ψ are incoherent enough.

The sample complexity mentioned in Theorem 4.7 incurs an extra (possibly parasitic) poly-logarithmic factor relative to the sample complexity of ONESHOT, stated in (4.4). However, the drawback of ONESHOT is that the sample complexity depends inversely on the estimation error κ, and therefore a very small target error would incur a high overhead in terms of number of samples.

Removing all the extra logarithmic factors remains an open problem in general (although some improvements can be obtained using the method of [47]). However, if we assume additional structure in the measurement matrix A, we can decrease the sample complexity even further. This corresponds to Case 2b.

Theorem 4.8. *(Sample complexity when the elements of A are subgaussian.) Assume that all assumptions and definitions in Theorem 4.6 holds except that the rows of matrix A are independent subgaussian isotropic random vectors. Then, in order to achieve the requisite RSS/RSC properties of Theorem 4.6, the number of samples needs to scale as:*

$$m = \mathcal{O}\left(s \log \frac{n}{s}\right),$$

provided that the bases Φ and Ψ are incoherent enough.

The leading big-Oh constant in the expression for m in Theorems 4.7 and 4.8 is somewhat complicated, and hides the dependence on the incoherence parameter ε, the mutual coherence ϑ, the RSC/RSS constants, and the growth parameters of the link function l_1 and l_2. Please see section 6 for more details.

In Theorem 4.6, we expressed the upper bounds on the estimation error in terms of the constituent vector, t. It is easy to translate these results in terms of the component vectors w and z using the triangle inequality:

$$\max\{\|w^0 - w^*\|_2, \|z^0 - z^*\|_2\} \leq \|t^0 - t^*\|_2 \leq \|w^0 - w^*\| + \|z^0 - z^*\|_2,$$

See Section 6 for proofs and futher details.

5 Experimental Results

In this section, we provide a range of numerical experiments for our proposed algorithms based on synthetic and real data. We compare the performance of OneShot and DHT with a LASSO-type technique for demixing, as well as a heuristic version of OneShot based on soft thresholding (inspired by the approach proposed in [48]). We call these methods *Nonlinear convex demixing with LASSO* or (NlcdLASSO), and *Demixing with Soft Thresholding* or DST, respectively. Before describing our simulation results, we briefly describe these two methods.

NlcdLASSO is a method proposed in [12], although it was not explicitly developed in the demixing context. Using our notation from Section 3 and 4, NlcdLASSO solves the following convex problem:

$$\begin{aligned} \min_{z,w} \quad & \|\widehat{x}_{\text{lin}} - (\Phi z + \Psi w)\|_2 \\ \text{subject to} \quad & \|w\|_1 \leq \sqrt{s}, \quad \|z\|_1 \leq \sqrt{s}. \end{aligned} \tag{5.1}$$

Here, $\widehat{x}_{\text{lin}}$ denotes the proxy of x (equal to $\frac{1}{m}A^T y$) and s denotes the sparsity level of signals w and z in basis Φ and Ψ, respectively. The constraints in problem (5.1) are convex penalties reflecting the knowledge that w and z are s-sparse and have unit ℓ_2-norm.The outputs of this algorithm are the estimates $\widehat{w}$, $\widehat{x}$, and $\widehat{x} = \Phi\widehat{w} + \Psi\widehat{z}$. To solve the optimization problem in (5.1), we have used SPGL1 [49, 50]. This solver can handle large-scale problems, which is the scenario that we have used in our experimental evaluations.

On the other hand, DST solves the optimization problem (4.6) via a convex relaxation of the sparsity constraint. In other words, this method attempts to solve the following relaxed version of the problem (4.6):

$$\min_t \quad \frac{1}{m}\sum_{i=1}^{m} \Theta(a_i^T \Gamma t) - y_i a_i^T \Gamma t + \beta\|t\|_1, \tag{5.2}$$

where $\|t\|_1$ represents l_1-norm of the constituent vector t and $\beta > 0$ denotes the tuning parameter. The solution of this problem at iteration k is given by soft thresholding operator as follows:

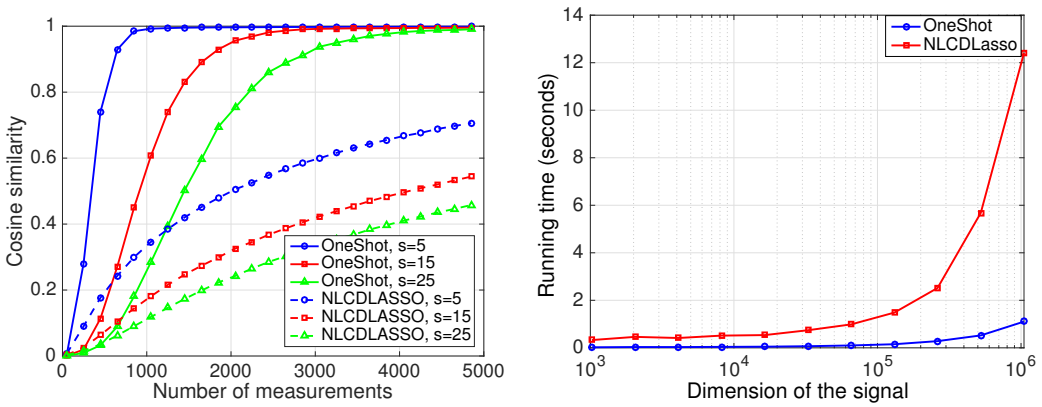
$$t^{k+1} = S_{\beta\eta'}(t^k - \eta'\nabla F(t^k)),$$

where η' denotes the step size, and the soft thresholding operator, $S_\lambda(.)$ is given by:

$$S_\beta(y) = \begin{cases} y - \beta\,, & \text{if } y > \beta \\ 0\,, & \text{if } |y| \leq \beta \\ y + \beta\,, & \text{if } y < -\beta. \end{cases}$$

Both OneShot and NlcdLASSO do not assume knowledge of the link function, and consequently return a solution up to a scalar ambiguity. Therefore, to compare performance across algorithms, we use the (scale-invariant) *cosine similarity* between the original superposition signal x and the output of a given algorithm $\widehat{x}$ defined as:

$$\cos(x, \widehat{x}) = \frac{x^T\widehat{x}}{\|x\|_2\|\widehat{x}\|_2}.$$



(a) Recovery

(b) Running time

Figure 1: *(a) Performance of* OneShot *and* NlcdLASSO *according to the* Cosine Similarity *for different choices of sparsity level s for $g(x) = sign(x)$. (b) Comparison of running times of* OneShot *with* NlcdLASSO.

5.1 Synthetic Data

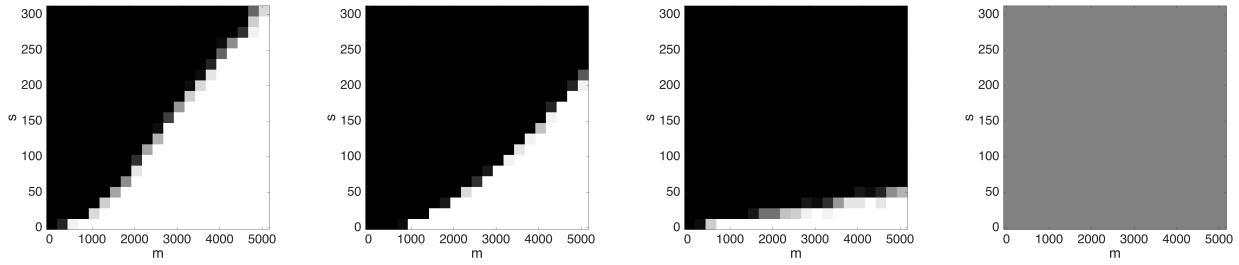
As discussed above, for successful recovery we require the constituent signals to be sufficiently incoherent. To achieve this, we choose Φ to be the 1D Haar wavelets basis, and Ψ to be the noiselet basis[6]. For the measurement operator A, we choose a partial DFT matrix. Such matrices are known to have similar recovery performance as random Gaussian matrices, but enable fast numerical operations [52]. Also, we present our experiments based on both non-smooth as well as differentiable link functions. For the non-smooth case, we choose $g(x) = \text{sign}(x)$; here, we only present recovery results using OneShot and NlcdLASSO since in our analysis DHT and DST can only handle smooth link functions.

The results of our first experiment are shown in Figure 1(a). The test signal is generated as follows: set length $n = 2^{20}$, and generate the vectors w and z by randomly selecting a signal support with s nonzero elements, and populating the nonzero entries with random Gaussian coefficients. The plot illustrates the performance of Oneshot and NlcdLASSO measured by the Cosine Similarity for different choices of sparsity level s, where the nonlinear link function is set to $f(x) = \text{sign(x)}$. The horizontal axis denotes an increasing number of measurements. Each data point in the plot is obtained by conducting a Monte Carlo experiment in which a new random measurement matrix A is generated, recording the cosine similarity between the true signal x and the reconstructed estimate and averaging over 100 trials.

As we can see, notably, the performance of NlcdLASSO is worse than OneShot for any fixed choice of m and s. Even when the number of measurements is high (for example, at $m = 4850$), we see that OneShot outperforms NlcdLASSO by a significant degree. In this case, NlcdLASSO has 30% worse in terms of signal estimation quality, while OneShot recovers the (normalized) signal perfectly. This result indicates the inefficiency of NlcdLASSO in the context of nonlinear demixing.

Next, we contrast the running time of both algorithms, illustrated in Figure 1(b). In this experiment, we measure the wall-clock running time of the two recovery algorithms (OneShot and NlcdLASSO), by varying signal size x from $n = 2^{10}$ to $n = 2^{20}$. Here, we set $m = 500$, $s = 5$, and the number of Monte Carlo trials to 1000. Also, the nonlinear link function is considered as $f(x) = \text{sign}(x)$. As we can see from the plot, OneShot is 12 times faster than NlcdLASSO when the size of signal equals to 2^{20}. Overall, OneShot is efficient even for large-scale nonlinear demixing problems. We mention that in the above setup, the main computational costs incurred in OneShot involve a matrix-vector multiplication followed by a thresholding step, both of which can be performed in time that is *nearly-linear* in terms of the signal length n for certain

[6]These bases are known to be maximally incoherent relative to each other [51]



(a) DHT (b) DST (c) OneShot (d) NlcdLASSO

Figure 2: *Phase transition plots of various algorithms for solving the demixing problem* (3.4) *as a function of sparsity level* s *and number of measurements* m *with cosine similarity as the criterion.*

choices of A, Φ, Ψ. In particular, we have experimentally verified that varying the sparsity level does not have any effect in running time.

Next, we turn to differentiable link functions. In this case, we generate the constituent signal coefficient vectors, w, z with $n = 2^{16}$, and compare performance of the four above algorithms. The nonlinear link function is chosen to be $g(x) = 2x + \sin(x)$; it is easy to check that the derivative of this function is strictly bounded between $l_1 = 1$ and $l_2 = 3$. The maximal number of iterations for both DHT and DST is set to to 1000 with an early stopping criterion if convergence is detected. The step size is hard to estimate in practice, and therefore is chosen by manual tuning such that both DHT and DST obtain the best respective performance.

Figure 2 illustrates the performance of the four algorithms in terms of *phase transition* plots, following [22]. In these plots, we varied both the sparsity level s and the number of measurements m. For each pair (s, m), as above we randomly generate the test superposition signal by choosing both the support and coefficients of x at random, as well as the measurement matrix. We repeat this experiment over 20 Monte Carlo trials. We calculate the empirical probability of successful recovery as the number of trials in which the output cosine similarity is greater than 0.99. Pixel intensities in each figure are normalized to lie between 0 and 1, indicating the probability of successful recovery.

As we observe in Fig. 2, DHT has the best performance among the different methods, and in particular, outperforms both the convex-relaxation based methods. The closest algorithm to DHT in terms of the signal recovery is DST, while the LASSO-based method fails to recover the superposition signal x (and consequently the constituent signals w and z). The improvements over OneShot are to be expected since as discussed before, this algorithm does not leverage the knowledge of the link function g and is not iterative.

In Fig. 3, we fix the sparsity level $s = 50$ and plot the probability of recovery of different algorithms with a varying number of measurements. The number of Monte Carlo trials is set to 20 and the empirical probability of successful recovery is defined as the number of trials in which the output cosine similarity is greater than 0.95. The nonlinear link function is set to be $g(x) = 2x + \sin(x)$ for figure (a) and $g(x) = \frac{1}{1+e^{-x}}$ for figure (b). As we can see, DHT has the best performance, while NlcdLASSO for figure (a) and Oneshot, and NlcdLASSO for figure (b) cannot recover the superposition signal even with the maximum number of measurements.

5.2 Real Data

In this section, we provide representative results on real-world 2D image data using Oneshot and NlcdLASSO for non-smooth link function given by $g(x) = \text{sign}(x)$. In addition, we illustrate results for all four algorithms using smooth $g(x) = \frac{1-e^{-x}}{1+e^{-x}}$ as our link function.

We begin with a 256×256 test image. First, we obtain its 2D Haar wavelet decomposition and retain the $s = 500$ largest coefficients, denoted by the s-sparse vector w. Then, we reconstruct the image based on these

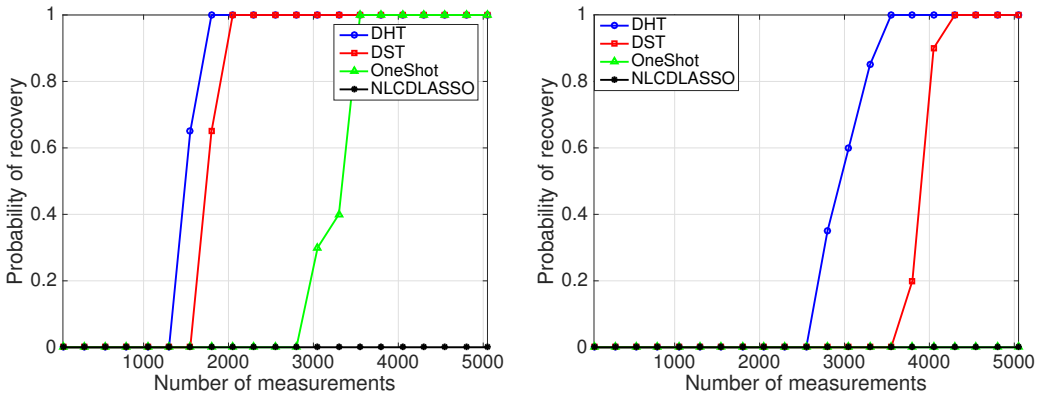


Figure 3: *Probability of recovery for four algorithms;* DHT, STM, ONESHOT, *and* NLCDLASSO. *Sparsity level is set to* $s = 50$. *(a)* $g(x) = 2x + \sin(x)$, *(b)* $g(x) = \frac{1}{1+e^{-x}}$.

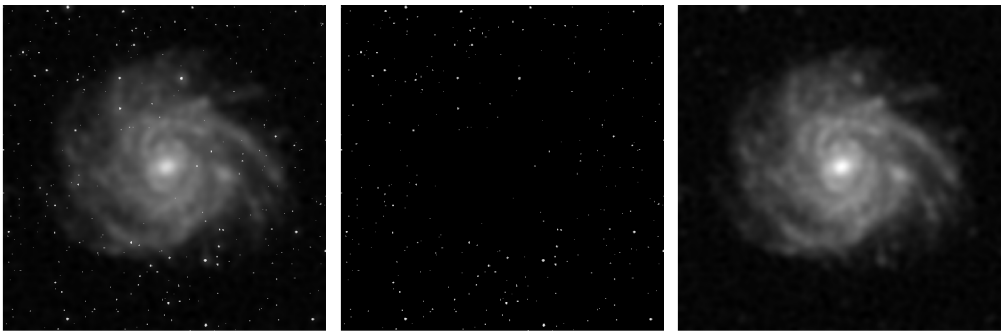
largest coefficients, denoted by $\widehat{x} = \Phi w$. Similar to the synthetic case, we generate a noise component in our superposition model based on 500 noiselet coefficients z. In addition, we consider a parameter which controls the strength of the noiselet component contributing to the superposition model. We set this parameter to 0.1. Therefore, our test image x is given by $x = \Phi w + 0.1\Psi z$.

x $\widehat{x}$ (ONESHOT) $\widehat{x}$ (NLCDLASSO)

Figure 4: *Comparison of* ONESHOT *and* NLCDLASSO *for real 2D image data from nonlinear under-sampled observations. Parameters:* $n = 256 \times 256, s = 500, m = 35000, g(x) = sign(x)$.

Figure 4 illustrates both the true and the reconstructed images x and $\widehat{x}$ using ONESHOT and NLCDLASSO. The number of measurements is set to 35000 (using subsampled Fourier matrix with $m = 35000$ rows). From visual inspection we see that the reconstructed image, $\widehat{x}$, using ONESHOT is better than the reconstructed image by NLCDLASSO. Quantitatively, we also calculate Peak signal-to-noise-ratio (PSNR) of the reconstructed images using both algorithms relative to the test image, x. We obtain PSNR of 19.8335 dB using ONESHOT, and a PSNR of 17.9092 dB using NLCDLASSO, again illustrating the superior performance of ONESHOT compared to NLCDLASSO.

Next, we show our results using a differentiable link function. For this experiment, we consider an astronomical image illustrated in Fig. 5. This image includes two components; the "stars" component, which can be considered to be sparse in the identity basis (Φ), and the "galaxy" component which are sparse when they are expressed in the discrete cosine transform basis (Ψ). The superposition image $x = \Phi w + \Psi z$



(a) Original x (b) $\Phi(\widehat{w})$ $\Psi(\widehat{z})$

Figure 5: *Successful demixing on a real 2-dimensional image from nonlinear under-sampled observations with* DHT. *Parameters:* $n = 512 \times 512, s = 1000, m = 15000, g(x) = \frac{1-e^{-x}}{1+e^{-x}}$. *Image credits: NASA and [23].*

is observed using a subsampled Fourier matrix with $m = 15000$ rows multiplied with a diagonal matrix with random ± 1 entries [53]. Further, each measurement is nonlinearly transformed by applying the (shifted) logistic function $g(x) = \frac{1-e^{-x}}{1+e^{-x}}$ as the link function. In the recovery procedure using DHT, we set the number of iterations to 1000 and step size η' to 150000. As is visually evident, our proposed DHT method is able to reliably recover the component signals.

6 Proofs

In this section, we derive the proofs of our theoretical results stated in Section 4.

6.1 Analysis of OneShot

Our analysis mostly follows the techniques of [12]. However, several additional complications in the proof arise due to the structure of the demixing problem. As a precursor, we need the following lemma from geometric functional analysis, restated from [12].

Lemma 6.1. *Assume K is a closed star-shaped set. Then for $u \in K$, and $a \in \mathbb{R}^n$, one has the following result $\forall\ t > 0$:*

$$\|\mathcal{P}_K(a) - u\|_2 \leq \max\left(t, \frac{2}{t}\|a - u\|_{K_t^o}\right). \tag{6.1}$$

We also use the following result of [12].

Claim 6.2. *(Orthogonal decomposition of a_i.) Suppose we decompose the rows of A, a_i, as:*

$$a_i = \langle a_i, \bar{x}\rangle \bar{x} + b_i, \tag{6.2}$$

where $b_i \in \mathbb{R}^n$ is orthogonal to $\bar{x}$. Then we have $b_i \sim \mathcal{N}(0, I_{x^\perp})$ since $a_i \sim \mathcal{N}(0, I)$. Also, $I_{x^\perp} = I - \bar{x}\bar{x}^T$. Moreover, the measurements y_i in equation (3.4) *and the orthogonal component b_i are statistically independent.*

Proof of Theorem 4.2. Observe that the magnitude of the signal x may be lost due to the action of the non-linear measurement function f (such as the sign$(\cdot)$ function). Therefore, our recovered signal $\widehat{x}$ approximates

the true signal modulo a scaling factor. Indeed, for μ defined in Lemma 4.1, we have:

$$\begin{aligned}\|\widehat{x} - \mu\bar{x}\|_2 &= \|\Phi\widehat{w} + \Psi\widehat{z} - \alpha\mu\Phi\bar{w} - \alpha\mu\Psi\bar{z}\|_2 \\ &\leq \|\Phi\|\|\widehat{w} - \mu\alpha\bar{w}\|_2 + \|\Psi\|\|\widehat{z} - \mu\alpha\bar{z}\|_2 \\ &\leq \left(t + \frac{2}{t}\|\Phi^*\widehat{x}_{\text{lin}} - \mu\alpha\bar{w}\|_{K_t^o}\right) + \left(t + \frac{2}{t}\|\Psi^*\widehat{x}_{\text{lin}} - \mu\alpha\bar{z}\|_{K_t^o}\right).\end{aligned}$$

The equality comes from the definition of $\bar{x}$. The first inequality results from an application of the triangle inequality and the definition of the operator norm of a matrix, while the second inequality follows from Lemma 6.1.

It suffices to derive a bound on the first term in the above expression (since a similar bound will hold for the second term.) This proves the first part of Theorem 4.2. We have:

$$\begin{aligned}\|\Phi^*\widehat{x}_{\text{lin}} - \mu\alpha\bar{w}\|_{K_t^o} &= \|\Phi^*\frac{1}{m}\Sigma_i(y_i\langle a_i, \bar{x}\rangle\bar{x} + y_i b_i) - \mu\alpha\bar{w}\|_{K_t^o} \\ &\leq \|\Phi^*\frac{1}{m}\Sigma_i(y_i\langle a_i, \bar{x}\rangle\bar{x}) - \mu\alpha\bar{w}\|_{K_t^o} + \|\Phi^*\frac{1}{m}\Sigma_i y_i b_i\|_{K_t^o} \\ &\leq \underbrace{\|\Phi^*\frac{1}{m}\Sigma_i(y_i\langle a_i, \bar{x}\rangle\bar{x}) - \mu\Phi^*\bar{x}\|_{K_t^o}}_{S_1} + \underbrace{\|\mu\alpha\Phi^*\Psi\bar{z}\|_{K_t^o}}_{S_2} + \underbrace{\|\Phi^*\frac{1}{m}\Sigma_i y_i b_i\|_{K_t^o}}_{S_3}. \qquad (6.3)\end{aligned}$$

The first equality follows from Claim 6.2, while the second and third inequalities result from the triangle inequality. We first bound S_1 as follows:

$$\begin{aligned}S_1 &= \|\Phi^*\frac{1}{m}\Sigma_i(y_i\langle a_i, \bar{x}\rangle\bar{x}) - \mu\Phi^*\bar{x}\|_{K_t^o} \\ &= \|(\frac{1}{m}\Sigma_i(y_i\langle a_i, \bar{x}\rangle - \mu))\Phi^*\bar{x}\|_{K_t^o} \\ &= |\frac{1}{m}\Sigma_i(y_i\langle a_i, \bar{x}\rangle - \mu)|\|\Phi^*\bar{x}\|_{K_t^o}.\end{aligned}$$

Therefore,

$$\mathbb{E}(S_1^2) = \mathbb{E}(|\frac{1}{m}\Sigma_i(y_i\langle a_i, \bar{x}\rangle - \mu)|^2\|\Phi^*\bar{x}\|_{K_t^o}^2).$$

Define $\gamma_i \stackrel{\Delta}{=} y_i\langle a_i, \bar{x}\rangle - \mu_i$. Then,

$$\begin{aligned}\mathbb{E}(|\frac{1}{m}\Sigma_i(y_i\langle a_i, \bar{x}\rangle - \mu)|^2) &= \mathbb{E}(\frac{1}{m^2}(\Sigma_i\gamma_i)^2) \\ &= \mathbb{E}(\frac{1}{m^2}(\sum_{i=}^{m}\gamma_i^2 + \Sigma_{i\neq j}\gamma_i\gamma_j)) \\ &= \frac{1}{m^2}(\sum_{i=1}^{m}\mathbb{E}\gamma_i^2) = \frac{1}{m}\mathbb{E}\gamma_1^2 \\ &= \frac{\sigma^2}{m}\end{aligned}$$

where σ^2 has been defined in Lemma 4.1. The third and last equalities follow from the fact that the y_i's are independent and identically distributed. Now, we bound $\|\Phi^*\bar{x}\|_{K_t^o}^2$ as follows::

$$
\begin{aligned}
\|\Phi^*\bar{x}\|_{K_t^o} &= \sup_{u\in(K-K)\cap tB_n^2} \langle \Phi^*\bar{x}, u\rangle \\
&= t \sup_{\substack{v_1\in\frac{1}{t}K, v_2\in\frac{1}{t}K \\ \|v_i\|\le 1, i=1,2}} \langle \Phi^*\bar{x}, v_1 - v_2\rangle \\
&\le 2t \sup_{\substack{\|a\|_0\le s \\ \|a\|\le 1}} |\langle \Phi^*\bar{x}, a\rangle| \\
&\le 2t\Big(\sup_{\substack{\|a\|_0\le s \\ \|a\|\le 1}} |\langle \alpha\bar{w}, a\rangle| + \sup_{\substack{\|a\|_0\le s \\ \|a\|\le 1}} |\langle \alpha\Phi^*\Psi\bar{z}, a\rangle|\Big) \\
&\le 2\alpha t(1 + \sup_{\substack{\|a\|_0\le s \\ \|a\|\le 1}} |\langle \alpha\Psi\bar{z}, \Phi a\rangle|) \\
&= 2\alpha t(1+\varepsilon).
\end{aligned}
$$

This implies that:

$$
\implies \mathbb{E}(S_1^2) \le 4\frac{\alpha^2 t^2\sigma^2}{m}(1+\varepsilon)^2. \tag{6.4}
$$

The second inequality follows from (3.2) and the triangle inequality. The last inequality is results from an application of the Cauchy-Schwarz inequality and the definition of ε. Similarly we can bound S_2 as follows:

$$
\begin{aligned}
\mathbb{E}(S_2) &= \mathbb{E}(\|\mu\alpha\Phi^*\Phi\bar{z}\|_{K_t^o}) \\
&= \mathbb{E}(|\mu\alpha|\|\Phi^*\Phi\bar{z}\|_{K_t^o}) \\
&= |\mu\alpha|\|\Phi^*\Phi\bar{z}\|_{K_t^o} \\
&= |\mu\alpha| \sup_{u\in(K-K)\cap tB_n^2} \langle \Psi\bar{z}, \Phi u\rangle \\
&= |\mu\alpha| t \sup_{\substack{v_1\in\frac{1}{t}K, v_2\in\frac{1}{t}K \\ \|v_i\|\le 1, i=1,2}} \langle \Psi\bar{z}, \Phi(v_1 - v_2)\rangle \\
&\le 2\mu\alpha t\varepsilon.
\end{aligned} \tag{6.5}
$$

Finally, we give the bound for S_3. Define $L \triangleq \frac{1}{m}\Sigma_i y_i b_i$. Then, we get:

$$
\mathbb{E}(S_3) = \mathbb{E}\|\Phi^*\frac{1}{m}\Sigma_i y_i b_i\|_{K_t^o} = \mathbb{E}\|\Phi^* L\|_{K_t^o}.
$$

Our goal is to bound $\mathbb{E}\|\Phi^* L\|_{K_t^o}$. Since y_i and b_i are independent random variables (as per Claim 6.2), we can use the law of conditional covariance and the law of iterated expectation. That is, we first condition on y_i, and then take expectation with respect to b_i.

By conditioning on y_i, we have $L \sim \mathcal{N}(0, \beta^2 I_{x^\perp})$ where $I_{x^\perp} = I - \bar{x}\bar{x}^T$ is the covariance of vector b_i according to claim 6.2 and $\beta^2 = \frac{1}{m^2}\Sigma_i y_i^2$. Define $g_{x^\perp} \sim \mathcal{N}(0, I_{x^\perp})$. Therefore, $L = \beta g_{x^\perp}$ Putting everything together, we get:

$$
\begin{aligned}
\mathbb{E}(S_3) &= \mathbb{E}\|\Phi^* L\|_{K_t^o} \\
&= \mathbb{E}\|\Phi^*\beta g_{x^\perp}\|_{K_t^o} \\
&= \beta\mathbb{E}\|\Phi^* g_{x^\perp}\|_{K_t^o}.
\end{aligned}
$$

We need to extend the support of distribution of $g_{x^\perp}$ and consequently L from $x^\perp$ to $\mathbb{R}^n$. This is done by the following claim in [12]:

Claim 6.3. *Let g_E be a random vector which is distributed as $\mathcal{N}(0, I_E)$. Also, assum that $\Gamma : \mathbb{R}^n \to \mathbb{R}$ is a convex function. Then, for any subspace E of $\mathbb{R}^n$ such that $E \subseteq F$, we have:*

$$\mathbb{E}(\Gamma(g_E)) \leq \mathbb{E}(\Gamma(g_F)).$$

Hence, we can orthogonally decompose $\mathbb{R}^n$ as $\mathbb{R}^n = D \oplus C$ where D is a subspace supporting $x^\perp$ and C is the orthogonal subspace onto it. Thus, $g_{\mathbb{R}^n} = g_D + g_C$ in distribution such that $g_D \sim \mathcal{N}(0, I_D)$, $g_C \sim \mathcal{N}(0, I_C)$. Also, $\|.\|_{K_t^o}$ is a convex function since it is a semi-norm. Hence,

$$\begin{aligned}
\mathbb{E}_D\|\Phi^* g_D\|_{K_t^o} &= \mathbb{E}_D\|\Phi^* g_D + \mathbb{E}_C(g_C)\|_{K_t^o} \\
&= \mathbb{E}_D\|\mathbb{E}_{C|D}(\Phi^* g_D + g_c)\|_{K_t^o} \\
&\leq \mathbb{E}_D\mathbb{E}_{C|D}\|\Phi^*(g_D + g_C)\|_{K_t^o} \\
&= \mathbb{E}\|\Phi^* g_{\mathbb{R}^n}\|_{K_t^o}.
\end{aligned}$$

The first inequality follows from Jensen's inequality, while the second inequality follows from the law of iterated expectation. Therefore, we get:

$$\begin{aligned}
\mathbb{E}\|\Phi^* L\|_{K_t^o} &= \mathbb{E}\|\Phi^* \beta g_{x^\perp}\|_{K_t^o} \\
&= \beta\mathbb{E}\|\Phi^* g_{x^\perp}\|_{K_t^o} \\
&\leq \beta\mathbb{E}\|\Phi^* g_{\mathbb{R}^n}\|_{K_t^o} \\
&= \beta \sup_{u \in (K-K)\cap tB_n^2} \langle \Phi^* \beta g_{\mathbb{R}^n}, u\rangle \\
&= \beta W_t(K).
\end{aligned}$$

The last equality follows from the fact that $\Phi^* g_{\mathbb{R}^n} \sim \mathcal{N}(0, I)$. The final step is to take an expectation with respect to y_i, giving us a bound on $\mathbb{E}(S_3)$:

$$\begin{aligned}
\mathbb{E}(S_3) &= \mathbb{E}\|\Phi^* L\|_{K_t^o} \\
&\leq \mathbb{E}(\beta) W_t(K) \\
&\leq \sqrt{\mathbb{E}(\beta^2)} W_t(K)\,,
\end{aligned}$$

where $\beta^2 = \frac{1}{m^2}\sum_{i=1}^m y_i^2$. Hence,

$$\mathbb{E}(S_3) \leq \frac{\eta}{\sqrt{m}} W_t(K)\,. \tag{6.6}$$

Putting together the results from (6.4), (6.5), and (6.6), we have:

$$\begin{aligned}
\mathbb{E}(\|\Phi^* \widehat{x}_{\text{lin}} - \mu\alpha\bar{w}\|_{K_t}) &\leq \mathbb{E}(S_1) + \mathbb{E}(S_2) + \mathbb{E}(S_3) \\
&\leq \sqrt{\mathbb{E}(S_1)} + \mathbb{E}(S_2) + \mathbb{E}(S_3) \\
&\leq \frac{2\alpha t\sigma}{\sqrt{m}}(1+\varepsilon) + 2\mu\alpha t\varepsilon + \frac{\eta}{\sqrt{m}} W_t(K).
\end{aligned}$$

Therefore, we obtain:

$$\mathbb{E}\|\widehat{w} - \mu\alpha\bar{w}\|_2 \leq t + \frac{2}{t}\mathbb{E}(\|\Phi^* \widehat{x}_{lin} - \mu\alpha\bar{w}\|_{K_t}) \leq t + \frac{4\alpha\sigma}{\sqrt{m}}(1+\varepsilon) + 4\mu\alpha\varepsilon + \frac{2\eta}{t\sqrt{m}} W_t(K). \tag{6.7}$$

Moreover, we can bound $\alpha = \frac{1}{\|\Phi\bar{w} + \Psi\bar{z}\|}$ as follows:

$$\begin{aligned}
\|\Phi\bar{w} + \Psi\bar{z}\|_2^2 &\geq \|\Phi\bar{w}\|_2^2 + \|\Psi\bar{z}\|_2^2 - 2|\langle \Phi\bar{w}, \Psi z\rangle| \\
&\geq 2 - 2\varepsilon, \quad \text{or,} \\
\alpha &\leq \frac{1}{\sqrt{2}\sqrt{1-\varepsilon}}.
\end{aligned}$$

By plugging α in (6.7), we obtain the desired result in Theorem 4.2. However, K is a closed star-shaped set (the set of s-sparse signals), and therefore $W_t(K) = tW_1(K)$ [12]. Now using (6.3), we can conclude the Corollary 4.3 (bound on the estimation error for superposition signal):

$$\mathbb{E}(\|\widehat{x} - \mu\bar{x}\|) \leq 2t + \frac{8\alpha\sigma}{\sqrt{m}}(1+\varepsilon) + 8\mu\alpha\varepsilon + \frac{4\eta}{t\sqrt{m}}W_t(K).$$

We can use Lemma 2.3 in [45] and plug in $W_t(K) \leq Ct\sqrt{s\log(2n/s)}$. Using the above bound on α and by letting $t \to 0$, we get:

$$\mathbb{E}\|\widehat{x} - \mu\bar{x}\| \leq \frac{4\sqrt{2}\sigma}{\sqrt{m}}\left(\frac{1+\varepsilon}{\sqrt{1-\varepsilon}}\right) + 4\sqrt{2}\mu\left(\frac{\varepsilon}{\sqrt{1-\varepsilon}}\right) + \frac{C\eta}{\sqrt{m}}\sqrt{s\log\frac{2n}{s}}\,, \tag{6.8}$$

where $C > 0$ is an absolute constant. This completes the proof of Corollary 4.3. □

We now prove the high-probability version of the main theorem. As a precursor, we need a few preliminary definitions and lemmas:

Definition 6.4. *(Subexponential random variable.) A random variable X is subexponential if it satisfies the following relation:*

$$\mathbb{E}\exp\left(\frac{cX}{\|X\|_{\psi_1}}\right) \leq 2,$$

where $c > 0$ is an absolute constant. Here, $\|X\|_{\psi_1}$ denotes the ψ_1-norm, defined as follows:

$$\|X\|_{\psi_1} = \sup_{p\geq 1}\frac{1}{p}(\mathbb{E}|X|^p)^{\frac{1}{p}}.$$

We should mention that there are other definitions for subexponential random variables (also for sub-Gaussian defined in Definition 3.3). Please see [43] for a detailed treatment.

Lemma 6.5. *Let X and Y be two subgaussian random variables. Then, XY is a subexponential random variable.*

Proof. According to the definition of the ψ_2-norm, we have:

$$(\mathbb{E}|XY|^p)^{\frac{1}{p}} = (\mathbb{E}|X|^p|Y|^p)^{\frac{1}{p}} \leq \left(\left(\mathbb{E}|X|^{2p}\right)^{\frac{1}{2p}}\left(\mathbb{E}|Y|^{2p}\right)^{\frac{1}{2p}}\right) \leq \sqrt{2}p\|X\|_{\psi_2}\|Y\|_{\psi_2}, \tag{6.9}$$

where the first inequality results from Cauchy-schwarz inequality, and the last inequality is followed by the subgaussian assumption on X and Y. This shows that the random variable XY is subexponential random variable according to Definition 6.4. □

Lemma 6.6. *(Gaussian concentration inequality) See [43, 54]. Let $(G_x)_{x\in T}$ be a centered gaussian process indexed by a finite set T. Then $\forall t > 0$:*

$$\mathbb{P}(\sup_{x\in T} G_x \geq \mathbb{E}\sup_{x\in T} G_x + t)) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right)$$

where $\sigma^2 = \sup_{x\in T}\mathbb{E}G_x^2 < \infty$.

Lemma 6.7. *(Bernstein-type inequality for random variables) [43]. Let $X_1, X_2, \ldots, X_n$ be independent sub-exponential random variables with zero-mean. Also, assume that $K = \max_i \|X_i\|_{\psi_1}$. Then, for any vector $a \in \mathbb{R}^n$ and every $t \geq 0$, we have:*

$$\mathbb{P}(|\Sigma_i a_i X_i| \geq t) \leq 2\exp\left(-c\min\left\{\frac{t^2}{K^2\|a\|_2^2}, \frac{t}{K\|a\|_\infty}\right\}\right).$$

where $c > 0$ is an absolute constant.

Proof of Theorem 4.5. We follow the proof given in [12]. Let $\beta = \frac{s}{2\sqrt{m}}$ for $0 < s < \sqrt{m}$ where m denotes the number of measurements. In (6.3), we saw that

$$\|\widehat{x} - \mu\bar{x}\|_2 \leq 2(t + \frac{2}{t}(S_1 + S_2 + S_3)). \tag{6.10}$$

We attempt to bound each term S_1, S_2, and S_3 with high probability, and then use a union bound to obtain the desired result.

For S_1, we have:

$$S_1 \leq |\frac{1}{m}\Sigma_i(y_i\langle a_i, \bar{x}\rangle - \mu)|\|\Phi^*\bar{x}\|_{K_t^o}.$$

We note that y_i is a sub-gaussian random variable (by assumption) and $\langle a_i, \bar{x}\rangle$ is a standard normal random variable. Hence, by Lemma 6.5, $y_i\langle a_i, \bar{x}\rangle$ is a sub-exponential random variable. Also, $y_i\langle a_i, \bar{x}\rangle$ for $i = 1, 2, \ldots, m$ are independent sub-exponential random variables that can be centered by subtracting their mean μ. Now, we can apply Lemma 6.7 on $|\frac{1}{m}\Sigma_i(y_i\langle a_i, \bar{x}\rangle - \mu)|$. Therefore:

$$\mathbb{P}(|\frac{1}{m}\Sigma_i(y_i\langle a_i, \bar{x}\rangle - \mu))| \geq \eta\beta) \leq 2\exp\left(-\frac{c\beta^2\eta^2 m}{\|y_1\|_{\psi_2}^2}\right).$$

Here, η and μ are as defined in 4.1. Using the bound on $\|\Phi^*\bar{x}\|_{K_t^o}$, we have:

$$S_1 \leq \sqrt{2}\eta\beta t\frac{1+\varepsilon}{\sqrt{1-\varepsilon}}, \tag{6.11}$$

with probability at least $1 - 2\exp(-\frac{c\beta^2\eta^2 m}{\|y_1\|_{\psi_2}^2})$ where $c > 0$ is some constant.

For S_2 we have:

$$S_2 \leq \sqrt{2}\mu\alpha t\frac{\varepsilon}{\sqrt{1-\varepsilon}}, \tag{6.12}$$

with probability 1 since S_2 is a deterministic quantity.

For S_3 we have:

$$S_3 \leq \|\Phi^*\frac{1}{m}\Sigma_i y_i b_i\|_{K_t^o}.$$

To obtain a tail bound for S_3, we are using the following:

$$S_3 \leq \frac{1}{m}(\Sigma_i y_i^2)^{1/2}\|\Phi^* g\|_{K_t^o}$$

We need to invoke the Bernstein Inequality (Lemma 6.7) for sub-exponential random variables $(y_i^2 - \eta^2)$ for $i = 1, 2, \ldots, m$ which are zero mean subexponential random variables in order to bound $\frac{1}{m}(\Sigma_i y_i^2)^{1/2}$. we have $\left|\frac{1}{m}\Sigma_i(y_i^2 - \eta^2)\right| \leq 3\eta^2$ with high probability $1 - 2\exp(-\frac{cm\eta^4}{\|y_1\|_{\psi_2}^4})$.

Next, we upper-bound $\|\Phi^* g\|$ (where $g \sim \mathcal{N}(0, I)$) with high probability. Since Φ is an orthogonal matrix, we have that $\Phi^* g \sim \mathcal{N}(0, I)$. Hence, we can use the Gaussian concentration inequality to bound $\Phi^* g$ as mentioned in Lemma 6.6. Putting these pieces together, we have:

$$S_3 \leq \frac{2\eta}{\sqrt{m}}\left(W_t(K) + t\beta\sqrt{m}\right), \tag{6.13}$$

with probability at least $1 - 2\exp(-\frac{cm\eta^4}{\|y_1\|_{\psi_2}^4}) - \exp(c\beta^2 m)$. Here, $W_t(K)$ denotes the local mean width for the set K_1 defining in Definition 3.1.

Now, combining (6.10), (6.11), (6.12), and (6.13) together with the union bound, we obtain:

$$\|\widehat{x} - \mu\bar{x}\|_2 \leq \frac{2\sqrt{2}\eta s}{\sqrt{m}}\left(\frac{1+\varepsilon}{\sqrt{1-\varepsilon}}\right) + 4\sqrt{2}\mu\left(\frac{\varepsilon}{\sqrt{1-\varepsilon}}\right) + \frac{C\eta}{\sqrt{m}}\sqrt{s\log\frac{2n}{s}} + 4\frac{\eta s}{\sqrt{m}},$$

with probability at least $1 - 4\exp(-\frac{cs^2\eta^4}{\|y_1\|_{\psi_2}^4})$ where $C, c > 0$ are absolute constants. Here, we have again used the well-known bound on the local mean width of the set of sparse vectors (for example, see Lemma 2.3 of [45]). This completes the proof. □

6.2 Analysis of DHT

Our analysis of DHT occurs in two stages. First, we define a loss function $F(t)$ that depends on the nonlinear link function g and the measurement matrix A. We first assume that $F(t)$ satisfies certain regularity conditions (restricted strong convexity/smoothness), and use this to prove algorithm convergence. The proof of Theorem 4.6 follows the proof of convergence of the iterative hard thresholding (IHT) algorithm in the linear case [40], and is more closely related to the work of [17] who extended it to the nonlinear setting. Our derivation here differs from these previous works in our specific notion of restricted strong convexity/smoothness, and is relatively more concise. Later, we will prove that the RSS/RSC assumptions on the loss function indeed are valid, given a sufficient number of samples that obey certain measurement models. We assume a variety of measurement models including isotropic row measurements as well as subgaussian measurements. To our knowledge, these derivations of sample complexity are novel.

First, we state the definitions for restricted strong convexity and restricted strong smoothness, abbreviated as *RSC* and *RSS*. The RSC and RSS was first proposed by [**?**, 55]; also, see [16].

Definition 6.8. *A function f satisfies the RSC and RSS conditions if one of the following equivalent definitions is satisfied for all t_1, t_2 such that $\|t_1\|_0 \leq 2s$ and $\|t_2\|_0 \leq 2s$:*

$$\frac{m_{4s}}{2}\|t_2 - t_1\|_2^2 \leq f(t_2) - f(t_1) - \langle \nabla f(t_1), t_2 - t_1\rangle \leq \frac{M_{4s}}{2}\|t_2 - t_1\|_2^2, \tag{6.14}$$

$$m_{4s}\|t_2 - t_1\|_2^2 \leq \langle \nabla f(t_2) - \nabla f(t_1), t_2 - t_1\rangle \leq M_{4s}\|t_2 - t_1\|_2^2, \tag{6.15}$$

$$m_{4s} \leq \|\nabla_\xi^2 f(t)\| \leq M_{4s}, \tag{6.16}$$

$$m_{4s}\|t_2 - t_1\|_2 \leq \|\nabla_\xi f(t_2) - \nabla_\xi f(t_1)\|_2 \leq M_{4s}\|t_2 - t_1\|_2, \tag{6.17}$$

where $\xi = supp(t_1) \cup supp(t_2)$, $\|\xi\|_0 \leq 4s$. Moreover, m_{4s} and M_{4s} are called the RSC-constant and RSS-constant, respectively. We note that $\nabla_\xi f(t)$ denotes the gradient f restricted to set ξ. In addition, $\nabla_\xi^2 f(t)$ is a $4s \times 4s$ sub-matrix of the Hessian matrix $\nabla^2 f(t)$ comprised of row/column indices indexed by ξ.

Proof. (Equivalence of Eqs. (6.14), (6.15), (6.16), (6.17)). The proof of above equivalent definitions only needs some elementary arguments and we state them here for completeness. If we assume that (6.14) is given, then by exchanging t_1 and t_2 in (6.14), we have:

$$\frac{m_{4s}}{2}\|t_1 - t_2\|_2^2 \leq f(t_1) - f(t_2) - \langle \nabla f(t_2), t_1 - t_2\rangle \leq \frac{M_{4s}}{2}\|t_1 - t_2\|_2^2, \tag{6.18}$$

by adding (6.18) with (6.14), inequality in (6.15) is resulted. Now, assume that (6.15) is given. Then we can set $t_2 = t_1 + \Delta(t_2 - t_1)$ in (6.15) and then letting $\Delta \to 0$ results (6.16) according to the definition of

second derivative. Next, if we assume that (6.16) is given, then we can invoke the *mean value theorem* [56] for twice-differentiable vector-valued multivariate functions:

$$\nabla_\xi f(t_2) - \nabla_\xi f(t_1) = \int_0^1 P_\xi^T \nabla^2 f(ct_2 + (1-c)t_1)(t_2 - t_1)dt.$$

where $c > 0$ and P_ξ denotes the identity matrix which its columns is restricted to set ξ with $\|\xi\|_0 \leq 2s$. It follows that:

$$\begin{aligned}\left\|\nabla_\xi f(t_2) - \nabla_\xi f(t_1)\right\| &\leq \int_0^1 \left\|P_\xi^T \nabla^2 f(ct_2 + (1-c)t_1)(t_2 - t_1)\right\| dt \\ &\leq M_{4s}\|(t_2 - t_1)\|.\end{aligned}$$

where the last inequality follows by (6.16). Similarly, we can establish the lower bound in (6.17) by invoking the Cauchy Schwartz inequality in (6.15).

Finally, suppose that (6.16) holds. We can establish (6.14) by performing a Taylor expansion of $f(t)$. For upper bound in (6.14) and some $0 \leq c \leq 1$, we have:

$$\begin{aligned}f(t_2) &\leq f(t_1) - \langle \nabla f(t_1), t_2 - t_1\rangle + \frac{1}{2}(t_2 - t_1)^T \nabla_\xi^2 f(ct_2 + (1-c)t_1)(t_2 - t_1) \\ &\leq f(t_1) - \langle \nabla f(t_1), t_2 - t_1\rangle + \frac{M_{4s}}{2}\|t_1 - t_2\|_2^2.\end{aligned}$$

The lower bound in (6.14) also follows similarly. □

We now give a proof that DHT enjoys the linear convergence, as stated in Theorem 4.6. Recall that as opposed to the commonly used least-squares loss function, we instead define a special objective function:

$$F(t) = \frac{1}{m}\sum_{i=1}^m \Theta(a_i^T \Gamma t) - y_i a_i^T \Gamma t,$$

where $\Gamma = [\Phi\ \Psi]$, $t = [w\ z]^T$, and $\Theta(x) = \int_{-\infty}^x g(u)du$ with $g(x)$. The gradient and Hessian of the objective function are given as follows:

$$\nabla F(t) = \frac{1}{m}\sum_{i=1}^m \Gamma^T a_i g(a_i^T \Gamma t) - y_i \Gamma^T a_i\,, \tag{6.19}$$

$$\nabla^2 F(t) = \frac{1}{m}\sum_{i=1}^m \Gamma^T a_i a_i^T \Gamma g'(a_i^T \Gamma t)\,. \tag{6.20}$$

We start with the projection step in Algorithm 2. In what follows, the superscript k denotes the k-th iteration. Let $t^{k+1} = [t_1^k\ t_2^k]^T \in \mathbb{R}^{2n}$ be the constituent vector as the k^{th} iteration. Hence,

$$t^{k+1} = \mathcal{P}_{2s}\left(t^k - \eta' \nabla F(t^k)\right),$$

where η' denotes the step size in Algorithm 2 and $\mathcal{P}_{2s}(.)$ denotes the hard thresholding operation. Furthermore, $\nabla F(t^k)$ is the gradient of the objective function at iteration k. Moreover, we define sets S^k, S^{k+1}, S^* as follows, each of whose cardinalities is no greater than $2s$:

$$\text{supp}(t^k) = S^k,\ \text{supp}(t^{k+1}) = S^{k+1},\ \text{supp}(t^*) = S^*.$$

Moreover, define $S^k \cup S^{k+1} \cup S^* = J$ such that $\|J\|_0 \leq 6s$.

Define $b = t^k - \eta' \nabla_J F(t^k)$. Then,

$$\|t^{k+1} - t^*\|_2 \leq \|t^{k+1} - b\|_2 + \|b - t^*\|_2 \leq 2\|b - t^*\|_2, \tag{6.21}$$

where $t^* = [t_1^* \; t_2^*] \in \mathbb{R}^{2n}$ such that $\|t^*\|_0 \leq 2s$ is the solution of the optimization problem in (4.6). The last inequality follows since t^{k+1} is generated by taking the $2s$ largest entries of $t^k - \eta' \nabla F(t^k)$; by definition of J, t^{k+1} also has the minimum Euclidean distance to b over all vectors with cardinality $2s$. Moreover:

$$\begin{aligned} \|b - t^*\|_2 &= \|t^k - \eta' \nabla_J F(t^k) - t^*\|_2 \\ &\leq \|t^k - t^* - \eta' \left(\nabla_J F(t^k) - \nabla_J F(t^*)\right) \|_2 + \eta' \|\nabla_J F(t^*)\|_2. \end{aligned} \tag{6.22}$$

Now, by invoking RSC and RSS in the Definition 6.8, we have:

$$\|t^k - t^* - \eta' \left(\nabla_J F(t^k) - \nabla_J F(t^*)\right) \|_2^2 \leq (1 + \eta'^2 M_{6s}^2 - 2\eta' m_{6s}) \|t^k - t^*\|_2^2,$$

where M_{6s} and m_{6s} denote the RSC and RSS constants. The above inequality follows by the upper bound of (6.17) and the lower bound of (6.15) in Definition 6.8 with the restriction set ξ chosen as J. Now let $q = \sqrt{1 + \eta'^2 M_{6s}^2 - 2\eta' m_{6s}}$. By (6.21) and (6.22), we have:

$$\|t^{k+1} - t^*\|_2 \leq 2q \|t^k - t^*\|_2 + 2\eta' \|\nabla_J F(t^*)\|_2. \tag{6.23}$$

In order for the algorithm to exhibit linear convergence, we need to have $2q < 1$. That is,

$$\eta'^2 M_{6s}^2 - 2\eta' m_{6s} + \frac{3}{4} < 0.$$

By solving this quadratic inequality with respect to η', we obtain that η', m_{6s}, and M_{6s} should satisfy

$$1 \leq \frac{M_{6s}}{m_{6s}} \leq \frac{2}{\sqrt{3}}, \quad \frac{0.5}{M_{6s}} < \eta' < \frac{1.5}{m_{6s}}.$$

Under these conditions, we obtain the following linear convergence by induction on k:

$$\|t^{k+1} - t^*\|_2 \leq (2q)^k \|t^0 - t^*\|_2 + \frac{2\eta'}{1 - 2q} \|\nabla_J F(t^*)\|_2, \tag{6.24}$$

where t^0 denotes the initial value for the constituent vector, t. The bound in (6.24) shows that after enough iterations the first term vanishes and the quality of estimation depends on the vanishing speed of the second term, $\frac{2\eta'}{1-2q} \|\nabla_J F(t^*)\|_2$ that is determined by the number of measurements.

To bound the gradient in second term, $\|\nabla_J F(t^k)\|_2$, we need the following lemma:

Lemma 6.9. *(Khintchine inequality [43].) Let X_i be a finite number of independent and zero mean sub-gaussian random variables with unit variance. Assume that $\|X_i\|_{\psi_2} \leq r$. Then, for any real b_i and $p \geq 2$:*

$$\left(\sum_i b_i^2 \right)^{\frac{1}{2}} \leq \left(\mathbb{E} |\sum_i b_i X_i|^p \right)^{\frac{1}{p}} \leq Cr\sqrt{p} \left(\sum_i b_i^2 \right)^{\frac{1}{2}}.$$

Recall that our measurement model is given by:

$$y_i = g(a_i^T \Gamma t) + e_i, \quad i = 1, \ldots, m.$$

As mentioned above, we assume that e_i represents the additive subgaussian noise with $\|e_i\|_{\psi_2} \leq \tau$ for $i = 1 \ldots m$.

We leverage the Khintchine inequality to bound $\mathbb{E}\|\nabla_J F(t^k)\|_2$ under the subgaussian assumption on e_i. Denoting by $(\nabla_J F(t))_k$ as the k^{th} entry of the gradient (restricted to set J), from the Khintchine inequality, and for each $k = 1, \ldots, |J|$, we have:

$$\begin{aligned}
\left(\mathbb{E}\left|(\nabla_J F(t))_k\right|^2\right)^{\frac{1}{2}} &\overset{r_1}{=} \left(\mathbb{E}\left(\frac{1}{m}\sum_{i=1}^{m}(\Gamma_J)_k^T\, a_i e_i\right)^2\right)^{\frac{1}{2}} \\
&\overset{r_2}{\leq} \frac{1}{m}\mathbb{E}\left(C\tau\sqrt{2}\left(\sum_{i=1}^{m}\left((\Gamma_J)_k^T\, a_i\right)^2\right)^{\frac{1}{2}}\right) \\
&\leq \frac{1}{m}C\tau\sqrt{2}\left(\sum_{i=1}^{m}(\Gamma_J)_k^T\,\mathbb{E}\left(a_i a_i^T\right)(\Gamma_J)_k\right)^{\frac{1}{2}} \\
&\overset{r_3}{=} \frac{C\tau\sqrt{2}}{\sqrt{m}},
\end{aligned} \tag{6.25}$$

where Γ_J denotes the restriction of the columns of the dictionary to set J with $\|J\|_0 \leq 6s$ such that $3s$ of the columns are selected from each basis of the dictionary. Here, r_1 follows from (6.19), r_2 follows from the Khintchine inequality with $p = 2$ and the fact that e_i are independent from a_i. Finally, r_3 holds since the rows of A are assumed to be isotropic random vectors. Now, we can bound $\mathbb{E}\|\nabla_J F(t^k)\|_2$ as follows:

$$\mathbb{E}\|\nabla_J F(t^k)\|_2 \leq \sqrt{\mathbb{E}\|\nabla_J F(t^k)\|_2^2} \leq C'\tau\sqrt{\frac{s}{m}}, \tag{6.26}$$

where $C' > 0$ is an absolute constant and the last inequality is followed by (6.25) and the fact that $\|J\|_0 \leq 6s$.

Proof of Theorem 4.6. By taking expectation from bound in (6.24) and using the bound stated in (6.25), we obtain the desired bound in Theorem 4.6 as follows:

$$\begin{aligned}
\|t^{k+1} - t^*\|_2 &\leq (2q)^k\,\|t^0 - t^*\|_2 + \frac{2\eta'}{1-2q}\|\nabla_J F(t^*)\|_2 \\
&\leq (2q)^k\,\mathbb{E}\|t^0 - t^*\|_2 + C\tau\sqrt{\frac{s}{m}},
\end{aligned} \tag{6.27}$$

where $C > 0$ is a constant which depends only on the step size, η' and q. In addition, in the noiseless case ($\tau = 0$), if we denote κ as the desired accuracy for solving optimization problem (4.6), then the number of iterations to achieve the accuracy κ is given by $N = \mathcal{O}(\log \frac{\|t^0 - t^*\|_2}{\kappa})$. □

In the above convergence analysis of DHT, we assumed that objective function in (4.6), $F(t)$ satisfies the RSC/RSS conditions. In this section, we validate this assumption via the proofs for Theorems 4.7 and 4.8. As discussed above, we separately analyze two cases.

6.2.1 Case (a): isotropic rows of A

We first consider the case where the rows of the measurement matrix A are sampled from an isotropic probability distribution in $\mathbb{R}^n$. Specifically, we make the following assumptions on A:

1. the rows of A are independent isotropic vectors. That is, $\mathbb{E}a_i a_i^T = I_{n\times n}$ for $i = 1 \ldots m$.
2. $\|a_i^T \Gamma_\xi\|_\infty \leq \vartheta$ for $i = 1 \ldots m$.

Remark 6.10. *Assumption 2 is unavoidable in our analysis, and indeed this is one of the cases where our derivation differs from existing proofs. The condition $||a_i^T\Gamma_\xi||_\infty \leq \vartheta$ requires that all entries in $A\Gamma_\xi$ are bounded by some number ϑ. In other words, ϑ captures the cross-coherence between the measurement matrix, A and the dictionary $\Gamma_\xi = [\Phi\ \Psi]_\xi$ and controls the interaction between these two matrices. Without this assumption, one can construct a counter-example with the Hessian of the objective to be zero with high probability (for instance, consider partial DFT matrix as the measurement matrix A and $\Gamma_\xi = [I\ \Psi]_\xi$ with Ψ being the inverse DFT basis).*

Modifying (6.20), we define the *restricted* Hessian matrix as a $4s \times 4s$ sub-matrix of the Hessian matrix:

$$\nabla^2_\xi F(t) = \frac{1}{m}\sum_{i=1}^m \Gamma_\xi^T a_i a_i^T \Gamma_\xi g'(a_i^T\Gamma t), \quad \|\xi\|_0 \leq 4s. \tag{6.28}$$

Here, Γ_ξ is the restriction of the columns of the dictionary $\Gamma = [\Phi\ \Psi]$ with respect to set ξ, such that $2s$ columns are selected from each basis. Let $S_i = \Gamma_\xi^T a_i a_i^T \Gamma_\xi g'(a_i^T\Gamma t), i = 1 \dots m$. As per our assumption in Section 3, the derivative of the link function, $g(x)$ satisfies $0 < l_1 \leq g'(x) \leq l_2$. By this assumption, it is guaranteed that $\lambda_{\min}(S_i) \geq 0, i = 1 \dots m$; this follows since $\Gamma_\xi^T a_i a_i^T \Gamma_\xi$ is a positive semidefinite matrix and $g' > 0$, we have $\lambda_{\min}(S_i) = \lambda_{\min}(\Gamma_\xi^T a_i a_i^T \Gamma_\xi) g' \geq 0$.

Let $\Lambda_{\max} = \max_\xi \lambda_{\max}(\nabla^2_\xi F(t))$ and $\Lambda_{\min} = \min_\xi \lambda_{\min}(\nabla^2_\xi F(t))$ where $\lambda_{\min}$ and $\lambda_{\max}$ denote the minimum and maximum eigenvalues of the restricted Hessian matrix. Furthermore, let U be any index set with $\|U\|_0 \leq 6s$ such that $\xi \subseteq U$. We have:

$$l_1 \min_U \lambda_{\min}\left(\frac{1}{m}\sum_{i=1}^m \Gamma_U^T a_i a_i^T \Gamma_U\right) \leq \Lambda_{\min} \leq \Lambda_{\max} \leq l_2 \max_U \lambda_{\max}\left(\frac{1}{m}\sum_{i=1}^m \Gamma_U^T a_i a_i^T \Gamma_U\right).$$

Here, Γ_U is the restriction of the columns of Γ with respect to a set U such that $3s$ columns is selected from each basis. By taking expectations, we obtain:

$$l_1 \mathbb{E}\min_U \lambda_{\min}\left(\frac{1}{m}\sum_{i=1}^m \Gamma_U^T a_i a_i^T \Gamma_U\right) \leq \mathbb{E}\Lambda_{\min} \leq \mathbb{E}\Lambda_{\max} \leq l_2 \mathbb{E}\max_U \lambda_{\max}\left(\frac{1}{m}\sum_{i=1}^m \Gamma_U^T a_i a_i^T \Gamma_U\right). \tag{6.29}$$

Inequality in (6.29) shows that for proving RSC and RSS, we need to bound the expectation of the maximum and minimum eigenvalues of $\frac{1}{m}\sum_{i=1}^m \Gamma_\xi^T a_i a_i^T \Gamma_U$ over sets U with $\|U\|_0 \leq 6s$. We should mention that (6.29) establishes RSC/RSS constants in expectation. One can establish RSC/RSS in tail probability using results in [54, 57].

As our main tool for bounding the RSC/RSS constants, we use the *uniform Rudelson's inequality* [43,57].

Lemma 6.11. *(Uniform Rudelson's inequality) Let x_i be vectors in $\mathbb{R}^n$ for $i = 1, \dots, m$ and $m \leq n$. Also assume that the entries of x_i's are bounded by ϑ, that is, $\|x_i\|_\infty \leq \vartheta$. Let h_i denote independent Bernoulli random variables (with parameter 1/2) for $i = 1 \dots m$. Then for every set $\Omega \subseteq [n]$, we have:*

$$\mathbb{E}\max_{|\Omega|\leq n}\left\|\sum_{i=1}^m h_i (x_i)_\Omega (x_i)_\Omega^T\right\| \leq C_\vartheta \sqrt{|\Omega|}\max_{|\Omega|\leq n}\left\|\sum_{i=1}^m (x_i)_\Omega (x_i)_\Omega^T\right\|^{\frac{1}{2}}, \tag{6.30}$$

where $(x_i)_\Omega$ denotes the restriction of x_i to Ω, $l = \log(|\Omega|)\sqrt{\log m}\sqrt{\log n}$, and C_ϑ denotes the dependency of C only on ϑ.

Before using the above result, we need to restate the uniform version of the standard symmetrization technique (Lemma 5.70 in [43]):

Lemma 6.12. *(Uniform symmetrization) Let x_{ik}, $i = 1 \dots m$ be independent random vectors in some Banach space where indexed by some set Ξ such that $k \in \Xi$. Also, assume that h_i, $i = 1 \dots m$ denote independent Bernoulli random variables (with parameter 1/2) for $i = 1 \dots m$. Then,*

$$\mathbb{E} \sup_{k \in \Xi} \Big\| \sum_{i}^{m} (x_{ik} - \mathbb{E}x_{ik}) \Big\| \leq 2\mathbb{E} \sup_{k \in \Xi} \Big\| \sum_{i}^{m} h_i x_{ik} \Big\|. \tag{6.31}$$

Now we apply the Uniform Rudelson's inequality on $\lambda_{\max}\left(\frac{1}{m}\sum_{i=1}^{m} \Gamma_U^T a_i a_i^T \Gamma_U\right)$ over all set U with $\|U\|_0 \leq 6s$. We have:

$$\begin{aligned} R \triangleq \mathbb{E} \max_U \Big\| \frac{1}{m} \sum_{i=1}^{m} \Gamma_U^T a_i a_i^T \Gamma_U - \Gamma_U^T \Gamma_U \Big\| &\overset{r_1}{\leq} 2\mathbb{E} \max_U \Big\| \frac{1}{m} \sum_{i=1}^{m} h_i \Gamma_U^T a_i a_i^T \Gamma_U \Big\| \\ &\overset{r_2}{\leq} \frac{C_\vartheta \sqrt{6s}}{\sqrt{m}} \mathbb{E} \max_U \Big\| \frac{1}{m} \sum_{i=1}^{m} \Gamma_U^T a_i a_i^T \Gamma_U \Big\|^{\frac{1}{2}}, \end{aligned} \tag{6.32}$$

where r_1 follows from Lemma 6.12 with h_i defined in this lemma and r_2 follows from (6.30). In addition $l = \log(6s)\sqrt{\log m}\sqrt{\log 2n}$. Then by application of a triangle inequality, we have:

$$\mathbb{E} \max_U \Big\| \frac{1}{m} \sum_{i=1}^{m} \Gamma_U^T a_i a_i^T \Gamma_U \Big\| \leq R + \max_U \big\| \Gamma_U^T \Gamma_U \big\|.$$

On the other hand by Cauchy-Schwarz inequality, we get:

$$\mathbb{E} \max_U \Big\| \frac{1}{m} \sum_{i=1}^{m} \Gamma_U^T a_i a_i^T \Gamma_U \Big\|^{\frac{1}{2}} \leq \left(\mathbb{E} \max_U \Big\| \frac{1}{m} \sum_{i=1}^{m} \Gamma_U^T a_i a_i^T \Gamma_U \Big\| \right)^{\frac{1}{2}}$$

By combining the above inequalities, we obtain:

$$R \leq \frac{C'_\vartheta l \sqrt{s}}{\sqrt{m}} \left(R + \max_U \big\| \Gamma_U^T \Gamma_U \big\| \right)^{\frac{1}{2}}, \tag{6.33}$$

where C'_ϑ depends only on ϑ. This inequality is a quadratic inequality in terms of R and is easy to solve. By noting $\beta = \max_U \big\| \Gamma_U^T \Gamma_U \big\|$, we can write (6.33) as $\frac{R}{\beta} \leq \frac{C'_\vartheta l \sqrt{s}}{\sqrt{m}} \frac{1}{\beta} \left(1 + \frac{R}{\beta}\right)^{\frac{1}{2}}$. Now we can consider two cases; either $\frac{R}{\beta} \leq 1$, or $\frac{R}{\beta} > 1$. As a result, we have:

$$R \leq \max \left(\delta_0 \left(\max_U \big\| \Gamma_U^T \Gamma_U \big\| \right)^{\frac{1}{2}}, \delta_0^2 \right), \tag{6.34}$$

where $\delta_0 = \frac{C'_\vartheta l \sqrt{s}}{\sqrt{m}}$. In addition, we can use the Gershgorin Circle Theorem [58] to bound $\lambda_{\max}(\Gamma_U^T \Gamma_U) = \|\Gamma_U^T \Gamma_U\|$ and $\lambda_{\min}(\Gamma_U^T \Gamma_U)$. This follows since:

$$\Gamma_U^T \Gamma_U = \begin{bmatrix} I & \Phi^T \Psi \\ \Psi^T \Phi & I \end{bmatrix}_{6s \times 6s},$$

and hence we have:

$$\Big| \lambda_i(\Gamma_U^T \Gamma_U) - 1 \Big| \leq (6s - 1)\gamma, \quad i = 1 \dots 6s,$$

where γ denotes the mutual coherence of Γ. Hence, the following holds for all index set U:

$$1 - (6s-1)\gamma \leq \lambda_{\min}(\Gamma_U^T \Gamma_U) \leq \lambda_{\max}(\Gamma_U^T \Gamma_U) \leq 1 + (6s-1)\gamma, \tag{6.35}$$

provided that $\gamma \leq \frac{1}{6s-1}$ to have nontrivial lower bound.

Proof of Theorem 4.7. If we choose $m \geq \left(\frac{C''_\vartheta}{\delta^2} s \log(n) \log^2 s \log\left(\frac{1}{\delta^2} s \log(n) \log^2 s\right)(1 + (6s-1)\gamma)\right)$ in (6.34), then we have $R \leq \delta$ for some $\delta \in (0,1)$ and $C''_\vartheta > 0$ which depends only on ϑ. If $s = o(1/\gamma)$, then we obtain the stated sample complexity in Theorem 4.7. □

6.2.2 Case (b): isotropic subgaussian rows of A

Now, suppose that the measurement matrix A has independent isotropic subgaussian rows. We show that under this assumption, one can obtain better sample complexity bounds compared to the previous case. We use the following argument (which is more or less standard; see [59–61]). Let $\Gamma = [\Phi\ \Psi]$, and let $B_U = A\Gamma_U$ for any fixed $|U| \leq 6s$, where $3s$ elements are chosen from each basis. According to the notation from Section 6.2, we have:

$$l_1 \min_U \lambda_{\min}\left(\frac{1}{m} B_U^T B_U\right) \leq \Lambda_{\min} \leq \Lambda_{\max} \leq l_2 \max_U \lambda_{\max}\left(\frac{1}{m} B_U^T B_U\right). \tag{6.36}$$

where l_1, l_2 are upper and lower bounds on the derivative of the link function. Therefore, all we need to do is to bound the maximum and minimum singular values of $\frac{1}{\sqrt{m}} B_U$. To do so, we use the fact that if the rows of A are m independent copies of an isotropic vector with bounded ψ_2 norm, then the following holds for any fixed vector $v \in \mathbb{R}^{2n}$:

$$\left|\frac{1}{m}\left\|Bx\right\|_2^2 - \left\|\Gamma x\right\|_2^2\right| \leq \frac{\varepsilon'}{2},$$

for some constant ε' with probability at least $1 - \exp(-Cm\varepsilon^2/2)$ for some absolute constant C [60]. Now fix any set U as above. Then, one can show using a covering number argument (for example, Lemma 2.1 in [59]) that with probability greater than $1 - 2(1 + \frac{2}{\varepsilon})^{6s} \exp(-c_1 m\varepsilon^2)$, we get for any $v \in U$:

$$(1-\varepsilon)\|\Gamma_U v\|_2^2 \leq \|A\Gamma_U v\|_2^2 \leq (1+\varepsilon)\|\Gamma_U v\|_2^2.$$

Taking a union bound over all possible subsets U with $|U| \leq 6s$, we get:

$$\max_U \left\|\frac{1}{m} B_U^T B_U - \Gamma_U^T \Gamma_U\right\| \leq \delta, \tag{6.37}$$

with probability at least $1 - 2\binom{n}{6s}(1 + 1/\delta)^{6s} \exp\left(-c_2 u^2 m\right)$. Therefore, for sufficiently large m (that we specify below), the following holds with high probability:

$$\lambda_{\min}\left(\Gamma_U^T \Gamma_U\right) - \delta \leq \lambda_{\min}\left(\frac{1}{m} B_U^T B_U\right) \leq \lambda_{\max}\left(\frac{1}{m} B_U^T B_U\right) \leq \lambda_{\max}\left(\Gamma_U^T \Gamma_U\right) + \delta$$

We use (6.35) to bound $\lambda_{\max}(\Gamma_U^T \Gamma_U) = \|\Gamma_U^T \Gamma_U\|$ and $\lambda_{\min}(\Gamma_U^T \Gamma_U)$; as a result,

$$1 - (6s-1)\gamma - \delta \leq \lambda_{\min}\left(\frac{1}{m} B_U^T B_U\right) \leq \lambda_{\max}\left(\frac{1}{m} B_U^T B_U\right) \leq 1 + (6s-1)\gamma + \delta \tag{6.38}$$

Thus, we obtain the desired bound in (6.36). That is:

$$l_1\left(1 - (6s-1)\gamma - \delta\right) \leq \Lambda_{\min} \leq \Lambda_{\max} \leq l_2\left(1 + (6s-1)\gamma - \delta\right). \tag{6.39}$$

holds with high probability for some $0 < \delta < 1 - (6s-1)\gamma$.

Proof of Theorem 4.8. The probability of failure of the above statement can be vanishingly small if we set $m \geq \frac{C'}{\delta^2} s \log \frac{n}{s}$ for some $\delta \in (0,1)$ and absolute constant $C' > 0$. Note that we only obtain nontrivial upper and lower bounds on $\Lambda_{\min}, \Lambda_{\max}$ if $\gamma \leq \frac{1}{6s-1}$. Assuming constant δ and coherence γ inversely proportional to s, we obtain the required sample complexity of DHT as: $m = \mathcal{O}\left(s \log \frac{n}{s}\right)$. □

For both cases (a) and (b), RSC and RSS constants follow by setting $M_{6s} \leq l_2\left(1+(6s-1)\gamma-\delta\right)$ and $m_{6s} \geq l_1\left(1-(6s-1)\gamma-\delta\right)$. As we discussed in the begging of section 6.2, we require that $\frac{0.5}{M_{6s}} < \eta' < \frac{1.5}{m_{6s}}$ in order to establish linear convergence of *DHT*. Hence, for linear convergence, the step size must satisfy:

$$\frac{0.5}{l_2\left(1+(6s-1)\gamma-\delta\right)} < \eta' < \frac{1.5}{l_1\left(1-(6s-1)\gamma-\delta\right)}$$

for some $0 < \delta < 1-(6s-1)\gamma$.

7 Conclusion

In this paper, we consider the problem of demixing sparse signals from their nonlinear measurements. We specifically study the more challenging scenario where only a limited number of nonlinear measurements of the superposition signal are available. As our primary contribution, we propose two fast algorithms for recovery of the constituent signals, and support these algorithms with the rigorous theoretical analysis to derive nearly-tight upper bounds on their sample complexity for achieving stable demixing.

We anticipate that the problem of demixing signals from nonlinear observations can be used in several different practical applications. As future work, we intend to extend our methods to more general signal models (including rank-sparsity models for matrix valued data), as well as robust recovery under more general nonlinear observation models.

References

[1] M. Soltani and C. Hegde. Demixing sparse signals from nonlinear observations. Technical report, Iowa State University, 2016.

[2] M. Soltani and C. Hegde. Demixing sparse signals from nonlinear observations. In *Proc. Asilomar Conf. Sig. Sys. Comp.*, Nov. 2016.

[3] M. Elad, J. Starck, P. Querre, and D. Donoho. Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA). *Appl. Comput. Harmonic Analysis*, 19(3):340–358, 2005.

[4] D. Donoho, M. Elad, and V. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. IEEE Trans. Inform. Theory, 52(1):6–18, 2006.

[5] E. Candès. Compressive sampling. In *Proc. Int. Congress of Math.*, Madrid, Spain, Aug. 2006.

[6] D. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289–1306, 2006.

[7] S. Foucart and H. Rauhut. *A mathematical introduction to compressive sensing*, volume 1. Springer.

[8] M. Wakin, J. Laska, M. Duarte, D. Baron, S. Sarvotham, D. Takhar, K. Kelly, and R. Baraniuk. An architecture for compressive imaging. In *Proc. IEEE Int. Conf. Image Processing (ICIP)*, Atlanta, GA, Oct. 2006.

[9] M. Mishali and Y. Eldar. From theory to practice: Sub-Nyquist sampling of sparse wideband analog signals. *IEEE J. Select. Top. Signal Processing*, 4(2):375–391, 2010.

[10] C. Hegde and R. Baraniuk. SPIN : Iterative signal recovery on incoherent manifolds. In *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, July 2012.

[11] C. Hegde and R. Baraniuk. Signal recovery on incoherent manifolds. *IEEE Trans. Inform. Theory*, 58(12):7204–7214, Dec. 2012.

[12] Y. Plan, R. Vershynin, and E. Yudovina. High-dimensional estimation with geometric constraints. *arXiv preprint arXiv:1404.3749*, 2014.

[13] C. Thrampoulidis, E. Abbasi, and B. Hassibi. LASSO with non-linear measurements is equivalent to one with linear measurements. In *Proc. Adv. Neural Inf. Proc. Sys (NIPS)*, 2015.

[14] Y. Plan and R. Vershynin. The generalized LASSO with nonlinear observations. *IEEE Trans. Inform. Theory*, 62(3):1528–1537, 2016.

[15] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal Statist. Soc B*, 58(1):267–288, 1996.

[16] S. Bahmani, B. Raj, and P. Boufounos. Greedy sparsity-constrained optimization. *J. Machine Learning Research*, 14(1):807–841, 2013.

[17] Xiaotong Yuan, Ping Li, and Tong Zhang. Gradient hard thresholding pursuit for sparsity-constrained optimization. In *Proc. Int. Conf. Machine Learning*, pages 127–135, 2014.

[18] P. Jain, A. Tewari, and P. Kar. On iterative hard thresholding methods for high-dimensional m-estimation. In *Adv. Neural Inf. Proc. Sys. (NIPS)*, pages 685–693, 2014.

[19] S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comp.*, 20(1):33–61, 1998.

[20] J. Bobin, J. Starck, J. Fadili, Y. Moudden, and D. Donoho. Morphological component analysis: An adaptive thresholding strategy. *IEEE Trans. Image Proc.*, 16(11):2675–2681, 2007.

[21] C. Studer, P. Kuppinger, G. Pope, and H. Bölcskei. Recovery of sparsely corrupted signals. *IEEE Trans. Inform. Theory*, 58(5):3115–3130, 2012.

[22] M. McCoy and J. Tropp. Sharp recovery bounds for convex demixing, with applications. *Foundations of Comp. Math.*, 14(3):503–567, 2014.

[23] M. McCoy, V. Cevher, Q. Dinh, A. Asaei, and L. Baldassarre. Convexity in source separation: Models, geometry, and algorithms. *IEEE Sig. Proc. Mag.*, 31(3):87–95, 2014.

[24] E. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):11, 2011.

[25] V. Chandrasekaran, S. Sanghavi, P. Parrilo, and A. S. Willsky. Sparse and low-rank matrix decompositions. In *Proc. Allerton Conf. on Comm., Contr., and Comp.*, pages 962–967, 2009.

[26] V. Chandrasekaran, S. Sanghavi, P. Parrilo, and A. S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM J. Opt.*, 21(2):572–596, 2011.

[27] V. Chandrasekaran, P. Parrilo, and A. S. Willsky. Latent variable graphical model selection via convex optimization. In *Proc. Allerton Conf. on Comm., Contr., and Comp.*, pages 1610–1613, 2010.

[28] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma. Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images. *IEEE Trans. Pattern Anal. Machine Intell.*, 34(11):2233–2246, 2012.

[29] P. Boufounos and R. Baraniuk. 1-bit compressive sensing. In *Int. Conf. Info. Sciences and Systems (CISS)*, pages 16–21. IEEE, 2008.

[30] Y. Plan and R. Vershynin. One-bit compressed sensing by linear programming. *Comm. Pure and Applied Math.*, 66(8):1275–1297, 2013.

[31] E. Candes, X. Li, and M. Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Trans. Inform. Theory*, 61(4):1985–2007, 2015.

[32] M. Davenport, Y. Plan, E. van den Berg, and M. Wootters. 1-bit matrix completion. *Information and Inference*, 3(3):189–223, 2014.

[33] R. Ganti, L. Balzano, and R. Willett. Matrix completion under monotonic single index models. In *Adv. Neural Inf. Proc. Sys. (NIPS)*, pages 1864–1872, 2015.

[34] A. Kalai and R. Sastry. The isotron algorithm: High-dimensional isotonic regression. In *COLT*, 2009.

[35] S. Kakade, V. Kanade, O. Shamir, and A. Kalai. Efficient learning of generalized linear and single index models with isotonic regression. In *Adv. Neural Inf. Proc. Sys. (NIPS)*, pages 927–935, 2011.

[36] R. Ganti, N. Rao, R. Willett, and R. Nowak. Learning single index models in high dimensions. *arXiv preprint arXiv:1506.08910*, 2015.

[37] X. Yi, Z. Wang, C. Caramanis, and H. Liu. Optimal linear estimation under unknown nonlinear transform. In *Adv. Neural Inf. Proc. Sys. (NIPS)*, pages 1549–1557, 2015.

[38] A. Beck and Y. Eldar. Sparsity constrained nonlinear optimization: Optimality conditions and algorithms. *SIAM Journal on Optimization*, 23(3):1480–1509, 2013.

[39] Z. Yang, Z. Wang, H. Liu, Y. Eldar, and T. Zhang. Sparse nonlinear regression: Parameter estimation and asymptotic inference. *J. Machine Learning Research*, 2015.

[40] T. Blumensath and M. Davies. Iterative hard thresholding for compressed sensing. *Appl. Comput. Harmon. Anal*, 27(3):265–274, 2009.

[41] D. Needell and J. Tropp. Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *Appl. Comput. Harmon. Anal*, 26(3):301–321, 2009.

[42] Sahand Negahban, Bin Yu, Martin J Wainwright, and Pradeep K Ravikumar. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *Adv. Neural Inf. Proc. Sys. (NIPS)*.

[43] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[44] E. Candes and J. Romberg. Sparsity and incoherence in compressive sampling. *Inverse problems*, 23(3):969, 2007.

[45] Y. Plan and R. Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Trans. Inform. Theory*, 59(1):482–494, 2013.

[46] J. Tropp. On the conditioning of random subdictionaries. *Appl. Comput. Harmon. Anal.*, 25(1):1–24, 2008.

[47] M. Cheraghchi, V. Guruswami, and A. Velingker. Restricted isometry of Fourier matrices and list decodability of random linear codes. *SIAM J. Comp.*, 42(5):1888–1914, 2013.

[48] M. Yang, N. Ahuja, and D. Kriegman. Face recognition using kernel eigenfaces. In *Proc. IEEE Int. Conf. Image Processing (ICIP)*, Vancouver, BC, Sept. 2000.

[49] E. van den Berg and M. Friedlander. Probing the pareto frontier for basis pursuit solutions. *SIAM J. Sci. Comp.*, 31(2):890–912, 2008.

[50] E. van den Berg and M. Friedlander. SPGL1: A solver for large-scale sparse reconstruction, June 2007. http://www.cs.ubc.ca/labs/scl/spgl1.

[51] R. Coifman, F. Geshwind, and Y. Meyer. Noiselets. *Appl. Comput. Harmonic Analysis*, 10(1):27–44, 2001.

[52] E. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, 52(2):489–509, 2006.

[53] F. Krahmer and R. Ward. New and improved johnson-lindenstrauss embeddings via the restricted isometry property. *SIAM Journal on Mathematical Analysis*, 43(3):1269–1281, 2011.

[54] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

[55] G. Raskutti, M. J Wainwright, and B. Yu. Restricted eigenvalue properties for correlated gaussian designs. *J. Machine Learning Research*, 11(Aug):2241–2259, 2010.

[56] R. McLeod. Mean value theorems for vector valued functions. *Proceedings of the Edinburgh Mathematical Society (Series 2)*, 14(03):197–209, 1965.

[57] M. Rudelson and R. Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Communications on Pure and Applied Mathematics*, 61(8):1025–1045, 2008.

[58] R. A Horn and C. Johnson. *Matrix analysis*. Cambridge university press, 2012.

[59] H. Rauhut, K. Schnass, and P. Vandergheynst. Compressed sensing and redundant dictionaries. *IEEE Trans. Inform. Theory*, 54(5):2210–2219, 2008.

[60] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. Uniform uncertainty principle for Bernoulli and subgaussian ensembles. *Constructive Approximation*, 28(3):277–289, 2008.

[61] E. Candes, Y. Eldar, D. Neeell, and R. Paige. Compressed sensing with coherent and redundant dictionaries. *Appl. Comput. Harmonic Analysis*, 31(1):59–73, 2011.