A FAST ITERATIVE ALGORITHM FOR DEMIXING SPARSE SIGNALS FROM NONLINEAR OBSERVATIONS

Mohammadreza Soltani and Chinmay Hegde

ECpE Department, Iowa State University, Ames, IA, 50010

ABSTRACT

In this paper, we propose an iterative algorithm based on hard thresholding for demixing a pair of signals from nonlinear observations of their superposition. We focus on the under-determined case where the number of available observations is far less than the ambient dimension of the signals. We derive nearly-tight upper bounds on the sample complexity of the algorithm to achieve stable recovery of the component signals. Moreover, we show that the algorithm enjoys a linear convergence rate. We provide a range of simulations to illustrate the performance of the algorithm both on synthetic and real data.

Index Terms— Demixing, sparse recovery, nonlinear measurements, linear convergence, incoherence.

1. INTRODUCTION

The problem of *demixing* a pair of signals from their superposition signal impacts several applications in signal and image processing, statistics, and data analysis [1, 2]. In the simplest setting, consider a length-n signal x that can be expressed as $x = \Phi w + \Psi z$, where Φ and Ψ are *incoherent* bases in $\mathbb{R}^n$, and $w, z \in \mathbb{R}^n$ are the basis coefficients. The goal of demixing is to reliably recover the constituent signals, w and z, given the superposition signal x. It is clear that even in this simple case, the demixing problem is highly ill-posed since the number of unknowns $(2n)$ is greater than the number of equations (n). To enable reliable recovery of the constituent signals, one has to assume some notion of incoherence between the bases of constituent signals, Φ and Ψ [3, 4, 5].

Now, consider the more challenging case where instead of the superposition signal x, we only have access to *linear* measurements $y = Ax$, where $A \in \mathbb{R}^{m \times n}$ denotes the measurement operator and where $m \ll n$. In this scenario, the demixing problem is further confounded by the fact that A possesses a nontrivial null space. Therefore, some additional *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 [6, 7, 8].

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

$$y_i = g(a_i^T(\Phi w + \Psi z)) + e_i,\ i = 1 \dots m. \tag{1.1}$$

Here, the superposition signal is given by $x = \Phi w + \Psi z$, and each observation is generated by the composition of a linear functional of the signal $\langle a_i, x \rangle$ with a nonlinear function g. Further, we assume that the observation y_i is corrupted by subgaussian additive noise. Here, g represents a nonlinear, smooth, and strictly monotonic function (sometimes called a *link*, or *transfer* function), and a_i denotes the i^{th} row of a linear measurement matrix A. Particularly, we are interested in recovering constituent signals which are s-*sparse* (i.e., they do not have more than s nonzero entries) from m nonlinear observations, when m is much smaller than the ambient dimension n.

In this paper, we introduce a novel approach for the demixing problem under the observation model (1.1). Our approach is based on two key ideas. First, we formulate our nonlinear demixing problem in terms of an *optimization problem* with respect to a special loss function that depends on the nonlinearity g. Second, for solving the proposed optimization problem, we provide an iterative algorithm based on *hard thresholding* for demixing of the constituent signals in (1.1) given the nonlinear observations y. The algorithm is assumed to possess oracle knowledge of the measurement matrix A, bases Φ and Ψ, and the link function g. In contrast with previous methods for nonlinear demixing [4, 9], we show that leveraging prior knowledge of the link function g can significantly improve the recovery performance. In our setup, we make the key assumption that the measurement vectors a_i are independent, isotropic random vectors that are *incoherent* with the bases Φ and Ψ. (See (3.6) for a precise definition of incoherence.) This assumption is more general than the i.i.d. Gaussian assumption on the measurement vectors made in [4, 9], and is applicable to a wider range of measurement models.

We support our algorithm with a rigorous analysis. Our analysis reveals upper bounds on the sample complexity of demixing with nonlinear observations (here, sample complexity denotes the required number of observations for reliably recovering the signal coefficients w and z). More precisely, we show that the sample complexity is upper-bounded by $m = \mathcal{O}(s \log^2 n \log^2 s \log(s \log n))$, provided that the bases are sufficiently incoherent with respect to each other. This matches the sample complexity of recovering s-sparse signals from linear observations up to polylogarithmic factors. Moreover, we show that the algorithm enjoys a *linear* convergence rate to the desired solution.

Furthermore, we provide a range of simulations to show the superior performance of our algorithm compared to previous algorithms [4, 9] both on synthetic and real data. Due to page-limit constraints, we merely state our theoretical claims and a few select experiments, and refer the reader to the supplementary of this paper [10] for full proofs.

2. RELATED WORK

Demixing problems of various flavors have been long studied in research areas spanning signal processing, statistics, and physics, and have been the focus of significant research in recent years. For example, Morphological Component Analysis (MCA) [11] and separation

This work is supported in part by the National Science Foundation under the grants CCF-1566281 and IIP-1632116.

of foreground and background in video [12] are two well-known examples of demixing applications in image processing.

Demixing from linear observations can be considered as a special class of *linear inverse problems* that have risen to the fore in the last decade, notably for compressive sensing applications [6, 7]. More recently, ideas from compressive sensing have been extended to inverse problems where the available observations are manifestly *nonlinear*. Instances include 1-bit compressive sensing [13, 14], phase retrieval [15], and nonlinear matrix completion [16, 17]. Similar problems have been studied in the statistical learning theory literature [18, 19, 20].

A natural fusion of the above streams of research is to consider the problem of signal demixing from nonlinear measurements. The paper [4] explicitly addresses this problem and introduces a fast, non-iterative algorithm (called ONESHOT) for recovering a pair of incoherent signals from a few nonlinear measurements. Their approach is an extension of a geometric argument proposed in [9]. Although this algorithm is very fast, the sparse components are recovered only up to an arbitrary unknown scale factor for general random measurement matrix. This can lead to high estimation errors in practice, and this can be unsatisfactory in applications. Moreover, the sample complexity of the algorithm is inversely dependent on the estimation error.

In this paper, we resolve these issues, and provide an algorithm with sample complexity that is *independent* of the estimation error and that is nearly-optimal. Our method is inspired by a recent line of efficient, iterative methods for signal estimation in high dimensions [21, 22, 23, 24, 25]. To the best of our knowledge, none of the above approaches explicitly consider the problem of *demixing* from nonlinear observations. Our algorithm (and theoretical analysis) leverages the algebraic structure of the demixing problem, and highlights the effect of incoherence both between the component bases, as well as between the measurements and the signal bases.

3. PRELIMINARIES

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 $A \in \mathbb{R}^{m\times n}$. Define the constituent vector, $t = [w^T\ z^T]^T \in \mathbb{R}^{2n}$. as the vector obtaining by stacking the coefficient vectors, w, z, of the component signals. Further, suppose that w and z are s-sparse vectors.

We use the following ideas from random matrix theory:

Definition 3.1. *(Subgaussian random variable.) A random variable X is called subgaussian if it satisfies the following:*

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

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

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

Definition 3.2. *(Isotropic random vectors.) A random vector $v \in \mathbb{R}^n$ is said to be isotropic if $\mathbb{E}vv^T = I_{n\times n}$.*

Also, we have the following definition from [26]:

Definition 3.3. *(RSC/RSS) A function f satisfies Restricted Strong Convexity/Smoothness (RSC/RSS) if:*

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

where $\xi = supp(t_1) \cup supp(t_2)$, for all $\|t_i\|_0 \le 2s$, where m_{4s} and M_{4s} are (respectively) the RSC and RSS constants. Also $\nabla^2_\xi f(t)$ denotes a $4s \times 4s$ sub-matrix of the Hessian matrix $\nabla^2 f(t)$ comprised of row/column indices indexed by ξ.

The underlying assumption in demixing problems of the form (1.1) is that the constituent bases are sufficiently *incoherent* as per the following definition:

Definition 3.4. *(ε-incoherence.) The orthonormal bases Φ and Ψ are said to be ε-incoherent if:*

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

The parameter ε is related to the more well-known *mutual coherence* of a matrix. Indeed, if we consider the dictionary $\Gamma = [\Phi\ \Psi]$, then the mutual coherence of Γ is given by $\gamma = \max_{i\neq j} |(\Gamma^T\Gamma)_{ij}|$, and one can show that $\varepsilon \le s\gamma$ [8].

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

$$y_i = g(a_i^T x) + e_i,\ i = 1 \dots m, \tag{3.2}$$

where $x \in \mathbb{R}^n$ is the superposition signal, given by $x = \Phi w + \Psi z$. Here, matrices $\Phi, \Psi \in \mathbb{R}^{n\times n}$ denote orthonormal bases, and $w, z \in \mathbb{R}^n$ denote the component s-sparse signals, and $g : \mathbb{R} \mapsto \mathbb{R}$ represents a (known) nonlinear, smooth, strictly monotonic function that we call a link function. We denote $\Theta(x) = \int_{-\infty}^{x} g(u)du$ as the integral of the link function g.

In this model, we assume that the observation y_i is corrupted by a subgaussian additive noise with $\|e_i\|_{\psi_2} \le \tau$ for $i = 1, \dots, m$. We also assume that the additive noise has zero mean when conditioned on a_i, i.e., $\mathbb{E}(e_i|a_i) = 0$ for $i = 1, \dots, m$. Also, we make the following (crucial) assumption on the link function:

Assumption 3.5. *There exist $l_1, l_2 > 0$ (resp. $l_1, l_2 < 0$) such that $0 < l_1 \le g'(x) \le l_2$ (resp. $l_1 \le g'(x) \le l_2 < 0$).*

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 $0 < l_1 \le g'(x) \le l_2$. The analysis of the complementary case is similar.

In our model, we assume that the vectors a_i (i.e., the rows of the measurement matrix A) are independent isotropic random vectors. In addition to incoherence between the component bases, we also need a measure of incoherence between the measurement matrix A and the dictionary Γ. The following notion of incoherence was introduced in the early literature of compressive sensing [27]:

Definition 3.6. *(Cross-coherence.) The cross-coherence between the measurement matrix A and the dictionary $\Gamma = [\Phi\ \Psi]$ is defined as follows:*

$$\vartheta = \max_{i,j} \frac{a_i^T \Gamma_j}{\|a_i\|_2}, \tag{3.3}$$

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 $\left\|a_i^T\Gamma_\xi\right\|_\infty \le \vartheta$ for $i = 1, \dots, m$ where Γ_ξ denotes the restriction of the columns of the dictionary to set $\xi \subset [2n]$, with $|\xi| \le 4s$ such that $2s$ columns are selected from each basis Φ and Ψ.

Algorithm 1 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}
$[w^k; z^k] \leftarrow \mathcal{P}_{2s}\left(\tilde{t}^k\right)$ {Projection}
$x^k \leftarrow \Phi w^k + \Psi z^k$ {Estimating $\widehat{x}$}
$k \leftarrow k + 1$
end while
Return: $(\widehat{w}, \widehat{z}) \leftarrow \left(w^N, z^N\right)$

4. ALGORITHM AND MAIN THEOREM

We now describe our algorithm and main theoretical results. First, we formulate our demixing problem as the minimization of a special loss function $F(t) : \mathbb{R}^{2n} \to \mathbb{R}$,

$$\begin{aligned} \min_{t\in\mathbb{R}^{2n}} \quad & F(t) = \frac{1}{m}\sum_{i=1}^{m} \Theta(a_i^T\Gamma t) - y_i a_i^T\Gamma t \\ \text{s. t.} \quad & \|t\|_0 \leq 2s. \end{aligned} \tag{4.1}$$

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). In fact, the objective function in (4.1) can be considered as the *sample* version of the problem:

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

where a, y and Γ satisfies the model (3.2). It is not hard to show that the solution of this problem satisfies $\mathbb{E}(y_i|a_i) = g(a_i^T\Gamma t)$ which coincides with the assumption on the subgaussian noise in section 3 [20].

The gradient of the loss function is given by:

$$\begin{aligned} \nabla F(t) &= \frac{1}{m}\sum_{i=1}^{m} \Gamma^T a_i g(a_i^T\Gamma t) - y_i\Gamma^T a_i, \\ &= \frac{1}{m}\Gamma^T A^T(g(A\Gamma t) - y). \end{aligned} \tag{4.2}$$

We now propose an *iterative* algorithm for solving (4.1) that we call it DEMIXING WITH HARD THRESHOLDING (DHT). The method is detailed in Algorithm 1. 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 (4.2). Next, it updates the current estimate according to the gradient update being determined in Algorithm 1. 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.

Implicitly, we have assumed that both component vectors w and z are s-sparse; however, in passing we note that the algorithm and results easily extend to different levels of sparsity in the two components. In Algorithm 1, $\mathcal{P}_{2s}$ denotes the projection of vector $\tilde{t}^k \in \mathbb{R}^{2n}$ on the set of $2s$ sparse vectors which is implemented through simple hard thresholding by retaining the $2s$ largest entries of t and setting the others to zero.

Now, we provide our main theorem 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).

Theorem 4.1. *Consider the measurement model* (3.2) *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 $|J| \leq 6s$ 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 true constituent vector, $t^* = [w^*; z^*] \in \mathbb{R}^{2n}$ satisfies the following upper bound (in expectation) for any $k \geq 1$:*

$$\|t^{k+1} - t^*\|_2 \leq (2q)^k \|t^0 - t^*\|_2 + C\tau\sqrt{\frac{s}{m}}, \tag{4.3}$$

where $q = 2\sqrt{1 + \eta'^2 M_{6s}^2 - 2\eta' m_{6s}}$ and $C > 0$ is a constant that depends on the step size η' and the convergence rate q.

Equation (4.3) indicates the linear convergence behavior of our proposed algorithm. In particular, for the noiseless case $\tau = 0$, this implies that Alg. 1 returns a solution with accuracy κ after $N = O(\log \frac{\|t^0 - t^*\|_2}{\kappa})$ iterations. The proof of Theorem 4.1 leverages the fact that the objective function $F(t)$ in (4.1) satisfy the RSC/RSS conditions specified in Definition 3.3. Please see [10] and [26] for a more detailed discussion.

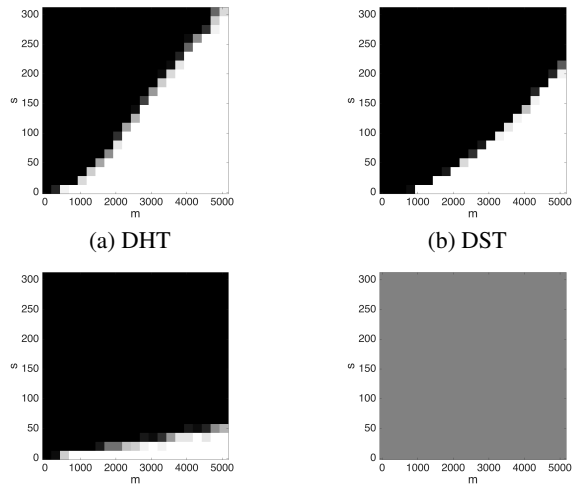
In the next theorem, we also provide the sample complexity of Alg. 1:

Theorem 4.2. *Under the assumptions in Theorem 4.1, the sample complexity or the required number of measurements to reliably recover constituent signals w and z is given by $m = \mathcal{O}(s\log n\log^2 s\log(s\log n))$, provided that the bases Φ and Ψ are incoherent enough.*

The leading constant in the expression for m is somewhat complicated, and hides the dependence on the incoherence parameter ε, the cross-coherence ϑ, the RSC/RSS constants, and the growth parameters of the link function l_1 and l_2. See [10] for more details.

5. NUMERICAL RESULTS

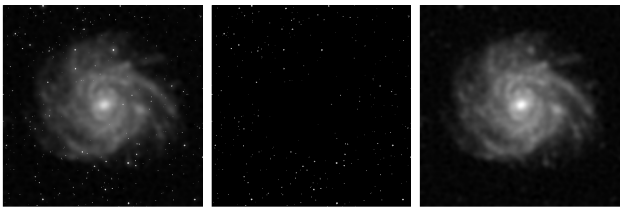
In this section, we provide some numerical experiments to illustrate the performance of the proposed algorithm both in synthetic and real data. We compare the algorithm with a convex relaxation-based heuristic version of the proposed algorithm that we dub as DEMIXING WITH SOFT THRESHOLDING (DST) which is a modification of the nonlinear recovery method in [25], and two other algorithms called NLCDLASSO and ONESHOT [9, 4] as discussed above in Section 2. In the experiments below, the initial estimate x^0 in both DHT and DST is set to be the solution returned by ONESHOT.



(a) DHT (b) DST

(c) ONESHOT (d) NLCDLASSO

Fig. 1: *Phase transition plots of various algorithms for solving the demixing problem* (3.2) *as a function of sparsity level* s *and number of measurements* m*.*



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

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

We first generate constituent signals, $w, z \in \mathbb{R}^n$ with $n = 2^{16}$. We consider 1D Haar wavelets and noiselet bases for Φ and Ψ respectively; these bases are known to be maximally incoherent relative to each other [28]. In addition, we choose a partial DFT matrix as the measurement matrix A. All three matrices (A, Φ, Ψ) are known to support fast matrix-vector multiplication operations and are well-suited for our application.

Both ONESHOT and NLCDLASSO do not assume knowledge of the link function and return a solution modulo 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 follows: $\cos(x, \widehat{x}) = \frac{x^T \widehat{x}}{\|x\|_2 \|\widehat{x}\|_2}$.

Figure 1 illustrates the performance of the four algorithms in terms of *phase transition* plots, following [1]. In these plots, we varied both the sparsity level s and the number of measurements m. For each pair (s, m), we randomly generate the test superposiition signal (by choosing both its support and coefficients at random) as well as the measurement matrix. We repeat this experiment over 20 Monte Carlo trials. The nonlinear link function is chosen as $g(x) = 2x + \sin(x)$; it is easy to check that the derivative of this function is strictly bounded between $l_1 = 1$ and $l_2 = 3$. The number of iterations for both DHT and DST is set to to 1000. The step size η' is hard to estimate in practice, and therefore is chosen by manual tuning such that both DHT and DST show the best performance. 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. 1, 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 the constituent signals w and z). The improvements over ONESHOT are to be expected since as discussed in [4], this algorithm does not leverage the knowledge of the link function g and is not iterative.

We also demonstrate the performance of our proposed algorithm on real-world 2D images. For this experiment, we consider an astronomical image illustrated in Fig. 2. 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 $x = \Phi w + \Psi z$ is observed using a subsampled Fourier matrix with $m = 15000$ rows multiplied with a diagonal matrix with random ± 1 entries [29]. Further, each measurement is nonlinearly transformed by applying the (shifted) logistic function $g(x) = \frac{1}{1+e^{-x}} - \frac{1}{2}$ 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. REFERENCES

[1] M. McCoy and J. Tropp, "Sharp recovery bounds for convex demixing, with applications," *Foundations of Comp. Math.*, vol. 14, no. 3, pp. 503–567, 2014.

[2] M. McCoy, V. Cevher, Q. Dinh, A. Asaei, and L. Baldassarre, "Convexity in source separation: Models, geometry, and algorithms," *IEEE Sig. Proc. Mag.*, vol. 31, no. 3, pp. 87–95, 2014.

[3] D. Donoho, M. Elad, and V. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," IEEE Trans. Inform. Theory, vol. 52, no. 1, pp. 6–18, 2006.

[4] M. Soltani and C. Hegde, "Demixing sparse signals from nonlinear observations," Tech. Rep., Iowa State University, 2016.

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

[6] E. Candès, "Compressive sampling," in *Proc. Int. Congress of Math.*, Madrid, Spain, Aug. 2006.

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

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

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

[10] M. Soltani and C. Hegde, "Fast algorithms for demixing sparse signals from nonlinear observations," *arXiv preprint arXiv:1608.01234*, 2016.

[11] J. Bobin, J. Starck, J. Fadili, Y. Moudden, and D. Donoho, "Morphological component analysis: An adaptive thresholding strategy," *IEEE Trans. Image Proc.*, vol. 16, no. 11, pp. 2675–2681, 2007.

[12] C. Studer, P. Kuppinger, G. Pope, and H. Bölcskei, "Recovery of sparsely corrupted signals," *IEEE Trans. Inform. Theory*, vol. 58, no. 5, pp. 3115–3130, 2012.

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

[14] Y. Plan and R. Vershynin, "One-bit compressed sensing by linear programming," *Comm. Pure and Applied Math.*, vol. 66, no. 8, pp. 1275–1297, 2013.

[15] E. Candes, X. Li, and M. Soltanolkotabi, "Phase retrieval via wirtinger flow: Theory and algorithms," *IEEE Trans. Inform. Theory*, vol. 61, no. 4, pp. 1985–2007, 2015.

[16] M. Davenport, Y. Plan, E. van den Berg, and M. Wootters, "1-bit matrix completion," *Information and Inference*, vol. 3, no. 3, pp. 189–223, 2014.

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

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

[19] 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)*, 2011, pp. 927–935.

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

[21] A. Beck and Y. Eldar, "Sparsity constrained nonlinear optimization: Optimality conditions and algorithms," *SIAM Journal on Optimization*, vol. 23, no. 3, pp. 1480–1509, 2013.

[22] S. Bahmani, B. Raj, and P. Boufounos, "Greedy sparsity-constrained optimization," *J. Machine Learning Research*, vol. 14, no. 1, pp. 807–841, 2013.

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

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

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

[26] S. Negahban, B. Yu, M. Wainwright, and P. Ravikumar, "A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers," in *Adv. Neural Inf. Proc. Sys. (NIPS)*.

[27] E. Candes and J. Romberg, "Sparsity and incoherence in compressive sampling," *Inverse problems*, vol. 23, no. 3, pp. 969, 2007.

[28] R. Coifman, F. Geshwind, and Y. Meyer, "Noiselets," *Appl. Comput. Harmonic Analysis*, vol. 10, no. 1, pp. 27–44, 2001.

[29] F. Krahmer and R. Ward, "New and improved johnson-lindenstrauss embeddings via the restricted isometry property," *SIAM Journal on Mathematical Analysis*, vol. 43, no. 3, pp. 1269–1281, 2011.