Nonlinear Structured Signal Estimation in High Dimensions via Iterative Hard Thresholding

Kaiqing Zhang
University of Illinois, Urbana-Champaign

Zhuoran Yang
Princeton University

Zhaoran Wang
Northwestern University and Tencent AI Lab

Abstract

We study the high-dimensional signal estimation problem with nonlinear measurements, where the signal of interest is either sparse or low-rank. In both settings, our estimator is formulated as the minimizer of the nonlinear least-squares loss function under a combinatorial constraint, which is obtained efficiently by the iterative hard thresholding (IHT) algorithm. Although the loss function is nonconvex due to the nonlinearity of the statistical model, the IHT algorithm is shown to converge linearly to a point with optimal statistical accuracy using arbitrary initialization. Moreover, our analysis only hinges on conditions similar to those required in the linear case. Detailed numerical experiments are included to corroborate the theoretical results.

1 Introduction

Signal recovery via linear measurements under the high-dimensional regime is extensively studied in the past two decades with fruitful results (Bühlmann and van de Geer, 2011). However, the linear model is too stringent for modeling real-world datasets where nonlinear models usually yield better performance. To relax the linear assumption, given a monotone and univariate function f, we study the nonlinear model

$$Y = f(\langle \boldsymbol{X}, \boldsymbol{\Theta}^* \rangle) + \epsilon, \tag{1.1}$$

where $Y \in \mathbb{R}$ is the response variable, $\boldsymbol{X}$ is the covariate, $\boldsymbol{\Theta}^*$ is the parameter of interest, and $\epsilon \in \mathbb{R}$ is the stochastic noise independent of $\boldsymbol{X}$. Here we assume f is known and $\boldsymbol{\Theta}^*$ is either a sparse vector or a low rank matrix and the inner product in (1.1) is the trace inner product in the matrix case. Given n independent and identically distributed (i.i.d.) observations $\{Y_i, \mathbf{X}_i\}_{i=1}^n$ of $(Y, \boldsymbol{X})$, our goal is to estimate $\boldsymbol{\Theta}^*$ in the high-dimensional setting where the ambient dimension is much larger than the sample size n.

When f is the identity function, model (1.1) reduces to the classical linear model. Our model is motivated by a broad family of nonlinear structured signal recovery problems that receive great research interest recently. See, e.g., Xu et al. (2011); Beck and Eldar (2013a); Blumensath (2013); Aksoylar and Saligrama (2014); Gulliksson and Oleynik (2016) and the references therein. Moreover, this model also finds applications in machine learning, for example, the training of (deep) neural networks (Hecht-Nielsen et al., 1988; Glorot and Bengio, 2010), where the activation functions are usually unknown and nonlinear.

Since f is known, a tempting method to handle the nonlinear model is to apply methods for the linear setting on the transformed data $\{f^{-1}(Y_i), \mathbf{X}_i\}_{i=1}^n$. Nonetheless, such an approach succeeds only in the noiseless case where ϵ is zero. In the noisy setting, the conditional distribution of $f^{-1}(Y)$ given $\boldsymbol{X}$ in general will not be centered at $\langle \boldsymbol{X}, \boldsymbol{\Theta}^* \rangle$ in presence of the stochastic noise. Thus, applying methods for the linear model usually incurs large estimation errors. Instead of avoiding the nonlinearity through transformation, we attack the estimation problem by minimizing the nonlinear least-squares loss function

$$\ell(\boldsymbol{\Theta}) := \frac{1}{2n} \sum_{i=1}^{n} \big[Y_i - f\left(\langle \boldsymbol{X}_i, \boldsymbol{\Theta} \rangle\right) \big]^2 \tag{1.2}$$

directly under a combinatorial constraint. Such a constraint enables us to obtain a sparse or low-rank solution for the vector and matrix cases, respectively. Specifically, when $\boldsymbol{\Theta}^*$ is a sparse vector, we solve the optimization problem in (2.1) subject to a cardinality constraint that $\boldsymbol{\Theta}$ has no more than s nonzero entries for some appropriate integer $s > 0$. Whereas we adopt the rank constraint $\mathrm{rank}(\boldsymbol{\Theta}) \leq s$ when $\boldsymbol{\Theta}^*$ is a low-rank matrix.

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84.

Due to the combinatorial constraint, such an optimization problem is not tractable. To obtain a computationally efficient estimator, we apply the iterative hard thresholding (IHT) algorithm, which is a special case of the projected gradient descent algorithm. At each iteration, the algorithm first performs a standard gradient descent step, then truncates the updated estimate such that the combinatorial constraint is satisfied. Specifically, for the cardinality constraint, the truncation step simply selects the largest s entries of the updated estimate in magnitude (Blumensath and Davies, 2009). As for the rank constraint, it reduces to computing the best rank-s approximation of the updated estimate (Tanner and Wei, 2013), which is achieved via singular value decomposition (SVD).

For linear models, such an algorithm is shown to converge linearly to an estimator with optimal statistical accuracy (Blumensath and Davies, 2009; Jain et al., 2014). Unlike the linear case in which the least-squares loss function is convex, in general settings the loss function in (1.2) can be highly nonconvex due to the existence of the nonlinear function f. Moreover, some standard assumptions on the loss function such as the restricted strong convexity (RSC) condition (Negahban et al., 2012) are too stringent to hold in general nonlinear settings. Hence, new theoretical guarantees are required for the estimator based on IHT algorithm.

In §3, we show that, despite nonconvexity, the IHT algorithm enjoys both computational efficiency and statistical accuracy. In particular, similar to the linear case, it converges linearly in terms of computation with arbitrary initialization and achieves optimal statistical rate of convergence after sufficient number of iterations. We summarize our main result as follows. Let $\{\boldsymbol{\Theta}^{(t)}, t \geq 0\}$ be the iterates of the IHT algorithm, under mild assumptions, there exist two absolute constants $0 < \mu < 1$ and $C > 0$ such that, with high probability,

$$\|\boldsymbol{\Theta}^{(t)} - \boldsymbol{\Theta}^*\|_\bullet \leq \mu^t \cdot \|\boldsymbol{\Theta}^{(0)} - \boldsymbol{\Theta}^*\|_\bullet + C \cdot \mathtt{Rate}^*. \quad (1.3)$$

Here $\|\cdot\|_\bullet$ stands for the ℓ_2-norm in the vector case and the Frobenius norm in the matrix case. Moreover, $\mathtt{Rate}^*$ in (1.3) denotes the minimax statistical rate of convergence in the linear model. The first term on the right-hand side of (1.3) is the optimization error, which converges to zero with linear rate; the second term is the statistical error, which establishes the optimality of our approach.

Related Work. The model in (1.1) is an extension of high-dimensional signal recovery with linear measurements, namely Compressed Sensing, which have been extensively studied. See Bühlmann and van de Geer (2011) for a thorough review of the literature. In this case, all projected gradient methods (Needell and Tropp, 2009; Blumensath and Davies, 2009; Garg and Khandekar, 2009; Foucart, 2011) are able to obtain optimal statistical rates of convergence.

In addition to Compressed Sensing, our model is also related to the Single Index Model (SIM), which assumes the nonlinear function in (1.1) is unknown. SIM has been studied in the low-dimensional settings where $d \ll n$. See, e.g., McCullagh et al. (1989); Härdle et al. (1993); Ichimura (1993); Sherman (1994); Xia and Li (1999); Delecroix et al. (2000, 2006); Horowitz (2000); Ganti et al. (2015) for details. Most of these works study M-estimators that are global optima of the nonconvex optimization problem, thus are known to be computationally intractable. For high-dimensional SIM with Gaussian covariates, Plan et al. (2017); Plan and Vershynin (2016); Thrampoulidis et al. (2015); Neykov et al. (2016a); Oymak and Soltanolkotabi (2016) study generalized Lasso estimators which enjoy sharp statistical rates of convergence. This method is later extended in Goldstein et al. (2016); Yang et al. (2017a) for non-Gaussian covariates. In addition, Han and Wang (2015) propose a method using rank-based statistics smoothing techniques, Yi et al. (2015) consider an estimator based on the method of moments, Chen and Banerjee (2017) propose robust estimators based on U-statistics. However, their results hinge on the assumption that the distribution of covariate is known. Hence the flexibility of SIM comes at the price of more stringent distributional assumptions on the data. Moreover, since $\|\boldsymbol{\Theta}^*\|_2$ is incorporated into f, these methods can only estimate the direction of $\boldsymbol{\Theta}^*$, i.e., $\boldsymbol{\Theta}^*/\|\boldsymbol{\Theta}^*\|_2$, instead of the parameter itself. In comparison to these works in the regime where f is known, our method is able to estimate $\boldsymbol{\Theta}^*$ directly with $\boldsymbol{X}$ following general distributions.

Moreover, the problem of sufficient dimension reduction is also relevant, where the goal is to recover a subspace $\mathcal{U}$ such that Y only depends on the projection of $\boldsymbol{X}$ onto $\mathcal{U}$. See, e.g., Li (1991, 1992); Cook (1998); Cook and Lee (1999) and the references therein. These estimators are based on similar symmetry assumptions and involve computing second-order (conditional and unconditional) moments which are difficult to estimate in high-dimensions without restrictive assumptions. Furthermore, Kalai and Sastry (2009); Kakade et al. (2011) propose iterative algorithms that estimate f and $\boldsymbol{\beta}^*$ alternatively, based on isotonic regression in the setting with $d \ll n$. However, theory for parameter estimation is not derived in their analysis. For the special case where the nonlinear function f is quadratic, the estimation problem is known as the phase retrieval problem, where the model is $Y = |\boldsymbol{X}^\top \boldsymbol{\beta}|^2 + \epsilon$ and $\boldsymbol{X} \in \mathbb{C}^d$ is a complex random vector. For high-dimensional settings, this problem

has been investigated under both noisy and noiseless settings. See, e.g., Jaganathan et al. (2012); Ohlsson et al. (2012); Li and Voroninski (2013); Candès et al. (2013); Eldar and Mendelson (2014); Ohlsson and Eldar (2014); Candès et al. (2015); Waldspurger et al. (2015); Eldar et al. (2015); Cai et al. (2016); Tu et al. (2016); Goldstein and Studer (2016); Dhifallah et al. (2017); Ma et al. (2017); Soltanolkotabi (2017) and the references therein. Moreover, Neykov et al. (2016b); Yang et al. (2017b) study the misspecified phase retrieval model in high dimensions, which can be viewed as a single index model with symmetric link functions. It is worth noting that all these works depend on the assumption that the distribution of covariate is known.

A more relevant line of works focuses on the sparsity-constrained optimization procedure for nonlinear problems. Shalev-Shwartz et al. (2010) studies a few greedy algorithms for minimizing the expected loss of a linear predictor. In their analysis, it is assumed that the loss function satisfies some smoothness and convexity condition, which does not hold for the least-squares loss function considered here. Liu et al. (2014) studies forward-backward selection algorithm for general convex smooth loss functions and Bahmani et al. (2013) considers the Gradient Support Pursuit algorithm and its variants. More relevant works are Yuan et al. (2014); Jain et al. (2014); Beck and Eldar (2013b) that provide analysis for the iterative hard thresholding algorithms. Although the convergence of the algorithms and the statistical error of the estimators are both derived in these works, they are limited to the case where the Restricted Strong Convexity and Restricted Smoothness Conditions are satisfied. This relatively stringent assumption forbids trivial extensions of these existing analysis to our problem with more general loss functions. For general Compressed Sensing problems, Blumensath (2013) proposes a formulation where the linear measurement operator which measures the signal is replaced by a general nonlinear operator. IHT algorithm is then advocated and the estimation error bound is obtained. However, the generality of the nonlinear operator comes at a price. Particularly, their analysis relies on both the Restricted Isometry Property (RIP) and Restricted Smoothness Conditions to hold for the Jacobian of the nonlinear operator. Moreover, the coefficient in the RIP condition is required to be less than 0.2, implying that the RIP assumption cannot be relaxed to the Restricted Strong Convexity condition. In contrast, we consider a more specific model and are able to derive optimal statistical rate of convergence under significantly weaker conditions.

Summary of Contributions. The main contribution of the present work is two-fold. First, we propose a unified treatment of the signal recovery problem with nonlinear statistical model in high dimensions. Second, we develop the iterative hard thresholding algorithm to efficiently achieve the recovery results for both sparse and low-rank signals. The IHT algorithm is guaranteed to achieve optimal statistical rates of convergence despite the model nonlinearity. In addition, the assumptions required for these guarantees are mild and similar to those required in the linear case.

Notation. We adopt the following notation throughout this paper. Let $\mathbb{N}, \mathbb{Z}$ and $\mathbb{R}$ be the set of natural numbers, integers, and real numbers. We write $\{1, \ldots, n\}$ as $[n]$ for any $n \in \mathbb{N}$, and $\lceil n \rceil$ as the smallest integer that is greater than n. For $0 \leq p \leq \infty$, we denote the ℓ_p-norm of $\mathbf{v}$ as $\|\mathbf{v}\|_p$, specifically, $\|\mathbf{v}\|_0$ denotes the number of nonzero entries in $\mathbf{v}$. For a matrix $\mathbf{M}$, let $\|\mathbf{M}\|_F$ and $\|\mathbf{M}\|_2$ be the Frobenius and operator norm of $\mathbf{M}$. Define $\|\mathbf{M}\|_{\max}$ as the max norm of $\mathbf{M}$, which is the largest absolute value of the elements in $\mathbf{M}$. We denote the inner product operation as $\langle \boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2 \rangle$. In the vector case, $\langle \boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2 \rangle = \boldsymbol{\Theta}_1^\top \boldsymbol{\Theta}_2$ whereas in the matrix case, $\langle \boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2 \rangle = \mathrm{tr}(\boldsymbol{\Theta}_1^\top \boldsymbol{\Theta}_2)$. For $\mathcal{S} \subseteq \mathbb{R}^d$, let $\mathbf{v}_\mathcal{S}$ denote the projection of $\mathbf{v}$ on the subspace $\mathcal{S}$ and similarly for $\mathcal{S} \subseteq \mathbb{R}^{m_1 \times m_2}$ and $\mathbf{M}_\mathcal{S}$. Let $|\mathcal{S}|$ denote the dimension of the subspace $\mathcal{S}$. We use $\mathrm{vec}(\mathbf{M})$ to denote the vectorized form of the matrix $\mathbf{M}$. A random vector $\boldsymbol{U} \in \mathbb{R}^d$ is sub-Gaussian with variance proxy τ^2 if $\mathbb{E}(\boldsymbol{U}) = \mathbf{0}$ and for all $\boldsymbol{A} \in \mathbb{R}^d$, $\mathbb{E}[\exp(\boldsymbol{A}^\top \boldsymbol{U})] \leq \exp(\|\boldsymbol{A}\|_2^2 \cdot \tau^2/2)$.

Organization. We introduce the combinatorial optimization problem and the estimation procedure in §2, and in §3 we present the theoretical results. The numerical experiments are illustrated in §4. Finally, we conclude the paper in §5 with further discussions. The theoretical guarantees of the proposed algorithm are proved in §B.

2 Parameter Estimation via Iterative Hard Thresholding

In this section, we introduce the proposed combinatorial optimization problem and the IHT algorithm for nonlinear structured signal estimation.

As stated in §1, we aim to estimate the high-dimensional signal $\boldsymbol{\Theta}^*$ given n i.i.d. observations $\{Y_i, \mathbf{X}_i\}_{i=1}^n$ of the model (1.1). It is assumed that s^*, d, m_1, and m_2 are positive integers satisfying $s^* \ll n \ll d$ and $s^* \ll \min\{m_1, m_2\}$. For sparse vector estimation, we assume $\boldsymbol{\Theta}^* \in \mathbb{R}^d$ with $\|\boldsymbol{\Theta}^*\|_0 = s^*$, while for low-rank matrix recovery, we consider $\boldsymbol{\Theta}^* \in \mathbb{R}^{m_1 \times m_2}$ with $\mathrm{rank}(\boldsymbol{\Theta}^*) = s^*$.

To recover the structured signal in high dimensions, we proposed a nonconvex optimization problem with a

Algorithm 1 The iterative hard thresholding (IHT) algorithm for nonlinear signal recovery in high dimensions

1: **Input:** Truncation level $s > 0$, step-size $\eta > 0$.
2: **Initialization:** set iteration counter $t \leftarrow 0$ and choose an initial estimator $\boldsymbol{\Theta}^{(0)}$, either a vector in $\mathbb{R}^d$ or a matrix in $\mathbb{R}^{m_1 \times m_2}$
3: **Output:** A sequence of estimators $\{\boldsymbol{\Theta}^{(t)}, t \geq 1\}$
4: **Repeat**
5: $\widetilde{\boldsymbol{\Theta}}^{(t+1)} \leftarrow \boldsymbol{\Theta}^{(t)} - \eta \cdot \nabla \ell(\boldsymbol{\Theta}^{(t)})$
6: $\boldsymbol{\Theta}^{(t+1)} \leftarrow \mathrm{Trunc}(\widetilde{\boldsymbol{\Theta}}^{(t+1)}, s)$. Solved by Algorithms 2 or 3 in §A
7: Update iteration counter $t \leftarrow t + 1$
8: **Until convergence**

combinatorial constraint, which reduces to a cardinality or a rank constraint. Specifically, for sparse vector recovery, the estimator is the minimizer of the combinatorial optimization problem

$$\min_{\boldsymbol{\Theta} \in \mathbb{R}^d} \ell(\boldsymbol{\Theta}) \ \text{ such that } \ \|\boldsymbol{\Theta}\|_0 \leq s, \tag{2.1}$$

where $\ell(\boldsymbol{\Theta})$ is as defined in (1.2), and $s > 0$ is the parameter that will be specified later. By definition, the optimization problem in (2.1) will always return a solution with no more than s nonzero entries. Likewise, for low-rank matrix recovery, we formulate an optimization problem with a rank constraint

$$\min_{\boldsymbol{\Theta} \in \mathbb{R}^{m_1 \times m_2}} \ell(\boldsymbol{\Theta}) \ \text{ such that } \ \mathrm{rank}(\boldsymbol{\Theta}) \leq s. \tag{2.2}$$

Due to the high computational complexity to directly solve (2.1) and (2.2) under the high-dimensional regime, we resort to the iterative hard thresholding algorithm to efficiently attack the combinatorial optimization problems. Such an algorithm iteratively generates a sequence of estimators $\{\boldsymbol{\Theta}^{(t)}, t \geq 1\}$ via a standard gradient descent followed by a truncation step. The unified algorithm for both sparse vector and low-rank matrix recovery via IHT is summarized in Algorithm 1.

In detail, let $\eta > 0$ be a fixed step-size. For each $t \geq 1$, at the t-th iteration, the algorithm first performs a gradient descent step

$$\widetilde{\boldsymbol{\Theta}}^{(t+1)} = \boldsymbol{\Theta}^{(t)} - \eta \cdot \nabla \ell(\boldsymbol{\Theta}^{(t)})$$

and then does a truncation step

$$\boldsymbol{\Theta}^{(t+1)} = \mathrm{Trunc}(\widetilde{\boldsymbol{\Theta}}^{(t+1)}, s), \tag{2.3}$$

where the positive integer $s > 0$ is the parameter of the algorithm controlling the truncation level. Specifically, for sparse vector recovery, the truncation function simply keeps the largest s entries of $\widetilde{\boldsymbol{\Theta}}^{(t+1)}$ in magnitude and shrinks the rest of the entries to zero. For low-rank matrix recovery, the truncation function computes the best s-rank approximation of $\widetilde{\boldsymbol{\Theta}}^{(t+1)}$ via singular value decomposition. The truncation step in (2.3) are presented in detail as Algorithms 2 and 3, respectively, in §A. Algorithm 1 iterates continuously until a convergence criteria is reached. For example, we can terminate the algorithm if $\|\boldsymbol{\Theta}^{(t+1)} - \boldsymbol{\Theta}^{(t)}\|_\bullet / \|\boldsymbol{\Theta}^{(t)}\|_\bullet \leq \varepsilon$ for some threshold $\varepsilon > 0$.

Although Algorithm 1 proceeds in the same way as its linear counterpart, i.e., the IHT algorithm for linear structured signal recovery (Bühlmann and van de Geer, 2011), there are still several questions remaining open. First, due to the nonlinearity of f, the loss function in (1.2) can be highly nonconvex. Thus it is not clear whether the IHT algorithm will converge; even if the algorithm does converge, nor is it clear about the rate of convergence and where the algorithm converges. In addition, for general nonconvex optimization, initialization plays a significant role in gradient-based algorithms since the algorithms can easily get stuck at a local minimum or saddle point with an initial point around them. Therefore, it is imperative to investigate whether our IHT algorithm requires nontrivial initialization.

Interestingly, we show in the next section that, under mild assumptions, the IHT algorithm is guaranteed to converge linearly for nonlinear structured signal estimation with proper step-size η under random initialization. Furthermore, the algorithm converges to a point with optimal statistical rate of convergence. Hence our proposed estimator achieves both computational feasibility and statistical accuracy.

3 Theoretical Results

In this section, we establish the convergence results for the iterative hard thresholding algorithm. As the basis for the ensuing analysis, an assumption on the nonlinear function f is first stated.

Assumption 3.1. We assume the nonlinear function $f : \mathbb{R} \to \mathbb{R}$ in (1.1) is monotone and differentiable. Additionally, there exist two positive constants a, and b such that $f'(u) \in [a, b]$ for all $u \in \mathbb{R}$.

Note that such an assumption holds in a significant machine learning application, i.e., the training of deep neural networks. In this case, the derivative f' is bounded as long as $\langle \boldsymbol{X}, \boldsymbol{\Theta}^* \rangle$ is bounded, whereas the latter has to be satisfied because otherwise the gradient of the cost function would either vanish or explode, making the training process slow and inefficient (Glorot and Bengio, 2010).

We acknowledge that such an assumption is stronger

than that in sparse single index models (Plan et al., 2017; Plan and Vershynin, 2016; Goldstein et al., 2016), where f only needs to satisfy that $\mathbb{E}[f'(\langle \boldsymbol{X}, \boldsymbol{\Theta}^* \rangle)] \neq 0$. However, their generality comes at a price. Specifically, all these works need to assume that the distribution of $\boldsymbol{X}$ is Gaussian or Elliptical, which is much stronger than our assumption on the covariate as we will provide later. In fact, for nonlinear signal recovery problems with general covariate, knowledge of f is required for consistent estimation. As such, the problems are solved in a case-by-case manner. See Ai et al. (2014); Loh and Wainwright (2015); Krahmer and Liu (2016) for results in generalized linear models, phase retrieval, and one-bit compressed sensing with sub-Gaussian covariate. By sacrificing some flexibility of f as in Assumption 3.1, we aim to present a more unified analysis for a rich class of nonlinear signal recovery problems, including both the sparse and low-rank cases, with milder assumptions on the covariate.

In the following, we present an assumption on the covariates $\{\mathbf{X}_i\}_{i=1}^n$, which are n independent observations of $\boldsymbol{X}$ in (1.1). Recall that $\boldsymbol{X}$ is a random vector in $\mathbb{R}^d$ for the sparse model and a random matrix in $\mathbb{R}^{m_1 \times m_2}$ for the low-rank case. The following assumption states that the Restricted Isometry Property (Candès and Tao, 2005; Candès, 2008) holds for the covariates.

Assumption 3.2. For $\{\mathbf{X}_i\}_{i=1}^n$, we assume that the RIP condition holds with parameter $2s+s^*$, where s^* is the sparsity or rank of $\boldsymbol{\Theta}^*$ and there exists $s \geq C_0 s^*$ for some sufficiently large constant C_0. For convenience of derivation, we further assume that $C_0 \geq 8$ in the ensuing analysis.

More specifically, for sparse vector recovery where $s^* = \|\boldsymbol{\Theta}^*\|_0$, we assume that for any $\mathbf{V} \in \mathbb{R}^d$ with $\|\mathbf{V}\|_0 \leq 2s + s^*$, there exists a constant $\delta(2s + s^*) \in [0, 1)$ such that

$$\left| \frac{1}{n} \sum_{i=1}^n \langle \mathbf{X}_i, \mathbf{V} \rangle^2 - \|\mathbf{V}\|_2^2 \right| \leq \delta(2s + s^*) \cdot \|\mathbf{V}\|_2^2. \quad (3.1)$$

Moreover, for the low-rank case where $s^* = \mathrm{rank}(\boldsymbol{\Theta}^*)$, we assume that for any $\mathbf{V} \in \mathbb{R}^{m_1 \times m_2}$ with $\mathrm{rank}(\mathbf{V}) \leq 2s + s^*$, we have

$$\left| \frac{1}{n} \sum_{i=1}^n \langle \mathbf{X}_i, \mathbf{V} \rangle^2 - \|\mathbf{V}\|_F^2 \right| \leq \delta(2s + s^*) \cdot \|\mathbf{V}\|_F^2, \quad (3.2)$$

where $\delta(2s + s^*) \in [0, 1)$ is a constant.

Note that the constant $\delta(2s + s^*)$ in (3.1) and (3.2) are not related; the one in (3.1) depends only on (n, d, s^*, s) whereas the one in (3.2) relies on (n, m_1, m_2, s^*, s). For ease of notation, we keep δ as a function of only $2s + s^*$.

The RIP condition is one of the earliest sufficient conditions for the success of compressed sensing, and has significant impact on the development of high-dimensional statistics. As shown in Vershynin (2010), in the vector case, the RIP condition is satisfied with high probability when $\boldsymbol{X}$ is a sub-Gaussian isotropic vector and that for the low-rank case holds when $\mathrm{vec}(\boldsymbol{X})$ is isotropic and sub-Gaussian. This includes the most common case assumed in low-rank matrix recovery where $\boldsymbol{X}$ has i.i.d. Gaussian entries (Negahban and Wainwright, 2011).

For estimating a high-dimensional sparse vector, the RIP condition can be relaxed to the sparse eigenvalue (SE) condition. Specifically, for any k-sparse $\mathbf{V}$, it only requires

$$\rho_-(k) \cdot \|\mathbf{V}\|_2^2 \leq \frac{1}{n} \sum_{i=1}^n |\mathbf{X}_i^\top \mathbf{V}|^2 \leq \rho_+(k) \cdot \|\mathbf{V}\|_2^2,$$

where $\rho_-(k)$ and $\rho_+(k)$ are two positive constants. Note that RIP requires $\rho_+(k) \leq 2$, thus is more stringent. The sparse eigenvalue condition and a closely related notion, the restricted strong convexity condition, have been studied extensively by Bickel et al. (2009); Raskutti et al. (2010); Zhang (2010); Negahban et al. (2012); Xiao and Zhang (2013); Bahmani et al. (2013); Wang et al. (2014); Loh and Wainwright (2015).

Remark 3.3. Note that the SE condition is only used for sparse vector recovery whereas the RSC condition is also used for low-rank matrix recovery (Negahban et al., 2012). It is not clear whether the counterpart of the SE condition for the low-rank case can be used for theoretical analysis. In two recent works, Carpentier and Kim (2015); Rauhut et al. (2016) consider the IHT algorithm for low-rank matrix recovery with linear measurements. Both of their theories hinge on the RIP condition; it is not clear whether such condition could be relaxed. Moreover, we note that in terms of sparse signal estimation, our theory also holds under the SE condition. We adopt the RIP condition in order to have a uniform treatment since the matrix model usually requires more delicate conditions.

In the following, we present the main results of this paper. For sparse vector recovery, an additional assumption on the regularity of the entries of $n^{-1} \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^\top$ is required for a more refined result.

Assumption 3.4. There exists an absolute constant D that does not depend on n, d, or s^* such that

$$\left\| \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^\top \right\|_{\max} \leq D.$$

This assumption is true if the distribution of $\boldsymbol{X}$ is regular. For example, if $\boldsymbol{X}$ is sub-Gaussian or sub-exponential, standard concentration inequality implies that

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^\top - \mathbb{E}\big(\boldsymbol{X}\boldsymbol{X}^\top\big) \right\|_{\max} \longrightarrow 0$$

as n goes to infinity. In this case Assumption 3.4 is naturally satisfied.

We thus obtain the following theorem in the case of sparse vector recovery.

Theorem 3.5. Under Assumptions 3.1, 3.2, and 3.4, if the step size η in Algorithm 1 satisfies

$$\frac{3}{7a^2[1-\delta(2s+s^*)]} < \eta < \frac{11}{7b^2[1+\delta(2s+s^*)]}$$

then for $\{\boldsymbol{\Theta}^{(t)}, t \geq 1\}$ obtained from Algorithm 1, it holds with probability at least $1-d^{-1}$ that for all $t \geq 1$,

$$\|\boldsymbol{\Theta}^{(t)} - \boldsymbol{\Theta}^*\|_2 \leq \underbrace{\mu_1^t \cdot \|\boldsymbol{\Theta}^{(0)} - \boldsymbol{\Theta}^*\|_2}_{\text{optimization error}} + \underbrace{C_1\sqrt{s^* \log d/n}}_{\text{statistical error}},$$

for the sparse vector recovery problem (2.1). $\mu_1 \in (0,1)$ and C_1 are absolute constants that do not depend on n, d, or s^*.

By Theorem 3.5, we see that the IHT algorithm converges linearly and yields an estimator with sharp statistical rate. In each step, the estimation error of $\boldsymbol{\Theta}^{(t)}$ is decomposed into two parts: the first part is a geometric sequence that converges to zero rapidly whereas the second part is the statistical error of order $\sqrt{s^* \log d/n}$. It is clear that after sufficient number of iterations, Algorithm 1 will output an estimator $\widehat{\boldsymbol{\Theta}}$ such that $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_2 = \mathcal{O}_{\mathbb{P}}(\sqrt{s^* \log d/n})$. This rate is of the same order as the minimax optimal rate of noisy compressed sensing and high-dimensional linear regression (Raskutti et al., 2011).

Furthermore, for low-rank matrix recovery, similar result on the convergence rate and statistical error is stated as follows.

Theorem 3.6. Under Assumptions 3.1 and 3.2, if the step size η in Algorithm 1 satisfies

$$\frac{1}{b^2} < \eta < \frac{11}{7b^2[1+2\delta(2s+s^*)]}, \quad \text{or}$$
$$\frac{3}{7[a^2-2b^2\delta(2s+s^*)]} < \eta < \frac{1}{a^2},$$

then for $\{\boldsymbol{\Theta}^{(t)}, t \geq 1\}$ obtained from Algorithm 1, it holds that for all $t \geq 1$,

$$\|\boldsymbol{\Theta}^{(t)} - \boldsymbol{\Theta}^*\|_F \leq \underbrace{\mu_2^t \cdot \|\boldsymbol{\Theta}^{(0)} - \boldsymbol{\Theta}^*\|_F}_{\text{optimization error}} + \underbrace{C_2\sqrt{s^*} \cdot \left\| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \mathbf{X}_i \right\|_2}_{\text{statistical error}},$$

for the low-rank matrix recovery problem (2.2). $\mu_2 \in (0,1)$ and C_2 are absolute constants that do not depend on m_1, m_2, n, or s^*.

Remark 3.7. To obtain a more general result, we drop the assumption on Gaussian covariate $\boldsymbol{X}$ as in most matrix sensing literature (Negahban et al., 2009; Recht et al., 2010). The order of statistical error, i.e., $\sqrt{s^*} \cdot \left\|n^{-1}\sum_{i=1}^{n} \epsilon_i \mathbf{X}_i\right\|_2$, is dependent on the assumption made on the covariate $\boldsymbol{X}$. In particular, if $\|\mathbf{X}_i\|_2 \leq R$ for some $R > 0$, then the statistical error has the order of $\mathcal{O}_{\mathbb{P}}(\sqrt{m_1+m_2} \cdot \sqrt{s^* \log(m_1+m_2)/n})$ with high probability (Tropp, 2012). As a special case, the same rate is achieved if $\mathbf{X}_i$ are assumed to be sampled from i.i.d. sub-Gaussian distribution. In addition, if we further assume $\mathbf{X}_i$ are sampled from the $\boldsymbol{\Sigma}$-ensemble for some positive definite $\boldsymbol{\Sigma}$ as in Negahban and Wainwright (2011), the rate can be improved to $\mathcal{O}_{\mathbb{P}}(\sqrt{m_1+m_2} \cdot \sqrt{s^*/n})$. In this case, the statistical rate of convergence attained by Algorithm 1 has the same order as the minimax optimal rate for linear low-rank matrix recovery shown in Negahban et al. (2009).

The proofs of both Theorem 3.5 and Theorem 3.6 are provided in the appendix.

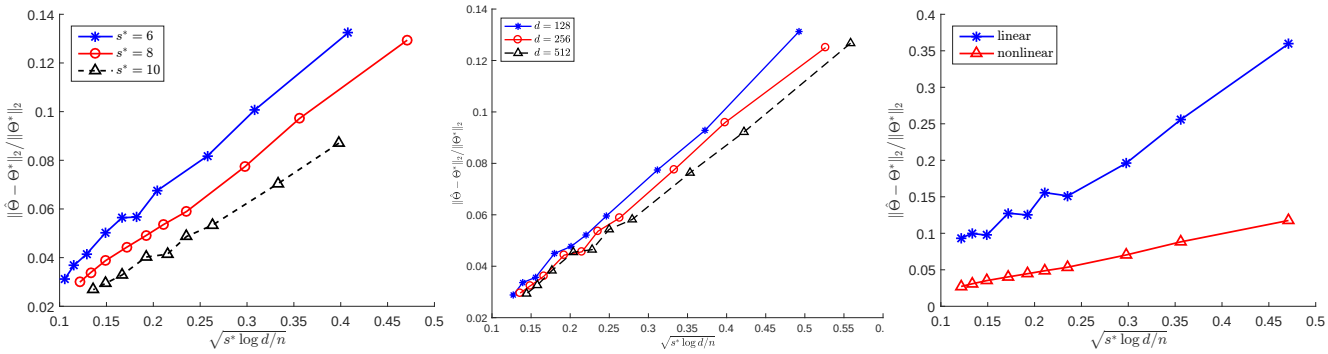
4 Numerical Experiments

We assess the finite sample performance of the proposed IHT algorithm for nonlinear structured signal estimation on both real and simulated data.

4.1 Tests on Simulated Data

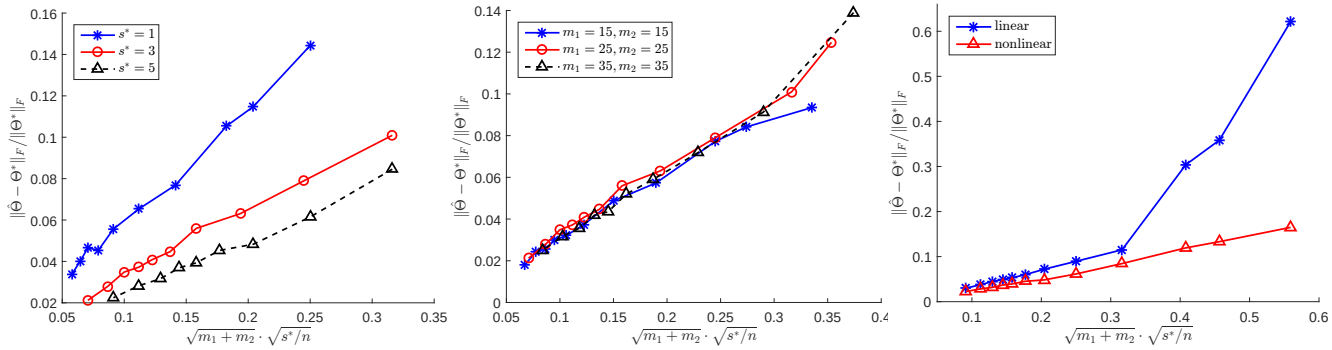
We first test the algorithm on simulated data for both sparse vector recovery and low-rank matrix recovery.

Data generation: We generate simulated data independently from model (1.1) with $\epsilon \sim \mathcal{N}(0,1)$. For sparse vector recovery, we generate $\mathbf{X}_i$ that follows $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ is a Toeplitz matrix with $\Sigma_{jk} = 0.95^{|j-k|}$ for any $1 \leq j \neq k \leq d$. The first s^* entries of $\boldsymbol{\Theta}^*$ are independently sampled from the uniform distribution on interval $[0,1]$ whereas the remaining entries are set zero, i.e., $\Theta_j^* \sim \mathrm{U}(0,1)$ for $1 \leq j \leq s^*$ and $\Theta_j^* = 0$ for $j > s^*$. For low-rank matrix recovery, the covariate $\mathbf{X}_i$ are sampled from $\boldsymbol{\Sigma}'$-ensemble. Specifically, the entries of $\mathrm{vec}(\mathbf{X}_i)$ follow $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}')$ with $\boldsymbol{\Sigma}'$ a Toeplitz matrix and $\Sigma'_{jk} = 0.5^{|j-k|}$ for any $1 \leq j \neq k \leq m_1 \times m_2$. We set $\boldsymbol{\Theta}^* = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^T$, where $\mathbf{U}$ and $\mathbf{V}$ are orthonormalized after the entries are i.i.d. drawn from $\mathcal{N}(0,1)$, and $\boldsymbol{\Lambda}$ is a diagonal matrix whose first s^* diagonal values are 1 and the remains are zero. The nonlinear function in model (1.1) is selected to be $f(x) = 2x + \cos(x)$ such that the derivative f' is bounded by $a = 1$ and $b = 4$. We sample $n = 30$ i.i.d. observations for both tests.



(a) Estimation with $d = 256$. (b) Estimation with $s^* = 10$. (c) Compare with Lasso.

Figure 1: Relative statistical error $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_2/|\boldsymbol{\Theta}^*\|_2$ for sparse vector recovery plotted against the value of $\sqrt{s^* \log d/n}$. In (a), $d = 256$ is fixed and the results with various values of s^* are compared. In (b), $s^* = 10$ is fixed and the results with various values of d are compared. In (c), the result of applying the IHT algorithm on the nonlinear measurements is compared with that of applying Lasso on the inverted linear measurements.



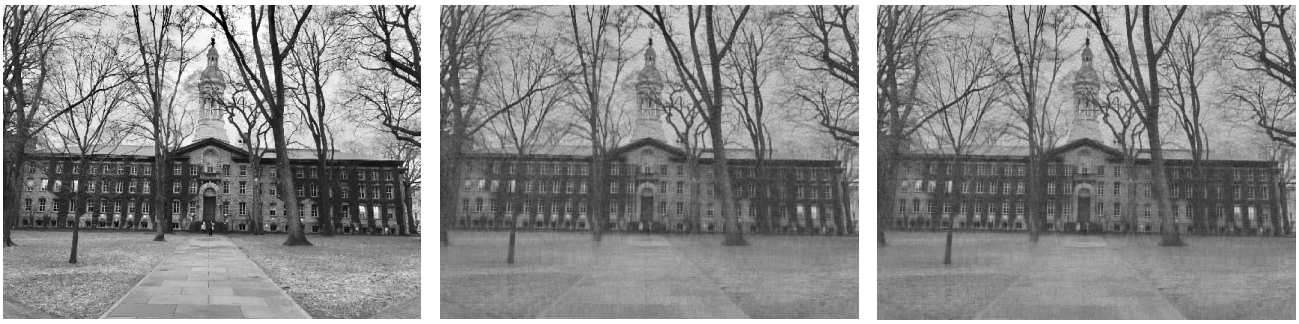
(a) Estimation with $m_1 = m_2 = 25$. (b) Estimation with $s^* = 3$. (c) Compare with NNLS.

Figure 2: Relative statistical error $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_F/|\boldsymbol{\Theta}^*\|_F$ for low-rank matrix recovery plotted against the value of $\sqrt{m_1 + m_2} \cdot \sqrt{s^*/n}$. In (a), $m_1 = m_2 = 25$ are fixed and the results with various values of s^* are compared. In (b), $s^* = 10$ is fixed and the results with various values of m_1 and m_2 are compared. In (c), the result of applying the IHT algorithm on the nonlinear measurements is compared with that of applying NNLS on the inverted linear measurements.

Sparse vector recovery: We compare the estimation error with $\sqrt{s^* \log d/n}$ under two different settings: (i) we fix $d = 256$, $s^* = 6, 8$, or 10, and vary n, and (ii) fix $s^* = 10$, $d = 128, 256$, or 512, and vary n. Here $\widehat{\boldsymbol{\Theta}}$ is the estimator produced by Algorithm 1. Given the data, we apply the IHT algorithm with $s = s^*$ and $\eta = 0.2$. With random initialization, we run $T = 1000$ iterations to obtain $\widehat{\boldsymbol{\beta}}$ and simulate the estimation procedure 30 times to evaluate the average error. As illustrated in Figure 1, the average estimation error $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_2$ grows linearly with $\sqrt{s^* \log d/n}$. This verifies our argument in Theorem 3.5 that $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_2 \le C_1 \sqrt{s^* \log d/n}$ for some absolute constant C_1.

Low-rank matrix recovery: Since the covariate $\mathbf{X}_i$ are sampled from $\boldsymbol{\Sigma}'$-ensemble, the statistical error is proved to be $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_F = \mathcal{O}_{\mathbb{P}}(\sqrt{m_1 + m_2} \cdot \sqrt{s^*/n})$ (See Remark 3.7). Here we select $s = s^*$ and $\eta = 0.1$. We plot the estimation error versus $\sqrt{m_1 + m_2} \cdot \sqrt{s^*/n}$ under two settings: (i) we fix $m_1 = 25, m_2 = 25$, $s^* = 1, 3$, or 5, and vary n, and (ii) fix $s^* = 3$, $m_1 = m_2 = 15, 25$, or 35, and vary n. In Figure 2, we show that $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_F$ grows (sub)linearly with $\sqrt{m_1 + m_2} \cdot \sqrt{s^*/n}$, which corroborates Theorem 3.6.

Comparison with linear estimators: Note that since the nonlinear function f is known, it is tempting to apply linear signal estimation techniques to the inverted data $\{Z_i, \mathbf{X}_i\}_{i=1}^n$ where $Z_i = f^{-1}(Y_i)$. However, since the mean of $Z = f^{-1}(Y) = f^{-1}[f(\langle \boldsymbol{X}, \boldsymbol{\Theta}^* \rangle) + \epsilon]$ is generally different from $\langle \boldsymbol{X}, \boldsymbol{\Theta}^* \rangle$ conditioned on $\boldsymbol{X}$, linear estimators may generate large estimation errors in noisy cases. We compare our IHT algorithm with two linear estimators in the experiment.



(a) The original image. (b) The compressed image. (c) The reconstructed image.

Figure 3: We apply the IHT algorithm to a sparse signal recovery problem based on a photo of the Nassau Hall at Princeton University. We show the original image in (a). The compressed image with $s^* = 80$ is shown in (b). In (c), we plot the reconstructed image using the IHT algorithm.

To this end, we first consider the setting where $d = 256$ and $s^* = 8$ for sparse vector recovery. Lasso is applied to the inverted data $\{Z_i, \mathbf{X}_i\}_{i=1}^n$ and the estimation error is reported in Figure 1 (c). It is shown that the proposed method indeed outperforms the linear method in terms of estimation errors. For low-rank matrix recovery, we apply the accelerated proximal gradient algorithm for nuclear norm regularized least squares (NNLS) (Toh and Yun, 2010) on the inverted data, where the setting $m_1 = m_2 = 25$ and $s^* = 5$ is considered. When the sample number n becomes relatively small, the NNLS algorithm explodes in estimation error while the proposed IHT algorithm maintains linear statistical error rate.

4.2 A Real-Data Example

We apply our algorithm for the sparse case to an image reconstruction example. The sparse signal is constructed from an image as follows. Let $\mathcal{I} \in \mathbb{R}^{h \times w}$ be the image with height $h \in \mathbb{N}$ and width $w \in \mathbb{N}$, where for simplicity we assume $h \leq w$. Let $\mathcal{I} = \sum_{j \in [h]} \sigma_j \cdot \mathbf{u}_j \mathbf{v}_j^\top$ be the singular value decomposition of $\mathcal{I}$, where $\sigma_1, \sigma_2, \ldots, \sigma_h$ are the singular values of $\mathcal{I}$ in the descending order, $\{\mathbf{u}_j\}_{j \in [h]} \subseteq \mathbb{R}^h$ and $\{\mathbf{v}_j\}_{j \in [h]} \subseteq \mathbb{R}^w$ are the left and right singular vectors, respectively. For a fixed integer s^*, let $\widetilde{\mathcal{I}} = \sum_{j \in [s^*]} \sigma_j \cdot \mathbf{u}_j \mathbf{v}_j^\top$ be the best rank-s^* approximation of $\mathcal{I}$. Finally, we let $\mathbf{b} = (\sigma_1, \ldots, \sigma_{s^*}, 0, \ldots, 0)^\top \in \mathbb{R}^h$ be the vector consisting of the top s^* singular values, and let the signal parameter be $\boldsymbol{\beta}^* = \mathbf{b}/\|\mathbf{b}\|_2$. We fix $\{\mathbf{u}_j, \mathbf{v}_j\}_{j \in [h]}$ and $\alpha = \|\mathbf{b}\|_2$. Given an estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}^*$, we reconstruct an image by $\widehat{\mathcal{I}} = \sum_{j \in [h]} \alpha \cdot \widehat{\beta}_j \cdot \sigma_j \cdot \mathbf{u}_j \mathbf{v}_j^\top$, which is an estimator of $\widetilde{\mathcal{I}}$.

In the experiment, we let $\mathcal{I}$ be a photo of the Nassau Hall at Princeton University (see Figure 3(a)) with $h = 1080$ and $w = 1440$. The signal parameter $\boldsymbol{\beta}^* \in \mathbb{R}^h$ is constructed with $s^* = 80$. To obtain the data, we sample $n = \lceil 5 s^* \log h \rceil$ i.i.d. observations of the nonlinear regression model $Y = f(\langle \boldsymbol{X}, \boldsymbol{\beta}^* \rangle) + \epsilon$, where the link function is $f(u) = 2u + \cos u$, the covariate is $\boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and the noise is $\epsilon \sim \mathcal{N}(0, 1)$. Given the data, we apply the IHT algorithm with $s = 100$ and $\eta = 0.005$. With random initialization, we run $T = 1000$ iterations to obtain $\widehat{\boldsymbol{\beta}}$. We observe that the ℓ_2-error of the iterates decays rapidly and converges to about 0.06, which indicates that we achieve a relative error of 6%. Moreover, the performance is not sensitive to the choice of s, η, and the initialization. The reconstructed image $\widehat{\mathcal{I}}$ is shown in Figure 3(c). Comparing $\widehat{\mathcal{I}}$ with the compressed image $\widetilde{\mathcal{I}}$ in Figure 3(b), we perceive very little visual difference, which demonstrates the success of our method.

5 Conclusions

In this paper, we consider a nonlinear structured signal estimation problem in high dimensions and propose an estimator that minimizes the nonlinear least squares loss function with combinatorial constraint. The iterative hard thresholding algorithm is leveraged to achieve an estimator for both the sparse and the low-rank models. Under mild assumptions similar to those required in the linear case, the IHT algorithm is guaranteed to converge linearly to a point which enjoys optimal statistical accuracy despite the model nonlinearity.

An interesting direction of future work is to extend the class of nonlinear functions in (1.1). Currently we assume f is monotonically increasing with derivative bounded from below and above. It would be interesting to incorporate functions such as $f(u) = u^2$ and $f(u) = \mathrm{sign}(u)$, thus making our analysis applicable to problems including phase retrieval and one-bit compressed sensing, where existing works require the covariate to be Gaussian distributed.

References

Ai, A., Lapanowski, A., Plan, Y. and Vershynin, R. (2014). One-bit compressed sensing with non-Gaussian measurements. *Linear Algebra and its Applications*, **441** 222–239.

Aksoylar, C. and Saligrama, V. (2014). Sparse recovery with linear and nonlinear observations: Dependent and noisy data. *arXiv preprint arXiv:1403.3109*.

Bahmani, S., Raj, B. and Boufounos, P. T. (2013). Greedy sparsity-constrained optimization. *Journal of Machine Learning Research*, **14** 807–841.

Beck, A. and Eldar, Y. C. (2013a). Sparse signal recovery from nonlinear measurements. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE.

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

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, **37** 1705–1732.

Blumensath, T. (2013). Compressed sensing with nonlinear observations and related nonlinear optimization problems. *IEEE Transactions on Information Theory*, **59** 3466–3474.

Blumensath, T. and Davies, M. E. (2009). Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, **27** 265–274.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Cai, T. T., Li, X., Ma, Z. et al. (2016). Optimal rates of convergence for noisy sparse phase retrieval via thresholded Wirtinger flow. *The Annals of Statistics*, **44** 2221–2251.

Candès, E. J. (2008). The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, **346** 589–592.

Candès, E. J., Eldar, Y. C., Strohmer, T. and Voroninski, V. (2015). Phase retrieval via matrix completion. *SIAM Review*, **57** 225–251.

Candès, E. J., Strohmer, T. and Voroninski, V. (2013). Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, **66** 1241–1274.

Candès, E. J. and Tao, T. (2005). Decoding by linear programming. *IEEE Transactions on Information Theory*, **51** 4203–4215.

Carpentier, A. and Kim, A. K. (2015). An iterative hard thresholding estimator for low rank matrix recovery with explicit limiting distribution. *arXiv preprint arXiv:1502.04654*.

Chen, S. and Banerjee, A. (2017). Robust structured estimation with single-index models. In *International Conference on Machine Learning*.

Cook, R. D. (1998). Principal Hessian directions revisited. *Journal of the American Statistical Association*, **93** 84–94.

Cook, R. D. and Lee, H. (1999). Dimension reduction in binary response regression. *Journal of the American Statistical Association*, **94** 1187–1200.

Delecroix, M., Hristache, M. and Patilea, V. (2000). Optimal smoothing in semiparametric index approximation of regression functions. Tech. rep., Discussion Papers, Interdisciplinary Research Project 373: Quantification and Simulation of Economic Processes.

Delecroix, M., Hristache, M. and Patilea, V. (2006). On semiparametric M-estimation in single-index regression. *Journal of Statistical Planning and Inference*, **136** 730–769.

Dhifallah, O., Thrampoulidis, C. and Lu, Y. M. (2017). Phase retrieval via linear programming: Fundamental limits and algorithmic improvements. *arXiv preprint arXiv:1710.05234*.

Eldar, Y. C. and Mendelson, S. (2014). Phase retrieval: Stability and recovery guarantees. *Applied and Computational Harmonic Analysis*, **36** 473–494.

Eldar, Y. C., Sidorenko, P., Mixon, D. G., Barel, S. and Cohen, O. (2015). Sparse phase retrieval from short-time Fourier measurements. *Signal Processing Letters, IEEE*, **22** 638–642.

Foucart, S. (2011). Hard thresholding pursuit: An algorithm for compressive sensing. *SIAM Journal on Numerical Analysis*, **49** 2543–2563.

Ganti, R. S., Balzano, L. and Willett, R. (2015). Matrix completion under monotonic single index models. In *Advances in Neural Information Processing Systems*.

Garg, R. and Khandekar, R. (2009). Gradient descent with sparsification: An iterative algorithm for sparse recovery with restricted isometry property. In *International Conference on Machine Learning*.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *International Conference on Artificial Intelligence and Statistics*.

Goldstein, L., Minsker, S. and Wei, X. (2016). Structured signal recovery from non-linear

and heavy-tailed measurements. *arXiv preprint arXiv:1609.01025*.

Goldstein, T. and Studer, C. (2016). Phasemax: Convex phase retrieval via basis pursuit. *arXiv preprint arXiv:1610.07531*.

Gulliksson, M. and Oleynik, A. (2016). Greedy Gauss-Newton algorithm for finding sparse solutions to nonlinear underdetermined systems of equations. *arXiv preprint arXiv:1610.03095*.

Han, F. and Wang, H. (2015). Provable smoothing approach in high dimensional generalized regression model. *arXiv preprint arXiv:1509.07158*.

Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single-index models. *The Annals of Statistics*, **21** 157–178.

Hecht-Nielsen, R. et al. (1988). Theory of the backpropagation neural network. *Neural Networks*, **1** 445–448.

Horowitz, J. L. (2000). *Semiparametric and Nonparametric Methods in Econometrics*, vol. 692. Springer.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*, **58** 71–120.

Jaganathan, K., Oymak, S. and Hassibi, B. (2012). On robust phase retrieval for sparse signals. In *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*. IEEE.

Jain, P., Tewari, A. and Kar, P. (2014). On iterative hard thresholding methods for high-dimensional M-estimation. In *Advances in Neural Information Processing Systems*.

Kakade, S. M., Kanade, V., Shamir, O. and Kalai, A. (2011). Efficient learning of generalized linear and single index models with isotonic regression. In *Advances in Neural Information Processing Systems*.

Kalai, A. T. and Sastry, R. (2009). The isotron algorithm: High-dimensional isotonic regression. In *Conference on Learning Theory*.

Krahmer, F. and Liu, Y.-K. (2016). Phase retrieval without small-ball probability assumptions. *arXiv preprint arXiv:1604.07281*.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, **86** 316–327.

Li, K.-C. (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *Journal of the American Statistical Association*, **87** 1025–1039.

Li, X. and Voroninski, V. (2013). Sparse signal recovery from quadratic measurements via convex programming. *SIAM Journal on Mathematical Analysis*, **45** 3019–3033.

Liu, J., Ye, J. and Fujimaki, R. (2014). Forward-backward greedy algorithms for general convex smooth functions over a cardinality constraint. In *International Conference on Machine Learning*.

Loh, P.-L. and Wainwright, M. J. (2015). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, **16** 559–616.

Ma, C., Wang, K., Chi, Y. and Chen, Y. (2017). Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion and blind deconvolution. *arXiv preprint arXiv:1711.10467*.

McCullagh, P., Nelder, J. A. and McCullagh, P. (1989). *Generalized linear models*, vol. 2. Chapman and Hall London.

Needell, D. and Tropp, J. A. (2009). Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, **26** 301–321.

Negahban, S. and Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, **39** 1069–1097.

Negahban, S., Yu, B., Wainwright, M. J. and Ravikumar, P. K. (2009). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems*.

Negahban, S. N., Ravikumar, P., Wainwright, M. J. and Yu, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, **27** 538–557.

Neykov, M., Liu, J. S. and Cai, T. (2016a). L1-regularized least squares for support recovery of high dimensional single index models with Gaussian designs. *Journal of Machine Learning Research*, **17** 1–37.

Neykov, M., Wang, Z. and Liu, H. (2016b). Agnostic estimation for misspecified phase retrieval models. In *Advances in Neural Information Processing Systems*.

Ohlsson, H. and Eldar, Y. C. (2014). On conditions for uniqueness in sparse phase retrieval. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE.

Ohlsson, H., Yang, A., Dong, R. and Sastry, S. (2012). Compressive phase retrieval from squared output measurements via semidefinite programming. In *IFAC Symposium on System Identification*.

Oymak, S. and Soltanolkotabi, M. (2016). Fast and reliable parameter estimation from nonlinear observations. *arXiv preprint arXiv:1610.07108*.

Plan, Y. and Vershynin, R. (2016). The generalized Lasso with non-linear observations. *IEEE Transactions on information theory*, **62** 1528–1537.

Plan, Y., Vershynin, R. and Yudovina, E. (2017). High-dimensional estimation with geometric constraints. *Information and Inference: A Journal of the IMA*, **6** 1–40.

Raskutti, G., Wainwright, M. J. and Yu, B. (2010). Restricted eigenvalue properties for correlated Gaussian designs. *The Journal of Machine Learning Research*, **11** 2241–2259.

Raskutti, G., Wainwright, M. J. and Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, **10** 6976–6994.

Rauhut, H., Schneider, R. and Stojanac, Z. (2016). Low rank tensor recovery via iterative hard thresholding. *arXiv preprint arXiv:1602.05217*.

Recht, B., Fazel, M. and Parrilo, P. A. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, **52** 471–501.

Shalev-Shwartz, S., Srebro, N. and Zhang, T. (2010). Trading accuracy for sparsity in optimization problems with sparsity constraints. *SIAM Journal on Optimization*, **20** 2807–2832.

Sherman, R. P. (1994). U-processes in the analysis of a generalized semiparametric regression estimator. *Econometric theory*, **10** 372–395.

Soltanolkotabi, M. (2017). Structured signal recovery from quadratic measurements: Breaking sample complexity barriers via nonconvex optimization. *arXiv preprint arXiv:1702.06175*.

Tanner, J. and Wei, K. (2013). Normalized iterative hard thresholding for matrix completion. *SIAM Journal on Scientific Computing*, **35** S104–S125.

Thrampoulidis, C., Abbasi, E. and Hassibi, B. (2015). Lasso with non-linear measurements is equivalent to one with linear measurements. In *Advances in Neural Information Processing Systems*.

Toh, K.-C. and Yun, S. (2010). An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pacific Journal of Optimization*, **6** 15.

Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, **12** 389–434.

Tu, S., Boczar, R., Simchowitz, M., Soltanolkotabi, M. and Recht, B. (2016). Low-rank solutions of linear matrix equations via Procrustes flow. In *International Conference on Machine Learning*.

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

Waldspurger, I., dAspremont, A. and Mallat, S. (2015). Phase recovery, maxcut and complex semidefinite programming. *Mathematical Programming*, **149** 47–81.

Wang, Z., Liu, H. and Zhang, T. (2014). Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *The Annals of Statistics*, **42** 2164–2201.

Xia, Y. and Li, W. (1999). On single-index coefficient regression models. *Journal of the American Statistical Association*, **94** 1275–1285.

Xiao, L. and Zhang, T. (2013). A proximal-gradient homotopy method for the sparse least-squares problem. *SIAM Journal on Optimization*, **23** 1062–1091.

Xu, W., Wang, M., Cai, J. and Tang, A. (2011). Sparse recovery from nonlinear measurements with applications in bad data detection for power networks. *arXiv preprint arXiv:1112.6234*.

Yang, Z., Balasubramanian, K. and Liu, H. (2017a). High-dimensional non-Gaussian single index models via thresholded score function estimation. In *International Conference on Machine Learning*.

Yang, Z., Balasubramanian, K., Wang, Z. and Liu, H. (2017b). Learning non-Gaussian multi-index model via second-order steins method. In *Advances in Neural Information Processing Systems*.

Yi, X., Wang, Z., Caramanis, C. and Liu, H. (2015). Optimal linear estimation under unknown nonlinear transform. In *Advances in Neural Information Processing Systems*.

Yuan, X., Li, P. and Zhang, T. (2014). Gradient hard thresholding pursuit for sparsity-constrained optimization. In *International Conference on Machine Learning*.

Zhang, T. (2010). Analysis of multi-stage convex relaxation for sparse regularization. *The Journal of Machine Learning Research*, **11** 1081–1107.