Robust Structured Estimation with Single-Index Models

Sheng Chen [1] **Arindam Banerjee** [1]

Abstract

In this paper, we investigate general single-index models (SIMs) in high dimensions. Based on U-statistics, we propose two types of robust estimators for the recovery of model parameters, which can be viewed as generalizations of several existing algorithms for one-bit compressed sensing (1-bit CS). With minimal assumption on noise, the statistical guarantees are established for the generalized estimators under suitable conditions, which allow general structures of underlying parameter. Moreover, the proposed estimator is novelly instantiated for SIMs with monotone transfer function, and the obtained estimator can better leverage the monotonicity. Experimental results are provided to support our theoretical analyses.

1. Introduction

In machine learning and statistics, a linear model of the form $y = \langle \boldsymbol{\theta}^*, \mathbf{x} \rangle + \epsilon$ is widely used to find the relationship between feature and response, which has gained overwhelming popularity for a very long time. Here $y \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^p$ is the pair of observed response and feature/measurement vector, ϵ is a zero-mean noise, and $\boldsymbol{\theta}^* \in \mathbb{R}^p$ is the unknown parameter to be estimated. The simplicity of linear model leads to its great interpretability and computational efficiency, which are often favored in practical applications. On theoretical side, even in high-dimensional regime where sample size is smaller than the problem dimension p, strong statistical guarantees have been established under mild assumptions for various estimators, such as Lasso (Tibshirani, 1996) and Dantzig selector (Candes & Tao, 2007). Despite its attractive merits, one main drawback of linear models is the stringent assumption of linear relationship between $\mathbf{x}$ and y, which may fail to hold in complicated scenarios. To introduce more flexibility, one option is to consider the general single-index models (SIMs) (Ichimura, 1993; Horowitz & Hardle, 1996),

$$\mathbb{E}[y|\mathbf{x}] = f^*(\langle \boldsymbol{\theta}^*, \mathbf{x} \rangle) , \tag{1}$$

where $f^* : \mathbb{R} \mapsto \mathbb{R}$ is an *unknown* univariate transfer function (a.k.a. link function). This class of models enjoys rich modeling power in the sense that it encompasses several useful models as special cases, which are briefly described below:

- **One-bit Compressed Sensing**: In one-bit compressed sensing (1-bit CS) (Boufounos & Baraniuk, 2008; Plan & Vershynin, 2013), the response y is restricted to be binary, i.e., $y \in \{+1, -1\}$, and the range of transfer function f^* is $[-1, 1]$. Given the measurement vector $\mathbf{x}$, one can generate y from the Bernoulli model,

  $$\frac{y+1}{2} \sim \text{Ber}\left(\frac{f^*(\langle \boldsymbol{\theta}^*, \mathbf{x} \rangle) + 1}{2}\right) . \tag{2}$$

  In the noiseless case, $f^*(z) = \text{sign}(z)$ and y always reflects the true sign of $\langle \boldsymbol{\theta}^*, \mathbf{x} \rangle$, while y can be incorrect for other f^* whose shape determines the noise level in some way.
- **Generalized Linear Models**: In generalized linear models (GLMs) (McCullagh, 1984), the transfer function is assumed to be *monotonically increasing* and conditional distribution of $y|\mathbf{x}$ belongs to exponential family. Different choices of f^* give rise to different members in GLMs. If f^* is identity function $f^*(z) = z$, one has the simple linear models, while the sigmoid function $f^*(z) = \frac{1}{1+e^{-z}}$ results in the logistic model for binary classification. In this work, however, we have *no access* to exact f^* other than knowing it is monotonic.
- **Noise in Monotone Transfer**: Instead of having the general expectation form of y as GLMs, one could directly introduce the noise inside monotone transfer $\tilde{f}$ to model the randomness of y (Plan et al., 2016),

  $$y = \tilde{f}\left(\langle \boldsymbol{\theta}^*, \mathbf{x} \rangle + \epsilon\right) . \tag{3}$$

  In this setting, the transfer function $\tilde{f}$ is slightly different from the f^* in (1), which are related by $f^*(z) = \mathbb{E}_\epsilon[\tilde{f}(z + \epsilon)|z]$.

[1] Department of Computer Science & Engineering, University of Minnesota-Twin Cities, Minnesota, USA. Correspondence to: Sheng Chen <shengc@cs.umn.edu>, Arindam Banerjee <banerjee@cs.umn.edu>.

Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017. Copyright 2017 by the author(s).

A key advantage of SIM is its robustness. First, allowing unknown f^* prevents the mis-specification of transfer function, which could otherwise lead to a poor estimate of $\boldsymbol{\theta}^*$. Secondly, the model in (1) makes minimal assumption on the distribution of y, thus being able to tolerate potentially heavy-tailed noise.

In order to estimate $\boldsymbol{\theta}^*$, we are given n measurements of $(\mathbf{x}, y) \in \mathbb{R}^p \times \mathbb{R}$, denoted by $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$. In this work, we focus on the $n < p$ regime. In such high-dimensional setting, the recovery of $\boldsymbol{\theta}^*$ is quite challenging as the problem is ill-posed even when f^* is given. Over the last decade, substantial progress has been made to address the challenge by exploiting the apriori structure of parameter $\boldsymbol{\theta}^*$, like sparsity (Tibshirani, 1996). For simple linear models or GLMs with known transfer, extensive studies have shown that sparse $\boldsymbol{\theta}^*$ can be consistently estimated under mild assumptions, with much lower sample complexity than p (Candes & Tao, 2007; Wainwright, 2009; Bickel et al., 2009; Kakade et al., 2010; Negahban et al., 2012; Yang et al., 2016). Recently the notion of structure has been suitably generalized beyond the unstructured sparsity (Bach et al., 2012), and *Gaussian width* (Gordon, 1985) has emerged as a useful measure to characterize the structural complexity which further determines the recovery guarantee of $\boldsymbol{\theta}^*$ (Chandrasekaran et al., 2012; Rao et al., 2012; Oymak et al., 2013; Amelunxen et al., 2014; Banerjee et al., 2014; Chatterjee et al., 2014; Vershynin, 2015; Tropp, 2015; Chen & Banerjee, 2016).

In the absence of exact f^*, though 1-bit CS and related variants were well-studied in recent years (Boufounos & Baraniuk, 2008; Jacques et al., 2013; Plan & Vershynin, 2013; Gopi et al., 2013; Zhang et al., 2014; Chen & Banerjee, 2015a; Zhu & Gu, 2015; Yi et al., 2015; Slawski & Li, 2015; Li, 2016; Slawski & Li, 2016), the exploration of general SIMs or the cases with monotone transfers is relatively limited, especially in the high-dimensional regime. Kalai & Sastry (2009) and Kakade et al. (2011) investigated the low-dimensional SIMs with monotone transfers, and they proposed perceptron-type algorithms to estimate both f^* and $\boldsymbol{\theta}^*$, with provable guarantees on prediction error. In high dimension, general SIMs were studied by Alquier & Biau (2013) and Radchenko (2015), in which only unstructured sparsity of $\boldsymbol{\theta}^*$ is considered. The algorithm developed in (Alquier & Biau, 2013) relies on reversible jump MCMC, which could be slow. In Radchenko (2015), a path fitting algorithm is designed to recover f^* and $\boldsymbol{\theta}^*$, but only asymptotic guarantees are provided. Ganti et al. (2015) considered the high-dimensional setting with monotone transfer, and their iterative algorithm is based on non-convex optimization, for which it is hard to establish the convergence. Besides, the prediction error bound they derived is also weak (in the sense that it is even worse than the initialization of the algorithm). Recently Oymak & Soltanolkotabi (2016) proposed a constrained least-squares method to estimate $\boldsymbol{\theta}^*$, with recovery error characterized by Gaussian width and related quantities. Though their analysis considered the general structure of $\boldsymbol{\theta}^*$, it only holds for noiseless setting where $y = f(\langle \boldsymbol{\theta}^*, \mathbf{x} \rangle)$. General structure of $\boldsymbol{\theta}^*$ was also explored in Vershynin (2015) and Plan et al. (2016). Other types of statistical guarantees for high-dimensional SIMs is also available, such as support recovery of $\boldsymbol{\theta}^*$ in Neykov et al. (2016). It is worth noting that all the aforementioned statistical analyses rely on sub-Gaussian noise or the transfer function being bounded or Lipschitz, which indicates that none of the results can immediately hold for heavy-tailed noise (or without Lipschitzness and boundedness).

In this paper, we focus on the parameter estimation of $\boldsymbol{\theta}^*$ instead of the prediction of y given new $\mathbf{x}$. In particular, we propose two families of generalized estimators, constrained and regularized, for model (1) under Gaussian measurement. The parameter $\boldsymbol{\theta}^*$ is assumed to possess certain low-complexity structure, which can be either captured by a constraint $\boldsymbol{\theta}^* \in \mathcal{K}$ or a norm regularization term $\|\boldsymbol{\theta}^*\|$. Our general approach is inspired by U-statistics and the advances in 1-bit CS, and subsumes several existing 1-bit CS algorithms as special cases. Similar to those algorithms, our estimator is simple and often admits closed-form solutions. Regarding the recovery analysis, there are two appealing aspects. First our results work for general structure, with error bound characterized by Gaussian width and some other easy-to-compute geometric measures. Instantiating our results with specific structure of $\boldsymbol{\theta}^*$ recovers previously established error bounds for 1-bit CS (Zhang et al., 2014; Chen & Banerjee, 2015a), which are sharper than those yielded by the general analysis in Plan & Vershynin (2013). Second, our analysis works with limited assumptions on the condition distribution of y. In particular, our estimator is robust to heavy-tailed noise and permit unbounded transfer functions f^* as well as non-Lipschitz ones. At the heart of our analysis is the *generic chaining* method (Talagrand, 2014), an advanced tool in probability theory, which has been successfully applied to sparse recovery (Koltchinskii, 2011) and dimensionality reduction (Dirksen, 2016), etc. Another key ingredient in our proof is a Hoeffding-type concentration inequality for U-statistics (Lee, 1990) with sub-Gaussian tails, which is less known yet generalizes the popular one for bounded U-statistics (Hoeffding, 1963). Apart from 1-bit CS, we particularly investigate the model (3), for which the generalized estimator is specialized in a novel way. The resulting estimator better leverages the monotonicity of the transfer function, which is also demonstrated through experiments. For the ease of exposition, whenever we say "monotone", it means "monotonically increasing" by default. Throughout the paper, we will use c, C, C', C_0, C_1 and so on to denote absolute constants, which may differ from context to context. Detailed proofs are deferred to the supplementary material due to page limit.

The rest of the paper is organized as follows. In Section 2, we introduce our estimators for SIMs along with their recovery guarantees. We also provide a few examples in 1-bit CS for illustration. Section 3 is focused on model (3), for which we instantiate the general results in a new way. Other structures of $\boldsymbol{\theta}^*$ beyond unstructured sparsity are also discussed. Section 4 provides the proof of our main results and the related lemmas. In Section 5, we complement our theoretical developments with some experiment results. The final section is dedicated to conclusions.

2. Generalized Estimation for Structured Parameter

2.1. Assumptions and Preliminaries

For the sake of identifiability, we assume w.l.o.g. that $\|\boldsymbol{\theta}^*\|_2 = 1$ throughout the paper. At the first glimpse of model (1), we may realize that it is difficult to recover $\boldsymbol{\theta}^*$ due to unknown f^*. In contrast, when f^* is given, the recovery guarantees of $\boldsymbol{\theta}^*$ can be established under mild assumptions of $\mathbf{x}$ and y, such as boundedness or sub-Gaussianity. If we know certain properties of the transfer function like the monotonicity introduced in GLMs and (3), the structure of f^* is largely restricted, and it is tempting to expect that similar results will continue to hold. Unfortunately, we first have the following claim, which indicates that without other constraints on f^* beyond strict monotonicity, $\boldsymbol{\theta}^*$ cannot be consistently estimated under general sub-Gaussian (or bounded) measurement, even in the noiseless setting of (3).

Claim 1 *Suppose that each element x_i of $\mathbf{x}$ is sampled i.i.d. from Rademacher distribution, i.e., $\mathbb{P}(x_i = 1) = \mathbb{P}(x_i = -1) = 0.5$. Under model (3) with noise $\epsilon = 0$, there exists a $\bar{\boldsymbol{\theta}} \in \mathbb{S}^{p-1}$ together with a monotone $\bar{f}$, such that $\operatorname{supp}(\bar{\boldsymbol{\theta}}) = \operatorname{supp}(\boldsymbol{\theta}^*)$ and $y_i = \bar{f}(\langle \bar{\boldsymbol{\theta}}, \mathbf{x}_i \rangle)$ for data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ with arbitrarily large sample size n, while $\|\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 > \delta$ for some constant δ.*

Now that consistent estimation of $\boldsymbol{\theta}^*$ is not possible for general sub-Gaussian measurement, it might be reasonable to focus on certain special cases. For this work, we assume that $\mathbf{x}$ is standard Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$. For SIM (1), we additionally assume that the distribution of y depends on $\mathbf{x}$ *only* through the value of $\langle \boldsymbol{\theta}^*, \mathbf{x} \rangle$, i.e., the distribution of $y|\mathbf{x}$ is fixed if $\langle \boldsymbol{\theta}^*, \mathbf{x} \rangle$ is given (no matter what the exact $\mathbf{x}$ is). This assumption is quite minimal, and it turns out that the examples we provide in Section 1 all satisfy it (if noise ϵ is independent of $\mathbf{x}$ in (3)). The same assumption is used in Plan et al. (2016) as well.

Under the assumptions above, given m i.i.d. observations $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$, we define

$$\mathbf{u}\left((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\right) = \sum_{i=1}^{m} q_i\left(y_1, \ldots, y_m\right) \cdot \mathbf{x}_i \ , \tag{4}$$

where all $q_i : \mathbb{R}^m \mapsto \mathbb{R}$ are bounded functions with $|q_i| \leq 1$, which are chosen along with m based on the properties of the transfer function. In Section 2.4 and 3.1, we will see examples for their choices. The vector $\mathbf{u} \in \mathbb{R}^p$ is critical due to the key observation below.

Lemma 1 *Suppose the distribution of y in model (1) depends on $\mathbf{x}$ through $\langle \boldsymbol{\theta}^*, \mathbf{x} \rangle$ and we define accordingly*

$$b_i\left(z_1, \ldots, z_m; \boldsymbol{\theta}^*\right) = \tag{5}$$
$$\mathbb{E}\left[q_i\left(y_1, \ldots, y_m\right) | \langle \boldsymbol{\theta}^*, \mathbf{x}_1 \rangle = z_1, \ldots, \langle \boldsymbol{\theta}^*, \mathbf{x}_m \rangle = z_m\right] \ .$$

With $\mathbf{x}$ being standard Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$, $\mathbf{u}$ defined in (4) satisfies

$$\mathbb{E}\left[\mathbf{u}\left((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\right)\right] = \beta \boldsymbol{\theta}^* \ , \tag{6}$$

where $\beta = \sum_{i=1}^m \mathbb{E}[b_i\left(g_1, \ldots, g_m; \boldsymbol{\theta}^*\right) \cdot g_i]$, and $g_1, \ldots, g_m$ are i.i.d. standard Gaussian.

Note that Lemma 1 is true for all choices of q_i, and the proof is given in the supplement. This lemma presents an insight towards the design of our estimator, that is, the direction of $\boldsymbol{\theta}^*$ can be approximated if we have a good sense about $\mathbb{E}\mathbf{u}$. As we will see in the sequel, the scalar β plays a key role in the estimation error bound, which can give us clues to the choice of q_i. We can assume w.l.o.g. that $\beta \geq 0$ since we can flip the sign of each q_i.

The recovery analysis is built on the notion of Gaussian width (Gordon, 1985), which is defined for any $\mathcal{A} \subseteq \mathbb{R}^p$ as $w(\mathcal{A}) = \mathbb{E}[\sup_{\mathbf{v} \in \mathcal{A}} \langle \mathbf{g}, \mathbf{v} \rangle]$, where $\mathbf{g}$ is a standard Gaussian random vector. Roughly speaking, $w(\mathcal{A})$ measures the scaled width of set $\mathcal{A}$ averaged over each direction.

2.2. Generalized Estimator

Inspired by Lemma 1, we define the vector $\hat{\mathbf{u}}$ for the observed data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$,

$$\hat{\mathbf{u}} = \frac{(n-m)!}{n!} \sum_{\substack{1 \leq i_1, \ldots, i_m \leq n \\ i_1 \neq \ldots \neq i_m}} \mathbf{u}\left((\mathbf{x}_{i_1}, y_{i_1}), \ldots, (\mathbf{x}_{i_m}, y_{i_m})\right) \ , \tag{7}$$

which is an unbiased estimator of $\mathbb{E}\mathbf{u}$, meaning that $\mathbb{E}\hat{\mathbf{u}} = \mathbb{E}\mathbf{u} = \beta \boldsymbol{\theta}^*$. When $m = 2$, we essentially have

$$\hat{\mathbf{u}} = \frac{1}{n(n-1)} \sum_{\substack{1 \leq i, j \leq n \\ i \neq j}} \mathbf{u}\left((\mathbf{x}_i, y_i), (\mathbf{x}_j, y_j)\right) \tag{8}$$

In fact, $\hat{\mathbf{u}}$ can be treated as a vector version of U-statistics with order m. Given $\hat{\mathbf{u}}$, a naive way to estimate $\boldsymbol{\theta}^*$ is to simply normalize $\hat{\mathbf{u}}$, i.e., $\hat{\boldsymbol{\theta}} = \hat{\mathbf{u}}/\|\hat{\mathbf{u}}\|_2$. In high-dimensional setting, $\boldsymbol{\theta}^*$ is often structured, but the naive estimator fails to take such information into account, which would lead to large error. To incorporate the prior knowledge on $\boldsymbol{\theta}^*$, we design two types of estimator, the constrained one and the regularized one.

Constrained Estimator: If we assume that $\boldsymbol{\theta}^*$ belongs to some structured set $\mathcal{K} \subseteq \mathbb{S}^{p-1}$, then the estimation of $\boldsymbol{\theta}^*$ is carried out via the constrained optimization

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\operatorname{argmin}} \; -\langle \hat{\mathbf{u}}, \boldsymbol{\theta}\rangle \quad \text{s.t.} \quad \boldsymbol{\theta} \in \mathcal{K} \, . \tag{9}$$

Here the set $\mathcal{K}$ can be non-convex, as long as the optimization can be solved globally. Since the objective function is very simple, we can often end up with a global minimizer. Similar estimator has been used in Plan et al. (2016), but they only focused on specific $\hat{\mathbf{u}}$.

Regularized Estimator: If we assume that the structure of $\boldsymbol{\theta}^*$ can be captured by certain norm $\|\cdot\|$, we may alternatively use the regularized estimator to find $\boldsymbol{\theta}^*$,

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\operatorname{argmin}} \; -\langle \hat{\mathbf{u}}, \boldsymbol{\theta}\rangle + \lambda\|\boldsymbol{\theta}\| \quad \text{s.t.} \quad \|\boldsymbol{\theta}\|_2 \le 1 \, . \tag{10}$$

The optimization is convex, thus the global minimum is always attained. Previously this estimator was used in 1-bit CS scenario with L_1 norm (Zhang et al., 2014).

2.3. Recovery Analysis

Regarding the constrained estimator, the recovery of $\boldsymbol{\theta}^*$ relies on the geometry of $\hat{\boldsymbol{\theta}}$, which is described by

$$\mathcal{A}_{\mathcal{K}}(\boldsymbol{\theta}^*) = \operatorname{cone}\left\{\mathbf{v} \;\middle|\; \mathbf{v} = \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*, \; \hat{\boldsymbol{\theta}} \in \mathcal{K}\right\} \bigcap \mathbb{S}^{p-1} \tag{11}$$

The set $\mathcal{A}_{\mathcal{K}}(\boldsymbol{\theta}^*)$ essentially contains all possible directions that error $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*$ could lie in. The following theorem characterizes the error of $\hat{\boldsymbol{\theta}}$.

Theorem 1 *Suppose that the optimization (9) can be solved to global minimum. Then the following error bound holds for the minimizer $\hat{\boldsymbol{\theta}}$ with probability at least $1 - C'' \exp\left(-w^2\left(\mathcal{A}_{\mathcal{K}}(\boldsymbol{\theta}^*)\right)\right)$,*

$$\left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|_2 \le \frac{C\kappa m^{\frac{3}{2}}}{\beta} \cdot \frac{w(\mathcal{A}_{\mathcal{K}}(\boldsymbol{\theta}^*)) + C'}{\sqrt{n}} \, , \tag{12}$$

where κ is the sub-Gaussian norm of a standard Gaussian random variable, and C, C', C'' are all absolute constant.

Remark: Note that estimator is consistent as long as $\beta \neq 0$. The error bound inversely depends on the scale of β, which implies that we should construct suitable q_i such that β is large according to its definition in Lemma 1. The choice of q_i further depends on the assumed property of f^*. Though dependency on m may prevent us from using higher-order $\mathbf{u}$, m is typically small in practice and can be treated as constant.

For regularized estimator, we can similarly establish the recovery guarantee in terms of Gaussian width.

Theorem 2 *Define the following set for any $\rho > 1$,*

$$\mathcal{A}_\rho\left(\boldsymbol{\theta}^*\right) = \operatorname{cone}\left\{\mathbf{v} \;\middle|\; \|\mathbf{v} + \boldsymbol{\theta}^*\| \le \|\boldsymbol{\theta}^*\| + \frac{\|\mathbf{v}\|}{\rho}\right\} \bigcap \mathbb{S}^{p-1}$$

If we set $\lambda = \rho\left\|\hat{\mathbf{u}} - \beta\boldsymbol{\theta}^*\right\|_* = O(\rho m^{3/2} w(\mathcal{B}_{\|\cdot\|})/\sqrt{n})$ and it satisfies $\lambda < \|\hat{\mathbf{u}}\|_*$, then with probability at least $1 - C' \exp\left(-w^2\left(\mathcal{B}_{\|\cdot\|}\right)\right)$, $\hat{\boldsymbol{\theta}}$ in (10) satisfies

$$\left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|_2 \le \frac{C(1+\rho)\kappa m^{\frac{3}{2}}}{\beta} \cdot \frac{\Psi\left(\mathcal{A}_\rho(\boldsymbol{\theta}^*)\right) \cdot w\left(\mathcal{B}_{\|\cdot\|}\right)}{\sqrt{n}} \, , \tag{13}$$

where $\Psi\left(\mathcal{A}_\rho(\boldsymbol{\theta}^*)\right) = \sup_{\mathbf{v}\in\mathcal{A}_\rho(\boldsymbol{\theta}^*)} \|\mathbf{v}\|$ and $\mathcal{B}_{\|\cdot\|} = \{\mathbf{v} \mid \|\mathbf{v}\| \le 1\}$ is the unit ball of norm $\|\cdot\|$.

Remark: The geometry of the regularized estimator is slightly different from the constrained one. Instead of having $\mathcal{A}_{\mathcal{K}}(\boldsymbol{\theta}^*)$, here the set $\mathcal{A}_\rho(\boldsymbol{\theta}^*)$ depends on the choice of the regularization parameter λ. The same phenomenon also appears in the (Banerjee et al., 2014). The geometric measure $\Psi\left(\mathcal{A}_\rho(\boldsymbol{\theta}^*)\right)$ is called *restricted norm compatibility*, which is non-random. For many interesting cases, it is easy to calculate (Negahban et al., 2012; Chen & Banerjee, 2015b).

2.4. Application to 1-bit CS

For 1-bit CS problem (2), the $\mathbf{u}$ defined in (4) can be chosen with $m = 1$ and $q_i = y_i$, ending up with

$$\mathbf{u}\left((\mathbf{x}, y)\right) = y\mathbf{x} \quad \text{and} \quad \hat{\mathbf{u}} = \frac{1}{n}\sum_{i=1}^{n} y_i \mathbf{x}_i \, . \tag{14}$$

By such choice of $\mathbf{u}$, the β defined in Lemma 1 is simply $\beta = \mathbb{E}[f^*(g)g]$ with g being standard Gaussian random vector. Under reasonably mild noise, y is likely to take the sign of the linear measurement, which means that $f^*(g)$ should be close to 1 (or -1) if g is positive (or negative). Thus we expect $f^*(g)g$ to be positive most of time and β to be large. Given the choice of $\mathbf{u}$, we can specialize our generalized constrained/regularized estimator to obtain previous results. If $\boldsymbol{\theta}^*$ is assumed to be s-sparse, for constrained estimator, we can choose a straightforward $\mathcal{K} = \{\boldsymbol{\theta} \mid \|\boldsymbol{\theta}\|_0 \le s\} \cap \mathbb{S}^{p-1}$, which results in the k-support norm estimator (Chen & Banerjee, 2015a),

$$\hat{\boldsymbol{\theta}}^{\text{ks}} = \underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\operatorname{argmin}} \; -\langle \hat{\mathbf{u}}, \boldsymbol{\theta}\rangle \;\; \text{s.t.} \;\; \|\boldsymbol{\theta}\|_0 \le s, \|\boldsymbol{\theta}\|_2 = 1 \tag{15}$$

Though $\mathcal{K}$ is non-convex, the global minimizer can actually be obtained in closed form,

$$\hat{\theta}_j^{\text{ks}} = \begin{cases} \hat{u}_j \; / \; \| |\hat{\mathbf{u}}|^{\downarrow}_{1:s} \|_2 \; , & \text{if } |\hat{u}_j| \text{ is in } |\hat{\mathbf{u}}|^{\downarrow}_{1:s} \\ 0 \; , & \text{otherwise} \end{cases} \tag{16}$$

where $|\hat{\mathbf{u}}|^{\downarrow}$ is the absolute-value counterpart of $\hat{\mathbf{u}}$ with entries sorted in descending order, and the subscript takes the top s entries. Similarly if the regularized estimator is instantiated with L_1 norm $\|\cdot\|_1$, we obtain the so-called passive algorithm introduced in Zhang et al. (2014),

$$\hat{\boldsymbol{\theta}}^{\text{ps}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\text{argmin}} \; - \langle \hat{\mathbf{u}}, \boldsymbol{\theta} \rangle + \lambda \|\boldsymbol{\theta}\|_1 \; \text{ s.t. } \; \|\boldsymbol{\theta}\|_2 \leq 1 \; , \tag{17}$$

whose solution is given by $\hat{\boldsymbol{\theta}}^{\text{ps}} = S(\hat{\mathbf{u}}, \lambda) / \|S(\hat{\mathbf{u}}, \lambda)\|_2$, where $S(\cdot, \cdot)$ is the elementwise soft-thresholding operator, $S_i(\hat{\mathbf{u}}, \lambda) = \max\{\text{sign}(\hat{u}_i)(|\hat{u}_i| - \lambda), 0\}$. Based on Theorem 1 and 2, we can easily obtain the error bound for both k-support norm estimator and passive algorithm.

Corollary 1 *Assume that $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ follow 1-bit CS model in (2) and $\hat{\mathbf{u}}$ is given as (14). For any s-sparse $\boldsymbol{\theta}^*$, with high probability, $\hat{\boldsymbol{\theta}}$ produced by both (15) and (17) (i.e., $\hat{\boldsymbol{\theta}}^{ks}$ and $\hat{\boldsymbol{\theta}}^{ps}$) satisfy*

$$\left\| \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \right\|_2 \leq O\left(\sqrt{\frac{s \log p}{n}} \right) \tag{18}$$

The proof is included in the supplementary material. The above result was shown by Slawski & Li (2015) and Zhang et al. (2014), but their analyses do not consider the general structure. Compared with $O(\sqrt[4]{s \log p / n})$ yielded by the general result in Plan & Vershynin (2013), our bound is much sharper.

3. A New Estimator for Monotone Transfer

In this section, we specifically study model (3). Here we further assume that $\tilde{f}$ is *strictly* increasing. What is worth mentioning is that the estimator we develop here can be applied to GLMs as well. To avoid the confusion with $\mathbf{u}$ and $\hat{\mathbf{u}}$ defined previously, we instead use new notations $\mathbf{h}$ and $\hat{\mathbf{h}}$ respectively in this section.

3.1. Estimator with Second-Order $\hat{\mathbf{h}}$

To motivate the design of $\mathbf{h}$, it is helpful to rewrite model (3) by applying the inverse of $\tilde{f}$ on both sides,

$$\tilde{f}^{-1}(y) = \langle \boldsymbol{\theta}^*, \mathbf{x} \rangle + \epsilon \; . \tag{19}$$

Note that the new formulation resembles the linear model except that we have no access to the value of $\tilde{f}^{-1}(y)$. Instead, all we know about $\mathbf{r} = [\tilde{f}^{-1}(y_1), \ldots, \tilde{f}^{-1}(y_n)]^T \in \mathbb{R}^n$ is that it preserves the ordering of $\mathbf{y} = [y_1, \ldots, y_n]^T$. Put in another way, $\mathbf{r}$ needs to satisfy the constraint that $r_i > r_j$ iff. $y_i > y_j$ and $r_i < r_j$ iff. $y_i < y_j$. To move one step further, it is equivalent to $\text{sign}(y_i - y_j) = \text{sign}(r_i - r_j) = \text{sign}(\langle \boldsymbol{\theta}^*, \mathbf{x}_i - \mathbf{x}_j \rangle + \epsilon_i - \epsilon_j)$ based on model assumption. Hence the information contained in sample $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ can be interpreted from the perspective of 1-bit CS, where $\text{sign}(y_i - y_j)$ reflects the perturbed sign of linear measurement $\langle \boldsymbol{\theta}^*, \mathbf{x}_i - \mathbf{x}_j \rangle$. Inspired by the $\mathbf{u}$ for 1-bit CS, we may choose $m = 2$ and define $\mathbf{h}$, $\hat{\mathbf{h}}$ as

$$\mathbf{h}\left((\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2)\right) = \text{sign}(y_1 - y_2) \cdot (\mathbf{x}_1 - \mathbf{x}_2) \; , \tag{20}$$

$$\hat{\mathbf{h}} = \frac{1}{n(n-1)} \sum_{\substack{1 \leq i,j \leq n \\ i \neq j}} \mathbf{h}\left((\mathbf{x}_i, y_i), (\mathbf{x}_j, y_j)\right) \; , \tag{21}$$

Given the definition of $\hat{\mathbf{h}}$, Lemma 1 directly implies the following corollary.

Corollary 2 *Suppose that $(\mathbf{x}_1, y_2)$ and $(\mathbf{x}_2, y_2)$ are generated by model (3), where $\mathbf{x}_1, \mathbf{x}_2$ follow Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$, and the noise ϵ_1, ϵ_2 are independent of $\mathbf{x}_1, \mathbf{x}_2$ and identically (but arbitrarily) distributed. Then the expectation of $\mathbf{h}\left((\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2)\right)$ satisfies*

$$\mathbb{E}\left[\mathbf{h}\left((\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2)\right)\right] = \sqrt{2}\beta' \boldsymbol{\theta}^* \; , \tag{22}$$

where $\beta' = \mathbb{E}_{g \sim \mathcal{N}(0,1)}\left[\text{sign}\left(g + (\epsilon_1 - \epsilon_2)/\sqrt{2}\right) \cdot g\right]$.

Remark: The scalar $\sqrt{2}\beta'$ serves as the role of β in Lemma 1, and β' is always guaranteed to be strictly positive regardless how the noise is distributed, which keeps $\boldsymbol{\theta}^*$ distinguishable all the time. To see this, let $\xi = (\epsilon_1 - \epsilon_2)/\sqrt{2}$. Note that ξ is symmetric, thus $\varepsilon\xi$ has the same distribution as ξ, where ε is a Rademacher random variable. Therefore

$$\begin{aligned} \beta' &= \mathbb{E}\left[\text{sign}\left(g + \xi\right) \cdot g\right] = \mathbb{E}_{g,\xi}\mathbb{E}_{\varepsilon}\left[\text{sign}\left(g + \varepsilon\xi\right) \cdot g\right] \\ &= \mathbb{E}_{\xi}\mathbb{E}_{g}\left[\frac{\text{sign}\left(g - \xi\right) + \text{sign}\left(g + \xi\right)}{2} \cdot g\right] \end{aligned}$$

Since $g(g - \xi) + g(g + \xi) = 2g^2 \geq 0$, it follows that $\text{sign}(g(g-\xi)) + \text{sign}(g(g+\xi)) = (\text{sign}(g-\xi) + \text{sign}(g+\xi)) \cdot \text{sign}(g) \geq 0$, thus $(\text{sign}(g - \xi) + \text{sign}(g + \xi)) \cdot g$ is always nonnegative. Find a large enough $M > 0$ such that $\mathbb{P}(|\xi| \leq M) = 0.5 > 0$, and we have

$$\begin{aligned} \beta' &= \mathbb{E}\left[\text{sign}\left(g + \xi\right) \cdot g\right] \geq \mathbb{E}_{\xi}\mathbb{E}_{g}\left[|g| \cdot \mathbb{I}\{|g| > |\xi|\}\right] \\ &\geq 0.5\mathbb{E}_{g}\left[|g| \cdot \mathbb{I}\{|g| > M\}\right] = \frac{M}{2} \cdot \mathbb{P}(|g| > M) > 0 \; . \end{aligned}$$

In the ideal noiseless case, β' achieve its maximum, $\beta'_{\max} = \mathbb{E}[\text{sign}(g)g] = \mathbb{E}[|g|] = \sqrt{2/\pi}$. In the worst case, if ϵ_1 and ϵ_2 are heavy-tailed and dominate g, then $\beta' \approx \mathbb{E}\left[\text{sign}\left((\epsilon_1 - \epsilon_2)/\sqrt{2}\right) \cdot g\right] \approx 0$.

Now we can instantiate the generalized estimator based on $\hat{\mathbf{h}}$. For example, if $\boldsymbol{\theta}^*$ is s-sparse, we estimate it by

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\operatorname{argmin}} \; -\langle \hat{\mathbf{h}}, \boldsymbol{\theta}\rangle \;\text{ s.t. }\; \|\boldsymbol{\theta}\|_0 \leq s, \|\boldsymbol{\theta}\|_2 = 1 \qquad (23)$$

which enjoys $O\left(\sqrt{s \log p/n}\right)$ error rate as shown in Corollary 1. The regularized estimator can also be obtained with the same $\hat{\mathbf{h}}$ according to (17). The bottleneck of computing $\hat{\boldsymbol{\theta}}$ lies in the calculation of $\hat{\mathbf{h}}$. A simple proposition below enables us to get $\hat{\mathbf{h}}$ in a fast manner.

Proposition 1 *Given $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, let $\pi^{\downarrow}$ be the permutation of $\{1,\ldots,n\}$ such that $y_{\pi_1^{\downarrow}} > y_{\pi_2^{\downarrow}} > \ldots > y_{\pi_n^{\downarrow}}$. Then we have*

$$\hat{\mathbf{h}} = \frac{2}{n(n-1)} \sum_{i=1}^{n} (n+1-2i) \cdot \mathbf{x}_{\pi_i^{\downarrow}} \qquad (24)$$

Remark: Based on the proposition above, $\hat{\mathbf{h}}$ can be efficiently computed in $O(np + n\log n)$ time, i.e., $O(n \log n)$ time for sorting $\mathbf{y}$ and $O(np)$ time for the weighted sum of all $\mathbf{x}_i$. This is a significant improvement compared with the the naive calculation using (21), which takes $O(n^2 p)$ time.

3.2. Beyond Unstructured Sparsity

So far we have illustrated the Gaussian width based error bounds, viz (12) and (13), only through unstructured sparsity of $\boldsymbol{\theta}^*$. Here we provide two more examples, non-overlapping group sparsity and fused sparsity.

Non-Overlapping Group Sparsity: Suppose the coordinates of $\boldsymbol{\theta}^*$ has been partitioned into K predefined disjoint groups $\mathcal{G}_1, \ldots, \mathcal{G}_K \subseteq \{1, 2, \ldots, p\}$, out of which only k groups are non-zero. If we use the regularized estimator with $L_{2,1}$ norm $\|\boldsymbol{\theta}\|_{2,1} = \sum_{i=1}^K \|\boldsymbol{\theta}_{\mathcal{G}_i}\|_2$, the optimal solution can be similarly obtained as (17), with elementwise soft-thresholding replaced by the groupwise one. The related geometric measures that appears in (13) can be found in Banerjee et al. (2014), which are given by

$$\Psi\left(\mathcal{A}_\rho(\boldsymbol{\theta}^*)\right) \leq O(\sqrt{k}) \qquad (25)$$

$$w\left(\mathcal{B}_{\|\cdot\|_{2,1}}\right) \leq O(\sqrt{\log K} + \sqrt{G}) \qquad (26)$$

Fused Sparsity: $\boldsymbol{\theta}^*$ is said to be *s-fused-sparse* if the cardinality of the set $\mathcal{F}(\boldsymbol{\theta}^*) = \{1 \leq i < p \mid \theta_i^* \neq \theta_{i+1}^*\}$ is smaller than s. If we resort to the constrained estimator (9) with $\mathcal{K} = \{\boldsymbol{\theta} \mid |\mathcal{F}(\boldsymbol{\theta})| \leq s, \; \|\boldsymbol{\theta}\|_2 = 1\}$, the associated optimization can be solved by dynamic programming (Bellman, 1961). The proposition below upper bounds the corresponding Gaussian width $w(\mathcal{A}_\mathcal{K}(\boldsymbol{\theta}^*))$ in (12).

Proposition 2 *For s-fused-sparse $\boldsymbol{\theta}^*$, the Gaussian width of set $\mathcal{A}_\mathcal{K}(\boldsymbol{\theta}^*)$ with $\mathcal{K} = \{\boldsymbol{\theta} \mid |\mathcal{F}(\boldsymbol{\theta})| \leq s, \; \|\boldsymbol{\theta}\|_2 = 1\}$ satisfies*

$$w(\mathcal{A}_\mathcal{K}(\boldsymbol{\theta}^*)) \leq O(\sqrt{s \log p}) \qquad (27)$$

The proof can be found in (Slawski & Li, 2016), and we provide a different one in supplementary material.

4. Lemmas and Proof Sketch of Theorem 1

Here we first present the important technical lemmas that will be used in the proof of Theorem 1. The first one is the Hoeffding-type inequality for sub-Gaussian U-statistics. In the literature, most of the studies are centered around bounded U-statistics, for which the celebrated concentration is established by Hoeffding (1963). Yet it is not easy to locate the counterpart for sub-Gaussian case. Therefore we provide the following result and attach a proof in the supplementary material.

Lemma 2 (Concentration for sub-Gaussian U-statistics) *Define the U-statistic*

$$U_{n,m}(h) = \frac{(n-m)!}{n!} \sum_{\substack{1 \leq i_1, \ldots, i_m \leq n \\ i_1 \neq i_2 \neq \ldots \neq i_m}} h\left(\mathbf{z}_{i_1}, \ldots, \mathbf{z}_{i_m}\right) \quad (28)$$

with order m and kernel $h : \mathbb{R}^{d\times m} \mapsto \mathbb{R}$ based on n independent copies of random vector $\mathbf{z} \in \mathbb{R}^d$, denoted by $\mathbf{z}_1, \cdots, \mathbf{z}_n$. If $h(\cdot, \ldots, \cdot)$ is sub-Gaussian with $\|h\|_{\psi_2} \leq \kappa$, then the following inequality holds for $U_{n,m}(h)$ with any $\delta > 0$,

$$\mathbb{P}\left(|U_{n,m}(h) - \mathbb{E}U_{n,m}(h)| > \delta\right) \leq 2\exp\left(-C\left\lfloor\frac{n}{m}\right\rfloor \cdot \frac{\delta^2}{\kappa^2}\right), \qquad (29)$$

in which C is an absolute constant.

As mentioned earlier in Section 1, generic chaining is the key tool that our analysis relies on. Specifically we utilize Theorem 2.2.27 from (Talagrand, 2014).

Lemma 3 (Generic chaining concentration) *Given metric space $(\mathcal{T}, s)$, if an associated stochastic process $\{Z_\mathbf{t}\}_{\mathbf{t}\in\mathcal{T}}$ has sub-Gaussian incremental, i.e., satisfies the condition*

$$\mathbb{P}\left(|Z_\mathbf{u} - Z_\mathbf{v}| \geq \delta\right) \;\leq\; C\exp\left(-\frac{C'\delta^2}{s^2(\mathbf{u},\mathbf{v})}\right), \;\forall\, \mathbf{u},\mathbf{v} \in \mathcal{T}, \qquad (30)$$

then the following inequality holds

$$\mathbb{P}\left(\sup_{\mathbf{u},\mathbf{v}\in\mathcal{T}} |Z_\mathbf{u} - Z_\mathbf{v}| \geq C_1\left(\gamma_2(\mathcal{T}, s) + \delta \cdot \operatorname{diam}(\mathcal{T}, s)\right)\right) \leq C_2 \exp\left(-\delta^2\right), \qquad (31)$$

where C, C', C_1 and C_2 are all absolute constants.

The definition of the above γ_2-functional $\gamma_2(\cdot,\cdot)$ is complicated, and is not of great importance. We refer interested reader to the books, Talagrand (2005; 2014). Loosely speaking, $\gamma_2(\mathcal{T}, s)$ can be thought of as a measure for the size of set $\mathcal{T}$ under metric s. What really matters is the following relationship between γ_2-functional and Gaussian width. (see Theorem 2.4.1 in Talagrand (2014))

Lemma 4 (Majorizing measures theorem) *For any set $\mathcal{T} \subseteq \mathbb{R}^p$, the γ_2-functional w.r.t. L_2-metric and Gaussian width satisfy the following inequality with an absolute constant C_0,*

$$\gamma_2\left(\mathcal{T}, \|\cdot\|_2\right) \leq C_0 \cdot w(\mathcal{T}) \tag{32}$$

Equipped with these lemmas, we are ready to present the proof sketch of Theorem 1. A complete proof is deferred to the supplementary material.

Proof Sketch of Theorem 1: We use the shorthand notation $\mathcal{A}_{\mathcal{K}}$ for the set $\mathcal{A}_{\mathcal{K}}(\boldsymbol{\theta}^*)$. As $\hat{\boldsymbol{\theta}}$ attains the global minimum of (9), we have

$$\langle \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*, \hat{\mathbf{u}} \rangle \geq 0 \iff \left\langle \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*, \frac{\hat{\mathbf{u}}}{\beta} - \boldsymbol{\theta}^* + \boldsymbol{\theta}^* \right\rangle \geq 0$$

$$\implies \langle \hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^* \rangle \geq 1 - \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \cdot \sup_{\mathbf{v} \in \mathcal{A}_{\mathcal{K}} \cup \{\mathbf{0}\}} \left\langle \mathbf{v}, \frac{\hat{\mathbf{u}}}{\beta} - \boldsymbol{\theta}^* \right\rangle$$

In order to bound the supremum above, we use the result from generic chaining. We define the stochastic process $\{Z_{\mathbf{v}} = \langle \mathbf{v}, \hat{\mathbf{u}}/\beta - \boldsymbol{\theta}^* \rangle\}_{\mathbf{v} \in \mathcal{A}_{\mathcal{K}} \cup \{\mathbf{0}\}}$. First, we need to check the process has sub-Gaussian incremental. For simplicity, we denote $\mathbf{u}\left((\mathbf{x}_{i_1}, y_{i_1}), \ldots, (\mathbf{x}_{i_m}, y_{i_m})\right)$ by $\mathbf{u}_{i_1,\ldots,i_m}$. By the definitions and properties of sub-Gaussian norm (Vershynin, 2012), it is not difficult to show that $\|\langle \mathbf{u}_{i_1,\ldots,i_m}, \mathbf{v} - \mathbf{w} \rangle\|_{\psi_2} \leq \kappa m \cdot \|\mathbf{v} - \mathbf{w}\|_2$ for any $\mathbf{v}, \mathbf{w} \in \mathcal{A}_{\mathcal{K}} \cup \{\mathbf{0}\}$. By Lemma 2, we have

$$\mathbb{P}\left(|Z_{\mathbf{v}} - Z_{\mathbf{w}}| > \delta\right) \leq 2 \exp\left(-C' \cdot \frac{n\beta^2\delta^2}{m^3\kappa^2 \cdot \|\mathbf{v} - \mathbf{w}\|_2^2}\right) .$$

Therefore we can conclude that $\{Z_{\mathbf{v}}\}$ has sub-Gaussian incremental w.r.t. the metric $s(\mathbf{v}, \mathbf{w}) \triangleq \kappa m^{\frac{3}{2}} \cdot \|\mathbf{v} - \mathbf{w}\|_2/\beta\sqrt{n}$. Now applying Lemma 3 to $\{Z_{\mathbf{v}}\}$ with a bit calculation, we can obtain

$$\mathbb{P}\Bigg(\sup_{\mathbf{v} \in \mathcal{A}_{\mathcal{K}} \cup \{\mathbf{0}\}} |Z_{\mathbf{v}}| \geq \frac{C_1 \kappa m^{\frac{3}{2}}}{\beta\sqrt{n}} \cdot \Big(\gamma_2\left(\mathcal{A}_{\mathcal{K}} \cup \{\mathbf{0}\}, \|\cdot\|_2\right) + 2\delta\Big)\Bigg) \leq C_2 \exp\left(-\delta^2\right)$$

Using $\gamma_2\left(\mathcal{A}_{\mathcal{K}} \cup \{\mathbf{0}\}, \|\cdot\|_2\right) \leq C_0 \cdot w\left(\mathcal{A}_{\mathcal{K}} \cup \{\mathbf{0}\}\right)$ implied by Lemma 4 and taking $\delta = w\left(\mathcal{A}_{\mathcal{K}} \cup \{\mathbf{0}\}\right)$, we get

$$\sup_{\mathbf{v} \in \mathcal{A}_{\mathcal{K}} \cup \{\mathbf{0}\}} \left\langle \mathbf{v}, \frac{\hat{\mathbf{u}}}{\beta} - \boldsymbol{\theta}^* \right\rangle \leq \frac{C_3 \kappa m^{\frac{3}{2}}}{\beta} \cdot \frac{w\left(\mathcal{A}_{\mathcal{K}}\right) + C_4}{\sqrt{n}}$$

with probability at least $1 - C_2 \exp\left(-w^2\left(\mathcal{A}_{\mathcal{K}}\right)\right)$. The inequality also uses the fact that $w\left(\mathcal{A}_{\mathcal{K}} \cup \{\mathbf{0}\}\right) \leq w\left(\mathcal{A}_{\mathcal{K}}\right) + C_4$, which is a result of Lemma 2 in Maurer et al. (2014) (See Lemma A in supplementary material). Lastly we turn to the quantity $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2$,

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \leq \sqrt{2 - 2\langle \hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^* \rangle} \leq \frac{2C_3 \kappa m^{\frac{3}{2}}}{\beta} \cdot \frac{w\left(\mathcal{A}_{\mathcal{K}}\right) + C_4}{\sqrt{n}} .$$

We finish the proof by letting $C = 2C_3$, $C' = C_4$ and $C'' = C_2$.

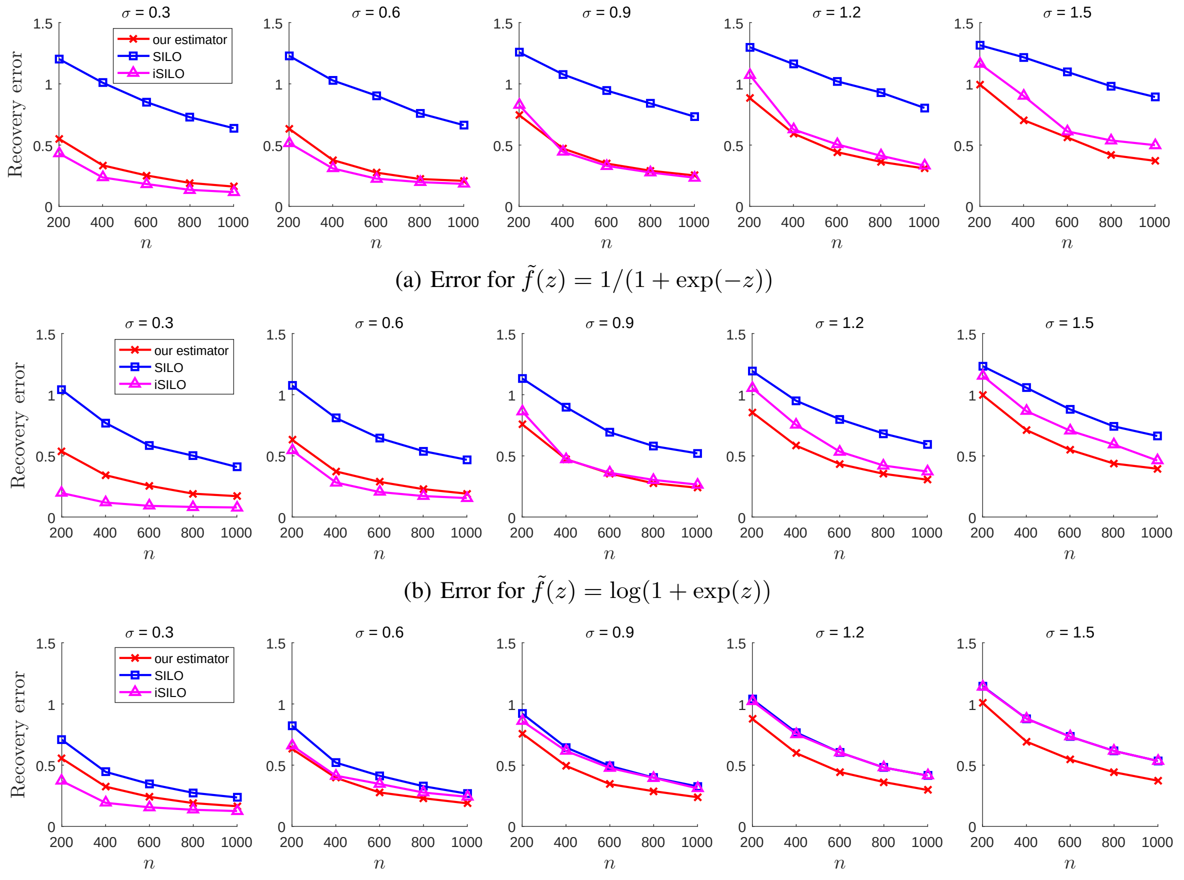
5. Experiment

In the experiment, we focus on model (3) with sparse $\boldsymbol{\theta}^*$. The problem dimension is fixed as $p = 1000$, and the sparsity of $\boldsymbol{\theta}^*$ is set to 10. Essentially we generate our data $(\mathbf{x}, y)$ from

$$y = \tilde{f}\left(\langle \boldsymbol{\theta}^*, \mathbf{x} \rangle + \epsilon\right) ,$$

where $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$. σ ranges from 0.3 to 1.5. We choose three monotonically increasing $\tilde{f}$, $\tilde{f}(z) = 1/(1 + \exp(-z))$ (which is bounded and Lipschitz), $\tilde{f}(z) = z^3$ (which is unbounded and non-Lipschitz), and $\tilde{f}(z) = \log(1 + \exp(z))$ (which is unbounded but Lipschitz). The sample size n varies from 200 to 1000. We use the estimator (23) in Section 3. The baselines we compare with is the SILO and iSILO algorithm introduced in (Ganti et al., 2015). SILO does not quite take the monotonicity in account. In fact, it is the special case of our generalized constrained estimator which uses the same choice of $\mathbf{u}$ as 1-bit CS. The original SILO use the constraint set $\{\boldsymbol{\theta} \mid \|\boldsymbol{\theta}\|_1 \leq \sqrt{s},\ \|\boldsymbol{\theta}\|_2 \leq 1\}$, which is computationally less efficient and statistically no better than $\mathcal{K} = \{\boldsymbol{\theta} \mid \|\boldsymbol{\theta}\|_0 \leq s\} \cap \mathbb{S}^{p-1}$ (Zhang et al., 2014; Chen & Banerjee, 2015a). Hence we also use $\mathcal{K}$ in SILO for a fair comparison. iSILO relies on a specific implementation of isotonic regression which explicitly restricts the Lipschitz constant of $\tilde{f}$ to be one. To fit iSILO into our setting, we remove the Lipschitzness constraint and perform the standard isotonic regression. Since the convergence is not guaranteed for the iterative procedure of iSILO, the number of its iterations is fixed to 100. The best tuning parameter of iSILO is obtained by grid search.

The experiment results are shown in Figure 1. Overall the iSILO algorithm works well under small noise, while our estimator has better performance when the variance of noise increases. To better demonstrate the robustness of our estimator to heavy-tailed noise, instead of Gaussian noise, we sample ϵ from the Student's t distribution with degrees of freedom equal to 3. We repeat the experiments for $\tilde{f}(z) = z^3$, and obtain the plots in Figure 2. We can see that the error of our estimator consistently decreases for all choice of σ as n increases. For SILO and iSILO, the errors are relatively large, and unable to shrink for large σ even when more data are provided.



(a) Error for $\tilde{f}(z) = 1/(1 + \exp(-z))$

(b) Error for $\tilde{f}(z) = \log(1 + \exp(z))$

(c) Error for $\tilde{f}(z) = z^3$

Figure 1. Recovery error vs. sample size (a) Our estimator has similar performance compared with iSILO, both of which outperform SILO by a large margin. (b) iSILO has smaller error when σ is small, while our estimator works better in high-noise regime (c) The error of SILO is reduced compared with other $\tilde{f}$, but iSILO fails to give further improvement over SILO when σ is large. Our estimator still outperforms them when $\sigma \geq 0.6$.

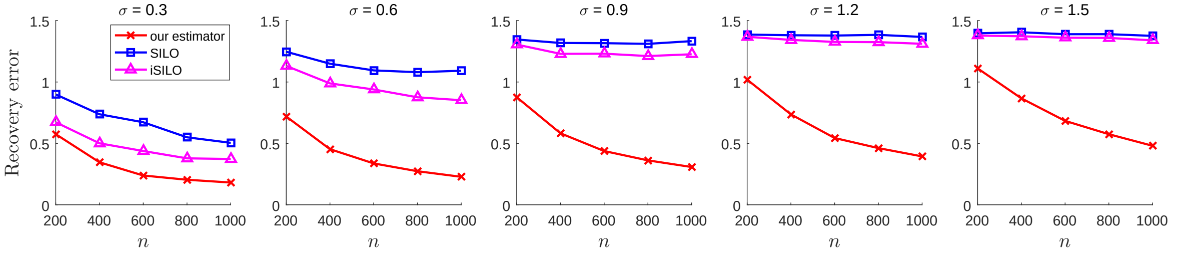


Figure 2. Recovery error vs. sample size, with $\tilde{f}(z) = z^3$ under heavy-tailed noise

6. Conclusion

In this paper, we study the parameter estimation for the high-dimensional single-index models. We propose two classes of robust estimators, which generalize previous works in two aspects. First we allow the diverse structure (e.g., binary, monotone and etc.) of the transfer function, which can help us customize the estimators. Secondly the structure of the true parameter can be general, either encoded by a constraint or a norm. With limited assumption on the noise, we can show that the estimation error can be bounded by simple geometric measures under Gaussian measurement, which subsumes the existing results for specific settings. The experiment results also validate our theoretical analyses.

Acknowledgements

We thank Sreangsu Acharyya for helpful discussions related to the paper. The research was supported by NSF grants IIS-1563950, IIS-1447566, IIS-1447574, IIS-1422557, CCF-1451986, CNS- 1314560, IIS-0953274, IIS-1029711, NASA grant NNX12AQ39A, and gifts from Adobe, IBM, and Yahoo.

References

Alquier, P. and Biau, G. Sparse single-index model. *Journal of Machine Learning Research*, 14:243–280, 2013.

Amelunxen, D., Lotz, M., McCoy, M. B., and Tropp, J. A. Living on the edge: Phase transitions in convex programs with random data. *Inform. Inference*, 3(3):224–294, 2014.

Bach, F., Jenatton, R., Mairal, J., and Obozinski, G. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2012.

Banerjee, A., Chen, S., Fazayeli, F., and Sivakumar, V. Estimation with norm regularization. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

Bellman, R. On the approximation of curves by line segments using dynamic programming. *Communications of the ACM*, 4(6):284, 1961.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

Boufounos, P. T and Baraniuk, R. G. 1-bit compressive sensing. In *Information Sciences and Systems, 2008. CISS 2008. 42nd Annual Conference on*, 2008.

Candes, E. and Tao, T. The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

Chandrasekaran, V., Recht, B., Parrilo, P. A., and Willsky, A. S. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

Chatterjee, S., Chen, S., and Banerjee, A. Generalized dantzig selector: Application to the k-support norm. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

Chen, S. and Banerjee, A. One-bit compressed sensing with the k-support norm. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2015a.

Chen, S. and Banerjee, A. Structured estimation with atomic norms: General bounds and applications. In *Advances in Neural Information Processing Systems*, 2015b.

Chen, S. and Banerjee, A. Structured matrix recovery via the generalized dantzig selector. In *Advances in Neural Information Processing Systems*, 2016.

Dirksen, S. Dimensionality reduction with subgaussian matrices: A unified theory. *Foundations of Computational Mathematics*, 16(5):1367–1396, 2016.

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

Gopi, S., Netrapalli, P., Jain, P., and Nori, A. One-bit compressed sensing: Provable support and vector recovery. In *Proceedings of The 30th International Conference on Machine Learning*, 2013.

Gordon, Y. Some inequalities for gaussian processes and applications. *Israel Journal of Mathematics*, 50(4):265–289, 1985.

Hoeffding, W. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Horowitz, J. L. and Hardle, W. Direct semiparametric estimation of single-index models with discrete covariates. *Journal of the American Statistical Association*, 91 (436):1632–1640, 1996.

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

Jacques, L., Laska, J. N, Boufounos, P. T, and Baraniuk, R. G. Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors. *IEEE Transactions on Information Theory*, 59(4):2082–2102, 2013.

Kakade, S., Shamir, O., Sindharan, K., and Tewari, A. Learning exponential families in high-dimensions: Strong convexity and sparsity. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010.

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

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

Koltchinskii, V. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Lecture Notes in Mathematics. Springer Berlin Heidelberg, 2011.

Lee, A. J. *U-Statistics: Theory and Practice*. Taylor & Francis, 1990.

Li, P. One scan 1-bit compressed sensing. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, 2016.

Maurer, A., Pontil, M., and Romera-Paredes, B. An Inequality with Applications to Structured Sparsity and Multitask Dictionary Learning. In *Conference on Learning Theory (COLT)*, 2014.

McCullagh, P. Generalized linear models. *European Journal of Operational Research*, 16(3):285–292, 1984.

Negahban, S., Ravikumar, P., Wainwright, M. J., and Yu, B. A unified framework for the analysis of regularized M-estimators. *Statistical Science*, 27(4):538–557, 2012.

Neykov, M., Liu, J. S., and Cai, T. L1-regularized least squares for support recovery of high dimensional single index models with gaussian designs. *J. Mach. Learn. Res.*, 17(1):2976–3012, 2016.

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

Oymak, S., Thrampoulidis, C., and Hassibi, B. The squared-error of generalized lasso: A precise analysis. In *Communication, Control, and Computing (Allerton), 2013 51st Annual Allerton Conference on*, 2013.

Plan, Y. and Vershynin, R. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory*, 59(1):482–494, 2013.

Plan, Y., Vershynin, R., and Yudovina, E. High-dimensional estimation with geometric constraints. *Information and Inference*, 2016.

Radchenko, P. High dimensional single index models. *Journal of Multivariate Analysis*, 139:266–282, 2015.

Rao, N., Recht, B., and Nowak, R. Universal Measurement Bounds for Structured Sparse Signal Recovery. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012.

Slawski, M. and Li, P. b-bit marginal regression. In *Advances in Neural Information Processing Systems*, 2015.

Slawski, M. and Li, P. Linear signal recovery from b-bit-quantized linear measurements: precise analysis of the trade-off between bit depth and number of measurements. *arXiv preprint arXiv:1607.02649*, 2016.

Talagrand, M. *The Generic Chaining*. Springer, 2005.

Talagrand, M. *Upper and Lower Bounds for Stochastic Processes*. Springer, 2014.

Tibshirani, R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

Tropp, J. A. Convex recovery of a structured signal from independent random linear measurements. In *Sampling Theory, a Renaissance*. 2015.

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. In Eldar, Y. and Kutyniok, G. (eds.), *Compressed Sensing*, chapter 5, pp. 210–268. Cambridge University Press, 2012.

Vershynin, R. *Estimation in High Dimensions: A Geometric Perspective*, pp. 3–66. Springer International Publishing, 2015.

Wainwright, M. J. Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming(Lasso). *IEEE Transactions on Information Theory*, 55:2183–2202, 2009.

Yang, Z., Wang, Z., Liu, H., Eldar, Y. C., and Zhang, T. Sparse nonlinear regression: Parameter estimation under nonconvexity. In *Proceedings of the 33nd International Conference on Machine Learning*, 2016.

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

Zhang, L., Yi, J., and Jin, R. Efficient algorithms for robust one-bit compressive sensing. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014.

Zhu, R. and Gu, Q. Towards a Lower Sample Complexity for Robust One-bit Compressed Sensing. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.