PAPER

Robust Bilinear Form Identification: A Subgradient Method with Geometrically Decaying Stepsize in the Presence of Heavy-Tailed Noise

Guowei YANG[†a)], ***Nonmember***

SUMMARY This paper delves into the utilisation of the subgradient method with geometrically decaying stepsize for Bilinear Form Identification. We introduce the iterative Wiener Filter, an l_2 regression method, and highlight its limitations when confronted with noise, particularly heavy-tailed noise. To address these challenges, the paper suggests employing the l_1 regression method with a subgradient method utilizing a geometrically decaying step size. The effectiveness of this approach is compared to existing methods, including the ALS algorithem. The study demonstrates that the l_1 algorithm, especially when paired with the proposed subgradient method, excels in stability and accuracy under conditions of heavy-tailed noise. Additionally, the paper introduces the standard rounding procedure and the $\mathcal{S}$-outlier bound as relaxations of traditional assumptions. Numerical experiments provide support and validation for the presented results.

key words: *bilinear, subgradient, l_1 regression*

1. Introduction

The investigation into bilinear forms has been a topic of exploration across various studies, particularly due to the versatile applications of bilinear models. These applications span a wide spectrum, such as object recognition [1], compressed sensing [2], digital filter synthesis [3], prediction problems [4], channel equalization [5], and echo cancellation [6]. In [7], the authors synthesized the findings of those studies and introduced a novel method known as the iterative Wiener Filter. The iterative Wiener Filter, categorized as an l_2 regression method, demonstrates commendable performance in the identification of bilinear forms. In [8], this method can also be referred to as the Alternated Least Squares (ALS) algorithm. However, this performance is contingent upon a strict limitation–namely, that the signal system is assumed to be in a noiseless environment or subjected to white Gaussian noise. Given the ubiquity of noise in real-world scenarios and the limited information available about its nature, the applicability of the filter is constrained. In the realm of compressed sensing, as noted in [9], l_2 regression methods excel in signal retrieval when the system operates in a noiseless or Gaussian noise environment. However, when confronted with heavy-tailed noise, l_2 regression struggles to converge effectively.

Manuscript received December 18, 2023.
Manuscript revised March 7, 2024.
Manuscript publicized May 6, 2024.

[†]School of Computer Science and Technology, Zhejiang Sci-Tech University, China.
a) E-mail: gwyang12@foxmail.com

DOI: 10.23919/transcom.2023EBP3210

To address system identification challenges under heavy-tailed noise conditions, we employ the l_1 regression method. The superiority of l_1 regression over l_2 regression is very intuitive in the presence of outlier observations, as l_1 regression is less affected by unusual observations due to its use of the absolute loss function. As far as we know, utilizing subgradient methods is the most practical approach to solve the l_1 regression problem. In [10], the authors discuss the Polyak subgradient method (which we will not consider in this paper since, under noiseless conditions, l_2 regression methods would be more effective) and the subgradient method with a geometrically decreasing step size. The convex version of the second algorithm can be traced back to Goffin [11]. Additionally, [12] analyzed these two methods for sharp weakly convex functions.

In this paper, we introduce the use of the subgradient method with a geometrically decaying stepsize, as introduced by Davis [12], as an effective l_1 algorithm for addressing the identification of bilinear forms under heavy-tailed noise conditions. To the best of our knowledge, the l_1 algorithm exhibits enhanced stability and attains greater accuracy when dealing with scenarios involving heavy-tailed noise. We have further demonstrated that a technique known as the standard rounding procedure [13] and an assumption, specifically the $\mathcal{S}$-outlier bound [14, Page 9], can be employed as a relaxation of conventional assumptions such as the Lipschitz bound and sharpness assumptions. Our numerical experiments have validated our results.

2. Identification of Bilinear Forms and ALS Algorithm

We consider the system with the bilinear forms given by:

$$y_i = \boldsymbol{\alpha}^T X_i \boldsymbol{\beta} + z_i, \quad i = 1, \ldots, p \tag{1}$$

in which $\boldsymbol{\alpha} \in \mathbb{R}^m$ is an unknown m-dimensional vector and $\boldsymbol{\beta} \in \mathbb{R}^n$ is also an unknown n-dimensional vector, $y_i, z_i \in \mathbb{R}$ are scalars, which denote the outcome of the system and the noise respectively. We assume $X_i = [(X_i)_1, (X_i)_2, \ldots, (X_i)_n]$ denotes an $m \times n$ matrix where $(X_i)_j$, $j = 1, \ldots, n$ are the m-dimensional column vectors of X_i. Throughout this paper, we will differentiate between scalars and vectors by denoting vectors as bold letters. For example, x represents a scalar, whereas $\boldsymbol{x}$ represents a vector.

The aim of this paper is to approximate the feasible solutions for both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ within their respective feasible sets.

Given that solving this problem is generally NP-hard and, therefore, computationally infeasible, our approach involves an approximate solution to the best subset problem.

Let $\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}$ be the estimations of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ respectively, $\boldsymbol{y} = [y_1, y_2, \cdots, y_p]^T$ is the p-dimensional vector of outcomes, and the estimation of $\boldsymbol{y}$ is $\hat{\boldsymbol{y}} = [\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_p]^T$.

Then we have

$$\hat{\boldsymbol{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_p \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{\alpha}}^T X_1 \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\alpha}}^T X_2 \hat{\boldsymbol{\beta}} \\ \vdots \\ \hat{\boldsymbol{\alpha}}^T X_p \hat{\boldsymbol{\beta}} \end{bmatrix}.$$

For further information and methods related to the system of bilinear forms, we recommend that readers refer to [15]. To facilitate our analysis, it will be very helpful to use the following relationships [7]:

$$\hat{\boldsymbol{y}} = \mathcal{X}\left(\hat{\boldsymbol{\alpha}} \otimes \hat{\boldsymbol{\beta}}\right) = \mathcal{X}\left(\hat{\boldsymbol{\alpha}} \otimes I_n\right)\hat{\boldsymbol{\beta}} = \mathcal{X}\left(I_m \otimes \hat{\boldsymbol{\beta}}\right)\hat{\boldsymbol{\alpha}}, \quad (2)$$

where

$$\mathcal{X} = \begin{bmatrix} \mathrm{Vec}\left(X_1\right)^T \\ \mathrm{Vec}\left(X_2\right)^T \\ \vdots \\ \mathrm{Vec}\left(X_p\right)^T \end{bmatrix},$$

and $\otimes$ denotes the Kronecker product, I_n and I_m are the identity matrices of sizes $n \times n$ and $m \times m$, respectively. We use operation $\mathrm{Vec}(X_i)$ to vectorize matrix X_i to a vector with mn entries, which means that stacking $(X_i)_j$ up. By employing well-established identities from the realm of linear algebra, these relationships can be readily derived.

Next we introduce the ALS algorithm as described in [8], and in [7] authors refer to this algorithm as the iterative Wiener filter. To avoid notation ambiguity between iterations and powers, we use $\boldsymbol{x}^{(k)}$ or $x^{(k)}$ to represent iterations (e.g. superscript enclosed in brackets), and x^k to represent x raised to the power of k. We use $\boldsymbol{x}^{(*)}$ to specifically denote that $\boldsymbol{x}$ belongs to the set of optimal solutions. We can define the function $G : (\mathbb{R}^m, \mathbb{R}^n) \to \mathbb{R}$ as:

$$\begin{aligned} G\left(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}\right) &:= \|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_2^2 \\ &= \left\|\boldsymbol{y} - \mathcal{X}\left(\hat{\boldsymbol{\alpha}} \otimes I_n\right)\hat{\boldsymbol{\beta}}\right\|_2^2 \\ &= \left\|\boldsymbol{y} - \mathcal{X}\left(I_m \otimes \hat{\boldsymbol{\beta}}\right)\hat{\boldsymbol{\alpha}}\right\|_2^2, \end{aligned}$$

the second and third equations follow from the Eq. (2), and then we can minimize G by the following update equations:

$$\hat{\boldsymbol{\alpha}}^{(k+1)} \leftarrow \left(\left(I_m \otimes \hat{\boldsymbol{\beta}}^{(k)}\right)^T \mathcal{X}^T \mathcal{X}\left(I_m \otimes \hat{\boldsymbol{\beta}}^{(k)}\right)\right)^{-1} \cdot \left(I_m \otimes \hat{\boldsymbol{\beta}}^{(k)}\right)\boldsymbol{y},$$

$$\hat{\boldsymbol{\beta}}^{(k+1)} \leftarrow \left(\left(\hat{\boldsymbol{\alpha}}^{(k)} \otimes I_n\right)^T \mathcal{X}^T \mathcal{X}\left(\hat{\boldsymbol{\alpha}}^{(k)} \otimes I_n\right)\right)^{-1} \cdot \left(\hat{\boldsymbol{\alpha}}^{(k)} \otimes I_n\right)\boldsymbol{y}.$$

Algorithm 1 Subgradient Method with A Geometrically Decaying Stepsize

Input The measurement matrix $\mathcal{X}$, observations $\boldsymbol{y}$, iteration times k, step size coefficient $\lambda_{\boldsymbol{\alpha}}^{(1)}, \lambda_{\boldsymbol{\beta}}^{(1)}$, initialized identifier $\hat{\boldsymbol{\alpha}}^{(1)}, \hat{\boldsymbol{\beta}}^{(1)}$.

1: Applying the Standard Rounding Procedure
2: **for** $i = 1$ to k **do**
3: Choose $\boldsymbol{h}_{\hat{\boldsymbol{\alpha}}}^{(i)} \in \partial F(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ and $\boldsymbol{h}_{\hat{\boldsymbol{\beta}}}^{(i)} \in \partial F(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$.
4: Set $\hat{\boldsymbol{\alpha}}^{(i+1)} = \hat{\boldsymbol{\alpha}}^{(i)} - \frac{\lambda_{\boldsymbol{\alpha}}^{(i)}}{\left\|\boldsymbol{h}_{\hat{\boldsymbol{\alpha}}}^{(i)}\right\|_2}\boldsymbol{h}_{\hat{\boldsymbol{\alpha}}}^{(i)}$.
5: Set $\hat{\boldsymbol{\beta}}^{(i+1)} = \hat{\boldsymbol{\beta}}^{(i)} - \frac{\lambda_{\boldsymbol{\beta}}^{(i)}}{\left\|\boldsymbol{h}_{\hat{\boldsymbol{\beta}}}^{(i)}\right\|_2}\boldsymbol{h}_{\hat{\boldsymbol{\beta}}}^{(i)}$.

The update equations above reveal that we iterate $\hat{\boldsymbol{\alpha}}^{(k)}$ and $\hat{\boldsymbol{\beta}}^{(k)}$ alternately. We will demonstrate that our subgradient method follows the same iterative procedure in next section.

3. Subgradient Method with a Geometrically Decaying Stepsize

We present our algorithm in detail as Algorithm 1. In the absence of the Lipschitz bound (3.2) and the μ-sharpness (3.1) assumptions, we employ the Standard Rounding Procedure in the first step. Subsequently, we select the subgradient of function F, with respect to parameters $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$. We delineated the function F in (3), and $\partial F(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ represents the set of subgradients of F. The determination of the step size is guided by Eqs. (6) and (7), with the method for calculating their coefficients $\lambda_{\boldsymbol{\alpha}}^{(1)}$ and $\lambda_{\boldsymbol{\beta}}^{(1)}$ provided immediately afterward.

In the rest of this section, we show how to use the subgradient method to solve the problem of identification with bilinear forms, defining function $F : (\mathbb{R}^m, \mathbb{R}^n) \to \mathbb{R}$ as:

$$F\left(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}\right) := \|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_1. \quad (3)$$

We can easily get the subgradient of $F(\boldsymbol{\alpha}, \boldsymbol{\beta})$ using the following proposition:

Proposition 3.1. *As for the aforementioned function F with fixed $\boldsymbol{\alpha}$,*

$$-\left(\mathcal{X}\left(\boldsymbol{\alpha} \otimes I_n\right)\right)^T \cdot \mathrm{sign}\left(F\left(\boldsymbol{\alpha}, \boldsymbol{\beta}\right)\right)$$

is a subgradient of F with respect to $\boldsymbol{\beta}$, and for a fixed $\boldsymbol{\beta}$,

$$-\left(\mathcal{X}\left(I_m \otimes \boldsymbol{\beta}\right)\right)^T \cdot \mathrm{sign}\left(F\left(\boldsymbol{\alpha}, \boldsymbol{\beta}\right)\right)$$

is a subgradient of F with respect to $\boldsymbol{\alpha}$.

Where $\mathrm{sign}(F(\boldsymbol{\alpha}, \boldsymbol{\beta}))$ denotes the sign of $F(\boldsymbol{\alpha}, \boldsymbol{\beta})$, that is a vector with the same dimensions as $F(\boldsymbol{\alpha}, \boldsymbol{\beta})$, but with a $+1$ entry when where $F(\boldsymbol{\alpha}, \boldsymbol{\beta})$ has an entry greater than zero, a -1 entry when $F(\boldsymbol{\alpha}, \boldsymbol{\beta})$ has an entry less than zero, and a zero entry where $F(\boldsymbol{\alpha}, \boldsymbol{\beta})$ has an entry equal to zero.

Proof. To simplify our proof and remove ambiguity, for a

fixed $\boldsymbol{\alpha}$, we will omit it in the expression $F(\boldsymbol{\alpha},\boldsymbol{\beta})$, and instead denoting it as $F(\boldsymbol{\beta})$.

Then for any $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ in domain of $F(\boldsymbol{\alpha},\boldsymbol{\beta})$, we have

$$\begin{aligned}
&F(\boldsymbol{\beta}_1) - F(\boldsymbol{\beta}_2)\\
=\ &\|Y - \mathcal{X}(\boldsymbol{\alpha}\otimes\boldsymbol{\beta}_1)\|_1 - \|Y - \mathcal{X}(\boldsymbol{\alpha}\otimes\boldsymbol{\beta}_2)\|_1\\
=\ &(Y - \mathcal{X}(\boldsymbol{\alpha}\otimes\boldsymbol{\beta}_1))^T \cdot \operatorname{sign}(Y - \mathcal{X}(\boldsymbol{\alpha}\otimes\boldsymbol{\beta}_1)) -\\
&(Y - \mathcal{X}(\boldsymbol{\alpha}\otimes\boldsymbol{\beta}_2))^T \cdot \operatorname{sign}(Y - \mathcal{X}(\boldsymbol{\alpha}\otimes\boldsymbol{\beta}_2))\\
\le\ &\left\{(Y - \mathcal{X}(\boldsymbol{\alpha}\otimes\boldsymbol{\beta}_1))^T - (Y - \mathcal{X}(\hat{\boldsymbol{\alpha}}\otimes\boldsymbol{\beta}_2))^T\right\}\cdot\\
&\operatorname{sign}(Y - \mathcal{X}(\boldsymbol{\alpha}\otimes\boldsymbol{\beta}_1))\\
=\ &-(\mathcal{X}(\boldsymbol{\alpha}\otimes I_n)(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2))^T \cdot \operatorname{sign}(Y - \mathcal{X}(\boldsymbol{\alpha}\otimes\boldsymbol{\beta}_1))\\
=\ &-\operatorname{sign}(F(\boldsymbol{\beta}_1))^T \cdot (\mathcal{X}(\boldsymbol{\alpha}\otimes I_n))(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)
\end{aligned}$$

Then we can use the same way to prove that the subgradient of $F(\boldsymbol{\alpha},\boldsymbol{\beta})$ with respect to $\boldsymbol{\alpha}$. □

Proposition 3.2. *The Kronecker product is a continuous mapping.*

Proof. Let $\boldsymbol{a}$ be any m-dimensional vector and $\boldsymbol{b}$ be a fixed n-dimensional vector, for any $\epsilon > 0$, there exists a $\delta = \frac{\epsilon}{2\sqrt{n}\|\boldsymbol{b}\|_\infty}$, and let $\boldsymbol{c}$ be a m-dimensional vector which satisfies that

$$\left\{\boldsymbol{c}\ \middle|\ \|\boldsymbol{a}-\boldsymbol{c}\|_\infty \le \frac{\epsilon}{2\sqrt{mn}\,\|\boldsymbol{b}\|_\infty}\right\}$$

We have

$$\begin{aligned}
\operatorname{dist}(\boldsymbol{a},\boldsymbol{c}) =\ &\|\boldsymbol{a}-\boldsymbol{c}\|_2\\
=\ &\sqrt{\sum_i^m (a_i - c_i)^2}\\
\le\ &\|\boldsymbol{a}-\boldsymbol{c}\|_\infty \sqrt{m} \le \delta
\end{aligned}$$

and

$$\begin{aligned}
\operatorname{dist}(\boldsymbol{a}\otimes\boldsymbol{b},\boldsymbol{c}\otimes\boldsymbol{b}) &= \sqrt{\sum_i^m\sum_j^n (a_i b_j - c_i b_j)^2}\\
&\le \sqrt{nm\,(\|\boldsymbol{b}\|_\infty)^2\,\|\boldsymbol{a}-\boldsymbol{c}\|_\infty^2}\\
&\le \frac{\epsilon}{2} \le \epsilon.
\end{aligned}$$

□

Let $\boldsymbol{a},\boldsymbol{b}$, and $\boldsymbol{c}$ be defined as described in the preceding proof of proposition. Then, there exists a scalar $\theta \in \mathbb{R}$. By applying the definition of a convex function, we obtain:

$$(\theta\boldsymbol{a} + (1-\theta)\boldsymbol{c})\otimes\boldsymbol{b} \le \theta(\boldsymbol{a}\otimes\boldsymbol{b}) + (1-\theta)(\boldsymbol{c}\otimes\boldsymbol{b}).$$

We observe that the Kronecker product constitutes a convex mapping, implying convexity in our objective function F. This assertion stems from the fact that the composition of a convex mapping (Kronecker product) with a convex function (l_1 norm) remains a convex function. The convergence guarantee for the subgradient method applied to convex functions can be found in [11]. Additionally, for weakly convex functions, the convergence assurance is established in [12].

Nevertheless, recent studies on the subgradient method often require assumptions about the objective function, such as the Lipschitz bound and sharpness assumptions.

Assumption 3.1. *(Restricted sharpness [14, Page 7]). A function $F(\cdot)$ is said to be μ-sharp with respect to $\boldsymbol{\xi}$ for some μ if*

$$F(\boldsymbol{\xi}) - F(\boldsymbol{\xi}^{(*)}) \ge \mu\left\|\boldsymbol{\xi} - \boldsymbol{\xi}^{(*)}\right\|_1 \tag{4}$$

holds for any $\boldsymbol{\xi} \in \mathbb{R}^{mn}$.

Assumption 3.2. *We assume that the function is L-Lipschitz continues, i.e. the function $F(\boldsymbol{\xi})$ satisfies*

$$\|F(\boldsymbol{\xi}_1) - F(\boldsymbol{\xi}_2)\|_2 \le L\,\|\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2\|_2. \tag{5}$$

In [14], not only the two properties mentioned earlier but also the properties of approximate restricted sharpness and mixed-norm restricted isometry property (RIP) are required. The RIP is widely used not only in the compressed sensing field but also in many other fields, as evidenced by studies such as [2], [9], [16], [17]. Here we introduce the rounding procedure [13], using the ellipsoid method to the polytope $P = P(\boldsymbol{x}) := \{\boldsymbol{x}|\ \|\mathcal{X}\boldsymbol{x}\|_1 \le 1\}$. For a given point $\boldsymbol{x} \notin P$ we using the hyperplane

$$\left\{\boldsymbol{y}\ \middle|\ (\boldsymbol{y}-\boldsymbol{x})^T\mathcal{X}^T\operatorname{sign}(\mathcal{X}\boldsymbol{x}) = ((\|\mathcal{X}\boldsymbol{x}\|_1) + 1)/2\right\}$$

to serve as a separation oracle, which separate $\boldsymbol{x}$ and P when $\|\mathcal{X}\boldsymbol{x}\|_1 > 1$.

In this procedure, we make use of the Gram-Schmidt method or an equivalent procedure, to orthogonalize the columns of $\mathcal{X}$ with respect to each other. Additionally, we normalize the columns of $\mathcal{X}$ such that they all have an l_1 norm of 1.

From [13, Theorem 2.1] we have that if $\|\boldsymbol{\xi}\|_2 \le \sqrt{mn}$, then $\|\boldsymbol{\xi}\|_1 \le 1$, and $\|\mathcal{X}\boldsymbol{\xi}\|_1 \le 1$ follows from columns scaling. After applying the rounding procedure, we can observe that a new version of matrices $\mathcal{X}$ possesses an essential nature. This condition states that the matrix $\mathcal{X}$ with the property that for any $\boldsymbol{\xi}$,

$$\|\boldsymbol{\xi}\|_1 \ge \|\mathcal{X}\boldsymbol{\xi}\|_1 \ge \frac{1}{mn\sqrt{mn}}\|\boldsymbol{\xi}\|_1.$$

A matrix $\mathcal{X}$ with this property will be known as the l_1-conditioned. This property provides insight into the behavior of $\mathcal{X}$ with respect to the ℓ_1 norm of its input vector $\boldsymbol{\xi}$.

With this rounding procedure, we can proof that we do not need the Lipschitz bound and sharpness assumptions.

Theorem 3.1. *In a noisy case, an l_1-conditioned matrix $\mathcal{X}$ ensures that the function $F(\boldsymbol{\xi}) = \|\mathcal{X}\boldsymbol{\xi} - \boldsymbol{y}\|_1$ Lipschitz continuous with a constant of $L = 1$.*

Proof.

$$\begin{aligned}
&|F(\boldsymbol{\xi}_1) - F(\boldsymbol{\xi}_2)| \\
=\ &|\|\boldsymbol{y}_1 - \boldsymbol{y}\|_1 - \|\boldsymbol{y}_2 - \boldsymbol{y}\|_1| \\
=\ &|\|\mathcal{X}\boldsymbol{\xi}_1 - \boldsymbol{y} + z\|_1 - \|\mathcal{X}\boldsymbol{\xi}_2 - \boldsymbol{y} + z\|_1| \\
\leq\ &\|\mathcal{X}\boldsymbol{\xi}_1 - \mathcal{X}\boldsymbol{\xi}_2\|_1 \\
\leq\ &\|\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2\|_1
\end{aligned}$$

where the second line equality follows from the presence of noise, specifically heavy-tailed noise, the third line inequality follows from the triangle inequality and the fourth inequality follows from the l_1-condition of $\mathcal{X}$. As a result, we have $L = 1$, and (5) follows. □

In prior literature, the RIP plays a crucial role in proving algorithm convergence. However, through the rounding procedure introduced here, we can directly obtain an l_1-conditioned measurement matrix $\mathcal{X}$. This property contributes to the establishment of the fourth inequality in Theorem 3.1.

Assumption 3.3. *($\mathcal{S}$-outlier bound [14, Page 9]) Matrix $X \in \mathbb{R}^{m\times n}$ is said to obey $\mathcal{S}$-outlier bound with respect to a set S with a constant δ if for all vectors $\boldsymbol{\alpha} \in \mathbb{R}^m, \boldsymbol{\beta} \in \mathbb{R}^n$, one has*

$$\delta \|\boldsymbol{\alpha} \otimes \boldsymbol{\beta}\|_1 \leq \left\|\boldsymbol{\alpha}^T X_{\mathcal{S}^c} \boldsymbol{\beta}\right\|_1 - \left\|\boldsymbol{\alpha}^T X_{\mathcal{S}} \boldsymbol{\beta}\right\|_1,$$

where $X_{\mathcal{S}^c}$ means $\{X_i\}_{i\in\mathcal{S}^c}$ and $X_{\mathcal{S}}$ means $\{X_i\}_{i\in\mathcal{S}}$.

Theorem 3.2. *In the presence of noise, if assumption 3.3 holds, a function $F(\boldsymbol{\xi}) = \|\mathcal{X}\boldsymbol{\xi} - \boldsymbol{y}\|_1$ is regarded as μ-sharp, with $\mu = c\delta$.*

Proof. If assumption 3.3 holds and noise is present, we have

$$\begin{aligned}
&F(\boldsymbol{\xi}) - F\left(\boldsymbol{\xi}^{(*)}\right) \\
=\ &\left\|\mathcal{X}\boldsymbol{\xi} - \mathcal{X}\boldsymbol{\xi}^{(*)} + z\right\|_1 - \|z\|_1 \\
=\ &\left\|\mathcal{X}_{\mathcal{S}^c}\boldsymbol{\xi} - \mathcal{X}_{\mathcal{S}^c}\boldsymbol{\xi}^{(*)}\right\|_1 + \\
&\sum_{i\in\mathcal{S}}\left(\left|\boldsymbol{\alpha}^T X_i \boldsymbol{\beta} - \left(\boldsymbol{\alpha}^{(*)}\right)^T X_i \boldsymbol{\beta}^{(*)} + z_i\right| - |z_i|\right) \\
\geq\ &\left\|\mathcal{X}_{\mathcal{S}^c}\boldsymbol{\xi} - \mathcal{X}_{\mathcal{S}^c}\boldsymbol{\xi}^{(*)}\right\|_1 - \\
&\sum_{i\in\mathcal{S}}\left(\left|\boldsymbol{\alpha}^T X_i \boldsymbol{\beta} - \left(\boldsymbol{\alpha}^{(*)}\right)^T X_i \boldsymbol{\beta}^{(*)}\right|\right) \\
\geq\ &c\delta\left\|\boldsymbol{\xi} - \boldsymbol{\xi}^{(*)}\right\|_1,
\end{aligned}$$

where the first equality arises from the presence of noise, with the former part simply unfolding the expression of the function F. The latter part arises from the fact that the term $\boldsymbol{\xi}^{(*)}$ is what we subtract within function F, and subtracting two identical terms leaves only a z. The second equality follows from the definition of $\mathcal{S}$, the third inequality follows from the triangle inequality, the last inequality follows from the $\mathcal{S}$-outlier bound, and c is a constant. Therefore, we have $\mu = c\delta$. □

Subsequently, it becomes apparent that we can regard the standard rounding procedure and the assumption 3.3 as a form of relaxation for the Lipschitz bound and sharpness assumptions.

Following this, we can employ a strategy akin to the one presented in [11] to ascertain the algorithm's step size:

$$t_{\boldsymbol{\alpha}}^{(k)} = \frac{\lambda_{\boldsymbol{\alpha}}^{(k)}}{\left\|\boldsymbol{h}_{\boldsymbol{\alpha}}^{(k)}\right\|_2}, \quad t_{\boldsymbol{\beta}}^{(k)} = \frac{\lambda_{\boldsymbol{\beta}}^{(k)}}{\left\|\boldsymbol{h}_{\boldsymbol{\beta}}^{(k)}\right\|_2}, \tag{6}$$

where $\lambda_{\boldsymbol{\alpha}}^{(k)} = \lambda_{\boldsymbol{\alpha}}^{(0)}\rho_{\boldsymbol{\alpha}}^q$ and $\lambda_{\boldsymbol{\beta}}^{(k)} = \lambda_{\boldsymbol{\beta}}^{(0)}\rho_{\boldsymbol{\beta}}^q$. We initialize $\lambda_{\boldsymbol{\alpha}}^{(0)} = R\mu/(mp)$ and $\lambda_{\boldsymbol{\alpha}}^{(0)} = R\mu/(np)$. Let $\rho_{\boldsymbol{\alpha}}$ and $\rho_{\boldsymbol{\beta}}$ satisfy that

$$\rho_{\boldsymbol{\alpha}} = \begin{cases} \sqrt{1-(\mu/m)^2} & \mu/m \leq \sqrt{2}/2 \\ \mu/(2m) & \mu/m \geq \sqrt{2}/2 \end{cases}, \tag{7}$$

$$\rho_{\boldsymbol{\beta}} = \begin{cases} \sqrt{1-(\mu/n)^2} & \mu/n \leq \sqrt{2}/2 \\ \mu/(2n) & \mu/n \geq \sqrt{2}/2 \end{cases}.$$

Here, R is a constant, and we assume that the iteration algorithm started in close proximity to the feasible solution set. This implies $\left\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{(*)}\right\|_2 + \left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(*)}\right\|_2 \leq R$.

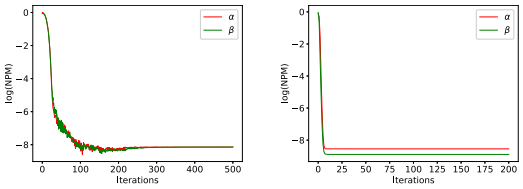
4. Numerical Experiment

This section presents the experimental settings and numerical results of our study. To evaluate the accuracy of the measurements, we use the normalized projection misalignment metric [7, Page 654].

$$\mathrm{NPM}(\boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}}) = 1 - \left(\frac{\boldsymbol{\alpha}^T\hat{\boldsymbol{\alpha}}}{\|\boldsymbol{\alpha}\|_2\|\hat{\boldsymbol{\alpha}}\|_2}\right)^2$$

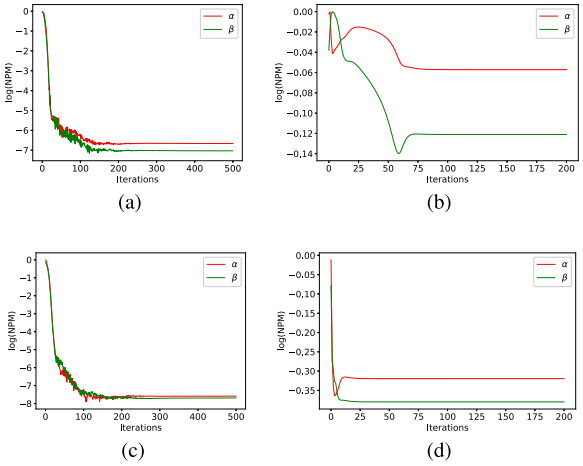
$$\mathrm{NPM}\left(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}\right) = 1 - \left(\frac{\boldsymbol{\beta}^T\hat{\boldsymbol{\beta}}}{\|\boldsymbol{\beta}\|_2\left\|\hat{\boldsymbol{\beta}}\right\|_2}\right)^2.$$

We generated the entries of vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ using the Bernoulli distribution with probability 1/2, while the matrix X_i was generated by independent identically distributed (i.i.d.) $N(0,1)$ random variables. We choose values for $\rho_{\boldsymbol{\alpha}}$ and $\rho_{\boldsymbol{\beta}}$ from the interval $[0.9, 1)$, and set the vector lengths to $m = 30$ and $n = 30$. Let us consider that there are $p = 200$ data samples available for estimating the vectors. In Fig. 1, it is evident that the l_2 regression model exhibits faster convergence than the l_1 regression model within the Gaussian noise structure. In Fig. 2, we examine a system exposed to Cauchy noise and another system subjected to heteroscedastic noise. The heteroscedastic noise structure is characterized by the following distribution: one-third of the entries conform to a Gaussian distribution, another third adhere to a Cauchy distribution, and the remaining entries follow a t-distribution. It is observable that for a system under Cauchy noise or heteroscedastic noise, the l_2 regression method struggles to converge, while the l_1 regression method



(a) (b)

Fig. 1 On the left (a), we have a system under Gaussian noise employing the l_1 regression model with the subgradient method and a geometrically decaying stepsize. On the right (b), we have a system under Gaussian noise using the l_2 regression model with the ALS algorithm.



(a) (b)

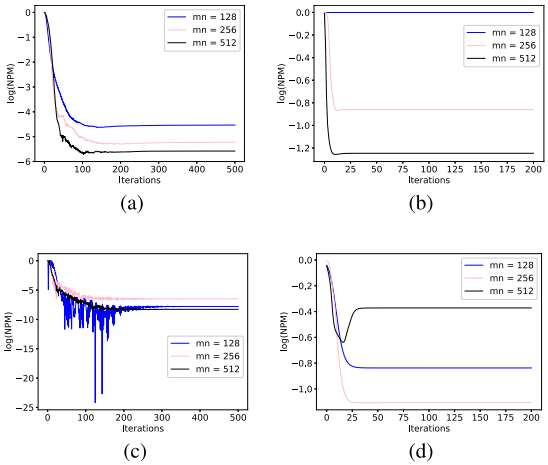
(c) (d)

Fig. 2 Figures (a) and (c) employ the l_1 regression model with the subgradient method and a geometrically decaying stepsize, while figures (b) and (d) utilize the ALS algorithm. Figures (a) and (b) illustrate the system under Cauchy noise, whereas figures (c) and (d) explore the impact of heteroscedastic noise.

continues to perform well.

Next, the algorithm's performance is evaluated from a system identification perspective. We generated the entries of $\boldsymbol{\alpha}$ according to the ITU-T G.168 Recommendation [18], and $\boldsymbol{\beta}$ are generated as $\beta_i = 2^{-(i-1)}$, with $i = 1, 2, \ldots, n$. In this simulation, as depicted in Fig. 3, the length of $\boldsymbol{\beta}$ varies with values of $m = 2, 4,$ and 8; consequently, the length of $\boldsymbol{\alpha}$ is fixed at $n = 64$. From a system identification standpoint, it is evident that under a heavy-tailed noise condition, the l_1 regression consistently outperforms the l_2 regression method. The variation observed in Figure (a) within Fig. 3 is attributed to the relatively short length of the vector $\boldsymbol{\beta}$. For a fixed value of p (e.g., the available data samples), it is evident that increasing the product of mn can contribute to achieving more accurate results.

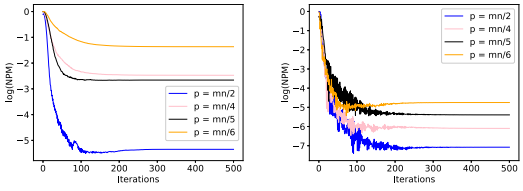
Finally, we will explore the impact of relatively small available data samples, denoted as p, on the algorithm. Notably, not only does the l_1 regression method converge when $p < mn$, but it also performs well when $p < mn/4$. Contrastingly, the l_2 regression method faces challenges in attaining satisfactory results under such conditions, primarily due to the influence of heavy-tailed noise and the limited availability of data samples. See Fig. 4.



(a) (b)

(c) (d)

Fig. 3 In this experiment, we demonstrate the convergence of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. In Figure (a) and figure (c), we apply the l_1 regression model using the subgradient method with a geometrically decaying step size. In Figure (b) and Figure (d), the l_2 regression model is employed. Figures (a) and (b) depict the NPM of $\boldsymbol{\alpha}$, while Figures (c) and (d) showcase the NPM of $\boldsymbol{\beta}$. We experiment with various combinations of m and n while maintaining a fixed $p = 200$ under a Cauchy noise condition.



(a) (b)

Fig. 4 Figure (a) illustrates the Normalized Power Mean (NPM) of $\boldsymbol{\alpha}$, while Figure (b) presents the NPM of $\boldsymbol{\beta}$. We conduct tests using the subgradient method with a geometrically decaying step size under a Cauchy noise scenario, varying the parameter p (representing available data samples). The vector lengths of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are set to $m = 8$ and $n = 64$, respectively.

5. Conclusion

To conclude, the utilization of l_1 regression methods presents the advantageous capability of generating robust solutions, a trait highly beneficial in diverse applications like phase retrieval and compressed sensing. This subgradient method exhibits relatively good performance when dealing with bilinear systems under heavy-tailed noise.

However, the selection between l_1 and l_2 regression methods hinges upon the distinct problem and noise characteristics at hand. In certain instances, the preference may lean towards l_2 regression, particularly when dealing with Gaussian noise and well-conditioned problems. In a broader perspective, the integration of subgradient methods for nonlinear problem-solving has demonstrated promising outcomes, holding significant potential for driving substantial advancements across various application domains.

Acknowledgments

I would like to extend my heartfelt gratitude and sincere

appreciation to Dr. Shen for their invaluable guidance, unwavering support, and insightful mentorship throughout my academic journey. I would also like to express my deep gratitude to Zhejiang Sci-Tech University for providing me with an exceptional learning environment and resources that have been crucial to my intellectual and personal growth.

References

[1] C. Zhang, J. Liu, Q. Tian, Y. Han, H. Lu, and S. Ma, "A boosting, sparsity-constrained bilinear model for object recognition," IEEE MultiMedia, vol.19, no.2, pp.58–68, 2012.
[2] P. Walk and P. Jung, "Compressed sensing on the image of bilinear maps," 2012 IEEE International Symposium on Information Theory Proceedings, pp.1291–1295, 2012.
[3] U. Forssen, "Adaptive bilinear digital filters," IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process., vol.40, no.11, pp.729–735, 1993.
[4] J. Lee and V.J. Mathews, "Adaptive bilinear predictors," Proc. ICASSP '94. IEEE International Conference on Acoustics, Speech and Signal Processing, vol.iii, pp.III/489–III/492, 1994.
[5] G. Ma, J. Lee, and V.J. Mathews, "A RLS bilinear filter for channel equalization," Proc. ICASSP '94. IEEE International Conference on Acoustics, Speech and Signal Processing, vol.iii, pp.III/257–III/260, 1994.
[6] R. Hu and H. Ahmed, "Echo cancellation in high speed data transmission systems using adaptive layered bilinear filters," IEEE Trans. Commun., vol.42, no.234, pp.655–663, 1994.
[7] J. Benesty, C. Paleologu, and S. Ciochină, "On the identification of bilinear forms with the wiener filter," IEEE Signal Process. Lett., vol.24, no.5, pp.653–657, 2017.
[8] G. Chen, M. Gan, S. Wang, and C.L.P. Chen, "Insights into algorithms for separable nonlinear least squares problems," IEEE Trans. Image Process., vol.30, pp.1207–1218, 2021.
[9] S. Li, D. Liu, and Y. Shen, "Adaptive iterative hard thresholding for least absolute deviation problems with sparsity constraints," J. Fourier Anal. Appl., vol.29, pp.1207–1218, 2022.
[10] V. Charisopoulos, D. Davis, M. Díaz, and D. Drusvyatskiy, "Composite optimization for robust rank one bilinear sensing," Information and Inference: A Journal of the IMA, vol.10, no.2, pp.333–396, Oct. 2020.
[11] J.L. Goffin, "On convergence rates of subgradient optimization methods," Mathematical Programming, vol.13, pp.329–347, 1977.
[12] D. Davis, D. Drusvyatskiy, K.J. MacPhee, and C. Paquette, "Subgradient methods for sharp weakly convex functions," J. Optim. Theory Appl., vol.179, pp.962–982, 2018.
[13] K.L. Clarkson, "Subgradient and sampling algorithms for l_1 regression," Proc. Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '05, USA, pp.257–266, 2005.
[14] T. Tong, C. Ma, and Y. Chi, "Low-rank matrix recovery with scaled subgradient methods: Fast and robust convergence without the condition number," IEEE Trans. Signal Process., vol.69, pp.2396–2409, 2021.
[15] D. Yang, "Solution theory for systems of bilinear equations," Ph.D. thesis, The College of William and Mary, April 2011.
[16] B. Recht, M. Fazel, and P.A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," SIAM Rev., vol.52, no.3, pp.471–501, 2010.
[17] Y. Chen, Y. Chi, and A.J. Goldsmith, "Exact and stable covariance estimation from quadratic sampling via convex programming," IEEE Trans. Inf. Theory, vol.61, no.7, pp.4034–4059, 2015.
[18] I.T. Union, "Digital network echo cancellers," ITU-T Recommendation G.168, ITU-T, 2002.

Guowei Yang earned his B.S. degree in Software Engineering from Xi'an University of Posts & Telecommunications, Xi'an, China, in 2020. Currently, he is pursuing a master's degree in Computer Technology at Zhejiang Sci-Tech University. His current research interests encompass optimization algorithms, compressed sensing, and phase retrieval.