

Bilinear Compressed Sensing Under Known Signs via Convex Programming

Alireza Aghasi¹, Ali Ahmed², Paul Hand, and Babhru Joshi³

Abstract—We consider the bilinear inverse problem of recovering two vectors, $x \in \mathbb{R}^L$ and $w \in \mathbb{R}^L$, from their entrywise product. We consider the case where x and w have known signs and are sparse with respect to known dictionaries of size K and N , respectively. Here, K and N may be larger than, smaller than, or equal to L . We introduce ℓ_1 -BranchHull, which is a convex program posed in the natural parameter space and does not require an approximate solution or initialization in order to be stated or solved. Under the assumptions that x and w satisfy a comparable-effective-sparsity condition and are S_1 - and S_2 -sparse with respect to a random dictionary, we present a recovery guarantee in a noisy case. We show that ℓ_1 -BranchHull is robust to small dense noise with high probability if the number of measurements satisfy $L \geq \Omega((S_1 + S_2) \log^2(K + N))$. Numerical experiments show that the scaling constant in the theorem is not too large. We also introduce variants of ℓ_1 -BranchHull for the purposes of tolerating noise and outliers, and for the purpose of recovering piecewise constant signals. We provide an ADMM implementation of these variants and show they can extract piecewise constant behavior from real images.

Index Terms—Inverse problems, deconvolution, blind source separation, compressed sensing, optimization.

I. INTRODUCTION

WE STUDY the problem of recovering two unknown vectors x and w in \mathbb{R}^L from observations

$$y = w \odot x \odot (1 + \xi), \quad (1)$$

where \odot denotes entrywise product and $\xi \in \mathbb{R}^L$ is noise. Let $B \in \mathbb{R}^{L \times K}$ and $C \in \mathbb{R}^{L \times N}$ such that $w = Bh$ and $x = Cm$ with $\|h\|_0 \leq S_1$ and $\|m\|_0 \leq S_2$. The bilinear inverse problem (BIP) we consider is to find w and x , up to the inherent scaling ambiguity, from y , B , C and sign(w).

Manuscript received August 31, 2019; revised March 22, 2020; accepted July 19, 2020. Date of publication September 14, 2020; date of current version November 20, 2020. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Yue M. Lu. This work was supported in part by NSF DMS under Grant 1848087. (Corresponding author: Babhru Joshi.)

Alireza Aghasi is with the J. Mack Robinson College of Business, Georgia State University, Atlanta, GA 30302 USA (e-mail: aaghasi@gsu.edu).

Ali Ahmed is with the Department of Electrical Engineering, ITU, Lahore 54000, Pakistan (e-mail: ali.ahmed@itu.edu.pk).

Paul Hand is with the Department of Mathematics and Khoury College of Computer Sciences, Northeastern University, Boston, MA 02115 USA (e-mail: p.hand@northeastern.edu).

Babhru Joshi is with the Department of Mathematics, University of British Columbia, Vancouver, BC V6T 1Z4, Canada (e-mail: b.joshi@math.ubc.ca).

Digital Object Identifier 10.1109/TSP.2020.3017929

BIPs, in general, have many applications in signal processing and machine learning and include fundamental practical problems like phase retrieval [1]–[3], blind deconvolution [4]–[7], non-negative matrix factorization [8], [9], self-calibration [10], blind source separation [11], dictionary learning [12], etc. These problems are in general challenging and suffer from identifiability issues that make the solution set non-unique and non-convex. A common identifiability issue, also shared by the BIP in (1), is the scaling ambiguity. In particular, if (w^\natural, x^\natural) solves a BIP, then so does $(cw^\natural, c^{-1}x^\natural)$ for any non-zero $c \in \mathbb{R}$. In this article, we resolve this scaling ambiguity by finding the point in the solution set closest to the origin with respect to the ℓ_1 norm.

Another identifiability issue of the BIP in (1) is if (w^\natural, x^\natural) solves (1), then so does $(1, w^\natural \odot x^\natural)$, where 1 is the vector of ones. In prior works like [4], this identifiability issue is resolved by assuming the unknown signals live in known subspaces. In contrast we resolve the identifiability issue by assuming the signals are sparse with respect to known bases or dictionaries. Natural choices for such bases include the standard basis, the Discrete Cosine Transform (DCT) basis, and a wavelet basis.

In addition to sparsity of the unknown vectors, w^\natural and x^\natural , with respect to known dictionaries, we also require sign information of w^\natural and x^\natural . Knowing the signs of these unknown vectors is justified in many imaging applications, where w^\natural or x^\natural correspond to images and contain non-negative pixel values. There are various problems in image/signal processing and optics, where blind demodulation is employed to recover or improve a desired signal:

- 1) In imaging applications, inspired by the Lambertian reflectance model, a well-know de-illumination technique considers decomposing an image into the product of an albedo pattern and a lighting map (e.g., see [13] or Section 9.2 of [14]). Multiplicative models also appear in image dehazing applications [15]–[17], where the image visibility is improved by removing the unwanted haze from the image.
- 2) In optical applications, and specifically terahertz time-domain spectroscopy, blind demodulation is used to remove sweep distortions from the actual signal. Such distortions are normally caused by the depth variations or the layered structure inter-reflections of the specimen [7]. We will highlight a stylized case of distortion removal from real images in the numerical experiments presented in Section III.

Recovering signals with known signs from bilinear measurements is also important in other structured inverse problems; one such scenario is blind deconvolution from phaseless measurements, which arises in visible light communications, and other imaging context, and is considered in [18]. Here one wants to recover signals by only observing the product of their phase-less measurements. The known sign assumption is naturally met in this context as the problem can be reframed as a bilinear recovery problem of positive-definite matrices.

In this article, we study a sparse bilinear inverse problem, where the unknown vectors are sparse with respect to known dictionaries. Similar sparse BIPs have been extensively studied in the literature and are known to be challenging. In particular, the best known algorithm for sparse BIPs that can provably recover \mathbf{h}^\natural and \mathbf{m}^\natural require measurements that scale quadratically, up to log factors, with respect to the sparsity levels, S_1 and S_2 . Recent work on sparse rank-1 matrix recovery problem in [19], which is motivated by considering the lifted version of the sparse blind deconvolution problem, provides an exact recovery guarantee of the sparse vectors \mathbf{h}^\natural and \mathbf{m}^\natural that satisfy a ‘peakiness’ condition, i.e. $\min\{\|\mathbf{h}^\natural\|_\infty, \|\mathbf{m}^\natural\|_\infty\} \geq c$ for some absolute constant $c \in \mathbb{R}$, using a non-convex approach. This result holds with high probability for random measurements if the number of measurements satisfy $L \geq \Omega(S_1 + S_2)$, up to a log factor. For general vectors that are not constrained to a class of sparse vectors like those satisfying the peakiness condition, the same work shows exact recovery is possible if the number of measurements satisfy $L \geq \Omega(S_1 S_2)$, up to a log factor.

The main contribution of the present article is to introduce a convex program in the natural parameter space for the sparse BIP described in (1) and show that it can stably recover sparse vectors, up to the global scaling ambiguity, provided they satisfy a comparable-effective-sparsity condition. Precisely, we say the sparse vectors \mathbf{h}^\natural and \mathbf{m}^\natural are ρ -comparable-effective-sparse, for some $\rho \geq 1$, if there exists an $\alpha \in \mathbb{R}$ that satisfies $\max(\alpha, 1/\alpha) \leq \rho$ and

$$\frac{\|\mathbf{h}^\natural\|_1}{\|\mathbf{h}^\natural\|_2} = \alpha \frac{\|\mathbf{m}^\natural\|_1}{\|\mathbf{m}^\natural\|_2}. \quad (2)$$

Note that the ratio of the ℓ_1 to ℓ_2 norm of a vector is an approximate notion of the square root of the sparsity of a vector. Additionally, we assume the noise in (1) does not change the sign of the measurements. Specifically, we consider noise $\boldsymbol{\xi} \in \mathbb{R}^L$ such that

$$\xi_\ell \geq -1 \text{ for all } \ell = 1, \dots, L. \quad (3)$$

Under the assumptions that the sparse vectors satisfy (2) and the noise is small as in (3), we show that the convex program stably recovers the unknown vectors, up to the global scaling ambiguity, with high probability provided \mathbf{B} and \mathbf{C} are random and the number measurements satisfy $L \geq \Omega(S_1 + S_2)$, up to log factors. Similar to the result in [19], this results has optimal sample complexity, up to log factors, for a class of sparse signals, namely those with comparable sparsity levels.

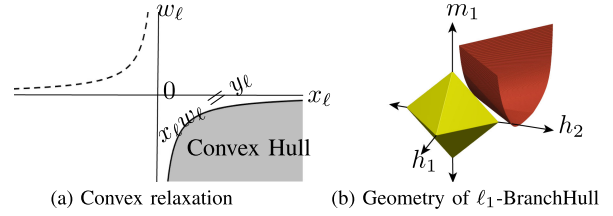


Fig. 1. Panel (a) shows the convex hull of the relevant branch of a hyperbola given a measurement y_ℓ and the sign information $\text{sign}(w_\ell)$. Panel (b) shows the interaction between the ℓ_1 -ball in the objective of (4) with its feasibility set. The feasibility set is ‘pointy’ along a hyperbola corresponding the scaling ambiguity, which allows for signal recovery where the ℓ_1 ball touches this hyperbola.

A. Convex Program and Main Results

We introduce a convex program written in the natural parameter space for the bilinear inverse problem described in (1). Let $(\mathbf{h}^\natural, \mathbf{m}^\natural) \in \mathbb{R}^K \times \mathbb{R}^N$ with $\|\mathbf{h}^\natural\|_0 \leq S_1$ and $\|\mathbf{m}^\natural\|_0 \leq S_2$. Let $w_\ell = \mathbf{b}_\ell^\top \mathbf{h}^\natural$, $x_\ell = \mathbf{c}_\ell^\top \mathbf{m}^\natural$ and $y_\ell = \mathbf{b}_\ell^\top \mathbf{h}^\natural \cdot \mathbf{c}_\ell^\top \mathbf{m}^\natural \cdot (1 + \xi_\ell)$, where \mathbf{b}_ℓ^\top and \mathbf{c}_ℓ^\top are the ℓ th row of \mathbf{B} and \mathbf{C} , respectively, and ξ_ℓ is the ℓ th entry of $\boldsymbol{\xi}$. Also, let $\mathbf{s} = \text{sign}(\mathbf{y})$ and $\mathbf{t} = \text{sign}(\mathbf{w}) = \text{sign}(\mathbf{B}\mathbf{h}^\natural)$. The convex program we consider to recover $(\mathbf{h}^\natural, \mathbf{m}^\natural)$ is the ℓ_1 -BranchHull program

$$\begin{aligned} \ell_1\text{-BH} : \quad & \underset{\mathbf{h} \in \mathbb{R}^K, \mathbf{m} \in \mathbb{R}^N}{\text{minimize}} \quad \|\mathbf{h}\|_1 + \|\mathbf{m}\|_1 \\ & \text{subject to} \quad \mathbf{s}_\ell (\mathbf{b}_\ell^\top \mathbf{h} \mathbf{c}_\ell^\top \mathbf{m}) \geq |y_\ell| \\ & \quad \quad \quad t_\ell \mathbf{b}_\ell^\top \mathbf{h} \geq 0, \quad \ell = 1, 2, \dots, L. \end{aligned} \quad (4)$$

The motivation for the feasible set in program (4) follows from the observation that each measurement $y_\ell = w_\ell \cdot x_\ell$ defines a hyperbola in \mathbb{R}^2 . As shown in Fig. 1(a), the sign information $t_\ell = \text{sign}(w_\ell)$ restricts (w_ℓ, x_ℓ) to one of the branches of the hyperbola. The feasible set in (4) corresponds to the convex hull of particular branches of the hyperbola for each y_ℓ . This also implies that the feasible set is convex as it is the intersection of L convex sets.

The objective function in (4) is an ℓ_1 minimization over (\mathbf{h}, \mathbf{m}) that finds a point $(\hat{\mathbf{h}}, \hat{\mathbf{m}})$ with $\|\hat{\mathbf{h}}\|_1 = \|\hat{\mathbf{m}}\|_1$.¹ Geometrically, this happens as the solution lies at the intersection of the ℓ_1 -ball and the hyperbolic curve (constraint) as shown in Fig. 1(b). So, the minimizer of (4) in the noiseless case, under successful recovery, is

$$(\hat{\mathbf{h}}, \hat{\mathbf{m}}) := \left(\mathbf{h}^\natural \sqrt{\frac{\|\mathbf{m}^\natural\|_1}{\|\mathbf{h}^\natural\|_1}}, \mathbf{m}^\natural \sqrt{\frac{\|\mathbf{h}^\natural\|_1}{\|\mathbf{m}^\natural\|_1}} \right). \quad (5)$$

We now present our main result which states that the ℓ_1 -BranchHull program (4) stably recovers \mathbf{w} and \mathbf{x} , up to the global scaling ambiguity, in the presence of small dense noise. We show that if \mathbf{w} and \mathbf{x} live in random subspaces with \mathbf{h}^\natural and

¹ If (\mathbf{h}, \mathbf{m}) is feasible, then so is $(c\mathbf{h}, \frac{1}{c}\mathbf{m})$ for $c > 0$. The value of c that minimizes $\|c\mathbf{h}\|_1 + \|\frac{1}{c}\mathbf{m}\|_1$ is such that both terms have equal ℓ_1 norm.

$\mathbf{m}^{\mathfrak{h}}$ containing at most S_1 and S_2 non zero entries, $\mathbf{h}^{\mathfrak{h}}$ and $\mathbf{m}^{\mathfrak{h}}$ satisfy comparable-effective-sparsity condition (2), the noise $\boldsymbol{\xi}$ satisfy (3), and there are at least $\Omega((S_1 + S_2) \log^2(K + N))$ number of measurements, then the minimizer of the ℓ_1 -BranchHull program is close to the bilinear ambiguity set $\{(c\hat{\mathbf{h}}, c^{-1}\hat{\mathbf{m}}) | c > 0\}$ with high probability. Moreover, in the noiseless case, the minimizer of the ℓ_1 -BranchHull is the point on this bilinear ambiguity set with equal ℓ_1 norm with high probability, i.e. the minimizer is $(\hat{\mathbf{h}}, \hat{\mathbf{m}})$ with high probability.

Theorem 1. (Noisy recovery): Fix $\rho \geq 1$. Fix $(\mathbf{h}^{\mathfrak{h}}, \mathbf{m}^{\mathfrak{h}}) \in \mathbb{R}^{K+N}$ that are ρ -comparable-effective-sparse as defined in (2) with $\mathbf{h}^{\mathfrak{h}} \neq \mathbf{0}$, $\|\mathbf{h}^{\mathfrak{h}}\|_0 \leq S_1$ and $\mathbf{m}^{\mathfrak{h}} \neq \mathbf{0}$, $\|\mathbf{m}^{\mathfrak{h}}\|_0 \leq S_2$. Let $\mathbf{B} \in \mathbb{R}^{L \times K}$ and $\mathbf{C} \in \mathbb{R}^{L \times N}$ have i.i.d. $\mathcal{N}(0, 1)$ entries. Let $\mathbf{y} \in \mathbb{R}^L$ contain measurements that satisfy (1) with noise $\boldsymbol{\xi} \in \mathbb{R}^L$ satisfying (3). If $L \geq C_\rho(\sqrt{S_1 + S_2} \log(K + N) + t)^2$ for any $t > 0$ then the ℓ_1 -BranchHull program (4) recovers $(\tilde{\mathbf{h}}, \tilde{\mathbf{m}})$ that satisfies

$$\min_{c>0} \frac{\|(\tilde{\mathbf{h}}, \tilde{\mathbf{m}}) - (c\hat{\mathbf{h}}, c^{-1}\hat{\mathbf{m}})\|_2}{\|(c\hat{\mathbf{h}}, c^{-1}\hat{\mathbf{m}})\|_2} \leq 37\sqrt{\|\boldsymbol{\xi}\|_\infty}$$

with probability at least $1 - e^{-c_t L}$. Here, C_ρ and c_t are constants that depend quadratically on ρ and t , respectively. Furthermore, $(\tilde{\mathbf{h}}, \tilde{\mathbf{m}}) = (\hat{\mathbf{h}}, \hat{\mathbf{m}})$ if $\boldsymbol{\xi} = \mathbf{0}$.

Theorem 1 shows that exact recovery, up to the global scaling ambiguity, of sparse vectors that satisfy the comparable-effective-sparsity condition is possible if the number of measurements satisfy $L \geq \Omega((S_1 + S_2) \log^2(K + N))$. This result is optimal, up to the log factors. Numerical simulation on synthetic data verify Theorem 1 in the noiseless case and show that the constant in the sample complexity is not too large. We also present the results of numerical simulation on two real images which shows that a total variation reformulation of the convex program (4) can successfully recover the piecewise constant part of an otherwise distorted image.

B. Prior Art for Bilinear Inverse Problems

Recent approaches to solving bilinear inverse problems like blind deconvolution and phase retrieval include lifting the problems into a low rank matrix recovery task or to formulate a convex or non-convex optimization programs in the natural parameter space. Lifting transforms the problem of recovering $\mathbf{h} \in \mathbb{R}^K$ and $\mathbf{m} \in \mathbb{R}^N$ from bilinear measurements to the problem of recovering a low rank matrix $\mathbf{h}\mathbf{m}^T$ from linear measurements. The low rank matrix can then be recovered using a semidefinite program. The result in [4] for blind deconvolution showed that if \mathbf{h} and \mathbf{m} are coefficients of the target signals with respect to Fourier and Gaussian bases, respectively, then the lifting method successfully recovers the low rank matrix. The recovery occurs with high probability under near optimal sample complexity. Unfortunately, solving the semidefinite program is prohibitively computationally expensive because they operate in high-dimension space. Also, it is not clear whether or not it is possible to enforce additional structure like sparsity of \mathbf{h} and \mathbf{m} in the lifted formulation in a way that allows optimal sample complexity [20], [21].

In comparison to the lifting approach for blind deconvolution and phase retrieval, methods that formulate an algorithm in the natural parameter space, such as alternating minimization and gradient descent based method, are computationally efficient and also enjoy rigorous recovery guarantees under optimal or near optimal sample complexity [19], [22]–[25]. The work in [19] for sparse blind deconvolution is based on alternating minimization. In the article, the authors use an alternating minimization that successively approximate the sparse vectors while enforcing the low rank property of the lifted matrix. However, because these methods are non-convex, convergence to the global optimal requires a good initialization [22], [26], [27].

Other approaches that operate in the natural parameter space include PhaseMax [28], [29] and BranchHull [30]. PhaseMax is a linear program which has been proven to find the target signal in phase retrieval under optimal sample complexity if a good anchor vector is available. As with alternating minimization and gradient descent based approach, PhaseMax requires a good initializer to even be stated. In PhaseMax the initialization is part of the objective function but in alternating minimization the initialization is part of the algorithmic implementation. BranchHull is a convex program which solves the BIP described in (1) in the dense signal case under optimal sample complexity. Like the ℓ_1 -BranchHull presented in this article, BranchHull does not require an initialization but does require the sign information of the signals.

The ℓ_1 -BranchHull program (4) combines strengths of both the lifting method and the gradient descent based method. Specifically, the ℓ_1 -BranchHull program is a convex program that operates in the natural parameter space, without a need for an initialization. These strengths are achieved at the cost of the sign information of the target signals \mathbf{w} and \mathbf{x} , which can be justified in imaging applications where the goal might be to recover pixel values of a target image, which are non-negative.

C. Discussion and Extensions

The ℓ_1 -BranchHull formulation is inspired by the BranchHull formation introduced in [30], which is a novel convex relaxation for the bilinear recovery of the entrywise product of vectors with known signs, and share many of its advantages and drawbacks. Like in BranchHull, ℓ_1 -BranchHull finds a point in the convex feasibility set that is closest to the origin. The important difference between these formulations is that BranchHull finds the point with the least ℓ_2 norm while ℓ_1 -BranchHull finds the point with the least ℓ_1 -norm. Another difference is that BranchHull enjoys recovery guarantee when those vectors belong to random real subspaces while ℓ_1 -BranchHull enjoys recovery guarantee under a much weaker condition of those vectors admitting a sparse representation with respect to known dictionaries.

Similar to BranchHull, ℓ_1 -BranchHull is a flexible and can be altered to tolerate large outliers. In this article, we show that the ℓ_1 -BranchHull formulation is stable to small dense noise and extend the work presented in [31] for the noiseless case. However, as stated formulation (4) is not robust to outliers. This is because the formulations is particularly susceptible to noise

that changes the sign of even a single measurement. For the bilinear inverse problem as described in (1) with small dense noise and arbitrary outliers, we propose the following robust ℓ_1 -BranchHull program by adding a slack variable.

$$\begin{aligned} \ell_1\text{-RBH: } & \underset{\mathbf{h} \in \mathbb{R}^K, \mathbf{m} \in \mathbb{R}^N, \boldsymbol{\xi} \in \mathbb{R}^L}{\text{minimize}} \quad \|\mathbf{h}\|_1 + \|\mathbf{m}\|_1 + \lambda \|\boldsymbol{\xi}\|_1 \\ & \text{subject to} \quad s_\ell(\mathbf{c}_\ell^\top \mathbf{m} + \xi_\ell) \mathbf{b}_\ell^\top \mathbf{h} \geq |y_\ell|, \\ & \quad t_\ell \mathbf{b}_\ell^\top \mathbf{h} \geq 0, \quad \ell = 1, \dots, L. \end{aligned} \quad (6)$$

For measurements y_ℓ with incorrect sign, the corresponding slack variables ξ_ℓ shifts the feasible set so that the target signal is feasible. In the outlier case, the ℓ_1 penalty promotes sparsity of slack variable $\boldsymbol{\xi}$. We implement a slight variation of the above program to remove distortions from an otherwise piecewise constant signal. In the case where $\mathbf{w} = \mathbf{B}\mathbf{h}^\dagger$ is a piecewise constant signal, $\mathbf{x} = \mathbf{C}\mathbf{m}^\dagger$ is a distortion signal and $\mathbf{y} = \mathbf{w} \odot \mathbf{x}$ is the distorted signal, the total variation version (7) of the robust BranchHull program (6), under successful recovery, produces the piecewise constant signal $\mathbf{B}\mathbf{h}^\dagger$, up to a scaling.

$$\begin{aligned} \text{TV BH: } & \underset{\mathbf{h} \in \mathbb{R}^K, \mathbf{m} \in \mathbb{R}^N, \boldsymbol{\xi} \in \mathbb{R}^L}{\text{minimize}} \quad \text{TV}(\mathbf{B}\mathbf{h}) + \|\mathbf{m}\|_1 + \lambda \|\boldsymbol{\xi}\|_1 \\ & \text{subject to} \quad s_\ell(\xi_\ell + \mathbf{c}_\ell^\top \mathbf{m}) \mathbf{b}_\ell^\top \mathbf{h} \geq |y_\ell| \\ & \quad t_\ell \mathbf{b}_\ell^\top \mathbf{h} \geq 0, \quad \ell = 1, 2, \dots, L. \end{aligned} \quad (7)$$

In (7), $\text{TV}(\cdot)$ is a total variation operator and is the ℓ_1 norm of the vector containing pairwise difference of neighboring elements of the target signal $\mathbf{B}\mathbf{h}$. We implement (7) to remove distortions from images in Section III and leave detailed theoretical analysis of robust ℓ_1 BranchHull (6) and its variant (7) to future work. It would also be interesting to develop convex relaxations in the natural parameter space that do not require sign information and to extend the analysis to the case when the phases of complex vectors are known and to the case of deterministic dictionaries instead of random dictionaries. All of these directions are left for future research.

D. Organization of the Article

The remainder of the article is organized as follows. In Section I-E, we present notations used throughout the article. In Section II, we present an Alternating Direction of Multipliers implementation of robust ℓ_1 -BranchHull program (6). In Section III, we observe the performance of ℓ_1 -BranchHull on synthetic random data and natural images. In Section IV, we present the proof of Theorem 1.

E. Notation

Vectors and matrices are written with boldface, while scalars and entries of vectors are written in plain font. For example, c_ℓ is the ℓ th entry of the vector \mathbf{c} . We write $\mathbf{1}$ as the vector of all ones with dimensionality appropriate for the context. We write \mathbf{I}_N as the $N \times N$ identity matrix. For any $x \in \mathbb{R}$, let $(x)_- \in \mathbb{Z}$ such that $x - 1 < (x)_- \leq x$. For any matrix \mathbf{A} , let $\|\mathbf{A}\|_F$ be the Frobenius norm of \mathbf{A} . For any vector \mathbf{x} , let $\|\mathbf{x}\|_0$ be the number of non-zero entries in \mathbf{x} . For $\mathbf{x} \in \mathbb{R}^K$ and $\mathbf{y} \in \mathbb{R}^N$, (\mathbf{x}, \mathbf{y}) is the

corresponding vector in $\mathbb{R}^K \times \mathbb{R}^N$, and $\langle (\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2) \rangle = \langle \mathbf{x}_1, \mathbf{x}_2 \rangle + \langle \mathbf{y}_1, \mathbf{y}_2 \rangle$. For a set $\mathcal{A} \subset \mathbb{R}^m$, and a vector $\mathbf{a} \in \mathbb{R}^m$, we define by $\mathbf{a} \oplus \mathcal{A}$, a set obtained by adding \mathbf{a} to every element of \mathcal{A} .

II. ALGORITHM

In this section, we present an Alternating Direction Method of Multipliers (ADMM) implementation of an extension of the robust ℓ_1 -BranchHull program (6). The use of ADMM in signal processing and machine learning applications is motivated by several features that it offers, including distributability, robustness, speed and ease of implementation [32]. As we see, using an ADMM framework allows carrying out the optimization by a chain of low cost and easy-to-implement operations. The ADMM implementation of the ℓ_1 -BranchHull program (4) is similar to the ADMM implementation of (8) and we leave it to the readers. The extension of the robust ℓ_1 -BranchHull program we consider is

$$\begin{aligned} & \underset{\mathbf{h} \in \mathbb{R}^K, \mathbf{m} \in \mathbb{R}^N, \boldsymbol{\xi} \in \mathbb{R}^L}{\text{minimize}} \quad \|\mathbf{P}\mathbf{h}\|_1 + \|\mathbf{m}\|_1 + \lambda \|\boldsymbol{\xi}\|_1 \\ & \text{subject to} \quad s_\ell(\xi_\ell + \mathbf{c}_\ell^\top \mathbf{m}) \mathbf{b}_\ell^\top \mathbf{h} \geq |y_\ell| \\ & \quad t_\ell \mathbf{b}_\ell^\top \mathbf{h} \geq 0, \quad \ell = 1, 2, \dots, L, \end{aligned} \quad (8)$$

where $\mathbf{P} \in \mathbb{R}^{J \times K}$. The above extension reduces to the robust ℓ_1 -BranchHull program if $\mathbf{P} = \mathbf{I}_K$. Recalling that $\mathbf{w} = \mathbf{B}\mathbf{h}$ and $\mathbf{x} = \mathbf{C}\mathbf{m}$, we form the block diagonal matrices $\mathbf{E} = \text{diag}(\mathbf{C}, \mathbf{B}, \lambda^{-1} \mathbf{I}_L)$ and $\mathbf{Q} = \text{diag}(\mathbf{I}_N, \mathbf{P}, \mathbf{I}_L)$ and define the vectors

$$\mathbf{u} = \begin{pmatrix} \mathbf{x} \\ \mathbf{w} \\ \boldsymbol{\xi} \end{pmatrix}, \quad \mathbf{v} = \begin{pmatrix} \mathbf{m} \\ \mathbf{h} \\ \lambda \boldsymbol{\xi} \end{pmatrix}.$$

Using this notation, our convex program can be compactly written as

$$\underset{\mathbf{v} \in \mathbb{R}^{N+K+L}, \mathbf{u} \in \mathbb{R}^{3L}}{\text{minimize}} \quad \|\mathbf{Q}\mathbf{v}\|_1 \quad \text{subject to} \quad \mathbf{u} = \mathbf{E}\mathbf{v}, \quad \mathbf{u} \in \mathcal{C}.$$

Here $\mathcal{C} = \{(\mathbf{x}, \mathbf{w}, \boldsymbol{\xi}) \in \mathbb{R}^{3L} \mid s_\ell(\xi_\ell + x_\ell)w_\ell \geq |y_\ell|, \quad t_\ell w_\ell \geq 0, \quad \ell = 1, \dots, L\}$ is the convex feasible set of (8). Introducing a new variable \mathbf{z} the resulting convex program can be written as

$$\underset{\mathbf{v}, \mathbf{u}, \mathbf{z}}{\text{minimize}} \quad \|\mathbf{z}\|_1 \quad \text{subject to} \quad \mathbf{u} = \mathbf{E}\mathbf{v}, \quad \mathbf{Q}\mathbf{v} = \mathbf{z}, \quad \mathbf{u} \in \mathcal{C}.$$

We may now form the scaled ADMM through the following steps

$$\mathbf{u}_{k+1} = \arg \min_{\mathbf{u} \in \mathcal{C}} \frac{\rho}{2} \|\mathbf{u} + \boldsymbol{\alpha}_k - \mathbf{E}\mathbf{v}_k\|^2 \quad (9)$$

$$\mathbf{z}_{k+1} = \arg \min_{\mathbf{z}} \|\mathbf{z}\|_1 + \frac{\rho}{2} \|\mathbf{z} + \boldsymbol{\beta}_k - \mathbf{Q}\mathbf{v}_k\|^2 \quad (10)$$

$$\begin{aligned} \mathbf{v}_{k+1} = \arg \min_{\mathbf{v}} \quad & \frac{\rho}{2} \|\boldsymbol{\alpha}_k + \mathbf{u}_{k+1} - \mathbf{E}\mathbf{v}\|^2 \\ & + \frac{\rho}{2} \|\boldsymbol{\beta}_k + \mathbf{z}_{k+1} - \mathbf{Q}\mathbf{v}\|^2, \end{aligned} \quad (11)$$

which are followed by the dual updates

$$\boldsymbol{\alpha}_{k+1} = \boldsymbol{\alpha}_k + \mathbf{u}_{k+1} - \mathbf{E}\mathbf{v}_{k+1},$$

$$\beta_{k+1} = \beta_k + \mathbf{v}_{k+1} - \mathbf{Q}\mathbf{v}_{k+1}.$$

We would like to note that the first three steps of the proposed ADMM scheme can be presented in closed form. The update in (9) is the following projection

$$\mathbf{u}_{k+1} = \text{proj}_{\mathcal{C}}(\mathbf{E}\mathbf{v}_k - \boldsymbol{\alpha}_k),$$

where $\text{proj}_{\mathcal{C}}(\mathbf{v})$ is the projection of \mathbf{v} onto \mathcal{C} . Details of computing the projection onto \mathcal{C} are presented in Section II-A. The update in (10) can be written in terms of the soft-threshold operator

$$\mathbf{z}_{k+1} = S_{1/\rho}(\mathbf{Q}\mathbf{v}_k - \beta_k),$$

where for $c > 0$,

$$(S_c(\mathbf{z}))_i = \begin{cases} z_i - c & z_i > c \\ 0 & |z_i| \leq c \\ z_i + c & z_i < -c \end{cases},$$

and $(S_c(\mathbf{z}))_i$ is the i th entry of $S_c(\mathbf{z})$. Finally, the update in (11) takes the following form

$$\mathbf{v}_{k+1} = (\mathbf{E}^\top \mathbf{E} + \mathbf{Q}^\top \mathbf{Q})^{-1} (\mathbf{E}^\top (\boldsymbol{\alpha}_k + \mathbf{u}_{k+1}) + \mathbf{Q}^\top (\beta_k + \mathbf{z}_{k+1})).$$

In our implementation of the ADMM scheme, we initialize the algorithm with the $\mathbf{v}_0 = \mathbf{0}$, $\boldsymbol{\alpha}_0 = \mathbf{0}$, $\beta_0 = \mathbf{0}$.

A. Evaluation of the Projection Operator

Given a point $(\mathbf{x}', \mathbf{w}', \boldsymbol{\xi}') \in \mathbb{R}^{3L}$, in this section we focus on deriving a closed-form expression for $\text{proj}_{\mathcal{C}}((\mathbf{x}', \mathbf{w}', \boldsymbol{\xi}'))$, where

$$\mathcal{C} = \{(\mathbf{x}, \mathbf{w}, \boldsymbol{\xi}) \in \mathbb{R}^{3L} \mid s_\ell(\xi_\ell + x_\ell)w_\ell \geq |y_\ell|, t_\ell w_\ell \geq 0, \ell = 1, \dots, L\}$$

is the convex feasible set of (8). It is straightforward to see that the resulting projection program decouples into L convex programs in \mathbb{R}^3 as

$$\arg \min_{x \in \mathbb{R}, w \in \mathbb{R}, \xi \in \mathbb{R}} \frac{1}{2} \left\| \begin{pmatrix} x \\ w \\ \xi \end{pmatrix} - \begin{pmatrix} x'_\ell \\ w'_\ell \\ \xi'_\ell \end{pmatrix} \right\|_2^2$$

subject to $|y_\ell| - s_\ell x w - s_\ell \xi w \leq 0, -t_\ell w \leq 0. \quad (12)$

Throughout this derivation we assume that $|y_\ell| > 0$ (derivation of the projection for the case y_ℓ is easy) and as a result of which the second constraint $-t_\ell w \leq 0$ is never active (because then $w = 0$ and the first constraint requires that $|y_\ell| \leq 0$). We also consistently use the fact that t_ℓ and s_ℓ are signs and nonzero.

Forming the Lagrangian as

$$\mathcal{L}(x, w, \xi, \mu_1, \mu_2) = \frac{1}{2} \left\| \begin{pmatrix} x \\ w \\ \xi \end{pmatrix} - \begin{pmatrix} x'_\ell \\ w'_\ell \\ \xi'_\ell \end{pmatrix} \right\|_2^2 + \mu_1 (|y_\ell| - s_\ell x w - s_\ell \xi w) - \mu_2 (t_\ell w).$$

Along with the primal constraints, the KKT optimality conditions are

$$\frac{\partial \mathcal{L}}{\partial x} = x - x'_\ell - \mu_1 s_\ell w = 0, \quad (13)$$

$$\frac{\partial \mathcal{L}}{\partial w} = w - w'_\ell - \mu_1 s_\ell x - \mu_1 s_\ell \xi - \mu_2 t_\ell = 0, \quad (14)$$

$$\frac{\partial \mathcal{L}}{\partial \xi} = \xi - \xi'_\ell - \mu_1 s_\ell w = 0, \quad (15)$$

$$\mu_1 \geq 0, \quad \mu_1 (|y_\ell| - s_\ell x w - s_\ell \xi w) = 0, \quad (16)$$

$$\mu_2 \geq 0, \quad \mu_2 (t_\ell w) = 0. \quad (17)$$

We now proceed with the possible cases.

Case 1. $\mu_1 = \mu_2 = 0$:

In this case we have $(x, w, \xi) = (x'_\ell, w'_\ell, \xi'_\ell)$ and this result would only be acceptable when $|y_\ell| - s_\ell x'_\ell w'_\ell - s_\ell \xi'_\ell w'_\ell \leq 0$ and $t_\ell w'_\ell \geq 0$.

Case 2. $\mu_1 = 0, t_\ell w = 0$:

In this case the first feasibility constraint of (12) requires that $|y_\ell| \leq 0$, which is not possible when $|y_\ell| > 0$.

Case 3. $|y_\ell| - s_\ell x w - s_\ell \xi w = 0, t_\ell w = 0$:

Similar to the previous case, this cannot happen when $|y_\ell| > 0$.

Case 4. $\mu_2 = 0, |y_\ell| - s_\ell x w - s_\ell \xi w = 0$:

In this case we have

$$|y_\ell| = s_\ell x w + s_\ell \xi w.$$

Now combining this observation with (13) and (15) yields

$$|y_\ell| = s_\ell (x'_\ell + \mu_1 s_\ell w) w + s_\ell (\xi'_\ell + \mu_1 s_\ell w) w, \quad (18)$$

and therefore

$$\mu_1 = \frac{|y_\ell| - s_\ell (x'_\ell + \xi'_\ell) w}{2 w^2}. \quad (19)$$

Similarly, (14) yields

$$w = w'_\ell + \mu_1 s_\ell (x'_\ell + \mu_1 s_\ell w) + \mu_1 s_\ell (\xi'_\ell + \mu_1 s_\ell w). \quad (20)$$

Knowing that $w \neq 0$, μ_1 can be eliminated between (18) and (20) to generate the following fourth order polynomial equation in terms of w :

$$2 w^4 - 2 w'_\ell w^3 + s_\ell |y_\ell| (x'_\ell + \xi'_\ell) w - y_\ell^2 = 0.$$

After solving this 4-th order polynomial equation (e.g., the root command in MATLAB) we pick the real root w which obeys

$$t_\ell w \geq 0, \quad |y_\ell| - s_\ell (x'_\ell + \xi'_\ell) w \geq 0. \quad (21)$$

Note that the second inequality in (21) warrants non-negative values for μ_1 thanks to (19). After picking the right root, we can explicitly obtain μ_1 using (20) and calculate the solutions x and ξ using (13) and (15). Technically, in using the ADMM scheme for each ℓ we solve a fourth-order polynomial equation and find the projection.

III. NUMERICAL EXPERIMENTS

In this section, we provide numerical experiments on synthetic random data and natural images where the signals follow the model in (1). The first experiment shows a phase portrait that

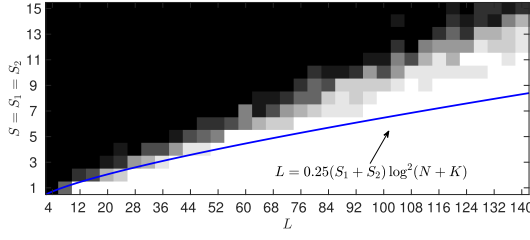


Fig. 2. The empirical recovery probability from synthetic data in the noiseless case with sparsity level S as a function of total number of measurements L . Each block correspond to the average from 10 independent trials. White blocks correspond to successful recovery and black blocks correspond to unsuccessful recovery. The area to the right of the line satisfies $L > 0.25(S_1 + S_2) \log^2(N + K)$.

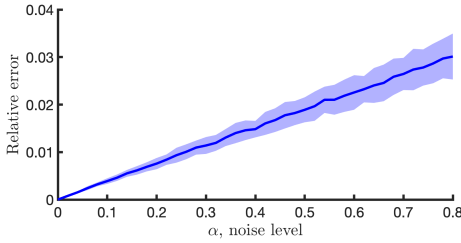


Fig. 3. The mean and standard deviation of relative error from 20 independent trials for different noise.

verifies Theorem 1 in the noiseless case. The second experiment shows that ℓ_1 -BranchHull program is robust to small dense noise. The third experiment compares the performance of ℓ_1 -BranchHull and Sparse Power Factorization (SPF), introduced in [19], in the noiseless and a noisy case. Lastly, we show that the total variation version (7) of robust ℓ_1 -BranchHull program can effectively remove distortion in real images. In these numerical experiments, we solve (4) and (7) using an ADMM implementation, as detailed in Section II.

The measurement details for numerical experiments on synthetic random data are as follows: fix $N, K, L \in \mathbb{Z}_+$ and noise level $\alpha \in [0, 1]$. Let the target signal $(\mathbf{h}^\natural, \mathbf{m}^\natural) \in \mathbb{R}^K \times \mathbb{R}^N$ be such that both \mathbf{h}^\natural and \mathbf{m}^\natural have $S = 0.05N$ non-zero entries with the non-zero indices randomly selected and set to ± 1 . Let $\mathbf{B} \in \mathbb{R}^{L \times K}$, $\mathbf{C} \in \mathbb{R}^{L \times N}$, and $\boldsymbol{\xi} \in \mathbb{R}^L$ such that $B_{ij} \sim \mathcal{N}(0, 1)$, $C_{ij} \sim \mathcal{N}(0, 1)$, and $\xi_i \sim \mathcal{N}(0, 1)$. Lastly, let $\mathbf{y} = \mathbf{B}\mathbf{h}^\natural \odot \mathbf{C}\mathbf{m}^\natural \odot (\mathbf{1} + \alpha \frac{\boldsymbol{\xi}}{\|\boldsymbol{\xi}\|_2})$ and $\mathbf{t} = \text{sign}(\mathbf{B}\mathbf{h}^\natural)$.

For the first experiment, fix $N \in \{20, 40, \dots, 300\}$, $L \in \{4, 8, \dots, 140\}$, $\alpha = 0$ and let $K = N$. Let $(\hat{\mathbf{h}}, \hat{\mathbf{m}})$ be the output of the ℓ_1 -BranchHull program (4) and let $(\tilde{\mathbf{h}}, \tilde{\mathbf{m}})$ be the candidate minimizer defined in (5). For each trial, we say (4) successfully recovers the target signal if $\|(\hat{\mathbf{h}}, \hat{\mathbf{m}}) - (\tilde{\mathbf{h}}, \tilde{\mathbf{m}})\|_2 < 10^{-8}$. Fig. 2 shows the fraction of successful recoveries from 10 independent trials using (4) for the bilinear inverse problem (1) from data as described above. Black squares correspond to no successful recovery and white squares correspond to 100% successful recovery. The line corresponds to $L = C(S_1 + S_2) \log^2(K + N)$ with $C = 0.25$ and indicates that the sample complexity constant in Theorem 1, in the noiseless case, is not very large in practice.

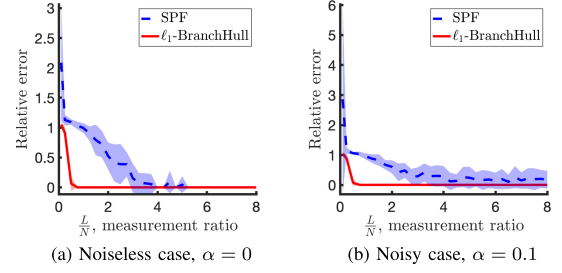


Fig. 4. Panel (a) compares the performance of ℓ_1 -BranchHull and the SPF algorithm in the noiseless case. For different measurement ratios, relative error from 20 independent trials are shown. Panel (b) shows the dependence of relative error on the measurement ratio in a noisy case of $\alpha = 0.1$.

For the second experiment, fix $N = K = 100$, $L = 150$, and noise level $\alpha \in \{0, 0.02, 0.04, \dots, 0.8\}$. Fig. 3 shows the empirical dependence of the relative error, with respect to the ℓ_2 -norm from 20 independent trials. The line corresponds to the mean relative error and the shaded region corresponds to the standard deviation. The figure confirms that ℓ_1 -BranchHull program (4) is robust to small dense noise. For the third experiment, fix $N = K = 100$, $L \in \{10, 25, 50, 75, \dots, 800\}$, and noise level $\alpha = 0$ or 1. Figs. 4(a) and 4(b) show the relative error, with respect to the Frobenius norm, from 20 independent trials using ℓ_1 -BranchHull program (4) and the SPF algorithm in the noiseless case and a noisy case. The SPF algorithm, introduced in [19], is an alternating minimization based algorithm that can estimate a rank-1 and sparse matrix $\mathbf{X}^\natural = \mathbf{h}^\natural \mathbf{m}^{\natural\top}$ from noisy linear measurement. The measurement $\mathbf{y} \in \mathbb{R}^L$ considered in [19] is $y_\ell = \langle \mathbf{X}^\natural, \mathbf{M}_\ell \rangle \cdot (1 + \xi_\ell)$, where $\mathbf{M}_\ell \in \mathbb{R}^{K \times N}$ are known L measurement matrices, $\boldsymbol{\xi}$ is noise, and $\langle \cdot, \cdot \rangle$ is the trace inner product operator. In the context of the bilinear problem (1) considered in this article, these measurements matrices are $\mathbf{M}_\ell = \mathbf{b}_\ell \mathbf{c}_\ell^\top$. Fig. 4(a) shows that ℓ_1 -BranchHull can exactly recover the unknown signals in the noiseless case with fewer number of measurements than the SPF algorithm. Following the recovery trend in the noiseless case, Fig. 4(b) shows that ℓ_1 -BranchHull outperforms SPF in a noisy case as well. This difference in performance is likely due to the fact that ℓ_1 -BranchHull is a convex program that is not sensitive to improper initialization. Also, the ℓ_1 -BranchHull program is posed in the natural parameter space and is able to exploit the rank-1 structure present in the measurement matrices \mathbf{M}_ℓ .

We now show the result of using the total variation BranchHull program (7) to remove distortions from real images $\tilde{\mathbf{y}} \in \mathbb{R}^{p \times q}$. In the experiments, The observation $\mathbf{y} \in \mathbb{R}^L$ is the column-wise vectorization of the image $\tilde{\mathbf{y}}$, the target signal $\mathbf{w} = \mathbf{B}\mathbf{h}$ is the vectorization of the piecewise constant image and $\mathbf{x} = \mathbf{C}\mathbf{m}$ corresponds to the distortions in the image. We use (7) to recover piecewise constant target images like in the foreground of Fig. 5(a) with $\text{TV}(\mathbf{B}\mathbf{h}) = \|\mathbf{D}\mathbf{B}\mathbf{h}\|_1$, where $\mathbf{D} = [\mathbf{D}_v; \mathbf{D}_h]$ in block form. Here, $\mathbf{D}_v \in \mathbb{R}^{(L-q) \times L}$ and $\mathbf{D}_h \in \mathbb{R}^{(L-p) \times L}$ with

$$(\mathbf{D}_v)_{ij} = \begin{cases} -1 & \text{if } j = i + \left(\frac{i-1}{p-1}\right)_- \\ 1 & \text{if } j = i + 1 + \left(\frac{i-1}{p-1}\right)_- \\ 0 & \text{otherwise} \end{cases},$$

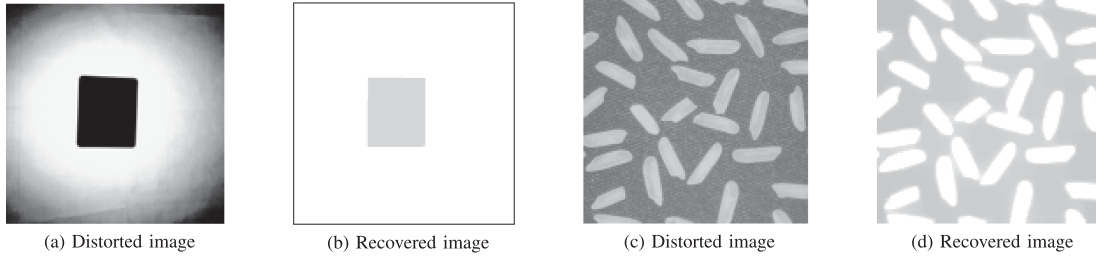


Fig. 5. Panel (a) shows an image of a mousepad with distortions and panel (b) is the piecewise constant image recovered using total variation ℓ_1 -BranchHull. Similarly, panel (d) shows an image containing rice grains and panel (e) is the recovered image.

$$(D_h)_{ij} = \begin{cases} -1 & \text{if } j = i \\ 1 & \text{if } j = i + p \\ 0 & \text{otherwise} \end{cases}$$

Lastly, we solve (7) using the ADMM algorithm detailed in Section II with $P = DB$.

We use the total variation BranchHull program on two real images. The first image, shown in Fig. 5(a), was captured using a camera and resized to a 115×115 image. The measurement $y \in \mathbb{R}^L$ is the vectorization of the image with $L = 13225$. Let B be the $L \times L$ identity matrix. Let F be the $L \times L$ inverse DCT matrix. Let $C \in \mathbb{R}^{L \times 300}$ with the first column set to 1 and remaining columns randomly selected from columns of F without replacement. The matrix C is scaled so that $\|C\|_F = \|B\|_F = \sqrt{L}$. The vector of known sign t is set to 1. Let $(\hat{h}, \hat{m}, \hat{\xi})$ be the output of (7) with $\lambda = 10^3$ and $\rho = 10^{-4}$. Fig. 5(b) corresponds to $B\hat{h}$ and shows that the object in the center was successfully recovered.

The second real image, shown in Fig. 5(c), is an image of rice grains. The size of the image is 128×128 . The measurement $y \in \mathbb{R}^L$ is the vectorization of the image with $L = 16384$. Let B be the $L \times L$ identity matrix. Let $C \in \mathbb{R}^{L \times 50}$ with the first column set to 1. The remaining columns of C are sampled from Bessel function of the first kind $J_\nu(\gamma)$ with each column corresponding to a fixed $\gamma \in \mathbb{R}$. Specifically, fix $g \in \mathbb{R}^L$ with $g_i = -9 + 14 \frac{i-1}{L-1}$. For each remaining column c of C , fix $\zeta \sim \mathcal{N}(0, I_3)$ and let $c_i = J_{\frac{g_i}{6+0.1|\zeta_1|}+5|\zeta_2|}(0.1+10|\zeta_3|)$. The matrix C is scaled so that $\|C\|_F = \|B\|_F = \sqrt{L}$. The vector of known sign t is set to 1. Let $(\hat{h}, \hat{m}, \hat{\xi})$ be the output of (7) with $\lambda = 10^3$ and $\rho = 10^{-7}$. Fig. 5(d) corresponds to $B\hat{h}$.

IV. PROOF OUTLINE

In this section, we provide a proof of Theorem 1 by considering a program similar to ℓ_1 -BranchHull program (4) with a different representation of the constraint set. Let $(\hat{w}_\ell, \hat{x}_\ell) = (b_\ell^\top \hat{h}, c_\ell^\top \hat{m})$, $\alpha_\ell = (|\hat{w}_\ell| + |\hat{x}_\ell|)/2$ and define a convex function

$$f(w_\ell, x_\ell) = \gamma(w_\ell, x_\ell) \left(\sqrt{4|y_\ell| + (w_\ell - s_\ell x_\ell)^2} - t_\ell(w_\ell + s_\ell x_\ell) \right) \quad (22)$$

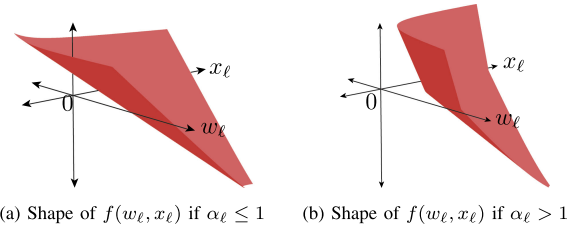


Fig. 6. Panels (a) and (b) shows the shape of the convex function $f(w_\ell, x_\ell)$ for $\alpha_\ell \leq 1$ and $\alpha_\ell > 1$, respectively. When $\alpha_\ell \leq 1$, $f(w_\ell, x_\ell)$ is differentiable everywhere. When $\alpha_\ell > 1$, $f(w_\ell, x_\ell)$ is not differentiable at (w_ℓ, x_ℓ) with $f(w_\ell, x_\ell) = 0$.

where $\gamma_\ell : \mathbb{R}^2 \rightarrow \mathbb{R}_{>0}$ is a piecewise constant function such that

$$\gamma(w_\ell, x_\ell) = \begin{cases} 1, & \text{if } \alpha_\ell > 1 \text{ and} \\ & \sqrt{4|y_\ell| + (w_\ell - s_\ell x_\ell)^2} - t_\ell(w_\ell + s_\ell x_\ell) \leq 0 \\ \alpha_\ell, & \text{otherwise.} \end{cases} \quad (23)$$

In order to see that f is convex, we first note that the function $\tilde{f}(w_\ell, x_\ell) = \sqrt{4|y_\ell| + (w_\ell - s_\ell x_\ell)^2} - t_\ell(w_\ell + s_\ell x_\ell)$ is convex. This is because the Hessian of \tilde{f} , given by

$$\nabla^2 \tilde{f}(w_\ell, x_\ell) = \frac{4}{(4|y_\ell| + (w_\ell - s_\ell x_\ell)^2)^{\frac{3}{2}}} \begin{bmatrix} |y_\ell| & -y_\ell \\ -y_\ell & |y_\ell| \end{bmatrix},$$

is positive semidefinite for all x_ℓ, w_ℓ . So, in the case where $0 < \alpha_\ell \leq 1$, $f(w_\ell, x_\ell) = \alpha_\ell \tilde{f}(w_\ell, x_\ell)$ is convex. In the case where $\alpha > 1$, the function $f(w_\ell, x_\ell)$ can be expressed as $\max\{\tilde{f}(w_\ell, x_\ell), \alpha_\ell \tilde{f}(w_\ell, x_\ell)\}$. Since, f is the point-wise maximum of two convex function, it itself is convex. Figs. 6(a) and 6(b) show typical $f(w_\ell, x_\ell)$ for $\alpha_\ell \leq 1$ and $\alpha_\ell > 1$, respectively.

Let $f_\ell(h, m) = f(b_\ell^\top h, c_\ell^\top m)$ with $\gamma_\ell(h, m) := \gamma(b_\ell^\top h, c_\ell^\top m)$. We note that $f_\ell : \mathbb{R}^{K+N} \rightarrow \mathbb{R}$ is also a convex function because its epigraph is a convex set. The epigraph is convex because it is the inverse image of a convex set over a linear map. Define a one-sided loss function

$$\mathcal{L}(h, m) := \frac{1}{L} \sum_{\ell=1}^L [f_\ell(h, m)]_+,$$

where $[\cdot]_+$ denotes the positive side. We analyze the following generalized version of the ℓ_1 -BranchHull program:

$$\ell_1\text{-GBH} : \underset{h \in \mathbb{R}^K, m \in \mathbb{R}^N}{\text{minimize}} \|h\|_1 + \|m\|_1$$

$$\text{subject to } \mathcal{L}(\mathbf{h}, \mathbf{m}) \leq 0. \quad (24)$$

Program (24) is equivalent to the ℓ_1 -BranchHull in the sense that the objective and the constraint set of both the programs are the same. Lemma 1 shows that the set defined by constraints $s_\ell(\mathbf{b}_\ell^\top \mathbf{h} \mathbf{c}_\ell^\top \mathbf{m}) \geq |y_\ell|$ with $t_\ell \mathbf{b}_\ell^\top \mathbf{h} \geq 0$ and the set defined by constraints $f_\ell(\mathbf{h}, \mathbf{m}) \leq 0$ are the same set.

Lemma 1: Fix $(\mathbf{h}^\dagger, \mathbf{m}^\dagger) \in \mathbb{R}^{K+N}$ such that $\mathbf{h}^\dagger \neq 0$ and $\mathbf{m}^\dagger \neq 0$. Let $\mathbf{B} \in \mathbb{R}^{L \times K}$, $\mathbf{C} \in \mathbb{R}^{L \times N}$ and $\boldsymbol{\xi} \in \mathbb{R}^L$. Let $\mathbf{y} \in \mathbb{R}^L$ contain measurements that satisfy (1). The set $\{(\mathbf{h}, \mathbf{m}) \in \mathbb{R}^{K+N} : s_\ell(\mathbf{b}_\ell^\top \mathbf{h} \mathbf{c}_\ell^\top \mathbf{m}) \geq |y_\ell|, t_\ell \mathbf{b}_\ell^\top \mathbf{h} \geq 0, \ell \in [L]\}$ is equal to the set $\{(\mathbf{h}, \mathbf{m}) \in \mathbb{R}^{K+N} : \mathcal{L}(\mathbf{h}, \mathbf{m}) \leq 0\}$.

Proof: Fix an $\ell \in \{1, \dots, L\}$. It is sufficient to show that the set $\mathcal{S}_{\ell,1} := \{(\mathbf{h}, \mathbf{m}) \in \mathbb{R}^{K+N} : s_\ell(\mathbf{b}_\ell^\top \mathbf{h} \mathbf{c}_\ell^\top \mathbf{m}) \geq |y_\ell|, t_\ell \mathbf{b}_\ell^\top \mathbf{h} \geq 0\}$ is equal to the set $\mathcal{S}_{\ell,2} := \{(\mathbf{h}, \mathbf{m}) \in \mathbb{R}^{K+N} : f_\ell(\mathbf{h}, \mathbf{m}) \leq 0\}$. Consider a $(\mathbf{h}, \mathbf{m}) \in \mathcal{S}_{\ell,1}$. We have

$$\begin{aligned} |y_\ell| &\leq s_\ell \mathbf{b}_\ell^\top \mathbf{h} \mathbf{c}_\ell^\top \mathbf{m}, \\ \Leftrightarrow 4|y_\ell| + (\mathbf{b}_\ell^\top \mathbf{h} - s_\ell \mathbf{c}_\ell^\top \mathbf{m})^2 &\leq (\mathbf{b}_\ell^\top \mathbf{h} + s_\ell \mathbf{c}_\ell^\top \mathbf{m})^2, \\ \Leftrightarrow \sqrt{4|y_\ell| + (\mathbf{b}_\ell^\top \mathbf{h} - s_\ell \mathbf{c}_\ell^\top \mathbf{m})^2} &\leq t_\ell(\mathbf{b}_\ell^\top \mathbf{h} + s_\ell \mathbf{c}_\ell^\top \mathbf{m}), \end{aligned} \quad (25)$$

$$\Leftrightarrow \gamma_\ell(\mathbf{h}, \mathbf{m}) \left(\sqrt{4|y_\ell| + (\mathbf{b}_\ell^\top \mathbf{h} - s_\ell \mathbf{c}_\ell^\top \mathbf{m})^2} - t_\ell(\mathbf{b}_\ell^\top \mathbf{h} + s_\ell \mathbf{c}_\ell^\top \mathbf{m}) \right) \leq 0, \quad (26)$$

where (25) holds because $t_\ell \mathbf{b}_\ell^\top \mathbf{h} \geq 0$ and $t_\ell s_\ell \mathbf{c}_\ell^\top \mathbf{m} \geq 0$ and (26) holds because $\gamma_\ell(\mathbf{h}, \mathbf{m}) \geq 0$. Thus, $\mathcal{S}_{\ell,1} \subseteq \mathcal{S}_{\ell,2}$. Now, consider a $(\mathbf{h}, \mathbf{m}) \in \mathcal{S}_{\ell,2}$. W.L.O.G. assume $\gamma_\ell(\mathbf{h}, \mathbf{m}) \neq 0$. The reverse implications above implies $|y_\ell| \leq s_\ell \mathbf{b}_\ell^\top \mathbf{h} \mathbf{c}_\ell^\top \mathbf{m}$. Also, $t_\ell(\mathbf{b}_\ell^\top \mathbf{h} + s_\ell \mathbf{c}_\ell^\top \mathbf{m}) \geq 0$ because $4|y_\ell| + (\mathbf{b}_\ell^\top \mathbf{h} - s_\ell \mathbf{c}_\ell^\top \mathbf{m})^2 \geq 0$. If $\mathbf{b}_\ell^\top \mathbf{h} = 0$, then $t_\ell \mathbf{b}_\ell^\top \mathbf{h} \geq 0$. If $\mathbf{b}_\ell^\top \mathbf{h} \neq 0$, then

$$t_\ell \mathbf{b}_\ell^\top \mathbf{h} \left(1 + \frac{s_\ell \mathbf{c}_\ell^\top \mathbf{m}}{\mathbf{b}_\ell^\top \mathbf{h}} \right) = t_\ell(\mathbf{b}_\ell^\top \mathbf{h} + s_\ell \mathbf{c}_\ell^\top \mathbf{m}) \geq 0.$$

So, $t_\ell \mathbf{b}_\ell^\top \mathbf{h} \geq 0$ because $s_\ell \mathbf{b}_\ell^\top \mathbf{h} \mathbf{c}_\ell^\top \mathbf{m} \geq |y_\ell|$ implies $(1 + \frac{s_\ell \mathbf{c}_\ell^\top \mathbf{m}}{\mathbf{b}_\ell^\top \mathbf{h}}) \geq 0$. Thus, $\mathcal{S}_{\ell,2} \subseteq \mathcal{S}_{\ell,1}$ as well, which proves that $\mathcal{S}_{\ell,1} = \mathcal{S}_{\ell,2}$. ■.

We will first show that if the noise $\boldsymbol{\xi}$ in the problem statement (1) satisfy $\xi_\ell \in [-1, 0]$ for all $\ell \in \{1, \dots, L\}$, then the ℓ_1 -Generalized BranchHull program (24) recovers a point close to the set $\{(c\hat{\mathbf{h}}, c^{-1}\hat{\mathbf{m}}) | c > 0\}$. We then extend the result to noise that satisfy condition (3). Since the ℓ_1 -Generalized BranchHull program (24) and the ℓ_1 -BranchHull program (4) are equivalent, the minimizer of ℓ_1 -BranchHull program is then also close to the set $\{(c\hat{\mathbf{h}}, c^{-1}\hat{\mathbf{m}}) | c > 0\}$. Our strategy will be to show that for any feasible perturbation $(\delta\mathbf{h}, \delta\mathbf{m}) \in \mathbb{R}^{K+N}$, the objective of the ℓ_1 -Generalized BranchHull program (24) strictly increases outside a curved cylinder of radius r centered at the bilinear ambiguity curve $\{(c\hat{\mathbf{h}}, c^{-1}\hat{\mathbf{m}}) | c > 0\}$, where the radius depends on the level of noise.

The subgradient of the ℓ_1 -norm at $(\hat{\mathbf{h}}, \hat{\mathbf{m}})$ is

$$\partial\|(\hat{\mathbf{h}}, \hat{\mathbf{m}})\|_1 := \left\{ \mathbf{g} \in \mathbb{R}^{K+N} : \begin{aligned} \|\mathbf{g}\|_\infty &\leq 1, \\ \mathbf{g}_{\Gamma_h} &= \text{sign}(\hat{\mathbf{h}}_{\Gamma_h}), \\ \mathbf{g}_{\Gamma_m} &= \text{sign}(\hat{\mathbf{m}}_{\Gamma_m}) \end{aligned} \right\},$$

where Γ_h , and Γ_m denote the support of non-zeros in $\hat{\mathbf{h}}$, and $\hat{\mathbf{m}}$, respectively. We first consider the following descent direction that are orthogonal to the set $\mathcal{N} := \text{span}(-\hat{\mathbf{h}}, \hat{\mathbf{m}})$

$$\begin{aligned} &\left\{ (\delta\mathbf{h}, \delta\mathbf{m}) \in \mathcal{N}_\perp : \begin{aligned} \langle \mathbf{g}, (\delta\mathbf{h}, \delta\mathbf{m}) \rangle &\leq 0, \\ \forall \mathbf{g} \in \partial\|(\hat{\mathbf{h}}, \hat{\mathbf{m}})\|_1 \end{aligned} \right\} \\ &\subseteq \left\{ (\delta\mathbf{h}, \delta\mathbf{m}) \in \mathcal{N}_\perp : \begin{aligned} \langle \mathbf{g}_{\Gamma_h}, \delta\mathbf{h}_{\Gamma_h} \rangle + \langle \mathbf{g}_{\Gamma_m}, \delta\mathbf{m}_{\Gamma_m} \rangle \\ + \|\delta\mathbf{h}_{\Gamma_h^c}, \delta\mathbf{m}_{\Gamma_m^c}\|_1 &\leq 0 \end{aligned} \right\} \\ &\subseteq \left\{ (\delta\mathbf{h}, \delta\mathbf{m}) \in \mathcal{N}_\perp : \begin{aligned} \|\delta\mathbf{h}_{\Gamma_h^c}, \delta\mathbf{m}_{\Gamma_m^c}\|_1 &\leq \\ \|\mathbf{g}_{\Gamma_h \cup \Gamma_m}\|_2 \|\delta\mathbf{h}_{\Gamma_h}, \delta\mathbf{m}_{\Gamma_m}\|_2 \end{aligned} \right\} \\ &= \left\{ (\delta\mathbf{h}, \delta\mathbf{m}) \in \mathcal{N}_\perp : \begin{aligned} \|\delta\mathbf{h}_{\Gamma_h^c}, \delta\mathbf{m}_{\Gamma_m^c}\|_1 &\leq \\ \sqrt{S_1 + S_2} \|\delta\mathbf{h}_{\Gamma_h}, \delta\mathbf{m}_{\Gamma_m}\|_2 \end{aligned} \right\} \\ &=: \mathcal{D} \end{aligned} \quad (27)$$

and show that descent direction from $(\hat{\mathbf{h}}, \hat{\mathbf{m}}) \oplus \mathcal{N}$ of large ℓ_2 norm is not feasible in (24). We do this by quantifying the “width” of the set \mathcal{D} through a Rademacher complexity, and a probability that one of the subgradients of the constraint functions lie in a certain half space. In the noiseless case, we show that the solution of (24) is in the set $(\hat{\mathbf{h}}, \hat{\mathbf{m}}) \oplus \mathcal{N}$. Since the only point in the set $(\hat{\mathbf{h}}, \hat{\mathbf{m}}) \oplus \mathcal{N}$ consistent with the constraint of (24) is $(\hat{\mathbf{h}}, \hat{\mathbf{m}})$, the minimizer of (24), in the noiseless case, is then $(\hat{\mathbf{h}}, \hat{\mathbf{m}})$. In the noisy case, we use the boundedness of the feasible directions from the line $(\hat{\mathbf{h}}, \hat{\mathbf{m}}) \oplus \mathcal{N}$ along with the observation that the feasible hyperbolic set diverges away from $(\hat{\mathbf{h}}, \hat{\mathbf{m}}) \oplus \mathcal{N}$ to conclude the solution of minimizer of (24) is close to the bilinear ambiguity curve $\{(c\hat{\mathbf{h}}, c^{-1}\hat{\mathbf{m}}) | c > 0\}$.

Recall that the constraint set of the ℓ_1 -BranchHull (4) has L hyperbolic constraints and L linear constraints. Let $\hat{\mathbf{y}} = \mathbf{B}\hat{\mathbf{h}} \odot \mathbf{C}\hat{\mathbf{m}}$ be the noiseless data. Using Lemma 1, each pair of hyperbolic and linear constraint $\{(\mathbf{h}, \mathbf{m}) : s_\ell \mathbf{b}_\ell^\top \mathbf{h} \mathbf{c}_\ell^\top \mathbf{m} \geq |y_\ell|, t_\ell \mathbf{b}_\ell^\top \mathbf{h} \geq 0\}$ can be expressed as $\hat{f}_\ell(\mathbf{h}, \mathbf{m}) \leq 0$ where

$$\begin{aligned} \hat{f}_\ell(\mathbf{h}, \mathbf{m}) &= \gamma_\ell(\mathbf{h}, \mathbf{m}) \left(\sqrt{4|\hat{y}_\ell| + (\mathbf{b}_\ell^\top \mathbf{h} - s_\ell \mathbf{c}_\ell^\top \mathbf{m})^2} \right. \\ &\quad \left. - t_\ell(\mathbf{b}_\ell^\top \mathbf{h} + s_\ell \mathbf{c}_\ell^\top \mathbf{m}) \right). \end{aligned}$$

First we note that a subgradient of $\hat{f}_\ell(\mathbf{h}, \mathbf{m})$ at $(\hat{\mathbf{h}}, \hat{\mathbf{m}})$ is $\mathbf{z}_\ell := -s_\ell(\mathbf{c}_\ell^\top \hat{\mathbf{m}} \mathbf{b}_\ell^\top, \mathbf{b}_\ell^\top \hat{\mathbf{h}} \mathbf{c}_\ell^\top)$. To see this, recall that $\alpha_\ell = (|\mathbf{b}_\ell^\top \hat{\mathbf{h}}| + |\mathbf{c}_\ell^\top \hat{\mathbf{m}}|)/2$ and let

$$\begin{aligned} \hat{g}_\ell(\mathbf{h}, \mathbf{m}) &= \alpha_\ell \left(\sqrt{4|\hat{y}_\ell| + (\mathbf{b}_\ell^\top \mathbf{h} - s_\ell \mathbf{c}_\ell^\top \mathbf{m})^2} \right. \\ &\quad \left. - t_\ell(\mathbf{b}_\ell^\top \mathbf{h} + s_\ell \mathbf{c}_\ell^\top \mathbf{m}) \right). \end{aligned}$$

When $\alpha_\ell \leq 1$, we have $\hat{f}_\ell = \hat{g}_\ell$ because $\gamma_\ell(\mathbf{h}, \mathbf{m}) = \alpha_\ell$ by definition of $\gamma(\mathbf{b}_\ell^\top \mathbf{h}, \mathbf{c}_\ell^\top \mathbf{m}) = \gamma_\ell(\mathbf{h}, \mathbf{m})$ in (23). When

$\alpha_\ell > 1$, \hat{f}_ℓ is non-differentiable at (\mathbf{h}, \mathbf{m}) where $\hat{f}_\ell(\mathbf{h}, \mathbf{m}) = 0$. Fig. 6(b) shows the shape of \hat{f}_ℓ when $\alpha_\ell > 1$. In this case, we have $\hat{f}_\ell(\hat{\mathbf{h}}, \hat{\mathbf{m}}) = 0$ and $\hat{f}_\ell(\mathbf{h}, \mathbf{m}) = \hat{g}_\ell(\mathbf{h}, \mathbf{m})$ for (\mathbf{h}, \mathbf{m}) that satisfy $\hat{f}_\ell(\mathbf{h}, \mathbf{m}) \leq 0$. Thus, $\nabla \hat{g}_\ell(\hat{\mathbf{h}}, \hat{\mathbf{m}}) \in \partial \hat{f}_\ell(\hat{\mathbf{h}}, \hat{\mathbf{m}})$ and $-s_\ell(\mathbf{c}_\ell^\top \hat{\mathbf{m}} \mathbf{b}_\ell^\top, \mathbf{b}_\ell^\top \hat{\mathbf{h}} \mathbf{c}_\ell^\top) \in \partial \hat{f}_\ell(\hat{\mathbf{h}}, \hat{\mathbf{m}})$ if $\nabla \hat{g}_\ell(\hat{\mathbf{h}}, \hat{\mathbf{m}}) = -s_\ell(\mathbf{c}_\ell^\top \hat{\mathbf{m}} \mathbf{b}_\ell^\top, \mathbf{b}_\ell^\top \hat{\mathbf{h}} \mathbf{c}_\ell^\top)$. So, consider

$$\begin{aligned} \frac{\partial \hat{g}_\ell}{\partial \mathbf{h}}(\hat{\mathbf{h}}, \hat{\mathbf{m}}) &= \alpha_\ell \left(\frac{2(\mathbf{b}_\ell^\top \hat{\mathbf{h}} - s_\ell \mathbf{c}_\ell^\top \hat{\mathbf{m}}) \mathbf{b}_\ell}{2\sqrt{4|\hat{y}_\ell| + (\mathbf{b}_\ell^\top \hat{\mathbf{h}} - s_\ell \mathbf{c}_\ell^\top \hat{\mathbf{m}})^2}} - t_\ell \mathbf{b}_\ell \right) \\ &= \alpha_\ell \left(\frac{(\mathbf{b}_\ell^\top \hat{\mathbf{h}} - s_\ell \mathbf{c}_\ell^\top \hat{\mathbf{m}}) \mathbf{b}_\ell}{\sqrt{(\mathbf{b}_\ell^\top \hat{\mathbf{h}} + s_\ell \mathbf{c}_\ell^\top \hat{\mathbf{m}})^2}} - t_\ell \mathbf{b}_\ell \right) \\ &= \alpha_\ell \left(\frac{(\mathbf{b}_\ell^\top \hat{\mathbf{h}} - s_\ell \mathbf{c}_\ell^\top \hat{\mathbf{m}}) \mathbf{b}_\ell}{|t_\ell(\mathbf{b}_\ell^\top \hat{\mathbf{h}} + s_\ell \mathbf{c}_\ell^\top \hat{\mathbf{m}})|} - t_\ell \mathbf{b}_\ell \right) \\ &= \alpha_\ell \left(\frac{(\mathbf{b}_\ell^\top \hat{\mathbf{h}} - s_\ell \mathbf{c}_\ell^\top \hat{\mathbf{m}}) \mathbf{b}_\ell - (\mathbf{b}_\ell^\top \hat{\mathbf{h}} + s_\ell \mathbf{c}_\ell^\top \hat{\mathbf{m}}) \mathbf{b}_\ell}{t_\ell(\mathbf{b}_\ell^\top \hat{\mathbf{h}} + s_\ell \mathbf{c}_\ell^\top \hat{\mathbf{m}})} \right) \\ &= \alpha_\ell \left(\frac{(\mathbf{b}_\ell^\top \hat{\mathbf{h}} - s_\ell \mathbf{c}_\ell^\top \hat{\mathbf{m}}) \mathbf{b}_\ell - (\mathbf{b}_\ell^\top \hat{\mathbf{h}} + s_\ell \mathbf{c}_\ell^\top \hat{\mathbf{m}}) \mathbf{b}_\ell}{t_\ell(\mathbf{b}_\ell^\top \hat{\mathbf{h}} + s_\ell \mathbf{c}_\ell^\top \hat{\mathbf{m}})} \right) \\ &= -s_\ell \mathbf{c}_\ell^\top \hat{\mathbf{m}} \mathbf{b}_\ell, \end{aligned}$$

where the second equality follows from the fact that $|\hat{y}_\ell| = s_\ell \mathbf{b}_\ell^\top \hat{\mathbf{h}} \mathbf{c}_\ell^\top \hat{\mathbf{m}}$, third and fourth equalities holds because $t_\ell \mathbf{b}_\ell^\top \hat{\mathbf{h}} \geq 0$ and $t_\ell s_\ell \mathbf{c}_\ell^\top \hat{\mathbf{m}} \geq 0$, and the last equality holds because $\alpha_\ell = t_\ell(\mathbf{b}_\ell^\top \hat{\mathbf{h}} + s_\ell \mathbf{c}_\ell^\top \hat{\mathbf{m}})/2$. Similarly, we have $\frac{\partial \hat{g}_\ell}{\partial \mathbf{m}}(\hat{\mathbf{h}}, \hat{\mathbf{m}}) = -s_\ell \mathbf{b}_\ell^\top \hat{\mathbf{h}} \mathbf{c}_\ell$ and $\nabla \hat{g}_\ell(\hat{\mathbf{h}}, \hat{\mathbf{m}}) = -s_\ell(\mathbf{c}_\ell^\top \hat{\mathbf{m}} \mathbf{b}_\ell^\top, \mathbf{b}_\ell^\top \hat{\mathbf{h}} \mathbf{c}_\ell^\top)$. Define the Rademacher complexity of a set $\mathcal{D} \subset \mathbb{R}^{K+N}$ as

$$\mathfrak{C}(\mathcal{D}) := \mathbb{E} \sup_{(\mathbf{h}, \mathbf{m}) \in \mathcal{D}} \frac{1}{\sqrt{L}} \sum_{\ell=1}^L \varepsilon_\ell \left\langle \mathbf{z}_\ell, \frac{(\mathbf{h}, \mathbf{m})}{\|(\mathbf{h}, \mathbf{m})\|_2} \right\rangle, \quad (28)$$

where $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_L$ are iid Rademacher random variables independent of everything else. For a set \mathcal{D} , the quantity $\mathfrak{C}(\mathcal{D})$ is a measure of width of \mathcal{D} around the origin in terms of the subgradients of the constraint functions. Our results also depend on a probability $p_\tau(\mathcal{D})$, and a positive parameter τ introduced below

$$p_\tau(\mathcal{D}) = \inf_{(\mathbf{h}, \mathbf{m}) \in \mathcal{D}} \mathbb{P} \left(\left\langle \mathbf{z}_\ell, \frac{(\mathbf{h}, \mathbf{m})}{\|(\mathbf{h}, \mathbf{m})\|_2} \right\rangle \geq \tau \right). \quad (29)$$

Intuitively, $p_\tau(\mathcal{D})$ quantifies the size of \mathcal{D} through the subgradient vector. For a small enough fixed parameter, a small value of $p_\tau(\mathcal{D})$ means that the \mathcal{D} is mainly invisible to the subgradient vector.

We now state a lemma which shows that if the noise $\boldsymbol{\xi}$ is such that

$$\begin{aligned} \xi_{\ell'} &= 0, \text{ for some } \ell' \in \{1, \dots, L\}, \text{ and} \\ \xi_\ell &\in [-1, 0] \text{ for all } \ell \in \{1, \dots, L\}, \end{aligned} \quad (30)$$

and the number of measurements $L \geq \frac{2\mathfrak{C}(\mathcal{D}) + t\tau}{\tau p_\tau(\mathcal{D})}$ for any $t > 0$, then the solution of (24) is close to the bilinear ambiguity curve $\{(c\hat{\mathbf{h}}, c^{-1}\hat{\mathbf{m}}) | c > 0\}$. The Proof of this lemma is based

on small ball method developed in [33], [34] and further studied in [35]–[37].

Lemma 2: Let \mathcal{D} be the set of descent directions, already characterized in (27), for which $\mathfrak{C}(\mathcal{D})$, and $p_\tau(\mathcal{D})$ can be determined using (28), and (29). Let the noise $\boldsymbol{\xi} \in \mathbb{R}^L$ be such that (30) is satisfied. Choose $L \geq \frac{2\mathfrak{C}(\mathcal{D}) + t\tau}{\tau p_\tau(\mathcal{D})}$ for any $t > 0$. Then the solution $(\tilde{\mathbf{h}}, \tilde{\mathbf{m}})$ of the BH in (4) satisfies

$$\|(\tilde{\mathbf{h}}, \tilde{\mathbf{m}}) - (c\hat{\mathbf{h}}, c^{-1}\hat{\mathbf{m}})\|_2 \leq 36\sqrt{\|\hat{\mathbf{h}}\|_2 \|\hat{\mathbf{m}}\|_2} \sqrt{\|\boldsymbol{\xi}\|_\infty}$$

for some > 0 with probability at least $1 - e^{-2Lt^2}$. Furthermore, $(\tilde{\mathbf{h}}, \tilde{\mathbf{m}}) = (\hat{\mathbf{h}}, \hat{\mathbf{m}})$ if $\boldsymbol{\xi} = \mathbf{0}$.

Proof: Without loss of generality, we analyze the ℓ_1 -Generalized BranchHull program (24). We note that $(\hat{\mathbf{h}}, \hat{\mathbf{m}})$ is feasible in (24) because the noise satisfy $\xi_\ell \in [-1, 0]$ for all $\ell \in \{1, \dots, L\}$. We first control the set of feasible descent direction $(\delta\mathbf{h}, \delta\mathbf{m}) \in \mathcal{D}$ from the set $(\hat{\mathbf{h}}, \hat{\mathbf{m}}) \oplus \mathcal{N}$. Since $(\delta\mathbf{h}, \delta\mathbf{m})$ is a feasible perturbation from a point $(\mathbf{h}^*, \mathbf{m}^*) := ((1 - \beta)\hat{\mathbf{h}}, (1 + \beta)\hat{\mathbf{m}})$ for some $\beta \in \mathbb{R}$, we have from (24)

$$\mathcal{L}(\mathbf{h}^* + \delta\mathbf{h}, \mathbf{m}^* + \delta\mathbf{m}) \leq 0. \quad (31)$$

Note that because of (31), $f_\ell(\mathbf{h}^* + \delta\mathbf{h}, \mathbf{m}^* + \delta\mathbf{m}) \leq 0$ for all $\ell \in \{1, \dots, L\}$ since, by definition, $\mathcal{L}(\mathbf{h}, \mathbf{m}) = \frac{1}{L} \sum_{\ell=1}^L [f_\ell(\mathbf{h}, \mathbf{m})]_+$. Thus, $\gamma_\ell(\mathbf{h}^* + \delta\mathbf{h}, \mathbf{m}^* + \delta\mathbf{m})$ in $f_\ell(\mathbf{h}^* + \delta\mathbf{h}, \mathbf{m}^* + \delta\mathbf{m})$ satisfies $0 \leq \gamma_\ell(\mathbf{h}^* + \delta\mathbf{h}, \mathbf{m}^* + \delta\mathbf{m}) \leq 1$. We now expand the loss function $\mathcal{L}(\mathbf{h}, \mathbf{m})$ at $(\mathbf{h}^* + \delta\mathbf{h}, \mathbf{m}^* + \delta\mathbf{m})$. Consider

$$\begin{aligned} &[f_\ell(\mathbf{h}^* + \delta\mathbf{h}, \mathbf{m}^* + \delta\mathbf{m})]_+ \\ &= \left[\gamma_\ell \left(\left((\mathbf{b}_\ell^\top (\mathbf{h}^* + \delta\mathbf{h}) - s_\ell \mathbf{c}_\ell^\top (\mathbf{m}^* + \delta\mathbf{m}))^2 + 4s_\ell \hat{y}_\ell \right. \right. \right. \\ &\quad \left. \left. \left. + 4s_\ell \hat{y}_\ell \xi_\ell \right)^{\frac{1}{2}} - t_\ell (\mathbf{b}_\ell^\top (\mathbf{h}^* + \delta\mathbf{h}) + s_\ell \mathbf{c}_\ell^\top (\mathbf{m}^* + \delta\mathbf{m})) \right) \right]_+ \\ &\geq \left[\gamma_\ell \left(\sqrt{4|\hat{y}_\ell| + (\mathbf{b}_\ell^\top (\mathbf{h}^* + \delta\mathbf{h}) - s_\ell \mathbf{c}_\ell^\top (\mathbf{m}^* + \delta\mathbf{m}))^2} \right. \right. \\ &\quad \left. \left. - t_\ell (\mathbf{b}_\ell^\top (\mathbf{h}^* + \delta\mathbf{h}) + s_\ell \mathbf{c}_\ell^\top (\mathbf{m}^* + \delta\mathbf{m})) \right. \right. \\ &\quad \left. \left. - \sqrt{[-4s_\ell \hat{y}_\ell \xi_\ell]_+} \right) \right]_+ \end{aligned} \quad (32)$$

$$\begin{aligned} &\geq \left[\hat{f}_\ell(\hat{\mathbf{h}} - \beta\hat{\mathbf{h}} + \delta\mathbf{h}, \hat{\mathbf{m}} + \beta\hat{\mathbf{m}} + \delta\mathbf{m}) \right]_+ \\ &\quad - 2\gamma_\ell \sqrt{[-s_\ell \hat{y}_\ell \xi_\ell]_+} \\ &\geq \left[\langle \mathbf{z}_\ell, (-\beta\hat{\mathbf{h}} + \delta\mathbf{h}, \beta\hat{\mathbf{m}} + \delta\mathbf{m}) \rangle \right]_+ - 2\sqrt{[-s_\ell \hat{y}_\ell \xi_\ell]_+} \end{aligned} \quad (33)$$

$$= [\langle \mathbf{z}_\ell, (\delta\mathbf{h}, \delta\mathbf{m}) \rangle]_+ - 2\sqrt{[-s_\ell \hat{y}_\ell \xi_\ell]_+}, \quad (34)$$

where in (32) we use $\text{sign}(y_\ell) = \text{sign}(\hat{y}_\ell)$ along with the fact that for $a \geq 0$ and $b < 0$ with $a + b \geq 0$, we have $\sqrt{a+b} \geq \sqrt{a} - \sqrt{-b}$ and for $a \geq 0$ and $b \geq 0$, we have $\sqrt{a+b} \geq \sqrt{a}$. Also, (33) holds because \hat{f}_ℓ is convex with $\mathbf{z}_\ell \in \partial \hat{f}_\ell(\hat{\mathbf{h}}, \hat{\mathbf{m}})$ and $0 \leq \gamma_\ell \leq 1$. Lastly, (34) holds because $\langle \mathbf{z}_\ell, (-\hat{\mathbf{h}}, \hat{\mathbf{m}}) \rangle = 0$. Combining

(31) and (34), we get

$$\begin{aligned} \underbrace{\frac{1}{L} \sum_{\ell=1}^L [\langle \mathbf{z}_\ell, (\delta \mathbf{h}, \delta \mathbf{m}) \rangle]_+}_I &\leq \frac{2}{L} \sum_{\ell=1}^L \sqrt{\hat{y}_\ell \xi_\ell} \\ &\leq \frac{2}{L} \sum_{\ell=1}^L \sqrt{\|\mathbf{b}_\ell^\top \hat{\mathbf{h}}\| \sqrt{\|\mathbf{c}_\ell^\top \hat{\mathbf{m}}\|}} \left(\sqrt{\|\boldsymbol{\xi}\|_\infty} \right) \\ &\leq \frac{2}{L} \sqrt{\|\mathbf{B} \hat{\mathbf{h}}\|_1 \|\mathbf{C} \hat{\mathbf{m}}\|_1} \left(\sqrt{\|\boldsymbol{\xi}\|_\infty} \right) \end{aligned} \quad (35)$$

$$\begin{aligned} &\leq \frac{2}{L} \sqrt{L \|\mathbf{B}\| \|\hat{\mathbf{h}}\|_2 \|\mathbf{C}\| \|\hat{\mathbf{m}}\|_2} \sqrt{\|\boldsymbol{\xi}\|_\infty} \\ &\leq 18 \sqrt{\|\hat{\mathbf{h}}\|_2 \|\hat{\mathbf{m}}\|_2} \sqrt{\|\boldsymbol{\xi}\|_\infty}, \end{aligned} \quad (36)$$

where (35) follows from Cauchy-Schwartz inequality and (36) holds with probability $1 - e^{-cL}$ because, by Corollary 5.35 in [38], $\|\mathbf{B}\| \leq 3\sqrt{L}$ with probability $1 - 2e^{-\frac{L}{2}}$ and $\|\mathbf{C}\| \leq 3\sqrt{L}$ with probability $1 - 2e^{-\frac{L}{2}}$ as well. We now lower bound I in (36). Let $\psi_t(s) := (s)_+ - (s-t)_+$. Using the fact that $\psi_t(s) \leq (s)_+$, and that for every $\alpha, t \geq 0$, and $s \in \mathbb{R}$, $\psi_{\alpha t}(s) = t\psi_\alpha(\frac{s}{t})$, we have

$$\begin{aligned} I &\geq \frac{1}{L} \sum_{\ell=1}^L \psi_{\tau \|(\delta \mathbf{h}, \delta \mathbf{m})\|_2} (\langle \mathbf{z}_\ell, (\delta \mathbf{h}, \delta \mathbf{m}) \rangle) \\ &= \|(\delta \mathbf{h}, \delta \mathbf{m})\|_2 \cdot \frac{1}{L} \sum_{\ell=1}^L \psi_\tau \left(\left\langle \mathbf{z}_\ell, \frac{(\delta \mathbf{h}, \delta \mathbf{m})}{\|(\delta \mathbf{h}, \delta \mathbf{m})\|_2} \right\rangle \right) \\ &= \|(\delta \mathbf{h}, \delta \mathbf{m})\|_2 \cdot \frac{1}{L} \left[\sum_{\ell=1}^L \mathbb{E} \psi_\tau \left(\left\langle \mathbf{z}_\ell, \frac{(\delta \mathbf{h}, \delta \mathbf{m})}{\|(\delta \mathbf{h}, \delta \mathbf{m})\|_2} \right\rangle \right) \right. \\ &\quad \left. - \sum_{\ell=1}^L \left(\mathbb{E} \psi_\tau \left(\left\langle \mathbf{z}_\ell, \frac{(\delta \mathbf{h}, \delta \mathbf{m})}{\|(\delta \mathbf{h}, \delta \mathbf{m})\|_2} \right\rangle \right) \right) \right. \\ &\quad \left. - \psi_\tau \left(\left\langle \mathbf{z}_\ell, \frac{(\delta \mathbf{h}, \delta \mathbf{m})}{\|(\delta \mathbf{h}, \delta \mathbf{m})\|_2} \right\rangle \right) \right]. \end{aligned} \quad (37)$$

The proof mainly relies on lower bounding the right hand side above uniformly over all $(\delta \mathbf{h}, \delta \mathbf{m}) \in \mathcal{D}$. To this end, define a centered random process $\mathcal{R}(\mathbf{B}, \mathbf{C})$ as follows

$$\begin{aligned} \mathcal{R}(\mathbf{B}, \mathbf{C}) &:= \sup_{(\delta \mathbf{h}, \delta \mathbf{m}) \in \mathcal{D}} \frac{1}{L} \sum_{\ell=1}^L \left[\mathbb{E} \psi_\tau \left(\left\langle \mathbf{z}_\ell, \frac{(\delta \mathbf{h}, \delta \mathbf{m})}{\|(\delta \mathbf{h}, \delta \mathbf{m})\|_2} \right\rangle \right) \right. \\ &\quad \left. - \psi_\tau \left(\left\langle \mathbf{z}_\ell, \frac{(\delta \mathbf{h}, \delta \mathbf{m})}{\|(\delta \mathbf{h}, \delta \mathbf{m})\|_2} \right\rangle \right) \right], \end{aligned}$$

and an application of bounded difference inequality [39] yields that $\mathcal{R}(\mathbf{B}, \mathbf{C}) \leq \mathbb{E} \mathcal{R}(\mathbf{B}, \mathbf{C}) + t\tau/\sqrt{L}$ with probability at least $1 - e^{-2Lt^2}$. It remains to evaluate $\mathbb{E} \mathcal{R}(\mathbf{B}, \mathbf{C})$, which after using a simple symmetrization inequality [40] yields

$$\mathbb{E} \mathcal{R}(\mathbf{B}, \mathbf{C}) \leq 2 \mathbb{E} \sup_{(\delta \mathbf{h}, \delta \mathbf{m}) \in \mathcal{D}} \frac{1}{L} \sum_{\ell=1}^L \varepsilon_\ell \psi_\tau \left(\left\langle \mathbf{z}_\ell, \frac{(\delta \mathbf{h}, \delta \mathbf{m})}{\|(\delta \mathbf{h}, \delta \mathbf{m})\|_2} \right\rangle \right), \quad (38)$$

where $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_L$ are independent Rademacher random variables. Using the fact that $\psi_t(s)$ is a contraction: $|\psi_t(\alpha_1) - \psi_t(\alpha_2)| \leq |\alpha_1 - \alpha_2|$ for all $\alpha_1, \alpha_2 \in \mathbb{R}$, we have from the Rademacher contraction inequality [41] that

$$\begin{aligned} \mathbb{E} \sup_{(\delta \mathbf{h}, \delta \mathbf{m}) \in \mathcal{D}} \frac{1}{L} \sum_{\ell=1}^L \varepsilon_\ell \psi_\tau \left(\left\langle \mathbf{z}_\ell, \frac{(\delta \mathbf{h}, \delta \mathbf{m})}{\|(\delta \mathbf{h}, \delta \mathbf{m})\|_2} \right\rangle \right) \\ \leq \mathbb{E} \sup_{(\delta \mathbf{h}, \delta \mathbf{m}) \in \mathcal{D}} \frac{1}{L} \sum_{\ell=1}^L \varepsilon_\ell \left\langle \mathbf{z}_\ell, \frac{(\delta \mathbf{h}, \delta \mathbf{m})}{\|(\delta \mathbf{h}, \delta \mathbf{m})\|_2} \right\rangle. \end{aligned} \quad (39)$$

In addition, using the facts that $t\mathbf{1}(s \geq t) \leq \psi_t(s)$ it follows

$$\begin{aligned} \mathbb{E} \psi_\tau \left(\left\langle \mathbf{z}_\ell, \frac{(\delta \mathbf{h}, \delta \mathbf{m})}{\|(\delta \mathbf{h}, \delta \mathbf{m})\|_2} \right\rangle \right) &\geq \tau \mathbb{E} \left[\mathbf{1} \left(\left\langle \mathbf{z}_\ell, \frac{(\delta \mathbf{h}, \delta \mathbf{m})}{\|(\delta \mathbf{h}, \delta \mathbf{m})\|_2} \right\rangle \geq \tau \right) \right] \\ &= \tau \mathbb{P} \left(\left\langle \mathbf{z}_\ell, \frac{(\delta \mathbf{h}, \delta \mathbf{m})}{\|(\delta \mathbf{h}, \delta \mathbf{m})\|_2} \right\rangle \geq \tau \right) \end{aligned} \quad (40)$$

Plugging (40), and (39) in (37), we have

$$\begin{aligned} I &\geq \tau \|(\delta \mathbf{h}, \delta \mathbf{m})\|_2 \mathbb{P} \left(\left\langle \mathbf{z}_\ell, \frac{(\delta \mathbf{h}, \delta \mathbf{m})}{\|(\delta \mathbf{h}, \delta \mathbf{m})\|_2} \right\rangle \geq \tau \right) \\ &\quad - \|(\delta \mathbf{h}, \delta \mathbf{m})\|_2 \left(2 \mathbb{E} \sup_{(\delta \mathbf{h}, \delta \mathbf{m}) \in \mathcal{D}} \frac{1}{L} \sum_{\ell=1}^L \varepsilon_\ell \left\langle \mathbf{z}_\ell, \frac{(\delta \mathbf{h}, \delta \mathbf{m})}{\|(\delta \mathbf{h}, \delta \mathbf{m})\|_2} \right\rangle \right. \\ &\quad \left. + \frac{t\tau}{\sqrt{L}} \right) \end{aligned} \quad (41)$$

Combining this with (36), we obtain the final result

$$\begin{aligned} \|(\delta \mathbf{h}, \delta \mathbf{m})\|_2 &\left[\tau \mathbb{P} \left(\left\langle \mathbf{z}_\ell, \frac{(\delta \mathbf{h}, \delta \mathbf{m})}{\|(\delta \mathbf{h}, \delta \mathbf{m})\|_2} \right\rangle \geq \tau \right) \right. \\ &\quad \left. - \left(2 \mathbb{E} \sup_{(\delta \mathbf{h}, \delta \mathbf{m}) \in \mathcal{D}} \frac{1}{L} \sum_{\ell=1}^L \varepsilon_\ell \left\langle \mathbf{z}_\ell, \frac{(\delta \mathbf{h}, \delta \mathbf{m})}{\|(\delta \mathbf{h}, \delta \mathbf{m})\|_2} \right\rangle + \frac{t\tau}{\sqrt{L}} \right) \right] \\ &\leq 18 \sqrt{\|\hat{\mathbf{h}}\|_2 \|\hat{\mathbf{m}}\|_2} \sqrt{\|\boldsymbol{\xi}\|_\infty} \end{aligned}$$

Using the definitions in (28), and (29), we can write

$$\begin{aligned} \|(\delta \mathbf{h}, \delta \mathbf{m})\|_2 &\left(\tau \mathbf{p}_\tau(\mathcal{D}) - \frac{(2\mathfrak{C}(\mathcal{D}) + t\tau)}{\sqrt{L}} \right) \\ &\leq 18 \sqrt{\|\hat{\mathbf{h}}\|_2 \|\hat{\mathbf{m}}\|_2} \sqrt{\|\boldsymbol{\xi}\|_\infty}. \end{aligned}$$

It is clear that choosing $L \geq \left(\frac{2\mathfrak{C}(\mathcal{D}) + t\tau}{\tau \mathbf{p}_\tau(\mathcal{D})} \right)^2$ implies the any feasible descent direction from $(\hat{\mathbf{h}}, \hat{\mathbf{m}}) \oplus \mathcal{N}$ is bounded by

$$\|(\delta \mathbf{h}, \delta \mathbf{m})\|_2 \leq 18 \sqrt{\|\hat{\mathbf{h}}\|_2 \|\hat{\mathbf{m}}\|_2} \sqrt{\|\boldsymbol{\xi}\|_\infty} \quad (42)$$

with probability at least $1 - e^{-c_t L}$. Here c_t is a constant that depends quadratically on t . Since $(\delta \mathbf{h}, \delta \mathbf{m}) \in \mathcal{N}_\perp$, the inequality above only gives us an element, $((1 - \beta_0)\hat{\mathbf{h}}, (1 + \beta_0)\hat{\mathbf{m}})$ for some $\beta_0 \in \mathbb{R}$, of the set $(\hat{\mathbf{h}}, \hat{\mathbf{m}}) \oplus \mathcal{N}$ obeys

$$\begin{aligned} \|(\tilde{\mathbf{h}}, \tilde{\mathbf{m}}) - ((1 - \beta)\hat{\mathbf{h}}, (1 + \beta)\hat{\mathbf{m}})\|_2 \\ \leq 18 \sqrt{\|\hat{\mathbf{h}}\|_2 \|\hat{\mathbf{m}}\|_2} \sqrt{\|\boldsymbol{\xi}\|_\infty}. \end{aligned} \quad (43)$$

That is, the solutions $(\tilde{\mathbf{h}}, \tilde{\mathbf{m}})$ cannot wander too far away from the line $(\hat{\mathbf{h}}, \hat{\mathbf{m}}) \oplus \mathcal{N}$. We call this norm cylinder constraint as the solution must lie within a cylinder, centered at a line $(\hat{\mathbf{h}}, \hat{\mathbf{m}}) \oplus \mathcal{N}$

and of radius given by the r.h.s. of the equation (43). Equivalently, a displacement $((1 - \beta)\hat{\mathbf{h}}, (1 + \beta)\hat{\mathbf{m}})$ of the ground truth $(\hat{\mathbf{h}}, \hat{\mathbf{m}})$ is sufficiently close to $(\tilde{\mathbf{h}}, \tilde{\mathbf{m}})$. Using this fact together with the fact that the feasible hyperbolic set diverges away from the line $(\hat{\mathbf{h}}, \hat{\mathbf{m}}) \oplus \mathcal{N}$ for large displacement β and touches the line at $\beta = 0$, we will conclude in the remaining proof that the Euclidean distance between $(\tilde{\mathbf{h}}, \tilde{\mathbf{m}})$ and the bilinear ambiguity curve corresponding to the ground truth $(\hat{\mathbf{h}}, \hat{\mathbf{m}})$ is bounded.

We first note that in the case when $\boldsymbol{\xi} = \mathbf{0}$, equation (43) implies that $(\tilde{\mathbf{h}}, \tilde{\mathbf{m}})$ must be on the line $(\hat{\mathbf{h}}, \hat{\mathbf{m}}) \oplus \mathcal{N}_\perp$. Since the only element in the line $(\hat{\mathbf{h}}, \hat{\mathbf{m}}) \oplus \mathcal{N}_\perp$ that is feasible is $(\hat{\mathbf{h}}, \hat{\mathbf{m}})$, we conclude that in the noiseless case $(\tilde{\mathbf{h}}, \tilde{\mathbf{m}}) = (\hat{\mathbf{h}}, \hat{\mathbf{m}})$.

Now, we use the fact that the noise $\boldsymbol{\xi}$ is such that $\xi_\ell \in [-1, 0]$ for every $\ell \in [m]$, and there exists an $\ell' \in [m]$ such that $\xi_{\ell'} = 0$. Trivially, the minimizer $(\tilde{\mathbf{h}}, \tilde{\mathbf{m}})$ must lie somewhere in the feasible set specified by the ℓ' constraint: $\text{sign}(y_{\ell'})\mathbf{b}_{\ell'}^\top \mathbf{h} \mathbf{c}_{\ell'}^\top \mathbf{m} \geq |y_{\ell'}|$ and $t_{\ell'}\mathbf{b}_{\ell'}^\top \mathbf{h} \geq 0$. Define the boundary \mathcal{B} of the feasible set as follows

$$\mathcal{B} := \{(\mathbf{h}, \mathbf{m}) : \mathbf{b}_{\ell'}^\top \mathbf{h} \mathbf{c}_{\ell'}^\top \mathbf{m} = y_{\ell'}, t_{\ell'}\mathbf{b}_{\ell'}^\top \mathbf{h} \geq 0\}. \quad (44)$$

The line $(\hat{\mathbf{h}}, \hat{\mathbf{m}}) \oplus \mathcal{N}_\perp$ only touches the feasible set \mathcal{B} at $(\hat{\mathbf{h}}, \hat{\mathbf{m}})$. For a fixed displacement $\beta \in \mathbb{R}$ from $(\hat{\mathbf{h}}, \hat{\mathbf{m}})$, define a segment of the norm cylinder in (43) as

$$\mathcal{C}_\beta := \left\{(\mathbf{h}, \mathbf{m}) : \begin{aligned} &\|(\mathbf{h}, \mathbf{m}) - ((1 - \beta)\hat{\mathbf{h}}, (1 + \beta)\hat{\mathbf{m}})\|_2 \\ &\leq 18\sqrt{\|\hat{\mathbf{h}}\|_2\|\hat{\mathbf{m}}\|_2\sqrt{\|\boldsymbol{\xi}\|_\infty}} \end{aligned} \right\}. \quad (45)$$

Clearly, there exists a β_0 such that $(\tilde{\mathbf{h}}, \tilde{\mathbf{m}}) \in \mathcal{C}_{\beta_0}$. Moreover, there exist a $c > 0$ such that $(c\hat{\mathbf{h}}, \frac{1}{c}\hat{\mathbf{m}}) \in \mathcal{C}_{\beta_0}$. This is because $(\tilde{\mathbf{h}}, \tilde{\mathbf{m}})$ must live in the convex hull of the set \mathcal{B} and the bilinear ambiguity curve corresponding to $(\hat{\mathbf{h}}, \hat{\mathbf{m}})$ is in the set \mathcal{B} . Since the distance between any two points in a cross-section of a cylinder is at most twice the radius of the cylinder, we have

$$\|(\tilde{\mathbf{h}}, \tilde{\mathbf{m}}) - (c\hat{\mathbf{h}}, \frac{1}{c}\hat{\mathbf{m}})\|_2 \leq 36\sqrt{\|\hat{\mathbf{h}}\|_2\|\hat{\mathbf{m}}\|_2\sqrt{\|\boldsymbol{\xi}\|_\infty}} \quad (46)$$

for some $c > 0$. ■

We now compute the Rademacher complexity $\mathfrak{C}(\mathcal{D})$ defined in (28) of the set of descent directions \mathcal{D} defined in (27).

Lemma 3: Fix $(\mathbf{h}^\mathbf{h}, \mathbf{m}^\mathbf{h}) \in \mathbb{R}^{K \times N}$. Let $\mathbf{B} \in \mathbb{R}^{L \times K}$ and $\mathbf{C} \in \mathbb{R}^{L \times N}$ have i.i.d. $\mathcal{N}(0, 1)$ entries. Let \mathcal{D} be as defined in (27). Then

$$\mathfrak{C}(\mathcal{D}) \leq C\sqrt{\left(\|\hat{\mathbf{m}}\|_2^2 + \|\hat{\mathbf{h}}\|_2^2\right)(S_1 + S_2)\log^2(K + N)} \quad (47)$$

where $C > 0$ is an absolute constant.

Proof: We start by evaluating

$$\mathfrak{C}(\mathcal{D})$$

$$\begin{aligned} &= \mathbb{E} \sup_{(\delta\mathbf{h}, \delta\mathbf{m}) \in \mathcal{D}} \frac{1}{\sqrt{L}} \sum_{\ell=1}^L \varepsilon_\ell \left\langle (\mathbf{c}_\ell^\top \hat{\mathbf{m}} \mathbf{b}_\ell, \mathbf{b}_\ell^\top \hat{\mathbf{h}} \mathbf{c}_\ell), \frac{(\delta\mathbf{h}, \delta\mathbf{m})}{\|(\delta\mathbf{h}, \delta\mathbf{m})\|_2} \right\rangle \\ &\leq \mathbb{E} \left\| \frac{1}{\sqrt{L}} \sum_{\ell=1}^L \varepsilon_\ell \left(\mathbf{c}_\ell^\top \hat{\mathbf{m}} \mathbf{b}_\ell|_{\Gamma_h}, \mathbf{b}_\ell^\top \hat{\mathbf{h}} \mathbf{c}_\ell|_{\Gamma_m} \right) \right\|_2 \end{aligned}$$

$$\begin{aligned} &\cdot \sup_{(\delta\mathbf{h}, \delta\mathbf{m}) \in \mathcal{D}} \left\| \frac{(\delta\mathbf{h}_{\Gamma_h}, \delta\mathbf{m}_{\Gamma_m})}{\|(\delta\mathbf{h}, \delta\mathbf{m})\|_2} \right\|_2 \\ &+ \mathbb{E} \left\| \frac{1}{\sqrt{L}} \sum_{\ell=1}^L \varepsilon_\ell \left(\mathbf{c}_\ell^\top \hat{\mathbf{m}} \mathbf{b}_\ell|_{\Gamma_h^c}, \mathbf{b}_\ell^\top \hat{\mathbf{h}} \mathbf{c}_\ell|_{\Gamma_m^c} \right) \right\|_\infty \\ &\cdot \sup_{(\delta\mathbf{h}, \delta\mathbf{m}) \in \mathcal{D}} \left\| \frac{(\delta\mathbf{h}_{\Gamma_h^c}, \delta\mathbf{m}_{\Gamma_m^c})}{\|(\delta\mathbf{h}, \delta\mathbf{m})\|_2} \right\|_1. \end{aligned} \quad (48)$$

First note that on set \mathcal{D} , we have

$$\begin{aligned} &\left\| \frac{(\delta\mathbf{h}_{\Gamma_h^c}, \delta\mathbf{m}_{\Gamma_m^c})}{\|(\delta\mathbf{h}, \delta\mathbf{m})\|_2} \right\|_1 \\ &\leq \sqrt{S_1 + S_2} \left\| \frac{(\delta\mathbf{h}_{\Gamma_h}, \delta\mathbf{m}_{\Gamma_m})}{\|(\delta\mathbf{h}, \delta\mathbf{m})\|_2} \right\|_2 \leq \sqrt{S_1 + S_2}. \end{aligned}$$

As for the remaining terms, we begin by writing

$$\begin{aligned} &\mathbb{E} \left\| \frac{1}{\sqrt{L}} \sum_{\ell=1}^L \varepsilon_\ell \left(\mathbf{c}_\ell^\top \hat{\mathbf{m}} \mathbf{b}_\ell|_{\Gamma_h}, \mathbf{b}_\ell^\top \hat{\mathbf{h}} \mathbf{c}_\ell|_{\Gamma_m} \right) \right\|_2 \\ &\leq \sqrt{\mathbb{E} \left\| \frac{1}{\sqrt{L}} \sum_{\ell=1}^L \varepsilon_\ell \left(\mathbf{c}_\ell^\top \hat{\mathbf{m}} \mathbf{b}_\ell|_{\Gamma_h}, \mathbf{b}_\ell^\top \hat{\mathbf{h}} \mathbf{c}_\ell|_{\Gamma_m} \right) \right\|_2^2} \\ &= \sqrt{\frac{1}{L} \sum_{\ell=1}^L \mathbb{E} \left(|\mathbf{c}_\ell^\top \hat{\mathbf{m}}|^2 \|\mathbf{b}_\ell|_{\Gamma_h}\|_2^2 + |\mathbf{b}_\ell^\top \hat{\mathbf{h}}|^2 \|\mathbf{c}_\ell|_{\Gamma_m}\|_2^2 \right)} \\ &= \sqrt{\|\hat{\mathbf{m}}\|_2^2 S_1 + \|\hat{\mathbf{h}}\|_2^2 S_2}, \end{aligned}$$

and the second term in (48) is

$$\begin{aligned} &\mathbb{E} \left\| \frac{1}{\sqrt{L}} \sum_{\ell=1}^L \varepsilon_\ell \left(\mathbf{b}_\ell^\top \hat{\mathbf{h}} \mathbf{c}_\ell|_{\Gamma_m^c}, \mathbf{c}_\ell^\top \hat{\mathbf{m}} \mathbf{b}_\ell|_{\Gamma_h^c} \right) \right\|_\infty \\ &\leq \sqrt{\mathbb{E} \left\| \frac{1}{\sqrt{L}} \sum_{\ell=1}^L \varepsilon_\ell \left(\mathbf{b}_\ell^\top \hat{\mathbf{h}} \mathbf{c}_\ell|_{\Gamma_m^c}, \mathbf{c}_\ell^\top \hat{\mathbf{m}} \mathbf{b}_\ell|_{\Gamma_h^c} \right) \right\|_\infty^2} \\ &\leq \left(2e \log(K + N) \cdot \frac{1}{L} \sum_{\ell=1}^L \mathbb{E} \max \{ |\mathbf{c}_\ell^\top \hat{\mathbf{m}}|^2 \|\mathbf{b}_\ell|_{\Gamma_h^c}\|_\infty^2, \right. \\ &\quad \left. |\mathbf{b}_\ell^\top \hat{\mathbf{h}}|^2 \|\mathbf{c}_\ell|_{\Gamma_m^c}\|_\infty^2 \} \right)^{\frac{1}{2}} \\ &\leq \left(2e \log(K + N) \mathbb{E} \max \{ |\mathbf{b}^\top \hat{\mathbf{h}}|^2 \|\mathbf{c}|_{\Gamma_m^c}\|_\infty^2, \right. \\ &\quad \left. |\mathbf{c}^\top \hat{\mathbf{m}}|^2 \|\mathbf{b}|_{\Gamma_h^c}\|_\infty^2 \} \right)^{\frac{1}{2}} \\ &\leq C \sqrt{\max\{\|\hat{\mathbf{h}}\|_2^2, \|\hat{\mathbf{m}}\|_2^2\} \log^2(K + N)}, \end{aligned}$$

where the second inequality by the application of Lemma 5.2.2 in [42], and the final equality is due to the fact that $\|\mathbf{c}|_{\Gamma_m^c}\|_\infty^2$ and $\|\mathbf{b}|_{\Gamma_h^c}\|_\infty^2$ are subexponential and using Lemma 3 in [43]. Plugging the bounds above back in (48), we obtain the upper

bound on the Rademacher complexity given below

$$\mathfrak{C}(\mathcal{D}) \leq C \sqrt{\left(\|\hat{\mathbf{m}}\|_2^2 + \|\hat{\mathbf{h}}\|_2^2\right) (S_1 + S_2) \log^2(K + N)}. \quad (49)$$

Next we compute the tail probability estimate $\mathfrak{p}_\tau(\mathcal{D})$ defined in (29).

Lemma 4: Fix $(\mathbf{h}^\dagger, \mathbf{m}^\dagger) \in \mathbb{R}^{K \times N}$. Let $\mathbf{B} \in \mathbb{R}^{L \times K}$ and $\mathbf{C} \in \mathbb{R}^{L \times N}$ have i.i.d. $\mathcal{N}(0, 1)$ entries. Let \mathcal{D} be as defined in (27) and set $\tau = \sqrt{\frac{1}{2} \min\{\|\hat{\mathbf{h}}\|_2^2, \|\hat{\mathbf{m}}\|_2^2\}}$. Then $\mathfrak{p}_\tau(\mathcal{D}) \geq \frac{1}{8c^4}$ for some absolute constant $c > 0$.

Proof: In order to evaluate

$$\mathfrak{p}_\tau(\mathcal{D}) = \inf_{(\delta\mathbf{h}, \delta\mathbf{m}) \in \mathcal{D}} \mathbb{P} \left(\left\langle \mathbf{c}_\ell^\top \hat{\mathbf{m}} \mathbf{b}_\ell^\top, \mathbf{b}_\ell^\top \hat{\mathbf{h}} \mathbf{c}_\ell^\top \right\rangle, \frac{(\delta\mathbf{h}, \delta\mathbf{m})}{\|(\delta\mathbf{h}, \delta\mathbf{m})\|_2} \right\rangle \geq \tau \right) \quad (50)$$

it suffice to estimate the probability $\mathbb{P}(|\mathbf{b}_\ell^\top \hat{\mathbf{h}} \mathbf{c}_\ell^\top \delta\mathbf{m} + \mathbf{b}_\ell^\top \delta\mathbf{h} \mathbf{c}_\ell^\top \hat{\mathbf{m}}| \geq \tau)$. Using Paley-Zygmund inequality, we have

$$\mathbb{P} \left(|\mathbf{b}_\ell^\top \hat{\mathbf{h}} \mathbf{c}_\ell^\top \delta\mathbf{m} + \mathbf{b}_\ell^\top \delta\mathbf{h} \mathbf{c}_\ell^\top \hat{\mathbf{m}}|^2 \geq \frac{1}{2} \mathbb{E} |\mathbf{b}_\ell^\top \hat{\mathbf{h}} \mathbf{c}_\ell^\top \delta\mathbf{m} + \mathbf{b}_\ell^\top \delta\mathbf{h} \mathbf{c}_\ell^\top \hat{\mathbf{m}}|^2 \right) \geq \frac{1}{4} \cdot \frac{\left(\mathbb{E} |\mathbf{b}_\ell^\top \hat{\mathbf{h}} \mathbf{c}_\ell^\top \delta\mathbf{m} + \mathbf{b}_\ell^\top \delta\mathbf{h} \mathbf{c}_\ell^\top \hat{\mathbf{m}}|^2 \right)^2}{\mathbb{E} |\mathbf{b}_\ell^\top \hat{\mathbf{h}} \mathbf{c}_\ell^\top \delta\mathbf{m} + \mathbf{b}_\ell^\top \delta\mathbf{h} \mathbf{c}_\ell^\top \hat{\mathbf{m}}|^4}.$$

Using norm equivalence of Gaussian random variables, we know that $(\mathbb{E} |\mathbf{b}_\ell^\top \hat{\mathbf{h}} \mathbf{c}_\ell^\top \delta\mathbf{m} + \mathbf{b}_\ell^\top \delta\mathbf{h} \mathbf{c}_\ell^\top \hat{\mathbf{m}}|^4)^{1/4} \leq c(\mathbb{E} |\mathbf{b}_\ell^\top \hat{\mathbf{h}} \mathbf{c}_\ell^\top \delta\mathbf{m} + \mathbf{b}_\ell^\top \delta\mathbf{h} \mathbf{c}_\ell^\top \hat{\mathbf{m}}|^2)^{1/2}$, this implies that

$$\mathbb{P} \left(|\mathbf{b}_\ell^\top \hat{\mathbf{h}} \mathbf{c}_\ell^\top \delta\mathbf{m} + \mathbf{b}_\ell^\top \delta\mathbf{h} \mathbf{c}_\ell^\top \hat{\mathbf{m}}|^2 \geq \frac{1}{2} \mathbb{E} |\mathbf{b}_\ell^\top \hat{\mathbf{h}} \mathbf{c}_\ell^\top \delta\mathbf{m} + \mathbf{b}_\ell^\top \delta\mathbf{h} \mathbf{c}_\ell^\top \hat{\mathbf{m}}|^2 \right) \geq \frac{1}{4} \cdot \frac{1}{c^4}.$$

Next, we show that $\mathbb{E} |\mathbf{b}_\ell^\top \hat{\mathbf{h}} \mathbf{c}_\ell^\top \delta\mathbf{m} + \mathbf{b}_\ell^\top \delta\mathbf{h} \mathbf{c}_\ell^\top \hat{\mathbf{m}}|^2 \geq \min\{\|\hat{\mathbf{h}}\|_2^2, \|\hat{\mathbf{m}}\|_2^2\} (\|\delta\mathbf{h}\|_2^2 + \|\delta\mathbf{m}\|_2^2)$. Consider

$$\begin{aligned} & \mathbb{E} |\mathbf{b}_\ell^\top \hat{\mathbf{h}} \mathbf{c}_\ell^\top \delta\mathbf{m} + \mathbf{b}_\ell^\top \delta\mathbf{h} \mathbf{c}_\ell^\top \hat{\mathbf{m}}|^2 \\ &= \mathbb{E}_b \mathbb{E}_c \hat{\mathbf{h}}^\top \mathbf{b}_\ell \mathbf{b}_\ell^\top \hat{\mathbf{h}} \delta\mathbf{m}^\top \mathbf{c}_\ell \mathbf{c}_\ell^\top \delta\mathbf{m} + \delta\mathbf{h}^\top \mathbf{b}_\ell \mathbf{b}_\ell^\top \delta\mathbf{h} \hat{\mathbf{m}}^\top \mathbf{c}_\ell \mathbf{c}_\ell^\top \hat{\mathbf{m}} \\ &\quad + 2 \mathbb{E}_b \mathbb{E}_c \delta\mathbf{h}^\top \mathbf{b}_\ell \mathbf{b}_\ell^\top \hat{\mathbf{h}} \delta\mathbf{m}^\top \mathbf{c}_\ell \mathbf{c}_\ell^\top \hat{\mathbf{m}} \\ &= \mathbb{E}_b \|\delta\mathbf{m}\|^2 \hat{\mathbf{h}}^\top \mathbf{b}_\ell \mathbf{b}_\ell^\top \hat{\mathbf{h}} + \|\hat{\mathbf{m}}\|^2 \delta\mathbf{h}^\top \mathbf{b}_\ell \mathbf{b}_\ell^\top \delta\mathbf{h} \\ &\quad + 2 \delta\mathbf{m}^\top \hat{\mathbf{m}} \delta\mathbf{h}^\top \mathbf{b}_\ell \mathbf{b}_\ell^\top \hat{\mathbf{h}} \\ &= \|\delta\mathbf{m}\|^2 \|\hat{\mathbf{h}}\|^2 + \|\hat{\mathbf{m}}\|^2 \|\delta\mathbf{h}\|^2 + 2 \delta\mathbf{m}^\top \hat{\mathbf{m}} \delta\mathbf{h}^\top \hat{\mathbf{h}} \\ &= \|\delta\mathbf{m}\|^2 \|\hat{\mathbf{h}}\|^2 + \|\hat{\mathbf{m}}\|^2 \|\delta\mathbf{h}\|^2 + 2 (\delta\mathbf{m}^\top \hat{\mathbf{m}})^2 \\ &\geq \min\{\|\hat{\mathbf{h}}\|_2^2, \|\hat{\mathbf{m}}\|_2^2\} (\|\delta\mathbf{h}\|_2^2 + \|\delta\mathbf{m}\|_2^2) \end{aligned}$$

where the second and third equalities follow because \mathbf{b}_ℓ and \mathbf{c}_ℓ contain i.i.d. $\mathcal{N}(0, 1)$ entries. The fourth equality follows from the fact $(\delta\mathbf{h}, \delta\mathbf{m}) \in \mathcal{D} \subset \mathcal{N}_\perp$, and hence $\mathcal{D} \perp \mathcal{N}$, which implies

that $\delta\mathbf{h}^\top \hat{\mathbf{h}} = \delta\mathbf{m}^\top \hat{\mathbf{m}}$. Normalizing by $\|(\delta\mathbf{h}, \delta\mathbf{m})\|_2$, and comparing with (50) directly shows that $\tau^2 = \frac{1}{2} \min\{\|\hat{\mathbf{h}}\|_2^2, \|\hat{\mathbf{m}}\|_2^2\}$, and $\mathfrak{p}_\tau(\mathcal{D}) \geq \frac{1}{8c^4}$. This completes the proof. ■

We now present a proof of Theorem 1. In Theorem 1, the noise satisfy $\xi_\ell \geq -1$ which is in contrast to $\xi_\ell \in [-1, 0]$ with $\xi_{\ell'} = 0$ for some $\ell' \in \{1, \dots, L\}$ in Lemma 2. The key idea is measurements with noise that satisfy $\xi_\ell \geq -1$ can be converted to measurements with noise in the interval $[-1, 0]$ with the noise for one of the measurement exactly equal to zero. In order to see this, let

$$s = \max_{\ell \in [L]} \frac{y_\ell}{\hat{y}_\ell} = 1 + \max_{\ell \in [L]} \xi_\ell \leq 1 + \|\xi\|_\infty, \quad (51)$$

$$\eta_\ell = \frac{1}{s} (1 - s + \xi_\ell). \quad (52)$$

We then consider the measurements $y_\ell = s\hat{y}_\ell(1 + \eta_\ell)$ for $\ell \in [L]$. Because $s\hat{y}_\ell(1 + \eta_\ell) = \hat{y}_\ell(1 + \xi_\ell)$, the noisy measurements are the same, however the noise may be different.

Proof of Theorem 1: As the noise of measurements $y_\ell = \hat{y}_\ell(1 + \xi_\ell)$ may not be one-sided as in (3), we consider equivalent measurements $y_\ell = s\hat{y}_\ell(1 + \eta_\ell)$, where s and η_ℓ are as defined in (51) and (52), respectively. This turns the ℓ_1 -BranchHull program (4) into

$$\begin{aligned} & \text{minimize}_{\mathbf{h} \in \mathbb{R}^K, \mathbf{m} \in \mathbb{R}^N} \|\mathbf{h}\|_1 + \|\mathbf{m}\|_1 \\ & \text{subject to } s_\ell \mathbf{b}_\ell^\top \mathbf{h} \mathbf{c}_\ell^\top \mathbf{m} \geq |s\hat{y}_\ell(1 + \eta_\ell)|, \\ & t_\ell \cdot \mathbf{b}_\ell^\top \mathbf{h} \geq 0, \ell \in \{1, \dots, L\}. \end{aligned} \quad (53)$$

First, we note that for all $\ell \in \{1, \dots, L\}$,

$$\eta_\ell = \frac{1}{s} (1 + \xi_\ell - s) \quad (54)$$

$$\leq \frac{1}{s} (s - s) \quad (55)$$

$$= 0, \quad (56)$$

where the first inequality holds because $1 + \xi_\ell \leq 1 + \max_{\ell \in [L]} \xi_\ell \leq s$. Second, we have $\eta_\ell \geq -1$ for all $\ell \in \{1, \dots, L\}$, which follows directly from $\xi_\ell \geq -1$ for all ℓ . Third, there exists a ℓ' such that $\eta_{\ell'} = 0$. Thus, the noise $\boldsymbol{\eta}$ satisfies (3) and by Lemma 2, the minimizer $(\tilde{\mathbf{h}}, \tilde{\mathbf{m}})$ of (53) is unique and if $L \geq \left(\frac{2c(\mathcal{D}) + t\tau}{\tau \mathfrak{p}_\tau(\mathcal{D})}\right)^2$, the minimizer satisfies

$$\begin{aligned} & \left(\|\tilde{\mathbf{h}} - c\sqrt{s}\hat{\mathbf{h}}\|_2^2 + \|\tilde{\mathbf{m}} - c^{-1}\sqrt{s}\hat{\mathbf{m}}\|_2^2 \right)^{\frac{1}{2}} \\ & \leq 36 \sqrt{s\|\hat{\mathbf{h}}\|_2 \|\hat{\mathbf{m}}\|_2} \sqrt{\|\boldsymbol{\eta}\|_\infty} \end{aligned} \quad (57)$$

for some $c > 0$ with probability at least $1 - e^{-c_\epsilon L}$. Furthermore, $c \rightarrow 1$ as $\boldsymbol{\eta} \rightarrow \mathbf{0}$. In (57),

$$\begin{aligned} s\|\boldsymbol{\eta}\|_\infty &= - \min_{\ell \in [L]} (1 - s + \xi_\ell) \\ &= s - 1 - \min_{\ell \in [L]} \xi_\ell \\ &= \max_{\ell \in [L]} \xi_\ell - \min_{\ell \in [L]} \xi_\ell \\ &\leq 2\|\xi\|_\infty \end{aligned} \quad (58)$$

where the first equality holds because $\eta_\ell \leq 0$ for all $\ell \in \{1, \dots, L\}$. We now compute

$$\begin{aligned} & \left(\left\| \tilde{\mathbf{h}} - \hat{\mathbf{c}}\hat{\mathbf{h}} \right\|_2^2 + \left\| \tilde{\mathbf{m}} - c^{-1}\hat{\mathbf{m}} \right\|_2^2 \right)^{\frac{1}{2}} \\ &= \left(\left\| \tilde{\mathbf{h}} - c\sqrt{s}\hat{\mathbf{h}} + c\sqrt{s}\hat{\mathbf{h}} - \hat{\mathbf{c}}\hat{\mathbf{h}} \right\|_2^2 \right. \\ & \quad \left. + \left\| \tilde{\mathbf{m}} - c^{-1}\sqrt{s}\hat{\mathbf{m}} + c^{-1}\sqrt{s}\hat{\mathbf{m}} - c^{-1}\hat{\mathbf{m}} \right\|_2^2 \right)^{\frac{1}{2}} \\ &\leq \left(\left\| \tilde{\mathbf{h}} - c\sqrt{s}\hat{\mathbf{h}} \right\|_2^2 + \left\| \tilde{\mathbf{m}} - c^{-1}\sqrt{s}\hat{\mathbf{m}} \right\|_2^2 \right)^{\frac{1}{2}} \\ & \quad + \left(\left\| c\sqrt{s}\hat{\mathbf{h}} - \hat{\mathbf{c}}\hat{\mathbf{h}} \right\|_2^2 + \left\| c^{-1}\sqrt{s}\hat{\mathbf{m}} - c^{-1}\hat{\mathbf{m}} \right\|_2^2 \right)^{\frac{1}{2}} \quad (59) \end{aligned}$$

$$\begin{aligned} &\leq 36 \left(s\|\boldsymbol{\eta}\|_\infty \|\hat{\mathbf{h}}\|_2 \|\hat{\mathbf{m}}\|_2 \right)^{\frac{1}{2}} \\ & \quad + (\sqrt{s} - 1) \left(\left\| \hat{\mathbf{c}}\hat{\mathbf{h}} \right\|_2^2 + \left\| c^{-1}\hat{\mathbf{m}} \right\|_2^2 \right)^{\frac{1}{2}} \quad (60) \end{aligned}$$

$$\begin{aligned} &\leq \left(36 \left(\frac{s\|\boldsymbol{\eta}\|_\infty}{2} \right)^{\frac{1}{2}} + (\sqrt{s} - 1) \right) \left(\left\| \hat{\mathbf{c}}\hat{\mathbf{h}} \right\|_2^2 + \left\| c^{-1}\hat{\mathbf{m}} \right\|_2^2 \right)^{\frac{1}{2}} \\ &\leq \left(36 (\|\boldsymbol{\xi}\|_\infty)^{\frac{1}{2}} + (\sqrt{s} - 1) \right) \left(\left\| \hat{\mathbf{c}}\hat{\mathbf{h}} \right\|_2^2 + \left\| c^{-1}\hat{\mathbf{m}} \right\|_2^2 \right)^{\frac{1}{2}} \quad (61) \end{aligned}$$

$$\begin{aligned} &\leq \left(36 (\|\boldsymbol{\xi}\|_\infty)^{\frac{1}{2}} + (\sqrt{1 + \|\boldsymbol{\xi}\|_\infty} - 1) \right) \left(\left\| \hat{\mathbf{c}}\hat{\mathbf{h}} \right\|_2^2 \right. \\ & \quad \left. + \left\| c^{-1}\hat{\mathbf{m}} \right\|_2^2 \right)^{\frac{1}{2}} \quad (62) \end{aligned}$$

$$\leq 37\sqrt{\|\boldsymbol{\xi}\|_\infty} \left(\left\| \hat{\mathbf{c}}\hat{\mathbf{h}} \right\|_2^2 + \left\| c^{-1}\hat{\mathbf{m}} \right\|_2^2 \right)^{\frac{1}{2}} \quad (63)$$

where (59) holds because of triangle inequality, (60) holds because of (57), (61) holds because of (58), (62) holds because of (51) and (63) holds because for $\sqrt{1 + \|\boldsymbol{\xi}\|_\infty} - 1 \leq \sqrt{\|\boldsymbol{\xi}\|_\infty}$. Lastly, we note that $\left(\frac{2\mathfrak{C}(\mathcal{D}) + t\tau}{\tau p_\tau(\mathcal{D})} \right)^2 \geq C_\rho (\sqrt{S_1} + S_2) \log(K + N) + t)^2$ because of Lemmas 3, 4 and the assumption that \mathbf{h}^\dagger and \mathbf{m}^\dagger are ρ -comparable-effective-sparse as in (2), for some $\rho \geq 1$. Here, C_ρ is constants that depends quadratically on ρ . This completes the proof. ■

REFERENCES

- [1] J. R. Fienup, "Phase retrieval algorithms: A comparison," *Appl. Opt.*, vol. 21, no. 15, pp. 2758–2769, 1982.
- [2] E. Candès, T. Strohmer, and V. Voroninski, "Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming," *Commun. Pure Appl. Math.*, vol. 66, no. 8, pp. 1241–1274, 2013.
- [3] E. Candès, T. Strohmer, and V. Voroninski, "Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming," *Commun. Pure Appl. Math.*, vol. 66, no. 8, pp. 1241–1274, 2013.
- [4] A. Ahmed, B. Recht, and J. Romberg, "Blind deconvolution using convex programming," *IEEE Trans. Inf. Theory*, vol. 60, no. 3, pp. 1711–1732, Mar. 2014.
- [5] T. G. Stockham, T. M. Cannon, and R. B. Ingebreetsen, "Blind deconvolution through digital signal processing," *Proc. IEEE*, vol. 63, no. 4, pp. 678–692, Apr. 1975.
- [6] D. Kundur and D. Hatzinakos, "Blind image deconvolution," *IEEE Signal Process. Mag.*, vol. 13, no. 3, pp. 43–64, May 1996.
- [7] A. Aghasi, B. Heshmat, A. Redo-Sanchez, J. Romberg, and R. Raskar, "Sweep distortion removal from terahertz images via blind demodulation," *Optica*, vol. 3, no. 7, pp. 754–762, 2016.
- [8] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *J. Mach. Learn. Res.*, vol. 5, pp. 1457–1469, 2004.
- [9] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, pp. 556–562.
- [10] S. Ling and T. Strohmer, "Self-calibration and biconvex compressive sensing," *Inverse Problems*, vol. 31, no. 11, 2015, Art. no. 115002.
- [11] P. D. O'Grady, B. A. Pearlmutter, and S. T. Rickard, "Survey of sparse and non-sparse methods in source separation," *Int. J. Imag. Syst. Technol.*, vol. 15, no. 1, pp. 18–33, 2005.
- [12] I. Tosic and P. Frossard, "Dictionary learning," *IEEE Signal Process. Mag.*, vol. 28, no. 2, pp. 27–38, Mar. 2011.
- [13] T. Chen, W. Yin, X. S. Zhou, D. Comaniciu, and T. S. Huang, "Total variation models for variable lighting face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 9, pp. 1519–1524, Sep. 2006.
- [14] B. Horn, B. Klaus, and P. Horn, *Robot Vision*. Cambridge, MA, USA: MIT Press, 1986.
- [15] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2341–2353, Dec. 2010.
- [16] R. T. Tan, "Visibility in bad weather from a single image," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2008, pp. 1–8.
- [17] R. Fattal, "Single image dehazing," *ACM Trans. Graph.*, vol. 27, no. 3, pp. 1–9, 2008.
- [18] A. Ahmed, A. Aghasi, and P. Hand, "Blind deconvolutional phase retrieval via convex programming," in *Proc. Adv. Neural Inform. Process. Syst.*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., Red Hook, NY, USA: Curran Associates, Inc., 2018, pp. 10 030–10 040.
- [19] K. Lee, Y. Wu, and Y. Bresler, "Near-optimal compressed sensing of a class of sparse low-rank matrices via sparse power factorization," *IEEE Trans. Inf. Theory*, vol. 64, no. 3, pp. 1666–1698, Mar. 2018.
- [20] X. Li and V. Voroninski, "Sparse signal recovery from quadratic measurements via convex programming," *SIAM J. Math. Anal.*, vol. 45, no. 5, pp. 3019–3033, 2013.
- [21] S. Oymak, A. Jalali, M. Fazel, Y. C. Eldar, and B. Hassibi, "Simultaneously structured models with application to sparse and low-rank matrices," *IEEE Trans. Inf. Theory*, vol. 61, no. 5, pp. 2886–2908, May 2015.
- [22] X. Li, S. Ling, T. Strohmer, and K. Wei, "Rapid, robust, and reliable blind deconvolution via nonconvex optimization," *Appl. Comput. Harmonic Anal.*, vol. 47, no. 3, pp. 893–934, 2019.
- [23] E. Candès, X. Li, and M. Soltanolkotabi, "Phase retrieval via wirtinger flow: Theory and algorithms," *IEEE Trans. Inf. Theory*, vol. 61, no. 4, pp. 1985–2007, Apr. 2015.
- [24] P. Netrapalli, P. Jain, and S. Sanghavi, "Phase retrieval using alternating minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2796–2804.
- [25] J. Sun, Q. Qu, and J. Wright, "A geometric analysis of phase retrieval," in *Proc. IEEE Int. Symp. Inf. Theory*, 2016, pp. 2379–2383.
- [26] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht, "Low-rank solutions of linear matrix equations via procrustes flow," in *Proc. Mach. Learn. Res.*, vol. 48, Jun. 2016, pp. 964–973, [Online]. Available: <https://proceedings.mlr.press/v48/tu16.html>
- [27] Y. Chen and E. Candes, "Solving random quadratic systems of equations is nearly as easy as solving linear systems," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 739–747.
- [28] S. Bahmani and J. Romberg, "Phase retrieval meets statistical learning theory: A flexible convex relaxation," in *Proc. 20th Int. Conf. Artif. Intell. Statist.*, (Proc. Mach. Learn. Res. Ser.), A. Singh and J. Zhu, Eds., vol. 54, Fort Lauderdale, FL, USA: PMLR, 20–22 Apr. 2017, pp. 252–260.
- [29] T. Goldstein and C. Studer, "Phasemax: Convex phase retrieval via basis pursuit," *IEEE Trans. Inf. Theory*, vol. 64, no. 4, pp. 2675–2689, Apr. 2018.
- [30] A. Aghasi, A. Ahmed, P. Hand, and B. Joshi, "Branchhull: Convex bilinear inversion from the entrywise product of signals with known signs," *Appl. Comput. Harmonic Anal.*, vol. 49, no. 2, pp. 636–654, 2020, [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1063520318300071>
- [31] A. Aghasi, A. Ahmed, P. Hand, and B. Joshi, "A convex program for bilinear inversion of sparse vectors," in *Proc. Adv. Neural Inf. Process. Syst.*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., Red Hook, NY, USA: Curran Associates, Inc., 2018, pp. 8548–8558.

- [32] S. Boyd *et al.*, “Distributed optimization and statistical learning via the alternating direction method of multipliers,” *Foundations Trends in Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.
- [33] V. Koltchinskii and S. Mendelson, “Bounding the smallest singular value of a random matrix without concentration,” *Int. Math. Res. Notices*, vol. 2015, no. 23, pp. 12 991–13 008, 2015.
- [34] S. Mendelson, “Learning without concentration,” in *Proc. Conf. Learn. Theory*, 2014, pp. 25–39.
- [35] G. Lecué *et al.*, “Regularization and the small-ball method i: Sparse recovery,” *Ann. Statist.*, vol. 46, no. 2, pp. 611–641, 2018.
- [36] G. Lecué and S. Mendelson, “Regularization and the small-ball method ii: Complexity dependent error rates,” *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 5356–5403, 2017.
- [37] S. Bahmani and J. Romberg, “Anchored regression: Solving random convex equations via convex programming,” 2017, *arXiv:1702.05327*.
- [38] R. Vershynin, “Introduction to the non-asymptotic analysis of random matrices,” in *Compressed Sensing: Theory and Applications*, Y. Eldar and G. Kutyniok, Eds., Cambridge, U.K.: Cambridge Univ. Press, 2012.
- [39] C. McDiarmid, “On the method of bounded differences,” *Surveys Combinatorics*, vol. 141, no. 1, pp. 148–188, 1989.
- [40] A. W. van der Vaart and J. A. Wellner, “Weak convergence and empirical processes with applications to statistics,” *J. Roy. Statistical Soc.-Series A Statist. Soc.*, vol. 160, no. 3, pp. 596–608, 1997.
- [41] M. Ledoux and M. Talagrand, *Probability Banach Spaces: Isoperimetry and Processes*. Berlin, Germany: Springer Science & Business Media, 2013.
- [42] M. G. Akritas, S. Lahiri, and D. N. Politis, *Topics in Nonparametric Statistics*. Berlin, Germany: Springer, 2016.
- [43] S. van de Geer and J. Lederer, “The Bernstein–Orlicz norm and deviation inequalities,” *Probability Theory and Related Fields*, vol. 157, no. 1–2, pp. 225–250, 2013.