Structured Least Squares Problems and Robust Estimators

Mert Pilanci, *Student Member, IEEE*, Orhan Arikan, *Member, IEEE*, and Mustafa C. Pinar

***Abstract*—A novel approach is proposed to provide robust and accurate estimates for linear regression problems when both the measurement vector and the coefficient matrix are structured and subject to errors or uncertainty. A new analytic formulation is developed in terms of the gradient flow of the residual norm to analyze and provide estimates to the regression. The presented analysis enables us to establish theoretical performance guarantees to compare with existing methods and also offers a criterion to choose the regularization parameter autonomously. Theoretical results and simulations in applications such as blind identification, multiple frequency estimation and deconvolution show that the proposed technique outperforms alternative methods in mean-squared error for a significant range of signal-to-noise ratio values.**

***Index Terms*—Blind identification, deconvolution, errors-in-variables, frequency estimation, least squares, robust least squares, structured total least squares.**

I. Introduction

In various signal processing applications including deconvolution, signal modeling, frequency estimation, blind channel identification and equalization, it is important to produce robust estimates for an unknown vector $\mathbf{x}$ from a set of measurements $\mathbf{y}$. Typically, a linear model is used to relate the unknowns to the available measurements: $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{w}$, where the matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ describes the linear relationship and $\mathbf{w}$ is additive measurement noise. Over the years, a multitude of techniques have been developed to obtain better estimates for $\mathbf{x}$. For instance, if $\mathbf{x}$ is a random vector with known first and second order statistics, the Wiener estimator, which minimizes the mean-squared error (MSE) over all linear estimators, can be used with proven success [1]. In the absence of such a statistical information on $\mathbf{x}$, the least squares (LS) criterion is commonly used. The well known LS method for solving the overdetermined linear equations $\mathbf{A}\mathbf{x} = \mathbf{y}$ for $m > n$, yields the maximum likelihood (ML) estimate of the deterministic unknown $\mathbf{x}$ when the observations are subject to independent identically distributed (i.i.d.) Gaussian noise and has the minimum MSE over all unbiased estimators [2]. In practice, the observation $\mathbf{y}$ is noisy and the elements of matrix $\mathbf{A}$ are also subject to errors since they may be results of some other measurements or obtained under some modeling assumptions. When the errors in $\mathbf{A}$ and $\mathbf{y}$ are zero-mean i.i.d. Gaussian random variables, the ML estimate can be obtained by the total least squares (TLS) technique, which "corrects" the system with minimum perturbation so that it becomes consistent [3], [4]. However in many applications, the linear system of equations has a structure, e.g., Toeplitz, Hankel, Vandermonde, hence the i.i.d. assumption on the errors is not valid. For that reason, the structured total least squares (STLS) techniques and its regularized versions (RSTLS) have been developed to obtain an accurate estimate by employing minimal norm structured perturbations on the original system until consistency is reached [5]–[7].

In two alternative min-max optimal approaches, the estimator that minimizes the worst case MSE: $E[||\mathbf{x} - \mathbf{x}_0||]$, [8], [9] or residual: $||\mathbf{A}\mathbf{x} - \mathbf{y}||$, [10] is sought respectively. Min-max approaches reduce to convex optimization problems. However, the worst case residual approach which is known as structured robust least squares (SRLS), can also be applied to any linear structured uncertainty. Furthermore, the SRLS problem can be efficiently solved using second-order cone programming [11]. The solution can be interpreted as a Tikhonov regularization in the unstructured case [12], [13]. When $\mathbf{A}$ is ill-conditioned, the min-max solution produces a biased $\hat{\mathbf{x}}$ to avoid the residual norm becoming unacceptably large. As a result the min-max approach may be overly conservative and its average performance is usually undesirable in many applications. Furthermore, the performance of the min-max techniques varies significantly based on the uncertainty bounds that might not be readily available.

In this paper, we propose and analyze a new method, Structured Least Squares with bounded data uncertainties (SLS-BDU), to provide a better tradeoff between the accuracy and robustness of the estimates for the solution to $\mathbf{A}\mathbf{x} = \mathbf{y}$ under structured and bounded uncertainty in $\mathbf{A}$ and $\mathbf{y}$. Unlike the SRLS technique that minimizes the worst case error, the proposed SLS-BDU technique minimizes the best case residual. For ill-conditioned problems, it is demonstrated both in theory and simulations that a small norm bound on the perturbation regularizes the solution and prevents numerical instability which is usually exhibited by the STLS estimator. The proposed estimator does not force the consistency of given equations, which is the primary reason of instability in practice. Instead, the most likely solution that is within the confidence bounds of the perturbations is found. There are important signal processing applications where such bounds on the perturbations are known. Hence, the proposed approach is well suited for such applications including array signal processing, channel estimation [14] and equalization [15], system identification [16], spectral estimation [17], signal modeling [18], where STLS is readily applied. When bounds on the perturbations are not available, the bound can be treated as a regularization

Manuscript received July 20, 2009; accepted December 30, 2009. Date of publication January 22, 2010; date of current version April 14, 2010. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Z. Jane Wang.

M. Pilanci and O. Arikan are with the Department of Electrical and Electronics Engineering, Bilkent University, Bilkent, Ankara TR-06800, Turkey (e-mail: pilanci@ee.bilkent.edu.tr; oarikan@ee.bilkent.edu.tr).

M. C. Pinar is with the Department of Industrial Engineering, Bilkent University, Bilkent, Ankara TR-06800, Turkey (e-mail: mustafap@bilkent.edu.tr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2010.2041279

parameter. For this case, we propose a simple strategy to determine the value of the bound that yields accurate and robust estimates.

The analysis of known estimators and solution of the proposed formulation relies mostly on the Fréchet derivatives of pseudoinverses which was studied in numerical optimization for nonlinear least squares fitting [19]. The geometry of gradient flow of the cost function reveals how the known techniques behave differently and their respective performance over different scenarios. The discussion on the gradient flow leads to a version of SLS-BDU that automatically chooses the bound parameter when it is not available to us. It is shown in numerical examples that the proposed estimator achieves smaller MSE than other alternatives for a large set of SNR values.

In the next section, we present a review of existing approaches for an overdetermined system of linear equations. In Section III, we introduce the SLS-BDU approach and provide an MSE bound. The proposed SLS-BDU technique and alternatives are studied on a simple example in Section IV. Section V presents an analysis of the gradient with Fréchet derivatives and states theorems to formalize the introduced ideas. Three alternatives to perform the proposed SLS-BDU optimization and a criterion to select the bound of optimization are discussed in Sections VI and VII. Finally, Section VIII presents the performance of the SLS-BDU technique in three signal processing applications.

II. Review of Existing Approaches

In this section, we provide a short review of algorithms that have been proposed for linear system of equations with errors in variables. The following approaches can be first divided in to two categories, namely the structured and unstructured perturbations. The total least squares and unstructured bounded errors in Variables approaches are in the first category. The structured total least squares approach is proposed to fulfill the goals of TLS in case of an existing structure. The structured robust least squares approach has been proposed to provide min-max optimal robust solutions to structured least squares problems. In the following, each approach will be briefly reviewed.

Throughout the paper, we denote by $\mathbf{A}^T$ and $\mathbf{A}^\dagger$ the transpose and Moore–Penrose pseudoinverse of the matrix $\mathbf{A}$ respectively. $\|\mathbf{A}\|_2$ is the spectral norm of $\mathbf{A}$, i.e., the largest singular value and $\sigma_{\min}(\mathbf{A})$ is the minimum singular value. For an integer $i,\ 1 \le i \le \operatorname{Rank}(\mathbf{A})$, $\sigma_i(\mathbf{A})$ is the ith largest singular value. $\|\mathbf{A}\|_F \triangleq \sqrt{\sum_i \sigma_i^2(\mathbf{A})}$ denotes the Frobenious norm of $\mathbf{A}$. $\mathbf{A} \odot \mathbf{B}$ denotes the Hadamard, i.e., elementwise product of two matrices of the same size. ∇ and $\mathbf{D}$ are the gradient and Fréchet derivative operators respectively. E denotes expectation of a random variable. $(\cdot)_+$ denotes the positive part of a real scalar and $(\cdot)_i$ denotes the ith subarray of an array of numbers.

A. The Total Least Squares Approach

In Total Least Squares (TLS) approach, the minimum norm perturbation $[\mathbf{\Delta A}\ \mathbf{\Delta y}]$ on $[\mathbf{A}\ \mathbf{y}]$ that results in a consistent system $(\mathbf{A}+\mathbf{\Delta A})\mathbf{x} = \mathbf{y}+\mathbf{\Delta y}$ is found. The TLS problem can be solved by using the singular value decomposition (SVD) as [3]

$$\mathbf{x}_{TLS} = (\mathbf{A}^T\mathbf{A} - \sigma_{n+1}^2\mathbf{I})^{-1}\mathbf{A}^T\mathbf{y} \tag{1}$$

where σ_{n+1} is the smallest singular value of $[\mathbf{A}\ \mathbf{y}]$. However, the subtraction of $\sigma_{n+1}^2\mathbf{I}$ from the diagonal of $\mathbf{A}^T\mathbf{A}$ deregulates the inverse operation, hence, results in an increased sensitivity to noise. It is known that the variance of the TLS estimator is always higher than that of the ordinary least squares estimator, and increases with the condition number of the true matrix $\mathbf{A}_0$ [20]. A weighted TLS solution provides the ML estimate for the random Gaussian linear model [4]. See [21] for other generalizations of the TLS.

B. Regularized-Structured Total Least Squares Approach

Often the imprecisions on $\mathbf{A}$ and $\mathbf{y}$ have a structure that is desired to be kept intact during the perturbations to obtain a consistent system. For this purpose, the STLS approaches have been proposed as a constrained optimization problem [5], [6], [22]:

$$\begin{aligned} \min_{\mathbf{\Delta A},\mathbf{\Delta y},\mathbf{x}} \quad & \|[\mathbf{\Delta A}\ \mathbf{\Delta y}]\|_F + \mu\|\mathbf{W}\mathbf{x}\|_2 \\ \text{s.t.} \quad & (\mathbf{A}+\mathbf{\Delta A})x = \mathbf{y}+\mathbf{\Delta y} \text{ and} \\ & [\mathbf{\Delta A}\ \mathbf{\Delta y}] \text{ has the same structure as } [\mathbf{A}\mathbf{y}] \end{aligned}$$

where for $\mu \ge 0$, $\mu\|\mathbf{W}\mathbf{x}\|$ is a penalty term that is used to regularize the solution. If the perturbations are such that the columns of $[\mathbf{\Delta A}\ \mathbf{\Delta y}]$ can be written as

$$[\mathbf{\Delta A}\ \mathbf{\Delta y}]_i = \mathbf{G}_i\mathbf{v}, \quad i = 1,\ldots,n+1 \tag{2}$$

where $\mathbf{v}$ is a white noise vector with variance σ^2, the RSTLS optimization can be reduced to the following nonlinear minimization [23], [24] :

$$\begin{pmatrix}\mathbf{x}\\-1\end{pmatrix}^T [\mathbf{A}\ \mathbf{y}]^T(\mathbf{H}_\mathbf{x}\mathbf{H}_\mathbf{x}^T)^{-1}[\mathbf{A}\ \mathbf{y}]\begin{pmatrix}\mathbf{x}\\-1\end{pmatrix} + \mu\|\mathbf{W}\mathbf{x}\|_2 \tag{3}$$

where

$$\mathbf{H}_x = \left(\sum_{i=1}^m \mathbf{x}_i\mathbf{G}_i\right) - \mathbf{G}_{m+1}. \tag{4}$$

Except for block circulant matrices [22], this optimization problem is nonconvex and the developed solution techniques are based on local optimization. In [24], it is shown that for high SNR the covariance matrix of the STLS $(\mu = 0)$ estimator can be approximated by

$$E[(\hat{\mathbf{x}}-\mathbf{x})(\hat{\mathbf{x}}-\mathbf{x})^T] \approx \sigma^2(\mathbf{A}_0^T(\mathbf{H}_\mathbf{x}\mathbf{H}_\mathbf{x}^T)^{-1}\mathbf{A}_0)^{-1}. \tag{5}$$

If $\mathbf{A}_0$ has a large condition number, the variance can be extremely large. It is usually noted in applications that at low SNR, the error variance is even larger than its approximation in (5) [25], [26].

C. Structured Robust Least Squares Approach

As a member of min-max class of techniques, the SRLS estimates $\mathbf{x}$ as the solution to the following optimization problem:

$$\min_{\mathbf{x}} \max_{\|\delta\|_2 \le \rho} \left\| \left(\mathbf{A} + \sum_{i=1}^p \delta_i\mathbf{A}_i\right)\mathbf{x} - \left(\mathbf{y} + \sum_{i=1}^p \delta_i\mathbf{y}_i\right)\right\|_2. \tag{6}$$

SRLS minimizes the worst case residual over a set of perturbations structured with constant matrices $\mathbf{A}_i$ and vectors $\mathbf{y}_i$. As the bound ρ gets larger, the obtained solutions become more

regularized. Hence, the SRLS approach trades accuracy for robustness. Since the min-max criterion is convex, the solution to the SRLS problem can be obtained efficiently by using convex, second-order cone programming [10].

D. Unstructured Bounded Errors-in-Variables (UBEV) Model

One of the important unstructured techniques is known as the bounded errors-in-variables approach, where the inner maximization of the unstructured robust least squares is replaced with a minimization over the allowed perturbations [27], [28]:

$$\min_{\mathbf{x}} \min_{\substack{\|[\mathbf{\Delta A}]\|_F \le \eta_A \\ \|[\mathbf{\Delta y}]\|_2 \le \eta_y}} \|(\mathbf{A}+\mathbf{\Delta A})\mathbf{x} - (\mathbf{y}+\mathbf{\Delta y})\|.$$

As opposed to the cautious approach taken by the min-max techniques, this technique has an optimistic approach and searches for the most favorable perturbation in the allowed set of perturbations. In this sense, it is closer to the TLS approach, but more robust since it does not pursue the consistency as in TLS resulting in sensitivity issues. However, unlike the min-max case, the min-min approach may be degenerate if the residual becomes zero [28]. The nondegenerate and unstructured case has the same form as the TLS solution

$$\mathbf{x}_{\mathrm{UBEV}} = (\mathbf{A}^T\mathbf{A} - \gamma\mathbf{I})^{-1}\mathbf{A}^T\mathbf{y},$$

for some positive valued γ which depends on the perturbation bounds and can be solved using secular equation techniques [29]. For small enough bounds on the perturbations, it can be shown that the value of γ is less than that of σ_{n+1}^2 in the TLS solution given in (1), resulting in less de-regularization than the TLS, hence more robust solutions.

The extended least squares (XLS) criterion [30], which is a blend of LS and STLS is another technique worth noting. In XLS and similar techniques [31], the model errors and measurement errors are distinguished using a weighted minimization.

III. Structured Least Squares With Bounded Data Uncertainties

We will consider the following deterministic relationship between the true variables of a linear system:

$$\mathbf{y}_0 = \mathbf{A}_0\mathbf{x} \tag{7}$$

where the true matrix $\mathbf{A}_0 \in \mathbb{R}^{m\times n}$ maps the unknowns $\mathbf{x}$ to $\mathbf{y}_0$. However neither $\mathbf{A}_0$ nor $\mathbf{y}_0$ is available to us directly. The measured $\mathbf{y}$ is related to $\mathbf{y}_0$ as

$$\mathbf{y} = \mathbf{y}_0 + \sum_{i=1}^{p}\mathbf{y}_i\theta_i + \mathbf{w} \tag{8}$$

where nonzero values of θ_i cause structured uncertainty and $\mathbf{w}$ is additive i.i.d. noise vector with variance σ_w^2. Furthermore, the observed untrue matrix $\mathbf{A}$ is a structurally perturbed version of $\mathbf{A}_0$:

$$\mathbf{A} = \mathbf{A}_0 + \sum_{i=1}^{p}\mathbf{A}_i\theta_i. \tag{9}$$

Here, both $\mathbf{A}_i$ and $\mathbf{y}_i$ are fixed matrices with known structure and θ_i is the ith element of the perturbation vector $\boldsymbol{\theta}$. Note that the structured errors in $\mathbf{A}$ and $\mathbf{y}$ may be correlated in this setup as in the case of linear prediction equations used in harmonic superresolution, AR and ARMA modeling [24], [32]. In those applications, such as deconvolution or system identification where no structure exists in the measurement vector, all $\mathbf{y}_i$'s can be set to zero.

A. The Proposed Optimization Problem

Borrowing the uncertainty set idea from the min-max framework we formulate the following optimization problem that is closer to the maximum likelihood solution in spirit:

$$\min_{\mathbf{x}} \min_{\|\mathbf{W}\boldsymbol{\alpha}\|_2 \le \rho} \left\| \left(\mathbf{A} + \sum_{i=1}^{p}\alpha_i\mathbf{A}_i\right)\mathbf{x} - \left(\mathbf{y} + \sum_{i=1}^{p}\alpha_i\mathbf{y}_i\right)\right\|_2^2 \tag{10}$$

which is a generalization of the bounded errors-in-variables model to the structured case [27]. Here, $\mathbf{W}$ is a positive-definite weighting matrix which may be used to incorporate prior knowledge of perturbations, e.g., imposing frequency domain constraints. Unlike the min-max case this optimization problem is nonconvex in general. In the following, we consider the cases of deterministic and random perturbations, and we will assume that ρ is small enough so that the objective of (10) is always positive.

1) Deterministic Perturbations: In Appendix A, given observations of $\mathbf{y}$ and $\mathbf{A}$, we show that there is no unbiased estimator of $\mathbf{x}$ with finite variance if $p > m - n$. This is because of the fact that for $p > m - n$ the Fisher Information Matrix is singular for a deterministic unknown vector $\boldsymbol{\theta}$. In particular this result applies to commonly encountered Toeplitz and Hankel structures which have $p = m+n-1$. If the uncertainty bounds of measurements are known beforehand, a reasonable biased estimate can be obtained even though the Cramér–Rao lower bound is infinite, by using the proposed constrained optimization. This case is demonstrated in the signal restoration application in Section VIII-B where the impulse response has an uncertainty with known bounds.

2) Random Perturbations: As a data preprocessing step, if the actual perturbation $\boldsymbol{\theta}$ is modeled as a random vector with nonzero-mean $\mathbf{m}_{\boldsymbol{\theta}}$ and positive-definite covariance matrix $\mathbf{\Sigma}$, one can define a new set of matrices and vectors:

$$\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{m}_{\boldsymbol{\theta}}\sum_{i=1}^{p}\mathbf{A}_i, \quad \tilde{\mathbf{y}} = \mathbf{y} + \mathbf{m}_{\boldsymbol{\theta}}\sum_{i=1}^{p}\mathbf{y}_i \tag{11}$$

$$\tilde{\mathbf{A}}_j = \sum_{i=1}^{p}P_{ij}\mathbf{A}_i, \quad \tilde{\mathbf{y}}_j = \sum_{i=1}^{p}P_{ij}\mathbf{y}_i \tag{12}$$

where $\mathbf{P}$ is the Cholesky factor of the covariance matrix, $\mathbf{\Sigma} = \mathbf{P}\mathbf{P}^T$. These new set of matrices enable us to use a whitened perturbation vector. Hence, without loss of generality, we can assume $\boldsymbol{\theta}$ is a zero-mean random vector containing independent identically distributed elements with variance σ^2. Then we have the expectation

$$E[\mathbf{A}^T\mathbf{A}] = \mathbf{A}_0^T\mathbf{A}_0 + \sum_i\sum_j \mathbf{A}_i^T\mathbf{A}_j E[\boldsymbol{\theta}_i\boldsymbol{\theta}_j] \tag{13}$$

$$= \mathbf{A}_0^T\mathbf{A}_0 + \sigma^2\sum_i \mathbf{A}_i^T\mathbf{A}_i. \tag{14}$$

For Toeplitz or Hankel structures, this expression can be further simplified to

$$E[\mathbf{A}^T\mathbf{A}] = \mathbf{A}_0^T\mathbf{A}_0 + m\sigma^2\mathbf{I}. \quad (15)$$

The above expression and also (14) illustrate the fact that, as a result of the diagonal loading term, even if $\mathbf{A}_0$ is an ill-conditioned matrix, the observed matrix may be well-conditioned. Hence, searching for a consistent system $\mathbf{A}_0\mathbf{x} = \mathbf{y}_0$ by employing perturbations on the observed system $(\mathbf{A}, \mathbf{y})$ could result in an inadmissible estimator with large variance. Adding a regularization term as in the RSTLS formulation may be a remedy for this problem. However as will be shown next, by using the proposed approach defined in (10), it is possible to find an estimator with smaller MSE.

B. The Mean Squared Error of the SLS-BDU Estimate

The proposed estimator falls into the class of biased estimators for the linear model where bias-variance tradeoff is of primary importance [33], [34]. To provide further insight, we next derive an MSE bound which indicates a similar tradeoff. We begin with the following definitions.

Definition 1: For a constant $\boldsymbol{\alpha} \in \mathbb{R}^p$ define functions,

$$\mathbf{A}(\boldsymbol{\alpha}) \triangleq \mathbf{A} + \sum_{i=1}^{p} \alpha_i \mathbf{A}_i, \quad \mathbf{y}(\boldsymbol{\alpha}) \triangleq \mathbf{y} + \sum_{i=1}^{p} \alpha_i \mathbf{y}_i. \quad (16)$$

Without loss of generality, we will assume that $\mathbf{y}_i = 0\ \forall i$ in the rest of the paper, since they can be embedded into $\tilde{\mathbf{A}}_i \triangleq [\mathbf{A}_i \mathbf{y}_i]$'s as follows:

$$\mathbf{A}(\boldsymbol{\alpha})\mathbf{x} - \mathbf{y}(\boldsymbol{\alpha}) = \mathbf{A} + \left(\sum_i [\mathbf{A}_i \mathbf{y}_i]\alpha_i\right)[\mathbf{x} - 1]^T - \mathbf{y}. \quad (17)$$

Then the following theorem characterizes the MSE of the proposed estimator.

Theorem 3.1: For $\mathbf{A}(\boldsymbol{\alpha})$ which is of full column rank for $\|\mathbf{W}\boldsymbol{\alpha}\|_2 \leq \rho$, the optimal $\hat{\mathbf{x}}$ for the proposed optimization in (10) has the following MSE upper bound:

$$E\left[\|\hat{\mathbf{x}} - \mathbf{x}\|_2^2\right] \leq \left(\|\mathbf{x}\|_2^2 E\,\|\mathbf{A}(\boldsymbol{\alpha}^*) - \mathbf{A}_0)\|_2^2 + n\sigma_{\mathbf{w}}^2\right) E\left[\frac{1}{\sigma_{\boldsymbol{\alpha}^*}^2}\right]$$

where $\boldsymbol{\alpha}^*$ is the optimal $\boldsymbol{\alpha}$ of (10) and $\sigma_{\boldsymbol{\alpha}^*}$ is the minimum singular value of $\mathbf{A}(\boldsymbol{\alpha}^*)$.

Proof: By analytically minimizing (10) over $\mathbf{x}$ for a fixed $\boldsymbol{\alpha}$ as an ordinary least squares problem, (10) reduces to

$$\min_{\|\mathbf{W}\boldsymbol{\alpha}\| \leq \rho} \left\|\mathbf{A}(\boldsymbol{\alpha})\mathbf{A}(\boldsymbol{\alpha})^\dagger \mathbf{y} - \mathbf{y}\right\|_2^2 = \min_{\|\mathbf{W}\boldsymbol{\alpha}\| \leq \rho} \left\|\mathbf{P}_{\boldsymbol{\alpha}}^{\perp}\mathbf{y}\right\|_2^2 \quad (18)$$

where $\mathbf{P}_{\boldsymbol{\alpha}}^{\perp} \triangleq \mathbf{I} - \mathbf{A}(\boldsymbol{\alpha})\mathbf{A}(\boldsymbol{\alpha})^\dagger$ is the projector matrix of the subspace perpendicular to the $\mathrm{Range}(\mathbf{A}(\boldsymbol{\alpha}))$ and we assumed $\mathbf{A}(\boldsymbol{\alpha})$ is of full column rank for $\|\mathbf{W}\boldsymbol{\alpha}\|_2 \leq \rho$. Thus, SLS-BDU estimator chooses the $\boldsymbol{\alpha}$ that minimizes the norm of the observation $\mathbf{y}(\boldsymbol{\alpha})$ which lies out of the range of $\mathbf{A}(\boldsymbol{\alpha})$.

The SLS-BDU estimate $\mathbf{x}$ which minimizes (10) can be written in terms of the optimal $\boldsymbol{\alpha}^*$ of (18) as

$$\hat{\mathbf{x}}_{SLS-BDU} = \mathbf{A}(\boldsymbol{\alpha}^*)^\dagger \mathbf{y}. \quad (19)$$

Since $\mathbf{y} = \mathbf{A}_0\mathbf{x} + \mathbf{w}$, the MSE of (19) can be written as [33]

$$\begin{aligned} E\left[\|\hat{\mathbf{x}} - \mathbf{x}\|_2^2\right] &= E\left[\left\|(\mathbf{A}(\boldsymbol{\alpha}^*)^\dagger\mathbf{A}_0 - \mathbf{I})\mathbf{x} + \mathbf{A}(\boldsymbol{\alpha}^*)^\dagger\mathbf{w}\right\|_2^2\right] \\ &= E\left[\left\|(\mathbf{A}(\boldsymbol{\alpha}^*)^\dagger\mathbf{A}_0 - \mathbf{I})\mathbf{x}\right\|_2^2\right] \\ &\quad + E\left[\mathrm{Tr}\{\mathbf{A}(\boldsymbol{\alpha}^*)^{\dagger T}\mathbf{A}(\boldsymbol{\alpha}^*)^\dagger \mathbf{w}\mathbf{w}^T\}\right]. \quad (20) \end{aligned}$$

Since $E[\mathrm{Tr}\{\mathbf{A}(\boldsymbol{\alpha}^*)^{\dagger T}\mathbf{A}(\boldsymbol{\alpha}^*)^\dagger \mathbf{w}\mathbf{w}^T\}] = \sigma_{\mathbf{w}}^2 E[\left\|\mathbf{A}(\boldsymbol{\alpha}^*)^\dagger\right\|_F^2]$, we get

$$E[\|\hat{\mathbf{x}} - \mathbf{x}\|_2^2] = E\left\|(\mathbf{A}(\boldsymbol{\alpha}^*)^\dagger\mathbf{A}_0 - \mathbf{I})\mathbf{x}\right\|_2^2 + \sigma_{\mathbf{w}}^2 E\left\|\mathbf{A}(\boldsymbol{\alpha}^*)^\dagger\right\|_F^2. \quad (21)$$

The following inequalities that are valid for full column rank matrices $\mathbf{F}$ and $\mathbf{G}$ help to obtain the desired upper bound:

$$\left\|\mathbf{F}^\dagger\mathbf{G} - \mathbf{I}\right\|_2 = \left\|\mathbf{F}^\dagger(\mathbf{G} - \mathbf{F})\right\|_2 \leq \left\|\mathbf{F}^\dagger\right\|_2 \|\mathbf{G} - \mathbf{F}\|_2,$$

$$\text{and} \quad \left\|\mathbf{F}^\dagger\right\|_F^2 = \sum_{i=1}^{n} \frac{1}{\sigma_i^2(\mathbf{F})} \leq \frac{n}{\sigma_{\min}^2(\mathbf{F})}.$$

Using the previous inequality, we can upper bound (21) using

$$\begin{aligned} &E\left[\left\|\mathbf{A}(\boldsymbol{\alpha}^*)^\dagger\right\|_2^2 \|\mathbf{A}(\boldsymbol{\alpha}^*) - \mathbf{A}_0\|_2^2 \|\mathbf{x}\|_2^2\right] + E\left[\sum_{i=1}^{n} \frac{\sigma_{\mathbf{w}}^2}{\sigma_{\boldsymbol{\alpha}^*,i}^2}\right] \\ &\qquad \leq \left(\|\mathbf{x}\|_2^2 E\,\|\mathbf{A}(\boldsymbol{\alpha}^*) - \mathbf{A}_0)\|_2^2 + n\sigma_{\mathbf{w}}^2\right) E\left[\frac{1}{\sigma_{\boldsymbol{\alpha}^*}^2}\right]. \end{aligned}$$

■

The obtained upper bound clearly shows that the MSE of the estimate has two parts: the part that increase with the difference between $\mathbf{A}_0$ and its estimate $\mathbf{A}(\alpha^*)$ and the part that increases with the Frobenious norm of the $\mathbf{A}^\dagger(\alpha^*)$. Since the Frobenious norm of $\mathbf{A}^\dagger(\alpha^*)$ can be very large for an ill-conditioned $\mathbf{A}_0$ when the estimate $\mathbf{A}(\alpha^*)$ gets close to $\mathbf{A}_0$, the second part of the bound can get extremely large. Therefore, the main idea behind the proposed estimator is to bound the allowed perturbations such that the MSE in (21) is near optimal. It is straightforward to show that when $\rho = 0$, the SLS-BDU solution is equal to the ordinary Least Squares solution. Since the STLS optimization seeks a minimal norm perturbation to zero out the cost function in (10), the solution given by STLS is identical to the SLS-BDU solution for a large enough value of the perturbation magnitude bound ρ. However that choice of ρ usually results a large MSE in (21) as previously noted in numerical results of [30].

C. MSE Comparison of SLS-BDU and STLS

Using the MSE bound in (3.1) we derive the condition in which the proposed estimator has smaller MSE then the Maximum Likelihood STLS estimator and interpret the result.

Theorem 3.2: For deterministic and bounded perturbations $\boldsymbol{\theta}$, let σ_A and σ_0 be the minimum singular values of $\mathbf{A}$ and $\mathbf{A}_0$, respectively, and define

$$S \triangleq \begin{cases} \sqrt{p}\max_i \|A_i\|_2 & \text{Arbitrary structure} \\ \max_i \|A_i\|_F & \text{Nonoverlapping structure} \\ \sqrt{n} & \text{Toeplitz or Hankel.} \end{cases} \quad (22)$$

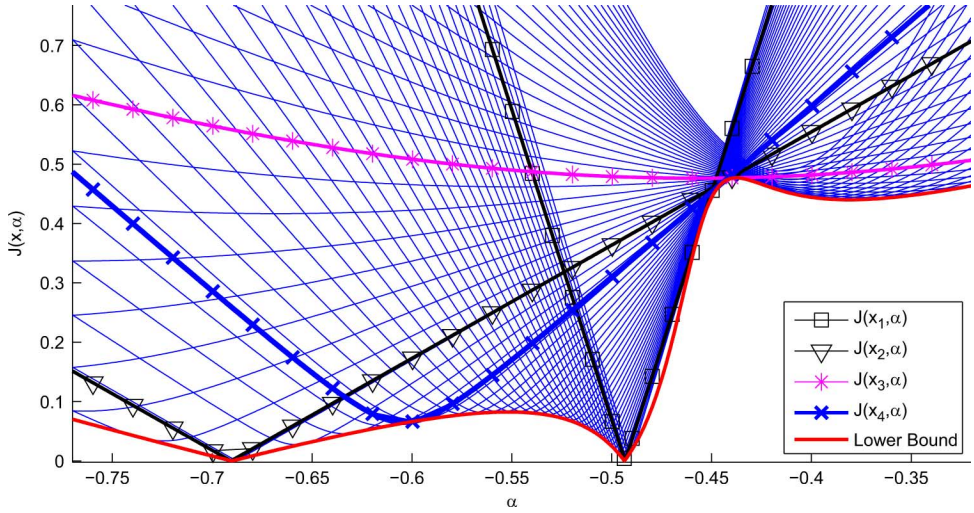

Fig. 1. Cost $J(\mathbf{x}, \boldsymbol{\alpha})$ in (26) plotted for a set of estimators on top of each other.

If the following holds:

$$(\rho + \|\theta\|)^2 S^2 \frac{\|\mathbf{x}_0\|_2^2}{n\sigma_w^2} + 1 \leq \left(\frac{\sigma_A - \rho S}{\sigma_0} \right)_+^2, \qquad (23)$$

then the asymptotically MSE of SLS-BDU with weight $\mathbf{W} = \mathbf{I}$, is strictly smaller than STLS.

Proof: See Appendix B.

Remark 1: Note that the expression $R \triangleq \|\mathbf{x}_0\|_2^2 / n\sigma_w^2$ in (23) denotes the signal-to-noise ratio (SNR), e.g., if $\mathbf{x}_0$ were a zero-mean Gaussian vector with variance σ_x^2, then $E[R] = \sigma_x^2/\sigma_w^2$.

Remark 2: The right-hand side of (23) is expected to be larger than 1 since, $\sigma_A \gg \sigma_0$ by the observation in (15).

Therefore, Theorem 3.2 asserts that, when SNR is sufficiently low, the condition in (23) is satisfied and the proposed SLS-BDU has smaller error than STLS. Furthermore, for ill-conditioned problems where σ_0 is small, the condition (23) may hold also for large SNR values. In Section VIII, we show that this theoretical result is in good agreement with numerical experiments.

IV. Analysis of Estimator Performance in an Illustrative Example

Consider the single parameter equation $\mathbf{A}(\alpha)x = \mathbf{y}(\alpha)$ below:

$$\begin{bmatrix} a_1 + \alpha \\ a_2 \end{bmatrix} x = \begin{bmatrix} y_1 \\ y_2 - \alpha \end{bmatrix}. \qquad (24)$$

The corresponding structures are

$$\mathbf{A}_1 = [1 \quad 0]^T, \quad \mathbf{y}_1 = [0 \quad -1]^T. \qquad (25)$$

Define the cost of x given α by

$$J(x, \alpha) \triangleq \|\mathbf{A}(\alpha)x - \mathbf{y}(\alpha)\|_2^2 \qquad (26)$$

which corresponds to a constant multiple of the negative log-likelihood given α for the observation $\mathbf{y}(\alpha) = \mathbf{A}(\alpha)\mathbf{x} + \mathbf{w}$ where $\mathbf{w}$ is a zero-mean Gaussian random variable. Fig. 1 depicts $J(x, \alpha)$ for several values of x plotted on top of each other for $\{a_1, a_2, y_1, y_2\} = \{0.46, 0.023, 0.38, -0.73\}$. The lower bound achievable for any $\mathbf{x}$ is given by

$$\min_x \|\mathbf{A}(\alpha)x - \mathbf{y}(\alpha)\|_2^2 = \left\|\mathbf{P}_{\boldsymbol{\alpha}}^{\perp}\mathbf{y}(\alpha)\right\|_2^2 \qquad (27)$$

which can be easily shown to be zero only for at most two values of α given by

$$\alpha_{1,2} = \frac{y_2 - a_1}{2} \pm \sqrt{\left(\frac{y_2 - a_1}{2}\right)^2 + a_1 y_2 - a_2 y_1}. \qquad (28)$$

By carefully inspecting Fig. 1, the two solutions of (28) $\alpha_1 = -0.69$ and $\alpha_2 = -0.49$ yields the following estimates for $\mathbf{x}$:

$$\begin{aligned} x_1 &= \mathbf{A}(\alpha_1)^{\dagger}\mathbf{y}(\alpha_1) = -1.62 \quad \text{and} \\ x_2 &= \mathbf{A}(\alpha_2)^{\dagger}\mathbf{y}(\alpha_2) = -10 \end{aligned} \qquad (29)$$

neither of which is robust since they have steeply rising linear costs for a small change in α. We utilize this observation later in Section VII by using the gradient of the lower bound as a measure of this sensitivity. Note that given any random or deterministic perturbation α, because of the consistency requirement, STLS and RSTLS methods produce either x_1 or x_2. If the system were consistent originally, i.e., $\mathbf{A}_0\mathbf{x} = \mathbf{y}_0$, the expected MSE and residual of such consistency constrained estimators would be large because of the distance $|x_1 - x_2|$. Note that the residual of x_1 is extremely large if α_2 is the true parameter.

In Fig. 1, the cost corresponding to a min-max solutions $x_3 = 0.75$ is also shown. Although the cost min-max solution is less sensitive to the variations in α, its average is considerably large.

However, the SLS-BDU solution given by (10) achieves the lower bound in (27) for some α^*, which corresponds to an inconsistent system $\{\mathbf{A}(\alpha^*), \mathbf{y}(\alpha^*)\}$, but balances robustness and accuracy by abandoning the consistency condition. An example of one such solution is given by $x_4 = -2.73$, which is neither over conservative as the min-max solution x_3 or over optimistic as the STLS solution x_1.

V. Fréchet Derivatives and Gradient Flow

In this section, Fréchet derivatives are introduced to analyze the gradient of the SLS-BDU cost function in detail. In addition, some analytical results on the rotation of the gradient around singularities, and the existence of consistencies as hyperplanes are presented.

A. Differentiation of Pseudoinverses and Projectors

The $m \times n$ matrix function $\mathbf{A}(\boldsymbol{\alpha}) = \mathbf{A} + \sum_{i=1}^{p} \alpha_i \mathbf{A}_i$ is a mapping between $\mathbb{R}^p$ and the space of linear transformations $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$. Assuming $\mathrm{Rank}(\mathbf{A}(\boldsymbol{\alpha}))$ is constant for $\|\mathbf{W}\boldsymbol{\alpha}\|_2 \leq \rho$, the pseudoinverse $\mathbf{A}(\boldsymbol{\alpha})^\dagger$ and the projector $\mathbf{P}_{\boldsymbol{\alpha}}^{\perp} = \mathbf{I} - \mathbf{A}(\boldsymbol{\alpha})\mathbf{A}(\boldsymbol{\alpha})^\dagger$ are both Fréchet differentiable with respect to $\boldsymbol{\alpha}$ and closed form formulas were derived in [19]. Formalism on Fréchet derivatives can be found in [35]. Here we provide some known facts as well as new results relevant to our application.

Definition 2: The Fréchet derivative of $\mathbf{A}(\boldsymbol{\alpha})$ denoted by $\mathbf{DA}(\boldsymbol{\alpha})$ is a tridimensional tensor, formed with p matrices of size $m \times n$ containing partial derivatives of the elements of $\mathbf{A}$ with respect to α_i, i.e., $[\mathbf{DA}(\boldsymbol{\alpha})]_i \triangleq \partial/\partial\alpha_i \mathbf{A}(\boldsymbol{\alpha})$.

The Fréchet derivative of $\mathbf{P}_{\boldsymbol{\alpha}}^{\perp}$ is given in [19] as

$$\begin{aligned}\mathbf{DP}_{\boldsymbol{\alpha}}^{\perp} &= -\mathbf{DP}_{\boldsymbol{\alpha}} \\ &= -\mathbf{P}_{\boldsymbol{\alpha}}^{\perp}\mathbf{DA}(\boldsymbol{\alpha})\mathbf{A}(\boldsymbol{\alpha})^\dagger - (\mathbf{P}_{\boldsymbol{\alpha}}^{\perp}\mathbf{DA}(\boldsymbol{\alpha})\mathbf{A}(\boldsymbol{\alpha})^\dagger)^T. \quad (30)\end{aligned}$$

The following lemma characterizes each entry in the gradient vector of the SLS-BDU cost function given in (18).

Lemma 5.1: Let $\mathbf{y}(\boldsymbol{\alpha})^\perp \triangleq \mathbf{P}_{\boldsymbol{\alpha}}^{\perp}\mathbf{y}(\boldsymbol{\alpha})$ and $\mathbf{x}_{\boldsymbol{\alpha}} \triangleq \mathbf{A}(\boldsymbol{\alpha})^\dagger \mathbf{y}(\boldsymbol{\alpha})$ then,

$$\frac{1}{2}\frac{\partial}{\partial\alpha_i}\left\|\mathbf{P}_{\boldsymbol{\alpha}}^{\perp}\mathbf{y}(\boldsymbol{\alpha})\right\|_2^2 = \left\langle \mathbf{y}(\boldsymbol{\alpha})^\perp,\ \mathbf{y}_i - \mathbf{A}_i\mathbf{x}_{\boldsymbol{\alpha}} \right\rangle. \quad (31)$$

Proof:

$$\begin{aligned}\nabla_{\boldsymbol{\alpha}}\left\|\mathbf{P}_{\boldsymbol{\alpha}}^{\perp}\mathbf{y}(\boldsymbol{\alpha})\right\|_2^2 &= \nabla_{\boldsymbol{\alpha}}\mathbf{y}(\boldsymbol{\alpha})^T\mathbf{P}_{\boldsymbol{\alpha}}^{\perp}\mathbf{y}(\boldsymbol{\alpha}) \\ &= \mathbf{Dy}(\boldsymbol{\alpha})^T\mathbf{P}_{\boldsymbol{\alpha}}^{\perp}\mathbf{y}(\boldsymbol{\alpha}) + \mathbf{y}(\boldsymbol{\alpha})^T\mathbf{DP}_{\boldsymbol{\alpha}}^{\perp}\mathbf{y}(\boldsymbol{\alpha}) \\ &\quad + \mathbf{y}(\boldsymbol{\alpha})^T\mathbf{P}_{\boldsymbol{\alpha}}^{\perp}\mathbf{Dy}(\boldsymbol{\alpha}) \\ &= 2\mathbf{y}(\boldsymbol{\alpha})^T\mathbf{P}_{\boldsymbol{\alpha}}^{\perp}\mathbf{Dy}(\boldsymbol{\alpha}) \\ &\quad - 2\mathbf{y}(\boldsymbol{\alpha})^T\mathbf{P}_{\boldsymbol{\alpha}}^{\perp}\mathbf{DA}(\boldsymbol{\alpha})\mathbf{A}(\boldsymbol{\alpha})^\dagger\mathbf{y}(\boldsymbol{\alpha}) \\ &= 2\mathbf{y}(\boldsymbol{\alpha})^T\mathbf{P}_{\boldsymbol{\alpha}}^{\perp}(\mathbf{Dy}(\boldsymbol{\alpha}) - \mathbf{DA}(\boldsymbol{\alpha})\mathbf{A}(\boldsymbol{\alpha})^\dagger\mathbf{y}(\boldsymbol{\alpha})) \\ &= 2\left\langle \mathbf{P}_{\boldsymbol{\alpha}}^{\perp}\mathbf{y}(\boldsymbol{\alpha}), \mathbf{y}_i - \mathbf{A}_i\mathbf{A}(\boldsymbol{\alpha})^\dagger\mathbf{y}(\boldsymbol{\alpha}) \right\rangle,\end{aligned}$$

since $[\mathbf{DA}(\boldsymbol{\alpha})]_i = \mathbf{A}_i$ and $[\mathbf{Dy}(\boldsymbol{\alpha})]_i = \mathbf{y}_i$. ∎

B. The Gradient Flow in a Simple Illustrative Case

Consider the following two parameter case:

$$\mathbf{A}(\boldsymbol{\alpha}) = [1+\alpha_1 \quad 1+\alpha_2]^T, \quad \mathbf{y}(\boldsymbol{\alpha}) = \mathbf{y} = [0 \quad 1]^T \quad (32)$$

which is consistent, i.e., $\mathbf{y}(\boldsymbol{\alpha}) \in \mathrm{R}(\mathbf{A}(\boldsymbol{\alpha}))$ for $\alpha_1 = -1$. The vector field $-\nabla_{\boldsymbol{\alpha}}\left\|\mathbf{P}_{\boldsymbol{\alpha}}^{\perp}\mathbf{y}(\boldsymbol{\alpha})\right\|_2^2$, which is calculated by (31) is shown in Fig. 2. The gradient norm is zero on two straight lines $\boldsymbol{\alpha}_1 = -1$ and $\boldsymbol{\alpha}_2 = -1$ denoting minimum and maximum of (18) which intersect at the singular point $(-1,-1)$. The gradient field rotates around the singularity by flowing from the maximum $(\alpha_2 = -1)$ to minimum $(\alpha_1 = -1)$ and the gradient norm increases gradually as $\boldsymbol{\alpha}$ gets closer to the singular point $(-1,-1)$. In Fig. 3, the solution of STLS and the proposed solution (10) are compared on a diagram for the example in (32). The points p_1 and p_2 denote the corrected vectors $\mathbf{A}(\boldsymbol{\alpha})$ for STLS and proposed SLS-BDU for a given ρ and $\mathbf{W} = \mathbf{I}$, respectively. p_1 denotes the closest consistent system while p_2 is the tangent point of the line passing through singularity to the circular boundary with radius ρ. This tangent point geometry was also encountered in unstructured min-min and min-max problems [29]. It is evident that with a small ρ, the corrected system is better conditioned with the proposed method. Note that for a larger value of ρ, the consistency lines will be in the allowed set of perturbations and the SLS-BDU and the STLS solutions would be identical.

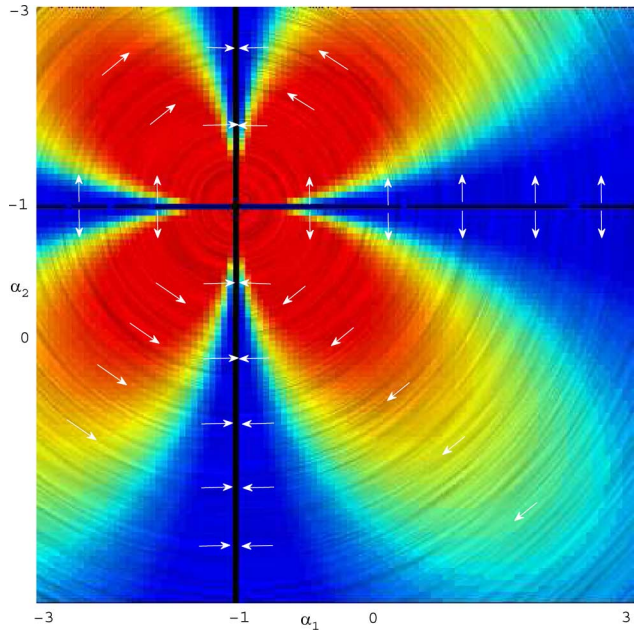

Fig. 2. Negative gradient field for the two parameter case in (32). All vectors rotate around the singularity at $(-1,-1)$.

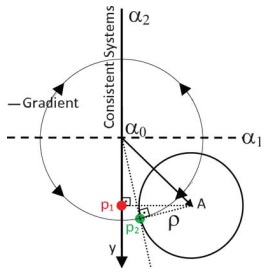

Fig. 3. Gradient flow diagram for the two parameter case in (32). The points p_1 and p_2 indicate the perturbations done by STLS and SLS-BDU, respectively.

C. Analytical Results on the Gradient Flow

In this section we present theoretical results which shed light on the interesting geometry of Fig. 2.

Theorem 5.2: Rotation Around a Singularity: If $\mathrm{Range}(\mathbf{A}(\boldsymbol{\alpha}_0)) \subset \mathrm{Range}(\mathbf{A}(\boldsymbol{\alpha}))$, the gradient field $\nabla_{\boldsymbol{\alpha}}\left\|\mathbf{P}_{\boldsymbol{\alpha}}^{\perp}\mathbf{y}(\boldsymbol{\alpha})\right\|$ is orthogonal to $\boldsymbol{\alpha} - \boldsymbol{\alpha}_0$, i.e.,

$$\left\langle \nabla_{\boldsymbol{\alpha}}\left\|\mathbf{P}_{\boldsymbol{\alpha}}^{\perp}\mathbf{y}(\boldsymbol{\alpha})\right\|_2^2,\ \boldsymbol{\alpha} - \boldsymbol{\alpha}_0 \right\rangle = 0. \quad (33)$$

Proof: By using Lemma 5.1, we get

$$\begin{aligned}
&-\frac{1}{2}\left\langle \nabla_{\boldsymbol{\alpha}} \left\|\mathbf{P}_{\boldsymbol{\alpha}}^{\perp}\mathbf{y}(\boldsymbol{\alpha})\right\|_2^2,\, \boldsymbol{\alpha}_0-\boldsymbol{\alpha}\right\rangle \\
&\quad=\sum_i \left\langle \mathbf{P}_{\boldsymbol{\alpha}}^{\perp}\mathbf{y}(\boldsymbol{\alpha}),\, \mathbf{A}_i\mathbf{A}(\boldsymbol{\alpha})^{\dagger}\mathbf{y}(\boldsymbol{\alpha})\right\rangle(\alpha_i-\alpha_{0i}) \\
&\quad=\left\langle \mathbf{P}_{\boldsymbol{\alpha}}^{\perp}\mathbf{y}(\boldsymbol{\alpha}),\, \sum_i \mathbf{A}_i[\alpha_{0i}-\alpha_i]\mathbf{A}(\boldsymbol{\alpha})^{\dagger}\mathbf{y}(\boldsymbol{\alpha})\right\rangle \\
&\quad=\mathbf{y}(\boldsymbol{\alpha})^T\mathbf{P}_{\boldsymbol{\alpha}}^{\perp}\left[\mathbf{A}(\boldsymbol{\alpha}_0)-\mathbf{A}(\boldsymbol{\alpha})\right]\mathbf{A}(\boldsymbol{\alpha})^{\dagger}\mathbf{y}(\boldsymbol{\alpha}). \qquad (34)
\end{aligned}$$

Because $\mathrm{Range}(\mathbf{A}(\boldsymbol{\alpha}_0)) \subset \mathrm{Range}(\mathbf{A}(\boldsymbol{\alpha}))$ implies $\mathrm{Range}(\mathbf{A}(\boldsymbol{\alpha}_0)-\mathbf{A}(\boldsymbol{\alpha})) \subset \mathrm{Range}(\mathbf{A}(\boldsymbol{\alpha}))$, $\mathbf{P}_{\boldsymbol{\alpha}}^{\perp}(\mathbf{A}(\boldsymbol{\alpha}_0)-\mathbf{A}(\boldsymbol{\alpha}))=0$, thus (34) is zero. ∎

Remark 3: Theorem 5.2 reveals the interesting geometry of Fig. 2, where all vectors absolutely rotate around the singularity $(-1,-1)$, since $\mathbf{A}(-1,-1)$ is of rank zero.

The next theorem states that every singularity is arbitrarily close to a consistency for a range of structures which are commonly encountered in applications.

Theorem 5.3: If there is no structure, or the structure is of Toeplitz or Hankel type, then, for $\mathbf{A}(\boldsymbol{\alpha})^T\mathbf{A}(\boldsymbol{\alpha})$ singular, there exists a vector $\boldsymbol{\epsilon}$ with arbitrarily small norm, satisfying $\mathbf{k} \in \mathrm{Range}(\mathbf{A}(\boldsymbol{\alpha}+\boldsymbol{\epsilon}))$ for any arbitrary $\mathbf{k} \in \mathbb{R}^m$.

Proof: First consider the unstructured case and let $\boldsymbol{v} \in \mathrm{Null}(\mathbf{A})$. Then

$$\left(\mathbf{A}(\alpha)+\frac{\epsilon}{\boldsymbol{v}^T\boldsymbol{v}}\boldsymbol{k}\boldsymbol{v}^T\right)\boldsymbol{v}=\epsilon\boldsymbol{k} \qquad (35)$$

which implies $\boldsymbol{k} \in \mathrm{Range}(\mathbf{A}(\alpha+\boldsymbol{\epsilon}))$. For the Toeplitz case, let $\boldsymbol{v} \in \mathrm{Null}(\mathbf{A}(\boldsymbol{\alpha}))$, then

$$\left(\mathbf{A}(\boldsymbol{\alpha})+\sum_i \theta_i\mathbf{A}_i\right)\mathbf{v}=\sum_i \theta_i\mathbf{v}_i=\mathbf{V}\boldsymbol{\theta} \qquad (36)$$

where $\boldsymbol{v}_i \triangleq \mathbf{A}_i\boldsymbol{v}$ and $\boldsymbol{V} \triangleq [\boldsymbol{v}_1 \cdots \boldsymbol{v}_{m+n-1}]$.

Because of the Toeplitz structure, it is straightforward to show that $\boldsymbol{V}$ is of full row rank if $\boldsymbol{v} \neq \mathbf{0}$ [36]. Then, for any $\boldsymbol{\epsilon}$, $\boldsymbol{\theta}=\epsilon\boldsymbol{V}^{\dagger}\boldsymbol{k}$ satisfies $\mathbf{A}(\boldsymbol{\alpha}+\theta)\boldsymbol{v}=\boldsymbol{V}\epsilon\boldsymbol{V}^{\dagger}\boldsymbol{k}=\epsilon\boldsymbol{k}$ as desired. The same argument follows similarly for the Hankel structure or any other structure for which $\boldsymbol{V}$ is of full row rank. ∎

Theorem 5.4: For Toeplitz or Hankel structured problems, every point $\boldsymbol{\alpha}$ such that $\mathbf{A}(\boldsymbol{\alpha})^T\mathbf{A}(\boldsymbol{\alpha})$ is singular, lies on an n-dimensional hyperplane of consistent systems.

Proof: Let $\boldsymbol{v} \in \mathrm{Null}(\mathbf{A})$ and $\mathbf{V}=\mathbf{U}[\boldsymbol{\Sigma}\,0][\mathbf{V}_1\,\mathbf{V}_2]^T$ be the SVD [37] of $\mathbf{V}$ defined after (36). Then $\mathbf{A}(\alpha+\boldsymbol{\epsilon})\mathbf{v}=\mathbf{V}\boldsymbol{\epsilon}=\beta_0\mathbf{y}$ has solution

$$\begin{aligned}
\boldsymbol{\epsilon}=\beta_0\mathbf{V}^{\dagger}\mathbf{y}+\mathbf{V}_2\boldsymbol{\beta}&=\beta_0\mathbf{V}_1\boldsymbol{\Sigma}^{-1}\mathbf{U}\mathbf{y}+\mathbf{V}_2\boldsymbol{\beta} \qquad (37)\\
&=[\mathbf{V}_1\boldsymbol{\Sigma}^{-1}\mathbf{U} \quad \mathbf{V}_2]\tilde{\boldsymbol{\beta}} \qquad (38)
\end{aligned}$$

for all $\tilde{\boldsymbol{\beta}}=[\beta_0 \quad \boldsymbol{\beta}]^T \in \mathbb{R}^n$. Therefore, since $\mathbf{V}_1$ and $\mathbf{V}_2$ are orthogonal, any vector $\mathbf{y}$ is in Range $(\mathbf{A}(\boldsymbol{\alpha}+\boldsymbol{\epsilon}))$ for any $\boldsymbol{\epsilon}$ which is in the n-dimensional columnspace of $[\mathbf{V}_1\boldsymbol{\Sigma}^{-1}\mathbf{U} \quad \mathbf{V}_2]$. ∎

Theorems 5.3 and 5.4 illustrate the ill-conditioned nature of the consistency constraints. Note that the structure in (32) is Toeplitz and the singularity lies in a one-dimensional plane of consistent systems. Theorems 5.2 and 5.4 show that the rotation property and the proximity of consistencies to singularities are valid for many systems of interest with arbitrary dimensions. Therefore, the above observations for the simple example (32) are commonly encountered in practice.

VI. Solving the SLS-BDU Optimization Problem

In this section, three iterative techniques are presented to solve the nonconvex optimization problem of the SLS-BDU approach.

A. Individual Optimization by Alternating Minimizations

Although the SLS-BDU cost function is nonconvex in $\mathbf{x}$ and $\boldsymbol{\alpha}$ together, it is convex for $\mathbf{x}$ and $\boldsymbol{\alpha}$ individually. It is easy to see that for a fixed $\boldsymbol{\alpha}$, the cost in (10) is convex over $\mathbf{x}$. The following derivation shows that for a fixed $\mathbf{x}$, the cost is convex over $\boldsymbol{\alpha}$ as well.

$$\left\|(\mathbf{A}\mathbf{x}-\mathbf{y})+\sum_i \alpha_i(\mathbf{A}_i\mathbf{x}-\mathbf{y}_i)\right\| = \|(\mathbf{A}\mathbf{x}-\mathbf{y})+\mathbf{U}(\mathbf{x})\boldsymbol{\alpha}\|,$$
$$\text{where } \mathbf{U}(\mathbf{x}) \triangleq [(\mathbf{A}_1\mathbf{x}-\mathbf{y}_1) \cdots (\mathbf{A}_p\mathbf{x}-\mathbf{y}_p)]$$

which is convex over $\boldsymbol{\alpha}$ for a fixed $\mathbf{x}$. Therefore, alternating minimizations, as in the minimization of extended least squares criterion [30], can be performed:

Algorithm 1: Alternating Minimizations

$\mathbf{x}^0 \leftarrow \mathbf{A}^{\dagger}\mathbf{y}$, $k \leftarrow 0$
while $\left\|\mathbf{x}^k-\mathbf{x}^{k-1}\right\| > \epsilon$ **do**
  $\boldsymbol{\alpha}^{k+1} \leftarrow \arg\min_{\|\mathbf{W}\boldsymbol{\alpha}\| \le \rho} \left\|(\mathbf{A}\mathbf{x}^k-\mathbf{y})+\mathbf{U}(\mathbf{x}^k)\boldsymbol{\alpha}\right\|$
  $\mathbf{x}^{k+1} \leftarrow \arg\min_{\mathbf{x}} \left\|\mathbf{A}(\boldsymbol{\alpha}^{k+1})\mathbf{x}-\mathbf{y}(\boldsymbol{\alpha}^{k+1})\right\|$
  $k \leftarrow k+1$
end while
$\mathbf{x}_{Min-Min} \leftarrow \mathbf{x}^k$

Note that for the $\boldsymbol{\alpha}$ update in the alternating minimizations, a Quadratically constrained quadratic program (QCQP) needs to be solved [38]. The advantage of this simple algorithm is that, the QCQP can be replaced with any other convex optimization and any choice of norm p, $1 \le p \le \infty$ can also be used. It is also possible to bound the perturbations by using multiple constraints of the form $\|\mathbf{W}_i\boldsymbol{\alpha}\| \le \epsilon_i$, $i=1,\ldots,P$, as well.

This alternating minimizations approach is widely used for optimizing a nonconvex function over two sets of variables in applications such as superresolution and image deblurring [39]. By Proposition 2.7.1 of [40], Algorithm 1 is guaranteed to converge globally to a stationary point of the problem.

B. Joint Optimization by Linearization

The SLS-BDU cost function can also be linearized around a given $(\mathbf{x},\boldsymbol{\alpha})$ for a small perturbation $[\boldsymbol{\Delta}\mathbf{x},\ \boldsymbol{\Delta}\boldsymbol{\alpha}]$ by ignoring second order terms as in [41]:

$$\begin{aligned}
&\|(\mathbf{A}(\boldsymbol{\alpha}+\boldsymbol{\Delta}\boldsymbol{\alpha}))(\mathbf{x}+\boldsymbol{\Delta}\mathbf{x})-\mathbf{y}\| \\
&\qquad \approx \|\mathbf{A}(\boldsymbol{\alpha})\mathbf{x}-\mathbf{y}+\mathbf{U}(\mathbf{x})\boldsymbol{\Delta}\boldsymbol{\alpha}+\mathbf{A}(\boldsymbol{\alpha})\boldsymbol{\Delta}\mathbf{x}\|. \qquad (39)
\end{aligned}$$

Then, the solution to the following optimization provides an update on the estimated $\mathbf{x}$ and $\boldsymbol{\alpha}$:

$$\min_{\substack{\Delta\mathbf{x},\Delta\boldsymbol{\alpha} \\ \|\mathbf{W}(\boldsymbol{\alpha}+\Delta\boldsymbol{\alpha})\|\leq\epsilon}} \left\| [\mathbf{A}(\boldsymbol{\alpha}) \quad \mathbf{U}(\mathbf{x})] \begin{bmatrix} \Delta\mathbf{x} \\ \Delta\boldsymbol{\alpha} \end{bmatrix} + (\mathbf{A}(\boldsymbol{\alpha})\mathbf{x}-\mathbf{y}) \right\|. \tag{40}$$

The following Newton iterations can be used to yield an estimate for the solution to the SLS-BDU problem in (10):

Algorithm 2: Newton's Method

$\mathbf{x}^0 \leftarrow \mathbf{A}^\dagger\mathbf{y}$, $\boldsymbol{\alpha}^0 \leftarrow 0$, $k \leftarrow 0$
while $\left\|\mathbf{x}^k - \mathbf{x}^{k-1}\right\| > \epsilon$ **do**
  Solve (40) for $\Delta\mathbf{x}$ and $\Delta\boldsymbol{\alpha}$ by using QCQP
  $\mathbf{x}^{k+1} \leftarrow \mathbf{x}^k + \Delta\mathbf{x}$
  $\boldsymbol{\alpha}^{k+1} \leftarrow \boldsymbol{\alpha}^k + \Delta\boldsymbol{\alpha}$
  $k \leftarrow k+1$
end while
$\mathbf{x}_{Min-Min} \leftarrow \mathbf{x}^k$

This algorithm is a hybrid of Gauss–Newton method and sequential quadratic programming (SQP). Assuming $\mathbf{A}(\boldsymbol{\alpha})$ is nonsingular for $\|\mathbf{W}\boldsymbol{\alpha}\| \leq \rho$, it converges locally quadratically to a stationary point by Theorem 12.4.1 [42].

C. Fixed Point Iteration Using the Fréchet Derivatives

By using Theorem 5.1, the gradient of the Lagrangian of problem (18) can be written as

$$\frac{1}{2}\nabla\mathcal{L}(\boldsymbol{\alpha},\lambda) = \mathbf{y}(\boldsymbol{\alpha})^T\mathbf{P}_{\boldsymbol{\alpha}}^{\perp}(\mathbf{y}_i - \mathbf{A}_i\mathbf{A}(\boldsymbol{\alpha})^\dagger\mathbf{y}(\boldsymbol{\alpha})) + \lambda\boldsymbol{\alpha}. \tag{41}$$

By solving λ under the constraint of $\|\mathbf{W}\boldsymbol{\alpha}\|_2 = \rho$, we obtain

$$\rho\mathbf{f}(\boldsymbol{\alpha}) = \boldsymbol{\alpha}\,\|\mathbf{W}\mathbf{f}(\boldsymbol{\alpha})\|_2 \tag{42}$$

where $\mathbf{f}_i(\boldsymbol{\alpha}) \triangleq \mathbf{y}(\boldsymbol{\alpha})^T\mathbf{P}_{\boldsymbol{\alpha}}^{\perp}(\mathbf{A}_i\mathbf{A}(\boldsymbol{\alpha})^\dagger\mathbf{y}(\boldsymbol{\alpha}) - \mathbf{y}_i)$, $i = 1,\ldots,p$. As given below, a fixed point iteration to solve (42) can be used to find the SLS-BDU estimate. Note that although this fixed point iteration converges faster, it can only be used for the Euclidean norm.

Algorithm 3: Fixed Point Iteration

$\boldsymbol{\alpha}^0 \leftarrow 0$, $k \leftarrow 0$
while $\left\|\boldsymbol{\alpha}^k - \boldsymbol{\alpha}^{k-1}\right\| > \epsilon$ **do**
  $\boldsymbol{\alpha}^{k+1} \leftarrow \rho\mathbf{f}(\boldsymbol{\alpha}^k)/\left\|\mathbf{W}\mathbf{f}(\boldsymbol{\alpha}^k)\right\|_2$
  $k \leftarrow k+1$
end while
$\boldsymbol{\alpha}^* \leftarrow \boldsymbol{\alpha}^k$, $\mathbf{x}_{Min-Min} \leftarrow \mathbf{A}(\alpha^*)^\dagger\mathbf{y}(\alpha^*)$

In our numerical experiments, we observed that this fixed point iteration has superior convergence. In the appendix we give a proof for the local Lipschitz continuity of $\nabla_{\boldsymbol{\alpha}}\left\|\mathbf{P}_{\boldsymbol{\alpha}}^{\perp}\mathbf{y}(\boldsymbol{\alpha})\right\|_2^2$ provided that there exists no singularity or consistency inside the constraint set $\|\mathbf{W}\boldsymbol{\alpha}\| \leq \rho$. Then, by Proposition A.26 of [40], Algorithm 3 convergences to a stationary point with geometric rate of convergence.

Remark 4: The convergence criterion of Algorithm 3 makes the need of such a norm constraint clearer. Note that the Lipschitz continuity would fail near a singularity.

VII. Choosing the Bound Parameter Based on the Gradient Norm

The SLS-BDU technique requires a bound on $\boldsymbol{\alpha}$. Such a bound may be readily available when uncertainty bounds on the matrix elements are known. However, for those cases when there exists no such descriptive information on the bound on $\boldsymbol{\alpha}$, it is desirable to have a robust scheme to determine the bound which yields a good tradeoff between $\|\mathbf{A}(\alpha^*) - \mathbf{A}_0\|$ and $\left\|\mathbf{A}^\dagger(\boldsymbol{\alpha})\right\|_F$. In this section we provide such a criterion based on the gradient norm. Inspecting the example of Section IV in Fig. 1, it can be concluded that an abrupt increase in the gradient norm of the lower bound results in estimates which are highly sensitive to $\boldsymbol{\alpha}$, loosing robustness. Hence, we investigated the following simple strategy in the choice of the bound ρ. As given in Algorithm 4, we start with $\rho = 0$ and increase it with small steps $\Delta\rho$ till the gradient norm $\|\mathbf{f}(\boldsymbol{\alpha})\|_2$ starts to increase. In a wide range of experiments we observed that this simple scheme provides highly effective results. In the next section, we illustrate its performance over a range of simulations conducted at different noise levels.

Algorithm 4: Automated Selection of Bound Parameter

$\rho^0 \leftarrow 0$, $k \leftarrow 0$, $\mathbf{f}(\boldsymbol{\alpha})^0 \leftarrow 0$, $\mathbf{f}(\boldsymbol{\alpha})^{-1} \leftarrow 1$
while $\left\|\mathbf{f}(\boldsymbol{\alpha})^k\right\|_2 < \left\|\mathbf{f}(\boldsymbol{\alpha})^{k-1}\right\|_2$ **do**
  $(\mathbf{x}^k, \mathbf{f}(\boldsymbol{\alpha})^k) \leftarrow$ **Algorithm 3**$(\rho^k, \mathbf{A}, \mathbf{y})$
  $\rho \leftarrow \rho + \Delta\rho$
  $k \leftarrow k+1$
end while
$\hat{\mathbf{x}} \leftarrow \mathbf{x}^k$

VIII. Applications and Simulations

A. Verification of Theorem 3.2

First, we verify the accuracy of our result in (23). A Toeplitz matrix $\mathbf{A}_0$ with smallest singular value σ_0 is generated and perturbed with an unknown $\boldsymbol{\theta}$ to obtain the measured matrix $\mathbf{A}$ as in (9). Based on the observation $\mathbf{y} = \mathbf{A}_0\mathbf{x}_0 + \mathbf{w}$ and $\mathbf{A}$ only, $\mathbf{x}_0$ is estimated using SLS-BDU and STLS for a range of σ_0 and $\mathrm{SNR} = \|x\|_2^2/n\sigma_w^2$ values while θ is fixed and $\|\theta\|_2 = 0.5$. The theorem specifies a region in (SNR, σ_0) plane where the MSE of SLS-BDU is smaller than STLS asymptotically as shown in Fig. 4(a). For comparison, the empirical probability of $\|\mathbf{x}_{\text{SLS-BDU}} - \mathbf{x}_0\| < \|\mathbf{x}_{\text{STLS}} - \mathbf{x}_0\|$ in 100 trials is shown in Fig. 4(b). Although the theoretical region is conservative, it clearly indicates the ill-conditioned small σ_0 and low SNR region where SLS-BDU outperforms with probability approaching one.

Next we discuss three signal processing applications of the SLS-BDU approach to illustrate its effectiveness in ill-conditioned problems.

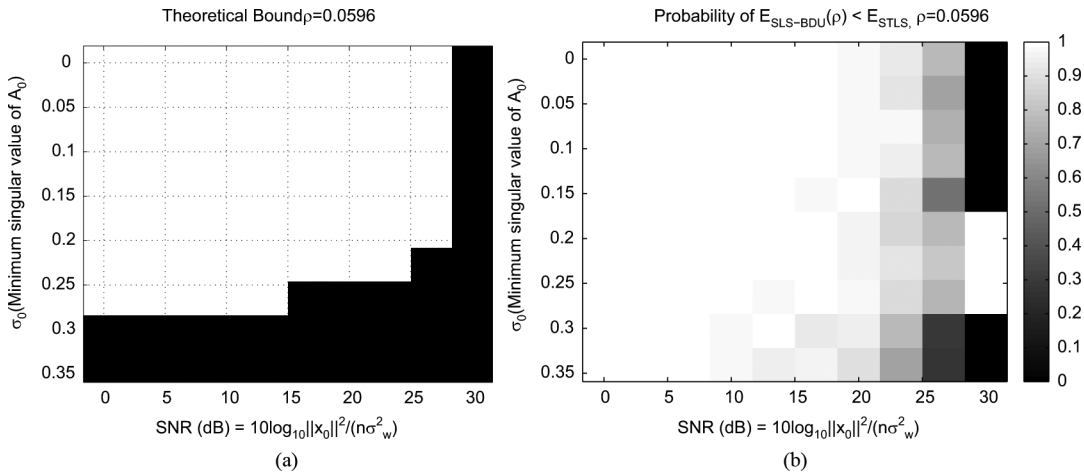

Fig. 4. Comparison of SLS-BDU and STLS: (a) Theoretical bound and (b) empirical probability.

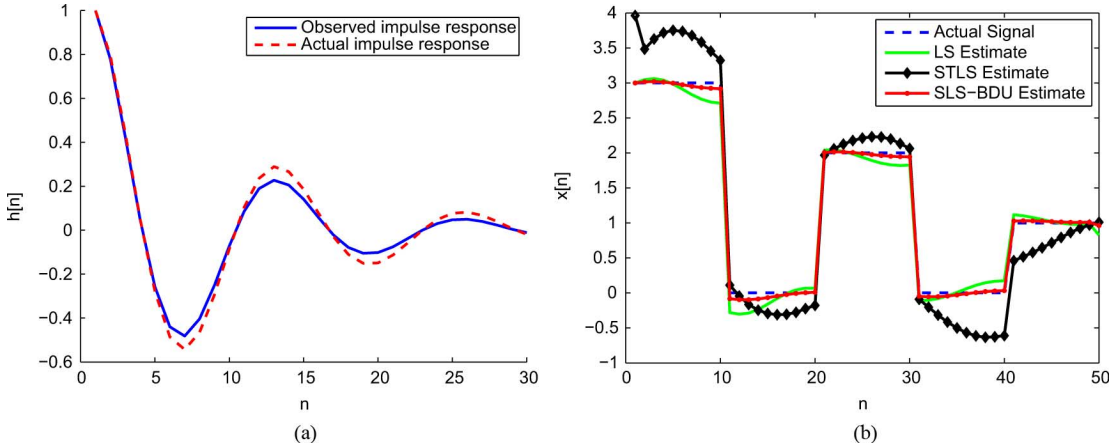

Fig. 5. Deconvolution under Impulse Response Uncertainties (a) True (dashed) and observed (solid) impulse responses (b) Actual and restored input signals $\mathbf{x}$ are shown in dashed and solid lines respectively.

B. Deconvolution Under Impulse Response Uncertainties

Suppose that the observed signal is the output of an LTI system with impulse response $h[n]$:

$$y[n] = \sum_{k=0}^{L-1} x[n-k]h[k] + w[n] \tag{43}$$

where and $w[n]$ is white Gaussian noise and

$$h[n] = \sum_{i=1}^{p} (a_i + \delta a_i) e^{-(b_i+\delta b_i)n} \cos(w_i n + \phi_i) \tag{44}$$

with bounded data uncertainties on coefficients $\mid \delta a_i \mid < \epsilon_{a_i}$ and damping terms $\mid \delta b_i \mid < \epsilon_{b_i}$, $i = 1, \ldots, p$. We want to recover $x[n]$ under this structured uncertainty on the impulse response $h[n]$. The uncertainties in b_i's can be linearized by a first order approximation, $e^{-(b_i+\delta b_i)n} \approx e^{-b_i n}(1 - \delta b_i n)$, to obtain the following:

$$\mathbf{y} = \left(\mathbf{H} + \sum_{i=1}^{p} \alpha_i \mathbf{H}_i \right) \mathbf{x} + \mathbf{w},$$

with the constraint $\|\mathbf{W}\boldsymbol{\alpha}\|_\infty \leq \epsilon$. Here, $\mathbf{H}$ and $\mathbf{H}_i$ are Toeplitz structured matrices which perform convolution operation with the terms in the summation of (44) and α_i's stand for the unknown perturbations $\delta a_i, \delta b_i$.

The impulse response $h[n]$ with uncertainties is shown in Fig. 5(a). As shown in Fig. 5(b), the SLS-BDU estimate closely approximates the actual input signal. Table I provides comparison results between the SLS-BDU and least squares estimates for both the input signal and the impulse response estimates at two different uncertainty levels. As expected based on Theorem 3.2, the tabulated results show that the SLS-BDU technique provides significantly better estimates for both the input and the impulse response. Note that STLS estimate is unsatisfactory since the perturbations are not bounded and linear approximation is not valid for large perturbations.

C. Frequency Estimation of Multiple Sinusoids

Consider the case where parameters of two complex sinusoids which are close in frequency need to be estimated with frequencies $f_1 = 0.12$ Hz and $f_2 = 0.10$ Hz in white noise w_n:

$$x(n) = \exp(2\pi j f_1 n) + \exp(2\pi j f_2 n) + w_n, \quad n = 0, 1, \ldots, 25. \tag{45}$$

TABLE I
$\mathbf{x}_{\text{true}}$, $\mathbf{x}_{\text{LS}}$ AND $\mathbf{x}_{\text{SLS-BDU}}$ CORRESPOND TO ACTUAL SIGNAL AND ESTIMATES, $\mathbf{H}_{\text{true}}$, $\mathbf{H}$, $\mathbf{H}_{\text{SLS-BDU}}$ CORRESPOND TO ACTUAL, NOMINAL AND CORRECTED MATRICES, RESPECTIVELY

ϵ_b/b_{true}	0.2	0.6
$\|\mathbf{x}_{true}-\mathbf{x}_{LS}\|\,/\,\|\mathbf{x}_{true}\|$	0.0820	0.2123
$\|\mathbf{x}_{true}-\mathbf{x}_{SLS-BDU}\|\,/\,\|\mathbf{x}_{true}\|$	0.0274	0.1279
$\|\mathbf{H}_{true}-\mathbf{H}_{SLS-BDU}\|_{\text{F}}\,/\,\|\mathbf{H}_{true}\|_{\text{F}}$	0.1072	0.2589
$\|\mathbf{H}_{true}-\mathbf{H}_{SLS-BDU}\|_{\text{F}}\,/\,\|\mathbf{H}_{true}\|_{\text{F}}$	0.0655	0.1284

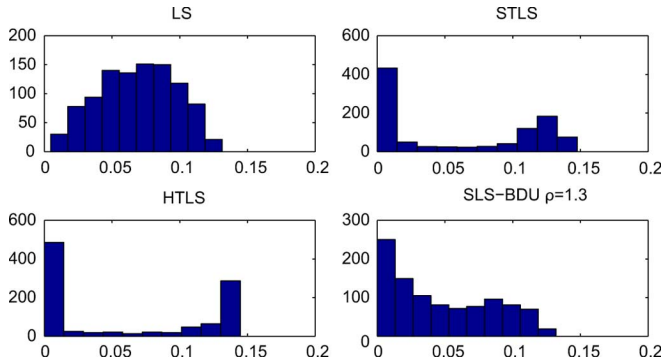

Fig. 6. Histogram of frequency estimation errors for LS, STLS, HTLS, and SLS-BDU. Note that the distribution of the estimation error is heavy-tailed for STLS and HTLS.

The following Linear prediction equations can be solved to estimate the parameters of L sinusoids [24]:

$$\begin{bmatrix} x_1 & x_2 & \cdots & x_L \\ x_2 & x_3 & \cdots & x_{L+1} \\ x_3 & \iddots & & \vdots \\ \vdots & & \iddots & x_{N-2} \\ x_{N-L} & \cdots & \cdots & x_{N-1} \end{bmatrix} \mathbf{z} = \begin{bmatrix} x_{L+1} \\ x_{L+2} \\ x_{L+3} \\ \vdots \\ x_N \end{bmatrix}. \quad (46)$$

The frequency estimation error defined by $\sqrt{(\hat{f}_1-f_1)^2+(\hat{f}_2-f_2)^2}$ is evaluated for the estimates with SLS-BDU with parameters $\rho=1.3$ and $\mathbf{W}=\mathbf{I}$ in 1000 independent trials at various SNR values. In Table II, a comparison of LS, TLS, STLS, HTLS [43], and SLS-BDU is given. Histograms of estimation errors are plotted in Fig. 6. As expected based on Theorem 3.2, the tabulated results and histograms reveal that the SLS-BDU estimator not only provides more accurate estimates on the average but it is also significantly more robust than the STLS estimator. As indicated by the obtained histograms, the errors of SLS-BDU estimates have higher concentration around zero, whereas STLS and HTLS estimates have heavy-tailed distributions.

D. System Identification

Consider the system identification setup depicted in Fig. 8. An input sequence u_0 is applied to the FIR filter $H(z)$ and the output y_0 is generated. Measurements of the input and the output contain noise w_i and w_o respectively. The identification of the filter $H(z)$ can be cast as the following regression problem [16]:

$$\mathbf{U}_0 h = y_0 \quad (47)$$

TABLE II
AVERAGE FREQUENCY ESTIMATION ERRORS FOR LS, TLS, STLS, HTLS AND SLS-BDU

SNR	4dB	7dB	10dB
LS	0.0347	0.0344	0.0346
TLS	0.0308	0.0295	0.0298
STLS	0.0297	0.0304	0.0321
HTLS	0.0311	0.0309	0.0275
SLS-BDU	0.0279	0.0249	0.0241

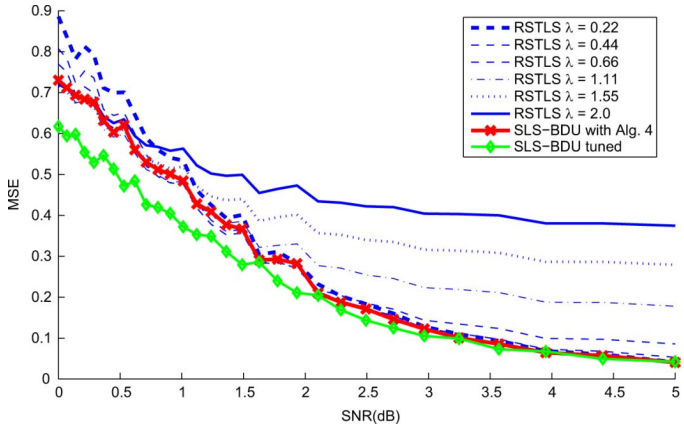

Fig. 7. MSE of Algorithm 4 and RSTLS solutions for a range of regularization parameters versus SNR.

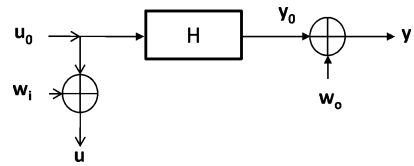

Fig. 8. System identification with noisy input u and noisy output y.

where $\mathbf{U}=\mathbf{U}_0+\mathbf{W}_i$ is the observed noisy Toeplitz matrix and $\mathbf{y}=\mathbf{y}_0+\mathbf{w}_o$ is the observed noisy output. The filter coefficients were set to $\mathbf{h}=[-0.3,-0.9,0.8]^T$, the training signal $\mathbf{u}_0$ was selected as a random sequence of ±1's and equal variance independent white noise was added to input and output. SLS-BDU estimates are generated with autonomously chosen bound ρ by using Algorithm 4. The MSE in 10 000 independent trials of the SLS-BDU estimator, and RSTLS for a range of regularization parameters are shown in Fig. 7. As seen from these results, the SLS-BDU estimator with autonomously chosen bound ρ provides lower MSE than the RSTLS estimates that are obtained with a range of regularization parameters. In this example, to illustrate the effectiveness of the criterion by which Algorithm 4 determines ρ, we included the performance of SLS-BDU estimates with hand tuned ρ as well. As seen from the obtained results, the autonomous choice provides performance results that are close to the hand tuned case.

The implementations of STLS and RSTLS used in numerical comparisons are [44], [45], respectively, and both available online. And for TLS and HTLS methods direct implementations of corresponding references are used.

IX. Conclusion

We considered linear regression problems with structured errors in all variables. A novel estimator, SLS-BDU is proposed in terms of a nonconvex optimization problem. The analysis of the MSE of the SLS-BDU estimator reveals the advantage over the alternative estimators. Three different methods are presented for iterative solution of the optimization problem. Among the three methods, the Fréchet gradient approach provides the fastest convergence. Furthermore, the gradient flow space enables us to study alternative approaches and be able to compare their performances. New theorems that characterize the gradient flow for practical cases of interest are proven. A simple but efficient criterion to select the optimization parameter based on the gradient norm is proposed. Extensive comparison results on the SLS-BDU estimator reveal the superior performance of the proposed technique in signal restoration, multiple frequency estimation and system identification applications. The automated selection of the optimization parameter adaptively regularizes the solution based on SNR and achieves improved MSE compared to the notable alternative RSTLS technique.

Appendix A
Singularity of the Fisher Information Matrix

It is known that for a singular Fisher information matrix, there exists no unbiased estimator with finite variance except under unusual circumstances [46]. In the following proof, we show that the information matrix is singular for the deterministic perturbation case when $p > m - n$.

Proof: The observation $\mathbf{y}$ is related to unknowns $\mathbf{x}$ and $\boldsymbol{\theta}$ as

$$\begin{aligned}\mathbf{y} &= \mathbf{A}_0\mathbf{x} + \sum_i \mathbf{y}_i\boldsymbol{\theta}_i + \mathbf{w} \\ &= \left(\mathbf{A} - \sum_i \mathbf{A}_i\boldsymbol{\theta}_i\right)\mathbf{x} + \sum_i \mathbf{y}_i\boldsymbol{\theta}_i + \mathbf{w}.\end{aligned}$$

Define $\mathbf{A}_{\boldsymbol{\theta}} \triangleq (\mathbf{A} - \sum_i \mathbf{A}_i\boldsymbol{\theta}_i)$ and $\mathbf{B} \triangleq [\mathbf{y}_1, \ldots, \mathbf{y}_p]$. Given that $\mathbf{w}$ is a zero-mean Gaussian random vector with covariance $\sigma^2\mathbf{I}$, the log-likelihood can be written as

$$\log p_{\boldsymbol{\theta}}(\mathbf{y}) = -\frac{m}{2}\log 2\pi - m\log\sigma - \frac{1}{2\sigma^2}\left\|\mathbf{y} - (\mathbf{A}_{\boldsymbol{\theta}}\mathbf{x} + \mathbf{B}\boldsymbol{\theta})\right\|_2^2.$$

Defining the vector of unknowns $\mathbf{z} \triangleq \begin{bmatrix}\mathbf{x} \\ \boldsymbol{\theta}\end{bmatrix}$ and $\mathbf{Q} \triangleq [\mathbf{A}_1\mathbf{x}, \ldots, \mathbf{A}_p\mathbf{x}]$, the gradient of the log-likelihood can be obtained as

$$\frac{\partial}{\partial \mathbf{z}} p_{\boldsymbol{\theta}}(\mathbf{y}) = \frac{-1}{\sigma^2}\begin{bmatrix}\mathbf{A}_{\boldsymbol{\theta}}^T(\mathbf{A}_{\boldsymbol{\theta}}\mathbf{x} + \mathbf{B}\boldsymbol{\theta} - \mathbf{y}) \\ (\mathbf{B} - \mathbf{Q})^T(\mathbf{A}_0\mathbf{x} - \mathbf{Q}\boldsymbol{\theta} + \mathbf{B}\boldsymbol{\theta} - \mathbf{y})\end{bmatrix}$$

and the corresponding Fisher information matrix can be expressed as

$$\begin{aligned}\mathbf{I}_{\mathbf{x},\boldsymbol{\theta}} &= E\left[\frac{\partial}{\partial \mathbf{z}} p_{\boldsymbol{\theta}}(\mathbf{y})\left(\frac{\partial}{\partial \mathbf{z}} p_{\boldsymbol{\theta}}(\mathbf{y})\right)^T\right] \\ &= \frac{1}{\sigma^4} E\begin{bmatrix}\mathbf{A}_{\boldsymbol{\theta}}^T\mathbf{w}\mathbf{w}^T\mathbf{A}_{\boldsymbol{\theta}} & \mathbf{A}_{\boldsymbol{\theta}}\mathbf{w}\mathbf{w}^T(\mathbf{B} - \mathbf{Q}) \\ (\mathbf{B} - \mathbf{Q})^T\mathbf{w}\mathbf{w}^T\mathbf{A}_{\boldsymbol{\theta}} & (\mathbf{B} - \mathbf{Q})^T\mathbf{w}\mathbf{w}^T(\mathbf{B} - \mathbf{Q})\end{bmatrix} \\ &= \frac{1}{\sigma^2}\begin{bmatrix}\mathbf{A}_{\boldsymbol{\theta}}^T\mathbf{A}_{\boldsymbol{\theta}} & \mathbf{A}_{\boldsymbol{\theta}}(\mathbf{B} - \mathbf{Q}) \\ (\mathbf{B} - \mathbf{Q})^T\mathbf{A}_{\boldsymbol{\theta}} & (\mathbf{B} - \mathbf{Q})^T(\mathbf{B} - \mathbf{Q})\end{bmatrix}.\end{aligned}$$

Next, we use the following fact: Assume $\mathbf{A}_{11}$ is invertible, the block matrix

$$\begin{bmatrix}\mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22}\end{bmatrix}$$

is invertible if and only if $\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$ is invertible. Since we assumed $\mathbf{A}_0^T\mathbf{A}_0 = \mathbf{A}_{\boldsymbol{\theta}}^T\mathbf{A}_{\boldsymbol{\theta}}$ is invertible, $\mathbf{I}_{\mathbf{x},\boldsymbol{\theta}}$ is invertible if and only if

$$\det\Big[(\mathbf{B} - \mathbf{Q})^T(\mathbf{B} - \mathbf{Q}) - (\mathbf{B} - \mathbf{Q})^T\mathbf{A}_{\boldsymbol{\theta}}(\mathbf{A}_{\boldsymbol{\theta}}^T\mathbf{A}_{\boldsymbol{\theta}})^{-1} \times \mathbf{A}_{\boldsymbol{\theta}}^T(\mathbf{B} - \mathbf{Q})\Big]$$

is nonzero. By using $\mathbf{P}_{\boldsymbol{\theta}}^{\perp} \triangleq \mathbf{I} - \mathbf{A}_{\boldsymbol{\theta}}(\mathbf{A}_{\boldsymbol{\theta}}^T\mathbf{A}_{\boldsymbol{\theta}})^{-1}\mathbf{A}_{\boldsymbol{\theta}}^T$, this condition can be simplified to

$$\det\Big[(\mathbf{B} - \mathbf{Q})^T\mathbf{P}_{\boldsymbol{\theta}}^{\perp}(\mathbf{B} - \mathbf{Q})\Big] \neq 0.$$

Therefore, $\mathbf{I}_{\mathbf{x},\boldsymbol{\theta}}$ is invertible if and only if $\mathbf{P}_{\boldsymbol{\theta}}^{\perp}(\mathbf{B} - \mathbf{Q}) \in \mathbb{R}^{m\times p}$ is full column rank. Since $\mathrm{Rank}(\mathbf{P}_{\boldsymbol{\theta}}^{\perp}) = m - \mathrm{Rank}(\mathbf{A}_{\boldsymbol{\theta}})$ it is easy to show that

$$\begin{aligned}\mathrm{Rank}(\mathbf{P}_{\boldsymbol{\theta}}^{\perp}(\mathbf{B} - \mathbf{Q})) &\leq \min\Big\{\mathrm{Rank}(\mathbf{P}_{\boldsymbol{\theta}}^{\perp}), \mathrm{Rank}(\mathbf{B} - \mathbf{Q})\Big\} \\ &\leq m - n\end{aligned}$$

which implies that $\mathbf{I}_{\mathbf{x},\boldsymbol{\theta}}$ is not invertible for $p > m - n$ and hence there exists no unbiased estimator with finite variance. ∎

Appendix B
Proof of Theorem 3.2

Proof: First, for any $\beta \in \mathbb{R}_p$, the following bounds can be obtained:

$$\begin{aligned}\left\|\sum_i \mathbf{A}_i\beta_i\right\|_2^2 &\leq \left(\sum_i \|\mathbf{A}_i\|_2 |\beta_i|\right)^2 \\ &\leq \max_i \|\mathbf{A}_i\|_2^2 \|\beta\|_1^2 \leq p \max_i \|\mathbf{A}_i\|_2^2 \|\beta\|_2^2. \quad (48)\end{aligned}$$

And for nonoverlapping structures, i.e., $\mathbf{A}_i \odot \mathbf{A}_j = 0$, $\forall i \neq j$:

$$\begin{aligned}\left\|\sum_i \mathbf{A}_i\beta_i\right\|_2^2 \leq \left\|\sum_i \mathbf{A}_i\beta_i\right\|_F^2 &= \sum_i \|\mathbf{A}_i\|_F^2 \beta_i^2 \\ &\leq \max_i \|\mathbf{A}_i\|_F^2 \|\beta\|_2^2. \quad (49)\end{aligned}$$

In particular Toeplitz and Hankel structures are nonoverlapping and both have $\max_i \|\mathbf{A}_i\|_F^2 = n$. Next, we use the bound $\|\alpha\|_2 \leq \rho$ of SLS-BDU and Weyl's theorem [47] and get

$$\frac{1}{\sigma_{\boldsymbol{\alpha}^*}^2} \leq \frac{1}{\left(\sigma_A - \left\|\sum_i \mathbf{A}_i\alpha_i\right\|_2\right)_+^2} \leq \frac{1}{(\sigma_A - \rho S)_+^2}. \quad (50)$$

Also, observe that

$$\|\mathbf{A}(\boldsymbol{\alpha}^*) - \mathbf{A}_0)\|_2^2 = \left\|\sum_i \mathbf{A}_i(\alpha_i + \theta_i)\right\|_2^2 \leq (\rho + \|\theta\|)^2 S^2. \quad (51)$$

Using (51) and (50) in Theorem 3.1, another MSE bound of SLS-BDU can be stated as follows:

$$E[\|\hat{\mathbf{x}}-\mathbf{x}\|_2^2] \leq \frac{(\rho+\|\theta\|)^2 S^2 \|\mathbf{x}_0\|_2^2 + n\sigma_w^2}{(\sigma_A-\rho S)_+^2}. \quad (52)$$

Since STLS is an ML estimator it is asymptotically unbiased and the asymptotic MSE is equivalent to the second part of (21) when $\mathbf{A}(\boldsymbol{\alpha}^*)$ is replaced by $\mathbf{A}_0$:

$$E[\|\mathbf{x}_{\mathrm{STLS}}-\mathbf{x}\|_2^2] = \sigma_w^2 \|\mathbf{A}_0\|_F^2 = \sum_{i=1}^{n} \frac{\sigma_{\mathbf{w}}^2}{\sigma_{\mathbf{A}_0,i}^2} > \frac{\sigma_w^2}{\sigma_0^2}. \quad (53)$$

Therefore, when (23) is satisfied, we get

$$E[\|\mathbf{x}_{\mathrm{STLS}}-\mathbf{x}\|_2^2] > E[\left\|\mathbf{x}_{\text{SLS-BDU}(\rho)}-\mathbf{x}\right\|_2^2],$$

asymptotically. ∎

Appendix C
Local Lipschitz Continuity

Proposition C.1: Assume $\mathbf{A}(\boldsymbol{\alpha})$ is of full column rank and $\left\|\mathbf{P}_{\boldsymbol{\alpha}}^{\perp}\mathbf{y}\right\| \neq 0$ for $\|\mathbf{W}\boldsymbol{\alpha}\| \leq \rho$, then $\mathbf{f}(\boldsymbol{\alpha}) \triangleq \nabla_{\boldsymbol{\alpha}} \left\|\mathbf{P}_{\boldsymbol{\alpha}}^{\perp}\mathbf{y}\right\|_2^2$ is locally Lipschitz continuous.

Proof: Let $\alpha, \beta \in \mathbb{R}^p$ be any two vectors satisfying $\mathrm{Rank}(\mathbf{A}(\boldsymbol{\alpha})) = \mathrm{Rank}(\mathbf{A}(\boldsymbol{\beta}))$. And let $\sigma_{\min}$ be the minimum singular value of $\mathbf{A}(\boldsymbol{\alpha})$ in $\|\mathbf{W}\boldsymbol{\alpha}\| \leq \rho$. Using Lemma 5.1, we get

$$\|\mathbf{f}(\boldsymbol{\alpha})-\mathbf{f}(\boldsymbol{\beta})\|_2^2 \quad (54)$$

$$\leq 4\sum_i (\mathbf{y}^T\mathbf{y})^2 \left\|\mathbf{P}_{\boldsymbol{\alpha}}^{\perp}\mathbf{A}_i\mathbf{A}(\boldsymbol{\alpha})^{\dagger} - \mathbf{P}_{\boldsymbol{\beta}}^{\perp}\mathbf{A}_i\mathbf{A}(\boldsymbol{\beta})^{\dagger}\right\|_2^2 \quad (55)$$

$$= 4\|y\|_2^4 \sum_i \left\| \frac{1}{2}\Big((\mathbf{P}_{\boldsymbol{\alpha}}^{\perp}-\mathbf{P}_{\boldsymbol{\beta}}^{\perp})\mathbf{A}_i(\mathbf{A}(\boldsymbol{\alpha})^{\dagger}+\mathbf{A}(\boldsymbol{\beta})^{\dagger}) + (\mathbf{P}_{\boldsymbol{\alpha}}^{\perp}+\mathbf{P}_{\boldsymbol{\beta}}^{\perp})\mathbf{A}_i(\mathbf{A}(\boldsymbol{\alpha})^{\dagger}-\mathbf{A}(\boldsymbol{\beta})^{\dagger})\Big)\right\|_2^2. \quad (56)$$

Now let $M^+ \triangleq \max(\|\mathbf{A}(\boldsymbol{\alpha})\|_2, \left\|\mathbf{A}(\boldsymbol{\beta})^{\dagger}\right\|_2)$ and $M^- \triangleq \min(\|\mathbf{A}(\boldsymbol{\alpha})\|_2, \left\|\mathbf{A}(\boldsymbol{\beta})^{\dagger}\right\|_2)$. In [48], the following are derived for pseudoinverses and projectors having the same rank:

$$\left\|\mathbf{A}(\boldsymbol{\alpha})^{\dagger}-\mathbf{A}(\boldsymbol{\beta})^{\dagger}\right\| \leq 3M^+ \|\mathbf{A}(\boldsymbol{\alpha})-\mathbf{A}(\boldsymbol{\beta})\| \quad (57)$$

$$\left\|\mathbf{P}_{\boldsymbol{\alpha}}^{\perp}-\mathbf{P}_{\boldsymbol{\beta}}^{\perp}\right\| \leq M^- \|\mathbf{A}(\boldsymbol{\alpha})-\mathbf{A}(\boldsymbol{\beta})\|. \quad (58)$$

Using the above bounds with (47) yields that $\nabla_{\boldsymbol{\alpha}} \left\|\mathbf{P}_{\boldsymbol{\alpha}}^{\perp}\mathbf{y}\right\|_2^2$ is Lipschitz continuous with constant

$$\|y\|_2^2 S\left(3M^+\left\|\mathbf{P}_{\boldsymbol{\alpha}}^{\perp}+\mathbf{P}_{\boldsymbol{\beta}}^{\perp}\right\| + M^-\left\|\mathbf{A}(\boldsymbol{\alpha})^{\dagger}+\mathbf{A}(\boldsymbol{\beta})^{\dagger}\right\|_2\right) \leq 2\|y\|_2^2 S\left(\frac{1}{\sigma_{\min}^2}+\frac{3}{\sigma_{\min}}\right). \quad (59)$$

∎

Using the above result, we will next prove that the Algorithm 3 converges geometrically provided that ρ is sufficiently small:

Theorem C.2: If ρ satisfies

$$2\rho\kappa\|y\|_2^2 S \leq \frac{\sigma_{\min}^2}{1+3\sigma_{\min}} \min_{\|\mathbf{W}\boldsymbol{\alpha}\leq\rho\|} \|\mathbf{W}\mathbf{f}(\boldsymbol{\alpha})\|_2 \quad (60)$$

where $\kappa \triangleq \|\mathbf{W}\|_2 \left\|\mathbf{W}^{\dagger}\right\|_2$ is the condition number of $\mathbf{W}$, then (42) is a contraction mapping and Algorithm 3 converges to a minimum of (26) with a geometric rate.

Proof: Define the contraction mapping of Algorithm 3 as $T(\boldsymbol{\alpha}) \triangleq \rho\mathbf{f}(\boldsymbol{\alpha})/\|\mathbf{f}(\boldsymbol{\alpha})\|_2$. Then, we have

$$\|T(\boldsymbol{\alpha})-T(\boldsymbol{\beta})\| = \left\| \frac{\|\mathbf{W}\mathbf{f}(\boldsymbol{\beta})\|\,\mathbf{f}(\boldsymbol{\alpha}) - \|\mathbf{W}\mathbf{f}(\boldsymbol{\alpha})\|\,\mathbf{f}(\boldsymbol{\beta})}{\|\mathbf{W}\mathbf{f}(\boldsymbol{\alpha})\|\,\|\mathbf{W}\mathbf{f}(\boldsymbol{\beta})\|} \right\| \quad (61)$$

$$= \frac{\|(\mathbf{f}(\boldsymbol{\alpha})-\mathbf{f}(\boldsymbol{\beta}))\,\|\mathbf{W}\mathbf{f}(\boldsymbol{\beta})\| + \mathbf{f}(\boldsymbol{\beta})(\|\mathbf{W}\mathbf{f}(\boldsymbol{\beta})\|-\|\mathbf{W}\mathbf{f}(\boldsymbol{\alpha})\|)\|}{\|\mathbf{W}\mathbf{f}(\boldsymbol{\alpha})\|\,\|\mathbf{W}\mathbf{f}(\boldsymbol{\beta})\|}$$

$$\leq \frac{\|\mathbf{W}\|\,\|\mathbf{f}(\boldsymbol{\beta})\|\,\|\mathbf{f}(\boldsymbol{\alpha})-\mathbf{f}(\boldsymbol{\beta})\|}{\|\mathbf{W}\mathbf{f}(\boldsymbol{\alpha})\|\,\|\mathbf{W}\mathbf{f}(\boldsymbol{\beta})\|} \leq \frac{\kappa\,\|\mathbf{f}(\boldsymbol{\alpha})-\mathbf{f}(\boldsymbol{\beta})\|}{\|\mathbf{W}\mathbf{f}(\boldsymbol{\alpha})\|}. \quad (62)$$

In Proposition A.26 of [40], it is shown that geometric convergence is assured when $\|T(\boldsymbol{\alpha})-T(\boldsymbol{\beta})\| \leq \gamma\|\boldsymbol{\alpha}-\boldsymbol{\beta}\|$ with $\gamma < 1$. Then, using (59) in (62), we arrive at (60) which satisfies the specified condition.

Acknowledgment

The authors would like to thank B. Oguz for his invaluable insight on the initial phase of our work.

References

[1] N. Wiener, *The Extrapolation, Interpolation and Smoothing of Stationary Time Series*. New York: Wiley, 1949.
[2] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[3] G. Golub and F. V. Loan, "An analysis of the total least squares problem," *SIAM J. Numer. Anal.*, vol. 17, no. 6, pp. 883–893, Dec. 1980.
[4] A. Wiesel, Y. Eldar, and A. Yeredor, "Linear regression with Gaussian model uncertainty: Algorithms and bounds," *IEEE Trans. Signal Process.*, vol. 56, no. 6, pp. 2194–2205, Jun. 2008.
[5] B. D. Moor, "Total least squares for affinely structured matrices and the noisy realization problem," *IEEE Trans. Signal Process.*, vol. 42, no. 11, pp. 3104–3113, Nov. 1994.
[6] A. Pruessner and D. O'Leary, "Blind deconvolution using a regularized structured total least norm algorithm," *SIAM J. Mat. Anal. Appl.*, vol. 24, no. 4, pp. 1018–1037, 2002.
[7] P. Lemmerling, "Structured total least squares: Analysis, algorithms and applications," Ph.D. dissertation, Katholieke Universiteit Leuven, Leuven, Belgium, 1999.
[8] Y. Eldar, A. Ben-Tal, and A. N. A. Beck, "Robust mean-squared error estimation in the presence of model uncertainties," *IEEE Trans. Signal Process.*, vol. 53, no. 1, pp. 168–181, Jan. 2005.
[9] A. Beck, Y. Eldar, and A. Ben-Tal, "Mean-squared error estimation for linear systems with block circulant uncertainty," *SIAM J. Mater. Anal. Appl.*, vol. 29, no. 3, pp. 712–730, 2007.
[10] L. El-Ghaoui and H. Lebret, "Robust solutions to least-squares problems with uncertain data," *SIAM J. Mater. Anal. Appl.*, vol. 18, no. 4, pp. 1035–1064, 1997.
[11] F. Alizadeh and D. Goldfarb, "Second-order cone programming," *Math. Programm.*, vol. 95, no. 1, pp. 1436–4646, Jan. 2003.
[12] A. N. Tikhonov and V. Y. Arsenin, *Solution of Ill-Posed Problems*. Washington, DC: Winston and Wiley, 1977.
[13] A. Sayed and S. Chandrasekaran, "Parameter estimation with multiple sources and levels of uncertainties," *IEEE Trans. Signal Process.*, vol. 48, no. 3, pp. 680–692, Mar. 2000.
[14] M. Guillaud, "Transmission and channel modeling techniques for multiple-antenna communication systems," Ph.D. dissertation, TELECOM ParisTech, Paris, France, 2005.
[15] R. DeGroat and E. Dowling, "The data least squares problem and channel equalization," *IEEE Trans. Signal Process.*, vol. 42, no. 1, pp. 407–411, Jan. 1993.

[16] I. Markovsky, J. C. Willems, and S. V. Huffel, "Application of structured total least squares for system identification and model reduction," *IEEE Trans. Autom. Control*, vol. 50, no. 10, pp. 1490–1500, Oct. 2005.
[17] H. Chen, S. V. Huel, and D. V. Ormondt, "Application of the Structured Total Least Norm technique in spectral estimation," in *Proc. 8th Eur. Signal Process. Conf.*, Trieste, Italy, 1996, pp. 706–709.
[18] J. Severson, "Modeling and frequency tracking of marine mammal whistle calls," Ph.D. dissertation, Massachusetts Inst. of Technology and Woods Hole Oceanographic Institution, Woods Hole, MA, 2009.
[19] G. Golub and V. Pereyra, "The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate," *SIAM J. Numer. Anal.*, vol. 10, no. 2, pp. 413–432, Apr. 1973.
[20] S. V. Huffel and J. Vandewalle, "The total least squares problem: Computational aspects and analysis," in *Frontiers in Applied Mathematics*. Philadelphia, PA: SIAM, 1991.
[21] I. Markovsky and S. V. Huffel, "Overview of total least-squares methods," *Signal Process.*, vol. 87, no. 10, pp. 2283–2302, Oct. 2007.
[22] A. Beck, A. Ben-Tal, and C. Kanzow, "A fast method for finding the global solution of the regularized structured total least squares problem for image deblurring," *SIAM J. Mater. Anal. Appl.*, vol. 30, no. 1, pp. 419–443, Feb. 2008.
[23] V. Z. Mesarovic, N. P. Galatsanos, and A. K. Katsaggelos, "Regularized constrained total least squares image restoration," *IEEE Trans. Image Process.*, vol. 4, no. 8, pp. 1096–1109, 1995.
[24] T. Abatzoglou, J. Mendel, and G. Harada, "The constrained total least squares technique and its application to harmonic superresolution," *IEEE Trans. Signal Process.*, vol. 39, no. 5, pp. 1070–1087, May 1991.
[25] M. Pilanci, O. Arikan, B. Oguz, and M. Pinar, "Structured least squares with bounded data uncertainties," in *IEEE Int. Conf. Acoust. Speech, Signal Processing (ICASSP)*, Apr. 2009, pp. 3261–3264.
[26] M. Pilanci, O. Arikan, B. Oguz, and M. Pinar, "A novel technique for a linear system of equations applied to channel equalization," in *Proc. IEEE 17th Signal Processing Communications Applications Conf.*, Apr. 2009, pp. 948–951.
[27] S. Chandrasekaran, G. Golub, M. Gu, and A. Sayed, "An efficient algorithm for a bounded errors-in-variables model," *SIAM J. Mater. Anal. Appl.*, vol. 20, no. 4, pp. 839–859, Oct. 1999.
[28] S. Chandrasekaran, M. Gu, A. Sayed, and K. E. Schubert, "The degenerate bounded errors-in-variables model," *SIAM J. Mater. Anal. Appl.*, vol. 23, no. 1, pp. 138–166, Oct. 2001.
[29] K. E. Schubert, "A new look at robust estimation and identification," Ph.D. dissertation, Univ. of California, Santa Barbara, Santa Barbara, CA, 2003.
[30] A. Yeredor, "The extended least squares criterion: Minimization algorithms and applications," *IEEE Trans. Signal Process.*, vol. 49, no. 1, pp. 74–86, Jan. 2000.
[31] P. Lemmerling and B. D. Moor, "Misfit versus latency," *Automatica*, vol. 37, pp. 2057–2067, 2001.
[32] M. Hayes, *Statistical Digital Signal Processing and Modeling*. New York: Wiley, 1996.
[33] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 42, no. 1, pp. 80–86, Sep. 2000.
[34] Y. C. Eldar, "Rethinking biased estimation: Improving maximum likelihood and the Cramér–Rao bound," *Found. Trends Signal Process.*, vol. 1, no. 4, pp. 305–449, 2008.
[35] J. Dieudonn, *Foundations of Modern Analysis*. New York: Academic, 1960.
[36] R. Gray, "Toeplitz and circulant matrices: A review," Stanford Univ. Inform. Sys. Lab., Stanford, CA, 1977.
[37] G. H. Golub and C. F. V. Loan, *Matrix Computations*. Baltimore, MD: The Johns Hopkins Univ. Press, 1996.
[38] G. H. Golub and U. von Matt, "Quadratically constrained least squares and quadratic problems," *Numer. Math.*, vol. 59, no. 1, pp. 561–580, Dec. 1991.
[39] F. Sroubek, G. Cristobal, and J. Flusser, "Unified approach to superresolution and multichannel blind deconvolution," *IEEE Trans. Image Process.*, vol. 16, no. 9, pp. 2322–2332, Sep. 2007.
[40] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed. Singapore: Athena Scientific, 1995.
[41] J. Rosen, H. Park, and J. Glick, "Formulation and solution of structured total least norm problems for parameter estimation," *IEEE Trans. Signal Process.*, vol. 44, no. 10, pp. 2464–2474, Oct. 1996.
[42] R. Fletcher, *Practical Methods of Optimization*, 2nd ed. New York: Wiley, 1987.
[43] S. V. Huffel, H. Chen, C. Decanniere, and P. V. Hecke, "Algorithm for time-domain nmr data fitting based on total least squares," *J. Magn. Reson.*, vol. 110, no. 2, pp. 228–237, 1994.
[44] I. Markovsky, S. V. Huffel, and R. Pintelon, "Software for structured total least squares estimation: User's guide," Electr. Eng. Dept., K. U. Leuven, Leuven, Belgium, Tech. Rep. 03–136, 2003.
[45] D. P. O'Leary, *Scientific Computing With Case Studies*. Philadelphia, PA: SIAM, 2009.
[46] P. Stoica and T. L. Marzetta, "Parameter estimation problems with singular information matrices," *IEEE Trans. Signal Process.*, vol. 49, no. 1, pp. 87–90, Jan. 2001.
[47] H. Weyl, "Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung)," (in German) *J. Mathematische Annalen*, vol. 71, no. 4, pp. 441–479, Dec. 1912.
[48] G. W. Stewart, "On the perturbation of pseudo-inverses, projections and linear least squares problems," *SIAM Rev.*, vol. 19, no. 4, pp. 634–662, Oct. 1977.

Mert Pilanci (S'06) was born in Eskisehir, Turkey, in 1987. He received the B.Sc. degree in electrical and electronics engineering from Bilkent University, Ankara, Turkey, in 2008. He is currently working towards the M.S. degree under the supervision of Prof. O. Arikan in the Department of Electrical and Electronics Engineering, Bilkent University.

His research interests are in applied linear algebra and numerical analysis, inverse problems and compressed sensing.

Orhan Arikan (M'91) was born in Manisa, Turkey, in 1964. He received the B.Sc. degree in electrical and electronics engineering from the Middle East Technical University, Ankara, Turkey, in 1986 and both the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Illinois, Urbana-Champaign, in 1988 and 1990, respectively.

Following his graduate studies, he worked for three years as a Research Scientist at Schlumberger-Doll Research, Ridgefield, CT. In 1993, he joined Bilkent University, Ankara, Turkey, where he is presently Professor of electrical engineering since 2006. His current research interests are in statistical signal processing, time-frequency analysis, and array signal processing.

Mustafa C. Pinar received the B.Sc. degree in industrial engineering from Bogazici University, Ankara, Turkey, in 1987 and the M.Sc. and Ph.D. degrees in systems engineering from the University of Pennsylvania, Philadelphia, in 1989 and 1992, respectively.

He is currently a Professor of industrial engineering at Bilkent University, Ankara, Turkey. His research interests are in applied numerical optimization.