This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2759201, IEEE Access

Bounded Perturbation Regularization for Linear Least Squares Estimation

Tarig Ballal *Member, IEEE*, Mohamed A. Suliman and Tareq Y. Al-Naffouri, *Member, IEEE*

***Abstract*—This paper addresses the problem of selecting the regularization parameter for linear least-squares estimation. We propose a new technique called *bounded perturbation regularization* (BPR). In the proposed BPR method, a perturbation with a bounded norm is allowed into the linear transformation matrix to improve the singular-value structure. Following this, the problem is formulated as a min-max optimization problem. Next, the min-max problem is converted to an equivalent minimization problem to estimate the unknown vector quantity. The solution of the minimization problem is shown to converge to that of the ℓ_2-regularized least squares problem, with the unknown regularizer related to the norm bound of the introduced perturbation through a nonlinear constraint. A procedure is proposed that combines the constraint equation with the *mean squared error* (MSE) criterion to develop an approximately optimal regularization parameter selection algorithm. Both direct and indirect applications of the proposed method are considered. Comparisons with different Tikhonov regularization parameter selection methods, as well as with other relevant methods, are carried out. Numerical results demonstrate that the proposed method provides significant improvement over state-of-the-art methods.**

***Index Terms*—Linear estimation, least squares, Tikhonov regularization, mean squared error.**

I. Introduction

This paper addresses the problem of estimating a vector quantity $\mathbf{x} \in \mathbb{C}^n$ from an observation vector

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{z}, \tag{1}$$

where $\mathbf{y} \in \mathbb{C}^m$ is an observed data vector, $\mathbf{A} \in \mathbb{C}^{m \times n}$ is a known full-rank linear transformation matrix and $\mathbf{z}$ is white noise with zero mean and unknown variance σ_z^2. We focus on the case where $\mathbf{x}$ is zero-mean stochastic vector with independent but not (necessarily) identically distributed (i.n.d.) elements. The covariance matrix of $\mathbf{x}$ is denoted as $\mathbf{C}_{xx}$ and is positive semi-definite. Problems based on the model (1) arise in many areas of science and engineering such as communications, signal processing, machine learning, geophysics, econometrics and control [1], [2].

This work was supported by the KAUST-KFUPM joint research initiative and the KAUST CRG3 funding. The authors are with the Electrical Engineering Department, King Abdullah University of Science and Technology (KAUST), Thuwal 23955, Saudi Arabia. E-mail: {Tarig.Ahmed, Mohamed.Suliman, Tareq.Alnaffouri}@kaust.edu.sa.

Owing to the lack of prior information on $\mathbf{x}$, usually, a *least squares* (LS) approach is pursued. Ordinary LS estimation attempts to find an estimate of $\mathbf{x}$ that minimizes the squared norm of the residual error, i.e.,

$$\min_{\hat{\mathbf{x}}} ||\mathbf{A}\hat{\mathbf{x}} - \mathbf{y}||_2^2, \tag{2}$$

where $||.||_2$ is the ℓ_2 norm. Solving this minimization problem results in the LS estimator [1], [3]:

$$\hat{\mathbf{x}}_{\text{LS}} = (\mathbf{A}^H\mathbf{A})^{-1}\mathbf{A}^H\mathbf{y}, \tag{3}$$

where $(.)^H$ is the Hermitian transpose of the matrix. The agnostic nature of the LS estimator (3) to the probabilistic structure of the data and its ease of implementation have made it very popular in many practical situations [3]. Adds to that is the fact that this simple LS estimator is unbiased.

However, the main issue with (3) is that, depending on the *singular-value* structure of $\mathbf{A}$, the solution can be very sensitive to noise. This makes (3) useless in many practical situations, e.g., when the *condition number* of $\mathbf{A}$ (the ratio of the largest singular value to the smallest) is very large [4], [5], [6]. To overcome this difficulty, usually *regularization* methods are applied. Existing regularization techniques include ridge regression [4], the shrunken estimator [7], the covariance shaping LS estimator [8], and Tikhonov regularization with its different forms [9], [10], [6]. In Tikhonov *regularized LS* (RLS) methods, a penalty term is added to (2) to *shrink* the elements of $\hat{\mathbf{x}}$ towards zero. Different penalties have been proposed for different kinds of problems. A popular form of the RLS problem is the ℓ_2 penalized form given by

$$\min_{\hat{\mathbf{x}}} ||\mathbf{A}\hat{\mathbf{x}} - \mathbf{y}||_2^2 + \lambda||\hat{\mathbf{x}}||_2^2, \tag{4}$$

where $\lambda \in \mathbb{R}^+$ is a regularization parameter. The solution of (4) can be obtained as [4], [11], [6]

$$\hat{\mathbf{x}}_{\text{RLS}} = (\mathbf{A}^H\mathbf{A} + \gamma\mathbf{I}_n)^{-1}\mathbf{A}^H\mathbf{y}, \tag{5}$$

where $\gamma = \lambda$ (always positive), and $\mathbf{I}_n$ is an identity matrix of dimension n. This widely used form of RLS is a special case of what is referred to as *Tikhonov regularization* [9], [10], [6]. For $\gamma = 0$, (5) reduces to the *ordinary* LS estimator (3). It is proved in [4] that there always exists a positive value γ such that the

estimator (5) offers lower *mean squared error* (MSE) than the ordinary LS estimator (3).

The key question in (5) is how to find a suitable value of the regularization parameter γ. Several methods exist for choosing the regularization parameter required in (5). These methods include the *l-curve* [12], the *generalized cross validation* (GCV) [13] and the *quasi-optimal* [14] methods, to mention a few. These methods use different criteria to select the regularization parameter. The performance of different regularization parameter selection methods varies significantly depending on the problem and the scenario under consideration. A comparison of regularization parameter selection methods is given in [15].

In this paper, we approach the regularization problem in a different way. Instead of starting from (4), the proposed method works by allowing for an artificial perturbation to improve the singular-value structure of the matrix $\mathbf{A}$. We formulate the problem as a min-max problem, which is converted to a minimization problem. The solution of the minimization problem is obtained assuming knowledge of the artificial perturbation second norm bound. A combination of the MSE criterion, ideas from stochastic programming and some approximations are utilized to set the perturbation bound in an optimal or near-optimal way, depending on the nature of $\mathbf{x}$. The proposed method is shown to converge to the Bayesian *linear minimum mean squared error* (LMMSE) estimator [3] when the elements of the vector $\mathbf{x}$ are i.i.d. The solution obtained based on proposed method coincides with the RLS form (5) where the parameter γ is obtained by solving a nonlinear equation in the parameter γ.

A. Notations

We use lower-case bold-face letters to denote (column) vectors (e.g., $\mathbf{x}$) and upper-case bold-face letters to denote matrices (e.g., $\mathbf{V}$). Notations such as $\boldsymbol{u}_i$ denotes the i'th column of a matrix $\mathbf{V}$. Lower-case letters denote scalars (e.g., γ). The i-th column of a matrix $\mathbf{V}$ is denoted by $\mathbf{v}_i$. The i-th element of a vector $\mathbf{x}$ is denoted by x_i. The notations $(\cdot)^H$ and $\mathrm{tr}(\cdot)$ are used to denote the vector/matrix Hermitian transpose operation and the trace of a matrix, respectively. For a vector argument, the operation denoted by $\mathrm{diag}(\cdot)$ returns the diagonal matrix whose diagonal entries are the vector's elements. For a matrix argument, $\mathrm{diag}(\cdot)$ returns the vector with the diagonal elements of the matrix as its entries. The statistical expectation operation is denoted by $\mathbb{E}(\cdot)$, while the estimated value is denoted by $\hat{(\cdot)}$. The notation $||\cdot||_2$ denotes the ℓ_2 norm in the case of a vector, or the 2-induced norm in the case of a matrix. The symbol $\mathbf{I}_n$ is the identity matrix of dimension n. The short form i.i.d. stands for "independent and identically distributed", while "independent but not identically distributed" is abbreviated as i.n.d.

II. Analysis of The MSE of the RLS Estimator

By applying the *singular-value decomposition* (SVD) [16]

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H, \tag{6}$$

(where $\mathbf{U} \in \mathbb{C}^{m\times n}$, $\mathbf{V} \in \mathbb{C}^{n\times n}$, and $\mathbf{\Sigma} = \mathrm{diag}([\sigma_1, \sigma_2, \cdots, \sigma_n]^T), \sigma_1 > \sigma_2 > \cdots > \sigma_n$), (5) can be written a

$$\hat{\mathbf{x}}_{\mathrm{RLS}} = \mathbf{V}\left(\mathbf{\Sigma}^2 + \gamma\mathbf{I}_n\right)^{-1}\mathbf{\Sigma}\mathbf{U}^H\mathbf{y}. \tag{7}$$

We define the overall MSE of the estimator (7) as

$$\mathrm{MSE} = \mathrm{tr}\left\{\mathbb{E}\left[(\hat{\mathbf{x}} - \mathbf{x})(\hat{\mathbf{x}} - \mathbf{x})^H\right]\right\}. \tag{8}$$

Based on (6), (1) can be written as $\mathbf{y} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H\mathbf{x} + \mathbf{z}$. By substituting this expression of $\mathbf{y}$ in (7), substituting the result in (8) and manipulating, we obtain

$$\begin{aligned}\mathrm{MSE} &= \sigma_z^2\mathrm{tr}\left[\mathbf{\Sigma}^2\left(\mathbf{\Sigma}^2 + \gamma\mathbf{I}_n\right)^{-2}\right] \\ &+ \gamma^2\mathrm{tr}\left[\left(\mathbf{\Sigma}^2 + \gamma\mathbf{I}_n\right)^{-2}\mathbf{V}^H\mathbf{C}_{xx}\mathbf{V}\right].\end{aligned} \tag{9}$$

where $\mathbf{C}_{xx} \triangleq \mathbb{E}\left(\mathbf{x}\mathbf{x}^H\right)$ is the covariance matrix of $\mathbf{x}$, which is diagonal with positive or zero elements. A similar MSE expression has been obtained in [4] for the deterministic case. As discussed in [4], the MSE function is the sum of a variance term (the first term) and a bias term (the second term). By examining the first derivative of each term, it is easy to see that the variance and the bias are, respectively, monotonically decreasing and monotonically increasing functions in γ. For $\gamma = 0$, i.e., for the ordinary LS, the bias term is zero. In other words, the effect of including a positive parameter γ in (7) is to reduce the variance of the estimator at the cost of introducing bias. As elaborated in [4], there always exist positive values of γ for which the MSE (variance + bias) is smaller than that of the ordinary LS estimator.

Theorem 1. *The MSE in* (9) *is a convex function in* γ.

Proof. To prove Theorem 1, we write (9) as the sum

$$\mathrm{MSE} = \sum_{i=1}^{n}\mathrm{MSE}_i = \sum_{i=1}^{n}\frac{\sigma_z^2\sigma_i^2 + \gamma^2\mathbf{v}_i^H\mathbf{C}_{xx}\mathbf{v}_i}{\left(\sigma_i^2 + \gamma\right)^2} \tag{10}$$

By considering each individual element of the sum and taking the gradient, we obtain

$$\mathrm{MSE}_i^{'} = \frac{2\sigma_i^2\left(\gamma\mathbf{v}_i^H\mathbf{C}_{xx}\mathbf{v}_i - \sigma_z^2\right)}{\left(\sigma_i^2 + \gamma\right)^3}. \tag{11}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2759201, IEEE Access

Form (11), it is obvious that each MSE_i has a unique critical point at

$$\gamma_i = \frac{\sigma_z^2}{\mathbf{v}_i^H \mathbf{C}_{xx} \mathbf{v}_i}. \tag{12}$$

Noting that $\mathbf{C}_{xx}$ is positive semi-definite implies that $\mathbf{v}_i^H \mathbf{C}_{xx} \mathbf{v}_i$ are positive (we do not consider the trivial case where all the diagonal elements of $\mathbf{C}_{xx}$ are zero). Based on this, it is easy to show that the second derivatives of MSE_i are always positive, which indicates that all γ_i are minima. Hence, all MSE_i are convex functions in γ. Consequently, the sum of these functions is also convex in γ. □

Corollary 1. *The minimum of the MSE in* (9) *as a function of* γ *is given by*

$$\tilde{\gamma} \in \left(\min_{i \in \{1,\cdots,n\}} \left(\frac{\sigma_z^2}{\mathbf{v}_i^H \mathbf{C}_{xx} \mathbf{v}_i} \right), \max_{i \in \{1,\cdots,n\}} \left(\frac{\sigma_z^2}{\mathbf{v}_i^H \mathbf{C}_{xx} \mathbf{v}_i} \right) \right). \tag{13}$$

Proof. Corollary 1 is a direct result of (12). □

Since the MSE function in (9) is convex, we can obtain the optimal value of γ by differentiating with respect to γ and equating to zero. After manipulating, this results in the following equation:

$$\begin{aligned} \text{MSE}^{'} &= -2\sigma_z^2 \text{tr}\left[\mathbf{\Sigma}^2 \left(\mathbf{\Sigma}^2 + \gamma \mathbf{I}_n \right)^{-3} \right] \\ &+ 2\gamma \text{tr}\left[\mathbf{\Sigma}^2 \left(\mathbf{\Sigma}^2 + \gamma \mathbf{I}_n \right)^{-3} \mathbf{V}^H \mathbf{C}_{xx} \mathbf{V} \right] = 0. \end{aligned} \tag{14}$$

From the above discussion, (14) has a unique positive root in the interval given in (13). The value of the root depends on the signal and noise second order statistics. Unfortunately, (14) does not admit a closed-form solution for γ in these statistics in the general case. To reach a closed form, we resort to approximation. Namely, we write (14) as

$$\begin{aligned} \text{MSE}^{'} &\approx -2\sigma_z^2 \text{tr}\left[\mathbf{\Sigma}^2 \left(\mathbf{\Sigma}^2 + \gamma \mathbf{I}_n \right)^{-3} \right] \\ &+ 2\gamma \frac{\text{tr}\left(\mathbf{C}_{xx} \right)}{n} \text{tr}\left[\mathbf{\Sigma}^2 \left(\mathbf{\Sigma}^2 + \gamma \mathbf{I}_n \right)^{-3} \right] = 0, \end{aligned} \tag{15}$$

which yields an approximately optimal value of γ given by

$$\tilde{\gamma} \approx \frac{n\sigma_z^2}{\text{tr}\left(\mathbf{C}_{xx} \right)}. \tag{16}$$

It is noted that for a vector $\mathbf{x}$ with zero mean and i.i.d. elements, (15) is *exactly* equivalent to (14); hence, (16) is the exact optimal value of γ for the i.i.d. case. When the elements of $\mathbf{x}$ are i.n.d., $\tilde{\gamma}$ is guaranteed to fall in the optimal interval defined in (13). It is clear that the smaller the optimal interval the more precise (13) is. The size of the optimal interval depends on the covariance matrix of $\mathbf{x}$ and the eigenvector matrix $\mathbf{V}$. The optimal interval is bounded unless a column (or more) of $\mathbf{V}$ is orthogonal to the vector of the diagonal elements of $\mathbf{C}_{xx}$, a peculiar setup that we do not consider in this paper. In the i.i.d. case, the optimal interval shrinks to a single point given by (16). In Appendix (A), we analyze the error of the approximation in (15) that leads to to the result in (16).

In the Section (III), we propose a method to obtain a regularization parameter value close to (16). In Section V, we demonstrate that (16) is sufficient to obtain near-optimal performance in different scenarios.

III. The Proposed Bounded Perturbation Regularization Approach

In the proposed *bounded perturbation regularization* (BPR) method, we perturb the data model (1) to

$$\mathbf{y} \approx (\mathbf{A} + \mathbf{\Delta})\, \mathbf{x} + \mathbf{z}, \tag{17}$$

where $\mathbf{\Delta} \in \mathbb{C}^{m \times n}$ is an unknown perturbation matrix. The aim of adding such perturbation is to alter the singular values of $\mathbf{A}$ such that the resultant model (17) has a better structure compared to (1). By adding this perturbation, we choose to tradeoff model accuracy for model stability. We envision that, for certain choices of $\mathbf{\Delta}$, the gain in robustness against noise outweighs the loss in model accuracy. It is obvious that the perturbation should somehow be bounded since extremely perturbing the model can destroy the model fidelity. Therefore, we bound the 2-induced norm of the matrix $\mathbf{\Delta}$ by a positive amount ζ, i.e.,

$$||\mathbf{\Delta}||_2 \leq \zeta. \tag{18}$$

This bound is generally not known and is a key subject of the proposed BPR method. For now, let us assume that ζ is a constant whose value is known. Later, we will make up for the lack of knowledge of the perturbation bound ζ and eliminate this parameter from the final solution.

To find an estimate of $\mathbf{x}$, we pursue the following min-max optimization:

$$\begin{aligned} \min_{\hat{\mathbf{x}}} \max_{\mathbf{\Delta}} & ||\mathbf{y} - (\mathbf{A} + \mathbf{\Delta})\hat{\mathbf{x}}||_2 \\ & \text{subject to: } ||\mathbf{\Delta}||_2 \leq \zeta. \end{aligned} \tag{19}$$

The rationale behind (19) is that we seek an estimate of $\mathbf{x}$ which minimizes the *maximum* residual error over all possible bounded perturbations Δ.

Now, using Minkowski's inequality [17] and the fact that $||\mathbf{\Delta}\hat{\mathbf{x}}||_2 \leq ||\mathbf{\Delta}||_2 ||\hat{\mathbf{x}}||_2 \leq \zeta ||\mathbf{\Delta}||_2$, we have

$$\begin{aligned} ||\mathbf{y} - (\mathbf{A} + \mathbf{\Delta})\hat{\mathbf{x}}||_2 &\leq ||\mathbf{y} - \mathbf{A}\hat{\mathbf{x}}||_2 + ||\mathbf{\Delta}\hat{\mathbf{x}}||_2 \\ &\leq ||\mathbf{y} - \mathbf{A}\hat{\mathbf{x}}||_2 + ||\mathbf{\Delta}||_2 ||\hat{\mathbf{x}}||_2 \\ &\leq ||\mathbf{y} - \mathbf{A}\hat{\mathbf{x}}||_2 + \zeta ||\hat{\mathbf{x}}||_2. \end{aligned} \tag{20}$$

We can see that $||\mathbf{y} - \mathbf{A}\hat{\mathbf{x}}||_2 + \zeta ||\hat{\mathbf{x}}||_2$ is a bound on $||\mathbf{y} - (\mathbf{A} + \mathbf{\Delta})\hat{\mathbf{x}}||_2$. By simple substitution and manipulation,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2759201, IEEE Access

it can easily be shown that this bound is attainable for

$$\Delta = \Delta_{\text{bound}} = \frac{-\zeta(\mathbf{y} - \mathbf{A}\hat{\mathbf{x}})\hat{\mathbf{x}}^H}{||\mathbf{y} - \mathbf{A}\hat{\mathbf{x}}||_2||\hat{\mathbf{x}}||_2}. \quad (21)$$

Hence, solving the min-max problem (19) is equivalent to solving the minimization problem

$$\min_{\hat{\mathbf{x}}} ||\mathbf{y} - \mathbf{A}\hat{\mathbf{x}}||_2 + \zeta||\hat{\mathbf{x}}||_2. \quad (22)$$

Comparing (22) and (4), we can see similarity between the two cost functions. However, the difference lies in that in (22) the norms are not squared, which makes the two functions mathematically inequivalent. Interestingly, the solution of (22) can be shown to take the RLS form (5) (or equivalently, (7)), with the additional constraint

$$\gamma = \frac{\zeta||\mathbf{y} - \mathbf{A}\hat{\mathbf{x}}||_2}{||\hat{\mathbf{x}}||_2}, \quad (23)$$

where $\hat{\mathbf{x}} = \hat{\mathbf{x}}_{\text{RLS}}$ as in (5). Compared to $\gamma = \lambda$ that corresponds to the cost function in (4), (23) provides a richer expression for γ. We can see that γ is directly proportional to the perturbation bound ζ, while it is also dependent on the residual error produced by the estimate $\hat{\mathbf{x}}$ and the ℓ_2 norm of that estimate. Note that both ζ and $\hat{\mathbf{x}}$ are not known. Subsequently, we will show how the formula (23), combined with the MSE criterion highlighted in Section II, can be utilized to find a value of the parameter γ that coincides with a near-optimal choice of the bound ζ.

Now, we have the RLS problem solution (5), with γ given by (23). It is obvious that with (5) and (23) in hand, we need only the value of the bound ζ in order to find the value of the regularization parameter γ. By substituting (5) in (23) and manipulating, we obtain the equation

$$\begin{aligned}&\gamma^2\mathbf{y}^H\mathbf{A}\left(\mathbf{A}^H\mathbf{A} + \gamma\mathbf{I}_n\right)^{-2}\mathbf{A}^H\mathbf{y} \\ &- \zeta^2\left[\mathbf{y}^H\mathbf{y} - \mathbf{y}^H\mathbf{A}\left(\mathbf{A}^H\mathbf{A} + \gamma\mathbf{I}_n\right)^{-1}\mathbf{A}^H\mathbf{y}\right. \\ &\left.-\gamma\mathbf{y}^H\mathbf{A}\left(\mathbf{A}^H\mathbf{A} + \gamma\mathbf{I}_n\right)^{-2}\mathbf{A}^H\mathbf{y}\right] = 0. \quad (24)\end{aligned}$$

Now, we apply the matrix decomposition in (6) and manipulate (24) to the form

$$\mathbf{y}^H\mathbf{U}\left(\mathbf{\Sigma}^2 - \zeta^2\mathbf{I}_n\right)\left(\mathbf{\Sigma}^2 + \gamma\mathbf{I}_n\right)^{-2}\mathbf{U}^H\mathbf{y} = 0. \quad (25)$$

Equation (25) can be solved to obtain the value of the regularization parameter required by (5). However, we need a value for the parameter ζ. This parameter dictates the amount of regularization introduced, and hence, it determines the quality of the final solution. Therefore, care has to be taken when the value of this parameter is chosen. In the following subsection, we will apply the results from Section II to set the bound ζ such that (25) yields an optimal or near-optimal value of the regularization parameter.

A. Setting The bound ζ

As presented earlier in this section, BPR seeks a bound on the norm of an artificial perturbation matrix $\mathbf{\Delta}$ that improves the LS problem solution. In this subsection, we propose a strategy for choosing ζ such that we approximately optimize performance in the MSE sense.

To start with, let us consider the optimal regularization parameter $\tilde{\gamma}$ in (16). This parameter is deterministic. However, $\tilde{\zeta}$ as in (25) is stochastic since it is dependent on $\mathbf{y}$, which is linearly dependent on two stochastic variables, $\mathbf{x}$ and the noise $\mathbf{z}$, as per Equation (1). Thus, it can be concluded that there exists no single deterministic value $\tilde{\zeta}$ that can produce $\tilde{\gamma}$ from Equation (25). To resolve this issue, we will draw on ideas from the field of stochastic programming [18]. First, we seek to obtain a deterministic relationship between $\tilde{\gamma}$ and $\tilde{\zeta}$. Instead of seeking a $\tilde{\zeta}$ that satisfies (25) for each individual realization of $\mathbf{y}$; we try to find a $\tilde{\zeta}$ that satisfies (25) *on average*, i.e., we consider

$$\mathbb{E}\left[\mathbf{y}^H\mathbf{U}\left(\mathbf{\Sigma}^2 - \tilde{\zeta}^2\mathbf{I}_n\right)\left(\mathbf{\Sigma}^2 + \tilde{\gamma}\mathbf{I}_n\right)^{-2}\mathbf{U}^H\mathbf{y}\right] = 0. \quad (26)$$

We can solve (26) for $\tilde{\zeta}$ and manipulate to obtain

$$\tilde{\zeta}^2 = \frac{\text{tr}\left[\mathbf{\Sigma}^2\left(\mathbf{\Sigma}^2 + \tilde{\gamma}\mathbf{I}_n\right)^{-2}\mathbf{U}^H\mathbf{C}_{\text{yy}}\mathbf{U}\right]}{\text{tr}\left[\left(\mathbf{\Sigma}^2 + \tilde{\gamma}\mathbf{I}_n\right)^{-2}\mathbf{U}^H\mathbf{C}_{\text{yy}}\mathbf{U}\right]}, \quad (27)$$

where $\mathbf{C}_{\text{yy}} = \mathbb{E}\left(\mathbf{y}\mathbf{y}^H\right)$. Equation (27) is a deterministic relationship between $\tilde{\gamma}$ and $\tilde{\zeta}$. If we know $\tilde{\gamma}$ and $\mathbf{C}_{\text{yy}}$, then we can calculate $\tilde{\zeta}$, which is optimal on average. Now, let us try to simplify (27). Based on (1) and (6), we can write

$$\begin{aligned}\mathbf{C}_{\text{yy}} = \mathbb{E}\left(\mathbf{y}\mathbf{y}^H\right) &= \mathbf{A}\mathbf{C}_{xx}\mathbf{A}^H + \sigma_z^2\mathbf{I}_n \\ &= \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H\mathbf{C}_{xx}\mathbf{V}\mathbf{\Sigma}\mathbf{U}^H + \sigma_z^2\mathbf{I}_n \quad (28)\end{aligned}$$

By substituting (28) in (27), we obtain (29).

Now, we can use $\tilde{\zeta}$ as a feasible bound on the perturbation added to achieve regularization. We can plug this bound into (25) and solve to obtain the value of the regularization parameter γ for the current observation vector $\mathbf{y}$. Unfortunately, the parameter $\tilde{\zeta}$ is dependent on the noise variance and the covariance matrix $\mathbf{C}_{xx}$, which are unknown. In the next step, we attempt to remove this dependence by exploiting the fact that $\tilde{\gamma}$ is also a function of the noise variance and the covariance matrix $\mathbf{C}_{xx}$, as per (16). To this end, we resort to an approximation similar to that in (15). Namely, we modify (29) into (30). An analysis of this approximation is presented in Appendix A. Equation (30) can easily be written in the form

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2759201, IEEE Access

$$\tilde{\zeta}^2 = \frac{\sigma_z^2 \text{tr}\left[\boldsymbol{\Sigma}^2\left(\boldsymbol{\Sigma}^2+\tilde{\gamma}\mathbf{I}_n\right)^{-2}\right] + \text{tr}\left[\boldsymbol{\Sigma}^4\left(\boldsymbol{\Sigma}^2+\tilde{\gamma}\mathbf{I}_n\right)^{-2}\mathbf{V}^H\mathbf{C}_{xx}\mathbf{V}\right]}{\sigma_z^2 \text{tr}\left[\left(\boldsymbol{\Sigma}^2+\tilde{\gamma}\mathbf{I}_n\right)^{-2}\right] + \text{tr}\left[\boldsymbol{\Sigma}^2\left(\boldsymbol{\Sigma}^2+\tilde{\gamma}\mathbf{I}_n\right)^{-2}\mathbf{V}^H\mathbf{C}_{xx}\mathbf{V}\right]} \tag{29}$$

$$\tilde{\zeta}^2 \approx \frac{\sigma_z^2 \text{tr}\left[\boldsymbol{\Sigma}^2\left(\boldsymbol{\Sigma}^2+\tilde{\gamma}\mathbf{I}_n\right)^{-2}\right] + \frac{\text{tr}(\mathbf{C}_{xx})}{n}\text{tr}\left[\boldsymbol{\Sigma}^4\left(\boldsymbol{\Sigma}^2+\tilde{\gamma}\mathbf{I}_n\right)^{-2}\right]}{\sigma_z^2 \text{tr}\left[\left(\boldsymbol{\Sigma}^2+\tilde{\gamma}\mathbf{I}_n\right)^{-2}\right] + \frac{\text{tr}(\mathbf{C}_{xx})}{n}\text{tr}\left[\boldsymbol{\Sigma}^2\left(\boldsymbol{\Sigma}^2+\tilde{\gamma}\mathbf{I}_n\right)^{-2}\right]} \tag{30}$$

$$\tilde{\zeta}^2 \approx \frac{\text{tr}\left[\boldsymbol{\Sigma}^2\left(\boldsymbol{\Sigma}^2+\tilde{\gamma}\mathbf{I}_n\right)^{-2}\left(\boldsymbol{\Sigma}^2+\frac{n\sigma_z^2}{\text{tr}(\mathbf{C}_{xx})}\mathbf{I}_n\right)\right]}{\text{tr}\left[\left(\boldsymbol{\Sigma}^2+\tilde{\gamma}\mathbf{I}_n\right)^{-2}\left(\boldsymbol{\Sigma}^2+\frac{n\sigma_z^2}{\text{tr}(\mathbf{C}_{xx})}\mathbf{I}_n\right)\right]}. \tag{31}$$

Based on (16), we can insert $\tilde{\gamma}$ to replace $\frac{n\sigma_z^2}{\text{tr}(\mathbf{C}_{xx})}$ and manipulate to obtain

$$\tilde{\zeta}^2 = \frac{\text{tr}\left[\boldsymbol{\Sigma}^2\left(\boldsymbol{\Sigma}^2+\tilde{\gamma}\mathbf{I}_n\right)^{-1}\right]}{\text{tr}\left[\left(\boldsymbol{\Sigma}^2+\tilde{\gamma}\mathbf{I}_n\right)^{-1}\right]}. \tag{32}$$

Equation (32) dictates the relationship between the best choice (on average and approximately) of the bound on the perturbation matrix norm ($\tilde{\zeta}$), on the one hand, and the corresponding $\tilde{\gamma}$ that minimizes the MSE, on the other hand. All the quantities involved are deterministic and the unknown perturbation bound is dependent on only one unknown quantity, which is the optimal regularization parameter $\tilde{\gamma}$. Substituting (32) in (25), and for simplicity of notations, replacing $\tilde{\gamma}$ with γ, we obtain

$$\begin{aligned} f(\gamma) = &\underbrace{\text{tr}\left[\left(\boldsymbol{\Sigma}^2+\gamma\mathbf{I}_n\right)^{-1}\right]\text{tr}\left[\boldsymbol{\Sigma}^2\left(\boldsymbol{\Sigma}^2+\gamma\mathbf{I}_n\right)^{-2}\mathbf{b}\mathbf{b}^H\right]}_{f_1(\gamma)} \\ &- \underbrace{\text{tr}\left[\boldsymbol{\Sigma}^2\left(\boldsymbol{\Sigma}^2+\gamma\mathbf{I}_n\right)^{-1}\right]\text{tr}\left[\left(\boldsymbol{\Sigma}^2+\gamma\mathbf{I}_n\right)^{-2}\mathbf{b}\mathbf{b}^H\right]}_{f_2(\gamma)} \\ &= 0, \end{aligned} \tag{33}$$

where $\mathbf{b} \triangleq \mathbf{U}^H\mathbf{y}$. We can manipulate (33) further to reach the form

$$\begin{aligned} f(\gamma) = &\underbrace{\text{tr}\left[\left(\boldsymbol{\Sigma}^2+\gamma\mathbf{I}_n\right)^{-1}\right]\text{tr}\left[\left(\boldsymbol{\Sigma}^2+\gamma\mathbf{I}_n\right)^{-1}\mathbf{b}\mathbf{b}^H\right]}_{\mathcal{Q}_1(\gamma)} \\ &- \underbrace{n\,\text{tr}\left[\left(\boldsymbol{\Sigma}^2+\gamma\mathbf{I}_n\right)^{-2}\mathbf{b}\mathbf{b}^H\right]}_{\mathcal{Q}_2(\gamma)} = 0, \end{aligned} \tag{34}$$

which will be referred to as *BPR equation*. Equation (33) is an equivalent BPR equation that will be utilized to establish some of the proofs in the following subsection.

Remark 1. *The MSE criterion has been incorporated into the derivation of* (34)*; hence we expect the solution of* (34) *to yield regularization parameter values that minimize the MSE of the RLS estimator when the elements of* $\mathbf{x}$ *are i.i.d., and that approximately minimize the MSE when the elements of* $\mathbf{x}$ *are i.n.d. An example of a related work that incorporates the MSE in the design of a regularized estimator is [19], where the Stein's unbiased risk estimate (SURE) [20] of the MSE is employed. A major difference between the regularized estimator in [19] and the proposed method is that the expression of the estimator in [19] involves the noise variance, whereas the proposed estimator does not require prior knowledge of the latter noise statistic.*

B. Properties of $f(\gamma)$

We are interested in finding a positive root of the function $f(\gamma)$. Before trying to do so, let us examine some of the properties of this function.

Property 1. *$f(\gamma)$ has n discontinuities at $\gamma = -\sigma_i^2, \forall i = 1, \cdots n$. These discontinuities are of no interest as far as BPR is concerned.*

Property 2. *$f(\gamma)$ is continuous in the interval $(-\sigma_n^2, +\infty)$.*

Property 3. $\lim_{\gamma\to+\infty} f(\gamma) = 0$.

Property 4. $\lim_{\gamma\to-a_n} f(\gamma) = -\infty$.

The proofs of property 1–3 are straightforward. The proof of property 4 is given in Appendix B.

Property 5. *The functions $f_1(\gamma)$, $f_2(\gamma)$, $\mathcal{Q}_1(\gamma)$ and $\mathcal{Q}_2(\gamma)$, as in (33) and (34), are completely monotonic in the interval $(-\sigma_n^2, +\infty)$.*

Proof. Since these functions are continuous in the interval of interest, this property can be proved simply by continuously differentiating each of these functions and, by induction, we will find that each function, say $g(\gamma)$, satisfies the complete monotonicity condition [21], [22]:

$$(-1)^k g^{(k)}(\gamma) \geq 0, \forall k \in \mathbb{N}, \tag{35}$$

where $(.)^{(k)}$ is the k-th derivative. □

Theorem 2. *The function $f(\gamma)$, defined in (33) and (34), has at most two roots in the interval $(-\sigma_n^2, +\infty)$.*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2759201, IEEE Access

Proof. The proof of Theorem 2 stems from two pieces of already-published results. First, it has been shown that a completely monotonic function can be represented as a sum of exponentials. Namely, a completely monotonic function $g(\gamma)$ can be uniformly approximated as [21], [22]

$$g(\gamma) \approx \sum_{i=1}^{z} \beta_i e^{-\alpha_i \gamma}, \tag{36}$$

where $\alpha_i, \beta_i \in \mathcal{R}$ and z is the number of exponentials required to achieve certain level of approximation accuracy. It is shown in [22] that *a best uniform* approximation of $g(\gamma)$ always exists. It is also shown that the error in this approximation gets smaller and smaller as z increases. For a large z value, finding the approximation parameters α_i and β_i becomes very challenging. However, herein, the interest in the general form (36) rather than the specific values involved. To sum up, any of the functions $f_1(\gamma)$, $f_2(\gamma)$, $\mathcal{Q}_1(\gamma)$ and $\mathcal{Q}_2(\gamma)$ (defined in (33) and (34)), can be written in the form (36) for a value of z which is sufficiently large for the approximation (36) to hold with a negligible amount of error.

Now, based on the results published in [23], two sum-of-exponential functions can intersect at most at two points. Consequently, $f_1(\gamma)$ and $f_2(\gamma)$ have a maximum of two possible intersections. The same applies to $\mathcal{Q}_1(\gamma)$ and $\mathcal{Q}_2(\gamma)$. Thus, the function $f(\gamma)$ can have at most two intersections in the interval $(-\sigma_n^2, +\infty)$. □

Theorem 3. *A sufficient condition for the function $f(\gamma)$, as in (33) and (34), to have at least one root in the interval $(-\sigma_n^2, +\infty)$ is give by*

$$n\, tr\left(\mathbf{\Sigma}^2 \mathbf{b}\mathbf{b}^H\right) > tr\left(\mathbf{\Sigma}^2\right) tr\left(\mathbf{b}\mathbf{b}^H\right). \tag{37}$$

The proof of Theorem 3 is relegated to Appendix C.

Theorem 4. *If (37) is satisfied, then the function $f(\gamma)$ has a unique root in the interval $(-\sigma_n^2, +\infty)$.*

Proof. We prove this theorem by contradiction. Based on Theorem 2 and Theorem 3, the function $f(\gamma)$ has either one or two roots when (37) is satisfied. Let us assume that two roots, γ_1 and γ_2, with $\gamma_1 < \gamma_2$, exist. Based on Rolle's Theorem [24], $f(\gamma)$ must have an extremum between γ_1 and γ_2 which occurs, say, at γ_m[1]. From Property 4, and since γ_1 is the first zero crossing in the interval $(-\sigma_n^2, +\infty)$, the function $f(\gamma)$ is increasing at γ_1. Hence, the extremum at γ_m is a maximum and the function is thus decreasing at γ_2. Considering Property 3, we expect the function to have a minimum with a negative value at some point in the interval $(\gamma_2, +\infty)$. However, from the proof of Theorem 3 (see Appendix C), the condition in (37) guarantees that the function $f(\gamma)$ approaches zero from the positive direction when γ approaches $+\infty$. Thus, using the intermediate value theorem [24], the function $f(\gamma)$ must have a zero crossing in the interval $(\gamma_2, +\infty)$. This means that the function must have a third root, which contradicts Theorem 2. The only way to avoid a third root is for the second root not to exist. Therefore, we conclude that, when (37) holds, the function $f(\gamma)$ will have one and only one root in the interval $(-\sigma_n^2, +\infty)$. □

[1]The case where $f(\gamma) = 0$ all over the interval $[\gamma_1, \gamma_2]$ is out of the question.

Corollary 2. *If $f(\gamma)$ has a positive root, γ^+, and (37) holds, then $f(\gamma)$ is always negative in the interval $[0, \gamma^+)$.*

Corollary 3. *If $f(\gamma)$ has a positive root, γ^+, and (37) holds, then $f(\gamma)$ is an increasing function in the interval $[0, \gamma^+]$.*

Corollary 4. *If $f(\gamma)$ has a negative root, γ^-, and (37) holds, then $f(\gamma)$ is always positive in the interval $(\gamma^-, 0]$.*

Corollaries 2–4 are direct results of Theorem 4, Property 4 and the intermediate value theorem. These corollaries will be used later to devise a method to find the root of $f(\gamma)$, and to establish the convergence of that method.

Remark 2. *From Theorem 3, the function $f(\gamma)$ may not have a positive root (when* (37) *does not hold). In such a case, the proposed BPR method will not be applicable. However, it will be shown that* (37) *is satisfied in many application scenarios.*

C. Finding The Root of $f(\gamma)$

Since the function $f(\gamma)$ is differentiable and the expression of the first derivative $f'(\gamma)$ is easily obtainable, Newton's method [25] can be applied to find the root in a straightforward manner. Starting from an initial value $\gamma^{t=0}$, we carry out the iterations

$$\gamma^{t+1} = \gamma^t - \frac{f(\gamma^t)}{f'(\gamma^t)}. \tag{38}$$

The process is stopped when $|f(\gamma^{t+1})| < \epsilon$, where ϵ is a sufficiently small positive quantity.

D. Convergence

In the case where a positive root γ^+ exists, the convergence of (38) can be proved using Corollary 2 and Corollary 3. As Corollary 2 states, $f(\gamma^t) \leq 0$ for $[0, \gamma^+]$. On the other hand, Corollary 3 emphasizes that $f'(\gamma^t) \geq 0$ for $[0, \gamma^+]$. Thus, starting from $\gamma^0 = 0$, (38) will produce a progressively increasing estimate of γ. Convergence occurs when $\gamma^t \to \gamma^+$, in which case $f(\gamma^t) \to 0$ and $\gamma^{t+1} \to \gamma^t$. It can also be shown

that starting from $\gamma^0 = 0$, (38) can also converge to the negative root. However, one can use Corollary 4, to abandon the whole process in (38) when the root is negative. Namely, we pursue (38) only after we confirm that $f(\gamma^0) < 0$ holds true.

E. BPR Algorithm Summary

Based on the previous subsections, the proposed BPR algorithm is summarized as follows:

1) Set $\gamma = 0$.
2) Calculate $f(0)$ based on the definition of $f(\gamma)$ in (34).
3) If $f(0) < 0$;
   3.1 apply Newton's method (38) till it converges to the root γ^+,
   3.2 set $\gamma = \gamma^+$;
4) Calculate $\hat{\mathbf{x}}$ using (7).

The computational complexity of the BPR algorithm is dominated by the computational complexity of calculating the SVD, which is roughly $\mathcal{O}\left(mn^2\right)$, and that of solving (34), which is $\mathcal{O}\left(mn + Tn\right)$, where T is the total number of iterations required for (38) to converge. A Matlab implementation of the BPR algorithm is available online[2].

IV. Applications

In this section, we discuss two main applications of the proposed regularization method. These two applications represent one direct application where the signal model complies with (1); and an indirect application where (1) is reached through a modification of an existing formula. These applications are *signal estimation* and *robust beamforming*.

A. Signal Estimation

In signal estimation for *multiple-input multiple-output* (MIMO) systems, the central problem is to estimate a transmitted symbol vector $\mathbf{x}$ from an observation vector $\mathbf{y}$ related to $\mathbf{x}$ through the linear transformation (1). In this case, $\mathbf{A}$ is the *channel matrix* [26], [27]. In communication literature, when the LS estimator (without regularization) is applied to solve this problem, it is usually termed the "zero-forcing" estimator. Different structures of the channel matrix are discussed in the literature. These include $\mathbf{A}$ being a Gaussian i.i.d. matrix, and $\mathbf{A}$ being a Gaussian i.i.d. matrix with left or/and right correlation [28].

[2]Check: https://www.researchgate.net/publication/319042479_bpr.

B. Robust Beamforming

As opposed to signal estimation, we apply the proposed method to the beamforming problem in a rather indirect manner. The beamforming problem is stated and discussed in many references, e.g., [29], [30], [31], [32], [33], [34], [35]. Solutions are proposed in the form of different beamforming algorithms. Here we focus on the *minimum variance distortionless* (MVDR) beamformer, which is also known as *Capon* beamformer [36]. In particular, we consider the case where uncertainty occurs in the *steering vector* of a *uniform linear array* and hence *robust* beamforming is required. The output of any beamformer for a linear array with n_e array elements, at a discrete time instant t, is given by

$$y_{\text{BF}}[t] = \mathbf{w}^H \mathbf{y}[t], \tag{39}$$

where $\mathbf{w} \in \mathbb{C}^{n_e}$ is a vector of beamformer weighting coefficients and $\mathbf{y}[t] \in \mathbb{C}^{n_e}$ is a vector that contains a spatial sample of the received signal at each element of the array. For the MVDR beamformer, the weighting coefficients are given by [36]

$$\mathbf{w}_{\text{MVDR}} = \frac{\hat{\mathbf{C}}_{yy}^{-1}\mathbf{a}}{\mathbf{a}^H \hat{\mathbf{C}}_{yy}^{-1}\mathbf{a}}, \tag{40}$$

where $\mathbf{a}$ is the array *steering vector* and $\hat{\mathbf{C}}_{yy}$ is a sample covariance matrix estimated from n_s *snapshots* using

$$\hat{\mathbf{C}}_{yy} = \frac{1}{n_s}\sum_{t=1}^{n_s} \mathbf{y}[t]\mathbf{y}[t]^H. \tag{41}$$

There are two problems associated with MVDR beamforming. The first problem is the invertibility of the matrix $\hat{\mathbf{C}}_{yy}$–the matrix obtained from a limited number of snapshots is usually ill-conditioned. A simple solution to this problem is *diagonal loading* of the sample covariance matrix [36]. The second problem is the uncertainty that occurs with the steering vector $\mathbf{a}$ such that it is not known precisely [33]. Several methods have been proposed to jointly solve the two problems associated with beamforming. These are typically referred to as *robust* beamforming methods. Here we propose the following method for robust beamforming using RLS. Based on (39) and (40), let us write the beamformer output for the snapshot at time t as

$$\begin{aligned}\mathbf{y}_{\text{BF}}[t] &= \frac{\hat{\mathbf{a}}^H \hat{\mathbf{C}}_{yy}^{-\frac{1}{2}} \hat{\mathbf{C}}_{yy}^{-\frac{1}{2}} \mathbf{y}[t]}{\mathbf{a}^H \hat{\mathbf{C}}_{yy}^{-\frac{1}{2}} \hat{\mathbf{C}}_{yy}^{-\frac{1}{2}} \mathbf{a}} = \frac{\hat{\mathbf{a}}^H \hat{\mathbf{C}}_{yy}^{-\frac{1}{2}} \hat{\mathbf{C}}_{yy}^{-\frac{1}{2}} \mathbf{y}[t]}{\mathbf{a}^H \hat{\mathbf{C}}_{yy}^{-\frac{1}{2}} \hat{\mathbf{C}}_{yy}^{-\frac{1}{2}} \mathbf{a}} \\ &= \frac{\mathbf{r}^H \mathbf{q}[t]}{\mathbf{r}^H \mathbf{r}},\end{aligned} \tag{42}$$

where $\mathbf{r} \triangleq \hat{\mathbf{C}}_{yy}^{-\frac{1}{2}}\mathbf{a}$ and $\mathbf{q}[t] \triangleq \hat{\mathbf{C}}_{yy}^{-\frac{1}{2}}\mathbf{y}[t]$. Based on these definitions we can think of the newly introduced variables, $\mathbf{r}$ and $\mathbf{q}$, as being respectively obtained from

the inverse of the linear models

$$\mathbf{a} = \hat{\mathbf{C}}_{yy}^{\frac{1}{2}}\mathbf{r}, \tag{43}$$

and

$$\mathbf{y}[t] = \hat{\mathbf{C}}_{yy}^{\frac{1}{2}}\mathbf{q}[t]. \tag{44}$$

Since the matrix $\hat{\mathbf{C}}_{yy}^{\frac{1}{2}}$ is ill-conditioned, direct inversion does not provide viable solutions. Therefore, we can apply a regularization method to obtain estimates of $\mathbf{r}$ and $\mathbf{q}$ based on the linear models (43) and (44). Given that $\mathbf{a}$ and $\mathbf{y}$ are noisy, each of the latter two models is equivalent to the linear model (1). In the proposed robust beamforming method, we apply the proposed regularization algorithm to estimate $\mathbf{r}$ and $\mathbf{q}$ given the models (43) and (44). The results are then substituted in (42). Note that in this case, since $\hat{\mathbf{C}}_{yy}$ is symmetric positive semi-definite, the *eigenvalue decomposition* $\hat{\mathbf{C}}_{yy} = \mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^H$ can be used instead of the SVD. Using (5) for $\mathbf{A} = \hat{\mathbf{C}}_{yy}^{\frac{1}{2}} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^H$, the beamformer output using RLS will take the form

$$y_{\text{BF-RLS}} = \frac{\mathbf{a}^H\mathbf{U}\left(\mathbf{\Sigma}^2+\gamma_r\mathbf{I}_n\right)^{-1}\left(\mathbf{\Sigma}^2+\gamma_q\mathbf{I}_n\right)^{-1}\mathbf{\Sigma}^2\mathbf{U}^H\mathbf{y}}{\mathbf{a}^H\mathbf{U}\left(\mathbf{\Sigma}^2+\gamma_r\mathbf{I}_n\right)^{-2}\mathbf{\Sigma}^2\mathbf{U}^H\mathbf{a}}, \tag{45}$$

where γ_r and γ_q are the regularization parameters pertaining to the linear systems (43) and (44), respectively. Equation (45) suggests that the weighting coefficients for the RLS approach to robust beamforming are given by

$$\mathbf{w}_{\text{BF-RLS}} = \frac{\mathbf{a}^H\mathbf{U}\left(\mathbf{\Sigma}^2+\gamma_r\mathbf{I}_n\right)^{-1}\left(\mathbf{\Sigma}^2+\gamma_q\mathbf{I}_n\right)^{-1}\mathbf{\Sigma}^2\mathbf{U}^H}{\mathbf{a}^H\mathbf{U}\left(\mathbf{\Sigma}^2+\gamma_r\mathbf{I}_n\right)^{-2}\mathbf{\Sigma}^2\mathbf{U}^H\mathbf{a}}. \tag{46}$$

To evaluate performance, the *signal-to-interference-and-noise ratio* (SINR) defined as [36]

$$\text{SINR} = \frac{\sigma_s^2|\mathbf{w}^H\mathbf{a}|^2}{\mathbf{w}^H\mathbf{C}_{i+n}\mathbf{w}}, \tag{47}$$

is usually used. In (47), σ_s^2 is the power of the signal of interest, and $\mathbf{C}_{i+n}$ is the covariance matrix of the signal made up of the interference and noise.

V. Numerical Results

The performance of the proposed BPR approach is evaluated via numerical simulations based on the two applications discussed in Section IV. In this section, we present the results obtained for each of these applications.

A. Signal Estimation Results

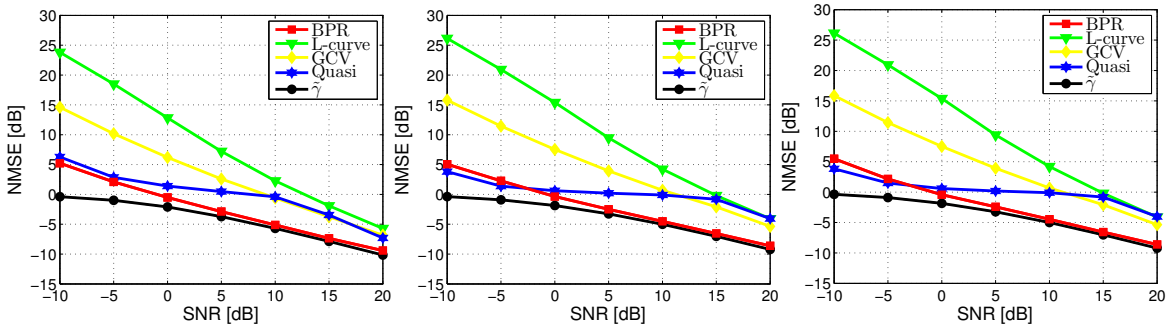
Signal estimation performance was evaluated in terms of the *normalized mean squared error* (NMSE), that is the MSE divided by $||\mathbf{x}||^2$. The NMSE was estimated from 10^4 simulation trials. Performance is presented as the NMSE versus the SNR (in dB) defined as $\text{SNR} = 10\log_{10}(||\mathbf{A}\mathbf{x}||_2^2/||\mathbf{z}||_2^2)$. The noise was assumed to be white Gaussian. The performance of the proposed BPR method is compared with other regularization parameter estimation methods. Namely, l-curve method [12], the GCV method [13] and the quasi-optimal method [14]. For reference purpose, we also present the (optimal) performance of the RLS estimator (7) for the optimal regularizer given in (16), which is equivalent to the LMMSE estimator when the elements of $\mathbf{x}$ are i.i.d. The LS estimator was excluded from the result plots on the grounds that it offers extremely poor performance that hinders good visualization of the other results.

Three different scenarios were identified from the literature [26], [27], [28]: (a) the elements of $\mathbf{x}$ and $\mathbf{A}$ are i.i.d. according to the standard complex Gaussian distribution; (b) $\mathbf{x}$ is i.i.d. Gaussian, while $\mathbf{A}$ is obtained by multiplying an i.i.d. Gaussian matrix with a left correlation matrix $\mathbf{R}_a^{\frac{1}{2}}$, where $\mathbf{R}_a[i,j] = \rho_a^{|i-j|}$, with $\rho_a = 0.5$; (c) $\mathbf{A}$ is the same as in (b), while $\mathbf{x}$ is obtained by (left) multiplication of an i.i.d. Gaussian vector with a correlation matrix $\mathbf{R}_x^{\frac{1}{2}}$, where $\mathbf{R}_x[i,j] = \rho_x^{|i-j|}$, with $\rho_x = 0.3$. Fig. 1 (a)–(c) shows performance comparison for these three different scenarios. From Fig. 1, it can be seen that the proposed BPR method significantly outperforms the three benchmark methods. The variation of the performance of various methods over the three test scenarios is quite small. In most cases, the proposed BPR method stays closest to the performance of the RLS estimator for the optimal value of γ, which is the LMMSE for scenario (a) and (b).

B. Robust Beamforming Results

To test the proposed regularization methods when applied to robust beamforming, we simulated a scenario with a linear array with 10 elements placed at half of the wavelength of the signal of interest and two interference signals. The *direction of arrival* (DOA) of each signal was generated from a uniform distribution in the interval $[-90^{\text{o}}, 90^{\text{o}}]$. The steering vector $\mathbf{a}$ was calculated from the true direction of arrival of the signal of interest plus an error which was uniformly distributed in the interval $[-5^{\text{o}}, 5^{\text{o}}]$. Randomizing both the angles and the error aimed to avoid scenarios that favour one algorithm or other. The SINR output of each algorithm was computed from 10^3 simulation trials. The proposed BPR algorithm is compared to the LCMV beamformer [37], [38], the response vector optimization LCMV (RVO LCMV) beamformer [35], the MVDR based robust adaptive beamformer (RAB MVDR) [30], and the robust adaptive beamformer based on semi-definite programming (RAB SDP) [39]. In addition, we plot the optimal performance obtained using the true covariance matrix, and the performance of standard MVDR beamformer. We have also considered applying existing regularization methods to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2759201, IEEE Access



(a) $\mathbf{x} = \boldsymbol{\nu}, \mathbf{A} = \boldsymbol{\Phi}$ (b) $\mathbf{x} = \boldsymbol{\nu}, \mathbf{A} = \mathbf{R}_a^{\frac{1}{2}}\boldsymbol{\Phi}$ (c) $\mathbf{x} = \mathbf{R}_x^{\frac{1}{2}}\boldsymbol{\nu}, \mathbf{A} = \mathbf{R}_a^{\frac{1}{2}}\boldsymbol{\Phi}$

Fig. 1: Signal estimation performance comparison for three different scenarios. In all cases, both the vector $\boldsymbol{\nu}$ and the matrix $\boldsymbol{\Phi}$ are Gaussian with i.i.d. elements. The matrices $\mathbf{R}_x$ and $\mathbf{R}_a$ are correlation matrices.

solve the two linear least-squares problems involved in the robust beamforming problem as formulated in this paper (see Equation (43) and (44)). We present the results from the best performing regularization method, which is the quasi-optimal [14].

Fig. 2 (a) plots the output SINR versus the input SNR for a number of snapshots $n_s = 30$. It can be seen that the proposed method outperforms all the other methods and provides SINRs that are close to the optimal values. Existing regularization methods provide very poor performance, especially in the high SNR regime. The best among the tested method is the quasi optimal method, which offers relatively low output SINRs.

Fig. 2 (b) plots the output SINR versus the number of snapshots n_s for a fixed SNR of 20 dB. It is shown clearly that the proposed BPR method outperforms all the benchmark methods and, again, stays nearest to the optimal performance, with the closest rival being the RAB SDP method. Again, the best RLS methods, the quasi-optimal method, exhibits unsatisfactory performance.

VI. Conclusions

Bounded perturbation regularization (BPR) was introduced in this paper. In BPR, the linear model matrix is perturbed using a matrix with a bounded norm. Based on the perturbed model, a min-max formulation was shown to produce a closed-form solution identical to that of the standard ℓ_2 Tikhonov regularized least squares problem. Finding a suitable value of the regularization parameter was shown to be equivalent to setting the value of the norm bound of the perturbation matrix in a non-linear equation. A procedure that employs the mean squared error criterion was proposed that resulted in eliminating the perturbation norm bound from the equation. Newton's method was proposed as a suitable method to solve the resultant equation to obtain a near-optimal value of the regularization parameter. Conditions for convergence of the Newton's method to the desired solution were established. The proposed BPR method was compared with existing regularization parameter selection methods, and other relevant methods, over two selected applications. Numerical results show that the proposed BPR method stays closest to the optimal performance in all the tested scenarios.

Appendix A
Error analysis of (15) and (30)

The three approximations in (15) and (30) take the form

$$\begin{aligned}&\mathrm{tr}\left[\boldsymbol{\Sigma}^{2k}\left(\boldsymbol{\Sigma}^2+\gamma\mathbf{I}_n\right)^{-p}\mathbf{V}^H\mathbf{C}_{xx}\mathbf{V}\right]\\&\quad\approx\frac{1}{n}\mathrm{tr}\left[\boldsymbol{\Sigma}^{2k}\left(\boldsymbol{\Sigma}^2+\gamma\mathbf{I}_n\right)^{-p}\right]\mathrm{tr}\left[\mathbf{C}_{xx}\right].\end{aligned}\tag{A.1}$$

Let us define $\mathbf{D} \triangleq \boldsymbol{\Sigma}^{2k}\left(\boldsymbol{\Sigma}^2+\gamma\mathbf{I}_n\right)^{-p}$, which is a diagonal matrix with elements

$$d_i = \frac{\sigma_i^{2k}}{(\sigma_i^2+\gamma)^p}.\tag{A.2}$$

Using the pair of inequalities in [40] (Equation (5)), we can see that

$$\lambda_{\min}(\mathbf{C}_{xx})\mathrm{tr}\left(\mathbf{D}\right) \le \mathrm{tr}\left(\mathbf{D}\mathbf{V}^T\mathbf{C}_{xx}\mathbf{V}\right) \le \lambda_{\max}(\mathbf{C}_{xx})\mathrm{tr}\left(\mathbf{D}\right),\tag{A.3}$$

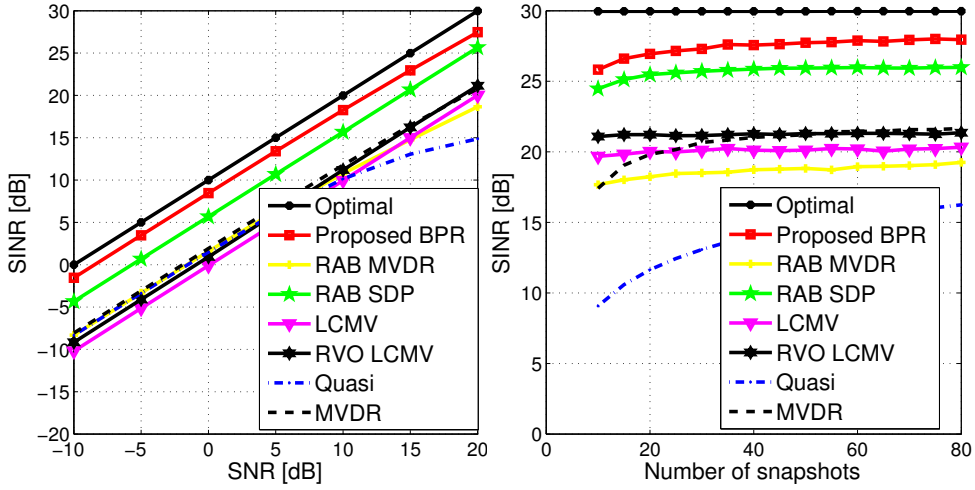
where $\lambda_{\min}$ and $\lambda_{\max}$ denote the minimum and maximum eigenvalues. Since $\mathbf{C}_{xx}$ is diagonal, $\lambda_{\min}$ and $\lambda_{\max}$ also represent the minimum and maximum diagonal elements, respectively. Similar to (A.3), we can write

$$\lambda_{\min}\left(\mathbf{D}\right)\mathrm{tr}\left(\mathbf{C}_{xx}\right) \le \mathrm{tr}\left(\mathbf{D}\mathbf{V}^T\mathbf{C}_{xx}\mathbf{V}\right) \le \lambda_{\max}\left(\mathbf{D}\right)\mathrm{tr}\left(\mathbf{C}_{xx}\right)\tag{A.4}$$

Now, let us define the normalized error in the approximation (A.1) as

$$\epsilon = \frac{\mathrm{tr}\left(\mathbf{D}\mathbf{V}^T\mathbf{C}_{xx}\mathbf{V}\right) - \frac{1}{n}\mathrm{tr}\left(\mathbf{D}\right)\mathrm{tr}\left(\mathbf{C}_{xx}\right)}{\frac{1}{n}\mathrm{tr}\left(\mathbf{D}\right)\mathrm{tr}\left(\mathbf{C}_{xx}\right)}.\tag{A.5}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2759201, IEEE Access



(a) (b)

Fig. 2: Robust beamforming performance: a) SINR versus input SNR; b) SINR versus number of signal snapshots.

Note that this represents the normalized error in using the quantity on the right-hand side of (A.1) to replace the quantity on the left-hand side. This way of defining the error is found more useful in carrying out the error analysis. We are interested in deriving a bound on $|\epsilon|$. To this end, we will derive a bound based on (A.3) and another bound using (A.4). Then we will combine the two bounds to form a single bound.

A. *Bound Based on The Inequality pair* (A.3)

Subtracting $\frac{1}{n}\text{tr}\,(\mathbf{D})\,\text{tr}\,(\mathbf{C}_{xx})$ from (A.3) and dividing by the same quantity, we obtain

$$\frac{\lambda_{\min}(\mathbf{C}_{xx})}{\lambda_{\text{avg}}(\mathbf{C}_{xx})} - 1 \leq \epsilon \leq \frac{\lambda_{\max}(\mathbf{C}_{xx})}{\lambda_{\text{avg}}(\mathbf{C}_{xx})} - 1, \qquad \text{(A.6)}$$

where $\lambda_{\text{avg}}(\mathbf{C}_{xx}) \triangleq \frac{1}{n}\text{tr}\,(\mathbf{C}_{xx})$. Hence $|\epsilon|$ can be bounded by a positive quantity according to

$$|\epsilon| \leq \mu_x = \max\left[1 - \frac{\lambda_{\min}(\mathbf{C}_{xx})}{\lambda_{\text{avg}}(\mathbf{C}_{xx})}, \frac{\lambda_{\max}(\mathbf{C}_{xx})}{\lambda_{\text{avg}}(\mathbf{C}_{xx})} - 1\right]. \qquad \text{(A.7)}$$

B. *Bound Based on The Inequality pair* (A.4)

Starting from (A.4) and applying a similar procedure to that used to derive (A.7), we obtain another bound

$$|\epsilon| \leq \mu_a = \max\left[1 - \frac{\lambda_{\min}(\mathbf{D})}{\lambda_{\text{avg}}(\mathbf{D})}, \frac{\lambda_{\max}(\mathbf{D})}{\lambda_{\text{avg}}(\mathbf{D})} - 1\right]. \qquad \text{(A.8)}$$

Using (A.2), (A.8) can be written as (A.9).

C. *Combined Bound*

Combining (A.7) and (A.9), we obtain the final bound

$$|\epsilon| \leq \mu = \min\left(\mu_x, \mu_a\right). \qquad \text{(A.10)}$$

D. *Evaluation*

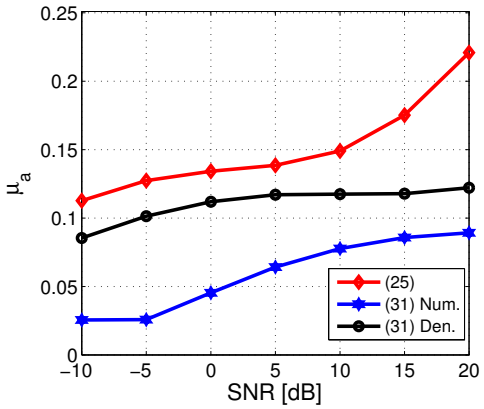
The value of μ_x depends only on the unknown covariance matrix of $\mathbf{x}$. For the i.i.d. case, we have $\mu_x = 0$. It is obvious that the closer the elements of $\mathbf{x}$ to being i.i.d., the smaller μ_x, and hence μ might also get smaller. On the other hand, the value of μ_a depends on the known matrix $\mathbf{A}$ and γ, which is unknown. To provide quantitative evaluation of μ_a, we exploit the fact that the optimal value of γ as in (16) is related to the SNR, and hence μ_a can be calculated for any SNR value of interest.

We evaluate μ_a for two different matrices over a range of SNR $\triangleq 10\log_{10}(||\mathbf{Ax}||^2/||\mathbf{z}||^2)$. For each SNR point, the value of γ needed to generate the matrix $\mathbf{D}$ defined in (A.2) is set equal to the optimal value as in (16). To simplify the analysis, and without loss of generality, the columns of $\mathbf{A}$ are normalized to have unit norm. In this case, the value of the optimal γ is exactly equal the SNR.

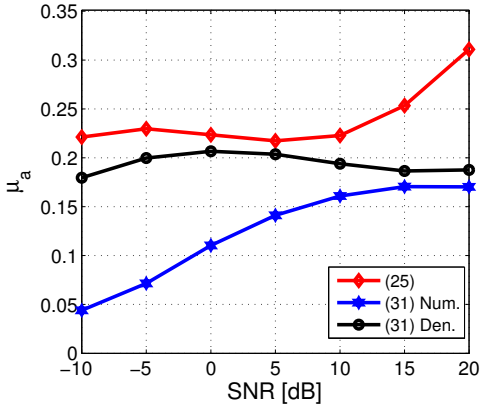
Fig. 3 plots μ_a, for the error in (15) and the numerator and denominator of (30), for the two different matrices used to generate the results in Fig. 1. We can see that μ_a, which is a somewhat pessimistic bound, stays within a range of reasonably small values for the SNR range of interest.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2759201, IEEE Access

$$\epsilon \leq \mu_a = \max\left[1 - \frac{\min\limits_{i\in\{1,\cdots,n\}} \left[\frac{\sigma_i^{2k}}{(\sigma_i^2+\gamma)^p}\right]}{\frac{1}{n}\sum_{i=1}^{n} \frac{\sigma_i^{2k}}{(\sigma_i^2+\gamma)^p}}, \frac{\max\limits_{i\in\{1,\cdots,n\}} \left[\frac{\sigma_i^{2k}}{(\sigma_i^2+\gamma)^p}\right]}{\frac{1}{n}\sum_{i=1}^{n} \frac{\sigma_i^{2k}}{(\sigma_i^2+\gamma)^p}} - 1\right] \tag{A.9}$$



(a) $\mathbf{A} = \mathbf{\Phi}$



(b) $\mathbf{A} = \mathbf{R}_a^{\frac{1}{2}}\mathbf{\Phi}$

Fig. 3: Error Bound for two different matrices. Both the vector $\boldsymbol{\nu}$ and the matrix $\mathbf{\Phi}$ are Gaussian with i.i.d. elements. The matrix $\mathbf{R}_a$ is the same correlation matrix used in Fig. 1.

Appendix B
Proof of Property 4 of $f(\gamma)$

First, let us write Equation (33) in the form

$$f(\gamma) = \left(\sum_{i=1}^{n} \frac{1}{\sigma_i^2+\gamma}\right)\left(\sum_{i=1}^{n} \frac{\sigma_i^2 b_i^2}{(\sigma_i^2+\gamma)^2}\right) - \left(\sum_{i=1}^{n} \frac{\sigma_i^2}{\sigma_i^2+\gamma}\right)\left(\sum_{i=1}^{n} \frac{b_i^2}{(\sigma_i^2+\gamma)^2}\right). \tag{B.1}$$

Equation (B.1) can be manipulated to the form

$$f(\gamma) = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \frac{\sigma_i^2-\sigma_j^2}{(\sigma_i^2+\gamma)(\sigma_j^2+\gamma)} \times \left(\frac{b_i^2}{\sigma_i^2+\gamma} - \frac{b_j^2}{\sigma_j^2+\gamma}\right), \tag{B.2}$$

which can be written as

$$f(\gamma) = \sum_{i=1}^{n-2}\sum_{j=i+1}^{n-1} \frac{\sigma_i^2-\sigma_j^2}{(\sigma_i^2+\gamma)(\sigma_j^2+\gamma)} \times \left(\frac{b_i^2}{\sigma_i^2+\gamma} - \frac{b_j^2}{\sigma_j^2+\gamma}\right) + \sum_{i=1}^{n-1} \frac{\sigma_i^2-\sigma_n^2}{(\sigma_i^2+\gamma)(\sigma_n^2+\gamma)} \times \left(\frac{b_i^2}{\sigma_i^2+\gamma} - \frac{b_n^2}{\sigma_n^2+\gamma}\right). \tag{B.3}$$

Taking the limit as $\gamma \to -\sigma_n^2$, we have

$$\lim_{\gamma\to-\sigma_n^2} f(\gamma) = \lim_{\gamma\to-\sigma_n^2} \sum_{i=1}^{n-2}\sum_{j=i+1}^{n-1} \frac{\sigma_i^2-\sigma_j^2}{(\sigma_i^2+\gamma)(\sigma_j^2+\gamma)} \times \left(\frac{b_i^2}{\sigma_i^2+\gamma} - \frac{b_j^2}{\sigma_j^2+\gamma}\right) + \lim_{\gamma\to-\sigma_n^2} \sum_{i=1}^{n-1} \frac{\sigma_i^2-\sigma_n^2}{(\sigma_i^2+\gamma)(\sigma_n^2+\gamma)} \times \left(\frac{b_i^2}{\sigma_i^2+\gamma} - \frac{b_n^2}{\sigma_n^2+\gamma}\right). \tag{B.4}$$

The first limit in the right-hand side of (B.4) is finite, say, equal to a value c_1. Accordingly, we can write:

$$\lim_{\gamma\to-\sigma_n^2} f(\gamma) = c_1 + \lim_{\gamma\to-\sigma_n^2} \sum_{i=1}^{n-1} \underbrace{\frac{\sigma_i^2-\sigma_n^2}{(\sigma_i^2+\gamma)(\sigma_n^2+\gamma)}}_{g_{in}(\gamma)} \times \left(\underbrace{\frac{b_i^2}{\sigma_i^2+\gamma}}_{q_i(\gamma)} - \underbrace{\frac{b_n^2}{\sigma_n^2+\gamma}}_{q_n(\gamma)}\right). \tag{B.5}$$

Using the fact that $\sigma_1^2 > \sigma_1^2 > \cdots > \sigma_n^2$, we can see that $g_{in}(\gamma = -\sigma_n^2)$ and $q_i(\gamma = -\sigma_n^2)$ are finite,

$\forall i = 1, \cdots, n-1$. On the other hand, $q_n(\gamma = -\sigma_n^2)$ goes to infinity as we take the limit. Consequently, the limit can be simplified to

$$\lim_{\gamma \to -\sigma_n^2} f(\gamma) = c_2 - \lim_{\gamma \to -\sigma_n^2} \sum_{i=1}^{n-1} \frac{\sigma_i^2 - \sigma_n^2}{(\sigma_i^2 + \gamma)(\sigma_n^2 + \gamma)} \times \left(\frac{b_n^2}{\sigma_n^2 + \gamma} \right). \tag{B.6}$$

From (B.6), it is easy to see that

$$\lim_{\gamma \to -\sigma_n^2} f(\gamma) = -\infty. \tag{B.7}$$

APPENDIX C
PROOF OF THEOREM 3

Proof. From Property 2–4, the function $f(\gamma)$ will take negative values somewhere in the interval $(-\sigma_n^2, +\infty)$. To guarantee that the function will have a root, based on the intermediate value theorem [24], the function must also take a positive value anywhere in the interval $(-\sigma_n^2, +\infty)$. Starting from Eq. (B.1), we have

$$f(\gamma) = \frac{1}{\gamma^3} \left(\sum_{i=1}^{n} \frac{1}{\frac{\sigma_i^2}{\gamma} + 1} \right) \left(\sum_{i=1}^{n} \frac{\sigma_i^2 b_i^2}{\left(\frac{\sigma_i^2}{\gamma} + 1 \right)^2} \right) - \frac{1}{\gamma^3} \left(\sum_{i=1}^{n} \frac{\sigma_i^2}{\frac{\sigma_i^2}{\gamma} + 1} \right) \left(\sum_{i=1}^{n} \frac{b_i^2}{\left(\frac{\sigma_i^2}{\gamma} + 1 \right)^2} \right). \tag{C.1}$$

Taking the limit of (C.1) as γ approaches $+\infty$, we obtain

$$\lim_{\gamma \to +\infty} f(\gamma) = \left(\lim_{\gamma \to +\infty} \frac{1}{\gamma^3} \right) \times \left\{ n \left(\sum_{i=1}^{n} \sigma_i^2 b_i^2 \right) - \left(\sum_{i=1}^{n} \sigma_i^2 \right) \left(\sum_{i=1}^{n} b_i^2 \right) \right\}. \tag{C.2}$$

It is obvious that the limit in (C.2) is equal to zero. Whether the function will approach zero from the positive or negative direction depends on the sign of the expression between the curly brackets. For the function to approach zero from the positive direction, the following condition must hold:

$$n \left(\sum_{i=1}^{n} \sigma_i^2 b_i^2 \right) > \left(\sum_{i=1}^{n} \sigma_i^2 \right) \left(\sum_{i=1}^{n} b_i^2 \right), \tag{C.3}$$

which is equivalent to (37). When (37) is satisfied, the function $f(\gamma)$ is guaranteed to take positive values in the interval of interest. Since the function also takes negative values (as per Property 4), the function will cross the abscissa at least once. □

REFERENCES

[1] T. Kailath, A.H. Sayed, and B. Hassibi, *Linear Estimation*, Prentice-Hall information and system sciences series. Prentice Hall, 1 edition, 2000.

[2] Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco, "On regularization algorithms in learning theory," *J. Complex.*, vol. 23, no. 1, pp. 52–72, Feb. 2007.

[3] S. M. Kay, *Fundamentals of Statistical Signal Processing*, Printice Hall, 1993.

[4] A. E. Hoerl and R.W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problem," *Technometrics*, vol. 12, pp. 55–67, Feb 1970.

[5] P. C. Hansen, "Regularization tools version 4.0 for matlab 7.3," *Numerical Algorithms*, vol. 46, pp. 189–194, Oct 2007.

[6] P. C. Hansen, *Discrete Inverse Problems*, Society for Industrial and Applied Mathematics, 2010.

[7] L. S. Mayer and T. A. Willke, "On biased estimation in linear models," *Technometrics*, vol. 15, pp. 497508, Aug 1973.

[8] Y.C. Eldar and A.V. Oppenheim, "Covariance shaping least-squares estimation," *IEEE Transactions on Signal Processing*, vol. 51, no. 3, pp. 686–697, Mar 2003.

[9] A. N. Tikhonov, "Solution of incorrectly formulated problems and the regularization method," *Soviet Math. Dokl.*, vol. 5, pp. 1035–1038, 1963.

[10] John B. Bell, "Reviewed work: Solutions of ill-posed problems. by A. N. Tikhonov, V. Y. Arsenin," *Mathematics of Computation*, vol. 32, no. 144, pp. 1320–1322, Oct 1978.

[11] W. Gander, "Least squares with a quadratic constraint," *Numerische Mathematik*, vol. 36, pp. 291–307, 1981.

[12] K. Miller, "Least squares methods for ill-posed problems with a prescribed bound," *SIAM Journal on Mathematical Analysis*, vol. 1, no. 1, pp. 52–74, 1970.

[13] G. Wahba, "Spline models for observational data," in *CBMS-NSF Regional Conference Series in Applied Mathematics*, 1990, vol. 59 of *SIAM, Philadelphia*.

[14] V. A. Morozov, *Methods for Solving Incorrectly Posed Problems*, Springer Verlag, New York, 1984.

[15] Frank Bauer and Mark A. Lukas, "Comparingparameter choice methods for regularization of ill-posed problems," *Mathematics and Computers in Simulation*, vol. 81, no. 9, pp. 1795 – 1841, 2011.

[16] Gene H. Golub and Charles F. Van Loan, *Matrix Computations (3rd Ed.)*, Johns Hopkins University Press, Baltimore, MD, USA, 1996.

[17] Lech Maligranda, "A simple proof of the hlder and the minkowski inequality," *The American Mathematical Monthly*, vol. 102, no. 3, pp. 256–259, 1995.

[18] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on Stochastic Programming*, Society for Industrial and Applied Mathematics, 2009.

[19] Y. C. Eldar, "Generalized sure for exponential families: Applications to regularization," *IEEE Transactions on Signal Processing*, vol. 57, no. 2, pp. 471–481, Feb 2009.

[20] W. James and C. Stein, *Estimation with Quadratic Loss*, Springer New York, New York, NY, 1992.

[21] David W. Kammler, "Chebyshev approximation of completely monotonic functions by sums of exponentials," *SIAM Journal on Numerical Analysis*, vol. 13, no. 5, pp. 761–774, 1976.

[22] David W. Kammler, "Least squares approximation of completely monotonic functions by sums of exponentials," *SIAM Journal on Numerical Analysis*, vol. 16, no. 5, pp. 801–818, 1979.

[23] Yuri K. Shestopaloff, *Sums of Exponential Functions and their New Fundamental Properties, with Applications to Natural Phenomena*, AKVY Press, Toronto, ON, 2008.

[24] Tom M. Apostol, *Calculus. Vol. I. , One-variable calculus, with an introduction to linear algebra*, J. Wiley & Sons, New York, 1967.

[25] C. Zarowski, *An Introduction to Numerical Analysis for Electrical and Computer Engineers*, John Wiley & Sons, 2004.

[26] H. Shin, M. Z. Win, J. H. Lee, and M. Chiani, "On the capacity of doubly correlated mimo channels," *IEEE Transactions on Wireless Communications*, vol. 5, no. 8, pp. 2253–2265, Aug 2006.

[27] A. M. Haimovich, R. S. Blum, and L. J. Cimini, "Mimo radar with widely separated antennas," *IEEE Signal Processing Magazine*, vol. 25, no. 1, pp. 116–129, 2008.

[28] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive mimo: Benefits and challenges," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 742–758, Oct 2014.

[29] H. Ruan and R. C. de Lamare, "Robust adaptive beamforming using a low-complexity shrinkage-based mismatch estimation algorithm," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 60–64, Jan 2014.

[30] A. Khabbazibasmenj, S. A. Vorobyov, and A. Hassanien, "Robust adaptive beamforming based on steering vector estimation with as little as possible prior information," *IEEE Transactions on Signal Processing*, vol. 60, no. 6, June 2012.

[31] R. G. Lorenz and S. P. Boyd, "Robust minimum variance beamforming," *IEEE Transactions on Signal Processing*, vol. 53, no. 5, pp. 1684–1696, May 2005.

[32] A. Mutapcic, S. J. Kim, and S. Boyd, "Beamforming with uncertain weights," *IEEE Signal Processing Letters*, vol. 14, no. 5, pp. 348–351, May 2007.

[33] S. A. Vorobyov, A. B. Gershman, and Zhi-Quan Luo, "Robust adaptive beamforming using worst-case performance optimization: a solution to the signal mismatch problem," *IEEE Transactions on Signal Processing*, vol. 51, no. 2, pp. 313–324, Feb 2003.

[34] S. A. Vorobyov, A. B. Gershman, Zhi-Quan Luo, and Ning Ma, "Adaptive beamforming with joint robustness against mismatched signal steering vector and interference nonstationarity," *IEEE Signal Processing Letters*, vol. 11, no. 2, pp. 108–111, Feb 2004.

[35] J. Xu, G. Liao, S. Zhu, and L. Huang, "Response vector constrained robust lcmv beamforming based on semidefinite programming," *IEEE Transactions on Signal Processing*, vol. 63, no. 21, pp. 5720–5732, Nov 2015.

[36] Jian Li, P. Stoica, and Zhisong Wang, "On robust capon beamforming and diagonal loading," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, April 2003, vol. 5, pp. V–337–40 vol.5.

[37] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935, Aug 1972.

[38] K. Takao, M. Fujita, and T. Nishi, "An adaptive antenna array under directional constraint," *IEEE Transactions on Antennas and Propagation*, vol. 24, no. 5, pp. 662–669, Sep 1976.

[39] Z. L. Yu, M. H. Er, and W. Ser, "A novel adaptive beamformer based on semidefinite programming (sdp) with magnitude response constraints," *IEEE Transactions on Antennas and Propagation*, vol. 56, no. 5, pp. 1297–1307, May 2008.

[40] Sheng-De Wang, Te-Son Kuo, and Chen-Fa Hsu, "Trace bounds on the solution of the algebraic matrix riccati and lyapunov equation," *IEEE Transactions on Automatic Control*, vol. 31, no. 7, pp. 654–656, Jul 1986.