Minimum Variance in Biased Estimation: Bounds and Asymptotically Optimal Estimators

Yonina C. Eldar, *Member, IEEE*

***Abstract*—We develop a uniform Cramér–Rao lower bound (UCRLB) on the total variance of any estimator of an unknown vector of parameters, with bias gradient matrix whose norm is bounded by a constant. We consider both the Frobenius norm and the spectral norm of the bias gradient matrix, leading to two corresponding lower bounds.**

We then develop optimal estimators that achieve these lower bounds. In the case in which the measurements are related to the unknown parameters through a linear Gaussian model, Tikhonov regularization is shown to achieve the UCRLB when the Frobenius norm is considered, and the shrunken estimator is shown to achieve the UCRLB when the spectral norm is considered. For more general models, the penalized maximum likelihood (PML) estimator with a suitable penalizing function is shown to *asymptotically* achieve the UCRLB. To establish the asymptotic optimality of the PML estimator, we first develop the asymptotic mean and variance of the PML estimator for any choice of penalizing function satisfying certain regularity constraints and then derive a general condition on the penalizing function under which the resulting PML estimator asymptotically achieves the UCRLB. This then implies that from all linear and nonlinear estimators with bias gradient whose norm is bounded by a constant, the proposed PML estimator asymptotically results in the smallest possible variance.

***Index Terms*—Asymptotic optimality, biased estimation, bias gradient norm, Cramér–Rao lower bound, penalized maximum likelihood, Tikhonov regularization.**

I. Introduction

ESTIMATION theory arises in a vast variety of areas in science and engineering including, for example, communication, economics, signal processing, seismology, and control. A common approach to developing well-behaved estimators in overparameterized estimation problems is to use regularization techniques, which where first systematically studied by Tikhonov [1], [2]. In general, regularization methods measure both the fit to the observed data and the physical plausibility of the estimate. In many cases, the use of regularization can reduce the variance of the resulting estimator at the expense of increasing the bias so that the design of such estimators is typically subject to a tradeoff between variance and bias.

Biased estimation methods are used extensively in a variety of different signal processing applications. Examples include regularization methods in image restoration [3], where the bias corresponds to spatial resolution, smoothing techniques in time series analysis [4], [5], and spectrum estimation [6], [7].

We consider the class of estimation problems in which we seek to estimate an unknown deterministic parameter vector $\mathbf{x}_0$ from some given measurements $\mathbf{y}$, where the relationship between $\mathbf{y}$ and $\mathbf{x}_0$ is described by the joint probability density function (pdf) $p(\mathbf{y};\mathbf{x}_0)$ of $\mathbf{y}$ characterized by $\mathbf{x}_0$.

It is well known that the total variance of any unbiased estimator of $\mathbf{x}_0$ is bounded by the Cramér–Rao lower bound (CRLB) [8]–[11]. In the case in which the measurements $\mathbf{y}$ are related to the unknowns $\mathbf{x}_0$ through a linear Gaussian model, the maximum likelihood (ML) estimate of $\mathbf{x}_0$, which is given by the value of $\mathbf{x}$ that maximizes $p(\mathbf{y};\mathbf{x})$, achieves the CRLB. Furthermore, when $\mathbf{x}_0$ is estimated from independent identically distributed (iid) measurements, under suitable regularity assumptions on the pdf $p(\mathbf{y};\mathbf{x}_0)$, the ML estimator is asymptotically unbiased and achieves the CRLB [9], [10], [12].

Since the estimators resulting from regularization methods are typically biased, their variance cannot be bounded by the CRLB. The total variance of any estimator with a given bias is bounded by the *biased CRLB* [13], which is an extension of the CRLB for unbiased estimators. It turns out that the biased CRLB does not depend directly on the bias but only on the bias gradient matrix, which makes intuitive sense. Indeed, any constant bias is removable, even if it is very large and, therefore, should not effect the performance of the estimator.

Given a desired bias gradient, the biased CRLB serves as a bound on the smallest attainable variance. However, in applications, it may not be obvious how to choose a particular bias gradient. In such cases, it would be useful to have a lower bound on the smallest attainable variance using any estimator whose bias gradient belongs to a suitable class. A bound of this form was first developed by Hero *et al.* [14], [15]. Specifically, they consider the problem of estimating a *scalar* function of a deterministic vector parameter. To quantify the fundamental tradeoff between bias and variance, they propose the *uniform CRLB (UCRLB)*, which is a bound on the smallest attainable variance that can be achieved using any estimator with bias gradient whose norm is bounded by a constant. In the case of a linear Gaussian model, they show that the UCRLB is achievable using a linear estimator. For a Poison model, the UCRLB is shown to be approximately achievable asymptotically. However, for more general models, the UCRLB is not shown to be achievable.

In this paper, we extend the results of [14] and [15] in two ways. First, we derive a UCRLB for vector parameters. Second,

Manuscript received December 15, 2002; revised August 26, 2003. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Xiaodong Wang.

The author is with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa, 32000, Israel (e-mail: yonina@ee.technion.ac.il).

Digital Object Identifier 10.1109/TSP.2004.828929

we develop a class of estimators that asymptotically achieve the UCRLB when estimating an unknown vector from iid vector measurements.

In the case in which it is desired to estimate an unknown vector $\mathbf{x}_0$, we may use the results in [14] and [15] to obtain bounds on the variance of the estimation error in each of the individual components to be estimated, subject to a constraint on the norm of the individual bias gradients. However, in many contexts, it is of interest to bound the total variance that is achievable in estimating the vector $\mathbf{x}_0$, subject to a constraint on the total bias gradient norm, rather than bounds on the individual variances subject to individual constraints. In order to obtain results on the total variance, in Sections III and IV, we extend the UCRLB to vector parameters. Specifically, we derive bounds on the total variance of any estimator $\hat{\mathbf{x}}$ of $\mathbf{x}_0$ with bias gradient matrix whose norm is bounded by a constant. We consider two different matrix norms which lead to two lower bounds; in Section III, we consider the Frobenius norm corresponding to an average bias gradient measure, and in Section IV, we consider the spectral norm corresponding to a worst-case bias gradient measure. As we show in Section II, these measures characterize the (possibly weighted) average and worst-case variation of the bias, respectively, over an ellipsoidal region around the true parameters $\mathbf{x}_0$.

In Sections III-A and V, we show that the estimator achieving the vector UCRLB can result in a smaller total variance than the estimator achieving the scalar UCRLB of [15] so that by treating the parameters to be estimated jointly, we can reduce the total variance in the estimation.

To establish the fact that the UCRLB is achievable, in Section V, we consider the case in which the measurements $\mathbf{y}$ are related to the unknown parameters $\mathbf{x}_0$ through a linear Gaussian model and derive linear estimators of $\mathbf{x}_0$ that achieve the UCRLB. In particular, we show that among all estimators with bias gradient matrix whose Frobenius norm is bounded by a constant, the ridge estimator proposed by Hoerl and Kennard [16] (also known as Tikhonov regularization [2]), with an appropriate regularization factor, minimizes the total variance. We also show that among all estimators with bias gradient matrix whose spectral norm is bounded by a constant, the shrunken estimator proposed by Mayer and Willke [17] with an appropriate shrinkage factor minimizes the total variance.

An important question is whether the UCRLB is achievable for more general, and not necessarily Gaussian, models. In Section VI, we consider the case of estimating $\mathbf{x}_0$ from iid measurements and develop a class of *penalized maximum likelihood (PML) estimators* that asymptotically achieve the UCRLB. Thus, we establish that asymptotically, the UCRLB is achievable in many cases.

The PML estimator was first proposed by Good and Gaskins [18], [19] as a modification of the ML estimator and is given by the value that maximizes a penalized likelihood function. This approach is equivalent to the maximum *a posteriori* (MAP) method in Baysian estimation if we interpret the penalizing factor as the log-likelihood of the prior pdf of $\mathbf{x}_0$. Note, however, that the analysis of the PML and MAP estimators is fundamentally different; whereas, in the Baysian approach, the unknown parameters are assumed to be random, in the PML approach, the unknown parameters are deterministic but unknown. Therefore, performance measures such as MSE average the performance over both the noise and the parameters in the Baysian approach, whereas in the PML approach, the performance is only averaged over the noise but not over the parameters, which are assumed to be fixed.

The PML method has been widely used in many engineering applications; see, e.g., [20]–[24]. We may interpret the PML approach as a method for obtaining biased estimators where the tradeoff between variance and bias depends on the penalizing function. Although various penalizing functions have been proposed for a variety of problems, no general assertions of optimality properties for the various choices of the penalizing functions are known. A possible approach is to choose the penalizing function to achieve an optimal bias-variance tradeoff in some sense.

In Section VI, we consider estimation of a vector parameter from iid measurements and develop the asymptotic bias and covariance of any PML estimator with a penalizing function that satisfies certain regularity constraints. Using these asymptotic results, we develop a condition on the penalizing function such that the resulting PML estimator achieves the UCRLB. In Section VII, we consider an example illustrating the asymptotic optimality properties of the PML estimator.

In the sequel, we denote vectors in $\mathbb{C}^m$ (m arbitrary) by boldface lowercase letters and matrices in $\mathbb{C}^{n\times m}$ by boldface uppercase letters. $\mathbf{I}$ denotes the identity matrix of appropriate dimension, $(\cdot)^*$ denotes the Hermitian conjugate of the corresponding matrix, and $\hat{(\cdot)}$ denotes an estimated vector or matrix. The ith column of the matrix $\mathbf{D}$ is denoted by $[\mathbf{D}]_i$, the ijth element of $\mathbf{D}$ is denoted by $[\mathbf{D}]_{ij}$, and the ith component of a vector $\mathbf{x}$ is denoted by x_i. The true value of an unknown vector parameter $\mathbf{x}$ is denoted by $\mathbf{x}_0$, and the true value of an unknown scalar parameter x is denoted by x_0. $\partial f(\mathbf{x}_0)/\partial \mathbf{x}$ denotes the gradient of the function $f(\mathbf{x})$ evaluated at the point $\mathbf{x}_0$ and is a row vector with j elements equal to $\partial f(\mathbf{x}_0)/\partial x_j$. The gradient of a vector $\partial \mathbf{b}(\mathbf{x}_0)/\partial \mathbf{x}$ is a matrix, with ijth element equal to $\partial b_i(\mathbf{x}_0)/\partial x_j$, i.e., the derivative of the ith component of the vector $\mathbf{b}(\mathbf{x}_0)$ with respect to x_j. Using the notation in [25], $\overset{a}{\sim}$ denotes "asymptotically distributed according to," and $\mathcal{N}(\mu,\sigma^2)$ denotes the Gaussian distribution with mean μ and variance σ^2. The matrix inequality $\mathbf{A} \preceq \mathbf{B} (\mathbf{A} \prec \mathbf{B})$ means that $\mathbf{B}-\mathbf{A}$ is non-negative (postive) definite].

II. Biased Cramér–Rao Lower Bound

We consider the problem of estimating an unknown deterministic parameter vector $\mathbf{x}_0 \in \mathbb{C}^m$ from given measurements $\mathbf{y} \in \mathbb{C}^n$, where the relationship between $\mathbf{y}$ and $\mathbf{x}_0$ is described by the pdf $p(\mathbf{y};\mathbf{x}_0)$ of $\mathbf{y}$ and is characterized by $\mathbf{x}_0$.

Under suitable regularity conditions on $p(\mathbf{y};\mathbf{x})$ (see, e.g., [8], [10]), the covariance of any unbiased estimator $\hat{\mathbf{x}}$ of $\mathbf{x}_0$ is bounded by the CRLB. A similar bound is also given for the covariance of a biased estimator, which is known as the biased CRLB [13]. Specifically, let $\hat{\mathbf{x}}$ denote an arbitrary estimator of $\mathbf{x}_0$ with bias

$$\mathbf{b}(\mathbf{x}_0) = E(\hat{\mathbf{x}}) - \mathbf{x}_0 \tag{1}$$

and covariance

$$\mathbf{C}_{\hat{\mathbf{x}}} = E\left\{[\hat{\mathbf{x}} - E(\hat{\mathbf{x}})][\hat{\mathbf{x}} - E(\hat{\mathbf{x}})]^*\right\}. \tag{2}$$

Then, the covariance $\mathbf{C}_{\hat{\mathbf{x}}}$ must satisfy

$$\mathbf{C}_{\hat{\mathbf{x}}} \succeq (\mathbf{I}+\mathbf{D})\,\mathbf{J}^{-1}\,(\mathbf{I}+\mathbf{D})^* \triangleq \mathbf{C}(\mathbf{D}) \tag{3}$$

where $\mathbf{J}$ is the Fisher information matrix defined by

$$\mathbf{J} = E\left\{\left[\frac{\partial \log p(\mathbf{y};\mathbf{x}_0)}{\partial \mathbf{x}}\right]^*\left[\frac{\partial \log p(\mathbf{y};\mathbf{x}_0)}{\partial \mathbf{x}}\right]\right\} \tag{4}$$

and is assumed to be nonsingular,[1] and $\mathbf{D}$ is the bias gradient matrix defined by

$$\mathbf{D} = \frac{\partial \mathbf{b}(\mathbf{x}_0)}{\partial \mathbf{x}}. \tag{5}$$

For a given bias gradient $\mathbf{D}$, the total variance that is achievable using any linear or nonlinear estimator with this bias gradient is bounded below by $\mathrm{Tr}(\mathbf{C}(\mathbf{D}))$, where the total variance $\sum_{i=1}^m E\left\{[\hat{x}_i - E(\hat{x}_i)]^2\right\}$ is the sum of the variances in estimating the individual components of $\mathbf{x}_0$. Typically, in estimation problems, there are two conflicting objectives that we would like to minimize: We would like to choose an estimator $\hat{\mathbf{x}}$ to achieve the smallest possible total variance *and* the smallest possible bias. However, generally, minimizing the bias results in an increase in variance and *vice versa*. To quantify the best achievable performance of any estimator $\hat{\mathbf{x}}$ of $\mathbf{x}_0$ taking both the bias and the total variance into account, we choose to minimize the total variance

$$C(\mathbf{D}) = \mathrm{Tr}(\mathbf{C}(\mathbf{D})) = \mathrm{Tr}\left((\mathbf{I}+\mathbf{D})\,\mathbf{J}^{-1}\,(\mathbf{I}+\mathbf{D})^*\right) \tag{6}$$

subject to a constraint on the bias gradient matrix $\mathbf{D}$. Note that $\mathbf{D}$ is invariant to a constant bias term so that in effect, it characterizes the part of the bias that cannot be removed.

A. Bias Gradient Matrix

To develop a meaningful constraint on $\mathbf{D}$, following [15], we first show that the norm of the bias gradient matrix is a measure of the sensitivity of the bias $\mathbf{b}(\mathbf{x})$ to changes in $\mathbf{x}$ over a neighborhood of $\mathbf{x}_0$.

Using a Taylor expansion, to the first-order approximation, we have that

$$\mathbf{b}(\mathbf{x}) - \mathbf{b}(\mathbf{x}_0) \approx \mathbf{D}(\mathbf{x}-\mathbf{x}_0) \triangleq \mathbf{D}\mathbf{u} \tag{7}$$

where $\mathbf{u} = \mathbf{x} - \mathbf{x}_0$. Therefore, the squared norm of the bias variation $\mathbf{b}(\mathbf{x}) - \mathbf{b}(\mathbf{x}_0)$ at a point $\mathbf{x}$ in the neighborhood of $\mathbf{x}_0$ is approximately given by

$$\|\mathbf{b}(\mathbf{x}) - \mathbf{b}(\mathbf{x}_0)\|^2 \approx \mathbf{u}^*\mathbf{D}^*\mathbf{D}\mathbf{u} \triangleq \mathcal{V}. \tag{8}$$

Let

$$\mathcal{S} = \{\mathbf{x} | (\mathbf{x}-\mathbf{x}_0)^*\mathbf{M}^{-1}(\mathbf{x}-\mathbf{x}_0) \le 1\} \tag{9}$$

be the set of vectors $\mathbf{x}$ that lie in the ellipsoidal region around $\mathbf{x}_0$ defined by $\mathbf{M}$, where $\mathbf{M}$ is an arbitrary positive definite weighting matrix. Then, the maximal variation of the bias norm over the region $\mathcal{S}$ is

$$\max_{\mathbf{u}\in\mathcal{S}} \mathcal{V} = \max_{\mathbf{z}^*\mathbf{z}\le 1} \mathbf{z}^*\mathbf{M}^{1/2}\mathbf{D}^*\mathbf{D}\mathbf{M}^{1/2}\mathbf{z} = \|\mathbf{D}\mathbf{M}^{1/2}\|^2 \tag{10}$$

where $\mathbf{z} = \mathbf{M}^{-1/2}\mathbf{u}$, and $\|\mathbf{A}\|$ denotes the spectral norm of the matrix $\mathbf{A}$ [26], i.e., the largest singular value of $\mathbf{A}$. The worst-case variation $\|\mathbf{D}\mathbf{M}^{1/2}\|^2$ occurs when $\mathbf{z}$ is chosen to be a unit-norm vector in the direction of the eigenvector corresponding to the largest eigenvalue of $\mathbf{M}^{1/2}\mathbf{D}^*\mathbf{D}\mathbf{M}^{1/2}$. It follows from (10) that the spectral norm $\|\mathbf{D}\mathbf{M}^{1/2}\|$ is approximately equal to the largest variation in the norm of the bias over the ellipsoid $\mathcal{S}$ and is therefore a reasonable worst-case bias measure.

To develop an average bias measure, instead of choosing $\mathbf{z}$ to be in the direction of the worst-case eigenvector, we may choose $\mathbf{z} = \sum_{i=1}^m a_i\mathbf{v}_i$, where $\mathbf{v}_i$, $1 \le i \le m$ are the eigenvectors of $\mathbf{M}^{1/2}\mathbf{D}^*\mathbf{D}\mathbf{M}^{1/2}$, and a_i are arbitrary coefficients satisfying $\sum_{i=1}^m a_i^2 = 1$ so that $\|\mathbf{z}\| = 1$. For this choice of $\mathbf{z}$

$$\mathcal{V} = \mathbf{z}^*\mathbf{M}^{1/2}\mathbf{D}^*\mathbf{D}\mathbf{M}^{1/2}\mathbf{z} = \sum_{i=1}^m a_i^2\lambda_i \tag{11}$$

where λ_i are the eigenvalues of $\mathbf{M}^{1/2}\mathbf{D}^*\mathbf{D}\mathbf{M}^{1/2}$. Denoting by $\mathbf{A}$ the diagonal matrix with diagonal elements a_i^2, we can express $\mathcal{V}$ of (11) as

$$\mathcal{V} = \mathrm{Tr}\left(\mathbf{V}\mathbf{A}\mathbf{V}^*\mathbf{M}^{1/2}\mathbf{D}^*\mathbf{D}\mathbf{M}^{1/2}\right) = \mathrm{Tr}\left(\mathbf{D}^*\mathbf{D}\mathbf{Q}\right) \tag{12}$$

where $\mathbf{V}$ is the matrix of eigenvectors $\mathbf{v}_i$, and $\mathbf{Q} = \mathbf{M}^{1/2}\mathbf{V}\mathbf{A}\mathbf{V}^*\mathbf{M}^{1/2}$. If follows from (12) that the weighted Frobenius norm $\mathrm{Tr}(\mathbf{D}^*\mathbf{D}\mathbf{Q})$ of $\mathbf{D}$ is a measure of the average variation in the norm of the bias over the ellipsoid $\mathcal{S}$ and is therefore a reasonable average bias measure. More generally, we can consider the weighted Frobenius norm $\mathrm{Tr}(\mathbf{D}^*\mathbf{D}\mathbf{W})$ for an arbitrary non-negative definite matrix $\mathbf{W}$ as an average bias measure.

We conclude that the weighted spectral norm and the weighted Frobenius norm of $\mathbf{D}$ measure the worst-case and average variation, respectively, in the bias norm over an ellipsoidal region around $\mathbf{x}_0$ and therefore represent reasonable measures of bias. Motivated by these observations, in our development, we consider the following two measures of bias gradient: an average bias gradient measure corresponding to a weighted squared Frobenius norm

$$D_{\mathrm{AVG}} = \mathrm{Tr}\left(\mathbf{D}^*\mathbf{D}\mathbf{W}\right) \tag{13}$$

where now, $\mathbf{W}$ is an arbitrary non-negative definite weighting matrix, and a worst-case bias gradient measure corresponding to a weighted squared spectral norm

$$D_{\mathrm{WC}} = \max_{\mathbf{z}\in\mathbb{C}^m, \|\mathbf{z}\|=1} \mathbf{z}^*\mathbf{S}\mathbf{D}^*\mathbf{D}\mathbf{S}\mathbf{z} \tag{14}$$

for some non-negative definite matrix $\mathbf{S}$.

In Section III, we develop the UCRLB with an average bias constraint, and in Section IV, we develop the UCRLB with a worst-case bias constraint. Which bound to use in practice depends strongly on the specific application. For example, in

[1]This assumption is made to simplify the derivations.

the context of image restoration, the bias gradient norm can be viewed as a measure of the geometric resolution of the estimator [15], [27]. In applications, we may wish to constraint the average geometric resolution, in which case, the UCRLB with average bias constraint is appropriate, or we may wish to constraint the worst-case geometric resolution, in which case, the UCRLB with worst-case bias constraint should be considered.

III. UCRLB with Average Bias Constraint

We first consider the problem of minimizing $C(\mathbf{D})$ of (6) subject to

$$D_{\mathrm{AVG}} = \mathrm{Tr}\left(\mathbf{D}^*\mathbf{D}\mathbf{W}\right) \leq \gamma. \tag{15}$$

If $\gamma \geq \mathrm{Tr}(\mathbf{W})$, then we can choose $\mathbf{D} = -\mathbf{I}$, which results in $C(\mathbf{D}) = 0$.

We next consider the case $\gamma < \mathrm{Tr}(\mathbf{W})$. To find the optimal $\mathbf{D}$, we form the Lagrangian

$$L = \mathrm{Tr}\left((\mathbf{I}+\mathbf{D})\,\mathbf{J}^{-1}\,(\mathbf{I}+\mathbf{D})^*\right) + \alpha\left(\mathrm{Tr}\left(\mathbf{D}^*\mathbf{D}\mathbf{W}\right) - \gamma\right) \tag{16}$$

where from the Karush–Kuhn–Tucker (KKT) conditions [28], we must have $\alpha \leq 0$. Since L is strictly convex, it has a unique minimum, which can be determined by setting the derivative of L to 0.

Differentiating[2] L with respect to $\mathbf{D}$ and equating to 0

$$(\mathbf{I}+\mathbf{D})\mathbf{J}^{-1} + \alpha\mathbf{D}\mathbf{W} = \mathbf{0} \tag{17}$$

so that the minimum of L is given by $\mathbf{D} = \widehat{\mathbf{D}}_{\mathrm{AVG}}$ with

$$\begin{aligned}\widehat{\mathbf{D}}_{\mathrm{AVG}} &= -\,\mathbf{J}^{-1}\left(\mathbf{J}^{-1} + \alpha\mathbf{W}\right)^{-1} \\ &= (\mathbf{I}+\alpha\mathbf{W}\mathbf{J})^{-1} \\ &= -\,\mathbf{I} + \alpha\,(\mathbf{I}+\alpha\mathbf{W}\mathbf{J})^{-1}\,\mathbf{W}\mathbf{J}\end{aligned} \tag{18}$$

where we used the Matrix Inversion Lemma [26].

If $\alpha = 0$, then $\widehat{\mathbf{D}}_{\mathrm{AVG}} = -\mathbf{I}$, which violates the constraint (15). Therefore, $\alpha > 0$, which, from the KKT conditions, implies that (15) must be satisfied with equality. Thus, the optimal $\mathbf{D}$ is $\mathbf{D} = \widehat{\mathbf{D}}_{\mathrm{AVG}}$ given by (18), where $\alpha > 0$ is chosen such that

$$\begin{aligned}&\mathrm{Tr}\left(\widehat{\mathbf{D}}^*_{\mathrm{AVG}}\widehat{\mathbf{D}}_{\mathrm{AVG}}\mathbf{W}\right) \\ &\qquad = \mathrm{Tr}((\mathbf{I}+\alpha\mathbf{J}\mathbf{W})^{-1}(\mathbf{I}+\alpha\mathbf{W}\mathbf{J})^{-1}\mathbf{W}) = \gamma.\end{aligned} \tag{19}$$

If $\mathbf{W}$ is positive definite, then

$$\mathrm{Tr}(\widehat{\mathbf{D}}^*_{\mathrm{AVG}}\widehat{\mathbf{D}}_{\mathrm{AVG}}\mathbf{W}) = \mathrm{Tr}((\mathbf{W}^{-1}+\alpha\mathbf{J})^{-2}\mathbf{W}^{-1}) \tag{20}$$

so that $\alpha > 0$ is chosen such that

$$\mathrm{Tr}((\mathbf{W}^{-1}+\alpha\mathbf{J})^{-2}\mathbf{W}^{-1}) = \gamma. \tag{21}$$

We now show that there is a unique $\alpha > 0$ satisfying (19). To this end, let

$$T(\alpha) = \mathrm{Tr}((\mathbf{I}+\alpha\mathbf{J}\mathbf{W})^{-1}(\mathbf{I}+\alpha\mathbf{W}\mathbf{J})^{-1}\mathbf{W}) - \gamma \tag{22}$$

[2]In our derivations, we use the following derivative: For any Hermitian $\mathbf{A}$, $\partial\mathrm{Tr}(\mathbf{B}\mathbf{A}\mathbf{B}^*)/\partial\mathbf{B} = 2\mathbf{B}\mathbf{A}$.

so that any α satisfying (19) is a root of $T(\alpha)$. We can immediately verify that $T(\alpha)$ is monotonically decreasing in α. Since $T(0) = \mathrm{Tr}(\mathbf{W}) - \gamma > 0$ and $T(\alpha) \to -\gamma < 0$ for $\alpha \to \infty$, there exists exactly one $\alpha > 0$ for which $T(\alpha) = 0$.

We conclude that the total variance of any estimator $\hat{\mathbf{x}}$ of $\mathbf{x}_0$ with bias gradient $\mathbf{D}$ satisfying (15) with $\gamma < \mathrm{Tr}(\mathbf{W})$ is bounded by

$$\begin{aligned}\mathrm{Tr}(\mathbf{C}_{\hat{\mathbf{x}}}) &\geq \mathrm{Tr}\left(\left(\mathbf{I}+\widehat{\mathbf{D}}_{\mathrm{AVG}}\right)\mathbf{J}^{-1}\left(\mathbf{I}+\widehat{\mathbf{D}}_{\mathrm{AVG}}\right)^*\right) \\ &= \alpha^2\mathrm{Tr}((\mathbf{I}+\alpha\mathbf{W}\mathbf{J})^{-1}\mathbf{W}\mathbf{J}\mathbf{W}(\mathbf{I}+\alpha\mathbf{J}\mathbf{W})^{-1})\end{aligned} \tag{23}$$

where $\alpha > 0$ is given by (19).

If $\mathbf{W}$ is positive definite, then

$$\mathrm{Tr}(\mathbf{C}_{\hat{\mathbf{x}}}) \geq \alpha^2\mathrm{Tr}((\mathbf{W}^{-1}+\alpha\mathbf{J})^{-2}\mathbf{J}) \tag{24}$$

where $\alpha > 0$ is given by (21).

We summarize our results in the following theorem.

Theorem 1: Let $\mathbf{x}_0$ denote an unknown deterministic parameter vector, let $\mathbf{y}$ denote measurements of $\mathbf{x}_0$, and let $p(\mathbf{y};\mathbf{x}_0)$ denote the pdf of $\mathbf{y}$ characterized by $\mathbf{x}_0$. Let $\mathbf{J}$ denote the Fisher information matrix and $\mathbf{D}$ denote the bias gradient matrix defined by (4) and (5), respectively, and let $\mathbf{W}$ be a non-negative Hermitian weighting matrix. Then, the total variance $C = C(\mathbf{D})$ defined by (6) of any estimator of $\mathbf{x}_0$ with bias gradient matrix $\mathbf{D}$ such that $\mathrm{Tr}(\mathbf{D}^*\mathbf{D}\mathbf{W}) \leq \gamma < \mathrm{Tr}(\mathbf{W})$ satisfies

$$C \geq \alpha^2\mathrm{Tr}((\mathbf{I}+\alpha\mathbf{W}\mathbf{J})^{-1}\mathbf{W}\mathbf{J}\mathbf{W}(\mathbf{I}+\alpha\mathbf{J}\mathbf{W})^{-1})$$

where $\alpha > 0$ is chosen such that

$$\mathrm{Tr}((\mathbf{I}+\alpha\mathbf{W}\mathbf{J})^{-1}\mathbf{W}(\mathbf{I}+\alpha\mathbf{J}\mathbf{W})^{-1}) = \gamma.$$

If, in addition, $\mathbf{W}$ is positive definite, then

$$C \geq \alpha^2\mathrm{Tr}((\mathbf{W}^{-1}+\alpha\mathbf{J})^{-2}\mathbf{J})$$

where $\alpha > 0$ is chosen such that

$$\mathrm{Tr}((\mathbf{W}^{-1}+\alpha\mathbf{J})^{-2}\mathbf{W}^{-1}) = \gamma.$$

A. Comparison with the Scalar UCRLB

In Section II, we developed a lower bound on the total variance attainable using an arbitrary estimator $\hat{\mathbf{x}}$ of $\mathbf{x}_0$ with average bias gradient bounded by a constant by treating the unknowns to be estimated jointly. Alternatively, we can obtain a lower bound on the total variance by using the scalar UCRLB of Hero *et al.* [14], [15] to bound the variance in estimating each of the individual components of $\mathbf{x}_0$. We now show that in general, the UCRLB of Theorem 1 on the total variance is *lower* than the bound on the total variance resulting from the scalar UCRLB. This implies that if the UCRLB is achievable, as it is, for example, in the case of a linear Gaussian model (see Section V), then we can obtain a lower variance when estimating the parameters jointly subject to a joint constraint than by estimating each of the components individually subject to individual constraints.

To develop a lower bound on the total variance of any estimator $\hat{\mathbf{x}}$ of $\mathbf{x}_0$ with bias gradient Frobenius norm that is bounded by a constant γ using the scalar UCRLB, denote by

$b_i = E(\hat{x}_i) - x_i$ the bias in estimating the ith component of $\mathbf{x}_0$ and by $\mathbf{d}_i = \partial b_i(\mathbf{x}_0)/\partial \mathbf{x}$ the corresponding bias gradient. The scalar UCRLB minimizes

$$[\mathbf{C}(\mathbf{D})]_{ii} = (\mathbf{[I]}_i^* + \mathbf{d}_i)\,\mathbf{J}^{-1}\,(\mathbf{[I]}_i + \mathbf{d}_i^*) \tag{25}$$

for each $1 \le i \le m$, subject to the constraint that

$$\mathbf{d}_i \mathbf{W} \mathbf{d}_i^* \le \gamma_i \tag{26}$$

for some non-negative definite matrix $\mathbf{W}$, where $\sum_{i=1}^m \gamma_i = \gamma$. The total variance in estimating $\mathbf{x}_0$ using any estimator $\hat{\mathbf{x}}$ with bias gradient vectors satisfying (26) is then bounded by

$$\sum_{i=1}^m \min_{\mathbf{d}_i} \left\{ (\mathbf{[I]}_i^* + \mathbf{d}_i)\,\mathbf{J}^{-1}\,(\mathbf{[I]}_i + \mathbf{d}_i^*) \right\} \tag{27}$$

which can equivalently be expressed as

$$\begin{aligned} &\min_{\mathbf{D}} \left\{ \sum_{i=1}^m (\mathbf{[I]}_i^* + \mathbf{d}_i)\mathbf{J}^{-1}(\mathbf{[I]}_i + \mathbf{d}_i^*) \right\} \\ &\quad = \min_{\mathbf{D}} \left\{ \mathrm{Tr}\left((\mathbf{I}+\mathbf{D})\mathbf{J}^{-1}(\mathbf{I}+\mathbf{D})^* \right) \right\} \\ &\quad = \min_{\mathbf{D}} C(\mathbf{D}) \end{aligned} \tag{28}$$

where $\mathbf{D}$ is the matrix with rows $\mathbf{d}_i$, subject to

$$[\mathbf{D}\mathbf{W}\mathbf{D}^*]_{ii} \le \gamma_i, \quad 1 \le i \le m. \tag{29}$$

In contrast, the vector UCRLB is obtained by minimizing (28) subject to

$$\sum_{i=1}^m [\mathbf{D}\mathbf{W}\mathbf{D}^*]_{ii} \le \sum_{i=1}^m \gamma_i = \gamma. \tag{30}$$

Note that any matrix $\mathbf{D}$ satisfying (29) also satisfies (30); however, the reverse implication is not true. Therefore, we have immediately that the vector UCRLB is no larger than the bound on the total variance resulting from the scalar UCRLB. In the case in which the vector UCRLB is achievable, this implies that a lower total variance may be achieved by treating the parameters to be estimated jointly, as we demonstrate in the context of a concrete example in Section V.

IV. UCRLB with Worst-Case Bias Constraint

We now consider the problem of minimizing $C(\mathbf{D})$ of (6) subject to

$$D_{\mathrm{WC}} = \max_{\mathbf{z}\in\mathbb{C}^m, \|\mathbf{z}\|=1} \mathbf{z}^*\mathbf{S}\mathbf{D}^*\mathbf{D}\mathbf{S}\mathbf{z} \le \gamma \tag{31}$$

for some non-negative definite matrix $\mathbf{S}$. In Section IV-A, we consider the case in which $\mathbf{S}$ is a positive definite matrix that has the same eigenvector matrix as $\mathbf{J}$. As we will show, in this case, there is a closed-form solution for the optimal bias gradient matrix $\mathbf{D}$. In Section IV-B, we consider an arbitrary weighting $\mathbf{S}$. In this case, the optimal $\mathbf{D}$ can be found as a solution to a semidefinite programming problem (SDP) [29]–[31], which is a convex optimization problem that can be solved very efficiently, e.g., using interior point methods [31], [32].

A. UCRLB with $\mathbf{S}$ and $\mathbf{J}$ Jointly Diagonalizable

We first consider the problem of minimizing $C(\mathbf{D})$ of (6) subject to (31), where $\mathbf{S}$ is positive definite and is jointly diagonalizable with $\mathbf{J}$. Specifically, let $\mathbf{J}^{-1}$ have an eigendecomposition $\mathbf{J}^{-1} = \mathbf{Q}\Lambda\mathbf{Q}^*$, where $\mathbf{Q}$ is a unitary $m \times m$ matrix, and Λ is a diagonal matrix with diagonal elements $\lambda_i > 0$, and let $\mathbf{q}_i$, $1 \le i \le m$ denote the columns of $\mathbf{Q}$. Then, we assume that $\mathbf{S}$ has the form $\mathbf{S} = \sum_{i=1}^m \beta_i \mathbf{q}_i \mathbf{q}_i^*$ for some $\beta_i > 0$.

We first note that we can express (31) as

$$\mathbf{z}^*\mathbf{S}\mathbf{D}^*\mathbf{D}\mathbf{S}\mathbf{z} \le \gamma, \quad \mathbf{z} \in \mathbb{C}^m, \quad \mathbf{z}^*\mathbf{z} = 1. \tag{32}$$

If $\gamma \ge \lambda_{\max}^2$, where $\lambda_{\max} = \max_i \beta_i$ is the largest eigenvalue of $\mathbf{S}$, then we can choose $\mathbf{D} = -\mathbf{I}$, which results in $C(\mathbf{D}) = 0$.

We next consider the case in which $\gamma < \lambda_{\max}$. In (32), we have infinitely may constraints on the matrix $\mathbf{D}$ so that the problem is hard to solve. Instead, we first consider the simpler problem of minimizing $C(\mathbf{D})$ subject to a *finite subset* of the constraints (32), i.e., we consider (32) for a finite set of choices $\mathbf{z}$. With $\hat{C}$ and $\hat{C}'$ denoting the minimum attainable total variance subject to (32) and a subset of (32), respectively, we have immediately that $\hat{C} \ge \hat{C}'$. Thus, our approach is to first find the optimal $\mathbf{D}$ that achieves the minimum total variance $\hat{C}'$ and then show that this optimal $\mathbf{D}$ also satisfies (32) so that $\hat{C} = \hat{C}'$.

Thus, we now consider minimizing $C(\mathbf{D})$ subject to

$$\mathbf{q}_i^*\mathbf{S}\mathbf{D}^*\mathbf{D}\mathbf{S}\mathbf{q}_i \le \gamma, \quad 1 \le i \le m. \tag{33}$$

Since $\mathbf{S}\mathbf{q}_i = \beta_i \mathbf{q}_i$, the constraints (33) become

$$\beta_i^2 \mathbf{q}_i^*\mathbf{D}^*\mathbf{D}\mathbf{q}_i \le \gamma, \quad 1 \le i \le m. \tag{34}$$

To find the optimal $\mathbf{D}$, we form the Lagrangian

$$\mathrm{Tr}\left((\mathbf{I}+\mathbf{D})\,\mathbf{J}^{-1}\,(\mathbf{I}+\mathbf{D})^* \right) + \sum_{i=1}^m \alpha_i \beta_i^2 \left(\mathbf{q}_i^*\mathbf{D}^*\mathbf{D}\mathbf{q}_i - \gamma \right) \tag{35}$$

where, from the KKT conditions, $\alpha_i \ge 0$. Differentiating with respect to $\mathbf{D}$ and equating to 0

$$(\mathbf{I}+\mathbf{D})\mathbf{J}^{-1} + \sum_{i=1}^m \alpha_i \beta_i^2 \mathbf{D}\mathbf{q}_i\mathbf{q}_i^* = \mathbf{0} \tag{36}$$

so that

$$\begin{aligned} \mathbf{D} &= -\,\mathbf{J}^{-1} \left(\mathbf{J}^{-1} + \sum_{i=1}^m \alpha_i \beta_i^2 \mathbf{q}_i\mathbf{q}_i^* \right)^{-1} \\ &= -\sum_{i=1}^m \frac{\lambda_i}{\lambda_i + \alpha_i\beta_i^2} \mathbf{q}_i\mathbf{q}_i^*. \end{aligned} \tag{37}$$

Let $\mathcal{I}$ denote the set of indices for which $\beta_i^2 > \gamma$. Since $\gamma < \lambda_{\max}^2$, the set $\mathcal{I}$ is not empty. If $\alpha_j = 0$ for some $j \in \mathcal{I}$, then

$$\beta_j^2 \mathbf{q}_j^*\mathbf{D}^*\mathbf{D}\mathbf{q}_j = \beta_j^2 > \gamma \tag{38}$$

which violates the jth constraint of (34). Therefore, for all $i \in \mathcal{I}$, $\alpha_i > 0$, and (34) is satisfied with equality, which implies that

$$\frac{\lambda_i}{\lambda_i + \alpha_i\beta_i^2} = \frac{\sqrt{\gamma}}{\beta_i}, \quad i \in \mathcal{I}. \tag{39}$$

For $j \notin \mathcal{I}$, the choice $\alpha_j = 0$ does not violate the constraints (34) so that we may choose $\alpha_j = 0$ or $\alpha_j > 0$, which implies that $\lambda_j/(\lambda_j+\alpha_j\beta_j^2) = \sqrt{\gamma}/\beta_j$. We can immediately verify that $C(\mathbf{D})$ is minimized for $\alpha_j = 0$ so that

$$\alpha_i = 0, \quad i \notin \mathcal{I}. \tag{40}$$

Substituting (39) and (40) into (37)

$$\begin{aligned}\mathbf{D} &= -\sqrt{\gamma}\sum_{i\in\mathcal{I}}\frac{1}{\beta_i}\mathbf{q}_i\mathbf{q}_i^* - \sum_{i\notin\mathcal{I}}\mathbf{q}_i\mathbf{q}_i^* \\ &= -\sqrt{\gamma}\mathbf{S}^{-1}\mathbf{P} - (\mathbf{I}-\mathbf{P}) \\ &= \left(\mathbf{I}-\sqrt{\gamma}\mathbf{S}^{-1}\right)\mathbf{P} - \mathbf{I}\end{aligned} \tag{41}$$

where $\mathbf{P} = \sum_{i\in\mathcal{I}}\mathbf{q}_i\mathbf{q}_i^*$ is the orthogonal projection onto the space spanned by the eigenvectors of $\mathbf{S}$ corresponding to eigenvalues $\beta_i^2 > \gamma$.

We conclude that the optimal $\mathbf{D}$ that minimizes the total variance $C(\mathbf{D})$ subject to (33) is $\mathbf{D} = \widehat{\mathbf{D}}_{\mathrm{WC}}$, where

$$\widehat{\mathbf{D}}_{\mathrm{WC}} = \left(\mathbf{I}-\sqrt{\gamma}\mathbf{S}^{-1}\right)\mathbf{P} - \mathbf{I}. \tag{42}$$

For this choice of bias gradient

$$\mathbf{S}\widehat{\mathbf{D}}_{\mathrm{WC}}^*\widehat{\mathbf{D}}_{\mathrm{WC}}\mathbf{S} = \gamma\mathbf{P} + \mathbf{S}^2(\mathbf{I}-\mathbf{P}). \tag{43}$$

Since for $i \notin \mathcal{I}$, $\gamma \geq \beta_i^2$, we have that $\mathbf{S}^2(\mathbf{I}-\mathbf{P}) \leq \gamma(\mathbf{I}-\mathbf{P})$, and

$$\mathbf{S}\widehat{\mathbf{D}}_{\mathrm{WC}}^*\widehat{\mathbf{D}}_{\mathrm{WC}}\mathbf{S} \leq \gamma\mathbf{P} + \gamma(\mathbf{I}-\mathbf{P}) = \gamma\mathbf{I} \tag{44}$$

so that (32) is satisfied. Therefore, $\widehat{\mathbf{D}}_{\mathrm{WC}}$ also minimizes the total variance subject to (32).

Thus, the total variance of any estimator $\hat{\mathbf{x}}$ of $\mathbf{x}_0$ with bias gradient $\mathbf{D}$ satisfying (32) with $\gamma < \lambda_{\max}$ is bounded by

$$\begin{aligned}\mathrm{Tr}(\mathbf{C}_{\hat{\mathbf{x}}}) &\geq \mathrm{Tr}\left(\left(\mathbf{I}-\sqrt{\gamma}\mathbf{S}^{-1}\right)\mathbf{P}\mathbf{J}^{-1}\mathbf{P}\left(\mathbf{I}-\sqrt{\gamma}\mathbf{S}^{-1}\right)\right) \\ &= \mathrm{Tr}\left(\left(\mathbf{I}-\sqrt{\gamma}\mathbf{S}^{-1}\right)^2\mathbf{P}\mathbf{J}^{-1}\right)\end{aligned} \tag{45}$$

where we used the fact that $\mathbf{J}^{-1}$, $\mathbf{P}$, and $\mathbf{S}^{-1}$ all commute.

In the special case in which $\mathbf{S} = \mathbf{I}$, all the eigenvalues of $\mathbf{S}$, which are equal to 1, are larger than γ, which is constrained to be smaller than $\lambda_{\max}^2 = 1$. Thus, $\mathbf{P} = \mathbf{I}$, and

$$\mathrm{Tr}(\mathbf{C}_{\hat{\mathbf{x}}}) \geq \mathrm{Tr}\left((1-\sqrt{\gamma})^2\mathbf{J}^{-1}\right). \tag{46}$$

B. UCRLB with Arbitrary $\mathbf{S}$

We now consider the problem of minimizing $C(\mathbf{D})$ of (6) subject to (31) for an arbitrary non-negative definite matrix $\mathbf{S}$. This problem can equivalently be expressed as

$$\min_{t,\mathbf{D}} t \tag{47}$$

subject to

$$\mathrm{Tr}\left((\mathbf{I}+\mathbf{D})\mathbf{J}^{-1}(\mathbf{I}+\mathbf{D})^*\right) \leq t \tag{48}$$

$$\mathbf{S}^*\mathbf{D}^*\mathbf{D}\mathbf{S} \preceq \gamma\mathbf{I}. \tag{49}$$

If $\gamma \geq \lambda_{\max}^2$, where $\lambda_{\max}^2$ denotes the largest eigenvalue of $\mathbf{S}$, then we can choose $\mathbf{D} = -\mathbf{I}$, which results in $t = 0$. We next consider the case in which $\gamma < \lambda_{\max}^2$.

As we now show, the problem of (47) subject to (48) and (49) can be formulated as a standard SDP [29]–[31], which is the problem of minimizing a linear functional subject to linear matrix inequality (LMI) constraints, i.e., matrix constraints in which the matrices involved depend *linearly* on the unknowns to be optimized. By exploiting the many well-known algorithms for solving SDPs [29], [30], e.g., interior point methods[3] [31], [32], the optimal $\mathbf{D}$ can be computed very efficiently in polynomial time. In addition, SDP-based algorithms are guaranteed to converge to the global optimum.

To formulate our problem as an SDP, let $\mathbf{g} = \mathrm{vec}(\mathbf{J}^{-1/2}(\mathbf{I}+\mathbf{D})^*)$, where $\mathbf{m} = \mathrm{vec}(\mathbf{M})$ denotes the vector obtained by stacking the columns of $\mathbf{M}$. With this notation, our problem reduces to minimizing (47) subject to the constraints

$$\begin{aligned}\mathbf{g}^*\mathbf{g} &\leq t \\ \mathbf{S}^*\mathbf{D}^*\mathbf{D}\mathbf{S} &\preceq \gamma\mathbf{I}.\end{aligned} \tag{50}$$

The constraints (50) are not in the form of LMIs because of the terms $\mathbf{g}^*\mathbf{g}$ and $\mathbf{S}^*\mathbf{D}^*\mathbf{D}\mathbf{S}$ in which the elements of $\mathbf{D}$ do not appear linearly. To express these inequalities as LMIs, we rely on the following lemma [26, p. 472]:

Lemma 1 (Schur's Complement): Let

$$\mathbf{M} = \begin{bmatrix}\mathbf{A} & \mathbf{B}^* \\ \mathbf{B} & \mathbf{C}\end{bmatrix}$$

be a Hermitian matrix. Then, with $\mathbf{C} \succ 0$, $\mathbf{M} \succeq 0$ if and only if $\Delta_{\mathbf{C}} \succeq 0$, where $\Delta_{\mathbf{C}}$ is the Schur complement of $\mathbf{C}$ in $\mathbf{M}$ and is given by

$$\Delta_{\mathbf{C}} = \mathbf{A} - \mathbf{B}^*\mathbf{C}^{-1}\mathbf{B}.$$

Using Lemma 1, we can express the constraints (50) as

$$\begin{aligned}\begin{bmatrix}t & \mathbf{g}^* \\ \mathbf{g} & \mathbf{I}\end{bmatrix} &\succeq 0 \\ \begin{bmatrix}\gamma\mathbf{I} & \mathbf{S}\mathbf{D}^* \\ \mathbf{D}\mathbf{S} & \mathbf{I}\end{bmatrix} &\succeq 0\end{aligned} \tag{51}$$

which are LMIs in t and $\mathbf{D}$.

We conclude that the problem of minimizing $C(\mathbf{D})$ of (6) subject to (31) is equivalent to the SDP problem of (47) subject to (51).

We summarize our results in the following theorem.

Theorem 2: Let $\mathbf{x}_0$ denote an unknown deterministic parameter vector, let $\mathbf{y}$ denote measurements of $\mathbf{x}_0$, and let $p(\mathbf{y};\mathbf{x}_0)$ denote the pdf of $\mathbf{y}$ characterized by $\mathbf{x}_0$. Let $\mathbf{J}$ denote the Fisher information matrix and $\mathbf{D}$ denote the bias gradient matrix defined by (4) and (5), respectively, let $\mathbf{S}$ denote an arbitrary nonnegative definite matrix, and let $\lambda_{\max}$ denote the largest eigenvalue of $\mathbf{S}$. Then, the total variance $C = C(\mathbf{D})$

[3]Interior point methods are iterative algorithms that terminate once a prespecified accuracy has been reached. A worst-case analysis of interior point methods shows that the effort required to solve an SDP to a given accuracy grows no faster than a polynomial of the problem size. In practice, the algorithms behave much better than predicted by the worst-case analysis, and in fact, in many cases the number of iterations is almost constant in the size of the problem.

of any estimator of $\mathbf{x}_0$ with bias gradient matrix $\mathbf{D}$ such that $\max_{\mathbf{z}\in\mathbb{C}^m,\|\mathbf{z}\|=1} \mathbf{z}^*\mathbf{S}\mathbf{D}^*\mathbf{D}\mathbf{S}\mathbf{z} \le \gamma < \lambda_{\max}^2$ satisfies $C \ge C_{\min}$, where $C_{\min}$ is the solution to the semidefinite programming problem

$$C_{\min} = \min_{t,\mathbf{D}} t$$

subject to

$$\begin{bmatrix} t & \mathbf{g}^* \\ \mathbf{g} & \mathbf{I} \end{bmatrix} \succeq 0$$
$$\begin{bmatrix} \gamma\mathbf{I} & \mathbf{S}\mathbf{D}^* \\ \mathbf{D}\mathbf{S} & \mathbf{I} \end{bmatrix} \succeq 0$$

where $\mathbf{g} = \mathrm{vec}(\mathbf{J}^{-1/2}(\mathbf{I}+\mathbf{D})^*)$.

If $\mathbf{S} = \sum_{i=1}^m \beta_i \mathbf{q}_i\mathbf{q}_i^*$ for some $\beta_i > 0$, where $\mathbf{q}_i$ are the eigenvectors of $\mathbf{J}$, then

$$C_{\min} = \mathrm{Tr}\left(\left(\mathbf{I} - \sqrt{\gamma}\mathbf{S}^{-1}\right)^2 \mathbf{P}\mathbf{J}^{-1}\right)$$

where $\mathbf{P} = \sum_{i\in\mathcal{I}} \mathbf{q}_i\mathbf{q}_i^*$ is the orthogonal projection onto the space spanned by the eigenvectors of $\mathbf{S}$ corresponding to eigenvalues $\beta_j, j \in \mathcal{I}$, where $\mathcal{I}$ is the set indices for which $\beta_j^2 > \gamma$. If, in addition, $\mathbf{S} = \mathbf{I}$, then

$$C_{\min} = \mathrm{Tr}\left((1-\sqrt{\gamma})^2\mathbf{J}^{-1}\right).$$

Note from Theorems 1 and 2 that as we expect, the two UCRLB bounds coincide for the scalar case.

Theorems 1 and 2 characterize the smallest possible total variance of any estimator with bias gradient matrix whose norm is bounded by a constant. However, the theorems do not guarantee that there exists estimators achieving these lower bounds. In Section V, we show that for the case of a linear Gaussian model, both lower bounds are achievable using a linear estimator. In Section VI, we consider more general, not necessarily Gaussian models, and develop a class of estimators that *asymptotically* achieve the UCRLB.

V. Optimal Estimators for the Linear Gaussian Model

We now consider the class of estimation problems represented by the linear model

$$\mathbf{y} = \mathbf{H}\mathbf{x}_0 + \mathbf{n} \tag{52}$$

where $\mathbf{x}_0 \in \mathbb{C}^m$ is a deterministic vector of unknown parameters, $\mathbf{H}$ is a known $n \times m$ matrix with rank m, and $\mathbf{n} \in \mathbb{C}^n$ is a zero-mean Gaussian random vector with positive definite covariance $\mathbf{C}_n$.

For the model (52), the Fisher information matrix is given by [25]

$$\mathbf{J} = \mathbf{H}^*\mathbf{C}_n^{-1}\mathbf{H}. \tag{53}$$

Let $\widehat{\mathbf{D}}$ denote the optimal gradient bias that minimizes $C(\mathbf{D})$ subject to (15) or (32) so that $\widehat{\mathbf{D}}$ is given by (18) or (42) with $\mathbf{J}$ given by (53). Then, the total variance of any linear or nonlinear estimator $\hat{\mathbf{x}}$ of $\mathbf{x}_0$ is bounded by

$$\mathrm{Tr}(\mathbf{C}_{\hat{\mathbf{x}}}) \ge \mathrm{Tr}\left((\mathbf{I}+\widehat{\mathbf{D}})\left(\mathbf{H}^*\mathbf{C}_n^{-1}\mathbf{H}\right)^{-1}(\mathbf{I}+\widehat{\mathbf{D}})^*\right). \tag{54}$$

We now derive a linear estimator $\hat{\mathbf{x}} = \mathbf{G}\mathbf{y}$ of $\mathbf{x}_0$ that achieves the bound (54). Let

$$\mathbf{G} = (\mathbf{I}+\widehat{\mathbf{D}})(\mathbf{H}^*\mathbf{C}_n^{-1}\mathbf{H})^{-1}\mathbf{H}^*\mathbf{C}_n^{-1}. \tag{55}$$

The bias of this estimator is $\mathbf{b} = (\mathbf{G}\mathbf{H} - \mathbf{I})\mathbf{x}_0$ so that the bias gradient matrix is

$$\mathbf{D} = \mathbf{G}\mathbf{H} - \mathbf{I} = \widehat{\mathbf{D}} \tag{56}$$

and therefore satisfies (15) or (32). The total variance of $\hat{\mathbf{x}} = \mathbf{G}\mathbf{y}$ is

$$\begin{aligned}\mathrm{Tr}(\mathbf{C}_{\hat{\mathbf{x}}}) &= \mathrm{Tr}(\mathbf{G}\mathbf{C}_n\mathbf{G}^*) \\ &= \mathrm{Tr}\left((\mathbf{I}+\widehat{\mathbf{D}})(\mathbf{H}^*\mathbf{C}_n^{-1}\mathbf{H})^{-1}(\mathbf{I}+\widehat{\mathbf{D}})^*\right)\end{aligned} \tag{57}$$

so that this estimator achieves the lower bound (54).

Note that from (55)–(57), it follows that the estimator of the form

$$\mathbf{G} = (\mathbf{I}+\mathbf{D})\left(\mathbf{H}^*\mathbf{C}_n^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^*\mathbf{C}_n^{-1} \tag{58}$$

achieves the biased CRLB for estimators with bias gradient $\mathbf{D}$. Thus, in the case of a linear Gaussian model, the biased CRLB is always achieved by a linear estimator.

We conclude that among all estimators with bias gradient $\mathbf{D}$ satisfying $\mathrm{Tr}(\mathbf{D}^*\mathbf{D}\mathbf{W}) \le \gamma < \mathrm{Tr}(\mathbf{W})$ for some non-negative Hermitian matrix $\mathbf{W}$, the estimator that results in the smallest possible total variance is $\hat{\mathbf{x}} = \mathbf{G}\mathbf{y}$, where $\mathbf{G}$ is given by (55) with $\widehat{\mathbf{D}} = \widehat{\mathbf{D}}_{\mathrm{AVG}}$. Thus

$$\hat{\mathbf{x}} = \begin{cases} (\mathbf{W}\mathbf{H}^*\mathbf{C}_n^{-1}\mathbf{H} + \delta\mathbf{I})^{-1}\mathbf{W}\mathbf{H}^*\mathbf{C}_n^{-1}\mathbf{y}, & 0 \le \gamma < \mathrm{Tr}(\mathbf{W}) \\ 0, & \gamma \ge \mathrm{Tr}(\mathbf{W}) \end{cases} \tag{59}$$

where the regularization parameter $\delta > 0$ is chosen such that $\mathrm{Tr}((\mathbf{I} + (1/\delta)\mathbf{W}\mathbf{H}^*\mathbf{C}_n^{-1}\mathbf{H})^{-1}\mathbf{W}(\mathbf{I} + (1/\delta)\mathbf{H}^*\mathbf{C}_n^{-1}\mathbf{H}\mathbf{W})^{-1}) = \gamma$.

In the case in which $\mathbf{W}$ is invertible, we have

$$\hat{\mathbf{x}} = \begin{cases} \left(\mathbf{H}^*\mathbf{C}_n^{-1}\mathbf{H} + \delta\mathbf{W}^{-1}\right)^{-1}\mathbf{H}^*\mathbf{C}_n^{-1}\mathbf{y}, & 0 \le \gamma < \mathrm{Tr}(\mathbf{W}) \\ 0, & \gamma \ge \mathrm{Tr}(\mathbf{W}) \end{cases} \tag{60}$$

where $\delta > 0$ is chosen such that

$$\mathrm{Tr}\left((\mathbf{W}^{-1} + (1/\delta)\mathbf{H}^*\mathbf{C}_n^{-1}\mathbf{H})^{-2}\mathbf{W}^{-1}\right) = \gamma.$$

The estimator $\hat{\mathbf{x}}$ of (60) is equal to the ridge estimator proposed by Hoerl and Kennard [16] (also known as Tikhonov regularization [2]) and is widely used for solving inverse problems [33] and ill-conditioned least-squares problems [34]. We therefore conclude that the ridge estimator has a strong optimal property: Among all linear and nonlinear estimators of $\mathbf{x}_0$ in the linear Gaussian model (52) with bounded average weighted bias gradient, the ridge estimator minimizes the total variance. A similar result was obtained in [15] for the scalar case.

It follows from our results that Tikhonov regularization also minimizes the total variance among all *linear* estimators with average bias gradient bounded by a constant for any noise distribution.

Similarly, among all estimators with bias gradient $\mathbf{D}$ satisfying $\mathbf{z}^*\mathbf{S}\mathbf{D}^*\mathbf{D}\mathbf{S}\mathbf{z} \le \gamma < \lambda_{\max}^2$ for all $\mathbf{z} \in \mathbb{C}^m$ such that $\mathbf{z}^*\mathbf{z} = 1$, where $\mathbf{S}$ is a positive definite matrix that commutes

with $\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}$ and with eigenvalues β_i, and $\lambda_{\max} = \max_i \beta_i$, the estimator that results in the smallest possible total variance is $\hat{\mathbf{x}} = \mathbf{G}\mathbf{y}$, where $\mathbf{G}$ is given by (55) with $\widehat{\mathbf{D}} = \widehat{\mathbf{D}}_{\mathrm{WC}}$. Thus, we have (61), shown at the bottom of the page, where $\mathbf{P}$ is an orthogonal projection onto the space spanned by the eigenvectors of $\mathbf{S}$ corresponding to eigenvalues $\beta_i^2 > \gamma$.

The estimator $\hat{\mathbf{x}}$ of (61) with $\mathbf{S} = \mathbf{I}$ is equal to the shrunken estimator proposed by Mayer and Willke [17], which is simply a scaled version of the least-squares estimator. We therefore conclude that the shrunken estimator also has a strong optimality property: Among all linear and nonlinear estimators of $\mathbf{x}_0$ in the linear Gaussian model (52) with bounded worst case bias gradient, the shrunken estimator minimizes the total variance. For more general choices of $\mathbf{S}$, the estimator of (61) can be viewed as a generalization of the shrunken estimator.

We note that the shrunken estimator of (61) also minimizes the worst-case bias gradient among all *linear* estimators in the case in which the noise vector $\mathbf{n}$ is not necessarily Gaussian.

A. Application to System Identification

We now compare the performance of the estimator achieving the UCRLB with an average bias constraint, and that of the estimator achieving the scalar UCRLB, in the context of a system identification problem.

Suppose we are given noisy measurements $y[k]$, $0 \le k \le n-1$ of a filtered signal, which is obtained by filtering a causal input sequence $u[k]$ with a length-n filter with unknown impulse response $h[k]$. Thus

$$\begin{aligned} y[k] &= h[k] * u[k] + \eta[k] \\ &= \sum_{m=0}^{n-1} h[m]u[k-m] + \eta[k], \quad 0 \le k \le n-1 \end{aligned} \tag{62}$$

where $*$ denotes discrete-time convolution, and $\eta[k]$ is an iid Gaussian noise process with variance σ^2.

Denoting by $\mathbf{y}$, $\mathbf{x}_0$, and $\mathbf{n}$ the length-n vectors with components $y[k]$, $h[k]$, and $\eta[k]$, respectively, and defining

$$\mathbf{H} = \begin{bmatrix} u[0] & 0 & 0 & \cdots & 0 \\ u[1] & u[0] & 0 & \cdots & 0 \\ u[2] & u[1] & u[0] & \cdots & 0 \\ & & \ddots & \ddots & \\ u[n-1] & u[n-2] & \cdots & u[1] & u[0] \end{bmatrix} \tag{63}$$

we can express (62) in the form of a linear model

$$\mathbf{y} = \mathbf{H}\mathbf{x}_0 + \mathbf{n}. \tag{64}$$

Our problem then is to estimate $\mathbf{x}_0$ from the measurements $\mathbf{y}$.

Since (64) is a linear Gaussian model, the UCRLB is achievable using a linear estimator. In Fig. 1, we plot the minimal attainable total variance for any estimator with bias gradient matrix $\mathbf{D}$ satisfying $\mathrm{Tr}(\mathbf{D}^*\mathbf{D}) \le \gamma$, as a function of γ, for the case in which

$$\begin{aligned} &u[0] = 0.4, \quad u[1] = 0.6, \quad u[2] = 0.5, \quad u[3] = 0.6 \\ &u[4] = 0.2, \quad u[5] = 0.3. \end{aligned} \tag{65}$$

We also plot the total variance resulting from the Tikhonov estimator (60), which, in our case, reduces to

$$\hat{\mathbf{x}} = \alpha(\alpha\mathbf{H}^*\mathbf{H} + \sigma^2\mathbf{I})^{-1}\mathbf{H}^*\mathbf{y} \tag{66}$$

where α is chosen such that $\mathrm{Tr}((\mathbf{I} + \alpha/\sigma^2\mathbf{H}^*\mathbf{H})^{-2}) = \gamma$. The variance is computed by averaging the performance over 1000 noise realizations, where the true parameters are chosen as

$$\mathbf{x}_0 = [\,1 \quad 0.6 \quad 0.5 \quad 0.3 \quad 0.2 \quad 0.1\,]^* \tag{67}$$

and $\sigma^2 = 0.3$. As we expect, the Tikhonov estimator achieves the UCRLB for all values of γ.

For comparison, we also plot the total variance using the Tikhonov estimator that achieves the scalar UCRLB, which is given by [15]

$$\hat{x}_i = \alpha[(\alpha\mathbf{H}^*\mathbf{H} + \sigma^2\mathbf{I})]_i^{-1}\mathbf{H}^*\mathbf{y} \tag{68}$$

where α is chosen such that $[(\mathbf{I} + \alpha/\sigma^2\mathbf{H}^*\mathbf{H})^{-2}]_{ii} = \gamma/n$. The total variance in estimating $\mathbf{x}$ using the scalar UCRLB where the bias gradient norm of each of the components $\hat{x}_i$ of $\hat{\mathbf{x}}$ is bounded by γ/n is depicted by the dashed line in Fig. 1. We see from Fig. 1 that by treating the parameters $\mathbf{x}_0$ to be estimated jointly, we can improve the estimation performance over individual estimation of each of the components.

VI. Asymptotic Optimality of the PML Estimator

In general, there is no guarantee that an estimator exists that achieves the UCRLB. In Section V, we showed that in the case of a linear Gaussian model, there exists a linear estimator achieving the UCRLB. When the average bias is considered, the estimator takes on the form of Tikhonov regularization. It is well known that Tikhonov regularization also maximizes the penalized log-likelihood

$$\begin{aligned} \hat{\mathbf{x}} &= \arg\max \left\{ \log p(\mathbf{y};\mathbf{x}) - \frac{\beta}{2}\mathbf{x}^*\mathbf{W}\mathbf{x} \right\} \\ &= \arg\min \left\{ (\mathbf{y} - \mathbf{H}\mathbf{x})^*\mathbf{C}_n^{-1}(\mathbf{y} - \mathbf{H}\mathbf{x}) + \beta\mathbf{x}^*\mathbf{W}\mathbf{x} \right\} \end{aligned} \tag{69}$$

where $p(\mathbf{y};\mathbf{x})$ is a Gaussian distribution with mean $\mathbf{H}\mathbf{x}$ and covariance $\mathbf{C}_n$. When the worst-case bias is considered with weighting $\mathbf{S} = \mathbf{I}$, the shrunken estimator achieves the UCRLB. We can immediately verify that the shrunken estimator also maximizes (69), with $\mathbf{W} = -\mathbf{H}^*\mathbf{H}$. A similar result holds for the case in which $\mathbf{S}$ has the same eigenvector matrix as $\mathbf{J}$. Thus, we conclude that in the case of a linear Gaussian model, the

$$\hat{\mathbf{x}} = \begin{cases} (\mathbf{I} - \sqrt{\gamma}\mathbf{S}^{-1})\mathbf{P}\left(\mathbf{H}^*\mathbf{C}_n^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^*\mathbf{C}_n^{-1}\mathbf{y}, & 0 \le \gamma < \lambda_{\max}^2 \\ 0, & \gamma \ge \lambda_{\max}^2 \end{cases} \tag{61}$$

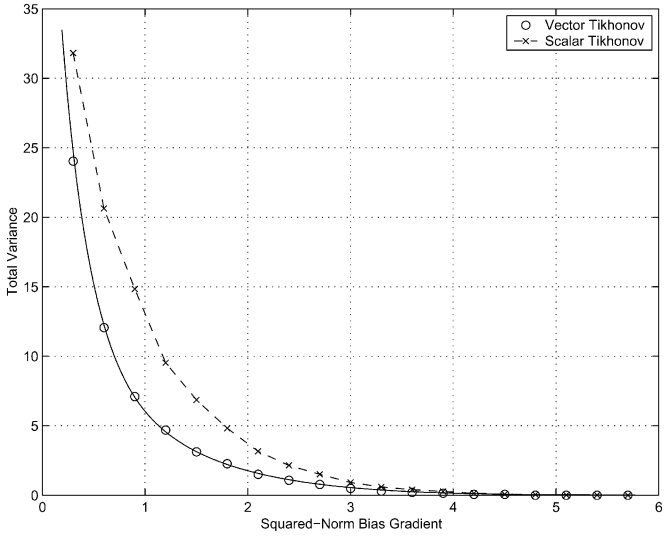

Fig. 1. Variance of the vector Tikhonov estimator (66) and the scalar Tikhonov estimator (68) as a function of the squared-norm bias gradient in comparison with the vector and scalar UCRLB. The line denotes the vector UCRLB, os denote the performance of the vector Tikhonov estimator, the dashed line denotes the scalar UCRLB, and the xs denote the scalar Tikhonov estimator.

PML estimator with an appropriate choice of penalizing function achieves the UCRLB.

In this section, we demonstrate that this optimality property of the PML estimator is more general. Specifically, we show that the PML estimator *asymptotically* achieves the UCRLB for many other statistical models. To this end, we first develop the asymptotic bias and variance of the PML estimator for a general class of penalizing functions. We then show that in many cases, we can choose the penalizing function such that the PML estimator asymptotically achieves the UCRLB.

A. PML Estimator

The PML estimator of $\mathbf{x}_0$, denoted $\hat{\mathbf{x}}^{\mathrm{PML}}$, is chosen to maximize the penalized log-likelihood function

$$\log p(\mathbf{y};\mathbf{x}) - \beta R(\mathbf{x}) \tag{70}$$

where $\beta > 0$ is a regularization parameter, and $R(\mathbf{x})$ is a penalizing function. The PML approach is equivalent to the maximum *a posteriori* method in Baysian estimation if we interpret $e^{-\beta R(\mathbf{x})}$ as the prior pdf of $\mathbf{x}_0$.

In the case in which we seek to estimate $\mathbf{x}_0$ from N iid (vector) measurements $\mathbf{y}_1,\ldots,\mathbf{y}_N$, $\hat{\mathbf{x}}^{\mathrm{PML}}$ is chosen to maximize

$$PL(x) = \sum_{i=1}^{N} \log p(\mathbf{y}_i;\mathbf{x}) - \beta_N R(\mathbf{x}) \tag{71}$$

where β_N is a regularization parameter that may depend on N.

Although many different choices of penalizing functions $R(\mathbf{x})$ have been proposed in the literature for various problems [20]–[24], no general assertions of optimality are known for these different choices.

In Section VI-B, we show that in many cases, the penalizing function $R(\mathbf{x})$ can be chosen such that the resulting PML estimator achieves the UCRLB. To this end, we first derive the asymptotic properties of the PML estimator. Specifically, we show that under certain regularity conditions, the PML estimator of $\mathbf{x}_0$ from N iid measurements is asymptotically Gaussian, and we derive explicit expressions for the asymptotic mean and variance.

B. Asymptotic Properties of the PML Estimator

Suppose we wish to estimate a vector $\mathbf{x}_0$ from N iid measurements $\mathbf{y}_1,\ldots,\mathbf{y}_N$. We consider the PML estimator $\hat{\mathbf{x}}^{\mathrm{PML}}$ that is chosen to maximize (71), where β_N is a parameter satisfying $\beta_N/N \to \beta_0$ for some constant β_0 as $N \to \infty$, and $R(\mathbf{x})$ is an arbitrary function of $\mathbf{x}$ such that $\partial^3 R(\mathbf{x})/\partial x_j \partial x_k \partial x_l$ is bounded for all j,k,l. To develop the asymptotic properties of the PML estimator, we make the following assumptions on the pdf $p(\mathbf{y};\mathbf{x})$:

Assumption 1: The derivatives $\partial \log p(\mathbf{y};\mathbf{x})/\partial x_j$, $\partial^2 \log p(\mathbf{y};\mathbf{x})/\partial x_j \partial x_k$ and $\partial^3 \log p(\mathbf{y};\mathbf{x})/\partial x_j \partial x_k \partial x_l$ exist for all j,k,l and $\mathbf{x} \in \mathcal{X}$, where $\mathcal{X}$ is an open interval including $\check{\mathbf{x}}$, with

$$\check{\mathbf{x}} = \arg\max \left\{ E\left\{\log p(\mathbf{y};\mathbf{x})\right\} - \beta_0 R(\mathbf{x}) \right\}. \tag{72}$$

Assumption 2: For each $\mathbf{x} \in \mathcal{X}$

$$\left| \frac{\partial^3 \log p(\mathbf{y};\mathbf{x})}{\partial x_j \partial x_k \partial x_l} \right| \le d(\mathbf{y}), \quad 1 \le j,k,l \le m \tag{73}$$

where $E_{\mathbf{x}}\{d(\mathbf{y})\} < \infty$ for all $\mathbf{x} \in \mathcal{X}$.

Assumption 3:

$$-E\left\{ \frac{\partial^2 \log p(\mathbf{y};\check{\mathbf{x}})}{\partial \mathbf{x}^2} \right\} + \beta_0 \frac{\partial^2 R(\mathbf{x})}{\partial \mathbf{x}^2} > 0. \tag{74}$$

Note that these assumptions are similar to the assumptions made on $p(\mathbf{y};\mathbf{x})$ in proving the asymptotic optimality of the ML estimator [10].

Under these assumptions, we have the following theorem.

Theorem 3: Let $\mathbf{x}_0$ denote an unknown deterministic parameter vector, let $\mathbf{y}_1,\ldots,\mathbf{y}_N$ denote N iid measurements of $\mathbf{x}_0$, and let $\hat{\mathbf{x}}^{\mathrm{PML}}$ denote the PML estimator of $\mathbf{x}_0$ from the measurements $\mathbf{y}_1,\ldots,\mathbf{y}_N$ that maximizes the penalized log-likelihood (71). Then, under Assumptions 1–3

$$\sqrt{N}(\hat{\mathbf{x}}^{\mathrm{PML}} - \check{\mathbf{x}}) \overset{a}{\sim} \mathcal{N}\left(0, (\mathbf{J}(\check{\mathbf{x}}) + \beta_0 \mathbf{M}(\check{\mathbf{x}}))^{-1}\ \mathbf{C}(\check{\mathbf{x}})\,(\mathbf{J}(\check{\mathbf{x}}) + \beta_0 \mathbf{M}(\check{\mathbf{x}}))^{-1}\right)$$

where $\beta_0 = \lim_{N\to\infty} \beta_N/N$

$$\check{\mathbf{x}} = \arg\max \left\{ E\left\{\log p(\mathbf{y};\mathbf{x})\right\} - \beta_0 R(\mathbf{x}) \right\}$$

$$\mathbf{C}(\check{\mathbf{x}}) = \mathrm{cov}\left\{ \frac{\partial \log p(\mathbf{y};\check{\mathbf{x}})}{\partial \mathbf{x}} \right\}$$

$$\mathbf{J}(\check{\mathbf{x}}) = -E\left\{ \frac{\partial^2 \log p(\mathbf{y};\check{\mathbf{x}})}{\partial \mathbf{x}^2} \right\}$$

and

$$\mathbf{M}(\check{\mathbf{x}}) = \frac{\partial^2 R(\check{\mathbf{x}})}{\partial \mathbf{x}^2}.$$

Proof: See Appendix A. □

C. PML Estimator and the UCRLB

From Theorem 3, the asymptotic total variance of $\hat{\mathbf{x}}^{\mathrm{PML}}$ is

$$\frac{1}{N}\mathrm{Tr}\left(\left(\mathbf{J}(\check{\mathbf{x}})+\beta_0\mathbf{M}(\check{\mathbf{x}})\right)^{-1}\mathbf{C}(\check{\mathbf{x}})\left(\mathbf{J}(\check{\mathbf{x}})+\beta_0\mathbf{M}(\check{\mathbf{x}})\right)\right) \quad (75)$$

and the asymptotic bias gradient $\mathbf{D}_{\mathrm{PML}}$ is

$$\mathbf{D}_{\mathrm{PML}}=\frac{\partial\check{\mathbf{x}}}{\partial\mathbf{x}_0}-\mathbf{I}. \quad (76)$$

To develop an expression for $\partial\check{\mathbf{x}}/\partial\mathbf{x}_0$, we note that from (72)

$$E\left\{\frac{\partial\log p(\mathbf{y};\check{\mathbf{x}})}{\partial\mathbf{x}}\right\}-\beta_0\frac{\partial R(\check{\mathbf{x}})}{\partial\mathbf{x}}=0. \quad (77)$$

Differentiating (77) with respect to $\mathbf{x}_0$

$$\left(E\left\{\frac{\partial^2\log p(\mathbf{y};\check{\mathbf{x}})}{\partial\mathbf{x}^2}\right\}-\beta_0\frac{\partial^2 R(\check{\mathbf{x}})}{\partial\mathbf{x}^2}\right)\frac{\partial\check{\mathbf{x}}}{\partial\mathbf{x}_0}+\frac{\partial}{\partial\mathbf{x}_0}E\left\{\frac{\partial\log p(\mathbf{y};\check{\mathbf{x}})}{\partial\mathbf{x}}\right\}=0 \quad (78)$$

or, equivalently

$$\left(\mathbf{J}(\check{\mathbf{x}})+\beta_0\mathbf{M}(\check{\mathbf{x}})\right)\frac{\partial\check{\mathbf{x}}}{\partial\mathbf{x}_0}=\frac{\partial}{\partial\mathbf{x}_0}E\left\{\frac{\partial\log p(\mathbf{y};\check{\mathbf{x}})}{\partial\mathbf{x}}\right\} \quad (79)$$

so that

$$\frac{\partial\check{\mathbf{x}}}{\partial\mathbf{x}_0}=\left(\mathbf{J}(\check{\mathbf{x}})+\beta_0\mathbf{M}(\check{\mathbf{x}})\right)^{-1}\frac{\partial}{\partial\mathbf{x}_0}E\left\{\frac{\partial\log p(\mathbf{y};\check{\mathbf{x}})}{\partial\mathbf{x}}\right\}. \quad (80)$$

With $\gamma=\mathbf{D}_{\mathrm{PML}}^{*}\mathbf{D}_{\mathrm{PML}}$, it follows from Theorem 1 that the total variance of any estimate of $\mathbf{x}_0$ with bias gradient $\mathbf{D}$ such that $\mathrm{Tr}(\mathbf{D}^{*}\mathbf{D})\le\mathrm{Tr}(\mathbf{D}_{\mathrm{PML}}^{*}\mathbf{D}_{\mathrm{PML}})$ satisfies

$$C\ge\frac{\alpha^2}{N}\mathrm{Tr}\left(\left(\mathbf{I}+\alpha\mathbf{J}_1\right)^{-2}\mathbf{J}_1\right) \quad (81)$$

where $\alpha>0$ is chosen such that

$$\mathrm{Tr}\left(\left(\mathbf{I}+\alpha\mathbf{J}_1\right)^{-2}\right)=\mathrm{Tr}\left(\left(\frac{\partial\check{\mathbf{x}}}{\partial\mathbf{x}_0}-\mathbf{I}\right)^{*}\left(\frac{\partial\check{\mathbf{x}}}{\partial\mathbf{x}_0}-\mathbf{I}\right)\right) \quad (82)$$

and

$$\mathbf{J}_1=E\left\{\left(\frac{\partial\log p(\mathbf{y}_1;\mathbf{x}_0)}{\partial\mathbf{x}}\right)^{*}\frac{\partial\log p(\mathbf{y}_1;\mathbf{x}_0)}{\partial\mathbf{x}}\right\} \quad (83)$$

is the Fisher information from a single observation. Therefore, if we can choose $R(\mathbf{x})$ such that

$$\mathrm{Tr}\left(\left(\mathbf{J}(\check{\mathbf{x}})+\beta_0\mathbf{M}(\check{\mathbf{x}})\right)^{-1}\mathbf{C}(\check{\mathbf{x}})\left(\mathbf{J}(\check{\mathbf{x}})+\beta_0\mathbf{M}(\check{\mathbf{x}})\right)^{-1}\right)=\alpha^2\mathrm{Tr}\left(\left(\mathbf{I}+\alpha\mathbf{J}_1\right)^{-2}\mathbf{J}_1\right) \quad (84)$$

where α is given by (82), with $\partial\check{\mathbf{x}}/\partial\mathbf{x}_0$ given by (80), then the corresponding PML estimator achieves the UCRLB with average bias constraint, so that asymptotically, there is no linear or non-linear estimator with bias gradient $\mathbf{D}$ satisfying $\mathrm{Tr}(\mathbf{D}^{*}\mathbf{D})\le\mathrm{Tr}(\mathbf{D}_{\mathrm{PML}}^{*}\mathbf{D}_{\mathrm{PML}})$ and with smaller total variance than that of the PML estimator.

From Theorem 2, the variance of any estimate of $\mathbf{x}_0$ with bias gradient $\mathbf{D}$ such that $\|\mathbf{D}\|^2\le\|\mathbf{D}_{\mathrm{PML}}\|^2$ satisfies

$$C\ge\frac{1}{N}\mathrm{Tr}\left(\left(1-\|\mathbf{D}_{\mathrm{PML}}\|\right)^2\mathbf{J}_1^{-1}\right)=\frac{1}{N}\mathrm{Tr}\left(\left(1-\left\|\frac{\partial\check{\mathbf{x}}}{\partial\mathbf{x}_0}-\mathbf{I}\right\|\right)^2\mathbf{J}_1^{-1}\right). \quad (85)$$

Thus, if we can choose $R(\mathbf{x})$ such that

$$\mathrm{Tr}\left(\left(\mathbf{J}(\check{\mathbf{x}})+\beta_0\mathbf{M}(\check{\mathbf{x}})\right)^{-1}\mathbf{C}(\check{\mathbf{x}})\left(\mathbf{J}(\check{\mathbf{x}})+\beta_0\mathbf{M}(\check{\mathbf{x}})\right)^{-1}\right)=\mathrm{Tr}\left(\left(1-\left\|\frac{\partial\check{\mathbf{x}}}{\partial\mathbf{x}_0}-\mathbf{I}\right\|\right)^2\mathbf{J}_1^{-1}\right) \quad (86)$$

where $\partial\check{\mathbf{x}}/\partial\mathbf{x}_0$ is given by (80), then the corresponding PML estimator achieves the UCRLB with worst-case bias constraint so that asymptotically, there is no linear or nonlinear estimator with bias gradient $\mathbf{D}$ satisfying $\|\mathbf{D}\|\le\|\mathbf{D}_{\mathrm{PML}}\|$ and with smaller total variance than that of the PML estimator.

The conditions (84) and (86) are not very insightful. To develop some intuition into the optimal choice of $R(\mathbf{x})$, we now consider the case in which we seek to estimate a scalar x_0 from N iid measurements. In this case, the average and worst-case UCRLB coincide so that the variance C of any estimate of x_0 with bias gradient D such that $D^2\le D_{\mathrm{PML}}^2=(\partial\check{x}/\partial x_0-1)^2$ satisfies

$$C\ge\left(1-\left|\frac{\partial\check{x}}{\partial x_0}-1\right|\right)^2\frac{1}{NJ_1}. \quad (87)$$

Here

$$J_1=E\left\{\left(\frac{\partial\log p(y;x_0)}{\partial x}\right)^2\right\} \quad (88)$$

$$\check{x}=\arg\max\left\{E\left\{\log p(\mathbf{y};x)\right\}-\beta_0R(x)\right\} \quad (89)$$

and

$$\frac{\partial\check{x}}{\partial x_0}=\frac{1}{J(\check{x})+\beta_0M(\check{x})}\frac{\partial}{\partial x_0}E\left\{\frac{\partial\log p(y;\check{x})}{\partial x}\right\} \quad (90)$$

with

$$C(\check{x})=\mathrm{var}\left\{\frac{\partial\log p(y;\check{x})}{\partial x}\right\}$$
$$J(\check{x})=-E\left\{\frac{\partial^2\log p(y;\check{x})}{\partial x^2}\right\}$$
$$M(\check{x})=\frac{\partial^2R(\check{x})}{\partial x^2}. \quad (91)$$

The asymptotic variance of the PML estimator is given from Theorem 3 by

$$C_{\mathrm{PML}}=\frac{C(\check{x})}{N(J(\check{x})+\beta_0M(\check{x}))^2}. \quad (92)$$

It thus follows that if we can choose $R(x)$ such that

$$\left(1-\left|\frac{\partial\check{x}}{\partial x_0}-1\right|\right)^2\frac{1}{J_1}=\frac{C(\check{x})}{(J(\check{x})+\beta_0M(\check{x}))^2} \quad (93)$$

where $\partial\check{x}/\partial x_0$ is given by (90), then the corresponding PML estimator achieves the UCRLB. In Appendix B, we develop a general condition under which (93) is satisfied, which is summarized in the following theorem.

Theorem 4: Let x_0 denote an unknown deterministic parameter, let $\mathbf{y}_1,\ldots,\mathbf{y}_N$ denote N iid vector measurements of x_0, and let $\hat{x}^{\mathrm{PML}}$ denote the PML estimator of x_0 from the measurements $\mathbf{y}_1,\ldots,\mathbf{y}_N$ that maximizes the penalized log-likelihood with penalizing function $R(x)$. Then, $\hat{x}^{\mathrm{PML}}$ asymptotically achieves the UCRLB if and only if $R(x)$ is chosen such that

$$\left(1-\left|\frac{\partial\check{x}}{\partial x_0}-1\right|\right)^2\frac{1}{J_1}=\frac{C(\check{x})}{(J(\check{x})+\beta_0 M(\check{x}))^2}$$

where J_1 is the Fisher information from a single observation given by (88); $\check{x}$ is defined in (89); $C(\check{x})$, $J(\check{x})$, and $M(\check{x})$ are defined in (91); and

$$\frac{\partial\check{x}}{\partial x_0}=\frac{1}{J(\check{x})+\beta_0 M(\check{x})}\frac{\partial}{\partial x_0}E\left\{\frac{\partial\log p(\mathbf{y};\check{x})}{\partial x}\right\}.$$

In addition, if $\partial\check{x}/\partial x_0\le 1$, then $\hat{x}^{\mathrm{PML}}$ asymptotically achieves the UCRLB if and only if $R(x)$ is chosen such that

$$\frac{\partial\log p(\mathbf{y};\check{x})}{\partial x}-E\left\{\frac{\partial\log p(\mathbf{y};\check{x})}{\partial x}\right\}=c\frac{\partial\log p(\mathbf{y};x_0)}{\partial x}\quad(94)$$

for some deterministic constant c.

In many cases, (94) is satisfied for all $R(x)$ so that any $R(x)$ such that $\partial\check{x}/\partial x\le 1$ is asymptotically optimal. For example, suppose we are given measurements $\mathbf{y}_i=\mathbf{m}+\sigma_0\mathbf{n}_i$, $1\le i\le N$, where $\mathbf{m}$ is a known length-n vector, $\mathbf{n}_i$ are iid random vectors with $\mathbf{n}_1\sim\mathcal{N}(0,\mathbf{I})$, and σ_0 is unknown. In this example

$$\frac{\partial\log p(\mathbf{y};\check{\sigma})}{\partial\sigma}=-\frac{n}{\check{\sigma}}+\frac{1}{\check{\sigma}^3}(\mathbf{y}-\mathbf{m})^*(\mathbf{y}-\mathbf{m}).\quad(95)$$

Since $E\{(\mathbf{y}-\mathbf{m})^*(\mathbf{y}-\mathbf{m})\}=n\sigma_0^2$, we have that

$$\begin{aligned}&\frac{\partial\log p(\mathbf{y};\check{x})}{\partial x}-E\left\{\frac{\partial\log p(\mathbf{y};\check{x})}{\partial x}\right\}\\&\quad=\frac{1}{\check{\sigma}^3}\left((\mathbf{y}-\mathbf{m})^*(\mathbf{y}-\mathbf{m})-n\sigma_0^2\right)\\&\quad=\frac{\check{\sigma}^3}{\sigma_0^3}\frac{\partial\log p(\mathbf{y};x_0)}{\partial x}\end{aligned}\quad(96)$$

so that (94) is satisfied for all $R(x)$. The same conclusion holds when estimating the mean $\mathbf{m}$, assuming σ_0 is known. Another, non-Gaussian example is considered in the next section.

VII. Example

We now consider an example illustrating the PML estimator and its asymptotic optimality.

Consider the case in which we are given N scalar iid measurements $y_1,\ldots,y_N$ of an exponential random variable with unknown mean $1/x_0>0$. Thus

$$p(y_i;x_0)=x_0e^{-y_ix_0},\quad 1\le i\le N.\quad(97)$$

The PML estimate $\hat{x}^{\mathrm{PML}}$ with penalizing function $R(x)$ is given by the value of x that maximizes

$$PL(x)=N\log x-x\sum_{i=1}^{N}y_i-\beta_N R(x)\quad(98)$$

for some parameter $\beta_N>0$ such that $\beta_N/N\to\beta_0$, as $N\to\infty$. We seek a penalizing function $R(x)$ that is optimal in the sense that the resulting estimator asymptotically achieves the UCRLB.

From (97)

$$\frac{\partial\log p(y;x)}{\partial x}=\frac{1}{x}-y\quad(99)$$

so that

$$E\left\{\frac{\partial\log p(y;\check{x})}{\partial x}\right\}=\frac{1}{\check{x}}-\frac{1}{x_0}.\quad(100)$$

Therefore

$$\frac{\partial\log p(y;\check{x})}{\partial x}-E\left\{\frac{\partial\log p(y;\check{x})}{\partial x}\right\}=\frac{1}{x_0}-y\quad(101)$$

and

$$\frac{\partial\log p(y;x_0)}{\partial x_0}=\frac{1}{x_0}-y.\quad(102)$$

Thus, from Theorem 4, it follows that for any choice of $R(x)$ such that $\partial\check{x}/\partial x_0\le 1$, the resulting PML estimator asymptotically achieves the UCRLB. Note, however, that for finite values of N, the performance of the PML estimator will depend on the specific choice of $R(x)$.

To compute the derivative $\partial\check{x}/\partial x_0$, we note that from (100)

$$\frac{\partial}{\partial x_0}E\left\{\frac{\partial\log p(y;\check{x})}{\partial x}\right\}=\frac{1}{x_0^2}.\quad(103)$$

Differentiating (99) with respect to x

$$\frac{\partial^2\log p(y;x)}{\partial x^2}=-\frac{1}{x^2}\quad(104)$$

so that

$$J(\check{x})=\frac{1}{\check{x}^2}.\quad(105)$$

Combining (80), (103), and (105)

$$\frac{\partial\check{x}}{\partial x_0}=\frac{\frac{1}{x_0^2}}{\frac{1}{\check{x}^2}+\beta_0 M(\check{x})}.\quad(106)$$

If $\partial R(\check{x})/\partial x,\partial^2R(\check{x})/\partial x^2\ge 0$, then from the definition of $\check{x}$

$$\frac{1}{\check{x}}=\frac{1}{x_0}+\beta_0\frac{\partial R(\check{x})}{\partial x}\ge\frac{1}{x_0}\quad(107)$$

so that

$$\frac{\partial\check{x}}{\partial x_0}=\frac{\frac{1}{x_0^2}}{\frac{1}{\check{x}^2}+\beta_0 M(\check{x})}\le 1\quad(108)$$

and the PML estimator is optimal.

As an example, suppose that $R(x) = x$. The resulting PML estimator is given by

$$\hat{x}^{\mathrm{PML}} = \arg\max \left\{ N \log x - x \left(\sum_{i=1}^{N} y_i + \beta_N \right) \right\} = \frac{N}{\sum_{i=1}^{N} y_i + \beta_N}. \tag{109}$$

Since $\partial R(\check{x})/\partial x = 1 \geq 0$ and $\partial^2 R(\check{x})/\partial x^2 = 0$, it follows that the estimator of (109) asymptotically achieves the UCRLB.

As another example, suppose that $R(x) = \log x$. In this case, $\partial^2 R(\check{x})/\partial x^2 \leq 0$. Nonetheless, as we now show, $\partial \check{x}/\partial x_0 < 1$, so that the resulting PML estimator is optimal.

From (107)

$$\check{x} = (1 - \beta_0) x_0 \tag{110}$$

so that from (106)

$$\frac{\partial \check{x}}{\partial x_0} = \frac{\frac{1}{x_0^2}}{\frac{(1-\beta_0)}{\check{x}^2}} = 1 - \beta_0 \leq 1. \tag{111}$$

We therefore conclude that the resulting PML estimator, which is given by

$$\hat{x}^{\mathrm{PML}} = \arg\max \left\{ (N - \beta_N) \log x - x \sum_{i=1}^{N} y_i \right\} = \frac{N - \beta_N}{\sum_{i=1}^{N} y_i} \tag{112}$$

asymptotically achieves the UCRLB.

We now compare the performance of the PML estimators of (109) and (112) with the UCRLB, for different values of N. To this end, we need to determine the variance $\sigma_{\hat{x}}^2$ of the estimators and the squared bias gradient D^2. Rather than attempting to determine these quantities analytically, we propose to estimate them from the measurements. Thus, for each value of γ, and each of the estimators, we generate an estimate $\hat{\sigma}^2$ of the estimator's variance $\sigma_{\hat{x}}^2$ and an estimate $\hat{D}^2$ of the squared bias gradient D^2.

To estimate the variance of each of the estimators, for each γ, we generate $L = 5000$ PML estimators, where each estimator is based on N iid measurements. Let $(\hat{x}^{\mathrm{PML}})^{(i)}$ denote the ith estimator. The variance is then estimated as

$$\hat{\sigma}^2 = \frac{1}{L} \sum_{i=1}^{L} \left((\hat{x}^{\mathrm{PML}})^{(i)} - \bar{x}_{\mathrm{PML}} \right)^2 \tag{113}$$

where $\bar{x}_{\mathrm{PML}}$ is the sample mean and is given by

$$\bar{x}_{\mathrm{PML}} = \frac{1}{L} \sum_{i=1}^{L} (\hat{x}^{\mathrm{PML}})^{(i)}. \tag{114}$$

To estimate the squared bias gradient of the estimator, we used the procedure detailed in [15]. Specifically, in [15], the authors propose to estimate the squared bias gradient of an estimator $\hat{x}$ of x as $\hat{D}^2$, where

$$\hat{D} = \frac{1}{L} \sum_{i=1}^{L} \left(\left(\hat{x}^{(i)} - \zeta^{(i)} \right) \sum_{j=1}^{N} \frac{\partial \log p\left(y_j^{(i)}; x \right)}{\partial x} \right) - 1. \tag{115}$$

Here

$$\zeta^{(i)} = \frac{1}{L-1} \left(\sum_{j=1}^{L} \hat{x}^{(j)} - \hat{x}^{(i)} \right) \tag{116}$$

and $y_j^{(i)}$ denotes the jth observation used in computing the ith estimator.

In our example

$$\frac{\partial \log p\left(y_j^{(i)}; x \right)}{\partial x} = \frac{1}{x} - y_j^{(i)} \tag{117}$$

so that

$$\hat{D} = \frac{1}{L} \sum_{i=1}^{L} \left(\left((\hat{x}^{\mathrm{PML}})^{(i)} - \zeta^{(i)} \right) \left(\frac{N}{x} - \sum_{j=1}^{N} y_j^{(i)} \right) \right) - 1. \tag{118}$$

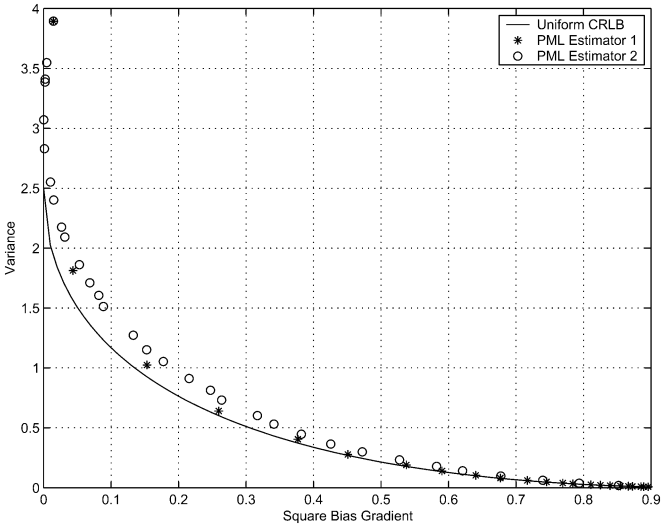

Fig. 2. Performance of the PML estimators (109) (denoted "1") and (112) (denoted "2") with $N = 10$ in comparison with the UCRLB. The line denotes the UCRLB, the circles denote the performance of the PML estimator 1, and the stars denote the performance of the PML estimator 2.

In Figs. 2–4, we plot the estimated variance of the PML estimators as a function of the estimated squared bias gradient for $N = 10$, 20 and 30, respectively. For comparison, we also plot the UCRLB. From the figures, it is apparent that even for small N, the UCRLB serves as a good approximation of the estimator's variance, particularly for large values of bias gradient norm. However, for small values of the squared bias gradient, the actual variance is larger than the bound. We note that the variance of the bias gradient estimate (118) is larger for small bias gradients, which may partially explain the large deviation in this regime. As we expect from our analysis, for increasing values of N, the variance of both estimators approaches that of the UCRLB for all values of squared bias gradient, as can be seen from Figs. 3 and 4. Note, however, that for small values of

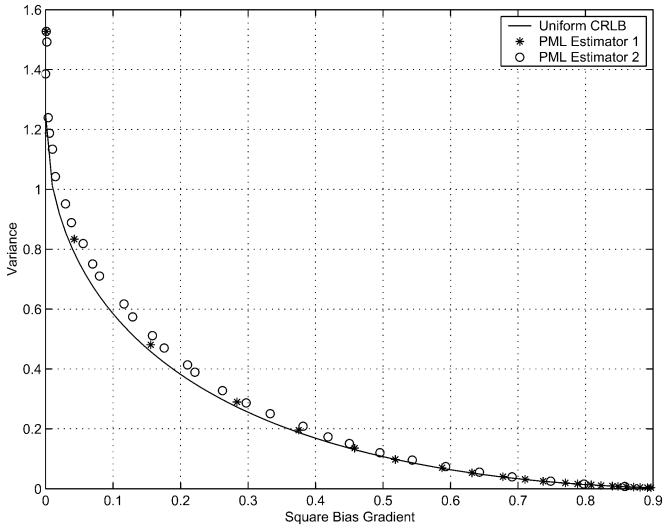

Fig. 3. Performance of the PML estimators (109) (denoted "1") and (112) (denoted "2") with $N = 20$ in comparison with the UCRLB. The line denotes the UCRLB, the circles denote the performance of the PML estimator 1, and the stars denote the performance of the PML estimator 2.

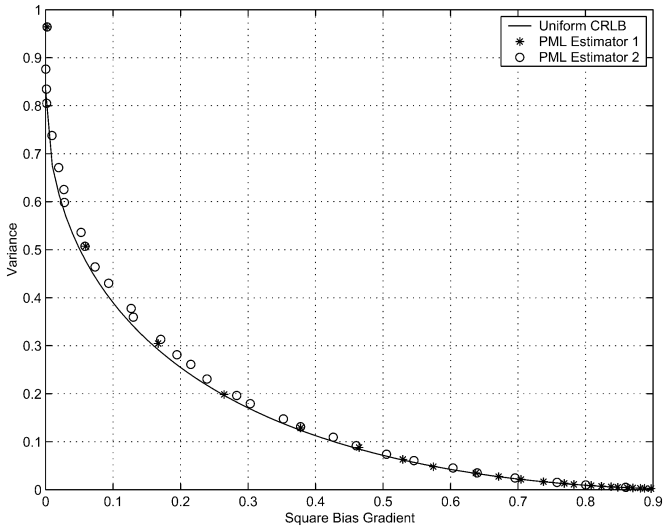

Fig. 4. Performance of the PML estimators (109) (denoted "1") and (112) (denoted "2") with $N = 30$, in comparison with the UCRLB. The line denotes the UCRLB, the circles denote the performance of the PML estimator 1, and the stars denote the performance of the PML estimator 2.

N, the performance of the two estimators is different. In particular, the estimator given by (109) results in a smaller variance than the estimator given by (112) for finite values of N.

VIII. Conclusion

In this paper, we characterized the fundamental tradeoff between variance and bias in estimating an unknown deterministic parameter vector by deriving lower bounds on the minimal achievable total variance subject to constraints on the norm of the bias gradient matrix. In the case in which the unknown deterministic parameters are related to the measurements through a linear Gaussian model, we demonstrated that the lower bounds are achievable using linear estimators. In particular, we showed that Tikhonov regularization minimizes the total variance from all estimators with a bounded average bias gradient, and the shrunken estimator minimizes the total variance from all estimators with a bounded worst-case bias gradient.

We then derived the asymptotic mean and covariance of the PML estimator when estimating an unknown vector from iid measurements and showed that for an appropriate choice of penalizing function, the PML estimator asymptotically achieves the UCRLB.

Although, in many cases, there are several PML estimators that asymptotically achieve the UCRLB, as we demonstrated in the context of a concrete example in Section VII, the performance of these estimators differ for finite values of the number N of measurements. An interesting direction for future research, therefore, is to analyze the performance of the PML estimator for finite values of N. Another interesting question is whether or not there are other cases besides the linear Gaussian model, in which the PML estimator achieves the UCRLB for all values of N. Finally, throughout the paper, we explicitly assume that the Fisher information matrix is nonsingular. It would also be of interest to extend the results to the case of a singular Fisher information matrix.

Appendix A
Proof of Theorem 3

The proof of Theorem 3 relies on the following lemma.

Lemma 2: Let $\mathbf{x}_0$ denote an unknown deterministic vector, let $\mathbf{y}_1, \ldots, \mathbf{y}_N$ denote N iid measurements of $\mathbf{x}_0$, let $\hat{\mathbf{x}}^{\mathrm{PML}}$ denote the PML estimator of $\mathbf{x}_0$ from the measurements $\mathbf{y}_1, \ldots, \mathbf{y}_N$ that maximizes the penalized likelihood (71), and let $\check{\mathbf{x}}$ be defined by (72). Then, $\hat{\mathbf{x}}^{\mathrm{PML}} \to \check{\mathbf{x}}$ as $N \to \infty$ with probability one.

Proof: For $N \to \infty$, we have that

$$\frac{1}{N} PL(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} \log p(\mathbf{y}_i; \mathbf{x}) - \frac{\beta_N}{N} R(\mathbf{x}) \to E\{\log p(\mathbf{y}; \mathbf{x})\} - \beta_0 R(\mathbf{x}). \quad (119)$$

Therefore

$$\hat{\mathbf{x}}^{\mathrm{PML}} \to \arg\max \left\{ \lim_{N\to\infty} \frac{1}{N} PL(\mathbf{x}) \right\} = \arg\max \{E\{\log p(\mathbf{y}; \mathbf{x})\} - \beta_0 R(\mathbf{x})\} = \check{\mathbf{x}}. \quad (120)$$

□

We now expand $\partial \log PL(\mathbf{x})/\partial \mathbf{x}$ about $\check{\mathbf{x}}$ using a Taylor expansion. Note that Assumption 1 ensures that such a Taylor expansion exists. By the mean value theorem, for each $1 \le j \le m$

$$\frac{\partial \log PL(\hat{\mathbf{x}}^{\mathrm{PML}})}{\partial x_j} = \frac{\partial \log PL(\check{\mathbf{x}})}{\partial x_j} + \sum_{k=1}^{m} \left(\hat{x}_k^{\mathrm{PML}} - \check{x}_k\right) \frac{\partial^2 \log PL(\check{\mathbf{x}})}{\partial x_j \partial x_k} + \frac{1}{2} \sum_{k,l=1}^{m} \left(\hat{x}_k^{\mathrm{PML}} - \check{x}_k\right)\left(\hat{x}_l^{\mathrm{PML}} - \check{x}_l\right) \frac{\partial^3 \log PL(\tilde{\mathbf{x}})}{\partial x_j \partial x_k \partial x_l} \quad (121)$$

where $\tilde{\mathbf{x}}$ is a point on the line segment connecting $\hat{\mathbf{x}}^{\mathrm{PML}}$ and $\check{\mathbf{x}}$. By the definition of $\hat{\mathbf{x}}^{\mathrm{PML}}$, we have that

$$\frac{\partial \log PL(\hat{\mathbf{x}}^{\mathrm{PML}})}{\partial \mathbf{x}} = 0 \tag{122}$$

so that from (121)

$$-\frac{1}{\sqrt{N}}\frac{\partial \log PL(\check{\mathbf{x}})}{\partial x_j} = \sqrt{N}\sum_{k=1}^{m}\left(\hat{x}_k^{\mathrm{PML}} - \check{x}_k\right)\left(\frac{1}{N}\frac{\partial^2 \log PL(\check{\mathbf{x}})}{\partial x_j \partial x_k} + \frac{1}{2N}\sum_{l=1}^{m}\left(\hat{x}_l^{\mathrm{PML}} - \check{x}_l\right)\frac{\partial^3 \log PL(\tilde{\mathbf{x}})}{\partial x_j \partial x_k \partial x_l}\right) \tag{123}$$

which can be expressed in vector form as

$$\mathbf{z} = \mathbf{A}\mathbf{u} \tag{124}$$

where

$$\begin{aligned}\mathbf{z} &= -\frac{1}{\sqrt{N}}\left(\frac{\partial \log PL(\check{\mathbf{x}})}{\partial \mathbf{x}}\right)^* \\ \mathbf{u} &= \sqrt{N}\left(\hat{\mathbf{x}}^{\mathrm{PML}} - \check{\mathbf{x}}\right) \\ \mathbf{A}_{jk} &= \frac{1}{N}\frac{\partial^2 \log PL(\check{\mathbf{x}})}{\partial x_j \partial x_k} + \frac{1}{2N}\sum_{l=1}^{m}\left(\hat{x}_l^{\mathrm{PML}} - \check{x}_l\right)\frac{\partial^3 \log PL(\tilde{\mathbf{x}})}{\partial x_j \partial x_k \partial x_l}.\end{aligned} \tag{125}$$

Here, $\mathbf{A}_{jk}$ denotes the jkth element of the matrix $\mathbf{A}$.

Now, from the strong law of large numbers, we have that

$$\begin{aligned}\frac{1}{N}\frac{\partial^2 \log PL(\check{\mathbf{x}})}{\partial \mathbf{x}^2} &= \frac{1}{N}\sum_{i=1}^{N}\frac{\partial^2 \log p(\mathbf{y}_i;\check{\mathbf{x}})}{\partial \mathbf{x}^2} - \frac{\beta_N}{N}\frac{\partial^2 R(\check{\mathbf{x}})}{\partial \mathbf{x}^2} \\ &\to E\left\{\frac{\partial^2 \log p(\mathbf{y};\check{\mathbf{x}})}{\partial \mathbf{x}^2}\right\} - \beta_0\frac{\partial^2 R(\check{\mathbf{x}})}{\partial \mathbf{x}^2} \\ &= -\mathbf{J}(\check{\mathbf{x}}) - \beta_0\mathbf{M}(\check{\mathbf{x}}).\end{aligned} \tag{126}$$

Similarly, from the strong law of large numbers and Assumption 2

$$\begin{aligned}&\frac{1}{N}\frac{\partial^3 \log PL(\tilde{\mathbf{x}})}{\partial x_j \partial x_k \partial x_l} = \\ &= \frac{1}{N}\sum_{i=1}^{N}\frac{\partial^3 \log p(\mathbf{y}_i;\tilde{\mathbf{x}})}{\partial x_j \partial x_k \partial x_l} - \frac{\beta_N}{N}\frac{\partial^3 R(\check{\mathbf{x}})}{\partial x_j \partial x_k \partial x_l} \\ &\to E\left\{\frac{\partial^3 \log p(y;\tilde{\mathbf{x}})}{\partial x_j \partial x_k \partial x_l}\right\} - \beta_0\frac{\partial^3 R(\check{\mathbf{x}})}{\partial x_j \partial x_k \partial x_l} < \infty\end{aligned} \tag{127}$$

with probability 1. From Lemma 2, $\hat{\mathbf{x}}^{\mathrm{PML}} - \check{\mathbf{x}} \to 0$ as $N \to \infty$, which implies that

$$\frac{(\hat{\mathbf{x}}^{\mathrm{PML}} - \check{\mathbf{x}})}{N}\frac{\partial^3 \log PL(\tilde{\mathbf{x}})}{\partial x_j \partial x_k \partial x_l} \to 0 \tag{128}$$

with probability 1. Therefore, the matrix $\mathbf{A}$ converges to $\mathbf{J}(\check{\mathbf{x}}) + \beta_0\mathbf{M}(\check{\mathbf{x}})$ with probability 1.

We now consider the asymptotic distribution of $\mathbf{z}$, which we express as

$$\begin{aligned}\mathbf{z} &= \frac{1}{\sqrt{N}}\left(\frac{\partial \log PL(\check{\mathbf{x}})}{\partial \mathbf{x}}\right)^* \\ &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{v}_i - \frac{\beta_N}{\sqrt{N}}\left(\frac{\partial R(\check{\mathbf{x}})}{\partial \mathbf{x}}\right)^* \\ &= \mathbf{t}_N - \frac{\beta_N}{\sqrt{N}}\left(\frac{\partial R(\check{\mathbf{x}})}{\partial \mathbf{x}}\right)^*\end{aligned} \tag{129}$$

where

$$\mathbf{v}_i = \left(\frac{\partial \log p(\mathbf{y}_i;\check{\mathbf{x}})}{\partial \mathbf{x}}\right)^* \tag{130}$$

and

$$\mathbf{t}_N = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{v}_i. \tag{131}$$

Since the random vectors $\mathbf{v}_i$ are iid, it follows from the multivariate central limit theorem [12] that $\mathbf{t}_N$ is asymptotically Gaussian. To complete the description of $\mathbf{t}_N$, we need to determine its mean and covariance. From (72), it follows that

$$\frac{\partial E\left\{\log p(\mathbf{y};\check{\mathbf{x}})\right\}}{\partial \mathbf{x}} - \beta_0\frac{\partial R(\check{\mathbf{x}})}{\partial \mathbf{x}} = 0 \tag{132}$$

so that

$$E\left\{\mathbf{t}_N\right\} = \sqrt{N}\beta_0\left(\frac{\partial R(\check{\mathbf{x}})}{\partial \mathbf{x}}\right)^*. \tag{133}$$

In addition

$$\begin{aligned}&E\left\{(\mathbf{t}_N - E\left\{\mathbf{t}_N\right\})(\mathbf{t}_N - E\left\{\mathbf{t}_N\right\})^*\right\} = \\ &\quad E\left\{(\mathbf{v}_1 - E\left\{\mathbf{v}_1\right\})^*(\mathbf{v}_1 - E\left\{\mathbf{v}_1\right\})\right\} = \mathbf{C}(\check{\mathbf{x}}).\end{aligned} \tag{134}$$

Thus, we conclude that

$$\mathbf{t}_N \overset{a}{\sim} \mathcal{N}\left(\sqrt{N}\beta_0\left(\frac{\partial R(\check{\mathbf{x}})}{\partial \mathbf{x}}\right)^*, \mathbf{C}(\check{\mathbf{x}})\right). \tag{135}$$

To develop the asymptotic distribution of $\mathbf{z}$, we rely on the following Lemma [35, p. 19].

Lemma 3: Let $\mathbf{t}_N$ denote a sequence of random vectors that converges in distribution to $\mathbf{t}$, and let $\mathbf{s}_N$ denote a sequence of random vectors that converges in probability to a finite vector $\mathbf{s}$. Then, $\mathbf{t}_N + \mathbf{s}_N$ converges in distribution to $\mathbf{t} + \mathbf{s}$.

We can express $\mathbf{z}$ as $\mathbf{z} = \mathbf{t}_N + \mathbf{s}_N$, where $\mathbf{s}_N = -(\beta_N/\sqrt{N})\left(\partial R(\check{\mathbf{x}})/\partial \mathbf{x}\right)^*$. It then follows from Lemma 3 and (135) that

$$\mathbf{z} \overset{a}{\sim} \mathcal{N}\left(0, \mathbf{C}(\check{\mathbf{x}})\right). \tag{136}$$

To complete the proof of Theorem 3, we rely on the fact that if $\mathbf{z} = \mathbf{A}\mathbf{u}$, where $\mathbf{A}$ converges in probability to an invertible matrix, then $\mathbf{u}$ converges in distribution to $\mathbf{A}^{-1}\mathbf{z}$ [12, p. 465]. Since $\mathbf{A}$ converges to $\mathbf{J}(\check{\mathbf{x}}) + \beta_0\mathbf{M}(\check{\mathbf{x}})$

$$\mathbf{u} = \sqrt{N}(\hat{\mathbf{x}}^{\mathrm{PML}} - \check{\mathbf{x}}) \overset{a}{\sim} \mathcal{N}\left(0, (\mathbf{J}(\check{\mathbf{x}}) + \beta_0\mathbf{M}(\check{\mathbf{x}}))^{-1}\mathbf{C}(\check{\mathbf{x}})(\mathbf{J}(\check{\mathbf{x}}) + \beta_0\mathbf{M}(\check{\mathbf{x}}))^{-1}\right). \tag{137}$$

Appendix B
Proof of Theorem 4

Using the equality

$$\begin{aligned}\frac{\partial}{\partial x_0}E\left\{\frac{\partial \log p(y;\check{x})}{\partial x}\right\} &= \int \frac{\partial \log p(y;\check{x})}{\partial x}\frac{\partial p(y;x_0)}{\partial x_0}dy \\ &= \int \frac{\partial \log p(y;\check{x})}{\partial x}\frac{\partial \log p(y;x_0)}{\partial x_0}p(y;x_0)dy \\ &= E\left\{\frac{\partial \log p(y;\check{x})}{\partial x}\frac{\partial \log p(y;x_0)}{\partial x_0}\right\}\end{aligned} \tag{138}$$

we have that

$$\frac{\partial \check{x}}{\partial x_0} = \frac{1}{J(\check{x}) + \beta_0 M(\check{x})} E\left\{ \frac{\partial \log p(y;\check{x})}{\partial x} \frac{\partial \log p(y;x_0)}{\partial x_0} \right\}. \tag{139}$$

Now, suppose that $\partial \check{x}/\partial x_0 \leq 1$. In this case, using (139), (93) becomes

$$E^2\left\{ \frac{\partial \log p(y;\check{x})}{\partial x} \frac{\partial \log p(y;x_0)}{\partial x_0} \right\} = J_1 C(\check{x}). \tag{140}$$

To see when there exists an $R(x)$ such that (140) is satisfied, define

$$\begin{aligned} A' &= \frac{\partial \log p(y;\check{x})}{\partial x} \\ A &= A' - E\{A'\} \\ B &= \frac{\partial \log p(y;x_0)}{\partial x_0}. \end{aligned} \tag{141}$$

We have immediately that $E\{B\} = 0$, $E\{B^2\} = J_1$, and $E\{A^2\} = C(\check{x})$. In addition, since $E\{B\} = 0$, $E(A'B) = E(AB)$. Therefore, (140) is equivalent to

$$E^2\{AB\} = E\{B^2\} E\{A^2\}. \tag{142}$$

From the Cauchy–Schwarz inequality, we have that for any random variables A and B

$$E^2\{AB\} \leq E\{B^2\} E\{A^2\} \tag{143}$$

with equality if and only if $A = cB$ for some deterministic constant c. It follows that (142) can be satisfied if and only if

$$\frac{\partial \log p(y;\check{x})}{\partial x} - E\left\{ \frac{\partial \log p(y;\check{x})}{\partial x} \right\} = c \frac{\partial \log p(y;x_0)}{\partial x} \tag{144}$$

for some deterministic constant c.

ACKNOWLEDGMENT

The author wishes to thank Prof. J. A. Fessler for many valuable comments and suggestions, for carefully reviewing several earlier versions of this paper, for referring her to related work in this field, and for motivating the discussion in Section VI, Prof. G. W. Wornell for fruitful discussions, and the anonymous reviewers for constructive comments that greatly improved the presentation of this paper.

REFERENCES

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

[2] A. N. Tikhonov and V. Y. Arsenin, *Solution of Ill-Posed Problems*. Washington, DC: V.H. Winston, 1977.

[3] G. Demoment, "Image reconstruction and restoration: Overview of common estimation structures and problems," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 2024–2036, Dec. 1989.

[4] D. M. Titterington, "Common structure of smoothing techniques in statistics," *Int. Statist. Rev.*, vol. 53, pp. 141–170, 1985.

[5] F. O'Sullivan, "A statistical perspective on ill-posed inverse problems," *Statist. Sci.*, vol. 1, no. 4, pp. 502–527, 1986.

[6] S. M. Kay, *Modern Spectral Estimation: Theory and Application*. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[7] K. S. Riedel and A. Sidorenko, "Minimum biased multitaper spectral estimation," *IEEE Trans. Signal Proc.*, vol. 43, pp. 188–195, Jan. 1995.

[8] H. Cramer, *Mathematical Methods of Statistics*. Princeton, NJ: Princeton Univ. Press, 1946.

[9] C. R. Rao, "Minimum variance and the estimation of several parameters," in *Proc. Cambridge Phil. Soc.*, 1946, pp. 280–283.

[10] ——, *Linear Statistical Inference and its Applications*, Second ed. New York: Wiley, 1973.

[11] H. V. Poor, *An Introduction to Signal Detection and Estimation*, Second ed. New York: Springer-Verlag, 1994.

[12] E. L. Lehmann and G. Casella, *Theory of Point Estimation*, Second ed. New York: Springer-Verlag, 1998.

[13] H. L. Van Trees, *Detection, Estimation, and Modulation Theory*. New York: Wiley, 1968.

[14] A. O. Hero, "A Cramer-Rao Type Lower Bound for Essentially Unbiased Parameter Estimation," Lincoln Lab., Mass. Inst. Technol., Lexington, MA, Tech. Rep. 890, DTIC AD-A246 666, 1992.

[15] A. O. Hero, J. A. Fessler, and M. Usman, "Exploring estimator bias-variance tradeoffs using the uniform CR bound," *IEEE Trans. Signal Processing*, vol. 44, pp. 2026–2041, Aug. 1996.

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

[17] L. S. Mayer and T. A. Willke, "On biased estimation in linear models," *Technometrics*, vol. 15, pp. 497–508, Aug. 1973.

[18] I. J. Good and R. A. Gaskins, "Nonparametric roughness penalties for probability densities," *Biometrika*, vol. 58, pp. 255–277, 1971.

[19] ——, "Density estimation and bump hunting by the penalized likelihood method exemplified by scattering and meteorite data (with discussion and rejoinder)," *J. Amer. Statist.*, vol. 10, pp. 811–824, 1980.

[20] L. Kaufman, "Maximum likelihood, least squares, and penalized least squares for PET," *IEEE Trans. Med. Imag.*, vol. 12, pp. 200–214, June 1993.

[21] J. A. Fessler and A. O. Hero, "Penalizes maximum-likelihood image reconstruction using space alternating EM algorithms," *IEEE Trans. Image Processing*, vol. 4, pp. 1417–1425, Oct. 1995.

[22] J. A. Fessler, "Mean and variance of implicity defined biased estimators (such as penalized maximum likelihood): Applications to tomography," *IEEE Trans. Image Processing*, vol. 5, pp. 493–506, Mar. 1996.

[23] T. J. Schulz, "Penalized maximum-likelihood estimation of covariance matrices with linear structure," *IEEE Trans. Neural Networks*, vol. 45, pp. 3027–3038, Dec. 1997.

[24] D. Ormoneit and V. Tresp, "Averaging, maximum penalized likelihood and Bayesian estimation for improving Gaussian mixture probability density estimates," *IEEE Trans. Neural Networks*, vol. 9, pp. 639–650, July 1998.

[25] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ: Prentice-Hall, 1993.

[26] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1985.

[27] J. A. Fessler and A. O. Hero, "Cramer-Rao bounds for biased estimators in image restoration," in *Proc. 36th IEEE Midwest Symp. Circuits Syst.*, Detroit, MI, Aug. 1993.

[28] D. P. Bertsekas, *Nonlinear Programming*, Second ed. Belmont, MA: Athena Scientific, 1999.

[29] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization*, ser. Optimization. Philadelphia, PA: MPS-SIAM, 2001.

[30] L. Vandenberghe and S. Boyd, "Semidefinite programming," *SIAM Rev.*, vol. 38, no. 1, pp. 40–95, Mar. 1996.

[31] Y. Nesterov and A. Nemirovski, *Interior-Point Polynomial Algorithms in Convex Programming*. Philadelphia, PEA: SIAM, 1994.

[32] F. Alizadeh, "Combinatorial Optimization With Interior Point Methods and Semi-Definite Matrices," Ph.D. dissertation, Univ. Minnesota, Minneapolis, MN, 1991.

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

[34] M. H. J. Gruber, *Regression Estimators: A Comparative Study*. San Diego, CA: Academic, 1990.

[35] R. J. Serfling, *Approximation Theorems of Mathematical Statistics*. New York: Wiley, 1980.

Yonina C. Eldar (S'98–M'02) received the B.Sc. degree in physics in 1995 and the B.Sc. degree in electrical engineering in 1996, both from Tel-Aviv University (TAU), Tel-Aviv, Israel, and the Ph.D. degree in electrical engineering and computer science in 2001 from the Massachusetts Institute of Technology (MIT), Cambridge.

From January 2002 to July 2002, she was a Postdoctoral fellow at the Digital Signal Processing Group at MIT. She is currently a Senior Lecturer with the Department of Electrical Engineering, Technion–Israel Institute of Technology, Haifa, Israel. She is also a Research Affiliate with the Research Laboratory of Electronics at MIT. Her current research interests are in the general areas of signal processing, statistical signal processing, and quantum information theory.

Dr. Eldar was in the program for outstanding students at TAU from 1992 to 1996. In 1998, she held the Rosenblith Fellowship for study in Electrical Engineering at MIT, and in 2000, she held an IBM Research Fellowship. She is currently a Horev Fellow of the Leaders in Science and Technology program at the Technion as well as an Alon Fellow.