Exploring Estimator Bias-Variance Tradeoffs Using the Uniform CR Bound

Alfred O. Hero, Jeffrey A. Fessler, and Mohammad Usman

Abstract

We introduce a plane, which we call the delta-sigma plane, that is indexed by the norm of the estimator bias gradient and the variance of the estimator. The norm of the bias gradient is related to the maximum variation in the estimator bias function over a neighborhood of parameter space. Using a uniform Cramer-Rao (CR) bound on estimator variance a delta-sigma tradeoff curve is specified which defines an "unachievable region" of the delta-sigma plane for a specified statistical model. In order to place an estimator on this plane for comparison to the delta-sigma tradeoff curve, the estimator variance, bias gradient, and bias gradient norm must be evaluated. We present a simple and accurate method for experimentally determining the bias gradient norm based on applying a bootstrap estimator to a sample mean constructed from the gradient of the log-likelihood. We demonstrate the methods developed in this paper for linear Gaussian and non-linear Poisson inverse problems.

Key Words: parametric estimation, performance bounds, bias-variance plane, unachievable regions, inverse problems, image reconstruction.

EDICS Number: SP 3.8.1

I. Introduction

The goal of this work is to quantify fundamental tradeoffs between the bias and variance functions for parametric estimation problems. Let $\underline{\theta} = [\theta_1, ..., \theta_n]^T \in \Theta$ be a vector of unknown and non-random parameters which parameterize the density $f_Y(y;\underline{\theta})$ of an observed random variable Y. The parameter space Θ is assumed to be an open subset of n-dimensional Euclidean space $\mathbb{R}^n$. For fixed $\underline{\theta}$ let $\hat{t} = \hat{t}(Y)$ be an estimator of the scalar $t_{\underline{\theta}}$, where $t : \Theta \to \mathbb{R}$ is a specified function. Let this estimator have bias $b_{\underline{\theta}} = E_{\underline{\theta}}[\hat{t}] - t_{\underline{\theta}}$ and variance $\sigma^2_{\underline{\theta}} = E_{\underline{\theta}}[(\hat{t} - E_{\underline{\theta}}[t_{\underline{\theta}}])^2]$. Bias is due to 'mismatch' between the average value of the estimator and the true parameter while variance arises from fluctuations in the estimator due to statistical sampling.

In most applications, estimator designs are subject to a tradeoff between bias and variance. For example, in non-parametric spectrum estimation [1], smoothing methods have long been used to reduce the variance of the periodogram at the expense of increased bias [2], [3]. In image restoration, regularization is frequently implemented to reduce noise amplification (variance) at the expense of reduced spatial resolution (bias) [4]. In multiple regression with multicollinearity, biased shrinkage estimators [5] and biased ridge estimators [6] are used to reduce variance of the ordinary least squares estimator. The quantitative study of estimator bias and variance has been useful for characterizing statistical performance for many statistical signal processing applications including: tomographic reconstruction [7], [8], [9], functional imaging [10], non-linear and morphological filtering [11], [12], and spectral estimation of time series [13], [14].

However, the plane parameterized by the bias and variance $b_{\underline{\theta}}$ and $\sigma^2_{\underline{\theta}}$ is not useful for studying fundamental tradeoffs since an estimator can always be found which makes both the bias and variance zero at a given point $\underline{\theta}$. Furthermore, the bias value $b_{\underline{\theta}}$ unfairly penalizes estimators that may have large but constant, and hence removable, biases. In this work we consider the plane parameterized by the norm or length of the bias gradient $\delta_{\underline{\theta}} = \|\nabla b_{\underline{\theta}}\|$ and the square root variance $\sqrt{\sigma^2_{\underline{\theta}}}$, which we call the delta-sigma or $\delta\sigma$ plane. The norm of the bias gradient is directly related to the maximal variation of the bias function over a neighborhood of $\underline{\theta}$ induced by the norm and is unaffected by constant estimator bias components. By appropriate choice of norm, the bias gradient length can be related to the overall bias variation over any prior region of parameter values. For the inverse problems studied here we select the norm to correspond to an *a priori* smoothness constraint on the object.

This paper provides a means for specifying unachievable regions in the $\delta\sigma$ plane via fundamental delta-sigma tradeoff curves. These curves are generated using an extension of the Cramer-Rao (CR) lower bound on the variance of biased estimators presented in [15]. This extension is called the uniform CR bound. In [15] the bound was derived only for an unweighted Euclidean norm on the bias gradient and for non-singular Fisher information. Therein the reader was cautioned that the resulting bound will generally depend on the units and dimensions used to express each of the parameters. It was also pointed out in [15] that the user should identify an ellipsoid of expected parameter variations, which will depend in the user's units, and perform a normalizing transformation of the ellipsoid to a spheroid prior to applying the bound. This parameter transformation is equivalent to using a diagonally weighted bias gradient norm constraint in the original untransformed parameter space. The uniform CR bound presented in this paper generalizes [15] to allow functional estimation, to

[1]A. Hero and J. Fessler are with the Dept. of Electrical Engineering and Computer Science, The University of Michigan, Ann Arbor, MI 48109-2122. M. Usman is with the Punjab Institute of Computer Science, Pakistan. This work was supported in part by National Science Foundation under grant BCS-9024370, a Government of Pakistan Postgraduate Fellowship, NIH grants CA-54362 and CA-60711, and DOE grant DE-FG02-87ER60561

cover the case of singular or ill-conditioned Fisher matrices, and to account for a general norm constraint on bias gradient. Some elements of the latter generalization were first presented in [16].

The methods described herein can be used for system optimization, i.e., to choose the system which minimizes the size of the unachievable region, when estimator unbiasedness is an overly stringent or unrealistic constraint [17], or they can be used to gauge the closeness to optimality of biased estimators in terms of their nearness to the unachievable region [18]. Alternatively, as discussed in more detail in [15], these results can be used to investigate the reliability of CR bound studies when small estimator biases may be present. Finally, these results can be used for validation of estimator simulations by empirically verifying that the simulations do not place estimator performance in the unachievable region of the $\delta\sigma$ plane.

In order to place an estimator on the $\delta\sigma$ plane we must calculate estimator variance and bias gradient norm. For most nonlinear estimators computation of these quantities is analytically intractable. We present a methodology for experimentally determining these quantities which uses the gradient of the log-likelihood function $\nabla \ln f_Y(y;\underline{\theta})$ and a bootstrap-type estimator to estimate the bias gradient norm.

We illustrate these methods for linear Gaussian and nonlinear Poisson inverse problems. Such problems arise in image restoration, image reconstruction, and seismic deconvolution, to name but a few examples. Note that even for the linear Gaussian problem there may not exist unbiased estimators when the system matrix is ill-conditioned or rank deficient [19]. For each model we compare the performance of quadratically penalized maximum likelihood estimators to the fundamental delta-sigma tradeoff curve. We show that the bias gradient $\nabla b_{\underline{\theta}}$ of these estimators is closely related to the point spread function of the estimator when one wishes to estimate a single component $t_{\underline{\theta}} = \theta_k$. For the full rank linear Gaussian case the quadratically penalized likelihood estimator achieves the fundamental delta-sigma tradeoff in the $\delta\sigma$ plane when the roughness penalty matrix is matched to the norm chosen by the user to measure bias gradient length. In this case the bias gradient norm constraint is equivalent to a constraint on bias variation over a roughness constrained neighborhood of $\underline{\theta}$. We thus have a very strong optimality property: the penalized maximum likelihood estimator minimizes variance over all estimators whose maximal bias variation is bounded over the neighborhood. For the rank deficient linear Gaussian problem the uniform CR bound is shown to be achievable by a different estimator under certain conditions. Finally, for the non-linear Poisson case an asymptotic analysis shows that the penalized maximum likelihood estimator of [20] achieves the fundamental delta-sigma tradeoff curve for sufficiently large values of the regularization parameter and a suitably chosen penalty matrix. We present simulation results that empirically validate our asymptotic analysis.

A. *Variance, Bias and Bias Gradient*

Let $\hat{t}$ be an estimator of the scalar differentiable function $t_{\underline{\theta}}$. The mean-square-error (MSE) is a widely used measure of performance for an estimator $\hat{t}$ and is simply related to the estimator bias $b_{\underline{\theta}}$ and the estimator variance $\sigma^2_{\underline{\theta}}$ through the relation $\text{MSE}_{\underline{\theta}} = \sigma^2_{\underline{\theta}} + b^2_{\underline{\theta}}$. While the MSE criterion is of value in many applications, the estimator bias and estimator variance provide a more complete picture of performance than the MSE alone. From $b_{\underline{\theta}}$ and $\sigma^2_{\underline{\theta}}$ one can derive other important measures such as signal-to-noise-ratio $SNR = |t_{\underline{\theta}} + b_{\underline{\theta}}|^2/\sigma^2_{\underline{\theta}}$, coefficient of variation $1/SNR$, and generalized MSE $= \alpha g_1(b_{\underline{\theta}}) + (1-\alpha) g_2(\sigma_{\underline{\theta}})$, where $\alpha \in [0,1]$ and g_1, g_2 are non-negative functions. The generalized MSE has been used in response surface design [21] and in minimum bias and variance estimation for nonlinear regression models [22], [23]. Furthermore, since they jointly specify the first two moments of the estimator probability distribution, the pair $(b_{\underline{\theta}}, \sigma^2_{\underline{\theta}})$ provides essential information for constructing and evaluating $\hat{t}$-based hypothesis tests and confidence intervals. Indeed the popular jackknife method was originally introduced by [24], [25] to estimate bias and variance of a statistic and to test whether the statistic has prespecified mean [26] .

An estimator $\hat{t}$ whose bias function $b : \Theta \to \mathbb{R}$ is constant is as good as unbiased since the bias can be removed without knowledge of $\underline{\theta}$. Therefore, in as far as one is interested in fundamental tradeoffs, it is the bias variation which will be of interest. When the density function $f_Y(y;\underline{\theta})$ is sufficiently smooth to guarantee existence of the Fisher information matrix (defined below), $b_{\underline{\theta}}$ is always differentiable regardless of the form of the estimator as long as $E_{\underline{\theta}}[\hat{t}^2]$ is upper bounded [27, Lemma 7.2]. In this case the bias gradient $\nabla b_{\underline{\theta}} : \Theta \to \mathbb{R}^n$ uniquely specifies the bias $b_{\underline{\theta}}$ up to an additive constant

$$b_{\underline{\theta}} = \sum_{k=1}^{n} \int_{\theta_k^o}^{\theta_k} \frac{\partial b_{\underline{u}}}{\partial u_k}\, du_k + b_{\underline{\theta}^o},$$

where $\underline{\theta}^o$ is a point such that the line segment connecting $\underline{\theta}^o$ and $\underline{\theta}$ is contained in Θ – such a point is guaranteed to exist when Θ is convex or star shaped. Thus the gradient function $\nabla b_{\underline{\theta}} = \left[\frac{\partial b_{\underline{\theta}}}{\partial \theta_1}, \ldots, \frac{\partial b_{\underline{\theta}}}{\partial \theta_1}\right]^T$ (a *column* vector) characterizes the unremovable bias component of the bias function.

A.1 Bias Gradient Norm and Maximal Bias

Define the norm or length of the bias gradient vector

$$\delta_{\underline{\theta}} = \|\nabla b_{\underline{\theta}}\|_C \tag{1}$$

where the norm $\|\bullet\|_C$ is is defined in terms of a symmetric positive definite matrix $\mathbf{C}$

$$\|\underline{u}\|_C^2 = \underline{u}^T \mathbf{C} \underline{u}. \tag{2}$$

We will use the notation $\|\underline{u}\|_2$ to denote the Euclidean norm obtained when $\mathbf{C} = \mathbf{I}$.

The norm of the bias gradient at a point $\underline{u} = \underline{\theta}$ is a measure of the sensitivity of the estimator mean $m_{\underline{u}} = E_{\underline{u}}[\hat{t}]$ to changes in $\underline{u}$ over a neighborhood of $\underline{\theta}$. Below we derive a relation between bias gradient norm and maximal bias variation over an arbitrary ellipsoidal neighborhood.

Define the ellipsoidal region of parameter variations $\mathcal{C} = \mathcal{D}(\underline{\theta}, \mathbf{C}) = \{\underline{u} : (\underline{u} - \underline{\theta})^T \mathbf{C}^{-1} (\underline{u} - \underline{\theta}) \leq 1\}$ where $\underline{\theta}$ is a point in Θ and $\mathbf{C}$ is a symmetric positive definite matrix. The maximal width of the ellipsoid is $2\sqrt{\lambda_M^C}$, where λ_M^C is the maximal eigenvalue of $\mathbf{C}$. Assume that the bias function $b_{\underline{u}}$ is continuously twice differentiable and that the magnitude of the eigenvalues of of the Hessian matrix $\nabla^2 b_{\underline{u}} = \nabla \nabla^T b_{\underline{u}}$ are upper bounded over $\underline{u} \in \mathcal{C}$ by a non-negative constant $\alpha < \infty$. Then, using (1) and the Taylor expansion with remainder, the maximal squared variation of the bias $b_{\underline{u}}$ over $\mathcal{D}$ is

$$\max_{\underline{u} \in \mathcal{C}} |b_{\underline{u}} - b_{\underline{\theta}}|^2 = \max_{\underline{u} \in \mathcal{C}} |\nabla^T b_{\underline{\theta}} \Delta \underline{u} + \tfrac{1}{2} \Delta \underline{u}^T \nabla^2 b_{\underline{\xi}} \Delta \underline{u}|^2 \tag{3}$$

where $\Delta \underline{u} = \underline{u} - \underline{\theta}$ and $\underline{\xi}$ is a point along the line segment joining $\underline{\theta}$ and $\underline{u}$. Now, expanding the square on the right hand side of (3) and collecting terms we obtain

$$\max_{\underline{u} \in \mathcal{C}} |b_{\underline{u}} - b_{\underline{\theta}}|^2 = \max_{\underline{u} \in \mathcal{C}} |\nabla^T b_{\underline{\theta}} \Delta \underline{u}|^2 (1 + \epsilon_o), \tag{4}$$

where $|\epsilon_o| \leq \rho(1 + 0.25\rho)$ and $\rho = \lambda_M^C \frac{\alpha}{\sqrt{\nabla^T b_{\underline{\theta}} \mathbf{C} \nabla b_{\underline{\theta}}}}$. Defining $\Delta \underline{\tilde{u}} = \mathbf{C}^{-\frac{1}{2}} \Delta \underline{u}$ and using the Cauchy-Schwarz inequality we obtain

$$\begin{aligned} \max_{\underline{u} \in \mathcal{C}} |\nabla^T b_{\underline{\theta}} \Delta \underline{u}|^2 &= \max_{\|\Delta \underline{\tilde{u}}\| \leq 1} |\nabla^T b_{\underline{\theta}} \mathbf{C}^{\frac{1}{2}} \Delta \underline{\tilde{u}}|^2 \\ &= \max_{\Delta \underline{\tilde{u}}} \frac{|\nabla^T b_{\underline{\theta}} \mathbf{C}^{\frac{1}{2}} \Delta \underline{\tilde{u}}|^2}{\|\Delta \underline{\tilde{u}}\|_2} \\ &= \nabla^T b_{\underline{\theta}} \mathbf{C} \nabla b_{\underline{\theta}}. \end{aligned} \tag{5}$$

Therefore, combining (5)-(3)

$$\|\nabla b_{\underline{\theta}}\|_C^2 = \max_{\underline{u} \in \mathcal{C}} |b_{\underline{u}} - b_{\underline{\theta}}|^2 (1 + \epsilon), \tag{6}$$

where $\epsilon = O(\rho[1 + 0.25\rho])$. Hence we see that when $\rho \ll 1$ the norm $\|\nabla b_{\underline{u}}\|_C$ is approximately equal to the maximal bias variation over the ellipsoidal neighborhood $\mathcal{D}(\underline{\theta}, \mathbf{C})$ of $\underline{\theta}$. Note that this occurs when when the product of the ellipsoid width λ_M^C and the ratio of the curvature α of the bias function to the bias gradient norm $\sqrt{\nabla^T b_{\underline{\theta}} \mathbf{C} \nabla b_{\underline{\theta}}}$ is small. For the special case where the bias is a linear function ($b_{\underline{\theta}} = \underline{L}^T \underline{\theta} - c$, $\underline{L} \in \mathbb{R}$), $\rho = 0$ in which case the relation between bias gradient norm (1) and maximal bias variation (3) is exact.

The above discussion suggests that the choice of norm $\| \bullet \|_C$ should reflect the range $\mathcal{D}$ of joint parameter variations which are of interest to the user. This will be illustrated in Section IV.

II. Unachievable Regions

For any estimator with bias gradient norm $\delta_{\underline{\theta}}$ and variance $\sigma_{\underline{\theta}}^2$ we plot the pair $(\delta_{\underline{\theta}}, \sigma_{\underline{\theta}})$ as a coordinate in the plane $\mathbb{R}^2$. We will call this parameterization of the plane the *delta-sigma* or $\delta\sigma$ plane. A region of the $\delta\sigma$ plane is called unachievable if no estimator can exist having coordinates in this region. While no non-empty unachievable region can exist in the bias-variance plane parameterized by $(b_{\underline{\theta}}, \sigma_{\underline{\theta}})$, we will show that interesting unachievable regions almost always exist in the delta-sigma plane.

A. The Biased CR Bound

The Cramer-Rao lower bound on estimator variance, first introduced in Frechet [28] and later in Darmois [29], Cramer [30], and Rao [31], is commonly used to bound the variance of unbiased estimators. For a biased estimator $\hat{t}$ of $t_{\underline{\theta}}$ with mean $m_{\underline{\theta}} = E_{\underline{\theta}}[\hat{t}]$ the CR bound has the following form, referred to here as the *biased CR bound*

$$\begin{aligned} \sigma_{\underline{\theta}}^2 &\geq [\nabla m_{\underline{\theta}}]^T \mathbf{F}_Y^+ [\nabla m_{\underline{\theta}}] \\ &= [\nabla t_{\underline{\theta}} + \nabla b_{\underline{\theta}}]^T \mathbf{F}_Y^+ [\nabla t_{\underline{\theta}} + \nabla b_{\underline{\theta}}], \end{aligned} \tag{7}$$

where $\mathbf{F}_Y = \mathbf{F}_Y(\underline{\theta})$ is the $n \times n$ Fisher information matrix

$$\mathbf{F}_Y = E_{\underline{\theta}} \left\{ [\nabla_{\underline{\theta}} \ln f_Y(Y; \underline{\theta})] \, [\nabla_{\underline{\theta}} \ln f_Y(Y; \underline{\theta})]^T \right\},$$

and $\mathbf{F}_Y^+$ denotes the Moore-Penrose pseudo-inverse matrix of the possibly singular matrix $\mathbf{F}_Y$.

The non-singular-$\mathbf{F}_Y$ form of the biased CR bound has been around for some time, e.g. [32]. The more general pseudo-inverse-$\mathbf{F}_Y$ form given in (7) is less well known but can be easily derived by identifying $U = \hat{t} - t_{\underline{\theta}}$ and $V = \nabla_{\underline{\theta}} \ln f_Y(Y; \underline{\theta})$ in the relation [33, Lemma 1]

$$E_{\underline{\theta}} [UU^T] \geq E_{\underline{\theta}} [UV^T] \left(E_{\underline{\theta}} [VV^T] \right)^+ E_{\underline{\theta}} [VU^T],$$

and using the well known identities $E_{\underline{\theta}}[\nabla_{\underline{\theta}} \ln f_Y(Y; \underline{\theta})] = \underline{0}$ and $E_{\underline{\theta}}[\hat{t} \nabla_{\underline{\theta}} \ln f_Y(Y; \underline{\theta})] = \nabla m_{\underline{\theta}}$ (easily derivable from (18) below).

The biased CR bound (7) only applies to the class of estimators $\hat{t}$ which have a particular bias gradient function $\nabla b_{\underline{\theta}}$. Therefore (7) cannot be used to simultaneously bound the variance of several estimators, each of which have different but comparable bias gradients.

B. The Uniform CR Bound

In [15] a "uniform" CR bound was presented as a way to study the reliability of the unbiased CR bound under conditions of very small estimator bias. In [34] this uniform bound was used to trace out curves over the sigma-delta plane which includes both large and small biases. The following theorem extends the results of [15] and [34] to allow singular Fisher information matrices, arbitrary weighted Euclidean norm $\| \bullet \|_C$, and arbitrary differentiable function $t_{\underline{\theta}}$. For a proof of this theorem see Appendix A.

Theorem 1: Let $\hat{t}$ be an estimator of the scalar differentiable function $t_{\underline{\theta}}$ of the parameter $\underline{\theta} = [\theta_1, \ldots, \theta_n]^T$. For a fixed $\delta \geq 0$ let the bias gradient of $\hat{t}$ satisfy the norm constraint $\|\nabla \underline{b}_\theta\|_C \leq \delta$, where $\mathbf{C}$ is an arbitrary $n \times n$ symmetric positive definite matrix. Define $\mathcal{P}_B$ as the $n \times n$ matrix which projects onto the column space of $\mathbf{B} = \mathbf{C}^{-\frac{1}{2}}\mathbf{F}_Y^{+}\mathbf{C}^{-\frac{1}{2}}$. Then the variance of $\hat{t}$ satisfies:

$$\sigma_{\underline{\theta}}^2 \geq B(\underline{\theta}, \delta), \tag{8}$$

where if $\delta^2 \geq \nabla^T t_{\underline{\theta}} \mathbf{C}^{\frac{1}{2}} \mathcal{P}_B \mathbf{C}^{\frac{1}{2}} \nabla t_{\underline{\theta}}$ then $B(\underline{\theta}, \delta) = 0$, while if $\delta^2 < \nabla^T t_{\underline{\theta}} \mathbf{C}^{\frac{1}{2}} \mathcal{P}_B \mathbf{C}^{\frac{1}{2}} \nabla t_{\underline{\theta}}$ then:

$$\begin{aligned} B(\underline{\theta}, \delta) &= \left[\nabla t_{\underline{\theta}} + \underline{d}_{min}\right]^T \mathbf{F}_Y^{+} \left[\nabla t_{\underline{\theta}} + \underline{d}_{min}\right], \\ &= \lambda^2 \nabla^T t_{\underline{\theta}} \mathbf{C} \left[\lambda \mathbf{C} + \mathbf{F}_Y^{+}\right]^{-1} \mathbf{F}_Y^{+} \left[\lambda \mathbf{C} + \mathbf{F}_Y^{+}\right]^{-1} \mathbf{C} \nabla t_{\underline{\theta}}, \end{aligned} \tag{9}$$

where in (9)

$$\underline{d}_{min} = -\left[\lambda \mathbf{C} + \mathbf{F}_Y^{+}\right]^{-1} \mathbf{F}_Y^{+} \nabla t_{\underline{\theta}}. \tag{10}$$

In (9) and (10) $\lambda > 0$ is determined by the unique positive solution of $g(\lambda) = \delta^2$ where

$$g(\lambda) = \nabla^T t_{\underline{\theta}} \mathbf{F}_Y^{+} \left[\lambda \mathbf{C} + \mathbf{F}_Y^{+}\right]^{-1} \mathbf{C} \left[\lambda \mathbf{C} + \mathbf{F}_Y^{+}\right]^{-1} \mathbf{F}_Y^{+} \nabla t_{\underline{\theta}}. \tag{11}$$

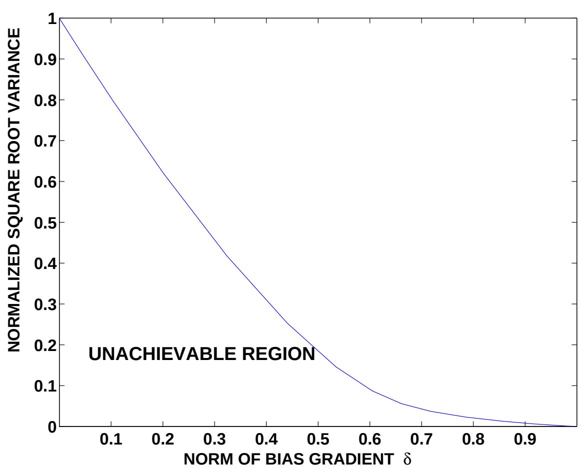

Fig. 1. The Normalized Uniform CR bound on the $\delta\sigma$ tradeoff plane for a specified value of $\underline{\theta}$.

By tracing out the family of points $\{(\delta, \sqrt{B(\underline{\theta}, \delta)}) : \delta \geq 0\}$ one obtains a curve in the $\delta\sigma$ plane for a particular $\underline{\theta} \in \Theta$. The curve is always monotone non-increasing in δ. Since $B(\underline{\theta}, \delta)$ is a lower bound on $\sigma_{\underline{\theta}}^2$ the region below the curve defines an unachievable region. Figure 1 shows a typical delta-sigma tradeoff curve plotted in terms of normalized standard deviation $\sigma = \sqrt{\frac{B(\underline{\theta},\delta)}{B(\underline{\theta},0)}}$. If an estimator lies on the curve then lower variance can only be bought at the price of increased bias gradient and vice versa. For this reason we call this curve the *delta-sigma tradeoff curve.*

It is important to point out that the delta-sigma tradeoff curve can be generated without solving the non-linear equation (11), which generally must be solved numerically. It is much easier to continuously vary λ over the range $(0, \infty)$ and sweep out the curve by using the λ-parameterizations of $g(\lambda) = \delta^2$ and $B(\underline{\theta}, \delta)$ specified by relations (11) and (9), respectively.

Comments:

- The uniform bound $B(\underline{\theta}, \delta)$ is always less than or equal to the unbiased CR bound $B(\underline{\theta}; 0) = \nabla^T t_{\underline{\theta}} \mathbf{F}_Y^{+} \nabla t_{\underline{\theta}}$. The slope of $B(\underline{\theta}, \delta)$ at $\delta = 0$ gives a bias sensitivity index η for the unbiased CR bound. For non-singular $\mathbf{F}_Y$ and single component estimation ($t_{\underline{\theta}} = \theta_1$) it is shown in [15] that $\eta = 2\sqrt{1 + \underline{c}^T \mathbf{F}_S^{-2} \underline{c}}$, where $\underline{c}$ is the first column of $\mathbf{F}_Y$ and $\mathbf{F}_S$ is the principal minor of $[\mathbf{F}_Y]_{11}$. Large values of this index indicate that the unbiased form of the CR bound is not reliable for estimators which may have very small, and perhaps even unmeasurable, biases.
- The orthogonal projection $\mathcal{P}_B$ can be expressed either as $\mathcal{P}_B = \mathbf{B}[\mathbf{B}^T\mathbf{B}]^{+}\mathbf{B}^T = \mathbf{B}^{+}\mathbf{B} = \mathbf{B}\mathbf{B}^{+}$, or via the eigendecomposition of $\mathbf{B}$ as $\mathcal{P}_B = \sum_{i=1}^{r} \underline{\xi}\underline{\xi}^T$, where r is the rank of $\mathbf{F}_Y$ and $\{\underline{\xi}\}_{i=1}^{r}$ are the orthonormal eigenvectors associated with the non-zero eigenvalues of $\mathbf{B}$. By using properties of the Moore-Penrose pseudo-inverse it can be shown that $\nabla^T t_{\underline{\theta}} \mathbf{C}^{\frac{1}{2}} \mathcal{P}_B \mathbf{C}^{\frac{1}{2}} \nabla t_{\underline{\theta}} = \nabla^T t_{\underline{\theta}} \mathbf{F}_Y^{+} [\mathbf{C}^{-\frac{1}{2}} \mathbf{F}_Y^{+}]^{+} \mathbf{C}^{\frac{1}{2}} \nabla t_{\underline{\theta}}$.
- When $\mathbf{F}_Y$ is non-singular, $\mathbf{F}_Y^{+} = \mathbf{F}_Y^{-1}$, $\mathcal{P}_B = \mathbf{I}$, $\nabla^T t_{\underline{\theta}} \mathbf{C}^{\frac{1}{2}} \mathcal{P}_B \mathbf{C}^{\frac{1}{2}} \nabla t_{\underline{\theta}} = \|\nabla t_{\underline{\theta}}\|_C^2$ and (9)-(11) of Theorem 1 reduce to

$$\begin{aligned} B(\underline{\theta}, \delta) &= [\nabla t_{\underline{\theta}} + \underline{d}_{min}]^T \mathbf{F}_Y^{-1} [\nabla t_{\underline{\theta}} + \underline{d}_{min}] \\ &= \lambda^2 \nabla^T t_{\underline{\theta}} \left[\mathbf{C}^{-1} + \lambda \mathbf{F}_Y\right]^{-1} \mathbf{F}_Y \left[\mathbf{C}^{-1} + \lambda \mathbf{F}_Y\right]^{-1} \nabla t_{\underline{\theta}}, \end{aligned} \tag{12}$$

  where,

$$\underline{d}_{min} = -\mathbf{C}^{-1} \left[\mathbf{C}^{-1} + \lambda \mathbf{F}_Y\right]^{-1} \nabla t_{\underline{\theta}}, \tag{13}$$

  and $\lambda > 0$ is given by the unique positive solution of $g(\lambda) = \delta^2$ where

$$g(\lambda) = \nabla^T t_{\underline{\theta}} \left[\mathbf{C}^{-1} + \lambda \mathbf{F}_Y\right]^{-1} \mathbf{C}^{-1} \left[\mathbf{C}^{-1} + \lambda \mathbf{F}_Y\right]^{-1} \nabla t_{\underline{\theta}} \tag{14}$$

  When $\mathbf{C} = \mathbf{I}$ and $t_{\underline{\theta}} = \theta_1$ these are identical to the results obtained in [15].
- In Theorem 1, $\underline{d}_{min}$ defined in (10) is an optimal bias gradient in the sense that it minimizes the biased CR bound (7) over all vectors $\nabla b_{\underline{\theta}}$ satisfying the constraint $\|\nabla b_{\underline{\theta}}\|_C \leq \delta$. The bound is independent of the particular estimator bias as long as the bias gradient norm constraint holds. From the proof of Theorem 1, if $\delta^2 \geq \nabla^T t_{\underline{\theta}} \mathbf{C}^{\frac{1}{2}} \mathcal{P}_B \mathbf{C}^{\frac{1}{2}} \nabla t_{\underline{\theta}}$, then the minimizing bias gradient is of the form $\underline{d}_{min} = -\mathcal{P}_B \mathbf{C}^{\frac{1}{2}} \nabla t_{\underline{\theta}} + \underline{\phi}$, where $\underline{\phi}$ is any vector satisfying $\mathbf{B}\underline{\phi} = \underline{0}$, and $\|\underline{\phi}\|^2 \leq \delta^2 - \nabla^T t_{\underline{\theta}} \mathbf{C}^{\frac{1}{2}} \mathcal{P}_B \mathbf{C}^{\frac{1}{2}} \nabla t_{\underline{\theta}}$. Thus for the case of singular $\mathbf{F}_Y$ there exist many optimal bias gradients.
- An estimator is said to locally achieve a bound in a neighborhood of a point $\underline{\theta}$ if the estimator achieves

the bound whenever the true parameter lies in the neighborhood. It has been shown [15] that if $\mathbf{F}_Y$ is non-singular, if δ is small, if $t_{\underline{\theta}} = \theta_1$, and if the unbiased matrix CR bound is locally achievable by an unbiased estimator $\hat{\underline{\theta}}^*$ in a neighborhood of a point $\underline{\theta}$, then one can construct an estimator that locally achieves the uniform bound in this neighborhood by introducing a small amount of bias into $\hat{\underline{\theta}}^*$. However, since unbiased estimators may not exist for singular $\mathbf{F}_Y$, the uniform CR bound for singular $\mathbf{F}_Y$ may not be locally achievable. An example where the bound is globally achievable over all Θ is presented in Section IV.

- While we will not use it in this paper, a more general form of Theorem 1 holds for the case that $\mathbf{C}$ may be non-negative definite. This situation is relevant for cases where the user does not wish to penalize the estimator for high bias variation over certain hyperplanes in the parameter space. For example when estimation of image contrast is of interest, spatially homogeneous biases may be tolerable and $\mathbf{C}$ may be chosen to be of rank $n-1$ having the vector $\underline{1} = [1, \ldots, 1]^T$ in its nullspace. Let $B(\underline{\theta}, \delta)$, $\underline{d}_{min}$ and $g(\lambda)$ be as defined in Theorem 1. Assume that $\mathbf{C}$ is non-negative definite but $\mathbf{F}_Y^+ + \lambda \mathbf{C}$ is positive definite for $0 < \lambda < \infty$. For fixed $\delta > 0$ let the bias gradient of $\hat{t}$ satisfy the semi-norm constraint $\|\nabla b_{\underline{\theta}}\|_C \leq \delta$. Then $\mathrm{var}_{\underline{\theta}}(\hat{t}) \geq B^*(\underline{\theta}, \delta)$, where

$$B^*(\underline{\theta}, \delta) = \begin{cases} B(\underline{\theta}, g(\infty)), & 0 \leq \delta^2 < g(\infty) \\ B(\underline{\theta}, \delta), & g(\infty) \leq \delta^2 \leq g(0) \\ 0, & g(0) < \delta^2 < \infty \end{cases}$$

and $g(\infty) = \lim_{\lambda \to \infty} g(\lambda)$ and $g(0) = \lim_{\lambda \to 0} g(\lambda)$.

C. Recipes for Uniform Bound Computation

As written in Theorem 1 expressions (9)-(11) are not in the most convenient form for computation as they involve several matrix multiplications and inversions. An equivalent form for the pair $B(\underline{\theta}, \delta)$ and $g(\lambda)$ in (9) and (11) was obtained in the process of proving the theorem ((46) and (47))

$$\begin{aligned} B(\underline{\theta}, \delta) &= \lambda^2 \nabla^T t_{\underline{\theta}} \mathbf{C}^{\frac{1}{2}} [\lambda \mathbf{I} + \mathbf{B}]^{-1} \mathbf{B} [\lambda \mathbf{I} + \mathbf{B}]^{-1} \mathbf{C}^{\frac{1}{2}} \nabla t_{\underline{\theta}} \\ g(\lambda) &= \nabla^T t_{\underline{\theta}} \left[\mathbf{C}^{\frac{1}{2}} \mathbf{B} (\lambda \mathbf{I} + \mathbf{B})^{-2} \mathbf{B} \mathbf{C}^{\frac{1}{2}} \right] \nabla t_{\underline{\theta}} = \delta^2, \end{aligned}$$

where $\mathbf{B} = \mathbf{C}^{-\frac{1}{2}} \mathbf{F}_Y^+ \mathbf{C}^{-\frac{1}{2}}$. If an eigendecomposition of the matrix $\mathbf{B}$ is available, the delta-sigma tradeoff curve can be efficiently computed by sweeping out λ in the following pair of weighted sums of inner products

$$B(\underline{\theta}, \delta) = \sum_{i=1}^n \frac{\lambda^2 \beta_i}{(\lambda + \beta_i)^2} |\nabla^T t_{\underline{\theta}} \mathbf{C}^{\frac{1}{2}} \underline{\xi}_i|^2 \quad (15)$$

$$\delta^2 = \sum_{i=1}^n \frac{\beta_i^2}{(\lambda + \beta_i)^2} |\nabla^T t_{\underline{\theta}} \mathbf{C}^{\frac{1}{2}} \underline{\xi}_i|^2, \quad (16)$$

where β_i and $\underline{\xi}_i$ denote an eigenvalue and eigenvector of $\mathbf{B}$.

When $\mathbf{F}_Y$ is ill-conditioned the computation of $\mathbf{B}$ may be numerically unstable. In the case of non-singular $\mathbf{F}_Y$ a simple algebraic manipulation in (12)-(14) yields

$$\begin{aligned} B(\underline{\theta}, \delta) &= \lambda^2 \nabla^T t_{\underline{\theta}} \mathbf{C}^{\frac{1}{2}} [\mathbf{I} + \lambda \mathbf{G}]^{-1} \mathbf{G} [\mathbf{I} + \lambda \mathbf{G}]^{-1} \mathbf{C}^{\frac{1}{2}} \nabla t_{\underline{\theta}}, \\ \delta^2 &= \nabla^T t_{\underline{\theta}} \mathbf{C}^{\frac{1}{2}} [\mathbf{I} + \lambda \mathbf{G}]^{-2} \mathbf{C}^{\frac{1}{2}} \nabla t_{\underline{\theta}}, \end{aligned} \quad (17)$$

where $\mathbf{G} = \mathbf{B}^{-1} = \mathbf{C}^{\frac{1}{2}} \mathbf{F}_Y \mathbf{C}^{\frac{1}{2}}$. Note that computation of the form (17) requires only one matrix inversion $[\mathbf{I} + \lambda \mathbf{G}]^{-1}$. Since $\lambda > 0$ and $\mathbf{F}_Y$ is positive definite this inversion is well conditioned except if λ is very large.

The eigendecomposition of $\mathbf{G}$ can be used in (17) to produce a pair of expressions similar to (15)-(16) for computing the delta-sigma tradeoff curve for non-singular $\mathbf{F}_Y$. Alternatively, the right hand sides of (17) can be approximated by using iterative equation solving methods such as Gauss-Seidel (GS) or preconditioned conjugate gradient (CG) algorithms [35]. See [45], [36] for a more detailed discussion of the application of iterative equation solvers to CR bound approximation. This approach can be implemented in the following sequence of steps.

1. Select $\lambda \in (0, \infty)$.
2. Compute $\underline{x} = [\mathbf{I} + \lambda \mathbf{F}_Y \mathbf{C}]^{-1} \nabla t_{\underline{\theta}}$ by applying CG or GS iterations to solve the following linear equation for $\underline{x}$:
$$[\mathbf{I} + \lambda \mathbf{F}_Y \mathbf{C}] \underline{x} = \nabla t_{\underline{\theta}}.$$
3. Compute $\underline{y} = \mathbf{C} \underline{x}$.
4. Compute the point $(\delta, B(\underline{\theta}, \delta))$ via
$$\begin{aligned} B(\underline{\theta}, \delta) &= \lambda^2 \underline{y}^T \mathbf{F}_Y \underline{y} \\ \delta &= \sqrt{\underline{x}^T \underline{y}}. \end{aligned}$$

Since step 2 must be repeated for each value of λ, this method is competitive when one is interested in evaluation of the curve $B(\underline{\theta}, \delta)$ at only a small number of values of $\delta = \delta(\lambda)$. When a denser sampling of the curve is desired an eigendecomposition method, e.g. as in (15)-(16), becomes more attractive since, once the quantities β_i and $|\nabla^T t_{\underline{\theta}} \mathbf{C}^{\frac{1}{2}} \underline{\xi}_i|^2$ are available, the curve can be swept out over λ without performing additional vector operations.

III. Estimation of Bias Gradient Norm

To be able to compare the performance of an estimator against the uniform CR bound of Theorem 1, we need to determine the estimator variance and the bias gradient length. In most cases the bias gradient cannot be determined analytically and it is therefore important to have a computationally efficient method to estimate it either experimentally or via simulations. A brute force estimate would be to estimate the finite difference approximation $\nabla b_{\underline{\theta}} \approx \frac{1}{\epsilon} \left[b_{\underline{\theta}+\epsilon \underline{e}_1} - b_{\underline{\theta}}, \ldots, b_{\underline{\theta}+\epsilon \underline{e}_n} - b_{\underline{\theta}} \right]$ but this requires performing a simulation run for each coordinate perturbation $\underline{\theta} + \epsilon \underline{e}_k$. In the following we describe a more direct method for estimating the bias gradient which does not require performing multiple simulation runs nor does it require making a finite difference approximation. The method is based on the fact that for any random variable Z with finite mean

$$\begin{aligned} E_{\underline{\theta}}\left[Z\frac{\partial \ln f_Z(Z;\underline{\theta})}{\partial\theta_j}\right] &= \int_Z z\frac{\partial f_Z(z;\underline{\theta})}{\partial\theta_j}\,dz \\ &= \frac{\partial}{\partial\theta_j}\int_Z z f_Z(z;\underline{\theta})\,dz \\ &= \frac{\partial}{\partial\theta_j}E_{\underline{\theta}}(Z)\,. \end{aligned} \tag{18}$$

Thus in particular we have the following relation

$$\nabla b_{\underline{\theta}} = E_{\underline{\theta}}\left[\hat{t}(Y)\nabla_{\underline{\theta}}\ln f_Y(Y;\underline{\theta})\right] - \nabla t_{\underline{\theta}}.$$

Since $E_{\underline{\theta}}\left[\nabla_{\underline{\theta}}\ln f_Y(Y;\underline{\theta})\right] = 0$, an equivalent relation is

$$\nabla b_{\underline{\theta}} = E_{\underline{\theta}}\left[(\hat{t}(Y)-\zeta)\nabla_{\underline{\theta}}\ln f_Y(Y;\underline{\theta})\right] - \nabla t_{\underline{\theta}}, \tag{19}$$

for any random variable ζ statistically independent of Y. As explained in the following discussion, the quantity ζ can be used to control the variance of the bias gradient estimate.

Substituting sample averages for ensemble averages in (19) we obtain the following unbiased and consistent estimator of the bias gradient vector $\nabla b_{\underline{\theta}}$

$$\widehat{\nabla b_{\underline{\theta}}} = \frac{1}{L}\sum_{i=1}^{L}\left(\hat{t}(Y_i)-\zeta_i\right)\nabla_{\underline{\theta}}\ln f_Y(Y_i;\underline{\theta}) - \nabla t_{\underline{\theta}}, \tag{20}$$

where $\{Y_i\}_{i=1}^L$ is a set of i.i.d. realizations from $f_Y(y;\underline{\theta})$. In (20) $\{\zeta_i\}_{i=1}^L$ is any sequence of i.i.d. random variables such that Y_i, ζ_i are statistically independent for each i.

It can be shown that when $\zeta_i = 0$ for all i the covariance matrix of $\widehat{\nabla b_{\underline{\theta}}}$ is the matrix sum

$$\begin{aligned} S(\widehat{\nabla b_{\underline{\theta}}}) &= \frac{1}{L}\mathrm{cov}_{\underline{\theta}}\left((\hat{t}(Y_i)-m_{\underline{\theta}})\nabla\ln f_{Y_i}(Y_i;\underline{\theta})\right) \\ &\quad+\frac{1}{L}\left[2m_{\underline{\theta}}\mathbf{R} + m_{\underline{\theta}}^2\mathbf{F}_Y\right] \end{aligned} \tag{21}$$

where
$\mathbf{R} = E_{\underline{\theta}}\left[(\hat{t}(Y_i)-m_{\underline{\theta}})[\nabla_{\underline{\theta}}\ln f_{Y_i}(Y_i;\underline{\theta})][\nabla_{\underline{\theta}}^T\ln f_{Y_i}(Y_i;\underline{\theta})]^T\right]$, and $\mathbf{F}_Y = E_{\underline{\theta}}\left([\nabla_{\underline{\theta}}\ln f_Y(Y_i;\underline{\theta})][\nabla_{\underline{\theta}}\ln f_Y(Y_i;\underline{\theta})]^T\right)$ is the single trial Fisher information. The first term on the right hand side (RHS) of (21) decays as $1/L$ and is independent of the mean $m_{\underline{\theta}}$. The second term (in brackets "[ ]) also decays as $1/L$ but is unbounded in the mean $m_{\underline{\theta}}$. It is easily shown that this term can be eliminated by setting $\zeta_i = m_{\underline{\theta}} =$ constant but this is not a practical since the mean $m_{\underline{\theta}} = E_{\underline{\theta}}[\hat{t}(Y_i)]$ is unknown to the user. However, we can use the punctured sample mean estimate of $m_{\underline{\theta}}$, $\zeta_i = \frac{1}{L-1}\sum_{\substack{j=1\\ j\neq i}}^L \hat{t}(Y_j)$ which, as required for the validity of (19), is statistically independent of Y_i. Substitution of this ζ_i into (20) and simplifying gives the following unbiased and consistent sample mean estimate of $\nabla b_{\underline{\theta}}$

$$\begin{aligned} &\widehat{\nabla b_{\underline{\theta}}} \\ &= \frac{1}{L-1}\sum_{i=1}^{L}\left(\hat{t}(Y_i) - \frac{1}{L}\sum_{j=1}^{L}\hat{t}(Y_j)\right)\nabla_{\underline{\theta}}\ln f_Y(Y_i;\underline{\theta}) - \nabla t_{\underline{\theta}}. \end{aligned} \tag{22}$$

A simple calculation shows that the covariance of (22) is the matrix sum

$$\begin{aligned} S(\widehat{\nabla b_{\underline{\theta}}}) &= \frac{1}{L}\mathrm{cov}_{\underline{\theta}}\left((\hat{t}(Y_i)-m_{\underline{\theta}})\nabla_{\underline{\theta}}\ln f_Y(Y_i;\underline{\theta})\right) \\ &+ \frac{1}{L(L-1)}\left[\mathrm{var}_{\underline{\theta}}(\hat{t}(Y_i))\mathbf{F}_Y + \nabla b_{\underline{\theta}}\nabla^T b_{\underline{\theta}}\right] \end{aligned} \tag{23}$$

Note that the second term in (23) depends on $m_{\underline{\theta}}$ only through its gradient and decreases to zero at the much faster asymptotic rate of $\frac{1}{L^2}$ as compared to the rate $\frac{1}{L}$ in (21).

A. A Bootstrap Estimator for Bias Gradient Norm

A natural "method-of-moments" estimate for $\delta_{\underline{\theta}}^2 = \|\nabla b_{\underline{\theta}}\|_C^2$ is the norm squared of the unbiased estimator $\hat{\delta}^2 = \|\widehat{\nabla b_{\underline{\theta}}}\|_C^2$ (20). It can easily be shown that this estimator is biased with bias equal to $E_{\underline{\theta}}\left[\|\widehat{\nabla b_{\underline{\theta}}} - \nabla b_{\underline{\theta}}\|_C^2\right] = \mathrm{trace}\{S(\widehat{\nabla b_{\underline{\theta}}})\}$ which, in view of (21) or (23), decays to zero only as $1/L$. Below we present a norm estimator based on the bootstrap resampling methodology whose bias decays at a faster rate.

Let $Y_1^*, \ldots, Y_L^*$ denote a *bootstrap sample* obtained by randomly resampling the realizations $Y_1 = y_1, \ldots, Y_L = y_L$ with replacement. Given the estimate $\hat{\delta}^2 = \hat{\delta}^2(y_1, \ldots, y_L) = \|\widehat{\nabla b_{\underline{\theta}}}\|_C^2$, the bootstrap estimate of $\delta_{\underline{\theta}}^2$ is defined as the expectation of $\hat{\delta}_*^2 = \hat{\delta}^2(Y_1^*, \ldots, Y_L^*)$ with respect to the resampling distribution [37]

$$E_*[\hat{\delta}_*^2] = \sum_*\hat{\delta}^2(Y_1^*, \ldots, Y_L^*)\binom{L}{c_1\ldots c_L}L^{-L}. \tag{24}$$

In (24) c_i is the number of times the value y_i appears in the set $\{Y_j^*\}$ and $\sum_*$ denotes a summation over all non-negative integers $c_1, \ldots, c_L$ satisfying $\sum_{i=1}^L c_i = L$. The bootstrap estimate of the bias of the estimator $\hat{\delta}^2$ is defined as $E_*[\hat{\delta}_*^2] - \hat{\delta}^2$ which leads to the bias corrected estimator $\hat{\delta}_c^2$

$$\hat{\delta}_c^2 = 2\hat{\delta}^2 - E_*[\hat{\delta}_*^2]. \tag{25}$$

Due to the simple quadratic dependence of $\hat{\delta}^2$ on the single sample quantities $\hat{t}(y_i)\nabla_{\underline{\theta}}\ln f_Y(y_i;\underline{\theta})$, $i = 1, \ldots, L$, the expectation (24) can be expressed in analytical form (see Appendix B) leading to the bias corrected estimate

$$\hat{\delta}_c^2 = \|\widehat{\nabla b_{\underline{\theta}}}\|_C^2 - \frac{1}{L^2}\sum_{i=1}^{L}\|\widehat{\nabla b_{\underline{\theta}}}(y_i) - \widehat{\nabla b_{\underline{\theta}}}\|_C^2, \tag{26}$$

where $\widehat{\nabla b_{\underline{\theta}}}(y_i)$ is the estimate (20) based on a single sample y_i $(L=1)$ with $\zeta_i = 0$. The bias of $\hat{\delta}_c^2$ is equal to

$$E_{\underline{\theta}}\left[\hat{\delta}_c^2\right] - \|\nabla b_{\underline{\theta}}\|_C^2 = \frac{1}{L}E_{\underline{\theta}}[\|\widehat{\nabla b_{\underline{\theta}}} - \nabla b_{\underline{\theta}}\|_C^2],$$

which, relative to the estimator $\|\widehat{\nabla b_{\underline{\theta}}}\|_C^2$, decays to zero at the much faster rate of $1/L^2$. However, if L is insufficiently large the bootstrap estimator $\hat{\delta}_c^2$ may take on negative values.

IV. Application to Inverse Problems

We use the theory developed above to perform a study of fundamental bias-variance tradeoffs for three general classes of inverse problems. First we consider well-posed linear Gaussian inverse problems which have non-singular Fisher information. Next we consider ill-posed Gaussian inverse problems where the Fisher matrix is singular. For these two linear applications an exact analysis is possible since all curves in the delta-sigma tradeoff plane have analytic expressions. Finally we study a non-linear Poisson inverse problem to illustrate the empirical bias-gradient norm approximations discussed in the previous section.

A. Linear Gaussian Model

Assume that the observation consists of a vector $Y = \underline{Y} \in \mathbb{R}^m$ which obeys the linear Gaussian model:

$$\underline{Y} = \mathbf{A}\underline{\theta} + \underline{\epsilon}, \tag{27}$$

where $\mathbf{A}$ is an $m \times n$ coefficient matrix, called the system matrix, $\underline{\theta}$ is an unknown source, and $\underline{\epsilon}$ is a vector of zero mean Gaussian random variables with positive definite covariance matrix $\mathbf{\Sigma}$. For concreteness we will refer to θ_i as the intensity of the source at pixel i. The Fisher information matrix has the well known form [19]

$$\mathbf{F}_Y = \mathbf{A}^T \mathbf{\Sigma}^{-1} \mathbf{A}. \tag{28}$$

This matrix is non-singular when $\mathbf{A}$ is of full column rank n. We will consider estimation of the linear combination $t_{\underline{\theta}} = \underline{h}^T \underline{\theta}$ where $\underline{h}$ is a fixed non-zero vector in $\mathbb{R}^n$. Since $\mathbf{F}_Y$ and $\nabla t_{\underline{\theta}} = \underline{h}$ are not functionally dependent on $\underline{\theta}$ the uniform bound $B(\underline{\theta}, \delta)$ will not depend on the specific form of the unknown source $\underline{\theta}$.

To demonstrate the achievability of the fundamental delta-sigma tradeoff curve we consider the quadratically penalized maximum likelihood (QPML) estimator. The QPML strategy is frequently used in order to obtain stable solutions in the presence of small variations in experimental conditions [38] and to incorporate parameter constraints or *a priori* information [39]. For the linear Gaussian problem (27) the QPML estimator of the linear combination $t_{\underline{\theta}} = \underline{h}^T \underline{\theta}$ is $\hat{t} = \underline{h}^T \underline{\hat{\theta}}$ where $\underline{\hat{\theta}}$ minimizes the following objective function over $\underline{\theta}$

$$[\underline{Y} - \mathbf{A}\underline{\theta}]^T \mathbf{\Sigma}^{-1} [\underline{Y} - \mathbf{A}\underline{\theta}] + \beta \underline{\theta}^T \mathbf{P} \underline{\theta}. \tag{29}$$

In the above $\beta > 0$ is a regularization parameter and $\mathbf{P}$ is a symmetric nonnegative-definite penalty matrix. For ill-conditioned or singular $\mathbf{A}$ the penalty improves the numerical stability of the matrix inversion $[\mathbf{F}_Y + \beta\,\mathbf{P}]^{-1}$ in (30) below by lowering its condition number. The simplest choice for the penalty matrix $\mathbf{P}$ is the identity $\mathbf{I}$, which yields a class of energy penalized least squares estimators variously known as Tikonov regularized least squares in the inverse problem literature [38], and shrinkage estimation or ridge regression in the multivariate statistics literature [6]. A popular choice in imaging applications is to use a non-diagonal differencing type operator to enforce smoothness constraints or roughness priors [40], [41].

The minimizer of (29) is the penalized weighted least squares (PLS) estimator

$$\underline{\hat{\theta}} = [\mathbf{F}_Y + \beta \mathbf{P}]^{-1} \mathbf{A}^T \mathbf{\Sigma}^{-1} \underline{Y}, \tag{30}$$

yielding the QPML estimator $\hat{t} = \underline{h}^T \underline{\hat{\theta}}$.

The estimator bias is

$$b_{\underline{\theta}} = \underline{h}^T \left[[\mathbf{F}_Y + \beta \mathbf{P}]^{-1} \mathbf{F}_Y - \mathbf{I} \right] \underline{\theta},$$

and its bias gradient is

$$\nabla b_{\underline{\theta}} = \left[\mathbf{F}_Y [\mathbf{F}_Y + \beta \mathbf{P}]^{-1} - \mathbf{I} \right] \underline{h} \tag{31}$$

$$= -\beta \mathbf{P} [\beta \mathbf{P} + \mathbf{F}_Y]^{-1} \underline{h}. \tag{32}$$

Finally, the variance of the QPML estimator $\hat{t}$ is

$$\sigma^2_{\underline{\theta}} = \frac{1}{\beta^2} \underline{h}^T \left[\mathbf{P} + \frac{1}{\beta} \mathbf{F}_Y \right]^{-1} \mathbf{F}_Y \left[\mathbf{P} + \frac{1}{\beta} \mathbf{F}_Y \right]^{-1} \underline{h}. \tag{33}$$

Consider the special case of estimation of a single component θ_k of $\underline{\theta}$ for which $\underline{h} = \underline{e}_k = [0, \ldots, 0, 1, 0, \ldots, 0]^T$. When the matrices $\mathbf{F}_Y$ and $\mathbf{P}$ commute, as occurs for example when $\mathbf{P} = \mathbf{I}$, the bias gradient (31) is seen to be equal to the difference between the mean response $[\beta \mathbf{P} + \mathbf{F}_Y]^{-1} \mathbf{F}_Y \underline{e}_k$ of the PLS estimator to a point source $\underline{\theta} = \underline{e}_k$, i.e., the point spread function of the estimator, and the ideal point response $\underline{e}_k$. Thus, under the commutative assumption the bias gradient norm can be viewed as a measure of the geometric resolution of $\hat{t}$ [16].

A.1 Non-Singular Fisher Matrix

Assume that $\mathbf{F}_Y$ is non-singular and compare (32) and (33) to the equations (13) and (12) for $\underline{d}_{min}$ and the bound $B(\underline{\theta}, \delta)$, respectively. Identifying $\nabla t_{\underline{\theta}} = \underline{h}$, $\lambda = 1/\beta$, it is clear that when $\mathbf{P}$ is chosen as $\mathbf{C}^{-1}$ the PLS estimator achieves the bound $B(\underline{\theta}, \delta)$ and has optimal bias gradient $\underline{d}_{min}$. Thus for linear functions $t_{\underline{\theta}}$ the uniform bound is achievable and the region above and including the fundamental delta-sigma tradeoff curve is an *achievable region.* Furthermore, since the bias gradient is a linear function, from relation (6) we have a very strong optimality property: the QPML estimator $\hat{t}$ is a minimum variance *biased* estimator in the sense that it is an estimator of minimum variance among estimators which satisfy the maximal bias constraint $\sup_{\underline{u} \in \mathcal{D}} |b_{\underline{u}} - b_{\underline{\theta}}| \leq \delta^2$, where $\delta^2 = g(1/\beta)$ and $\mathcal{C}$ is the ellipsoid defined above (3).

We used the Computational Recipe presented in Section II to trace out the delta-sigma tradeoff curve (uniform bound) parametrically as a function of $\lambda > 0$. Figure 2 shows the delta-sigma tradeoff curve for the case of pixel intensity estimation ($\underline{h} = \underline{e}_{67}$) and a well conditioned full rank discrete Gaussian system matrix. Specifically we generated a 128×128 matrix $\mathbf{A}$ with elements $a_{ij} = \frac{1}{\sqrt{2\pi}w} e^{\frac{-(i-j)^2}{2w^2}}$ and $w = 0.5$. The condition number

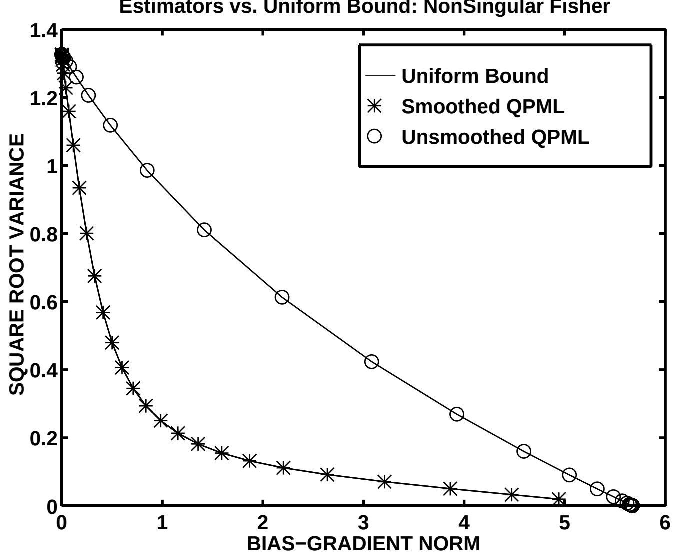

Fig. 2. Bias variance tradeoff study for pixel-intensity estimation and non-singular Fisher information. The smoothed PLS estimator (labeled smoothed QPML) exactly achieves the uniform bound.

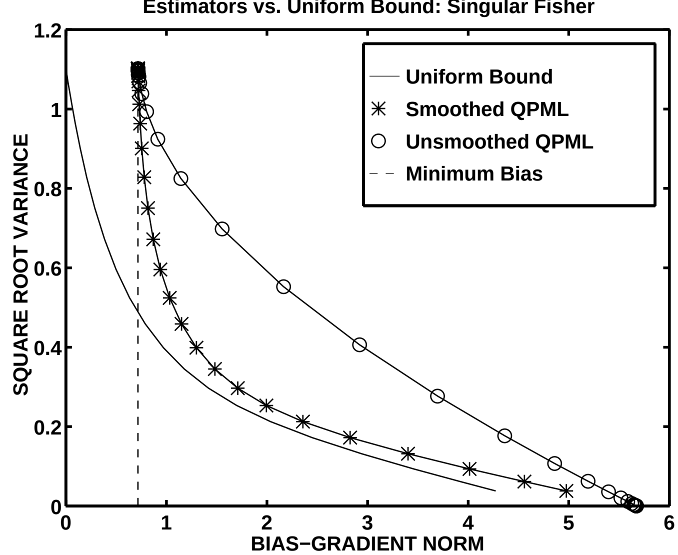

Fig. 3. Bias variance tradeoff study for pixel intensity estimation and singular Fisher information. Neither of the QPML estimators achieve the uniform bound.

of this $\mathbf{A}$ is 1.7. The matrix $\mathbf{C}$ in the norm $\|\nabla b_{\underline{\theta}}\|_C$ was selected as the inverse of the second order (Laplacian) differencing matrix

$$\mathbf{C}^{-1} = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & 0 & \\ & \ddots & \ddots & \ddots & \\ 0 & -1 & 2 & -1 & \\ & & -1 & 2 & \end{bmatrix}. \quad (34)$$

With this norm the restriction $\|\nabla b_{\underline{\theta}}\|_C \leq \delta$ corresponds to a constraint on maximal bias variation $\max_{\underline{\theta}\in\mathcal{C}} |\Delta b_{\underline{\theta}}|$ over a roughness constrained neighborhood $\mathcal{C}(\underline{\theta}, \mathbf{C})$ of $\underline{\theta}$ (recall relation (6)). Also plotted in Fig. 2 are the performance curves $(\|\nabla b_{\underline{\theta}}\|_2, \sigma_{\underline{\theta}})$ for two PLS pixel intensity estimators (30), one using the smoothing matrix $\mathbf{P} = \mathbf{C}^{-1}$, called the *smoothed QPML estimator*, and another using the diagonal "energy penalty" $\mathbf{P} = \mathbf{I}$, called the *unsmoothed QPML estimator*. These curves were traced out in the bias variance tradeoff plane by varying β in the parametric descriptions of estimator variance (33) and estimator bias gradient (32).

A.2 Singular Fisher Matrix

When $\mathbf{A}$ has rank less than n, $\mathbf{F}_Y$ is singular and unbiased estimators may not exist for all linear functions $t_{\underline{\theta}}$ of $\underline{\theta}$ [19], [42]. A lower bound on the norm of the bias gradient can derived (see Appendix C) using the relation (6) between the norm and the maximal bias variation over a region of parameter space. Since the uniform CR bound is finite and equal to the unbiased CR bound at $\delta = 0$, we cannot expect the delta-sigma tradeoff curve to be achievable for all δ as in the non-singular case.

To illustrate we repeat the study of Fig. 2 with a rank deficient Gaussian kernel matrix $\mathbf{A}_r$ obtained by decimating the rows of a full rank Gaussian kernel matrix $\mathbf{A}$ ($w = 2$) by a factor of 4. This yields the ill-posed problem of estimating a vector of 128 pixel intensities $\underline{\theta}$ based on only 32 observations $\underline{Y}$. We used the singular value decomposition of $\mathbf{A}$ to compute the delta-sigma tradeoff curve and the minimal bias gradient norm. The results for pixel intensity estimation ($t_{\underline{\theta}} = \underline{e}_{67}$) are plotted in Figure 3 along with the performance curves associated with the smoothed QPML ($\mathbf{P} = \mathbf{C}^{-1}$ of (34)) and unsmoothed QPML ($\mathbf{P} = \mathbf{I}$) estimators. Note that neither of the estimators achieve the uniform bound for any value of the parameter β. The minimal bias gradient norm is an asymptote on estimator performance which forces a sharp knee in the estimator performance curves. At points close to this knee maximal reduction in bias is only achieved at the price of significant increase in the variance.

For comparison, in Fig. 5 we plot the analogous curves for smoothed and unsmoothed QPML estimation of the contrast function defined as $t_{\underline{\theta}} = \underline{h}^T\underline{\theta}$, where the elements of $\underline{h}$ are plotted in Fig. 4. Observe that the smoothed QPML estimator of contrast comes much closer to the uniform bound than does the smoothed QPML estimator of pixel intensity shown in Fig. 3.

Under certain conditions the uniform CR bound is exactly achievable even for singular $\mathbf{F}_Y$, although generally not by a QPML estimator of the form (30) and generally not for all δ. Consider the estimator

$$\hat{\underline{\theta}} = \underline{h}^T \left[\mathbf{I} + \beta \mathbf{F}_Y^{+}\mathbf{P}\right]^{-1} \mathbf{F}_Y^{+}\mathbf{A}^T\mathbf{\Sigma}^{-1}\underline{Y}. \quad (35)$$

This estimator reduces to the previous estimator (30) for the case of non-singular $\mathbf{F}_Y$. The estimator bias gradient is

$$\begin{aligned} \nabla b_{\underline{\theta}} &= \left(\mathbf{F}_Y\mathbf{F}_Y^{+}\left[\mathbf{I} + \beta\mathbf{P}\mathbf{F}_Y^{+}\right]^{-1} - \mathbf{I}\right)\nabla t_{\underline{\theta}} \\ &= -\left[\frac{1}{\beta}\mathbf{P}^{-1} + \mathbf{F}_Y^{+}\right]^{-1}\mathbf{F}_Y^{+}\underline{h} \\ &\quad - \left[\mathbf{I} - \mathbf{F}_Y\mathbf{F}_Y^{+}\right]\left[\mathbf{I} + \beta\mathbf{P}\mathbf{F}_Y^{+}\right]^{-1}\underline{h}, \quad (36) \end{aligned}$$

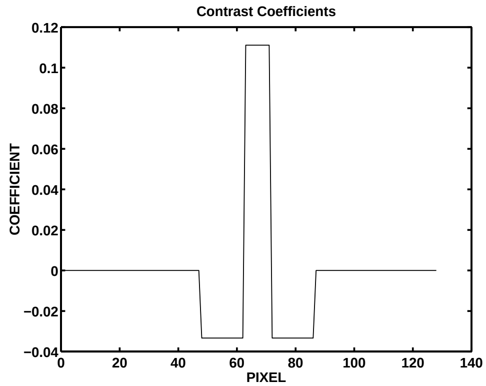

Fig. 4. Coefficients $\underline{h}$ of contrast function $t_{\underline{\theta}} = \underline{h}^T\underline{\theta}$ used for bias variance tradeoff study shown in Fig. 5.

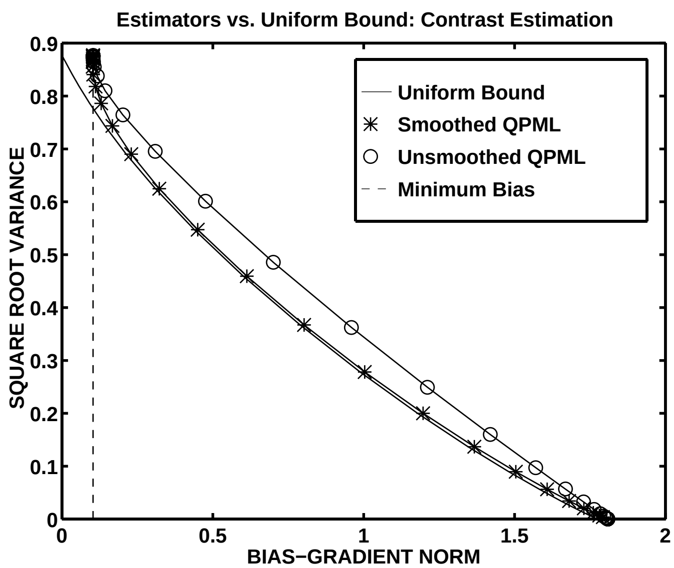

Fig. 5. Bias variance tradeoff study for estimation of contrast function (illustrated in Fig. 4) for singular Fisher information. The smoothed QPML estimator of contrast virtually achieves the uniform bound for bias-gradient norm δ greater than 0.2.

and the estimator variance is

$$\sigma^2_{\underline{\theta}} = \underline{h}^T\mathbf{P}^{-1}\left[\mathbf{P}^{-1}+\beta\mathbf{F}_Y^+\right]^{-1}\mathbf{F}_Y^+\left[\mathbf{P}^{-1}+\beta\mathbf{F}_Y^+\right]^{-1}\mathbf{P}^{-1}\underline{h}, \quad (37)$$

where in (37) we have used the property $\mathbf{F}_Y^+\mathbf{F}_Y\mathbf{F}_Y^+ = \mathbf{F}_Y^+$ [43]. Noting that here $\nabla t_{\underline{\theta}} = \underline{h}$, we conclude that the estimator variance is equal to the lower bound expression $B(\underline{\theta},\delta)$ given in (9) when $\mathbf{P} = \mathbf{C}^{-1}$ and $\beta = 1/\lambda$. Furthermore, under these conditions the bias gradient (36) differs from the optimal bias gradient $\underline{d}^T_{min}$, given in (10), only by the presence of the second additive term on the right hand side of (36). Thus the estimator (35) with $\mathbf{P} = \mathbf{C}^{-1}$ is an optimal biased estimator when this second additive term is equal to zero.

We summarize these results in a theorem which applies to both singular and non-singular $\mathbf{F}_Y$.

Theorem 2: Let $\mathbf{B} = \mathbf{P}^{\frac{1}{2}}\mathbf{F}_Y^+\mathbf{P}^{\frac{1}{2}}$ where $\mathbf{F}_Y$ is the possibly singular Fisher information matrix. If

1. $\delta^2 < \underline{h}^T\mathbf{P}^{\frac{1}{2}}\mathcal{P}_B\mathbf{P}^{\frac{1}{2}}\underline{h}$, and
2. the vector $\underline{h}$ lies in the nullspace of $[\mathbf{I}-\mathbf{F}_Y\mathbf{F}_Y^+][\mathbf{I}+\beta\mathbf{P}\mathbf{F}_Y^+]^{-1}$,

then the estimator $\hat{t}$ given by (35) achieves the fundamental delta-sigma tradeoff in the sense of having minimum variance over all estimators of $t_{\underline{\theta}} = \underline{h}^T\underline{\theta}$ satisfying $\|\nabla b_{\underline{\theta}}\|^2_{P^{-1}} \leq \delta^2 = g(1/\beta)$ where $g(\bullet)$ is the function given in (11).

Recognizing the matrix $\mathbf{I}-\mathbf{F}_Y^+\mathbf{F}_Y = \mathbf{I}-\mathbf{F}_Y\mathbf{F}_Y^+$ as the operator which projects onto the null space of $\mathbf{F}_Y$, an equivalent condition to (2) is that $[\mathbf{I}+\beta\mathbf{P}\mathbf{F}_Y^+]^{-1}\underline{h}$ lie in the range space of $\mathbf{F}_Y$. For the special case of $\beta = 0$, condition (2) of Theorem 2 reduces to the well known necessary condition for achievability of the unbiased CR bound: the vector $\underline{h}$ must lie in the range space of the Fisher information $\mathbf{F}_Y$. In order for the uniform CR bound $B(\underline{\theta},\delta)$ to be achievable for all values of δ, condition (2) must hold for all $\beta > 0$. This is a much stronger condition except when the eigenspace of $[\mathbf{I}+\beta\mathbf{P}\mathbf{F}_Y^+]^{-1}$ is independent of β, which occurs for example when $\mathbf{P} = \mathbf{I}$. This suggests that when estimation of any fixed $t_{\underline{\theta}}$ is of interest and the Fisher information is singular, the uniform bound will rarely be achievable everywhere in the delta-sigma plane.

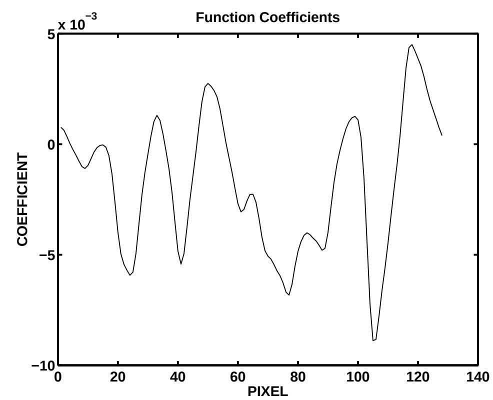

Fig. 6. Coefficients $\underline{h}$ of the linear compound $t_{\underline{\theta}} = \underline{h}^T\underline{\theta}$ satisfying condition 2 of Theorem 2 and used for computing the curves in Fig. 7.

To illustrate Theorem 2 we selected a small value of β and found a vector $\underline{h}$ lying in the nullspace of the matrix $[\mathbf{I}-\mathbf{F}_Y^+\mathbf{F}_Y][\mathbf{I}+\beta\mathbf{P}\mathbf{F}_Y^+]^{-1}$ via singular value decomposition. This vector is shown in Fig. 6. In view of Theorem 2 we know that the estimator (35) of $\underline{h}^T\underline{\theta}$ should achieve the uniform bound for the chosen value of β. In Fig. 7 we plot the uniform bound for estimators of $\underline{h}^T\underline{\theta}$ and the performance curve of two estimators of the form (35), one smoothed ($\mathbf{P} = \mathbf{C}^{-1}$) and one unsmoothed ($\mathbf{P} = \mathbf{I}$). Observe that the smoothed estimator essentially achieves the uniform bound for $\delta < 0.2$.

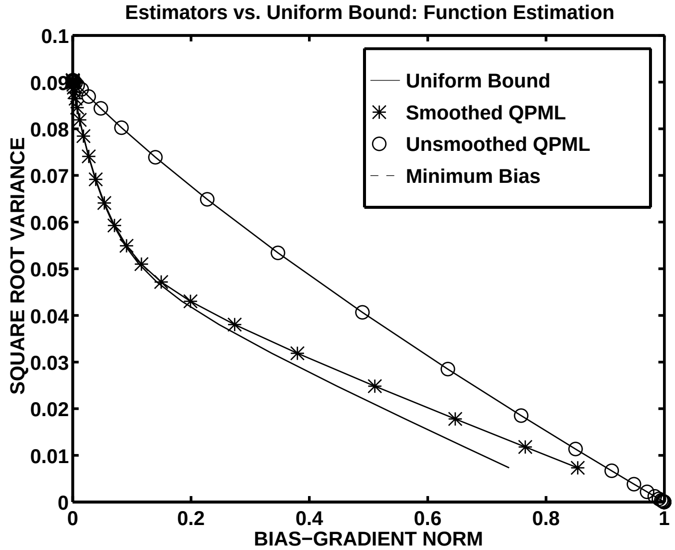

Fig. 7. Bias variance tradeoff study for estimation of nullspace compound $\underline{h}^T\underline{\theta}$ for singular Fisher information ($\underline{h}$ is illustrated in Fig. 6). The bound is exactly achieved by the smoothed QPML estimator at the point $\delta = 5.4 \times 10^{-6}$.

B. Poisson Model

In some applications the observations $\underline{Y}$ are given by the linear model (27) but with non-Gaussian additive noise. Here we consider the case of Poisson noise which arises in emission computed tomography and other quantum limited inverse problems [44]. The observation $\underline{Y} = [Y_1, \ldots, Y_m]^T$ is a vector of integers or counts with a vector of means $\underline{\mu} = [\mu_1, \ldots, \mu_m]^T$. This vector of counts obeys independent Poisson statistics with log-likelihood

$$\ln f_{\underline{Y}}(\underline{y};\underline{\theta}) \quad = \quad \sum_{j=1}^{m} (y_j \ln(\mu_j(\underline{\theta})) - \mu_j(\underline{\theta})) + c. \qquad (38)$$

In (38) c is a constant independent of the unknown source $\underline{\theta}$ and the mean number of counts is assumed to obey the linear model

$$\underline{\mu}(\underline{\theta}) = \mathbf{A}\,\underline{\theta} + \underline{r}. \qquad (39)$$

In emission computed tomography $\underline{\mu}$ is a vector of mean object projections measured over m detectors, $\mathbf{A}$ is a $m \times n$ system matrix that depends on the tomographic geometry, $\underline{\theta}$ is an unknown image intensity vector, and $\underline{r}$ is a $m \times 1$ vector representing background noise due to randoms and scattered photons.

The Fisher information has the form [45]

$$\mathbf{F}_Y(\underline{\theta}) \quad = \quad \sum_{j=1}^{m} \frac{1}{\mu_j(\underline{\theta})} \mathbf{A}_{j*}\mathbf{A}_{j*}^T, \qquad (40)$$

where $\mathbf{A}_{j*}^T$ is the j-th row of $\mathbf{A}$.

To investigate the achievability of the region above the delta-sigma tradeoff curve, and to illustrate the empirical computation of bias gradient, we consider again the QPML strategy. The QPML estimator studied is $\hat{t} = t_{\underline{\hat{\theta}}}$ where $\underline{\hat{\theta}}$ is the vector $\underline{\theta}$ which maximizes the penalized likelihood function

$$J(\underline{\theta}) = \ln f_{\underline{Y}}(y;\underline{\theta}) - \frac{\beta}{2}\underline{\theta}^T\mathbf{P}\underline{\theta}, \qquad (41)$$

where $\mathbf{P}$ is a nonnegative definite matrix. In the simulations below we used $\mathbf{P} = \mathbf{C} = \mathbf{I}$.

Exact analytic expressions for the variance, bias, and bias gradient of the QPML estimator are intractable. However, it will be instructive to consider asymptotic approximations to these quantities. In Appendix D expressions for asymptotic bias, bias gradient and variance are derived under the assumption that the difference between the projection $\mathbf{A}E_{\underline{\theta}}[\underline{\hat{\theta}}]$ of the mean QPML image and the projection $\mathbf{A}\underline{\hat{\theta}}$ of the true image is small – frequently a very good approximation in image restoration and tomography. Specializing the results (57)-(59) in Appendix D to the case of linear functions $t_{\underline{\theta}} = \underline{h}^T\underline{\theta}$, we obtain the following expressions for the asymptotic variance of $\hat{t}$

$$\sigma_{\underline{\theta}}^2 = \underline{h}^T \left[\mathbf{F}_Y(\underline{\theta}) + \beta\mathbf{P}\right]^{-1} \mathbf{F}_Y(\underline{\theta}) \left[\mathbf{F}_Y(\underline{\theta}) + \beta\mathbf{P}\right]^{-1} \underline{h}, \quad (42)$$

and the asymptotic bias gradient

$$\nabla b_{\underline{\theta}} \quad = \quad -\mathbf{P}\left[\mathbf{P} + \frac{1}{\beta}\mathbf{F}_Y(\underline{\theta})\right]^{-1} \underline{h} + O\left(\frac{1}{\beta}\right). \quad (43)$$

where $O(1/\beta)$ is a remainder term of order $1/\beta$.

When we identify $\mathbf{P} = \mathbf{C}^{-1}$ and $\beta = 1/\lambda$ we see that the estimator variance is identical to the optimal variance (12), and that for linear $t_{\underline{\theta}}$ the bias gradient is identical to the optimal bias gradient (13) to order $O(1/\beta)$. Therefore, assuming the bias gradient and variance approximations (43) and (42) are accurate, for linear $t_{\underline{\theta}}$ we can expect that the fundamental delta-sigma tradeoff curve will be approximately achieved by the QPML estimator for large values of the regularization parameter β if $\mathbf{P} = \mathbf{C}^{-1}$.

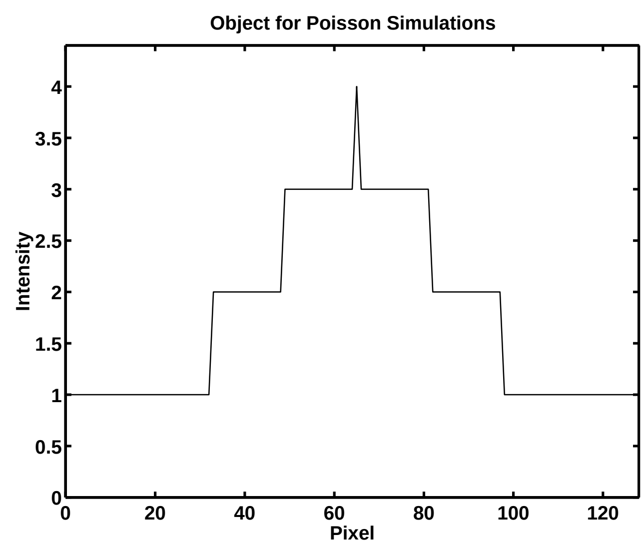

Fig. 8. Emission source $\underline{\theta}$ used for Poisson simulations. The spike in the center was the pixel of interest.

To examine the performance of the methods for estimating bias gradient norm described in Section III, and to verify the asymptotic bias and variance performance predictions, we generated simulated Poisson measurements with means given by (39). In these simulations $\mathbf{A}$ was a 128×128 tri-diagonal blurring matrix with kernel $(0.23, 0.54, 0.23)$, for which the condition number is 12.5. The source intensity $\underline{\theta}$ is shown in Fig. 8. The function of interest was chosen as $t_{\underline{\theta}} = \theta_{65}$, the intensity of pixel 65 in Fig. 8. We generated $L = 1000$ realizations of the measurements each having a mean total of $\sum_{j=1}^{m} \mu_j(\underline{\theta}) = 2100$ counts, including a 5% background representing random coincidences [20].

We computed three types of estimates of $\underline{\theta}$: the quadratically penalized maximum likelihood estimator using the "energy penalty" ($\mathbf{P} = \mathbf{I}$), a truncated SVD estimator, and a "deconvolve/shrink" estimator. We maximized the non-quadratic penalized likelihood objective using the PML-SAGE algorithm, a variant of the iterative space alternating generalized expectation-maximization (SAGE) algorithm of [20] adapted for penalized maximum likelihood image reconstruction [46]. We initialized PML-SAGE with an unweighted penalized least-squares estimate: $(\mathbf{A}^T\mathbf{A} + \beta^\star\mathbf{I})^{-1}\mathbf{A}'(Y - \underline{r})$, which is linear so can be computed noniteratively. Here $\beta^\star = \beta \sum_j \mathbf{A}_{jk} / (\sum_j \mathbf{A}_{jk}/Y_j)$ for $k = 65$ (cf [47], [48]). By so initializing, only 30 iterations were needed to ensure convergence to a precision well below the estimate standard deviation. For the truncated SVD estimator, we computed the singular value decomposition (SVD) of $\mathbf{A}$, and computed approximate pseudo-inverses of $\mathbf{A}$ by excluding the 10 smallest eigenvalues. The form of the "deconvolve/shrink" estimator is:

$$\hat{\underline{\theta}}(Y) = \beta(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T(Y - \underline{r}),$$

where β ranges from 0 to 1.

We applied each estimator to the $L = 1000$ measurement realizations and computed the standard sample variance $\hat{\sigma}^2 = \frac{1}{L-1}\sum_{i=1}^{L}(\hat{t}(Y_i) - \bar{\hat{t}})^2$ where $\bar{\hat{t}} = \frac{1}{L}\sum_{i=1}^{L}\hat{t}(Y_i)$ is the estimator sample mean. We estimated the estimator bias gradient length (BGL) (the norm $\|\cdot\|_C$ with $\mathbf{C} = \mathbf{I}$) via the methods described in Section III. We traced out the estimator performance curves in the delta-sigma plane by varying the regularization parameter β.

Figure 9 illustrates the benefits of using the bootstrap estimate of BGL as compared with the ordinary method-of-moments BGL estimator for the identity penalized likelihood estimator. Included are standard error bars (twice the length gives 95% confidence intervals) for bias (horizontal lines) and variance (vertical lines smaller than plotting symbol) of the bootstrap BGL estimator for $L = 500$ and $L = 1000$ realizations. The BGL error bars were computed under a large L Gaussian approximation to the bias gradient estimates and a square root transformation. In general as the smoothing parameter β is decreased QPML estimator bias decreases while QPML estimator variance increases. This increase in variance produces an increasingly large positive bias in the ordinary BGL estimator causing the curve to abruptly diverge to the right. However, the bias of the bootstrap BGL estimator remains small as β decreases so that it extends the range of reliable estimation of the ordinary BGL estimator.

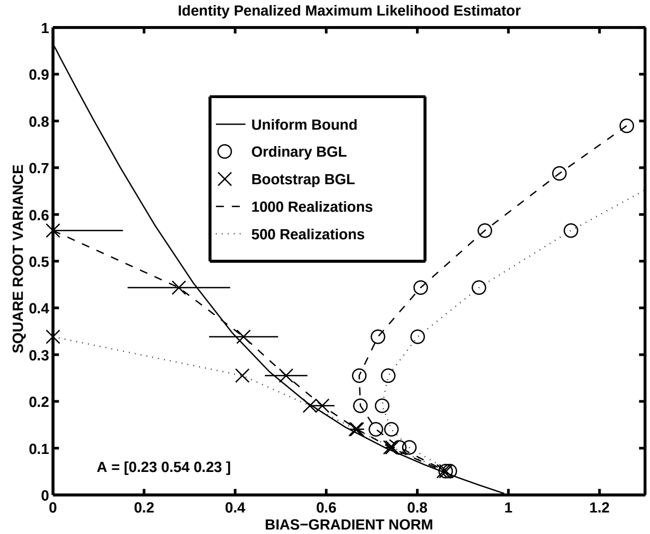

Fig. 9. Performance of penalized likelihood estimator compared to uniform CR bound. Bias gradient length (BGL) estimates were computed using both the standard method-of-moments estimate and the bootstrap estimate described in Section III. Data points to left fall below bound due to insufficient number of realizations for reliable BGL estimation.

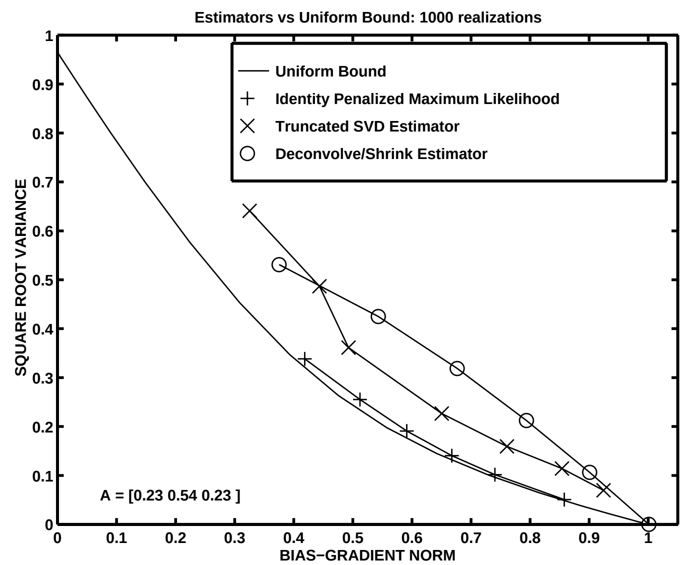

Fig. 10. Performance of three different estimators compared to uniform CR bound.

In Figure 10 we compare the three different estimators to the uniform CR bound. As predicted by the asymptotic analysis the uniform bound is virtually achieved by the identity penalized likelihood estimator in the high bias and low variance region (large β). The identity penalized maximum likelihood estimator visibly outperforms the other two estimators. Unfortunately, for fixed $L = 1000$ as the estimator performance curves approach the left side of the delta-sigma plane, the bootstrap BGL estimates become increasingly variable (recall error bars in Figure 9), so an increasingly large number of realizations is required to make reliable comparisons between the estimator performance and the bound. On the other hand ECT images corresponding to such highly variable estimates of $\underline{\theta}$ are unlikely to be of much practical interest.

V. Conclusions

We have presented a method for specifying a lower bound in the delta-sigma plane defined as the set of pairs $(\delta_{\underline{\theta}}, \sigma_{\underline{\theta}})$ where $\delta_{\underline{\theta}}$ is the estimator bias gradient norm and $\sigma^2_{\underline{\theta}}$ is the estimator variance. For two inverse problems, one linear and one non-linear, we have established that the bound is achievable under certain circumstances.

There remain several open problems. In ill-posed problems the Fisher matrix is singular and an eigendecomposition appears to be required to compute the bound. For small ill-posed problems this is not a major impediment. However, for large problems with many parameters, which includes many image reconstruction and image restoration problems, the eigendecomposition is not practical and faster numerical methods are needed. Another problem is that the variance of the bootstrap estimator for bias gradient norm increases rapidly with the number of unknown parameters. Since the bootstrap estimator is not guaranteed to be non-negative this high variance can make the estimator useless for estimating small valued bias gradient norms. In such cases, asymptotic bias and variance formulas may be useful and can be derived along similar lines as described in Appendix D. Finally, we established a general relation between bias gradient norm and maximal bias variation. Although for general estimation problems the interpretation of the bias gradient norm may be difficult, for the two applications considered in this paper, the bias gradient norm was interpreted as a measure of spatial resolution of the estimator.

Acknowledgement

The authors gratefully acknowledge the anonymous reviewers for their comments and suggestions, and W.L. Rogers and Neal Clinthorne of the Division of Nuclear Medicine, University of Michigan Hospital, for many helpful discussions of this work.

Appendix A: Proof of Theorem 1

Proof of Theorem 1: For a fixed $\delta > 0$ we perform constrained minimization of the biased form of the CR bound (7) over the feasible set $\nabla b_{\underline{\theta}} : \|\nabla b_{\underline{\theta}}\|_C \leq \delta$ of bias gradient vectors

$$\begin{array}{rcl} \mathrm{var}_{\underline{\theta}}(\hat{t}) & \geq & [\nabla t_{\underline{\theta}} + \nabla b_{\underline{\theta}}]^T \mathbf{F}_Y^+ [\nabla t_{\underline{\theta}} + \nabla b_{\underline{\theta}}] \\ & \geq & \min\limits_{\underline{d}:\|\underline{d}\|_C \leq \delta} Q(\underline{d}), \end{array}$$

where

$$Q(\underline{d}) = [\nabla t_{\underline{\theta}} + \underline{d}]^T \mathbf{F}_Y^+ [\nabla t_{\underline{\theta}} + \underline{d}],$$

and $\underline{d}$ is a vector in $\mathbb{R}^n$. Defining $\underline{\tilde{d}} = \mathbf{C}^{\frac{1}{2}} \underline{d}$, where $\mathbf{C}^{\frac{1}{2}}$ is a square root factor of $\mathbf{C}$, the minimization of $Q(d)$ is equivalent to

$$\min_{\underline{\tilde{d}}:\|\underline{\tilde{d}}\|_2 \leq \delta} \left[\mathbf{C}^{\frac{1}{2}} \nabla t_{\underline{\theta}} + \underline{\tilde{d}}]^T \mathbf{B} [\mathbf{C}^{\frac{1}{2}} \nabla t_{\underline{\theta}} + \underline{\tilde{d}}\right], \tag{44}$$

where $\mathbf{B} = \mathbf{C}^{-\frac{1}{2}} \mathbf{F}_Y^+ \mathbf{C}^{-\frac{1}{2}}$.

First we consider the case where the unconstrained minimum $Q(\underline{d}) = 0$ occurs in the interior of the constraint set $\|\underline{d}\|_C \leq \delta$. From (44) it is clear that $Q(\underline{d})$ can be zero if and only if $\mathbf{C}^{\frac{1}{2}} \nabla t_{\underline{\theta}} + \underline{\tilde{d}}$ lies in the null space of $\mathbf{B}$. Such a solution $\underline{\tilde{d}}^o$ must have the form

$$\underline{\tilde{d}}^o = -\mathcal{P}_B \mathbf{C}^{\frac{1}{2}} \nabla t_{\underline{\theta}} + \underline{\phi}$$

where $\underline{\phi}$ is an arbitrary vector in the null space of $\mathbf{B}$. But for $\underline{\tilde{d}}^o$ to be a feasible solution it must satisfy $\|\underline{\tilde{d}}^o\|_2 \leq \delta$ so that, by orthogonality of $\mathcal{P}_B \mathbf{C}^{\frac{1}{2}} \nabla t_{\underline{\theta}}$ and $\underline{\phi}$

$$\delta^2 \geq \|\underline{\tilde{d}}^o\|_2^2 = \|\mathcal{P}_B \mathbf{C}^{\frac{1}{2}} \nabla t_{\underline{\theta}}\|_2^2 + \|\underline{\phi}\|_2^2 \geq \|\mathcal{P}_B \mathbf{C}^{\frac{1}{2}} \nabla t_{\underline{\theta}}\|_2^2.$$

We conclude that $\min_{\underline{d}:\|\underline{d}\|_C \leq \delta} Q(\underline{d}) = 0$ iff $\|\mathcal{P}_B \mathbf{C}^{\frac{1}{2}} \nabla t_{\underline{\theta}}\|_2^2 = \nabla^T t_{\underline{\theta}} \mathbf{C}^{\frac{1}{2}} \mathcal{P}_B \mathbf{C}^{\frac{1}{2}} \nabla t_{\underline{\theta}} \leq \delta^2$.

If $\nabla^T t_{\underline{\theta}} \mathbf{C}^{\frac{1}{2}} \mathcal{P}_B \mathbf{C}^{\frac{1}{2}} \nabla t_{\underline{\theta}} = 0$ then we have nothing left to prove. Otherwise, assume δ lies in the range: $0 \leq \delta^2 < \nabla^T t_{\underline{\theta}} \mathbf{C}^{\frac{1}{2}} \mathcal{P}_B \mathbf{C}^{\frac{1}{2}} \nabla t_{\underline{\theta}}$. In this case the minimizing $\underline{\tilde{d}}$ lies on the boundary and satisfies the equality constraint $\|\underline{\tilde{d}}\|_2 = \delta$. We thus need solve the unconstrained minimization of the Lagrangian:

$$\min_{\underline{\tilde{d}}} \left[\left[\mathbf{C}^{\frac{1}{2}} \nabla t_{\underline{\theta}} + \underline{\tilde{d}}]^T \mathbf{B} [\mathbf{C}^{\frac{1}{2}} \nabla t_{\underline{\theta}} + \underline{\tilde{d}}\right] + \lambda(\underline{\tilde{d}}^T \underline{\tilde{d}} - \delta^2)\right], \tag{45}$$

where we have introduced the undetermined multipler $\lambda \geq 0$. Assuming for the moment that λ is strictly positive, the matrix $\lambda \mathbf{I} + \mathbf{B}$ is positive definite and the completion of the square in the Lagrangian in (45) gives

$$\begin{array}{l} \left[\underline{\tilde{d}} + (\lambda \mathbf{I} + \mathbf{B})^{-1} \mathbf{B} \mathbf{C}^{\frac{1}{2}} \nabla t_{\underline{\theta}}\right]^T (\lambda \mathbf{I} + \mathbf{B}) \left[\underline{\tilde{d}} + (\lambda \mathbf{I} + \mathbf{B})^{-1} \mathbf{B} \mathbf{C}^{\frac{1}{2}} \nabla t_{\underline{\theta}}\right] \\ \qquad + \nabla^T t_{\underline{\theta}} \mathbf{F}_Y^+ \nabla t_{\underline{\theta}} - \nabla^T t_{\underline{\theta}} \left[\mathbf{C}^{\frac{1}{2}} \mathbf{B} (\lambda \mathbf{I} + \mathbf{B})^{-1} \mathbf{B} \mathbf{C}^{\frac{1}{2}}\right] \nabla t_{\underline{\theta}} - \lambda \delta^2. \end{array}$$

It follows immediately from (46) that

$$\underline{\tilde{d}} = \underline{\tilde{d}}_{min} = -(\lambda \mathbf{I} + \mathbf{B})^{-1} \mathbf{B} \mathbf{C}^{\frac{1}{2}} \nabla t_{\underline{\theta}}$$

achieves the minimum. Noting that $\underline{d}_{min} = \mathbf{C}^{-\frac{1}{2}} \underline{\tilde{d}}_{min}$, expressing $\mathbf{B}$ in terms of $\mathbf{C}$ and $\mathbf{F}_Y$, and performing simple matrix algebra, we obtain (10). Substituting the expression for $\underline{\tilde{d}}_{min}$ into (44):

$$\begin{array}{rl} & \min\limits_{\underline{d}:\|\underline{d}\|_C \leq \delta} Q(\underline{d}) \\ = & [\nabla t_{\underline{\theta}} + \underline{d}_{min}]^T \mathbf{F}_Y^+ [\nabla t_{\underline{\theta}} + \underline{d}_{min}] \\ = & \nabla^T t_{\underline{\theta}} \mathbf{C}^{\frac{1}{2}} \left[\mathbf{I} - [\lambda \mathbf{I} + \mathbf{B}]^{-1} \mathbf{B}\right]^T \mathbf{B} \left[\mathbf{I} - [\lambda \mathbf{I} + \mathbf{B}]^{-1} \mathbf{B}\right] \mathbf{C}^{\frac{1}{2}} \nabla t_{\underline{\theta}} \\ = & \lambda^2 \nabla^T t_{\underline{\theta}} \mathbf{C}^{\frac{1}{2}} [\lambda \mathbf{I} + \mathbf{B}]^{-1} \mathbf{B} [\lambda \mathbf{I} + \mathbf{B}]^{-1} \mathbf{C}^{\frac{1}{2}} \nabla t_{\underline{\theta}} \end{array} \tag{46}$$

which, after simple matrix manipulations, gives the expression (9).

The Lagrange multiplier λ is determined by the equality constraint

$$\delta^2 = \underline{\tilde{d}}_{min}^T \underline{\tilde{d}}_{min} = \nabla^T t_{\underline{\theta}} \left[\mathbf{C}^{\frac{1}{2}} \mathbf{B}[\lambda \mathbf{I} + \mathbf{B}]^{-2} \mathbf{B} \mathbf{C}^{\frac{1}{2}} \right] \nabla t_{\underline{\theta}} = g(\lambda). \quad (47)$$

Substitution of $\mathbf{B}$ with $\mathbf{C}^{-\frac{1}{2}} \mathbf{F}_Y^+ \mathbf{C}^{-\frac{1}{2}}$ yields the expression (11) after simple matrix algebra.

Let the non-negative definite symmetric matrix $\mathbf{B}$ have eigendecomposition $\mathbf{B} = \sum_{i=1}^r \beta_i \underline{\xi}_i \underline{\xi}_i^T$ where β_i are positive eigenvalues, $\underline{\xi}_i$ are eigenvectors, and $r > 0$ is the rank of $\mathbf{B}$. With these definitions the function $g(\lambda)$ (47) has the equivalent form

$$g(\lambda) = \sum_{i=1}^r \frac{\beta_i^2}{(\lambda + \beta_i)} |\nabla^T t_{\underline{\theta}} \mathbf{C}^{\frac{1}{2}} \underline{\xi}_i|^2. \quad (48)$$

Since by assumption $\nabla^T t_{\underline{\theta}} \mathbf{C}^{\frac{1}{2}} \mathcal{P}_B \mathbf{C}^{\frac{1}{2}} \nabla t_{\underline{\theta}} > 0$, $\mathbf{C}^{\frac{1}{2}} \nabla t_{\underline{\theta}}$ does not lie in the nullspace of $\mathbf{B}$ and thus $|\nabla^T t_{\underline{\theta}} \mathbf{C}^{\frac{1}{2}} \underline{\xi}_i|^2 > 0$ for at least one i, $i = 1, \ldots, r$. Therefore, from (48), it is obvious that the function $g(\lambda)$ is continuous monotone decreasing over $\lambda \geq 0$ with $\lim_{\lambda \to \infty} g(\lambda) = 0$ and $\lim_{\lambda \to 0} g(\lambda) = \nabla t_{\underline{\theta}} \mathbf{C}^{\frac{1}{2}} \mathcal{P}_B \mathbf{C}^{\frac{1}{2}} \nabla t_{\underline{\theta}}$. Hence there exists a unique strictly positive λ such that $g(\lambda) = \delta^2$ for any value $\delta^2 \in [0, \nabla t_{\underline{\theta}} \mathbf{C}^{\frac{1}{2}} \mathcal{P}_B \mathbf{C}^{\frac{1}{2}} \nabla t_{\underline{\theta}})$. □

Appendix B: Bootstrap Derivation

We start with the following simple estimator $\widehat{\nabla b_{\underline{\theta}}} = \widehat{\nabla b_{\underline{\theta}}}(Y_1, \ldots, Y_L)$ of the bias gradient $\nabla b_{\underline{\theta}}$

$$\widehat{\nabla b_{\underline{\theta}}} = \frac{1}{L} \sum_{i=1}^L \underline{z}(Y_i),$$

where $\underline{z}$ is the column vector

$$\underline{z}(Y_i) = \hat{t}(Y_i) \nabla_{\underline{\theta}}^T \ln f_Y(Y_i; \underline{\theta}) - \nabla t_{\underline{\theta}}.$$

The (biased) sample magnitude estimator of the norm squared $\delta^2 = \|\nabla b_{\underline{\theta}}\|_C^2$ is

$$\hat{\delta}^2(Y_1, \ldots, Y_L) = \|\widehat{\nabla b_{\underline{\theta}}}\|_C^2 = \frac{1}{L^2} \|\sum_{i=1}^L \underline{z}(Y_i)\|_C^2.$$

Now given the random sample $Y_1 = y_1, \ldots, Y_L = y_L$, the resampled estimate $\hat{\delta}_*^2 = \hat{\delta}_*^2(Y_1^*, \ldots, Y_L^*)$ is [37]

$$\begin{aligned} \hat{\delta}_*^2 &= \frac{1}{L^2} \|\sum_{i=1}^L c_i\, \underline{z}(y_i)\|_C^2 \\ &= \frac{1}{L^2} \sum_{i=1}^L \sum_{j=1}^L c_i c_j < \underline{z}(y_i), \underline{z}(y_j) >_C, \quad (49) \end{aligned}$$

where $< \underline{u}, \underline{v} >_C = \underline{u}^T \mathbf{C} \underline{v}$ is defined as the (weighted) inner product of column vectors $\underline{u}$ and $\underline{v}$. Define $\underline{c} = [c_1, \ldots, c_L]^T$ and let $\mathbf{H} = [[< \underline{z}(Y_i), \underline{z}(Y_j) >_C]]$ denote a $L \times L$ matrix of inner products. Then the resampled estimate (49) has the equivalent form

$$\begin{aligned} \hat{\delta}_*^2 &= \frac{1}{L^2} \underline{c}^T \mathbf{H} \underline{c} \quad (50) \\ &= \frac{1}{L^2} \text{trace}\{\mathbf{H}\, \underline{c}\,\underline{c}^T\}. \quad (51) \end{aligned}$$

The resampling outcomes $c_1, \ldots, c_L$ are multinomial distributed with equal cell probabilities $p_1 = \ldots = p_L = 1/L$ and $\sum_{i=1}^L c_i = L$. Averaging (51) over $\underline{c}$ we obtain the bootstrap estimate of the mean

$$\begin{aligned} E_*[\hat{\delta}_*^2] &= \sum_* \hat{\delta}_*^2 \binom{L}{c_1 \ldots c_L} L^{-L} \quad (52) \\ &= \frac{1}{L^2} \text{trace}\left\{\mathbf{H}\, E_*[\underline{c}\,\underline{c}^T]\right\}. \end{aligned}$$

From the mean and covariance of the multinomial distribution [49, Sec. 3.2]:

$$E_*[\underline{c}\,\underline{c}^T] = \mathbf{I} + \tfrac{L-1}{L} \underline{1}\,\underline{1}^T$$

where $\underline{1}$ is a vector of ones. Applying this identity to (52) and noting that trace$\{\mathbf{H}\} = \sum_{i=1}^L \|\underline{z}(Y_i)\|_C^2$ and $\underline{1}^T \mathbf{H} \underline{1} = \|\sum_{i=1}^L \underline{z}(Y_i)\|_C^2$, we obtain

$$\begin{aligned} E_*[\hat{\delta}_*^2] &= \frac{1}{L^2} \text{ trace}\left\{\mathbf{H}\, E_*[\underline{c}\,\underline{c}^T]\right\} \\ &= \frac{1}{L^2} \left[\text{trace}\{\mathbf{H}\} + \frac{L-1}{L} \underline{1}^T \mathbf{H} \underline{1} \right] \\ &= \frac{1}{L^2} \left[\sum_{i=1}^L \|\underline{z}(Y_i)\|_C^2 - L\|\overline{\underline{z}}\|_C^2 + L^2 \|\overline{\underline{z}}\|_C^2 \right] \\ &= \frac{1}{L^2} \sum_{i=1}^L \|\underline{z}(Y_i) - \overline{\underline{z}}\|_C^2 + \hat{\delta}^2, \end{aligned}$$

where $\overline{\underline{z}} = \overline{\nabla b_{\underline{\theta}}} = \frac{1}{L} \sum_{i=1}^L \underline{z}(Y_i)$ is the sample mean and we have identified $\hat{\delta}^2 = \|\overline{\underline{z}}\|_C^2$. Plugging this last expression into (25) we obtain:

$$\begin{aligned} \hat{\delta}_c^2 &= 2\hat{\delta}^2 - E_*[\hat{\delta}_*^2] \\ &= \hat{\delta}^2 - \frac{1}{L^2} \sum_{i=1}^L \|\underline{z}(Y_i) - \overline{\underline{z}}\|_C^2, \end{aligned}$$

which is identical to the expression (26).

Appendix C: Lower Bound on Bias Gradient

Here we derive a simple lower bound on the maximal bias variation over the region $\mathcal{C} = \{\underline{u} : (\underline{u} - \underline{\theta})^T \mathbf{C}^{-1} (\underline{u} - \underline{\theta}) \leq 1\}$ under the assumptions: (i) $\mathbf{F}_Y(\underline{u})$ is constant over $\underline{u} \in \mathcal{C}$; and (ii) the functional $t_{\underline{u}}$ to be estimated is linear over $\underline{u} \in \mathcal{C}$. Define: 1) $\mathcal{P}_F = \mathbf{F}_Y \mathbf{F}_Y^+ = \mathbf{F}_Y^+ \mathbf{F}_Y$ as the symmetric matrix which orthogonally projects onto the range space of $\mathbf{F}_Y$, 2) $\mathcal{N}_F = \{\underline{u} : \mathcal{P}_F \underline{u} = \underline{0}\}$ as the nullspace of $\mathbf{F}_Y$, 3) $m_{\underline{u}} = E_{\underline{u}}[\hat{t}]$. Under assumption (i) the parameter $\underline{u}$ is not identifiable for $\underline{u} \in \mathcal{N}_F \cap \mathcal{C}$ and it follows that $m_{\underline{u}} - m_{\underline{\theta}} = 0$. Therefore, we obtain the lower bound

$$\begin{aligned} \max_{\underline{u}\in\mathcal{C}} |b_{\underline{u}} - b_{\underline{\theta}}|^2 &= \max_{\underline{u}\in\mathcal{C}} |m_{\underline{u}} - t_{\underline{\theta}} - m_{\underline{\theta}} + t_{\underline{\theta}}|^2 \\ &\geq \max_{\underline{u}\in\mathcal{C}\cap\mathcal{N}_F} |m_{\underline{u}} - m_{\underline{\theta}} + t_{\underline{\theta}} - t_{\underline{u}}|^2 \\ &= \max_{\underline{u}\in\mathcal{C}\cap\mathcal{N}_F} |t_{\underline{u}} - t_{\underline{\theta}}|^2 \\ &= \max_{\underline{u}\in\mathcal{C}\cap\mathcal{N}_F} |\underline{h}^T \Delta\underline{u}|^2 \\ &= \max_{\underline{u}\in\mathcal{C}} |\underline{h}^T [\mathbf{I} - \mathcal{P}_F]\Delta\underline{u}|^2, \quad (53) \end{aligned}$$

where $\Delta\underline{u} = \underline{u} - \underline{\theta}$. Now, using an extremal property of the Rayleigh quotient, the right hand side of (53) is

$$\begin{aligned} &\max_{\underline{u}\in\mathcal{C}} |\underline{h}^T [\mathbf{I} - \mathcal{P}_F]\Delta\underline{u}|^2 \\ &= \max_{\Delta\underline{u}} \frac{\Delta\underline{u}^T [\mathbf{I} - \mathcal{P}_F]\underline{h}\underline{h}^T[\mathbf{I} - \mathcal{P}_F]\Delta\underline{u}}{\Delta\underline{u}^T \mathbf{C}^{-1}\Delta\underline{u}} \\ &= \max_{\underline{u}} \frac{\underline{u}^T \mathbf{C}^{\frac{1}{2}}[\mathbf{I} - \mathcal{P}_F]\underline{h}\underline{h}^T[\mathbf{I} - \mathcal{P}_F]\mathbf{C}^{\frac{1}{2}}\underline{u}}{\underline{u}^T\underline{u}} \\ &= \underline{h}^T[\mathbf{I} - \mathcal{P}_F]\mathbf{C}[\mathbf{I} - \mathcal{P}_F]\underline{h} \quad (54) \end{aligned}$$

where in the third line $\underline{u} = \mathbf{C}^{-\frac{1}{2}}\Delta\underline{u}$.

In view of the relation (6) combination of (53) and (54) yields the following lower bound on the norm $\|\nabla b_{\underline{u}}\|_C$

$$\|\nabla b_{\underline{u}}\|_C \geq \sqrt{\underline{h}^T[\mathbf{I} - \mathcal{P}_F]\mathbf{C}[\mathbf{I} - \mathcal{P}_F]\underline{h}} + \epsilon \quad (55)$$

For the case that the bias $b_{\underline{u}}$ is linear over $\underline{u} \in \mathcal{C}$ ϵ in (55) is equal to zero and we have an exact bound $\|\nabla b_{\underline{u}}\|_C \geq \sqrt{\underline{h}^T[\mathbf{I} - \mathcal{P}_F]\mathbf{C}[\mathbf{I} - \mathcal{P}_F]\underline{h}}$. In view of the relation (53), replacing $\underline{h}$ with $\nabla t_{\underline{\theta}}$ in this bound will give an approximate bound when $t_{\underline{u}}$ is non-linear but smooth over $\underline{u} \in \mathcal{C}$. The bound (55) with $\epsilon = 0$ will probably be fairly tight when the range space components of $\underline{u}$ can be estimated without bias, which is true for linear models $Y = \mathbf{A}\underline{u} + \underline{w}$. However, the reader is cautioned that for non-linear models the bound (55) may not be very tight since unbiased estimators may not exist even for components lying in the range space of $\mathbf{F}_Y$ [42].

Appendix D: Asymptotic Approximation of Bias, Bias Gradient, and Variance for Poisson QPML

Define the vector

$$\begin{aligned} \underline{z} &= [\mathbf{F}_Y(\underline{\theta}) + \beta\mathbf{P}]^{-1}\mathbf{F}_Y(\underline{\theta})\underline{\theta}, \\ &= \left[\mathbf{I} - \left[\mathbf{P} + \frac{1}{\beta}\mathbf{F}_Y(\underline{\theta})\right]^{-1}\mathbf{P}\right]\underline{\theta}. \quad (56) \end{aligned}$$

Here we derive the following asymptotic formulas for variance, bias, and bias gradient of the Poisson QPML estimator of a general differentiable function $t_{\underline{\theta}}$.

Asymptotic Variance:

$$\sigma^2_{\underline{\theta}} = \nabla^T t_{\underline{z}} \left[\mathbf{F}_Y(\underline{\theta}) + \beta\mathbf{P}\right]^{-1} \mathbf{F}_Y(\underline{\theta}) \left[\mathbf{F}_Y(\underline{\theta}) + \beta\mathbf{P}\right]^{-1} \nabla t_{\underline{z}}. \quad (57)$$

Asymptotic Bias:

$$b_{\underline{\theta}} = t_{\underline{z}} - t_{\underline{\theta}}, \quad (58)$$

Asymptotic Bias Gradient:

$$\nabla b_{\underline{\theta}} = \mathbf{F}_Y(\underline{\theta})\left[\beta\mathbf{P} + \mathbf{F}_Y(\underline{\theta})\right]^{-1} \nabla t_{\underline{z}} - \nabla t_{\underline{\theta}} - O\left(\frac{1}{\beta}\right) \quad (59)$$

where

$$\begin{aligned} O\left(\frac{1}{\beta}\right) &= \frac{1}{\beta}\sum_{j=1}^{m} \underline{\gamma}_j(\underline{\theta})\mathbf{A}^T_{j*}\mathbf{B}^{-1}\nabla t_{\underline{z}} \quad (60) \\ \underline{\gamma}_j(\underline{\theta}) &= \frac{\mathbf{A}_{j*}\underline{\theta}^T\mathbf{P}\mathbf{B}^{-1}\mathbf{A}_{j*}}{\mu^2_j(\underline{\theta})}, \quad (61) \end{aligned}$$

and $\nabla t_{\underline{z}}$ denotes the evaluation of the gradient of $t_{\underline{\theta}}$ at the point $\underline{\theta} = \underline{z}$.

Define the ambiguity function $a(\underline{u}, \underline{\theta}) = E_{\underline{\theta}}[J(\underline{u})]$ and let $\underline{u} = \underline{z} = \underline{z}(\underline{\theta})$ be the root of the equation $\underline{\lambda}(\underline{u}) = 0$ where $\underline{\lambda}(\underline{u}) = \nabla^{10}a(\underline{u}, \underline{\theta})$. Assuming the technical conditions underlying [50, Corollary 3.2, Sec. 6.3] are satisfied[2] we have the following approximation: in the limit of large observation time the estimator $\hat{\underline{\theta}}$ is asymptotically normal with mean $\underline{z}$ and covariance matrix $\mathbf{\Sigma} = \left[\nabla^{20}a(\underline{z}, \underline{\theta})\right]^{-1}\ \mathbf{G}_{\underline{z}}\ \left[\nabla^{20}a(\underline{z}, \underline{\theta})\right]^{-T}$ where $\mathbf{G}_{\underline{z}} = \mathrm{cov}_{\underline{\theta}}(\nabla J(\underline{z}))$. Furthermore, assuming that the function $t_{\underline{\theta}}$ has nonzero derivative at $\underline{\theta} = \underline{z}$, the estimator $\hat{t} = t_{\hat{\underline{\theta}}}$ is asymptotically normal with mean $t_{\underline{z}}$ and variance $\nabla^T t_{\underline{z}}\mathbf{\Sigma}\nabla t_{\underline{z}}$ [51, p. 122, Theorem A]. This gives the asymptotic expression for bias $b_{\underline{\theta}} = E_{\underline{\theta}}[\hat{t} - t_{\underline{\theta}}] = t_{\underline{z}} - t_{\underline{\theta}}$, as given in (58), and an asymptotic expression for variance

$$\mathrm{var}_{\underline{\theta}}(\hat{t}) = \nabla^T t_{\underline{z}} \left[\nabla^{20}a(\underline{z}, \underline{\theta})\right]^{-1}\ \mathbf{G}_{\underline{z}}\ \left[\nabla^{20}a(\underline{z}, \underline{\theta})\right]^{-1}\nabla t_{\underline{z}}. \quad (62)$$

Since the penalized Poisson likelihood function $J(\underline{u})$ in (41) is linear in the observations $\underline{Y}$, and Y_i are independent Poisson random variables, it is simple to derive the following expression for the covariance matrix of $\nabla J(\underline{z}) = \sum_{j=1}^T \mathbf{A}_{j*}\left(\frac{Y_j}{\mu_j(\underline{z})} - 1\right) - \beta\mathbf{P}\underline{z}$:

$$\begin{aligned} \mathrm{cov}_{\underline{\theta}}[\nabla J(\underline{z})] &= \mathbf{F}(\underline{z}, \underline{\theta}) \\ &= \sum_{j=1} \mathbf{A}_{j*}\mathbf{A}^T_{j*}\frac{1}{\mu_j(\underline{\theta})}\left(\frac{\mu_j(\underline{\theta})}{\mu_j(\underline{z})}\right)^2 \\ &= \mathbf{F}_Y(\underline{\theta}) + o\left(\underline{\mu}(\underline{z} - \underline{\theta})\right), \quad (63) \end{aligned}$$

where $\mathbf{A}^T_{j*}$ is the j-th row of $\mathbf{A}$, $\mathbf{F}_Y(\underline{\theta})$ is the Fisher matrix (28), and $o\left(\underline{\mu}(\underline{z} - \underline{\theta})\right)$ is a remainder term which is close to zero when the difference between the projections $\underline{\mu}(\underline{z})$ and $\underline{\mu}(\underline{\theta})$ is small. To obtain the expression (63) with remainder term we used the series development

$$\frac{\mu_j(\underline{\theta})}{\mu_j(\underline{z})} = 1 - \frac{1}{\mu_j(\underline{\theta})}(\underline{z} - \underline{\theta})^T\mathbf{A}_{j*} + o\left((\underline{z} - \underline{\theta})^T\mathbf{A}_{j*}\right).$$

[2] Among other things these conditions involve showing that the gradient function $\nabla J(\underline{\theta})$ converges a.s. to zero as the observation time increases.

The ambiguity function is

$$a(\underline{u},\underline{\theta}) = \sum_{j=1}^{m} \left(\mu_j(\underline{\theta}) \ln \mu_j(\underline{u}) - \mu_j(\underline{u})\right) - \frac{\beta}{2}\underline{u}^T \mathbf{P}\underline{u}.$$

Differentiation of the ambiguity function with respect to $\underline{u}$ yields

$$\begin{aligned} \nabla^{10} a(\underline{u},\underline{\theta}) &= \sum_{j=1}^{m} \mathbf{A}_{j*}\left(\frac{\mu_j(\underline{\theta})}{\mu_j(\underline{u})} - 1\right) - \beta\mathbf{P}\underline{u} \qquad (64) \\ &= -\left(\sum_{j=1}^{m} \frac{1}{\mu_j(\underline{\theta})}\mathbf{A}_{j*}^T\mathbf{A}_{j*}\right)(\underline{u}-\underline{\theta}) - \beta\mathbf{P}\underline{u} \\ &\quad + o\left(\underline{\mu}(\underline{u}-\underline{\theta})\right) \\ &= -\mathbf{F}_Y(\underline{\theta})(\underline{u}-\underline{\theta}) - \beta\mathbf{P}\underline{u} + o\left(\underline{\mu}(\underline{u}-\underline{\theta})\right), \end{aligned}$$

and similarly

$$\nabla^{20} a(\underline{u},\underline{\theta}) = -\mathbf{F}_Y(\underline{\theta}) - \beta\mathbf{P} + O\left(\underline{\mu}(\underline{u}-\underline{\theta})\right). \qquad (65)$$

Neglecting the $O\left(\underline{\mu}(\underline{u}-\underline{\theta})\right)$ remainder terms, multiplication of the inverse of (65) the covariance (63) and the inverse transpose of (65), yields the asymptotic variance expression (62). Likewise, neglecting the remainder term in (64) and solving for the root $\underline{u} = \underline{z}$ of the equation $\nabla^{10} a(\underline{u},\underline{\theta}) = 0$ yields the expression for the root (56).

We next derive the expression (59) for the bias gradient. Applying the chain rule to differentiate the bias function (58) we obtain

$$\nabla b_{\underline{\theta}} = \nabla \underline{z}_{\underline{\theta}}^T \, \nabla t_{\underline{z}} - \nabla t_{\underline{\theta}}, \qquad (66)$$

where $\nabla \underline{z}_{\underline{\theta}}^T$ is an $n \times n$ matrix of derivatives of $\underline{z} = \underline{z}_{\underline{\theta}}$. From (56) the k-th row of this matrix is

$$\begin{aligned} \frac{d}{d\theta_k}\underline{z}_{\underline{\theta}}^T &= \underline{e}_k^T\left[\mathbf{I} - \mathbf{P}\left[\mathbf{P} + \frac{1}{\beta}\mathbf{F}_Y(\underline{\theta})\right]^{-1}\right] \qquad (67) \\ &\quad - \underline{\theta}^T\mathbf{P}\frac{d}{d\theta_k}\left[\mathbf{P} + \frac{1}{\beta}\mathbf{F}_Y(\underline{\theta})\right]^{-1} + O\left(\underline{\mu}(\underline{u}-\underline{\theta})\right). \end{aligned}$$

Define the matrix $\mathbf{B} = \mathbf{P} + \frac{1}{\beta}\mathbf{F}_Y(\underline{\theta})$. From the differentiation formula $\frac{d}{d\theta_k}\mathbf{B}(\theta)^{-1} = -\mathbf{B}^{-1}\frac{d}{d\theta_k}\mathbf{B}(\theta)\mathbf{B}^{-1}$, and from expression (40) for the Fisher information matrix $\mathbf{F}_Y(\underline{\theta})$ we have

$$\begin{aligned} &-\underline{\theta}^T\mathbf{P}\frac{d}{d\theta_k}\left[\mathbf{P} + \frac{1}{\beta}\mathbf{F}_Y(\underline{\theta})\right]^{-1} \\ &= \frac{1}{\beta}\underline{\theta}^T\mathbf{P}\mathbf{B}^{-1}\,\frac{d}{d\theta_k}\mathbf{F}_Y(\underline{\theta})\,\mathbf{B}^{-1} \\ &= -\frac{1}{\beta}\sum_{j=1}^{m}\frac{\mathbf{A}_{jk}\underline{\theta}^T\mathbf{P}\mathbf{B}^{-1}\mathbf{A}_{j*}}{\mu_j^2(\underline{\theta})}\,\mathbf{A}_{j*}^T\mathbf{B}^{-1} \\ &= -\frac{1}{\beta}\sum_{j=1}^{m}\gamma_{jk}(\underline{\theta})\,\mathbf{A}_{j*}^T\mathbf{B}^{-1}, \qquad (68) \end{aligned}$$

where $\gamma_{jk}(\underline{\theta})$ is the k-th element of the vector $\underline{\gamma}_j$ defined in (61). Combining the results (67) and (68) we obtain

$$\nabla \underline{z}_{\underline{\theta}} = \left[\mathbf{I} - \mathbf{P}\left[\mathbf{P} + \frac{1}{\beta}\mathbf{F}_Y(\underline{\theta})\right]^{-1}\right] + O\left(\frac{1}{\beta}\right),$$

which, when substituted into (66) and neglecting the remainder term $O\left(\underline{\mu}(\underline{u}-\underline{\theta})\right)$ yields the bias gradient expression (59).

References

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

[2] P. J. Daniell, "Discussion of paper by m. s. bartlett," *J. Royal Statistical Society, Ser. B*, vol. 8, pp. 27, 1946.

[3] R. B. Blackman and J. W. Tukey, "The measurement of power spectra from the point of view of communication engineering," *Bell Syst. Tech. Journ.*, vol. 37, pp. 183–282, 1958.

[4] R. Lagendijk and J. Biemond, *Iterative identification and restoration of images*, Kluwer Academic Publishers, Boston, 1991.

[5] C. M. Stein, "Multiple regression," in *Contributions to probability and statistics: Essays in honor of Harold Hotelling*, I. Olkin, S. Ghurge, W. Hoeffding, W. Nadow, and H. Mann, editors, pp. 424–443, Stanford University Press, 1960.

[6] A. E. Hoerl and R. W. Kennard, "Ridge regression: biased estimation of non-orthogonal components," *Technometrics*, vol. 12, pp. 55–67, 1970.

[7] J.-S. Liow and S. C. Strother, "Practical tradeoffs between noise, quantitation, and number of iterations for maximum likelihood-based reconstructions," *IEEE Trans. on Medical Imaging*, vol. 10, no. 4, pp. 563–571, 1991.

[8] S. J. Lee, G. R. Gindi, I. G. Zubal, and A. Rangarajan, "Using ground-truth data to design priors in Bayesian SPECT reconstruction," in *Information Processing in Medical Imaging*, Y. Bizais, C. Barillot, and R. D. Paola, editors, Kluwer, 1995.

[9] J. A. Fessler, "Penalized weighted least-squares image reconstruction for positron emission tomography," *IEEE Trans. on Medical Imaging*, vol. 13, no. 2, pp. 290–300, June 1994.

[10] R. E. Carson, Y. Yan, B. Chodkowski, T. K. Yap, and M. E. Daube-Witherspoon, "Precision and accuracy of regional radioactivity quantitation using the maximum likelihood EM reconstruction algorithm," *IEEE Trans. on Medical Imaging*, vol. 13, no. 3, pp. 526–537, September 1994.

[11] D. Wand, V. Haese-Coat, A. Bruno, and J. Ronsin, "Some statistical properties of mathematical morphology," *Signal Processing*, vol. 43, no. 8, pp. 1955–65, August 1995.

[12] N. Himayat and S. A. Kassam, "Approximate performance analysis of edge preserving fil ters," *Signal Processing*, vol. 41, no. 9, pp. 2764–2777, September 1993.

[13] M. B. Woodroofe and J. W. Van Ness, "The maximum deviation of sample spectral densities," *Ann. Math. Statist.*, vol. 38, pp. 1558–1569, 1967.

[14] P. Bloomfield, *Fourier Analysis of Time Series*, Wiley, New York, 1976.

[15] A. O. Hero, "A Cramer-Rao type lower bound for essentially unbiased parameter estimation," Technical Report 890, M.I.T. Lincoln Laboratory, Lexington, MA, Lexington, MA 02173-0073, Jan., 1992. DTIC AD-A246666.

[16] J. Fessler and A. Hero, "Cramer-Rao bounds for biased estimators in image restoration," in *Proc. of 36th IEEE Midwest Symposium on Circuits and Systems*, Detroit, MI, Aug. 1993.

[17] M. Usman, A. Hero, J. A. Fessler, and W. Rogers, "Bias-variance tradeoffs analysis using uniform CR bound for a SPECT system," in *Proc. of IEEE Nuclear Science Symposium and Medical Imaging Conf*, pp. 1463–1467, San Francisco, November 1993.

[18] M. Usman, A. Hero, and J. A. Fessler, "Bias-variance tradeoffs analysis using uniform CR bound for images," in *Proc. of IEEE Image Processing Conference*, pp. 835–839, Austin, TX, November 1994.

[19] C. R. Rao, *Linear Statistical Inference and Its Applications*, Wiley, New York, 1973.

[20] J. A. Fessler and A. O. Hero, "Space-alternating generalized EM algorithm," *IEEE Trans. on Signal Processing*, vol. SP-42, no. 10, pp. 2664–2677, Oct. 1994.

[21] G. E. P. Box and N. R. Draper, "A basis for the selection of a response surface design," *J. Am. Statist. Assoc.*, vol. 54, pp. 622–654, 1959.

[22] M. A. Ali, "On a class of shrinkage estimators of the vector of shrinkage coefficients," *Commun. Statist.-Theory Meth.*, vol. 18, no. 12, pp. 4491–4500, 1981.

[23] T. Bednarski, "On minimum bias and variance estimation for parametric models with shrinking contamination," *Probability and Math. Statist.*, vol. 6, no. 2, pp. 121–129, 1985.

[24] M. Quenouille, "Approximate tests of correlation in time series," *J. Royal Statistical Society, Ser. B*, vol. 11, pp. 18–84, 1949.

[25] J. Tukey, "Bias and confidence in not quite large samples," *Ann. Math. Statist.*, vol. 29, pp. 614, 1956.

[26] B. Efron, "Bootstrap methods: another look at the jacknife," *Annals of Statistics*, vol. 7, pp. 1–26, 1979.

[27] I. A. Ibragimov and R. Z. Has'minskii, *Statistical estimation: Asymptotic theory*, Springer-Verlag, New York, 1981.

[28] M. Fréchet, "Sur l'extension de certaines évaluations statistiques de petits echantillons," *Rev. Inst. Int. Statist*, vol. 11, pp. 128–205, 1943.

[29] G. Darmois, "Sur les lois limites de la dispersion de certaines estimations," *Rev. Inst. Int. Statist*, vol. 13, pp. 9–15, 1945.

[30] H. Cramér, "A contribution to the theory of statistical estimation," *Skand. Aktuaries Tidskrift*, vol. 29, pp. 458–463, 1946.

[31] C. R. Rao, "Minimum variance and the estimation of several parameters," *Proc. Cambridge Phil. Soc.*, vol. 43, pp. 280–283, 1946.

[32] H. L. Van-Trees, *Detection, Estimation, and Modulation Theory: Part I*, Wiley, New York, 1968.

[33] J. D. Gorman and A. O. Hero, "Lower bounds for parametric estimation with constraints," *IEEE Trans. on Inform. Theory*, vol. IT-36, pp. 1285–1301, Nov. 1990.

[34] M. Usman, "Biased and unbiased Cramer-Rao bounds: computational issues and applications," PhD Thesis, Dept EECS, The University of Michigan, Aug. 1994.

[35] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery, *Numerical recipes in C: The art of scientific computing (2nd Ed.)*, Cambridge University Press, 1992.

[36] A.O. Hero, M. Usman, A.C. Sauve and J.A. Fessler, "Recursive algorithms for computing the Cramer-Rao bound," *IEEE Trans. on Signal Processing*, in review.

[37] B. E. Efron, *The jackknife, the bootstrap, and other resampling plans*, SIAM Press, Philadelphia PA, 1982.

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

[39] K. Lange, "Convergence of EM image reconstruction algorithms with Gibbs smoothing," *IEEE Trans. on Medical Imaging*, vol. 9, no. 4, pp. 439–446, June 1991.

[40] T. Hebert and R. Leahy, "A generalized EM algorithm for 3-D Bayesian reconstruction from Poisson data using Gibbs priors," *IEEE Trans. on Medical Imaging*, vol. 8, no. 2, pp. 194–203, 1989.

[41] V. Johnson, W. Wong, X. Hu, and C. Chen, "Image restoration using Gibbs priors: boundary modeling, treatment of blurring, and selection of hyperparameters," *IEEE Trans. on Pattern Anal. and Machine Intell.*, vol. PAMI-13, no. 5, pp. 413–425, 1991.

[42] R. C. Liu and L. D. Brown, "Nonexistence of informative unbiased estimators in singular problems," *Annals of Statistics*, vol. 21, no. 1, pp. 1–13, 1993.

[43] G. H. Golub and C. F. Van Loan, *Matrix Computations (2nd Edition)*, The Johns Hopkins University Press, Baltimore, 1989.

[44] D. L. Snyder and M. I. Miller, *Random Point Processes in Time and Space*, Springer-Verlag, New York, 1991.

[45] A. O. Hero and J. A. Fessler, "A recursive algorithm for computing CR-type bounds on estimator covariance," *IEEE Trans. on Inform. Theory*, vol. 40, pp. 1205–1210, July 1994.

[46] J. A. Fessler and A. O. Hero, "Penalized maximum likelihood image reconstruction using space alternating generalized EM algorithms," *IEEE Trans. on Image Processing*, vol. 3, , to appear Oct. 1995.

[47] J. A. Fessler and W. L. Rogers, "Uniform quadratic penalties cause nonuniform image resolution (and sometimes vice versa)," in *Proc. of IEEE Nuclear Science Symposium*, volume 4, pp. 1915–1919, 1994.

[48] J. A. Fessler, "Resolution properties of regularized image reconstruction methods," Technical Report 297, Comm. and Sig. Proc. Lab. (CSPL), Dept. EECS, University of Michigan, Ann Arbor, August 1995.

[49] E. B. Manoukian, *Modern Concepts and Theorems of Mathematical Statistics*, Springer-Verlag, New York N.Y., 1986.

[50] P. J. Huber, *Robust Statistics*, Wiley, New York, 1981.

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

BIOGRAPHIES

Alfred O. Hero III was born in Boston, MA. in 1955. He received the B.S. *summa cum laude* from Boston University (1980) and the Ph.D from Princeton University (1984), both in Electrical Engineering. He held the G.V.N. Lothrop Fellowship in Engineering at Princeton University. He is presently Associate Professor of Electrical Engineering and Computer Science and Area Chair for Signal Processing at the University of Michigan, Ann Arbor. He has held positions of Visiting Scientist at M.I.T. Lincoln Laboratory, Lexington, MA (1987 - 1989), Visiting Professor at Ecole Nationale de Techniques Avancées (ENSTA), Paris, France (1991), and William Clay Ford Fellow at Ford Motor Company, Dearborn. MI (1993). His research interests are in the areas of detection and estimation theory applied to statistical signal and image processing. Alfred Hero is a member of Tau Beta Pi, the American Statistical Association, and Commission C of the International Union of Radio Science (URSI). He is currently an Associate Editor for the IEEE Transactions on Information Theory. He was Chairman for Publicity for the 1986 IEEE International Symposium on Information Theory. He was General Chairman for the 1995 IEEE International Conference on Acoustics, Speech, and Signal Processing.

Jeff Fessler received the BSEE degree from Purdue University in 1985, the MSEE degree from Stanford University in 1986, and the M.S. degree in Statistics from Stanford University in 1989. From 1985 to 1988 he was a National Science Foundation Graduate Fellow at Stanford, where he earned a Ph.D. in electrical engineering in 1990. He has worked at the University of Michigan since then. From 1991 to 1992 he was a Department of Energy Alexander Hollaender Post-Doctoral Fellow in the Division of Nuclear Medicine. From 1993 to 1995 he was an Assistant Professor in Nuclear Medicine and the Bioengineering Program. Since 1995 he has been with the Department of Electrical Engineering and Computer Science. His research interests are in statistical aspects of imaging problems.

Mohammad Usman received his B.S. degree in electrical engineering from the University of Engineering and Technology, Lahore, Pakistan in 1987, M.S. in electronics engineering from Philips International Institute, Eindhoven, the Netherlands in 1989, and Ph.D. in electrical engineering and computer science from the University of Michigan, Ann Arbor MI, USA in 1994. From 1988 to 1989 he was at the Philips Research Laboratories, Eindhoven, the Netherlands. From 1989 to 1990 he was at the Advanced Development Center, Philips Industrial Activities, Leuven, Belgium. Since 1995 he is on the faculty of FAST Institute of Computer Science, Lahore, Pakistan, where he is an Assistant Professor. He is also a consultant to Cressoft, Denver, CO. His current research interests include image reconstruction, speech compression, and code optimization. He is a member of IEEE.