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 nonlinear 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 \( \mathbf{\theta} = [\theta_1, ..., \theta_n]^T \in \Theta \) be a vector of unknown and non-random parameters which parameterize the density \( f_Y(y; \mathbf{\theta}) \) of an observed random variable \( Y \). The parameter space \( \Theta \) is assumed to be an open subset of \( n \)-dimensional Euclidean space \( \mathbb{R}^n \). For fixed \( \mathbf{\theta} \) let \( t = t(Y) \) be an estimator of the scalar \( t_\mathbf{\theta} \), where \( t : \Theta \rightarrow \mathbb{R} \) is a specified function. Let this estimator have bias \( b_\mathbf{\theta} = E_{\mathbf{\theta}}[t] - t_\mathbf{\theta} \) and variance \( \sigma^2_\mathbf{\theta} = E_{\mathbf{\theta}}[(t - E_\mathbf{\theta}[t])^2] \). 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 \( b_\mathbf{\theta} \) and \( \sigma^2_\mathbf{\theta} \) 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 \( \mathbf{\theta} \). Furthermore, the bias value \( b_\mathbf{\theta} \) 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 \( \delta_\mathbf{\theta} = ||\nabla b_\mathbf{\theta}|| \) and the square root variance \( \sqrt{\sigma^2_\mathbf{\theta}} \), 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 \( \mathbf{\theta} \) 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. 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
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 $\nabla \ln f(y; \hat{\theta})$ 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 $\nabla b_\theta$ of these estimators is closely related to the point spread function of the estimator when one wishes to estimate a single component $t_k = \theta_k$. 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 $\hat{\theta}$. 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 $\hat{\theta}$ be an estimator of the scalar differentiable function $t_\theta$. The mean-square-error (MSE) is a widely used measure of performance for an estimator $\hat{\theta}$ and is simply related to the estimator bias $b_\theta$ and the estimator variance $\sigma_\theta^2$ through the relation $\text{MSE}_{\hat{\theta}} = b_\theta^2 + \sigma_\theta^2$. 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 $b_\theta$ and $\sigma_\theta^2$ one can derive other important measures such as signal-to-noise-ratio $\text{SNR} = \frac{|t_\theta + b_\theta|^2}{\sigma_\theta^2}$, coefficient of variation $1/\sqrt{\text{SNR}},$ and generalized MSE $= \alpha g_1(b_\theta) + (1 - \alpha) g_2(\sigma_\theta^2)$, 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 $(b_\theta, \sigma_\theta^2)$ provides essential information for constructing and evaluating $t$-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 $\hat{\theta}$ whose bias function $b : \Theta \to \mathbb{R}$ is constant is as good as unbiased since the bias can be removed without knowledge of $\Theta$. 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 $f_\theta(y; \theta)$ is sufficiently smooth to guarantee existence of the Fisher information matrix (defined below), $b_\theta$ is always differentiable regardless of the form of the estimator as long as $E_{\theta}[f^2]$ is upper bounded [27, Lemma 7.2]. In this case the bias gradient $\nabla b_\theta : \Theta \to \mathbb{R}^n$ uniquely specifies the bias $b_\theta$ up to an additive constant

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

where $\theta^*$ is a point such that the line segment connecting $\theta^*$ and $\theta$ is contained in $\Theta$ such a point is guaranteed to exist when $\Theta$ is convex or star shaped. Thus the gradient function $\nabla b_\theta = \left[ \frac{\partial b_\theta}{\partial \theta_1}, \ldots, \frac{\partial b_\theta}{\partial \theta_n} \right]^T$ (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

$$\delta_\theta = \| \nabla b_\theta \|_C$$

(1)

where the norm $\| \cdot \|_C$ is defined in terms of a symmetric positive definite matrix $C$

$$\|u\|_C^2 = u^T C u$$

(2)

We will use the notation $\|u\|_b$ to denote the Euclidean norm obtained when $C = I$. 

The norm of the bias gradient at a point \( \hat{\eta} = \hat{\theta} \) is a measure of the sensitivity of the estimator mean \( m_\hat{\eta} = E[\hat{\eta}] \) to changes in \( \eta \) over a neighborhood of \( \theta \). 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 \( C = \mathcal{D}(\theta, C) = \{ \eta : (\eta - \theta)^T C^{-1} (\eta - \theta) \leq 1 \} \) where \( \hat{\theta} \) is a point in \( \Theta \) and \( C \) is a symmetric positive definite matrix. The maximal eigenvalue of the ellipsoid is \( 2\sqrt{\lambda_C} \), where \( \lambda_C \) is the maximal eigenvalue of \( C \). Assume that the bias function \( b_\hat{\eta} \) is continuously twice differentiable and that the magnitude of the eigenvalues of the Hessian matrix \( \nabla^2 b_\hat{\eta} = \nabla \nabla^T b_\hat{\eta} \) are upper bounded over \( \eta \in C \) by a non-negative constant \( \alpha < \infty \). Then, using (1) and the Taylor expansion with remainder, the maximal squared variation of the bias \( b_\hat{\eta} \) over \( C \) is

\[
\max_{\omega \in C} |b_{\hat{\eta}} - b_\omega|^2 = \max_{\omega \in C} |\nabla^T b_{\omega} \Delta \omega + \frac{1}{2} \nabla^T b_{\omega} \nabla^2 b_{\omega} \Delta \omega|^2
\]

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

\[
\max_{\omega \in C} |b_{\hat{\eta}} - b_\omega|^2 = \max_{\omega \in C} |\nabla^T b_{\omega} \Delta \omega|^2 (1 + \epsilon),
\]

where \( \epsilon = O(1) \). Hence we see that when \( \epsilon \ll 1 \) the norm \( \| b_{\hat{\eta}} \|_C \) is approximately equal to the maximal bias variation over the ellipsoidal neighborhood \( \mathcal{D}(\theta, C) \) of \( \hat{\theta} \). Note that this occurs when when the product of the ellipsoid width \( \sqrt{\lambda_C} \) and the ratio of the curvature \( \alpha \) of the bias function to the bias gradient norm \( \sqrt{\nabla^T b_{\omega} \nabla b_{\omega}} \) is small. For the special case where the bias is a linear function \( b_{\hat{\eta}} = L^T \hat{\theta} - c, L \in \mathbb{R} \) and \( \eta = 0 \) 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 \( \| \cdot \|_C \) 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 \( \Delta_{\hat{\theta}} \) and variance \( \sigma^2_{\hat{\theta}} \) we plot the pair \( (\Delta_{\hat{\theta}}, \sigma^2_{\hat{\theta}}) \) 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 \( (\Delta_{\hat{\theta}}, \sigma^2_{\hat{\theta}}) \), 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 \( \hat{\theta} \) with mean \( \hat{\theta} = E[\hat{\theta}] \) the CR bound has the following form, referred to here as the biased CR bound

\[
\frac{\sigma_{\hat{\theta}}^2}{\lambda} \geq \frac{[\nabla f(Y; \hat{\theta})]^T F_Y^+ [\nabla f(Y; \hat{\theta})]}{\lambda}
\]

where \( F_Y = F_Y(\hat{\theta}) \) is the \( n \times n \) Fisher information matrix

\[
F_Y = E_{\hat{\theta}} \left\{ [\nabla f(Y; \hat{\theta}) f(Y; \hat{\theta})]^T \right\}
\]

and \( F_Y^+ \) denotes the Moore-Penrose pseudo-inverse matrix of the possibly singular matrix \( F_Y \).

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

\[
E_{\hat{\theta}} \left\{ UU^T \right\} \geq E_{\hat{\theta}} \left\{ UVV^T \right\} \left( E_{\hat{\theta}} \left\{ VV^T \right\} \right)^+ E_{\hat{\theta}} \left\{ VU^T \right\},
\]

and using the well known identities \( E_{\hat{\theta}} \left\{ \nabla \ln f(Y; \hat{\theta}) \right\} = \Theta \) and \( E_{\hat{\theta}} \left\{ \nabla \ln f_Y(Y; \hat{\theta}) \right\} = \nabla_{\hat{\theta}} \).\( \left\{ \right\} \) easily derivable from (18) below.

The biased CR bound (7) only applies to the class of estimators \( \hat{\eta} \) which have a particular bias gradient function \( \nabla_{\hat{\theta}} \). 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 \( \| \cdot \|_C \), and arbitrary differentiable function \( t_{\hat{\theta}} \). For a proof of this theorem see Appendix A.

**Theorem 1:** Let \( \hat{\eta} \) be an estimator of the scalar differentiable function \( t_{\hat{\theta}} \) of the parameter \( \hat{\theta} = [\theta_1, \ldots, \theta_n]^T \). For a
fixed $\delta \geq 0$ let the bias gradient of $\hat{t}$ satisfy the norm constraint $||\nabla b_h||_C \leq \delta$, where $C$ is an arbitrary $n \times n$ symmetric positive definite matrix. Define $P_B$ as the $n \times n$ matrix which projects onto the column space of $B = C^{-1/2}F_Y C^{-1/2}$. Then the variance of $\hat{t}$ satisfies:

$$\sigma^2_x \geq B(\hat{t}, \delta),$$

where if $\delta^2 \geq \nabla^T t_x C^{-1/2} P_B C^{1/2} \nabla t_x$ then $B(\hat{t}, \delta) = 0$, while if $\delta^2 < \nabla^T t_x C^{-1/2} P_B C^{1/2} \nabla t_x$ then:

$$B(\hat{t}, \delta) = [\nabla^T t_x + d_{\text{min}}]^T F_Y^+ [\nabla^T t_x + d_{\text{min}}],$$

$$= \lambda^2 \nabla^T t_x C [\lambda C + F_Y^+]^T F_Y^+ [\lambda C + F_Y^+]^{-1} C \nabla t_x,$$

where in (9)

$$d_{\text{min}} = -[\lambda C + F_Y^+]^{-1} F_Y^+ \nabla t_x.$$  \hspace{1cm} (10)

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

$$g(\lambda) = \nabla^T t_x F_Y^+ [\lambda C + F_Y^+]^{-1} C [\lambda C + F_Y^+]^{-1} F_Y^+ \nabla t_x$$

\hspace{1cm} (11)

![Figure 1](image.png)

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

By tracing out the family of points \{$(\delta, \sqrt{B(\hat{t}, \delta)}) : \delta \geq 0$\} one obtains a curve in the $\delta \sigma$ plane for a particular $\hat{t} \in \Theta$. The curve is always monotone non-increasing in $\delta$. Since $B(\hat{t}, \delta)$ is a lower bound on $\sigma^2_x$ 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 $\sigma = \sqrt{B(\hat{t}, \delta)/B(\hat{t}, 0)}$. 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 $\lambda$ over the range $(0, \infty)$ and sweep out the curve by using the $\lambda$-parameterizations of $g(\lambda) = \delta^2$ and $B(\hat{t}, \lambda)$ specified by relations (11) and (9), respectively.

**Comments:**

- The uniform bound $B(\hat{t}, \lambda)$ is always less than or equal to the unbiased CR bound $B(\hat{t}, 0) = \nabla^T t_x F_Y^+ \nabla t_x$. The slope of $B(\hat{t}, \lambda)$ at $\lambda = 0$ gives a bias sensitivity index $\eta$ for the unbiased CR bound. For non-singular $F_Y$ and single component estimation ($t_x = \theta_1$) it is shown in [15] that $\eta = 2\sqrt{1 + C^2 \bar{f}^2 C}$, where $C$ is the first column of $F_Y$ and $\bar{f}$ is the principal minor of $[F_Y]_{11}$. 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 $P_B$ can be expressed either as $P_B = B [B^T B]^+ B^T = B^+ B = B B^+$, or via the eigendecomposition of $B$ as $P_B = \sum_{i=1}^r \xi_i \xi_i^T$, where $r$ is the rank of $F_Y$ and $\{\xi_i\}_{i=1}$ are the orthonormal eigenvectors associated with the non-zero eigenvalues of $B$. By using properties of the Moore-Penrose pseudo-inverse it can be shown that $\nabla^T t_x C^{-1} P_B C^{-1} \nabla t_x = \nabla^T t_x F_Y^+ [C^{-1} F_Y^{-1}]^+ C^{-1} \nabla t_x$.

- When $F_Y$ is non-singular, $F_Y^+ = F_Y^{-1}$, $P_B = I$, $\nabla^T t_x C^{-1} P_B C^{-1} \nabla t_x = ||\nabla t_x||^2_C$ and (9)-(11) of Theorem 1 reduce to

$$B(\hat{t}, \lambda) = [\nabla^T t_x + d_{\text{min}}]^T F_Y^{-1} [\nabla^T t_x + d_{\text{min}}],$$

$$= \lambda^2 \nabla^T t_x [C^{1/2} + \lambda F_Y]^{-1} F_Y [C^{1/2} + \lambda F_Y]^{-1} \nabla t_x,$$

where

$$d_{\text{min}} = -C^{-1/2} \lambda C^{-1} F_Y^{-1} \nabla t_x.$$  \hspace{1cm} (13)

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

$$g(\lambda) = \nabla^T t_x [C^{1/2} + \lambda F_Y]^{-1} C^{-1} [C^{1/2} + \lambda F_Y]^{-1} \nabla t_x$$

When $C = I$ and $t_x = \theta_1$ these are identical to the results obtained in [15].

- In Theorem 1, $d_{\text{min}}$ defined in (10) is an optimal bias gradient in the sense that it minimizes the biased CR bound (7) over all vectors $\nabla b_h$ satisfying the constraint $||\nabla b_h||_C \leq \delta$. 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 $\delta^2 \geq \nabla^T t_x C^{-1} P_B C^{-1} \nabla t_x$, then the minimizing bias gradient is of the form $d_{\text{min}} = -P_B C^{1/2} \nabla t_x + \phi$, where $\phi$ is any vector satisfying $B \phi = 0$, and $||\phi||^2 \leq \delta^2 - \nabla^T t_x C^{-1} P_B C^{1/2} \nabla t_x$. Thus for the case of singular $F_Y$ there exist many optimal bias gradients.

- An estimator is said to locally achieve a bound in a neighborhood of a point $\hat{t}$ if the estimator achieves
the bound whenever the true parameter lies in the neighborhood. It has been shown [15] that if $F_Y$ is non-singular, if $\delta$ is small, if $t_2 = \theta_1$, and if the unbiased matrix CR bound is locally achievable by an unbiased estimator $\hat{\theta}^*$ in a neighborhood of a point $\hat{\theta}$, then one can construct an estimator that locally achieves the uniform bound in this neighborhood by introducing a small amount of bias into $\hat{\theta}^*$. However, since unbiased estimators may not exist for singular $F_Y$, the uniform CR bound for singular $F_Y$ may not be locally achievable. An example where the bound is globally achievable over all $\Theta$ 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 $C$ 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 locally tolerable and $C$ may be chosen to be of rank $n-1$ having the vector $\underline{1} = [1, \ldots, 1]^T$ in its nullspace. Let $B(\hat{\theta}, \delta), d_{\text{min}}$ and $g(\lambda)$ as defined in Theorem 1. Assume that $C$ is non-negative definite but $F_Y^{-1} + \lambda C$ is positive definite for $0 < \lambda < \infty$. For fixed $\delta > 0$ let the bias gradient of $\hat{t}$ satisfy the semi-norm constraint $\| \nabla b_2 \|_C \leq \delta$. Then

$$\text{var}_{\hat{t}}(\hat{t}) \geq B^*(\hat{\theta}, \delta),$$

where $B^*(\hat{\theta}, \delta) = \begin{cases} B(\hat{\theta}, g(\lambda)), & \text{if } 0 \leq \delta^2 \leq g(\lambda) \\ B(\hat{\theta}, \delta), & \text{if } g(\lambda) \leq \delta^2 \leq g(0) \\ 0, & \text{if } g(0) < \delta^2 < \infty \end{cases}$

and $g(\lambda) = \lim_{\lambda \to \infty} g(\lambda)$ 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 $B(\hat{\theta}, \delta)$ and $g(\lambda)$ in (9) and (11) was obtained in the process of proving the theorem (46) and (47)

$$B(\hat{\theta}, \delta) = \lambda^2 \nabla t_2 C^T \left[ |I + \lambda G|^{-1} G |I + \lambda G|^{-1} C^T \nabla t_2 \right]$$

$$g(\lambda) = \nabla t_2 \left[ C^T B |I + \lambda B|^{-2} B C^T \right] \nabla t_2 = \delta^2,$$

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

$$B(\hat{\theta}, \delta) = \sum_{i=1}^n \frac{\lambda^2 \beta_i}{(\lambda + \beta_i)^2} \| \nabla t_2 C_{\lambda i} \xi_i \|^2$$

$$\delta^2 = \sum_{i=1}^n \frac{\beta_i^2}{(\lambda + \beta_i)^2} \| \nabla t_2 C_{\lambda i} \xi_i \|^2,$$

where $\beta_i$ and $\xi_i$ denote an eigenvalue and eigenvector of $B$.

When $F_Y$ is ill-conditioned the computation of the matrix $B$ may be numerically unstable. In the case of non-singular $F_Y$ a simple algebraic manipulation in (12)-(14) yields

$$B(\hat{\theta}, \delta) = \lambda^2 \nabla t_2 C^T \left[ |I + \lambda G|^{-1} G |I + \lambda G|^{-1} C^T \nabla t_2 \right]$$

$$\delta^2 = \nabla t_2 C^T \left[ |I + \lambda G|^{-2} C^T \nabla t_2 \right]$$

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

The eigendecomposition of $G$ 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 $F_Y$. 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 $x = |I + \lambda F_Y C|^{-1} \nabla t_2$ by applying CG or GS iterations to solve the following linear equation for $x$:

$$|I + \lambda F_Y C| x = \nabla t_2.$$

3. Compute $y = C x$
4. Compute the point $(\delta, B(\hat{\theta}, \delta))$ via

$$B(\hat{\theta}, \delta) = \lambda^2 y^T F_Y y$$

$$\delta^2 = \sqrt{\frac{y^T y}{\lambda}}.$$

Since step 2 must be repeated for each value of $\lambda$, this method is competitive when one is interested in evaluation of the curve $B(\hat{\theta}, \delta)$ 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 $\beta_i$ and $\| |I + \lambda G|^{-1} C^T \nabla t_2 \|^2$ are available, the curve can be swept out over $\lambda$ 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 $\nabla b_2 \approx \frac{1}{\delta} \left[ b_{i+\delta} - b_i, \ldots, b_{i+\delta - \epsilon} - b_i \right]$ but this requires performing a simulation run for each coordinate perturbation $\hat{\theta} + \epsilon_i$. 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
\[
E_Z \left[ Z \theta \ln f_Z(Z; \theta) \right] = \int_Z z \frac{\partial f_Z(z; \theta)}{\partial \theta} \, dz = \frac{\partial}{\partial \theta} \int_Z z f_Z(z; \theta) \, dz = \frac{\partial}{\partial \theta} E_Z(Z).
\]
(18)
Thus in particular we have the following relation
\[
\nabla \theta \bar b = E_\theta \left[ \left( \hat Y(Y) - \zeta \right) \nabla \ln f_Y(Y; \theta) \right] - \nabla \theta \underline b.
\]
Since \( E_\theta \left[ \nabla \ln f_Y(Y; \theta) \right] = 0 \), an equivalent relation is
\[
\nabla \theta \bar b = E_\theta \left[ \left( \hat Y(Y_i) - \zeta \right) \nabla \ln f_Y(Y_i; \theta) \right] - \nabla \theta \underline b
\]
(19)
for any random variable \( \zeta \) statistically independent of \( Y \). As explained in the following discussion, the quantity \( \zeta \) 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 \( \nabla \theta \bar b \)
\[
\nabla \theta \bar b = \frac{1}{L} \sum_{i=1}^L \left( \hat Y(Y_i) - \zeta \right) \nabla \ln f_Y(Y_i; \theta) - \nabla \theta \underline b
\]
(20)
where \( \{Y_i\}_{i=1}^L \) is a set of i.i.d. realizations from \( f_Y(y_i; \theta) \). In (20) \( \{\zeta_i\}_{i=1}^L \) is any sequence of i.i.d. random variables such that \( Y_i, \zeta_i \) are statistically independent for each \( i \).

It can be shown that when \( \zeta_i = 0 \) for all \( i \) the covariance matrix of \( \nabla \theta \bar b \) is the matrix sum
\[
S(\nabla \theta \bar b) = \frac{1}{L} \text{cov}_\theta \left( (\hat Y_i - \bar Y) \nabla \ln f_Y(Y_i; \theta) \right) + \frac{1}{L} \left[ 2m \theta \theta + m \theta \theta F_Y \right]
\]
(21)
where
\[
\theta \theta = E_\theta \left[ (\hat Y(Y_i) - \bar Y)^2 \nabla \ln f_Y(Y_i; \theta) \right] \nabla \ln f_Y(Y_i; \theta) ]^T, \quad \text{and} \quad F_Y = E_\theta \left[ \nabla \ln f_Y(Y_i; \theta) \nabla \ln f_Y(Y_i; \theta) ]^T \right]
\]

A simple calculation shows that the covariance of (22) is the matrix sum
\[
S(\nabla \theta \bar b) = \frac{1}{L} \text{cov}_\theta \left( (\hat Y_i - \bar Y)^2 \nabla \ln f_Y(Y_i; \theta) \right) + \frac{1}{L(L-1)} \left[ \text{var}_\theta ((\hat Y_i - \bar Y)^2) \right] \nabla \theta \nabla \theta \nabla \theta \nabla \theta \nabla \theta \nabla \theta \nabla \theta \nabla \theta
\]
(23)
Note that the second term in (23) depends on \( m \theta \) only through its gradient and decreases to zero at the much faster asymptotic rate of \( 1/n^2 \) as compared to the rate \( 1/n \) in (21).

A. A Bootstrap Estimator for Bias Gradient Norm

A natural “method-of-moments” estimator for \( \delta^2 = \| \nabla \bar b \|_2^2 \) is the norm squared of the unbiased estimator \( \delta^2 = \| \nabla \bar b \|_2^2 \). (20). It can easily be shown that this estimator is biased with bias equal to \( E_\theta \left[ \| \nabla \bar b \|_2^2 - \| \nabla \bar b \|_2^2 \right] = \text{trace} \{ S(\nabla \bar b) \} \) 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 \( \delta^2 = \delta^2(y_1, \ldots, y_L) = \| \nabla \bar b \|_2^2 \), the bootstrap estimate of \( \delta^2 \) is defined as the expectation of \( \delta^2 = \delta^2(Y_1, \ldots, Y_L) \) with respect to the resampling distribution \( [37] \)
\[
E_\theta \left[ \delta^2 \right] = \sum_{i} \delta^2(Y^*_i, \ldots, Y^*_L) \left( c_1 \ldots c_L \right)^{-L}. \quad (24)
\]
In (24) \( c_i \) is the number of times the value \( y_i \) appears in the set \( \{Y^*_i\} \) and \( \sum_i \) denotes a summation over all non-negative integers \( c_1, \ldots, c_L \) satisfying \( \sum c_i = L \). The bootstrap estimate of the bias of the estimator \( \delta^2 \) is defined as \( E_\theta \left[ \delta^2 \right] - \delta^2 \) which leads to the bias corrected estimator \( \tilde \delta^2 \)
\[
\tilde \delta^2 = 2\delta^2 - E_\theta \left( \tilde \delta^2 \right). \quad (25)
\]

Due to the simple quadratic dependence of \( \delta^2 \) on the single sample quantities \( \hat \bar Y \nabla \ln f_Y(y_i; \theta) \), \( i = 1, \ldots, L \), the expectation (24) can be expressed in analytical form (see Appendix B) leading to the biased corrected estimate
\[
\delta^2 = \| \nabla \bar b \|_2^2 - \frac{1}{L^2} \sum_{i=1}^L \| \nabla \bar b(y_i) - \nabla \bar b \|_2^2, \quad (25)
\]
where \( \nabla \bar b(y_i) \) is the estimate of (20) based on a single sample \( y_i \) \( (L = 1) \) with \( \zeta_i = 0 \). The bias of \( \delta^2 \) is equal to
\[
E_\theta \left[ \delta^2 \right] - \| \nabla \bar b \|_2^2 = \frac{1}{L} E_\theta \left[ \| \nabla \bar b - \nabla \bar b \|_2^2 \right],
\]
which, relative to the estimator \( \| \nabla \bar b \|_2^2 \), decays to zero at the much faster rate of \( 1/L^2 \). However, if \( L \) is insufficiently large the bootstrap estimator \( \delta^2 \) 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 \( \mathbf{Y} = \mathbf{Y} \in \mathbb{R}^m \) which obeys the linear Gaussian model:

\[
\mathbf{Y} = \mathbf{A} \mathbf{\theta} + \mathbf{\epsilon},
\]

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

\[
\mathbf{F}_Y = \mathbf{A}^T \Sigma^{-1} \mathbf{A}.
\]

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

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 \( t_\delta = h^T \mathbf{\theta} \) is \( \hat{t} = \overline{h^T \mathbf{\theta}} \) where \( \overline{\mathbf{\theta}} \) minimizes the following objective function over \( \mathbf{\theta} \):

\[
[\mathbf{Y} - \mathbf{A} \overline{\mathbf{\theta}}^T \Sigma^{-1} \mathbf{Y}] + \beta \overline{\mathbf{\theta}}^T \mathbf{P} \overline{\mathbf{\theta}}
\]

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 conditioning 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 Tikhonov 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

\[
\overline{\mathbf{\theta}} = [\mathbf{F}_Y + \beta \mathbf{P}]^{-1} \mathbf{A}^T \Sigma^{-1} \mathbf{Y},
\]

yielding the QPML estimator \( \hat{t} = \overline{h^T \mathbf{\theta}} \).

The estimator bias is

\[
b_\perp = h^T [\mathbf{F}_Y + \beta \mathbf{P}]^{-1} \mathbf{F}_Y - \mathbf{I} \overline{\mathbf{\theta}},
\]

and its bias gradient is

\[
\nabla b_\perp = [\mathbf{F}_Y [\mathbf{F}_Y + \beta \mathbf{P}]^{-1} - \mathbf{I}] h = -\beta \mathbf{P} [\beta \mathbf{P} + \mathbf{F}_Y]^{-1} h.
\]

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

\[
\sigma^2 = \frac{1}{\beta^2} h^T [\mathbf{P} + \frac{1}{\beta} \mathbf{F}_Y]^{-1} \mathbf{F}_Y [\mathbf{P} + \frac{1}{\beta} \mathbf{F}_Y]^{-1} \mathbf{F}_Y h.
\]

Consider the special case of estimation of a single component \( \theta_k \) of \( \mathbf{\theta} \) for which \( h = \overline{e_k} = [0, \ldots, 0, 1, 0, \ldots, 0]^T \). 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 \( [\beta \mathbf{P} + \mathbf{F}_Y]^{-1} \mathbf{F}_Y \overline{e_k} \) of the PLS estimator to a point source \( \mathbf{\theta} = \overline{e_k} \), i.e., the point spread function of the estimator, and the ideal point response \( \overline{e_k} \). 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 \( \overline{d}_{\text{min}} \) and the bound \( B(\mathbf{\theta}, \delta) \), respectively. Identifying \( \nabla t_\delta = h \), \( \lambda = 1/\beta \), it is clear that when \( \mathbf{P} \) is chosen as \( \mathbf{C}^{-1} \), the PLS estimator achieves the bound \( B(\mathbf{\theta}, \delta) \) and has optimal bias gradient \( \overline{d}_{\text{min}} \). Thus for linear functions \( t_\delta \) the uniform bound is achievable and the region above 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 \( \sup_{\mathbf{\theta} \in \mathcal{C}} |b_{\perp} - b_\parallel| \leq \delta^2 \), 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 (\( h = \overline{e_{\text{pixel}}} \)) and a well conditioned full rank discrete Gaussian system matrix. Specifically we generated a \( 128 \times 128 \) matrix \( \mathbf{A} \) with elements \( a_{ij} = \frac{1}{\sqrt{2\pi w}} e^{-\frac{w^2}{2}} \) and \( w = 0.5 \). The condition number
of this \( A \) is 1.7. The matrix \( C \) in the norm \( \| \nabla b_d \|_C \) was selected as the inverse of the second order (Laplacian) differencing matrix

\[
C^{-1} = \begin{bmatrix}
2 & -1 & & \\
-1 & 2 & -1 & 0 \\
& & \ddots & \ddots \\
0 & -1 & 2 & -1 \\
\end{bmatrix}.
\] (34)

With this norm the restriction \( \| \nabla b_d \|_C \leq \delta \) corresponds to a constraint on maximal bias variation \( \max_{d \in \mathcal{C}} \| \Delta b_d \|_C \) over a roughness constrained neighborhood \( \mathcal{C} \). Also plotted in Fig. 2 are the performance curves \( (\| \nabla b_d \|_C, \sigma_d) \) for two PLS pixel intensity estimators (30), one using the smoothing matrix \( P = C^{-1} \), called the smooth QPML estimator, and another using the diagonal “energy penalty” \( P = I \), called the unsmoothed QPML estimator. These curves were traced out in the bias variance tradeoff plane by varying \( \beta \) in the parametric descriptions of estimator variance (33) and estimator bias gradient (32).

A.2 Singular Fisher Matrix

When \( A \) has rank less than \( n \), \( F_Y \) is singular and unbiased estimators may not exist for all linear functions \( t_{\mathbf{u}} \) of \( \hat{\theta} \) [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 \( \delta \) 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 \( \hat{\theta} \) based on only 32 observations \( \mathbf{Y} \). 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 \( t_{\mathbf{u}} = \mathbf{A}_r \mathbf{Y} \) are plotted in Figure 3 along with the performance curves associated with the smooth QPML (\( P = C^{-1} \)) and unsmoothed QPML (\( P = I \)) estimators. Note that neither of the estimators achieve the uniform bound for any value of the parameter \( \beta \). 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 \( t_{\mathbf{u}} = \mathbf{A}_r \mathbf{Y} \) where the elements of \( \mathbf{A}_r \) are plotted in Fig. 4. Observe that the smooth QPML estimator of contrast comes much closer to the uniform bound than does the smooth QPML estimator of pixel intensity shown in Fig. 3.

Under certain conditions the uniform CR bound is exactly achievable even for singular \( F_Y \), although generally not by a QPML estimator of the form (30) and generally not for all \( \delta \). Consider the estimator

\[
\hat{\theta} = \mathbf{A}_r^T \left[ I + \beta \mathbf{F}_Y^+ \mathbf{P}^{-1} \right]^{-1} \mathbf{F}_Y^+ \mathbf{P}^{-1} \mathbf{A}_r^T \mathbf{Y} \mathbf{Y}^{-1} \mathbf{Y} \] (35)

This estimator reduces to the previous estimator (30) for the case of non-singular \( F_Y \). The estimator bias gradient is

\[
\nabla b_{\hat{\theta}} = \left( \mathbf{F}_Y \mathbf{F}_Y^+ \left[ I + \beta \mathbf{P} \mathbf{F}_Y^+ \right]^{-1} - \mathbf{I} \right) \nabla b_{\mathbf{Y}}
\]

\[
= - \left[ \frac{1}{\beta} \mathbf{F}_Y \mathbf{F}_Y^+ \right]^{-1} \mathbf{F}_Y^+ \mathbf{h}
\]

\[
= - \left[ I - \mathbf{F}_Y \mathbf{F}_Y^+ \right] \left[ I + \beta \mathbf{P} \mathbf{F}_Y^+ \right]^{-1} \mathbf{h} \] (36)
s, the vector \( \hat{h} \) lies in the nullspace of
\[
[I + \beta P F_Y^+ F_Y^+ I]^{-1} \hat{h}
\]
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 \( t_0 = \frac{h^T \theta}{\| \nabla b_\theta \|_{\mu-1}^2} \) satisfying
\[
\| \nabla b_\theta \|_{\mu-1}^2 \leq \delta^2 = g(1/\beta)
\]
where \( g(\cdot) \) is the function given in (11).

Recognizing the matrix \( I - F_Y^+ F_Y = I - F_Y F_Y^+ \) as the operator which projects onto the null space of \( F_Y \), an equivalent condition to (2) is that \( [I + \beta P F_Y^+]^{-1} \hat{h} \) lie in the range space of \( F_Y \). 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 \( \hat{h} \) must lie in the range space of the Fisher information \( F_Y \). In order for the uniform bound \( B(\theta, \delta) \) to be achievable for all values of \( \delta \), condition (2) must hold for all \( \beta > 0 \). This is a much stronger condition except when the eigenspace of \( [I + \beta P F_Y^+]^{-1} \) is independent of \( \beta \), which occurs for example when \( P = I \). This suggests that when estimation of any fixed \( t_0 \) is of interest and the Fisher information is singular, the uniform bound will rarely be achievable everywhere in the delta-sigma plane.

and the estimator variance is
\[
\sigma^2_\hat{t} = h^T P^{-1} \left[ P^{-1} + \beta F_Y^+ F_Y \right]^{-1} F_Y^T P^{-1} \left[ P^{-1} + \beta F_Y^+ F_Y \right]^{-1} \hat{h} \tag{37}
\]
where in (37) we have used the property \( F_Y^+ F_Y F_Y^+ = F_Y^+ \) [43]. Noting that here \( \nabla t_0 = \hat{h} \), we conclude that the estimator variance is equal to the lower bound expression \( B(\theta, \delta) \) given in (9) when \( P = C^{-1} \) and \( \beta = 1/\lambda \). Furthermore, under these conditions the bias gradient (36) differs from the optimal bias gradient \( \nabla \epsilon_{\text{opt}} \) given in (10), only by the presence of the second additive term on the right hand side of (36). Thus the estimator (35) with \( P = C^{-1} \) 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 \( F_Y \).

**Theorem 2:** Let \( B = P^+ F_Y^+ P^+ \) where \( F_Y \) is the possibly singular Fisher information matrix. If

1. \( \delta^2 < \frac{1}{\lambda} \) \( P^+ P_B P^+ \) and
2. the vector \( \hat{h} \) lies in the nullspace of
\[
[I + \beta P F_Y^+ F_Y^+ I]^{-1} \hat{h}
\]
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 \( t_0 = \frac{h^T \theta}{\| \nabla b_\theta \|_{\mu-1}^2} \) satisfying
\[
\| \nabla b_\theta \|_{\mu-1}^2 \leq \delta^2 = g(1/\beta)
\]
where \( g(\cdot) \) is the function given in (11).

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

In some applications the observations 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 $Y = [Y_1, \ldots, Y_m]^T$ is a vector of integrals or counts with a vector of means $\mu = [\mu_1, \ldots, \mu_m]^T$. This vector of counts obeys independent Poisson statistics with log-likelihood

$$\ln f_Y(y; \theta) = \sum_{j=1}^m (y_j \ln(\mu_j(\theta)) - \mu_j(\theta)) + c.$$  \hspace{1cm} (38)

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

$$\mu(\theta) = A \theta + \eta$$  \hspace{1cm} (39)

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

The Fisher information has the form [45]

$$F_Y(\theta) = \sum_{j=1}^m \frac{1}{\mu_j(\theta)} A_{j*} A_{j*}^T,$$  \hspace{1cm} (40)

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

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 $\hat{t} = \hat{t}_q$ where $\theta$ is the vector $\theta$ which maximizes the penalized likelihood function

$$J(\theta) = \ln f_Y(y; \theta) - \frac{\beta}{2} \theta^T P \theta.$$  \hspace{1cm} (41)

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

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 $A E_0 \hat{\theta}$ of the mean QPML image and the projection $A \hat{\theta}$ 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 $t_0 = \hat{t}^T \theta$, we obtain the following expressions for the asymptotic variance of $\hat{t}$

$$\sigma_n^2 = \hat{t}^T \left[ F_Y(\theta) + \beta P \right]^{-1} F_Y(\theta) \left[ F_Y(\theta) + \beta P \right]^{-1} \hat{t},$$  \hspace{1cm} (42)

and the asymptotic bias gradient

$$\nabla b_n = -P \left[ P + \frac{1}{\beta} F_Y(\theta) \right]^{-1} \hat{t} + O\left(\frac{1}{\beta}\right).$$  \hspace{1cm} (43)

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

When we identify $P = C^{-1}$ and $\beta = 1/\lambda$ we see that the estimator variance is identical to the optimal variance (12), and that for linear $t_0$ 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 $t_0$ 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 $\beta$ if $P = C^{-1}$.

Fig. 8. Emission source $\theta$ 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, $A$ was a $128 \times 128$ tri-diagonal blurring matrix with kernel $(0.23, 0.54, 0.23)$, for which the condition number is $12.5$. The source intensity $\theta$ is shown in Fig. 8. The function of interest was chosen as $f = \theta_{65}$, the intensity of pixel 65 in Fig. 8. We generated $L = 1000$ realizations of the measurements each having a mean total of $\sum_{j=1}^{M} \mu_j(\theta) = 2100$ counts, including a 5% background representing random coincidences [20].

We computed three types of estimates of $\theta$: the quadratically penalized maximum likelihood estimator using the “energy penalty” ($P = I$), 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: $(A^T A + \beta I)^{-1} A^T (Y - r)$, which is linear so it can be computed noniteratively. Here $\beta^* = \beta \sum_{j} A_{jk} / (\sum_{j} A_{jk}^2)$ 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 $A$, and computed approximate pseudoinverses of $A$ by excluding the 10 smallest eigenvalues. The form of the “deconvolve/shrink” estimator is:

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

where $\beta$ ranges from 0 to 1.

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

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 $\beta$ 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 $\beta$ decreases so that it extends the range of reliable estimation of the ordinary BGL estimator.

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 $\beta$). 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 \( \hat{d} \) 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 \((\delta, \sigma_\theta)\) where \(\delta\) is the estimator bias gradient norm and \(\sigma_\theta^2\) is the estimator variance. For two inverse problems, one linear and one non-linear, we have established that the bound is achievable under certain conditions.

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, this 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 \( \nabla b_\perp : \|\nabla b_\perp\|_C \leq \delta \) of bias gradient vectors

\[
\text{var}_\perp(d) \geq \frac{\|\nabla b_\perp + \nabla b_\parallel d^T F_\perp |\nabla b_\perp + \nabla b_\parallel d\|}{\text{min}_{d : \|d\|_C \leq \delta} Q(d)},
\]

where

\[
Q(d) = \nabla t_\perp + d^T F_\parallel |\nabla t_\perp + d|,
\]

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

\[
\min_{d : \|d\|_C \leq \delta} \left[ C^\frac{1}{2} \nabla t_\perp + d^T B \left[C^\frac{1}{2} \nabla t_\perp + d\right] \right],
\]

(44)

where \( B = C^{-\frac{1}{2}} F_\parallel C^{-\frac{1}{2}} \).

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

\[
\hat{d} = -P_B C^{\frac{1}{2}} \nabla t_\perp + \phi,
\]

where \( \phi \) is an arbitrary vector in the null space of \( B \). But for \( \hat{d} \) to be a feasible solution it must satisfy \( \|\hat{d}\|_2 \leq \delta \) so that, by orthogonality of \( P_B C^{\frac{1}{2}} \nabla t_\perp \) and \( \phi \)

\[
\delta^2 \geq \|\hat{d}\|^2 = \|P_B C^{\frac{1}{2}} \nabla t_\perp\|^2 + \|\phi\|^2 \geq \|P_B C^{\frac{1}{2}} \nabla t_\perp\|^2.
\]

We conclude that \( \min_{d : \|d\|_C \leq \delta} Q(d) = 0 \) if \( \|P_B C^{\frac{1}{2}} \nabla t_\perp\|^2 = \nabla^T t_\perp C^{\frac{1}{2}} P_B C^{\frac{1}{2}} \nabla t_\perp \leq \delta^2 \). If \( \nabla^T t_\perp C^{\frac{1}{2}} P_B C^{\frac{1}{2}} \nabla t_\perp = 0 \) then we have nothing left to prove. Otherwise, assume \( \delta \) lies in the range: \( 0 \leq \delta^2 < \nabla^T t_\perp C^{\frac{1}{2}} P_B C^{\frac{1}{2}} \nabla t_\perp \). In this case the minimizing \( \hat{d} \) lies on the boundary and satisfies the equality constraint \( \|\hat{d}\|_2 = \delta \).

We thus need solve the unconstrained minimization of the Lagrangian:

\[
\min_{\hat{d}} \left[ \frac{1}{2} \left( C^{\frac{1}{2}} \nabla t_\perp + \hat{d}^T B \left[C^{\frac{1}{2}} \nabla t_\perp + \hat{d}\right] \right) + \lambda (\hat{d}^T \hat{d} - \delta^2) \right],
\]

(45)

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

\[
\min_{\hat{d}} \left[ \left( \hat{d} + (\lambda I + B)^{-1} B C^{\frac{1}{2}} \nabla t_\perp \right)^T (\lambda I + B) \left( \hat{d} + (\lambda I + B)^{-1} B C^{\frac{1}{2}} \nabla t_\perp \right) \right]
\]

\[
+ \nabla^T t_\perp C^{\frac{1}{2}} (I - (\lambda I + B)^{-1} B) \left[ C^{\frac{1}{2}} B (\lambda I + B)^{-1} B C^{\frac{1}{2}} \nabla t_\perp \right] \nabla t_\perp - \lambda \delta^2.
\]

It follows immediately from (46) that

\[
\hat{d} = \hat{d}_{\min} = -(\lambda I + B)^{-1} B C^{\frac{1}{2}} \nabla t_\perp
\]

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

\[
\min_{d : \|d\|_C \leq \delta} Q(d)
\]

\[
= \left[ \nabla t_\perp + \hat{d}_{\min} \right]^T F_\parallel \left[ \nabla t_\perp + \hat{d}_{\min} \right]
\]

\[
= \nabla^T t_\perp C^{\frac{1}{2}} \left[ I - (\lambda I + B)^{-1} B \right] B \left[ I - (\lambda I + B)^{-1} B \right] \left[ I - (\lambda I + B)^{-1} B \right] C^{\frac{1}{2}} \nabla t_\perp
\]

\[
= \lambda^2 \nabla^T t_\perp C^{\frac{1}{2}} \left[ I - (\lambda I + B)^{-1} B \right] \left[ I - (\lambda I + B)^{-1} B \right] C^{\frac{1}{2}} \nabla t_\perp
\]

(46)
which, after simple matrix manipulations, gives the expression (9).

The Lagrange multiplier $\lambda$ is determined by the equality constraint

$$\delta^2 = \min \sum \nabla^2 t_\lambda = \nabla^T t_\lambda \left[ C \frac{1}{2} B \lambda + B \right]^{-1} 2BC \nabla t_\lambda = g(\lambda). \quad (47)$$

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

Let the non-negative definite symmetric matrix $B$ have eigendecomposition $B = \sum \beta_i \xi_i \xi_i^T$ where $\beta_i$ are positive eigenvalues, $\xi_i$ are eigenvectors, and $r > 0$ is the rank of $B$. With these definitions the function $g(\lambda)$ (47) has the equivalent form

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

Since by assumption $\nabla^T t_\lambda \frac{1}{2} C \frac{1}{2} \frac{1}{2} P_B \frac{1}{2} \frac{1}{2} \nabla t_\lambda > 0$, $\frac{1}{2} \frac{1}{2} C \frac{1}{2} \frac{1}{2} \nabla t_\lambda$ does not lie in the nullspace of $B$ and thus $|\nabla^T t_\lambda \frac{1}{2} C \frac{1}{2} \xi_i^2|^2 > 0$ 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 $\lim_{\lambda \to 0} g(\lambda) = 0$ and $\lim_{\lambda \to \infty} g(\lambda) = \nabla^T t_\lambda \frac{1}{2} C \frac{1}{2} \frac{1}{2} P_B \frac{1}{2} \frac{1}{2} \nabla t_\lambda$. Hence there exists a unique strictly positive $\lambda$ such that $g(\lambda) = \delta^2$ for any value $\delta^2 \in [0, \nabla^T t_\lambda \frac{1}{2} P_B \frac{1}{2} \frac{1}{2} \nabla t_\lambda]$. □

**Appendix B: Bootstrap Derivation**

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

$$\hat{\nabla} b_\lambda = \frac{1}{L} \sum_{i=1}^L \hat{z}(Y_i),$$

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

$$\hat{z}(Y_i) = \frac{1}{L} \nabla^T \ln f(Y_i; \theta) - \nabla t_\lambda.$$

The (biased) sample mean estimator of the norm squared $\delta^2 = |\nabla b_\lambda|^2$ is

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

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

$$\bar{\delta}^2 = \frac{1}{L} \sum_{i=1}^L c_i \hat{z}(y_i) ||\hat{z}(y_i)||_C^2$$

$$= \frac{1}{L} \sum_{i=1}^L \sum_{j=1}^L c_i c_j < \hat{z}(y_i), \hat{z}(y_j) >_C, \quad (49)$$

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

$$\bar{\delta}^2 = \frac{1}{L} \frac{1}{L} \nabla^T H \mathbf{e}$$

$$= \frac{1}{L} \text{trace}\{H \mathbf{e} \mathbf{e}^T\}. \quad (50)$$

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

$$E, [\hat{\delta}^2] = \sum \bar{\delta}^2 \left( \frac{L}{c_1 + \ldots + c_L} \right)^{-L}$$

$$= \frac{1}{L} \text{trace} \{H E, [\mathbf{e} \mathbf{e}^T]\}. \quad (52)$$

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

$$E, [\mathbf{e}^T] = \mathbf{I} + \mathbf{e} \mathbf{T} \mathbf{1}^T$$

$$\frac{1}{L} \text{trace} \{H E, [\mathbf{e} \mathbf{e}^T]\}$$

$$= \frac{1}{L} \left[ \text{trace} \{H\} + \frac{L - 1}{L} \frac{1}{L} \mathbf{T} \mathbf{1}^T \right]$$

$$= \frac{1}{L} \left[ \sum_{i=1}^L ||\hat{z}(Y_i)||_C^2 - L ||\mathbf{e}||_C^2 + L^2 ||\mathbf{e}||_C^2 \right]$$

$$= \frac{1}{L} \sum_{i=1}^L ||\hat{z}(Y_i) - \mathbf{e}||_C^2 + \delta^2,$$

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

$$\bar{\delta}^2 = 2 \delta^2 - E, [\hat{\delta}^2]$$

$$= \delta^2 - \frac{1}{L} \sum_{i=1}^L ||\hat{z}(Y_i) - \mathbf{e}||_C^2,$$

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

\[
\max_{\underline{u} \in C} ||b_{\underline{u}} - \underline{b}_{\hat{\theta}}||^2 = \max_{\underline{u} \in C} ||m_{\underline{u}} - t_{\underline{u}} - m_{\theta} + t_{\theta}||^2 \\
\geq \max_{\underline{u} \in \mathcal{N} \cap \mathcal{C}} ||m_{\underline{u}} - m_{\theta} + t_{\theta} - t_{\underline{u}}||^2 = \max_{\underline{u} \in \mathcal{N} \cap \mathcal{C}} ||t_{\underline{u}} - t_{\theta}||^2 = \max_{\underline{u} \in \mathcal{N} \cap \mathcal{C}} ||h^T[I - \mathcal{P}_{\theta}]\Delta \underline{u}||^2 = \max_{\underline{u} \in C} ||h^T[I - \mathcal{P}_{\theta}]\Delta \underline{u}||^2, \quad (53)
\]

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

\[
||\nabla \underline{b}_{\underline{u}}||_C \geq \sqrt{h^T[I - \mathcal{P}_{\theta}]\Delta \underline{u}^T[I - \mathcal{P}_{\theta}]\Delta \underline{u}^T}[\mathcal{C}[I - \mathcal{P}_{\theta}]h + \epsilon \quad (55)
\]

For the case that the bias \( b_{\underline{u}} \) is linear over \( \underline{u} \in C \) in (55) is equal to zero and we have an exact bound

\[
||\nabla \underline{b}_{\underline{u}}||_C \geq \sqrt{h^T[I - \mathcal{P}_{\theta}]\Delta \underline{u}^T[I - \mathcal{P}_{\theta}][\Delta \underline{u}^T]h}. \quad (55)
\]

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

\[
||\nabla \underline{b}_{\underline{u}}||_C \geq \sqrt{\frac{h^T[I - \mathcal{P}_{\theta}]\Delta \underline{u}^T[I - \mathcal{P}_{\theta}]h}{\Delta \underline{u}^T[I - \mathcal{P}_{\theta}]h} + \epsilon. \quad (55)
\]

Asymptotic Bias:

\[
b_{\underline{u}} = t_{\underline{u}} - t_{\theta}, \quad (58)
\]

Asymptotic Bias Gradient:

\[
\nabla b_{\underline{u}} = F_Y(\theta)[\beta P + F_Y(\theta)]^{-1} \nabla t_{\underline{u}} - \nabla t_{\theta} - O\left(\frac{1}{\beta}\right)g \]

where

\[
O\left(\frac{1}{\beta}\right) = \frac{1}{\beta} \sum_{j=1}^{m} \gamma_{j}(\theta) A_j^T B^{-1} \nabla t_{\underline{u}} \quad (60)
\]

\[
\gamma_j(\theta) = \frac{A_j \theta^T P B^{-1} A_j}{\mu_j^2(\theta)} \quad (61)
\]

and \( \nabla t_{\underline{u}} \) denotes the evaluation of the gradient of \( t_{\underline{u}} \) at the point \( \theta = \hat{\theta} \).

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

\[
\text{var}_{\underline{u}}(\underline{z}) = \nabla^T \nabla t_{\underline{u}} \left[\nabla^2 a(\underline{z}, \theta)^{-1}\right] \Sigma \nabla t_{\underline{u}} \left[\nabla^2 a(\underline{z}, \theta)^{-1}\right]^T \nabla t_{\underline{u}} \quad (62)
\]

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

\[
\text{cov}_{\underline{u}}[\nabla J(\underline{z})] = F(\underline{z}, \theta) = \sum_{j=1}^{\mu_j(\underline{z})} A_j^T \frac{1}{\mu_j(\underline{z})} \left(\mu_j(\underline{z})\right)^2 = F_Y(\theta) + o(\mu_j(\underline{z}) \theta), \quad (63)
\]

where \( A_j \) is the \( j \)-th row of \( A \), \( F_Y(\theta) \) is the Fisher matrix (28), and \( o(\mu_j(\underline{z}) \theta) \) is a remainder term which is close to zero when the difference between the projections \( \mu_j(\underline{z}) \) and \( \mu_j(\theta) \) is small. To obtain the expression (63) with remainder term we used the series development

\[
\frac{\mu_j(\theta)}{\mu_j(\underline{z})} = 1 - \frac{1}{\mu_j(\underline{z})} (\underline{z} - \theta)^T A_j + o((\underline{z} - \theta)^T A_j). \]

Among other things these conditions involve showing that the gradient function \( \nabla J(\theta) \) converges a.s. to zero as the observation time increases.

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

Define the vector

\[
\underline{z} = \left[ F_Y(\theta) + \beta P \right]^{-1} F_Y(\theta) \theta = \left[I - \left[P + \frac{1}{\beta} F_Y(\theta)\right]^{-1} P \right] \theta \quad (56)
\]

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

Asymptotic Variance:

\[
\sigma^2_{\underline{u}} = \nabla^T t_{\underline{u}} [F_Y(\theta) + \beta P]^{-1} F_Y(\theta) [F_Y(\theta) + \beta P]^{-1} \nabla t_{\underline{u}} \quad (57)
\]
The ambiguity function is

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

Differentiation of the ambiguity function with respect to \( u \) yields

\[ \nabla^\text{\text{20}} a(u, \theta) = -F_Y(\theta)(u - \theta) - \beta P u + O \left( \mu(u - \theta) \right), \]

and similarly

\[ \nabla^\text{\text{10}} a(u, \theta) = -F_Y(\theta)(u - \theta) - \beta P u \]

Neglecting the \( O(\mu(u - \theta)) \) 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 \( u = z \) of the equation \nabla^\text{\text{10}} a(u, \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

\[ \nabla b_u = \nabla z_u^T \nabla t_u - \nabla t_u, \]

where \( \nabla z_u^T \) is an \( n \times n \) matrix of derivatives of \( z = z_u \). From (56) the \( k \)-th row of this matrix is

\[ \frac{d}{d\theta_k} z_u^T = z_u^T \left[ I - P \left[ P + \frac{1}{\beta} F_Y(\theta) \right]^{-1} \right] \]

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

\[ -\frac{d}{d\theta} P \frac{d}{d\theta} \left[ P + \frac{1}{\beta} F_Y(\theta) \right]^{-1} = \frac{1}{\beta} P \left( P + \frac{1}{\beta} F_Y(\theta) \right) B^{-1} \]

\[ = \frac{1}{\beta} \sum_{j=1}^{m} A_{ij} P B^{-1} A_j, \]

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

\[ \nabla b_u = \left[ I - P \left[ P + \frac{1}{\beta} F_Y(\theta) \right]^{-1} \right] + O \left( \frac{1}{\beta} \right), \]

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

REFERENCES

Denver, CO. His current research interests include image reconstruction, speech compression, and code optimization. He is a member of IEEE.