Bias-Variance Trade-offs For Parametric Estimation Problems Using the Uniform CR Bound

Mohammad Usman*, Alfred O. Hero*, and Jeffrey A. Fessler**
* Dept. of Electrical Engineering and Computer Science, and ** Dept. of Internal Medicine,
The University of Michigan, Ann Arbor, MI 48109-2122

Abstract

We introduce a plane, which we call the bias-variance 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 bias-variance tradeoff curve is specified which defines an "unachievable region" of the bias-variance plane for a specified statistical model. In order to place an estimator on this plane for comparison to the bias-variance 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 two inverse problems: penalized linear least squares estimation under Gaussian statistics and smoothness constrained image reconstruction under Poisson statistics.

Key Words: Non-linear estimation, bias-variance plane, unachievable regions, inverse problems, image reconstruction.

EDICS Number: SP 3.8.1

1 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 of an observed random variable Y. The parameter space Θ is assumed to be an open subset of n-dimensional Euclidean space $I\!R^n$. Let $\hat{\theta}_1 = \hat{\theta}_1(\mathrm{Y})$ be an estimator of the scalar θ_1 and let this estimator have bias function $b = b_{\underline{\theta}} = E_{\underline{\theta}}[\hat{\theta}_1] - \theta_1$ and variance function $\sigma^2_{\underline{\theta}} = E_{\underline{\theta}}[(\hat{\theta}_1 - \theta_1)^2]$. Bias is due to 'mismatch' between the average value of the estimator and the true parameter while variance is due to fluctuations in the estimator due to statistical sampling.

Any estimator design is subject to the well known tradeoff between bias and variance. In image restoration, regularization is implemented to reduce noise amplification (variance) at the expense of reduced spatial resolution (bias) [1]. In multiple regression with multicollinearity, biased shrinkage estimators [2] and biased ridge estimators [3] are used to reduce variance of the ordinary least squares estimator. In emission computed tomography, curves which relate bias to variance, such as bias versus square-root-of-variance curves [4] and contrast vs. SNR curves [5], are used to investigate the resolution and sensitivity characteristics of a specific reconstruction algorithm.

However, for fixed $\underline{\theta}$ the plane parameterized by the bias and variance $b_{\underline{\theta}}$ and $\sigma^2_{\underline{\theta}}$ is not usefull for studying fundamental tradeoffs since an estimator can always be found which makes both the bias and variance zero at a point $\underline{\theta}$. Furthermore, the bias value $b_{\underline{\theta}}$ unfairly penalizes estimators that 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 variance $\sigma^2_{\underline{\theta}}$, which we call the bias-variance or $\delta\sigma$ plane. The norm of the bias gradient is directly related to the maximal variation of the bias function over the $\underline{\theta}$-region induced by the norm and is unaffected by constant estimator bias components. By appropriate choice of norm, the bias gradient length can be directly related to the overall bias variation over any prior region of parameter values. As discussed here and in [6], the choice of norm should depend on the specific units and dimensions used to express each of the parameters. In inverse problems, such as image restoration and image reconstruction, we show that the norm should be be selected to correspond to *a priori* smoothness constraints on the object.

This paper provides a means for specifying unachievable regions in the $\delta\sigma$ plane via a uniform extension of the Cramer-Rao (CR) lower bound to biased estimators presented in [6]. 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 constraint, or they can be used to gauge the closeness to optimality of a biased estimator in terms of its nearness to the unachievable region. These results can also be used to validate the correctness 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 estimator variance and bias gradient norm are required. For many estimators computation of these quantities is analytically intractable. We present a methodology for experimentally determining the variance and bias gradient norm which only requires performing a single sequence of L experiments and uses the measured gradient of the log-likelihood function $\nabla \ln f_{\mathrm{Y}}(\mathrm{y}; \underline{\theta})$ and a bootstrap-type estimator to

[1]This work was supported in part by National Science Foundation under grant BCS-9024370, a Government of Pakistan Postgraduate Fellowship, NIH grant CA-60711, and DOE grant DE-FG02-87ER60561

estimate the bias gradient norm.

We will illustrate these methods for a linear Gaussian and a non-linear Poisson inverse problems utilizing varying amounts of smoothing. Even for the linear inverse problem there may not exist unbiased estimators when the convolution kernel matrix is ill-conditioned or rank deficient. For each case we compare the performance of quadratically penalized maximum likelihood algorithms to the fundamental bias-variance tradeoff curve. The bias gradient $\nabla b_{\underline{\theta}}$ of these algorithms is directly related to the point spread function of the estimator. For the linear Gaussian case we show that the quadratically penalized likelihood estimator achieves the optimal bias-variance 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. Since the bias-gradient is linear, the bias-gradient norm constraint is equivalent to a constraint on maximal bias variation and we have a very strong optimality property: the penalized maximum likelihood estimator minizes variance over all estimators whose maximal bias variation is bounded over a $\underline{\theta}$-region. Here the shape and size of the $\underline{\theta}$-region is determined by the roughness penalty matrix and smoothing parameter. For the non-linear Poisson case we show that the penalized maximum likelihood algorithm of [7] locally achieves the optimal bias-variance tradeoff curve for suitably chosen penalty matrix.

2 The Bias-Variance (ffiœ) Plane

The mean-square-error (MSE) is a widely used measure of precision for an estimator $\hat{\theta}_1$ and is simply related to the estimator bias $b_{\underline{\theta}}$ and the estimator variance $\sigma^2_{\underline{\theta}}$ by: $\mathrm{MSE}_{\underline{\theta}} = \sigma^2_{\underline{\theta}} + b^2_{\underline{\theta}}$. 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. From $b_{\underline{\theta}}$ and $\sigma_{\underline{\theta}}$ one can derive other important measures such as signal-to-noise-ratio $SNR = |\theta_1 + b_{\underline{\theta}}|^2/\sigma^2_{\underline{\theta}}$, coefficient of variation $1/SNR$, and generalized $\mathrm{MSE} = \alpha g_1(b_{\underline{\theta}}) + (1-\alpha) g_2(\sigma_{\underline{\theta}})$, where $\alpha \in [0,1]$ and g_1, g_2 are non-negative functions. The generalized MSE has been used in response surface design [8] and minimum bias and variance estimation for non-linear regression models [9, 10]. 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 $\hat{\theta}_1$-based hypothesis tests and confidence intervals.

An estimator $\hat{\theta}_1$ whose bias function $b_{\underline{\theta}} : \Theta \rightarrow I\!R$ is a 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 $f_{\mathbf{Y}}(\mathrm{y}; \underline{\theta})$ 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 $E[\hat{\theta}_1^2] < \infty$ [11, Lemma 7.2]. In this case the bias gradient $\nabla b_{\underline{\theta}} : \Theta \rightarrow I\!R^n$ uniquely specifies the bias $b_{\underline{\theta}}$ up to an undetermined 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 for all $\underline{\theta} \in \Theta$ – such a point is guaranteed to exist when Θ is convex or star shaped. Thus the gradient function $\nabla b_{\underline{\theta}} = \left[\frac{\partial b_{\underline{\theta}}}{\partial \theta_1}, \ldots, \frac{\partial b_{\underline{\theta}}}{\partial \theta_1}\right]$ characterizes the unremovable bias component of $b_{\underline{\theta}}$.

Define the norm or length of the bias gradient vector

$$\delta_{\underline{\theta}} = \|\nabla b_{\underline{\theta}}\|_D$$

where the $\|\bullet\|_D$ is is defined in terms of a symmetric positive definite matrix D:

$$\|\underline{u}\|^2_D = \underline{u}^T \mathbf{D} \underline{u}. \qquad (1)$$

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

The bias gradient norm is a measure of the influence of each component parameter $\theta_1, ..., \theta_n$ on the mean $m_{\underline{\theta}}$ of the estimator $\hat{\theta}_1$. Ideally, to be close to unbiased, one would like $m_{\underline{\theta}} \approx \theta_1$, i.e. $m_{\underline{\theta}}$ should be sensitive to the variations in θ_1 but should be insensitive to the variations in the other parameters $\theta_2, ..., \theta_n$. Alternatively, since $b_{\underline{\theta}} = m_{\underline{\theta}} - \theta_1$, it is desirable that the components $\frac{\partial}{\partial \theta_k} b_{\underline{\theta}}$ be of small magnitude, $k = 2, ..., n$. The bias gradient therefore provides important information about parameter coupling to the estimator mean. However, the bias gradient norm is in general only indirectly related to the estimator bias, with the exception that $\nabla b_{\underline{\theta}} = 0$ implies $b_{\underline{\theta}} =$ constant. On the other hand, as will be shown below, a direct relation between bias gradient norm and bias variation over a region Θ..

Bias Gradient Norm and Maximal Bias

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

$$\sup_{\underline{\theta} \in \mathcal{D}} |b_{\underline{\theta}} - b_{\underline{\theta}^o}|^2 = \sup_{\underline{\theta} \in \mathcal{D}} |\nabla b_{\underline{\theta}^o} \Delta\underline{\theta} + \Delta\underline{\theta}^T \nabla^2 b_{\underline{\xi}} \Delta\underline{\theta}|^2 \qquad (2)$$

where $\Delta\underline{\theta} = \underline{\theta} - \underline{\theta}^o$ and $\underline{\xi}$ is a point along the line segment joining $\underline{\theta}^o$ and $\underline{\theta}$. Now, expanding the square on the right hand side of Eq. (2) and collecting terms:

$$\sup_{\underline{\theta} \in \mathcal{D}} |b_{\underline{\theta}} - b_{\underline{\theta}^o}|^2 = \sup_{\underline{\theta} \in \mathcal{D}} |\nabla b_{\underline{\theta}^o} \Delta\underline{\theta}|^2 + \epsilon, \qquad (3)$$

where $|\epsilon| = O\left(\lambda_M^D \cdot \alpha\right)$. Defining $\Delta\underline{\tilde{\theta}} = \mathbf{D}^{-\frac{1}{2}}\Delta\underline{\theta}$ and using the Cauchy-Schwarz inequality we obtain:

$$\begin{aligned}\sup_{\underline{\theta}\in\mathcal{D}} |\nabla b_{\underline{\theta}^o}\Delta\underline{\theta}|^2 &= \sup_{\|\Delta\underline{\theta}\|\le 1} |\nabla b_{\underline{\theta}^o}\mathbf{D}^{\frac{1}{2}}\Delta\underline{\tilde{\theta}}|^2 \\ &= \sup_{\Delta\underline{\theta}} \frac{|\nabla b_{\underline{\theta}^o}\mathbf{D}^{\frac{1}{2}}\Delta\underline{\tilde{\theta}}|^2}{\|\Delta\underline{\theta}\|_2} \\ &= \nabla b_{\underline{\theta}^o}\mathbf{D}\nabla b_{\underline{\theta}^o}^T. \end{aligned} \quad (4)$$

Therefore, combining (4) and (2)

$$\sup_{\underline{\theta}\in\mathcal{D}} |b_{\underline{\theta}} - b_{\underline{\theta}^o}|^2 = \|\nabla b_{\underline{\theta}^o}\|_D^2 + \epsilon. \quad (5)$$

For the special case of linear estimation $\epsilon = 0$ and the relation between bias gradient norm and maximal bias variation is exact for all ellipsoidal regions $\mathcal{D}$. In general, the norm of the bias gradient is approximately equal to the maximal bias variation over a region $\mathcal{D}$ when the product of the width of the region and the curvature of the bias function $b_{\underline{\theta}}$ over the region is small. Thus the norm $\|\bullet\|^2$ should be selected in such a way that the *a priori* range of uncertainty $\mathcal{D}$ adequately reflects expected joint parameter variations.

3 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 *bias-variance* or $\delta\sigma$ plane. A region of the $\delta\sigma$ plane is called unachievable if no estimator can exist having coordinates (δ, σ) in this region. While no non-empty unachievable region can exist for the parameterization $(b_{\underline{\theta}}, \sigma_{\underline{\theta}})$ of the plane, we will show that interesting unachievable regions exist for the parameterization $(\delta_{\underline{\theta}}, \sigma_{\underline{\theta}})$ of the plane.

The Biased CR Bound

Let $m_{\underline{\theta}} = E_{\underline{\theta}}[\hat{\theta}_1]$ be the mean of the estimator $\hat{\theta}_1$. The standard Cramer-Rao lower bound on estimator variance [12] only applies to unbiased estimators. For a biased estimator the following *biased CR bound* on variance $\sigma_{\underline{\theta}}^2$ is available:

$$\begin{aligned}\sigma_{\underline{\theta}}^2 &\ge [\nabla m_{\underline{\theta}}]\, \mathbf{F}_Y^+\, [\nabla m_{\underline{\theta}}]^T \\ &= \left[\underline{e}_1^T + \nabla b_{\underline{\theta}}\right] \mathbf{F}_Y^+ \left[\underline{e}_1^T + \nabla b_{\underline{\theta}}\right]^T, \end{aligned} \quad (6)$$

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

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

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

The non-singular form of the biased CR bound has been around for some time [13]. The more general pseudo-inverse form given in (6) is less well known but can be easily derived by identifying $U = \hat{\theta}_1 - \theta_1$ and $V = \nabla \ln f_{\mathrm{Y}}(\mathrm{Y};\underline{\theta})$ in the relation [14, Lemma 1]:

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

and using the well known facts: $E_{\underline{\theta}}[\nabla_{\underline{\theta}} \ln f_{\mathrm{Y}}(\mathrm{Y};\underline{\theta})] = 0$ and $E_{\underline{\theta}}[\hat{\theta}_1 \nabla_{\underline{\theta}} \ln f_{\mathrm{Y}}(\mathrm{Y};\underline{\theta})] = \nabla m_{\underline{\theta}}$.

The biased CR bound (6) only applies to the class of estimators $\hat{\theta}_1$ which have a particular bias gradient function $\nabla b_{\underline{\theta}}$. Therefore it cannot be used to simultaneously bound the variance of several estimators, each of which have different but perhaps acceptable bias gradients.

The Uniform CR Bound

In [6] a "uniform" CR bound was presented as a way to study the reliability of the standard CR bound under conditions of very small estimator bias. The following theorem extends the results of [6] to allow singular Fisher information matrices and arbitrary weighted Euclidean norm $\|\bullet\|_D$.

Theorem 1 *Let $\hat{\theta}_1$ be an estimator of the first component of the unknown parameter vector $\underline{\theta} = [\theta_1, \ldots, \theta_n]^T$. For a fixed scalar $\delta \ge 0$ let the bias gradient of $\hat{\theta}_1$ satisfy the norm constraint $\|\nabla b_{\underline{\theta}}\|_D \le \delta$, where $\boldsymbol{D}$ is an arbitrary $n \times n$ non-singular matrix. Define $\mathcal{P}_B$ as the $n \times n$ matrix which projects onto the column space of $\boldsymbol{B} = \boldsymbol{D}^{-\frac{T}{2}}\boldsymbol{F}_Y^+\boldsymbol{D}^{-\frac{1}{2}}$, and let $\underline{e}_1 = [1, 0, ..., 0]^T$ be the n-element unit vector. Then the variance of $\hat{\theta}_1$ satisfies:*

$$\sigma_{\underline{\theta}}^2 \ge B(\underline{\theta}, \delta), \quad (7)$$

where if $\underline{e}_1^T \boldsymbol{D}^{\frac{T}{2}}\mathcal{P}_B\boldsymbol{D}^{\frac{1}{2}}\underline{e}_1 \le \delta^2$ then $B(\underline{\theta}, \delta) = 0$, while if $\underline{e}_1^T \boldsymbol{D}^{\frac{T}{2}}\mathcal{P}_B\boldsymbol{D}^{\frac{1}{2}}\underline{e}_1 \le \delta^2$ then:

$$\begin{aligned} B(\underline{\theta}, \delta) &= \left[\underline{e}_1 + \underline{d}_{min}\right]^T \boldsymbol{F}_Y^+ \left[\underline{e}_1 + \underline{d}_{min}\right], \quad (8) \\ &= \lambda^2 \underline{e}_1^T \boldsymbol{D}\left[\lambda\boldsymbol{D} + \boldsymbol{F}_Y^+\right]^{-1} \boldsymbol{F}_Y^+ \left[\lambda\boldsymbol{D} + \boldsymbol{F}_Y^+\right]^{-1} \boldsymbol{D}\underline{e}_1, \quad (9) \end{aligned}$$

where in (8)

$$\underline{d}_{min} = -\left[\lambda\boldsymbol{D} + \boldsymbol{F}_Y^+\right]^{-1} \boldsymbol{F}_Y^+ \underline{e}_1, \quad (10)$$

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

$$g(\lambda) \stackrel{def}{=} \underline{e}_1^T \boldsymbol{F}_Y^+ \left[\lambda\boldsymbol{D} + \boldsymbol{F}_Y^+\right]^{-1} \boldsymbol{D}\left[\lambda\boldsymbol{D} + \boldsymbol{F}_Y^+\right]^{-1} \boldsymbol{F}_Y^+ \underline{e}_1 \quad (11)$$

By tracing out the family of points $\{(\delta, B(\underline{\theta}, \delta)) : \delta \ge 0\}$ one obtains a curve in the $\delta\sigma$ plane. 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 bias-variance trade-off curve plotted in terms of normalized standard deviation $\sqrt{\frac{B(\underline{\theta},\delta)}{B(\underline{\theta},0)}}$. If an estimator lies on the curve then lower variance can only be bought at the price of increased

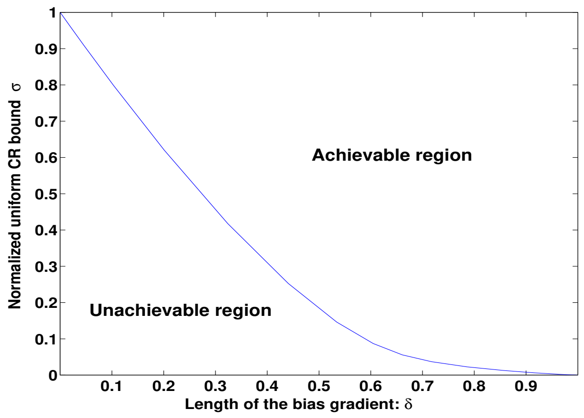

Figure 1: The Normalized Uniform CR bound on the $\delta\sigma$ trade-off plane.

bias and vice versa. For this reason we call this curve the *bias-variance tradeoff curve*.

It is important to point out that the bias-variance tradeoff curve can be generated without solving the non-linear equation (11), which generally must be solved numerically. It is much easier to continously vary λ over the range $(0, \infty)$ and sweep out the curve by using the λ-parameterizations of δ 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 $B(\underline{\theta}; 0) = \underline{e}_1 \mathbf{F}_Y^+ \underline{e}_1$ with equality at $\delta = 0$. The slope of $B(\underline{\theta}, \delta)$ at $\delta = 0$ gives a bias sensitivity index η for the CR bound. For non-singular $\mathbf{F}_Y$ it is shown in [6] that $\eta = 2\sqrt{1 + \underline{c}^T \mathbf{F}_S^{-1} \underline{c}}$, 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 a reliable lower bound for estimators which may have very small and perhaps unmeasurable biases.

- The orthogonal projection $\mathcal{P}_\mathbf{B}$ can be expressed either as $\mathcal{P}_\mathbf{B} = \mathbf{B}[\mathbf{B}^T\mathbf{B}]^+\mathbf{B}^T$, or via the eigendecomposition of $\mathbf{B}$ as $\mathcal{P}_\mathbf{B} = \sum_{i=1}^r \underline{\xi}\underline{\xi}^T$, 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 $\underline{e}_1^T \mathbf{D}^{\frac{T}{2}} \mathcal{P}_B \mathbf{D}^{\frac{1}{2}} \underline{e}_1 = \underline{e}_1^T \mathbf{F}_Y^+ [\mathbf{D}^{-\frac{T}{2}} \mathbf{F}_Y^+]^+ \mathbf{D}^{\frac{1}{2}} \underline{e}_1$ which simplifies to $\underline{e}_1^T \underline{e}_1 = 1$ when $\mathbf{F}_Y$ is positive definite.

- When $\mathbf{F}_Y$ is non-singular, $F_Y^+ = F_Y^{-1}$, $\mathcal{P}_B = \mathbf{I}$, $\underline{e}_1^T \mathbf{D}^{\frac{T}{2}} \mathcal{P}_B \mathbf{D}^{\frac{T}{2}} \underline{e}_1 = \|\underline{e}_1\|_D^2$ and (8)-(11) of Theorem 1 reduce to

$$
\begin{aligned}
B(\underline{\theta}, \delta) &= [\underline{e}_1 + \underline{d}_{min}]^T \mathbf{F}_Y^{-1} [\underline{e}_1 + \underline{d}_{min}] \qquad (12)\\
&= \lambda^2 \cdot \underline{e}_1^T \left[\mathbf{D}^{-1} + \lambda \mathbf{F}_Y\right]^{-1} \mathbf{F}_Y \left[\mathbf{D}^{-1} + \lambda \mathbf{F}_Y\right]^{-1} \underline{e}_1
\end{aligned}
$$

where,

$$
\underline{d}_{min} = -\mathbf{D}^{-1} \left[\mathbf{D}^{-1} + \lambda\, \mathbf{F}_Y\right]^{-1} \underline{e}_1, \qquad (13)
$$

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

$$
g(\lambda) = \underline{e}_1^T \left[\mathbf{D}^{-1} + \lambda \mathbf{F}_Y\right]^{-1} \mathbf{D}^{-1} \left[\mathbf{D}^{-1} + \lambda \mathbf{F}_Y\right]^{-1} \underline{e}_1. \quad (14)
$$

When $\mathbf{D} = \mathbf{I}$ these are identical to the results obtained in [6].

- In Theorem 1, $\underline{d}_{min}$ defined in (10) is an optimal bias gradient in the sense that it minimizes the biased CR bound (6) over all vectors $\nabla b_{\underline{\theta}}$ under the constraint $\delta < 1$. The bound is independent of the particular estimator bias as long as the bias gradient norm constraint is satisfied. From the proof of Theorem 1, if $\underline{e}_1^T \mathbf{D}^{\frac{T}{2}} \mathcal{P}_\mathbf{B} \mathbf{D}^{\frac{1}{2}} \underline{e}_1 \le \delta^2$, then the minimizing bias gradient is of the form $\underline{d}_{min} = -\mathcal{P}_\mathbf{B} \mathbf{D}^{\frac{1}{2}} \underline{e}_1 + \underline{\phi}$, where $\underline{\phi}$ is any vector satisfying $\mathbf{B}\underline{\phi} = \underline{0}$, and $\|\underline{\phi}\|^2 \le \delta^2 - \underline{e}_1^T \mathbf{D}^{\frac{T}{2}} \mathcal{P}_\mathbf{B} \mathbf{D}^{\frac{1}{2}} \underline{e}_1$. Thus for the case of singular $\mathbf{F}_Y$ there exist many optimal bias gradients for which the biased CR bound is minimized.

- It has been shown [6] that if $\mathbf{F}_Y$ is non-singular, and the unbiased CR bound is achievable by an unbiased estimator $\hat{\theta}^*$, then one can construct an estimator that locally achieves the uniform bound by introducing a small amount of bias in $\hat{\theta}^*$. However, since unbiased estimators may not exist for singular $\mathbf{F}_Y$, the uniform CR bound for singular $\mathbf{F}_Y$ may not always be achievable.

Computational Issues

As written in Theorem 1 expressions (9)-(11) are not in the most convenient form for computation as they involve several matrix multiplications/inversions and may be numerically unstable for ill conditioned $\mathbf{F}_Y$. In this section we discuss methods for computation.

An equivalent form for the pair $g(\lambda)$ and $B(\underline{\theta}, \delta)$ in (11) and (9) was obtained in the process of proving Theorem 1

$$
\begin{aligned}
B(\underline{\theta}, \delta) &= \lambda^2 \underline{e}_1^T \mathbf{D}^{\frac{T}{2}} \left[\lambda\mathbf{I} + \mathbf{B}\right]^{-1} \mathbf{B} \left[\lambda\mathbf{I} + \mathbf{B}\right]^{-1} \mathbf{D}^{\frac{1}{2}} \underline{e}_1\\
g(\lambda) &= \underline{e}_1^T \left[\mathbf{D}^{\frac{T}{2}} \mathbf{B} (\lambda \cdot \mathbf{I} + \mathbf{B})^{-2} \mathbf{B} \mathbf{D}^{\frac{1}{2}}\right] \underline{e}_1 = \delta^2.
\end{aligned}
$$

If the eigendecomposition of the matrix $\mathbf{B}$ is available, the bias-variance tradeoff curve can be efficiently computed by sweeping out λ in the following weighted sum of inner products

$$
\begin{aligned}
B(\underline{\theta}, \delta) &= \sum_{i=1}^n \frac{\beta_i}{(\lambda + \beta_i)^2} |\underline{\tilde{e}}_1^T \xi_i|^2 \qquad (15)\\
\delta^2 &= \sum_{i=1}^n \frac{\beta_i^2}{(\lambda + \beta_i)^2} |\underline{\tilde{e}}_1^T \xi_i|^2, \qquad (16)
\end{aligned}
$$

where β_i and ξ_i denote eigenvalues and eigenvectors of $\mathbf{B}$ and $\tilde{\underline{e}}_1 = D^{half}\underline{e}_1$.

When $\mathbf{F}_Y$ is non-singular but has an ill-conditioned inverse, the eigendecomposition of $\mathbf{B}$ may be numerically unstable. A simple algebraic manipulation in (12)-(14) yields

$$\begin{aligned} B(\underline{\theta}, \delta) &= \lambda^2 \underline{e}_1^T \mathbf{D}^{\frac{T}{2}} [\mathbf{I} + \lambda\mathbf{C}]^{-1} \mathbf{C} [\mathbf{I} + \lambda\mathbf{C}]^{-1} \mathbf{D}^{\frac{1}{2}} \underline{e}_1 \quad (17) \\ g(\lambda) &= \underline{e}_1^T \mathbf{D}^{\frac{T}{2}} [\mathbf{I} + \lambda\mathbf{C}]^{-2} \mathbf{D}^{\frac{1}{2}} \underline{e}_1, \end{aligned}$$

where $\mathbf{C} = \mathbf{B}^{-1} = \mathbf{D}^{\frac{1}{2}} \mathbf{F}_Y \mathbf{D}^{\frac{T}{2}}$. Note that the form (18) only requires inversion of the matrix $\mathbf{I} + \lambda\mathbf{C}$. Since $\lambda > 0$ and $\mathbf{F}_Y$ is positive definite this inversion is well conditioned.

For non-singular $\mathbf{F}_Y$, it is possible to significantly reduce the required number of computations by using the conjugate gradient algorithm (CG) applied to solving a system of equations $\mathbf{A}\underline{t} = \underline{c}$ where $\mathbf{A}$ is symmetric positive definite [15].

CG Uniform Bound Computation

Assume the uniform CR bound for the k-th component θ_k of an $n \times 1$ parameter vector $\underline{\theta} = [\theta_1, \ldots, \theta_n]^T$. Then the numerically stable form of the uniform CR bound (18) for non-singular $\mathbf{F}_Y$ can be computed as follows.

1. Compute $\mathbf{C} = \mathbf{D}^{\frac{1}{2}} \mathbf{F}_Y \mathbf{D}^{\frac{T}{2}}$ and $\underline{c} = \mathbf{D}^{\frac{1}{2}} \underline{e}_1$.
2. Interchange 1-st and k-th row of $\underline{c}$ and 1-st and k-th row and column of $\mathbf{C}$.
3. Select $\lambda \in (0, \infty)$.
4. Compute $\underline{t} = [\mathbf{I} + \lambda\mathbf{C}]^{-1} \mathbf{D}^{\frac{1}{2}} \underline{e}_1$ by applying the CG algorithm to solve the following for $\underline{t}$:
$$[\mathbf{I} + \lambda\mathbf{C}]\, \underline{t} = \underline{c}.$$
5. Compute the point $(\delta, B(\underline{\theta}, \delta))$ via
$$\begin{aligned} B(\underline{\theta}, \delta) &= \lambda^2 \underline{t}^T \mathbf{C} \underline{t} \\ \delta &= \|\underline{t}\|_2. \end{aligned}$$

For each λ the above algorithm requires a single application of the conjugate gradient algorithm. Applying this algorithm for several values of λ allows one to trace the curve $B(\underline{\theta}, \delta)$ in the $\delta\sigma$ trade-off plane of Figure 1.

4 Estimation of Bias Gradient Norm

In order 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. If we could generate L i.i.d. realizations $\{\mathrm{Y}_i^j\}_{i=1}^L$ from each of the densities $\{f_{\mathrm{Y}}(\mathrm{y}; \underline{\theta} + \epsilon\underline{e}_j)\}_{j=0}^n$, where $\underline{e}_0 = \underline{0}$, $\underline{e}_j$ $(j > 0)$ is the j-th unit vector in $\mathbb{R}^n$, and ϵ is a small but non-zero scalar, then we could estimate the bias gradient vector by the finite difference approximation

$$\widehat{\nabla b_{\underline{\theta}}} = \frac{1}{\epsilon} \left[\widehat{b_{\underline{\theta}+\epsilon\underline{e}_1}} - \widehat{b_{\underline{\theta}}}, \ldots, \widehat{b_{\underline{\theta}+\epsilon\underline{e}_n}} - \widehat{b_{\underline{\theta}}} \right], \qquad (18)$$

where, for $j = 0, \ldots, n$, $\widehat{b_{\underline{\theta}+\epsilon\underline{e}_j}}$ is the sample mean estimator

$$\widehat{b_{\underline{\theta}}} = \frac{1}{L} \sum_{i=1}^{L} \hat{\theta}(\mathrm{Y}_i^j) - \underline{\theta}.$$

The estimator (18) is a consistent and unbiased estimate of the finite difference approximation to $\nabla b_{\underline{\theta}}$. However, it requires generation of nL realizations of Y which is not practical when the dimension n of $\underline{\theta}$ is large. For example, in image reconstruction a 32×32 image would require performing $n - 1 = 1023$ additional simulations as compared to the single simulation required for computation of sample bias and sample variance of $\hat{\theta}_1$.

The following is a direct method for estimating the bias gradient which does not require performing $n-1$ additional simulations 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

$$\begin{aligned} E_{\underline{\theta}}\left[\mathrm{Z} \frac{\partial \ln f_{\mathrm{Z}}(\mathrm{Z}; \underline{\theta})}{\partial \theta_j} \right] &= \int_{\mathrm{Z}} z \frac{\partial f_{\mathrm{Z}}(z; \underline{\theta})}{\partial \theta_j}\, dz \\ &= \frac{\partial}{\partial \theta_j} \int_{\mathrm{Z}} z f_{\mathrm{Z}}(z; \underline{\theta})\, dz \\ &= \frac{\partial}{\partial \theta_j} E_{\underline{\theta}}(\mathrm{Z}). \end{aligned} \qquad (19)$$

Thus in particular we have the following relation

$$\nabla b_{\underline{\theta}} = E_{\underline{\theta}} \left[\hat{\theta}_1(\mathrm{Y}) \nabla_{\underline{\theta}} \ln f_{\mathrm{Y}}(\mathrm{Y}; \underline{\theta}) \right] - \underline{e}_1.$$

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

$$\nabla b_{\underline{\theta}} = E_{\underline{\theta}} \left[(\hat{\theta}_1(\mathrm{Y}) - \zeta_i) \nabla_{\underline{\theta}} \ln f_{\mathrm{Y}}(\mathrm{Y}; \underline{\theta}) \right] - \underline{e}_1, \qquad (20)$$

for any random variable ζ_i statistically independent of Y. As explained in the following discussion, the quantity ζ_i is introduced in (20) to control the variance of the bias gradient estimate.

Substituting sample averages for ensemble averages in (20) we obtain the following unbiased and consistent estimator of the bias gradient vector $\nabla b_{\underline{\theta}}$

$$\widehat{\nabla b_{\underline{\theta}}} = \frac{1}{L} \sum_{i=1}^{L} \left(\hat{\theta}_1(\mathrm{Y}_i) - \zeta_i \right) \nabla_{\underline{\theta}} \ln f_{\mathrm{Y}}(\mathrm{Y}_i; \underline{\theta}) - \underline{e}_1^T. \qquad (21)$$

where Y_i is a set of i.i.d. realizations from $f_{\mathrm{Y}}(\mathrm{y}; \underline{\theta})$.

The covariance matrix of $\widehat{\nabla b_{\underline{\theta}}}$ is

$$S(\widehat{\nabla b_{\underline{\theta}}}) = \frac{1}{L}\mathrm{cov}_{\underline{\theta}}\left((\hat{\theta}_1(\mathrm{Y}_i) - m_{\underline{\theta}})\nabla \ln f_{\mathrm{Y}}(\mathrm{Y}_i;\underline{\theta})\right) + \frac{1}{L^2}\sum_{j=1}^{L} E_{\underline{\theta}}\left[(m_{\underline{\theta}} - \zeta_j)^2\right] \mathbf{F}_Y, \tag{22}$$

where $m_{\underline{\theta}} = E_{\underline{\theta}}[\hat{\theta}_1]$ and $\mathbf{F}_Y = E_{\underline{\theta}}\left([\nabla \ln f_{\mathrm{Y}_i}(\mathrm{Y}_i;\underline{\theta})]^T [\nabla \ln f_{\mathrm{Y}_i}(\mathrm{Y}_i;\underline{\theta})]\right)$ is the single trial Fisher information. Assuming the ζ_i are i.i.d., in which case $\frac{1}{L^2}\sum_{j=1}^{L} E_{\underline{\theta}}\left[(m_{\underline{\theta}} - \zeta_j)^2\right] \mathbf{F}_Y = \frac{1}{L}E_{\underline{\theta}}\left[(m_{\underline{\theta}} - \zeta_j)^2\right] \mathbf{F}_Y$ both terms on RHS of (22) decreases to zero at rate $\frac{1}{L}$. However, unless ζ_i is chosen appropriately, the second term on the RHS of (22) can dominate the expression, as shown in [16] for the weighted least-squares estimator. Since $m_{\underline{\theta}}$ is unknown we cannot eliminate this second term by choosing $\zeta_i = m_{\underline{\theta}} =$ constant for all i. However, we can use the punctured sample mean estimate of $m_{\underline{\theta}}$: $\zeta_i = \frac{1}{L-1}\sum_{j=1, j\neq i}^{L} \hat{\theta}(\mathrm{Y}_j)$. which, as required for the validity of (20), is statistically independent of Y_i. Substitution of this ζ_i into (21) gives the following unbiased and consistent sample mean estimate of $\nabla b_{\underline{\theta}}$

$$\widehat{\nabla b_{\underline{\theta}}} \tag{23}$$

$$= \frac{1}{L-1}\sum_{i=1}^{L}\left(\hat{\theta}_1(\mathrm{Y}_i) - \sum_{j=1}^{L}\hat{\theta}(\mathrm{Y}_j)\right) \nabla_{\underline{\theta}} \ln f_{\mathrm{Y}}(\mathrm{Y}_i;\underline{\theta}) - \underline{e}_1^T.$$

A simple calculation shows that the covariance of (23)

$$S(\widehat{\nabla b_{\underline{\theta}}}) = \frac{1}{L}\mathrm{cov}_{\underline{\theta}}\left((\hat{\theta}_1(\mathrm{Y}_i) - m_{\underline{\theta}})\nabla_{\underline{\theta}} \ln f_{\mathrm{Y}}(\mathrm{Y}_i;\underline{\theta})\right) + \frac{1}{L(L-1)}\mathrm{var}_{\underline{\theta}}(\hat{\theta}_1(\mathrm{Y}_i)) \cdot \mathbf{F}_Y. \tag{24}$$

Note that now the second term in (24) decreases to zero at the much faster asymptotic rate of $\frac{1}{L^2}$ as compared to the rate $\frac{1}{L}$ in (22).

A Bootstrap Estimator for Bias Gradient Norm

A natural estimate for $\delta_{\underline{\theta}} = \|\nabla b_{\underline{\theta}}\|_D^2$ is the norm squared of the sample mean $\hat{\delta}^2 = \|\widehat{\nabla b_{\underline{\theta}}}\|_D^2$. It can easily be shown that the bias of this estimate is equal to the quantity $E_{\underline{\theta}}\left[\|\widehat{\nabla b_{\underline{\theta}}} - \nabla b_{\underline{\theta}}\|_D^2\right] = trace\{S(\widehat{\nabla b_{\underline{\theta}}})\}$ which, in view of (22), decays to zero only as $1/L$. Below we present a reduced bias estimator based on the bootstrap resampling methodology.

Let $\mathrm{Y}_1^*, \ldots, \mathrm{Y}_L^*$ denote a *bootstrap sample* obtained by randomly resampling the data points $\mathrm{Y}_1, \ldots, \mathrm{Y}_L$ with replacement. Given the estimator $\hat{\delta}^2 = \hat{\delta}^2(\mathrm{Y}_1, \ldots, \mathrm{Y}_L) = \|\widehat{\nabla b_{\underline{\theta}}}\|_D^2$, the bootstrap estimate of $\delta_{\underline{\theta}}^2$ is defined as the expectation of $\hat{\delta}_*^2 = \hat{\delta}^2(\mathrm{Y}_1^*, \ldots, \mathrm{Y}_L^*)$ with respect to the resampling distribution [17]:

$$E_*[\hat{\delta}_*^2] = \sum_* \hat{\delta}^2(\mathrm{Y}_1^*, \ldots, \mathrm{Y}_L^*)\binom{L}{c_1 \ldots c_L} L^{-L} \tag{25}$$

where $c_1, \ldots, c_L$ are non-negative integers satisfying $\sum_{i=1}^{L} c_i = L$ and $\sum_*$ denotes a summmation over all possible resamplings (with replacement) of the L quantities $\mathrm{Y}_1, \ldots, \mathrm{Y}_L$. 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

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

Due to the simple quadratic dependence of $\hat{\delta}^2$ on Y_i, the expectation (25) can be expressed in analytical form leading to the bias corrected estimate:

$$\hat{\delta}_c^2 = \|\widehat{\nabla b_{\underline{\theta}}}\|_D^2 - \frac{1}{L^2}\sum_{i=1}^{L}\|\widehat{\nabla b_{\underline{\theta}}}(\mathrm{Y}_i) - \widehat{\nabla b_{\underline{\theta}}}\|_D^2,$$

where $0\widehat{\nabla b_{\underline{\theta}}}(\mathrm{Y}_i)$ is the estimate (23) based on a single sample Y_i ($L = 1$). The bias of $\hat{\delta}_c^2$ is equal to

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

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

5 Application to Inverse Problems

Linear Gaussian Noise Model

Assume that the observations consist of a vector $\mathrm{Y} = \underline{\mathrm{Y}} \in I\!R^m$ which obeys the linear Gaussian model:

$$\underline{\mathrm{Y}} = \mathbf{A}\underline{\theta} + \underline{\epsilon}, \tag{26}$$

where $\mathbf{A}$ is an $m \times n$ matrix, called the convolution kernel matrix, $\underline{\theta}$ is an unknown source to be estimated, and $\underline{\epsilon}$ is a vector of zero mean Gaussian random variables with positive definite covariance matrix Σ. The Fisher information matrix has the well known form [18]

$$\mathbf{F}_Y = \mathbf{A}^T\Sigma^{-1}\mathbf{A}. \tag{27}$$

This matrix is non-singular when $\mathbf{A}$ is of full rank n. We use Theorem 1 to trace out the bias-variance tradeoff curve in the $\delta\sigma$ plane by evaluating $\delta^2 = g(\lambda)$ in (11) and $B(\underline{\theta}, \delta)$ in (9) and plotting $(\delta^2, B(\underline{\theta}, \delta))$ parametrically as a function of $\lambda > 0$. Since $\mathbf{F}_Y$ is not functionally dependent on $\underline{\theta}$ this curve will not depend on the specific form of the source.

To investigate the achievability of the bias-variance 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 [19] and to incorporate parameter constraints or a priori information [20]. The QPML estimator minimizes the objective function

$$
\begin{aligned}
&[\underline{Y} - \mathbf{A}\underline{\theta}]^T \Sigma^{-1} [\underline{Y} - \mathbf{A}\underline{\theta}] + \alpha \underline{\theta}^T \mathbf{P} \underline{\theta} = \\
&\left[\underline{\theta} - [\mathbf{F}_Y + \alpha \mathbf{P}]^{-1} \mathbf{A}^T \Sigma^{-1} \underline{Y}\right]^T (\mathbf{F}_Y + \alpha \mathbf{P}) \\
&\left[\underline{\theta} - [\mathbf{F}_Y + \alpha \mathbf{P}]^{-1} \mathbf{A}^T \Sigma^{-1} \underline{Y}\right] \\
&+ \underline{Y}^T \left[\Sigma^{-1} - \Sigma^{-1} \mathbf{A} (\mathbf{F}_Y + \alpha \mathbf{P}) \mathbf{A}^T \Sigma^{-1}\right] \underline{Y},
\end{aligned} \tag{28}
$$

where $\alpha > 0$ is a regularization or smoothing parameter and $\mathbf{P}$ is a symmetric positive-definite penalty matrix. For ill conditioned or singular $\mathbf{A}$ the regularization parameter $\alpha > 0$ improves the numerical stability of the matrix inversion $(\mathbf{A}^T \Sigma^{-1} \mathbf{A} + \alpha \mathbf{P})^{-1}$ by lowering its condition number at the expense of introducing bias. The simplest choice for the penalty matrix $\mathbf{P}$ is the identity $\mathbf{I}$ which yields a class of energy constrained least squares estimators variously known as Tikonov regularized least squares in the inverse problem literature [19], or ridge regression in the multivariate statistics literature [3]. A popular choice in imaging applciations is to use a non-diagonal differencing operator to enforce smoothness constraints or roughness priors [21, 22].

The solution $\underline{\theta} = \hat{\underline{\theta}}$ which minimizes (28) is the penalized least squares (PLS) estimator

$$
\hat{\underline{\theta}} = [\mathbf{F}_Y + \alpha \mathbf{P}]^{-1} \mathbf{A}^T \Sigma^{-1} \underline{Y}. \tag{29}
$$

Since the estimator is linear its bias and variance can be computed in closed form. The bias of the first component $\hat{\theta}_1$ of the PLS estimator is

$$
b_{\underline{\theta}} = \underline{e}_1^T \left[[\mathbf{F}_Y + \alpha \mathbf{P}]^{-1} \mathbf{F}_Y - \mathbf{I} \right] \underline{\theta},
$$

and its bias gradient is

$$
\begin{aligned}
\nabla b_{\underline{\theta}} &= \underline{e}_1^T \left[[\mathbf{F}_Y + \alpha \mathbf{P}]^{-1} \mathbf{F}_Y - \mathbf{I} \right] && (30) \\
&= -\alpha \underline{e}_1^T [\mathbf{F}_Y + \alpha \mathbf{P}]^{-1} \mathbf{P}. && (31)
\end{aligned}
$$

From expression (30) the bias gradient (transpose) is seen to be equal to the difference between the mean response $\mathbf{F}_Y [\mathbf{F}_Y + \alpha \mathbf{P}]^{-1} \underline{e}_1$ of the PLS estimator to a point source $\underline{\theta} = \underline{e}_1$, i.e., the point spread function, and the ideal point response $\underline{e}_1$. Thus the bias gradient can be viewed as a measure of the geometric resolution of the estimator.

The variance of the first component $\hat{\theta}_1$ of the PLS estimator is

$$
\sigma_{\underline{\theta}}^2 = \underline{e}_1^T \left[\mathbf{P} + \frac{1}{\alpha} \mathbf{F}_Y \right]^{-1} \mathbf{F}_Y \left[\mathbf{P} + \frac{1}{\alpha} \mathbf{F}_Y \right]^{-1} \underline{e}_1. \tag{32}
$$

Assume that $\mathbf{F}_Y$ is non-singular and compare (31) and (32) to the simplified equations (13) and (12) for $\underline{d}_{min}$ and the bound $B(\underline{\theta}, \delta)$, respectively. Identifying $\lambda = 1/\alpha$, it is clear that when $\mathbf{P}$ is chosen as $\mathbf{D}^{-1}$ the PLS estimator achieves the bound $B(\underline{\theta}, \delta)$. Thus the region above and including the bias-variance tradeoff curve in the $\delta\sigma$ plane is an achievable region. Furthermore, since the bias-gradient is linear, relation (5) implies that this PLS estimator is a minimum variance biased estimator: it minimizes variance over all estimators satisfying a constraint on maximal bias variation over the region $\sup_{\underline{\theta} \in \mathcal{D}} |b_{\underline{\theta}} - b_{\underline{\theta}^o}| \leq \delta^2$, where α and δ are related through the relation (14), $g(1/\alpha) = \delta^2$. Thus we conclude that for the case of non-singular $\mathbf{F}_Y$ the PLS estimator (29) introduces bias in an optimal manner, i.e. the bias gradient achieves the most variance reduction for its allowed amount of bias.

Figure 2 shows the bias-variance tradeoff curve and achievable/unachievable regions for $\mathbf{D} = I$ and for $\mathbf{A} = ((a_{ij}))$ a full rank 300×300 matrix corresponding to the discrete Gaussian kernel $a_{ij} = \frac{1}{\sqrt{2\pi}} e^{\frac{-(i-j)^2}{2w^2}}$ $(w = 5)$, a kernel frequently used for ill-conditioned discrete deconvolution problems [23]. The asterisks denote the bias-variance curve $(\|\nabla b_{\underline{\theta}}\|_2, \sigma_{\underline{\theta}})$ associated with the PLS estimator.

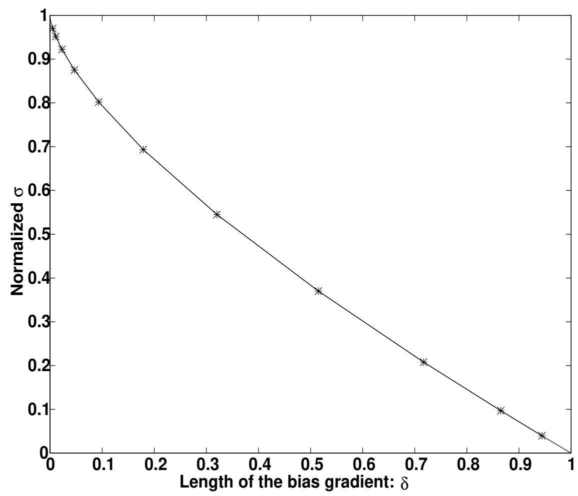

Figure 2: Unconstrained WLSE with ideal weights exactly achieves the uniform bound.

When $\mathbf{A}$ has rank less than n, $\mathbf{F}_Y$ is singular and there generally exists no unbiased estimator of the vector $\underline{\theta}$ [18, 24]. Thus in this case we cannot expect the bias-variance tradeoff curve to be achievable for all δ. In Figure 3 we plot the bias-variance tradeoff curve along with the bias-variance curve associated the PLS estimator (29) under the same conditions as in Figure 2 except that the Gaussian kernel matrix $\mathbf{A}$ is rank deficient of dimension 150×300. This is a case of "super resolution with an up-sampling factor of 2." The bound $B(\underline{\theta}, \delta)$ is equal to zero for $\delta > 0.81$ since $\underline{e}_1^T \mathcal{P}_{\mathbf{F}_Y} \underline{e}_1 \leq \delta^2$ in that region. Note that the PLS estimator (29) now fails to achieve the bound for any value of α.

Under certain conditions the uniform CR bound is achievable for singular $\mathbf{F}_Y$, although not by a PLS esti-

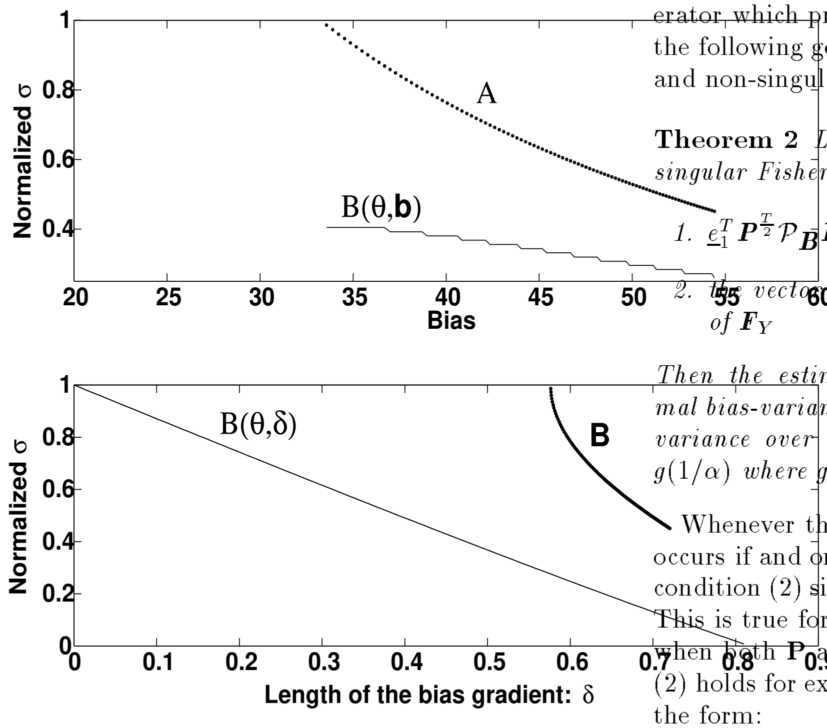

Figure 3: Super resolution problem: WLSE fails to achieve the uniform bound at all points. The unbiased CR bound $B(\underline{\theta}, 0)$ is $\approx 50\%$ of the same in 2.

mator of the form (29). Consider the estimator

$$\hat{\theta} = \mathbf{P}^{-1}\left[\mathbf{P}^{-1} + \alpha \mathbf{F}_Y^{+}\right]^{-1} \mathbf{F}_Y^{+} \mathbf{A}^T \Sigma^{-1} \underline{Y}, \tag{33}$$

where as before α is a non-negative constant. The first component of $\hat{\theta}$ has bias gradient

$$\begin{aligned} \nabla b_{\underline{\theta}} &= \underline{e}_1^T \left(\mathbf{P}^{-1}\left[\mathbf{P}^{-1} + \alpha \mathbf{F}_Y^{+}\right]^{-1} \mathbf{F}_Y^{+}\mathbf{F}_Y - \mathbf{I}\right) \\ &= -\underline{e}_1^T \mathbf{F}_Y^{+} \left[\frac{1}{\alpha}\mathbf{P}^{-1} + \mathbf{F}_Y^{+}\right]^{-1} \\ &\quad -\underline{e}_1^T \mathbf{P}^{-1}\left[\mathbf{I} + \alpha \mathbf{F}_Y^{+}\mathbf{P}\right]^{-1}\left[\mathbf{I} - \mathbf{F}_Y^{+}\mathbf{F}_Y\right]. \end{aligned} \tag{34}$$

and variance

$$\begin{aligned} &\sigma^2_{\underline{\hat{\theta}}} \\ &= \underline{e}_1^T \mathbf{P}^{-1}\left[\mathbf{P}^{-1} + \alpha \mathbf{F}_Y^{+}\right]^{-1} \mathbf{F}_Y^{+} \left[\mathbf{P}^{-1} + \alpha \mathbf{F}_Y^{+}\right]^{-1} \mathbf{P}^{-1} \underline{e}_1 \end{aligned} \tag{35}$$

where in (35) we have used the property $\mathbf{F}_Y^{+}\mathbf{F}_Y\mathbf{F}_Y^{+} = \mathbf{F}_Y^{+}$ [25]. It is obvious that the estimator variance is equal to the lower bound expression $B(\underline{\theta}, \delta)$ given in (8) after making the identifications $\mathbf{P} = \mathbf{D}^{-1}$ and $1/\alpha = \lambda$. Furthermore, the bias gradient (34) differs from the optimal bias gradient (transpose) $\underline{d}_{min}^T$, given in (10), only by the presence of the second additive term on the right hand side of (34). Thus the estimator (33) is an optimal biased estimator when this second additive term is equal to zero.

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 $\mathbf{F}_Y$ we obtain the following general result which applies to both singular and non-singular $\mathbf{F}_Y$.

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

1. *$\underline{e}_1^T \boldsymbol{P}^{\frac{T}{2}} \mathcal{P}_{\boldsymbol{B}} \boldsymbol{P}^{\frac{1}{2}} \underline{e}_1 \leq \delta^2$, and*
2. *the vector $[\boldsymbol{I} + \boldsymbol{F}_Y^{+}\boldsymbol{P}]^{-1}\boldsymbol{P}^{-1}\underline{e}_1$ lies in the range space of $\boldsymbol{F}_Y$*

Then the estimator $\hat{\theta}_1$ given by (33) achieves the optimal bias-variance tradeoff in the sense of having minimum variance over all estimators satisfying $\|\nabla b_{\underline{\theta}}\|_{P^{-1}} \leq \delta = g(1/\alpha)$ where $g(\bullet)$ is the function given in (11).

Whenever the two matrices $\mathbf{P}$ and $\mathbf{F}_Y$ commute, which occurs if and only if they have the same eigenvectors, [26], condition (2) simplifies to: $\underline{e}_1$ lies in the range space of $\mathbf{F}_Y$. This is true for example when $\mathbf{P} = a\mathbf{I}$, for some $a > 0$, or when both $\mathbf{P}$ and $\mathbf{F}_Y$ are diagonal. In this case condition (2) holds for example when $\mathbf{F}_Y$ a block diagonal matrix of the form:

$$\mathbf{F}_Y = \begin{bmatrix} \mathbf{A} & 0 \\ 0 & \mathbf{B} \end{bmatrix},$$

where $\mathbf{A}$ is positive definite.

Deconvolution for a Poisson Noise Model

In some applications the observations $\underline{\mathrm{Y}}$ are given by the linear model (26) 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 [27]. The log-likelihood for the observations $\underline{\mathrm{Y}} = [\mathrm{Y}_1, \ldots, \mathrm{Y}_m]^T$ has the form

$$\ln f_{\underline{\mathrm{Y}}}(\underline{y}; \underline{\theta}) = \sum_{j=1}^{m} (y_j \ln(\mu_j) - \mu_j) + c. \tag{36}$$

In (36) c is a constant independent of $\underline{\theta}$ and $\underline{\mu} = [\mu_1, \ldots, \mu_m]^T$ is

$$\underline{\mu} = \mathbf{A}\,\underline{\theta} + \underline{s}. \tag{37}$$

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

The FIM has the form [28]

$$\mathbf{F}_Y = \mathbf{A}^T\left[\mathrm{diag}(\mu)\right]^{-1}\mathbf{A}. \tag{38}$$

We computed the bias-variance tradeoff curve for a rank n system matrix $\mathbf{A}$ corresponding to a parallel ray tomographic projection system arising in single photon emission

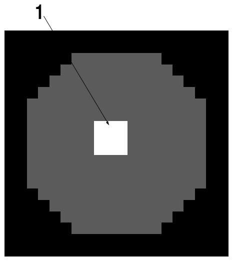

Figure 4: 20×20 image phantom used for bound calculation is a disk of intensity 1 with a high intensity region (hot spot) in the center having intensity 3. Pixel of interest is marked **1**.

tomography (SPECT), and a phantom image $\underline{\theta}$ shown in Figure 4. The pixel of interest with intensity θ_1 is at the top of the hot spot in the Figure 4.

To investigate the achievability of the region just above the bias-variance tradeoff curve, and to illustrate the empirical computation of bias gradient, we implemented a variant of the iterative space alternating generalized expectation-maximization (SAGE) of [7] adapted for penalized maximum likelohood image reconstruction (PML-SAGE) [29]. This algorithm iterates to the maximum of the penalized likelihood function

$$\ln f_{\underline{Y}}(\mathrm{y};\underline{\theta}) - \alpha\, \underline{\theta}^T \mathbf{P} \underline{\theta},$$

where $\underline{\theta}^T \mathbf{P}\underline{\theta} = \sum_j \sum_{k \in N_j} w_{jk}(\theta_j - \theta_k)^2$ is a roughness penalty chosen to reflect the expected roughness of the image (w_{jk} is a nearest neighbor weighting and N_j is the 8 pixel neighborhood of pixel j). Setting the smoothing parameter $\alpha = 0$ corresponds to maximum likelihood reconstruction with no image smoothing.

Since the PML-SAGE algorithm is non-linear, exact analytic expressions of the variance, bias and bias gradient are intractable. However, it will be instructive to consider the local approximation for variance derived in [30]

$$\sigma^2_{\underline{\theta}} \approx \underline{e}_1^T \left[\mathbf{F}_Y + \alpha \mathbf{P}\right]^{-1} \mathbf{F}_Y \left[\mathbf{F}_Y + \alpha \mathbf{P}\right]^{-1} \underline{e}_1.$$

Comparison to (12) makes it evident that if $\mathbf{P}$ is identified with $\mathbf{D}^{-1}$ the variance approximation for PML-SAGE is identical to the lower bound on the variance. However, this does not guarantee that the bias-variance tradeoff curve is achieved unless the bias gradient matches the optimal bias-gradient expression (13). A local-asymptotic approximation to the bias of PML-SAGE can be derived in a similar manner as in [31] to yield the expression

$$b_{\underline{\theta}} \approx \underline{e}_1^T \left[\nabla^{11} J(\underline{\theta},\underline{\theta})\right]^{-1} \nabla^{01} J(\underline{\theta},\underline{\theta})$$

where

$$J(\underline{\theta},\underline{u}) = E_{\underline{u}}\left[\ln f_{\underline{Y}}(\mathrm{y};\underline{\theta}) - \alpha\, \underline{\theta}^T \mathbf{P}\underline{\theta}\right].$$

Since $\nabla^{11} J(\underline{\theta},\underline{\theta}) = -\nabla^{20} J(\underline{\theta},\underline{\theta}) = \mathbf{F}_Y + \alpha \mathbf{P}$ and $\nabla^{01} J(\underline{\theta},\underline{\theta}) = -\alpha \mathbf{P}\underline{\theta}$ we obtain

$$b_{\underline{\theta}} \approx -\underline{e}_1^T \left[\mathbf{P} + \frac{1}{\alpha}\mathbf{F}_Y\right]^{-1} \mathbf{P}\underline{\theta}.$$

Hence, taking the gradient

$$\nabla b_{\underline{\theta}} = -\underline{e}_1^T \left[\mathbf{P} + \frac{1}{\alpha}\mathbf{F}_Y\right]^{-1} \mathbf{P} + O\left(\frac{1}{\alpha}\right)$$

When we identify $\mathbf{P} = \mathbf{D}^{-1}$ and $\alpha = 1/\lambda$ we see that the bias gradient is identical to the (transpose) optimal bias gradient $\underline{d}_{min}$ in (13) up to order $O(1/\alpha)$. Therefore, as long as the above local variance and bias approximations are accurate, we can expect that the bias-variance tardeoff curve will be achieved by the PML-SAGE algorithm for high values of the smoothing parameter α.

To verify the local bias and variance performance predictions we simulated the PML-SAGE algorithm using $L = 400$ realizations of the estimator $\hat{\theta}$ to estimate the bias gradient and the variance via the methods described in Section 4. For each realization the average total number of detected photons $\sum_{i=1}^m \mu_i$ was one million. The PML-SAGE recursions were initialized by a uniform disk of intensity 1. We traced out the estimator bias-variance curve by varying the smoothing parameter α (Figure 5). The ellipsoidal confidence regions (95%) are not shown in the figures since they are smaller than the size of the plotting symbol '$*$'. In Figure 5 the PML-SAGE algorithms were terminated after 20 iterations for all values of α studied. Note that the bound, denoted by $B(\underline{\theta},\delta)$ in Figure 5, is virtually achieved for high bias and low variance, a region corresponding to overly smoothed image reconstructions. For $\alpha = 2^1$ and smaller, the curve 'B' tends to deviate from the lower bound and saturate i.e. lower α does not decrease the bias gradient. On the other hand, it was observed that the bias decreases to nearly zero. At points close to the knee of the curve, maximal reduction in bias is achieved at the price of significant increase in the variance.

In Figure 6 we have plotted the MSE against α, denoted 'A', along with the function $B(\underline{\theta},\delta) + b^2_{\underline{\theta}}$. Since this latter function is a bound on the MSE for the PML-SAGE estimator, it is useful as a validation of the correctness of our simulations. Observe that the smallest value of α gives minimum MSE which corresponds to a minimum amount of smoothing and, in view of Figure 5, to maximum image reconstruction variance. Since unsmoothed high variance images are unacceptable for most medical diagnostics, this reconfirms that minimizing MSE is not a suitable imaging objective.

Appendix: Proof of Theorem 1

Proof of Theorem 1: We perform constrained minimization of the biased form of the CR bound (6) over the feasible set of bias gradient vectors:

$$\mathrm{var}_{\underline{\theta}}(\hat{\theta}_1) \geq [\underline{e}_1 + \nabla^T b_{\underline{\theta}}]^T \mathbf{F}_Y^{+} [\underline{e}_1 + \nabla^T b_{\underline{\theta}}]$$

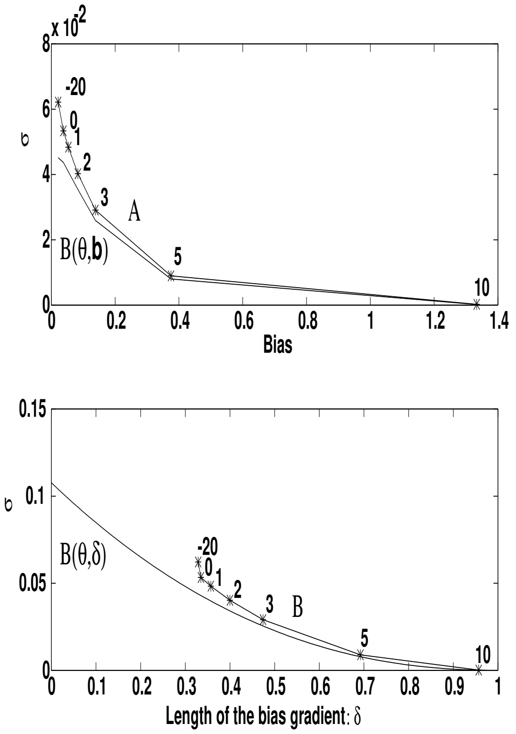

Figure 5: Effect of changing the smoothing parameter on the uniform bound. The numbers associated with the curves indicate the value of the smoothing parameter used ($\log_2(\alpha)$).

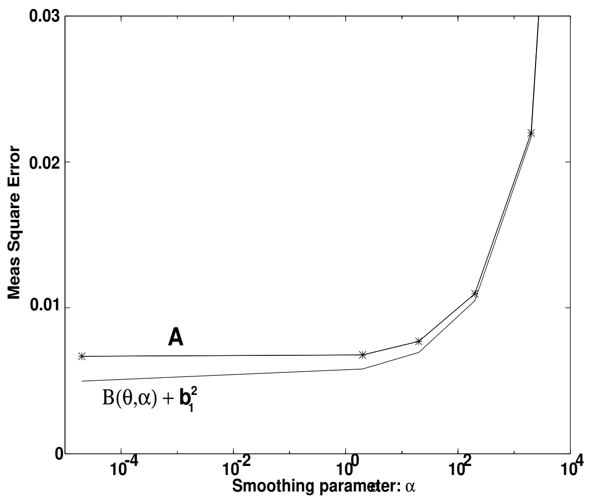

Figure 6: A bound on MSE and the estimator MSE performance.

$$\begin{aligned} &\geq \min_{\underline{d}:\|\underline{d}\|_D\leq\delta} [\underline{e}_1 + \underline{d}]^T \mathbf{F}_Y^+ [\underline{e}_1 + \underline{d}] \\ &= \min_{\underline{d}:\|\underline{d}\|_D\leq\delta} Q(\underline{d}), \end{aligned}$$

where $\underline{d}$ is a vector in $I\!R^n$. Defining $\tilde{\underline{d}} = \mathbf{D}^{\frac{1}{2}}\nabla^T b_{\underline{\theta}}$, where $\mathbf{D}^{\frac{1}{2}}$ is a square root factor of $\mathbf{D}$, this minimization is equivalent to:

$$\min_{\tilde{\underline{d}}:\|\tilde{\underline{d}}\|_2\leq\delta} \left[\mathbf{D}^{\frac{1}{2}}\underline{e}_1 + \tilde{\underline{d}}\right]^T \mathbf{B} [\mathbf{D}^{\frac{1}{2}}\underline{e}_1 + \tilde{\underline{d}}\right], \qquad (39)$$

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

First we consider the case where the unconstrained minimum $Q(\underline{d}) = 0$ occurs in the interior of the constraint set $\|\underline{d}\|_D \leq \delta$. From (39) it is clear that $Q(\underline{d})$ can be zero if and only if $\mathbf{D}^{\frac{1}{2}}\underline{e}_1 + \tilde{\underline{d}}$ lies in the null space of $\mathbf{B}$. Such a solution $\tilde{\underline{d}}$ must have the form:

$$\tilde{\underline{d}} = -\mathcal{P}_B\mathbf{D}^{\frac{1}{2}}\underline{e}_1 + \underline{\phi}$$

where $\underline{\phi}$ is an arbitrary vector in the null space of $\mathbf{B}$. But for $\tilde{\underline{d}}$ to be a feasible solution it must satisfy $\|\tilde{\underline{d}}\|_2 \leq \delta$ so that, by orthogonality of $\mathcal{P}_B\mathbf{D}^{\frac{1}{2}}\underline{e}_1$ and $\underline{\phi}$:

$$\delta^2 \geq \|\tilde{\underline{d}}\|_2^2 = \|\mathcal{P}_B\mathbf{D}^{\frac{1}{2}}\underline{e}_1\|_2^2 + \|\underline{\phi}\|_2^2 \geq \|\mathcal{P}_B\mathbf{D}^{\frac{1}{2}}\underline{e}_1\|_2^2.$$

We conclude that $\min_{\underline{d}:\|\underline{d}\|_D\leq\delta} Q(\underline{d}) = 0$ iff $\|\mathcal{P}_B\mathbf{D}^{\frac{1}{2}}\underline{e}_1\|_2^2 = \underline{e}_1^T\mathbf{D}^{\frac{T}{2}}\mathcal{P}_B\mathbf{D}^{\frac{1}{2}}\underline{e}_1 \leq \delta^2$.

If $\underline{e}_1^T\mathbf{D}^{\frac{T}{2}}\mathcal{P}_B\mathbf{D}^{\frac{1}{2}}\underline{e}_1 = 0$ then we have nothing left to prove. Otherwise, it remains to consider δ in the range: $0 \leq \delta^2 < \underline{e}_1^T\mathbf{D}^{\frac{T}{2}}\mathcal{P}_B\mathbf{D}^{\frac{1}{2}}\underline{e}_1$. In this case the minimizing $\tilde{\underline{d}}$ lies on the boundary and satisfies the equality constraint $\|\tilde{\underline{d}}\|_2 = \delta$. We thus need solve the unconstrained minimization of the Lagrangian:

$$\min_{\tilde{\underline{d}}} \left[\left[\mathbf{D}^{\frac{1}{2}}\underline{e}_1 + \tilde{\underline{d}}\right]^T \mathbf{B}[\mathbf{D}^{\frac{1}{2}}\underline{e}_1 + \tilde{\underline{d}}\right] + \lambda(\tilde{\underline{d}}^T\tilde{\underline{d}} - \delta^2)\right], \qquad (40)$$

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 (40) has the form:

$$\left[\tilde{\underline{d}} + (\lambda\mathbf{I}+\mathbf{B})^{-1}\mathbf{B}\mathbf{D}^{\frac{1}{2}}\underline{e}_1\right]^T (\lambda\mathbf{I}+\mathbf{B}) \left[\tilde{\underline{d}} + (\lambda\mathbf{I}+\mathbf{B})^{-1}\mathbf{B}\mathbf{D}^{\frac{1}{2}}\underline{e}_1\right]$$

$$+\underline{e}_1^T\mathbf{F}_Y^+\underline{e}_1 - \underline{e}_1^T\left[\mathbf{D}^{\frac{T}{2}}\mathbf{B}(\lambda\mathbf{I}+\mathbf{B})^{-1}\mathbf{B}\mathbf{D}^{\frac{1}{2}}\right]\underline{e}_1 - \lambda\delta^2 \qquad (41)$$

It follows immediately from (41) that

$$\tilde{\underline{d}} = \tilde{\underline{d}}_{min} = -(\lambda\mathbf{I}+\mathbf{B})^{-1}\mathbf{B}\mathbf{D}^{\frac{1}{2}}\underline{e}_1$$

achieves the minimum. Noting that $\underline{d}_{min} = \mathbf{D}^{-\frac{1}{2}}\tilde{\underline{d}}_{min}$, expressing $\mathbf{B}$ in terms of $\mathbf{D}$ and $\mathbf{F}_Y$, and performing simple algebraic manipulations, we obtain (10). Substituting the expression for $\tilde{\underline{d}}_{min}$ into (39):

$$\begin{aligned}\min_{\underline{d}:\|\underline{d}\|_D\leq\delta} Q(\underline{d}) &= [\underline{e}_1+\underline{d}_{min}]^T\mathbf{F}_Y^+[\underline{e}_1+\underline{d}_{min}] \\ &= \underline{e}_1^T\mathbf{D}^{\frac{T}{2}}\left[\mathbf{I}-[\lambda\mathbf{I}+\mathbf{B}]^{-1}\mathbf{B}\right]^T\mathbf{B}\left[\mathbf{I}-[\lambda\mathbf{I}+\mathbf{B}]^{-1}\mathbf{B}\right]\mathbf{D}^{\frac{1}{2}}\underline{e}_1 \\ &= \lambda^2\underline{e}_1^T\mathbf{D}^{\frac{T}{2}}[\lambda\mathbf{I}+\mathbf{B}]^{-1}\mathbf{B}[\lambda\mathbf{I}+\mathbf{B}]^{-1}\mathbf{D}^{\frac{1}{2}}\underline{e}_1\end{aligned}$$

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

In the above λ is determined by the equality constraint:

$$\delta^2 = \tilde{\underline{d}}_{min}^T\tilde{\underline{d}}_{min} = \underline{e}_1^T\left[\mathbf{D}^{\frac{T}{2}}\mathbf{B}(\lambda\mathbf{I}+\mathbf{B}]^{-2}\mathbf{B}\mathbf{D}^{\frac{1}{2}}\right]\underline{e}_1 = g(\lambda) \quad (42)$$

After simple matrix algebra we obtain the expression (11). By considering the eigendecomposition of $\mathbf{B}(\lambda\mathbf{I}+\mathbf{B})^{-2}\mathbf{B}^T$ it is easily seen that the function $g(\lambda)$ is continuous monotone decreasing over $\lambda\geq 0$. Furthermore $lim_{\lambda\to\infty}g(\lambda)=0$ and $lim_{\lambda\to 0}g(\lambda)=\underline{e}_1\mathbf{D}^{\frac{T}{2}}\mathcal{P}_B\mathbf{D}^{\frac{1}{2}}\underline{e}_1$. Therefore there exists a unique strictly positive λ such that $g(\lambda)=\delta^2$ for any value $\delta^2\in[0,\underline{e}_1\mathbf{D}^{\frac{T}{2}}\mathcal{P}_B\mathbf{D}^{\frac{1}{2}}\underline{e}_1)$. 2

References

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

[2] 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.

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

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

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

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

[14] 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.

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

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

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

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

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

[20] 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.

[21] Z. Liang, R. J. Jaszczak, and K. Greer, "On Bayesian image reconstruction from projections: Uniform and nonuniform a priori source information," *IEEE Trans. on Medical Imaging*, vol. MI-8, no. 3, pp. 227–235, 1989.

[22] 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.

[23] J. Philip, "The most ill-posed non-negative kernels in discrete deconvolution," *Inverse Problems*, vol. 3, no. 2, pp. 309–328, May 1987.

[24] 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.

[25] F. A. Graybill, *Matrices with Applications in Statistics*, Wadsworth Publishing Co., Belmont CA, 1983.

[26] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge, 1985.

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

[28] 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.

[29] 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.

[30] J. A. Fessler, "Mean and variance of implicitly defined estimators (such as penalized maximum likelihood): applications to tomography," *IEEE Trans. on Image Processing*, to appear 1995.

[31] A. O. Hero, "Timing estimation for a filtered Poisson process in Gaussian noise," *IEEE Trans. on Inform. Theory*, vol. 37, no. 1, pp. 92–106, Jan. 1991.