Uniform CR Bound: Implementation Issues And Applications

M. Usman, A.O. Hero and J.A. Fessler
University of Michigan

Abstract

We apply a uniform Cramer-Rao (CR) bound [1] to study the bias-variance trade-offs in single photon emission computed tomography (SPECT) image reconstruction. The uniform CR bound is used to specify achievable and unachievable regions in the bias-variance trade-off plane. The image reconstruction algorithms considered in this paper are: 1) Space alternating generalized EM and 2) penalized weighted least-squares.

I. Introduction

The mean-square error (MSE) is an important measure of precision of a scalar component $\widehat{\theta_1}$ of an estimator $\widehat{\underline{\theta}}$. It is well known that the MSE is a function of both the bias, denoted $\text{bias}_{\underline{\theta}}(\widehat{\theta_1})$ and the variance, denoted $\text{var}_{\underline{\theta}}(\widehat{\theta_1})$ of the scalar estimator:

$$\text{MSE}_{\underline{\theta}}(\widehat{\theta_1}) = \text{var}_{\underline{\theta}}(\widehat{\theta_1}) + \text{bias}^2_{\underline{\theta}}(\widehat{\theta_1}).$$

Obviously increases in MSE can be due to increases in either the bias or variance of $\widehat{\theta_1}$. Bias and variance are complementary in nature. While bias is due to 'mismatch' between the average value of the estimator and the true parameter, variance is due to statistical fluctuations in the estimator. There usually exists a tradeoff between bias and variance of the estimated parameter. For example in image reconstruction, implementation of the maximum likelihood algorithm with a smoothness penalty reduces the variance only at the expense of introducing bias. Different estimators can be effectively compared by plotting their performance on a bias-variance trade-off plane. The classical or the unbiased CR bound has been previously applied to compare different estimators [2, 3]. However, in most image processing applications the estimators are biased and their variance is not bounded by the unbiased CR bound. For biased estimators a biased CR bound is available [4] which is only applicable to estimators with fixed bias gradient $\nabla_{\underline{\theta}}\text{bias}_{\underline{\theta}}(\widehat{\theta_1})$, hence it is unable to give a meaningful comparison of different biased estimators that have acceptable bias but different bias gradients. We use uniform CR bound [1] on the variance of biased estimators which divides the bias-variance trade-off plane δ–σ into achievable and unachievable regions. Different estimators can be placed in the achievable region of the δ–σ plane and their performance can be effectively compared.

This work was supported in part by National Science Foundation under grant BCS-9024370 and a Government of Pakistan Postgraduate Fellowship.

II. Unbiased CR bound

Consider the problem of estimation of an n-dimensional parameter $\underline{\theta} = [\theta_1, ..., \theta_n]^T$ given an observation of a vector of random variables $\underline{Y}$ with probability density function (pdf) $f_{\underline{Y}}(\underline{y}; \underline{\theta})$.The Cramer-Rao lower bound on the variance of unbiased parameter estimator $\widehat{\theta_1}$ is given by the upper-left (1,1) element of the inverse of an $n \times n$, symmetric, positive definite Fisher information matrix (FIM) $F_Y = F_Y(\underline{\theta})$:

$$\text{var}_{\underline{\theta}}(\hat{\theta}_1) \geq \underline{e}_1^T \; F_Y^{-1} \; \underline{e}_1, \tag{1}$$

where,

$$F_Y = E_{\underline{\theta}}[\nabla^T_{\underline{\theta}} \ln f_{\underline{Y}}(\underline{Y}; \underline{\theta}) \; \nabla_{\underline{\theta}} \ln f_{\underline{Y}}(\underline{Y}; \underline{\theta})],$$

$\nabla_{\underline{\theta}}$ denotes the (row) gradient vector $[\frac{\partial}{\partial\theta_1}, ..., \frac{\partial}{\partial\theta_n}]$, and $\underline{e}_1 = [1, 0, ..., 0]^T$ is an n-element unit vector.

While the unbiased CR bound (1) is known to be asymptotically achievable for large number of independent identically distributed measurements, in practice, most estimation algorithms are biased and the unbiased CR bound is inapplicable.

III. Uniform CR bound

For a biased estimator $\widehat{\theta_1}$ the following form of the biased CR bound is well known [4]:

$$\text{var}_{\underline{\theta}}(\widehat{\theta_1}) \geq [\nabla_{\underline{\theta}} m_1] \; F_Y^{-1} \; [\nabla_{\underline{\theta}} m_1]^T, \tag{2}$$

where $\nabla_{\underline{\theta}} m_1 = \nabla_{\underline{\theta}} m_1(\underline{\theta}) = \nabla_{\underline{\theta}} b_1 + \underline{e}_1$ is an n element row vector of the gradient of the mean $E_{\underline{\theta}}(\widehat{\theta_1}) = m_1(\underline{\theta})$. The application of the biased CR bound (2) is very restricted due to the fact that it is only applicable to estimators with a given bias gradient $\nabla_{\underline{\theta}} b_1$. In [1] Hero gives a 'uniform' CR bound on the variance of a single parameter θ_1 for non-singular F_Y. This bound is applicable to all biased estimators whose bias gradient length $||\nabla_{\underline{\theta}} b_1||$ satisfies:

$$||\nabla_{\underline{\theta}} b_1||^2 \leq \delta^2 < 1. \tag{3}$$

The following theorem is proven in [1].

Theorem 1 *Let $\widehat{\theta}_1$ be an estimator with bias $b_1(\underline{\theta})$ whose n-element bias gradient vector $\nabla_{\underline{\theta}} b_1$ satisfies (3). Assume that the FIM F_Y is non-singular. Then the variance of $\widehat{\theta}_1$ is given by:*

$$var_{\underline{\theta}}(\widehat{\theta}_1) \geq B(\underline{\theta}, \delta), \quad (4)$$

where $B(\underline{\theta}, \delta)$ is equal to:

$$\begin{aligned} B(\underline{\theta}, \delta) &= [\underline{e}_1 + \underline{d}_{min}]^T F_Y^{-1} [\underline{e}_1 + \underline{d}_{min}], & (5) \\ &= \lambda^2 \underline{e}_1^T \left[I + \lambda F_Y\right]^{-1} F_Y \left[I + \lambda F_Y\right]^{-1} \underline{e}_1 & (6) \end{aligned}$$

where $\underline{e}_1 = [1, 0, ..., 0]^T$ is an n-element unit vector and:

$$\underline{d}_{min} = -[I + \lambda\, F_Y]^{-1} \underline{e}_1, \quad (7)$$

and λ is given by the unique non-negative solution of the following equation involving the monotone decreasing, strictly convex function $g(\lambda) \in [0, 1]$:

$$g(\lambda) = \underline{d}_{min}^T \underline{d}_{min} = \delta^2 \qquad \lambda \geq 0. \quad (8)$$

A more general version of Theorem 1, which will not be required here, is given in [5] and applies to singular F_Y. Note that since $\lambda \geq 0$ and $F_Y \geq 0$, the use of the expression (6) does not suffer from any ill-conditioning of the FIM F_Y. In Theorem 1, $\underline{d}_{min}$ defined in (7) is an optimal bias gradient in the sense that it minimizes the biased CR bound (2) over all vectors $\nabla_{\underline{\theta}} b_1$.

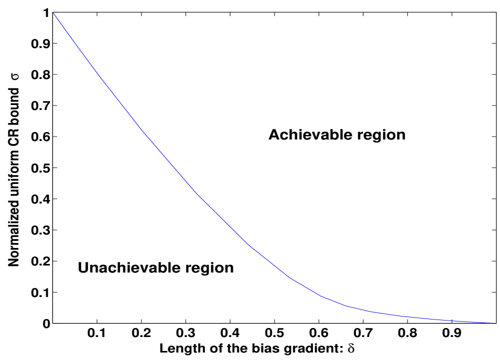

Figure 1: The Normalized Uniform CR bound.

Figure 1 shows a typical bias-variance trade-off curve in the δ–σ plane. The region above and including the curve is the so called 'achievable' region where all the realizable estimators exist. Note that if an estimator lies on the curve then lower variance can only be bought at the price of increased bias and vice versa. At $\delta = 1$ the variance goes to zero. This corresponds to the trivial case $\widehat{\theta}_1$=Constant for which $\nabla_{\underline{\theta}} b_1 = \underline{e}_1$.

Estimation of the Bias Gradient

To compare a particular estimator to the uniform bound of Theorem 1 we require the length of the estimator bias gradient so that the estimator can be placed somewhere within the achievable region of Figure 1. In most cases the bias and the bias-gradient are analytically intractable. The method of moments is the standard method for experimentally determining bias and covariance of $\widehat{\underline{\theta}}$ which is based on forming the sample mean and sample covariance statistics for a sequence of L repeated experiments $\{\underline{Y}_i\}_{i=1}^L$ each generated from the density $f_{\underline{Y}}(\underline{y}_i; \underline{\theta})$. The method of moments for estimating the bias-gradient would require n additional sequences of L repeated experiments, each generated for a particular perturbation of a different component of the parameter vector $\underline{\theta}$. Such a direct approach is impractical. In [5] a method for experimentally determining the bias-gradient of an estimator $\widehat{\underline{\theta}}$ is presented that requires a single simulation of the same type as that commonly used to determine bias and covariance of $\widehat{\underline{\theta}}$. The unbiased estimate of the bias gradient for the estimate of $\widehat{\theta}_1$ is given by [5]: $\widehat{\nabla_{\underline{\theta}} b_1} =$

$$\frac{1}{L-1} \sum_{i=1}^{L} \left(\widehat{\theta}_1(\underline{Y}_i) - \frac{1}{L} \sum_{j=1}^{L} \widehat{\theta}_1(\underline{Y}_j) \right) \nabla_{\underline{\theta}} \ln f_{\underline{Y}}(\underline{Y}_i; \underline{\theta}) - \underline{e}_1^T. \quad (9)$$

A few comments about the bias gradient are in order. The bias gradient $\nabla_{\underline{\theta}} b_1$ is a measure of the influence of each component parameter $\theta_1, ..., \theta_n$ on the mean $m_1(\underline{\theta})$ of the estimator $\widehat{\theta}_1$. Ideally, to be close to unbiased one would like $m_1(\underline{\theta})$ to be insensitive to the variations in the other parameters $\theta_2, ..., \theta_n$. Alternatively, since $b_1(\underline{\theta}) = m_1(\underline{\theta}) - \theta_1$, it is desirable that the components $\frac{\partial}{\partial \theta_k} b_1(\underline{\theta})$ be of small magnitude, $k = 2, ..., n$. The bias gradient therefore provides important information about the parameter coupling to the estimator mean. The bias gradient is in general only indirectly related to the estimator bias, with the exception that $\nabla_{\underline{\theta}} b_1 = 0$ implies $b_1(\underline{\theta}) =$ constant. An estimator that has a constant bias independent of $\underline{\theta}$ is removable, and therefore $\nabla_{\underline{\theta}} b_1 = 0$ implies that the estimation can be performed without bias. Conversely, a non-zero bias gradient implies non-removable estimator bias that is dependent on the estimator parameters. On the other hand, one can have a large bias gradient even though the bias is very small. Therefore the bias and the bias gradient together give a more complete picture of estimator behavior.

Bias-Variance Trade-Off Plane

When accurate estimates $\widehat{b}_1$, $\widehat{\nabla_{\underline{\theta}} b_1}$ and $\widehat{\sigma}^2$ of the estimator bias, bias gradient, and variance are available for a given estimator $\widehat{\theta}_1$ of θ_1, the uniform CR bound lying in the δ–σ plane can be easily mapped into the b–σ plane of variance and biases. This is accomplished by using the ordered triplet $(\widehat{b}_1, \widehat{\nabla_{\underline{\theta}} b_1}, \widehat{\sigma}^2)$ as a mapping between the δ–σ and the b–σ planes. The uniform CR bound on the variance as a function of bias is simply the ordered pair: $\left(\widehat{b}_1, \left[\underline{e}_1 + \widehat{\nabla_{\underline{\theta}}^T b_1}\right]^T F_Y^+ \left[\underline{e}_1 + \widehat{\nabla_{\underline{\theta}}^T b_1}\right]\right)$, denoted $B(\underline{\theta}; b)$ in the sequel.

IV. Applications

We will apply the uniform CR bound to study the bias-variance trade-offs for: 1) Space alternating generalized expectation-maximization (SAGE) algorithm and 2) penalized weighted least-squares estimator (WLSE).

SPECT Image Reconstruction

System Description

The system used in this paper is shown in Figure 2 and is called the SPRINT II system [6]. The system was designed specifically for brain imaging and consists of a ring of detectors and a ring of collimators. The function of the collimator is to reduce the uncertainty associated with the emission location of a γ-ray to a line or a strip in the field of view (Figure 2). During imaging time, the collimator ring is rotated through small steps about the source. A γ-ray photon passing through one of the collimator slits at one of the rotation angles is counted as an event acquired in one 'detector bin'. For reconstruction the source domain is divided into n small regions, called pixels. The detection process is governed by Poisson statistics: $\underline{Y} = [Y_1, ..., Y_d]^T$.

$$f_{\underline{Y}}(\underline{y}; \underline{\theta}) = \prod_{j=1}^{d} \frac{\mu_j^{Y_j}}{Y_j!} e^{\mu_j}. \qquad (10)$$

In (10) θ_i is the average γ-ray intensity of the i-th pixel; $i = 1, ...p$, Y_j is number of γ-rays detected at the j-th detector, and μ_j is the average γ-ray intensity of the j-th detector; $j = 1, ..., d$: $\underline{\mu} = A\,\underline{\theta}$, where A is the $d \times p$ system matrix that depends on the tomographic geometry.

The objective is to reconstruct the object intensity of each pixel $\underline{\theta} = [\theta_1, ..., \theta_n]^T$ given the set of observations $\underline{Y}$. It can be easily shown that the FIM is of the form:

$$F_Y(\underline{\theta}) = A^T\,[\mathrm{diag}(\mu)]^{-1}\,A \qquad (11)$$

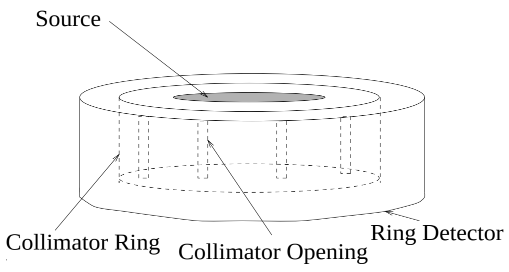

Figure 2: The SPRINT II system. Not drawn to scale.

The system parameters are given in Appendix A and unless otherwise specified are those used in the simulations.

In the following simulations the effect of attenuation was neglected. The total number of detected γ-ray counts were 10^9. Noise due to scatter were 5% of the total counts. Since the algorithm considered in this section is non-linear, an analytic expression for the bias gradient is intractable, and therefore the bias gradient was estimated using (9).

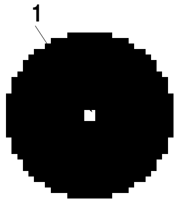

Figure 3: The object used in the simulations. The object dimensions are 32×32. The black pixels are of intensity 1 while the white pixels are of intensity 2.

We used $L = 400$ realizations of the projection data $\underline{Y}$ to ensure the statistical accuracy of our estimator bias, bias gradient and variance. The object is a disk of uniform intensity 1 with a high intensity region of 4 pixels in the center of uniform intensity 2, called the hot spot. The pixel of interest was the pixel at the upper edge of the hot spot, marked '1'. The diameter of the disk is 32 pixels. In the following simulation, the algorithm was initialized by a uniform disk of intensity 1 and diameter 32 pixels.

Space Alternating Generalized EM

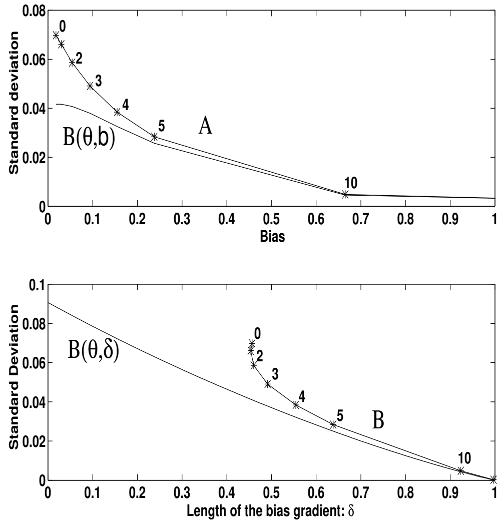

Figure 4: Performance of PML-SAGE as a function of α.

The iterative space alternating generalized expectation-maximization (PML-SAGE) algorithm maximizes a penalized likelihood objective function of the form:

$$\widehat{\underline{\theta}}(\underline{y}) = \underset{\underline{\theta} \in \Theta}{\mathrm{argmax}} \left\{ \ln f_{\underline{Y}}(\underline{y}; \underline{\theta}) - \alpha P(\underline{\theta}) \right\},$$

where $P(\underline{\theta})$ is the penalty function and α is the smoothing parameter. Setting $\alpha = 0$ corresponds to no image smoothing while a large value of α corresponds to a significant amount of smoothing. PML-SAGE uses an intelligent choice of 'hidden data spaces' such that the E and

M steps are analytically tractable. A detailed description of the PML-SAGE algorithm is given in [7].

It is easy to show that for the Poisson model

$$\nabla_{\underline{\theta}} \ln f_{\underline{Y}}(\underline{y};\underline{\theta}) = A^T \left[-\underline{1} + \underline{y} \oslash \underline{\mu}\right],$$

where $\oslash$ is a vector operation denoting element-by-element division, and $\underline{1} = [1, 1, ..., 1]^T$.

For the first set of simulations the smoothing parameter α was varied (Figure 4). Points on the curves in Figures 4 are labeled by the exponent of α. The bias, bias gradient and variance were estimated and the uniform bound was plotted over the bias gradient length-variance trade-off plane, denoted δ–σ, and the bias-variance trade-off plane, denoted b–σ. The PML-SAGE algorithms were terminated after 100 iterations for each of the $L = 400$ trials. The ellipsoidal confidence regions are not shown in the figure since they are smaller than the size of the plotting symbol '$*$'. Note that the bound, denoted by $B(\underline{\theta};\delta)$ in Figure 4, is achieved for large biases, i.e. large α. For α small, the curve 'B' tends to deviate more from the lower bound and saturate, i.e. lower α does not decrease the bias gradient. On the other hand the bias decreases to an asymptote near zero.

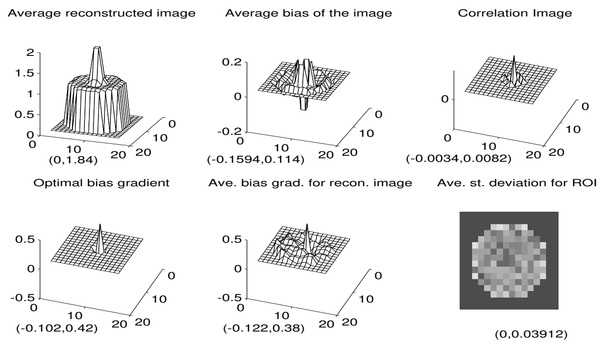

Figure 5: PML-SAGE: different image quantities of interest. An ordered pair with each curve indicates the (minimum, maximum) value associated with that image. The images in (a) and (b), from top left to bottom right, are: Average reconstructed image, average bias of the reconstructed image, correlation image, optimal bias gradient d_{min}, average bias gradient for the reconstructed image, and average standard deviation.

At points close to the unbiased point, i.e. the leftmost corner of the horizontal axis, in curve 'A', maximal reduction in bias is achieved at the price of significant increase in the variance.

Figures 5 and 6 show several graphs of reconstruction quantities for $\alpha = 2^4$, and $\alpha = 2^{10}$, respectively. For clarity in the figures, we down-sampled all the images by a factor of 2. For each image in Figures 5 and 6 the ordered pair at bottom indicates the minimum and maximum values for that image. In Figure 5, the mean reconstructed image is very close to the true image except around the

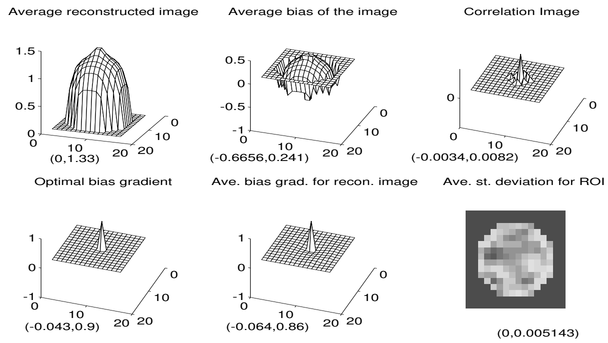

Figure 6: PML-SAGE: different graphs of reconstruction quantities for $\log_2(\alpha) = 10$.

edges. The correlation image, i.e. the column of F_Y^{-1} corresponding to the pixel of interest, θ_{ROI}, shows a strong correlation with the neighboring pixels. This implies that to estimate θ_{ROI} we must also estimate the strongly correlated neighboring pixels accurately, while the influence of the far pixels can be ignored. Ideally, one would like the correlation between the pixels to be zero so that the estimate of a certain pixel, θ_{ROI}, is independent of the estimates of all other pixels. The plot for the theoretically optimal bias gradient $\underline{d}_{min}$ shows a similar strong influence from the neighboring pixels.

The average bias gradient $\nabla_{\underline{\theta}} b_1$ for the reconstructed image is different from the theoretically optimal bias gradient $\underline{d}_{min}$. Thus the PML-SAGE image reconstruction algorithm does not take best advantage of its bias allocation since it is only by using the optimal bias gradient $\underline{d}_{min}$ given by (7) that the minimum bias length is achieved.

Figure 6 shows the same set of images as in Figure 5 but for $\alpha = 2^{10}$. Due to very high regularization, the hot spot is almost entirely smoothed out. Also, neither $\underline{d}_{min}$ nor the average bias gradient $\nabla_{\underline{\theta}} b_1$ for the reconstructed image show significant coupling between the pixel of interest and the neighboring pixels. This is to be expected since in the overly smoothed case the bias is principally determined by the smoothness penalty as opposed to the projection data.

Weighted Least-Squares Estimator

Similar to the PML-SAGE, the WLSE is penalized for roughness, but minimized over a quadratic objective function. The WLSE is given by [8]:

$$\widehat{\underline{\theta}}(\underline{y}) = \underset{\underline{\theta} \in \Theta,\ \underline{\theta} \geq 0}{\operatorname{argmin}} \left\{ \frac{1}{2} \left[(\widehat{\underline{y}} - A\underline{\theta})^T \Sigma^{-1} (\widehat{\underline{y}} - A\underline{\theta}) \right] + \alpha P(\underline{\theta}) \right\},$$

where Σ is a weight matrix, $P(\underline{\theta})$ is a regularization penalty, and A is the system matrix. We use a penalty function described in [8] which is imposed on the 8 neighboring pixels for each pixel of interest. The weight matrix Σ is diagonal, consisting of the covariance estimate of the observations. It is shown in [5] that a WLSE with an identity penalty function and ideal weight matrix

$\Sigma = \mathrm{diag}_i(\mu_i)$ exactly achieves the uniform CR bound for all biases.

Figure 7 shows the δ–σ and b–σ plots for the WLSE. The WLSE estimator follows the uniform CR bound closely for high bias and low variance, but tends to deviate away from the bound An interesting point to note is that both the PML-SAGE and the WLSE have similar bias-variance trade-off curves. However, the uniform bound on bias $B(\underline{\theta}, b)$ is different for PML-SAGE than that for WLSE since the bound on bias is indexed by algorithm bias gradient which is obviously algorithm dependent.

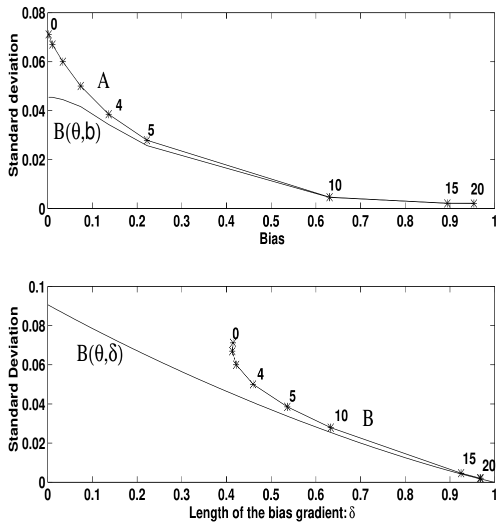

Figure 7: Performance of WLSE as a function of α.

Figures 8 and 9 show several graphs of reconstruction quantities using the WLSE for $\alpha = 2^4$ and $\alpha = 2^{10}$. The comments for Figures 5 and 6 are valid here. The only exception being that the WLSE fails to accurately estimate the edges for small α. This is due to the fact that the estimates of covariance involving the projections that graze the image edges are less accurate.

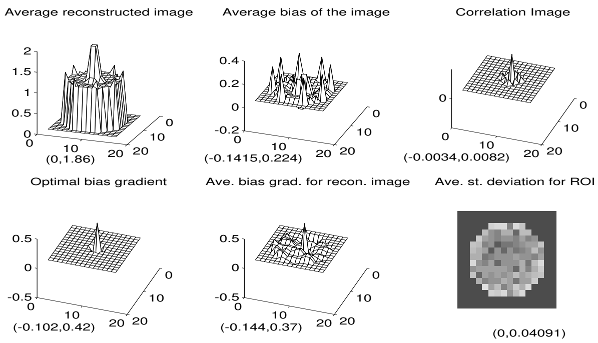

Figure 8: WLSE: different graphs of reconstruction quantities for $\log_2(\alpha) = 4$.

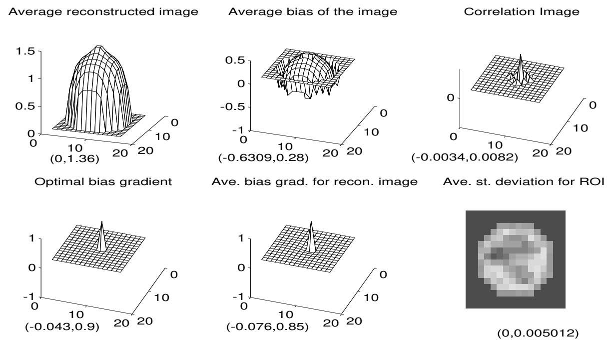

Figure 9: WLSE: different graphs of reconstruction quantities for $\log_2(\alpha) = 10$.

I. System specifications

Radius of the detector ring	25 cms
Number of detectors	512
Radius of the collimator ring	17 cms
Number of collimator slits	10 (uniformly spaced)
Slit Width	2.4 mm

References

[1] A. O. Hero
A Cramer-Rao Type Lower Bound for Essentially Unbiased parameter Estimation, Technical Report 890, Lincoln Laboratory, MIT 1992.

[2] M. Usman, A. O. Hero and W. L. Rogers
Performance Gain Analysis for Adding Vertex View to a Standard SPECT, MWSCS, August 1993, Detroit, MI.

[3] R. Kakarala and A.O. Hero
On Achievable Accuracy in Edge Localization, IEEE Transactions on Pattern Analysis and Machine Intelligence, 14:7, pp. 777-78 1, July, 1992.

[4] H.L. van Trees
Detection, Estimation and Modulation Theory, (Part I), John Wiley and Sons. 1968.

[5] M. Usman, A.O. Hero, and J.A. Fessler
Bias-Variance Trade-offs For Parametric Estimation Problems Using Uniform CR Bound, To be submitted to IEEE Transactions on Signal Processing.

[6] W.L. Rogers, N.H. Clinthorne, L.Shao, P.Chiao, J.Stamos, and K.F.Koral
SPRINT II, A Second Generation Single Photon Ring Tomograph, IEEE Transactions on Medical Imaging, 7:4, pp. 291-297, 1988.

[7] J.A. Fessler and A.O. Hero
Space-Alternating Generalized Expectation-Maximization Algorithm, To appear in IEEE Transactions on Signal Processing, Oct., 1994.

[8] J. A. Fessler
Penalized Weighted Least-Squares Image Reconstruction for Positron Emission Tomography, IEEE Transactions on Medical Imaging, 13:2, June 1994, In Press.