Achievable Regions in the Bias-Variance Plane for Parametric Estimation Problems [1]

Alfred Hero* and Mohamed Usman**

*Dept. of EECS, The University of Michigan, Ann Arbor, Michigan, USA

**Punjab Institute of Computer Science, Lahore, Pakistan

Abstract — **In this paper we use the uniform Cramer-Rao (CR) lower bound [1] to generate bias-variance tradeoff curves which separate achievable from unachievable regions in the estimator bias variance plane.**

I. Introduction

Let $\underline{\theta} = [\theta_1, ..., \theta_n]^T \in \Theta$ be a vector of unknown parameters which parameterize the distribution of an observed random variable $\mathbf{Y}$. Let $\hat{\theta}_1 = \hat{\theta}_1(\mathbf{Y})$ be an estimator of the scalar θ_1 and define the estimator bias function $b_1 = b_1(\underline{\theta}) = E_{\underline{\theta}}[\hat{\theta}_1] - \theta_1$ and the variance function $\sigma^2 = \sigma^2(\underline{\theta}) = E_{\underline{\theta}}[(\hat{\theta}_1 - \theta_1)^2]$. The goal of this work is to quantify fundamental tradeoffs between the bias and variance functions for any parametric estimation problem. When considered as surfaces over the parameter space Θ, the bias and variance provide a very informative description of estimator performance, for example they jointly specify the MSE.However, since comparison of performance surfaces over a large set Θ is usually impractical, the bias and variance in a small neighborhood is of greater interest. In this case, the bias gradient $\nabla_{\underline{\theta}} b_1$ is more useful since it is insensitive to constant and hence removable biases. It can be shown that $\nabla_{\underline{\theta}} b_1$ is directly related to the width of the point spread function for penalized maximum likelihood deconvolution problems [2]. The weighted norm of the bias gradient is directly related to the variation of the bias function over Θ by: $|\Delta b_1(\underline{\theta})| = \|\nabla_{\underline{\theta}} b_1\|_D + o(det|D|)$, where $\|\underline{u}\|_D^2 = \underline{u}^T D^T D \underline{u}$ and D is an invertible matrix whose determinant is inversely proportional to the volume of the region.

II. The Bias-Variance Tradeoff Curve

The tradeoff curve is derived from a generalization of the bound on estimator variance presented in [1]. Unlike the bound of [1], this bound applies to the case of singular Fisher information matrices (FIM), an important case arising in deconvolution problems, and permits use of any weighted l_2 norm of the bias gradient.

Theorem 1 *For a fixed scalar $\delta \in [0,1]$ let $\hat{\theta}_1$ be an estimator whose bias gradient satisfies the norm constraint $\|\nabla_{\underline{\theta}} b_1\|_D^2 = \underline{u}^T D^T D \underline{u} \le \delta^2$, where D is an arbitrary non-singular matrix. Define the oblique projection operator ($n \times n$ matrix) $\mathcal{P}_{\mathcal{F}_Y} = \mathcal{F}_Y[\mathcal{F}_Y \mathcal{D}^T \mathcal{D} \mathcal{F}_Y]^+ \mathcal{F}_Y \mathcal{D}^T \mathcal{D}$ which maps n-dimensional space onto the column space of the FIM F_Y, and define the n-element unit vector $\underline{e}_1 = [1, 0, ..., 0]^T$. Then the variance of $\hat{\theta}_1$ satisfies:*

$$var_{\underline{\theta}}(\hat{\theta}_1) \ge B(\underline{\theta}, \delta), \qquad (1)$$

where if $\|\mathcal{P}_{\mathcal{F}_Y} \underline{\ }_{\infty}\|_{\mathcal{D}} \le \delta$ then $B(\underline{\theta}, \delta) = 0$, while if $\|\mathcal{P}_{\mathcal{F}_Y} \underline{\ }_{\infty}\|_{\mathcal{D}} > \delta$ then:

$$B(\underline{\theta}, \delta) = \underline{e}_1^T F_Y^+ \underline{e}_1 - \underline{e}_1^T \left[F_Y^+ (\lambda \cdot D^T D + F_Y^+)^{-1} F_Y^+ \right] \underline{e}_1 - \lambda \delta^2 \qquad (2)$$

In (2) $\lambda > 0$ is determined by the unique non-negative solution of the following equation:

$$g(\lambda) = \underline{e}_1^T \left[F_Y^+ \left(\lambda \cdot D^T D + F_Y \right)^{-2} F_Y^+ \right] \underline{e}_1 = \delta^2. \qquad (3)$$

By calculating the family of points $\{(B(\underline{\theta}, \delta), \delta) : \delta \in [0,1]\}$ we sweep out a curve in the *bias-variance plane* which lower bounds any estimator plotted in the plane. Figure 1 illustrates this curve for a simple one dimensional Gaussian deconvolution problem and the unweighted l_2 norm (D=identity) [2]. The region above and including the curve is the so called 'achievable' region where all the realizable pairs of estimator variance and bias-gradients 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. For this example the regularized least squares estimator attains optimal bias-variance tradeoff, i.e. it hits the lower bound for all values of δ [2]. In this case the bias gradient norm δ was swept out by varying the smoothing (regularization) parameter of the estimator.

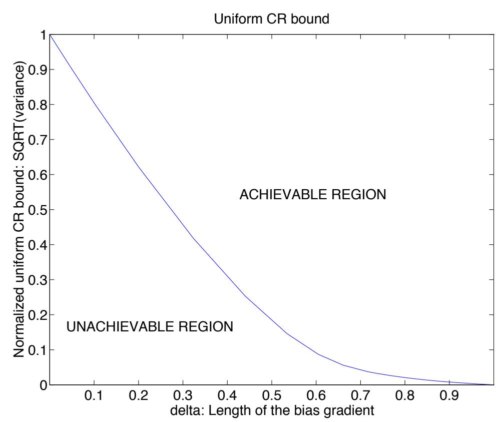

Figure 1: Bias-Variance Plane and Lower Bound.

In general to place an estimator somewhere within the achievable region of Figure 1 requires the variance and length of the estimator bias gradient. In most cases the variance and the bias-gradient length are analytically intractable and must be empirically estimated. Since the sample mean estimate of the bias gradient norm has severe positive bias some form of bias correction is necessary. We have developed a bootstrap estimator and a $(1-\alpha)\%$ lower confidence bound for this purpose.

References

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

[1] This work was partially supported by NSF Grant BCS-9024370

[2] M. Usman and A.O. Hero, "Bias-variance tradeoffs for parametric estimation problems using the uniform CR bound," in revision for publication in *IEEE Trans. on Signal Processing*.