Minimum Variance in Biased Estimation: Bounds and Asymptotically Optimal Estimators

Author
Yonina C. Eldar

Presented by
Aristeidis Daskalopoulos (10640)
Georgios Rousomanis (10703)

IEEE Transactions on Signal Processing, July 2004

Motivation and Background

- A common approach to developing well-behaved estimators in overparameterized estimation problems is to use *regularization techniques*.
- Regularization reduces variance but introduces bias.
- Cramér–Rao Lower Bound (CRLB) assumes unbiased estimators.
- Need for bounds applicable to biased estimators.

Key Goals of the Paper

- Develop Uniform Cramér–Rao Lower Bound (UCRLB) for biased estimators.
- Use Frobenius and spectral norms of the bias gradient matrix.
- Construct estimators (Tikhonov, shrunken, PML) that achieve the bounds.

Classical and Biased CRLB

- Unbiased CRLB: $\text{Var}(\hat{\mathbf{x}}) \geq \mathbf{J}^{-1}$
- Biased CRLB:

$$\begin{aligned}
\mathbf{b}(\mathbf{x}_0) &= \mathbb{E}[\hat{\mathbf{x}}] - \mathbf{x}_0 \\
\text{Cov}(\hat{\mathbf{x}}) &\geq (\mathbf{I} + \mathbf{D})\mathbf{J}^{-1}(\mathbf{I} + \mathbf{D})^* \triangleq \mathbf{C}(\mathbf{D}) \\
\mathbf{D} &= \frac{\partial \mathbf{b}(\mathbf{x}_0)}{\partial \mathbf{x}}
\end{aligned}$$

- Biased CRLB does not depend directly on the bias but only on the bias gradient matrix!
- D is invariant to a constant bias term so that in effect, it characterizes the part of the bias that cannot be removed

Bias Gradient Matrix

- Given a desired bias gradient, the biased CRLB serves as a bound on the smallest attainable variance.
- How to choose $\mathbf{D}$?

Bias Gradient Matrix

- Given a desired bias gradient, the biased CRLB serves as a bound on the smallest attainable variance.
- How to choose $\mathbf{D}$?
- Instead, use norms of $\mathbf{D}$ to constrain variance
  - Frobenius norm (average bias)
  - Spectral norm (worst-case bias)
- Uniform CRLB (UCRLB): is a bound on the smallest attainable variance that can be achieved using any estimator with bias gradient whose norm is bounded by a constant.

Bias Gradient Matrix (cont.)

- Generally, minimizing the bias results in an increase in variance and vice versa.
- Tradeoff between bias and variance
- Minimize $\mathrm{Tr}[\mathbf{C}(\mathbf{D})]$ subject to some constraint on $\mathbf{D}$.
- How to develop a meaningful constraint on $\mathbf{D}$?

Bias Gradient Matrix (cont.)

- Generally, minimizing the bias results in an increase in variance and vice versa.
- Tradeoff between bias and variance
- Minimize $\mathrm{Tr}[\mathbf{C}(\mathbf{D})]$ subject to some constraint on $\mathbf{D}$.
- How to develop a meaningful constraint on $\mathbf{D}$?
  - Worst Case Bias Constraint:

$$D_{WC} = \max_{\mathbf{z}\in\mathbb{C}^m,\, ||\mathbf{z}||=1} \mathbf{z}^*\mathbf{S}\mathbf{D}^*\mathbf{D}\mathbf{S}\mathbf{z}, \quad S \geq 0$$

  - Average Bias Constraint:

$$D_{AVG} = \mathrm{Tr}(\mathbf{D}^*\mathbf{D}\mathbf{W}), \quad \mathbf{W} \geq 0$$

UCRLB with Average Bias Constraint

- Minimize total variance $C(\mathbf{D}) = \mathrm{Tr}((\mathbf{I}+\mathbf{D})\mathbf{J}^{-1}(\mathbf{I}+\mathbf{D})^*)$
- Subject to average bias constraint:

$$D_{\mathrm{AVG}} = \mathrm{Tr}(\mathbf{D}^*\mathbf{D}\mathbf{W}) \leq \gamma$$

- $\mathbf{W}$ is a non-negative Hermitian weighting matrix
- $\mathbf{J}$ is the Fisher information matrix
- $\mathbf{D}$ is the bias gradient matrix

Key Insight: The norm of the bias gradient matrix measures sensitivity of bias to parameter changes

Theorem 1: UCRLB with Average Bias

Theorem (Theorem 1)

For $\gamma < Tr(\mathbf{W})$, the total variance C of any estimator with $Tr(\mathbf{D}^*\mathbf{D}\mathbf{W}) \leq \gamma$ satisfies:

$$C \geq \alpha^2 Tr((\mathbf{I} + \alpha\mathbf{W}\mathbf{J})^{-1}\mathbf{W}\mathbf{J}^{-1}\mathbf{W}(\mathbf{I} + \alpha\mathbf{J}\mathbf{W})^{-1})$$

where $\alpha > 0$ is chosen such that:

$$Tr((\mathbf{I} + \alpha\mathbf{W}\mathbf{J})^{-1}\mathbf{W}\mathbf{J}^{-1}(\mathbf{I} + \alpha\mathbf{J}\mathbf{W})^{-1}) = \gamma$$

- If $\mathbf{W} > 0$:

$$C \geq \alpha^2 \mathrm{Tr}((\mathbf{W}^{-1} + \alpha\mathbf{J})^{-2}\mathbf{J})$$

where $\alpha > 0$ is chosen such that:

$$\mathrm{Tr}((\mathbf{W}^{-1} + \alpha\mathbf{J})^{-2}\mathbf{W}^{-1}) = \gamma.$$

Comparison with UCRLB

- Until know we minimized the joint variance
- Scalar UCRLB minimizes variance for each component separately:

$$[\mathbf{C}(\mathbf{D})]_{ii} = ([\mathbf{I}]_i^* + \mathbf{d}_i)\mathbf{J}^{-1}([\mathbf{I}]_i + \mathbf{d}_i^*) \quad \text{s.t. } \mathbf{d}_i \mathbf{W} \mathbf{d}_i^* \leq \gamma_i$$

- The total variance is:

$$\min_{\mathbf{D}} \left\{ \sum_{i=1}^{m} ([\mathbf{I}]_i^* + \mathbf{d}_i)\mathbf{J}^{-1}([\mathbf{I}]_i + \mathbf{d}_i^*) \right\} = \min_{\mathbf{D}} \left\{ \mathrm{Tr}\left((\mathbf{I} + \mathbf{D})\mathbf{J}^{-1}(\mathbf{I} + \mathbf{D})^* \right) \right\}$$

$$= \min_{\mathbf{D}} C(\mathbf{D})$$

Thus, we have the same optimization problem!

Comparison with UCRLB (cont.)

- Scalar UCRLB minimizes $\mathbf{C}(\mathbf{D})$ s.t.

$$[\mathbf{DWD}^*]_{ii} \leq \gamma_i, \quad 1 \leq i \leq m.$$

- Vector UCRLB minimizes $\mathbf{C}(\mathbf{D})$ s.t.

$$\sum_{i=1}^{m} [\mathbf{DWD}^*]_{ii} \leq \sum_{i=1}^{m} \gamma_i = \gamma, \quad 1 \leq i \leq m.$$

- Scalar constraints are tighter for the same optimization problem.
- Joint optimization over **D** yields lower overall variance
- Cross-correlation allows bias in one component to reduce total variance by compensating for another.

UCRLB with Worst-Case Bias Constraint

- Minimize total variance $C(\mathbf{D}) = \mathrm{Tr}((\mathbf{I}+\mathbf{D})\mathbf{J}^{-1}(\mathbf{I}+\mathbf{D})^*)$
- Subject to worst-case bias constraint:

$$D_{\mathrm{WC}} = \max_{\mathbf{z}\in\mathbb{C}^m, \|\mathbf{z}\|=1} \mathbf{z}^*\mathbf{S}\mathbf{D}^*\mathbf{D}\mathbf{S}\mathbf{z} \leq \gamma$$

- Two solution approaches:
  1. **Jointly diagonalizable case**: Analytical solution via eigen-decomposition

$$\mathbf{J}^{-1} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^*, \quad \mathbf{S} = \mathbf{Q}\mathbf{\Gamma}\mathbf{Q}^*$$

  2. **Arbitrary S case**: Requires numerical SDP optimization
- Why consider jointly diagonalizable **S** and **J**?

UCRLB with Worst-Case Bias Constraint

- Minimize total variance $C(\mathbf{D}) = \mathrm{Tr}((\mathbf{I}+\mathbf{D})\mathbf{J}^{-1}(\mathbf{I}+\mathbf{D})^*)$
- Subject to worst-case bias constraint:

$$D_{\mathrm{WC}} = \max_{\mathbf{z}\in\mathbb{C}^m, \|\mathbf{z}\|=1} \mathbf{z}^*\mathbf{S}\mathbf{D}^*\mathbf{D}\mathbf{S}\mathbf{z} \leq \gamma$$

- Two solution approaches:
  1. **Jointly diagonalizable case**: Analytical solution via eigen-decomposition

$$\mathbf{J}^{-1} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^*, \quad \mathbf{S} = \mathbf{Q}\boldsymbol{\Gamma}\mathbf{Q}^*$$

  2. **Arbitrary S case**: Requires numerical SDP optimization
- Why consider jointly diagonalizable **S** and **J**?
  - Enables **closed-form solution** via eigendecomposition
  - Simplifies analysis by decoupling constraints along eigenvectors

Worst-Case Constraint: Jointly Diagonalizable

- Assume **S** and **J** share eigenvectors
- Optimal bias gradient matrix:

$$\hat{\mathbf{D}}_{\mathsf{WC}} = (\mathbf{I} - \sqrt{\gamma}\mathbf{S}^{-1})\mathbf{P} - \mathbf{I}$$

  where **P** is projection onto eigenvectors of **S** with $\beta_j^2 > \gamma$

- Resulting total variance bound:

$$\mathrm{Tr}(\mathbf{C}_{\hat{\mathbf{x}}}) \geq \mathrm{Tr}((\mathbf{I} - \sqrt{\gamma}\mathbf{S}^{-1})^2\mathbf{P}\mathbf{J}^{-1})$$

- Special case $\mathbf{S} = \mathbf{I}$:

$$\mathrm{Tr}(\mathbf{C}_{\hat{\mathbf{x}}}) \geq \mathrm{Tr}((1 - \sqrt{\gamma})^2\mathbf{J}^{-1})$$

Worst-Case Constraint: Arbitrary S

- Assume that **S** is an arbitrary non-negative definite matrix
- We formulate the minimization of $\mathbf{C}(\mathbf{D})$ as a semidefinite programming (SDP) problem:

$$\min_{t,\mathbf{D}} t$$

$$\text{subject to} \begin{bmatrix} t & \mathbf{g}^* \\ \mathbf{g} & \mathbf{I} \end{bmatrix} \succeq 0$$

$$\begin{bmatrix} \gamma\mathbf{I} & \mathbf{SD}^* \\ \mathbf{DS} & \mathbf{I} \end{bmatrix} \succeq 0$$

where $\mathbf{g} = \text{vec}(\mathbf{J}^{-1/2}(\mathbf{I}+\mathbf{D})^*)$

- Efficiently solvable using interior point methods

Theorem 2: UCRLB with Worst-Case Bias

Theorem (Theorem 2)

For $\gamma < \lambda^2_{\max}$, the total variance C of any estimator with $\|\mathbf{DS}\|^2 \leq \gamma$ satisfies $C \geq C_{\min}$, where $C_{\min}$ is the solution to the SDP problem.
For $\mathbf{S} = \sum \beta_i \mathbf{q}_i \mathbf{q}_i^*$ (same eigenvectors as $\mathbf{J}$):

$$C_{\min} = Tr((\mathbf{I} - \sqrt{\gamma}\mathbf{S}^{-1})^2 \mathbf{P}\mathbf{J}^{-1})$$

For $\mathbf{S} = \mathbf{I}$:

$$C_{\min} = Tr((1 - \sqrt{\gamma})^2 \mathbf{J}^{-1})$$

Note: From Theorems 1 and 2 the two UCRLB bounds coincide for the scalar case.

Optimal Estimators for the Linear Gaussian Model

- The described theorems, 1 and 2, characterize the smallest possible total variance of any estimator – *with bias gradient matrix whose norm is bounded by a constant*.
- They do not guarantee that there exists estimators achieving these lower bounds.
- Now, we will show that for the case of a linear Gaussian model *both lower bounds* are achievable using a linear estimator.

Optimal Estimators for the Linear Gaussian Model

- The described theorems, 1 and 2, characterize the smallest possible total variance of any estimator – *with bias gradient matrix whose norm is bounded by a constant*.
- They do not guarantee that there exists estimators achieving these lower bounds.
- Now, we will show that for the case of a linear Gaussian model *both lower bounds* are achievable using a linear estimator.

So, we consider the class of estimation problems represented by the linear model:

$$\mathbf{y} = \mathbf{H}\mathbf{x_0} + \mathbf{n}$$

- $\mathbf{x_0} \in \mathbb{C}^n$ is a deterministic vector of unknown parameters
- $\mathbf{H}$ is a known $n \times m$ matrix with rank m
- $\mathbf{n} \in \mathbb{C}^n$ is a zero-mean Gaussian random vector with positive definite covariance $\mathbf{C_n}$.

Optimal Estimators for the Linear Gaussian Model (cont.)

Linear Model: $\mathbf{y} = \mathbf{H}\mathbf{x_0} + \mathbf{n}$

- For this model, the Fisher information matrix is given by:

$$\mathbf{J} = \mathbf{H}^*\mathbf{C_n}^{-1}\mathbf{H}.$$

- Let $\hat{\mathbf{D}}$ denote the optimal gradient bias that minimizes $\mathbf{C}(\mathbf{D})$ subject to the Worst-Case *or* Average Bias constrains.
- Then, the total variance of any linear or nonlinear estimator $\hat{\mathbf{x}}$ of $\mathbf{x_0}$ is bounded by:

$$\mathrm{Tr}(\mathbf{C}_{\hat{\mathbf{x}}}) \geq \mathrm{Tr}\left((\mathbf{I} + \hat{\mathbf{D}})(\mathbf{H}^*\mathbf{C_n}^{-1}\mathbf{H})^{-1}(\mathbf{I} + \hat{\mathbf{D}})^*\right).$$

- We now derive a linear estimator $\hat{\mathbf{x}} = \mathbf{G}\mathbf{y}$ of $\mathbf{x}_0$ that achieves the above bound.
- Let: $\mathbf{G} = (\mathbf{I} + \hat{\mathbf{D}})(\mathbf{H}^*\mathbf{C}_n^{-1}\mathbf{H})^{-1}\mathbf{H}^*\mathbf{C}_n^{-1}$.

Optimal Estimators for the Linear Gaussian Model (cont.)

$$\text{Linear Model:}\quad \mathbf{y} = \mathbf{H}\mathbf{x_0} + \mathbf{n}$$

$$\text{Linear Estimator:}\quad \hat{\mathbf{x}} = \mathbf{G}\mathbf{y}, \quad \text{with}\quad \mathbf{G} = (\mathbf{I} + \hat{\mathbf{D}})(\mathbf{H}^*\mathbf{C}_n^{-1}\mathbf{H})^{-1}\mathbf{H}^*\mathbf{C}_n^{-1}$$

- The bias of this estimator is $\mathbf{b} = (\mathbf{G}\mathbf{H} - \mathbf{I})\mathbf{x}_0$ so that the bias gradient matrix is:

  $$\mathbf{D} = \mathbf{G}\mathbf{H} - \mathbf{I} = \hat{\mathbf{D}}$$

  which satisfies the Worst-Case *or* Average Bias constrains.
- The total variance of $\hat{\mathbf{x}} = \mathbf{G}\mathbf{y}$ is:

  $$\mathrm{Tr}(\mathbf{C}_{\hat{\mathbf{x}}}) = \mathrm{Tr}(\mathbf{G}\mathbf{C}_n\mathbf{G}^*) = \mathrm{Tr}\left((\mathbf{I} + \hat{\mathbf{D}})(\mathbf{H}^*\mathbf{C}_n^{-1}\mathbf{H})^{-1}(\mathbf{I} + \hat{\mathbf{D}})^*\right)$$

  so this estimator achieves the lower bound.
- Note that the chosen estimator achieves the biased CRLB for estimators with bias gradient $\mathbf{D}$. Thus, in the case of a linear Gaussian model, the biased CRLB is always achieved by a linear estimator.

Optimal Estimator in the Linear Gaussian Model (AVG)

- Goal: Minimize total variance of estimators $\hat{\mathbf{x}}$ under a constraint on the **Average** weighted bias gradient.
- Constraint: $\mathrm{Tr}(\mathbf{D}^*\mathbf{D}\mathbf{W}) \leq \gamma < \mathrm{Tr}(\mathbf{W})$
- Optimal estimator:

$$\hat{\mathbf{x}} = \begin{cases} \left(\mathbf{H}^*\mathbf{C}_n^{-1}\mathbf{H} + \delta\mathbf{W}^{-1}\right)^{-1}\mathbf{H}^*\mathbf{C}_n^{-1}\mathbf{y}, & 0 \leq \gamma < \mathrm{Tr}(\mathbf{W}) \\ 0, & \gamma \geq \mathrm{Tr}(\mathbf{W}) \end{cases},$$

with regularization parameter $\delta > 0$ chosen such that:

$$\mathrm{Tr}\left((\mathbf{W}^{-1} + \frac{1}{\delta}\mathbf{H}^*\mathbf{C}_n^{-1}\mathbf{H})^{-2}\mathbf{W}^{-1}\right) = \gamma$$

- Conclusion: This is the **ridge estimator** (Tikhonov regularization), which:
  - Minimizes total variance among all (linear and nonlinear) estimators *with bounded average bias gradient*.
  - Remains optimal (among linear estimators) for any noise distribution.

Optimal Estimator in the Linear Gaussian Model (WC)

- Goal: Minimize total variance under **Worst-Case** bias gradient constraint:

$$\mathbf{z}^*\mathbf{S}\mathbf{D}^*\mathbf{D}\mathbf{S}\mathbf{z} \le \gamma < \lambda_{\text{max}}^2, \quad \forall \mathbf{z},\ \|\mathbf{z}^*\mathbf{z}\| = 1$$

  where $\mathbf{S} \succ 0$ and commutes with $\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}$.

- Optimal estimator:

$$\hat{\mathbf{x}} = \begin{cases} (\mathbf{I} - \sqrt{\gamma}\mathbf{S}^{-1})\mathbf{P}\left(\mathbf{H}^*\mathbf{C_n}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^*\mathbf{C_n}^{-1}\mathbf{y}, & 0 \le \gamma < \lambda_{\text{max}}^2 \\ 0, & \gamma \ge \lambda_{\text{max}}^2 \end{cases},$$

  - **P**: projection onto eigenspace of **S** with eigenvalues s.t. $\beta_i^2 > \gamma$
  - **G**: defined as in earlier optimal estimator formula, with $\hat{\mathbf{D}} = \hat{\mathbf{D}}_{\text{WC}}$

- Special case: If $\mathbf{S} = \mathbf{I}$, this becomes the **shrunken estimator** (scaled version of least-squares estimator).
- Conclusion:
  - The above estimator minimizes total variance among all (linear and nonlinear) estimators *with bounded worst-case bias gradient*.
  - Remains optimal (among linear estimators) for any noise distribution.
  - For general **S**, this is a generalization of the shrunken estimator.

Application to System Identification

- **Model**: Noisy measurements of filtered signal

$$y[k] = \sum_{m=0}^{n-1} h[m]u[k-m] + \eta[k], \quad 0 \le k \le n-1$$

- Matrix form: $\mathbf{y} = \mathbf{H}\mathbf{x}_0 + \mathbf{n}$ where:
  - $\mathbf{H}$ is lower triangular convolution matrix
  - $\mathbf{x}_0 = [h[0], \ldots, h[n-1]]^T$ (unknown impulse response)
  - $\mathbf{n} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ (white Gaussian noise)
- **Note:** Since we have a linear Gaussian model, the UCRLB is achievable using a linear estimator.
- **Two Estimators**:
  1. Vector Tikhonov (joint estimation):

  $$\hat{\mathbf{x}}_{\text{vec}} = \alpha(\alpha\mathbf{H}^*\mathbf{H} + \sigma^2\mathbf{I})^{-1}\mathbf{H}^*\mathbf{y}, \text{ where '}a\text{' chosen s.t. } \mathrm{Tr}((\mathbf{I}+\alpha/\sigma^2\mathbf{H}^*\mathbf{H})^{-2})=\gamma.$$

  2. Scalar Tikhonov (component-wise):

  $$\hat{x}_i = \alpha[(\alpha\mathbf{H}^*\mathbf{H} + \sigma^2\mathbf{I})^{-1}]_i^{-1}\mathbf{H}^*\mathbf{y}, \text{ where '}a\text{' chosen s.t. } [(\mathbf{I}+\alpha/\sigma^2\mathbf{H}^*\mathbf{H})^{-2}]_{ii}=\frac{\gamma}{n}.$$

Application to System Identification (cont.)

Observations:

- Vector estimator achieves lower total variance for same bias gradient norm
- Performance gap increases with constraint tightness (γ)
- Empirical results match theoretical UCRLB predictions for both scalar and vector estimator

Key Insight: Joint parameter estimation with vector constraint outperforms component-wise scalar constraints

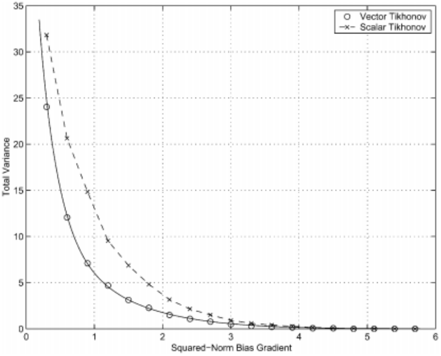

Fig. 1. Variance of the vector Tikhonov estimator (66) and the scalar Tikhonov estimator (68) as a function of the squared-norm bias gradient in comparison with the vector and scalar UCRLB. The line denotes the vector UCRLB, os denote the performance of the vector Tikhonov estimator, the dashed line denotes the scalar UCRLB, and the xs denote the scalar Tikhonov estimator.

Asymptotic Optimality of the PML Estimator

- In general, the UCRLB may not be achievable.
- In the linear Gaussian model:
  - With average bias constraint, the **Tikhonov (ridge)** *estimator achieves the UCRLB*:

$$\hat{\mathbf{x}} = \arg\max\left\{\log p(\mathbf{y};\mathbf{x}) - \frac{\beta}{2}\mathbf{x}^*\mathbf{W}\mathbf{x}\right\}$$

  - Equivalent to minimizing: $(\mathbf{y}-\mathbf{H}\mathbf{x})^*\mathbf{C}_n^{-1}(\mathbf{y}-\mathbf{H}\mathbf{x}) + \beta\mathbf{x}^*\mathbf{W}\mathbf{x}$
  - With worst-case bias and $\mathbf{S}=\mathbf{I}$ (or if $\mathbf{S}$ has same eigenvectors as $\mathbf{J}$), **shrunken estimator** also achieves the UCRLB (with: $\mathbf{W}=-\mathbf{H}^*\mathbf{H}$).
- Conclusion: In linear Gaussian models, the PML estimator with suitable penalty achieves the UCRLB.
- Extension: The PML estimator asymptotically achieves the UCRLB in many *general statistical models*, given an appropriate penalizing function.

A. PML Estimator

- The PML estimator maximizes a penalized log-likelihood:

$$\hat{\mathbf{x}}^{\mathrm{PML}} = \arg\max \left\{ \log p(\mathbf{y};\mathbf{x}) - \beta R(\mathbf{x}) \right\}$$

  where $\beta > 0$ for regularization, and $R(\mathbf{x})$ is a penalizing function.
- Interpretation: Equivalent to MAP estimation with prior pdf of $\mathbf{x}_0$

$$p(\mathbf{x}) \propto e^{-\beta R(\mathbf{x})}.$$

- For N iid measurements $\mathbf{y}_1, \ldots, \mathbf{y}_N$, the PML becomes:

$$\hat{\mathbf{x}}^{\mathrm{PML}} = \arg\max \left\{ \sum_{i=1}^{N} \log p(y_i;\mathbf{x}) - \beta_N R(\mathbf{x}) \right\}$$

- Many penalization forms exist, but...
  general optimality is **not** guaranteed for these different choices!
- However, under certain regularity conditions:
  - The PML estimator, from $\mathbf{y}_1, \ldots, \mathbf{y}_N$ iid, is *asymptotically Gaussian*.
  - Explicit asymptotic mean and variance expressions can be derived.
  - The PML asymptotically achieves the UCRLB.

B. Asymptotic Properties of the PML Estimator

- Goal: Estimate deterministic vector $\mathbf{x}_0$ from N iid measurements $\mathbf{y}_1, \dots, \mathbf{y}_N$ via the PML estimator, for:
  - β_N s.t. $\beta_N / N \to \beta_0$ for some constant β_0 as $N \to \infty$
  - $R(\mathbf{x})$ s.t. $\partial^3 R(\mathbf{x}) / \partial x_j \partial x_k \partial x_l$ is bounded for all j, k, l

B. Asymptotic Properties of the PML Estimator

- Goal: Estimate deterministic vector $\mathbf{x}_0$ from N iid measurements $\mathbf{y}_1, \ldots, \mathbf{y}_N$ via the PML estimator, for:
  - β_N s.t. $\beta_N / N \to \beta_0$ for some constant β_0 as $N \to \infty$
  - $R(\mathbf{x})$ s.t. $\partial^3 R(\mathbf{x}) / \partial x_j \partial x_k \partial x_l$ is bounded for all j, k, l
- Assumptions on the pdf $p(\mathbf{y}; \mathbf{x})$:
  - **A1:** First, second, and third$(\partial^3 \log p(\mathbf{y}; \mathbf{x}) / \partial x_j \partial x_k \partial x_l)$ derivatives of $\log p(\mathbf{y}; \mathbf{x})$ exist on open $\mathcal{X} \ni \check{\mathbf{x}}$, where

    $$\check{\mathbf{x}} = \arg\max \left\{ E\left[\log p(\mathbf{y}; \mathbf{x})\right] - \beta_0 R(\mathbf{x}) \right\}$$

  - **A2:** Third derivatives of $\log p(\mathbf{y}; \mathbf{x})$ are bounded by $d(\mathbf{y})$ with $E_{\mathbf{x}}[d(\mathbf{y})] < \infty$, $\forall \mathbf{x} \in \mathcal{X}$
  - **A3:** $-E\left[\frac{\partial^2 \log p(\mathbf{y}; \check{\mathbf{x}})}{\partial \mathbf{x}^2}\right] + \beta_0 \frac{\partial^2 R(\mathbf{x})}{\partial \mathbf{x}^2} > 0$

B. Asymptotic Properties of the PML Estimator

- Goal: Estimate deterministic vector $\mathbf{x}_0$ from N iid measurements $\mathbf{y}_1, \ldots, \mathbf{y}_N$ via the PML estimator, for:
  - β_N s.t. $\beta_N/N \to \beta_0$ for some constant β_0 as $N \to \infty$
  - $R(\mathbf{x})$ s.t. $\partial^3 R(\mathbf{x})/\partial x_j \partial x_k \partial x_l$ is bounded for all j, k, l
- Assumptions on the pdf $p(\mathbf{y}; \mathbf{x})$:
  - **A1:** First, second, and third$(\partial^3 \log p(\mathbf{y}; \mathbf{x})/\partial x_j \partial x_k \partial x_l)$ derivatives of $\log p(\mathbf{y}; \mathbf{x})$ exist on open $\mathcal{X} \ni \check{\mathbf{x}}$, where

$$\check{\mathbf{x}} = \arg\max \left\{ E\left[\log p(\mathbf{y}; \mathbf{x})\right] - \beta_0 R(\mathbf{x}) \right\}$$

  - **A2:** Third derivatives of $\log p(\mathbf{y}; \mathbf{x})$ are bounded by $d(\mathbf{y})$ with $E_{\mathbf{x}}[d(\mathbf{y})] < \infty$, $\forall \mathbf{x} \in \mathcal{X}$
  - **A3:** $-E\left[\frac{\partial^2 \log p(\mathbf{y}; \check{\mathbf{x}})}{\partial \mathbf{x}^2}\right] + \beta_0 \frac{\partial^2 R(\mathbf{x})}{\partial \mathbf{x}^2} > 0$
- **Theorem 3:** Under A1–A3, the PML estimator is asymptotically normal:

$$\sqrt{N}(\hat{\mathbf{x}}^{\mathrm{PML}} - \check{\mathbf{x}}) \overset{a}{\sim} \mathcal{N}\left(0, (\mathbf{J}(\check{\mathbf{x}}) + \beta_0 \mathbf{M}(\check{\mathbf{x}}))^{-1} \mathbf{C}(\check{\mathbf{x}}) (\mathbf{J}(\check{\mathbf{x}}) + \beta_0 \mathbf{M}(\check{\mathbf{x}}))^{-1}\right)$$

where $\beta_0 = \lim_{N \to \infty} \beta_N / N$.

C. PML Estimator and the UCRLB

- From Theorem 3, the *asymptotic* total variance of $\hat{\mathbf{x}}^{\text{PML}}$ is

$$\frac{1}{N}\text{Tr}\left((\mathbf{J}(\check{\mathbf{x}})+\beta_0\mathbf{M}(\check{\mathbf{x}}))^{-1}\mathbf{C}(\check{\mathbf{x}})(\mathbf{J}(\check{\mathbf{x}})+\beta_0\mathbf{M}(\check{\mathbf{x}}))\right), \quad \text{where:}$$

$$\check{\mathbf{x}} = \arg\max\{E\{\log p(\mathbf{y};\mathbf{x})\} - \beta_0 R(\mathbf{x})\},$$

$$\mathbf{C}(\check{\mathbf{x}}) = \text{cov}\left\{\frac{\partial \log p(\mathbf{y};\check{\mathbf{x}})}{\partial \mathbf{x}}\right\}, \quad \mathbf{J}(\check{\mathbf{x}}) = -E\left\{\frac{\partial^2 \log p(\mathbf{y};\check{\mathbf{x}})}{\partial \mathbf{x}^2}\right\},$$

$$\text{and} \quad \mathbf{M}(\check{\mathbf{x}}) = \frac{\partial^2 R(\check{\mathbf{x}})}{\partial \mathbf{x}^2}.$$

- From that we have that the *asymptotic* bias gradient $\mathbf{D}_{\text{PML}}$ is

$$\mathbf{D}_{\text{PML}} = \frac{\partial \check{\mathbf{x}}}{\partial \mathbf{x}_0} - \mathbf{I}.$$

- After calculations we find the $\partial\check{\mathbf{x}}/\partial\mathbf{x}_0 = (\partial\check{\mathbf{x}}/\partial\mathbf{x})\cdot(\partial\mathbf{x}/\partial\mathbf{x}_0)$ to be:

$$\frac{\partial \check{\mathbf{x}}}{\partial \mathbf{x}_0} = (\mathbf{J}(\check{\mathbf{x}})+\beta_0\mathbf{M}(\check{\mathbf{x}}))^{-1}\frac{\partial}{\partial \mathbf{x}_0}E\left\{\frac{\partial \log p(\mathbf{y};\check{\mathbf{x}})}{\partial \mathbf{x}}\right\}.$$

C. PML Estimator and the UCRLB (& Theorem 1)

- Let $\gamma = \mathbf{D}^*_{\mathsf{PML}}\mathbf{D}_{\mathsf{PML}}$. Then, from Theorem 1, any estimator with bias gradient $\mathbf{D}$ satisfying $\mathrm{Tr}(\mathbf{D}^*\mathbf{D}) \leq \mathrm{Tr}(\gamma)$ must satisfy:

$$C \geq \frac{\alpha^2}{N}\mathrm{Tr}\left((\mathbf{I}+\alpha\mathbf{J}_1)^{-2}\mathbf{J}_1\right),$$

  where:
  - $\alpha > 0$ s.t. $\mathrm{Tr}\left((\mathbf{I}+\alpha\mathbf{J}_1)^{-2}\right) = \mathrm{Tr}\left(\left(\frac{\partial\check{\mathbf{x}}}{\partial\mathbf{x}_0}-\mathbf{I}\right)^*\left(\frac{\partial\check{\mathbf{x}}}{\partial\mathbf{x}_0}-\mathbf{I}\right)\right)$,
  - $\mathbf{J}_1 = E\left\{\left(\frac{\partial\log p(\mathbf{y}_1;\mathbf{x}_0)}{\partial\mathbf{x}}\right)^*\left(\frac{\partial\log p(\mathbf{y}_1;\mathbf{x}_0)}{\partial\mathbf{x}}\right)\right\}$.
    (*Fisher information from a single observation*)

- Thus, if $R(\mathbf{x})$ is chosen such that

$$\mathrm{Tr}\left((\mathbf{J}(\check{\mathbf{x}})+\beta_0\mathbf{M}(\check{\mathbf{x}}))^{-1}\mathbf{C}(\check{\mathbf{x}})(\mathbf{J}(\check{\mathbf{x}})+\beta_0\mathbf{M}(\check{\mathbf{x}}))^{-1}\right) =$$

$$\alpha^2\mathrm{Tr}\left((\mathbf{I}+\alpha\mathbf{J}_1)^{-2}\mathbf{J}_1\right),$$

  then the PML estimator achieves the UCRLB under the *average* bias constraint, and asymptotically no estimator (linear nor non-linear) with $\mathbf{D}$ st: $\mathrm{Tr}(\mathbf{D}^*\mathbf{D}) \leq \mathrm{Tr}(\gamma)$ has lower total variance.

C. PML Estimator and the UCRLB (& Theorem 2)

- From Theorem 2, the variance of any estimate of $\mathbf{x}_0$ with bias gradient $\mathbf{D}$ such that $\|\mathbf{D}\|^2 \le \|\mathbf{D}_{\mathrm{PML}}\|^2$ satisfies:

$$C \ge \frac{1}{N} \operatorname{Tr}\left((1 - \|\mathbf{D}_{\mathrm{PML}}\|)^2 \mathbf{J}_1^{-1}\right), \text{ where: } \mathbf{D}_{\mathrm{PML}} = \frac{\partial \check{\mathbf{x}}}{\partial \mathbf{x}_0} - \mathbf{I}.$$

- We choose $R(\mathbf{x})$ such that:

$$\operatorname{Tr}\left((\mathbf{J}(\check{\mathbf{x}}) + \beta_0 \mathbf{M}(\check{\mathbf{x}}))^{-1} \mathbf{C}(\check{\mathbf{x}}) (\mathbf{J}(\check{\mathbf{x}}) + \beta_0 \mathbf{M}(\check{\mathbf{x}}))^{-1}\right)$$

$$= \operatorname{Tr}\left(\left(1 - \left\|\frac{\partial \check{\mathbf{x}}}{\partial \mathbf{x}_0} - \mathbf{I}\right\|\right)^2 \mathbf{J}_1^{-1}\right), \textit{ where } \partial\check{\mathbf{x}}/\partial\mathbf{x}_0 \textit{ known.}$$

- Based on the above, the corresponding PML estimator achieves the UCRLB under the *worst-case* bias constraint.
- Asymptotically, no other linear or nonlinear estimator exists with:
  - A bias gradient $\mathbf{D}$ such that $\|\mathbf{D}\| \le \|\mathbf{D}_{\mathrm{PML}}\|$, and
  - A smaller total variance than that of the PML estimator.

C. PML Estimator and the UCRLB (for scalar x_0)

Observation

The conditions that are found for each one of the Theorems 1 and 2 are not very insightful when it comes to choosing $R(\mathbf{x})$, so we seek now to estimate a *scalar* x_0 from N iid measurements.

- In this analysis, the average and worst-case UCRLB coincide, and the variance C of any estimate of x_0 with bias gradient D such that:

$$D^2 \leq D^2_{\text{PML}} = (\partial \check{x}/\partial x_0 - 1)^2 ,$$

satisfies the bound:

$$C \geq \left(1 - \left|\frac{\partial \check{x}}{\partial x_0} - 1\right|\right)^2 \cdot \frac{1}{NJ_1}, \text{ where:}$$

$$J_1 = E\left\{\left(\frac{\partial \log p(y;x_0)}{\partial x}\right)^2\right\}, \check{x} = \arg\max\left\{E\left\{\log p(y;x)\right\} - \beta_0 R(x)\right\}.$$

C. PML Estimator and the UCRLB (for scalar x_0 – cont.)

- Recall the sensitivity of the regularized estimator $\check{x}$:

$$\frac{\partial \check{x}}{\partial x_0} = \frac{1}{J(\check{x}) + \beta_0 M(\check{x})} \cdot \frac{\partial}{\partial x_0} E\left\{ \frac{\partial \log p(y;\check{x})}{\partial x} \right\},$$

where:

$$J(\check{x}) = -E\left\{\partial^2 \log p(y;\check{x})/\partial x^2\right\}, \quad M(\check{x}) = \partial^2 R(\check{x})/\partial x^2,$$

$$C(\check{x}) = \mathrm{var}\left\{\partial \log p(y;\check{x})/\partial x\right\}.$$

- From Theorem 3, the asymptotic variance of the PML estimator is:

$$C_{\mathrm{PML}} = \frac{C(\check{x})}{N(J(\check{x}) + \beta_0 M(\check{x}))^2}.$$

- So, if we can choose $R(x)$ such that:

$$\left(1 - \left|\frac{\partial \check{x}}{\partial x_0} - 1\right|\right)^2 \cdot \frac{1}{J_1} = \frac{C(\check{x})}{(J(\check{x}) + \beta_0 M(\check{x}))^2},$$

then the PML estimator *achieves* the UCRLB for scalar x_0.

Note: A general condition under which the above is satisfied can be found in Appendix B.

Example: Exponential Mean Estimation

- Consider estimating an unknown deterministic scalar parameter from N iid measurements of an exponential random variable.
- The pdf is $p(y_i; x_0) = x_0 e^{-y_i x_0}, \quad 1 \leq i \leq N$, and $1/x_0 > 0$.
- The PML estimate $\hat{x}^{\mathrm{PML}}$ is obtained by maximizing:

$$PL(x) = N \log x - x \sum_{i=1}^{N} y_i - \beta_N R(x),$$

  for $\beta_N > 0$ with $\beta_N / N \to \beta_0$ as $N \to \infty$.
- The goal is to choose $R(x)$ such that $\hat{x}^{\mathrm{PML}}$ asymptotically achieves the UCRLB.
- After calculations we have that:

$$\frac{\partial \log p(y; \check{x})}{\partial x} - E\left\{ \frac{\partial \log p(y; \check{x})}{\partial x} \right\} = \frac{1}{x_0} - y,$$

$$\frac{\partial \log p(y; x_0)}{\partial x_0} = \frac{1}{x_0} - y.$$

Example: Exponential Mean Estimation (cont.)

- From the mathematical analysis, if $\partial \check{x} / \partial x_0 \leq 1$, then the resulting PML estimator asymptotically achieves the UCRLB.
- However, for finite N, the choice of $R(x)$ still significantly impacts performance.
- We have, after computations:

$$\frac{\partial \check{x}}{\partial x_0} = \frac{\frac{1}{x_0^2}}{\frac{1}{\check{x}^2} + \beta_0 M(\check{x})}$$

- If $\partial R(\check{x}) / \partial x, \partial^2 R(\check{x}) / \partial x^2 \geq 0$, then from the definition of $\check{x}$ we have:

$$\frac{1}{\check{x}} = \frac{1}{x_0} + \beta_0 \frac{\partial R(\check{x})}{\partial x} \geq \frac{1}{x_0},$$

  and by differentiating that we indeed conclude to $\partial \check{x} / \partial x_0 \leq 1$ which means the PML estimator is optimal.
- For instance, by choosing $R(x) = x$ or $R(x) = \log x$ we have in both cases estimators that asymptotically achieve the UCRLB.

Example: Compare the Performance of PML Estimators

- For $R(x) = x$ we have:

$$\hat{x}^{\text{PML}} = \arg\max \left\{ N \log x - x \left(\sum_{i=1}^{N} y_i + \beta_N \right) \right\} = \frac{N}{\sum_{i=1}^{N} y_i + \beta_N}.$$

- For $R(x) = \log x$ we have:

$$\hat{x}^{\text{PML}} = \arg\max \left\{ (N - \beta_N) \log x - x \sum_{i=1}^{N} y_i \right\} = \frac{N - \beta_N}{\sum_{i=1}^{N} y_i}.$$

We will compare the performance of the PML estimators above with the UCRLB, for different values of N.

Example: Empirical Evaluation of PML Estimators

- We compare two PML estimators with the UCRLB across different sample sizes N by estimating:
  - Estimator variance $\hat{\sigma}^2$.
  - Squared bias gradient $\hat{D}^2$.

  Rather than attempting to determine these quantities analytically, we propose to estimate them from the measurements.
- To estimate the variance of each of the estimators (for each γ), we generate $L = 5000$ PML estimators.
- Variance is estimated as:

$$\hat{\sigma}^2 = \frac{1}{L} \sum_{i=1}^{L} \left((\hat{x}^{\mathrm{PML}})^{(i)} - \bar{x}_{\mathrm{PML}} \right)^2 ,$$

where $\bar{x}_{\mathrm{PML}}$ is the sample mean of $(\hat{x}^{\mathrm{PML}})^{(i)}$ (which denotes and i-th estimator).

Example: Empirical Evaluation of PML Estimators (cont.)

- We compare two PML estimators with the UCRLB across different sample sizes N by estimating:
  - Estimator variance $\hat{\sigma}^2$.
  - Squared bias gradient $\hat{D}^2$.

  Rather than attempting to determine these quantities analytically, we propose to estimate them from the measurements.
- Squared bias gradient is estimated by:

$$\hat{D} = \frac{1}{L}\sum_{i=1}^{L}\left(\left((\hat{x}^{\mathrm{PML}})^{(i)} - \zeta^{(i)}\right)\left(\frac{N}{x} - \sum_{j=1}^{N} y_j^{(i)}\right)\right) - 1,$$

$$\text{where:}\quad \zeta^{(i)} = \frac{1}{L-1}\left(\sum_{j=1}^{L}\hat{x}^{(j)} - \hat{x}^{(i)}\right).$$

- Conclusion: This empirical framework enables a practical assessment of how well each PML estimator approaches the UCRLB, especially for finite N.

Example: Comparison of PML Estimators

- The next figures show the estimated variance vs. squared bias gradient for the two PML estimators, and the UCRLB, for $N = 10, 20, 30$.
- **Key observations:**
  - Even for small N, the UCRLB closely approximates the estimator variance, especially when the squared bias gradient is large.
  - For small bias gradients, actual variance often exceeds the bound – partially due to higher estimation variance in this regime.
  - As N increases, both PML estimators converge to the UCRLB across all bias levels.
- For small N, the estimator with $R(x) = x$ shows consistently lower variance compared to the other one, indicating better finite-sample performance.

Example: Plots (for $N = 10$ and $N = 20$)

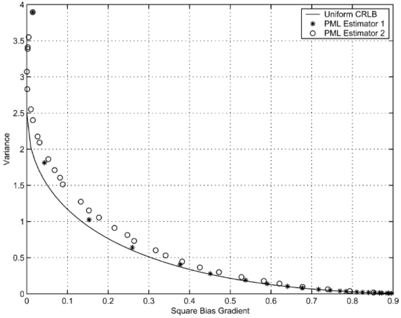

Fig. 2. Performance of the PML estimators (109) (denoted "1") and (112) (denoted "2") with $N = 10$ in comparison with the UCRLB. The line denotes the UCRLB, the circles denote the performance of the PML estimator 1, and the stars denote the performance of the PML estimator 2.

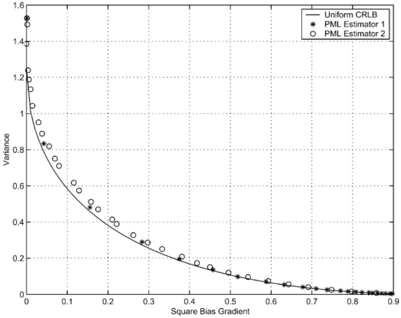

Fig. 3. Performance of the PML estimators (109) (denoted "1") and (112) (denoted "2") with $N = 20$ in comparison with the UCRLB. The line denotes the UCRLB, the circles denote the performance of the PML estimator 1, and the stars denote the performance of the PML estimator 2.

Example: Plots (for $N = 20$ and $N = 30$)

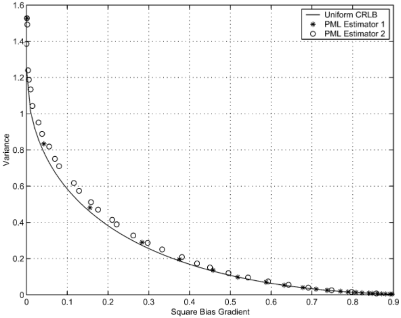

Fig. 3. Performance of the PML estimators (109) (denoted "1") and (112) (denoted "2") with $N = 20$ in comparison with the UCRLB. The line denotes the UCRLB, the circles denote the performance of the PML estimator 1, and the stars denote the performance of the PML estimator 2.

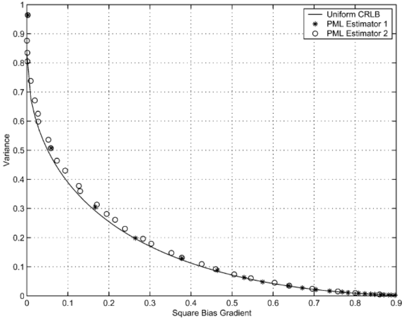

Fig. 4. Performance of the PML estimators (109) (denoted "1") and (112) (denoted "2") with $N = 30$, in comparison with the UCRLB. The line denotes the UCRLB, the circles denote the performance of the PML estimator 1, and the stars denote the performance of the PML estimator 2.

Conclusions

- Developed lower bounds on the **total variance** of biased estimators by constraining the **bias gradient norm**.
- For the **linear Gaussian model**:
  - **Tikhonov estimator** minimizes variance under average bias constraint.
  - **Shrunken estimator** minimizes variance under worst-case bias constraint.
- **PML estimator** with appropriate regularization:
  - Asymptotically achieves the UCRLB.
  - Varies in performance for finite sample sizes.
- **Open Questions**:
  - Can the PML achieve the UCRLB beyond the linear Gaussian model?
  - How does the PML perform for *finite* data?
  - Can these results extend to cases where the Fisher Information Matrix is singular?

References

- Hero et al., IEEE TSP, 1996
- Tikhonov, Sov. Math. Dokl., 1963
- Kay, Fundamentals of Statistical Signal Processing, 1993
- Eldar, IEEE TSP, 2004
- Fessler et al., IEEE Trans., multiple years

Q & A

Thank you for your attention!
Questions and Discussion