[19] J. Snyders and Y. Be'ery, "Maximum likelihood soft decoding of binary block codes and decoders for the Golay codes," *IEEE Trans. Inform. Theory*, vol. 35, pp. 963–975, Sept. 1989.
[20] G. Solomon and H. C. A. van Tilborg, "A connection between block and convolutional codes," *SIAM J. Appl. Math.*, pp. 358–369, 1979.
[21] K. S. Trivedi, *Probability and Statistics with Reliability, Queuing, and Computer Science Applications.* Englewood Cliffs, NJ: Prentice-Hall, 1982.
[22] A. Vardy and Y. Be'ery, "Bit-level soft decision decoding of Reed–Solomon codes," *IEEE Trans Commun.*, vol. 37, pp. 440–445, 1991.
[23] ——, "More efficient soft-decision decoding of the Golay codes," *IEEE Trans. Inform. Theory*, vol. 37, pp. 667–672, 1991.
[24] A. J. Viterbi, "Error bound for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Inform. Theory*, vol. IT-13, pp. 260–269, Apr. 1967.
[25] J. K. Wolf, "Efficient maximum likelihood decoding of linear block codes using a trellis," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 76–80, Jan. 1978.

A Proof of the Fisher Information Inequality via a Data Processing Argument

Ram Zamir, *Member, IEEE*

***Abstract*— The Fisher information $J(X)$ of a random variable X under a translation parameter appears in information theory in the classical proof of the Entropy-Power Inequality (EPI). It enters the proof of the EPI via the De-Bruijn identity, where it measures the variation of the differential entropy under a Gaussian perturbation, and via the convolution inequality $J(X+Y)^{-1} \geq J(X)^{-1} + J(Y)^{-1}$ (for independent X and Y), known as the Fisher Information Inequality (FII). The FII is proved in the literature directly, in a rather involved way. We give an alternative derivation of the FII, as a simple consequence of a "data-processing inequality" for the Cramer–Rao lower bound on parameter estimation.**

***Index Terms*—Cramer–Rao bound, data processing inequality, entropy-power inequality, Fisher information, linear modeling, non-Gaussian noise, prefiltering.**

I. Introduction

The data processing inequality (or the data processing theorem) is used in information theory for proving the converse channel-coding theorem [4, Secs. V.3, V.4], [6, Secs. II.8, VIII.9]. This inequality asserts that if the random variables $W - X - Y$ form a Markov chain in this order, then the mutual informations between them satisfy

$$I(W;Y) \leq I(W;X). \tag{1}$$

In the special case where Y is given by a deterministic function ϕ of X, (1) becomes

$$I(W;\phi(X)) \leq I(W;X) \tag{2}$$

with equality if $W - \phi(X) - X$ form a Markov chain, e.g., if $\phi(\cdot)$ is an invertible function. The proof of (1) follows straightforwardly from the chain rule and the positivity of the mutual information [6].

The name "data processing inequality" apparently came from the analogy to the problem of optimal filtering. Suppose that W, X, Y are real variables. In analogy with (1) and (2), it is clear and easy to verify that the conditional variance, i.e., the mean-squared error of the conditional mean estimator of W, satisfies the data processing inequalities

$$\mathrm{VAR}(W|Y) \geq \mathrm{VAR}(W|X)$$

and

$$\mathrm{VAR}(W|\phi(X)) \geq \mathrm{VAR}(W|X) \tag{3}$$

where

$$\mathrm{VAR}(W|X) \triangleq E[W - E(W|X)]^2.$$

When the estimated quantity is a parameter $\boldsymbol{\theta}$ (i.e., not a random variable), it is impossible to use the conditional variance as a measure for the goodness of the optimal estimator. Instead, it is common to use the Fisher Information matrix (FI) of the measurement $\boldsymbol{X}$ relative to the parameter vector $\boldsymbol{\theta}$, defined as [4], [6], [10]

$$\begin{aligned}\boldsymbol{J}(\boldsymbol{X};\boldsymbol{\theta}) &\triangleq \mathrm{COV}\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\ \ln(f_{\boldsymbol{\theta}}(\boldsymbol{X}))\right\} \\ &= \int \frac{1}{f_{\boldsymbol{\theta}}(\boldsymbol{x})}\left(\frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial\boldsymbol{\theta}}\right)\cdot\left(\frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial\boldsymbol{\theta}}\right)^t d\boldsymbol{x}\end{aligned} \tag{4}$$

where $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_m)$, the set $\{f_{\boldsymbol{\theta}}(\boldsymbol{x})\}$ is a family of densities of $\boldsymbol{X}$ parameterized by $\boldsymbol{\theta}$, $\partial/\partial\boldsymbol{\theta}$ denotes the gradient (i.e., a column vector of partial derivatives) with respect to the parameters $\theta_1, \cdots, \theta_m$, $\ln(\cdot)$ denotes the natural logarithm, and $\mathrm{COV}\{\cdot\}$ denotes the $m \times m$ covariance matrix calculated relative to the distribution of $\boldsymbol{X}$. Here $\boldsymbol{X}$ may either be a single measurement or a vector of n measurements. The importance of the matrix $\boldsymbol{J}(\boldsymbol{X};\boldsymbol{\theta})$ follows from the Cramer–Rao Bound (CRB), [4], [6], [10], saying that for any unbiased estimator $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\boldsymbol{x})$ (i.e., estimator for which $E\{\hat{\boldsymbol{\theta}}(\boldsymbol{X})\} = \boldsymbol{\theta}$) the error vector $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}$ satisfies

$$\mathrm{COV}\{\hat{\boldsymbol{\theta}}(\boldsymbol{X})\} \geq \boldsymbol{J}(\boldsymbol{X};\boldsymbol{\theta})^{-1} \tag{5}$$

where throughout the correspondence an inequality between (nonnegative definite) matrices means that the difference matrix is nonnegative definite. As it turns out (see Lemma 3 below), the notion of data processing extends easily to the FI; if $\boldsymbol{\theta} - \boldsymbol{X} - \boldsymbol{Y}$ satisfy a chain relation of the form $f(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{\theta}) = f_{\boldsymbol{\theta}}(\boldsymbol{x})f(\boldsymbol{y}|\boldsymbol{x})$ (i.e., the conditional distribution of Y given X is independent of $\boldsymbol{\theta}$), then we have the data processing inequality

$$\boldsymbol{J}(\boldsymbol{Y};\boldsymbol{\theta}) \leq \boldsymbol{J}(\boldsymbol{X};\boldsymbol{\theta}) \tag{6}$$

whose deterministic version (in analogy with (2)) is

$$\boldsymbol{J}(\phi(\boldsymbol{X});\boldsymbol{\theta}) \leq \boldsymbol{J}(\boldsymbol{X};\boldsymbol{\theta}). \tag{7}$$

Equality in (7) holds if $\phi(X)$ is a *sufficient statistic* relative to the family $\{f_{\boldsymbol{\theta}}(x)\}$, i.e., $\boldsymbol{\theta} - \phi(\boldsymbol{X}) - \boldsymbol{X}$ form a chain [6, Sec. II.10].[1]

In the context of information-theoretic inequalities, e.g., in the derivation of the Entropy Power Inequality (EPI), there appears a

Manuscript received June 1, 1995; revised October 1, 1997. This work was supported in part by the Wolfson Research Awards administered by the Israel Academy of Science and Humanities. The material in this correspondence was presented in part at the Information Theory Workshop on Multiple Access and Queuing, St. Louis, MO, April 1995.

The author is with the Department of Electrical Engineering–Systems, Tel Aviv University, Tel Aviv 69978, Israel.

Publisher Item Identifier S 0018-9448(98)02377-3.

[1] An alternative ("Bayesian") way to express the equality condition is that $\boldsymbol{\Theta} - \phi(\boldsymbol{X}) - \boldsymbol{X}$ for a Markov chain for any distribution on the parameter θ.

special form of the FI matrix, namely, the FI of a random vector with respect to a *translation parameter*

$$\boldsymbol{J}(\boldsymbol{N}) \triangleq \boldsymbol{J}(\boldsymbol{\theta}+\boldsymbol{N};\boldsymbol{\theta}) = \mathrm{COV}\left\{\frac{\partial}{\partial \boldsymbol{N}}\ \ln f(\boldsymbol{N})\right\} = \int \frac{1}{f(\boldsymbol{n})}\left(\frac{\partial f(\boldsymbol{n})}{\partial \boldsymbol{n}}\right)\cdot\left(\frac{\partial f(\boldsymbol{n})}{\partial \boldsymbol{n}}\right)^{t} d\boldsymbol{n} \tag{8}$$

where $f(\boldsymbol{n})$ is the density function of the vector $\boldsymbol{N}$ ($f(n)$ is independent of $\boldsymbol{\theta}$), and $\boldsymbol{J}(\boldsymbol{N})$ is a square matrix whose dimension equals that of $\boldsymbol{N}$; see [3], [4], [6], and [8]. Unlike the general case (4), this form of the FI is a function of the density of the random vector alone, and not of its parameterization.[2]

The FI under translation (8) exhibits some well-known properties [1], [7], e.g.,

$$\boldsymbol{J}(A\boldsymbol{N}) = A^{-t}\boldsymbol{J}(\boldsymbol{N})A^{-1} \tag{9}$$

for any nonsingular square matrix A, and

$$\boldsymbol{J}(\boldsymbol{N}) \geq \mathrm{COV}(\boldsymbol{N})^{-1} \tag{10}$$

with equality iff $\boldsymbol{N}$ is Gaussian. Another property which is of particular interest for us is a convolution inequality, called the *Fisher Information Inequality* (FII). Let N_1 and N_2 be statistically independent random variables. Then

$$J(N_1+N_2)^{-1} \geq J(N_1)^{-1} + J(N_2)^{-1} \tag{11}$$

with equality iff N_1 and N_2 are Gaussian. Vector, matrix and "convex" versions of (11) exist in the literature [7], [8], some of which will be mentioned in the sequel. The FII (11), together with the De-Bruijn identity,[3] consist the key tools in the classical proof of the EPI [3], [4], [8]. Both FII and EPI relate to the tendency towards Gaussianity of the sum of independent random variables [1], [9], [16].

Existing proofs of the FII (11) [3], [4] involve a direct calculation of the convolution of the densities of $\boldsymbol{N}_1$ and $\boldsymbol{N}_2$ and application of the Cauchy–Schwartz inequality, and they are rather technical. In this correspondence we show that the FII follows from the Fisher information data processing inequality given in (7). We derive the FII by applying the data processing inequality to a suitable linear model relating the measurements and the parameters. This model provides an interesting interpretation to the difference between the two sides of inequality (11): $J(N_1+N_2)^{-1} - J(N_1)^{-1} - J(N_2)^{-1}$ amounts to the *loss in the CRB after optimal linear estimation*. If N_1 and N_2 are Gaussian, linear estimation is globally optimum, and the CRB loss is zero. However, if N_1 and N_2 are not Gaussian, noninvertible linear operation may increase the CRB.

In our proof, we consider a generalized form of the FII (11), namely, the matrix form of the FII, which was presented in [15] and [16]. The new derivation of the FII (11) is given in Section II. Some additional properties of the matrix form of the FII are given in Section III. In Section IV we use the matrix-FII to analyze the loss in FI (or in CRB) due to prefiltering in a certain linear model for parameter estimation. This part of the work appeared originally in [14].

II. Derivation of Results

In this section we prove a matrix form of the FII using the Fisher information data processing inequality (7). For completeness we give also a proof for (7), for which we could not find a reference in the literature. We then show that the matrix form of the FII implies the form in (11).

[2] In some references the FI of $\boldsymbol{N}$ is defined as $K(\boldsymbol{N}) = \mathrm{trace}\{\boldsymbol{J}(\boldsymbol{N})\}$.

[3] The scalar form of the De-Bruijn identity is

$$(d/dt)h(X+\sqrt{t}Z) = (1/2)J(X+\sqrt{t}Z)$$

where Z is a standard normal variable.

The derivation follows a sequence of lemmas. Below we assume that $\{f_{\boldsymbol{\theta}}(x,y)\}$ is a family of density functions parameterized by $\boldsymbol{\theta}$, where the first and second derivatives of $f_{\boldsymbol{\theta}}(x,y)$ with respect to $\boldsymbol{\theta}$ exist and are absolutely integrable (see [10, p. 66, eq. (c)]).

Lemma 1 (Chain Rule for the FI Matrix):

$$\boldsymbol{J}(X,Y;\boldsymbol{\theta}) = \boldsymbol{J}(X;\boldsymbol{\theta}) + \boldsymbol{J}(Y;\boldsymbol{\theta}|X) \tag{12}$$

where

$$\boldsymbol{J}(Y;\boldsymbol{\theta}|X) = E_X\{\boldsymbol{J}(Y;\boldsymbol{\theta}|X=x)\} = E_{XY}\left\{\left(\frac{\partial\ \ln\ f_{\boldsymbol{\theta}}(Y|X)}{\partial\boldsymbol{\theta}}\right)\left(\frac{\partial\ \ln\ f_{\boldsymbol{\theta}}(Y|X)}{\partial\boldsymbol{\theta}}\right)^{t}\right\}. \tag{13}$$

Note that the argument of the first expectation in (13), $\boldsymbol{J}(Y;\boldsymbol{\theta}|X=x)$, is the FI of Y relative to $\boldsymbol{\theta}$, calculated with respect to the conditional density of Y given a specific value $X=x$.

Proof: We prove assuming θ is scalar. The generalization to a vector $\boldsymbol{\theta}$ is straightforward. Since

$$J(X,Y;\theta) = E\{[\partial\ \ln\ f_\theta(X,Y)/\partial\theta]^2\}$$

and

$$\ln\ f_\theta(x,y) = \ln\ f_\theta(x) + \ln\ f_\theta(y|x)$$

we see that $J(X,Y;\theta)$ is given by the sum in the right-hand side of (12), plus a cross term which is twice

$$E\left\{\frac{\partial\ \ln\ f_\theta(X)}{\partial\theta}\cdot\frac{\partial\ \ln\ f_\theta(Y|X)}{\partial\theta}\right\} = E_X\left\{\frac{\partial\ \ln\ f_\theta(X)}{\partial\theta}\cdot E_Y\left\{\left.\frac{\partial\ \ln\ f_\theta(Y|X)}{\partial\theta}\right| X\right\}\right\} \tag{14}$$

where the right-hand side follows by iterating the expectation. This cross term is zero since the inner expectation in the right-hand side of (14) is zero for each value of X (see e.g. [10, p. 67, eq. (190)]).[4] □

Lemma 2 (Data Refinement Inequality):

$$\boldsymbol{J}(X,Y;\boldsymbol{\theta}) \geq \boldsymbol{J}(X;\boldsymbol{\theta}) \tag{15}$$

with equality if X is a sufficient statistic relative to the family $\{f_{\boldsymbol{\theta}}(x,y)\}$, i.e., if $\boldsymbol{\theta} - X - (X,Y)$ form a chain.

Proof: The inequality follows from Lemma 1 by the nonnegativity of the FI $\boldsymbol{J}(Y;\boldsymbol{\theta}|X=x)$. The equality condition implies that $\boldsymbol{J}(Y;\boldsymbol{\theta}|X=x)=0$ for each x. □

Lemma 3 (Data Processing Inequality):

$$\boldsymbol{J}(X;\boldsymbol{\theta}) \geq \boldsymbol{J}(\phi(X);\boldsymbol{\theta}) \tag{16}$$

with equality if $\phi(X)$ is a sufficient statistic relative to the family $\{f_{\boldsymbol{\theta}}(x)\}$, e.g., if $\phi(\cdot)$ is an invertible function.

Proof: By Lemma 2 we have

$$\boldsymbol{J}(\phi(X);\boldsymbol{\theta}) \leq \boldsymbol{J}(\phi(X),X;\boldsymbol{\theta}) = \boldsymbol{J}(X;\boldsymbol{\theta})$$

where the second equality follows from the chain rule (12) since $\phi(X)$ is deterministic given X thus $\boldsymbol{J}(\phi(X);\boldsymbol{\theta}|X)=0$. The inequality becomes equality if $\boldsymbol{J}(X;\boldsymbol{\theta}|\phi(X))=0$, i.e., if $\phi(X)$ is a sufficient statistic. □

[4] As pointed out by A. Yeredor, the proof follows even more easily from the equivalent formula for FI $J(X,Y;\theta) = -E\{\partial^2\ \ln\ f_\theta(X,Y)/\partial\theta^2\}$.

Lemma 4 (Parameter Transformation): Let $\{f_{\boldsymbol{\phi}}(\boldsymbol{x})\}$ be a family of densities parameterized by $\boldsymbol{\phi}$, and suppose that the vector of parameters $\boldsymbol{\phi} \in \mathcal{R}^n$ is a function of the vector $\boldsymbol{\theta} \in \mathcal{R}^m$. Then

$$\boldsymbol{J}(X;\boldsymbol{\theta}) = \left(\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{\theta}}\right)^t \cdot \boldsymbol{J}(X;\boldsymbol{\phi}(\boldsymbol{\theta})) \cdot \left(\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{\theta}}\right) \tag{17}$$

where

$$\partial \frac{\boldsymbol{\phi}}{\partial \boldsymbol{\theta}} = \left\{\frac{\partial \phi_i}{\partial \theta_j}\right\}, \qquad i = 1 \cdots n, j = 1 \cdots m$$

denotes the matrix of partial derivatives of the vector function $\boldsymbol{\phi}(\boldsymbol{\theta})$. In the linear case $\boldsymbol{\phi} = Q^t\boldsymbol{\theta}$, where Q is some $m \times n$ matrix, we have

$$\boldsymbol{J}(X;\boldsymbol{\theta}) = Q \cdot \boldsymbol{J}(X;\phi = Q^t\boldsymbol{\theta}) \cdot Q^t. \tag{18}$$

Proof: By the chain rule for derivatives, we have

$$\frac{\partial \ \ln \ f_{\boldsymbol{\phi}(\boldsymbol{\theta})}(x)}{\partial \boldsymbol{\theta}} = \left(\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{\theta}}\right)^t \frac{\partial \ \ln \ f_{\boldsymbol{\phi}}(x)}{\partial \boldsymbol{\phi}}$$

which when substituted in the definition of $J(X;\boldsymbol{\theta})$ (4) results in (17). □

We now define a linear model relating the estimated parameters $\boldsymbol{\theta}$ with their measurements $\boldsymbol{X}$ and $\boldsymbol{Y}$. Let the random vector $\boldsymbol{N} \in \mathcal{R}^n$ have some joint density $f_{\boldsymbol{N}}$, and let Q and P be some $m_\theta \times n$ and $m_y \times n$ matrices, respectively, where $m_\theta \le m_y \le n$. Let $\boldsymbol{\theta} \in \mathcal{R}^{m_\theta}$ be a vector of parameters, and consider the linear model

$$\boldsymbol{X} = Q^t \cdot \boldsymbol{\theta} + \boldsymbol{N} \quad \text{and} \quad \boldsymbol{Y} = P \cdot \boldsymbol{X} \tag{19}$$

where $\boldsymbol{X} \in \mathcal{R}^n$ and $\boldsymbol{Y} \in \mathcal{R}^{m_y}$.

Lemma 5: For the model in (19)

$$\boldsymbol{J}(\boldsymbol{X};\boldsymbol{\theta}) = Q\boldsymbol{J}(\boldsymbol{N})Q^t \quad \text{and} \quad \boldsymbol{J}(\boldsymbol{Y};\boldsymbol{\theta}) = QP^t\boldsymbol{J}(P\boldsymbol{N})PQ^t \tag{20}$$

where $\boldsymbol{J}(\boldsymbol{N})$ and $\boldsymbol{J}(P\boldsymbol{N})$ are the $n \times n$ and the $m_y \times m_y$ FI matrices under a translation parameter (8) of the vectors $\boldsymbol{N}$ and $P\boldsymbol{N}$, respectively.

Proof: Let $\phi = Q^t\boldsymbol{\theta}$. Using Lemma 4 we have

$$\boldsymbol{J}(\boldsymbol{X};\theta) = \boldsymbol{J}(\phi + \boldsymbol{N};\boldsymbol{\theta}) = Q\boldsymbol{J}(\phi + \boldsymbol{N};\phi)Q^t = Q\boldsymbol{J}(\boldsymbol{N})Q^t$$

where the last equality follows from the definition of FI under a translation parameter (8). The expression for $\boldsymbol{J}(\boldsymbol{Y})$ follows similarly. □

Lemma 6: For the model in (19)

$$\boldsymbol{J}(\boldsymbol{X};\boldsymbol{\theta}) \ge \boldsymbol{J}(\boldsymbol{Y};\boldsymbol{\theta}). \tag{21}$$

Proof: The inequality follows from the data processing inequality (16) since $\boldsymbol{Y} = P\boldsymbol{X}$. □

Lemmas 5 and 6 above imply three inequalities regarding the effect of linear transformation on the FI under translation:

Corollary 1:

a) For any $m \times n$ matrix A and random vector $\boldsymbol{N} \in \mathcal{R}^n$

$$A^t\boldsymbol{J}(A\boldsymbol{N})A \le \boldsymbol{J}(\boldsymbol{N}). \tag{22}$$

b) For any $m \times n$ matrix Λ with orthonormal rows, i.e., $\Lambda\Lambda^t = I_{m\times m} \triangleq$ the $m \times m$ identity matrix, and any $\boldsymbol{N}$

$$\boldsymbol{J}(\Lambda\boldsymbol{N}) \le \Lambda\boldsymbol{J}(\boldsymbol{N})\Lambda^t. \tag{23}$$

c) For any $m \times n$ matrix A with a full row rank, and random vector $\boldsymbol{N} \in \mathcal{R}^n$ with nonsingular FI matrix $\boldsymbol{J}(\boldsymbol{N})$

$$\boldsymbol{J}(A\boldsymbol{N}) \le (A\boldsymbol{J}(\boldsymbol{N})^{-1}A^t)^{-1} \tag{24}$$

with equality if $m = n$ or if $\boldsymbol{N}$ is Gaussian.

Remark: Inequality (24) was shown in [13] and [16] for the case where the components of $\boldsymbol{N}$ are statistically independent.

Proof:

a) Assume $m_\theta = 1$. Combining (20) and (21) implies
$$Q \cdot [\boldsymbol{J}(\boldsymbol{N}) - P^t\boldsymbol{J}(P\boldsymbol{N})P]Q^t \ge 0$$
for any $1 \times n_\theta$ vector Q. This implies that the inner difference is a nonnegative definite matrix.

b) Substitute $m_\theta = m_y = m$ and $Q = P = \Lambda$ in (20), use $\Lambda\Lambda^t = I_{m\times m}$, and substitute in (21).

c) Substitute $m_\theta = m_y = m,\ P = A$, and
$$Q = (A\boldsymbol{J}(\boldsymbol{N})^{-1}A^t)^{-1}A\boldsymbol{J}(\boldsymbol{N})^{-1}$$
in (20), and substitute in (21) to obtain the inequality. The sufficient conditions for equality follow from (9) and (10), noting that $\mathrm{COV}\,(A\boldsymbol{N}) = A\mathrm{COV}\,(\boldsymbol{N})A^t$. See Proposition 3 in the next section for a necessary and sufficient condition for $N_1 \cdots N_n$ statistically independent. □

In the model above the vector $\boldsymbol{N}$ was arbitrary. We now specialize to a vector with *independent components* and give our main result.

Theorem 1 (FII): Let N_1 and N_2 be statistically independent random variables. Then

$$J(N_1 + N_2) \le (J(N_1)^{-1} + J(N_2)^{-1})^{-1} \tag{25}$$

with equality if N_1 and N_2 are Gaussian. The inverse of the left- and right-hand sides of (25) are equal to the CRB's when a single parameter θ is estimated from $X_1 + X_2$ and from (X_1, X_2), respectively, where $X_i = \alpha_i\theta + N_i$ and

$$\alpha_i = 1 - \frac{J(N_i)}{J(N_1) + J(N_2)}, \qquad i = 1, 2.$$

Proof: Since N_1 and N_2 are independent, $\boldsymbol{J}([N_1, N_2])$ is a 2×2 diagonal matrix whose diagonal elements are $J(N_1)$ and $J(N_2)$. Hence the inequality and the equality condition follow by substituting $m = 1$, $n = 2$, and $A = [1, 1]$ in (24).

The interpretation of $J(N_1 + N_2)^{-1}$ and $J(N_1)^{-1} + J(N_2)^{-1}$ as CRB's can be seen by substituting $P = [1, 1]$ and $Q = [\alpha_1, \alpha_2]$ in the linear model (19), and then using (20) to calculate the corresponding CRB's $J(X_1 + X_2;\ \theta)$ and $J(X_1,\ X_2;\ \theta)$. □

We thus proved that the FII follows from the Fisher information data processing inequality. Furthermore, the FII corresponds to the loss in CRB due to "filtering" in a certain linear additive-noise model for parameter estimation. This loss is due to the non-Gaussianity of the noise and vanishes if the noise is Gaussian. A certain drawback in this alternative derivation of the FII is that the necessity of the equality condition does not follow easily, and requires some additional effort. See Proposition 3 in the next section.

III. Additional Results Regarding the Matrix Form of the FII

For completeness, we review below additional results regarding the matrix form of the FII, some of which appeared elsewhere. We start with a convex-matrix form of the FII (11) which was presented in [12] and [15].

Proposition 1 (Convex-Matrix Form of FII): Let

$$\boldsymbol{N} = (N_1, \cdots, N_n)$$

be a vector with statistically independent components, and $\Lambda = \{\lambda_{i,j}\}$ be an $m \times n$ matrix with orthonormal rows, i.e., $\Lambda\Lambda^t = I_{m\times m}$. Then

$$\mathrm{trace}\,\{\boldsymbol{J}(\Lambda\boldsymbol{N})\} \le \sum_{i=1}^{m}\sum_{j=1}^{n} \lambda_{i,j}^2 J(N_j). \tag{26}$$

In particular, if $J(N_1) = \cdots = J(N_n) = J_N$, the inequality becomes $(1/m)\ \mathrm{trace}\{\boldsymbol{J}(\Lambda\boldsymbol{N})\} \le J_N$, since the orthonormality of the rows implies

$$\sum_{j=1}^{n} \lambda_{i,j}^2 = 1, \qquad \text{for } i = 1 \cdots m.$$

Proof: Since $\boldsymbol{N}$ is an independent vector, the matrix $\boldsymbol{J}(\boldsymbol{N})$ is diagonal

$$\boldsymbol{J}(\boldsymbol{N}) = \mathrm{diag}\,[J(N_1), \cdots, J(N_n)].$$

Thus the $i,\ i$th component of the matrix $\Lambda\boldsymbol{J}(\boldsymbol{N})\Lambda^t$ in the right-hand side of (23) is given by

$$\sum_{j=1}^{n} \lambda_{ij}^2 J(N_j).$$

Taking the trace of both sides of (23) results in (26). □

We turn to a convex-matrix version of the Entropy-Power Inequality (EPI) which also appears in [12] and [15].

Proposition 2 (Convex-Matrix-EPI): Let $\boldsymbol{N}$ and Λ be as in Proposition 1 above. Then

$$h(\Lambda\boldsymbol{N}) \ge \sum_{i=1}^{m}\sum_{j=1}^{n} \lambda_{i,j}^2 h(N_j) \tag{27}$$

where $h(\cdot)$ denotes (joint) differential entropy.

Proof: The proof is based on the convex-matrix form of the FII in Proposition 1 and the De-Bruijn identity, and follows the line of the proof of Theorem 7 in [8, p. 1509]. □

From (27) it is easy to obtain the matrix form of the EPI presented in [16]; see also [12] and [15].

We end this section with a special case of (24), regarding a linear transformation of an *independent* vector, for which we can prove a necessary and sufficient condition. The proof can be found in [13].

Proposition 3 (Necessary and Sufficient Condition for Equality in the Matrix-FII for an Independent Vector): Let A be an $m \times n$ matrix $m \le n$. Let $\mathcal{I}_R(A)$ denote the set of indices $j \in \{1 \cdots n\}$ for which n_j is uniquely determined by $A\boldsymbol{n}$, and let $\mathcal{I}_0(A)$ denote the set of indices of the all-zero columns of A. Assume that $\boldsymbol{N} \in \mathcal{R}^n$ is a random vector with independent components, each having finite FI. Then, inequality (24), $\boldsymbol{J}(A\boldsymbol{N})^{-1} \ge A\boldsymbol{J}(\boldsymbol{N})^{-1}A^t$, holds with equality if and only if N_j is Gaussian for all $j \notin \mathcal{I}_R(A) \cup \mathcal{I}_0(A)$.

The necessary condition of Proposition 3 asserts that the *joint* FI can be used as a "contrast" (or "objective") function for blind deconvolution/signal separation [5], [9].

Clearly, if $m < n$ and A does not have all-zero columns, then there is at least one index j not in $\mathcal{I}_R(A)$. We thus have the following corollary of Proposition 3.

Corollary 2 If the $m \times n$ matrix A does not have all-zero columns, and $N_1 \cdots N_n$ are independent and identically distributed (i.i.d.) random variables with $J(N_i) = J_N$, then $J(A\boldsymbol{N})^{-1} \ge J_N^{-1}AA^t$, with equality if and only if $m = n$ or the N_i's are identical Gaussians.

IV. Parameter Estimation from Prefiltered Measurements

Some insight into the problem of parameter estimation in the presence of *non-Gaussian* measurement noise can be gained by investigating the properties of the FI in the model $\boldsymbol{X} = Q^t\boldsymbol{\theta} + \boldsymbol{N}$ and $\boldsymbol{Y} = P\boldsymbol{X}; \boldsymbol{\theta} \in \mathcal{R}^{m_\theta}; \boldsymbol{X}, \boldsymbol{N} \in \mathcal{R}^n; \boldsymbol{Y} \in \mathcal{R}^{m_y}$ given in (19). In many practical situations, a large number of noisy measurements $X_1 \cdots X_n$ is taken to estimate a small number of parameters $\theta_1 \cdots \theta_m$. For example, the linear relation $\boldsymbol{X} = Q^t\boldsymbol{\theta} + \boldsymbol{N}$ may represent a radar application, where $X_1 \cdots X_n$ are the outputs of an n-element phased array which observes the targets $\theta_1 \cdots \theta_m$, where $n \gg m$; here the entry q_{ij} of the matrix Q is the (complex) gain of the jth element towards the ith target [11]. Another possible application is that of AR parameter estimation, where a long "training" sequence $X_1 \cdots X_n$ is modeled concisely by

$$X_j = \theta_1 X_{j-1} + \cdots + \theta_m X_{j-m} + N_j, \qquad j = 1 \cdots n.$$

Here $q_{ij} = X_{j-i}$.

In the applications above $\boldsymbol{Y} = P\boldsymbol{X}$ represents a prefiltered version of the measurements, from which we wish to estimate $\boldsymbol{\theta}$. The matrix P thus represents a *noninvertible* linear operation. For instance, we may perform such a prefiltering operation to reduce the dimensionality and hence the complexity of the (possibly nonlinear) estimator. See [11], [2], and the references therein. A case of interest would be when the matrix P is a projection matrix (i.e., $PP^t = I_{m_y \times m_y}$) whose rows span the row space of Q (i.e., $Q = GP$ for some $m \times m$ matrix G). This guarantees that in the noisless ($\boldsymbol{N} \equiv 0$) case, $\boldsymbol{Y}$ contains the same information about $\boldsymbol{\theta}$ as $\boldsymbol{X}$. Furthermore, it makes sense to assume that the noise samples $N_1 \cdots N_n$ are i.i.d., and that the n columns of Q have a fixed, say, unit norm, i.e.,

$$\sum_{i=1}^{m_\theta} q_{ij}^2 = 1 \forall j.$$

(The latter condition corresponds to a fixed total power gain per each element in the radar phased array application above.)

Under these quite natural assumptions, one may wonder how does the quality of the estimation vary with n and with the noise properties, and how much do we loose (do we?) by applying the projection operation P prior to estimation.

In order to isolate the effect of the signal-to-noise ratio (SNR) on the performance, we introduce a noise gain parameter α, and consider the model

$$\boldsymbol{X} = Q^t\boldsymbol{\theta} + \alpha\boldsymbol{N} \quad \text{and} \quad \boldsymbol{Y} = P\boldsymbol{X}.$$

Using Lemmas 5 and 6, and the identity $QP^tPQ^t = QQ^t$ implied by the model above, it is easy to prove that for any $m_\theta \le m_y \le n$

$$\boldsymbol{J}(\boldsymbol{X};\boldsymbol{\theta}) = \frac{J_N}{\alpha^2} QQ^t$$

and

$$\boldsymbol{J}(\boldsymbol{Y};\boldsymbol{\theta}) = QP^t\boldsymbol{J}(P\boldsymbol{N})PQ^t \le \frac{J_N}{\alpha^2} QQ^t \tag{28}$$

where J_N is the FI of the (i.i.d.) noise samples. Furthermore, since

$$\sum_{i=1}^{m_\theta} q_{ij}^2 = 1, \qquad \text{for } j = 1 \cdots n$$

it follows that

$$n\frac{\sigma_N^{-2}}{\alpha^2} \le \mathrm{trace}\,\{\boldsymbol{J}(\boldsymbol{Y};\boldsymbol{\theta})\} \le \mathrm{trace}\,\{\boldsymbol{J}(\boldsymbol{X};\boldsymbol{\theta})\} = n\frac{J_N}{\alpha^2} \tag{29}$$

where σ_N^2 is the variance of the noise, and the lower bound follows from (10). Equality holds in both inequalities in (29) if N is Gaussian (in which case $J_N = \sigma_N^{-2}$). Furthermore, for any noise distribution, the right inequality holds with equality if $n = m_y = m_\theta$. Otherwise the inequalities are strict; see Corollary 2.

One simple implication of (29) is that without prefiltering the total FI increases *linearly with the number of the measurements*. The same is true even after projection by P if the measurement noise is Gaussian. However, for non-Gaussian noise projection causes a loss of FI whenever $m_y < n$. Thus from the FI/CRB point of view, the

optimal estimator cannot be decomposed into projection followed by nonlinear processing.

This phenomena can be explained by the tendency towards Gaussianity of the sum of independent random variables. Projection, which is a noninvertible linear transformation, makes the residual noise *more Gaussian* and thus less favorable for estimation. A similar phenomena causes *increase of entropy* after noninvertible filtering [9], [16].

In [14], we suggested another way to interpret (29), namely, as an accuracy–quantity tradeoff relation. Notice that $1/\alpha^2$ represents the accuracy (or the resolution) of the measurements. Thus without prefiltering, keeping the quantity/accuracy product n/α^2 fixed keeps the FI constant. The same is true for a Gaussian noise even *after* (appropriate) prefiltering, but not true when the noise is not Gaussian. Thus if prefiltering (projection) is used prior to estimation in the presence of a non-Gaussian noise, *it is better to take few accurate measurements than many noisy ones*.

ACKNOWLEDGMENT

My joint work with Meir Feder on information-theoretic inequalities formed the basis for this correspondence. I also wish to thank Toby Berger, Jean-Francois Cardoso, Hagit Messer, Nadav Shulman, Yossef Steinberg, Tony Weiss, and Arie Yeredor for helpful discussions.

REFERENCES

[1] A. R. Barron, "Entropy and the central limit theorem," *Ann. Prob.*, vol. 14, no. 1, pp. 336–342, 1986.
[2] A. Bartov and H. Messer, "Lower bound on the achievable DSP performance for localizing step-like continuous signals in noise," *IEEE Trans. Signal Processing*, to be published.
[3] N. M. Blachman, "The convolution inequality for entropy powers," *IEEE Trans. Inform. Theory*, vol. IT-11, pp. 267–271, 1965.
[4] R. E. Blahut, *Principles and Practice of Information Theory*. Reading, MA: Addison Wesley, 1987.
[5] J.-F. Cardoso, "Blind signal separation: A review," *Proc. IEEE*, submitted for publication.
[6] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[7] A. Dembo, "Information inequalities and uncertainty principles," Dept. Statistics, Stanford Univ., Tech. Rep. 75, Stanford, CA, July 1990.
[8] A. Dembo, T. M. Cover, and J. A. Thomas, "Information theoretic inequalities," *IEEE Trans. Inform. Theory*, vol. 37, pp. 1501–1518, Nov. 1991.
[9] D. Donoho, "On minimum entropy deconvolution," *Applied Time Series Analysis II*. NY: Academic, 1981, pp. 565–608.
[10] H. L. Van Trees, *Detection Estimation and Modulation theory (Part I)*. New York: Wiley, 1968.
[11] A. J. Weiss and B. Friedlander, "Preprocessing for direction finding with minimal variance degradation," *IEEE Trans. Signal Processing*, vol. 42, pp. 1478–1485, June 1994.
[12] R. Zamir, "Universal encoding of signals by entropy coded dithered quantization," Ph.D. dissertation, Tel-Aviv Univ., Tel–Aviv, Israel, 1994.
[13] ——, "A necessary and sufficient condition for equality in the matrix Fisher-information-inequality," Tech. Rep., Tel Aviv Univ., Dept. Elec. Eng.–Syst., 1997.
[14] ——, "It is better to take few accurate measurements than many noisy ones," in *Proc. Information Theory Workshop on Information Theory, Multiple Access and Queueing* (St. Louis, MO, Apr. 1995), p. 35.
[15] R. Zamir and M. Feder, "A generalization of information theoretic inequalities to linear transformations of independent vector," in *Proc. 6th Joint Swedish–Russian Int. Workshop on Information Theory* (Molle, Sweden, Aug. 1993), pp. 254–258.
[16] ——, "A generalization of the Entropy Power Inequality with applications," *IEEE Trans. Inform. Theory*, vol. 39, pp. 1723–1727, Sept. 1993.

Zero-Error Capacity for Models with Memory and the Enlightened Dictator Channel

Rudolf Ahlswede, Ning Cai, and Zhen Zhang, *Senior Member, IEEE*

***Abstract*— We present a general class of zero-error capacity problems with memory covering known cases such as coding for error correction and many new cases. This class can be incorporated into a model of channels with memory, which thus are shown to give a unification of a multitude of seemingly very different coding problems. In this correspondence we analyze a seemingly basic channel in this class.**

***Index Terms*— Finite-state channels, independence number, memory, 0-error capacity.**

I. INTRODUCTION

A GENERAL 0-ERROR PROBLEM WITH MEMORY

For the space $\mathcal{Z}^n$ of words of length n over alphabet $\mathcal{Z}$, there are several interesting graphs $\mathcal{G} = (\mathcal{Z}^n, \mathcal{E}_n)$ with vertex set $\mathcal{Z}^n$ and an edge set $\mathcal{E}_n$ reflecting string properties.

Examples are, the strong graph product (Shannon's product graph) and the case $\mathcal{Z} = \{0,1\}$ with $(x^n, x'^n) \in \mathcal{E}$ if and only if (iff) for no two components s,t $x_s = 1 \neq x'_s$ and $x_t = 0 \neq x'_t$.

The product space structure makes it particularly interesting to investigate $\alpha(\mathcal{G}_n)$, the maximal size of cocliques, as a function of n. Then the coclique of the graph in the first example is Shannon's well-known zero-error code and the coclique of the graph in the second example is the well-known Sperner system or antichain (c.f., e.g., [6, Ch. 1]). We propose here a quite general class of such problems, which we term "0-error ∞-memory capacity problems," because they generalize Shannon's well-known zero-error capacity problems and concentrate on a new aspect, namely, memory. Those problems arose for instance in [2].

Definition We call any pair of words from $\mathcal{Z}^\ell$ a separator and any set $\mathcal{S} \subset (\mathcal{Z}^\ell)^2$ of pairs of words of length ℓ a set of separators.

For any $n \geq \ell$ we consider the associated graph $\mathcal{G}^n_{\mathcal{S}} = (\mathcal{Z}^n, \mathcal{E}(\mathcal{S})_n)$, where $(x^n, x'^n) \in \mathcal{E}(\mathcal{S})_n$ iff for no $(s^\ell, s'^\ell) \in \mathcal{S}$ there is an index set $I = \{i_1, \cdots, i_\ell\} \subset \{1, \cdots, n\}$ with $x_{i_j} = s_j$, $x'_{i_j} = s'_j$ $(i_1 < i_2 < \cdots < i_\ell)$.

In the examples above $\mathcal{S}$ is symmetric, that is, $(s^\ell, s'^\ell) \in \mathcal{S}$ implies $(s'^\ell, s^\ell) \in \mathcal{S}$. Here the graphs can be viewed as undirected graphs. In the sequel we assume that $\mathcal{S}$ is symmetric. Thus $\mathcal{S}$ can be viewed as a set of unordered pairs of subsequences.

This covers also t-error correcting codes for $\mathcal{S} = \{(0^{2t+1}, 1^{2t+1})\}$.

.

II. CONSECUTIVE SEPARATING PAIRS

Another associated graph $\mathcal{G}^*_{\mathcal{S}} = (\mathcal{Z}^n, \mathcal{E}^*(\mathcal{S})_n)$ is obtained by limiting I in the previous definition to intervals in $\{1, 2, \cdots, n\}$. $\mathcal{S}$ plays here the role of a set of *consecutive* separating pairs of words of length ℓ. Here the problem is to find a maximal $\mathcal{C} \subset \mathcal{Z}^n$ such that for all $c^n, c'^n \in \mathcal{C}$ there is an $\{\alpha, \beta\} \in \mathcal{S}$ and an $i \in \{1, 2, \cdots, n - \ell + 1\}$ such that

$$\{(c_i, c_{i+1}, \cdots, c_{i+\ell-1}), (c'_i, c'_{i+1}, \cdots, c'_{i+\ell-1})\} = \{\alpha, \beta\}.$$

Manuscript received December 12, 1996; revised November 10, 1997.

R. Ahlswede and N. Cai are with Fakultät für Mathematik, Universität Bielefeld, 33501 Bielefeld, Germany.

Z. Zhang is with the Department of Electrical Engineering–Systems, University of Southern California, Los Angeles, CA 90089 USA.

Publisher Item Identifier S 0018-9448(98)02370-0.