Mutual Kolmogorov-Sinai Entropy Approach to Nonlinear Estimation

Bing-Pei Wu
Department of Control Engineering
National Chiao Tung University
Hsinchu, Taiwan

Edmond A. Jonckheere
Department of Electrical Engineering and Systems
University of Southern California
Los Angeles, CA 90089-2533

Abstract
For a general nonlinear estimation problem, we develop an up-
per bound on the correlation coefficient in terms of the mutual
Kolmogorov-Sinai entropy. The upper bound may be reached by
means of a nonlinear transformation that, after transformation,
the processes are jointly gaussian.

1 Introduction
The mutual KS entropy between two processes is introduced in this pa-
per. The amount of mutual information between two processes can be
measured by the mutual KS entropy which is related to the singular val-
ues of the canonical correlation coefficient matrix. In the general estima-
tion problem, the mutual entropy is a measure of predictability of an un-
known sequence $y$ given a process $x$. We want to maximize the mutual
entropy between $x$ and $y$ to get the largest information. We will show that,
if the mutual KS entropy is finite and the correlation coefficient matrix
$C_{xy}$ is invertible, the global maximum of the mutual KS entropy over all
probability distributions $P(x,y)$ can be achieved by a jointly gaussian
density probability function $f_{xy}$, and the global maximum of the mutual
KS entropy is $-\log 2$.

If we assume that $y$ is a stationary white noise and $x$ is a stationary process,
we will show that the mutual KS entropy $K_{xy}$ is equal to the mutual
information $I(x,y)$, which is a measure of predictability from $x$ to $y$.

Moreover, we can give an algorithm in order to find the global maximum of
the mutual KS entropy. The algorithm is based on a method of ascent.
In this paper, we will present an algorithm for finding the global maximum
of the mutual KS entropy for the general estimation problem.

2 Background
Consider the experiment of observing $x_k$, where $x_k$ is constrained to lay in a
set $S$. Define the discrete sample space

$$
\mathcal{X} = \{x_k : x_k = 1, 2, \ldots \mid x_k \in S, k = 1, 2, \ldots \}
$$

where $S$ can be finite or countable.

A sequence of experimental data can be represented by a dynamical system
$(\{\tilde{X}_k\}, \tilde{S})$, where $\tilde{S}$ is the product of $S$ and the $x_k$-sequence.
This is the $x_k$-sequence $\{\tilde{X}_k\}$ which is a measure of points in the
sequence $\{x_k\}$ and to emphasize the latter sequence, we will, with a slight abuse, write the
above dynamical system as $(\{\tilde{X}_k\}, \tilde{S})$.

In this paper, we developed the connection between the KS entropy and
Shannon entropy for stochastic processes. We will discuss the relation-
ship between the mutual KS entropy and the mutual-Shannon Entropy
in the following section.

2.1 Mutual KS Entropy and Mutual Shannon Entropy
Let $X$ and $Y$ be two stationary random variables with zero means,
$X = (x_1, x_2, \ldots)$ and $Y = (y_1, y_2, \ldots)$, and with discrete sample space
$\mathcal{S}_X \times \mathcal{S}_Y$, respectively. Fix $S^x$ and $S^y$, respectively.第一部 process $X$ and $Y$ have
zero mean and finite variance.

Let $T_i^x$ be the $i^{th}$ moment of $X$, and $T_i^y$ be the $i^{th}$ moment of $Y$.

The mutual KS entropy $K_{xy}$ is equal to the mutual information $I(x,y)$,
which is a measure of predictability from $x$ to $y$.

Therefore, we will show that the mutual KS entropy $K_{xy}$ is equal to the
mutual information $I(x,y)$.

In this paper, we will present an algorithm for finding the global maximum
of the mutual KS entropy for the general estimation problem.

Throughout this paper, $K(1, \varepsilon)$ refers to entropy and entropy rate, re-
spectively.

*Research supported in part by NSF Grant ECS-9112088.

$\begin{align*}
R(1) & = K(1, \varepsilon) \\
R(\varepsilon) & = K(1, \varepsilon) - K(1, \varepsilon + \varepsilon)
\end{align*}$

The mutual KS entropy between two partitions is defined as

$$
R(\varepsilon) = \sum_{i=1}^{n} \log (a_i) / a_i
$$
\[ E \left( X \mid Y \right) = \sum_{x} \left( \sum_{y} p_{X|Y}(x|y) \log \frac{p_{X|Y}(x|y)}{p_{Y}(y)} \right) = \mathcal{H}(X) - \mathcal{H}(X \mid Y) = H(Y) - H(Y \mid X) \]

We define the mutual entropy between \( X \) and \( Y \) in the sense of Kolmogorov and Stamp as

\[ H(X;Y) = \sum_{x} \sum_{y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \]

We choose the following two-partitions, \( C \in \mathcal{C} \) and \( C' \in \mathcal{C}' \), such as \( C \) is said to be overrating in the sense that

\[ \sup_{Y} H(\mathcal{C} \mid \mathcal{X}) \leq H(\mathcal{C} \mid \mathcal{Y}^{*}) \]

From Jourdain and Wu, we have equality between the Kolmogorov-Sinai and Shannon entropies.

\[ H(\mathcal{C} \mid \mathcal{X}) = X(e_{1}, \ldots, e_{n}) \]

The following equality is derived in Billingsley [1, pp. 144-146].

\[ \sup_{Y} H(fl, \mathcal{C} \mid \mathcal{X}) = \mathcal{H}(x_{1}, x_{2}, \ldots, x_{n}) \]

By Eq. (3), we get

\[ H(\mathcal{C} \mid \mathcal{X}) = H(\mathcal{C} \mid \mathcal{Y}) = H(\mathcal{C} \mid \mathcal{Y}^{*}) \]

Thus, by Eq. (4)-(7), we obtain

\[ H(X;Y) = I(X;Y) \]

Consider the case of stationary distribution on the continuum. The continuum distribution is approximated by a discrete distribution. From the definition of \( I(X;Y) \), we have

\[ p(x_{1}, x_{2}, \ldots, x_{n}, \ldots) \rightarrow p(x_{1}, x_{2}, \ldots, x_{n}, \ldots) \times \frac{\Delta x_{1} \times \Delta x_{2} \times \ldots}{\Delta x_{1} \times \Delta x_{2} \times \ldots} \times \frac{\delta(x_{1})}{\Delta x_{1} \times \Delta x_{2} \times \ldots} \times \frac{\delta(x_{2})}{\Delta x_{1} \times \Delta x_{2} \times \ldots} \times \frac{\delta(x_{n})}{\Delta x_{1} \times \Delta x_{2} \times \ldots} \times \ldots \]

where \( \delta(x) \) is the probability density function, assuming that it exists.

Shannon defined the mutual entropy between \( X \) and \( Y \) with continuous distribution as follows:

\[ I(X;Y) \]

\[ \mathcal{H}(X) = \mathcal{H}(Y) = \mathcal{H}(X \mid Y) = \mathcal{H}(Y \mid X) = \mathcal{H}(X;Y) = \mathcal{H}(Y;X) \]

\[ \mathcal{H}(X;Y) = \mathcal{H}(X) + \mathcal{H}(Y) - \mathcal{H}(X \mid Y) - \mathcal{H}(Y \mid X) \]

Thus, for \( X \) and \( Y \), we can define

\[ \bar{E} \geq \mathcal{H}(X;Y) = \mathcal{H}(X) + \mathcal{H}(Y) - \mathcal{H}(X \mid Y) - \mathcal{H}(Y \mid X) \]

Theorem 1

Let \( Y \) and \( Z \) be two stationary stochastic processes and let \( \mathbb{E}[Y_{i}Z_{j}] \) be the correlation coefficient of \( Y_{i} \) and \( Z_{j} \) for \( i \) and \( j \) be two different values. The mutual entropy defined by Shannon is equivalent to the one defined by Kolmogorov and Stamp.

\[ I(X;Y) = \mathcal{H}(X) + \mathcal{H}(Y) - \mathcal{H}(X \mid Y) - \mathcal{H}(Y \mid X) \]

Theorem 2

The mutual entropy is achieved whenever the two random variables \( X \) and \( Y \) are jointly Gaussian, that is, \( \mathcal{H}(X;Y) = \mathcal{H}(X) + \mathcal{H}(Y) - \mathcal{H}(X \mid Y) - \mathcal{H}(Y \mid X) \).
where \( P(x,y) \) is the joint probability density function of \( X \) and \( Y \).

Proof. We prove this result by the Calculus of Variations. Let us define the augmented functional

\[
I = \int \left[ F(X,Y)P(X,Y) \right] dX dY
\]

where

\[
F(X,Y) = P(X) - P(Y) + P(X)P(Y) - \lambda_1\sum_{x,y} P(x,y) + \lambda_2\sum_{x} P(x) + \lambda_3\sum_{y} P(y)
\]

\[
= X^2 + \lambda_1 X + \lambda_2 Y + \lambda_3 Y^2 + \lambda_4 Y + \lambda_5 X Y + \lambda_6 X Y^2 + \lambda_7 Y X^2 + \lambda_8 Y X^3
\]

where \( \lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6, \lambda_7, \lambda_8 \) are constant vectors, and \( \lambda_1, \lambda_2, \lambda_3, \lambda_4 \) are constant matrices. These \( \lambda_1, \lambda \) and \( \lambda_4 \) are Lagrange multipliers.

By the Euler-Lagrange Equation [13, pp. 423-424], \( \frac{\partial F}{\partial P} = 0 \), we obtain the necessary condition for optimality.

\[
P_{\text{opt}} = P_{\text{opt}} \exp \left\{ \frac{1}{2} \sum \lambda_i \lambda_i^T \right\}
\]

(17)

where

\[
\lambda_i = \begin{bmatrix} \lambda_1 & \lambda_2 & \lambda_3 & \lambda_4 \end{bmatrix}
\]

The last step in the derivation follows from the integral

\[
\int \left( \frac{1}{2} \sum \lambda_i \lambda_i^T \right) dX dY = \frac{1}{2} \sum \lambda_i \lambda_i^T
\]

(18)

Consider the following gaussian joint density function as candidate solution:

\[
P_{\text{Gaussian}} = \frac{1}{(2\pi)^{1/2}} \exp \left\{ -\frac{1}{2} \sum \lambda_i \lambda_i^T \right\}
\]

where

\[
\lambda_1 = \begin{bmatrix} \frac{\partial F}{\partial X} \\ \frac{\partial F}{\partial Y} \end{bmatrix}
\]

and

\[
\frac{\partial F}{\partial X} = \begin{bmatrix} \lambda_1 X + \lambda_2 Y + \lambda_3 Y^2 + \lambda_4 Y + \lambda_5 X Y + \lambda_6 X Y^2 + \lambda_7 Y X^2 + \lambda_8 Y X^3 \end{bmatrix}
\]

(19)

Theorem 3: Cramér-Wold Theorem:

There is no information loss under a nonlinear homeomorphic, which, by definition, transforms the coordinate in a one-to-one continuous as a continuously invertible function, because the mutual entropy is independent of the coordinate system. Therefore, for the mutual entropy in the sense of Shannon, we have \( H(X|Y) = H(X|g(Y)) \). Now, let us go back to the definition of the mutual Kolmogorov-Joint entropy. The gap in the definition is taken over all partitions \( \mathcal{E} \). The supremum ensures that the mutual entropy is also invariant under nonlinear homeomorphism.

\[ H(X|Y) = H(X|g(Y)) \]

This is in full agreement with \( H(X|Y) = H(Y|X) \). Thus, the right-hand side of Eq.(18) can’t be changed by nonlinear transformation. However, it does change the canonical correlation coefficient matrix \( R_{XY} \), i.e., the left-hand side of Eq.(18) is changed. Therefore, if we set \( \lambda(X|Y) \) and \( g(Y) \) to be jointly gaussian by nonlinear transformations \( f \) and \( g \), then the left-hand side of Eq.(18) reduces to the mutual.

\[ H(X|Y) = H(X|g(Y)) = H(Y|X) \]

(20)

3.1 An Example

The following example shows how to compute the mutual KS entropy between \( x \), uniformly distributed over \([0,1]\), and \( y \), zero-mean, unit variance, gaussian distribution.

We use the two variables having correlation coefficient \( \rho \), where \( \rho \) is the zero-mean, \( \rho \) is the standard deviation, and \( \rho \) is the correlation coefficient. Let \( \rho = 0 \).

To show that this specific \( P_{\text{Gaussian}} \) minimizes \( I(X,Y) \), we need to prove that the second variation \( \delta^2 \) relative to a perturbed solution \( P(x,y) \) is positive for this gaussian joint density function where \( \delta \) is constant vector absolute value is taken small and \( X,Y \) are any order free-discrete function such that \( f(X,Y) \) is one-to-one continuous function. Applying this to our problem, we obtain

\[ I'_{\text{opt}} = \int F(X,Y) P_{\text{opt}} dX dY
\]

(21)

where

\[ F(X,Y) = \frac{1}{2} \sum \lambda_i \lambda_i^T
\]

(22)

Computing the section order derivative, it follows that

\[ P = \frac{1}{P_{\text{opt}}} \exp \left\{ -\frac{1}{2} \sum \lambda_i \lambda_i^T \right\}
\]

Therefore, the P in Eq.(18) minimizes \( I(X,Y) \) and the minimum of \( I(X,Y) \) is given by Eq.(12).

We need to show that this minimum is the global minimum. It is easily observed that \( I(X,Y) \) is a convex function of \( P_{\text{opt}} \) which is defined on \([0,1] \times [0,1] \) convex set. Also, the constraint functions are all convex in \( P_{\text{opt}} \). So, \( I(X,Y) \) is convex. Our problem, then, becomes a convex programming problem [4, p. 216]. Convex functions have a special property. A convex function may have many stationary points, however, the optimum of these stationary points is the global optimum. In general, however, the stationary function, the local minimum is also the global minimum [8, p. 191]. Therefore, since \( \lambda \in \mathbb{R} \), we conclude that \( -\ln \det (-D^2 F(X,Y)) \) is the global minimum of \( I(X,Y) \).

Corollary 1

If \( X \) and \( Y \) are stationary random processes with continuous distributions, then we have the following relationship subject to the constraints of Eq.(18)-(15)

\[ H(Y|X) = \frac{1}{2} \ln \left( 1 - \left( 1 - H(X|Y) \right) \right) \]

(23)

The equality is achieved if \( X \) and \( Y \) are jointly gaussian.

Proof. The theorem follows from Theorem 2, Theorem 1(b).

Lemma 1

Let \( z \) be a uniformly distributed random variable on the unit interval, \( p \) be a zero-mean, zero-variance gaussian random variable, and \( \rho \) be the correlation coefficient. The joint probability density function of \( x \) and \( y \), diagonal equivalent to Gaussian, is as follows:

\[ P_{\text{Gaussian}} = \frac{1}{\pi} \exp \left\{ -\frac{1}{2} \left( 1 - \rho^2 \right) x^2 \right\}
\]

(24)

where

\[ \rho = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)} \sqrt{\text{var}(Y)}}
\]

(25)

It is observed that \( H(Y|X) = 1 - H(Y|p) \) if \( \rho = 0 \). Therefore, for scalar random variables and \( p \), we can easily obtain the mutual KS entropy of \( X \) and \( Y \) by transferring \( p \) into \( P(X,Y)p \), respectively, such that \( \rho = 0 \) and \( \rho \) are jointly normal.

Lemma 2

The extreme values for functions \( X \) and \( Y \) are the following:

\[ P_{\text{Gaussian}} = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left( x^2 - \rho \right) \right\}
\]

(26)

Theorem 4

Theorem 4: Theorem of Mutual Entropy

There is no information loss under a nonlinear homeomorphic, which, by definition, transforms the coordinate in a one-to-one continuous as a continuously invertible function, because the mutual entropy is independent of the coordinate system. Therefore, for the mutual entropy in the sense of Shannon, we have \( H(X|Y) = H(X|g(Y)) \). Now, let us go back to the definition of the mutual Kolmogorov-Joint entropy. The gap in the definition is taken over all partitions \( \mathcal{E} \). The supremum ensures that the mutual entropy is also invariant under nonlinear homeomorphism.

\[ H(X|Y) = H(X|g(Y)) \]

This is in full agreement with \( H(X|Y) = H(Y|X) \). Thus, the right-hand side of Eq.(18) can’t be changed by nonlinear transformation. However, it does change the canonical correlation coefficient matrix \( R_{XY} \), i.e., the left-hand side of Eq.(18) is changed. Therefore, if we set \( \lambda(X|Y) \) and \( g(Y) \) to be jointly gaussian by nonlinear transformations \( f \) and \( g \), then the left-hand side of Eq.(18) reduces to the mutual.

\[ H(X|Y) = H(X|g(Y)) = H(Y|X) \]

(27)
where $A \triangleq I - D'D$, $G : R^n \rightarrow R^d$ defined by $X \equiv G(U)$, and $D \equiv DA_{-1/2}$.

Let $u$ and $x$ be the $i$-th entry of $U$ and $X$, respectively. Thus, $u_i = \frac{1}{\sqrt{2\pi}^d} \int \exp(-\frac{1}{2} (x_i - E)^2) dx_i$.

$$D = DA_{-1/2}U = \sqrt{\frac{1}{2}} D_{XX}U$$ \hspace{1cm} (23)

Eq (23) is derived from direct computation of the canonical correlation co-efficients matrix. And the maximum singular value of $D$ is less than or equal to $1$. In vector case, $X, Y, F(X)$ denote $G(U)$ and $G(U) \equiv Y$. The initial KS estimate is

$$H(X,Y) = \frac{1}{2} \ln(1 - \frac{1}{2} D_{XX} D_{YY})$$

### 3.2 An Approach to maximize $-\frac{1}{2} \ln(1 - D_{XX} D_{YY})$

The problem is now: For dimensionally equivalent Gaussian processes $u, y$, we need to develop an approach to solve $D_E = \frac{\partial \mathcal{L}}{\partial u}$ where $u$ and $y$ are zero-mean, unit variance jointly gaussian random variables. We can always transform $u$ to be a standard uniform random variable over $[0, 1]$, i.e.,

$$u \overset{d}{=} F(u)$$

where $F(u)$ is the probability distribution function of $u$ and is invertible.

Again, consider the following transformation where $u$ is derived from gaussian $v$.

$$u \overset{d}{=} G(v)$$

$G$ is invertible too. Then $F(x) = u \overset{d}{=} G(v)$ implies that $v = G^{-1}(F(x))$.

So, the desired nonlinear transformation $f(x)$ is

$$f(x) = G^{-1}(F(x))$$

Similarly, $g(y) = G^{-1}(F(y))$.

Thus, $R(x,y) = \frac{1}{2} \ln(1 - \frac{1}{2} D_{xx} D_{yy})$. In the notation of Eq (21), $G(x) \equiv v$.

For example, if $x$ is the Gaussian distribution with positive constant $c$, rather than being uniformly distributed, we want to transform $e$ to $f(x)$ such that $f(x)$ is a gaussian:

$$F(x) = \int_{-\infty}^{c} \exp(-\frac{1}{2} x^2) dx = 1 - e^{-c^2}$$

Then $f(x) = \sqrt{-\ln(1 - e^{-c^2})}$.

For the vector case $X, Y$, the approach is the same. $F(x), F(y)$ are of vector form and $G(v), G(w)$ are of.

Thus:

$$f(x) \equiv G^{-1}(F(x))$$

For simplicity, we define two functions $G_1^{-1}(x) : R^n \rightarrow R^d$ and $G_2^{-1}(x) \rightarrow R^n$ which transform Gaussian diagonally equivalent distributions $X$ and $Y$ to be gaussian, respectively.

$$G_1^{-1}(x) \overset{d}{=} G_1^{-1}(G_2^{-1}(y))$$

and each entry of $G_1^{-1}(x)$ and $G_2^{-1}(y)$ is a zero-mean and unit-correlated gaussian random variable. If $X$ is uniform, then $G_1^{-1}(x)$ reduces to $G_1^{-1}(x)$, $f(x)$ and $g(y)$ are zero-mean and unit-correlated. However, we can transform $f(x)$ and $g(y)$ to possess arbitrary properties as desired by nonstationary linear transformation and translation. Therefore, $f(x)$ and $g(y)$ maximizing $-\frac{1}{2} \ln(1 - D_{xx} D_{yy})$ could be chosen as follows and are still gaussian.

$$f(x) = \frac{1}{\sqrt{2\pi}^d} \int \exp(-\frac{1}{2} (x_i - E)^2) dx_i$$

where $E$ and $B$ are nonconstant and $g$ and $h$ are the expected values of $f(x)$ and $g(y)$, respectively.

$$D_{xx} D_{yy} \overset{d}{=} D_{xx} D_{yy}$$

is invariant under nonstationary linear transformation and translation. Thus, $f(x)$ and $g(y)$ maximizing $-\frac{1}{2} \ln(1 - D_{xx} D_{yy})$ is achieved as is Eq (24).

### 4 Linear Estimation

Given a sequence of data $(x_1, x_2, \ldots, x_M, y_1, y_2, \ldots, y_M)$, define $X = (x_1, x_2, \ldots, x_M)$, $Y = (y_1, y_2, \ldots, y_M)$, the future data, both with zero mean, and let $X_{XX}, X_{XY}, X_{YY}$ denote the correlation $XX$, $XY$, $YX$, and $YY$, respectively. Can we optimally predict $Y$ from $X$ in the minimum weighted-mean-square (WMS) error sense?

In general, the best estimate $\hat{Y}$ of $Y$ given $X$ is the linear estimate $E[Y|X]$. However, this estimate is not easy to compute because conditional probability is necessary. Since the singular value of $D_{xx}$ are invariant under linear nonstationary transformation of $X$ and $Y$, linear functions of $X$ and $Y$ and ($x, y$), are considered to improve our ability to predict $Y$. So, consider the linear function $f(X) = \phi(x) \psi(y)$, then the WMS error of an approximative linear estimate of $\phi(Y) \psi(x)$ given $f(X)$ is defined as

$$E[(\phi(Y) - \phi(\hat{Y}))^2 \mid \phi(x) \mid \psi(y)]$$

where $\phi(Y) \psi(x)$ is the function of $f(X)$.

Since $\phi$ is a function of $f(X)$, then

$$E[(\phi(Y) - \phi(\hat{Y}))^2 \mid \phi(x) \mid \psi(y)] = A(f(X))$$

Thus, from the above discussion, we obtain

$$\phi(Y) \psi(x) = \max_{f(x)} E[(\phi(Y) - \phi(\hat{Y}))^2 \mid \phi(x) \mid \psi(y)]$$

where the maximum is taken over all measurable functions $f$ and $g$ with $E[g(Y)] = 0$.

### 5 Nonlinear Solvers for Maximizing $D_{xx} D_{yy}$

We consider the real scalar case first i.e., $x, y \in R, f(x)$ and $g(y)$ are both real scalar random variables and $D_{xx} D_{yy}$ is real. The problem is to find $f$ and $g$ such that

$$\phi, \psi \max_{f, g} D_{xx} D_{yy} = \max_{f, g} \int f(x)g(y)f(x)g(y)dx \, dy$$

under the following constraints

$$\int f(x)g(y)dx = 0, \quad \int f(x)^2 dx = 1, \quad \int g(y)^2 dy = 1, \quad \int f(x)g(y)dx = 0$$

This problem can be solved by the Calculus of Variations. Let's define the quantity $L_{ij}$ by the integral

$$L_{ij} \equiv \int f(x)g(y)dx$$

where

$$L_{ij} \equiv \int f(x)g(y)f(x)g(y)dx \, dy + 2h_jf(x)g(y) dx + 2h_jf(x)g(y) dy$$

is clearly the segmental functional and $L_{ij}$ and $L_{ij}$ are Lagrange multipliers.

By Euler-Lagrange Equation [2], pp. 527-433, $\frac{\partial L}{\partial f(x)} = 0$, we get

$$\frac{\partial L}{\partial f(x)} = 0, \quad h_j = 0$$

Integrating Eq (25) for $X$ on both sides, we obtain, via Eq (21) of [2], $h_j = 0$. Again, multiplying Eq (25) by $\phi(x)$ and integrating for $X$, we get $h_j = -\cdot \phi(x)$. Rewriting Eq (25) as

$$\frac{\partial L}{\partial g(y)} = 0, \quad \phi(x) = 0$$

To show that this specific $f(x)$ maximizes $D_{xx} D_{yy}$, we use Bola's criteria [2, p. 40] introduced in Section 2. Since $\phi(x) = 0$, $f(x) = -\cdot \phi(x) \neq 0$.

It follows that $f(x)$ maximizes $D_{xx} D_{yy}$.

Certainly, $h_j$ can be rewritten as follows

$$L_{ij} = \int G(Y) \phi(x)$$

where

$$G(Y) = \int f(x)g(y)f(x)g(y)dx) + 2h_jf(x)g(y) dx + 2h_jf(x)g(y) dy$$

Interchanging the roles of $X$ and $Y$, $f(x)$ and $g(y)$, and following the same approach as above, we get

$$L_{ij} = \int G(Y) \phi(x)$$
\[ f(X) \leq f(Y) = 0 \] (36)

Similarly, the \( f(Y) \) maximizes \( D_{TV} \). Subtracting Eq. (35) from Eq. (34) to eliminate \( f(Y) \), it turns out that the nonlinear programming function \( f(X) \) is given by the following integral operator eigenvalue problem:

\[ K_0(X) \leq \int K(X,Y)K(X)^2 = \lambda(X) \] (36)

where the kernel function \( K(X,Y) \) of the operator \( K \) is defined as

\[ K(X,Y) \leq \int_{P(Y \setminus P(X))} P(X,Y) dX \] (37)

Eq. (36) is the necessary condition for \( f(X) \) to locally maximize \( D_{TV}(X,Y) \) in Eq. (26). Since \( D_{TV}(X,Y) \) is a convex function of \( f(X) \) and the mixed functions \( f(X,Y) \) and Eq. (26) are also convex in \( f(X) \), it follows, by the argument of the proof of Theorem 2, that this is also a convex programming problem. Therefore, the local maximum is the same as the global maximum.

Similarly, we have another integral equation yielding \( g(Y) \):

\[ f(Y) \leq \int K(X,Y)K(Y)^2 = \lambda(Y) \] (38)

where the kernel \( K(X,Y) \) is defined as

\[ K(X,Y) \leq \int_{P(X \setminus P(Y))} P(Y,X) dY \] (37)

\[ D_{TV}(X,Y) \) is also a convex function of \( f(Y) \) and the constraint functions of Eq. (31) are also convex in \( g(Y) \). Similar to the case of \( f(X) \), \( g(Y) \), the solution of Eq. (38), also globally maximizes Eq. (26). Therefore, we conclude that we can globally maximize \( D_{TV}(X,Y) \) by the integral equation \( g(Y) \) which satisfy Eq. (32) and Eq. (38), respectively.

Solve the integral equation \( K_0(X) \) as \( f(X) \) and \( K_0(Y) = \lambda(Y) \), then \( f(X) \) is an eigenfunction of \( f(X) \) and \( K_0(X) \) and \( f(Y) \) are the eigenfunctions of \( K(X,Y) \) and \( f(Y) \), respectively. We obtain the same result as Luttrell and Brillouin [7], but via a different approach. Moreover, we have also shown that these \( f(X) \) and \( g(Y) \) globally maximize the correlation coefficient.

### 6 Solutions of Integral Equations

In this section, we assume that \( f(X,Y) \) is convex the vectors \( X,Y \) to scales, respectively. To solve these two equations, consider the following four examples:

(i) \( X \) and \( Y \) are independent then, by Eq. (37), \( K(X,Y) = P(X) \). So \( K_0(X) = P(X) \). That is, the estimation error can't be reduced on matter what kinds of nonlinear functions we use, if \( X \) and \( Y \) are independent.

(ii) \( X \) and \( Y \) are jointly normal with zero mean and correlation \( 
\begin{align*}
\hat{P}(X) &= \frac{1}{2} \sqrt{\hat{P}(X)} \exp \left[ \frac{1}{2} \left( D_{TV}(X,Y) - D_{TV}(X,Y) \right) \right] \\
\end{align*}
\]

where \( D_{TV}(X,Y) \) is the total variation distance between \( X \) and \( Y \). This is the same as Eq. (31) for the case of independent \( X \) and \( Y \).

(iii) \( f(X) \) is a maximum singular value of \( D_{TV}(X,Y) \) and \( g(Y) \) as the singular value of \( D_{TV}(X,Y) \). If \( f(X) = g(Y) \), then \( K(X,Y) = \lambda(X) \) and \( K(Y,X) = \lambda(Y) \).

Therefore, the maximum of Eq. (32) can be achieved if we choose

\[ \mu = \nu \] (41)

that is, the maximum singular value of \( D_{TV}(X,Y) \) and the corresponding eigenvalue for \( g(Y) \). Applying the boundary conditions in Eq. (31) (32), we need
coefficient matrix $D_k(Y)$ to jointly Gaussian processes $f(X)$ and $g(Y)$ with the marginal correlation coefficient $r(X,Y)$. 

Proof. Let $X,Y$ be scalars or vectors. Let $u$ and $v$ be zero-mean, unit variance, jointly Gaussian random variables with correlation coefficient $r_{uv}$ and let $u,v$ be the linear transforms that can be generated by $x = g(x)$ and $y = g(y)$ wherever $g(x)$ and $g(y)$ are invertible. Thus,

$$P_u = P_{g(x)} \quad \text{and} \quad P_v = P_{g(y)}$$

where $P_u$ is the joint probability density function of the Gaussian random variables $u$ and $v$, which is known and is of closed form. We have

$$K(x,x') = \int P_u(x)P_v(x')dx = \int P_{g(x)}(x)P_{g(x')}(x')dx$$

Since $P_u$, $P_v$, $P_u$, and $P_v$ are given, it follows that

$$K(x,x') = \frac{1}{\sqrt{2\pi \sigma_1^2}} \exp \left\{ -\frac{(x'-\mu_1)^2}{2\sigma_1^2} \right\} \quad \text{for} \quad f(x) = \frac{1}{\sqrt{2\pi \sigma_2^2}} \exp \left\{ -\frac{(x'-\mu_2)^2}{2\sigma_2^2} \right\}$$

Therefore, for $x = G_1(x)$ and $u = G_1(x)$, the first integral equation becomes

$$K(x,x') = \frac{1}{\sqrt{2\pi \sigma_1^2}} \exp \left\{ -\frac{1}{2\sigma_1^2} (x'-\mu_1)^2 \right\} f(u)(x')dx' = \tilde{A}(x')f(x)$$

Obviously, the solutions are

$$f(x) = f(G_1(x)) = \frac{1}{\sqrt{2\pi \sigma_1^2}} \exp \left\{ -\frac{1}{2\sigma_1^2} (x'-\mu_1)^2 \right\} \quad \text{for} \quad u = G_1(x) = \mu_1$$

Similarly, we can solve the second integral equation

$$K(y,y') = \frac{1}{\sqrt{2\pi \sigma_2^2}} \exp \left\{ -\frac{1}{2\sigma_2^2} (y'-\mu_2)^2 \right\} f(v)(x')dx' = \tilde{A}(y')f(y)$$

$$f(y) = f(G_2(y)) = \frac{1}{\sqrt{2\pi \sigma_2^2}} \exp \left\{ -\frac{1}{2\sigma_2^2} (y'-\mu_2)^2 \right\} \quad \text{for} \quad v = G_2(y) = \mu_2$$

$$\Rightarrow f(x) = f(G_1(x)) = \frac{1}{\sqrt{2\pi \sigma_1^2}} \exp \left\{ -\frac{1}{2\sigma_1^2} (x'-\mu_1)^2 \right\}$$

Thus, $f(x)$ and $f(y)$ are jointly Gaussian with correlation coefficient $r$. Since $\sigma$ is the maximum correlation coefficient, it follows that the correlation coefficient $r_{uv}$ is reduced by the transformations $G_1(x)$ and $G_2(y)$.

The vector case $X,Y$ for these two integral equation kernels is as follows:

$$D = D_{g(X,Y)} \quad \text{and} \quad D' = D_{g(X,Y)}'$$

$$K(X,X') = \int \exp \left\{ -\frac{1}{2\sigma_1^2} (x'-\mu_1)^2 \right\} f(G_1(x'))dx'$$

$$K(Y,Y') = \int \exp \left\{ -\frac{1}{2\sigma_2^2} (y'-\mu_2)^2 \right\} f(G_2(y'))dx'$$

$$\Rightarrow f(x) = f(G_1(x)) = \frac{1}{\sqrt{2\pi \sigma_1^2}} \exp \left\{ -\frac{1}{2\sigma_1^2} (x'-\mu_1)^2 \right\}$$

Therefore, after some manipulations, we have the following integral equations:

$$K(x,x') = \int \exp \left\{ -\frac{1}{2\sigma_1^2} (x'-\mu_1)^2 \right\} f(G_1(x'))dx'$$

$$\Rightarrow f(x) = f(G_1(x)) = \frac{1}{\sqrt{2\pi \sigma_1^2}} \exp \left\{ -\frac{1}{2\sigma_1^2} (x'-\mu_1)^2 \right\}$$

$$\Rightarrow f(x) = f(G_1(x)) = \frac{1}{\sqrt{2\pi \sigma_1^2}} \exp \left\{ -\frac{1}{2\sigma_1^2} (x'-\mu_1)^2 \right\} \quad \text{for} \quad u = G_1(x) = \mu_1$$

$$\Rightarrow f(y) = f(G_2(y)) = \frac{1}{\sqrt{2\pi \sigma_2^2}} \exp \left\{ -\frac{1}{2\sigma_2^2} (y'-\mu_2)^2 \right\} \quad \text{for} \quad v = G_2(y) = \mu_2$$

Therefore, we conclude that Eq. (43) and Eq. (44) are the precise solutions of two integral equations Eq. (26) and Eq. (36), respectively.

As a remark, observe that we don't need to solve these two complicated integral equations. What we need to do is to transform the processes $X$ and $Y$ to Gaussian processes $f(X)$ and $g(Y)$ by Eq. (43) and Eq. (44), respectively. If this can be done, i.e., if $X$ and $Y$ are diagonal equivalent to Gaussian, a procedure to make $G_1(x)$, $G_2(y)$ $[Y]$ Gaussian is given in Section 3.2.

7 Nonlinear Estimation

We extend the problem of Section 5 to vector nonlinear functions, $f(x)$ and $g(y)$, which have $n$ components rather than one component, to maximize the correlation. We need to maximize the $D_{g(x),g(y)}(X,Y)$, instead of $D_{g(x),g(x)}(X,X)$, for nonlinear $g$. Let $u = f(x)$ and $v = g(y)$ by step based on the method of finding $f$ and $g$ in case they are linear. However, our approach, based on the Colonial of Variations, allow us to use $f$ and $g$ in vector form. No sequential steps are necessary.

For a given positive integer $n$, find the vector functions $f(x)$ and $g(y)$, which have $n$ components rather than one component, to maximize the correlation of $D_{g(x),g(y)}(x,x)$, which is achieved, that is,

$$\Rightarrow \max_{f(x),g(y)} D_{g(x),g(y)}(X,Y)$$

subject to the constraints

$$\int f(x)g(x)dx = 0, \quad \int f(x)g' (x)dx = 0$$

$$\int f(x)g(x)dx = 1, \quad \int f(x)g' (x)dx = 1$$

This problem can be solved by the Colonial of Variations. Let's define the quantity $t_n$ by the integral

$$t_n = \int f(x)g(x)dx$$

where

$$f(x), g(x) = \begin{cases} f(x), & f(x) \text{ is a constant multiplier and } \lambda_0 \text{ is a constant vector multiplier} \\
0, & \text{otherwise} \end{cases}$$

Using the same approach as in Section 5, we have

$$t_n\Rightarrow D_{g(x),g(y)}(x,x) = \lambda_0$$

where

Note that $D_{g(x),g(x)}(X,Y)$ is a constant related to $X$ and $Y$. To show that this specific integral equation is equivalent to Eq. (26) in Section 6.1, we need to use some results from Section 6.1. Let $f(x)$ and $g(y)$ be $n$ components, and let $f(x)$ and $g(y)$ be $n$ components. Then, we have,

$$D_{g(x),g(y)}(X,Y) = \lambda_0$$

Therefore, we have the following integral equation

$$\Rightarrow \int f(x)g(x)dx - D_{g(x),g(y)}(x,x) = \lambda_0$$

Using the same approach as in Section 5 but substitute it to $f(x)$. It follows that $f(x)$ maximizes $D_{g(x),g(y)}(X,Y)$ if and only if

$$\Rightarrow \int f(x)g(x)dx = D_{g(x),g(y)}(x,x) = \lambda_0$$

Substituting Eq. (46) into Eq. (43), we have, after some manipulation, the "decoupled" equation

$$\Rightarrow \lambda_0$$

$$\Rightarrow \int f(x)g(x)dx = D_{g(x),g(y)}(x,x) = \lambda_0$$

Using the singular value decomposition of $D_{g(x),g(y)}(x,x)$, we have

$$= \lambda_0$$
Then, Eq. (31) becomes

$$\kappa_i(x) = \int k(x,x')k(x'x'')d\mu(x'') = \xi^2(x)$$  

(52)

where \( R(x,x') = \Xi(x) \) and \( \Xi = \text{diag}(\xi_1, \ldots, \xi_r) \). Therefore

$$\kappa_i(x) = \int k(x,x')k(x'x'')d\mu(x'') = \xi^2(x) = 1, \ldots, r$$

where \( \xi_i \) is the \( i \)-th element of the vector \( \xi \).

The sum of the eigenvalues of Eq. (52) is the maximum of \( \mu_k(x,x')\mu_k(x,x'')d\mu(x'') \) and the eigenvalues of Eq. (56) is the nonlinear function \( k(x,x') \) max-

imizes the constraint of Eq. (44) (44). Therefore, without loss of generality, we can assume that \( \Xi = \xi \). On the other hand, we have another integral equation.

$$
\kappa(x,y) \equiv \int k(x,x')k(x'y')d\mu(x')d\mu(y') \\
\Xi^2(x) \Xi^2(y) \\
\Xi^2(xy) = \Xi^2(xy) \\
(53)

Without loss of generality, \( \Xi = \xi \). The nonlinear function \( \mu_k(x,x') \) maximizes the constraint of Eq. (44) and is the eigenfunction of Eq. (56).

Similar to the argument used for \( \kappa_i(x) = \xi_i^2(x) \), we get a similar component of \( \kappa_i(x,y) \), respectively, \( \Xi^2(x) \) and \( \Xi^2(y) \), respectively, \( \Xi^2(xy) \). Therefore, the maximum of \( \mu_k(x,x') \) over all \( \mu_k(x,y') \) is given by \( \Xi^2(x) \) and \( \Xi^2(y) \) are known, using the singular value decomposition, we can get

$$
\kappa_i(x) = \Xi^2(x) \\
\Xi^2(y) = \Xi^2(xy) \\
(54)

where \( \Xi = \text{diag}(\xi_1, \ldots, \xi_r) \).

From Eq. (44) and Eq. (45), it follows that

$$
\kappa_i(x) = \mu_i^2(x) \\
\mu_i^2(y) = \mu_i^2(xy) \\
(55)

where \( \mu_i \) is the \( i \)-th column of \( U \) and \( \mu_i \) is the \( i \)-th column of \( V \). Therefore, the maximum of \( \mu_k(x,x') \) over all \( \mu_k(x,x'') \) is given by \( \Xi^2(x) \). Similar to Section 3, \( X \) and \( Y \) are not restricted to be non-negative and unit-covarianced.

If \( \kappa \equiv \text{dim}(X) \) is odd, then \( X \) and \( Y \) are just non-

negative linear transformations of \( \mu^2_1(X), \mu^2_1(Y) \), respectively. Furthermore, since \( X(x,y) \) and \( Y(y,x) \) are non-negative and unit-covarianced, we have non-

negative linear transformations and translations on Eq. (55), respectively, in that

$$
\kappa_i(x) = \mu_i^2(x) \\
\mu_i^2(y) = \mu_i^2(xy) \\
(56)

where \( \mu_i \) and \( \mu_j \) are non-negative \( \mu_i \) and \( \mu_j \) are the expected value of \( x \) and \( y \), respectively. Therefore, we can let \( \mu_i \) be the means of \( X \) and \( \mu_j \) be the \( k \)-th element of \( X \). Similarly, \( \mu_i^2 = \mu_i^2 \Xi \Xi^2 \).

It follows that \( \mu_i^2(x) \) and \( \mu_i^2(xy) \) and the singular values of these functions. Comparing Eq. (31) with Eq. (52) in Section 3, we find that \( \kappa_i(x) \) and \( \kappa_i(xy) \) are obtained in the same case but not the same; however, the singular values of the non-

negative linear covarianced matrices are same. That is, we can use the same \( k(x,x') \), \( k(x',x'') \), and \( \mu_i \) to evaluate the non-

negative linear covarianced over \( X, Y, \) and \( X, Y \), respectively.

8 \( H(X,Y) \) and MMSE Error

Lemma 3

If \( 0 \leq \varepsilon \leq 1 \), then

$$
-\ln(1 - \varepsilon^2) \geq \frac{1}{2} \log \frac{1}{\varepsilon^2}
$$

where the equality holds for \( \varepsilon = 0 \).