ON LIKELIHOOD-RATIOS, MUTUAL INFORMATION AND ESTIMATION ERROR FOR THE ADDITIVE GAUSSIAN CHANNEL

Moshe Zakai

Department of Electrical Engineering
Technion–Israel Institute of Technology

July ’04
The additive Gaussian channel

\[ y'(\omega) = \rho x'(\omega) + n(\omega) \]

detection, estimation of parameters, filtering, smoothing, communication, smoothing of images on \([0, T] \times [0, T]\).
1) detection \( \mu_W \) no signal \( \ell(y_0^T) = \frac{d\mu_Y}{d\mu_W}(y_0^T) \) \( \mu_Y \) signal present

2) estimation error \( \hat{x}_t = E(x_t|y_0^t) \) causal \( \hat{x}_t = E(x_t|y_0^T) \) non-causal

a) \( E(x_t - \hat{x}_t)^2, E \int_0^T (x_t - \hat{x}_t)^2 dt; E(x_t - \hat{x}_t)^2, E \int_0^T (x_t - \hat{x}_t)^2 dt \)

b) The mutual information between \( x \) and \( y \): \( I(x; y) \)

\[
I(x; y) = E \log \frac{d\mu_{x;y}}{d(\mu_x \times \mu_y)}
\]
The purpose of this talk is to present relations between $\ell(y_0^T)$ and $\hat{x}$ and between $E(x - \hat{x})^2$ and $I(x; y)$ in a general setup.

Early results:

$$\ell(y_0^T) = \exp \left\{ \int_0^T \hat{x}_t(y) dy_t - \frac{1}{2} \int_0^T (\hat{x}_t)^2 dt \right\}$$

$$I(x; y) = \frac{1}{2} \int_0^T E(x_t - \hat{x}_t)^2 dt$$
The Abstract Wiener space (AWS)

The classical Wiener process

1. \( W = \{ w(t), \quad t \in [0, 1]; \text{continuous, } w(0) = 0 \} \)

\[ \left\{ W, \sup_{t \in [0,1]} |w(t)| \right\} \text{ is a Banach space} \]

2. The Cameron-Martin space \( H \) (Hilbert): \( h = \int_0^\cdot h'_s \, ds, \)

\[ |h|_H^2 = \int_0^1 (h'_s)^2 \, ds \]

The Wiener integral: \( \delta h = \int_0^1 h'_s \, dw_s = N(0, |h|_H^2). \)

3. \( W^* = \{ e \} \hookrightarrow H \hookrightarrow W. \)
The AWS (L. Gross)

(i) $W = \{w\}$ be separable Banach space

(ii) $W^* = \{e\}$ the dual space to $W$.

(iii) The Cameron-Martin space $H$: a Hilbert space such that

$$W^* \hookrightarrow H \hookrightarrow W$$

Then (L. Gross) there exists a probability $\mu$ on $W$ such that,

$$\delta e = \langle w, e \rangle_{W^*} \equiv N(0, |e|_H^2).$$

By Ito-Nisio if $e_i$ is a C.O.N.B. on $H$ then:

$$\left| w - \sum_{1}^{n} \langle w, e_i \rangle e_i \right|_W \rightarrow 0$$
Recall some pathologies of the classical Wiener process:
Continuous but nowhere differentiable; quadratic variation

\[ \mu_W \perp \mu_{\alpha W} \quad |\alpha| \neq 1 \]

The Ito calculus was able to “overcome” these pathologies by introducing the assumption of “adapted integrands”.

The Malliavin Calculus (or “the stochastic calculus of variations”) is based on the Cameron Martin theorem: \( \mu_W \sim \mu_{W+h} \) (\( h \) non random)
iff \( h \in \) Cameron Martin space then

\[
\frac{d \mu_{W+h}}{d \mu_W}(w) = \exp \left( \delta h - \frac{1}{2}|h|^2_H \right)
\]
A crash course in the Malliavin calculus

The Gradient Operator

Let $S = \{ f(w) = \varphi(\delta e_1, \cdots, \delta e_n), n = 1, 2, \cdots \}$

$\varphi \in C_b^\infty(\mathbb{R}^n); \quad e_i \in W^*, \quad \delta e = W^* \langle e, w \rangle_W$

For any $f \in S$, set

$$\nabla f = \sum_{1}^{n} \frac{\partial \varphi}{\partial i}(\delta e_1, \cdots, \delta e_n)e_i.$$ 

Then

$$\frac{df(w + \varepsilon e)}{d\varepsilon} \bigg|_{\varepsilon=0} = \sum_{1}^{n} \frac{\partial \varphi}{\partial i}(\delta e_1, \cdots, \delta e_n)(e_i, e) = (\nabla f, e)$$
Now, define the class $\mathcal{D}_{2,1}$ of functionals to be the closure of $S$ under the Sobolev norm:

$$\|f\|_{2,1} = E^{\frac{1}{2}} \left( |f(w)|^2 + |\nabla f|_{H}^2 \right)$$

Then

(a) $\nabla \psi(f(w)) = \psi'(f(w))\nabla f$

(b) $\nabla_h \delta e = \frac{d}{d\varepsilon} \langle e, w + \varepsilon h \rangle \big|_{\varepsilon=0}$

or

$$\nabla \delta e = \int_0^\cdot e' ds = e$$
The divergence (Skorohod integral)

Let $u(w)$ be an $H$-valued r.v. in $(W, H, \mu)$, $u$ will be said to be in the $\text{dom}_2 \delta$ if $E|u(w)|^2_H < \infty$ and there exists a random variable, say $\delta u$, such that for all smooth functionals $f(w) = \varphi(\delta e_1, \ldots, \delta e_n)$ then “integration by parts formula”

$$E(\nabla f, u)_H = E(f \cdot \delta u)$$

holds. $u \in \text{dom}_2 \delta$ iff for all $f \in S$

$$\left| E(\nabla f, u)_H \right| \leq K \cdot E^{\frac{1}{2}} |f|^2$$

($K$ may depend on $u$). $\delta u$ is called the divergence (S. integral) of $u$. 
The main properties of the divergence \((E(\nabla f, u)_H = E(f \cdot \delta u))\)

- \(\delta h = \langle h, w \rangle\)
- \(E \delta u = 0\)
- \(\delta(a(w)u(w)) = a(w)\delta u - (\nabla a, u)_H\)
- \(\langle u(w), w \rangle = \delta u + \text{trace} \nabla u\)
- \(E(\delta u)^2 = E|u_H|^2 - \text{trace}(\nabla u)^2\)
- If (on the classical Wiener space) \(u\) is adapted \(\delta u = \int_0^1 u'_s dw_s\) (Ito)

Note while \(\nabla \varphi\) is invariant under an abs continuous transformation of measure, \(\delta u\) is *not*.
Relations between the estimation error and the likelihood ratio

Let \( y = \rho x + w, \ y, w \in W \).

Notation: \( \mu_Y, \mu_X, \mu_{Y|X}, \mu_{X|Y} \)

Further assume that the \( H \)-valued r.v. \( x \) is assumed to be \( W^* \) valued, and \( \exp \alpha(x, h)_H \in L^1(\mu_X) \) for all real \( \alpha \) and all \( h \in W^* \).

By the Cameron-Martin theorem and since \( x \) and \( w \) are independent, we have

\[
\frac{d\mu_{Y|X}}{d\mu_W}(w) = \exp \left( \rho \langle w, x \rangle - \frac{\rho^2}{2} |x|^2_H \right), \quad w \in W
\]

Hence

\[
\ell(w) = \frac{d\mu_Y}{d\mu_W}(w) = \int_H \frac{d\mu_{Y|X}}{d\mu_W}(w, x) \mu_X(dx)
\]
Theorem: Under these assumptions it holds that

(a) a.s. \( \mu_W \)

\[ \nabla \ell = \rho \ell(w) \hat{x} \]

(b) \[ \left( \nabla^n \ell(w), h_1 \otimes \cdots \otimes h_n \right)_{H^\otimes n} = \ell(w) \rho^n \left( \prod_{i=1}^n (h_i, x) \right)^{\sim} \] \hspace{1cm} (1)

(c) in particular trace \( \nabla^2 \ell(w) \) exists and a.s. \( \mu_W \)

\[ \nabla^2_{h_1, h_2} \ell(w) = \rho^2 \ell(w) \left( (h_1, x) \cdot (h_2, x) \right)^{\sim} \]

\[ \nabla^2_{h, h} \log \ell(w) = \rho^2 \left( \left( (x, h)^2 \right)^{\sim} - (\hat{x}, h)^2 \right) \]

\[ \text{trace} \nabla^2 \log \ell = \rho^2 E |x - \hat{x}|^2_H \]

(d) \[ \left( \prod_{i=1}^n (h_i, x) \right)^{\sim} = (h_n, \hat{x}) \left( \prod_{i=1}^{n-1} (h_i, x) \right)^{\sim} + \nabla_{h_n} \left( \prod_{i=1}^{n-1} (h_i, x) \right) \hat{x} \]
Lemma: Assume that $\mu_Y$ and $\mu_{Y|X}$ are absolutely continuous with respect to $\mu_W$ then for all bounded and measurable functions $\psi$

$$
\int_{X \times W} \psi(x; y) \frac{d\mu_{Y|X}}{d\mu_W}(y, x) \mu_X(dx) \mu_W(dy)
= \int_{X \times W} \psi(x; y) \frac{d\mu_Y}{d\mu_W}(y) \mu_{X|Y}(dx, y) \mu_W(dy).
$$
Outline of the proof of the theorem:

\[ \nabla_h \ell(w) = \int_H \rho(\nabla_h \langle w, x \rangle) \exp \left( \rho \langle w, x \rangle - \frac{\rho^2}{2} |x|_H^2 \right) \mu_X(dx) \]

\[ = \int_H \rho(h, x)_H \exp \left( \rho \langle w, x \rangle - \frac{\rho^2}{2} |x|_H^2 \right) \mu_X(dx) \]

\[ = \int_H \rho(h, x)_H \frac{d\mu_Y|_X}{d\mu_W}(w, x) \mu_X(dx). \]

Thus, by Lemma

\[ \nabla_h \ell(w) = \int_X \rho(h, x) \frac{d\mu_Y}{d\mu_W}(w) \mu_X|_Y(dx, w) \]

\[ = \rho \ell(w)(h, \hat{x}) \]

proving (1).
by repeated differentiation:

\[ \left( \nabla^n \ell(w), h_1 \otimes \cdots \otimes h_n \right)_{H^\otimes n} = \rho^n \int_x \left( \prod_{i=1}^n (h_i, x) \right) \frac{d\mu_Y}{d\mu_W}(w) \mu_{X|Y}(dx, w), \]

which yields (2). Hence

\[ \nabla^2_{h_1, h_2} \ell(w) = \rho^2 \ell(w) \left( (h_1, x) \cdot (h_2, x) \right) \hat{\nabla} \]

\[ \nabla^2_{h, h} \log \ell(w) = \frac{1}{\ell(w)} \nabla^2_{h, h} \ell(w) - \left| \nabla_h \log \ell(w) \right|^2 \]

\[ = \rho^2 \left[ \left( (x, h)^2 \right)^{\hat{\nabla}} - \left( (x, h)^1 \right)^2 \right] \]

proving (3).

\[ \delta \hat{x} = \frac{1}{\rho} \delta \nabla \log \ell(w); \quad \tilde{\delta} \hat{x} = \delta \hat{x} - \rho \left| \hat{x} \right|^2_H \]
The Guo-Shamai-Verdu relation between the mutual information and the mean square of the estimation error

The mutual information between \( x \) and \( y \) is defined as

\[
I(X; Y) = \int_{H \times W} \log \frac{d\mu_{X;Y}}{d(\mu_X \times \mu_Y)}(x; y) \, \mu_{X;Y}(dx, dy).
\]

\( E_0 \) will denote expectation w.r. to the Wiener measure and \( E_1 \) will denote expectation w.r. to the measure on \( W \) induced by \( y \) (hence \( Ef(y) = E_1f(w) = E_0\ell(w)f(w) \)).

**Theorem**

\[
\frac{dI(X;Y)}{d\rho} = \rho E\left( |x|_H^2 - |\hat{x}|_H^2 \right) = \rho E| x - \hat{x} |_H^2
\]
Proof: \( \frac{d\mu_{X;Y}}{d(\mu_X \times \mu_Y)} = \frac{d\mu_Y|_X}{d\mu_W} \cdot \frac{d\mu_W}{d\mu_Y} \) hence

\[
I(X; Y) = \int_{H \times W} \left\{ \log \frac{d\mu_Y|_X}{d\mu_W}(x; y) - \log \frac{d\mu_Y}{d\mu_W}(y) \right\} \mu(dx, dy)
\]

\[
= E \left( \rho \langle y, x \rangle - \frac{\rho^2}{2} \frac{|x|}{H}^2 \right) - E \log \ell(w).
\]

Note that \( E \rho \langle y, x \rangle = \rho^2 E |x|_H^2 \), hence

\[
I(X; Y) = \frac{\rho^2}{2} E |x|_H^2 - E_1 \log \ell(w)
\]

and

\[
\frac{dI(X; Y)}{d\rho} = \rho E |x|_H^2 - \frac{d}{d\rho} E_0 \ell(w) \log \ell(w)
\]

\[
= \rho E |x|_H^2 - E_0 \left( \frac{d\ell(w)}{d\rho} \cdot \log \ell(w) \right) - 0.
\]
Now,

\[
\frac{d\ell(w)}{d\rho} = \int_X \left( \langle x, w \rangle - \rho |x|_H^2 \right) \frac{d\mu_Y|X(x)}{d\mu_W}(w) \mu_X(dx).
\]

By lemma

\[
\frac{d\ell(w)}{d\rho} = \int_X \left( \langle x, w \rangle - \rho |x|_H^2 \right) \frac{d\mu_Y}{d\mu_W}(w) \mu_X|Y(dx)
\]

\[
= \left( \langle \hat{x}, w \rangle - \rho |x|_H^2 \hat{\rangle} \right) \ell(w).
\]

which yields

\[
\frac{dI}{d\rho} = \rho E|x|_H^2 - E_0 \left\{ \left( \langle \hat{x}, w \rangle - \rho |x|_H^2 \right) \ell(w) \log \ell(w) \right\}. \quad (*)
\]
By similar calculations

\[ E_0 \ell \log \ell(\hat{x}, w) = E_0 \left( \frac{1}{\rho} \log \ell(\nabla \ell, w) \right) \]

by (\cdot) \quad \Rightarrow \quad E_0 \frac{1}{\rho} \log \ell(\delta \nabla \ell + \text{trace } \nabla^2 \ell) \]

\[ = \rho E_0 \ell(w)(\hat{x}, \hat{x}) + E_0 \rho \ell(w) \log \ell(w)(|x|^2) \]

Substituting into (*) yields

\[ \frac{dI}{d\rho} = \rho E|x|^2 - \rho E|\hat{x}|^2 + E_0 \rho \ell \log \ell(|x|^2) - E_0 \rho \ell \log \ell(|x|^2) \]

\[ \square \]
An extension of the main result

\[ \text{assume } m \perp_{x} w \]

\[ I(m, y) = I(x; y) - I(x; y | m) \]

\[ \frac{d}{d\rho} I(m, y) = \rho E|\mathbf{x} - \hat{x}|^2_H - \rho E|\mathbf{x} - \mathbf{E}(\mathbf{x}|y, \mathbf{m})|^2_H \]
A generalized version of the De Bruijn identity

The Fisher information matrix $J$ associated with a smooth probability density $p(y_1, \ldots, y_n), y \in \mathbb{R}^n$ is defined as

$$J = \left( \frac{\partial^2 \log p(y_1, \ldots, y_n)}{\partial y_i \partial y_j} \right)_{1 \leq i,j \leq n}$$

and then

$$E \text{ trace } J = -E \left\{ \left| \nabla \log p \right|^2_{\mathbb{R}^n} \right\}, \quad (**)$$

where $E$ is the expectation with respect to the $p$ density. The De Bruijn identity deals with the case where $y = x + \sqrt{t}w$ where $w = w_1, w_2, \ldots, w_n$ and the $w_j, \quad j = 1, \ldots, n$ are i.i.d. $N(0, 1)$ and $x$ is an $\mathbb{R}^n$ random variable independent of $w$. It states that

$$\frac{d}{dt} E \log p(y) = \frac{1}{2} E \left\{ \left| \nabla \log p(y) \right|^2_{\mathbb{R}^n} \right\}.$$
The Fisher information matrix cannot be extended directly to the case where \( y \) is infinite dimensional. However, the results here yield some similar relations, we have

\[
\frac{d}{d\rho} E_1 \log \ell(w) = \rho E_1 |\hat{x}|^2_H
\]

\[
= \frac{1}{\rho} E_1 |\nabla \log \ell(w)|^2_H,
\]

which may be considered a generalized De Bruijn identity. Note that \( E_1 \log \ell(w) \) is the relative entropy of \( \mu_Y \) relative to \( \mu_W \), also,

\[
\rho \frac{dI(x; y)}{d\rho} = E_1 \text{trace } \nabla^2 \log \ell(w) = \rho^2 E|x|^2_H - E_1 |\nabla \log \ell(w)|^2_H,
\]

which is different from (**) by the \( \rho^2 E|x|^2_H \) term.

[P.S. The insertion of time in the AWS].