On likelihood-ratios, mutual information and estimation error for the additive Gaussian channel

Moshe Zakai∗

Abstract

This paper considers the model of an arbitrary distributed signal $x$ observed in additive Gaussian noise $w$, $y = x + w$. New relations between the minimal mean square error of the non-causal estimator and the likelihood ratio between $y$ and $w$ are derived. These results are applied to prove an extended version of a recently derived relation between the mutual information $I(x,y)$ and the minimal mean square error. The derivation of the presented results is based on the Malliavin calculus.

Keywords: Mutual information, Gaussian channel, minimal mean square estimation error, relative entropy, Malliavin calculus.

May 2, 2004

*Department of Electrical Engineering, Technion–Israel Institute of Technology, Haifa 32000, Israel. Email: zakai@ee.technion.ac.il
1 Introduction

Let \( w_t, 0 \leq t \leq T \) denote the standard \( d \)-dimensional Wiener process and \( w'_t \) the related white noise. The white noise channel is, roughly speaking, defined by \( y'(t) = x'(t) + w'_t, 0 \leq t \leq T \) where \( x'(t) \) is a signal independent of the white noise process \( w'_t \).

In the context of detection theory, the key entity is \( \ell(y) \), the likelihood ratio, i.e. the Radon-Nikodym derivative of the measure induced by the \( \{y'(t), t \in [0,T]\} \) process with respect to the measure induced by the white noise \( w'_t, t \in [0,T] \). In the context of filtering theory the key entities are the causal and the non causal estimates, i.e. the conditional mean \( E(x'_t|y'_\eta, 0 \leq \eta \leq t) \) or \( E(x'_t|y'_\eta, 0 \leq \eta \leq T) \) respectively. In addition to this pair of random entities, there is also the averaged minimal errors which amount to \( \int_0^T E(x'_t - E(x'_t|y'_\eta, \eta \in [0,t]))^2 dt \), or \( \int_0^T E(x'_t - E(x'_t|y'_\eta, \eta \in [0,T]))^2 dt \) on one hand, and on the other hand the mutual information between the paths \( \{y = (y_\eta, \eta \in [0,T])\} \) and \( \{x = (x_\eta, \eta \in [0,T])\} \), i.e.

\[
E \log \frac{dP(x,y)}{dP(x)P(y)}
\]

where the expectation is w.r. to the \( P(x,y) \) measure, and the relative entropy \( E\ell(y) \). Relations between the likelihood ratio \( \ell(y) \) and the causal conditional expectation were discovered in the late 60’s and this was soon followed by a relation between the mutual information and the causal mean square error [6], [2]. These relations which involved causal mean square errors were based on the Ito calculus. Similar problems for the non causal estimator were also considered [3], [5]. The formulation and results in the non causal case were restricted to the finite dimensional time discrete model of the Gaussian channel. Recently, however, Guo, Shamai and Verdú (GSV) [4] applied information theoretic arguments to derive new interesting results relating the mutual information with non causal estimation in Gaussian channels.

The Ito calculus which has proved to be a powerful tool for the relations associated with causal estimation could not be applied to problems related to non causal problems which explains the slow progress in the direction of relations for non causal estimates. However, the development of the Malliavin calculus, namely, the stochastic calculus of variation which was introduced in the mid 70’s led in the early 80’s to results which prove to be a very useful tool for the non causal type of problems.

The purpose of this paper is to apply the Malliavin calculus in order to derive the
extension of the finite discrete time results relation between non causal estimation
with likelihood ratios to continuous time (section 4) and to present an alternative
method for proving the results of [4] in an extended setup (sections 5 and 6). We will
often make some technical assumptions which are stronger than necessary in order to
simplify the proofs.

In the next section we define the Abstract Wiener Space which generalizes the
classical $d$-dimensional Wiener process, and formulate the additive Gaussian channel
which will be considered in the paper. Also, the problems considered in sections 4 and
5 are outlined in this section. Section 3 is a very short introduction to the Malliavin
calculus. Section 4 presents the results relating likelihood ratios (R-N derivatives)
with non-causal least square estimates and in section 5 we derive an extended version
of the GSV results. In section 6 the results of section 5 are presented in terms of
the related relative entropy. The modelling of the additive Gaussian channel on the
abstract Wiener space (in contrast to the $d$-dimensional Wiener process on the time
interval $[0, T]$) yield in sections 4, 5 and 6 results of wide applicability, e.g. for the
case of random fields.

Acknowledgement: We wish to express our thanks to Shlomo Shamai for call-
ing our attention to the problems considered in this paper and providing us with a
preliminary version of [4], and to Ofer Zeitouni for useful comments.

2 The underlying Wiener space and the additive
channel model

Consider, first, the classical $d$-dimensional Wiener process, i.e. “the integral of
white noise”. Let $W$ denote the Banach space of the collection of continuous functions
$w(t) \in \mathbb{R}^d, t \in [0, 1]$, endowed with the norm $|w|_W = \sup |w(t)|_{\mathbb{R}^d}$ over $0 \leq t \leq 1$.
Next consider the Cameron-Martin space $H$ which is the Hilbert space $\mathbb{R}^d$ valued
$h(t), t \in [0, 1]$ such that for some $h'(s)$,

$$h(t) = \int_0^t h'(s)ds \quad \text{and} \quad |h|^2_H =: \int_0^1 |h'(s)|^2_{\mathbb{R}^d}ds < \infty.$$
Then for any non random \( h \in H \) the Wiener integral

\[
\delta h = \int_0^1 h'(s) dw(s)
\]

is a well defined Gaussian random variable taking values in \( \mathbb{R}^d \), \( E\delta h = 0 \) and \( E|\delta h|^2_H = |h|^2_H \).

The Abstract Wiener Space (AWS) is an abstraction of this model where \( W \) is a separable Banach space and \( H \), the Cameron-Martin space, is an Hilbert space densely and continuously embedded in \( W \). The dual space to \( W \) (the space of continuous linear functionals on \( W \)) is denoted \( W^* \) and assumed to be continuously and densely embedded in \( H \). The Abstract Wiener Space \((W, H, \mu_W)\) supports a \( W \)-valued random variable \( w \) such that for every \( e \in W^* \), \( \delta e := \langle w, e \rangle_{W^*} \) is a \( \mathcal{N}(0, |e|^2_H) \) random variable. Cf. e.g. [12], or appendix B of [13] and the references therein, for further information on the AWS.

In the case of the classical Wiener space on \( \mathbb{R}^1 \), the space \( W^* \) of all continuous linear functionals on \( W \) can be represented by the class of real valued functions \( \{k(s), s \in [0, 1]\} \) such that \( \int_0^1 k^2(s)ds < \infty \) and \( k(s) = \nu(s, 1] \), where \( \nu \) is a signed measure on \( [0, 1] \) and \( \nu(\{0\}) = 0 \), then

\[
\langle w, k \rangle_{W^*} = \int_0^1 w(s) \nu(ds)
\]

which by integration by parts yields

\[
\langle w, e \rangle_{W^*} = \int_0^1 \nu(s, 1]w(ds) = \delta \left( \int_0^1 k(\theta)d\theta \right).
\]

cf. [10] for details. Note that, unlike the classical case, the Abstract Wiener Space does not have any time-like parameter (this however can be added cf. e.g. section 2.6 of [13]).

In order to introduce the general setup of the additive Gaussian channel, let \((W, H, \mu_W)\) be an abstract Wiener space and let \((H, \sigma(H), \mu_X)\) be a probability space on the Cameron-Martin Hilbert space \( H \) which is induced by an \( H \)-valued r.v. \( X \). Let
\( \theta = (x, w), x \in H \) and \( w \in W \), set \( \Theta = \{ \theta \} \) and consider the combined probability space

\[
(\Theta, \mathcal{F}, \mathcal{P}) = \left( \Theta, \sigma(H) \lor \sigma(W), \mu_X \times \mu_W \right)
\]

which is the space of the mutually independent ‘signals’ \( x \) and ‘noise’ \( w \). Now, since \( H \) is continuously embedded in \( W \) we can identify \( x \) with its image in \( W \) and defined the additive Gaussian channel as

\[
y(\theta) = \rho x + w,
\]

where \( \rho \) is a free scalar ‘signal to noise’ parameter which will become relevant in Section 5. We will denote by \( X \) and \( Y \) the sigma fields induced on \( W \) by the r.v.’s \( x \) and \( y \) respectively. Note that \( y \) and \( w \) are \( W \) valued, \( x \) is \( H \) valued and we identify \( x \) with its image in \( W \). In fact we will make throughout this paper, just for reasons of simplicity, the additional assumption that \( x \) is \( W^* \) valued. As mentioned earlier, since \( W^* \subset H \subset W \) we can also consider \( x \) to be \( H \) or \( W \) valued.

In section 4 we will be interested in the relation between two types of objects. The first class of objects is

\[
E\left((x, e)_H | Y\right) \quad \text{and} \quad E\left((x, e_1)_H \cdot (x, e_2)_H | Y\right)
\]

for \( e, e_1, e_2 \in W^* \) or globally

\[
\hat{x} = E(x|Y) \quad \text{and} \quad (x, x)_H = E\left((x, x)_H | Y\right).
\]

The second class of objects being the likelihood ratio (the R-N derivative) between the measures induced by \( y \) and the one induced by \( w \) on \( W \). This likelihood ratio will be denoted \( \ell(w) \), \( w \in W \). Note that if \( W \) is infinite dimensional then the measure induced by \( x \) is singular with respect to the measure induced by \( w \), (since \( x \in H \) while \( w \notin H \)).

In section 5 we will consider the relation between \( I(x, y) \) or rather \( dI(x, y)/d\rho \) and the non-causal filtering error:

\[
\left\{ E|x|^2_H - \left| E(x|Y) \right|^2_H \right\} = E\left\{ |x|^2_H - |\hat{x}|^2_H \right\}.
\]

A related result for \( d(E \log \ell(w))/d\rho \) is considered in section 6 and shown to be an extended version of the De Bruijn identity.
3 A short introduction to the Malliavin calculus

For further information cf. e.g. [7], [9], [12] or appendix B of [13].

(a) The gradient

Let \((W, H, \mu)\) be an AWS and let \(e_i, i = 1, 2, \ldots \) be a sequence of elements in \(W^*\). Assume that the image of \(e_i\) in \(H\) form a complete orthonormal base in \(H\). Let \(f(x_1, \ldots, x_n)\) be a smooth function on \(\mathbb{R}^n\) and denote by \(f'_i\) the partial derivative of \(f\) with respect to the \(i\)-th coordinate.

For cylindrical smooth random variables \(F(w) = f(\delta e_1, \ldots, \delta e_n)\), define \(\nabla_h F = \frac{dF(w+\epsilon h)}{d\epsilon} \big|_{\epsilon=0}\). Therefore we set the following: \(\nabla_h F = (\nabla F, h)\) where \(\nabla F\) is \(H\)-valued, where for \(F(w) = \delta e\), \(\nabla F = e\), and

\[
\nabla F = \sum_{i=1}^n f'_i(\delta e_1, \ldots, \delta e_n) \cdot e_i .
\]

It can be shown that this definition is closable in \(L^p(\mu)\) for any \(p > 1\). We will restrict ourselves to \(p = 2\), consequently the domain of the \(\nabla\) operation can be extended to all functions \(F(w)\) for which there exists a sequence of smooth cylindrical functions \(F_m\) such that \(F_m \rightarrow F\) in \(L_2\) and \(\nabla F_m\) is Cauchy in \(L^2(H, \mu)\). In this case set \(\nabla F\) to be the \(L_2(\mu, H)\) limit of \(\nabla F_m\). This class of r.v. will be denoted \(D_{2,1}\). It is a closed linear space under the norm

\[
\|F\|_{2,1} = E^{\frac{1}{2}}[|F|^2] + E^{\frac{1}{2}}[|\nabla F|^2_H] .
\]

Similarly let \(K\) be an Hilbert space and \(k_1, k_2, \ldots\) a complete orthonormal base in \(K\). Let \(\varphi\) be the smooth \(K\)-valued function \(\varphi = \sum_j f_j(\delta e_1, \ldots, \delta e_n)k_j\) define

\[
\nabla \varphi = \sum_{j=1}^n \sum_{i=1}^n (f'_j)(\delta e_1, \ldots, \delta e_n)e_i \otimes k_j
\]

and denote by \(D_{2,1}(K)\) the completion of \(\nabla \varphi\) under the norm

\[
\|\varphi\|_{2,1} = E^{\frac{1}{2}}[|\varphi|^2_K] + E^{\frac{1}{2}}[|\nabla \varphi|^2_{H \times K}] .
\]

Note that this enables us to define recursively \(\nabla^n F(w)\) for \(n > 1\).
(b) The divergence (the Skorohod integral)

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

\[
E\left( \nabla f, u(w) \right)_H = E(f \cdot \delta u)
\]  

is satisfied. \( \delta u \) is called the divergence or Skorohod integral. Note that while the definition of \( \nabla f \) (at least for smooth functionals) is invariant under an absolutely continuous change of measure, this is not the case for the divergence which involves expectation in the definition. For non-random \( h \in W^* \), \( \delta h = \langle h, w \rangle \), setting \( f = 1 \) in (3.5) yields that \( E\delta h = 0 \). It can be shown that if \( u \in \mathbb{D}_{2,1}(H) \) then \( u \in \text{dom}_2 \delta \).

Also, for smooth \( f(w) \) it can be verified directly that

\[
\delta (f(w)h) = f(w)\delta h - \left( \nabla f, h \right)_H
\]

and more generally under proper restrictions

\[
\delta (f(w)u(w)) = f(w)\delta u - \left( \nabla f, u(w) \right)_H .
\]  

(3.6)

Consequently, if \( E|\nabla \varphi|_H^2 < \infty \), \( \nabla \varphi \) is \( H \)-valued and \( \nabla^2 \varphi \) is of trace class then

\[
_w(\nabla \varphi, w) = \delta \nabla \varphi + \text{trace} \nabla^2 \varphi .
\]  

(3.7)

where for an operator \( A \) on \( H \) and \( e_i, i = 1, 2, \ldots \) a CONB on \( H \), define

\[
\text{trace} A = \sum_{1}^{\infty} (e_i, Ae_i)
\]

provided the series converges absolutely and in this case \( A \) is said to be of trace class. Among the interesting facts about the divergence operator, let us also note that for the classical Brownian motion and if

\[
(u(w))(\cdot) = \int_{0}^{\cdot} u_s'(w)ds
\]  

7
and \( u'(w) \) is adapted and square integrable then \( \delta u \) coincides with the Ito integral i.e. \( \delta u = \int_0^1 u'_s(w) dw_s \).

(c) Let \((W, H, \mu)\) be an abstract Wiener space and let \( \mu_1 \) be another probability measure on the same Banach space \((W, \sigma\{W\})\). Assume that \( \mu_1 \) is absolutely continuous with respect to \( \mu \). Set

\[
\ell(w) = \frac{d\mu_1}{d\mu}(w) \quad \text{and} \quad Q(w) = \{ w : \ell(w) > 0 \}
\]

\( E_1 \) and \( E_0 \) will be used to denote the expectation with respect to the measures \( \mu_1 \) and \( \mu \) respectively. We will use the convention \( 0 \log 0 = 0 \) throughout the paper.

Following the definition in 3(b), we define the divergence with respect to \( \mu_1 \) to be as follows. The \( H \)-valued random variable \( u(w) \) will be said to be in \( \text{dom} \tilde{\delta} \) if there exists a r.v., say \( \tilde{\delta}u \), which is \( L^2 \) under \( u_1 \) and such that for all smooth r.v.s \( f(w) \), it holds that

\[
E_1\left( f(w) \cdot \tilde{\delta}u \right) = E_1(u, \nabla f)_H.
\]

The relation between \( \tilde{\delta}u \) and \( \delta u \) is given by the following lemma.

**Lemma 3.1** Assume that \( \ell(w) \in \mathbb{D}_{2,1}, u \in \text{dom}_2 \delta, \ell \cdot \delta u \in L_2 \) and \( \ell \cdot \nabla u \in \mathbb{D}_{2,0}(H) \) and \( \mu_1 \ll \mu_W \) (where \( \mathbb{D}_{2,0}(H) \) is the completion of (3.3) under the \( H \)-norm). Then \( u \in \text{dom}_2 \tilde{\delta} \) and

\[
\tilde{\delta}u = 1_Q(w)(\delta u - \left( \nabla \log \ell(w), u(w) \right)_H)
\]

(3.8)

**Proof:** Since \( f(w) \) is a smooth r.v., \( \ell \cdot f \cdot \delta u - f(\nabla f, u)_H \) is in \( L_1 \) and \( \ell(w) \nabla \log \ell(w) = \nabla \ell(w) \) a.s.-\( \mu \). Hence

\[
E_1\left( f(w) \delta u - f(w) \left( \nabla \log \ell(w), u \right)_H \right) = E_0\left( \ell \cdot f \cdot \delta u - \ell f(\nabla \log \ell, u)_H \right)
\]

\[
= E_0\left( (\nabla (\ell \cdot f), u)_H - f(\nabla \ell, u) \right)
\]

\[
= E_0\left( \ell(\nabla f, u)_H \right)
\]

\[
= E_1(\nabla f, u)_H
\]

\( \square \)
4 Relations between the estimation error and the likelihood ratio

Let \((W, H, \mu), (H, \sigma(H), \mu_X), (\Theta, \mathcal{F}, \mathcal{P})\) and \(y(\theta) = \rho x + w\) be as in section 2. We will further assume that the \(H\)-valued r.v. \(x\) is actually \(W^*\) valued, and \(\exp \alpha(x, h)_H \in L^1(\mu_X)\) for all real \(\alpha\) and all \(h \in W^*\). The measures induced by \(y\) and \(x\) on \(W\) will be denoted \(\mu_Y\) and \(\mu_X\) respectively. The conditional probability induced on \(W\) by \(y(\theta)\) conditioned on \(x\) will be denoted by \(\mu_Y|_X\). Similarly, \(\mu_X|Y\) will denote the conditional probability induced on \(W^*\) of \(x\) conditioned on \(y\) cf. e.g. [1] for the existence of these conditional probabilities).

By the Cameron-Martin theorem (cf. e.g. [13]) and since \(x\) and \(w\) are independent, we have
\[
\frac{d\mu_{Y|X}}{d\mu_W}(w) = \exp \left( \rho \langle w, x \rangle - \frac{\rho^2}{2} |x|^2_H \right), \quad w \in W
\] (4.1)
which by our assumptions belongs to \(L^p\) for all \(p > 0\). Hence, denoting by \(\mu_X(dx)\) the restriction of \(\mathcal{P}\) to \(H\):
\[
\ell(w) = \int_H \frac{d\mu_Y}{d\mu_W}(w) \frac{d\mu_{Y|X}}{d\mu_W}(w, x) \mu_X(dx)
= \int_H \exp \left( \rho \langle w, x \rangle - \frac{\rho^2}{2} |x|^2_H \right) \mu_X(dx)
\] (4.2)

**Proposition 4.1** Under these assumptions it holds that

(a)
\[
(\nabla \ell, h)_H =: \nabla_h \ell(w) = \rho \ell(w) \langle \hat{x}, h \rangle_H, \quad \forall h \in H \quad \text{hence} \quad \nabla \ell = \rho \ell(w) \hat{x}
\] (4.3)
a.s. \(\mu_W\)

(b)
\[
\left( \nabla^n \ell(w), h_1 \otimes \cdots \otimes h_n \right)_{H^\otimes n} = \ell(w) \rho^n \left( \prod_{i=1}^n \langle h_i, x \rangle \right)^\otimes
\] (4.4)
where (a) $\hat{\cdot}$ denotes the conditional expectation conditioned on $Y$.

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

\[
\nabla^2_{h_1, h_2} \ell(w) = \rho^2 \ell(w) \left( (h_1, x) \cdot (h_2, x) \right)^\hat{\cdot} \tag{4.5}
\]

and

\[
\nabla^2_{h, h} \log \ell(w) = \rho^2 \left( (x, h)^2 \right)^\hat{\cdot} - (\hat{x}, h)^2 \tag{4.6}
\]

where $\nabla^2_{h_1, h_2} \varphi =: (\nabla(\nabla \varphi, h_2), h_1)_H$.

(d)

\[
\left( \prod_{i=1}^{n} (h_i, x) \right)^\hat{\cdot} = (h_n, \hat{x}) \left( \prod_{i=1}^{n-1} (h_i, x) \right)^\hat{\cdot} + \nabla_{h_n} \left( \prod_{i=1}^{n-1} (h_i, x) \right) \hat{x}. \tag{4.7}
\]

**Remark 1:** Let $E_1$ denote the measure induced by $y$ on $W$ and $E$ will denote expectation w.r. to the measure in (2.3). For an operator $A$ on $H$ and $e_i, i = 1, 2, \ldots$ a CONB on $H$, define

\[
\text{trace } A = \sum_{i=1}^\infty (e_i, Ae_i)
\]

provided the series converges. Consequently, we have from (4.6) that

\[
E_1 \text{ trace } \nabla^2 \log \ell(w) = \rho^2 E |x - \hat{x}|_H^2 \tag{4.8}
\]

(c.f. also equations (6.3) and (6.4)).

**Remark 2:** Consider the case where the abstract Wiener space is a classical Wiener space $\mathbb{R}^n$, then $u \in H$ is of the form $\int_0^t u'(s)ds$, $x \in H$ is of the form $\int_0^t x'(s)ds$ where $x'(s) \in \mathbb{R}_n$ and $\int_0^T |x'(s)|^2_{\mathbb{R}_n} ds < \infty$. Further assume that $E \int_0^T |x'(s)|^2_{\mathbb{R}_n} ds < \infty$ and $x'(s)$ is a.s. continuous on $[0, T]$. Then given some $t \in [0, T]$, one can consider a sequence of linear functionals $h_n$ such that $(h_n, x)$ converges in $L^2$ to $x(t, w)$ and extend the results of proposition 4.1 to $\hat{x}_t := E(x'(t)|Y)$ for any $t \in [0, T]$. We will not follow this path.

10
The following lemma will be needed in the proof of Proposition 4.1:

**Lemma 4.1** Assume that $\mu_Y$ and $\mu_{Y|X}$ are absolutely continuous with respect to $\mu_W$ then for all bounded and measurable functions $\psi$ on $\Theta$

$$
\int_{X \times W} \psi(x, y) \frac{d\mu_{Y|X}}{d\mu_W}(y, x) \mu_X(dx) \times \mu_W(dy) = \int_{X \times W} \psi(x, y) \frac{d\mu_Y}{d\mu_W}(y) \mu_{X|Y}(dx, y) \mu_W(dy).
$$

**Proof of Lemma:** Let

$$L = \int_{X \times Y} \psi(x, y) \mu_{X,Y}(dx, dy).$$

Then, by Fubini’s theorem

$$L = \int_{X \times Y} \psi(x, w) \mu_{Y|X}(dw, x) \mu_X(dx)$$

$$= \int_{X \times Y} \psi(x, y) \frac{d\mu_{Y|X}}{d\mu_W}(y, x) \mu_X(dx) \mu_W(dy).$$

Since the conditional probability $\mu_{X|Y}$ is regular (cf e.g. theorem 10.2.2 of [1]) we also have

$$L = \int_{X \times Y} \psi(x, y) \mu_{X|Y}(dx, y) \mu_Y(dy)$$

$$= \int_{X \times Y} \psi(x, y) \frac{d\mu_Y}{d\mu_W}(y) \mu_{X|Y}(dx, y) \mu_W(dy).$$

**Proof of Proposition:** From (4.1) and (4.2) and since by our assumptions we may (by dominated convergence) interchange the order of integration and differentiation

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

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

$$= \int_H \rho(h, x) \frac{d\mu_Y}{d\mu_W}(w, x) \mu_X(dx).$$
Thus, by Lemma 4.1
\[
\nabla_h \ell(w) = \int_X \rho(h, x) \frac{d\mu_Y}{d\mu_W}(w) \mu_{X|Y}(dx, w) \\
= \rho \ell(w)(h, \hat{x})
\]
proving (4.3). The same arguments also hold for repeated differentiation
\[
\left( \nabla^n \ell(w), h_1 \otimes \cdots \otimes h_n \right)_{H^\otimes n} = \rho^n \int_X (h_1, x) \cdots (h_n, x) \frac{d\mu_Y|X}{d\mu_W}(w, x) \mu_X(dx) \\
= \rho^n \int_x \left( \prod_{i=1}^n (h_i, x) \right) \frac{d\mu_Y}{d\mu_W}(w) \mu_{X|Y}(dx, w),
\]
which yields (4.4). (4.5) follows directly from (4.4) since
\[
\nabla^2_{h_1, h_2} \ell(w) = \rho^2 \ell(w) \left( (h_1, x) \cdot (h_2, x) \right) = \nabla^2_{h, h} \ell(w) = \rho^2 \ell(w) \left( (h_1, x) \cdot (h_2, x) \right)
\]
proving (4.5) and (4.6). From (4.4) we have
\[
\rho^n \ell(w) \left( \prod_{i=1}^n (h_i, x) \right) = \nabla_{h_n} \left( \nabla^{n-1} \ell(w), h_1 \otimes \cdots \otimes h_{n-1} \right)_{H^\otimes(n-1)} \\
= \nabla_{h_n} \left( \ell(w) \cdot \rho^{n-1} \cdot \left( \prod_{i=1}^n (h_i, x) \right) \right)
\]
and (4.7) follows.

We conclude this section with some results for $\delta \hat{x}$ and $\tilde{\delta} \hat{x}$ (cf. part (c) of section 3). By the assumptions of this section $\hat{x} \in \text{dom}_2 \delta$ and $\hat{x} \in \text{dom}_2^1 \delta$. Therefore by (4.3)
\[
\delta \hat{x} = \frac{1}{\rho} \delta \nabla \log \ell(w).
\]
Note that $\mathcal{L} = \delta \nabla$ is the number operator, i.e. if $\alpha(w)$ is a square integrable r.v. of the Wiener space and $\alpha = \sum_{n=1} I_n$, where $I_n$ is the Wiener chaos decomposition of $x$; then, formally, $\mathcal{L} \alpha = \sum_{n=1} n I_n$. Therefore if $\alpha(w) \in \mathcal{L}^2$ and $E\alpha(w) = 0$ then $\mathcal{L}^{-1}$ is well behaved, consequently it holds that

$$\ell(w) = \exp \rho \mathcal{L}^{-1} \delta \hat{x} . \tag{4.10}$$

For $\tilde{\delta} \hat{x}$ we have

**Lemma 4.2**

$$\tilde{\delta} \hat{x} = \frac{1}{\rho \ell(w)} \delta \nabla \ell(w) .$$

**Proof:** By (4.3)

$$\delta \nabla \ell = \rho \delta (\ell(w) \hat{x})$$

$$= \rho \ell(w) \delta \hat{x} - \rho (\hat{x}, \nabla \ell)_H$$

$$= \rho \ell(w) \delta \hat{x} - \rho^2 \ell(w)(\hat{x}, \hat{x})_H$$

and

$$\delta \hat{x} = \frac{1}{\rho \ell(w)} \delta \nabla \ell(w) + \rho(\hat{x}, \hat{x})_H .$$

Hence by Lemma 3.1

$$\tilde{\delta} \hat{x} = \delta \hat{x} - (\nabla \log \ell(w), \hat{x})_H$$

$$= \delta \hat{x} - \rho |\hat{x}|^2_H$$

$$= \frac{\mathcal{L} \ell(w)}{\rho \ell(w)}$$

\[\square\]
5 The GSV relation between the mutual information and the mean square of the estimation error

Consider the setup and assumptions in the first paragraph of section 4. The mutual information between $x$ and $y$ is defined as

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

$E$ will denote expectation w.r. to the measure in (2.3), $E_0$ 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 $Ef(y) = E_1f(w) = E_0\ell(w)f(w)$).

**Proposition 5.1** Under the assumptions of the previous section, it holds that

$$\frac{dI(X, Y)}{d\rho} = \rho E\left(|x|_H^2 - |\hat{x}|_H^2\right)$$

(5.1)

$$= \rho E|x - \hat{x}|_H^2$$

Proof: By our assumptions we have

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

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

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

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

(5.2)

and

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

(5.3)

$$= \rho E|x|_H^2 - E_0 \left( \frac{d\ell(w)}{d\rho} \cdot \log \ell(w) \right) - 0.$$
Now,
\[
\frac{d\ell(w)}{d\rho} = \int_X \left( \langle x, w \rangle - \rho |x_H|^2 \right) \frac{d\mu_Y(x)}{d\mu_W}(w) \mu_X(dx).
\]

By lemma 4.1
\[
\frac{d\ell(w)}{d\rho} = \int_X \left( \langle x, w \rangle - \rho |x_H|^2 \right) \frac{d\mu_Y}{d\mu_W}(w) \mu_X(dx) = \left( \langle \hat{x}, w \rangle - \rho |x_H|^2 \right) \ell(w).
\]

Substituting in (5.4) yields
\[
\frac{dI}{d\rho} = \rho E|x|^2_H - E_0 \left\{ \left( \langle \hat{x}, w \rangle - \rho |x_H|^2 \right) \ell(w) \log \ell(w) \right\}.
\]

Now, by (4.3)
\[
E_0 \ell \log \ell \langle \hat{x}, w \rangle = E_0 \left( \frac{1}{\rho} \log \ell \langle \nabla \ell, w \rangle \right)
\]
\[
\quad \quad \quad \quad \text{by (3.7) } E_0 \frac{1}{\rho} \log \ell \left( \delta \nabla + \text{trace } \nabla^2 \ell \right)
\]
\[
\quad \quad \quad \quad \text{by (3.6) } E_0 \frac{1}{\rho} \delta (\log \ell \nabla \ell) - E_0 \frac{1}{\rho} \left( \nabla \ell \right. \left. \nabla \log \ell \right) + E_0 \frac{1}{\rho} (\log \ell \text{ trace } \nabla^2 \ell)
\]
\[
\quad \quad \quad \quad \text{by (3.5) and (4.9) } 0 - \frac{1}{\rho} E_0 \frac{1}{\ell} \left( \nabla \ell, \nabla \ell \right) + E_0 \rho \left( \ell \log \ell(|x_H|^2) \right)
\]
\[
\quad \quad \quad \quad = \rho E_0 \ell(w)(\hat{x}, \hat{x}) + E_0 \rho \ell(w) \log \ell(w)(|x_H|^2).
\]

Substituting into (5.5) yields
\[
\frac{dI}{d\rho} = \rho E|x_H|^2 - E_0 \rho E|\hat{x}|^2_H + E_0 \rho \ell(w) \log \ell(|x_H|^2) - E_0 \rho \ell(w) \log \ell(|x_H|^2).
\]

\[\square\]

**Remark:** Consider the following generalizations to the additive Gaussian channel. Let \(M\) be “the space of messages which generate the signals” \(x\), i.e. \((M, B, P)\) is...
a probability space and \( x = g(m) \), \( m \in M \), where \( g \) is a measurable from \((M, \mathcal{B}_\mu)\) to \( H \). Then obviously \( I(M, Y) = I(X, Y) \). More generally, consider the case where \( x \) and \( m \) are related by some joint probability on \( M \times H \) and \( w \) and \( m \) are conditionally independent conditioned on \( x \). The extension of proposition 5.1 in this context follows along the same arguments as in theorem 13 of [4] and therefore omitted.

6 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

\[
E \text{trace } J = -E \left\{ \left| \nabla \log p \right|^2_{\mathbb{R}^n} \right\},
\]

where \( E \) is the expectation with respect to the \( p \) density. The De Bruijn identity (cf [4] and the references therein) deals with the case where \( y = x + \sqrt{t}w \) where \( w = w_1, w_2, \ldots, w_n \) and the \( w_j, \ j = 1, \ldots, n \) are i.i.d. \( N(0, 1) \) and \( x \) is an \( \mathbb{R}_n \) random variable independent of \( w \). It states that

\[
\frac{d}{dt} E \log p(y) = \frac{1}{2} E \left\{ \left| \nabla \log p(y) \right|^2_{\mathbb{R}^n} \right\}.
\]

(6.2)

The Fisher information matrix cannot be extended directly to the case where \( y \) is infinite dimensional. However, the results of sections 4 and 5 yield some similar relations. Under the assumptions of section 5, comparing (5.1) with (5.3) we have

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

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

(6.3)

which is “similar” to (6.2) and may be considered a generalized De Bruijn identity. Note that \( E_1 \log \ell(w) \) is the relative entropy of \( \mu_w \) relative to \( \mu_W \), also, note the
difference between the $\rho$ and the $t$ parametrizations. Comparing (5.1) with (4.8) yields
\[
E_1 \text{trace } \nabla^2 \log \ell(w) = \rho \frac{dI(x,y)}{d\rho} = \rho^2 E|\mathbf{x}|_H^2 - E_1|\nabla \log \ell(w)|_H^2,
\]
which is different from (6.1) by the $\rho^2 E|\mathbf{x}|_H^2$ term.

References


