INTERNATIONAL CENTRE FOR MECHANICAL SCIENCES

GYULA KATONA
MATHEMATICAL INSTITUTE
HUNGARIAN ACADEMY OF SCIENCES, BUDAPEST

GENERAL THEORY OF NOISELESS CHANNELS

LECTURES HELD AT THE DEPARTMENT
FOR AUTOMATION AND INFORMATION
JUNE 1970

UDINE 1970

COURSES AND LECTURES - No. 31

ISBN 978-3-211-81167-2 ISBN 978-3-7091-2872-5 (eBook)
DOI 10.1007/978-3-7091-2872-5

Originally published by Springer Vienna in 1970

P R E F A C E

"Noiseless channels" is an expression like "rose without a thorn". In practical cases we have almost always noisy channels. However, it is useful to examine the noiseless channels because there are channels with small noise, we may consider them to be noiseless channels.

On the other hand, studying noiseless channels we can get directions about the properties of noisy channels, which are more complicated, thus it is much more difficult to study them directly.

This short survey paper is the written form of my 8 lectures organized by CISM in June 1970.

In the lecture notes only the elementary probability theory is used and some elementary properties of the information-type functions. These properties are proved in the "Preliminaries" written by Professor I. Csiszár.

I would like to express my thanks to the CISM for giving me this opportunity.

I hope this paper will be an almost noiseless channel from the author to the readers.

Udine, June 22, 1970. *G. Katona*

Preliminaries

In this section we summarize some basic definitions and relations which will be used freely in the sequel : the simple proofs will be sketched only.

The term "random variable" will be abbreviated as RV ; for the sake of simplicity, attention will be restricted to the case of discrete RV's,i.e., to RV's with values in a finite or countably infinite set.

ξ, η, ζ will denote RV's with values in the (finite or countably infinite) sets X, Y, Z.

All random variables considered at the same time will be assumed to be defined on the same probability space. Recall that a probability space is a triplet $(\Omega, \mathcal{F}, P)$ where Ω is a set (the set of all conceivable outcomes of an experiment), $\mathcal{F}$ is a σ-algebra of subsets of Ω (the class of observable events) and P is a measure (non-negative countably additive set function) defined on $\mathcal{F}$ such that $P(\Omega)=1$. Rv's ξ, η etc. are functions $\xi(\omega)$, $\eta(\omega)$ etc. $(\omega \in \Omega)$. The probability $P\{\xi = x\}$ is the measure of the set of those ω's for which $\xi(\omega)=x$; similarly, $P\{\xi=x, \eta=y\}$ is the measure of the set of those ω's for which $\xi(\omega)=x$ and $\eta(\omega)=y$.

The conditional probability $P\{\xi=x \mid \eta=y\}$ is defined as $\frac{P\{\xi=x, \eta=y\}}{P\{\eta=y\}}$ (if $P\{\eta=y\}=0$, $P\{\xi=x \mid \eta=y\}$ is undefined).

Definition 1. The RV's defined by

(1) $\iota_\xi = -\log_2 P\{\xi = x\}$ if $\xi = x$

(2) $\iota_{\xi\wedge\eta} = \log_2 \frac{P\{\xi=x, \eta=y\}}{P\{\xi=x\}P\{\eta=y\}}$ if $\xi = x, \eta = y$

are called the entropy density of ξ and the information density of ξ and η, respectively.

(3) $\iota_{\xi|\eta} = -\log_2 P\{\xi=x \mid \eta=y\}$ if $\xi=x, \eta=y$

(4) $\iota_{\xi|\eta,\zeta} = -\log_2 P\{\xi=x \mid \eta=y, \zeta=z\}$ if $\xi=x, \eta=y, \zeta=z$

are conditional entropy densities and

(5) $\iota_{\xi\wedge\eta|\zeta} = \log_2 \frac{P\{\xi=x, \eta=y \mid \zeta=z\}}{P\{\xi=x \mid \zeta=z\}P\{\eta=y \mid \zeta=z\}}$ if $\xi=x, \eta=y, \zeta=z$

is a conditional information density.

Remark. Entropy density is often called "self-information" and information density "mutual information". In our terminology, the latter term will mean the expectation of $\iota_{\xi\wedge\eta}$.

Definition 2. The quantities

$$E(\xi) \stackrel{\text{def}}{=} E\iota_{\xi} = -\sum_{x\in X} P\{\xi=x\}\log_2 P\{\xi=x\} \tag{6}$$

$$I(\xi\wedge\eta) \stackrel{\text{def}}{=} E\iota_{\xi\wedge\eta} = \sum_{x\in X,\,y\in Y} P\{\xi=x,\eta=y\}\log_2 \frac{P\{\xi=x,\eta=y\}}{P\{\xi=x\}P\{\eta=y\}} \tag{7}$$

are called the entropy of ξ and the mutual information of ξ and η, respectively.

The quantities

$$H(\xi|\eta) \stackrel{\text{def}}{=} E\iota_{\xi|\eta} = -\sum_{x\in X,\,y\in Y} P\{\xi=x,\eta=y\}\log_2 P\{\xi=x|\eta=y\} \tag{8}$$

$$H(\xi|\eta,\zeta) \stackrel{\text{def}}{=} E\iota_{\xi|\eta,\zeta} = -\sum_{x\in X,\,y\in Y,\,z\in Z} P\{\xi=x,\eta=y,\zeta=z\}\log_2 P\{\xi=x|\eta=y,\zeta=z\} \tag{9}$$

are called conditional entropies and

$$I(\xi\wedge\eta|\zeta) \stackrel{\text{def}}{=} E\iota_{\xi\wedge\eta|\zeta} = \sum_{x\in X,\,y\in Y,\,z\in Z} P\{\xi=x,\eta=y,\zeta=z\}\log_2 \frac{P\{\xi=x,\eta=y|\zeta=z\}}{P\{\xi=x|\zeta=z\}P\{\eta=y|\zeta=z\}} \tag{10}$$

is called conditional mutual information.

Here terms like $0\log_2 0$ or $0\log_2\frac{0}{0}$ are meant to be 0.

The quantities (6)-(10) are always non-negative (for (7) and (10) this requires proof ; see (17), (18)) but they may be infinite. The latter contingency should be kept in mind ; in particular, identities like $I(\xi\wedge\eta) = H(\xi) - H(\xi|\eta)$ (cf. (21)) are valid only under the condition that they do not contain the undefined expression $+\infty - \infty$.

$H(\xi)$ is interpreted as the measure of the average amount of information contained in spec-

ifying a particular value of ξ; $I(\xi \wedge \eta)$ is a measure of the average amount of information obtained with respect to the value of η when specifying a particular value of ξ. Conditional entropy and conditional mutual information are interpreted similarly. Logarithms to the basis 2 (rather than natural logarithms) are used to ensure that the amount of information provided by a binary digit (more exactly, by a random variable taking on the values 0 and 1 with probabilities 1/2) be unity. This unit of the amount of information is called bit.

The interpretation of the quantities (6)-(10) as measures of the amount of information is not merely a matter of convention ; rather, it is convincingly suggested by a number of theorems of information theory as well as by the great efficency of heuristic reasonings based on this interpretation. There is much less evidence for a similar interpretation of the entropy and information densities. Thus we do not insist on attaching any intuitive meaning to the latters ; they will be used simply as convenient mathematical tools.

A probability distribution, to be abbreviated as PD, on the set X is a non-negative valued function $p(x)$ on X with $\sum_{x \in X} p(x) = 1$; PD's will be denoted by script letters, e.g. $\mathcal{P} = \{ p(x), x \in X \}$.

Definition 3. The I-divergence of two PD's $\mathcal{P} = \{p(x),\ x \in X\}$ and $\mathcal{Q} = \{q(x), x \in X\}$ is defined as

$$I(\mathcal{P}\|\mathcal{Q}) = \sum_{x \in X} p(x) \log_2 \frac{p(x)}{q(x)} \quad . \tag{11}$$

Here terms of the form $a \log_2 \frac{a}{0}$ with $a > 0$ are meant to be $+\infty$.

Lemma 1. Using the notations $p(A) = \sum_{x \in A} p(x)$, $q(A) = \sum_{x \in A} q(x)$ we have for an arbitrary subset A of X

$$\sum_{x \in A} p(x) \log_2 \frac{p(x)}{q(x)} \geq p(A) \log_2 \frac{p(A)}{q(A)} \ ; \tag{12}$$

if $q(A) > 0$ the equality holds iff*) $p(x) = \frac{p(A)}{q(A)} q(x)$ for every $x \in A$. In particular, setting $A = X$:

$$I(\mathcal{P}\|\mathcal{Q}) \geq 0 \ , \qquad \text{equality iff } \mathcal{P} = \mathcal{Q} . \tag{13}$$

Proof. The concavity of the function $f(t) = \ln t$ implies $\ln t \leq t-1$, with equality iff $t = 1$. Setting now $t = \frac{q(x)}{p(x)} \frac{p(A)}{q(A)}$ one gets $\ln \frac{q(x)}{p(x)} \leq \ln \frac{q(A)}{p(A)} + \frac{q(x)}{p(x)} \frac{p(A)}{q(A)} - 1$ whenever $p(x)\, q(x) > 0$, with equality iff $\frac{q(x)}{p(x)} = \frac{q(A)}{p(A)}$. Multiplying by $p(x)$ and summing for every $x \in A$ with $p(x) > 0$ (one may obviously assume that then $q(x) > 0$ too) (12) follows, including the condition for equality. The choice of the basis of the logarithms being clearly immaterial. The I-divergence $I(\mathcal{P}\|\mathcal{Q})$ is a measure of how different the PD $\mathcal{P}$ is from the PD $\mathcal{Q}$ (however note, that in general $I(\mathcal{P}\|\mathcal{Q}) = I(\mathcal{Q}\|\mathcal{P})$). If $\mathcal{P}$ and $\mathcal{Q}$ are two

*) Iff is an abbreviation for "if and only if".

hypothetical PD's on X then $I(\mathcal{P}\|\mathcal{Q})$ may be interpreted as the average amount of information in favour of $\mathcal{P}$ and against $\mathcal{Q}$, obtained from observing a randomly chosen element of X, provided that the PD $\mathcal{P}$ is the true one.

The distribution of a RV ξ is the PD $\mathcal{P}_\xi$ defined by

$$(14) \qquad \mathcal{P}_\xi = \left\{ p_\xi(x),\ x \in X \right\},\quad p_\xi(x) = P\left\{ \xi = x \right\}.$$

The joint distribution $\mathcal{P}_{\xi\eta}$ of the RV's ξ and η is defined as the distribution of the RV (ξ,η) taking values in $X \times Y$ i.e. $\mathcal{P}_{\xi\eta} = \left\{ p_{\xi\eta}(x,y), x\in X, y\in Y \right\}$, $p_{\xi\eta}(x,y) = P\left\{ \xi = x,\ \eta = y \right\}$.

From (7) and (11) it follows

$$(15) \qquad I(\xi \wedge \eta) = I(\eta \wedge \xi) = I\left(\mathcal{P}_{\xi\eta} \| \mathcal{P}_\xi \times \mathcal{P}_\eta\right)$$

where $\mathcal{P}_\xi \times \mathcal{P}_\eta = \left\{ p_\xi(x)\, p_\eta(y),\ x \in X,\ y \in Y \right\}$ and also

$$(16) \qquad I(\xi \wedge \eta) = \sum_{x\in X} p_\xi(x)\, I\left(\mathcal{P}_{\eta|\xi=x} \| \mathcal{P}_\eta\right)$$

where $\mathcal{P}_{\eta|\xi=x} = \left\{ p_x(y),\ y \in Y \right\}$, $p_x(y) = P\left\{ \eta = y \quad \xi = x \right\}$.
(15) and (13) yield

(17) $I(\xi \wedge \eta) \geq 0$, equality iff ξ and η are independent.
By a comparison of (7) and (10), this implies

(18) $I(\xi \wedge \eta \mid \zeta) \geq 0$, equality iff ξ and η are condition

ally independent for ζ given.

Let us agree to write $\iota_{\xi,\eta}$ for $\iota_{(\xi,\eta)}$ (entropy density of the RV (ξ,η)), $\iota_{\xi,\eta\wedge\zeta}$ for $\iota_{(\xi,\eta)\wedge\zeta}$ (information density of the RV's (ξ,η) and ζ) etc. ; omitting the brackets will cause no ambiguities.

Theorem 1. (Basic identities)

$$\iota_{\xi,\eta} = \iota_{\xi|\eta} + \iota_{\eta} \qquad H(\xi,\eta) = H(\xi|\eta) + H(\eta) \tag{19}$$

$$\iota_{\xi,\eta|\zeta} = \iota_{\xi|\eta,\zeta} + \iota_{\eta|\zeta} \qquad H(\xi,\eta|\zeta) = H(\xi|\eta,\zeta) + H(\eta|\zeta) \tag{20}$$

$$\iota_{\xi} = \iota_{\xi|\eta} + \iota_{\xi\wedge\eta} \qquad H(\xi) = H(\xi|\eta) + I(\xi\wedge\eta) \tag{21}$$

$$\iota_{\xi|\zeta} = \iota_{\xi|\eta,\zeta} + \iota_{\xi\wedge\eta|\zeta} \qquad H(\xi|\zeta) = H(\xi|\eta,\zeta) + I(\xi\wedge\eta|\zeta) \tag{22}$$

$$\iota_{\xi_1,\xi_2\wedge\eta} = \iota_{\xi_1\wedge\eta} + \iota_{\xi_2\wedge\eta|\xi_1} \; ; \quad I(\xi_1,\xi_2\wedge\eta) = I(\xi_1\wedge\eta) + I(\xi_2\wedge\eta|\xi_1) \tag{23}$$

$$\iota_{\xi_1,\xi_2\wedge\eta|\zeta} = \iota_{\xi_1\wedge\eta|\zeta} + \iota_{\xi_2\wedge\eta|\xi_1,\zeta} \; ; \quad I(\xi_1,\xi_2\wedge\eta|\zeta) = I(\xi_1\wedge\eta|\zeta) + I(\xi_2\wedge\eta|\xi_1,\zeta) \tag{24}$$

Proof. Immediate from definitions 1 and 2

Theorem 2. (Basic inequalities)

The information quantities (6)-(10) are non-negative ;

$$H(\xi,\eta) \geq H(\xi) \;, \quad H(\xi,\eta|\zeta) \geq H(\xi|\zeta) \tag{25}$$

$$H(\xi|\eta,\zeta) \leq H(\xi|\eta) \leq H(\xi) \tag{26}$$

$$I(\xi_1, \xi_2 \wedge \eta) \geqq I(\xi_1 \wedge \eta);\ I(\xi_1, \xi_2 \wedge \eta \mid \zeta) \geqq$$

(27) $$\geqq I(\xi_1 \wedge \eta \mid \zeta)$$

(28) $$I(\xi \wedge \eta) \leqq H(\xi),\ I(\xi \wedge \eta \mid \zeta) \leqq H(\xi \mid \zeta).$$

If ξ has at most r possible values then

(29) $$H(\xi) \leqq \log_2 r.$$

If ξ has at most $r(y)$ possible values when $\eta = y$ then

(30) $$H(\xi \mid \eta) \leqq E \log_2 r(\eta).$$

Proof. (25)-(28) are direct consequences of (19)-(24). (29) follows from (13) setting $\mathcal{P} = \mathcal{P}_\xi$, $\mathcal{Q} = \left\{\frac{1}{r}, \ldots, \frac{1}{r}\right\}$; on comparison of (6) and (8), (29) implies (30).

Remark. $I(\xi \wedge \eta \mid \zeta) \leqq I(\xi \wedge \eta)$ is not valid; in general. E. g., if ξ and η are independent but not conditionally independent for a given ζ, then

$$I(\xi \wedge \eta) = 0 < I(\xi \wedge \eta \mid \zeta).$$

Theorem 3. (Substitutions in the information quantities).

For arbitrary functions $f(x)$, $f(y)$ or $f(x,y)$ defined on X, Y or $X \times Y$, respectively, the following inequalities hold

(31) $$H(f(\xi)) \leqq H(\xi);\quad I(f(\xi) \wedge \eta) \leqq I(\xi \wedge \eta)$$

$$H(\xi|f(\eta)) \geq H(\xi|\eta) \tag{32}$$

$$H(f(\xi,\eta)|\eta) \leq H(\xi|\eta). \tag{33}$$

If f is one-to-one, or $f(x,y)$ as a function of x is one-to-one for every fixed $y\in Y$, respectively, the equality signs are valid. In the second half of (31) and in (32) the equality holds also if ξ and η are conditionally independent for given $f(\xi)$ or $f(\eta)$, respectively.

Proof. In the one-to-one case, the validity of (31)-(33) with the equality sign is obvious from definition 2. In the general case, apply this observation for $\tilde{f}$ instead of f where $\tilde{f}(x)=(x,f(x))$, $\tilde{f}(y)=(y,f(y))$ or $\tilde{f}(x,y)=(x,f(x,y))$, respectively ; then theorem 2 gives rise to the desired inequalities. The last statements follow from (18) and the identities :

$$I(\xi\wedge\eta) = I(\xi,f(\xi)\wedge\eta) = I(f(\xi)\wedge\eta) + I(\xi\wedge\eta\ f(\xi))$$

$$H(\xi) = H(\xi,f(\xi)) \geq H(f(\xi))$$

$$H(\xi|\eta) = H(\xi|\eta,f(\eta)) \leq H(\xi|f(\eta))$$

$$H(\xi|\eta) = H(\xi,f(\xi,\eta)|\eta) \geq H(f(\xi,\eta)|\eta)$$

respectively.

Theorem 4. (Convexity properties).

Consider the entropy and the mutual information as a function of the distribution of ξ, in the latter case keeping the conditional distributions $\mathcal{P}_{\eta|\xi=x} = \{ p_x(y), y \in Y \}$ fixed :

$$(34) \qquad H(\mathcal{P}) = - \sum_{x \in X} p(x) \log_2 p(x)$$

$$(35) \qquad I(\mathcal{P}) = \sum_{x \in X, y \in Y} p(x)\, p_x(y) \log_2 \frac{p_x(y)}{q_{\mathcal{P}}(y)} ; \quad q_{\mathcal{P}}(y) = \sum_{x \in X} p(x) p_x(y).$$

Then $H(\mathcal{P})$ and $I(\mathcal{P})$ are concave functions of the PD $\mathcal{P} = \{ p(x), x \in X \}$ i.e., if $\mathcal{P}_1 = \{ p_1(x) = x \in X \}$, $\mathcal{P}_2 = \{ p_2(x), x \in X \}$ and $\mathcal{P} = a\mathcal{P}_1 + (1-a)\mathcal{P}_2 = \{ a p_1(x) + (1-a) p_2(x), x \in X \}$ where $0 < a < 1$ is arbitrary, then

$$(36) \qquad H(\mathcal{P}) \geq a H(\mathcal{P}_1) + (1-a) H(\mathcal{P}_2), \quad I(\mathcal{P}) \geq a I(\mathcal{P}_1) + (1-a) I(\mathcal{P}_2).$$

Proof. The function $f(t) = - t \log_2 t$ is concave hence so is $H(\mathcal{P})$ as well. Since the PD $\mathcal{Q}_{\mathcal{P}} = \{ q_{\mathcal{P}}(y), y \in Y \}$ depends linearly on the PD $\mathcal{P}$, the concavity of $f(t) = - t \log_2 t$ also implies that

$$\sum_{x \in X} p(x)\, p_x(y) \log_2 \frac{p_x(y)}{q_{\mathcal{P}}(y)} =$$

$$= - q_{\mathcal{P}}(y) \log_2 q_{\mathcal{P}}(y) + \sum_{x \in X} p(x) p_x(y) \log_2 p_x(y)$$

is a concave function of $\mathcal{P}$, for every fixed $y \in Y$. Summation for all $y \in Y$ shows that $I(\mathcal{P})$ is concave, too.

Theorem 5. (Useful estimates with the I-divergence).

Let $\mathcal{P} = \{ p(x), x \in X \}$ and $\mathcal{Q} = \{ q(x), x \in X \}$ be two PD's on X. Then

$$\sum_{x \in X} | p(x) - q(x) | \leq \sqrt{\frac{2}{\log_2 e} I(\mathcal{P} \| \mathcal{Q})} \tag{37}$$

$$\sum_{x \in X} p(x) \left| \log_2 \frac{p(x)}{q(x)} \right| \leq I(\mathcal{P} \| \mathcal{Q}) + \min \left(\frac{2 \log_2 e}{e}, \sqrt{2 \log_2 e - I(\mathcal{P} \| \mathcal{Q})} \right). \tag{38}$$

Proof. Let $A = \{ x : p(x) \leq q(x) \}$, $B = \{ x : p(x) > q(x) \}$; put $p(A) = p$, $q(A) = q$. Then $p \leq q$, $p(B) = 1 - p$, $q(B) = 1 - q$,

$$\sum_{x \in X} | p(x) - q(x) | = 2 (q - p), \tag{39}$$

while from (11) and (12) it follows

$$I(\mathcal{P} \| \mathcal{Q}) \geq p \log_2 \frac{p}{q} + (1 - p) \log_2 \frac{1 - p}{1 - q}. \tag{40}$$

A simple calculation shows that

$$p \log_2 \frac{p}{q} + (1 - p) \log_2 \frac{1 - p}{1 - q} - 2 \log_2 e \cdot (p - q)^2 \geq 0 \qquad (0 \leq p \leq q \leq 1) \tag{41}$$

(for $p = q$ the equality holds and the derivative of the left hand side of (41) with respect to p is ≤ 0 if $0 < p \leq q < 1$).

The relations (39), (40), (41) prove (37).

From (11) and (12) it also follows

$$\sum_{x \in X} p(x) \left| \log_2 \frac{p(x)}{q(x)} \right| = I(\mathcal{P} \| \mathcal{Q}) - 2 \sum_{x \in A} p(x) \log_2 \frac{p(x)}{q(x)} \leq \leq I(\mathcal{P} \| \mathcal{Q}) - 2 p \log_2 \frac{p}{q} = I(\mathcal{P} \| \mathcal{Q}) + 2 p \log_2 \frac{q}{p}. \tag{42}$$

Here

$$2p \log_2 \frac{q}{p} = 2 \log_2 e \cdot p \ln \frac{q}{p} \leqq 2 \log_2 e \cdot p \ln \frac{1}{p} \leqq \frac{2 \log_2 e}{e}$$

(since $f(t) = t \ln \frac{1}{t}$ takes on its maximum for $t = \frac{1}{e}$); further, more as $\ln \frac{q}{p} = \ln \left(1 + \frac{q-p}{p}\right) \leqq \frac{q-p}{p}$, we also have (using (39)) $2p \log_2 \frac{q}{p} \leqq 2 \log_2 e \cdot (q-p) = \log_2 e \cdot \sum_{x \in X} |p(x) - q(x)|$. In view of these estimates, (42) and (37) imply (38).