Entropy, **2001**, *3*, 162–190

entropy
ISSN 1099-4300
www.mdpi.org/entropy/

Basic Concepts, Identities and Inequalities – the Toolkit of Information Theory

F. Topsøe*

Department of Mathematics; University of Copenhagen; Denmark
email: topsoe@math.ku.dk

Received: 12 September 2001/ Accepted: 18 September 2001 / Published: 30 September 2001

Abstract: Basic concepts and results of that part of Information Theory which is often referred to as "Shannon Theory" are discussed with focus mainly on the discrete case. The paper is expository with some new proofs and extensions of results and concepts.
Keywords: Entropy, divergence, mutual information, concavity, convexity, datareduction.

---

*Research supported by the Danish Natural Science Research Council, by INTAS (project 00-738) and by IMPAN–BC, the Banach Center, Warsaw.

1 Codes

Though we are not interested in technical coding, the starting point of Information Theory may well be taken there. Consider Table 1. It shows a codebook pertaining to the first six letters of the alphabet. The code this defines maps the letters to binary code words. The efficiency is determined by the code word lengths, respectively 3,4,3,3,1 and 4. If the frequencies of individual letters are known, say respectively 20,8,15,10, 40 and 7 percent, efficiency can be related to the average code length, in the example equal to 2,35 measured in bits (binary digits). The shorter the average code length, the higher the efficiency. Thus average code length may be taken as the key quantity to worry about. It depends on the distribution (P) of the letters and on the code (κ). Actually, it does not depend on the internal structure of the code words, only on the associated code word lengths. Therefore, we take κ to stand for the map providing these lengths ($\kappa(a) = 3, \cdots, \kappa(f) = 4$). Then average code length may be written, using bracket notation, as $\langle \kappa, P \rangle$.

a	1	0	0	
b	1	1	1	0
c	1	0	1	
d	1	1	0	
e	0			
f	1	1	1	1

Table 1: A codebook

Clearly, not every map which maps letters to natural numbers is acceptable as one coming from a "sensible" code. We require that the code is *prefix–free*, i.e. that no codeword in the codebook can be the beginning of another codeword in the codebook. The good sense in this requirement may be realized if we imagine that the binary digits in a codeword corresponding to an initially unknown letter is revealed to us one by one e.g. by a "guru" as replies to a succession of questions: "is the first digit a 1?", "is the second digit a 1?" etc. The prefix–free property guarantees the "instantaneous" nature of the procedure. By this we mean that once we receive information which is consistent with one of the codewords in the codebook, we are certain which letter is the one we are looking for.

The code shown in Table 1 is *compact* in the sense that we cannot make it more efficient, i.e. decrease one or more of the code lengths, by simple operations on the code, say by deleting one or more binary digits. With the above background we can now prove the key result needed to get the theory going.

Theorem 1.1 (Kraft's inequality). *Let $\mathbb{A}$ be a finite or countably infinite set, the* alphabet, *and κ any map of $\mathbb{A}$ into $\mathbb{N}_0 = \{0, 1, 2, \dots\}$. Then the necessary and sufficient condition that there exists a prefix–free code of $\mathbb{A}$ with code lenghts as prescribed by κ is that* Kraft's inequality

$$\sum_{i \in \mathbb{A}} 2^{-\kappa(i)} \le 1 \tag{1.1}$$

holds. Furthermore, Kraft's equality

$$\sum_{i \in \mathbb{A}} 2^{-\kappa(i)} = 1 \tag{1.2}$$

holds, if and only if there exists no prefix–free code of $\mathbb{A}$ with code word lenghts given by a function ρ such that $\rho(i) \le \kappa(i)$ for all $i \in \mathbb{A}$ and $\rho(i_0) < \kappa(i_0)$ for some $i_0 \in \mathbb{A}$.

Proof. With every binary word we associate a binary interval contained in the unit interval $[0; 1]$ in the "standard" way. Thus, to the empty codeword, which has length 0, we associate $[0; 1]$, and if $J \subseteq [0; 1]$ is the binary interval associated with $\varepsilon_1 \cdots \varepsilon_k$, then we associate the left half of J with $\varepsilon_1 \cdots \varepsilon_k 0$ and the right half with $\varepsilon_1 \cdots \varepsilon_k 1$. To any collection of possible code words associated with the elements ("letters") in $\mathbb{A}$ we can then associate a family of binary sub–intervals of $[0; 1]$, indexed by the letters in $\mathbb{A}$ – and vice versa. We realize that in this way the prefix–free property corresponds to the property that the associated family of binary intervals consists of pairwise disjoint sets. A moments reflection shows that all parts of the theorem follow from this observation. $\square$

In the sequal, $\mathbb{A}$ denotes a finite or countably infinite set, the *alphabet*.

We are not interested in combinatorial or other details pertaining to actual coding with binary codewords. For the remainder of this paper we idealize by allowing arbitrary non–negative numbers as code lengths. Then we may as well consider e as a base for the exponentials occuring in Kraft's inequality. With this background we define a *general code* of $\mathbb{A}$ as a map $\kappa : \mathbb{A} \to [0; \infty]$ such that

$$\sum_{i \in \mathbb{A}} e^{-\kappa_i} \le 1 \,, \tag{1.3}$$

and a *compact code* as a map $\kappa : \mathbb{A} \to [0; \infty]$ such that

$$\sum_{i \in \mathbb{A}} e^{-\kappa_i} = 1 \,. \tag{1.4}$$

The set of general codes is denoted $\sim\!K(\mathbb{A})$, the set of compact codes $K(\mathbb{A})$. The number κ_i above is now preferred to $\kappa(i)$ and referred to as the *code length* associated with i. The compact codes are the most important ones and, for short, they are referred to simply as *codes*.

2 Entropy, redundancy and divergence

The set of *probability distributions* on $\mathbb{A}$, just called *distributions* or sometimes *sources*, is denoted $M_+^1(\mathbb{A})$ and the set of non–negative measures on $\mathbb{A}$ with total mass at most 1, called *general distributions*, is denoted $\tilde{}M_+^1(\mathbb{A})$. Distributions in $\tilde{}M_+^1(\mathbb{A}) \setminus M_+^1(\mathbb{A})$ are *incomplete distributions*. For $P, Q, \cdots$ in $\tilde{}M_+^1(\mathbb{A})$ the corresponding point probabilities are denoted by $p_i, q_i, \cdots$.

There is a natural bijective correspondence between $\tilde{}M_+^1(\mathbb{A})$ and $\tilde{}K(\mathbb{A})$, expressed notationally by writing $P \leftrightarrow \kappa$ or $\kappa \leftrightarrow P$, and defined by the formulas

$$\kappa_i = -\log p_i\,, \quad p_i = e^{-\kappa_i}\,.$$

Here, log is used for natural logarithms. Note that the values $\kappa_i = \infty$ and $p_i = 0$ correspond to eachother. When the above formulas hold, we call (κ, P) a *matching pair* and we say that κ is *adapted* to P or that P is the general distribution which *matches* κ.

As in Section 1, $\langle \kappa, P \rangle$ denotes *average code length*. We may now define *entropy* as minimal average code length:

$$H(P) = \min_{\kappa \in \tilde{}K(\mathbb{A})} \langle \kappa, P \rangle\,, \tag{2.5}$$

and *redundancy* $D(P\|\kappa)$ as actual average code length minus minimal average code length, i.e.

$$D(P\|\kappa) = \langle \kappa, P \rangle - H(P)\,. \tag{2.6}$$

Some comments are in order. In fact (2.6) may lead to the indeterminate form $\infty - \infty$. Nevertheless, $D(P\|\kappa)$ may be defined as a definite number in $[0;\infty]$ in all cases. Technically, it is convenient first to define *divergence* $D(P\|Q)$ between a probability distribution P and a, possibly incomplete, distribution Q by

$$D(P\|Q) = \sum_{i \in \mathbb{A}} p_i \log \frac{p_i}{q_i}\,. \tag{2.7}$$

Theorem 2.1. *Divergence between $P \in M_+^1(\mathbb{A})$ and $Q \in \tilde{}M_+^1(\mathbb{A})$ as given by* (2.7) *is a well defined number in* $[0;\infty]$ *and* $D(P\|Q) = 0$ *if and only if* $P = Q$.

Entropy defined by (2.5) *also makes sense and the minimum is attained for the code adapted to* P, *i.e.*

$$H(P) = -\sum_{i \in \mathbb{A}} p_i \log p_i\,. \tag{2.8}$$

If $H(P) < \infty$, *the minimum is only attained for the code adapted to* P.

Finally, for every $P \in M_+^1(\mathbb{A})$ and $\kappa \in {}^\sim K(\mathbb{A})$, the following identity holds with Q the distribution matching κ:

$$\langle \kappa, P \rangle = H(P) + D(P \| Q). \tag{2.9}$$

Proof. By the inequality

$$x \log \frac{x}{y} = -x \log \frac{y}{x} \geq x - y \tag{2.10}$$

we realize that the sum of negative terms in (2.7) is bounded below by -1, hence $D(P\|Q)$ is well defined. The same inequality then shows that $D(P\|Q) \geq 0$. The discussion of equality is easy as there is strict inequality in (2.10) in case $x \neq y$.

The validity of (2.9) with $-\sum p_i \log p_i$ in place of $H(P)$ then becomes a triviality and (2.8) follows. □

The simple identity (2.9) is important. It connects three basic quantities: entropy, divergence and average code length. We call it the *linking identity*. Among other things, it shows that in case $H(P) < \infty$, then the definition (2.6) yields the result $D(P\|\kappa) = D(P\|Q)$ with $Q \leftrightarrow \kappa$. We therefore now define *redundancy* $D(P\|\kappa)$, where $P \in M_+^1(\mathbb{A})$ and $\kappa \in {}^\sim K(\mathbb{A})$ by

$$D(P\|\kappa) = D(P\|Q)\,;\ Q \leftrightarrow \kappa\,. \tag{2.11}$$

Divergence we think of, primarily, as just a measure of discrimination between P and Q. Often it is more appropriate to think in terms of redundancy as indicated in (2.6). Therefore, we often write the linking identity in the form

$$\langle \kappa, P \rangle = H(P) + D(P\|\kappa)\,. \tag{2.12}$$

3 Some topological considerations

On the set of general distributions ${}^\sim M_+^1(\mathbb{A})$, the natural topology to consider is that of pointwise convergence. When restricted to the space $M_+^1(\mathbb{A})$ of probability distributions this topology coincides with the topology of convergence in total variation (ℓ^1–convergence). To be more specific, denote by $V(P,Q)$ the *total variation*

$$V(P,Q) = \sum_{i \in \mathbb{A}} |p_i - q_i|\,. \tag{3.13}$$

Then we have

Lemma 3.1. *Let $(P_n)_{n\geq 1}$ and P be probability distributions over $\mathbb{A}$ and assume that $(P_n)_{n\geq 1}$ converges pointwise to P, i.e. $p_{n,i} \to p_i$ as $n \to \infty$ for every $i \in \mathbb{A}$. Then P_n converges to P in total variation, i.e. $V(P_n, P) \to 0$ as $n \to \infty$.*

Proof. The result is known as Scheffé's lemma (in the discrete case). To prove it, consider $P_n - P$ as a function on $\mathbb{A}$. The negative part $(P_n - P)^-$ converges pointwise to 0 and for all n, $0 \leq (P_n - P)^- \leq P$, hence, e.g. by Lebesgue's dominated convergence theorem, $\sum_{i\in\mathbb{A}}(P_n - P)^-(i) \to 0$. As, for the positive part, $\sum_{i\in\mathbb{A}}(P_n - P)^+(i) = \sum_{i\in\mathbb{A}}(P_n - P)^-(i)$, we find that also $\sum_{i\in\mathbb{A}}(P_n - P)^+(i)$ converges to 0. As $V(P_n, P) = \sum_{i\in\mathbb{A}} |P_n - P|(i)$ and, generally, $|x| = x^+ + x^-$, we now conclude that $V(P_n, P) \to 0$. □

We denote convergence in $M_+^1(\mathbb{A})$ by $P_n \xrightarrow{V} P$. As the lemma shows, it is immaterial if we here have the topology of pointwise convergence or the topology of convergence in total variation in mind.

Another topological notion of convergence in $M_+^1(\mathbb{A})$ is expected to come to play a significant role but has only recently been discovered. This is the notion defined as follows: For $(P_n)_{n\geq 1} \subseteq M_+^1(\mathbb{A})$ and $P \in M_+^1(\mathbb{A})$, we say that $(P_n)_{n\geq 1}$ *converges in divergence* to P, and write $P_n \xrightarrow{D} P$, if $D(P_n \| P) \to 0$ as $n \to \infty$. The new and somewhat unexpected observation is that this is indeed a topological notion. In fact, there exists a strongest topology on $M_+^1(\mathbb{A})$, the *information topology*, such that $P_n \xrightarrow{D} P$ implies that $(P_n)_{n\geq 1}$ converges in the topology to P and for this topology, convergence in divergence and in the topology are equivalent concepts. We stress that this only holds for ordinary sequences and does not extend to generalized sequences or nets. A subset $\mathcal{P} \subseteq M_+^1(\mathbb{A})$ is open in the information topology if and only if, for any sequence $(P_n)_{n\geq 1}$ with $P_n \to P$ and $P \in \mathcal{P}$, one has $P_n \in \mathcal{P}$, eventually. Equivalently, $\mathcal{P}$ is closed if and only if $(P_n)_{n\geq 1} \subseteq \mathcal{P}$, $P_n \xrightarrow{D} P$ implies $P \in \mathcal{P}$.

The quoted facts can either be proved directly or they follow from more general results, cf. [7] or [1]. We shall not enter into this here but refer the reader to [6].

Convergence in divergence is, typically, a much stronger notion than convergence in total variation. This follows from Pinsker's inequality $D(P\|Q) \geq \frac{1}{2}V(P, Q)^2$ which we shall prove in Section 4. In case $\mathbb{A}$ is finite, it is easy to see that the convergence $P_n \xrightarrow{D} P$ amounts to usual convergence $P_n \xrightarrow{V} P$ and to the equality $\mathrm{supp}\,(P_n) = \mathrm{supp}\,(P)$ for n sufficiently large. Here, "supp" denotes *support*, i.e. the set of elements in $\mathbb{A}$ with positive probability.

We turn to some more standard considerations regarding lower semi–continuity. It is an important fact that entropy and divergence are lower semi–continuous, even with respect to the usual topology. More precisely:

Theorem 3.2. *With respect to the usual topology, the following continuity results hold:*

(i) *The entropy function* $H : M_+^1(\mathbb{A}) \to [0;\infty]$ *is lower semi–continuous and, if* $\mathbb{A}$ *is finite,* H *is continuous.*

(ii) *Divergence* $D : M_+^1(\mathbb{A}) \times^{\sim} M_+^1(\mathbb{A}) \to [0;\infty]$ *is jointly lower semi–continuous.*

Proof. We need a general abstract result: *Let* X *be a topological space and let* $(\varphi_n)_{n\geq 1}$ *be a sequence of lower semi–continuous functions* $\varphi_n : X \to]-\infty;\infty]$ *and assume that* $\varphi = \sum_1^\infty \varphi_n$ *is a well defined function* $\varphi : X \to]-\infty;\infty]$. *Assume also that there exist continuous minorants* $\psi_n : X \to]-\infty;\infty[$, *i.e.* $\psi_n \leq \varphi_n$; $n \geq 1$, *such that the sum* $\psi = \sum_1^\infty \psi_n$ *is a well defined continuous function* $\psi : X \to]-\infty;\infty[$. *Then* φ *is lower semi–continuous.*

To prove this auxiliary result, let (x_ν) be an ordinary or generalized sequence and x an element of X such that $x_\nu \to x$. We have to prove that $\liminf \varphi(x_\nu) \geq \varphi(x)$. Fix $N \in \mathbb{N}$ and use the fact that a finite sum of lower semi–continuous functions is lower semi–continuous to conclude that

$$\begin{aligned}
\liminf \varphi(x_\nu) &= \liminf\left((\varphi-\psi)(x_\nu) + \psi(x_\nu)\right)\\
&= \liminf(\varphi-\psi)(x_\nu) + \psi(x)\\
&\geq \liminf \sum_{n=1}^{N}(\varphi_n - \psi_n)(x_\nu) + \psi(x)\\
&\geq \sum_{n=1}^{N}(\varphi_n - \psi_n)(x) + \psi(x)\,.
\end{aligned}$$

As this holds for all $N \in \mathbb{N}$, we conclude that

$$\liminf \varphi(x_\nu) \geq \sum_{n=1}^{\infty}(\varphi_n - \psi_n)(x) + \psi(x) = \varphi(x)\,,$$

as desired.

In particular, *a sum of non-negative real valued lower semi–continuous functions is lower semi–continuous.* The statement (i) follows from this fact as $x \curvearrowright -x \log x$ is non–negative and continuous on $[0;1]$.

We now turn to the proof of (ii). First we remark that we may restrict attention to D defined on the space $M_+^1(\mathbb{A}) \times M_+^1(\mathbb{A})$. To see this, take any (ordinary or generalized) sequence $(P_\nu, Q_\nu)_\nu$ in $M_+^1(\mathbb{A}) \times^{\sim} M_+^1(\mathbb{A})$ which converges in that space to (P, Q). By taking a proper subsequence if necessary, we may assume that the sequence $(D(P_\nu\|Q_\nu))_\nu$ is convergent and also that $(Q_\nu(\mathbb{A}))_\nu$ converges. Then we may add a point – the "point at infinity" – to the space $\mathbb{A}$ and extend all measures considered to the new space in a natural way such that all measures become probability distributions. Denoting the extended measures with a star, we then find that $(P_\nu^*, Q_\nu^*)_\nu \to (P^*, Q^*)$ and we have $\liminf_\nu D(P_\nu^*\|Q_\nu^*) \geq D(P^*\|Q^*)$, provided lower semi-continuity has been estableshed for true probability distributions. Then $\lim_\nu D(P_\nu \| Q_\nu) \geq D(P\|Q)$ follows and we conclude that the desired semi-continuity property also holds if the Q's are allowed to be improper distributions.

To prove (ii) we may thus restrict attention to the space $M_+^1(\mathbb{A}) \times M_+^1(\mathbb{A})$. Further, we may assume that $\mathbb{A} = \mathbb{N}$. For each n, denote by φ_n the map $M_+^1(\mathbb{A}) \times M_+^1(\mathbb{A}) \to]-\infty;\infty]$ defined by $(P,Q) \curvearrowright p_n \log \frac{p_n}{q_n}$. Then φ_n is lower semi–continuous. Denote by ψ_n the map $(P,Q) \curvearrowright p_n - q_n$. Then ψ_n is a continuous minorant to φ_n and $\sum_1^\infty \psi_n = 0$. As $D = \sum_1^\infty \varphi_n$, the auxiliary result applies and the desired conclusion follows. □

In Section 5 we shall investigate further the topological properties of H and D. For now we point out one simple continuity property of divergence which has as point of departure, not so much convergence in the space $M_+^1(\mathbb{A})$, but more so convergence in $\mathbb{A}$ itself. The result we have in mind is only of interest if $\mathbb{A}$ is infinite as it considers approximations of $\mathbb{A}$ with finite subsets. Denote by $\mathcal{P}_0(\mathbb{A})$ the set of finite subsets of $\mathbb{A}$, ordered by inclusion. Then $(\mathcal{P}_0(\mathbb{A}), \subseteq)$ is an upward directed set and we can consider convergence along this set, typically denoted by $\lim_{A\in\mathcal{P}_0(\mathbb{A})}$.

Theorem 3.3. *For any* $P, Q \in M_+^1(\mathbb{A})$

$$\lim_{A\in\mathcal{P}_0(\mathbb{A})} D(P|A\|Q) = D(P\|Q),$$

$P|A$ *denoting as usual the conditional distribution of* P *given* A.

Proof. First note that the result makes good sense as $P(A) > 0$ if A is large enough. The result follows by writing $D(P|A\|Q)$ in the form

$$D(P|A\|Q) = \log \frac{1}{P(A)} + \frac{1}{P(A)} \sum_{x\in A} P(x) \log \frac{P(x)}{Q(x)}$$

since $\lim_{A\in\mathcal{P}_0(\mathbb{A})} P(A) = 1$ and since

$$\lim_{A\in\mathcal{P}_0(\mathbb{A})} \sum_{x\in A} P(x) \log \frac{P(x)}{Q(x)} = D(P\|Q).$$

□

4 Datareduction

Let, again, $\mathbb{A}$ be the alphabet and consider a decomposition θ of $\mathbb{A}$. We shall think of θ as defining a *datareduction*. We often denote the classes of θ by A_i with i ranging over a certain index set which, in pure mathematical terms, is nothing but the quotient space of $\mathbb{A}$ w.r.t. θ. We denote this quotient space by $\partial\mathbb{A}$ – or, if need be, by $\partial_\theta\mathbb{A}$ – and call $\partial\mathbb{A}$ the *derived* alphabet (the alphabet *derived from the datareduction* θ). Thus $\partial\mathbb{A}$ is nothing but the set of classes for the decomposition θ.

Now assume that we have also given a source $P \in M_+^1(\mathbb{A})$. By ∂P (or $\partial_\theta P$) we denote the *derived source*, defined as the distribution $\partial P \in M_+^1(\partial\mathbb{A})$ of the quotient map $\mathbb{A} \to \partial\mathbb{A}$ or, if you prefer, as the image measure of P under the quotient map. Thus, more directly, ∂P is the probability distribution over $\partial\mathbb{A}$ given by

$$(\partial P)(A) = P(A)\,; \quad A \in \partial\mathbb{A}.$$

If we choose to index the classes in $\partial\mathbb{A}$ we may write $\partial P(A_i) = P(A_i)$, $i \in \partial\mathbb{A}$.

Remark. Let $\mathbb{A}_0$ be a basic alphabet, e.g. $\mathbb{A}_0 = \{0,1\}$ and consider natural numbers s and t with $s < t$. If we take $\mathbb{A}$ to be the set of words $x_1 \cdots x_t$ of length t from the alphabet $\mathbb{A}_0$, i.e. $\mathbb{A} = \mathbb{A}_0^t$, and θ to be the decomposition induced by the projection of $\mathbb{A}$ onto $\mathbb{A}_0^s$, then the quotient space $\partial\mathbb{A}_0^t$ can be identifyed with the set $\mathbb{A}_0^s$. The class corresponding to $x_1 \cdots x_s \in \mathbb{A}_0^s$ consists of all strings $y_1 \cdots y_t \in \mathbb{A}_0^t$ with $x_1 \cdots x_s$ as prefix.

In this example, we may conveniently think of $x_1 \cdots x_s$ as representing the past (or the known history) and $x_{s+1} \cdots x_t$ to represent the future. Then $x_1 \cdots x_t$ represents past + future.

Often, we think of a datareduction as modelling either conditioning or given information. Imagine, for example, that we want to observe a random element $x \in \mathbb{A}$ which is govorned by a distribution P, and that direct observation is impossible (for practical reasons or because the planned observation involves what will happen at some time in the future, cf. Example **??**). Instead, partial information about x is revealed to us via θ, i.e. we are told which class $A_i \in \partial\mathbb{A}$ the element x belongs to. Thus "$x \in A_i$" is a piece of information (or a condition) which partially determines x.

Considerations as above lie behind two important definitions: By the *conditional entropy* of P *given* θ we understand the quantity

$$H^\theta(P) = \sum_{i \in \partial\mathbb{A}} P(A_i) H(P|A_i). \tag{4.14}$$

As usual, $P|A_i$ denotes the conditional distribution of P given A_i (when well defined). Note that when $P|A_i$ is undefined, the corresponding term in (4.14) is, nevertheless, well defined (and $= 0$).

Note that the conditional entropy is really the *average* uncertainty (entropy) that remains after the information about θ has been revealed.

Similarly, the *conditional divergence* between P and Q *given* θ is defined by the equation

$$D^{\theta}(P\|Q) = \sum_{i \in \partial \mathbb{A}} P(A_i) D(P|A_i \| Q|A_i). \tag{4.15}$$

There is one technical comment we have to add to this definition: It is possible that for some i, $P(A_i) > 0$ whereas $Q(A_i) = 0$. In such cases $P|A_i$ is welldefined whereas $Q|A_i$ is not. We agree that in such cases, $D^{\theta}(P\|Q) = \infty$. This corresponds to an extension of the basic definition of divergence by agreering that the divergence between a (well defined) distribution and some undefined distribution is infinite.

In analogy with the interpretation regarding entropy, note that, really, $D^{\theta}(P\|Q)$ is the *average* divergence after information about θ has been revealed.

We also note that $D^{\theta}(P\|Q)$ does not depend on the full distribution Q but only on the family $(Q|A_i)$ of conditional distributions (with i ranging over indices with $P(A_i) > 0$). Thinking about it, this is also quite natural: If Q is conceived as a predictor then, if we know that information about θ will be revealed to us, the only thing we need to predict is the conditional distributions given the various A_i's.

Whenever convenient we will write $H(P|\theta)$ in place of $H^{\theta}(P)$ whereas a similar notation for divergence appears awquard and will not be used.

From the defining relations (4.14) and (4.15) it is easy to identify circumstances under which $H^{\theta}(P)$ or $D^{\theta}(P\|Q)$ vanish. For this we need two new notions: We say that P is *deterministic modulo* θ, and write $P = 1 \pmod{\theta}$, provided the conditional distribution $P|A_i$ is deterministic for every i with $P(A_i) > 0$. And we say that Q *equals* P *modulo* θ, and write $Q = P \pmod{\theta}$, provided $Q|A_i = P|A_i$ for every i with $P(A_i) > 0$. This condition is to be understood in the sense that if $P(A_i) > 0$, the conditional distribution $Q|A_i$ must be well defined (i.e. $Q(A_i) > 0$) and coincide with $P|A_i$. The new notions may be expressed in a slightely different way as follows:

$$P = 1\ (\mathrm{mod}\theta) \Leftrightarrow \forall i \exists x \in A_i : P(A_i \setminus \{x\}) = 0$$

and

$$Q = P\ (\mathrm{mod}\ \theta) \Leftrightarrow \forall P(A_i) > 0 \exists c > 0 \forall x \in A_i : Q(x) = c \cdot P(x).$$

It should be noted that the relation "equality mod θ" is not symmetric: The two statements $Q = P \pmod{\theta}$ and $P = Q \pmod{\theta}$ are only equivalent if, for every i, $P(A_i) = 0$ if and only if $Q(A_i) = 0$.

We leave the simple proof of the following result to the reader:

Theorem 4.1. *(i)* $H^{\theta}(P) \geq 0$, *and a necessary and sufficient condition that* $H^{\theta}(P) = 0$ *is that* P *be deterministic modulo* θ.

(ii) $D^\theta(P\|Q) \geq 0$, *and a necessary and sufficient condition that* $D^\theta(P\|Q) = 0$, *is that* Q *be equal to* P *modulo* θ.

Intuitively, it is to be expected that entropy and divergence decrease under datareduction: $H(P) \geq H(\partial P)$ and $D(P\|Q) \geq D(\partial P\|\partial Q)$. Indeed, this is so and we can even identify the amount of the decrease in information theoretical terms:

Theorem 4.2 (datareduction identities). *Let* P *and* Q *be distributions over* $\mathbb{A}$ *and let* θ *denote a datareduction. Then the following two identities hold:*

$$H(P) = H(\partial P) + H^\theta(P), \tag{4.16}$$

$$D(P\|Q) = D(\partial P\|\partial Q) + D^\theta(P\|Q). \tag{4.17}$$

The identity (4.16) is called *Shannon*'s *identity* (most often given in a notation involving random variables, cf. Section 7).

Proof. Below, sums are over i with $P(A_i) > 0$. For the right hand side of (4.16) we find the expression

$$-\sum_i P(A_i)\log P(A_i) - \sum_i P(A_i)\sum_{x\in A_i}\frac{P(x)}{P(A_i)}\log\frac{P(x)}{P(A_i)}$$

which can be rewritten as

$$-\sum_i\sum_{x\in A_i} P(x)\log P(A_i) - \sum_i\sum_{x\in A_i} P(x)\log\frac{P(x)}{P(A_i)},$$

easily recognizable as the entropy $H(P)$.

For the right hand side of (4.17) we find the expression

$$\sum_i P(A_i)\log\frac{P(A_i)}{Q(A_i)} + \sum_i P(A_i)\sum_{x\in A_i}\frac{P(x)}{P(A_i)}\log\left(\frac{P(x)}{P(A_i)}\Big/\frac{Q(x)}{Q(A_i)}\right)$$

which can be rewritten as

$$\sum_i\sum_{x\in A_i} P(x)\log\frac{P(A_i)}{Q(A_i)} + \sum_i\sum_{x\in A_i} P(x)\log\left(\frac{P(x)Q(A_i)}{Q(x)P(A_i)}\right),$$

easily recognizable as the divergence $D(P\|Q)$. □

Of course, these basic identities can, more systematically, be written as $H(P) = H(\partial_\theta P) + H^\theta(P)$ and similarly for (4.17). An important corollary to Theorems 4.1 and 4.2 is the following

Corollary 4.3 (datareduction inequalities). *With notation as above the following results hold:*

(i). $H(P) \geq H(\partial P)$ and, in case $H(P) < \infty$, equality holds if and only if P is deterministic modulo θ.

(ii). $D(P\|Q) \geq D(\partial P\|\partial Q)$ and, in case $D(P\|Q) < \infty$, equality holds if and only if Q equals P modulo θ.

Another important corollary is obtained by emphazising conditioning instead of datareduction in Theorem 4.2:

Corollary 4.4 (inequalities under conditioning). *With notation as above, the following results hold:*

(i). (Shannon's inequality for conditional entropy). $H(P) \geq H^{\theta}(P)$ and, in case $H(P) < \infty$, equality holds if and only if the support of P is contained in one of the classes A_i defined by θ.

(ii). $D(P\|Q) \geq D^{\theta}(P\|Q)$ and, in case $D(P\|Q) < \infty$, equality holds if and only if $\partial P = \partial Q$, i.e. if and only if, for all classes A_i defined by θ, $P(A_i) = Q(A_i)$ holds.

We end this section with an important inequality mentioned in Section 3:

Corollary 4.5 (Pinsker's inequality). *For any two probability distributions,*

$$D(P\|Q) \geq \frac{1}{2}V(P,Q)^2 \,. \tag{4.18}$$

Proof. Put $A^+ = \{i \in \mathbb{A} \mid p_i \geq q_i\}$ and $A^- = \{i \in \mathbb{A} \mid p_i < q_1\}$. By Corollary 4.3, $D(P\|Q) \geq D(\partial P\|\partial Q)$ where ∂P and ∂Q refer to the datareduction defined by the decomposition $\mathbb{A} = A^+ \cup A^-$. Put $p = P(A^+)$ and $q = P(A^-)$. Keep p fixed and assume that $0 \leq q \leq p$. Then

$$D(\partial P\|\partial Q) - \frac{1}{2}V(P,Q)^2 = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q} - 2(p-q)^2$$

and elementary considerations via differentiation w.r.t. q (two times!) show that this expression is non–negative for $0 \leq q \leq p$. The result follows. □

5 Approximation with finite partition

In order to reduce certain investigations to cases which only involve a finite alphabet and in order to extend the definition of divergence to general Borel spaces, we need a technical result on approximation with respect to finer and finer partitions.

We leave the usual discrete setting and take an arbitrary Borel space $(\mathbb{A}, \mathcal{A})$ as our basis. Thus $\mathbb{A}$ is a set, possibly uncountable, and $\mathcal{A}$ a Borel structure (the same as a σ-algebra) on $\mathbb{A}$. By

$\Pi_\sigma(\mathbb{A},\mathcal{A})$ we denote the set of countable decompositions of $\mathbb{A}$ in measurable sets (sets in $\mathcal{A}$), ordered by subdivision. We use "$\prec$" to denote this ordering, i.e. for $\pi,\rho\in\Pi_\sigma(\mathbb{A},\mathcal{A})$, $\pi\prec\rho$ means that every class in π is a union of classes in ρ. By $\pi\vee\rho$ we denote the coarsest decomposition which is finer than both π and ρ, i.e. $\pi\vee\rho$ consists of all non–empty sets of the form $A\cap B$ with $A\in\pi$, $B\in\rho$.

Clearly, $\Pi_\sigma(\mathbb{A},\mathcal{A})$ is an upward directed set, hence we may consider limits based on this set for which we use natural notation such as $\lim_\pi$, $\liminf_\pi$ etc.

By $\Pi_0(\mathbb{A},\mathcal{A})$ we denote the set of finite decompositions in $\Pi_\sigma(\mathbb{A},\mathcal{A})$ with the ordering inherited from $\Pi_\sigma(\mathbb{A},\mathcal{A})$. Clearly, $\Pi_0(\mathbb{A},\mathcal{A})$ is also an upward directed set.

By $M_+^1(\mathbb{A},\mathcal{A})$ we denote the set of probability measures on $(\mathbb{A},\mathcal{A})$. For $P\in M_+^1(\mathbb{A},\mathcal{A})$ and $\pi\in\Pi_\sigma(\mathbb{A},\mathcal{A})$, $\partial_\pi P$ denotes the *derived distributions* defined in consistency with the definitions of the previous section. If $\mathcal{A}_\pi$ denotes the σ–algebra generated by π, $\partial_\pi P$ may be conceived as a measure in $M_+^1(\mathbb{A},\mathcal{A}_\pi)$ given by the measures of the atoms of $\mathcal{A}_\pi$: $(\partial_\pi P)(A)=P(A)$ for $A\in\pi$. Thus

$$H(\partial_\pi P)=-\sum_{A\in\pi}P(A)\log P(A)\,,$$
$$D(\partial_\pi P\|\partial_\pi Q)=\sum_{A\in\pi}P(A)\log\frac{P(A)}{Q(A)}\,.$$

In the result below, we use Π_0 to denote the set $\Pi_0(\mathbb{A},\mathcal{A})$.

Theorem 5.1 (continuity along finite decompositions). *Let* $(\mathbb{A},\mathcal{A})$ *be a Borel space and let* $\tau\in\Pi_\sigma(\mathbb{A},\mathcal{A})$.

(i) For any $P\in M_+^1(\mathbb{A},\mathcal{A})$,

$$H(\partial_\tau P)=\lim_{\pi\in\Pi_0,\pi\prec\tau}H(\partial_\pi P)=\sup_{\pi\in\Pi_0,\pi\prec\tau}H(\partial_\pi P)\,. \tag{5.19}$$

(ii) For $P,Q\in M_+^1(\mathbb{A},\mathcal{A})$,

$$D\left(\partial_\tau P\|\partial_\tau Q\right)=\lim_{\pi\in\Pi_0,\pi\prec\tau}D\left(\partial_\pi P\|\partial_\pi Q\right)=\sup_{\pi\in\Pi_0,\pi\prec\tau}D\left(\partial_\pi P\|\partial_\pi Q\right)\,. \tag{5.20}$$

Proof. We realize that we may assume that $(\mathbb{A},\mathcal{A})$ is the discrete Borel structure $\mathbb{N}$ and that τ is the decomposition of $\mathbb{N}$ consisting of all singletons $\{n\}$; $n\in\mathbb{N}$.

For any non–empty subset A of $\mathbb{A}=\mathbb{N}$, denote by x_A the first element of A, and, for $P\in M_+^1(\mathbb{A})$ and $\pi\in\Pi_0$, we put

$$P_\pi=\sum_{A\in\pi}P(A)\delta_{x_A}$$

with δ_x denoting a unit mass at x. Then $P_\pi \xrightarrow{V} P$ along the directed set Π_0 and $H(\partial_\pi P) = H(P_\pi)$; $\pi \in \Pi_0$.

Combining lower semi–continuity and the datareduction inequality (i) of Corollary 4.3, we find that

$$\begin{aligned} H(P) &\le \liminf_{\pi\in\Pi_0} H(P_\pi) = \liminf_{\pi\in\Pi_0} H(\partial_\pi P) \le \limsup_{\pi\in\Pi_0} H(\partial_\pi P) \\ &\le \sup_{\pi\in\Pi_0} H(\partial_\pi P) \le H(P)\,, \end{aligned}$$

and (i) follows. The proof of (ii) is similar and may be summarized as follows:

$$\begin{aligned} D(P\|Q) &\le \liminf_{\pi\in\Pi_0} D(P_\pi\|Q_\pi) = \liminf_{\pi\in\Pi_0} D(\partial_\pi P\|\partial_\pi Q) \\ &\le \limsup_{\pi\in\Pi_0} D(\partial_\pi P\|\partial_\pi Q) \le \sup_{\pi\in\Pi_0} D(\partial_\pi P\|\partial_\pi Q) \le D(P\|Q)\,. \end{aligned}$$

□

As the sequences in (5.19) and in (5.20) are weakly increasing, we may express the results more economically by using the sign "↑" in a standard way:

$$\begin{aligned} &H(\partial_\pi P) \uparrow H(\partial_\tau P) \text{ as } \pi \in \Pi_0.\pi \prec \tau\,, \\ &D(\partial_\pi P\|\partial_\pi Q) \uparrow D(\partial_\tau P\|\partial_\tau Q) \text{ as } \pi \in \Pi_0, \pi \prec \tau\,. \end{aligned}$$

The type of convergence established also points to martingale-type of considerations, cf. [8] [1]

Motivated by the above results, we now extend the definition of divergence to cover probability distributions on arbitrary Borel spaces. For $P, Q \in M_+^1(\mathbb{A}, \mathcal{A})$ we simply define $D(P\|Q)$ by

$$D(P\|Q) = \sup_{\pi\in\Pi_0} D(\partial_\pi P\|\partial_\pi Q)\,. \tag{5.21}$$

By Theorem 5.1 we have

$$D(P\|Q) = \sup_{\pi\in\Pi_\sigma} D(\partial_\pi P\|\partial_\pi Q) \tag{5.22}$$

with $\Pi_\sigma = \Pi_\sigma(\mathbb{A}, \mathcal{A})$. The definition given is found to be the most informative when one recalls the separate definition given earlier for the discrete case, cf. (2.6) and (2.7). However, it is also important to note the following result which most authors use as definition. It gives a direct analytical expression for divergence which can be used in the discrete as well as in the general case.

[1] the author had access to an unpublished manuscript by Andrew R. Barron: *Information Theory and Martingales*, presented at the 1991 International Symposium on Information Theory in Budapest where this theme is pursued.

Theorem 5.2 (divergence in analytic form). *Let* $(\mathbb{A}, \mathcal{A})$ *be a Borel space and let* $P, Q \in M_+^1(\mathbb{A}, \mathcal{A})$. *Then*

$$D(P\|Q) = \int_{\mathbb{A}} \log \frac{dP}{dQ} dP \tag{5.23}$$

where $\frac{dP}{dQ}$ *denotes a version of the Radon–Nikodym derivative of* P *w.r.t.* Q. *If this derivative does not exist, i.e. if* P *is not absolutely continuous w.r.t.* Q, *then* (5.23) *is to be interpretated as giving the result* $D(P\|Q) = \infty$.

Proof. First assume that P is not absolutely continuous w.r.t. Q. Then $P(A) > 0$ and $Q(A) = 0$ for some $A \in \mathcal{A}$ and we see that $D(\partial_\pi P\|\partial_\pi Q) = \infty$ for the decomposition $\pi = \{A, \complement A\}$. By (5.21), $D(P\|Q) = \infty$ follows.

Then assume that P is absolutely continuous w.r.t. Q and put $f = \frac{dP}{dQ}$. Furthermore, put $I = \int \log f dP$. Then I can also be written as $\int \varphi(f) dQ$ with $\varphi(x) = x \log x$. As φ is convex, $\frac{1}{Q(A)} \int_A \varphi(f) dQ \geq \varphi\left(\frac{P(A)}{Q(A)}\right)$ for every $A \in \mathcal{A}$. It is then easy to show that $I \geq D(\partial_\pi P\|\partial_\pi Q)$ for every $\pi \in \Pi_\sigma$, thus $I \geq D(P\|Q)$.

In order to prove the reverse inequality, let $t < I$ be given and choose $s > t$ such that $I - \log s > t$. As $P(\{f = 0\}) = 0$, we find that

$$I = \sum_{n=-\infty}^{\infty} \int_{A_n} \log f \, dP$$

with

$$A_n = \{s^n \leq f < s^{n+1}\} \, ; \; n \in \mathbb{Z} \, .$$

Then, from the right–hand inequality of the double inequality

$$S^n Q(A_n) \leq P(A_n) \leq s^{n+1} Q(A_n) \, ; \; n \in \mathbb{Z} \, , \tag{5.24}$$

we find that

$$I \leq \sum_{n=-\infty}^{\infty} \log s^{n+1} \cdot P(A_n)$$

and, using also the left–hand inequality of (5.24), it follows that

$$I \leq \log s + D(\partial_\pi P\|\partial_\pi Q)$$

with $\pi = \{A_n | n \in \mathbb{Z}\} \cup \{f = 0\}$. It follows that $D(\partial_\pi P\|\partial_\pi Q) \geq t$. As $\pi \in \Pi_\sigma$, and as $t < I$ was arbitrary, it follows from (5.22) that $D(P\|Q) \leq I$. □

For the above discussion and results concerning divergence $D(P\|Q)$ between measures on arbitrary Borel spaces we only had the case of probability distributions in mind. However, it is easy to extend the discussion to cover also the case when Q is allowed to be an imcomplete distribution. Detals are left to the reader.

It does not make sense to extend the basic notion of entropy to distributions on general measure spaces as the natural quantity to consider, $\sup_\pi H(\partial_\pi P)$ with π ranging over Π_0 or Π_σ can only yield a finite quantity if P is essentially discrete.

6 Mixing, convexity properties

Convexity properties are of great significance in Information Theory. Here we develop the most important of these properties by showing that the entropy function is concave whereas divergence is convex.

The setting is, again, a discrete alphabet $\mathbb{A}$. On $\mathbb{A}$ we study various probability distributions. If $(P_\nu)_{\nu\geq 1}$ is a sequence of such distributions, then a *mixture* of these distributions is any distribution P_0 of the form

$$P_0 = \sum_{\nu=1}^{\infty} \alpha_\nu P_\nu \tag{6.25}$$

with $\alpha = (\alpha_\nu)_{\nu\geq 1}$ any probability vector ($\alpha_\nu \geq 0$ for $\nu \geq 1$, $\sum_1^\infty \alpha_\nu = 1$). In case $\alpha_\nu = 0$, eventually, (6.25) defines a normal *convex combination* of the P_ν's. The general case covered by (6.25) may be called an *ω-convex combination*[2].

As usual, a non-negative function $f : M_+^1(\mathbb{A}) \to [0,\infty]$ is *convex* (*concave*) if $f(\sum \alpha_\nu P_\nu) \leq \sum \alpha_\nu f(P_\nu)$ ($f(\sum \alpha_\nu P_\nu) \geq \sum \alpha_\nu f(P_\nu)$) for every convex combination $\sum \alpha_\nu P_\nu$. If we, instead, extend this by allowing ω-convex combinations, we obtain the notions we shall call *ω-convexity*, respectively *ω-concavity*. And f is said to be *strictly ω-convex* if f is ω-convex and if, provided $f(\sum \alpha_\nu P_\nu) < \infty$, equality only holds in $f(\sum \alpha_\nu P_\nu) \leq \sum \alpha_\nu f(P_\nu)$ if all the P_ν with $\alpha_\nu > 0$ are identical. Finally, f is *strictly ω-concave* if f is ω-concave and if, provided $\sum \alpha_\nu f(P_\nu) < \infty$, equality only holds in $f(\sum \alpha_\nu P_\nu) \geq \sum \alpha_\nu f(P_\nu)$ if all the P_ν with $\alpha_\nu > 0$ are identical.

It is an important feature of the convexity (concavity) properties which we shall establish, that the inequalities involved can be deduced from identities which must then be considered to be the more basic properties.

[2] "ω" often signals countability and is standard notation for the first infinite ordinal.

Theorem 6.1 (identities for mixtures). *Let $P_0 = \sum_1^\infty \alpha_\nu P_\nu$ be a mixture of distributions $P_\nu \in M_+^1(\mathbb{A})$; $\nu \geq 1$. Then*

$$H(\sum_{\nu=1}^{\infty} \alpha_\nu P_\nu) = \sum_{\nu=1}^{\infty} \alpha_\nu H(P_\nu) + \sum_{\nu=1}^{\infty} \alpha_\nu D(P_\nu \| P_0). \tag{6.26}$$

And, if a further distribution $Q \in M_+^1(\mathbb{A})$ is given,

$$\sum_{\nu=1}^{\infty} \alpha_\nu D(P_\nu \| Q) = D(\sum_{\nu=1}^{\infty} \alpha_\nu P_\nu \| Q) + \sum_{\nu=1}^{\infty} \alpha_\nu D(P_\nu \| P_0). \tag{6.27}$$

Proof. By the linking identity, the right hand side of (6.26) equals

$$\sum_{\nu=1}^{\infty} \alpha_\nu \langle \kappa_0, P_\nu \rangle$$

where κ_0 is the code adapted to P_0, and this may be rewritten as

$$\langle \kappa_0, \sum_{\nu=1}^{\infty} \alpha_\nu P_\nu \rangle \,,$$

i.e. as $\langle \kappa_0, P_0 \rangle$, which is nothing but the entropy of P_0. This proves (6.26).

Now add the term $\sum \alpha_\nu D(P_\nu \| Q)$ to each side of (6.26) and you get the following identity:

$$H(P_0) + \sum_{\nu=1}^{\infty} \alpha_\nu D(P_\nu \| Q) = \sum_{\nu=1}^{\infty} \alpha_\nu (H(P_\nu) + D(P_\nu \| Q)) + \sum_{\nu=1}^{\infty} \alpha_\nu D(P_\nu \| P_0).$$

Conclude from this, once more using the linking identity, that

$$H(P_0) + \sum_{\nu=1}^{\infty} \alpha_\nu D(P_\nu \| Q) = \sum_{\nu=1}^{\infty} \alpha_\nu \langle \kappa, P_\nu \rangle + \sum_{\nu=1}^{\infty} \alpha_\nu D(P_\nu \| P_0),$$

this time with κ adapted to Q. As $\sum \alpha_\nu \langle \kappa, P_\nu \rangle = \langle \kappa, P_0 \rangle = H(P_0) + D(P_0 \| Q)$, we then see upon subtracting the term $H(P_0)$, that (6.27) holds provided $H(P_0) < \infty$. The general validity of (6.27) is then deduced by a routine approximation argument, appealing to Theorem **??**. □

Theorem 6.2 (basic convexity/concavity properties). *The function $P \curvearrowright H(P)$ of $M_+^1(\mathbb{A})$ into $[0, \infty]$ is strictly ω-concave and, for any fixed $Q \in M_+^1(\mathbb{A})$, the function $P \curvearrowright D(P \| Q)$ of $M_+^1(\mathbb{A})$ into $[0, \infty]$ is strictly ω-convex.*

This follows from Theorem 6.1.

The second part of the result studies $D(P\|Q)$ as a function of the first argument P. It is natural also to look into divergence as a function of its second argument. To that end we introduce the *geometric mixture* of the probability distributions $Q_\nu, \nu \geq 1$ w.r.t. the weights $(\alpha_\nu)_{\nu\geq 1}$ (as usual, $\alpha_\nu \geq 0$ for $\nu \geq 1$ and $\sum \alpha_\nu = 1$). By definition, this is the incomplete probability distribution Q_0^g, notationally denoted $\sum^g \alpha_\nu Q_\nu$, which is defined by

$$Q_0^g(x) = \exp(\sum_{\nu=1}^{\infty} \alpha_\nu \log Q_\nu(x)) \quad x \in \mathbb{A}. \tag{6.28}$$

In other words, the point probabilities $Q_0^g(x)$ are the geometric avarages of the corresponding point probabilities $Q_\nu(x)$; $\nu \geq 1$ w.r.t. the weights α_ν; $\nu \geq 1$.

That Q_0^g is indeed an incomplete distribution follows from the standard inequality connecting geometric and arithmetic mean. According to that inequality, $Q_0^g \leq Q_0^a$, where Q_0^a denotes the usual arithmetic mixture:

$$Q_0^a = \sum_{\nu=1}^{\infty} \alpha_\nu Q_\nu.$$

To distinguish this distribution from Q_0^g, we may write it as $\sum^a \alpha_\nu Q_\nu$.

If we change the point of view by considering instead the adapted codes: $\kappa_\nu \leftrightarrow Q_\nu$, $\nu \geq 1$ and $\kappa_0 \leftrightarrow Q_0^g$, then, corresponding to (6.28), we find that

$$\kappa_0 = \sum_{\nu=1}^{\infty} \alpha_\nu \kappa_\nu,$$

which is the usual arithemetic average of the codes κ_ν. We can now prove:

Theorem 6.3 (2nd convexity identity for divergence). *Let P and $O_\nu; \nu \geq 1$be probability distributions over $\mathbb{A}$ and let $(\alpha_\nu)_{\nu\geq 1}$ be a sequence of weights. Then the identity*

$$\sum_{\nu=1}^{\infty} \alpha_\nu D(P\|Q_\nu) = D(P\|\sum_{\nu=1}^{\infty}{}^{g} \alpha_\nu P_\nu) \tag{6.29}$$

holds.

Proof. Assume first that $H(P) < \infty$. Then, from the linking identity, we get (using notation as above):

$$\begin{aligned}\sum_{\nu=1}^{\infty}\alpha_\nu D(P\|Q_\nu) &= \sum_{\nu=1}^{\infty}\alpha_\nu\left(\langle\kappa_\nu, P\rangle - H(P)\right)\\ &= \langle\sum_{\nu=1}^{\infty}\alpha_\nu\kappa_\nu, P\rangle - H(P)\\ &= \langle\kappa_0, P\rangle - H(P)\\ &= D(P\|Q_0^g),\end{aligned}$$

so that (6.29) holds in this case.

In order to establish the general validity of (6.29) we first approximate P with the conditional distributions$P|A$; $A \in \mathcal{P}_0(\mathbb{A})$ (which all have finite entropy). Recalling Theorem3.3, and using the result established in the first part of this proof, we get:

$$\begin{aligned}\sum_{\nu=1}^{\infty}\alpha_\nu D(P\|Q_\nu) &= \sum_{\nu=1}^{\infty}\alpha_\nu \lim_{A\in\mathcal{P}_0(\mathbb{A})} D(P|A\|Q_\nu)\\ &\le \lim_{A\in\mathcal{P}_0(\mathbb{A})}\sum_{\nu=1}^{\infty}\alpha_\nu D(P|A\|Q_\nu)\\ &= \lim_{A\in\mathcal{P}_0(\mathbb{A})} D(P|A\|Q_0^g)\\ &= D(P\|Q_0^g).\end{aligned}$$

This shows that the inequality "$\le$" in (6.29) holds, quite generally. But we can see more from the considerations above since the only inequality appearing (obtained by an application of Fatou's lemma, if you wish) can be replaced by equality in case $\alpha_\nu = 0$, eventually (so that the α's really determine a finite probability vector). This shows that (6.29) holds in case $\alpha_\nu = 0$, eventually.

For the final step of the proof, we introduce, for each n, the approximating finite probability vector $(\alpha_{n1}, \alpha_{n2}, \cdots, \alpha_{nn}, 0, 0, \cdots)$ with

$$\alpha_{n\nu} = \frac{\alpha_\nu}{\alpha_1 + \alpha_2 + \cdots + \alpha_n}; \quad \nu = 1, 2, \cdots, n.$$

Now put

$$Q_{n0}^g = \sum_{\nu=1}^{\infty}{}^{g}\,\alpha_{n\nu}Q_\nu = \sum_{\nu=1}^{n}{}^{g}\,\alpha_{n\nu}Q_\nu.$$

It is easy to see that $Q^g_{n0} \to Q^g_0$ as $n \to \infty$. By the results obtained so far and by lower semi-continuity of D we then have:

$$\begin{aligned} D(P\|Q^g_0) &= D(P\| \lim_{n\to\infty} Q^g_{n0}) \\ &\leq \lim_{n\to\infty} D(P\|Q^g_{n0}) \\ &= \lim_{n\to\infty} \sum_{\nu=1}^{n} \alpha_{n\nu} D(P\|Q_\nu) \\ &= \sum_{\nu=1}^{\infty} \alpha_\nu D(P\|Q_\nu), \end{aligned}$$

hereby establishing the missing inequality "$\geq$" in (6.29). □

Corollary 6.4. *For a distribution $P \in M^1_+(\mathbb{A})$ and any ω-convex combination $Q^a_0 = \sum_1^\infty \alpha_\nu Q_\nu$ of distributions $Q_\nu \in M^1_+(\mathbb{A})$; $\nu \geq 1$, the identity*

$$\sum_{\nu=1}^{\infty} \alpha_\nu D(P\|Q) = D(P\| \sum_{\nu=1}^{\infty} \alpha_\nu Q_\nu) + \sum_{x\in\mathbb{A}} P(x) \log \frac{Q^a_0(x)}{Q^g_0(x)}$$

holds with Q^g_0 denoting the geometric mexture $\sum^g \alpha_\nu Q_\nu$.

This is nothing but a convenient reformulation of Theorem6.3. By the usual inequality connecting geometric and arithemetic mean – and by the result concerning situations with equality in this inequality – we find as a further corollary that the following convexity result holds:

Corollary 6.5 (Convexity of D in the second argument). *For each fixed $P \in M^1_+(\mathbb{A})$, the function $Q \curvearrowright D(P\|Q)$ defined on $M^1_+(\mathbb{A})$ is strictly ω-convex.*

It lies nearby to investigate joint convexity of $D(\cdot\|\cdot)$ with both first and second argument varying.

Theorem 6.6 (joint convexity divergence). *$D(\cdot\|\cdot)$ is jointly ω-convex, i.e. for any sequence $(P_n)_{n\geq 1} \subseteq M^1_+(\mathbb{A})$, for any sequence $(Q_n)_{n\geq 1} \subseteq {}^\sim M^1_+(\mathbb{A})$ and any sequence $(\alpha_n)_{n\geq 1}$ og weights $(\alpha_n \geq 0\,;\, n \geq 1\,,\, \sum \alpha_n = 1)$, the following inequality holds*

$$D\left(\sum_{n=1}^{\infty} \alpha_n P_n \| \sum_{n=1}^{\infty} \alpha_n Q_n\right) \leq \sum_{n=1}^{\infty} \alpha_n D(P_n\|Q_n)\,. \tag{6.30}$$

In case the left–hand side in (6.30) *is finite, equality holds in* (6.30) *if and only if either there exists P and Q such that $P_n = P$ and $Q_n = Q$ for all n with $\alpha_n > 0$, or else, $P_n = Q_n$ for all n with $\alpha_n > 0$.*

Proof. We have

$$\begin{aligned}\sum_{n=1}^{\infty} \alpha_n D(P_n \| Q_n) &= \sum_{n=1}^{\infty} \sum_{i \in \mathbb{A}} \alpha_n p_{n,i} \log \frac{p_{n,i}}{q_{n,i}} \\ &= \sum_{i \in \mathbb{A}} \sum_{n=1}^{\infty} \alpha_n p_{n,i} \log \frac{p_{n,i}}{q_{n,i}} \\ &\geq \sum_{i \in \mathbb{A}} \left(\sum_{n=1}^{\infty} \alpha_n p_{n,i} \right) \log \frac{\sum_{n=1}^{\infty} \alpha_n p_{n,i}}{\sum_{n=1}^{\infty} \alpha_n q_{n,i}} \\ &= D\left(\sum_{n=1}^{\infty} \alpha_n P_n \middle\| \sum_{n=1}^{\infty} \alpha_n Q_n \right) .\end{aligned}$$

Here we used the well–known "log–sum inequality":

$$\sum x_\nu \log \frac{x_\nu}{y_\nu} \geq \left(\sum x_\nu \right) \log \frac{\sum x_\nu}{\sum y_\nu} .$$

As equality holds in this inequality if and only if (x_ν) and (y_ν) are proportional we see that, under the finiteness condition stated, equality holds in (6.30) if and only if, for each $i \in \mathbb{A}$ there exists a constant c_i such that $q_{n,i} = c_i p_{n,i}$ for all n. From this observation, the stated result can be deduced. □

7 The language of the probabilist

Previously, we expressed all definitions and results via probability distributions. Though these are certainly important in probability theory and statistics, it is often more suggestive to work with random variables or, more generally – for objects that do not assume real values – with random elements. Recall, that a *random element* is nothing but a measurable map defined on a probability space, say $X : \Omega \to S$ where $(\Omega, \mathcal{F}, P)$ is a probability space and $(S, \mathcal{S})$ a Borel space. As we shall work in the discrete setting, S will be a discrete set and we will then take $\mathcal{S} = \mathcal{P}(S)$ as the basic σ-algebra on S. Thus a *discrete random element* is a map $X : \Omega \to S$ where $\Omega = (\Omega, \mathcal{F}, P)$ is a probability space and S a discrete set. As we are accustomed to, there is often no need to mention explicitly the "underlying" probability measure. If misunderstanding is unlikely, "P" is used as the generic letter for "probability of". By P_X we denote the *distribution* of X:

$$P_X(E) = P(X \in E); \quad E \subseteq S.$$

If several random elements are considered at the same time, it is understood that the underlying probability space $(\Omega, \mathcal{F}, P)$ is the same for all random elements considered, whereas the discrete sets where the random elements take their values may, in general, vary.

The *entropy* of a random element X is defined to be the entropy of its distribution: $H(X) = H(P_X)$. The *conditional entropy* $H(X|B)$ *given an event* B with $P(B) > 0$ then readily makes sense as the entropy of the conditional distribution of X given B. If Y is another random element, the *joint entropy* $H(X,Y)$ also makes sense, simply as the entropy of the random element (X,Y) : $\omega \curvearrowright (X(\omega), Y(\omega))$. Another central and natural definition is the *conditional entropy* $H(X|Y)$ *of* X *given the random element* Y which is defined as

$$H(X|Y) = \sum_y P(Y = y) H(X|Y = y). \tag{7.31}$$

Here it is understood that summation extends over all possible values of Y. We see that $H(X|Y)$ can be interpretated as *the average entropy of* X *that remains after observation of* Y. Note that

$$H(X|Y) = H(X,Y|Y). \tag{7.32}$$

If X and X' are random elements which take values in the same set, $D(X\|X')$ is another notation for the divergence $D(P_X\|P_{X'})$ between the associated distributions.

Certain extreme situations may occur, e.g. if X and Y are independent random elements or, a possible scenario in the "opposite" extreme, if X is *a consequence of* Y, by which we mean that, for every y with $P(Y = y)$ positive, the conditional distribution of X given $Y = y$ is deterministic.

Often we try to economize with the notation without running a risk of misunderstanding, cf. Table 2 below

short notation	full notation or definition
$P(x),\ P(y),\ P(x,y)$	$P(X = x),\ P(Y = y),\ P((X,Y) = (x,y))$
$P(x\|y),\ P(y\|x)$	$P(X = x\|Y = y),\ P(Y = y\|X = x)$
$X\|y$	conditional distribution of X given $Y = y$
$Y\|x$	conditional distribution of Y given $X = x$

Table 2

For instance, (7.31) may be written

$$H(X|Y) = \sum_y P(y) H(X|y).$$

Let us collect some results formulated in the language of random elements which follow from results of the previous sections.

Theorem 7.1. *Consider discrete random elements $X, Y, \cdots$. The following results hold:*

(i) $0 \leq H(X) \leq \infty$ *and a necessary and sufficient condition that $H(X) = 0$ is, that X be deterministic.*

(ii) $H(X|Y) \geq 0$ *and a necessary and sufficient condition that $H(X|Y) = 0$ is, that X be a consequence of Y.*

(iii) *(Shannon's identity):*

$$H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y). \tag{7.33}$$

(iv) $H(X) \leq H(X, Y)$ *and, in case $H(X) < \infty$, a necessary and sufficient condition that $H(X) = H(X, Y)$ is, that Y be a consequence of X.*

(v)

$$H(X) = H(X|Y) + \sum_y P(y) \cdot D(X|y \| X). \tag{7.34}$$

(vi) *(Shannon's inequality):*

$$H(X|Y) \leq H(X) \tag{7.35}$$

and, in case $H(X|Y) < \infty$, a necessary and sufficient condition that $H(X|Y) = H(X)$ is, that X and Y be independent.

Proof. (i) is trivial.

(ii): This inequality, as well as the discussion of equality, follows by (i) and the defining relation (7.31).

(iii) is really equivalent to (4.16), but also follows from a simple direct calculation which we shall leave to the reader.

(iv) follows from (ii) and (iii).

(v): Clearly, the distribution of X is nothing but the mixture of the conditional distributions of X given $Y = y$ w.r.t. the weights $P(y)$. Having noted this, the identity follows directly from the identity for mixtures, (6.26) of Theorem 6.1.

(vi) follows directly from (v) (since independence of X and Y is equivalent to coincidence of all conditional distributions of X given $Y = y$ whenever $P(y)$ is positive). □

8 Mutual information

The important concept of mutual information is best introduced using the language of random elements. The question we ask ourselves is this one: Let X and Y be random elements. How much information about X is revealed by Y? In more suggestive terms: We are interested in the value of X but cannot observe this value directly. However, we do have access to an auxillary random element Y which can be observed directly. How much information about X is contained in an observation of Y?

Let us suggest two different approaches to a sensible aswer. Firstly, we suggest the following general principle:

$$\boxed{\text{Information gained} = \text{decrease in uncertainty}}$$

There are two states involved in an observation: START where nothing has been observed and END where Y has been observed. In the start position, the uncertainty about X is measured by the entropy $H(X)$. In the end position, the uncertainty about X is measured by the conditional entropy $H(X|Y)$. The decrease in uncertainty is thus measured by the difference

$$H(X) - H(X|Y) \tag{8.36}$$

which then, according to our general principle, is the sought quantity "information about X contained in Y".

Now, let us take another viewpoint and suggest the following general principle, closer to coding:

$$\boxed{\text{Information gained} = \text{saving in coding effort}}$$

When we receive the information that Y has the value, say y, this enables us to use a more efficient code than was available to us from the outset since we will then know that the distribution of X should be replaced by the conditional distribution of X given y. The saving realized we measure by the divergence $D(X|y\|X)$. We then find it natural to measure the overall saving in coding effort by the average of this divergence, hence now we suggest to use the quantity

$$\sum_y P(y) D(X|y\|X) \tag{8.37}$$

as the sought measure for the "information about X contained in Y".

Luckily, the two approaches lead to the same quantity, at least when $H(X|Y)$ is finite. This follows by (7.34) of Theorem 7.1. From the same theorem, now appealing to (7.33), we discover that, at least when $H(X|Y)$ and $H(Y|X)$ are finite, then "the information about X contained in Y" is the same as "the information about Y contained in X". Because of this symmetry, we choose to use a more "symmetric" terminology rather than the directional "information about X contained in Y". Finally, we declare a preference for the "saving in coding effort-definition" simply because it is quite general as opposed to the "decrease in uncertainty-definition" which leads to (8.36) that could result in the indeterminate form $\infty - \infty$.

With the above discussion in mind we are now prepared to define $I(X \wedge Y)$, the *mutual information of X and Y* by

$$I(X \wedge Y) = \sum_y P(y) D(X|y \| X). \tag{8.38}$$

As we saw above, mutual information is symmetric in case $H(X|Y)$ and $H(Y|X)$ are finite. However, symmetry holds in general as we shall now see. Let us collect these and other basic results in one theorem:

Theorem 8.1. *Let X and Y be discrete random elements with distributions P_X, respectively P_Y and let $P_{X,Y}$ denote the joint distribution of (X, Y). Then the following holds:*

$$I(X, Y) = D(P_{X,Y} \| P_X \otimes P_Y), \tag{8.39}$$

$$I(X \wedge Y) = I(Y \wedge X), \tag{8.40}$$

$$H(X) = H(X|Y) + I(X \wedge Y), \tag{8.41}$$

$$H(X, Y) + I(X \wedge Y) = H(X) + H(Y), \tag{8.42}$$

$$I(X \wedge Y) = 0 \Leftrightarrow X \text{ and } Y \text{ are independent}, \tag{8.43}$$

$$I(X \wedge Y) \leq H(X), \tag{8.44}$$

$$I(X \wedge Y) = H(X) \Leftrightarrow I(X \wedge Y) = \infty \vee X \text{ is a consequence of } Y. \tag{8.45}$$

Proof. We find that

$$
\begin{aligned}
I(X \wedge Y) &= \sum_y P(y) \sum_x P(x|y) \log \frac{P(x|y)}{P(x)} \\
&= \sum_{x,y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \\
&= D(P_{X,Y} \| P_X \otimes P_Y)
\end{aligned}
$$

and (8.39) follows. (8.40) and (8.43) are easy corollaries – (8.43) also follows directly from the defining relation (8.38).

Really, (8.41) is the identity (6.26) in disguise. The identity (8.42) follows by (8.41) and Shannon's identity (7.33). The two remaining results, (8.44) and (8.45) follow by (8.41) and (ii) of Theorem (7.1). □

It is instructive to derive the somewhat surprising identity (8.40) directly from the more natural datareduction identity (4.17). To this end, let $X : \Omega \to \mathbb{A}$ and $Y : \Omega \to \mathbb{B}$ be the random variables concerned, denote their distributions by P_1, respectively P_2, and let P_{12} denote their joint distribution. Further, let $\pi : \mathbb{A} \times \mathbb{B} \to \mathbb{A}$ be the natural projection. By (4.17) it then follows that

$$
\begin{aligned}
D(P_{12} \| P_1 \otimes P_2) &= D(\partial_\pi P_{12} \| \partial_\pi (P_1 \otimes P_2)) + D^\pi(P_{12} \| P_1 \otimes P_2) \\
&= D(P_1 \| P_1) + \sum_{x \in \mathbb{A}} P_{12}(\pi^{-1}(x)) D(P_{12}|\pi^{-1}(x) \| P_1 \otimes P_2|\pi^{-1}(x)) \\
&= 0 + \sum_{x \in \mathbb{A}} P_1(x) D(Y|x \| Y) \\
&= I(Y \wedge X)\,.
\end{aligned}
$$

By symmetry, e.g. by considering the natural projection $\mathbb{A} \times \mathbb{B} \to \mathbb{B}$ instead, we then see that $I(X \wedge Y) = I(Y \wedge X)$.

9 Information Transmission

An important aspect of many branches of mathematics is that to a smaller or larger extent one is free to choose/design/optimize the system under study. For key problems of information theory this freedom lies in the choice of a distribution or, equivalently, a code. In this and in the next section we look at two models which are typical for optimization problems of information theory. A detailed study has to wait until later chapters. For now we only introduce some basic concepts and develop their most fundamental properties.

Our first object of study is a very simple model of a communication system given in terms of a *discrete Markov kernel* $(\mathbb{A}, \mathbb{P}, \mathbb{B})$. This is a triple with $\mathbb{A}$ and $\mathbb{B}$ discrete sets, referred to respectively, as the *input alphabet* and the *output alphabet.* And the *kernel* $\mathbb{P}$ itself is a map $(x,y) \curvearrowright P(y|x)$ of $\mathbb{A}\times\mathbb{B}$ into $[0,1]$ such that, for each fixed $x \in \mathbb{A}$, $y \curvearrowright \mathbb{P}(y|x)$ defines a probability distribution over $\mathbb{B}$. In suggestive terms, if the letter x is "sent", then $\mathbb{P}(\cdot|x)$ is the (conditional) distribution of the letter "received". For this distribution we also use the notation $\mathbb{P}_x$. A distribution $P \in M_+^1(\mathbb{A})$ is also referred to as an *input distribution.* It is this distribution which we imagine we have a certain freedom to choose.

An input distribution P *induces* the *output distribution* Q defined as the mixture

$$Q = \sum_x P(x)\mathbb{P}_x.$$

Here, and below, we continue to use simple notation with x as the generic element in $\mathbb{A}$ and with y the generic element in $\mathbb{B}$.

When P induces Q, we also express this notationally by $P \rightsquigarrow Q$.

If P is the actual input distribution, this defines in the usual manner a random element (X,Y) taking values in $\mathbb{A}\times\mathbb{B}$ and with a distribution determined by the point probabilities $P(x)\mathbb{P}(y|x)$; $(x,y) \in \mathbb{A}\times\mathbb{B}$. In this way, X takes values in $\mathbb{A}$ and its distribution is P, whereas Y takes values in $\mathbb{B}$ and has the induced distribution Q.

A key quantity to consider is the *information transmission rate,* defined to be the mutual information $I(X \wedge Y)$, thought of as the information about the sent letter (X) contained in the received letter (Y). As the freedom to choose the input distribution is essential, we leave the language expressed in terms of random elements and focus explicitly on the distributions involved. Therefore, in more detail, the *information transmission rate with P as input distribution* is denoted $I(P)$ and defined by

$$I(P) = \sum_x P(x)D(\mathbb{P}_x\|Q) \tag{9.46}$$

where $P \rightsquigarrow Q$. In taking this expression as the defining quantity, we have already made use of the fact that mutual information is symmetric (cf. (8.40)). In fact, the immediate interpretation of (9.46) really concerns information about what was received given information about what was sent rather than the other way round.

The first result we want to point out is a trivial translation of the identity (6.27) of Theorem 6.1:

Theorem 9.1 (The compensation identity). *Let P be an input distribution and Q the induced output distribution. Then, for any output distribution Q^*,*

$$\sum_x P(x)D(\mathbb{P}_x\|Q^*) = I(P) + D(Q\|Q^*). \tag{9.47}$$

The reason for the name given to this identity is the following: Assume that we commit an error when computing $I(P)$ by using the distribution Q^* in (9.46) instead of the induced distribution Q. We still get $I(P)$ as result if only we "compensate" for the error by adding the term $D(Q\|Q^*)$.

The identity holds in a number of cases, e.g. with D replaced by squared Euclidean distance, with socalled Bregman divergencies or with divergencies as they are defined in the quantum setting (then the identity is known as "Donald's identity"). Possibly, the first instance of the identity appeared in [9]

In the next result we investigate the behaviour of the information transmission rate under mixtures.

Theorem 9.2 (information transmission under mixtures). *Let $(P_\nu)_{\nu\geq 1}$ be a sequence of input distributions and denote by $(Q_\nu)_{\nu\geq 1}$ the corresponding sequence of induced output distributions. Furthermore, consider an ω-convex combination $P_0 = \sum_1^\infty s_\nu P_\nu$ of the given input distributions and let Q_0 be the induced output distribution: $P_0 \rightsquigarrow Q_0$. Then:*

$$I\left(\sum_{\nu=1}^{\infty} s_\nu P_\nu\right) = \sum_{\nu=1}^{\infty} s_\nu I(P_\nu) + \sum_{\nu=1}^{\infty} s_\nu D(Q_\nu\|Q_0). \tag{9.48}$$

Proof. By (9.47) we find that, for each $\nu \geq 1$,

$$\sum_x P_\nu(x) D(\mathbb{P}_x\|Q_0) = I(P_\nu) + D(Q_\nu\|Q_0).$$

Taking the proper mixture, we then get

$$\sum_\nu s_\nu \sum_x P_\nu(x) D(\mathbb{P}_x\|Q_0) = \sum_\nu s_\nu I(P_\nu) + \sum_\nu s_\nu D(Q_\nu\|Q_0).$$

As the left hand side here can be written as

$$\sum_x P_0(x) D(\mathbb{P}_x\|Q_0),$$

which equals $I(P_0)$, (9.48) follows. □

Corollary 9.3 (concavity of information transmission rate). *The information transmission rate $P \curvearrowright I(P)$ is a concave function on $M_+^1(\mathbb{A})$. For a given ω-convex mixture $P_0 = \sum s_\nu P_\nu$ for which $\sum s_\nu I(P_\nu)$ is finite, equality holds in the inequality*

Acknowledgements

Most results and proofs have been discussed with Peter Harremoës. This concerns also the findings involving the information topology. More complete coverage of these aspects will be published by Peter Harremoës, cf. [6].

References

[1] P. Antosik, "On a topology of convergence," *Colloq. Math.*, vol. 21, pp. 205–209, 1970.

[2] T.M. Cover and J.A. Thomas, *Information Theory*, New York: Wiley, 1991.

[3] I. Csiszár, "I-divergence geometry of probability distributions and minimization problems," *Ann. Probab.*, vol. 3, pp. 146–158, 1975.

[4] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems.* New York: Academic Press, 1981.

[5] R.G. Gallager, *Information Theory and reliable Communication.* New York: Wiley, 1968.

[6] P. Harremoës, "The Information Topology," in preparation

[7] J. Kisyǹski, "Convergence du type L," *Colloq. Math.*, vol. 7, pp. 205–211, 1959/1960.

[8] A. Perez, "Notions généralisées d´incertitude, d´entropie et d´information du point de vue de la théorie de martingales," *Transactions of the first Prague conference on information theory, statistical decision functions and random processes*, pp. 183–208. Prague: Publishing House of the Czechoslovak Academy of Sciences, 1957.

[9] F. Topsøe, "An information theoretical identity and a problem involving capacity," *Studia Sci. Math. Hungar.*, vol. 2, pp.291–292, 1967.