Codeword Distinguishability in Minimum Diversity Decoding

Andrea Sgarro and Luca Bortolussi

***Abstract*— We re-take a coding-theoretic notion which goes back to Cl. Shannon: codeword distinguishability. This notion is standard in zero-error information theory, but its bearing is definitely wider and it may help to better understand new forms of coding, as we argue below. In our approach, the underlying decoding principle is very simple and very general: one decodes by trying to minimise the diversity (in the simplest case the Hamming distance) between a codeword and the output sequence observed at the end of the noisy transmission channel. Symmetrically and equivalently, minimum-diversity decoders and codeword distinguishabilities may be replaced by maximum-similarity decoders and codeword confusabilities. The operational meaning of codeword distinguishability is made clear by a reliability criterion, which generalises the well-known criterion on minimum Hamming distances for error-correction codes. We investigate the formal properties of distinguishabilities versus diversities; these two notions are deeply related, and yet essentially different. An encoding theorem is put forward, which supports and suggests old and new code constructions. In a list of case studies, we examine channels with crossovers and erasures, or with crossovers, deletions and insertions, a channel of cryptographic interest, and the case of a few "odd distances" taken from DNA word design.**

***Index Terms*— : codeword distinguishability, codeword confusability, minimum diversity decoding, maximum similarity decoding, zero-error information theory, erasure channels, edit distance, simple substitution ciphers, DNA string distances.**

I. Introduction

This paper is basically a re-working of a coding-theoretic notion which goes back to Cl. Shannon [1]: *codeword distinguishability*, or, symmetrically and equivalently, *codeword confusability*. This notion is used in *zero-error information theory*, but, as we argue below, its bearing is definitely wider and might prove valuable to better understand new forms of coding, as is DNA word design; cf. below Section V (for an excellent overview of zero-error information theory cf. [2]; cf. [3], [4] for overviews of DNA word design in molecular computation). In our approach, the underlying decoding principle is very simple and very general: one decodes by trying to minimise the *diversity* between codewords and the output word observed at the end of the noisy transmission channel, or, symmetrically, by maximising the corresponding *similarity*; as for diversities or similarities between input codewords and channel outputs, one does not need to make any assumption about their nature, save the fact that they must be *comparable*. The idea is to adopt *a channel model* which is *as uncommittal as possible*: the fact that the number of assumptions one has to make is low allows one to give a compact view of a large number of situations, as those examined in Sections III to V. If one wants to correct "small" diversities, the parameter to check is precisely codeword distinguishability, as made clear by the *reliability criterion* in Section II; symmetrically, one has to check codeword confusability to ensure that "large" similarities are corrected. It is the reliability criterion which gives an *operational meaning* to the notion of codeword distinguishability, or symmetrically codeword confusability, by relating it to the key notion of the *error correction capability* of a code. Below we shall concentrate on diversities and distinguishabilities, rather than similarities and confusabilities, only because they compare better to the notion of distance, and in particular of Hamming distance; our choice does not bring about any loss in generality, cf. Remark 2, Section II.

Codeword distinguishability is expressed as a boolean function of diversities between inputs and outputs; its general expression (1) is in terms of an optimisation problem over the output space; we stress that distinguishability, unlike diversity, is a *global* notion, in the sense that it involves the *entire* output space. In some lucky cases the optimisation problem can be explicitly solved and, correspondingly, the formula for distinguishability is drastically simplified; this is what happens in the case of "usual codes" (e.g. algebraic codes), those based on Hamming distances, henceforth called for simplicity *Hamming-distance codes*. In general, even if the notions of diversity and distinguishability are tightly related, they are *not* interchangeable, and in some unruly cases, as those met in Shannon's zero-error coding, or in the example on a six-element space discussed at length in Sections III and IV, the difference can be dramatic.

In Section III, we deal with the case when the input space and the output space coincide: this allows comparing directly the notion of diversity (distortion, metric distance) with that of distinguishability; we discuss the formal properties of distinguishabilities as opposed to diversities, and put forward upper and lower bounds. In particular, the (lower) *metric bound* turns out to be a convenient tool to solve the minimisation problem in (1), and obtain explicit forms for the distinguishability function; cf. the case studies of Section V. In Section IV we discuss optimal code constructions with respect to the very general reliability criterion of Section II, and exhibit a whole class of situations when forgetting about distinguishabilities in favour of diversities, as currently done in algebraic coding and more generally in minimum Hamming-distance coding, is an admissible policy, because the two notions practically collapse

Partly supported by GNCS, INdAM.

A. Sgarro is with the Department of Mathematics and Computer Science, University of Trieste, 34100 Trieste, Italy. E-mail: sgarro@units.it

L. Bortolussi is with the Department of Mathematics and Computer Science, University of Udine, 33100 Udine, Italy. E-mail: luca.bortolussi@dimi.uniud.it

into one.

After re-taking shortly zero-error codes and usual Hamming-distance codes, some case studies are covered in Section V: we compute the distinguishability function for channels with crossovers and erasures (henceforth called simply erasure channels), channels with crossovers, deletions and insertions (unlike erasures, deletions are not re-traceable, and so the decoder does not know where they have occurred), for a channel of cryptographic interest, and for four channels based on "odd DNA distances". The explicit computation of the distinguishability functions shows that only *four* types of combinatorial code constructions are actually needed; in particular, our arguments support the use of Hamming-distance codes on channels with crossovers and erasures, but *not* on channels with crossovers, deletions and insertions.

When a code construction is already available, our approach can be used to devise *channel models* and *decoding procedures* which solve a sort of *inverse problem*, in the sense that they "explain" a posteriori that construction. Something similar is found in most textbooks on algebraic coding, where one "explains" the classical code constructions where one checks the Hamming distance between codewords by introducing a symmetric memoryless channel with a small crossover probability; note that in Section V we are able to "explain" these very same Hamming-distance constructions in alternative ways. In ongoing work, we are trying to apply systematically this point of view to DNA word design, e.g. to code constructions as those given in [5].

In the paper, facts which we think the reader may skip at a quick reading are relegated to remarks and appendices. These include a comparison between decoding by minimum diversity or by maximum similarity (the two approaches are fully interchangeable; cf. Remark 2, Section II), and a comparison between decoding by maximum similarity or by maximum likelihood (the latter approach is definitely more restrictive; cf. Remark 3, Section II, and Appendix B). Save for quick mentions, below we shall not cover two important subjects of channel transmission: error *detection* rather than error correction (cf. however Remark 5, Section II, and Remark 9, Section IV), and the *asymptotic* point of view which is typical of Shannon theory (cf. however Remark 10, Section IV).

II. Decoding by minimum diversity

Consider the following situation: a list of primary objects is given, one of them is selected, but in its place a secondary object is observed. Now, to each couple made up of a primary object a and a secondary object z, a *diversity* measure $d(a,z)$ is associated which belongs to a set $\mathcal{D}$; this set, which is usually made up of non-negative integers, is bound to be *totally ordered* and so diversities can be compared. Assume that, in a very broad sense of the word "likely", the situation is such that smaller diversities between the selected primary object a and the observed secondary object z are more likely to occur. Then a rational behaviour is to decide for a primary object in the list which minimises its diversity from the observed object. This situation is typical of channel coding, and from channel coding we shall borrow our terms, to be now introduced.

Let an *input space* $\mathcal{A}$ and an *output space* $\mathcal{B}$ be given; further, let a diversity measure d be assigned, $d : \mathcal{A} \times \mathcal{B} \to \mathcal{D}$. The three sets $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{D}$ are assumed to be finite, and so we shall take freely maxima and minima (a generalization to the countable case would be straightforward). Whatever their nature, the elements of $\mathcal{A}$ will be called *input words*, while the elements of $\mathcal{B}$ will be called *output words*. The following definition introduces our main notion:

Definition 1: If a and b are two input words their *distinguishability* is defined as

$$\delta(a,b) \;=\; \min_{z\in\mathcal{B}} \big[d(a,z) \vee d(b,z)\big] \;\in\; \mathcal{D} \tag{1}$$

(In our formulas the symbols $\wedge$ and $\vee$ stand for minima and maxima, respectively, and are used only because of graphic clarity). We stress that, unlike for the diversity d, *both* the arguments of the distinguishability $\delta(a,b)$ belong to the *input* space. Whatever the diversity d, the distinguishability δ is symmetric:

$$\delta(a,b) \;=\; \delta(b,a)$$

In channel coding an input word is selected and is sent through a noisy medium called the *transmission channel*; an output word is observed at the other end of the channel, which is a noisy version of the input transmitted. Let us choose a *codebook*, or simply a *code*, $\mathcal{C} \subset \mathcal{A}$, which is the primary list referred to at the beginning of this section; the words c in $\mathcal{C}$ are called *codewords*. The diversity d tells the degree of "distortion" which has been brought about by noise during channel transmission; lighter diversities are more likely, and so we assume *minimum diversity decoding*: if the output word y has been observed, one has to decode to a codeword c which minimises the diversity measure $d(c,y)$, $c \in \mathcal{C}$ (our decoders are *deterministic* mappings from the output space to the codebook $\mathcal{C}$; cf. however Remark 5). A relevant notion to describe the performance of a code is the following:

Definition 2: Given a codebook $\mathcal{C}$, its *minimum distinguishability* is defined as:

$$\delta_m \;=\; \delta_m(\mathcal{C}) \;=\; \min_{c,\,c'\in\mathcal{C},\,c\neq c'} \delta(c,c') \tag{2}$$

The following reliability criterion is soon proven, and actually it is quite akin to the one used by coding theorists in the special case when the diversity measure is Hamming distance; cf. Section V. It gives the operational meaning of the minimum distinguishability for code $\mathcal{C}$: in a good code distinct codewords have high distinguishability, and in this sense they are "wide apart" from each other. We stress that the very general reliability criterion below is formulated in terms of both diversities and distinguishabilities, the latter through δ_m; we also stress that being "wide apart" refers to distinguishability, and *not* to diversity.

Theorem 1 (reliability criterion for code $\mathcal{C}$): Whatever the codeword which has been sent through the channel, all diversities $\leq \tau$ are corrected if and only if $\delta_m > \tau$. Equivalently: all diversities $< \tau$ are corrected iff $\delta_m \geq \tau$.

Corollary 1: When the diversities are consecutive integers, all diversities $\leq \tau - 1$ are corrected iff $\delta_m \geq \tau$.

Proof: assume $\delta_m \leq \tau$, and say δ_m is obtained at the "triangle" c, c', z, with $c, c' \in \mathcal{C}$, $z \in \mathcal{B}$: if $d(c,z) \neq d(c',z)$, e.g. $d(c,z) < d(c',z) = \delta_m \leq \tau$, an error occurs whenever c'

is sent and z is received: now, the corresponding diversity is $\leq \tau$; if instead $d(c,z) = d(c',z)$, the decoder cannot ensure that both c and c' are properly recovered whenever one of the two is sent and z is received. Conversely, assume a decoding error occurs with $d(c,z) \leq \tau$; then a distinct codeword c' must exist with $d(c',z) \leq d(c,z)$; but then $\delta_m \leq \delta(c,c') \leq \tau$. The obvious corollary is quoted explicitly only because it covers the situation usually met in our examples. ■

Due to the reliability criterion, one may equivalently define the minimum distinguishability as follows:

Assume one decodes by minimum diversity: then the minimum distinguishability of a code $\mathcal{C}$ is the smallest diversity which may bring about decoding errors.

A warning to avoid misunderstandings: in the following, to get rid of trivial specifications, we shall always "expurgate" our ordered set $\mathcal{D}$ of all values s which are useless, in the sense that one never has $d(a,z) = s$. Actually, when $\mathcal{D}$ is numeric, one often writes a bound like $d(a,z) \geq s$, even if s is unattainable: below such a bound will be replaced by $d(a,z) \geq \lceil s \rceil_{\mathcal{D}}$, where $\lceil s \rceil_{\mathcal{D}}$ is the smallest value in $\mathcal{D}$ which is not smaller than s, i.e. the usual integer ceiling when $\mathcal{D}$ is made up of consecutive integers.

Remark 1 (order-preserving transformations): If $f : \mathcal{D} \to \mathcal{T}$ is an order isomorphism between the two totally ordered sets $\mathcal{D}$ and $\mathcal{T}$, i.e. $d < d'$ in $\mathcal{D}$ iff $f(d) < f(d')$ in $\mathcal{T}$, nothing special changes if one goes from $\mathcal{D}$ to $\mathcal{T}$: $\delta(a,b)$ becomes $f\big(\delta(a,b)\big)$, and the reliability criterion holds with $f(\tau)$ in lieu of τ. In this sense, the "nature" of $\mathcal{D}$ is irrelevant, provided the order relation is retained. The same holds true for similarities, as dealt with in the next remark.

Remark 2 (decoding by maximum similarity): If one has an ordered set $\mathcal{S}$ of *similarities*, rather than an ordered set $\mathcal{D}$ of diversities, the natural policy is decoding by *maximum similarity* rather than minimum diversity, i.e. one maximises the similarity $s(c,z)$. Nothing really new happens: it will be enough to replace $\mathcal{S}$ by $\mathcal{D}$, which is obtained from $\mathcal{S}$ by simply inverting the order, $d < d'$ in $\mathcal{D}$ iff $d > d'$ in $\mathcal{S}$. If one decodes by maximum similarity, distinguishabilities are replaced by *confusabilities* $\gamma(a,b) = \max_z \big[s(a,z) \wedge s(b,z)\big]$; the value of the distinguishability $\delta(a,b) \in \mathcal{D}$ is the same as the value of the confusability $\gamma(a,b) \in \mathcal{S}$. One defines the *maximum confusability* of the code $\gamma_M = \gamma_M(\mathcal{C})$ by taking the maximum confusability between distinct codewords, cf. (2); as soon checked, $\gamma_M = \delta_m$. In the reliability criterion, one may refer indifferently[1] to the correction of all the similarities $\geq \tau \in \mathcal{S}$ iff $\gamma_M < \tau$, or of all the diversities $\leq \tau \in \mathcal{D}$ iff $\delta_m > \tau$.

Remark 3 (decoding by maximum likelihood): Decoding by *maximum likelihood* may be seen as a special case of maximum similarity decoding. However, since only the ordering counts (cf. Remark 1), the following question is relevant: given a similarity matrix $s(a,z) \in \mathcal{S}$, can one replace it by a stochastic matrix $\psi(z|a) \in [0,1]$ without changing the ordering? (By so saying we mean that the order relation between $s(a,z)$ and $s(b,v)$ in $\mathcal{S}$ must always be the same as the order relation between the channel transition probabilities $\psi(z|a)$ and $\psi(v|b)$). This is well-known to be true in the Hamming case, cf. Section V, but it is *not* the case in general; the whole point will be deepened in Appendix B, devoted to *stochastic-like*[2] similarities.

Remark 4 (what is a decoding error?): So far, we have implicitly assumed that an error occurs iff codeword c is sent over the channel, while codeword $c' \neq c$ is decoded to. So, the *error set* $\mathcal{E}$ is made up of all the couples of distinct input words. In some situations a more flexible approach may be needed (cf. Remark 7, Section III, and the case of substitution ciphers in Section V), and the error set might be *any* subset $\mathcal{E}$ of unordered couples of distinct input words. Everything works, after setting:

$$\delta_m \;=\; \delta_m(\mathcal{C}) \;=\; \min_{c,c' \in \mathcal{C},\, (c,c') \in \mathcal{E}} \delta(c,c')$$

which is a more flexible definition than the one in (2). The statement in the reliability criteria "all diversities $\leq \tau$ are corrected" should be now understood as follows: whenever c is sent and z is received with $d(c,z) \leq \tau$, then z is decoded to c' such that $(c,c') \notin \mathcal{E}$.

Remark 5 (breaking ties): Our decoders are all deterministic, and so ties $d(c,z) = d(c',z)$ are always broken in a fixed way. Non-deterministic decoders, such as to randomly select the decoded codeword out of those which minimise the diversity to the received output word, would not make any change with respect to out reliability criterion, which requires that diversities $\leq \tau$ should be *always* corrected. We stress that whenever there is a tie, the corresponding diversity d is one of those whose correction is *not* ensured, and so its value is greater or equal to the minimum distinguishability of the code, $d \geq \delta_m$. Below we shall not cover the case of *detected errors*, when the decoder may refuse decoding, save for the fleeting mention in Remark 9, Section IV.

III. Distortions and distances versus distinguishabilities

In this section, to better compare the two notions of diversity and distinguishability, we assume that the input and the output spaces coincide, and that diversities are non-negative numbers; in symbols: $\mathcal{A} = \mathcal{B}$, $\mathcal{D} \subset \mathbf{R}^+$. Below we shall bound the distinguishability δ in terms of the diversity d, and we shall discuss the friendly cases when the lower or the upper bound, respectively, are met with equality. We shall put forward un unfriendly example where the distinguishability oscillates between the two bounds; it will be used in this section to provide a counterexample, and in Section IV to better contrast distinguishability and diversity when one wants to construct optimal codes.

[1] Actually, if $\mathcal{S}$ is made up of non-negative numbers as is usually the case, just inverting the order is definitely misleading. Recalling remark 1, a more user-friendly policy would be to go from the similarity $s(a,z)$ to the diversity $d(a,z) = s_M - s(a,z)$, s_M being the largest similarity in store; by so doing, the natural order between numbers is kept. With this choice, distinguishability and confusability sum to s_M; in the reliability criterion, one may refer indifferently to the correction of all the diversities $\leq \tau$ iff $\delta_m > \tau$, or of all the similarities $\geq s_M - \tau$ iff $\gamma_m < s_M - \tau$.

[2] Be the similarity matrix stochastic-like or not, we stress once more that the term likely, as used in the main body of this section, should not be taken as a technical term of probability theory such as maximum *likeli*hood, but rather as a "loose" term as those used in natural languages. Cf. [6] for an approach to coding based on multi-valued logic, rather than probability theory.

Definition 3: A diversity measure is a *(symmetric) distortion* when $d(a,b) = d(b,a) \geq d(a,a) = 0$ for all words a and b, a distortion is a *pseudometric distance*, or simply a pseudometric, when the triangle inequality $d(a,b) \leq d(a,z) + d(z,b)$ is always verified, and a pseudometric is a *metric distance*, or simply a metric, when further $d(a,b) = 0$ implies $a = b$.

For any distortion d, one has the *upper* bound:

$$\delta(a,b) \ \leq \ d(a,b) \tag{3}$$

To see this, just take in (1) a test z which is equal to a, or to b. Further, if d satisfies the triangle inequality, one has the lower *metric bound*:

$$\delta(a,b) \ \geq \ \frac{1}{2}\, d(a,b) \ \ , \quad d \ \text{triangular}$$

which can be strengthened to

$$\delta(a,b) \ \geq \ \left\lceil \frac{1}{2}\, d(a,b) \right\rceil_{\mathcal{D}} \ \ , \quad d \ \text{triangular} \tag{4}$$

Recall that the $\mathcal{D}$-ceiling has been introduced in Section II before the remarks; Section V contains a rather more general form of the two bounds. To prove the metric bound just observe that $d(a,z) \vee d(b,z)$ as in (1) is $\geq \frac{1}{2}\big(d(a,z) + d(b,z)\big)$. The lower bound in (4) is achieved with equality by Hamming distance (in this case the generalised ceiling is the usual integer ceiling). Cf. Section V, where the metric bound turns out to be a convenient short way to solve for the minimum in (1), and which contain other examples when the lower bound is met with equality. As for the upper bound (3), we shall now characterise the distortions (actually, the pseudometrics) which achieve it with equality, and, by so doing, fully trivialise the distinction between distinguishabilities and diversities.

The limit case of coarseness indices: We begin by putting forward a simple lemma. By saying that three (not necessarily distinct) words a, b, and c form a *thin-isosceles triangle* we mean that two of the "side-lengths" $d(a,b)$, $d(b,c)$, $d(c,a)$ are equal, and that the third side-length is the smallest of the three (the triangle might also be equilateral). To see why the lemma below works, first observe that in a thin-isosceles triangle the triangle inequality always holds; conversely, just observe that, whenever the triangle a, b, c is not thin-isosceles and $d(a,c)$, say, is the largest side, one has $\delta(a,c) \leq d(a,b) \vee d(b,c) < d(a,c)$, and so $\delta(a,c) \neq d(a,c)$.

Lemma 1: The upper bound $\delta \leq d$ is achieved with equality, $\delta = d$, iff all triangles are thin-isosceles.

Recall that an equivalence relation $\mathcal{R}$ is *coarser* than another equivalence relation $\mathcal{R}'$, $\mathcal{R} \vdash \mathcal{R}'$, when it "joins" equivalence classes of the latter, i.e. when $a\mathcal{R}'b$ implies $a\mathcal{R}b$; one also says that $\mathcal{R}'$ *refines* $\mathcal{R}$, $\mathcal{R}' \dashv \mathcal{R}$. We shall consider a chain of (distinct) $s+1$ equivalences, which are coarser and coarser:

$$\mathcal{R}_0 \ \dashv \mathcal{R}_{r_1} \ \ldots \ \dashv \mathcal{R}_{r_s}$$

and which are indexed to real numbers $r_0 = 0 < r_1 < \ldots < r_s$; $\mathcal{R}_{r_s}$ is bound to be the "very coarse" (and trivial) relation under which every two words are equivalent. Define the *coarseness index* $d(a,b)$ by setting it equal to the smallest relation index r_i such that a and b are still equivalent; a large coarseness index corresponds to words which are equivalent only under coarse relations.

As an example take the rests modulo 8, $\mathcal{A} = \{0,1,2,3,4,5,6,7\}$; consider the equivalences $\mathcal{R}_i$, $i = 0,1,2,3$, with $a\mathcal{R}_i b$ iff a and b are congruent modulo 2^{3-i}, i.e. modulo $8, 4, 2, 1$, respectively: the equivalence classes to which the rest 0 belongs are $\{0\}$, $\{0,4\}$, $\{0,2,4,6\}$, $\{0,1,2,3,4,5,6,7\}$, respectively. Consider the triangle $0, 4, 7$: one has $d(0,4) = 1$, $d(0,7) = d(4,7) = 3$.

Any coarseness index d is a pseudometric which verifies the thin-isosceles property, and so $d = \delta$; to check this, given three elements a, b and c, think of the smallest equivalence class which contains two of them, a and b, say: if r_i is the corresponding coarseness index, one has $d(a,b) = r_i \leq d(a,c) = d(b,c) = r_j$, $j \geq i$, $r_j \geq r_i$. The pseudometric d is also a metric iff the equivalence relation $\mathcal{R}_0$ is the (trivial) relation under which no distinct words are equivalent. Conversely, let $d = \delta$. Just set $a\mathcal{R}_{r_s}b$ whenever $d(a,b) \leq r_s$ to obtain a chain of equivalences as above, which yields back the pseudometric d, as soon checked. Consequently:

Theorem 2 (equality in the upper bound (3)): $d = \delta$ iff the underlying distortion d is a coarseness index.

To continue our example, consider the codebook made up of the three rests $\{0,2,3\}$. The minimum distance between distinct codewords, and so *also the distinguishability of the codebook*, is $d(0,2) = 2$. The minimum coarseness-index decoder corrects all outputs at coarseness index 1 from the transmitted input. E.g. say 0 is sent and 4 is received: one has $d(0,4) = 1 < d(2,4) < d(3,4)$, and so decoding is correct. This example will be continued in Section IV, where we discuss optimality.

An unruly example: The triangle inequality for the distinguishability δ does not imply the triangle inequality for the corresponding distortion d: just think of a ternary space $\{a,b,c\}$ with $d(a,c) > d(a,b) = d(b,c) = 0$ and so $\delta(a,c) = \delta(a,b) = \delta(b,c) = 0$. What is more interesting, below we shall discuss at length an example which shows that not even the inverse implication holds; $\mathcal{A} = \{x1, x2, x3, x4, x5, x6\}$.

$$d = \begin{array}{cccccc}
 & 1 & 1 & 1 & 1 & \frac{1}{2} \\
1 & & 1 & 1 & 1 & 1 \\
1 & 1 & & \frac{1}{4} & \frac{1}{2} & 1 \\
1 & 1 & \frac{1}{4} & & \frac{1}{4} & \frac{3}{4} \\
1 & 1 & \frac{1}{2} & \frac{1}{4} & & \frac{1}{2} \\
\frac{1}{2} & 1 & 1 & \frac{3}{4} & \frac{1}{2} &
\end{array}
\qquad
\delta = \begin{array}{cccccc}
 & 1 & 1 & \frac{3}{4} & \frac{1}{2} & \frac{1}{2} \\
1 & & 1 & 1 & 1 & 1 \\
1 & 1 & & \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \\
\frac{3}{4} & 1 & \frac{1}{4} & & \frac{1}{4} & \frac{1}{2} \\
\frac{1}{2} & 1 & \frac{1}{4} & \frac{1}{4} & & \frac{1}{2} \\
\frac{1}{2} & 1 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} &
\end{array}$$

$$\delta/d = \begin{array}{cccccc}
 & 1 & 1 & \frac{3}{4} & \frac{1}{2} & 1 \\
1 & & 1 & 1 & 1 & 1 \\
1 & 1 & & 1 & \frac{1}{2} & \frac{1}{2} \\
\frac{3}{4} & 1 & 1 & & 1 & \frac{2}{3} \\
\frac{1}{2} & 1 & \frac{1}{2} & 1 & & 1 \\
1 & 1 & \frac{1}{2} & \frac{2}{3} & 1 &
\end{array}$$

(to help readability, we did not write down the main diagonal, which is all-zero in the first two matrices, and undefined in the third). The first matrix describes a metric d; to see why

the triangle inequality holds it is enough to check that there are no triangles of side lengths 1/4, 1/2, 1, or 1/4, 1/4, 3/4, or 1/4, 1/4, 1. The second matrix exhibits the corresponding distinguishabilities. Since $\delta(x1, x5) = 1/2$ and $\delta(x5, x3) = 1/4$, while $\delta(x1, x3) = 1$, one has:

The distinguishability δ may violate the triangle inequality even if the diversity d is a metric distance.

The third matrix contains the ratios δ/d; it appears that in this "unruly" case *the distinguishability $\delta(a, b)$ spans the whole range from the lower bound $1/2\ d(a, b)$ to the upper bound $d(a, b)$*. This unruly example will be used again in Section IV.

Remark 6 (zero-distortion and zero-distinguishability): We are not assuming that zero-distortion $d(a, b) = 0$ and/or zero-distinguishability $\delta(a, b) = 0$ are transitive relations, and so one may have $d(a, b) = d(b, c) = 0$, and yet $d(a, c) \neq 0$, or $\delta(a, b) = \delta(b, c) = 0$, and yet $\delta(a, c) \neq 0$. When zero-distortion is a transitive relation, so is zero-distinguishability: just use bound (3). The inverse implication is not necessarily true; think again of a triangle a, b, c with $d(a, c) > d(a, b) = d(b, c) = 0$, and so with $\delta(a, c) = \delta(a, b) = \delta(b, c) = 0$. By the way, we observe that, if zero-distortion is transitive, then $\delta(a, b) = 0$ implies $d(a, b) = 0$; just think of the triangle which achieves $\delta(a, b)$. The inverse implication $d(a, b) = 0 \Rightarrow \delta(a, b) = 0$ is true for any distortion d, cf. bound (3). Cf. in Section V zero-error codes, which are interesting precisely when zero-distinguishability is *not* transitive (actually, zero-error diversities are not even bound to be as regular as distortions are, cf. Section V).

Remark 7 (is zero-distortion an error?): Re-take Remark 4, Section II. If there are distinct words at zero-distortion, one may wish to define the error set $\mathcal{E}$ as the set made up of all couples such that their distortion is positive: no error occurs if one decodes to a codeword at zero-distortion from the codeword actually sent; cf. below the case of substitution cipher distances, Section V. Correspondingly, $\delta_m(c)$ is written as

$$\delta_m(c) \ = \min_{c,c' \in \mathcal{C},\, d(c,c') \neq 0} \delta(c, c')$$

IV. Optimal codes

Let us now consider the classical *channel coding optimisation problem*: one fixes a threshold $\tau \in \mathcal{D}$, and has to find a maximum-size code $\mathcal{C}$ such as to satisfy the constraint $\delta_m \geq \tau$, and so ensure that all diversities $< \tau$ are corrected:

$$\max_{\delta_m \geq \tau} |\mathcal{C}|$$

Observe that this optimisation problem is being stated in the very general "abstract" frame of Section II: one imposes the constraint that all "small" diversities must be corrected, and, subject to this constraint, one maximises the size of the code. When the input words are n-length sequences, one equivalently maximises the *transmission rate* $n^{-1} \log_2 |\mathcal{C}|$. We stress that, even if $\mathcal{A} = \mathcal{B}$, the reliability constraint *must* be expressed in terms of the minimum distinguishability δ_m, and *not* in terms of the *minimum diversity* d_m (*minimum distortion, minimum distance*) between distinct codewords, as one is accustomed to do with Hamming-distance codes. However, even if only for a formal curiosity, one can also consider the corresponding size maximisation problem with a constraint of the type $d_m \geq \alpha$ rather than $\delta_m \geq \tau$. To make our point we re-take the examples of Section III.

The case of coarseness indices revisited: In this case diversities and distinguishabilities are fully interchangeable. Fix the constraint $\delta_m = d_m \geq r_j$, with r_j as in Section III; $0 < j \leq s$. Clearly, the constraint may be re-stated by requiring that no two codewords should be equivalent under relation $\mathcal{R}_{r_{j-1}}$. Then an optimal (maximum size) codebook $\mathcal{C}$ is obtained by selecting a representative out of each equivalence class in $\mathcal{R}_{r_{j-1}}$, and the optimal size is the corresponding number of equivalence classes, henceforth denoted by $\xi_{r_{j-1}}$. Going back to rests modulo 8 as in Section III, and choosing the constraint $\delta_m \geq 2$, the equivalence classes under $\mathcal{R}_1$ are $\xi_1 = 4$, namely $\{0, 4\}$, $\{1, 5\}$, $\{2, 6\}$, $\{3, 7\}$. An optimal codebook is, say, $\mathcal{C} = \{0, 2, 3, 5\}$, which strictly includes the *non*-optimal codebook used in Section III.

The unruly example revisited: Here the notions of distinguishability and diversity are definitely wide apart, and replacing distinguishability by diversity (a metric distance, actually) leads to bad code constructions. By an exhaustive search one is able to write down all the optimal codes, i.e. all the maximum size word subsets, which one finds with constraints expressed in terms of either d_m or δ_m. A list follows; observe that, whatever the numerical value of the constraint, the optimal code sizes are 6, 5 and 3 for d, while they are 6, 4 and 3 for δ. So, not even the optimal sizes coincide (cf. below Remark 8).

$d \geq 1/4$ or $\delta \geq 1/4$: $\{x1, x2, x3, x4, x5, x6\}$, optimal size 6, corrects no errors (each output word is a codeword and is decoded to itself)

$d \geq 1/2$: $\{x1, x2, x3, x5, x6\}$, optimal size 5

$\delta \geq 1/2$: $\{x1, x3, x5, x6\}$, $\{x1, x3, x4, x6\}$, $\{x1, x2, x5, x6\}$, $\{x1, x2, x4, x6\}$,

$\{x1, x2, x3, x6\}$, optimal size 4, correct all error distances equal to $1/4$

$d \geq 3/4$ or $d = 1$: $\{x1, x2, x3\}$, $\{x1, x2, x5\}$, ... , optimal size 3

$\delta \geq 3/4$ or $\delta = 1$: $\{x1, x2, x3\}$, ... , optimal size 3, correct all error distances $\leq 1/2$.

Fortunately, in many "friendly" cases, as are Hamming distances, distinguishability *is* a notion that one can safely forget about. We deepen this point. Let a diversity d be given with $\mathcal{A} \subseteq \mathcal{B}$, so that $d(a, b)$ is defined whenever so is $\delta(a, b)$; apart from the inclusion between the input and the output space, in the rest of this section we are assuming the general "abstract" setting[3] of Section II. Assume the following condition F holds true:

F: $\delta(a, b)$ is a non-decreasing function of $d(a, b)$, $\delta(a, b) = f\big(d(a, b)\big)$ with $f : \mathcal{D} \to \mathcal{D}$

Below we set $f^{-1}(\beta) = \min\{\alpha : f(\alpha) = \beta)\}$; f^{-1} is a genuine inverse function only when f is strictly in-

[3]The error set may be given in the general sense of Remark 4, Section II.

creasing, else one may have $\alpha > f^{-1}\big(f(\alpha)\big)$; of course $\beta = f\big(f^{-1}(\beta)\big)$ always holds with equality.

Theorem 3 (encoding theorem): Let condition F hold true: then the two optimisation problems which require to maximise the codebook size $|\mathcal{C}|$ under the constraints $\delta_m \geq \beta$ and $d_m \geq f^{-1}(\beta)$, respectively, have the same solutions.

Corollary 2: Let the function f in condition F be strictly increasing; then the optimisation problems for the constraints $d_m \geq \alpha$ and $\delta_m \geq f(\alpha)$, respectively, have the same solutions.

Proof: It will be enough to show that the inequalities $\delta(a,b) \geq \beta$ and $d(a,b) \geq f^{-1}(\beta)$ are verified by the same couples (a,b). If $\delta(a,b) \geq \beta$, then $d(a,b) \geq \alpha$ with $f(\alpha) = \beta$, and so $d(a,b) \geq \alpha = f^{-1}(\beta)$. The inverse is obvious: if $d(a,b) \geq f^{-1}(\beta)$, then $\delta(a,b) \geq \beta$. The corollary is equally obvious. ■

The following equality is straightforward, and will be frequently used in Section V: if F holds, for the minimum distinguishability and the minimum diversity of codebook $\mathcal{C}$ one has:

$$\delta_m(\mathcal{C}) = f\big(d_m(\mathcal{C})\big)$$

We observe that, unless f is *strictly* increasing, the problem in terms of d is more general than the problem in terms of δ, and *spurious* solutions may be obtained, which correspond to threshold values $\alpha > f^{-1}\big(f(\alpha)\big)$; cf. below Hamming-distance codes, Section V, where it appears that "spurious" solutions are not necessarily "useless" solutions.

Before getting rid of distinguishabilities, one should always check that a shortcut to diversities, as the one ensured by condition F, is legitimate. The request in F that $\delta(a,b)$ depends on a and b only through $d(a,b)$ may appear to be very stringent, indeed. Instead, this request is often fulfilled in practice (cf. next section), because many sequence spaces of interest are highly symmetric, or "isotropic".

We stress a fact which will come out clearly in the examples below on DNA distances. A code construction can be optimal for two types of channels based on two different types of diversity. This simply means that the codebook is the same in both cases, i.e. the optimal *encoder* is the same. However, as for the two *decoders*, they are generally different, since, on receiving the same output sequence, they decode by minimising a different quantity.

Remark 8 (reliability without optimality): If $\delta(a,b) \geq f(a,b)$ for some function f, one may devise code constructions which are optimal with respect to constraints like $f(a,b) \geq \gamma$. These code constructions are *reliable*, in the sense that they ensure $\delta(a,b) \geq \gamma$, but in general they are *not* optimal. Think of the case when the underlying diversity d verifies the triangular inequality, and so $\delta(a,b) \geq f(a,b)$ with $f = \frac{1}{2}\,d$: the non optimality of the constructions which one derives in terms of f (or equivalently in terms of d) have already been commented upon in the unruly example as revisited in this section.

Remark 9 (error detection): Assume $\alpha > f^{-1}\big(f(\alpha)\big)$ is a spurious threshold with the following property: $\delta(a,b) = d(a,z) = \alpha$ implies $d(b,z) = \alpha$; loosely, this means that any $z \in \mathcal{B}$ which yields $\delta(a,b) = \alpha$ is at "exactly half way" between a and b. This is the case of the Hamming geometry when α is an even integer, and can be used to endow spurious code constructions with error *detection* capabilities, as explained in any textbook on coding theory; cf. our comment on Hamming-distance codes in Section V.

Remark 10 (the "capacity" of a coarse channel): In this remark we take the asymptotic point of view which is typical of Shannon theory. Think e.g. of coarseness indices: two "single letters", e.g. two rests modulo 8 in our numeric example, are "similar" when they are equivalent even under "severe" relations, those for which the number ξ_{r_j} of equivalence classes is high; cf. above. This single-letter diversity may be extended to letter sequences of the same length n in several "reasonable" ways, e.g. by deciding that two sequences are similar when they are so in each position; this gives the diversity:

$$d_n(x_1x_2\,\ldots\,x_n, y_1y_2\,\ldots\,y_n) \;=\; \max_{1\leq i\leq n}\, d(x_i,y_i)$$

One soon checks that d_n is itself a coarseness index with the same values as $d = d_1$. With rests modulo 8 one has, e.g., $d_4(0272,0432) = \max\{0,2,1,0\} = 2$. Once the reliability threshold r_j has been chosen, an optimal codebook $\mathcal{C}_n$ for sequences is soon obtained from an optimal single-letter codebook $\mathcal{C} = \mathcal{C}_1$ by considering all the n-sequences over $\mathcal{C}$: $\mathcal{C}_n = \mathcal{C}_1^n$ (the proof is straightforward). So, the optimal code-rate is independent of the sequence length n:

$$\frac{1}{n}\,\log_2|\mathcal{C}_n| \;=\; \log_2\xi_{r_{j-1}}\;,\quad \text{for}\;\; \delta_m \geq r_j$$

and the *asymptotic optimal rate* when n goes to infinity remains equal to $\log_2\xi_{r_{j-1}}$. Whether to call or not to call this asymptotic optimal rate the channel's *capacity* is a matter of personal taste. We stress that this "capacity" depends on the reliability threshold r_j and depends of course on the "channel's model", i.e. on the way how we derive sequence diversities from letter diversities. As for the dependence on the reliability threshold, observe that also Shannon's capacity has *two* distinct values, one when the allowed error probability is positive even if negligible (this corresponds to capacity without further specifications, cf. [7] or [8]), the other when the allowed error probability is strictly zero (this corresponds to zero-error capacity, or *graph capacity*, cf. [8] or [2]). In the case of Hamming-distance codes the dependence of optimal asymptotic rates on the reliability threshold, in this case the *correction rate*, is a well-known fact; cf. e.g. [9]. Cf. footnote 5 for the relationship between capacity with respect to coarseness indices and graph capacity.

V. Some relevant case studies

In this section we put to work our machinery on several examples which we deem to be relevant. Apart from zero-error codes, the diversities which we examine are all "Hamming-like", and so it does not come as a surprise that distinguishabilities are obtained by output sequences $\underline{z}$ which achieve the lower bound (5); cf. instead the codebook based on the coarseness index in Section III and the "unruly example" in Sections III and IV.

In this section we assume, as we did in Section IV, $\mathcal{A} \subseteq \mathcal{B}$; we also assume that the diversity d is a non-negative real number and it is symmetric when it is restricted to the input set: $d(a,b) = d(b,a)$, $a, b \in \mathcal{A}$. With these assumptions, the *generalised triangle inequality* $d(a,z) + d(b,z) \geq d(a,b)$ $\forall a, b \in \mathcal{A}$, $\forall z \in \mathcal{B}$ might hold true. This inequality, as soon checked, implies the useful lower bound, or *extended metric bound*:

$$\delta(a,b) \;\geq\; \frac{1}{2}\, d(a,b)$$

which can be strengthened to

$$\delta(a,b) \;\geq\; \left\lceil \frac{1}{2}\, d(a,b) \right\rceil_{\mathcal{D}} \tag{5}$$

by recalling that $1/2\; d(a,b)$ is not necessarily a value that the diversity can take; recall that the $\mathcal{D}$-ceiling reduces to the integer ceiling whenever $\mathcal{D}$ is made up of consecutive integers.[4] Erasure channels with a low parameter ρ (cf. below) provide a non-metric example, where the lower bound (5) does not hold. The diversities we shall be dealing with may have or may have not certain metric properties; to speed up presentation, the related discussion is relegated to Appendix A (which is a sort of "container" for uninteresting computations).

A. Back to Shannon

Zero-error codes: The binary case $\mathcal{D} = \{0, 1\}$ belongs to Shannon's zero-error information theory and leads to the asymptotic and fascinating notion of *graph capacity*; cf. [2]. In practice, in the zero-error case one sets $d(a,z) = 0$ if the transition from the input word a to the output word z is *possible* (in Shannon's case this happens when the transition probability is *positive*, even if very small), else one sets $d(a,z) = 1$ (the transition is *impossible*, its probability is *zero*). It is in this context that the notion of distinguishability/confusability was developed: two input words a and b are *zero-error confusable* whenever there is an output word z such that the transitions from a to z and from b to z are both possible. In our setting, zero-error confusability translates to $\delta(a,b) = 0$; in a reliable zero-error code no two codewords should be zero-error confusable. Shannon's zero-error codes are trivial when confusability, i.e. distinguishability equal to zero, is a transitive relation[5] on the input space, but they are quite fascinating in the opposite case; cf. again [2]. As a non-metric (and nasty) example, take $\mathcal{A} = \mathcal{B} = \{a, b, c\}$, and assume that the only possible transitions are from a to b, from b to c, from c to a and c. The distinguishable couples are a, b and a, c. One has $\delta(a,b) = 1 > d(a,b) = d(a,b) \wedge d(b,a) = 0$; cf. instead the general lower bound in footnote 4. Here the (non-symmetric) diversity is not even a (non-symmetric) distortion, since $d(b,b) = 1 > d(a,b) = 0$.

B. Hamming geometry, theme and variations

Hamming-distance codes: Input and output sequences are of the same length n; $\mathcal{A} = \mathcal{B} = \mathcal{X}^n$; $\mathcal{X}$ is the single-letter alphabet, or *ground alphabet*. By using the metric bound (5), one soon gets:

$$\delta_H(\underline{x}, \underline{x}') \;=\; \left\lceil \frac{d_H(\underline{x}, \underline{x}')}{2} \right\rceil$$

(The minimising z in (1) coincides with the two sequences in those positions where they are equal, elsewhere it is equal to the first sequence in exactly half of the positions when the Hamming distance is even, in half ± 1 of the positions when the Hamming distance is odd). The dependence of $\delta = \delta_H$ on $d = d_H$ is non-increasing, and so one can optimise with respect to Hamming distances, as every coding theorist has always been doing (use the encoding theorem of Section IV). Below we tabulate the first values of the function $\delta = f(d)$:

d	1	2	3	4	5	6	...
δ	1	1	2	2	3	3	...

Spurious constructions correspond to an even d_m. For example, the constraint $d_m \geq 2$ gives parity-bit codes, which are quite useless from the point of view or error *correction*, as done in this paper. Needless to say, optimal code constructions corresponding to an even d_m may be made good use of in error *detection*; cf. [9], cf. also our Remarks 5 and 8.

In a probabilistic approach to coding, where one decodes by maximum likelihood, decoding by minimum Hamming distance is re-obtained if one assumes that the transmission channel is *symmetric stationary memoryless* and has a "small" *crossover probability* (smaller than $1/2$ in the binary case; the crossover probability is the probability of receiving an output symbol which is different from the input symbol). We observe that in a fully probabilistic approach the "natural" reliability criterion is to bound the decoding error probability, as Shannon did: however, this leads to *different* optimal code constructions from those obtained by bounding the minimum distance between codewords (it also leads to the notion of Shannon capacity as maximal mutual information). For these well-known facts cf. e.g. [7], [8] or [9].

Erasure channels: With respect to the Hamming case, add a new output symbol $*$ called the *erasure*, or the *star*: the output sequences are n-length sequences over the extended ground alphabet $\mathcal{X} \cup \{*\}$. We stress that erasures, unlike *deletions* as below, are *retraceable*, i.e. the decoder knows the exact positions where they have occurred. After setting $d_E(a, *) = \rho$ in each position where the output sequence has an erasure, $\rho > 0$, one may extend Hamming distances in the usual additive way; in practice, the diversity is $t + s\rho$ when there are t crossovers and s erasures. For example $d_E(0101, 1 * *1) = 1 + 2\rho$. For a probabilistic interpretation of ρ cf. Appendix C, part A. Unfortunately, d_E verifies the

[4] We are not insisting on unsymmetric diversities only because they are not needed in our examples; recall however that distinguishabilities are symmetric even if the underlying diversity is not. For $\mathcal{A} \subseteq \mathcal{B}$, the latter observation allows one to obtain "non-symmetric versions" of the metric bound and also of the upper bound (3). If $0 \leq d(a,a) \leq d(b,a)$ whatever a and b, the upper bound (3) generalises to $\delta(a,b) \leq d(a,b) \wedge d(b,a)$; zero-error codes are sometimes so unruly that even this very general upper bound may fall, cf. below. If d is a real number and if the extended triangle inequality holds true, the extended metric bound (5) generalises to $\delta(a,b) \geq \left\lceil \frac{1}{2} d(a,b) \right\rceil_{\mathcal{D}} \vee \left\lceil \frac{1}{2} d(b,a) \right\rceil_{\mathcal{D}}$.

[5] Readers may have noticed that, when zero-distinguishability is an equivalence relation, zero-error codes can be re-described in terms of coarseness indices with just two equivalence relations, the coarser of the two being the trivial relation when everything is equivalent. So, there is an intersection between graph capacity and the "capacity" as in Remark 10.

triangle inequality only for $\rho \geq 1/2$; cf. Appendix A on metric properties. Below angular brackets denote a binary logical value, $0 =$ false, $1 =$ true; one has:

$$\delta_E(\underline{x}, \underline{x}') \ = \rho \, d_H(\underline{x}, \underline{x}') \ , \ \ \rho \leq \frac{1}{2} \ ;$$

$$\delta_E(\underline{x}, \underline{x}') \ = \left\lceil \frac{d_H(\underline{x}, \underline{x}')}{2} \right\rceil \ , \ \ \rho \geq 1 \ ;$$

$$\delta_E(\underline{x}, \underline{x}') = \left\lfloor \frac{d_H(\underline{x}, \underline{x}')}{2} \right\rfloor + \rho \Big\langle d_H(\underline{x}, \underline{x}') \text{ is odd} \Big\rangle, \ \frac{1}{2} \leq \rho \leq 1$$

We prove these equalities in Appendix A; actually, for $\rho \geq 1/2$ and $d_E(\underline{x}, \underline{y})$ even, one just has to use the metric bound (5). In practice, minimisation as in (1) is achieved by the all-star sequence when $\rho \leq 1/2$, by the no-star sequence when $\rho \geq 1$, while for the intermediate values of ρ one uses one or no stars according whether the number of crossovers is odd or even.

Whatever ρ, the distinguishability $\delta = \delta_E$ is soon checked to depend monotonically on $d = d_E$, and so optimal code constructions may be obtained by minimising with respect to Hamming distances rather than distinguishabilities. This implies that "good" algebraic codes remain good also when used on channels with erasures, in accordance to what a practitioner would be doing anyway. For $\rho < 1$, δ is a *strictly* increasing function of $d = d_E$ and so there are no spurious solutions. We tabulate the first values of the function $\delta = f(d)$, assuming $\frac{1}{2} \leq \rho < 1$:

d	1	2	3	4	5	6	...
δ	ρ	1	$1+\rho$	2	$2+\rho$	3	...

For example, when using a parity bit code, the minimum Hamming distance between codewords is 2 and all single erasures are properly corrected, as ensured by the reliability criterion in Section II. As for the *decoder*, observe that it knows exactly the s positions where the received sequence $\underline{z}$ has stars; since those positions give a constant contribution to the diversity $d_E(\underline{c}, \underline{z})$ to be minimised, the decoder will simply have to minimise Hamming distances limited to the $n - s$ positions where $\underline{z}$ has no stars. In particular, it behaves exactly as a Hamming-distance decoder when $s = 0$ (for $s = 0$ the erasure distance and the Hamming distance coincide). As for error correction capabilities, un unpleasant feature of our diversity is that it mixes together erasures and crossovers, while the decoder knows exactly which erasures it has to correct. A finer analysis, based on a generalised version of the reliability criterion, is given in Appendix C, part B; there we also re–assess the capabilities to correct erasures and crossovers for the famous low-rate code with $n = 32$ and $d_m = 16$ used by Mariner 1969.

C. Edit distance: replacements, deletions and insertions

The reader is referred e.g. to [10] for a formal definition of the edit distance, or Levenštejn distance, and for the dynamic programming techniques used to compute it; cf. also Appendix A. We shall consider a situation when the input sequences have all a fixed length n, while the output space is made up of sequences over the input alphabet which have a bounded[6] length: $\mathcal{B} = \cup \mathcal{X}^i$ with $0 \leq i \leq m$, $m \geq n$. To compute the edit distance $d_{edit}(\underline{x}, \underline{y})$ one needs a number to be interpreted as the "cost" of each letter *deletion*, of each letter *insertion*, and of each letter *replacement* (letter crossover): for simplicity we shall assume, as one often does, that deletions, insertions and replacements have all the same cost 1. Loosely speaking, the edit distance is the smallest additive cost of a path which, by use of deletions, insertions and replacements, changes one sequence into the other. With our choice of costs, one has $d_{edit}(\underline{x}, \underline{y}) \leq d_H(\underline{x}, \underline{y})$ whenever $\underline{x}$ and $\underline{y}$ have the same length; to prove this, just think of a replacement in each position where the two words differ. Instead, there is *no* serious upper bound for d_H in terms of d_{edit}: take e.g. two sequences $a\underline{x}$ and $\underline{x}a$ where the letter a does not occur in $\underline{x}$, and $\underline{x}$ has length $n - 1$; one has $d_{edit}(a\underline{x}, \underline{x}a) = 2$, while $d_H(a\underline{x}, \underline{x}a)$ can take any value from 2 to n, according to the composition of $\underline{x}$. The edit distance is known to verify the triangle inequality; cf. Appendix A. Using the lower bound (5), one gets:

$$\delta_{edit}(\underline{x}, \underline{y}) \ = \ \left\lceil \frac{1}{2} \, d_{edit}(\underline{x}, \underline{y}) \right\rceil$$

(cf. again Appendix A). Since $d_{edit} \leq d_H$, optimal code constructions for edit distances are a *fortiori* reliable for Hamming distances; cf. Remark 8, Section IV. Unfortunately, the inverse is not true, and so Hamming-distance codes cannot be recommended: entirely new encoders and new decoders are needed. The combinatorial problem of devising clever code constructions by checking the edit distance between codewords appears to be hard indeed, and so practitioners may decide to use Hamming-distance codes even on channels with deletions and insertions.

As a toy example, consider the codebook $\{aaaa, abbb, cccb\}$. The minimum edit distance between distinct codewords is soon checked to be 3, and so the minimum distinguishability is $\lceil 3/2 \rceil = 2$: the decoder will correct all diversities ≤ 1, i.e. a single crossover, or a single deletion, or a single insertion. If, say, the codeword $abbb$ is sent, and it is received as $acbb$, or as abb, or as $abbab$, then it is correctly decoded by a minimum edit-distance decoder. Correction is not ensured when there two "accidents" during transmission: e.g. if $abbb$ is received as $aaab$ (two crossovers), the decoder will incorrectly decode to $aaaa$.

D. A cryptographic diversion

A substitution-cipher pseudometric: Set $d_{SC}(\underline{x}, \underline{y}) = \min_\sigma d_H\big(\underline{x}, \sigma(\underline{y})\big)$ where the minimum is taken over all the permutations σ of the ground alphabet $\mathcal{X}$; the single-letter "substitution" σ extends component-wise to sequences.

[6]Assuming that output sequences have a bounded length simply means that if an input letter is fed to the channel, at most m/n output letters can be produced at the output, a very reasonable assumption in practical cases. If the channel is used more than once (if several codewords are sent in turn), the block-length property of codewords, all of which have the *same* length n, goes completely lost, and decoding is hopeless unless we assume that there is some way to parse the flow of output words, some of whom might even be empty. This is the case when codewords are sent at regular time intervals, each time interval being long enough to produce m letters at the output. In practice, we are ignoring synchronisation problems.

For example, with the Latin alphabet as ground alphabet, $d_{SC}(\mathrm{CARRVM}, \mathrm{BELLIS}) = 0$: just take any σ which maps C to B, A to E, R to L, V to I, M to S, and so CARRUM to BELLIS. Think of a situation when what one observes are not the actual letters which make up the output sequence, but only letter differences, i.e. the fact whether in two positions the output sequence has or has not the same letter; this situation is met with cryptograms obtained by use of a simple substitution cipher; cf. e.g. [11] or [12]. Clearly, d_{SC} is a pseudometric; cf. also Appendix A. One has $d_{SC}(\underline{x}, \underline{y}) \leq d_H(\underline{x}, \underline{y})$; in the example above $d_{SC}(\underline{x}, \underline{y}) = 0$ and $d_H(\underline{x}, \underline{y}) = n$: this shows that there is no serious way of bounding from above d_H in terms of d_{SC}. By using the metric bound (5) the distinguishability is readily shown to be:

$$\delta_{SC}(\underline{x}, \underline{y}) \;=\; \left\lceil \frac{1}{2}\, d_{SC}(\underline{x}, \underline{y}) \right\rceil$$

(cf. Appendix A). Optimal code constructions may be obtained by minimising with respect to the underlying pseudometric d_{SC}, but *not* with respect to Hamming distance d_H. Thus, in this case, genuinely *new* code constructions are needed. Going back to Remark 7, if the error set comprises the couples at positive distance, $\mathcal{E} = \{\underline{x}, \underline{x}' : d_{SC}(\underline{x}, \underline{x}') \neq 0\}$, the situation is as follows: the codeword is enciphered by a nasty guy and then it is sent through a noisy symmetric channel as those found in algebraic coding; the decoder tries to recover the cryptogram cleansed of channel noise, so that the legitimate receiver may (hopefully) decrypt it by statistical cryptanalysis, as is standard for simple substitution ciphers. If instead the error set comprises *all* distinct couples, one expects that the decoder does the whole job, inclusive of decryption. For this to be possible, the codebook will have to be definitely smaller; in particular, no two codewords can have the same "abstract structure" (can be a cryptogram of each other, as are CARRVM and BELLIS above).

As an example, consider the (trivial) codebook $\{aaaa, abcd\}$. The substitution-cipher distance between the two codewords is 3, which gives a distinguishability equal to $\lceil 3/2 \rceil = 2$. If we are willing to take care of statistical cryptanalysis after decoding (by the way, a bad idea given the shortness of the codewords), we can freely add all the cryptograms which one may obtain from $aaaa$ and $abcd$ by use of a simple substitution cipher: if the ground alphabet is $\{a, b, c, d\}$, this will give a new codebook of 28 codewords, $aaaa$, $bbbb$, $cccc$, $dddd$ (the possible cryptograms for $aaaa$), and the $4! = 24$ permutations of $abcd$ (the possible cryptograms for $abcd$). The minimum distinguishability computed as in Remark 7, Section III, is still $\lceil 3/2 \rceil = 2$. If $abcd$ is sent and it is received unscathed, it might be decoded to any of its possible cryptograms, $dcba$, say, but this would be no error, since $d_{SC}(abcd, dcba) = 0$.

E. DNA distances

Below we shall deal with code constructions based on DNA string distances derived from DNA word design; we observe that DNA word design is a research domain where the need for a solid and comprehensive theoretical foundation is felt by many, and where the Shannonian notions of distinguishability and confusability might turn out to be valuable. The examples to follow are only a very preliminary step; our purpose here is simply to put to work our (hopefully flexible) tools in an unusual context, based on "odd" string distances. The channels below are too naive to explain code constructions as the ones in [5], say, in which one checks the reverse Hamming distance between codewords, and which are justified through biological arguments. One might think of more elaborate diversities, e.g. one might choose a diversity which is large when *both* the usual Hamming distance *and* the reverse Hamming distance between input and output are large; this would point to code constructions where *both* these distances must be kept high when choosing codewords. One will have to carefully *understand the error correction capabilities* of codes as those in [5], before trying to devise diversities which fit the corresponding constructions on the basis of the reliability criterion in Section II.

Reverse Hamming distance: After denoting by $\underline{x}^*$ the mirror image of $\underline{x}$, the reverse Hamming distance $d_{RH}(\underline{x}, \underline{y})$ between two strings $\underline{x}$ and $\underline{y}$ of the same length n is simply the Hamming distance $d_H(\underline{x}, \underline{y}^*)$. One has $d_{RH}(\underline{x}, \underline{x}) = 0$ iff $\underline{x}$ is a palindrome. This "distance" is symmetric, but the triangle inequality falls (cf. Appendix A). Even without the metric bound, one can soon compute the distinguishability δ_{RH}, which is exactly the same as in the standard Hamming case:

$$\begin{aligned}\delta_{RH}(\underline{x}, \underline{y}) &= \min_{\underline{z}} \left[d_{RH}(\underline{x}, \underline{z}) \vee d_{RH}(\underline{y}, \underline{z}) \right] \\ &= \min_{\underline{z}} \left[d_H(\underline{x}, \underline{z}^*) \vee d_H(\underline{y}, \underline{z}^*) \right] = \delta_H(\underline{x}, \underline{y})\end{aligned}$$

(Minimising over all $\underline{z}$ is the same as minimising over all $\underline{z}^*$). Since the distinguishability function is the same as in the standard case of Hamming distance coding, nothing new is required from the point of view of code constructions. Although encoding does not change, decoding is different: to ensure that certain errors are corrected, as in reliability criterion, one must decode by minimising the reverse Hamming distance, and *not* the usual Hamming distance; cf. our comment at the end of Section IV. In the case of reverse Hamming distances and reverse complement Hamming distances to be introduced below, distances and distinguishabilities are *not* minimised by taking $\underline{x} = \underline{y}$; in particular, they are not even distortions as in Definitions 3, Section III. However, the "operational" results of Sections II and IV do not need any special assumptions, and so they can be freely applied. Since the natural alphabet of DNA sequences has four "letters", A, C, G, T (adenine, cytosine, guanine and thymine), quaternary constructions are relevant as those given in [13].

Reverse complement Hamming distance: A *complement* is an *involutive* permutation σ of the ground alphabet $\mathcal{X}$, to be extended componentwise to strings; in the case of the DNA complement, $\sigma(A) = T$, $\sigma(C) = G$ (A adenine, C cytosine, G guanine and T thymine), and so $\sigma(AAGT) = TTCA$. One sets: $d_{RCH}(\underline{x}, \underline{y}) = d_H\big(\underline{x}, \sigma(\underline{y}^*)\big)$; observe that mirror image and complement commute: $\sigma(\underline{x}^*) = \big(\sigma(\underline{x})\big)^*$; cf. Appendix A. Once more the distinguishability δ_{RCH} is exactly the same as

for Hamming distances:

$$\delta_{RCH}(\underline{x},\underline{y}) = \min_{\underline{z}} \left[d_H\big(\underline{x},\sigma(\underline{z}^*)\big) \vee d_H\big(\underline{y},\sigma(\underline{z}^*)\big) \right] = \delta_H(\underline{x},\underline{y})$$

(Minimising over all $\underline{z}$ or over all $\sigma(\underline{z}^*)$ is the same). As for codebooks and decoders, cf. our comment to reverse Hamming distances.

Shifting Hamming distance: Even if it is not present in the list of DNA distances given in [4], we use this distance because it is a "simplified version" of the genuine DNA distance to follow, and so requires shorter computations. Take an integer k; given an n-length sequence $\underline{y}$, consider its circular shift $k(\underline{y})$ of k positions, clockwise if k is positive, counterclockwise if k negative; actually, shifts k with $0 \leq k < n-1$ will do. The shifting Hamming distance $d_{SH}(\underline{x},\underline{y})$ is obtained by first computing the n usual Hamming distances between $\underline{x}$ and the k-shifts of $\underline{y}$, $0 \leq k < n-1$, and by then selecting the smallest of them; it is a pseudometric (cf. Appendix A). The distinguishability δ_{SH} can be computed directly, and shows that the metric lower bound is attained:

$$\delta_{SH}(\underline{x},\underline{y}) \;=\; \left\lceil \frac{1}{2}\, d_{SH}(\underline{x},\underline{y}) \right\rceil$$

To prove this, we simply swap the minimum over the output space and the minima over the shifts, and so express $\delta_{SH}(\underline{x},\underline{y})$ in terms of the usual Hamming distinguishability δ_H:

$$\begin{aligned}
&\delta_{SH}(\underline{x},\underline{y}) = \\
&\quad = \min_h \min_k \left[\min_{\underline{z}} \left[d_H\big(h(\underline{x}),\underline{z}\big) \vee d_H\big(k(\underline{y}),\underline{z}\big) \right] \right] = \\
&\quad = \min_h \min_k \delta_H\big(h(\underline{x}),k(\underline{y})\big) = \\
&\quad = \min_k \delta_H\big(\underline{x},k(\underline{y})\big) = \min_k \left\lceil \frac{d_H\big(\underline{x},k(\underline{y})\big)}{2} \right\rceil = \\
&\quad = \left\lceil \frac{\min_k d_H\big(\underline{x},k(\underline{y})\big)}{2} \right\rceil = \left\lceil \frac{1}{2}\, d_{SH}(\underline{x},\underline{y}) \right\rceil
\end{aligned}$$

We stress that in the case of the shifting Hamming distinguishability the underlying distance d_{SH} has *no* significant lower-bound in terms of the usual Hamming distance (the Hamming distance between a sequence and one of its shifts may be as large as n): so, genuinely *new* code constructions are needed. Since $\delta_{SH}(\underline{x},\underline{y}) \leq \delta_H(\underline{x},\underline{y})$, possibly strictly, the new codebooks are sparser than Hamming-distance codebooks: if a code construction is optimal for shifting Hamming distances, it is *a fortiori* reliable for Hamming distances, for reverse Hamming distances and for reverse complement Hamming distances; cf. Remark 8, Section IV.

Reverse complement shifting Hamming distance: This is a combined version of the shifting and the reverse complement versions of the Hamming distance, and is called also H-measure in the literature [14]. To obtain $\delta_{SRH}(\underline{x},\underline{y})$ one proceeds as in the case of the shifting Hamming distances, but applies the n shifts not directly to $\underline{y}$, but rather to its reverse complement $\sigma(\underline{y}^*) = \big(\sigma(\underline{y})\big)^*$. The distinguishability can be again computed directly and turns out to be exactly the same as for shifting Hamming distances, and so the same code constructions are needed as for shifting Hamming distances. We omit the lengthy but uninteresting computations.

VI. Conclusion

The approach to channel coding taken in this paper may be seen as a *multi-level generalisation* of the two-level approach which is typical of Shannon's zero-error information theory. It makes it explicit that diversity is a local notion which involves just *one* output sequence at a time, while distinguishability is a global notion which involves *all* output sequences (equivalently: something is similarity, something else is confusability). Diversities and distinguishabilities are quite neatly contrasted in the text of the very general reliability criterion of Section II, even if a sequence space may be so regular that in practice the two notions collapse into one. The "coding principle" underlying the reliability criterion is quite flexible and may either lead to new code constructions, or may help to better understand available code constructions. In on-going work we plan to investigate systematically the significance of the present approach to the coding-theoretic problems posed by molecular computation as based on DNA string distances.

Appendix A
Metric properties

A. Erasures

The Hamming distance as extended to erasures verifies the triangle inequality for $\rho \geq 1/2$; this is soon checked by just splitting the three sequences into the s positions where $\underline{z}$ has a star, and the remaining $n-s$ positions. Instead, if $\rho < 1/2$ the triangle inequality is violated e.g. when $\underline{x}$ is a run of 0's, $\underline{y}$ is a run of 1's and $\underline{z}$ is a run of n erasures; observe that one has $\delta_E(\underline{x},\underline{y}) = \rho n < 1/2\; n = 1/2\; d_E(\underline{x},\underline{y})$, and so also the lower metric bound (5) is violated.

Let us move on to distinguishabilities. Nothing is needed for $\rho \geq 1/2$ when the Hamming distance between the input sequences is even, because one can use directly the metric bound; else we have to make a few computations, which are easy but a bit lengthy. We compute $\delta_E(\underline{x},\underline{y})$ assuming for the moment that the two strings have length h and differ in each of the h positions (recall that no stars occur in them). Let us solve the minimisation problem (1) with the *additional constraint* that $\underline{z}$ must have exactly s stars, $0 \leq s \leq h$; later, we shall have to minimise also with respect to s. The *constrained distinguishability* is then $\Delta_s = \Delta_s(\underline{x},\underline{y}) = s\,\rho + \lceil \frac{h-s}{2} \rceil$; actually, because of the additive nature of the erasure distance, we can as well think that the stars are in any fixed s positions, and then use the Hamming-distance metric bound for the remaining $h-s$ positions, where there cannot be any stars. Assume now $\rho \leq 1/2$. We compare Δ_s, $s < h$, with Δ_h, obtained by the all-star sequence. One soon checks that the difference $\Delta_s - \Delta_h = (h-s)\,(1/2-\rho) + 1/2\langle h-s \text{ odd}\rangle$ is non negative. So, for $\rho \leq 1/2$, minimising Δ_s over s gives $s = h$ and $\delta_E(\underline{x},\underline{y}) = h\,\rho$: the best thing to do is to use only stars. Now we go to the case $\rho \geq 1/2$; we assume that h is odd (nothing else will be needed). For s even (then $h-s$ is odd), let us compare Δ_s, $s \geq 2$, with the no-star sequence, i.e. with Δ_0; one soon checks that the difference $\Delta_s - \Delta_0 = s\,(\rho - 1/2)$ is non negative, and so the best thing to do is using no stars at all. For s odd (then $h-s$ is

even), let us compare Δ_s, $s \geq 3$, with Δ_1; one soon checks that the difference $\Delta_s - \Delta_1 = (s-1)(\rho - 1/2)$ is non negative and then the best thing to do is using exactly one star. So, it is enough to compare Δ_1 with Δ_0. The difference $\Delta_1 - \Delta_0 = \rho - 1$ is positive for $\rho > 1$ (do not use any star), zero for $\rho = 1$ (use one or no stars, indifferently), negative for $1/2 \leq \rho < 1$ (use exactly one star). All this proves the formulas for distinguishabilities given in Section V: just take the h positions where the two n-length sequences differ, $h = d_H(\underline{x}, \underline{y})$, since in the remaining $n - t$ positions the minimising z will of course coincide with both $\underline{x}$ and $\underline{y}$.

B. Edit distance

In general one has three (positive and finite) costs, ρ_d, ρ_i and ρ_r, for letter deletions, letter insertions and letter replacements, respectively (sometimes further types of error are considered, e.g. *twiddling* two adjacent letters). Take a path which converts $\underline{x}$ into $\underline{y}$, each step of the path corresponding to one letter deletion, one letter insertion, or one letter replacement; compute the overall cost of the path by addition of the single costs. The edit distance $d_{edit}(\underline{x}, \underline{y})$ is the lowest possible cost for converting $\underline{x}$ into $\underline{y}$. Since an optimal path from $\underline{x}$ to $\underline{z}$ followed by a an optimal path from $\underline{z}$ to $\underline{y}$ gives a not necessarily optimal path from $\underline{x}$ to $\underline{y}$, one always has the triangular property $d_{edit}(\underline{x}, \underline{z}) + d_{edit}(\underline{z}, \underline{y}) \geq d_{edit}(\underline{x}, \underline{y})$; the arguments must be specified in this order because the edit distance is *not* symmetric when $\rho_d \neq \rho_i$: e.g. $d_{edit}(a, \lambda) = \rho_d \neq d_{edit}(\lambda, a) = \rho_i$, λ being the zero-length sequence and a being any letter. If $\rho_d = \rho_i$, the edit distance is a metric; in Section V we have assumed $\rho_d = \rho_i = \rho_r = 1$. Any "reasonable" path from $\underline{x}$ to $\underline{y}$ (one which does not delete inserted letters, say) may be represented by a *transcript* over the quaternary alphabet $\{i, d, k, r\}$ (i = insert, d =delete, k = keep , r = replace); the length ℓ of the transcript is at most equal to the maximum length $\ell(\underline{x}) \vee \ell(\underline{y})$; e.g. an optimal transcript which transforms MOSHE into MOSES is $kkkdki$, while an optimal transcript which transforms MUSA into MOSES is $krkri$, and so the edit distances are 2 and 3, respectively. In our case, one uses exactly as many deletions as insertions since both input sequences have the same length n; one has $d_{edit}(\underline{x}, \underline{y}) = u + v$ when in the optimal transcript there are $u/2$ deletions d, $u/2$ insertions i, and v replacements r. The metric bound is obtained by the "short" sequence z obtained from $\underline{x}$ by deleting letters corresponding to d in the optimal transcript and by performing $\lceil v/2 \rceil$ of the required replacements.

C. Substitution ciphers

One deals with a pseudometric, as soon checked; let us e.g. verify the triangular property:

$$\begin{aligned} &d_{SC}\big(\underline{x}, \underline{z}\big) + d_{SC}\big(\underline{z}, \underline{y}\big) = \\ &\quad = d_H\big(\underline{x}, \sigma(\underline{z})\big) + d_H\big(\underline{z}, \tau(\underline{y})\big) = \\ &\quad = d_H\big(\underline{x}, \sigma(\underline{z})\big) + d_H\big(\sigma(\underline{z}), \sigma(\tau(\underline{y}))\big) \geq \\ &\quad \geq d_H\big(\underline{x}, \sigma(\tau(\underline{y}))\big) \geq d_{SC}\big(\underline{x}, \underline{y}\big) \end{aligned}$$

In the first equality σ and τ solve for the minimisation in the definition of the substitution cipher distance. One has $d_{SC}(\underline{x}, \underline{y}) = 0$ iff $\underline{x}$ and $\underline{y}$ have the same "structure", in the sense that in any two positions i and j the letters in $\underline{x}$ are equal iff so are the corresponding letters in $\underline{y}$, $1 \leq i < j \leq n$. In turn, this happens iff $\underline{y}$ can be obtained as the cryptogram of $\underline{x}$ by use of a simple substitution cipher; more generally, $d_{SC}(\underline{x}, \underline{y}) = d_{SC}\big(\underline{x}, \sigma(\underline{y})\big)$ for any permutation σ of the ground alphabet. As for the metric bound (5):

$$\begin{aligned} &\delta_{SC}\big(\underline{x}, \underline{y}\big) = \\ &\quad = \delta_{SC}\big(\underline{x}, \sigma(\underline{y})\big) \leq \delta_H\big(\underline{x}, \sigma(\underline{y})\big) = \\ &\quad = \left\lceil \frac{d_H\big(\underline{x}, \sigma(\underline{y})\big)}{2} \right\rceil = \left\lceil \frac{d_{SC}\big(\underline{x}, \underline{y}\big)}{2} \right\rceil \end{aligned}$$

The first equality for distinguishabilities soon derives from the corresponding equality for distances; we choose σ equal to the "minimising" permutation for which Hamming distance and substitution-cipher distance coincide; cf. the last equality. So the same sequence $\underline{z}$ yields both the metric bound for the substitution-cipher distance and the metric bound for the Hamming distance, however with $\sigma(\underline{y})$ instead of $\underline{y}$. Of course, in the derivation the inequality is actually an equality, else the metric bound would be violated.

D. Reverse Hamming distance

One has $d_{RH}(\underline{x}, \underline{y}) = 0$ iff $\underline{y} = \underline{x}^*$, and so one has $d_{RH}(\underline{x}, \underline{x}) = 0$ iff $\underline{x}$ is a palindrome. The reverse Hamming distance is symmetric, as soon checked; unfortunately, the triangle inequality falls: just take the triangle $\underline{x}, \underline{x}^*, \underline{x}$ when $\underline{x}$ is not a palindrome. In the terminology of Section III, the reverse Hamming distance is not even a symmetric distortion.

E. Reverse complement Hamming distance

One soon checks that mirror image * and complement σ commute: $\sigma(\underline{x}^*) = \big(\sigma(\underline{x})\big)^*$. One has $d_{RCH}(\underline{x}, \underline{y}) = 0$ iff $\underline{y} = \sigma(\underline{x}^*)$; one has $d_{RCH}(\underline{x}, \underline{x}) = 0$ iff $\underline{x}$ is a palindrome and its letters are all *fixed points* for the involutive permutation σ; this never happens in the case of the genuine DNA complement. As in the case of reverse Hamming distances, this distance is symmetric, but it does not verify the triangle inequality.

F. Shifting Hamming distance and reverse complement shifting Hamming distance

The shifting Hamming distances is a pseudometric, as soon checked. Let us e.g. prove the triangle inequality; below $k(\underline{y})$ denotes the sequence obtained by shifting $\underline{y}$ of k positions; for suitable h and k:

$$\begin{aligned} &d_{SH}(\underline{x}, \underline{y}) + d_{SH}(\underline{y}, \underline{z}) = \\ &\quad = d_H\big(\underline{x}, h(\underline{y})\big) + d_H\big(\underline{y}, k(\underline{z})\big) = \\ &\quad = d_H\big(\underline{x}, h(\underline{y})\big) + d_H\big(h(\underline{y}), h(k(\underline{z}))\big) \geq \\ &\quad \geq d_H\big(\underline{x}, h(k(\underline{z}))\big) \geq d_{SH}(\underline{x}, \underline{z}) \end{aligned}$$

In the second side of the first equality, h and k are optimal shifts which yield the minima; we have also used the fact that Hamming distance is insensitive to the same shift of its two arguments, and the fact that the composition of two shifts is itself a shift, not necessarily optimal. Also the reverse complement shifting Hamming distance, or Hamming measure, is a pseudometric; for the routine but lengthier proof the reader is referred to [14].

APPENDIX B
STOCHASTIC-LIKE SIMILARITIES

In this appendix we discuss the following problem, which has been motivated in Section II, remark 3. Let a similarity matrix $\{s_{ij}\}$ be given with K rows and H columns, $1 \leq i \leq K$, $1 \leq j \leq H$. The problem is the following: does a stochastic matrix Ψ_{ij} exist such that the order relation is the same in the two matrices, i.e. $s_{ij} < s_{uv}$ iff $\Psi_{ij} < \Psi_{uv}$, $s_{ij} = s_{uv}$ iff $\Psi_{ij} = \Psi_{uv}$, and consequently $s_{ij} > s_{uv}$ iff $\Psi_{ij} > \Psi_{uv}$ for all values of the indices? If the answer is yes, the similarity matrix is called *stochastic-like*. In fact, we are coping with a standard problem of linear programming: we have to understand whether the solution set for the system in the KH unknowns Ψ_{ij}

$$\Psi_{ij} \diamond \Psi_{uv} \ , \ \sum_j \Psi_{ij} = 1$$

consisting of $\binom{KH}{2} + K$ equalities/inequalities is or is not empty; above $\diamond$ stands for the required equality/inequality sign. By the way, we did not "forget" about the KH inequalities $\Psi_{ij} \geq 0$: if the "short" system as we have specified it admits of solutions, one soon checks that also the complete system does: just increment all the solutions obtained for the short system by a constant quantity so as to have non-negativity, and then normalise to obtain a stochastic matrix as required.

In the following we shall have to permute the H components of each row i of the original matrix $\{s_{ij}\}$ to obtain a new matrix $\{s^*_{ij}\}$, where $j < v$ implies $s^*_{ij} \leq s^*_{iv}$. We say that in two rows of a similarity matrix $\{s_{ij}\}$, say i and u, there is an *inversion* when the following happens: there are two columns j and v such that $s^*_{ij} < s^*_{uj}$ while $s^*_{iv} > s^*_{uv}$. If the similarity matrix $\{s_{ij}\}$ one starts with is already a stochastic matrix, in each two rows there must be at least one inversion (unless the two rows are a permutation of each other), else they could not both sum to one. Keeping this in mind, the following fact is obvious:

Necessary condition. For a similarity matrix $\{s_{ij}\}$ to be stochastic-like, there must be at least one inversion in each couple of rows, apart from couples of rows which are equal up to a permutation of their H entries.

The following counter-example shows that *this condition is not sufficient* already for three-row matrices. Take the similarity matrix

$$\begin{matrix} a & a & d & d \\ b & c & c & c \\ a & c & c & d \end{matrix}$$

with $a < b < c < d$; the three rows are already properly ordered. In rows 1 and 2 there is an inversion in positions (columns) 1 and 3, in rows 1 and 3 there is an inversion in positions 2 and 3, while in rows 2 and 3 there is an inversion in positions 1 and 4. However, the system above becomes

$$a < b < c < d \ , \ \ 2a+2d = 1 \ , \ \ b+3c = 1 \ , \ \ a+2c+d = 1 \ ,$$

whose solution set is empty: actually, the last two equations (after replacing $a + d$ by 1/2, cf. the first equation) give $b = c = 1/4$, while one should have $b < c$.

For what this can matter, the necessary condition is also sufficient when there are only two rows, $K = 2$; a proof is sketched. We can assume that the two rows are already ordered, $a_1 \leq a_2 \leq \ldots \leq a_H$, $b_1 \leq b_2 \leq \ldots \leq b_H$, $A = \sum a_i$, $B = \sum b_i$. If the two row-sums A and B are equal, just normalise; from now on we assume $A \neq B$. Our target is to transform the two rows, without changing the ordering in each row and between the two rows, in such a way that they will sum to the same number: then it will be enough to normalise. We shall make use of order-preserving continuous transformations. Since columns with $a_j = b_j$ are irrelevant to our target, we shall assume that they have been deleted. Assume $a_1 < b_1$, say, and assume i is the first index such that $a_i < b_i$ but $a_{i+1} > b_{i+1}$; one may have $b_u = b_i$ for some indices $u \leq i$ and $b_v = b_{i+1}$ for some indices $v > i + 1$; one may also have $b_i = b_{i+1}$; all these equalities must be kept in the transformations to follow. Assume that all the entries $< b_i$, e.g. a_i, have been "squeezed" into the open interval $]0, \epsilon[$, while all the entries $> b_{i+1}$, e.g. a_{i+1}, have been "squeezed" into the open interval $]\chi, \chi + \epsilon[$; take $\chi \geq H\epsilon$. Observe that in the first interval there are exactly i entries from the first row (the A-row), while in the second interval there are exactly $H - i$ entries from the A-row, and so $(H - i)\chi < A < H\epsilon + (H - i)\chi$. If $A > B$, move all the b_i outside the two intervals, e.g. b_i and b_{i+1}, into the second interval; then for the new sum B' one has $B' > (H-i+1)\chi$, and so $B' > A$ (use the upper bound on A); one just has to "stop" the continuous transformation at the "time-instant" when the second sum is exactly equal to A (out of the metaphor, $A-B$ is a continuous function of some of the B-row entries which takes on both positive and negative values: then, as elementary calculus tells, there must be a zero-point in the connected set where $A-B$ is defined). If $A < B$, move all the b_i outside the two intervals into the first interval; then for the new sum B'' one has $B'' < (i+1)\epsilon + (H - i - 1)(\chi + \epsilon)$, and so $B'' > A$ (use the lower bound on A); again, one just has to "stop" the continuous transformation at the right "time-instant".

APPENDIX C
ERASURES CHANNELS

A. Maximum likelihood decoding

One may put forward a *symmetric memoryless and stationary channel* such that decoding by maximum likelihood gives back our decoding rule for erasures channels, in much the same way as decoding by maximum likelihood on usual symmetric channels gives back decoding by minimum Hamming distance. The obvious computations are sketched below. Say η is the probability of an erasure and ϵ is the probability of a crossover; in a "real" channel both these probabilities will

be presumably small, and so we shall feel free to assume $0 < \epsilon, \eta < 1/3$. We assume for the moment that the input alphabet is binary; a generalisation to the q-ary case, $q \geq 3$, is straightforward, but does not add anything really new, cf. below. As for the likelihood $P^n(\underline{z}|\underline{x})$, i.e. the transition probability from the input word $\underline{x}$ to the output word $\underline{z}$, one has

$$P^n(\underline{z}|\underline{x}) = \eta^s \epsilon^d (1-\eta-\epsilon)^{n-s-d} \propto \Big(\frac{\eta}{1-\eta-\epsilon}\Big)^s \Big(\frac{\epsilon}{1-\eta-\epsilon}\Big)^d$$

where s is the number of erasures in $\underline{z}$, and d is the number of crossovers. On the right we have omitted constant factors and used a proportionality sign. Under our assumptions $\eta, \epsilon < 1/3$, the two powers on the right are strictly decreasing functions of their exponents. Maximising likelihooods amounts to minimising the "diversity"

$$\alpha s + \beta d,\ \alpha = \log_2 \frac{1-\eta-\epsilon}{\eta},\ \beta = \log_2 \frac{1-\eta-\epsilon}{\epsilon},\ \alpha, \beta > 0$$

or also the diversity $\rho s + d$ with $\rho = \alpha/\beta$, which is what we did in Section V. To have $\rho \leq 1$, one must have $\alpha \leq \beta$, i.e. the erasure probability η is at least as large as the crossover probability ϵ. Instead, "small" values of ρ, $\rho < \frac{1}{2}$, are obtained when the erasure probability η is "definitely larger" than the crossover probability ϵ; more precisely, as shown by elementary algebra, one must have $\eta > \eta^*$ with $\eta^* = \frac{1}{2}\big(-\epsilon + \sqrt{\epsilon^2 + 4\epsilon(1-\epsilon)}\big)$; if $\epsilon \approx 0$, $\eta^* \approx \sqrt{\epsilon}$. In general, if $q \geq 2$, we assume that our channel is fully symmetric, and so the probability of each specific crossover is $\epsilon' = \frac{\epsilon}{q-1}$. As soon checked, the diversity to minimise is the same, only one defines β by writing ϵ' rather than ϵ in the denominator; $\rho \leq 1$ for $\eta \geq \epsilon'$; $\rho < \frac{1}{2}$ for $\eta > \eta^*$ with $\eta^* = \frac{1}{2}\big(-\epsilon' + \sqrt{\epsilon'^2 + 4\epsilon'(1-\epsilon)}\big)$; if $\epsilon \approx 0$, $\eta^* \approx \sqrt{\epsilon'}$.

B. Error correction capabilities

As happens with Hamming-distance coding, one might cope with erasure channels without mentioning explicitly distinguishabilities; using them, as we do below, allows us to stress that they are a general mould for a wide variety of coding problems, and gives us the chance of introducing a slightly more general version of the reliability criterion of Section II.

Going back to abstract setting of Section II, if $\mathcal{U}$ is a subset of the output space, we can define the *constrained* or *conditional distinguishability* $\delta(a, b|z \in \mathcal{U})$ by restricting the minimisation in (1) to output words z which are constrained to belong to $\mathcal{U}$. Correspondingly, we can define the *constrained* or *conditional minimum distinguishability* $\delta_m(\mathcal{C}|z \in \mathcal{U})$ as in (2). In practice, it is as if we were replacing the original output space by a new output space (which happens to be a restriction of the old one), and so the arguments in Sections III and IV carry over to the new situation. In particular, the reliability criterion of Section II generalises to: if the received sequence belongs to $\mathcal{U}$, all diversities $\leq \tau$ are corrected iff the conditional minimum distinguishability is $> \tau$.

In the case of erasures channels, we shall consider the output spaces $\mathcal{Y}_0, \mathcal{Y}_1, \ldots, \mathcal{Y}_n$ where the sequences in $\mathcal{Y}_s$ are constrained to have exactly s stars, $0 \leq s \leq n$. If $s \leq d_H(\underline{x}, \underline{y})$, the conditional distinguishability has been already computed in Appendix A: $\Delta_s(\underline{x}, \underline{y}) = s\,\rho + \lceil \frac{d_H - s}{2} \rceil$ with $d_H = d_H(\underline{x}, \underline{y})$. If $s \geq d_H(\underline{x}, \underline{y})$, clearly $\Delta_s(\underline{x}, \underline{y}) = s\,\rho$.

Let us compute the conditional minimum distinguishabilities $\Delta_{m,s}$, obtained by minimising $\Delta_s(\underline{c}, \underline{c}')$, with $\underline{c}, \underline{c}' \in \mathcal{C}$, $\underline{c} \neq \underline{c}'$. If $s \leq d_m$, s is less than *any* Hamming distance between codewords, and so $\Delta_{m,s} = s\,\rho + \lceil \frac{d_m - s}{2} \rceil$ (recall that d_m is the minimum Hamming distance between codewords). If instead $s \geq d_m$, some codewords (at least two, those which yield d_m, and possibly all of them) have distinguishability $s\,\rho$, while some other codewords, those, if any, at Hamming distance $> s$, have a higher distinguishability; so $\Delta_{m,s} = s\,\rho$. In general, using a binary logical value in angular brackets:

$$\Delta_{m,s} = s\,\rho + \lceil \frac{d_m - s}{2} \rceil \langle s \leq d_m \rangle$$

Assuming $s < d_m$, the generalised criterion becomes (cf also the corollary 1, Section II): if the received sequence $\underline{z}$ has exactly s erasures with $s < d_m$, the decoder corrects all the s erasures and up to $t = \lceil \frac{d_m - s}{2} \rceil - 1$ crossovers. If instead $s \geq d_m$ there are noisy codewords which will be incorrectly decoded, even if there are no crossovers at all; just think of the two codewords which yield the minimum distance, when all the positions where they differ are erased during transmission.

As an example, consider a famous binary code construction, the Hadamard low-rate code with $n = 32$ and $d_m = 16$ used by Mariner 1969; cf. [9]. Each column gives the number s of erasures and the number t of crossovers whose proper correction is ensured:

s	0	1	2	3	4	5	6	7	8
t	7	7	6	6	5	5	4	4	3
s	9	10	11	12	13	14	15		
t	3	2	2	1	1	0	0		

REFERENCES

[1] C.E. Shannon "The Zero-Error Capacity of a Noisy Channel" *IRE Transactions on Information Theory*, Vol. **2**, pp 8-19, 1956.

[2] J. Körner, A. Orlitsky, "Zero-error Information Theory" *IEEE Transactions on Information Theory*, Vol. **44.6**, pp 2207-2229, 1998.

[3] A. Brenneman, A. Condon, Strand Design for Bio-Molecular Computation, (Survey Paper), *Theoretical Computer Science*, Vol. **287:1**, pp 39-58, 2002.

[4] G. Mauri, C. Ferretti, Word Design for Molecular Computing: A Survey, 9th International Workshop on DNA Based Computers, *DNA 2003* 37-46 (electronic edition), 2003.

[5] A. Condon, R.M. Corn, A. Marathe, On Combinatorial DNA Word Design, *J. Computational Biology*, Vol. **8:3**, pp. 201-220, 2001.

[6] A. Sgarro, "Possibilistic Information Theory: a Coding-theoretic Approach", *Fuzzy Sets and Systems*, Vol. **132-1**, pp. 11-32, 2002.

[7] T.M. Cover, J.A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).

[8] I. Csiszár, J. Körner, *Information Theory* (Academic Press, New York, 1981).

[9] J. van Lint, *Introduction to Coding Theory* (Springer Verlag, Berlin, 1999).

[10] D. Gusfield, *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology* (Cambridge University Press, NY, 1997).

[11] R.J. Blom, "Bounds on Key Equivocation for Simple Substitution Ciphers" *IEEE Transactions on Information Theory*, Vol. **25.1**, pp 8-18, 1979.

[12] A. Sgarro “Error Probabilities for Simple Substitution Ciphers” *IEEE Transactions on Information Theory*, Vol. **30.2**, pp 190-198, 1983.
[13] G.T. Bogdanova, A.E. Brouwer, St.N. Kapralov, P.R.J. Östergård. Error-correcting codes over an alphabet of four elements. *Designs, Codes and Cryptography*, Vol. **23(3)**, pp 333-342, 2001.
[14] M. Garzon, R. Deaton, P. Neathery. D.R. Franceschetti, R.C. Murphy, A New Metric for DNA Computing, *Proc. Second Genetic Programming Conf.*, Morgan-Kaufmann, Stanford. CA, pp. 472-478, 1997.