Possibilistic Coding: Error Detection vs. Error Correction

Luca Bortolussi and Andrea Sgarro

Abstract. Possibilistic information theory is a flexible approach to old and new forms of coding; it is based on possibilities and patterns, rather than pointwise probabilities and traditional statistics. Here we fill up a gap of the possibilistic approach, and extend it to the case of error detection, while so far only error correction had been considered.

1 Introduction

The possibilistic approach to source and channel coding (to compression codes and error-correcting codes) arose as a formal game in which pointwise probabilities, as currently used in Shannon's information theory, were replaced by possibilities, so as to find a possibilistic equivalent for probabilistic notions as are error probability, source entropy and channel capacity, cf. [14, 15]. The formal game has proved to be more stimulating than was expected: the possibilistic approach could be applied to the design of error-correcting phone keyboards [11], to more theoretic questions like defining the nature of channel noise in biological computation [1, 2], or introducing adequate "operational" information measures [8, 13, 15], and also, more recently, to the construction of codes which correct *twiddles*, i.e. transpositions between consecutive letters, inadvertently made [3]. It turns out that possibilistic information theory includes as sub-cases both Shannon's zero-error information theory [9] and the standard approach to error correcting codes [10] based on checking Hamming distances between codewords. As for the first inclusion, suffice it to say that the possibilistic approach may be seen as a multi-step generalisation of

Luca Bortolussi and Andrea Sgarro
Dept. of Mathematics and Informatics, University of Trieste, Italy

Luca Bortolussi and Andrea Sgarro
Centre for Biomolecular Medicine, Area Science Park, Trieste, Italy
e-mail: luca@dmi.units.it, sgarro@units.it

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 41–48.
springerlink.com

Shannon's approach, which is two-step only, without intermediate degrees of possibility between possible and impossible. As for the second inclusion, the probabilistic layer (or at least the traditional pointwise probabilistic layer) has always been felt to be rather thin, so that resorting to probabilistic symmetric channels, as done in introductory textbooks, e.g. [10], might appear to be a homage[1] paid to Shannon, rather than an intrinsic need of coding theory (as opposed to information theory proper, or Shannon theory).

The possibilistic approach to coding is rigorously Shannon-theoretic: based as it is on patterns rather than traditional statistics, it responds to a general need of new approaches to information theory felt in the computer science community [4]. A basic Shannon-theoretic notion like *channel noise* can be safely exported to the new setting, as we did in [1, 2] in the case of DNA word design (codewords are DNA strings, cf. [5]). One might want to mimic also in the DNA case what one successfully does for standard codes (the noise of symmetric channels is probabilistic), but we have proven in [2] that nothing like this holds in the DNA case, where *no* probabilistic description of channel noise is feasible. By this we have been able to give a remarkable example where channel noise is *intrinsically possibilistic and non-probabilistic*. Clearly, the probability which is ruled out here is *pointwise* probability: the interpretation of possibilities as *upper probabilities* suggests instead the feasibility of a probabilistic approach to information theory and coding which might prove to be quite comprehensive, even if its impact on coding practice remains to be assessed.

In this paper we emend a fault of the possibilistic approach, which appears to be able to deal only with *error correction* and not with *error detection*: so, at least seemingly, it has a weak point with respect not only to standard codes but also to DNA word design as covered in the literature. Below we give a solution to the problem of possibilistic error detection which is fully general, and which is based on the notion of *even codeword couples*, to be defined in Section 4; Sections 2 and 3 introduce our problem and ensure self-readability, while the problem itself is tackled in the final Section 4.

2 Distinguishability and Confusability

Let $\mathscr{X}$ be a finite metric space and let $d(x,y)$ be the corresponding metric distance. The idea is that an element $x \in \mathscr{X}$ is fed to a *transmission channel* and at the other end of the channel an element $z \in \mathscr{X}$ is observed, which might be different from x due to *channel noise*. The aim is to recover the correct input x from the observed output z. A *codebook* $\mathscr{C}$, or for short a

[1] Shannon's original approach is probabilistic even in the zero-error case: for him *possible* means that the probability is positive, however small it may be. Note that the notions of codeword distinguishability and codeword confusability are already present in the zero-error theory, even if not in the general form as below, Section 2, and so are due to Shannon.

code, is simply a non-void subset of elements called *codewords*, to be used as possible inputs to the channel.

Once we have a distance on $\mathscr{X}$ with maximal value N, a corresponding transition possibility from x to z can be obtained in a "canonical" way:

$$\mathrm{Poss}(z|x) \;=\; 1 - N^{-1} d(x,z) \tag{1}$$

These transition possibilities can be arranged into an $|\mathscr{X}| \times |\mathscr{X}|$ *possibility matrix*: the entries in each row of a possibility matrix, rather than summing up to 1 as in a stochastic matrix, have a 1 as their maximum, as typical of possibility theory, which is maxitive rather than additive (for possibility theory cf. e.g. [7]). In channel coding, a transition possibility as (1) can be interpreted as follows: the possibility of receiving z when x is sent over the noisy channel is high or low according whether the "pattern similarity" between input x and output z is high or low. We stress that a *possibilistic noisy channel* is completely described by a possibility matrix. The distance-based (geometric) approach will be more palatable to coding theorists, but an explicit use of possibilities[2] has the advantage of better emphasising the links with information theory, on the base of the opposition probability vs. possibility. Even if overlooked in the literature on standard coding and DNA word design, and this for reasons explained below, basic coding-theoretic notions are codeword distinguishability or, equivalently, codeword confusability.

Definition 1. *The distinguishability of a couple* (x,y) *is defined to be*

$$\delta(x,y) \;=\; \min_{z \in \mathscr{X}} \max\{d(x,z), d(y,z)\}$$

From now on, unless otherwise specified, distances are assumed to be consecutive integers, $d(x,y) \in \{0,1,\ldots,N\}$. One soon proves the following[3] bounds:

$$\left\lceil \frac{d(x,y)}{2} \right\rceil \;\leq\; \delta(x,y) \;\leq\; d(x,y) \tag{2}$$

In a possibilistic setting one might prefer to deal with *confusabilities*

$$\gamma(x,y) = \max_{z \in \mathscr{X}} \min\{\mathrm{Poss}(z|x), \mathrm{Poss}(z|y)\} = 1 - N^{-1}\,\delta(x,y) \tag{3}$$

The rightmost equality assumes (1). Nothing much changes, and so in the following we shall stick to distinguishabilities. A situation when the lower bound in (2) is achieved is the following:

[2] The possibilistic framework can be readily and naturally enlarged to more general situations, e.g. when the input alphabet and the output alphabet are distinct, cf. [12, 14]: then dissimilarities (which take up the role of distances) are between input and output, while distinguishabilities involve two inputs.

[3] The lower bound follows from the triangle inequality; if the distances are not constrained to be integers the integer ceiling must be understood as the smallest available distance which is $\geq d(x,y)/2$, c.f. [12].

Definition 2. *An integer metric space is* dense *when, whatever the couple* (x,y)*, for any integer* $m \in [0, d(x,y)]$ *one can find an element* z *at distance* m *from* x *and at distance* $d(x,y) - m$ *from* y*.*

An obvious example is given by Hamming distances for strings of the same length, a less obvious example is DNA word design, cf. [1, 2]. Instead, the upper bound in (2) is always achieved if and only if the (not necessarily integer) metric is an *ultrametric*, i.e. if and only if the fuzzy triangle inequality $\max\{d(x,z), d(z,y)\} \geq d(x,y)$ is always verified, cf. [12]. Cf. [3] for the case of significant string distances, e.g. variants of the edit distance or Spearman footrule [6], which might be used to correct twiddles, as hinted at in Section 1; the distinguishabilities corresponding to these distances are sometimes equal to the lower bound, sometimes to the upper bound, and sometimes have intermediate values, depending on (x,y), cf. [3].

3 Two Equivalent Approaches to Coding

For the moment being we deal only with error-correcting codes and so ignore error detection; the definition below might be equivalently given in term of confusabilities (3).

Definition 3. Optimal codes*: once the integer threshold* Δ *is chosen, construct maximum-size codes with* guaranteed minimum distinguishability Δ*, i.e. with* $\delta(x,y) \geq \Delta$ *for all couples of distinct codewords in* $\mathscr{C}$ *(*$0 < \Delta \leq \max_{x,y} \delta(x,y) \leq N$*).*

Whatever the code size, when threshold Δ is guaranteed one proves the following reliability criterion, given in two equivalent phrasings, cf. [12, 14]. To enhance self-readability, a quick proof is given; for more details cf. [12, 14].

Reliability criterion 1. *Decode to a codeword* x *which maximises the transition possibility* $\mathrm{Poss}(z|x)$ *to the output* z*: the error possibility*[4] *is at most equal to* $1 - \Delta/N$*.*

Reliability criterion 2. *Once the output string* z *is received, decode to a codeword* x *which minimises the distance* $d(x,z)$ *between input and output: if the input string* x *was such that* $d(x,z) < \Delta$*, decoding is successful.*

Proof. If x is sent, z is received, and $y \neq x$ is decoded to, then $d(y,z) \leq d(x,z)$ and so, by definition 1, $\delta(x,y) \leq d(x,z)$: by comparison, $d(x,z) \geq \Delta$. □

If two codewords x and y have distinguishability $\delta(x,y) = \Delta$, one can provide an output z at distance $d(x,z) = \Delta$ from x and at distance $d(y,z) \leq \Delta$ from y,

[4] For each codeword y sent over the channel, its error possibility is the possibility of the set of the outputs z which lead to a decoding error, and so, according to the maxitive rules of possibility theory, it is the maximum possibility $\mathrm{Poss}(z|y)$ of such z's.

or the other way round, which will bring about a decoding error of "weight" Δ and of possibility $1-\Delta/N$: in this sense, the Reliability criterion cannot be improved.

Actually, optimal codes of both standard coding and DNA word design are constructed by choosing a threshold T and checking directly distances rather than distinguishabilities. In general, ignoring distinguishabilities can lead to inconsistent results, in the sense that the resulting codebooks are nice combinatorial constructions devoid of error-correcting capabilities, cf. [1]. This is not so in the standard case or in the DNA case, because standard and DNA distinguishabilities are a *monotone*[5] function of the corresponding distances (recall that the lower bound (2) is always achieved in the case of dense spaces, Definition 2). When monotonicity is strict, everything is fine: here, however, monotonicity is only weak, and so the reader will object that one ends up "losing" all optimal codebooks which had been obtained by constraining the minimum distance $d(x,y)$ against an *even* integer threshold T. As a matter of fact, the Reliability criteria soon imply that even bounds on distances are completely useless if one insists on *hard* decoding (the decoder decides to a single codeword, however fishy the situation might be). Instead, even bounds on distances are quite relevant in *error detection*, when a *soft* decoder is used; in the next section we show how the possibilistic approach can deal with error detection quite in general.

4 Even Couples in Error Detection

Definition 4. *The couple* (x,y) *is an* even couple *when any* z *achieving* $\delta(x,y)$ *as in Definition 1 is at the same distance from both* x *and* y*, it is an* odd couple *if for any such* z *the two distances from* x *and* y *are distinct, else it is a* mixed couple.

Snags with error detection occur with mixed couples, i.e. when one can provide a *skew quadruple* (x,y,u,w) where u and w both achieve $\delta(x,y)$, but $\delta(x,y)=d(x,w)=d(w,y)$ while $\delta(x,y)=d(u,y)>d(u,x)$. Set $d(x,y)=d$, $\delta(x,y)=\rho$, $d(x,u)=\mu$, $d(u,w)=\xi$.

Lemma 1. *Four positive real numbers* d,ρ,μ,ξ *as above are the lengths of a skew quadruple in a metric space of size 4 if and only if they verify the constraints* $\lceil d/2\rceil \le \rho \le d$, $\rho \ne d/2$, $d-\rho \le \mu < \rho$, $\rho-\mu \le \xi \le \rho+\mu$.

Proof. Choose d; as for ρ the bounds (2) must hold. Forget ξ for the moment being: one is left with two triangles, and a check of the corresponding triangle inequalities gives $d-\rho \le \mu \le d+\rho$, but since μ should be strictly smaller

[5] The reader will have appreciated that, thinking of optimal codes and reliability, the possibilistic approach is basically invariant with respect to strictly monotone transformations of the transition possibilities involved, or of the distances involved.

than ρ one ends up imposing $d-\rho \leq \mu < \rho$. To avoid that this interval be void one must rule out the value $d/2$ for ρ. Adding ξ gives two more triangles, and a check of the corresponding triangle inequalities completes the proof. □

The proof does not assume that the distances should be integers; if it is so, the constraints on ρ can be subsumed to $\lceil (d+1)/2 \rceil < \rho \leq d$. The lemma allows one, after choosing d, to find ρ, μ, ξ in this order; it can be used in spaces of any size to spot mixed couples. E.g. take $d=2$ to find the three integer solutions $\rho = 2$, $\mu = 1$, $\xi \in \{1,2,3\}$. Of these, the third gives the distance matrix below; transition possibilities are soon obtained from (1) setting $N=3$; distinguishabilities are equal to distances, as soon checked, save $\delta(u,w) = 3 = \lceil d(u,w)/2 \rceil$:

	x	y	u	w
x	0	2	1	2
y	2	0	2	2
u	1	2	0	3
w	2	2	3	0

Theorem 1. *If x and y achieve the lower bound* (2) *and $d(x,y)$ is an integer, the couple (x,y) is even or odd according whether their distance $d(x,y)$ is even or odd, respectively. If x and y achieve the upper bound* (2)*, the couple (x,y) cannot be even.*

Proof. The first claim in the theorem below is a straightforward by-product of the lemma, which rules out mixed couples for $d = \delta/2$; the rest soon follows from the triangle inequality. As for the second claim just think that x and y are both minimizing z's as in definition 1. □

So, the set of even couples (x,y) is made up of *all* the couples at an even distance in the dense case, but in general it is not so. E.g., in spite of even distances, there are no even couples with $x \neq y$ for the distance matrix above, as soon checked (out of the six couples two are odd, four are mixed).
We modify optimality and reliability as follows; we assume $\Delta < N$ to no loss of generality (use theorem 1: if $\delta(x,y) = N$ then the upper bound in (2) is achieved and so (x,y) cannot be even).

Definition 5. Optimal codes for error detection*: once the integer threshold Δ is chosen, construct maximum-size codes as in definition 3, but adding the constraint: if $d(x,y) = \Delta$ then (x,y) is an even couple.*

Reliability criterion 3 for error detection. *Decode to the single codeword x which maximises the transition possibility* $\mathrm{Poss}(z|x)$ *to the output z, and in case of ties declare a detected error: the undetected error possibility is strictly less than $1-\Delta/N$.*

Reliability criterion 4 for error detection. *Once the output string z is received, decode to the single codeword x which minimises the distance $d(x,z)$*

between input and output; in case of ties declare instead a detected error. No undetected error occurs if the input string x was such that $d(x,z) \leq \Delta$.

Proof. Re-take the proof of the criterion for error correction. The equality $d(y,z) \leq d(x,z)$ must be modified to a strict inequality $d(y,z) < d(x,z)$, else a detected error would have been declared. Once more one gets $\Delta \leq \delta(x,y) \leq d(x,z)$; however, if $\Delta = d(x,z)$ and so $\Delta = \delta(x,y) = d(x,z)$, z would achieve $\delta(x,y)$ and (x,y) would be an odd or mixed couple. □

Criteria 3 and 4 do not require that the codes are maximum-size (optimal). In practice, whenever the lower bound (2) holds, the Reliability criterion 4 can be re-stated in a way which is quite familiar to coding-theorists, just use Theorem 1 (requiring that a couple (x,y) is even amounts to requiring that its distance should be an even integer). At the other end of the spectrum, we have ultrametric spaces, where error detection does not offer any advantage, and so can be safely ignored, use again Theorem 1. As for the intermediate and stimulating case of the string distances for twiddles mentioned in Section 2, to construct error-detecting codes one will have to carefully understand the structure of even couples in the corresponding string "geometry".

The gap of error detection having been filled, possibilistic information theory stands out as a full-fledged approach to information theory and coding, able to deal with situations, as is channel noise in DNA word design or the correction of twiddles, where the traditional probabilistic and distance-based approach falls short of the mark.

Acknowledgements. Work partially supported by FIRB-RBLA039M7M_005 LIBI Project on Bioinformatics.

References

1. Bortolussi, L., Sgarro, A.: Possibilistic channels for DNA word design. In: Lawry, J., Miranda, E., Bugarin, A., Li, S., Gil, M.A., Grzegorzewski, P., Hryniewicz, O. (eds.) Soft Methods for Integrated Uncertainty Modelling. Advances in Soft Computing, vol. 37, pp. 327–335. Springer, Heidelberg (2006)
2. Bortolussi, L., Sgarro, A.: Noise of DNA word design is not stochastic (submitted for publication, 2010), `www.dmi.units.it/~sgarro/nostochasticDNA.pdf`
3. Bortolussi, L., Dinu, L.P., Sgarro, A.: Twiddle correction and codeword distinguishability (in preparation, 2010), Preliminary version at: `www.dmi.units.it/~sgarro/rankCODES.pdf`
4. Brooks Jr., F.P.: Three great challenges for half-century-old computer science. J. ACM. 50(1), 25–26 (2003)
5. Condon, A., Corn, R.M., Marathe, A.: On combinatorial dna word design. J. Comput. Biol. 8(3), 201–220 (2001)
6. Deza, E., Deza, M.M.: Dictionary of Distances. Elsevier, Amsterdam (2006)
7. Dubois, D., Prade, H.: Fundamentals of Fuzzy Sets. The Handbooks of Fuzzy Sets Series. Kluwer Academic Publishers, Dordrecht (2000)

8. Guiaşu, S.: Comments on: "On possibilistic entropies" by Sgarro, A., Dinu, L.P. Internat. J. Uncertain. Fuzziness Knowledge-Based Systems 10(6), 655–657 (2002)
9. Körner, J., Orlitsky, A.: Zero-error information theory. IEEE Trans. Inform. Theory 44(6), 2207–2229 (1998)
10. van Lint, J.: Introduction to Coding Theory. Springer, Berlin (1999)
11. Luccio, F., Sgarro, A.: Fuzzy graphs and error-proof keyboards. In: Proceedings of the 9th International Conference on Information Processing and Management of Uncertainty in Knowledge Based Systems, IPMU 2002, Annecy, France, pp. 1503–1508 (2002)
12. Sgarro, A., Bortolussi, L.: Codeword distinguishability in minimum diversity decoding. J. Discrete Math. Sci. Cryptogr. 9(3), 487–502 (2006)
13. Sgarro, A., Dinu, L.P.: Possibilistic entropies and the compression of possibilistic data. Internat. J. Uncertain. Fuzziness Knowledge-Based Systems 10(6), 635–653 (2002)
14. Sgarro, A.: Possibilistic information theory: a coding-theoretic approach. Fuzzy Sets Syst. 132(1), 11–32 (2002)
15. Sgarro, A.: An axiomatic derivation of the coding-theoretic possibilistic entropy. Fuzzy Sets Syst. 143, 335–353 (2003)