Manuscript submitted to
AIMS' Journals
Volume **XXX**, Number **0XXX**, XXX **2007** Website: http://AIMsciences.org
pp. **XXX–XXX**

# THE LZ-DISTANCE SATISFIES THE METRIC AXIOMS

M. Christen, T. Ott and R. Stoop

Institute of Neuroinformatics
University / ETH Zürich
8052 Zürich, Switzerland

(Communicated by ???)

Abstract. Since the principles of biological information transfer have remained elusive, for the comparison of spike-trains, several concurring measures are in use. It is, however, rarely verified whether any of these measures satisfy the axioms of a metric. In this contribution, we prove that a recent Lempel-Ziv-complexity based spike-train distance measure indeed satisfies the axioms of a metric.

## 1. Introduction.

Biological neurons communicate by means of electrical pulses, called spikes. The exact way of how information is encoded in the emitted sequences of spikes called spike trains, however, has remained elusive. It has become clear, however, that the form of a spike train reflects the nature of the information to be transmitted, and the conditions under which it has been generated, i.e. the architecture of the involved neuronal circuitry alike.

The discipline of the analysis of the nature and the functions of the spike train signals recorded from ensembles of neurons is usually termed spike train analysis. One major focus of spike train analysis is the classification of a whole ensemble of neurons (usually recorded by means of an array of electrodes) into classes of similar firing, which may provide information about the functional connectivity of a probed neuronal network. To solve this problem, a variety of distance measures [1, 5, 6, 8, 10, 11, 12, 14, 15] have been used, usually in combination with some clustering procedure. The solutions offered, however, generally suffer from very basic shortcomings: First, it is unknown which of these measures should be considered relevant (since the nature of the neuronal information is still unknown). Second, and related to the first point, most of them introduce a strong bias in the form of predefined analysis parameters. Lastly, it is usually not shown whether the measures used are a metrics in the mathematical sense. This is not only dissatisfying from a theoretical point of view. It might also have practical consequences, as algorithmically, the triangle inequality – which is the essential step to be mastered in such a proof – is the key tool for the clustering of very similar data into subsets [2].

The optimal spike-train analysis therefore consists of two steps that should be as much independendent from any form of predefined notions of encoding and transmission of information among neurons: An unbiased measure of similarity, and an

2000 *Mathematics Subject Classification.* 94A17.
*Key words and phrases.* Information theory, spike train distance measure, metric axioms.

unbiased clustering procedure. Whereas for the first problem, we have already arrived at a satisfactory solution [**?**, **?**], for the second problem we have only recently proposed a spike train distance measure that is entirely based on fundamental notions of information theory. Our Lempel-Ziv-complexity [7] distance measure [3] does not require the choice of arbitrary analysis parameters, is easy to implement, and computationally cheap. The Lempel-Ziv-distance (LZ-distance) considers spike trains with similar but possibly delayed firing patterns as close and is therefore considerably noise-robust – which are important aspects to be taken care of when applying distance measures to biological data. In this contribution, we show that the LZ-distance also satisfies the axioms of a metric.

2. **The Lempel-Zif-Distance.** For our analysis, spike trains given as sequences of neuronal spike-times $\mathrm{t} = \{t_1, \dots, t_n\}$ are translated into bitstrings. For this translation, the time interval $[0, T]$ accross which the measurement is taken is partitioned into $n$ bins of width $\Delta\tau$ ($n\Delta\tau = T$). If at least one spike falls into the $i$-th bin, the letter "1" (and otherwise the letter "0") is written to the $i$-th position of the string. Usually, $\Delta\tau$ is chosen so that maximally one spike falls into one bin. This is achieved by setting $\Delta\tau \sim 1$ ms, because of the neuronal refractory period. The resulting bitstring is denoted by $\mathsf{X}_n$. A substring starting at position $i$ and ending at position $j$ will be denoted by $\mathsf{X}_n(i, j)$. Such bitstrings can be viewed as generated by an information source. For this source, we want to find the optimal coding [4, 13]. Such a coding is provided by a parsing that partitions the string into non-overlapping substrings called *phrases*. The set of phrases [4] that result from a parsing of a bitstring $\mathsf{X}_n$ is denoted by $\mathsf{P}_{\mathsf{X}_n}$. We will use a coding that sequentially parses the original string so that the new phrase is not yet contained in the set of phrases generated so far [16]. This henceforth termed *LZ-coding* can be defined as follows:

Definition 2.1. *Let $c(\mathsf{X}_n)$ denote the number of phrases that results from the LZ-coding of a file $\mathsf{X}_n$. For a bitstring $\mathsf{X}_n$, the* Lempel-Ziv-complexity[16] $K(\mathsf{X}_n)$ *of $\mathsf{X}_n$ is defined as*

$$K(\mathsf{X}_n) = \frac{c(\mathsf{X}_n) \log c(\mathsf{X}_n)}{n}.$$

To explain the LZ-distance, we start from two strings $\mathsf{X}_n$, $\mathsf{Y}_n$ of equal length $n$. From the perspective of LZ-complexity, the amount of information $\mathsf{Y}_n$ provides about $\mathsf{X}_n$ is given as $K(\mathsf{X}_n) - K(\mathsf{X}_n|\mathsf{Y}_n)$, where $c(\mathsf{X}_n|\mathsf{Y}_n)$ is the size of the difference set $\mathsf{P}_{\mathsf{X}_n} \setminus \mathsf{P}_{\mathsf{Y}_n}$. If $\mathsf{Y}_n$ provides no information about $\mathsf{X}_n$, then the sets $\mathsf{P}_{\mathsf{X}_n}$ and $\mathsf{P}_{\mathsf{Y}_n}$ are disjoint, and $K(\mathsf{X}_n) - K(\mathsf{X}_n|\mathsf{Y}_n) = 0$. If $\mathsf{Y}_n$ provides complete information about $\mathsf{X}_n$, then $\mathsf{P}_{\mathsf{X}_n} \setminus \mathsf{P}_{\mathsf{Y}_n} = \emptyset$ and $K(\mathsf{X}_n) - K(\mathsf{X}_n|\mathsf{Y}_n) = K(\mathsf{X}_n)$. The LZ-complexity approximates the Kolmogorov complexity $K_K(\mathsf{X}_n)$ of a bitstring and a theorem of the complexity theory by Kolmogorov implies that $K_K(\mathsf{X}_n) - K_K(\mathsf{X}_n|\mathsf{Y}_n) \approx K_K(\mathsf{Y}_n) - K_K(\mathsf{Y}_n|\mathsf{X}_n)$ [9]. In practical applications, where we deal with bitstrings of finite length, this symmetry does not hold. Therefore, we need to calculate $K(\mathsf{X}_n) - K(\mathsf{X}_n|\mathsf{Y}_n)/K(\mathsf{X}_n)$ as well as $K(\mathsf{Y}_n) - K(\mathsf{Y}_n|\mathsf{X}_n)/K(\mathsf{Y}_n)$, where we take the minimum in order to ensure $d(\mathsf{X}_n, \mathsf{X}_m) > 0$ for $n \neq m$. Furthermore, the expression $K(\mathsf{X}_n) - K(\mathsf{X}_n|\mathsf{Y}_n)$ is normalized by $K(\mathsf{X}_n)$ (and by $K(\mathsf{Y}_n)$, respectively) so that the distance $d(\mathsf{X}_n, \mathsf{Y}_n)$ ranges between 0 and 1. This leads to the following definition of the *LZ-distance*:

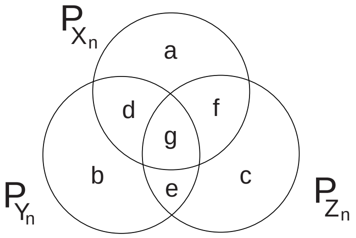

FIGURE 1. Coding used for the subsets for set of phrases obtained for the strings $\mathsf{X}_n$, $\mathsf{Y}_n$, $\mathsf{Z}_n$.

DEFINITION 2.2. *For two bitstrings of equal length $\mathsf{X}_n$ and $\mathsf{Y}_n$, the Lempel-Ziv distance* [3] *$d(\mathsf{X}_n, \mathsf{Y}_n)$ is defined by*

$$d(\mathsf{X}_n, \mathsf{Y}_n) = 1 - \min\left\{\frac{K(\mathsf{X}_n) - K(\mathsf{X}_n|\mathsf{Y}_n)}{K(\mathsf{X}_n)}, \frac{K(\mathsf{Y}_n) - K(\mathsf{Y}_n|\mathsf{X}_n)}{K(\mathsf{Y}_n)}\right\}.$$

3. **The LZ-distance is a metric.** We now prove that the LZ-distance is a metric. I.e., for strings $\mathsf{X}_n$, $\mathsf{Y}_n$, $\mathsf{Z}_n$, $d(\cdot,\cdot)$ satisfies the following axioms:

$$\begin{array}{rc} \text{a)} & d(\mathsf{X}_n,\mathsf{Y}_n) > 0 \quad \text{for} \quad \mathsf{X}_n \neq \mathsf{Y}_n \\ \text{b)} & d(\mathsf{X}_n,\mathsf{X}_n) = 0 \\ \text{c)} & d(\mathsf{X}_n,\mathsf{Y}_n) = d(\mathsf{Y}_n,\mathsf{X}_n) \\ \text{d)} & d(\mathsf{X}_n,\mathsf{Y}_n) \leq d(\mathsf{X}_n,\mathsf{Z}_n) + d(\mathsf{Z}_n,\mathsf{Y}_n) \end{array}$$

For the proof, we will need two Lemma, and, in order to simplify the arguments, the following general shorthand coding for the set of phrases $\mathsf{P}_{\mathsf{X}_n}$, $\mathsf{P}_{\mathsf{Y}_n}$, $\mathsf{P}_{\mathsf{Z}_n}$, resulting from the parsing of the strings $\mathsf{X}_n$, $\mathsf{Y}_n$, $\mathsf{Z}_n$ (see Fig. 1):

$$\begin{array}{rrr} c(\mathsf{X}_n) = a+d+f+g, & c(\mathsf{X}_n|\mathsf{Y}_n) = a+f, & c(\mathsf{Y}_n|\mathsf{X}_n) = b+e \\ c(\mathsf{Y}_n) = b+d+e+g, & c(\mathsf{X}_n|\mathsf{Z}_n) = a+d, & c(\mathsf{Y}_n|\mathsf{Z}_n) = b+d \\ c(\mathsf{Z}_n) = c+e+f+g, & c(\mathsf{Z}_n|\mathsf{Y}_n) = c+f, & c(\mathsf{Z}_n|\mathsf{X}_n) = c+e. \end{array}$$

LEMMA 3.1. *Consider the function $f(x) = x\log x$ and $x \in \mathbb{N}_0$. It follows:*

$$f(x) \leq f(y) + f(z), \quad y \wedge z \neq 0 \quad \Rightarrow \quad x \leq y+z.$$

*Proof:* Proof by contradiction. We assume $x > y+z$ and transform this equation as follows:

$$x^x > x^{y+z} = x^y x^z \overset{x>y,x>z}{>} y^y z^z.$$

But this would be inconsistent with the following transformation of the left hand side of the lemma, where we use the fact that $f(x)$ is monotonically increasing in $\mathbb{N}_0$:

$$x\log x \leq y\log y + z\log z \Rightarrow e^{x\log x} \leq e^{y\log y + z\log z} \Rightarrow x^x \leq y^y z^z.$$

Thus, the lemma is correct. Observe that the lemma does not hold on $\mathbb{R}_0^+$.

LEMMA 3.2. *Assume that $K(\mathsf{X}_n) \geq K(\mathsf{Y}_n)$. Then it follows that*

$$\frac{K(\mathsf{X}_n) - K(\mathsf{X}_n|\mathsf{Y}_n)}{K(\mathsf{X}_n)} \leq \frac{K(\mathsf{Y}_n) - K(\mathsf{Y}_n|\mathsf{X}_n)}{K(\mathsf{Y}_n)}.$$

*Proof:* We define $x_1 := a+f$, $x_2 := b+e$ and $x_3 := d+g$. Using Lemma 3.1, we obtain

$$K(\mathsf{X}_n) \geq K(\mathsf{Y}_n) \Rightarrow x_1 + x_3 \geq x_2 + x_3 \Rightarrow x_1 \geq x_x.$$

Using the assumption of the lemma and the fact, that $K(X_n)$ is monotonically increasing in $\mathbb{N}_0$, we transform the righthand side to

$$\begin{aligned} \tfrac{K(\mathsf{X}_n)-K(\mathsf{X}_n|\mathsf{Y}_n)}{K(\mathsf{X}_n)} &\leq \tfrac{K(\mathsf{Y}_n)-K(\mathsf{Y}_n|\mathsf{X}_n)}{K(\mathsf{Y}_n)}, \\ \tfrac{K(\mathsf{X}_n|\mathsf{Y}_n)}{K(\mathsf{X}_n)} &\geq \tfrac{K(\mathsf{Y}_n|\mathsf{X}_n)}{K(\mathsf{Y}_n)}, \\ K(\mathsf{X}_n|\mathsf{Y}_n) &\geq K(\mathsf{Y}_n|\mathsf{X}_n), \\ x_1 &\geq x_2. \end{aligned}$$

That axioms a), b) and c) of a metric are fulfilled follows straightforwardly from Def. 2.2. It remains to prove that the triangle inequality d) holds as well. To this end, we show that we can insert Def. 2.2 into axiom d) and then transform the inequality so that it will be true for all possible choices for $\mathsf{X}_n, \mathsf{Y}_n$ and $\mathsf{Z}_n$. Without loss of generality, we may assume $n > 0$, $c(\mathsf{X}_n) \geq c(\mathsf{Y}_n)$, and hence (as $K(X_n)$ is monotonically increasing in $\mathbb{N}_0$), $K(\mathsf{X}_n) \geq K(\mathsf{Y}_n)$. The following three cases cover all possible relations of $c(\mathsf{Z}_n)$ with $c(\mathsf{X}_n)$ and with $c(\mathsf{Y}_n)$: I) $c(\mathsf{Z}) \geq c(\mathsf{X})$, II) $c(\mathsf{X}) \geq c(\mathsf{Z}) \geq c(\mathsf{Y})$ and III) $c(\mathsf{Y}) \geq c(\mathsf{Z})$. We will consider each case separately.

*Proof of case I:* We have to prove that $d(\mathsf{X}_n, \mathsf{Y}_n) \leq d(\mathsf{X}_n, \mathsf{Z}_n) + d(\mathsf{Z}_n, \mathsf{Y}_n)$ under the assumption $c(\mathsf{Z}_n) \geq c(\mathsf{X}_n)$. Using Def. 2.2 and Lemma 3.2, we transform the triangle inequality into

$$K(\mathsf{Z}_n)\frac{K(\mathsf{X}_n|\mathsf{Y}_n)}{K(\mathsf{X}_n)} \leq K(\mathsf{Z}_n|\mathsf{X}_n) + K(\mathsf{Z}_n|\mathsf{Y}_n),$$

where $0 \leq \frac{K(\mathsf{X}_n|\mathsf{Y}_n)}{K(\mathsf{X}_n)} \leq 1$, because $0 \leq c(\mathsf{X}_n|\mathsf{Y}_n) \leq c(\mathsf{X}_n)$. We now assume the most difficult extral case $\frac{K(\mathsf{X}_n|\mathsf{Y}_n)}{K(\mathsf{X}_n)} = 1$ and apply Lemma 3.1, which gives us

$$c(\mathsf{Z}_n) \leq c(\mathsf{Z}_n|\mathsf{X}_n) + c(\mathsf{Z}_n|\mathsf{Y}_n).$$

Using our shorthand-coding, this translates into

$$\begin{aligned} c+e+f+g &\leq (c+e)+(c+f), \\ g &\leq c, \end{aligned}$$

which is correct under the assumption $\frac{K(\mathsf{X}_n|\mathsf{Y}_n)}{K(\mathsf{X}_n)} = 1$, because from $K(\mathsf{X}_n|\mathsf{Y}_n) = K(\mathsf{X}_n)$ it follows that $P_{\mathsf{X}_n} \cap P_{\mathsf{Y}_n} = \emptyset$ and thus $g = 0$. This concludes the proof of case I.

*Proof of case II:* We have to prove that $d(\mathsf{X}_n, \mathsf{Y}_n) \leq d(\mathsf{X}_n, \mathsf{Z}_n) + d(\mathsf{Z}_n, \mathsf{Y}_n)$, under the assumption $c(\mathsf{X}_n) \geq c(\mathsf{Z}_n) \geq c(\mathsf{Y}_n)$. Using Def. 2.2 and Lemma 3.2, we transform the triangle inequality into

$$\frac{K(\mathsf{Z}_n)}{K(\mathsf{X}_n)}(K(\mathsf{X}_n|\mathsf{Y}_n) - K(\mathsf{X}_n|\mathsf{Z}_n)) \leq K(\mathsf{Z}_n|\mathsf{Y}_n),$$

where $0 \leq \frac{K(\mathsf{Z}_n)}{K(\mathsf{X}_n)} \leq 1$, if $c(\mathsf{X}_n) \geq c(\mathsf{Z}_n)$. We again consider the extremal case $\frac{K(\mathsf{Z}_n)}{K(\mathsf{X}_n)} = 1$ and apply Lemma 3.1, which yields

$$c(\mathsf{X}_n|\mathsf{Y}_n) - c(\mathsf{X}_n|\mathsf{Z}_n) \leq c(\mathsf{Z}_n|\mathsf{Y}_n).$$

Using our shorthand-coding, this transforms into

$$\begin{gathered}(a+f)-(a+d) \leq (c+f)\\ 0 \leq c+d,\end{gathered}$$

which is correct for all choices of $\{c, d\}$. This concludes the proof of case II.

*Proof of case III:* We have to prove that $d(\mathsf{X}_n, \mathsf{Y}_n) \leq d(\mathsf{X}_n, \mathsf{Z}_n) + d(\mathsf{Z}_n, \mathsf{Y}_n)$ under the assumption $c(\mathsf{Y}_n) \geq c(\mathsf{Z}_n)$. Using Def. 2.2 and Lemma 3.2, we transform the triangle inequality into

$$\frac{K(\mathsf{Y}_n)}{K(\mathsf{X}_n)}(K(\mathsf{X}_n|\mathsf{Y}_n) - K(\mathsf{X}_n|\mathsf{Z}_n)) \leq K(\mathsf{Y}_n|\mathsf{Z}_n),$$

where $0 \leq \frac{K(\mathsf{Y}_n)}{K(\mathsf{X}_n)} \leq 1$, for $c(\mathsf{X}_n) \geq c(\mathsf{Y}_n)$. Again, we consider $\frac{K(\mathsf{Y}_n)}{K(\mathsf{X}_n)} = 1$ and apply Lemma 3.1, which yields

$$c(\mathsf{X}_n|\mathsf{Y}_n) - c(\mathsf{X}_n|\mathsf{Z}_n) \leq c(\mathsf{Y}_n|\mathsf{Z}_n).$$

Using our shorthand-coding, this transforms into

$$\begin{gathered}(a+f)-(a+d) \leq b+d,\\ c+e+f+g \leq b+c+2d+e+g,\\ c(\mathsf{Z}_n) \leq c(\mathsf{Y}_n) + c + d,\end{gathered}$$

which holds since $c(\mathsf{Y}_n) \geq c(\mathsf{Z}_n)$. This concludes the proof of case III. Thus, the LZ-distance fulfills all the axioms required for a metric.

## REFERENCES

[1] C.H. Bennett, P. Gács, M. Li, P.M.B. Vitányi, W.H. Zurek, Information Distance, IEEE Trans. Inform. Theory, 44(4) (1998): 1407–1423.

[2] E. Chavez, G. Navarro, R. Baeza-Yates, J.L. Marroquin, Searching in metric spaces, ACM Comput. Surv. 33 (2001): 273-321.

[3] M. Christen, A. Kohn, T. Ott, R. Stoop, Measuring spike pattern reliability with the LempelZiv-distance, J. Neurosci. Methods 156(1-2) (2006): 342–350.

[4] T.M. Cover, J.A. Thomas, Elements of Information Theory, John Wiley & Sons Inc., New York, 1991.

[5] G.L. Gerstein, D.H. Perkel, J.E. Dayhoff, Cooperative firing activity in simultaneously recorded populations of neurons: detection and measurement, J. Neurosci., 5(4) (1985): 881–889.

[6] D.H. Johnson, C.M. Gruner, K. Baggerly, C. Seshagiri, Information-theoretic analysis of neural coding, J. Comput. Neurosci., 10 (2001): 47–69.

[7] A. Lempel, J. Ziv, On the complexity of finite sequences, IEEE Trans. Inform. Theory, IT-22 (1976): 75–81.

[8] M. Li, J.H. Badger, X. Chen, S. Kwong, P. Kearney, H. Zhang, An information-based sequence distance and its application to whole mitochondrial genome phylogeny, Bioinformatics, 17(2) (2001): 149–154.

[9] M. Li, P. Vitányi, An introduction to Kolmogorov complexity and its applications, Springer Verlag, Berlin, 1997.

[10] D.H. Perkel, G.L. Gerstein, G.P. Moore, Neuronal spike trains and stochastic point processes II. Simultaneous spike trains, Biophys. J., 7 (1967): 419–440.

[11] J.M. Samonds, J.D. Allison, H.A. Brown, A.B. Bonds, Cooperation between area 17 neuron pairs enhances fine discrimination of orientation, J. Neurosci., 23 (2003): 2416–2425.

[12] S. Schreiber, J.-M. Fellous, D. Whitmer, P. Tiesinga, T.J. Sejnowski, A new correlation-based measure of spike timing reliability, Neurocomputing, 52-54 (2004): 925–931.

[13] W.H. Steeb, R. Stoop, Exact complexity of the logistic map, Int. J. Theor. Phys., 36 (1997): 943.

[14] M.C.W. Van Rossum, A novel spike distance, Neural Comput., 13 (2001): 751–763.

[15] J.D. Victor, K.P. Purpura KP, Metric-space analysis of spike trains: theory, algorithms and application, Network: Comp. Neural, 8 (1997): 127–164.

[16] J. Ziv, A. Lempel, Compression of individual sequences by variable rate coding, IEEE Trans. Inform. Theory, IT-24 (1978): 530–536.

Received XXX; revised XXX.

*E-mail address*: {markus,tott,ruedi}@ini.phys.ethz.ch