Anomaly Detection and Approximate Matching via Entropy Divergences

Russell Leidich
https://agnentropy.blogspot.com

Revised December 2, 2017.
Originally published October 23, 2017.

Keywords: entropy, shannon, jensen, kullback, leibler, divergence, exodivergence, transform, comparison, anomaly

0. Abstract

The Jensen-Shannon divergence (JSD) quantifies the “information distance” between a pair of probability distributions. (A more generalized version, which is beyond the scope of this paper, is given in [1]. It extends this divergence to arbitrarily many such distributions. Related divergences are presented in [2], which is an excellent summary of existing work.)

A couple of novel applications for this divergence are presented herein, both of which involving sets of whole numbers constrained by some nonzero maximum value. (We’re primarily concerned with discrete applications of the JSD, although it’s defined for analog variables.) The first of these, which we can call the “Jensen-Shannon divergence transform” (JSDT), involves a sliding “sweep window” whose JSD with respect to some fixed “needle” is evaluated at each step as said window moves from left to right across a superset called a “haystack”.

The second such application, which we can call the “Jensen-Shannon exodivergence transform” (JSET), measures the JSD between a sweep window and an “exosweep”, that is, the haystack minus said window, at all possible locations of the latter. The JSET turns out to be exceptionally good at detecting anomalous contiguous subsets of a larger set of whole numbers.

We then investigate and attempt to improve upon the shortcomings of the JSD and the related Kullback-Leibler divergence (KLD).

1. Background

The JSET would be useful, for example, in detecting bursts of ultrawideband data in a background of Gaussian noise. (Taken to the extreme, ultrawideband becomes what I call “entropy modulation”, which is a signal encoding technique which utilizes continuous changes in information density to convey bit streams, at a cost of many samples per single bit in exchange for arbitrarily high redundancy against interference.)

Ideally, it would be useful to the Search for Extraterrestrial Intelligence (SETI), as there are fundamental information theory reasons to assert that, as a putative intelligent civilization evolved evermore complicated technology, it would dedicate a progressively larger proportion of its total communication power to higher entropy encoding schemes. Such is quite clearly the case on Earth, if one looks at the evolution of digital radio technology, for example. Such schemes are not easily detected via conventional periodicity analysis such as Fourier transforms, wavelets, or phase folding. Of course, those methods are still necessary to filter out interference of mundane origin, but are of unknown utility for the purpose of combing through the resulting residue in search of alien signals.

Fortunately, there is precedent for the use of information theory in related fields, for example the work of Kimberly Cartier in exoplanets [3] or Jason Wright in “artifact” SETI [4]. However, there is much room for optimization. This paper is published in the hopes that it will prove useful to someone with the patience to actually test the techniques, which should prove elementary to any signal processing engineer, whether or not in the SETI field. Moreover, all the work has already been done for you in the open-source Agnentro entropy toolkit, which is available at the web address at the top of this paper.

Finally, I must be clear that my reason for publishing on Vixra is directness of public access. I am doing this as public service, period. I don’t receive any compensation for my work, and have no interest in seeing it paywalled by a

professional journal, regardless of whatever reputational upgrade that might afford. I can be contacted at the aforementioned web address.

2. The Jensen-Shannon Divergence (JSD)

An intuitive, physical way to think of the JSD is to assume that a pair of empirically derived probability distributions actually originate from the same unobservable “generator” distribution, itself a probability distribution, onto which noise has been superimposed, resulting in said pair of (apparently) different such distributions. The closer the JSD to zero, the more likely that this hypothesis is true in a particular case. This concept is easily generalized to arbitrarily many such probability distributions. But let’s start with a review the JSD in its simplest form.

Suppose that we have a pair of probability distributions, each consisting of a set of probabilities of occurrence of various masks (symbols), each of which assuming a value on the (inclusive) interval [0, (Z-1)], (Z>1). We then refer to Z as the “mask span”.

We denote these distributions as N and S, each of Z components of which having a subscript, M, denoting its corresponding mask:

$$N \equiv \left\{ N_0, N_1 \ldots N_{Z-1} \right\}$$
$$S \equiv \left\{ S_0, S_1 \ldots S_{Z-1} \right\}$$

(The literal letters, N and S, will make sense later. For now, consider them as “any old” probability distributions, the sums of which over all allowed values of M thus necessarily being unity.)

The JSD between N and S is then given, in units of bits, by

$$JSD(N,S,Z) \equiv \frac{1}{2 \ln 2} \sum_{M=0}^{Z-1} \left\{ N_M \ln \frac{2N_M}{N_M + S_M} + S_M \ln \frac{2S_M}{N_M + S_M} \right\}$$

as adapted fom units of nats as presented in [2]. But we can isolate all the fractions of (ln 2), which then sum to (2 ln 2), yielding

$$\equiv 1+\frac{1}{2\ln 2}\sum_{M=0}^{Z-1}\left\{N_M \ln\frac{N_M}{N_M+S_M}+S_M \ln\frac{S_M}{N_M+S_M}\right\}$$

$$JSD(N,S,Z)\equiv 1-\frac{1}{2\ln 2}\sum_{M=0}^{Z-1}\left\{(N_M+S_M)\ln(N_M+S_M)-N_M \ln N_M-S_M \ln S_M\right\}$$

where, to emphasize, the units are bits. Typically, I prefer to use nats, but the use of bits has the delightful property of normalizing both sides to real values on the unit interval, [0, 1]. Zero implies that N is identical to S; one implies that they're orthogonal (have a dot product of zero); all intermediate values provide an estimate of how similar they are to their mean, which is the JSD's model of the generator distribution, with lesser values implying greater similarity.

Note that while M may assume any value on the aforementioned domain, it's possible that some such values have zero probability, which creates singularities that require special handling. Specifically, (0 ln 0) must be taken as zero wherever it occurs. This is *not* arbitrary because, formally, the JSD should be expressed as the limit of the above sum as its constituent probabilities morph smoothly from equality to their actual values, but we disregard this and impose the foregoing rule for the sake of brevity.

The above form of the JSD is computationally friendly because the sum operands are all nonnegative, and a value on the unit interval is being subtracted from one, thereby yielding a result on the same.

Before we continue, it's helpful to develop some further intuition for what the JSD actually quantifies. Informally, it's the additional number of bits per mask which would be required to arithmetically encode the following with asymptotic efficiency: (1) a set of Q masks (Q>0) on [0, (Z-1)] adhering exactly to the probabilities given by N, followed by (2) an identically constrained set adhering to S; in both cases, using the average of N_M and S_M as the probability input required to arithmetically encode mask M whenever it occurs.

Note that achieving such asymptotically efficient encoding is actually impossible – and perhaps by many bits – *because it does not account for the encoding cost of the probability values themselves*. This is the entire point of agnentropy and its related transforms, as I discussed in “Introduction to Agnentropy”.

Note also that both notional sets contain Q masks. In other words, the JSD expresses the normalized (Q-agnostic) “information distance” (which is, formally, a divergence and *not* a metric) between the sets based on the assumption that they are of equal size. In the general case in which they may be of different sizes, the JSD measures the divergence between their implied probability distributions, wherein the “probability” of mask M occurring in a mask list is simply its frequency of occurrence, F_M, divided by the total number of masks, Q, contained therein. That is:

$$N_M \equiv \frac{F_{NM}}{Q_N}$$
$$S_M \equiv \frac{F_{SM}}{Q_S}$$

such that N_M and S_M are the probabilities of discovering mask M in the sets with distributions N and S, and “mask counts” (total number of masks) Q_N and Q_S, respectively. The JSD may then be expressed as

$$JSD(N,S,Z) \equiv 1 - \frac{1}{2\ln 2}\sum_{M=0}^{Z-1}\left\{\left(\frac{F_{NM}}{Q_N}+\frac{F_{SM}}{Q_S}\right)\ln\left(\frac{F_{NM}}{Q_N}+\frac{F_{SM}}{Q_S}\right) - \frac{F_{NM}}{Q_N}\ln\frac{F_{NM}}{Q_N} - \frac{F_{SM}}{Q_S}\ln\frac{F_{SM}}{Q_S}\right\}$$

$$JSD(N,S,Z) \equiv 1 - \frac{1}{2\ln 2}\sum_{M=0}^{Z-1}\left\{\left(\frac{F_{NM}}{Q_N}+\frac{F_{SM}}{Q_S}\right)\ln\left(F_{NM}Q_S+F_{SM}Q_N\right) - \frac{F_{NM}}{Q_N}\ln F_{NM}Q_S - \frac{F_{SM}}{Q_S}\ln F_{SM}Q_N\right\}$$

...but we can pull out ($\ln Q_N Q_S$) because all the fractional pieces of ($\ln Q_N$) and ($\ln Q_S$) add up to the former:

$$JSD(N,S,Z) \equiv 1 - \frac{1}{2\ln 2}\left\{\sum_{M=0}^{Z-1}\left\{\left(\frac{F_{NM}}{Q_N}+\frac{F_{SM}}{Q_S}\right)\ln\left(F_{NM}Q_S+F_{SM}Q_N\right) - \frac{F_{NM}}{Q_N}\ln F_{NM} - \frac{F_{SM}}{Q_S}\ln F_{SM}\right\} - \ln Q_N Q_S\right\}$$

As always, (0 ln 0) must be taken as zero. This is sufficient to guarantee that the JSD ends up on [0, 1]. This form is useful in practice because all of the log operands are whole, which means that they are to some extent amenable to acceleration via caching. Furthermore the divisions are actually just multiplications by reciprocals which remain constant throughout the entire summation process. Agnentro takes advantage of both, although for the sake of consistency with other normalized entropy quantifiers, it outputs (1-JSD) instead of literal JSD.

3. The Jensen-Shannon Divergence Transform (JSDT)

We now derive a sliding window transform – a “sweep transform” – which computes the JSD between a contiguous subset of a “haystack” of masks, called a “sweep window”; and a constant “needle” of masks in a separate set, at every step as the sweep window marches across the haystack. In practice, this means that we’re searching for approximate matches to the needle within the haystack.

The aforementioned sets are constrained as follows: (1) the haystack contains at least as many masks as the sweep window; (2) the needle and the haystack consist of masks on [0, (Z-1)]; (3) the haystack consists of Q_H masks ($Q_H>0$); (4) the needle consists of Q_N masks ($Q_N>0$); and (5) the sweep window consists of Q_S masks ($0<Q_S<=Q_H$), where Q_S is called the “sweep”. Note that there is no other relationship assumed between Q_N and Q_S, although in practice they’re usually equal.

Implicitly, the output of the JSDT is a vector containing (Q_H-Q_S+1) components sorted ascending by the zero-based base index of the sweep window. So for example the first component of the JSDT output is the JSD between the sweep window based at index zero (the leftmost index) of the haystack, and the needle; the second component is the JSD between the sweep window based at index one, and the needle; etc. Formally

$$JSDT(H,J,N,Z) \equiv JSDT_J \equiv JSD(N,S(J),Z), 0 \le J \le (Q_H - Q_S)$$

where S(J) is the probability distribution (with sum one) of masks derived from the frequencies with which they occur at indexes J through (J+Q_S-1) of the haystack, H; and H itself consists of masks H_K, where K is on [0, (Q_H-1)]. We can then we can define $S(J)_M$ as the "probability" of finding mask M within sweep window S(J):

$$S(J) \equiv \left[S(J)_0, S(J)_1 \ldots S(J)_{Z-1} \right]$$
$$S(J)_M \equiv \frac{1}{Q_S} \sum_{K=J}^{J+Q_S-1} (H_K = M)$$

where the value of the expression

$$(H_K = M)$$

is one if true, else zero. This is sufficient information to allow us to compute the JSDT. However, when (Q_H>Q_S), the computational complexity of repeatedly computing the JSD between the sweep window and the needle is burdensome. Fortunately, in that case, we can shortcut the process by considering the net probability flux due to the exit of mask M_0 from the left side of the sweep window, counterbalanced by the entry of mask M_1 into its right side, which occurs in the course of a single step (incrementation of J). To be clear

$$M_0 = H_J$$
$$M_1 = H_{J+Q_S}$$
$$0 \le J \le (Q_H - Q_S - 1)$$

(Note that M_1 is the first mask *after* the sweep window, whereas M_0 is the first mask *inside* it.) Initially, we need only evaluate

$$JSDT_0 \equiv JSD(N, S(0), Z)$$

Then, starting from (J=0): if (M_0=M_1), then $JSDT_{J+1}$ is simply equal to $JSDT_J$. Otherwise, we need to evaluate

$$\Delta JSDT_J \equiv JSD(N, S(J+1), Z) - JSD(N, S(J), Z)$$

then add this difference to $JSDT_J$ in order to obtain $JSDT_{J+1}$. From the definition of the JSD, we find that

$$
\begin{aligned}
&\Delta JSDT_J(2\ln 2)\equiv \\
&-\frac{F_{SM0}}{Q_S}\ln F_{SM0}Q_N \\
&-\frac{F_{SM1}}{Q_S}\ln F_{SM1}Q_N \\
&+\left(\frac{F_{SM0}}{Q_S}-\frac{1}{Q_S}\right)\ln\left(F_{SM0}Q_N-Q_N\right) \\
&+\left(\frac{F_{SM1}}{Q_S}+\frac{1}{Q_S}\right)\ln\left(F_{SM1}Q_N+Q_N\right) \\
&+\left(\frac{F_{NM0}}{Q_N}+\frac{F_{SM0}}{Q_S}\right)\ln\left(F_{NM0}Q_S+F_{SM0}Q_N\right) \\
&+\left(\frac{F_{NM1}}{Q_N}+\frac{F_{SM1}}{Q_S}\right)\ln\left(F_{NM1}Q_S+F_{SM1}Q_N\right) \\
&-\left(\frac{F_{NM0}}{Q_N}+\frac{F_{SM0}}{Q_S}-\frac{1}{Q_S}\right)\ln\left(F_{NM0}Q_S+F_{SM0}Q_N-Q_N\right) \\
&-\left(\frac{F_{NM1}}{Q_N}+\frac{F_{SM1}}{Q_S}+\frac{1}{Q_S}\right)\ln\left(F_{NM1}Q_S+F_{SM1}Q_N+Q_N\right)
\end{aligned}
$$

Note that the above expressions are to be computed *before* decrementing the frequency F_{SM0} (of M_0) or incrementing the frequency F_{SM1} (of M_1) within the sweep window. As always, (0 ln 0) must be treated as zero, although this can only occur with *some* of the log terms; therefore checking for zero operands is not necessary in every case.

Finally, multiplying both sides by Q_NQ_S yields

$$
\begin{aligned}
&\Delta JSDT_J(2Q_NQ_S\ln 2)\equiv \\
&-F_{SM0}Q_N\ln F_{SM0}Q_N \\
&-F_{SM1}Q_N\ln F_{SM1}Q_N \\
&+\left(F_{SM0}Q_N-Q_N\right)\ln\left(F_{SM0}Q_N-Q_N\right) \\
&+\left(F_{SM1}Q_N+Q_N\right)\ln\left(F_{SM1}Q_N+Q_N\right) \\
&+\left(F_{NM0}Q_S+F_{SM0}Q_N\right)\ln\left(F_{NM0}Q_S+F_{SM0}Q_N\right) \\
&+\left(F_{NM1}Q_S+F_{SM1}Q_N\right)\ln\left(F_{NM1}Q_S+F_{SM1}Q_N\right) \\
&-\left(F_{NM0}Q_S+F_{SM0}Q_N-Q_N\right)\ln\left(F_{NM0}Q_S+F_{SM0}Q_N-Q_N\right) \\
&-\left(F_{NM1}Q_S+F_{SM1}Q_N+Q_N\right)\ln\left(F_{NM1}Q_S+F_{SM1}Q_N+Q_N\right)
\end{aligned}
$$

which provides an expression for the delta in terms of more computationally friendly whole numbers, and also grows in expression complexity in order to show reusable terms.

Note that the constant coefficient on the lefthand side does not change the ordering of JSDT items. Therefore division of the righthand side by this constant can be deferred until just prior to result issuance. This can save time in the event that only the greatest or least however-many JSDT items are to be returned, as is usually the case with Agnentro Find, for example.

4. The Jensen-Shannon Exodivergence Transform (JSET)

Suppose that we now remove the needle entirely, and focus instead on the JSD between the sweep window and (the haystack without said sweep window). We call the latter the “exosweep”, in the sense of “the stuff outside of the sweep window”. Provided that ($Q_S<Q_H$) – because otherwise the JSET is defined to be zero – we now have a means of quantifying the difference between the probability distributions of the sweep window and its exosweep.

In this case, as with the JSDT, we have exiting and entering masks, M_0 and M_1, respectively. But this time when M_0 exits the sweep window, it *enters* the exosweep; when M_1 enters the sweep window, it *exits* the exosweep.

Initially, we must compute the JSD between the leftmost sweep window and its exosweep. We can use the very same expression for the JSD between a needle and a sweep window, as given above, except that in this case the values F_{NM} are taken from the exosweep and ($Q_N=(Q_H-Q_S)$). (We can simply accumulate the mask frequencies of the haystack and the sweep window, then subtract the latter distribution from the former to find F_N. This is exactly what Agnentro Scan does, for example.)

Having done that, we must once again compute successive deltas:

$$\Delta JSET_J \equiv JSD(N, S(J+1), Z) - JSD(N, S(J), Z)$$

but this time for ($0<=J<Q_N$). Building on the previous analysis, the following expression arises straightforwardly from symmetry:

$$
\begin{aligned}
&\Delta JSET_J(2\ln 2)\equiv \\
&-\frac{F_{NM0}}{Q_N}\ln\frac{F_{NM0}}{Q_N} \\
&-\frac{F_{NM1}}{Q_N}\ln\frac{F_{NM1}}{Q_N} \\
&-\frac{F_{SM0}}{Q_S}\ln\frac{F_{SM0}}{Q_S} \\
&-\frac{F_{SM1}}{Q_S}\ln\frac{F_{SM1}}{Q_S} \\
&+\left(\frac{F_{NM0}}{Q_N}+\frac{1}{Q_N}\right)\ln\left(\frac{F_{NM0}}{Q_N}+\frac{1}{Q_N}\right) \\
&+\left(\frac{F_{NM1}}{Q_N}-\frac{1}{Q_N}\right)\ln\left(\frac{F_{NM1}}{Q_N}-\frac{1}{Q_N}\right) \\
&+\left(\frac{F_{SM0}}{Q_S}-\frac{1}{Q_S}\right)\ln\left(\frac{F_{SM0}}{Q_S}-\frac{1}{Q_S}\right) \\
&+\left(\frac{F_{SM1}}{Q_S}+\frac{1}{Q_S}\right)\ln\left(\frac{F_{SM1}}{Q_S}+\frac{1}{Q_S}\right) \\
&+\left(\frac{F_{NM0}}{Q_N}+\frac{F_{SM0}}{Q_S}\right)\ln\left(\frac{F_{NM0}}{Q_N}+\frac{F_{SM0}}{Q_S}\right) \\
&+\left(\frac{F_{NM1}}{Q_N}+\frac{F_{SM1}}{Q_S}\right)\ln\left(\frac{F_{NM1}}{Q_N}+\frac{F_{SM1}}{Q_S}\right) \\
&-\left(\frac{F_{NM0}}{Q_N}+\frac{F_{SM0}}{Q_S}+\frac{1}{Q_N}-\frac{1}{Q_S}\right)\ln\left(\frac{F_{NM0}}{Q_N}+\frac{F_{SM0}}{Q_S}+\frac{1}{Q_N}-\frac{1}{Q_S}\right) \\
&-\left(\frac{F_{NM1}}{Q_N}+\frac{F_{SM1}}{Q_S}-\frac{1}{Q_N}+\frac{1}{Q_S}\right)\ln\left(\frac{F_{NM1}}{Q_N}+\frac{F_{SM1}}{Q_S}-\frac{1}{Q_N}+\frac{1}{Q_S}\right)
\end{aligned}
$$

$$
\begin{aligned}
&\Delta JSET_J(2Q_N Q_S \ln 2) \equiv \\
&-F_{NM0} Q_S \ln F_{NM0} Q_S \\
&-F_{NM1} Q_S \ln F_{NM1} Q_S \\
&-F_{SM0} Q_N \ln F_{SM0} Q_N \\
&-F_{SM1} Q_N \ln F_{SM1} Q_N \\
&+(F_{NM0} Q_S + Q_S) \ln (F_{NM0} Q_S + Q_S) \\
&+(F_{NM1} Q_S - Q_S) \ln (F_{NM1} Q_S - Q_S) \\
&+(F_{SM0} Q_N - Q_N) \ln (F_{SM0} Q_N - Q_N) \\
&+(F_{SM1} Q_N + Q_N) \ln (F_{SM1} Q_N + Q_N) \\
&+(F_{NM0} Q_S + F_{SM0} Q_N) \ln (F_{NM0} Q_S + F_{SM0} Q_N) \\
&+(F_{NM1} Q_S + F_{SM1} Q_N) \ln (F_{NM1} Q_S + F_{SM1} Q_N) \\
&-(F_{NM0} Q_S + F_{SM0} Q_N + Q_S - Q_N) \ln (F_{NM0} Q_S + F_{SM0} Q_N + Q_S - Q_N) \\
&-(F_{NM1} Q_S + F_{SM1} Q_N - Q_S + Q_N) \ln (F_{NM1} Q_S + F_{SM1} Q_N - Q_S + Q_N)
\end{aligned}
$$

where, as always, all frequency values are to be measured *prior* to their incrementation or decrementation.

As with the JSDT, the output of the JSET is a vector containing (Q_H-Q_S+1) components. (In the case that (Q_H=Q_S), we could perhaps consider the (null) exosweep as a uniform distribution, reached via some limit under analytic continuation of the JSET, from which to compute the JSD to the sweep window. But this is an exercise in transfinite mathematics beyond the scope of this paper. Moreover, the result would have no practical significance. So we just take the easy way out and define the JSET to be zero in this case!)

The JSET is of particular use in the discovery of anomalous bursts of data within larger sets, for example, prolate spheroidal waves [5] or solitons [6] buried in noise, both of which having obvious application to ultrawideband communication. Technologically superior aliens would, almost by definition, be expected to use even higher entropy signalling which would therefore be even less amenable to conventional oscillation analysis.

5. JSET vs. Exoelasticity vs. Exoentropy

In my paper entitled “Introduction to Entropy Transforms”, I introduced the concepts of exoelasticity and exoentropy. They’re similar to the JSET in that they both involve the divergence between a sweep window and its exosweep.

Agnentro includes SETI Demo, which is a signal injection test program designed to shed some light on the question of which entropy tools are most effective at detecting a signal. The approach is crude, involving the injection of a minimum-amplitude square wave into real Gaussian noise obtained from a radio telescope. At best, we can hope to derive qualitative information about the relative utility of various methods.

One surprising result is that searching for said signal using exoelasticity (mode bit one) is fully half as sensitive as cheating by knowing the exact signal topology in advance (the “Fourier” technique, mode bit 7). In particular, the former seems to need roughly 1300 samples to detect the faint signal, whereas the latter needs about half as many. But in general, it’s overwhelmingly likely that “the” signal will be either too short or too long, in which case both methods would work equally well (or not). And for the aforementioned evolutionary reasons, it would be wise for SETI to employ at least one method based on information theory.

JSET (mode bit 9) performs comparably to exoelasticity – apparently a bit worse at the 50% detection threshold, and slightly better at the 99.9% detection threshold. It currently takes about 10 times as long, however; as such, it’s merely intended as a reference demo as of this writing. My statistics are unfortunately weak with regards to this comparison, owing to insufficient compute time, but suffice to say that the methods are of comparable utility in this specific case. The results might be quite different with different types of injected signals, however. Fundamentally, exoelasticity is concerned with how well the exosweep predicts the sweep, whereas the JSET is concerned with the credibility of the notion that both regions originated from the same underlying phenomenon. In this regard, the former assumes that the statistical significance of the exosweep is dominant, whereas the latter assumes it to be of equivalent value, to that of the sweep window. In the signal injection simulation, the exosweep is many times larger, which therefore perhaps explains the apparent advantage enjoyed by exoelasticity.

For its part, exoentropy (mode bit 2) performs moderately worse, but is roughly twice as fast as exoelasticity. Agnentropy (mode bit zero) performs worse still, requiring essentially quadruple the number of samples as the

cheat mode, but executes on the order of 100 times as fast as the JSET, so for the purposes of realtime signal analysis with the expectation of a sufficiently long pulse, it could make practical sense.

There is, however, a method which outperforms all of the foregoing on SETI Demo, which we'll introduce later.

Practically speaking, however, the only way to know is to try. Agnentro is available on Github.

6. Caveats of the Kullback-Leibler Divergence (KLD)

The KLD [7] appears to be more popular than the JSD for some reason. As such, it deserves some analysis. Given a prior (presumed) probability distribution Q, with individual mask probabilities Q_M, and a posterior (empirical) probability distribution P, with individual mask probabilities P_M, then KLD(P||Q) (which counterinuitively means "the KLD from Q to P") is given, in units of bits, by

$$KLD(P\|Q) \equiv \frac{1}{\ln 2} \sum_{M=0}^{Z-1} P_M \ln \frac{P_M}{Q_M}$$

provided that (Q_M=0) implies (P_M=0) for all M, and in which case the term in question is to be taken as zero, just as with (0 ln 0).

But therein lies the first problem: what if Q_M is zero, but P_M isn't? Sorry, it's just undefined! This isn't practical because in reality stuff happens in set P that never occurred in set Q. The concepts of agnentropy (and agnostic frequency in particular) are a practical if imperfect means to asymptotically escape this problem. But then, so is every other entropy tool mentioned in this paper because, by design, they can't blow up like this.

The second issue with the KLD is that it's a poor basis of comparison between a pair of distributions, the reason being that it considers the prior distribution as "the" source of truth. In the real world, both distributions are

usually informative to some extent. Indeed, the JSD has the opposite weakness: it treats both as *equally* informative.

A better approach would seem to be to interpolate between these extremes…

7. The Leidich Divergence (LD)

Consider the Shannon entropy E_N, in bits, of a mask list with a mask count of Q, a mask span of Z, and mask frequencies F_{NM} implied by mask probabilities N_M:

$$E_N(N,Z) \equiv \frac{1}{\ln 2}\left\{ Q_N \ln Q_N - \sum_{M=0}^{Z-1} F_{NM} \ln F_{NM} \right\}$$

where, as always:

$$N_M \equiv \frac{F_{NM}}{Q_N}$$

Now suppose that we had a second such entropy, E_S:

$$E_S(S,Z) \equiv \frac{1}{\ln 2}\left\{ Q_S \ln Q_S - \sum_{M=0}^{Z-1} F_{SM} \ln F_{SM} \right\}$$

then (E_N+E_S) is the lower bound cost, in bits, of encoding both mask lists independently (without accounting for the overhead of storing all F_{NM} and F_{SM}). If instead both mask lists were encoded from the joint distribution, the resulting total Shannon entropy E_J of both mask lists would be

$$E_J(N,S,Z) \equiv \frac{1}{\ln 2}\left\{ (Q_N+Q_S)\ln(Q_N+Q_S) - \sum_{M=0}^{Z-1} (F_{NM}+F_{SM})\ln(F_{NM}+F_{SM}) \right\}$$

We then define the Leidich divergence as the additional cost, in bits per mask, incurred by defining the entropy as E_J instead of (E_N+E_S):

$$LD(N,S,Z) \equiv \frac{E_J - (E_N+E_S)}{Q_N+Q_S}$$

$$LD(N,S,Z)\equiv\frac{1}{(Q_N+Q_S)\ln 2}\left\{\Delta-\sum_{M=0}^{Z-1}\left((F_{NM}+F_{SM})\ln(F_{NM}+F_{SM})-F_{NM}\ln F_{NM}-F_{SM}\ln F_{SM}\right)\right\}$$

where

$$\Delta\equiv(Q_N+Q_S)\ln(Q_N+Q_S)-Q_N\ln Q_N-Q_S\ln Q_S$$

Voila! We now have a sample-size-weighted divergence between a pair of frequency lists (that is, a pair of Z-tuples consisting of mask frequencies). Note that when ($Q_N>>Q_S$), LD(N, S, Z) approaches KLD(S||N). On the other hand, as Q_N and Q_S approach equality, the LD approaches JSD(N, S, Z). (As always, (0 ln 0) must be treated as zero. However, unlike with the KLD, there is no requirement that F_{SM} be zero when F_{NM} is zero, or the other way round.)

8. The Generalized Leidich Divergence (GLD)

The LD is easily extended from a pair of such lists to K of them (K>1), yielding the GLD:

$$GLD(K,N_0,N_1\ldots N_{K-1},Z)\equiv\frac{1}{Q\ln K}\left\{\Delta_K-\sum_{M=0}^{Z-1}\left\{\left(\sum_{J=0}^{K-1}F_{JM}\right)\ln\left(\sum_{J=0}^{K-1}F_{JM}\right)-\sum_{J=0}^{K-1}F_{JM}\ln F_{JM}\right\}\right\}$$

where

$$\Delta_K\equiv Q\ln Q-\sum_{J=0}^{K-1}Q_J\ln Q_J$$

and Q is the sum of all Q_J:

$$Q\equiv\sum_{J=0}^{K-1}Q_J$$

and Q_J is just the mask count of frequency list J:

$$Q_J\equiv\sum_{M=0}^{Z-1}F_{JM}$$

and F_{JM} is analagous to F_M:

$$N_{JM} \equiv \frac{F_{JM}}{Q_J}$$

Note that the LD is normalized for all permitted values of K. This implies that, regardless of K and all the Q_J values, a GLD of zero implies that all distributions are identical, whereas one implies that they're all orthogonal. As such, the units are neither bits nor nats when (K>2). (For nats, multiply the GLD by (ln K); for bits, multiply by ($\log_2$ K).)

Is the GLD actually new? Not really. It's simply a degenerate case of the generalized Jensen-Shannon divergence (GJSD) as presented in [1], in which the weights have been tuned so as to be proportional to the mask counts of the distributions in question, with the result then normalized for the sake of meaningful comparison.

And let's be honest: the LD *still* doesn't account for the cost of encoding the probability values themselves; like the JSD and the Shannon entropy itself, it does not precisely express encoded arithmetic compression size, or any change in that size. For that, we'll need something like agnentropy; perhaps at some point we'll need to consider agnostic *divergences* as well. For small data sets or large ones deeply buried in noise, the distinction between asymptotically accurate divergences and actual differences in encoding size can make or break signal detection; information comes down to encoding size, not cheap asymptotic approximations thereof.

Moreover, even assuming the existence of a common generator underlying all observed (N_{JM})s, our estimation of the former is inevitably flawed due to the finite amount of information in the latter (in particular, the (F_{JM})s and (Q_J)s that we happened to observe). Clearly, such an approach can produce only whole frequencies and thus rational probabilities, but there is no reason to assume that the generator probabilities are rational. Put another way, there's an uncertainty in the generator, and thus also the divergence therefrom to the observed distributions, which even infinite precision cannot cure. Thus the

LD quantifies the normalized divergence from an *approximation* of the generator to the empirical distributions which gave rise to it.

I hope that the very fact that I've been pompous enough to name a mathematical object after myself will annoy you enough to motivate the search for a precursor in the literature. The act of doing so should at least create a constructive debate over its utility. I don't really care about the inevitable *ad hominem* attacks because hardly anyone seems to take me seriously anyway. (For example, I've been labeled as a pseudoscientist, which is particularly odd because my arguments are strictly mathematical – not scientific – and thus straightforwardly falsifiable.) But of course, if any such precursor should be found, I would be happy to revise this paper.

9. The Leidich Divergence Transform (LDT)

Analogously to the JSDT, we define the LDT between a sweep window based at index J of a haystack H, having probability distribution S(J); and a needle with probability distribution N as

$$LDT(H,J,N,Z) \equiv LDT_J \equiv LD(N,S(J),Z), 0 \le J \le (Q_H - Q_S)$$

so implicitly

$$\Delta LDT_J \equiv LD(N,S(J+1),Z) - LD(N,S(J),Z)$$

which works out as follows:

$$\begin{gathered}
\Delta LDT_J (Q_N + Q_S) \ln 2 \equiv \\
-F_{SM0} \ln F_{SM0} \\
-F_{SM1} \ln F_{SM1} \\
+(F_{NM0} + F_{SM0}) \ln (F_{NM0} + F_{SM0}) \\
+(F_{NM1} + F_{SM1}) \ln (F_{NM1} + F_{SM1}) \\
+(F_{SM0} - 1) \ln (F_{SM0} - 1) \\
+(F_{SM1} + 1) \ln (F_{SM1} + 1) \\
-(F_{NM0} + F_{SM0} - 1) \ln (F_{NM0} + F_{SM0} - 1) \\
-(F_{NM1} + F_{SM1} + 1) \ln (F_{NM1} + F_{SM1} + 1)
\end{gathered}$$

$$\begin{gathered}\Delta LDT_J(Q_N+Q_S)\ln 2\equiv \\ -\ln F_{SM0} \\ +\ln(F_{SM1}+1) \\ +\ln(F_{NM0}+F_{SM0}) \\ -\ln(F_{NM1}+F_{SM1}+1) \\ -(F_{SM0}-1)\Delta ln(F_{SM0}-1) \\ +F_{SM1}\Delta ln\, F_{SM1} \\ +(F_{NM0}+F_{SM0}-1)\Delta ln(F_{NM0}+F_{SM0}-1) \\ -(F_{NM1}+F_{SM1})\Delta ln(F_{NM1}+F_{SM1})\end{gathered}$$

where

$$\Delta ln(F)\equiv \ln(F+1)-\ln F, F>0$$

which is faster and more precise if a single unified Taylor series is used instead of computing the difference literally as stated.

The LDT is most useful at searching for needles in haystacks in cases in which the relative number of masks in the needle and the sweep window matters for the sake of statistical significance.

10. The Leidich Exodivergence Transform (LET)

Analogously to the JSET, the LET produces a vector providing LD between a sweep window and its exosweep as the former moves from left to right across a haystack. All constraints and definitions are identical for the LET as the JSET. And as with the JSET, the first step in computing the LET is to evaluate the LD between the leftmost sweep window and its exosweep. In the trivial case where (Q_H=Q_S), the output is defined to be zero, and we're done. Otherwise, we can compute the stepwise changes in the LET in the same manner as with the JSET:

$$\Delta LET_J\equiv LD(N,S(J+1),Z)-LD(N,S(J),Z)$$

This works out as follows:

$$
\begin{aligned}
&\Delta LET_J(Q_N+Q_S)\ln 2\equiv\\
&-F_{NM0}\ln F_{NM0}\\
&-F_{NM1}\ln F_{NM1}\\
&-F_{SM0}\ln F_{SM0}\\
&-F_{SM1}\ln F_{SM1}\\
&+(F_{NM0}+1)\ln(F_{NM0}+1)\\
&+(F_{NM1}-1)\ln(F_{NM1}-1)\\
&+(F_{SM0}-1)\ln(F_{SM0}-1)\\
&+(F_{SM1}+1)\ln(F_{SM1}+1)
\end{aligned}
$$

$$
\begin{aligned}
&\Delta LET_J(Q_N+Q_S)\ln 2\equiv\\
&+\ln(F_{NM0}+1)\\
&-\ln F_{NM1}\\
&-\ln F_{SM0}\\
&+\ln(F_{SM1}+1)\\
&+F_{NM0}\Delta ln\, F_{NM0}\\
&-(F_{NM1}-1)\Delta ln(F_{NM1}-1)\\
&-(F_{SM0}-1)\Delta ln(F_{SM0}-1)\\
&+F_{SM1}\Delta ln\, F_{SM1}
\end{aligned}
$$

which is rapidly computable using the cached logs and logdeltas of whole numbers.

11. Remarks

The JSD quantifies the normalized cost, in additional bits per mask, of preserving the information in a pair of probability distributions, starting with the assumption that they are in fact merely noisy manifestations the same underlying generator distribution which is simply their mean. As such, the JSD worldview is one in which both empirical distributions are equally informative. This can be useful for finding approximate matches across asymmetric scales, for example, similar color distributions in photos of different sizes, or similar temperature distributions during time periods of mismatched duration. However, when one of the distributions is much more statistically significant than the other, the JSD can provide suboptimal results.

The KLD “solves” this problem by computing the divergence from the “certain” distribution to the “noisy” one. In reality, of course, no empirically derived distribution is precisely accurate, and *all* empirical data provides *some* information about the state of the world. Consequently, the KLD explodes when it encounters an “impossible” mask in the posterior distribution.

The GJSD builds on the JD by assuming that the generator distribution is a linear combination – not necessarily equally weighted – of arbitrarily many probability distributions.

The LD then attempts to bridge the gap between the JSD and the KLD by tuning the weights of the GJSD so as to arrive at a sample-size-weighted relative valuation of the information contained in each distribution. In signal injection testing with SETI Demo, it outperformed every other detection method except Fourier cheat mode.

Will SETI try this? Hopefully someone somewhere someday, will.

12. Bibliography

[1] https://en.wikipedia.org/wiki/Jensen%E2%80%93Shannon_divergence

[2] Cha, Sung-Hyuk, "Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions", International Journal of Mathematical Models and Methods in Applied Sciences, 2007.

[3] https://sites.psu.edu/astrolady/background/transit-entropy

[4] http://sites.psu.edu/astrowright/2013/03/09/artifact-seti

[5] https://en.wikipedia.org/wiki/Prolate_spheroidal_wave_function

[6] https://en.wikipedia.org/wiki/Soliton

[7] https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence