Topological Detection of Wideband Anomalies

Russell Leidich
https://agnentropy.blogspot.com

April 11, 2018

Keywords: wideband, ultrawideband, anomaly, noise, signal, detection, logfreedom, dyspoissonism

0. Abstract

Successive real-valued measurements of any physical chaotic oscillator can serve as entropy inputs to a random number generator (RNG) with correspondingly many whole numbered outputs of arbitrarily small bias, assuming that no correlation exists between successive such measurements apart from what would be implied by their probability distribution function (AKA the oscillator's analog "generator", which is constant over time and thus asymptotically discoverable).

Given some historical measurements (a "snapshot") of such an oscillator, we can then train the RNG to expect inputs distributed uniformally among the real intervals defined by those measurements and spanning the entire real line. Each interval thus implies an index in sorted order, starting with the leftmost which maps to zero; the RNG does nothing more than to perform this mapping. We can then replace that first oscillator with a second one presumed to abide by the same generator. It would then be possible to characterize the accuracy of that presumption by quantifying the ensuing change in quality of the RNG.

Randomness quality is most accurately expressed via dyspoissonism, which is a normalized equivalent of the log of the number of ways in which a particular distribution of frequencies (occurrence counts) of masks (whole numbers) can occur. Thus the difference in dyspoissonism between the RNG output sets will serve to estimate the information divergence between their

respective generators, which in turn constitutes a ranking quantifier for the purpose of anomaly detection.

1. From Generator Outputs to RNG Intervals

Consider such an oscillator. Reading its outputs requires measurement of a physical property, such as mass; a composite thereof, such as the power in a particular frequency band of the radio spectrum; or even a financial quantifier such as transaction size. With negligible loss of generality, we can assume that all such measurements are unique. (If this is not the case, then trivial accommodations can be made to this method.)

Assume that we have a snapshot consisting of Z measurements (Z>1), all of which on the (closed) interval $[X_0, X_{Z-1}]$ where X_0 and X_{Z-1} are the minimum and maximum, respectively. Furthermore, we assume that all such X values are sorted such that X_0 is the closest to negative infinity and X_{Z-1} is closest to positive infinity. (We can forget the order of the particular measurements without incurring information loss because they are assumed to be uncorrelated. That is, there is nothing to be learned from the snapshot apart from the asymptotic convergence of our *model* of the generator with the *actual* one.) Therefore, again with negligible loss of generality, we can assume the existence of (Z+1) orthogonal intervals of nonzero length, the union of which being the entire real line, namely:

$$\{(-\infty, X_0), [X_0, X_1), [X_1, X_2)... [X_{Z-2}, X_{Z-1}), [X_{Z-1}, \infty)\}$$

where square brackets and parentheses indicate closed (inclusive) and open (exclusive) interval ends, respectively. We can then replace each interval with its respective whole numbered index such that $(-\infty, X_0)$ corresponds to zero and $[X_{Z-1}, \infty)$ corresponds to Z. But we also interpret Z as zero because symmetry implies that the expected probability contribution from the union of the infinite intervals equals the expected contribution from each *finite* interval. Therefore the intervals are better denoted as:

$$\{[X_{Z-1}, X_0), [X_0, X_1), [X_1, X_2)... [X_{Z-2}, X_{Z-1})\}$$

where $[X_{Z-1}, X_0)$ – an "antiinterval" – denotes the union of $(-\infty, X_0)$ and $[X_{Z-1}, \infty)$. (One *could*, of course, assign the end intervals to distinct index values, at the cost of some statistical distortion. This is a design tradeoff to be evaluated upon implementation.)

Given the set of N intervals obtained as described above, we can then map any measurement (which might be digitized as a signed integer, floating-point, or fixed-point value) to a whole number on [0, (Z-1)]. Because each interval corresponds to one measurement, the width of each one is thus a crude approximation of an equiprobable generator slice. Therefore, to the extent that the snapshot is unbiased, the RNG should be unbiased. But just how unbiased is the snapshot expected to be?

2. Poisson Jitter Limitations to Snapshot Bias

Ideally, each snapshot interval corresponds to a probability slice of the generator (hereinafter simply "slice") with area (1/Z).

But this is not so. The intervals are established by random measurements of the generator, akin to the distribution of typos made by a writer with a constant error rate. In the limit of infinitely many pages, the number of typos on each page is an archetype of Poisson noise.

Metaphorically, the "pages" are equiprobable slices and the "typos" are measurements. Therefore the number of measurements "owned" by each equiprobable slice is also Poisson noise.

But the same logic would apply again if we were to map at least Z *additional* measurements of the same oscillator to their corresponding interval indexes, resulting in a "Poisson squared bias" of the RNG. If instead we used a *different* oscillator, the bias would be *at least* as bad, asymptotically speaking. Quantifying the extent of that bias requires counting the number of members in certain topological equivalence classes, which is the subject of the next section.

3. A Quick Primer on Logfreedom and Dyspoissonism

3.1. The Terminology of Mask Lists

We presume that the reader is familiar with the terminology of mask lists [1]. To briefly summarize:

“Masks” are the whole numbers which occur in a “mask list” (set of masks) which is indexed from zero, sensitive to order, and may contain repeats, like the outputs of the RNG described above. Conventionally, the minimum mask is presumed to be zero, and the maximum, (Z-1), where (Z>1). Z is called the “mask span”. The nonzero number of masks in such a list, denoted Q, is called the “mask count”. (In the examples above, we’ve assumed that (Q=Z), but in general this is not required.)

A “frequency list” consists of the frequencies (occurrence counts) of corresponding masks on [0, (Z-1)]. The sum of the items in a frequency list is thus always Q.

A “population list” consists of the populations (frequencies by another name so as to avoid confusion) of the frequencies from zero through Q (because Q is the greatest possible frequency). The sum of the items in a population list is thus always Z.

In the limit of large Q, a mask list generated by an unbiased random number generator will give rise to a population list which, pursuant to area normalization, approaches a Poisson distribution. In the specific case that (Q=Z), the result will be a lambda-one Poisson distribution (LOPD), which has mean and variance both equal to (1/Z). Suffice to say that there are theoretical reasons to believe that LOPDs provide particularly good entropy contrast, that is, differentiation between interesting and uninteresting mask lists.

The number of *distinct* mask lists having the same population list is called the “way count” of that population list. The natural log of the way count is called the “logfreedom” of that list. Logfreedom is a useful concept simply because way counts tend to exceed the native numerical precision supported in

common microprocessors. Finally, note that the way count is just the multiplicity of a given population list over the set of all mask lists with some particular Q and Z (but much less verbose!).

3.2. The Logfreedom Formula

As derived in [2], the logfreedom L of a mask list with mask count Q and mask span Z is (exactly) given by:

$$L \equiv \ln Q! + \ln Z! - \ln H[0]! - \sum_{F=1}^{Q} \ln H[F]! - \sum_{F=1}^{Q} H[F] \ln F!$$

where H[F] denotes “the population of frequency F”. (Recall that zero factorial is one, so there are no infinities to worry about.) Note that the original formula had the summations going from (F=1) to K, where K is the greatest frequency with nonzero population; Q is more intuitive, if a bit obtuse. Note also that the logs of factorials are efficiently computable using interval arithmetic for safety’s sake, as explained in [3].

Recall that logfreedom is the log of the number of members of an equivalence class uniquely associated with a particular population list. The more members, the greater the probability of generating a member of that class via an unbiased RNG.

It’s best to think of logfreedom as a randomness quantifier, rather than a form of entropy, although it does have a fairly direct connection to the latter: assuming that the population list corresponding to a given mask list has zero information cost, then logfreedom measures the number of bits (indirectly, via conversion from nats) required to encode the “way” corresponding uniquely to that mask list. A “way” is just a whole number on the interval [0, (W-1)], where W is the way count of the population list. In the limit of infinite Q, but with fixed Z, the ratio of logfreedom to Shannon entropy approaches one (assuming that both are expressed in the same units).

Finally, consider an illustrative example of the difference between logfreedom and entropy when (Q=Z): whereas the maximum entropy mask

list would contain exactly one instance of each mask, the maximum logfreedom one would asymptotically correspond to an LOPD.

3.3. Comparing RNGs with Logfreedom

Suppose we have a set of mask lists of identical Q and Z (but with Q and Z not necessarily equal), each one having been generated by a different RNG. We can then compute the logfreedom of each mask list, then sort them. The result will be a list of absolute values of logs of probability slices of population lists. We would then conclude that the RNG which generated the mask list with the maximum logfreedom was probably the least biased. But this analysis is problematic because differences in logfreedom are absolute differences in information, which need to be normalized to be sensible. Dyspoissonism would be more useful in that regard, which is the subject of the next section.

3.4. Comparing RNGs with Dyspoissonism

Briefly, dyspoissonism measures the fractional data compression due to the replacement of a mask list with a series of bits which encode its way. (Recall that a way is just a whole number.) That is, if we assume that the corresponding population list is provided at zero information cost, then dyspoissonism measures the fractional savings afforded by expressing the mask list as its way. Therefore, considering that all mask lists are equiprobable outputs of an unbiased RNG, we can think of dyspoissonism as representing the statistically fair fractional overhead due to encoding the population list, which we have assumed as given for free. Thus the larger the dyspoissonism, the less common the population list in question must be, and therefore the “less random” the mask list.

Given a mask list with mask count Q, mask span Z, and logfreedom L, its dyspoissonism D is given by:

$$D \equiv 1 - \frac{L}{Q \ln Z}$$

where (Q ln Z) is called the “raw entropy” (in nats) of the mask list because it quantifies the amount of information necessary to encode that list, given only Q and Z for free. (Raw entropy is just the log of (Z^Q), which is the way count associated with all unique mask lists constrained only by Q and Z.)

Given the requirement that (Z>1), which is simply to avoid discussion of the trivial case, D cannot exceed (½), but doubling it would misrepresent its role as a compression fraction. D can never (quite) be zero, due to the nonzero fair information overhead of population lists. Its minimum possible value for some particular Q and Z is denoted $D_0(Q, Z)$. Implicitly, D_0 occurs when L is maximized. Unfortunately, L maximization seems to be intractable in the general case, even though binary searches for local maxima can rapidly converge on a good approximation thereof. (Dyspoissometer [4] has functions to do this in addition to a whole suite of dyspoissonism tools. Agnentro [5] implements dyspoissonism using interval arithmetic.)

Dyspoissonism effectively allows us to measure the logscale difference between a pair of mask lists in terms of fractional compressibility. This then allows us to compare various RNGs on an apples-to-apples basis. However, Q and Z should be kept constant so as to avoid distortions due to differences in D_0. As Q increases, an unbiased RNG will cause D to approach zero. If this fails to occur, then the extent of bias is monotonically manifested via the asymptotic nonzero value of D, which answers the problem above involving the futility of logfreedom ranking alone.

3.5. Dyspoissonism is Topological and Empirical

Dyspoissonism is essentially a topological measurement of randomness because it’s concerned with states as opposed to ordered quantities. In other words, it makes no difference that index 5 is greater than index 3; they’re just different states. We could, in principle, hunt for wideband anomalies using floating-point values to analyze a distribution of magnitudes relative to expectation. However, this sort of approach invites unfounded delusions of interpolation, as literally all we know about the generator is the snapshot. Perhaps, in practice, there is some theoretical presumption that the generator is purely Gaussian (normal) or lognormal in its distribution, but reality never

follows such elegant rules. We should therefore use as much information as possible, but without overstepping the bounds of what is actually known. Dyspoissonism fits the bill because it refers only to intervals in the abstract sense, but makes no assumptions about the behavior of the generator within them.

4. Wideband Anomaly Ranking via Dyspoissonism

4.1. Wideband: Definition and Constraints

“Wideband” describes a real-valued vector of physical origin, each scalar component of which assumed to have resulted from a constant – but not necessarily identical – generator. Despite its common use in the context of electromagnetic signals, the term could just as easily apply to the lowest temperatures in Chicago on each day in June. It would not be unreasonable to assume that June is brief enough relative to an entire year that the same underlying statistics govern each day’s temperature. An analog thermometer in the city is thus a chaotic oscillator which is constrained by an essentially constant distribution, but otherwise produces random real-valued measurements from day to day.

Equally so, we could compare those temperatures to the same in any other city, allowing us to establish some notion of thermal distance between Chicago and Mumbai in June, for example. We could do so using Euclidean distance in 30 dimensions, but this technique could fail to detect the difference between the following: good agreement with occasional spikey differences; and overall slightly worse agreement due to pervasive noise, but without such extreme spikes. That distinction is critical: it’s the difference between detecting a single scalar that sticks out like a sore thumb, and detecting a very subtle wideband signal spread across many scalar components of the same vector.

Furthermore, as always, apples-to-apples matters. It would not, for example, be reasonable to compare June to July, and not only because the generators are always materially different even in Chicago itself, but because the

number of measurements would differ – (Q=30) and (Q=31), respectively. In theory, we can compensate for differences in Q by accounting for the implied differences in the distribution of dyspoissonism itself, but doing so is beyond the scope of this paper.

4.2. Preparation

The first step is to measure the oscillator (physical system) repeatedly. Ideally, this should be done a number of times which is a multiple of Z. In practice, (Z=(2^8)) makes sense because in that case each mask will fit in exactly one byte.

Next we need to sort the measurements, resulting in a useful snapshot of the generator. Duplicate measurements are expected to be rare, given sufficient numerical precision. If they occur, they can probably be deleted and replaced with other measurements at the cost of some presumably trivial bias.

If in fact a multiple M of (2^8) measurements are produced, then each successive M measurements can be consolidated into one interval, resulting in a snapshot with (2^8) items. If the total number of measurements is *not* a multiple of Z, then some sort of error distribution or interpolation will be necessary, hopefully without excessively damaging sensitivity.

4.3. Measuring the Dyspoissonism of Each Oscillator

For the moment, we assume that each scalar of each vector is a manifestation of the same generator.

As snapshots, each containing Q measurements, arrive from subsequent oscillators, convert their constituent measurements to interval indexes on [0, (Z-1)] using the “training” snapshot obtained in the previous step. (Q need not be the same as during training, and will generally be much smaller, as comparatively vast amounts of information are required to develop an accurate impression of the generator in the first place.) Then measure the dyspoissonism of each resulting mask list. Finally, rank them in descending order so that the most anomalous wideband signals will appear at the top of

the list.

If in fact all of the subsequent oscillators share the same generator as the first one, then, in the limit of large Q, we expect all the dyspoissonisms to be on the order of the minimum threshold, D_T, given by:

$$D_T \approx 1-(1-D_0)^2$$

which is the expected minimum of D after accounting for the aforementioned Poisson squared bias (hence the power of 2), and where D_0 is as defined above.

This result has been verified over the course of 30,000 trials using a physical ("true") RNG with (Q=(2^16)) and (Z=(2^8)) – not that that amounts to a mathematical proof, but it passes the smell test. Note that the expression for D_T is only approximate because even if D_0 were known exactly (which is generally intractable), there are other intractably complicated combinatorial effects to account for. Nevertheless, in that experiment, its geometric mean was correct to one part in 1000.

4.4. Anomaly Detection with Multiple Distinct Generators

What if we assume that each scalar of a vector originates from a distinct generator? In this case, given some fixed Q and Z, we can still make meaningful comparisons among the implied mask lists. We just need to train on a separate snapshot for each index. Doing so by default could actually result in *increased* sensitivity if in fact we would otherwise have mistakenly assumed generator commonality. After conversion to interval indexes, the process of ranking by dyspoissonism still applies, as does the computation of D_T, with one notable exception.

Because each vector index is associated with a distinct snapshot, biases at different indexes will manifest as preferences for different mask subsets. So for example, in the presence of a wideband anomaly, 3 and 8 might be very common at index zero; whereas 1, 7, and 9 might be very common at index

one. In the aggregate, over all Q vector indexes, we might fail to see any bias at all.

While there is no surefire fix to this problem, reassigning interval indexes by length will improve sensitivity. In other words, within each training snapshot, interval zero is made the shortest, one the next shortest, and so on, such that the infinite antiinterval becomes index (Z-1). The reason is that if there is a sudden disagreement between the generator used for training and some future generator assigned to the same vector index, then this is likely to manifest by new measurements falling disproportionately within the longer intervals defined by the training measurements. Put another way, different generators are not likely to share precisely overlapping spikes, all else being equal.

By the way, in this case, it’s even more important to preprocess the snapshot so as to make it more uniform in distribution. For example, if the measurements are expected to be lognormal, then take their logs first. Otherwise the spikes *will* tend to overlap.

The net effect of all this is that anomalous measurements will preferentially map to higher interval indexes, and quite possibly the highest ones in the case of essentially orthogonal generators. Now we have some crudely consistent behavior among all the vector indexes involved in the production of anomalous measurements. This bias should manifest as a deficit in expected dyspoissonism.

4.5. A Thought Experiment in Radio Astronomy

Imagine that we have a snapshot of the magnitudes of each of Q frequencies. Each magnitude is a measure of the average power received from its corresponding carrier frequency within the same given time window. We expect, absent any other information, that the magnitudes will vary lognormally over time. Therefore we measure them as their logs in order to create a more uniform distribution. But because we have no further

knowledge of what to expect, we proceed to train on a billion separate snapshots of Q measurements each.

We decide to create 1000 different slices, so each slice contains a million sorted measurements, all of which but the extrema being discarded for the sake of speed. We then sort them by length as described above, ending the list with the compulsory infinite antiinterval as a “catchall” for unprecedented extreme values. Thus when each subsequent vector arrives during future observation, each of its Q components will be mapped to a value on [0, 999] by virtue of its respective snapshot residue. (It’s a “residue” because it’s what remains after discarding the measurements other than the extrema which define the intervals.) To be clear, each residue consists of 1000 interval boundary points, and there are Q such residues.

Now, by taking the dyspoissonism of each resulting mask list, we can rank entire spectra by how anomalous they are with respect to the entire set of Q residues. Due to the very nature of dyspoissonism, it’s most likely that high-ranking anomalies will feature unusual behavior spread across many components (frequencies) rather than just one, and perhaps without any of them being anomalously loud in isolation. This isn’t a problem in practice because finding anomalies in a single component is generally straightforward, for example, by comparing the power to some particular threshold.

5. Remarks

When informed by a sufficiently large history of measurements of a chaotic oscillator, the outputs of which being determined by a hidden constant generator, it’s possible to accurately reconstruct that generator. Having done so, any vector of future measurements can be converted into a set of whole numbers representing respective generator slices. Multiple such vectors can then be ranked according to their dyspoissonism in order to detect wideband anomalies, subject to some loss of sensitivity due to a pair of serialized

Poisson jitter processes. If each vector index is associated with its own sufficiently distinct generator, then the same process applied to each index may well yield a more accurate comparison. For maximum accuracy, the training snapshots should be as large as possible, but in any event their residues employed for the sake of vector translation during subsequent observations should have *at most* as many components as the vectors themselves. Otherwise all dyspoissonisms will appear to be minimal, on account of maximum entropy.

Acknowledgements

Special thanks to Greg Hellbourg and Chitwan Kaudan, who work for Breakthrough Listen at the University of California at Berkeley. I would not have bothered to produce this paper, absent their zeal for finding deeply buried intelligent signals. And indeed to Stuart Christmas and Michael Irwin for their relentless quest to find refuge from artificial intelligence in its final frontiers, namely noise and common sense.

Bibliography

[1] “The Terminology of Mask Lists”, http://dyspoissonism.blogspot.com/2015/05/the-terminology-of-mask-lists.html

[2] “The Logfreedom Formula”, http://dyspoissonism.blogspot.com/2015/05/the-logfreedom-formula.html

[3] http://vixra.org/abs/1609.0210

[4] “Source Code”, http://dyspoissonism.blogspot.com

[5] “Source Code”, http://agnentropy.blogspot.com