

# Learning Patterns for Detection with Multiscale Scan Statistics

**James Sharpnack**

*1 Shields Ave. Davis, CA, 95616*

JSHARPNA@UCDAVIS.EDU

**Editors:** Sebastien Bubeck, Vianney Perchet and Philippe Rigollet

## Abstract

This paper addresses detecting anomalous patterns in images, time-series, and tensor data when the location and scale of the pattern and the pattern itself is unknown a priori. The multiscale scan statistic convolves the proposed pattern with the image at various scales and returns the maximum of the resulting tensor. Scale corrected multiscale scan statistics apply different standardizations at each scale, and the limiting distribution under the null hypothesis—that the data is only noise—is known for smooth patterns. We consider the problem of simultaneously learning and detecting the anomalous pattern from a dictionary of smooth patterns and a database of many tensors. To this end, we show that the multiscale scan statistic is a subexponential random variable, and prove a chaining lemma for standardized suprema, which may be of independent interest. Then by averaging the statistics over the database of tensors we can learn the pattern and obtain Bernstein-type error bounds. We will also provide a construction of an  $\epsilon$ -net of the location and scale parameters, providing a computationally tractable approximation with similar error bounds.

**Keywords:** Detection, multiscale scan statistics, pattern adaptation, generic chaining

## 1. Introduction

Detection is the statistical task of determining if there is some structured signal within noisy data. If classification answers the question, “what am I seeing?”, detection answers the question, “do I see anything at all?”. In a sensor network (see [Culler et al. \(2004\)](#)), one is often interested in the dual problems of noticing an anomaly (detection) and then determining its location and extent (classification). Sensor networks are deployed in natural environments for contaminant detection, [Yang et al. \(2009\)](#); [White et al. \(2008\)](#), real-time surveillance ([Caron et al. \(2002\)](#)), radiation monitoring ([Brennan et al. \(2004\)](#)), and fire detection ([Pozo et al. \(1997\)](#)). In medical imaging, the critical task is often to test the existence of an anomaly ([Moon et al. \(2002\)](#); [James et al. \(2001\)](#)). Quick detection of outbreaks of pathogens, [Heffernan et al. \(2004\)](#); [Rotz and Hughes \(2004\)](#), can lead to early intervention. Yet, given that this is such a fundamental task in many applications, the development of detection methodology lags behind the sophisticated tools for classification.

In images, time series, and tensors, it is natural to assume that there is a structured signal, such as blob-like objects, but we do not know its location or scale. In [Figure 1](#), we can see a chemical plume from a multispectral image where each pixel value is lighter if a certain spectral signature is present (see [Manolakis and Shaw \(2002\)](#)). We did not know a-priori where this chemical would appear, nor how large it would be. A natural approach to detecting signals of this type would be to center a regular shape, or pattern, such as a square or an ellipse around every pixel and detect anomalously large quantities of the spectral signature. In this work, we will address the dual problem of detecting anomalous patterns in tensors, and learning regularly occurring patterns in a tensor dataset from a

dictionary. Let’s begin by outlining multiscale scan statistics and the recent advances in adaptive detection.

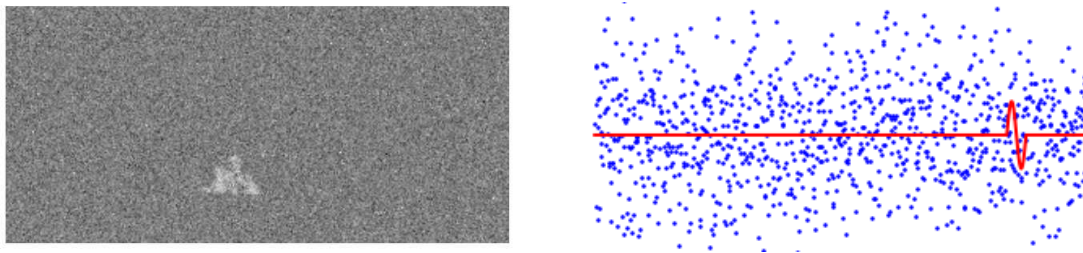


Figure 1: An image with an anomalous region of contaminant (left) and a simulated time series with an embedded sinusoidal signal with values on the  $y$ -axis (right).

The aptly named scan statistic is a test statistic that is based on scanning the image or time series ( $X$ ) for a pattern ( $f$ ) that may be centered anywhere in the domain. For each location ( $t$ ), one can form a likelihood ratio, and test if this statistic exceeds some predetermined threshold (thus, it is a generalized likelihood ratio test). Scan statistics are widely used in spatial detection applications; see [Glaz et al. \(2001\)](#) for a thorough introduction to the topic. They are commonly used to detect patterns in point clouds, [Naus \(1965\)](#); [Neill \(2012\)](#), a closely related problem to our own. Because we often do not know the size as well as the location of the anomaly, we will scan all locations ( $t$ ) and scales ( $h$ ) simultaneously—this is called the *multiscale scan statistic*. Hence, the multiscale scan will translate various scaled versions of a pattern, such as rectangles with many options of side lengths or circles of varying radii. In Neyman-Pearson testing, we attempt to control the probability of false rejection under a null hypothesis, when our data only consists of noise. To this end, we either must approximate the distribution of the scan statistic asymptotically, or set the detection threshold by simulation or resampling. [Siegmund and Worsley \(1995\)](#) provided a weak limit for the scan consisting of all intervals in 1 dimension (other approximations can be found in [Naus and Wallenstein \(2004\)](#) and [Pozdnyakov et al. \(2005\)](#)). In 2 dimensions, [Glaz and Zhang \(2004\)](#); [Haiman and Preda \(2006\)](#); [Wang and Glaz \(2014\)](#); [Kablichko \(2011\)](#), provided approximations of the null distribution for the multiscale scan statistic. [Arias-Castro et al. \(2005, 2011\)](#) analyzed scan statistics for blob-like patterns and determined thresholds for detectability in this context.

It was observed that if one naively tests all scales in the multiscale scan at the same threshold, then the rejection events will be dominated by the finest scale ([Chan and Walther \(2013\)](#)). It was shown in [Dumbgen and Spokoiny \(2001\)](#) that by separately standardizing the 1-dimensional scan at each scale one can detect with a signal-to-noise ratio that adapts to the scale of the anomalous signal. This scale correction for multiscale scan statistics with rectangular shapes was further studied in [Walther \(2010\)](#); [Sharpnack and Arias-Castro \(2016\)](#). [Proksch et al. \(2016\)](#) showed that similar results can be achieved for a large class of patterns that satisfy an average Hölder continuity assumption.

In traditional detection applications, a tensor is scanned for a known function or for specific blob-like patterns. In a database of many tensors with a repeated anomalous pattern in a database then it may be possible to simultaneously test for all of the patterns in a class  $\mathcal{F}$ . The catch is that each time-series or image—more generally, tensor—will have the anomalous pattern in a different

location and at a different scale. For example, we have many time series, all of which have some embedded smooth signal (such as the sinusoid in Figure 1) where the sinusoid begins at various time points, and has different periodicities. Without knowing a priori that we are looking for a sinusoid this problem seems intractable, but given that we know enough about the signal (such as it comes from a finite dictionary of functions), then with enough data we may both learn and detect the pattern within very noisy data.

Previous works have also studied the detection of any pattern from a large class of patterns in a single noisy tensor and other more complicated detection scenarios. The main difference between these works and our own is in the order of operations, namely our methodology proposes a pattern  $f$ , then tests if this patterns occurs in the tensors  $X^i$  for all  $i$  in a multiscale fashion, and repeats this for all patterns in a class; most methods will consider a tensor  $X^i$  (or a subset of tensors) then tests for any pattern within a class. Arias-Castro et al. (2011) studied the detection of Lipschitz deformations of the Euclidean ball in a single image. One can also consider a time series of images to form a single order 3 tensor; this is studied in Kifer et al. (2004), where they detect temporal changepoints of any patterns within a VC class. Detecting intervals in multiple time-series has also been studied in Chan et al. (2015), but this setting differs from our own because it assumes that only a sparse subset of the time series will contain an anomalous interval.

## 1.1. Contributions

We will introduce the multiscale scan statistic, with scale correction, in Section 2, and describe an  $\epsilon$ -net construction using repeated dilation and convolution operations. We begin our theoretical analysis, in Section 3, by proving a chaining result which shows that the standardized supremum of certain subGaussian random fields is a subexponential random variable. This is then used, in Section 4 to provide a finite sample bound on the scale corrected multiscale scan statistic (until now it was only known that it was A.S. bounded). We conclude by demonstrating that our  $\epsilon$ -net construction is indeed correct and a type 2 error control in Section 5.

## 2. Method and Model

### 2.1. Continuous scan statistic

Let's begin with the basic scan statistic over an image. For an image, which can be represented by a matrix,  $\{Y_{k,l} : k, l = -L, \dots, L\}$  we can convolve a pattern  $\{P_{k,l} : k, l = -H, \dots, H\}$  with the image,

$$(P \star Y)_{k,l} = \sum_{k',l'=-H}^H Y_{k-k',l-l'} P_{k',l'}, \quad k, l = -L + H, \dots, L - H.$$

Then the simple scan statistic is  $\max_{k,l} (P \star Y)_{k,l}$  (one can take the absolute value by also considering  $P \leftarrow -P$ ). In the case that  $P_{k,l} = 1/(2H)$  then this scans a square activation pattern over the image. A common assumption in this problem, is that under the null hypothesis,  $Y_{k,l}$  are independent standard subGaussian random variables. For an arbitrary, pattern matrix  $P$  we would like to scale both dimensions, so that  $H \leftarrow H'_j$  in dimension  $j$  for some  $1 \leq H'_j \leq L$  (e.g. stretching the square to form a rectangle). For general patterns, this can be difficult, and one approach is to scale the dimensions for a continuous function,  $f$ , and then rasterize it. This analysis can become

very cumbersome and not terribly enlightening, so we will approximate the pixels with a continuous domain and the image with a random field.

In order to implement our scan, we begin by proposing a given pattern, which is a function  $f \in \mathcal{F}$  over the  $d$ -dimensional domain. For images,  $d = 2$ , and for time-series,  $d = 1$ , but we will only assume that the  $d$  is fixed in our asymptotics. We assume that for every  $f \in \mathcal{F}$ ,  $\|f\|_{L_2} = 1$  and is supported over  $\Omega := [-1, 1]^d$ . We will further assume that  $f$  has continuous gradient ( $\mathcal{F} \subset C^1(\Omega)$ ). For a given field,  $X^i$ , over  $\Omega_L := [-L, L]^d$ , we propose a scale parameter,

$$h \in \mathcal{H} := \times_j [1, L],$$

such that  $h_j$  is the scale parameter for dimension  $j$ . (Throughout  $j = 1, \dots, d$  and  $j$  will always indicate dimension.) Given an  $h \in \mathcal{H}$ , we select

$$t \in \mathcal{T}_h := \times_j [-(L - h_j), L - h_j]$$

respectively, and test if the pattern  $f$  centered at  $t$  and scaled by  $h$  is hidden within image  $X^i$ . This is accomplished by convolving the field,  $X^i$ , with the scaled function  $f_h := h_\bullet^{-1/2} f(\cdot/h)$ ,

$$(f_h \star dX^i)(t) = \int_{\Omega_L} f_h(\tau) dX^i(t - \tau) = \int_{\Omega} \frac{1}{\sqrt{h_\bullet}} f(\tau) dX^i(t - h\tau),$$

where  $h_\bullet = \prod_j h_j$  and vector operations are such that  $t/h$  are performed elementwise. (Throughout,  $f_h(t) = 0$  if  $t$  is outside of the domain of  $f_h$ .) The multiscale scan statistic takes the form,

$$s(X^i; f) := \max_{h \in \mathcal{H}} v_h \left( \max_{t \in \mathcal{T}_h} (f_h \star dX^i)(t) - v_h \right). \quad (1)$$

where in this work we will take  $v_h = \sqrt{2 \sum_j \log(L/h_j)}$ . Hence, when the scale is coarse ( $h_j$  is large) then  $v_h$  is smaller. At the finest scale,  $v_h$  is large, the maximum at this scale concentrates about  $v_h$  with a rate parameter  $v_h$  (meaning that this concentrates more tightly than the pixel noise variance of 1).

Throughout, we will denote the set of all valid scale and location parameters as

$$\mathcal{D} := \{(t, h) : h \in \mathcal{H}, t \in \mathcal{T}_h\}.$$

We will assume that the fields,  $X^i$ , are independent and have additive white noise terms,  $W^i$ . Our noise model is such that  $\int f dW^i$  is a zero mean subGaussian random field indexed by  $f$  with variance  $\|f\|_{L_2}^2$  (see Section 3 for a definition). We will test if the field  $X^i$  is just noise (null hypothesis) or the noise is added to the function,  $f$ , which is translated by  $t^i$  and scaled by  $h^i$ ,

$$\begin{aligned} H_0 : dX^i(\tau) &= dW^i(\tau), i = 1, \dots, n \\ H_1 : dX^i(\tau) &= \mu f_{h^i}(t^i - \tau) d\tau + dW^i(\tau) \\ &\text{for some } f \in \mathcal{F}, \text{ and } (t^i, h^i) \in \mathcal{D}, i = 1, \dots, n. \end{aligned}$$

Critically, the location and scale parameters  $t^i, h^i$ , differ for each image. Notice that  $\|f_h\|_{L_2} = 1$ , so under the null hypothesis, the convolution  $(f_h \star dX^i)(t)$  has mean 0 and variance 1. Because each statistic,  $s(X^i; f)$ , is independent and standardized by  $v_h$ , then we can average these in order

to increase the power of our final test statistic. To this end, let's define the pattern adapted multiscale scan statistic (PAMSS),

$$S_n(X; \mathcal{F}) := \max_{f \in \mathcal{F}} \frac{1}{\sqrt{n}} \sum_{i=1}^n s_n(X^i; f). \quad (2)$$

We will show that  $s_n(X^i; f)$  is subexponential, and so we can obtain probabilistic bounds on  $S_n$  with the subexponential Bernstein-type inequality.

## 2.2. Multiscale $\epsilon$ -net construction

In practice, data is discrete, and the scan statistic must be computed over a finite set of scales and locations. Suppose that for sample,  $X^i$ , we have a draw from the null hypothesis, so that  $\mathbb{E}dX^i(\tau) = \mu f_{h^i}(t^i - \tau)$ . Instead of scanning over all  $(t, h) \in \mathcal{D}$ , we use a finite subset,  $\mathcal{D}_{\text{net}} \subset \mathcal{D}$ . Let  $t', h'$  be values in  $\mathcal{D}_{\text{net}}$  that are close to  $t^i, h^i$  in some sense. Then the expectation of the scan at this approximating location and scale is

$$\mathbb{E}(f_{h'} \star dX^i)(t') = \mu \int f_{h'}(\tau) f_{h^i}(\tau + t^i - t') d\tau.$$

If we define the shift operator  $(S_t f)(\tau) := f(\tau + t)$  then we see that this expectation is  $\mu \langle S_{t'} f_{h'}, S_{t^i} f_{h^i} \rangle_{L_2}$ . Incidentally, this metric appears when we consider the variation of our scan statistic under the null hypothesis,

$$\nu_f((t, h), (t', h')) := \|S_t f_h - S_{t'} f_{h'}\|_{L_2} = (\nabla((f_h \star dW)(t) - (f_{h'} \star dW)(t')))^{\frac{1}{2}},$$

where  $dW$  satisfies our noise assumptions (such as the standard multivariate Brownian motion). This fact will be useful for when we provide type 1 error control for our scan statistic. With this metric we will say that a finite subset  $\mathcal{D}_{\text{net}} \subset \mathcal{D}$  is an  $\epsilon$ -net if for any  $f \in \mathcal{F}$  and any point  $(t, h) \in \mathcal{D}$  there exist points  $(t', h') \in \mathcal{D}_{\text{net}}$  such that  $\nu_f((t, h), (t', h')) \leq \epsilon$ .

The sensitivity of  $\nu_f$  to small changes in  $t, h$  will depend on the smoothness of  $f$ . If  $f$  has large gradients then a small shift,  $S_t$ , can misalign the function with the unshifted version. To this end, we will consider two different notions of smoothness for the functions in our dictionary,  $\mathcal{F}$ . Define the isotropic total variation (recall that the functions have continuous gradients),

$$\|f\|_{\text{TV}} := \int_{\Omega} \|\nabla f(u)\|_2 du.$$

Then we may assume that all of the function are of bounded variation,

$$\exists \gamma_1 > 0 \text{ s.t. } \forall f \in \mathcal{F}, \quad \|f\|_{\text{TV}} \leq \gamma_1. \quad (\text{TVC})$$

The bounded variation assumption is consistent with the assumption in [Dumbgen and Spokoiny \(2001\)](#), although we generalize this work to the multidimensional case. Another notion of smoothness is the average Hölder condition of [Proksch et al. \(2016\)](#). Define the Hölder functional,

$$A_{t,s}(f) := \int_{\Omega_L} |f(t-z) - f(s-z)|^2 dz.$$

Then an alternative to the bounded variation assumption is that

$$\exists 0 < \gamma_2 \leq 1 \text{ s.t. } \forall f \in \mathcal{F}, \quad A_{t,s}(f) \leq c_A \|t - s\|_2^{2\gamma_2} \quad (\text{AHC})$$

where  $c_A$  is some constant. Throughout the paper we will refer to both conditions, and denote  $\gamma$  as either  $\gamma_1$  or  $\gamma_2$  depending on context.

Given that our functions  $\mathcal{F}$  satisfy some smoothness assumptions, we can specify an  $\epsilon$ -net construction. For a given  $\beta > 1$  define  $\ell_{\max} = \lfloor \log_{\beta} L \rfloor$ . We will begin with a subset of scales,

$$\mathcal{H}_{\beta} = \{\beta^{\ell} : \ell = \{0, \dots, \ell_{\max}\}^d\}.$$

where  $\beta > 1$  is some parameter. At a given scale, we will consider a grid of evenly spaced locations, where the distances between grid points increases with the scale,  $h$ ,

$$\mathcal{T}_{\alpha,h} := (\times_j (\alpha h_j \cdot \mathbb{Z})) \cap \mathcal{T}_h.$$

The spacing,  $\alpha > 0$ , is a tuning parameter as well in this construction. With  $\beta, \alpha$  specified, we can consider the  $\epsilon$ -net to be

$$\mathcal{D}_{\beta,\alpha} = \{(t, h) : h \in \mathcal{H}_{\beta}, t \in \mathcal{T}_{\alpha,h}\}.$$

(Notationally, we exchange  $\mathcal{D}_{\text{net}} \leftarrow \mathcal{D}_{\beta,\alpha}$ .) This is similar to the construction in [Sharpnack and Arias-Castro \(2016\)](#), which has a fast implementation on GPUs using a hierarchy of convolution and downsampling layers. The main idea is that instead of expanding the function  $f \rightarrow f_h$  can instead downsample the tensor (but the details on a finite image are somewhat arduous). We expect that a similar implementation is possible in this context, when the functions are adaptively rasterized, but such developments are outside of the scope of this paper.

Given an  $\epsilon$ -net, we can compute the approximate scan by restricting the evaluations to the finite set of location and scales,

$$e_{\beta,\alpha}(X^i; f) := \max_{(t,h) \in \mathcal{D}_{\beta,\alpha}} v_h((f_h \star X^i)(t) - v_h).$$

Then we similarly can define the  $\epsilon$ -net pattern adapted multiscale scan statistic ( $\epsilon$ -PAMSS),

$$E_n(X; \mathcal{F}) = \max_{f \in \mathcal{F}} \frac{1}{\sqrt{n}} \sum_{i=1}^n e_{\beta,\alpha}(X^i; f).$$

It is clear that  $E_n \leq S_n$  since we are maximizing over a strict subset of the continuous scan. Hence, any type 1 error bound on  $S_n$  is also conferred to  $E_n$ .

### 3. A chaining bound for standardized suprema

The scale correction in (1) is based on a precise characterization of the rate and location of a supremum of the random field resulting from the convolution. This standardization complicates the analysis significantly, and until now, it was only known that  $s(X^i; f)$  was almost surely bounded. We initially select the location and scale parameters  $(t, h) \in \mathcal{D}$  and take the supremum over this selection in forming the scan of  $f$  over  $X^i$ . Because under  $H_0$ , the random variables  $s(X^i; f)$  form iid copies of the same random variable, we will seek to exponential concentration inequalities bounding  $s$ . With this in hand, we can hope to obtain PAC-style bounds on their average

and the resulting selection of pattern from a finite function class  $\mathcal{F}$ . We will restrict this work to  $|\mathcal{F}| < \infty$ , but a continuous class of location and scale parameters,  $\mathcal{D}$ . While this bar seems to be set pretty low compared to the rich developments in classification, such as Vapnik-Chernovenkis theory, as we will see, controlling our statistic under finite function classes is a challenging first step to a more complete understanding of learning patterns in detection. Let us begin with a formal definition of a subGaussian random field, and recall that this is what we assume for the field,  $\{\int f_h dW^i(t - \cdot) : t, h \in \mathcal{D}\}$ , which satisfied when  $W^i$  is the  $d$ -dimensional Weiner process.

**Definition 1** *We say that a random field,  $\{Z(\iota)\}_{\iota \in \mathcal{I}}$ , is a (zero mean) standard subGaussian process if there exists a constant  $u_0 > 0$  such that*

$$\mathbb{P}\{|Z(\iota_0) - Z(\iota_1)| \geq u\} \leq 2 \exp\left(-\frac{u^2}{2d_Z(\iota_0, \iota_1)}\right), \quad (3)$$

$$\mathbb{P}\{Z(\iota_0) \geq u\} \leq \exp\left(-\frac{u^2}{2}\right), \quad (4)$$

for any  $\iota_0, \iota_1 \in \mathcal{I}$ ,  $u > u_0$ , and  $d_Z(\iota_0, \iota_1) = \sqrt{\mathbb{E}(Z(\iota_0) - Z(\iota_1))^2}$ , is the canonical distance.

Our noise model assumption can be formally stated as  $\{(f_h \star dW^i)(t) : (t, h) \in \mathcal{D}\}$  is a subGaussian random field with canonical distance  $\nu_f$ .

The generic chaining is a tool for bounding the suprema of random fields with subGaussian tails, [Talagrand \(2006\)](#) (it also has generalizations to other types of concentration). For general subGaussian random fields, the expectation of the supremum is bounded by a quantity,  $\mathbb{E} \sup Z \leq T_Z$ , and one can also show that  $(\sup Z - T_Z)/T_Z$  is a standard subGaussian random variable (using the standard chain construction). Hence, as the expectation bound,  $T_Z$ , increases the scale can also grow with  $T_Z$  (meaning that the supremum becomes more dispersed). This is in contrast to what we know from extreme value theory, where the maximum of independent random variables tends to concentrate more tightly (with a lower asymptotic variance), not less. For example, let  $\{z_i\}_{i=1}^N$  be iid standard normal random variables. Then we know that  $b_N(\max_i z_i - b_N)$  approaches a Gumbel distribution where  $b_N = \sqrt{2 \log n} + o(1)$  is a specific sequence (see [De Haan and Ferreira \(2007\)](#)). Notice that the scale of the max decreases like  $1/b_N$ , which is in contrast to what we obtain from the standard construction in the generic chaining, which has increasing scale. Multiscale scan statistics also have a Gumbel limiting distribution, [Sharpnack and Arias-Castro \(2016\)](#), and so we know that the scale of the statistic should be decreasing, and not increasing. The following theorem significantly modifies the construction in the generic chaining in order to provide an exponential inequality under VC-style conditions on the entropy of the random field,  $Z$ .

**Theorem 2** *Let  $Z(\iota)$  be a standard subGaussian process over an index set  $\mathcal{I}$ . Suppose that the metric space  $(\mathcal{I}, d_Z)$  has the following bound on the  $\epsilon$ -covering number ( $\rho > 0$ ),*

$$\mathcal{N}(\mathcal{I}, d_Z, \epsilon) \leq \Gamma \epsilon^{-\rho}. \quad (5)$$

*Then there exists an  $\Gamma_0 > 0$  such that for any  $\Gamma \geq \Gamma_0$ , the following supremum is bounded in probability,*

$$\mathbb{P}\left\{\sqrt{c_0 \log \Gamma} \left(\sup_{\iota \in \mathcal{I}} Z(\iota) - \sqrt{2 \log \Gamma}\right) - a_0 \log \log \Gamma > u\right\} \leq e^{-u}, \quad (6)$$



for  $u > u_0$  where  $u_0, c_0, a_0$  are constant depending on  $\rho$  (but not on  $\Gamma$ ). In words, the supremum of such a subGaussian process is subexponential with location and rate parameter,  $(2 \log \Gamma)^{1/2}$ .

**Remark 3** For a VC class of sets with metric  $L_2(Q)$  for some measure  $Q$  and VC dimension  $d_V$ , the above entropy bound holds with  $\rho = 2d_V - 1$  and  $\log \Gamma = d_V \log(4e) + \log d_V + K$  for some constant  $K$  (see Thm 2.6.4 of *Van der Vaart and Wellner (1996)*).

**Proof Sketch.** The generic chaining consists of a clever use of the union bound, subGaussian concentration, and a detailed chain construction. In order to illustrate how we can obtain an exponential inequality for the max of subGaussian random variables, let us consider the max of iid standard Gaussian random variables,  $\{z_i\}_{i=1}^N$ . Notice that by union bound and subGaussian concentration,

$$\mathbb{P} \left\{ \max_i z_i > \sqrt{2 \log N + u^2} \right\} \leq N e^{-\log N - \frac{u^2}{2}} = e^{-\frac{u^2}{2}}.$$

Now, instead of bounding,  $\sqrt{2 \log N + u^2} \leq u + (\log N)/u$  as is done in the generic chaining, we will use the bound  $\sqrt{2 \log N + u^2} \leq \sqrt{2 \log N} + u^2 / (2\sqrt{2 \log N})$  (we Taylor expand the square root around  $2 \log N$  instead of around  $u^2$ ). Hence, we obtain,

$$\mathbb{P} \left\{ 2\sqrt{2 \log N} \left( \max_i z_i - \sqrt{2 \log N} \right) > u \right\} \leq e^{-u}.$$

This is the main technique that we use for obtaining subexponential bounds from the max of independent subGaussian random variables.

The chain construction refers to a sequence of partitionings of the space  $\mathcal{I}$ . Given a partition  $\mathcal{A}_k$  of  $\mathcal{I}$ , we let  $A_k(\iota) \in \mathcal{A}_k$  be the element that contains the point  $\iota \in \mathcal{I}$  and  $\Delta(A_k(\iota))$  be its radius. In the standard generic chaining, a partition is called admissible if  $|\mathcal{A}_k| \leq 2^{2^k}$ . Then the supremum is controlled by uniformly bounding the centers of the single element  $A_0 \in \mathcal{A}_0$ , and then the differences between the centers of  $A_k \in \mathcal{A}_k$  and their closest centers in at level  $k - 1$ . The standard result is a bound on the supremum based on a functional of the radii  $\Delta(A_k(\iota))$ . For our modifications, the subexponential bound above requires a growing number of independent points to work, so we begin our chain at a deeper level than at  $k = 0$ . Furthermore, we have to modify the definition of an admissible partition to be  $|\mathcal{A}_k| \leq a^{2^k}$  where  $a \rightarrow 1$  as  $\Gamma \rightarrow \infty$ . Technical details regarding the chain introduce the  $\log \log \Gamma$  term. See the appendix for a complete proof. Although, we assume a specific form for the covering numbers, it may be possible to generalize this technique to other entropy bounds.

#### 4. Type 1 error guarantees for learning patterns

A type 1 error—detecting an anomaly under the null hypothesis—is typically the first error to be controlled in the Neyman-Pearson testing framework. Our statistic  $S_n$  will be compared to a threshold, which is determined through calibration, simulation, or theoretical guarantees. In this section, we will provide a finite-sample probabilistic bound for the multiscale scan statistic with exponential tail probability. We will then use this result to obtain a finite sample bound on the PAMSS,  $S_n$ , which increases logarithmically with the number of functions,  $\log |\mathcal{F}|$ .



**Lemma 4** Suppose that  $f \in \mathcal{F}$  satisfies either (TVC) or (AHC). Let  $\ell \in \{0, \dots, \lfloor \log_2 L \rfloor\}^d$ , and  $\mathcal{H}_2(\ell) = \times_j [2^{\ell_j}, 2^{\ell_j+1}]$ . Then when  $L \geq L_0$ ,

$$\mathbb{P} \left\{ c_1 \cdot \max_{h \in \mathcal{H}_\ell, t \in \mathcal{T}_h} v_h((f_h \star dX^i)(t) - v_h) - a_1 \log \log L > u \right\} \leq e^{-u} \quad (7)$$

for constants  $L_0, a_1, c_1 > 0$  depending on  $\gamma, d$  only.

**Proof** Let  $\mathcal{D}'$  denote the index set in the above display. Throughout, let  $c_1$  and  $a_1$  denote arbitrary constants depending on  $d, \gamma$  alone. By assumption,  $\{(f_h \star dW)(t) : (t, h) \in \mathcal{D}'\}$  is a subGaussian random field with canonical distance  $\nu_f$ . An  $\epsilon$ -net of  $\mathcal{D}$ , by definition, will be an  $\epsilon$ -covering of  $\mathcal{D}' \subset \mathcal{D}$ , so we just need to bound the size of the  $\epsilon$ -net,  $\mathcal{D}'_{\beta, \alpha} := \mathcal{D}' \cap \mathcal{D}_{\beta, \alpha}$ . By construction,

$$|\mathcal{D}'_{\beta, \alpha}| = \sum_{h \in \mathcal{H}_2(\ell) \cap \mathcal{H}_\beta} |\mathcal{T}_{\alpha, h}| \leq c_1 \sum_{h \in \mathcal{H}_2(\ell) \cap \mathcal{H}_\beta} \prod_j \frac{L}{\alpha h_j} \leq c_1 \frac{(L \log_\beta 2)^d}{\alpha^d 2^{\sum_j \ell_j}}.$$

and we can take  $\alpha, \beta$  as specified in the proof of Theorem 7. Furthermore, notice that  $2_j^\ell$  is within a factor of 2 of any  $h_j$  in  $\mathcal{H}_2(\ell)$ . Then, there are constants  $\tilde{C}, \tilde{c}$  such that  $|\mathcal{D}'_{\beta, \alpha}| \leq \tilde{C} L^d / (h_\bullet \epsilon^{\tilde{c}})$ . We can see this because  $\alpha^d (\log \beta)^d \leq \alpha^d (\beta - 1)^d \leq \epsilon^{\tilde{c}}$  for some constant depending on  $d$ . By Theorem 2, we have the subexponential bound for the random variable,

$$\sup_{(t, h) \in \mathcal{D}'} y_h((f_h \star dW)(t) - y_h) \text{ where } y_h = \sqrt{2 \sum_j \log(L/h_j) + 2 \log \tilde{C}}$$

Notice that  $y_h = v_h + O(1/v_h)$  which gives us the result, along with  $\log v_h \leq C \log \log L$ .  $\blacksquare$

From here, we simply apply the union bound with the bound in Lemma 4, for every element in the partitioning  $\mathcal{H} = \cup_\ell \mathcal{H}_2(\ell)$ . There are on the order of  $\log_2(d \log_2 L) = \log_2 \log L + O(1)$  such elements. This gives us

$$\mathbb{P} \left\{ c_2 \cdot \frac{s_n(X^i, f)}{\log \log L} - a_2 > u \right\} \leq e^{-u},$$

for some constants  $c_2, a_2$  depending on  $d, \gamma$ . The  $\log \log L$  is reminiscent of the law of iterated logarithm, and indeed this is a LIL result (but in multiple dimensions). Hence,  $z_i(f) := s_n(X^i, f) / \log \log L$  is subexponential with  $K := \|z_i(f)\|_{\psi_1}$  only depending on  $d, \gamma$  ( $\|\cdot\|_{\psi_1}$  is the Orlicz 1-norm). Thus, by the subexponential Bernstein inequality (Prop. 5.16 in Vershynin (2010)),

$$\mathbb{P} \left\{ \sum_{i=1}^n z_i(f) \geq t \right\} \leq \exp \left( -c_3 \min \left\{ \frac{t^2}{nK^2}, \frac{t}{K} \right\} \right),$$

where  $c_3$  is an absolute constant. This gives us our main result (we absorb  $c_3$  into  $K$  below).

**Theorem 5** Let  $\mathcal{F}$  be finite and assume that either all functions in  $\mathcal{F}$  satisfy either (TVC) or (AHC). Let

$$F_n(\delta) := \begin{cases} \sqrt{K \log \left( \frac{|\mathcal{F}|}{\delta} \right)}, & \log |\mathcal{F}| \leq \frac{n}{K} + \log \delta \\ \frac{K}{\sqrt{n}} \log \left( \frac{|\mathcal{F}|}{\delta} \right), & \log |\mathcal{F}| > \frac{n}{K} + \log \delta \end{cases} \quad (8)$$

then for some constants  $K, L_0$  depending on  $d, \gamma$ , and  $L > L_0$ ,

$$\mathbb{P} \{ S_n(X, \mathcal{F}) > F_n(\delta) \cdot \log \log L \} \leq \delta. \quad (9)$$

Theorem 5 proves this paper's main hypothesis, that we can learn patterns from a finite dictionary where the type 1 error bound increases logarithmically with  $|\mathcal{F}|$ . In fact if  $\log |\mathcal{F}| = o(n)$  then we have subGaussian concentration of the final test statistic  $S_n$ .

## 5. $\epsilon$ -net approximation and type 2 error

We have provided a construction of the  $\epsilon$ -net with parameters  $\alpha > 0$  and  $\beta > 1$ , but we did not specify the selection of either or prove the veracity of our claim that this indeed produces an  $\epsilon$ -net. We used the construction of our  $\epsilon$ -net in the proof of Lemma 4, so we will be careful in this section to prove the correctness of our construction from first principles. The following technical lemma is the main driver of these results.

**Lemma 6** *There is a constant  $C$  depending on  $d$  alone such that*

1. *Suppose that (TVC) holds for the class  $\mathcal{F}$ , then*

$$\nu_f((t, h), (t', h'))^2 \leq C\gamma_1 \left( \left\| \frac{t - t'}{h} \right\|_2^2 + \left\| \frac{h - h'}{h} \right\|_2^2 + \left( \sqrt{\frac{h'_\bullet}{h_\bullet}} - 1 \right)^2 \right).$$

2. *[Proksch et al. (2016)] Suppose that (AHC) holds for the class  $\mathcal{F}$ , then*

$$\nu_f((t, h), (t', h'))^2 \leq C \left( \sum_{j=1}^d \left| \frac{t_j - t'_j}{h_j} \right|^{2\gamma_2} + \sum_{j=1}^d \left| \frac{h_j - h'_j}{\sqrt{h_j h'_j}} \right|^2 + \sum_{j=1}^d \left| \frac{h_j - h'_j}{\sqrt{h_j h'_j}} \right|^{2\gamma_2} \right).$$

Lemma 6.1 is proven using the fact that the metric  $\nu_f$  is the canonical metric for  $S_t f_h$  convolved with the Wiener process. Then by a strategic use of integration by parts we can bound the variance form of  $\nu_f$ . It is immediately clear that we can use this result to prove that  $\mathcal{D}_{\beta, \alpha}$  forms an  $\epsilon$ -net of the space  $(\mathcal{D}, \nu_f)$ .

**Theorem 7** *Suppose that either one of (TVC) or (AHC) holds. Let  $\epsilon > 0$ , then there exists a  $\beta > 1, \alpha > 0$  depending on  $d, \epsilon$  such that for any  $(t, h) \in \mathcal{D}$ , there exists  $(t', h') \in \mathcal{D}_{\beta, \alpha}$  (the  $\epsilon$ -net) with  $\nu_f((t, h), (t', h')) \leq \epsilon$ .*

**Proof** Throughout, let  $\gamma$  mean either  $\gamma_1, \gamma_2$  depending on context. By the triangle inequality let us bound,

$$\nu_f((t, h), (t', h')) \leq \nu_f((t, h'), (t', h')) + \nu_f((t, h), (t, h')).$$

In our construction, notice that for any  $t$  there is a grid point in  $\mathcal{T}_{\alpha, h'}$  that is within  $\alpha h'_j$  from it in dimension  $j$ . Hence,

$$\sum_{j=1}^d \left| \frac{t_j - t'_j}{h_j} \right|^{2\gamma_2} \leq d\alpha^{2\gamma_2}; \quad \sum_{j=1}^d \left| \frac{t_j - t'_j}{h'_j} \right|^2 \leq d\alpha^2.$$

Furthermore, by the  $\epsilon$ -net construction, there exists an  $h'_j$  such that  $|\log h_j/h'_j| \leq \log \beta$  for every  $j$ . Thus,

$$\left| \frac{h_j - h'_j}{\sqrt{h_j h'_j}} \right| \leq \beta - 1; \quad \left| \sqrt{\frac{h'_\bullet}{h_\bullet}} - 1 \right| \leq \beta^{\frac{d}{2}} - 1.$$

Thus there is a constant,  $C$ , depending on  $\gamma$ , such that for  $\epsilon$  small enough  $\alpha = C\epsilon^{1/\gamma_2}$  and  $\beta = C(1 + \epsilon)^{2/d}$  is sufficient.  $\blacksquare$

Suppose that we are under the alternative hypothesis,  $H_1$ , so that there is some embedded signal  $f$  in each image and that the noise is a standard Wiener process. Consider evaluating the scan at the true location and scale,  $(t^i, h^i)$  for a given field  $X^i$ , then  $(f_{h^i} \star dX^i)(t^i)$  is normally distributed with mean  $\mu$  and variance 1. In the event that we scan over an  $\epsilon$ -net, then by arguments in Section 2, there is an element in the approximate scan with mean  $\mu(1 - \epsilon^2/2)$ . Hence, we have the following type 1 error bound, by summing the resulting  $n$  normal random variables.

**Proposition 8** *Suppose that  $\{X^i\}_{i=1}^n$  are drawn from  $H_1$  (with possibly different location and scale parameters) where the noise random field,  $W^i$ , is a standard Wiener process in  $d$  dimensions. Then define  $V_n = \sum_i v_{h^i}^2$  and  $M_n = \sum_i v_{h^i}$ ,*

$$\mathbb{P} \left\{ E_n(X; \mathcal{F}) - \frac{1}{\sqrt{n}} \left( \mu \cdot \left( 1 - \frac{\epsilon^2}{2} \right) M_n - V_n \right) < u \sqrt{\frac{V_n}{n}} \right\} \leq \Phi(u)$$

where  $\Phi$  is the standard Normal CDF. The above display is also true if we let  $\epsilon = 0$  and substitute  $E_n \leftarrow S_n$ .

In order to have diminishing type 1 error probability, we set a threshold for  $S_n$  at  $F_n(\delta) \cdot \log \log L$  for  $\delta \rightarrow 0$ . Assume that  $\epsilon \rightarrow 0$  (however slowly) then to have the type 2 error probability to decrease as well, we require that

$$\frac{\mu M_n - V_n}{\sqrt{n}} - F_n(\delta) \log \log L = \omega \left( \sqrt{\frac{V_n}{n}} \right).$$

For comparison purposes, let's derive other conditions for the PAMSS to be asymptotically powerful (diminishing type 1 and 2 error). Suppose that  $n = 1$  and  $|\mathcal{F}| = 1$  (the standard multiscale scan setting), then this would require

$$\mu - v_{h^1} - \frac{K}{v_{h^1}} \log \frac{1}{\delta} \cdot \log \log L \rightarrow \infty,$$

which is consistent up to constant  $K$  and lower order additive terms with previously known rates, Proksch et al. (2016). Critically, the dependence on  $\delta$  is logarithmic, and this provides the first such exponential finite sample bound (Dumbgen and Spokoiny (2001) provided only an almost sure bound; Proksch et al. (2016) gave a polynomial bound).

We can see that if  $|\mathcal{F}| = 1$  and  $n \rightarrow \infty$  then  $F_n(\delta) = O(1)$ ; if in addition,  $h_j^i \leq L^c$  for some  $0 \leq c < 1$ , then  $V_n/n = \Omega(\log L)$  and  $\log \log L$  is lower order. Then if these conditions hold, then a condition for the PAMSS to be asymptotically powerful is  $\mu - V_n/M_n \rightarrow \infty$ . We find that we can make a similar conclusion even if  $|\mathcal{F}|$  grows in  $n$ .

**Corollary 9** *Suppose that  $\log |\mathcal{F}| = o(n)$ , and recall that under the alternative hypothesis,  $H_1$ ,  $X^i$  has an embedded pattern  $f$  at scale  $h^i$  and  $v_{h^i}^2 = \sum_j \log(L/h_j^i)$ , and the noise is a standard*

Wiener process. Suppose also that  $h_j^i \leq L^c$  for some  $0 \leq c < 1$  for all  $i, j$ , then the PAMSS is asymptotically powerful (has diminishing probability of type 1 and type 2 error) if

$$\mu - \sqrt{2} \cdot \frac{\sum_{i=1}^n v_{h^i}^2}{\sum_{i=1}^n v_{h^i}} \rightarrow \infty. \quad (10)$$

We take this result to mean that as long as the function class,  $|\mathcal{F}|$ , does not grow exponentially in  $n$ , we achieve asymptotic power under the same conditions as if  $|\mathcal{F}| = 1$ .

**Proof** We can see this because, under the assumptions,  $V_n \geq Cn \log L$  for some  $C > 0$  and  $\sqrt{V_n}/M_n \rightarrow 0$ . Furthermore,  $F_n(\delta) = o(\sqrt{n})$  and so  $M_n/\sqrt{n} = \omega(F_n(\delta) \log \log L)$  making the term involving  $|\mathcal{F}|$  lower order. Evaluating  $M_n, V_n$  gives us the result. ■

Let us conclude with a remark about the restrictiveness of the assumption that we have a finite function class  $\mathcal{F}$ . It is known that functions of bounded variation have Haar wavelet coefficients that are bounded in a weak  $\ell_1$  norm, [Cohen et al. \(2003\)](#). It is reasonable to discretize the allowed coefficient values and then restrict our function class to functions with  $k$ -sparse wavelet coefficients of  $m$  then the log-size of the class scales like  $k \log m$  which is very manageable. One advantage with this approach is that the sparse Haar wavelets will naturally satisfy condition (TVC). It is outside of the scope of this work to extend the result to infinite function classes, but this would present a very interesting and important extension.

## 6. Conclusions

We have addressed learning and detecting patterns from a function class,  $\mathcal{F}$ , using multiscale scan statistics. We have introduced the multiscale scan statistic and proved a subexponential concentration bound for it, which relied on a novel chaining result for standardized suprema of subGaussian random fields (a result that may be of independent interest). We introduced the pattern adapted multiscale scan statistic, that can learn patterns in a database of tensors (when the locations and scales vary). This result allowed us to prove Bernstein-type concentration for the PAMSS, meaning that we can learn finite function classes that grow exponentially with the sample size,  $n$ . With evidence that representation learning and detection are not incompatible, we anticipate that efficient methods for learning functions in this setting will emerge, by using modern tools from deep learning and multiscale methods.

**Acknowledgements:** JS is supported in part by NSF DMS-1712996.

## References

- E. Arias-Castro, D.L. Donoho, and X. Huo. Near-optimal detection of geometric objects by fast multiscale methods. *IEEE Trans. Inform. Theory*, 51(7):2402–2425, 2005.
- Ery Arias-Castro, Emmanuel J. Candès, and Arnaud Durand. Detection of an anomalous cluster in a network. *Ann. Statist.*, 39(1):278–304, 2011.
- Sean M. Brennan, Angela M. Mielke, David C. Torney, and Arthur B. Maccabe. Radiation detection with distributed sensor networks. *Computer*, 37(8):57–59, 2004. ISSN 0018-9162. doi: <http://doi.ieeecomputersociety.org/10.1109/MC.2004.103>.

- Y. Caron, P. Makris, and N. Vincent. A method for detecting artificial objects in natural environments. In *Proceedings 16th International Conference on Pattern Recognition*, volume 1, pages 600–603. IEEE Comput. Soc., 2002.
- Hock Peng Chan and Guenther Walther. Detection with the scan and the average likelihood ratio. *Statistica Sinica*, pages 409–428, 2013.
- Hock Peng Chan, Guenther Walther, et al. Optimal detection of multi-sample aligned sparse signals. *The Annals of Statistics*, 43(5):1865–1895, 2015.
- Albert Cohen, Wolfgang Dahmen, Ingrid Daubechies, Ronald DeVore, et al. Harmonic analysis of the space  $bv$ . *Revista Matemática Iberoamericana*, 19(1):235–263, 2003.
- D. Culler, D. Estrin, and M. Srivastava. Overview of sensor networks. *IEEE Computer*, 37(8):41–49, 2004.
- Laurens De Haan and Ana Ferreira. *Extreme value theory: an introduction*. Springer Science & Business Media, 2007.
- Lutz Dumbgen and Vladimir G Spokoiny. Multiscale testing of qualitative hypotheses. *Annals of Statistics*, pages 124–152, 2001.
- Joseph Glaz and Zhenkui Zhang. Multiple window discrete scan statistics. *Journal of Applied Statistics*, 31(8):967–980, 2004.
- Joseph Glaz, Joseph I Naus, and Sylvan Wallenstein. *Scan statistics*. Springer, 2001.
- G Haiman and C Preda. Estimation for the distribution of two-dimensional discrete scan statistics. *Methodology and Computing in Applied Probability*, 8(3):373–382, 2006.
- R. Heffernan, F. Mostashari, D. Das, A. Karpati, M. Kulldorff, and D. Weiss. Syndromic surveillance in public health practice, New York City. *Emerging Infectious Diseases*, 10(5):858–864, 2004.
- D. James, B. D. Clymer, and P. Schmalbrock. Texture detection of simulated microcalcification susceptibility effects in magnetic resonance imaging of breasts. *Journal of Magnetic Resonance Imaging*, 13(6):876–881, 2001.
- Zakhar Kabluchko. Extremes of the standardized gaussian noise. *Stochastic Processes and their Applications*, 121(3):515–533, 2011.
- Daniel Kifer, Shai Ben-David, and Johannes Gehrke. Detecting change in data streams. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, pages 180–191. VLDB Endowment, 2004.
- Dimitris Manolakis and Gary Shaw. Detection algorithms for hyperspectral imaging applications. *IEEE signal processing magazine*, 19(1):29–43, 2002.
- Nathan Moon, Elizabeth Bullitt, Koen van Leemput, and Guido Gerig. Automatic brain and tumor segmentation. In *MICCAI '02: Proceedings of the 5th International Conference on Medical Image Computing and Computer-Assisted Intervention-Part I*, pages 372–379, London, UK, 2002. Springer-Verlag. ISBN 3-540-44224-3.

- Joseph I. Naus. The distribution of the size of the maximum cluster of points on a line. *J. Amer. Statist. Assoc.*, 60:532–538, 1965. ISSN 0162-1459.
- Joseph I Naus and Sylvan Wallenstein. Multiple window and cluster size scan procedures. *Methodology and Computing in Applied Probability*, 6(4):389–400, 2004.
- Daniel B Neill. Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):337–360, 2012.
- Vladimir Pozdnyakov, Joseph Glaz, Martin Kulldorff, and J Michael Steele. A martingale approach to scan statistics. *Annals of the Institute of Statistical Mathematics*, 57(1):21–37, 2005.
- D. Pozo, F.J. Olmo, and L. Alados-Arboledas. Fire detection and growth monitoring using a multi-temporal technique on AVHRR mid-infrared and thermal channels. *Remote Sensing of Environment*, 60(2):111–120, 1997.
- Katharina Proksch, Frank Werner, and Axel Munk. Multiscale scanning in inverse problems. *arXiv preprint arXiv:1611.04537*, 2016.
- L.D. Rotz and J.M. Hughes. Advances in detecting and responding to threats from bioterrorism and emerging infectious disease. *Nature Medicine*, pages S130–S136, 2004.
- James Sharpnack and Ery Arias-Castro. Exact asymptotics for the scan statistic and fast alternatives. *Electronic Journal of Statistics*, 10(2):2641–2684, 2016.
- David O Siegmund and Keith J Worsley. Testing for a signal with unknown location and scale in a stationary gaussian random field. *The Annals of Statistics*, pages 608–639, 1995.
- Michel Talagrand. *The generic chaining: upper and lower bounds of stochastic processes*. Springer Science & Business Media, 2006.
- AW Van der Vaart and JA Wellner. *Weak convergence and empirical processes*. Springer, New York, 1996.
- Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.
- Guenther Walther. Optimal and fast detection of spatial clusters with scan statistics. *Ann. Statist.*, 38(2):1010–1033, 2010. ISSN 0090-5364. doi: 10.1214/09-AOS732.
- Xiao Wang and Joseph Glaz. Variable window scan statistics for normal data. *Communications in Statistics-Theory and Methods*, 43(10-12):2489–2504, 2014.
- Brian A White, Antonios Tsourdos, Immanuel Ashokaraj, S Subchan, and Rafal Zbikowski. Contaminant cloud boundary monitoring using network of uav sensors. *IEEE Sensors Journal*, 8(10):1681–1692, 2008.
- Y Jeffrey Yang, Roy C Haught, and James A Goodrich. Real-time contaminant detection and classification in a drinking water pipe using conventional water quality sensors: Techniques and experimental results. *Journal of environmental management*, 90(8):2494–2506, 2009.

## Appendix A. Proof of Theorem 2

**Proof** We will follow the construction of generic chains as in Talagrand (2006), but will be significantly more careful about the details of the construction. This proof is similar in spirit to Dumbgen and Spokoiny (2001), but we will get probabilistic bounds and prefer this proof because it uses only first principles. Throughout this proof we will call variables that dependent only on  $\rho$ , constants, and some like  $C$  may change from line to line. Variables from the main body of the paper, other than those defined in Theorem 2, may appear in this proof and mean something different (we suppose that they are in a different scope).

Let's begin by defining  $G = \Gamma^{1/\rho}$  and

$$a = (1 - \log^{-1} G)^{-1} > 1.$$

Let an admissible partition,  $\mathcal{A}_k$ , be any partition of  $\mathcal{I}$  of size at most  $a^{a^k}$ . Let  $A_k(\iota)$  be the element of the partition containing  $\iota$  and let the center of this element be

$$\tau_n(\iota) := \inf_{\tau \in A_n(\iota)} \sup_{\iota' \in A_n(\iota)} d_Z(\tau, \iota'),$$

and  $\Delta(A_k(t)) = \sup_{\iota \in A_k(t)} d_Z(\tau_k(\iota), \iota)$  be its radius. Let  $k_0 = 1, \dots, \infty$  then  $Z(\iota) \leq Z(\tau_{k_0}(\iota)) + \sum_{k=k_0}^{\infty} |Z(\tau_{k+1}(\iota)) - Z(\tau_k(\iota))|$ . By the union bound and (4),

$$\begin{aligned} \mathbb{P} \left\{ \exists \iota \in \mathcal{I} : Z(\tau_{k_0}(\iota)) > \sqrt{2 \log a \cdot a^{k_0} + u^2} \right\} &\leq a^{a^{k_0}} \exp \left( -\frac{1}{2} \left( 2 \log a \cdot a^{k_0} + u^2 \right) \right) \\ &= e^{-\frac{u^2}{2}}. \end{aligned}$$

Let

$$\epsilon_k = G \cdot a^{-\frac{a^k}{\rho}}, \quad (11)$$

then there exists an admissible partition where the radius of the balls are  $\epsilon_k$  satisfy  $\mathcal{N}(\mathcal{I}, d_Z, \epsilon_k) \leq a^{a^k}$ . Hence,

$$\begin{aligned} \mathbb{P} \left\{ \exists \iota \in \mathcal{I} : |Z(\tau_{k+1}(\iota)) - Z(\tau_k(\iota))| > \epsilon_k \sqrt{2 \log a \cdot a^{k+1} + d_k u^2} \right\} \\ \leq 2a^{a^{k+1}} \exp \left( -\frac{1}{2} \left( 2 \log a \cdot a^{k+1} + d_k u^2 \right) \right) \\ = 2e^{-\frac{d_k u^2}{2}}, \end{aligned}$$

where

$$d_k = e^{\frac{1}{2}(1+\log a) \cdot (k-k_0)}. \quad (12)$$

Define the quantities,

$$\begin{aligned} A_{k_0} &= \sqrt{2 \log a \cdot a^{k_0} + u^2} + \sum_{k=k_0}^{\infty} \epsilon_k \sqrt{2 \log a \cdot a^{k+1} + d_k u^2}, \\ B_{k_0} &= \exp \left( -\frac{u^2}{2} \right) + \sum_{k=k_0}^{\infty} \exp \left( -d_k \frac{u^2}{2} \right), \end{aligned}$$



so that

$$\mathbb{P}\{\sup_{\iota \in \mathcal{I}} Z(\iota) > A_{k_0}\} \leq 2B_{k_0}. \quad (13)$$

Because  $\sqrt{b_1 + b_2} \leq \sqrt{b_1} + b_2/(2\sqrt{b_1})$  for  $b_1, b_2 > 0$  we have

$$A_{k_0} \leq a_0 a^{k_0/2} + \frac{a^{-k_0/2}}{2a_0} u^2 + \sum_{k=k_0}^{\infty} \left( a_0 \epsilon_k a^{\frac{k+1}{2}} + \frac{a^{-\frac{k+1}{2}}}{2a_0} \epsilon_k d_k u^2 \right) \quad (14)$$

for  $a_0 = \sqrt{2 \log a}$ . Let  $k_0$  satisfy

$$a^{k_0} \leq \rho(\log_a G + k_0) \leq a^{k_0+1}, \quad (15)$$

such that

$$k_0 \log a \leq C' \log \log G$$

for some constant  $C'$  possibly depending on  $\rho$ , which is guaranteed by Lemma 10 for  $\Gamma$  large enough. Then we have that

$$a_0 a^{\frac{k_0}{2}} \leq \sqrt{2 \log \Gamma + 2\rho k_0 \log a} \leq \sqrt{2 \log \Gamma} + \frac{\rho k_0 \log a}{\sqrt{2 \log \Gamma}} \leq \sqrt{2 \log \Gamma} + \frac{C' \log \log \Gamma}{\sqrt{2 \log \Gamma}} \quad (16)$$

Furthermore, by (11) and (15)

$$\begin{aligned} \sum_{k=k_0}^{\infty} a^{\frac{k+1}{2}} \epsilon_k &\leq G \sum_{k=k_0}^{\infty} a^{-\frac{a^k}{\rho} + \frac{k+1}{2}} = G a^{\frac{k_0+1}{2}} \sum_{j=0}^{\infty} a^{-a^{k_0} \frac{a^j}{\rho} + \frac{j}{2}} \\ &\leq G a^{\frac{k_0+1}{2}} \sum_{j=0}^{\infty} a^{-(\log_a G + k_0) a^{j-1} + \frac{j}{2}} = a^{\frac{1-k_0}{2}} \sum_{j=0}^{\infty} a^{(\log_a G + k_0)(1-a^{j-1}) + \frac{j}{2}} \\ &= a^{\frac{1-k_0}{2}} \left( (G a^{k_0})^{1-a^{-1}} + \sqrt{a} \sum_{j=0}^{\infty} a^{(\log_a G + k_0)(1-a^j) + \frac{j}{2}} \right). \end{aligned}$$

The first term is

$$a^{\frac{1-k_0}{2}} (G a^{k_0})^{1-a^{-1}} = e \cdot a^{\frac{1-k_0}{2}} a^{\frac{k_0}{\log G}}.$$

Furthermore, for any  $C < 1$  there exists a  $G_0 > 0$  such that  $\log a \geq C/\log G$  for all  $G \geq G_0$ , thus

$$a^{\frac{k_0}{\log G}} \leq (\rho(\log_a G + k_0))^{\frac{1}{\log G}} \leq (\rho(C \log^2 G + k_0))^{\frac{1}{\log G}} \rightarrow 1,$$

as  $G \rightarrow \infty$ . Considering the second term,

$$\sum_{j=0}^{\infty} a^{(\log_a G + k_0)(1-a^j) + \frac{j}{2}} \leq 1 + \sum_{j=1}^{\infty} a^{(\log_a G)(1-a^j) + \frac{j}{2}}.$$

Using the first order approximation,

$$a^j - 1 = \left( 1 - \frac{1}{\log G} \right)^{-j} - 1 \geq \frac{j}{\log G}.$$

So we may bound the series,

$$\sum_{j=0}^{\infty} a^{(\log_a G)(1-a^j)+\frac{j}{2}} \leq \sum_{j=0}^{\infty} \exp\left(\left(\frac{\log a}{2} - 1\right)j\right) \leq \left(1 - \exp\left(\frac{\log a}{2} - 1\right)\right)^{-1} \rightarrow \frac{1}{1 - e^{-1}},$$

as  $G \rightarrow \infty$ . Hence, there are constants  $C, G_0 > 0$  depending on  $\rho$  such that for  $G > G_0$ ,

$$\sum_{k=k_0}^{\infty} a_0 \epsilon_k a^{\frac{k+1}{2}} \leq C a_0 a^{-\frac{k_0}{2}} \leq C \sqrt{\frac{2a \log a}{\rho(\log_a G + k_0)}} \leq \frac{C\sqrt{2a} \log a}{\sqrt{2\rho \log G}}. \quad (17)$$

Also,

$$\frac{a^{-\frac{k_0}{2}}}{a_0} \leq \frac{C}{\sqrt{2\rho \log G}}. \quad (18)$$

Consider the fourth term in  $A_{k_0}$ ,

$$\begin{aligned} \sum_{k=k_0}^{\infty} a^{-\frac{k+1}{2}} \epsilon_k d_k &= G \sum_{k=k_0}^{\infty} a^{-\frac{k+1}{2}} a^{-\frac{a^k}{\rho}} e^{\frac{1}{2}(1+\log a)\cdot(k-k_0)} \\ &\leq G a^{-\frac{1}{2}(k_0+1)} \sum_{j=0}^{\infty} a^{-\frac{j}{2}} a^{-(\log_a G + k_0)a^{j-1}} e^{\frac{1}{2}(1+\log a)\cdot j} \\ &\leq a^{-\frac{3}{2}k_0 - \frac{1}{2}} \sum_{j=0}^{\infty} a^{-\frac{j}{2}} a^{(\log_a G + k_0)\cdot(1-a^{j-1})} e^{\frac{1}{2}(1+\log a)\cdot j} \end{aligned}$$

Isolating the summation,

$$\begin{aligned} \sum_{j=1}^{\infty} a^{(\log_a G + k_0)\cdot(1-a^{j-1})} e^{\frac{j}{2}} &\leq \sqrt{e} \sum_{j=0}^{\infty} \exp\left(\left(\frac{1}{2} - \frac{\log G + k_0 \log a}{\log G}\right) \cdot j\right) \\ &\leq \frac{\sqrt{e}}{1 - 1/\sqrt{e}} \leq 4.2. \end{aligned}$$

Furthermore,

$$a^{(\log_a G + k_0)\cdot(1-a^{-1})} \leq e^{-\left(1 + \frac{k_0}{\log_a G}\right)} \leq \frac{1}{e} < 0.4.$$

Hence, there exists a  $G_0 > 0$  such that

$$\frac{u^2}{a_0} \sum_{k=k_0}^{\infty} a^{-\frac{k+1}{2}} \epsilon_k d_k \leq u^2 \frac{4.6}{\sqrt{2 \log a}} \cdot a^{-\frac{3}{2}k_0 - \frac{1}{2}} \leq \frac{u^2}{\sqrt{\rho \log G}}, \quad (19)$$

for  $G \geq G_0$ . Hence, by (14) combined with (16), (17), (18), and (19) yields

$$A_{k_0} \leq \sqrt{2 \log \Gamma} + \frac{a_0 \log \log \Gamma + u^2/2}{\sqrt{c_0 \log \Gamma}},$$

for constants  $C_0, C_1, G_0$  with  $G \geq G_0$ . Consider  $B_{k_0}$ ,

$$\exp\left(\frac{u^2}{2}\right) \cdot \sum_{k=k_0}^{\infty} \exp\left(-d_k \frac{u^2}{2}\right) = \sum_{j=0}^{\infty} \exp\left(\left(1 - \exp\left(\frac{1}{2}(1 + \log a) \cdot j\right)\right) \cdot \frac{u^2}{2}\right) \quad (20)$$

$$\leq \sqrt{e} \sum_{j=0}^{\infty} \exp\left(-\frac{1}{2} \exp\left(\frac{1}{2}j\right)\right) = \sqrt{e} \sum_{j=0}^{\infty} b^{-b^j} \leq 2.4, \quad (21)$$

$$(22)$$

where  $b = \sqrt{e}$ . Hence,

$$B_{k_0} \leq C \cdot e^{-\frac{1}{2}u^2}.$$

Recalling (13) gives us our result. ■

**Lemma 10** *Let  $\rho > 0$ , and define the function*

$$a(G) = \left(1 - \frac{1}{\log G}\right)^{-1}. \quad (23)$$

*There exists a constant  $G_0$  depending on  $\rho$  alone such that if  $G > G_0$  then there exists an integer  $k_0$  satisfying*

$$a^{k_0} \leq \rho(\log_a G + k_0) \leq a^{k_0+1}, \quad (24)$$

*and there is a constant  $C$  depending on  $\rho$  such that  $k_0 \log a \leq C \log \log G$ .*

**Proof** Set  $G > 0$  and  $a = a(G)$ . Let  $\kappa_1$  be the root of the following function,

$$f(\kappa) := a^\kappa - \rho(\log_a G + \kappa).$$

Notice that  $f$  is strictly convex and so let  $\kappa_0$  be its unique minimizer.

$$\kappa_1 = \log_a(\rho(\log_a G + \kappa_1)) \geq \log_a(\rho \log_a G),$$

Notice that  $\log a \leq 1/\log G$ , hence,

$$\log_a G = \frac{\log G}{\log a} \geq \log^2 G,$$

and

$$\kappa_1 \geq \frac{\log \rho}{\log a} + 2 \frac{\log \log G}{\log a} \geq \log G \cdot (\log \rho + 2 \log \log G) =: x_1.$$

Moreover,  $\kappa_0$  satisfies  $\kappa_0 a^{\kappa_0} = \rho$ ,

$$(\kappa_1 - 1)a^{\kappa_1-1} = \frac{\rho}{a}(\kappa_1 - 1)(\log_a G + \kappa_1) \geq \frac{\rho}{a}(x_1 - 1)(\log^2 G + x_1).$$

The limit of the right hand side approaches  $+\infty$  as  $G$  increases. Hence, there exists an  $G_0 > 0$  depending on  $\rho$  only such that for any  $G \geq G_0$ ,

$$(\kappa_1 - 1)a^{\kappa_1-1} \geq \rho = \kappa_0 a^{\kappa_0}.$$

Because  $\kappa a^\kappa$  is a monotone increasing function of  $\kappa > 0$ , this implies that  $\kappa_1 - 1 \geq \kappa_0$ .

Let  $k_0 = \lfloor \kappa_1 \rfloor$ . We know that  $f$  is increasing and convex over  $[\kappa_0, \infty)$ . Let  $G \geq G_0$ . Because  $\kappa_0 \leq k_0 \leq \kappa_1 \leq k_0 + 1$  we have that  $f(k_0) \leq 0 \leq f(k_0 + 1)$ . Hence,

$$a^{k_0} - \rho(\log_a G + k_0) \leq 0 \leq a^{k_0+1} - \rho(\log_a G + k_0 + 1) \leq a^{k_0+1} - \rho(\log_a G + k_0),$$

which proves the first result. Finally,

$$\log(a^{2k_0}) \leq \log\left(a^{k_0} - \rho k_0\right) \leq \log(\rho \log_a G)$$

for some  $G(k_0)$  large enough. Notice that  $\log_a G \leq C(\log G)^2$  by the limit of  $\log a \rightarrow 1/\log G$ , demonstrating the second result.  $\blacksquare$

## Appendix B. Remaining proofs

**A note on the derivation of Theorem 5.** Notice that  $t^2/nK < t/K$  when  $t < n$ , and that let

$$t = \sqrt{\frac{nK}{c_3} \log \frac{|\mathcal{F}|}{\delta}}.$$

Then

$$|\mathcal{F}| \exp\left(-c_3 \frac{t^2}{nK^2}\right) \leq \delta$$

and  $t < n$  when

$$\log |\mathcal{F}| \leq \frac{n}{c_3 K} + \log \delta.$$

In the other case, we define  $t$  so that the probabilistic bound holds, and absorb  $c_3$  into  $K$ .

**Proof of Lemma 6.** (2) is proven in Proksch et al. (2016) (pg. 32) so we will focus on (1). We will partially follow the arguments in Dumbgen and Spokoiny (2001) (pg. 145). Let  $W$  denote a standard Wiener process in  $d$  dimensions. Notice that we can define  $\nu_f$  as,

$$\nu_f((t, h), (t', h'))^2 := \mathbb{V} \left( \int_{\Omega_L} f_h(t - z) dW(z) - \int_{\Omega_L} f_{h'}(t' - z) dW(z) \right).$$

Recall that

$$(f_h \star dW)(t) = \frac{1}{\sqrt{h_\bullet}} \int_{\Omega} f(u) dW(t - hu).$$

By integration by parts,

$$\frac{1}{\sqrt{h_\bullet}} \int_{\Omega} f(u) dW(t - hu) = \int_{\Omega} \frac{W(t - hu)}{\sqrt{h_\bullet}} \cdot \nabla f(u) du.$$

Therefore,

$$\begin{aligned}
 (f_h \star dW)(t) - (f_{h'} \star dW)(t') &= \int_{\Omega} \left( \frac{W(t-hu)}{\sqrt{h_{\bullet}}} - \frac{W(t'-h'u)}{\sqrt{h'_{\bullet}}} \right) \cdot \nabla f(u) du \\
 &\leq \int_{\Omega} \left\| \frac{W(t-hu)}{\sqrt{h_{\bullet}}} - \frac{W(t'-h'u)}{\sqrt{h'_{\bullet}}} \right\|_2 \cdot \|\nabla f(u)\|_2 du \\
 &\leq \sup_{u \in \Omega} \left\| \frac{W(t-hu)}{\sqrt{h_{\bullet}}} - \frac{W(t'-h'u)}{\sqrt{h'_{\bullet}}} \right\|_2 \cdot \int_{\Omega} \|\nabla f(u)\|_2 du.
 \end{aligned}$$

Hence,

$$\nu_f((t, h), (t', h'))^2 \leq \|f\|_{\text{TV}}^2 \cdot \mathbb{E} \sup_{u \in \Omega} \left\| \frac{W(t-hu)}{\sqrt{h_{\bullet}}} - \frac{W(t'-h'u)}{\sqrt{h'_{\bullet}}} \right\|_2^2.$$

Decompose the supremum term,

$$\begin{aligned}
 &\mathbb{E} \sup_{u \in \Omega} \left\| \frac{W(t-hu)}{\sqrt{h_{\bullet}}} - \frac{W(t'-h'u)}{\sqrt{h'_{\bullet}}} \right\|_2 \\
 &\leq \mathbb{E} \sup_{u \in \Omega} \left( \left\| \frac{W(t-hu) - W(t'-h'u)}{\sqrt{h_{\bullet}}} \right\|_2 + \|W(t'-h'u)\|_2 \left| \frac{1}{\sqrt{h_{\bullet}}} - \frac{1}{\sqrt{h'_{\bullet}}} \right| \right)
 \end{aligned}$$

By Brownian scaling,  $W(t-hu) - W(t'-h'u) / \sqrt{h_{\bullet}} \stackrel{d}{=} W\left(\frac{t}{h} - u\right) - W\left(\frac{t'}{h} - \frac{h'}{h}u\right)$ . Hence, the  $j$ th coordinate in the above LHS is a 1-dimensional Brownian motion equal in distribution to,  $B\left(\frac{t_j - t'_j}{h_j} - \left(1 - \frac{h'_j}{h_j}\right)u_j\right)$ . By the reflection principle, we have that for some constant  $c'$ ,

$$\mathbb{E} \sup_{u \in \Omega} B\left(\frac{t_j - t'_j}{h_j} - \left(1 - \frac{h'_j}{h_j}\right)u_j\right)^2 \leq c' \left( \left| \frac{t_j - t'_j}{h_j} \right|^2 + \left| \frac{h_j - h'_j}{h_j} \right|^2 \right).$$

Hence,

$$\mathbb{E} \sup_{u \in \Omega} \left\| \frac{W(t-hu) - W(t'-h'u)}{\sqrt{h_{\bullet}}} \right\|_2^2 \leq c' \left( \sum_j \left| \frac{h_j - h'_j}{h_j} \right|^2 + \left| \frac{t_j - t'_j}{h_j} \right|^2 \right).$$

Without loss of generality, we can select  $t = 0$  by translation invariance of  $\nu_f$ . Furthermore by a similar argument,  $\mathbb{E} \sup_u \|W(h'u)\|_2^2 \leq c'h'_{\bullet}$ , for another universal constant  $c'$ . Hence,

$$\mathbb{E} \sup_{u \in \Omega} \|W(t'-h'u)\|_2 \left| \frac{1}{\sqrt{h_{\bullet}}} - \frac{1}{\sqrt{h'_{\bullet}}} \right| \leq c' \left| \sqrt{\frac{h'_{\bullet}}{h_{\bullet}}} - 1 \right|.$$

Combining these bounds completes the proof. ■