On high-dimensional support recovery and signal detection

On High-dimensional Support Recovery and Signal Detection

Research supported by the Netherlands Organisation for Scientific Research (NWO) through the NWO Grant 613.001.114.

On High-dimensional Support Recovery and Signal Detection / Tánczos, Ervin

A catalogue record is available from the Eindhoven University of Technology Library.
ISBN: 978-90-386-4129-4

Printed by Gildeprint Drukkerijen - www.gildeprint.nl

On High-Dimensional Support Recovery and Signal Detection

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Technische Universiteit Eindhoven, op gezag van de rector magnificus prof.dr.ir. F.P.T. Baaijens, voor een commissie aangewezen door het College voor Promoties, in het openbaar te verdedigen op dinsdag 13 september 2016 om 16:00 uur

door

Ervin Tamás Tánczos

geboren te Szeged, Hongarije

Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

voorzitter:	prof.dr.ir. B. Koren
1^{e} promotor:	prof.dr. R.W. van der Hofstad
copromotor:	dr. R.M. Pires da Silva Castro
leden:	prof.dr. G. Lugosi (Universitat Pompeu Fabra)
	prof.dr. R.D. Nowak (University of Wisconsin-Madison)
	prof.dr. P. Grünwald (Universiteit Leiden)
	prof.dr. J.H. van Zanten (Universiteit van Amsterdam)
	prof.dr. E.R. van der Heuvel

Het onderzoek of ontwerp dat in dit proefschrift wordt beschreven is uitgevoerd in overeenstemming met de TU/e Gedragscode Wetenschapsbeoefening.

Acknowledgments

Here goes...

First I would like to thank my supervisors, Rui Castro and Remco van der Hofstad, for the opportunity to work with you in Eindhoven. I can't even enumerate all the things this journey has given me. So I won't. In short, these past years have been a blast, and I'm really grateful to you for making this happen.

Also, I have to mention that my scientific career would have probably come to an early halt if it had not been for the folks at the University of Szeged: Boda Kriszta, Bari Ferenc, Kincses Tamás and Szabó Botond. Thus, I am thankful for all you've done for me in the years I've spent at SZTE DMI.

Research-wise I was mostly in contact with Rui. Thanks for throwing a bunch of interesting stuff at me to think about, for being a supportive, inspiring and friendly supervisor, and for making me believe what my grandpa used to say - that I'm not as stupid as I appear to be. I learned a lot while working with you and I hope that we continue collaborating in the future. And also share a few more glasses of whiskey.

I spent a considerable time traveling during the PhD project and I feel inclined to thank Rob Nowak and Ery Arias-Castro for being my hosts in the US. During this long trip I've been able to cheer for the Pack at Lambeau and watch the Blue Angels zoom above my head from a balcony in San Diego. To be fair we also did some math while I was there... but these memories will stick forever.

I'm also indebted to the members of my defense committee, Gábor Lugosi, Robert Nowak, Peter Grünwald, Harry van Zanten and Edwin van der Heuvel. Thank you for accepting to be on the committee, reading my thesis and providing valuable feedback.

I'd like to thank everyone at the STO group of TU/e for the pleasant environment at work. Special thanks to my roommates, Botond and Paolo, and later

Sanyi and Robert, for putting up with me always opening the windows.

While in Eindhoven, I had the privilege to meet a fellow PhD student (now doctor), Emil, and his girlfriend (now wife), Kati. Thank you both for the fun times - the dinners, the games and for keeping us company here.

A round of gratitude goes out to friends at home with whom we were able to keep things going despite the distance between us. To Robi for the afternoon naps. To Tádé for thrashing. To Dzsúl, Robi, Peti and Kincses for all the főmanó's. To Fumiga for the poetic evenings we spent enjoying high-class movies. To Johnny for the Impetus battles.

A huge thanks goes to all members of my family for a bunch of miscellaneous things throughout not just the past four years, but all three decades I've been around. I'd like to give special credit to my mother for making us feel at home every time we went back to Szeged. And the apple pie. Very special credit for the apple pie.

Preserving my sanity was an effort mostly undertaken by Dóri (and Madve). Obviously, this feat warrants the biggest gratitude on my part. Thank you for the time we spent together, and I for one am really looking forward to what is to come.

Notations

$[n]$	the set of natural numbers $\{1, \ldots, n\}$
$[i, j]$	the set of natural numbers $\{i, \ldots, j\}$
$\langle u, v \rangle$	the scalar product of the vectors u and v
$\mathbf{1}\{A\}$	the indicator of the event A
$\mathbf{1}_S$	the indicator vector of the set S
$\mathrm{Unif}(S)$	the Uniform distribution on the set S
$\mathrm{Ber}(p)$	the Bernoulli distribution with parameter p
$\mathrm{Geom}(p)$	the Geometric distribution with parameter p
$\mathrm{Bin}(n, p)$	the Binomial distribution with parameters n, p
$N(\mu, \sigma^2)$	the Normal distribution with mean μ and variance σ^2
$TV(\mathbb{P}, \mathbb{Q})$	the Total Variation distance between distributions $\mathbb{P}$ and $\mathbb{Q}$
$D(\mathbb{P} \Vert \mathbb{Q})$	the Kullback-Leibler divergence between distributions $\mathbb{P}$ and $\mathbb{Q}$
$\chi^2(\mathbb{P}, \mathbb{Q})$	the Chi-squared divergence between distributions $\mathbb{P}$ and $\mathbb{Q}$

Contents

Chapter 1

Introduction

1.1 Primer on main notions

Support Recovery and Signal Detection

Suppose we have a large number of individuals, some of whom might have a certain disease without any visible symptoms. Though the disease might not be producing any symptoms yet, if left unchecked it can activate resulting in undesired consequences. Let us imagine we are allowed to take blood samples from the individuals, which can later be subject to analysis. One question we might ask is whether there is anyone among them carrying the infection. We refer to such questions as a *signal detection* problem, as we are only interested in the presence of infection, but not the identity of infected subjects. This can be thought of as a first step in examining such a system, as we are only aim to raise a red flag in case we see a deviation from the nominal state - which in this case is the absence of infection. A second step would be to determine exactly who is infected and who is not, which we call a *support recovery* problem.

The first difficulty in such tasks lies in the fact that we usually can not observe the true state of the individuals directly, but instead there is some sort of noise contaminating our observations. For instance, in the example above such uncertainty can arise during the analysis of the blood samples. The second problem one often runs into is that typically there is a very large number of people that need to be screened, but in reality only a few of them are infected. This combined with measurement uncertainty makes it easier for infected individuals to "hide in the

crowd". These are the types of questions we deal with throughtout this thesis.

Let us now discuss the nomenclature for the main objects in such problems. Let $\boldsymbol{x} \in \mathbb{R}^n$ be an unknown vector, which we refer to as the *signal*. The signal $\boldsymbol{x}$ depicts the true state of affairs. In the example above for instance, $\boldsymbol{x}_i$, the ith component of $\boldsymbol{x}$, would tell us the level of infection of the ith individual. In a number of applications the *extrinsic signal dimension* $n \in \mathbb{N}$ is large, but the signal is assumed to be living in a much lower dimensional subspace of $\mathbb{R}^n$. There are multiple ways of modeling such a phenomenon and the one we will be dealing with throughout this work is *sparsity*, which simply means that the majority of the components of $\boldsymbol{x}$ are zero. This is often a reasonable assumption to make. For instance, in the example above we need to screen a large number of individuals but probably only a few of them are actually infected. Also, when trying to identify genes which regulate a certain biological process, although there might be tens of thousands of genes in total, we believe that only a few dozen or hundred actually play a part in that particular process. Formally, we assume $\boldsymbol{x}_i = 0$ if and only if $i \notin S$ for some $S \subset [n]$ that we call the *support* of the signal. S is an element of a *class of supports* $\mathcal{C}$ with the additional property that the cardinality of the sets belonging to $\mathcal{C}$ is small compared to n. For now we can think of $\mathcal{C}$ as the collection of all subsets of $[n]$ with a given sparsity s, and s being much smaller than n.

As noted before, $\boldsymbol{x}$ can not be observed directly, but instead through some kind of a measurement mechanism. Since this represents some sort of uncertainty, it seems natural to model this as having access to a random variable $\mathbf{Y} \in \mathcal{Y}$ that we call the *observations* or *sample*. Note that although we simply use one symbol $\mathbf{Y}$ for the observations, this can (and often does) represent multiple measurements. We denote the distribution generating the observations by $\mathbb{P}_S$. Note that this depends on the signal $\boldsymbol{x}$ and thus also on the support S in some way, and since S is our main object of interest we emphasize the dependence with a subscript. For now we do not specify these objects any further, but if the reader would like to have something concrete in mind, recall the example above and think of $\mathbf{Y}$ as an element of $\mathbb{R}^n$, each component corresponding to the result of the blood sample analysis of one individual.

Now we are in a position to formulate the problems under consideration. Our main object of interest is the support of $\boldsymbol{x}$, but this is not directly observable to us. However, as seen before our observations do depend in some way on the true support, so our inference about S will be based on $\mathbf{Y}$. We first consider *signal*

detection. Here we are faced with two possibilities, which are called *hypotheses* in the statistics literature, and we wish to decide which of the two is true in reality. The default hypothesis states that $S = \emptyset$, which means all components of $\boldsymbol{x}$ are in fact zero. In the above example this corresponds to the situation where none of the tested individuals has the disease. The alternative hypothesis is that $S \in \mathcal{C}$, that is the support is non-empty and is one of the elements of $\mathcal{C}$. The decision between the two is made by a *test* $\Psi : \mathcal{Y} \to \{0, 1\}$, which is a function specifying which hypothesis we accept based on the observations.

Naturally, if we have a test then we would like to evaluate its performance in some way. Since $\mathbf{Y}$ is assumed to come from a stochastic model, a natural way to do this is to consider the probability of error of Ψ. One can make an erroneous decision in two ways. A so-called type I error is made when in reality $\boldsymbol{x} = 0$, yet we decide that some components of $\boldsymbol{x}$ are non-zero. In the above example this corresponds to declaring the presence of infection when in fact this is not the case. A type II error is made when in reality $S \neq \emptyset$, yet we decide that the alternative hypothesis is true. In the above example this corresponds to declaring that no infection is present, when in fact it is. Note that the two types of errors can have very different consequences in practice, hence their role is often not symmetric.

Note that a type II error can be made regardless of which $S \in \mathcal{C}$ is the true support. Thus, while the type I error probability is simply $\mathbb{P}_\emptyset(\Psi = 1)$, the type II error probability is usually defined by aggregating the quantities $\mathbb{P}_S(\Psi = 0),\ S \in \mathcal{C}$ in some way, common methods being to take the average or the maximum. Having defined the two error probabilities we often combine them to form a single metric of error, most commonly by summation or taking their maximum. Note that it is important to consider type I and II error probabilities together, since one can always make $\mathbb{P}_\emptyset(\Psi = 1) = 0$ by setting $\Psi(\mathbf{Y}) = 0$ for all $\mathbf{Y} \in \mathcal{Y}$ or vice versa.

In *support recovery* we wish to know exactly which one of the possible sets in the class $\mathcal{C}$ is the true support. Again, the decision is made based on the sample by a function $\widehat{S} : \mathcal{Y} \to \mathcal{C}$ which we often refer to as an *estimator*. As before, we have numerous options to evaluate the performance of $\widehat{S}$ and we now mention the most common ones. Given a fixed $S \in \mathcal{C}$, one possibility is simply to compute the error probability $\mathbb{P}_S(\widehat{S} \neq S)$. One could also count the number of erroneously classified components, giving rise to the so-called expected Hamming-distance $\mathbb{E}_S(|\widehat{S} \triangle S|)$. A frequently used and less demanding metric is the expected normalized number of false discoveries and non-discoveries $\mathbb{E}_S(|\widehat{S} \setminus S|/|\widehat{S}|) + \mathbb{E}_S(|S \setminus \widehat{S}|/|S|)$, called

the False Discovery Rate (FRD) and Non-Discovery Rate (NDR). Once we have settled on a measure of effectiveness for a fixed $S \in \mathcal{C}$, we often aggregate them into a single error metric in some way. Choices are taking the average or the maximum of the selected quantity over the possible supports $S \in \mathcal{C}$.

The mathematical framework outlined above fits a wide range of problems, and to back up this claim we now mention a few. In [66] Dorfman describes a project whose objective was to weed out all syphilitic men enlisted to armed service. The decision whether an individual was infected or not was based on a blood test, and Dorfman developed a methodology now widely known as group testing in an effort to solve the problem in an economically feasible way. Biology and medicine are fields of science with a rich collection of problems fitting this framework, for instance the analysis of DNA microarray data. Yoon et al. [147] describe a setting where one wishes to identify genes exhibiting high expression levels under certain conditions (disease, drugs, etc.), whereas Moore et al. [112] show an instance when the goal is to discover new phenotypes of a disease based on common symptoms exhibited by patients. During the analysis of medical imaging data such as tomography or MRI one often wishes to find regions of activity or anomalies in the scanned tissue (see for instance Moon et al. [111]).

Analyzing imaging data is a task that also arises in the field of computer vision (see Zhong et al. [152]), for instance when trying to identify objects such as roads in satellite imagery (see Geman & Jedynak [74]). As explained by Culler, Estrin & Srivastava in [55], several surveillance tasks using sensor networks can be categorized as signal detection or support recovery problems as well. In such settings, sensors are spread out across an area (such as a structure, or a geographical region) and we wish to monitor the system based on observations collected by the sensors. Some examples include the detection of radioactive materials (see Brennan et al. [30]), or biological/chemical substances (see Cui et al. [54]) or target tracking (see Zhang et al. [149]).

Other monitoring tasks include the detection of disease outbreaks based on data from hospital emergency visits and pharmacy sales of drugs (see Kulldorff et al. [100]) or the spread of viruses in computer networks (see Szor [137]). Similar tasks may arise during the analysis of social networks as well. As an example consider the problem of finding communities based on observing the existing connections in a social network. A possible model is to assume connections are formed with higher likelihood inside a community than otherwise. In this interpretation the

community detection problem can be viewed as a signal detection task (see for instance Arias-Castro& Verzelen [15] and Arias-Castro & Grimmett [13]).

When facing a support recovery or signal detection task we can take two points of view. The theorist inside us would like to know the fundamental difficulty of these problems, that is to unravel the intricate interplay between a chosen error metric, the measurement model $\mathbb{P}_S$ and the class of support sets $\mathcal{C}$. When this is understood, we can formulate conditions $\boldsymbol{x}$ needs to satisfy such that the task at hand can be solved reliably in terms of our chosen performance metric. On the other hand, the practitioner inside us would like to actually solve the problems. That is, he wishes to construct specific tests Ψ or estimators $\widehat{S}$, and then prove that these work if $\boldsymbol{x}$ meets certain conditions. Ideally, the two halves meet in the middle and complement each other. If we have results stating necessary conditions for $\boldsymbol{x}$ so that the problem can be solved reliably and also have specific methods that solve the problem whenever these conditions are met, we have a good understanding of the problem at hand, and ultimately this is what we hope to achieve.

In this thesis the focus is more often on the theoretical side, in the sense that often the procedures we develop are aimed more at corroborating our theoretical findings and less at being readily applicable in real-life situations.

Coordinate-wise Sampling and Compressive Sensing

As noted before, the measurement model $\mathbb{P}_S$ plays a pivotal role in our ability to perform the tasks described above, and in what follows we briefly discuss the two models considered in this work. Probably the most natural observation model in the context described above is one we refer to as *coordinate-wise sampling*. In this case we observe $\boldsymbol{x}$ component by component, each observation perturbed by measurement noise. The observations of the components are independent and their distribution depends on whether the component in question is in the support or not. Note that the defining characteristic of this sampling scheme is that we observe components of $\boldsymbol{x}$ one-by-one, with measurement uncertainty affecting each observation. This does not necessarily mean however that we observe every component exactly once, in fact we can observe some components multiple times or not at all and still have a coordinate-wise measurement scheme. Nonetheless, a common special case of this setup is when we do observe every component of $\boldsymbol{x}$ once and our observation for a component is its value perturbed by independent additive noise, that is $Y_t = \boldsymbol{x}_t + W_t,\ t \in [n]$ and W_t are i.i.d. random variables. In

the example above this model would correspond to analyzing every blood sample separately.

Probably the most prominent case of the setup described above is the normal means model in which $W_t \sim N(0,1)$ and are independent for all $t \in [n]$. This model has received a large amount of attention, mostly in the detection setting (see Ingster & Suslina [87, 88], Baraud [21], Donoho & Jin [64], Arias-Castro & Sharpnack [14]) also with extensions to correlated noise (see Hall & Jin [78]). The popularity of the model is due to the fact that it arises naturally in many situations. For instance the this model is reasonable when samples are collected independently from different locations, or when measurements are made of separate objects. Examples of the former case include measurements made in sensor networks (see for instance Cheung et al. [47], Guerriero, Willett & Glaz [75]) or imaging applications (in this case we can think of each pixel of an image as a separate sensor, see Desolneux, Moisan & Morel [61]). An example for the latter case is for instance measuring expression levels of different genes (see for instance Ernst et al. [69], Pawitan et al. [119]).

To introduce the second measurement model, let us revisit the example above. Suppose that instead of analyzing each blood sample separately, we could choose to combine samples before subjecting them to analysis. Depending on the exact mechanics of the blood test this can be advantageous as we might be able to create a stronger signal by pooling infected blood together, or conversely if non-infected blood is pooled together, multiple healthy subjects can be identified using only one lab test. This is a simple cartoon for the measurement model known as *compressive sensing*.

Formally, our observations are of the form $Y_t = \langle A_t, \boldsymbol{x} \rangle + W_t, t \in [m]$ where $A_t \in \mathbb{R}^n$ is some sensing vector, $\langle ., . \rangle$ denotes the scalar product, and $\{W_t\}_{t\in[m]}$ are i.i.d. random variables. A commonly considered setup is when $\|A_t\|_2 \leq 1$, where $\|.\|_2$ denotes the Euclidean norm, and $W_t \sim N(0,1)$ and independent, $t \in [m]$. The main difference compared to the previous setup is that now we can aggregate information from different components of $\boldsymbol{x}$ before measurement uncertainty takes effect. This results in a more flexible sensing scheme, which can be seen formally by noting that by taking $A_t = \mathbf{1}_{\{t\}}$, where $\mathbf{1}_{\{t\}}$ is the singleton vector whose tth coordinate is equal to one, we essentially recover coordinate-wise sampling.

This sensing model arises naturally when there is some sort of data compression happening in the physical domain before the signal hits the sensors. One such application is the so-called Single-pixel camera introduced by Duarte et al. in [68]

where the camera has one photon receptor, and a mechanism composed of tiny mirrors inside the camera reflects light from different locations in the scene into the sensor. Hence each observation can be viewed as a linear combination of light coming from different parts of the scene. Another natural setting is medical imaging, most notably tomography and MRI (see for instance Panych & Jolesz [118], Deutsch, Averbush & Dekel [62]), where the sensors observe projection data from the object being imaged. Due to its practical relevance, the underlying mathematics of compressive sensing has also been widely studied in recent years, see for instance Candès & Tao [35], Donoho [65], Candès & Tao [34], Candès& Wakin [36], Wainwright [140] and Foucart & Rauhut [72].

Adaptive Sensing

Let us revisit our example of screening for a disease using blood samples in a setting where the blood sample of each individual needs to be analyzed separately (that is we are in the coordinate-wise sampling setup). Suppose that the lab test for analyzing a particular blood sample is such that the more time we allocate to analyzing that sample the more accurate the result will be[1]. On the other hand, suppose the time it takes to analyze a particular sample is directly associated with a cost - the more time the test takes, the more expensive it is.

If we have a pre-determined budget to carry out the screening, and need to decide how to allocate our resources beforehand, the sensible thing to do is to analyze every blood sample for the same cost per sample (provided we have no prior information about the identity of the infected individuals). However, if there are only a relatively small number of infected people this results in most of our budget being allocated to analyzing non-infected blood samples, which seems to be somewhat wasteful. Hence the question arises if there is a way to allocate our resources more efficiently by also using information we gather as we perform the lab tests?

Consider the following two stage design as an example. In the first stage we allocate half of our budget to a coarse screening, and perform all lab tests as before, only now with half the cost and hence half the accuracy per lab test. Based on the results, we select individuals that are susceptible to being infected and carry them on to the second stage. For these individuals we perform a new lab test using the

[1]For instance when testing for the presence of bacteria, a possible method is to grow bacteria in a Petri dish. In this case the more time we let the bacteria to grow, the more accurate our assessment of the bacteria-content of that particular sample will be.

remaining half of the total budget. Imagine that we need to screen a million people, but we have reason to believe the number of infected are in the hundreds. Our hope is that by calibrating the decision who to include in the second stage properly (say for the sake of example we select a few thousand), everyone infected will be selected to the second stage. Notice that in the second stage we use the same budget as in the first stage, however for only a few thousand tests instead of a million, hence we have more accuracy per person and hence more detection/estimation power. To summarize, we hope that when properly performed, such a scheme would provide more bang for our buck compared to the previous one.

The paradigm where samples are collected sequentially and adaptively based on information gleaned from previous measurements is known as *adaptive sensing* or *active learning*. To be a bit more concrete, consider the coordinate-wise sensing model described above. In a non-adaptive sensing scheme the decision which components of $\boldsymbol{x}$ to measure needs to be made before any observations are collected. Without any prior information about the support of $\boldsymbol{x}$ this gives rise to strategies where each component is measured with the same accuracy, leading essentially to the normal means model described above, or something similar. Contrasting this, in adaptive sensing the decision which component to sample next is allowed to depend on our previous observations. How to go about designing a good sampling strategy in this case is not obvious. Considering compressive sensing, in a non-adaptive scheme one needs to design the sensing vectors $A_1, \ldots, A_m$ before any observations are made. How to do so is not immediately clear, but perhaps unsurprisingly, it turns out that randomized designs that allocate roughly equal amounts of "sensing energy" to each component are the way to go, see the above references. In an adaptive strategy, when designing the sensing vector A_t, we could use all the information learned up to that point. Again, the main question here is how to do this in an efficient way.

The above example outlines both the appeals and bottlenecks of such schemes. The first potential benefit to adaptive sensing is increased statistical power. The hope is that closing the loop between data collection and inference would produce more informative samples. This would result in more accurate inference for the same amount of resources, or from a different point of view we would be able to perform the same task in a more cost-efficient way, compared to non-adaptive sensing.

The second appeal of adaptive sensing is in connection with computational

complexity. Certain statistical inference tasks are known to be computationally demanding or even intractable. One such example being submatrix estimation. This is a support recovery problem where the class of possible supports are submatrices of a given matrix (see Balakrishnan et al. [18], Berthet & Rigollet [24]). In such situations, even though we might have a formal mathematical solution to the problem at hand, computing it for a real dataset would be impossible because of the huge amount of arithmetics involved. It turns out that often this phenomenon is due to the fact that the relevant information in the non-adaptively collected data is swamped by the large number of non-informative observations. However, using adaptive sensing we can potentially tailor the sample collection process to facilitate inference, thus shaking the computational burdens mentioned above.

Finally, adaptive sensing might also bear an impact on the design of data acquisition devices. Data is being collected at an ever increasing pace due to rapid technological improvement, a prominent example being the field of astronomy (see Chamberlin, Pesnell & Thompson [45]). In some cases the volume of data is so huge that it can not all be stored in the long run. To make matters worse, often most of the data collected is completely irrelevant or uninteresting in hindsight. Hence new methodology is being developed in an effort to perform inference on-the-go (see for instance Diehl & Hampshire [63], Thompson et al. [138]). Taking this a step further, one could also use on-line methods to inform decisions about future sample collection, thus making better use of the available hardware.

Though these potential benefits are appealing, adaptive sensing has its drawbacks. First of all it is not guaranteed that any of the advantages above actually manifest in a particular inference task. Second, it is clear that the bottleneck of adaptive sampling schemes is the sample collection process. When not done with care, adaptive sample collection may introduce bias and steer inference in the wrong direction early on, which carries the danger that algorithms that seem sound based on heuristics fail dramatically and unexpectedly. Third, the feedback loop in the data acquisition makes understanding the fundamental difficulties of the inference task quite challenging.

Nonetheless, due to the potential benefits adaptive sensing promises, it has received considerable interest in the past. Probably the first such scheme is the Sequential Probability Ratio Test (SPRT) introduced by Abraham Wald in the forties in [141]. The aim of the SPRT is to decide between two simple hypotheses using a random number of observations instead of a fixed sample size. In every step the

decision maker collects a new observation and selects between two options. Either the test is terminated and a decision is made or the data is deemed insufficient to make a decision and a new observation is collected. Wald showed that the SPRT minimizes the expected number of samples needed to reach the correct decision (in some sense).

A next step to take is to consider settings where the learner not only decides whether to acquire further samples or not, but also influences the distribution the sample comes from. Such situations arise in many different statistical inference tasks of which we mention a few. Considering the topic of experimental design the decision maker might have a set of experiments to choose from and each time can select the one deemed most useful based on the result of the previous ones. Such settings were investigated by Chernoff [46], Fedorov [70] and Nitinawarat, Atia & Veeravalli [117], to name a few.

In many statistical learning tasks the aim is to predict the value of a response variable based on a number of features. A simple example is the automated classification of handwritten digits, where the learner's task is to devise a rule that takes an image of a handwritten digit and outputs the corresponding digit (i.e. $0, 1, \ldots, 9$). As information is gathered the learner might value samples with particular features more than others and so there might be benefits to collecting samples adaptively. Hence the role of adaptive sensing in similar problems has received considerable attention, for instance in the context of classification (see for instance Cohn, Ghahramani & Jordan [52], Dasgupta et al. [58, 57, 56], Castro & Nowak [42], Balcan, Beygelzimer & Langford [19], Koltchinskii [96], Hanneke [80]), regression (see Willett, Nowak & Castro [146], Hall & Molchanov [79]) and pattern recognition (see Blanchard & Geman [27]).

Adaptive sensing has also been investigated in the context of support recovery and signal detection problems, both in coordinate-wise sampling models (see Malloy & Nowak [106, 105, 109], Haupt, Castro & Nowak [81], Castro [40]) and compressive sensing models (see Haupt et al. [82], Malloy & Nowak [108, 107], Arias-Castro, Candès & Davenport [7]).

Structured Supports

In a number of real-life examples of support recovery and signal detection, the object of interest is not an arbitrary subset of the signal vector. In fact in a lot of the examples mentioned earlier these objects have a certain structure to them. In gene

expression studies the data is in the form of a gene-expression matrix, whose (i, j)th element is the expression level of gene j of the ith individual. Often times the aim is to find genes co-expressed under the influence of certain conditions such as diseases or drugs. Hence the support in this case is a submatrix of the data matrix rather than an arbitrary subset of its elements. Similar patterns arise in computer science when aiming to classify malware (see Jang, Brumley & Venkataraman [89]).

Computer viruses spreading from host to host in a computer network give rise to star-shaped patterns on the network graph (see Szor [137]). Communities in networks are often modeled as cliques or clusters in the network (see Arias-Castro & Verzelen [15]). In computer vision and imaging one often aims to identify objects in images, and these appear as contiguous "blobs" not as pixels scattered about randomly (see Moon et al. [111]). The same phenomenon applies when one aims to find geographic regions with certain properties, as in Cheung et al. [47].

These structural assumptions are encoded in the class of supports $\mathcal{C}$. As noted before, $\mathcal{C}$ plays an important role in both the difficulty of support recovery and signal detection problems and the design of tests and estimators that solve these tasks. Hence understanding the impact different structured classes have on these tasks is both mathematically interesting and useful in practice.

Most work on structured supports has been done in the context of non-adaptive sensing in the normal means model. Structures investigated include intervals or blocks (see [90, 32, 33]), submatrices (see Shabalin et al. [127], Butucea & Ingster [31], Balakrishna et al. [18], Berthet & Rigollet [24]) and various structures embedded in graphs such as clusters (see Neill & Moore [115], Arias-Castro et al. [8, 13], Neill [113], Sharpnack, Krishnamurthy & Singh [131], Qian, Saligrama & Chen [124]), stars, cliques and matchings (see Addario-Berry et al. [1], Arias-Castro & Verzelen [15]) and paths (see Arias-Castro et al. [9, 12]).

The properties of compressive sensing for estimating tree-structured activations has been studied by Baraniuk et al. [20] in the non-adaptive setting. In the adaptive sensing setting Soni & Haupt [134, 135] investigate the previous structured class, while Krishnamuthy, Sharpnack & Singh [97] deal with supports having low cut-size and Balakrishnan et al. [17] investigate the problem of recovering blocks.

Dynamic Signals

So far we have discussed signals that are static over time. In particular we have assumed that the support remains the same throughout the sampling process.

Though this assumption is reasonable in a lot of examples, there are situations where the story might be different. For instance a certain disease might spread quickly and thus infection hot spots might change rapidly over time. This intuitively increases the difficulty of the detection of hot spots as by the time we look at a certain place the disease might have already moved to a different location.

The detection of covert communications could also provide a similar example. Suppose some frequencies in a certain range are used for covert communications, those frequencies exhibiting higher power. However, in an effort to remain undetected, the frequencies to communicate through are changed from time to time. Our task, that is detecting if such communication is taking place, is clearly hindered by the fact that the frequencies used are being changed. In such situations it is important to understand how the dynamic aspect of the signal affects the fundamental difficulty of the problem and how to adjust procedures that perform the statistical inference to deal with this phenomenon.

The problem outlined above is essentially that of spectrum scanning in a cognitive radio system (see Li [103], Caromi, Xin & Lai [37]). When monitoring computer systems one goal is to detect malicious activity such as break-ins or frauds in the system (see for instance Gwadera, Atallah & Szpankowski [76], Phoha [121]). In these so-called intrusion detection problems one has access to a huge volume of data arriving in a streaming fashion and anomalies can pop up at different locations and times and then disappear.

Similar problems arise in the field of image processing, notably video surveillance, such as identifying traffic violators on the road or detecting suspicious activity around a structure or at an airport (see Diehl & Hampshire [63] and Pokrajac, Lazarevic & Latecki [122]). The detection of momentary astronomical events such as supernovae or solar flares also serves as an example for an application of anomaly detection in dynamical signals (see Thompson et al. [138]). Such problems also arise when observing more complex systems such as monitoring the spread of anomalous behavior or identifying the source of propagating information (e.g. viruses or gossips) in networks (see Wang et al. [144], Shah & Zaman [128], Luo & Tay [104], Zhu & Ying [153]).

Distribution-free tests

A useful first step in understanding statistical learning problems is to investigate them within clear cut distributional models. For instance the popular normal means

model states $Y_t = \boldsymbol{x}_t + W_t,\ t \in [n]$ and $\{W_t\}_{t\in[n]}$ are i.i.d. standard normal. The noise terms in compressive sensing are also often assumed to be i.i.d. and normally distributed. In such frameworks we can more readily understand the fundamental difficulty of the problems at hand and can develop tests and estimators that solve them in an optimal fashion. However, in practice these assumptions are either hard to verify or outright violated. So while the fundamental understanding we gain by assuming specific noise distributions is valuable, developing inference methods not relying heavily on such assumptions is very important from a practical standpoint. These are often referred to as *distribution-free* methods.

The practical relevance of distribution-free methods is probably best illustrated by their widespread application. In the first half of the twentieth century Wilcoxon [145] and Mann & Whitney [110] suggested the use of ranks to assess whether samples from two populations have the same mean or not. In the same time period Friedman [73] introduced a method also relying on ranks, as a counterpart to the analysis of variance. A similar method was developed by Kruskal & Wallis [98]. In all cases, the use of ranks was motivated by trying to avoid the normality assumption inherent in the commonly used inference procedures of the time. Today, all the methods referenced above are well known and widely used.

In a more modern context, distribution-free tests are used in the areas of disease surveillance (see Jung & Cho [92], Kulldorff et al. [100]), the analysis of fMRI data (see Holmes et al. [116, 85]) and other imaging applications (see Flenner & Hewer [71]) to name a few.

Though these methods benefit from not relying on distributional assumptions and thus being more widely applicable, there is a tradeoff involved. The price we pay for greater flexibility is decreased statistical power. Hence it is imperative to understand how much we lose in terms of our ability to perform accurate inference when using non-parametric methods. In an ideal situation it is possible to derive non-parametric tests for which this loss in power is only minute, giving us methods that are powerful but also robust to distributional misspecification.

1.2 Overview of the Thesis

Most of the thesis is aimed at gaining a better understanding of adaptive sensing protocols. In the coming three chapters we investigate several different support recovery and signal detection scenarios. We wish to obtain a fundamental under-

standing about adaptive sensing in these settings, in particular how it compares to non-adaptive sensing.

Chapters 2 and 3 deal with the problem of structured support recovery, first in a coordinate-wise sensing model and then in a compressive sensing model. We examine several different structured support classes that serve as good cartoons for structured supports encountered in real applications. We develop adaptive sensing procedures that reliably recover such supports, and show their near-optimality by deriving necessary conditions the signal needs to satisfy in order for any adaptive sensing algorithm to succeed. For comparison purposes we also derive necessary conditions for non-adaptive procedures. We show that, in general, adaptive sensing protocols outperform non-adaptive ones in the context of structured support recovery, though by how much depends on the specific structured class at hand.

In Chapter 4 we return to coordinate-wise sampling and examine the task of detecting sparse signals that change in time. To model the temporal aspect of the signal we will use a simple stochastic model. In a nutshell, at each time step a biased coin is flipped for each element of the support and if the coin comes up tails that component moves to a different location. Traditionally in such problems, data is collected non-adaptively and then analyzed, for instance by using data mining methods. As noted before, the volume of data being collected is ever increasing, which in turn can render inference methods designed for passively acquired samples unfeasible in practice. Hence we will take a slightly different point of view than usual when examining this problem. We will consider a situation where the signal can be measured at each point in time, but instead of being able to observe the entire signal we can only observe one component of $\boldsymbol{x}$ at each time. We wish to gain a fundamental understanding about how this phenomenon coupled with the temporal aspect of the signal affects the inference task. We develop algorithms and show their near-optimality both under the adaptive sensing and non-adaptive sensing paradigms. We show that adaptive sensing has an edge over non-adaptive sensing in such a setup as well, with the gains being less pronounced when the the signal components are changing rapidly.

In all previous chapters the problems are set up using rigorous distributional assumptions. Though we will make some comments on possibilities to relax these, it would be valuable from a practical standpoint to design and characterize adaptive sensing procedures that are robust to misspecification of the underlying distribution. As a first step we need to understand the properties of such procedures in the

non-adaptive sensing context, and this is the topic of Chapter 5. Here we revisit the coordinate-wise sampling setting and develop non-parametric methods for the detection of structured supports. We show that when the data we observe comes from the exponential family there is essentially no loss in power asymptotically when using the proposed non-parametric methods instead of ones that rely on full knowledge of the data distribution. We also present simulation results to gain some understanding about the finite sample behavior of the proposed methods.

Finally some concluding remarks are provided in Chapter 6. On one hand we summarize the results presented in previous chapters with an emphasis on the lessons learned about adaptive sensing. On the other hand we also aim to give an outlook for future research by discussing the shortcomings of the results in the thesis as well as describing other problems that might be worth pursuing.

Chapter 2

Adaptive Sensing for Structured Support Recovery

This chapter is based on joint work with Rui Castro. The results presented here can also be found in Castro & T. [44].

2.1 Introduction

In this chapter we consider the problem of recovering the support set of a sparse signal through noisy coordinate-wise measurements. Under non-adaptive sensing paradigms, the most natural way to collect data is to measure each coordinate of the vector with the same accuracy (that is, provided each coordinate of the vector is equally likely to be in the support set). However, what if we have the additional flexibility of also choosing the precision and location of each measurement based on the data collected so far? It is not immediately clear how much can be gained by these adaptive sensing strategies over the non-adaptive ones.

In addition to the sparsity assumption one might consider further structural restrictions on the unknown support set, as described in Chapter 1. Can we use such additional structural knowledge to further increase the performance of support recovery algorithms? If so, to what extent does such structural knowledge help us?

We aim to address the questions above in the framework considered by Haupt, Castro & Nowak in [81] and [40]. In particular we shed light on how adaptive sensing can capitalize on structural information, by providing general and practical algorithms endowed with performance guarantees. Furthermore we show that these algorithms are essentially optimal as we give matching performance lower bounds.

The classes that we consider fall into two categories: (i) all support sets of cardinality s, which we call *s-sets*. This is the maximal class for a given level of sparsity s, thus we refer to this class, or the union of such classes with different values of s, as the *unstructured case*. In contrast, other classes we consider are more stringent as the sets are structurally restricted. For instance, the class of s-intervals, that consist of sets of the form $[i, i+s-1]$ for $i \in [n-s+1]$. This class of sets is much smaller than the class of s-sets, and therefore we expect the support recovery task to be significantly easier. We use the umbrella term *structured case* for such classes. In particular, we study the following classes:

- **s-sets:** any subset of $[n]$ of size s;
- **s-intervals:** sets consisting of s consecutive elements of $[n]$;
- **unions of s-intervals:** unions of k disjoint s-intervals;
- **s-stars:** any star of size s in a complete graph[1];
- **unions of s-stars:** unions of k disjoint s-stars;
- **s-submatrices:** any submatrix of size s of a $\sqrt{n} \times \sqrt{n}$ matrix.

The structured classes above serve as a good starting point for understanding the effect structure has on the problem of adaptive support recovery, because while being simple, they are good proxies to structured signals arising in practice. Supports resembling intervals or unions of those arise for instance in gene-expression studies where certain genes are known to activate simultaneously (see for instance Balakrishnan et al. [17]). When the gene-expression of several subjects are stacked on top of one another to form a signal matrix, certain individuals might have elevated expression levels on the same genes, giving rise to submatrix shaped signal supports (see Shabalin et al. [127]). Finally, a star shaped pattern in a graph can be thought of as a simple cartoon for the initial stage of the spreading of a disease

[1] In the complete graph $G = (V, E)$ whose edges are identified with $[n]$, a set $S \subset E$ is a star iff $\exists v \in V : v \in e \ \forall e \ \in S$, where $v \in e$ denotes the incidence relation.

Table 2.1: Summary of scaling laws for the signal magnitude.

	Non-Adaptive Sensing (necessary)	Adaptive Sensing (necessary)	Adaptive Sensing (sufficient)
s-sets	$\mu \sim \sqrt{\frac{n}{m}\log n}$	$\mu \sim \sqrt{\frac{n}{m}\log s}$	$\mu \sim \sqrt{\frac{n}{m}\log s}$
unions of k disjoint s-intervals	$\mu \sim \sqrt{\frac{n}{sm}\log \frac{n}{ks}}$	$\mu \sim \sqrt{\frac{n}{sm}\log ks}$	$\mu \sim \sqrt{\frac{n}{sm}\log ks}$
unions of k disjoint s-stars	$\mu \sim \sqrt{\frac{n}{m}\log \frac{\sqrt{n}}{ks}}$	$\mu \sim \sqrt{\frac{n}{sm}\log ks}$	$\mu \sim \sqrt{\frac{n}{sm}\log^2 ks}$
s-submatrices of an $\sqrt{n} \times \sqrt{n}$ matrix	$\mu \sim \sqrt{\frac{n}{\sqrt{sm}}\log \frac{n}{s}}$	$\mu \sim \sqrt{\frac{n}{sm}\log s}$	$\mu \sim \sqrt{\frac{n}{sm}\log^2 s}$

Scaling laws for the signal magnitude μ (constants omitted) that are necessary/sufficient for $\max_{S\in\mathcal{C}} \mathbb{E}(\widehat{S}\triangle S) \to 0$ as $n \to \infty$, where $\mathcal{C}$ denotes the corresponding class of support sets and m denotes the total amount of precision available for our measurements in expectation (see (2.2)). All the results assume sparsity, meaning both $s = o(\sqrt{n})$ and $ks = o(\sqrt{n})$ as $n \to \infty$.

or a computer virus on a network, when a single infected node spreads the infection to some of its neighbors.

We examine the problem of structured support recovery in a coordinate-wise sensing setting. We are allowed to make measurements of the components of the signal, each measurement being perturbed by measurement noise. In an adaptive sensing setting, the decision which component to measure and the precision of the measurement in any given step can depend on the observations gathered previously. In a non-adaptive sensing setting, the sampling strategy needs to be fixed before any observations are made. In either case, there is a constraint on the total precision we are allowed to use (in expectation), which we denote by m. For instance, $m = n$ means that, on average, we have unit precision per signal entry.

Table 2.1 summarizes the results obtained in this chapter, stated in terms of asymptotic behavior when the signal dimension n is large and the support set (of size s or ks) is small. Note that most results in the chapter are not asymptotic in nature, furthermore the constant factors in the scaling laws are also accounted for. Nevertheless the results become easier to state and interpret in asymptotic terms.

A first point to notice is that the necessary condition for non-adaptive sensing always includes a $\sqrt{\log n}$ factor, regardless of the class considered. This factor is essentially due to the extreme value properties of Gaussian random variables. Note, however, that for adaptive sensing that factor is replaced by a $\sqrt{\log |S|}$ term (where $|S|$ is the sparsity of the support). This means that adaptive sensing can better

mitigate the effect of measurement noise. This is particularly interesting when $m = n$ (or more generally m is proportional to n) meaning that one can make, on average, one measurement of precision one per signal entry. In that case the dependence on the extrinsic dimension n vanishes completely when considering adaptive sensing, as opposed to non-adaptive sensing where the factor $\sqrt{\log n}$ is ever-present. However, the gains of adaptive sensing when structure is present can sometimes be much more remarkable. For discussion purposes consider the case $m = n$: for the class of unions of disjoint s-stars one gets that $\mu \sim \sqrt{\log n}$ is necessary for non-adaptive sensing, but it suffices that $\mu \sim \sqrt{(1/s)\log^2(ks)}$ for adaptive sensing. Therefore, apart from logarithmic factors, there is also a factor $\sqrt{1/s}$ reduction on signal magnitude with adaptive sensing. This can be rather beneficial, for instance when $s \sim n^\beta$ for some $0 < \beta < 1/2$. These gains stem from the strong structural constraints in the class, which can be exploited by adaptive sensing strategies. However, as the cardinality of this class is still very large it renders the structural information almost useless for non-adaptive sensing. A similar situation happens for the s-submatrices class, although the gains there are less dramatic (apart from logarithmic factors there is a factor $s^{-1/4}$ reduction in signal magnitude). Finally, for the class of unions of s-intervals such structural gains are not present (although the logarithmic factors are still significantly improved). In summary, adaptive sensing can both remove the dependence on the extrinsic dimension n due to noise (which is reflected in the logarithmic terms), and further improve the signal magnitude scaling laws (compared to non-adaptive sensing) when further structural information is present.

Remark 2.1. *In this chapter we consider only Gaussian observation noise. However, all the results in this chapter can be generalized to non-Gaussian noise models, which leads to different scaling laws. Nevertheless, the qualitative comparison between adaptive and non-adaptive sensing remains essentially the same. See e.g. the works of Malloy & Nowak [106, 105]*

Related work: Naturally, the choice of support set class $\mathcal{C}$ plays a crucial role. There is a wide range of available literature exploring the effect structure has on detecting and estimating signal supports. Most of the work on the topic so far has considered the non-adaptive sensing setting. In [1] Addario-Berry et al. consider the problem of testing for the presence of a signal, when the signal is known to have a structured support. They prove a general lower bound for the

Bayes risk[2] and demonstrate its sharpness for several structured support classes defined on graphs. Arias-Castro, Candès & Durand also investigate the detection of structured supports in graphs in [8]. They consider connected subgraphs in a lattice and show that a type of scan statistic is optimal for the class of supports under consideration, but its performance depends on a condition that measures the thickness of the support. In [129, 131] Sharpnack et al. consider the problem of detection of clusters in a graph, defining the class of supports for the alternative as a collection of clusters that have a small cut-size, and provide computationally tractable methods to solve the detection problem.

Taking a different approach, Saligrama, Quian & Chen investigate the detection problem of various types of supports on several different graphs in [124] and [123]. They show lower bounds for the different setups and a computationally tractable procedure attaining those bounds. In [15] Arias-Castro & Verzelen derive a sharp detection boundary for the problem of detecting a community in a random network, and provide a test attaining the best possible performance.

Moving away from structures on graphs, Butucea & Ingster explores the problem of detecting a sparse submatrix of a given size in a matrix in [31]. They provide lower bounds for the detection boundary and a test procedure matching the lower bound. They also provide a test for the case when the size of the submatrix is unknown. Arias-Castro, Candès & Plan in [10] consider the framework of linear models, and investigate the problem of deciding whether the parameter vector of the model is zero or some sparsely supported vector. Their results show that for moderate sparsity levels of the parameter vector a global test is optimal, whereas under stronger sparsity constraints a scan-type test is the optimal one.

In addition to models with added mean, models with added covariance were also in the focus of much work in the non-adaptive sensing setting. In [6] Arias-Castro et al. investigate the problem of deciding whether the components of a high-dimensional vector are correlated or not. In their model the covariance matrix is the identity under the null, and under the alternative there is a subset of components with a common positive correlation. The structural assumptions are incorporated through the subset of correlated components, that is, these components form a structured subset of the original vector. The authors provide lower bounds for the

[2]The Bayes risk of a test Ψ is defined by considering the average of the errors probabilities for different supports in the class as the type II error, and then adding type I and type II error probabilities together (see Chapter 1). Formally, the Bayes risk is $\mathbb{P}_0(\Psi = 1) + \frac{1}{|\mathcal{C}|}\sum_{S\in\mathcal{C}} \mathbb{P}_S(\Psi = 0)$.

detection problem and investigate the performance of several testing procedures as well. In [41] this problem was addressed in an adaptive sensing scenario by Castro, Lugosi & Savalle, and it was shown that adaptive sensing can yield significant gains when structural assumptions are made. However, it is still not known whether adaptive sensing can provide gains in the absence of such structural assumptions. Amini & Wainwright [4] and Berthet & Rigollet [25] consider the detection of sparse principal components in the spiked covariance model. They provide lower bounds for the difficulty of the problem while providing a computationally efficient near-optimal test using convex relaxations.

In contrast to the previously cited work, the main focus of this chapter is the estimation of signal supports as opposed to detection of the presence of signal, and the possible gains of adaptive sensing schemes compared to non-adaptive sensing. Related questions have also been investigated by several authors in the compressive sensing setting. Since Chapter 3 deals with that setup, overview of the related literature will be provided there. Considering coordinate-wise observation models in the adaptive sensing framework, Nitinawarat, Atia & Veeravalli investigates a multi-hypothesis testing problem in [117], where the decision-maker has the ability to select an experiment from a finite set of experiments in each measurement step. The authors show the best attainable asymptotic error exponent for this problem while providing tests attaining these exponents, extending the previous work of Chernoff [46].

In [105, 106] Malloy & Nowak discuss the problem of sparse support recovery in a setting where the decision which coordinates of the signal to sample can be made adaptively (but not the precision of the measurements). In these works the authors propose a sequential thresholding method and characterize its sample complexity. They also show that this sample complexity is asymptotically optimal for sparse signals.

The setting we consider in this chapter was introduced in [81], where Haupt, Castro & Nowak provide a simple and efficient adaptive sensing algorithm for support recovery (without structural assumptions). Castro also provides lower bounds for this problem in [40], and this work can be viewed as the extension of those results to structured support recovery.

Organization: This chapter is organized as follows. Section 2.2 describes the framework that we are considering in detail. In Section 2.3 we introduce a general

procedure for support set estimation. We analyze the performance of the procedure in Section 2.4. The performance limitations of any support estimation recovery procedure is investigated in Section 2.5. We present a small numerical experiment to corroborate our theoretical findings in Section 2.6. Finally we provide concluding remarks in Section 2.7.

2.2 Problem Setting

Let $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ be a vector of the form

$$x_i = \begin{cases} \mu & \text{if } i \in S, \\ 0 & \text{if } i \notin S, \end{cases} \tag{2.1}$$

where $\mu > 0$, and $S \subseteq [n]$. We refer to the vector $\boldsymbol{x}$ as the *signal*, and to S as the *signal support* or the *significant components* of the signal. The latter is our main object of interest, as neither $\boldsymbol{x}$ and S are directly available. We are allowed to collect multiple noisy measurements of each individual component of $\boldsymbol{x}$, namely

$$Y_t = \boldsymbol{x}_{A_t} + \Gamma_t^{-1/2} W_t \ , \ t = 1, 2, \ldots \ .$$

For each measurement we can choose A_t, the entry of $\boldsymbol{x}$ to be measured, and the corresponding *precision* of the measurement $\Gamma_t > 0$. Finally $W_t \sim N(0,1)$ are independent and identically distributed standard normal random variables. Also for any given t, W_t is independent of $\{A_j, \Gamma_j\}_{j=1}^t$. Under the adaptive sensing paradigm A_t and Γ_t are allowed to be functions of the past observations $\{Y_j, A_j, \Gamma_j\}_{j=1}^{t-1}$. This model is only interesting if one includes some constraint on the total amount of precision available. Let $\mathbb{P}_S$ denote the joint probability distribution of $\{Y_t, A_t, \Gamma_t\}_{t \geq 1}$ and $\mathbb{E}_S$ denote the expectation with respect to $\mathbb{P}_S$. We require that

$$\mathbb{E}_S \left(\sum_t \Gamma_t \right) \leq m \ , \tag{2.2}$$

where m is our total precision budget, specified in advance. This constraint arises naturally in many practical settings, and can be viewed as a total time constraint in sensing modalities where precision is directly proportional to the amount of time necessary to collect a measurement (see for instance Castro [39]). Finally, using the collected data we construct an estimator $\widehat{S} \equiv \widehat{S}(\{Y_t, A_t, \Gamma_t\}_{t \geq 1})$ that is desirably

as close to S as possible.

Note that in our setup it is possible to make infinitely many measurements provided that the budget condition in (2.2) is satisfied. Although this might seem strange from a practical point of view it does enable a clear and insightful explanation of the tradeoff between adaptive and non-adaptive sensing, by decoupling the issue of sample complexity and sensing budget in a natural way. Furthermore we show that, despite this flexibility, one can devise optimal procedures that also have good sample complexity properties (see Remark 2.3).

At a first glance it might seem that the model (2.1) is overly restrictive, as all the significant components of $\boldsymbol{x}$ have exactly the same value μ. However, the results in this chapter can be generalized to sparse signals with non-zero significant components of arbitrary signs and magnitudes, provided the minimum magnitude of these is large enough. For the sake of clarity and simplicity we do not consider this extension here, but refer the reader to Castro [40] for details on how this can be done.

Note that it is also possible to consider algorithms satisfying an exact energy constraint as opposed to the expected energy constraint in (2.2). This requires an extension of the arguments in the body of the paper, but yields essentially the same results. For the sake of completeness we provide details in Appendix 2.A.

In this work the primary focus is on adaptive sensing algorithms. However, for comparison purposes, we will also consider non-adaptive sensing inference, which means that $\{A_t, \Gamma_t\}_{t\geq 1}$ must be chosen before any observations are collected. In other words, non-adaptive sensing requires $\{A_t, \Gamma_t\}_{t\geq 1}$ to be independent from $\{Y_t\}_{t\geq 1}$.

2.2.1 Inference Goals

Since our goal is to characterize the fundamental limitations of adaptive sensing, we will assume that μ is known, in addition to n and m. Therefore, the only unknown quantity is the signal support S. Our aim is to construct adaptive sensing methodologies that are able to estimate S. This is only possible when the signal magnitude μ is large enough. Furthermore, it is reasonable and desirable to make some concrete assumptions about S, namely that the signal has a sparse support (meaning that the cardinality of S is small) and also extra structural assumptions. All these can be formalized by assuming that S belongs to some class $\mathcal{C}$ of subsets of $[n]$. Note that we also assume the class $\mathcal{C}$ is known, which in particular means

that we assume knowledge of the size of the support.

There are various ways one can define *reliable* estimation of a support set (see Castro [39]). In this work we consider the worst-case Hamming-distance as an error metric of our estimator. Let $\widehat{S} \equiv \widehat{S}(\{Y_t, A_t, \Gamma_t\}_{t\geq 1})$ be a specific estimator. We wish to ensure that for a given $\varepsilon > 0$,

$$\max_{S\in\mathcal{C}} \mathbb{E}_S(|\widehat{S}\triangle S|) \leq \varepsilon \ , \tag{2.3}$$

where $\widehat{S}\triangle S$ is the symmetric set difference of $\widehat{S}$ and S, and $|\cdot|$ denotes the cardinality of a set. In words, we require the expected number of errors to be less than ε, regardless of the true unknown support set S.

One can also consider a slightly less stringent metric, namely the probability of falsely identifying the support set, that is $\mathbb{P}_S(\widehat{S} \neq S)$. Note that we have

$$\mathbb{P}_S(\widehat{S} \neq S) \leq \mathbb{E}_S(|\widehat{S}\triangle S|) \leq 2|S|\ \mathbb{P}_S(\widehat{S} \neq S) \ , \tag{2.4}$$

where the second inequality holds provided $|\widehat{S}| = |S|$ (this property holds for all the estimators that we that consider). According to this we are able to control the expected number of errors of a procedure by controlling the probability of error. This is exactly what we do, so the analysis of the procedure we propose will be applicable to both error metrics. In addition, we also derive lower bounds in terms of expected Hamming-distance. Whenever we can, we also provide lower bound in terms of probability of error.

2.3 A General Adaptive Sensing Estimation Procedure

At the core of the problem setting we have described is the issue of noise and measurement uncertainty, which is embodied by the precision budget in (2.2). Without this restriction the inference task is much easier, in fact it is merely combinatorial in nature, as one can make noiseless (infinite precision) measurements. Nevertheless, a sharp distinction between adaptive and non-adaptive sensing is still present for noiseless procedures, meaning that one can devise adaptive sensing procedures that outperform non-adaptive ones in this combinatorial setup. This gives rise to a simple, yet very powerful idea: take the *noiseless* adaptive sensing procedures and

transform them to be robust to noise. Our general approach hinges precisely on this "robustification" of noiseless procedures (which we refer to as *noiseless-case algorithms*) as follows. When a noiseless case algorithm observes an entry of $\boldsymbol{x}$, we take multiple noisy measurements of that entry and perform a sequential hypothesis test to decide whether the entry in question is zero or not. Then we use the result of the test as a surrogate for the noiseless observation. If we ensure that these tests have small enough probabilities of error, we can recover the support set with high probability. By carefully controlling these error probabilities we can also control the expected Hamming-distance of the devised estimator. To better illustrate the ideas we will make use of two running scenarios (corresponding to two different classes $\mathcal{C}$): (i) the class of *s-sets*, that is when the support set S is an arbitrary subset of $[n]$ with cardinality s; the class of *s-intervals*, that is all sets consisting of s consecutive elements of $[n]$.

2.3.1 Noiseless-case algorithms

An algorithm based on coordinate-wise sampling for the noiseless case can be described as follows. In each step $j \in \mathbb{N}$, the algorithm either collects an observation of a coordinate of $\boldsymbol{x}$, or stops and returns the estimate $\widehat{S}$ for the support set $S \in \mathcal{C}$. The observation collected in step j is denoted by $\widetilde{Y}_j = \mathbf{1}\{x_{Q_j} \neq 0\}$, where $Q_j \in [n]$ determines the coordinate of $\boldsymbol{x}$ that we sample in step j, and $\mathbf{1}\{\cdot\}$ denotes the usual indicator function. We call Q_j the *query* in step j, and it plays a role analogous to that of A_t in the problem description. In case the component indexed by Q_j is a signal component the value of $\widetilde{Y}_j$ is 1, otherwise it is 0. Note that if $\mathbf{1}_S(\cdot) : [n] \to \{0, 1\}$ denotes the indicator function of the support set S, the observations can be written as $\widetilde{Y}_j = \mathbf{1}_S(Q_j)$. After taking a number of observations we may decide to stop and return the estimator of the support set. T denotes the stopping time for the procedure and $\widehat{S}$ denotes the estimate of the support set $S \in \mathcal{C}$.

To fully describe such an algorithm, one needs to give the queries $\{Q_j\}_{j \geq 1}$, a stopping time T and a rule for constructing $\widehat{S}$. The query Q_j is a measurable function of $\{Q_i, \widetilde{Y}_i\}_{i=1}^{j-1}$ mapping to $[n]$. It determines the coordinate of $\boldsymbol{x}$ that we wish to sample in step j. We also consider randomized procedures, so Q_j need not be deterministic. Note that because the observations are noiseless, it is unnecessary to sample any coordinate of $\boldsymbol{x}$ more than once, and therefore we only consider procedures satisfying this property.

The stopping time T is the possibly random time at which we stop sampling and return an estimate of the support set. Thus T is an $\mathbb{N}$-valued measurable function of the filtration generated by $\{Q_j, \widetilde{Y}_j\}_{j\geq 1}$. Bearing in mind that later we wish to transform our noiseless-case algorithm to be robust to noise, we consider the following two possible definitions of T. The first definition is simply

$$T = \inf\left\{j : \text{ there is at most one } S' \in \mathcal{C} : \widetilde{Y}_i = \mathbf{1}_{S'}(Q_i)\ \forall i \in [j]\right\}. \tag{2.5}$$

This means that we consider procedures that stop sampling when there is a unique set in $\mathcal{C}$ that agrees with all the observations, or if there is no such set. Note that T is well defined, and since it is unnecessary to sample any coordinate of $\boldsymbol{x}$ more than once in the noiseless case, we have $T \leq n$. Furthermore in the noiseless case the procedure stops when there is exactly one set in $\mathcal{C}$ in line with our observations (since we assume $S \in \mathcal{C}$). Recall however that we will later modify the procedure to be able to handle noise and thus there will be a chance of making errors. Because of this, it is possible that there will be no set in $\mathcal{C}$ in agreement with all our observations. For this reason we enforce the procedure to stop when this happens to ensure T remains well-defined after the modification.

The second possibility is much more straightforward: we can simply take $T = n$. This will be useful for the unstructured example of s-sets, since it is easy to see that no matter what sampling strategy $\{Q_j\}_{j\geq 1}$ we use, in the worst case we have to sample every coordinate of $\boldsymbol{x}$. Hence for this class we might as well stick to simply defining $T = n$, which will make the transition of the noiseless-case algorithm to the noisy case much more fluent.

The estimator $\widehat{S}$ is then defined in the noiseless case as the unique set $S' \in \mathcal{C}$ that agrees with our observations (and clearly we have $\widehat{S} = S$). However, remembering that we wish to transform our procedures to be able to handle noise, we need to extend this definition to ensure that $\widehat{S}$ also remains well-defined after the aforementioned modification. First, consider the following definition, which we will use in general:

$$\widehat{S} = \begin{cases} S' & \text{if } \exists!\ S' \in \mathcal{C} :\ \widetilde{Y}_j = \mathbf{1}_{S'}(Q_j)\ \forall j \in [T], \\ \emptyset & \text{otherwise.} \end{cases} \tag{2.6}$$

This definition is in line with the first definition of the stopping time T. However, with a slight abuse of notation, we can also define $\widehat{S} = \{Q_j : \widetilde{Y}_j = 1\}$. That

is, $\widehat{S}$ can also be the collection of those coordinates for which the result of the query at that coordinate is 1. Note that now $\widehat{S}$ need not even be in $\mathcal{C}$, and this definition only makes sense when we use the choice $T = n$ for our stopping time. Indeed, this will be the case, and we will only use this definition for the class of s-sets and merely to make the transition from the noiseless case to the noisy case more seamless.

To illustrate what such procedures may look like, consider the examples of *s-sets* and *s-intervals*. In the first case consider a deterministic procedure, that samples every coordinate one after the other. That is let $Q_j = j$, $j \in [T]$. The procedure will stop after sampling every component of $\boldsymbol{x}$, that is $T = n$. Once sampling has stopped, the estimate of the support set is the collection of components for which the result of the query is 1, formally $\widehat{S} = \{Q_j : \widetilde{Y}_j = 1\} = \{j : \widetilde{Y}_j = 1\}$. Next, consider the class of *s-intervals*. Consider a randomized procedure consisting of two phases. In the first phase sample random coordinates of the vector $\boldsymbol{x}$ until a non-zero coordinate is found. In the second phase we search for the left endpoint of the interval by sampling coordinates to the left of the previously found signal component one after the other. The interval S is exactly determined either when a 0 is found in the second phase, or when all the s signal components are found. Formally, denoting by Unif the discrete uniform distribution, the procedure can be written as $Q_j \sim \text{Unif}\left([n] \setminus \{Q_i\}_{i\in[j-1]}\right)$, $\forall j \leq T'$, where

$$T' = \inf\left\{j : \widetilde{Y}_j = 1 \text{ and } \widetilde{Y}_i = 0 \quad \forall i \in [j-1]\right\} ,$$

and $Q_j = Q_{j-1} - 1$, $\forall j = [T' + 1, T]$. The estimator $\widehat{S}$ is defined as before as the unique set compatible with the observations. Note that no claim is made about whether this procedure is optimal in any sense. In particular it is possible to construct a procedure that takes less steps in expectation than this one, for instance by performing a binary search in the second phase.

2.3.2 From the noiseless to the noisy case

Assume now that one has a noiseless-case procedure. The next step is to translate this procedure to the noisy case, to handle the situation when the observations are contaminated by noise as in (2.1), and there is a total precision budget as defined in (2.2). The main idea is to replace each query Q_j by multiple observations of the entry of $\boldsymbol{x}$ indexed by Q_j, and perform a hypothesis test to assess whether

the component corresponding to that query is zero or not. Specifically, we will set type I and type II error probabilities α_j and β_j for each Q_j, perform a Sequential Likelihood Ratio Test (SLRT[3]) with these error probabilities, and use its result as a surrogate for $\widetilde{Y}_j$. How to properly choose the error probabilities α_j and β_j depends on the specific problem at hand, but for now assume these are simply given to us.

The procedures we propose have the nice property that all observations are made with the same precision Γ, namely $\Gamma_t = \Gamma > 0 \ \forall t \in \mathbb{N}$. This is not at all restrictive, provided Γ is relatively small, as justified by Proposition 2.1 below. For the first query Q_1 set the target type I and type II error probabilities to be α_1 and β_1 respectively. The SLRT collects observations

$$Y_t = x_{Q_1} + \Gamma^{-1/2} W_t \ , \ t = 1, \ldots, N_1 \ ,$$

where N_1 is an appropriate stopping time defined as follows. Let $f_0(\cdot)$ and $f_1(\cdot)$ denote the density of the observations when $Q_1 \notin S$ and $Q_1 \in S$ respectively. Define the log-likelihood ratio

$$\bar{z}_k = \sum_{t=1}^{k} \log \frac{f_1(Y_t)}{f_0(Y_t)} \ . \tag{2.7}$$

The stopping time N_1 is defined as

$$N_1 = \inf \{ k \in \mathbb{N} : \bar{z}_k \notin (l_1, u_1) \} \ ,$$

where $l_1 < 0 < u_1$ are chosen so that both $\mathbb{P}(Z_{N_1} \geq u_1 | Q_1 \notin S) \leq \alpha_1$ and $\mathbb{P}(Z_{N_1} \leq l_1 | Q_1 \in S) \leq \beta_1$[4]. Once N_1 observations have been collected a decision is made regarding whether or not Q_1 belongs to the support set. Namely we define the test function Ψ_1 as

$$\Psi_1 = \begin{cases} 0 & \text{if } \bar{z}_{N_1} \leq l_1, \\ 1 & \text{if } \bar{z}_{N_1} \geq u_1. \end{cases}$$

[3]In the literature this sequential procedure is also referred to as the Sequential Probability Ratio Test (SPRT) (see e.g., Wald [141]). We feel, however, that the use of the term "likelihood ratio" is perhaps more appropriate, as in most settings one is computing a ratio between densities and not probabilities.

[4]Ideally we would like to choose the stopping boundaries in a way that the true error probabilities are equal to the nominal ones. However, as we will see later, it is more simple to choose stopping boundaries that satisfy these inequalities.

We use the value of Ψ_1 as a surrogate for $\widetilde{Y}_1$ in the noiseless-case procedure. This then determines the next query Q_2. Again we perform an SLRT by taking observations of the coordinate $\boldsymbol{x}_{Q_2}$. We set the type I and type II error probabilities to be α_2 and β_2, determine upper and lower stopping boundaries l_2, u_2, perform the test resulting in Ψ_2 which we use as a surrogate for $\widetilde{Y}_2$, and so on. We continue in this manner until the condition for the stopping time T of the noiseless case procedure is met, and return the corresponding estimate $\widehat{S}$. The whole procedure is summarized in Algorithm 1.

Algorithm 1: General Adaptive Sensing Support Estimation

Input:

- A noiseless procedure characterized by: queries $\{Q_j\}_{j\geq 1}$, stopping time T, and estimator $\widehat{S}$
- Precision parameter $\Gamma > 0$
- Type I and II error probabilities α_j and β_j corresponding to query Q_j

for $j \leftarrow 1$ **to** $\cdots$ **do**

  Perform an SLRT for entry x_{Q_j} with error probabilities α_j, β_j resulting in Ψ_j

  Set $\widetilde{Y}_j = \Psi_j$

  If $T = j$ stop and return $\widehat{S}$

end

Remark 2.2. *Note that for a fixed time $j \geq 1$ the value of T need not be computable. Nonetheless, the logical value of the expression $T = j$ can be evaluated for every $j \geq 1$ as T is a stopping time.*

It is important to notice that the procedure is well defined. In particular, each of the SLRTs terminate with probability one, as shown in Proposition 2.1 below. Furthermore, by the definition of T (see (2.5)) the entire procedure is guaranteed to terminate with probability one, even if some of the SLRTs result in errors (meaning $\Psi_j \neq \mathbf{1}_S(Q_j)$). Finally, the definition (2.6) ensures $\widehat{S}$ is also well defined in the event of errors.

2.4 Performance Upper Bounds

In this section we use the procedure outlined in the previous section to characterize attainable inference limits in various settings. The SLRT is at the heart of our

procedure, and therefore we begin by deriving some important properties these satisfy. We then move on to the analysis of the full procedure.

2.4.1 Analysis of the SLRTs

Most tools used in our analysis stem from the seminal work by Wald [141]. However, some of these results have to be specialized for our setting. Consider a SLRT that we use to decide between the two simple hypotheses H_0 and H_1. We collect independent and identically distributed measurements $y_1, y_2, \ldots$, where $y_i \sim N(0, \Gamma^{-1})$ under H_0 and $y_i \sim N(\mu, \Gamma^{-1})$ under H_1. We set α and β as target type I and type II error probabilities respectively. These determine upper and lower stopping boundaries which we denote by $l = \log \frac{\beta}{1-\alpha}$ and $u = \log \frac{1-\beta}{\alpha}$. Note that for the latter to make sense we need $\alpha, \beta < 1/2$. We will in fact assume this for all error probabilities throughout the chapter. Recall the definition of the log-likelihood ratio in (2.7), and define the stopping time N_Γ as

$$N_\Gamma = \inf \{k \in \mathbb{N} : \ \bar{z}_k \notin (l, u)\} \ ,$$

where f_0 and f_1 are the densities of y_1 under H_0 and H_1 respectively, and the subscript Γ is meant to emphasize the dependence in Γ. Finally define the test Ψ as

$$\Psi = \begin{cases} 0 & \text{if } \bar{z}_{N_\Gamma} \leq l, \\ 1 & \text{if } \bar{z}_{N_\Gamma} \geq u. \end{cases}$$

We know from the theory of SLRTs that $\mathbb{P}(N_\Gamma < \infty) = 1$ (see Wald [141]), so the data collection terminates almost surely. We also know that

$$\mathbb{E}_0(N_\Gamma) \geq \frac{1}{-D(\mathbb{P}_0 \| \mathbb{P}_1)} \left((1-\alpha) \log \frac{\beta}{1-\alpha} + \alpha \log \frac{1-\beta}{\alpha} \right) \ ,$$

and

$$\mathbb{E}_1(N_\Gamma) \geq \frac{1}{D(\mathbb{P}_1 \| \mathbb{P}_0)} \left((1-\beta) \log \frac{1-\beta}{\alpha} + \beta \log \frac{\beta}{1-\alpha} \right) \ ,$$

where $\mathbb{P}_0$ and $\mathbb{P}_1$ are the distributions of y_1 under H_0 and H_1 respectively, $\mathbb{E}_0$ and $\mathbb{E}_1$ are the expectations with respect to $\mathbb{P}_0$ and $\mathbb{P}_1$ respectively and $D(\cdot \| \cdot)$ is the Kullback-Leibler divergence of two distributions[5]. Since $\mathbb{P}_0$ and $\mathbb{P}_1$ are normal

[5]Let F and G be two distributions with densities f and g w.r.t. a common dominating measure ν. The Kullback-Leibler divergence between F and G is $D(F \| G) = \int f \log \frac{f}{g} \, \mathrm{d}\nu$ when F is absolutely continuous w.r.t. G, and ∞ otherwise.

distributions we have $D(\mathbb{P}_1\|\mathbb{P}_0) = D(\mathbb{P}_0\|\mathbb{P}_1) = \Gamma\mu^2/2$ and therefore

$$\Gamma\mathbb{E}_0(N_\Gamma) \geq \frac{2}{\mu^2}\left((1-\alpha)\log\frac{1-\alpha}{\beta} + \alpha\log\frac{\alpha}{1-\beta}\right) \tag{2.8}$$

and

$$\Gamma\mathbb{E}_1(N_\Gamma) \geq \frac{2}{\mu^2}\left((1-\beta)\log\frac{1-\beta}{\alpha} + \beta\log\frac{\beta}{1-\alpha}\right) . \tag{2.9}$$

The derivation of these lower bounds goes roughly as follows. The cumulative log-likelihood $\{\bar{z}_k\}_{k\geq 1}$ is a discrete-time stochastic process. The process terminates when it leaves the interval (l, u). By assuming that the process hits the boundaries of this interval exactly we can get the lower bounds above. In reality the log-likelihood ratio will never be exactly equal to l and u. However, when the precision Γ is small, the increments to the stochastic process $\bar{z}_k$ are also small, and so this process will nearly hit the exact boundaries of the interval (l, u). This in turn means that the above lower bounds should be attainable when Γ approaches zero. This is indeed the case, as stated in the following result, which is proved in Appendix 2.B:

Proposition 2.1. *Let $\alpha_\Gamma = \mathbb{P}_0(\Psi = 1)$ and $\beta_\Gamma = \mathbb{P}_1(\Psi = 0)$ be, respectively, the type I and II error probabilities of the SLRT. Then*

$$\alpha_\Gamma \to \alpha \quad \text{and} \quad \beta_\Gamma \to \beta$$

as $\Gamma \to 0$. Furthermore

$$\Gamma\mathbb{E}_0(N_\Gamma) \to \frac{2}{\mu^2}\left((1-\alpha)\log\frac{1-\alpha}{\beta} + \alpha\log\frac{\alpha}{1-\beta}\right)$$

and

$$\Gamma\mathbb{E}_1(N_\Gamma) \to \frac{2}{\mu^2}\left((1-\beta)\log\frac{1-\beta}{\alpha} + \beta\log\frac{\beta}{1-\alpha}\right) ,$$

as $\Gamma \to 0$.

Remark 2.3. *Proposition 2.1 considers the setting when the precision of each measurement is made arbitrarily small. This is a suitable assumption to make from a theoretical standpoint and in fact makes the presentation of the results that follow insightful and clear. However, from a practical standpoint this might not be satisfactory. To this end, we present a version of the proposition above in Appendix 2.C (Proposition 2.14) where the precision $\Gamma > 0$ is fixed, and follows from slightly more detailed writing of the proof of Proposition 2.1. When the precision is fixed, the*

condition on the precision budget (2.2) *becomes a condition on the sample complexity, hence the results of Proposition 2.14 are stated in terms of* $\mathbb{E}_0(N_\Gamma)$ *and* $\mathbb{E}_1(N_\Gamma)$.

Proposition 2.1 shows that the lower bounds on the expected amount of precision used by the SLRT with error probabilities α, β can be achieved in the limit when $\Gamma \to 0$. Thus when analyzing the performance of our procedures in terms of expected precision used, we can use these lower bounds to calculate the expected precision used by the SLRTs. Note that we are interested in the case when α and β are small. Thus, to make the discussion less cumbersome we note that when α and β are both at most $1/2$ we have

$$\frac{2}{\mu^2}\Big((1-\alpha)\log\frac{1-\alpha}{\beta} + \alpha\log\frac{\alpha}{1-\beta}\Big) < \frac{2}{\mu^2}\log\frac{1}{\beta} \tag{2.10}$$

and

$$\frac{2}{\mu^2}\Big((1-\beta)\log\frac{1-\beta}{\alpha} + \beta\log\frac{\beta}{1-\alpha}\Big) < \frac{2}{\mu^2}\log\frac{1}{\alpha}\ , \tag{2.11}$$

since the last terms are negative, and the first terms can be upper bounded trivially by $1-\alpha \le 1$ and $1-\beta \le 1$ respectively. When α and β are small, the inequalities above are essentially tight. This means that when calculating the expected precision used by a SLRT, we do not lose much by using the expressions on right hand sides above. By Proposition 2.1, for fixed α, β, we can choose a small enough Γ such that the quantities on the right hand side above upper bound the expected precision used by the SLRT.

2.4.2 General Analysis of Algorithm 1

Now we turn our attention to the analysis of the general procedure of Section 2.3. Recall that a procedure for the noiseless case is characterized by queries $\{Q_j\}_{j\ge1}$, a stopping time T that indicates the time when we stop sampling, and the estimator $\widehat{S}$. Unless stated otherwise we consider the definition of the two last quantities given by (2.5) and (2.6). The queries Q_j will be defined separately for each special case.

Given a certain noiseless-case procedure we translate it to the noisy case by replacing the outcome of each noiseless query Q_j by a surrogate SLRT Ψ_j. This requires the specification of type I and type II error probabilities α_j and β_j for each of the tests $\{\Psi_j\}_{j\ge1}$. Naturally, α_j and β_j can be, in general, functions of

$\{Q_i, \Psi_i\}_{i=1}^{j-1}$, and we wish to choose them to ensure that the final estimator $\widehat{S}$ satisfies $\mathbb{E}_S(|\widehat{S} \triangle S|) \leq \varepsilon, \ \forall S \in \mathcal{C}$ on one hand, and that the total precision budget (2.2) is not exceeded. Clearly, to meet the former goal, α_j and β_j need to be small enough, while if these are too small the latter goal might not be attained. Therefore we need to make a compromise in setting these error probabilities. How to optimally choose $\{\alpha_j, \beta_j\}_{j \geq 1}$ depends on the specific procedure under consideration (and the class of possible support sets), and it is difficult to get a general answer. However, we will see that in many interesting cases simple and intuitive choices for α_j, β_j yield near-optimal results.

We illustrate the analysis of the procedure by first considering the unstructured case of all s-sets. In the unstructured case the near optimal procedure is very simple, and our choice of α_j, β_j does not depend on j, which greatly facilitates the analysis. Formally the class of s-sets is defined as

$$\mathcal{C} = \{S \subseteq [n] : \ |S| = s\} \ .$$

A simple procedure for the noiseless case is defined by taking $Q_j = j, \ j \in T$, and $T = n$. Then we use the definition $\widehat{S} = \{Q_j : \widetilde{Y}_j = 1\}$ which translates to the noisy case (also with our specific choice of Q_j) as $\widehat{S} = \{j : \Psi_j = 1\}$. In words, we simply estimate the support as the collection of components whose SLRT accepts the alternative. Because of the sparsity of the signal we expect the majority of the coordinates that we sample to be zero, and we know that there are exactly s that are non-zero. So it is sensible to take $\alpha_j \approx \varepsilon/n$ and $\beta_j \approx \varepsilon/s$. We will take the following concrete choice $\alpha_j = \varepsilon/2n$ and $\beta_j = \varepsilon/2s$, $j \in n$.

In the worst case, for any $S \in \mathcal{C}$ we query all the entries of $\boldsymbol{x}$. Using this crude upper bound we get

$$\mathbb{E}_S(|\widehat{S} \triangle S|) = \sum_{j=1}^{n} \mathbb{P}_S(\Psi_j \neq \mathbf{1}_S(Q_j)) \leq \sum_{j \notin S} \alpha_j + \sum_{j \in S} \beta_j \leq n\frac{\varepsilon}{2n} + s\frac{\varepsilon}{2s} \leq \varepsilon \ .$$

Since the inequality above holds for all $S \in \mathcal{C}$ we conclude that the expected number of errors for any $S \in \mathcal{C}$ is at most ε. Furthermore the total amount of precision

used in expectation by this procedure is

$$\mathbb{E}_S\left(\sum_t \Gamma_t\right) \leq \sum_{j\notin S} \frac{2}{\mu^2}\log\frac{2s}{\varepsilon} + \sum_{j\in S}\frac{2}{\mu^2}\log\frac{2n}{\varepsilon}$$
$$\leq \frac{2n}{\mu^2}\log\frac{2s}{\varepsilon} + \frac{2s}{\mu^2}\log\frac{2n}{\varepsilon}\ ,$$

where we used (2.10) and (2.11), and took Γ small enough. Note that the total amount of precision increases when the signal magnitude μ decreases. Combining this result with the bound on the total precision available (2.2) we can characterize the conditions on μ for which this procedure fits all the requirements outlined in Section 2.2.

Proposition 2.2. *Let $\mathcal{C}$ denote the class of all s-sets. Whenever*

$$\mu \geq \sqrt{\frac{2n}{m}\log\frac{2s}{\varepsilon} + \frac{2s}{m}\log\frac{2n}{\varepsilon}}\ , \tag{2.12}$$

the estimator $\widehat{S}$ resulting from the procedure above satisfies $\max_{S\in\mathcal{C}} \mathbb{E}_S(|\widehat{S}\triangle S|) \leq \varepsilon$, and the precision budget of (2.2)*.*

Since $s \leq n$ the first term on the right hand side of (2.12) is always as large as the second term. Thus the scaling of μ as a function of n, m, s and ε is determined by the first term. Therefore we have the following corollary.

Corollary 2.1 (s-sets)**.** *Consider the setting of Proposition 2.2, and let $\omega_n \to \infty$ be arbitrary (as $n \to \infty$). Whenever*

$$\mu \geq \sqrt{\frac{2n}{m}(\log s + \omega_n)}\ ,$$

the above procedure produces an estimator $\widehat{S}$ satisfying $\lim_{n\to\infty}\max_{S\in\mathcal{C}} \mathbb{E}_S(|\widehat{S}\triangle S|) = 0$, and that satisfies the precision budget of (2.2)*.*

We know from Castro [40] that, apart from constants, this is the best performance we can hope for when considering the expected Hamming-distance of the estimator (when $s \ll n$). The procedure presented by Malloy & Nowak in [109] essentially has the same performance as this one, and is also a coordinate-wise method that it is based on sequential thresholding. However, it is parameter adaptive and agnostic about s for a wide range of values.

We now turn our attention to a number of special cases, where the sets belonging to the class $\mathcal{C}$ have some sort of structure. As before, the starting point is some procedure for the noiseless case, specified by $\{Q_j\}_{j\geq 1}$. We will make no claim about whether the procedure we define for the noiseless case is optimal in any sense, although in most cases these do give rise to optimal scaling limits.

All the noiseless procedures that we consider consist of two phases. They begin with a *search* phase, where one identifies the general spatial location of the support set. In this phase we sample components of $\boldsymbol{x}$ according to some searching method, until we find a certain number $l_1 \leq |S|$ of signal entries. Then we switch to a *refinement* phase, where we exploit the structure of the support set to find a number of entries of S. In some cases the proposed procedures alternate between these two phases. Consider the following procedure for the class of s-intervals. The search phase simply scans the components until an element of S is found, and the refinement phase explores the coordinates in the neighborhood of the element of S found earlier.

The exact form of queries $\{Q_j\}_{j\geq 1}$ depends on the specific class under consideration. Likewise, the number of search phases K and how many components to find in each search phase $\eta_1, \ldots, \eta_K$ depends on the class of possible support sets. In the previous example for the s-intervals $K = 1$ and $\eta_1 = 1$. In what follows we denote the total number of signal entries we wish to find throughout the search phases as $\eta = \sum_{k=1}^{K} \eta_k$.

To translate the noiseless-case procedures to the noisy case we must specify α_j, β_j for each test $\Psi_j,\ j \in [T]$ to ensure that the overall probability of error of the procedure is small[6]. Afterwards we turn our attention to the expected precision used by the procedure. Combining the latter bound with the total amount of precision available as in (2.2) we get a condition on the minimal signal strength μ that is sufficient to ensure the support is recovered accurately. For the control of the overall error probability we can take advantage of the two phases. Suppose we want to keep the probability of error to be less than δ. First, note that since the noiseless case procedure does not sample any coordinate of $\boldsymbol{x}$ more than once, we perform at most n tests, thus the conservative choice $\alpha_j \approx \delta/n,\ j \in [T]$ suffices. Now note that throughout the search phases we plan to encounter no more than η non-zero coordinates of $\boldsymbol{x}$, so it is reasonable to set $\beta_j \approx \delta/\eta$ in such phases. Finally, since there are at most $|S|$ significant components we can observe, in the

[6] As noted in Section 2.2.1, it is enough to control the probability of error of a procedure, as then we can also control the expected Hamming-distance using (2.4)

refinement phase we take $\beta_j \approx \delta/|S|$.

It is crucial to note that for a given j, α_j, β_j are in general functions of $\{Q_i, \Psi_i\}_{i=1}^{j-1}$. This means that when defining the error probabilities we can only use the results of the tests carried out so far, but not the true identity of the entries that we sampled. It is important to keep this in mind in the analysis of the procedure. Also, note that the choices above are likely not optimal. For some classes to be considered later on, one can immediately improve the α_j, β_j of the next proposition (e.g. for the s-intervals we will perform at most n/s tests in the first phase so setting $\alpha_j = s\delta/n$ for the search phase suffices). Nevertheless, these choices for the probabilities of type I and II errors are simple and general, and yield essentially optimal results.

Proposition 2.3. *Suppose the noiseless case procedure is of the form described above, and let* $\alpha_j = \delta/4n,\ j \in [T]$, $\beta_j = \delta/2\eta$ *for the search phase and* $\beta_j = \delta/4|S|$ *for the refinement phases. Then*

$$\mathbb{P}_S(\widehat{S} \neq S) \leq \delta,\ \forall S \in \mathcal{C}\ .$$

Proof. Consider a noiseless case procedure given by $\{Q_j\}_{j\geq 1}$, and any support set $S \in \mathcal{C}$. Let $\mathcal{E}_j$ denote the event that we make an error in the test Ψ_j, meaning $\Psi_j \neq \mathbf{1}_S(Q_j)$. Let $\overline{\mathcal{E}}_j$ denote the complement of $\mathcal{E}_j$. In what follows we compute the probability that no errors are made.

The support set will be correctly identified if all tests are correct. Consequently,

$$\begin{aligned}
\mathbb{P}_S(\widehat{S} \neq S) &= 1 - \mathbb{P}_S(\widehat{S} = S) \\
&\leq 1 - \mathbb{P}_S\left(\bigcap_{j=1}^{T} \overline{\mathcal{E}_j}\right) \\
&= 1 - \mathbb{P}_S\left(\overline{\mathcal{E}_1}\right) \mathbb{P}_S\left(\overline{\mathcal{E}_2} \middle| \overline{\mathcal{E}_1}\right) \cdots \mathbb{P}_S\left(\overline{\mathcal{E}_T} \middle| \bigcap_{j=1}^{T-1} \overline{\mathcal{E}_j}\right)\ .
\end{aligned}$$

The above expression upper bounds the probability of error, by considering the case where all the test results coincide with the noiseless case. Since there are at most n zero components being measured in the entire noiseless-case procedure, η significant components being measured in the search phase, and at most $|S|$

significant components being measured in the refinement phase we conclude that

$$\begin{aligned}\mathbb{P}_S(\widehat{S} \neq S) &\leq 1 - \left(1 - \frac{\delta}{4n}\right)^n \left(1 - \frac{\delta}{2\eta}\right)^\eta \left(1 - \frac{\delta}{4|S|}\right)^{|S|} \\ &\leq \delta \ .\end{aligned}$$

The last inequality follows from a simple Taylor expansion of

$$g(\delta) = 1 - (1 - \delta/4n)^n (1 - \delta/2\eta)^\eta (1 - \delta/4|S|)^{|S|}$$

around $\delta = 0$, since $g(0) = 0$, $g'(0) = 1/4 + 1/2 + 1/4 = 1$ and $g''(\delta) \leq 0$ for every $\delta \in [0, 1]$. □

Proposition 2.3 ensures that the noisy case procedure has a probability of error that is sufficiently small. The next step is to evaluate the total expected precision used (considering that the precision Γ of each measurement is arbitrarily small). This quantity depends crucially on the noiseless case procedure we use for the specific class under consideration. For that reason this calculation is done separately for each case considered.

s-intervals

Consider the class of intervals of length s. Formally,

$$\mathcal{C} = \{S \subseteq [n] : \ S = [i, i + s - 1], \ i \in [n - s + 1]\}\ ^7.$$

For sake of simplicity assume n/s is an integer. This is merely to ease notation in the calculations that follow. The first step is to define a procedure for the noiseless case. Our choice consists of one search and one refinement phase. In the search phase we sample coordinates $1, s + 1, 2s + 1, \ldots$, until we find a non-zero coordinate. This gives us the approximate position of the interval. Then we move to the refinement phase to find the left endpoint of the interval by sequentially sampling coordinates of $\boldsymbol{x}$ to the left of the previously found non-zero coordinate[8].

[7] Note that, with a slight abuse of notation, we will use the same symbol $\mathcal{C}$ for different structured classes throughout this work. However, the class we are referring to will always be clear from the context.

[8] There are more efficient ways of finding the left endpoint, for instance using binary search. However, we stick to the simple method outlined above, as it is easier to formally describe and does not result in a significant loss of performance, since we are considering the sparse setting when $s \ll n$.

Note that in the second phase we query at most $s-1$ coordinates.

Formally $Q_j = (j-1)s+1$ for $j \in [T']$, where $T' = \inf\{j : \widetilde{Y}_j = 1\}$, and $Q_j = Q_{j-1} - 1$ for $j = [T'+1, T]$, where T is defined in general in (2.5). The estimator $\widehat{S}$ is defined in (2.6) as usual. Note that this is an instance of the general procedure described in the setting of Proposition 2.3 with $K = 1$ and $\eta_1 = \eta = 1$. Taking the corresponding choices for $\{\alpha_j, \beta_j\}_{j \geq 1}$ we ensure that $\mathbb{P}_S(\widehat{S} \neq S) \leq \delta$, $\forall S \in \mathcal{C}$. As for the expected precision we can make use of Proposition 2.1 and the choices of α_j and β_j to conclude that

$$\begin{aligned}\mathbb{E}_S\left(\sum_t \Gamma_t\right) &\leq \mathbb{E}_S\Bigg(\underbrace{\sum_{j=1}^{T'-1} \frac{2}{\mu^2}\log\frac{1}{\beta_j} + \frac{2}{\mu^2}\log\frac{1}{\min\{\alpha_j,\beta_j\}}}_{\text{search}} \\ &\qquad + \underbrace{\sum_{j=T'+1}^{T} \frac{2}{\mu^2}\log\frac{1}{\min\{\alpha_j,\beta_j\}}}_{\text{refinement}}\Bigg) \\ &\leq \frac{2}{\mu^2}\left(\frac{n}{s}\log\frac{2}{\delta} + s\log\frac{4n}{\delta}\right) .\end{aligned}$$

Combining this with the bound on the total precision available (2.2) we get the following result:

Proposition 2.4. *Let $\mathcal{C}$ denote the class of s-intervals, and suppose*

$$\mu \geq \sqrt{\frac{2n}{sm}\log\frac{2}{\delta} + \frac{2s}{m}\log\frac{4n}{\delta}} .$$

Then the procedure above results in an estimator $\widehat{S}$ satisfying $\max_{S\in\mathcal{C}} \mathbb{P}_S(\widehat{S} \neq S) \leq \delta$ and the precision budget (2.2).

In addition we can also control the expected Hamming-distance $\mathbb{E}_S(|\widehat{S}\triangle S|)$ by recalling (2.4). To guarantee that $\mathbb{E}_S(|\widehat{S}\triangle S|) \leq \varepsilon$ we simply have to be slightly more conservative, and require the probability of error δ to be at most ε/s. An analogous result to that of Proposition 2.4 follows immediately. In case signals are sufficiently sparse, meaning $s \ll n$, the first term inside the square root dominates the bound. Therefore we have the following result.

Corollary 2.2 (s-intervals)**.** *Consider the setting of Proposition 2.4. Suppose that $s = o\left(\sqrt{n/\log n}\right)$ as $n \to \infty$, and let $\omega_n \to \infty$ be arbitrary.*

(i) When

$$\mu \geq \omega_n \sqrt{\frac{n}{sm}} \ ,$$

the procedure above gives an estimator $\widehat{S}$ *such that* $\lim_{n\to\infty} \max_{S\in\mathcal{C}} \mathbb{P}_S(\widehat{S} \neq S) = 0$*, and that satisfies* (2.2)*.*

(ii) When

$$\mu \geq \sqrt{\frac{2n}{sm}\left(\log s + \omega_n\right)} \ ,$$

the procedure above gives an estimator $\widehat{S}$ *such that* $\lim_{n\to\infty} \max_{S\in\mathcal{C}} \mathbb{E}_S(|\widehat{S} \triangle S|) = 0$*, and that satisfies* (2.2)*.*

Unions of s-intervals

Now we consider the class whose elements are the union of k disjoint s-intervals, where s-intervals were defined in the previous subsection. Formally let $\mathcal{C}'$ be the class of s-intervals as defined previously. Then

$$\mathcal{C} = \left\{ S \subseteq [n] : \ S = \bigcup_{i=1}^{k} S_i, \ S_i \in \mathcal{C}' \ \forall i, \ S_i \cap S_j = \emptyset \ \forall i \neq j \right\} .$$

Again, assume n/s is an integer for simplicity. Note that the cardinality of the support sets belonging to this class is ks. In case $s = 1$ and $k = s$ this class is the same as the class of s-sets considered in Proposition 2.2. When we choose $k = 1$ this is the class of s-intervals described in Section 2.4.2. In that sense this class can be viewed as a bridge between the two previous classes.

The procedure for the noiseless case will again consist of one search and one refinement phase. In the search phase we sample coordinates $1, s+1, 2s+1, \ldots$ until we find k non-zero coordinates. Then in the refinement phase, we sample coordinates to the left of the previously found non-zero coordinates to find the left endpoints of all k intervals. Note that we make at most $k(s-1)$ queries in the second phase. This procedure is an instance of that described in the setting of Proposition 2.3 with $K = 1$ and $\eta_1 = \eta = k$. Taking the corresponding choices for α_j, β_j ensures $\mathbb{P}_S(\widehat{S} \neq S) \leq \delta, \ \forall S \in \mathcal{C}$. As for the expected precision used we can

write

$$\mathbb{E}_S\left(\sum_t \Gamma_t\right) \le \mathbb{E}_S\left(\sum_{j=1}^{T'} \frac{2}{\mu^2}\log\frac{1}{\beta_j} + |S|\frac{2}{\mu^2}\log\frac{1}{\min\{\alpha_j,\beta_j\}}\right)$$
$$\le \frac{2}{\mu^2}\left(\frac{n}{s}\log\frac{2k}{\delta} + ks\log\frac{4n}{\delta}\right) .$$

Combining this with the bound on the total precision available (2.2) we arrive to the following result:

Proposition 2.5. *Let $\mathcal{C}$ denote the class of unions of s-intervals as defined above, and suppose*

$$\mu \ge \sqrt{\frac{2n}{sm}\log\frac{2k}{\delta} + \frac{2ks}{m}\log\frac{4n}{\delta}} .$$

The procedure above results in an estimator $\widehat{S}$ satisfying both $\max_{S\in\mathcal{C}} \mathbb{P}_S(\widehat{S} \ne S) \le \delta$ and the precision budget (2.2).

In case of sparse signals, that is, when both s and k are small, the first term on the right side dominates this bound. More precisely we have the following result.

Corollary 2.3 (Unions of s-intervals)**.** *Consider the setting of Proposition 2.5. Assume $k \ge 2$ and $s \ge 1$ such that $s = o\left(\sqrt{\frac{n\log k}{k\log n}}\right)$ as $n \to \infty$. Let $\omega_n \to \infty$ be arbitrary.*

(i) When

$$\mu \ge \sqrt{\frac{2n}{sm}\left(\log k + \omega_n\right)} ,$$

the procedure above gives an estimator $\widehat{S}$ such that $\lim_{n\to\infty}\max_{S\in\mathcal{C}} \mathbb{P}_S(\widehat{S} \ne S) = 0$, and that satisfies (2.2).

(ii) When

$$\mu \ge \sqrt{\frac{2n}{sm}\left(\log ks + \omega_n\right)} ,$$

then the procedure above gives an estimator $\widehat{S}$ such that $\lim_{n\to\infty}\max_{S\in\mathcal{C}} \mathbb{E}_S(|\widehat{S}\triangle S|) = 0$, and that satisfies (2.2).

s-stars

Consider a setting when the coordinates of $\boldsymbol{x}$ correspond to edges of a complete undirected graph $G = (V, E)$ with p vertices. We call a support set S an s-star if

the edges in G corresponding to S form a star in G (see Figure 1 in Addario-Berry et al. [1]). Formally, let $e_i \in E$ denote the edge of G corresponding to coordinate i of $\boldsymbol{x}$ for all $i \in [n]$. The class of s-stars is defined as

$$\mathcal{C} = \left\{ S \subseteq [n] : \bigcap_{i \in S} e_i = v \in V,\ |S| = s \right\} ,$$

where $e_i \cap e_j$ is the set of common vertices of edges $e_i, e_j \in E$. Unlike what was done for the previous classes we use a randomized procedure for the noiseless case. Like was done for s-intervals the procedure consists of one search and one refinement phase. In the search phase we randomly search the coordinates of $\boldsymbol{x}$ until we find a non-zero coordinate. In the refinement phase we sample the coordinates of $\boldsymbol{x}$ that correspond to edges that share a vertex with the non-zero coordinate found in the search phase.

Define $Q_j \sim \text{Unif}\left([n] \setminus \{Q_i\}_{i \in [j-1]}\right)$ for $j \in [T']$ with $T' = \inf\{j : \widetilde{Y}_j = 1\}$, and $Q_j \sim \text{Unif}\left(\widetilde{X} \setminus \{Q_i\}_{i \in [j-1]}\right)$ for $j \in [T'+1, T]$, where $\widetilde{X} = \left\{i \in [n] :\ e_i \cap e_{T'} \neq \emptyset\right\}$. The stopping time T and estimator $\widehat{S}$ are defined as usual in (2.5) and (2.6). Note that this is an instance of the general procedure described in the setting of Proposition 2.3 with $K = 1$ and $\eta_1 = 1$.

The expected amount of precision used is now a bit more tedious to calculate due to the randomness in the search phase, which in the noisy case is prone to errors. For this reason we slightly modify the above procedure to greatly simplify the analysis. The modification is that in the search phase we only take at most J queries. Therefore, when J is small one might end the search phase without finding a star. However, we choose J large enough such that the probability of not querying a signal component is small. If we adjust the error probabilities α_j, β_j accordingly, we can still ensure that the probability of error of the procedure is small. More precisely, we choose J such that $\mathbb{P}_S(\forall j \in [J] :\ Q_j \notin S) \leq \delta/2$. Since

$$\mathbb{P}_S\left(\forall j \in [J] :\ Q_j \notin S\right) = \frac{\binom{n-s}{J}}{\binom{n}{J}} \leq \left(1 - \frac{s}{n}\right)^J ,$$

choosing $J = (n/s)\log(2/\delta)$ ensures that the probability above is less than $\delta/2$. Now choosing α_j, β_j according to Proposition 2.3 with δ replaced by $\delta/2$ ensures $\mathbb{P}_S(\widehat{S} \neq S) \leq \delta,\ \forall S \in \mathcal{C}$. With this modification the expected amount of precision

is bounded by

$$\begin{aligned}\mathbb{E}_S\left(\sum_t \Gamma_t\right) &\le \mathbb{E}_S\left(\sum_{j=1}^{T'} \frac{2}{\mu^2}\log\frac{1}{\beta_j} + |S|\frac{2}{\mu^2}\log\frac{1}{\alpha_j} + \sum_{j=T'+1}^{T}\frac{2}{\mu^2}\log\frac{1}{\beta_j}\right)\\ &\le \frac{2}{\mu^2}\left(J\log\frac{4}{\delta} + |S|\log\frac{8n}{\delta} + 2(p-2)\log\frac{8s}{\delta}\right)\\ &\le \frac{2}{\mu^2}\left(\frac{n}{s}\left(\log\frac{4}{\delta}\right)^2 + s\log\frac{8n}{\delta} + \sqrt{8n}\log\frac{8s}{\delta}\right)\ .\end{aligned}$$

Combining this with the bound on the total precision available (2.2) we get the following result:

Proposition 2.6. *Let $\mathcal{C}$ be the class of s-stars as defined above and suppose*

$$\mu \ge \sqrt{\frac{2n}{sm}\left(\log\frac{4}{\delta}\right)^2 + \frac{2s}{m}\log\frac{8n}{\delta} + \frac{\sqrt{32n}}{m}\log\frac{8s}{\delta}}\ .$$

The procedure above results in an estimator $\widehat{S}$ satisfying both $\max_{S\in\mathcal{C}} \mathbb{P}_S(\widehat{S}\neq S) \le \delta$ and the precision budget (2.2).

In case of sparse signals the first term on the right hand side dominates this bound. More precisely we have the following result.

Corollary 2.4 (s-stars)**.** *Consider the setting of Proposition 2.6. Suppose $n\to\infty$ such that $s = o(\sqrt{n}/\log n)$. Let $\omega_n\to\infty$ be arbitrary.*

(i) When

$$\mu \ge \omega_n\sqrt{\frac{n}{sm}}\ ,$$

the procedure above gives an estimator $\widehat{S}$ such that $\lim_{n\to\infty}\max_{S\in\mathcal{C}}\mathbb{P}_S(\widehat{S}\neq S) = 0$, and that satisfies (2.2).

(ii) When

$$\mu \ge \sqrt{\frac{2n}{sm}(\log^2 s + \omega_n)}\ ,$$

the procedure above gives an estimator $\widehat{S}$ such that $\lim_{n\to\infty}\max_{S\in\mathcal{C}}\mathbb{E}_S(|\widehat{S}\triangle S|) = 0$, and that satisfies (2.2).

Unions of s-stars

The unions of k non-intersecting s-stars is a generalization of the class of s-stars defined in the previous section. Suppose for technical reasons that $k < s$. Let $\mathcal{C}'$ be the class of s-stars as defined previously. Then

$$\mathcal{C} = \left\{ S \subseteq [n] : \ S = \bigcup_{i=1}^{k} S_i, \ S_i \in \mathcal{C}' \ \forall i, \ S_i \cap S_j = \emptyset \ \forall i \neq j \right\}.$$

Note that the cardinality of the support sets belonging to this class is ks.

In contrast to what we have done before, the proposed noiseless procedure will consist of alternating search and refinement phases. In the search phases we randomly search coordinates of $\boldsymbol{x}$ until we find a signal coordinate. Then we switch to the a refinement phase and sample every coordinate of $\boldsymbol{x}$ that corresponds to edges that share a vertex with the non-zero coordinate found previously. After doing so it may happen that we find signal entries corresponding to more than one star, which we would know by having identified a number of active components that is not a multiple of s. When there are stars left partly explored, we continue sampling edges that possibly belong to not yet fully explored stars (the candidates are edges adjacent to any previously found active component). When there are no partly explored stars, we switch back to the search phase. We keep iterating until we have found all k stars of the graph.

Formally $Q_j \sim \text{Unif}\left([n] \setminus \{Q_i\}_{i\in[j-1]}\right)$ in the search phases. Let $\widetilde{X}_j$ denote the set of edges that can belong to partly explored stars up to time j. Then the queries of the refinement phase can be defined as $Q_j \sim \text{Unif}\left(\widetilde{X}_j \setminus \{Q_i\}_{i\in[j-1]}\right)$. Note that this still fits the setting of Proposition 2.3 with $K \leq k$ being random and $\eta_1 = \eta_2 = \cdots = \eta_K = 1$.

Analogously to what was done for s-stars we consider a simple modification to facilitate the analysis: each time we are in a search phase we take at most J queries. We choose J such that the noiseless case procedure fails with small probability. Note that we perform at most k search phases, and in each of them there are at least s unexplored signal components. Thus, using essentially the same calculation as before, we get that by choosing $J = (n/s)\log(2k/\delta)$ we ensure that the probability of not querying a signal coordinate in any of the search phases is at most $\delta/2$. Finally, choosing α_j, β_j according to Proposition 2.3 with δ replaced by $\delta/2$ yields $\mathbb{P}_S(\widehat{S} \neq S) \leq \delta, \ \forall S \in \mathcal{C}$.

Note that the number of queries we perform in all of the search and refinement

phases is at most kJ and $2ksp$, respectively. However, for the expected number of queries performed throughout the search phases we can get a slightly better upper bound, which is necessary to get a more accurate dependence on the parameter k. Recall that $\mathcal{E}_j$ denotes the event that we make an error in the test Ψ_j, i.e. $\Psi_j \neq \mathbf{1}_S(Q_j)$. Also, let $\mathcal{E}_0$ denote the event that there is a search phase in which we do not query any coordinate containing a signal, and T_A denote the total number of queries in the search phases. Finally, let the number of queries in the ith search phase be $T_A^{(i)}$. Using the mean of the negative hypergeometric distribution, we have $\mathbb{E}_S\left(T_A^{(i)} \left| \cap_{j=0}^{T} \overline{\mathcal{E}}_j \right.\right) \leq n/(sk_i)$, where k_i is the number of unexplored stars in search phase i. Noting that $k_1 = k$ and $k_{i+1} < k_i,\ i = 1, \ldots, K$ we obtain the bound

$$\begin{aligned}\mathbb{E}_S\left(T_A \middle| \bigcap_{j=0}^{T} \overline{\mathcal{E}}_j\right) &= \sum_{i=1}^{K} \mathbb{E}_S\left(T_A^{(i)} \middle| \bigcap_{j=0}^{T} \overline{\mathcal{E}}_j\right) \\ &\leq \sum_{i=1}^{k} \frac{n}{is} \leq \frac{n}{s}(\log k + 1)\ .\end{aligned}$$

Finally, through the law of total expectation we get

$$\mathbb{E}_S\left(T_A\right) \leq \frac{n}{s}(\log k + 1) + \delta k J\ .$$

We are now ready to compute a bound on the precision used by the procedure.

$$\begin{aligned}\mathbb{E}_S\left(\sum_t \Gamma_t\right) &\leq \mathbb{E}_S\left(T_A \frac{2}{\mu^2}\log\frac{4k}{\delta} + |S|\frac{2}{\mu^2}\log\frac{8n}{\delta} + 2ks(p-2)\frac{2}{\mu^2}\log\frac{8ks}{\delta}\right) \\ &\leq \frac{2}{\mu^2}\left(\left(\frac{n}{s}(\log k + 1) + \delta k J\right)\log\frac{4k}{\delta} + ks\log\frac{8n}{\delta} + ks\sqrt{8n}\log\frac{8ks}{\delta}\right) \\ &\leq \frac{2}{\mu^2}\left(\frac{n(1+\delta k)}{s}\left(\log\frac{4k}{\delta}\right)^2 + ks\log\frac{8n}{\delta} + ks\sqrt{8n}\log\frac{8ks}{\delta}\right)\ .\end{aligned}$$

Combining this with the bound on the total precision available (2.2) we get the following result:

Proposition 2.7. *Let $\mathcal{C}$ be the class of unions of k disjoint s-stars as defined above*

and suppose

$$\mu \geq \sqrt{\frac{2n(1+\delta k)}{sm}\left(\log\frac{4k}{\delta}\right)^2 + \frac{2ks}{m}\log\frac{8n}{\delta} + \frac{ks\sqrt{32n}}{m}\log\frac{8ks}{\delta}}.$$

The procedure above results in an estimator $\widehat{S}$ *satisfying both* $\max_{S\in\mathcal{C}} \mathbb{P}_S(\widehat{S} \neq S) \leq \delta$ *and the precision budget* (2.2).

The result of the above proposition is perhaps a bit difficult to digest, but provided s and k are small relative to n the first term in the right side dominates the bound.

Corollary 2.5 (unions of s-stars)**.** *Consider the setting of Proposition 2.7. Suppose* $s = o\left(\sqrt{\frac{\sqrt{n}\log k}{k\log n}}\right)$ *as* $n \to \infty$*. Let* $\omega_n \to \infty$ *be arbitrary.*

(*i*) *When*

$$\mu \geq \sqrt{\frac{2n}{sm}(\log^2 k + \omega_n)}\ ,$$

the procedure above gives an estimator $\widehat{S}$ *such that* $\lim_{n\to\infty}\max_{S\in\mathcal{C}} \mathbb{P}_S(\widehat{S} \neq S) = 0$*, and that satisfies* (2.2).

(*ii*) *When*

$$\mu \geq \sqrt{\frac{2n}{sm}(\log^2 ks + \omega_n)}\ ,$$

the procedure above gives an estimator $\widehat{S}$ *such that* $\lim_{n\to\infty}\max_{S\in\mathcal{C}} \mathbb{E}_S(|\widehat{S}\triangle S|) = 0$*, and that satisfies* (2.2).

s-submatrices

In this setting the components of $\boldsymbol{x}$ are identified with the elements of a matrix $M \in \mathbb{R}^{n_1\times n_2}$ (the number of elements in the matrix be $n = n_1 n_2$). We assume that the support set S is a subset of $[n_1] \times [n_2]$ and furthermore we assume that it corresponds to a submatrix of size s. Formally, the class of all s-submatrices is defined as

$$\mathcal{C} = \{S_1 \times S_2 : S_1 \subseteq [n_1], S_2 \subseteq [n_2], \text{ and } |S_1|\cdot|S_2| = s\}\ ,$$

where $S_1 \times S_2$ denotes the cartesian product of S_1 and S_2. Note that if either n_1 or n_2 is of the same order as n, then this setting becomes similar to the unstructured case, but if $n_1, n_2 \approx \sqrt{n}$ there is a significant amount of structure one can take advantage of. Consider the following simple noiseless support recovery procedure: in a first phase randomly search the coordinates of $\boldsymbol{x}$ to find a non-zero coordinate. Once such a coordinate is found explore coordinates of $\boldsymbol{x}$ corresponding to the row and column of the non-zero coordinate found previously. Clearly, this fits the general procedure described in the setting of Proposition 2.3, with $K = 1$ and $l_1 = l = 1$. Like in the case of s-stars we stop the random search in the first phase after $J = (n/s)\log(2/\delta)$ queries to facilitate the analysis. For the expected amount of precision used we have

$$\begin{aligned}\mathbb{E}_S\left(\sum_t \Gamma_t\right) &\leq \frac{2}{\mu^2}\left(J\log\frac{4}{\delta} + s\log\frac{8n}{\delta} + (n_1+n_2)\log\frac{8s}{\delta}\right) \\ &\leq \frac{2}{\mu^2}\left(\frac{n}{s}\left(\log\frac{4}{\delta}\right)^2 + s\log\frac{8n}{\delta} + (n_1+n_2)\log\frac{8s}{\delta}\right).\end{aligned}$$

Proposition 2.8. *Let $\mathcal{C}$ denote the class of submatrices as defined above with and suppose*

$$\mu \geq \sqrt{\frac{2n}{sm}\left(\log\frac{4}{\delta}\right)^2 + \frac{2s}{m}\log\frac{8n}{\delta} + \frac{2(n_1+n_2)}{m}\log\frac{8s}{\delta}}\,.$$

The procedure above results in an estimator $\widehat{S}$ satisfying both $\max_{S\in\mathcal{C}} \mathbb{P}_S(\widehat{S} \neq S) \leq \delta$ and the precision budget (2.2).

In case of sparse signals, that is when $s \ll n$, and both $n_1 \approx \sqrt{n}$ and $n_2 \approx \sqrt{n}$ the first term on the right side dominates this bound. When $\max\{n_1, n_2\}$ is at the order of n, the situation becomes similar to the unstructured case, and the third term dominates the bound (so one recovers essentially the result in Corollary 2.1). Concerning the former case one has the following result:

Corollary 2.6 (s-submatrices)**.** *Consider the setting of Proposition 2.8. Assume $n_1 = n_2 = \sqrt{n}$ and $s = o(\sqrt{n}/\log n)$ as $n \to \infty$. Let $\omega_n \to \infty$ be arbitrary.*

(i) When

$$\mu \geq \omega_n\sqrt{\frac{n}{sm}}\,,$$

the procedure above gives an estimator $\widehat{S}$ such that $\lim_{n\to\infty}\max_{S\in\mathcal{C}} \mathbb{P}_S(\widehat{S} \neq S) = 0$, and that satisfies (2.2).

(ii) When

$$\mu \geq \sqrt{\frac{2n}{sm}(\log^2 s + \omega_n)} \ ,$$

the procedure above gives an estimator $\widehat{S}$ *such that* $\lim_{n\to\infty} \max_{S\in\mathcal{C}} \mathbb{E}_S(|\widehat{S}\triangle S|) = 0$*, and that satisfies* (2.2)*.*

Remark 2.4. *In all settings considered one can get exactly the same results with some degree of adaptivity to sparsity level, characterized by* s *and* k*. For instance, if one considers the class of all* k *and* $k-1$ *unions of* s*-intervals or* s*-stars then the results of Corollaries 2.3 and 2.5 still hold. Likewise mild adaptivity to* s *is also possible. Furthermore, all the results above will hold if the empty set is added to the class* $\mathcal{C}$ *under consideration.*

2.5 Lower Bounds

In this section we derive bounds for the signal strength μ for each special case considered earlier, such that if μ falls below these bounds, reliably recovering the support set $S \in \mathcal{C}$ is impossible. First we derive bounds for non-adaptive sensing for comparison purposes.

For the non-adaptive case we derive lower bounds considering the error metric $\mathbb{P}_S(\widehat{S} \neq S)$. These are lower bounds for the error metric $\mathbb{E}_S(|\widehat{S}\triangle S|)$ as well, since the latter dominates the former. The bounds we present for the non-adaptive case may not be sharp, particularly when the signal is not sparse. Nonetheless in the sparse setting they capture the essence of the difficulty of support recovery and illustrate well the gains one achieves by using adaptive sensing procedures.

For sharper bounds, and more discussion on lower bounding techniques for structured support sets in the non-adaptive setting, the reader is referred to Addario-Berry et.al. [1], Arias-Castro, Candès & Durand [8], Arias-Castro & Verzelen [15], Butucea & Ingster [31]. However, caution must be taken when comparing the previous results with the ones presented here, as the aforementioned results are bounds for the problem of detection, whereas those presented here concern the problem of estimation.

Afterwards, we derive lower bounds for adaptive sensing, which will show the near-optimality of the procedures proposed previously. In this setting, whenever we can, we prove bounds both for the Hamming-distance and the probability of error. The proofs of the results of this section make use of tools from Castro [40]

and Tsybakov [139].

2.5.1 Non-Adaptive Sensing

In this subsection we consider non-adaptive sensing for support recovery. The problem setting is the same as in Section 2.2, the only difference being that in the non-adaptive setting we have to specify $\{A_t, \Gamma_t\}_{t\geq 1}$ before any observations are made. All the bounds presented here are based on Proposition 2.3 in Tsybakov [139]. Recall that $D(\cdot\|\cdot)$ denotes the Kullback-Leibler divergence. The result states the following:

Lemma 2.1 (Proposition 2.3 in Tsybakov [139])**.** *Let $\mathbb{P}_0, \ldots, \mathbb{P}_M$ be probability measures on $(\mathcal{X}, \mathcal{A})$ satisfying*

$$\frac{1}{M}\sum_{j=1}^{M} D(\mathbb{P}_j\|\mathbb{P}_0) \leq \alpha \ ,$$

with $0 < \alpha < \infty$. For any $\mathcal{A}$-measurable function $\Psi : \mathcal{X} \to \{0, \ldots, M\}$

$$\max_{j=0,\ldots,M} \mathbb{P}_j(\Psi \neq j) \geq \sup_{0<\tau<1} \left(\frac{\tau M}{1+\tau M} \left(1 + \frac{\alpha + \sqrt{\alpha/2}}{\log \tau} \right) \right) .$$

We can use this result directly to get general lower bounds for μ in the non-adaptive setting. First let $\mathbb{P}_0, \ldots, \mathbb{P}_M$ be the probability measures induced by the sampling $\boldsymbol{x}$ with parameters $\{A_t, \Gamma_t\}_{t\geq 1}$, when the support sets are $S_0, \ldots, S_M$ respectively, where $S_i \in \mathcal{C}$. We take $S_0, \ldots, S_M$ to be all the support sets in $\mathcal{C}$, so that $M = |\mathcal{C}| - 1$. For fixed $S_k, S_l,\ k \neq l$ we have

$$D(\mathbb{P}_k\|\mathbb{P}_l) = \frac{\mu^2}{2} \sum_{t:A_t \in S_k \triangle S_l} \Gamma_t \ .$$

Let us define $b_i = \sum_{t:A_t=i} \Gamma_t$. Then

$$\sum_{j=1}^{M} D(\mathbb{P}_j\|\mathbb{P}_0) = \frac{\mu^2}{2} \sum_{S' \in \mathcal{C}\setminus\{S_0\}} \sum_{i \in S_0 \triangle S'} b_i \ .$$

We need to evaluate the quantity above. Since we can choose $S_0 \in \mathcal{C}$ freely, we can choose the one that makes the hypothesis test the hardest. On the other hand,

the measurement budget constraint (2.2) implies that $\sum_i b_i \leq m$. This yields the following optimization problem

$$\max_{\mathbf{b}\in\mathbb{R}^n_{+,0}:\ ||\mathbf{b}||_1\leq m} \ \min_{S\in\mathcal{C}} \ \sum_{S'\in\mathcal{C}\setminus\{S\}} \ \sum_{i\in S\triangle S'} b_i \ .$$

where $\mathbf{b} = (b_1, \ldots, b_n)^T$. The solution of this problem can be found explicitly if the class $\mathcal{C}$ under consideration has the following symmetry property (as introduced by Castro in [40]):

Definition 2.1. *Let $S \in \mathcal{C}$ be drawn uniformly at random. If $\mathbb{P}(i \in S) = s/n$ for all $i \in [n]$, then the class $\mathcal{C}$ is symmetric.*

We have the following proposition, proved in Appendix 2.D.

Proposition 2.9. *Suppose $\mathcal{C}$ is symmetric. Then*

$$\max_{\mathbf{b}\in\mathbb{R}^n_{+,0}:\ ||\mathbf{b}||_1\leq m} \ \min_{S\in\mathcal{C}} \ \sum_{S'\in\mathcal{C}\setminus\{S\}} \ \sum_{i\in S\triangle S'} b_i$$

is attained when $b_i = m/n,\ i \in [n]$.

We are now in position to prove the proposition which we can use to get lower bounds for μ in our special cases.

Proposition 2.10. *Let $\mathcal{C}$ be symmetric and suppose $1+\sqrt{2} \leq (1-2\varepsilon)\log(|\mathcal{C}|-1)$. If*

$$\mu^2 \leq (1-2\varepsilon)\frac{n}{2|S|m}\log(|\mathcal{C}|-1) \ ,$$

then no non-adaptive procedure can satisfy

$$\mathbb{P}_S(\widehat{S} \neq S) \leq \varepsilon,\ \forall S \in \mathcal{C} \ .$$

Proof. Let $\mathbb{P}_0, \ldots, \mathbb{P}_M$ be the probability measures induced by the sampling $\boldsymbol{x}$ with parameters $\{A_t, \Gamma_t\}_{t\geq 1}$, when the support sets are $S_0, \ldots, S_M$ respectively, where $S_i \in \mathcal{C}$. We take $S_0, \ldots, S_M$ to be all the support sets in $\mathcal{C}$, that is $M = |\mathcal{C}| - 1$. By Proposition 2.9 we know $b_i = m/n,\ i \in [n]$ is the optimal choice for distributing the precision in the non-adaptive setting for symmetric $\mathcal{C}$. From this we have

$$\frac{1}{M}\sum_{j=1}^{M} D(\mathbb{P}_j\|\mathbb{P}_0) \leq \max_{j=1,\ldots,M} D(\mathbb{P}_j\|\mathbb{P}_0) = |S|\frac{m}{n}\mu^2 := \alpha \ .$$

Then, by Lemma 2.1,

$$\sup_{S\in\mathcal{C}} \mathbb{P}_S(\widehat{S} \neq S) \geq \sup_{0<\tau<1} \left(\frac{\tau M}{1+\tau M} \left(1 + \frac{\alpha + \sqrt{\alpha/2}}{\log \tau} \right) \right) .$$

Setting $\tau = 1/M$ we get

$$\sup_{S\in\mathcal{C}} \mathbb{P}_S(\widehat{S} \neq S) \geq \frac{1}{2} \left(1 - \frac{\alpha + \sqrt{\alpha/2}}{\log M} \right) .$$

The right side of the above expression is bounded below by ε whenever

$$\alpha + \sqrt{\alpha/2} \leq (1-2\varepsilon) \log M \tag{2.13}$$

Plugging the values of a and M into the above inequality immediately yields bounds for μ. However, to make the bound more transparent, assume $\mathcal{C}$ is such that $1 + \sqrt{2} \leq (1-2\varepsilon)\log(|\mathcal{C}|-1)$. Then every α satisfying

$$2\alpha \leq (1-2\varepsilon) \log M$$

also satisfies (2.13). The statement now follows. □

Note that the condition $1+\sqrt{2} \leq (1-2\varepsilon)\log(|\mathcal{C}|-1)$ is not necessary to get the bound for μ, its role is merely to make the bound more transparent. Furthermore, it simply requires $\mathcal{C}$ to be large enough compared to ε. Since we are interested in cases where $\mathcal{C}$ is large and ε is small we can always safely assume this condition holds provided ε is small enough. The result of Proposition 2.10 is remarkably simple, as the lower bound depends exclusively on the cardinality of the class under consideration. With this in hand it is immediate to get non-adaptive lower bounds for all the classes considered in this chapter.

Theorem 2.1. *A necessary condition to ensure that any non-adaptive procedure satisfies* $\max_{S\in\mathcal{C}} \mathbb{P}_S(\widehat{S} \neq S) \leq \varepsilon,\ \forall S \in \mathcal{C}$ *is given by the following expressions, for the different classes* $\mathcal{C}$*:*

- *s-sets:* $\mu \geq \sqrt{(1-2\varepsilon)\frac{n}{2sm} \log\left(\binom{n}{s} - 1\right)}$.
- *s-intervals:* $\mu \geq \sqrt{(1-2\varepsilon)\frac{n}{2sm} \log\left(\frac{n}{s} - 1\right)}$.

- *unions of k disjoint s-intervals:*

$$\mu \geq \sqrt{(1-2\varepsilon)\frac{n}{2ksm}\log\left(\binom{n/s}{k}-1\right)}\,.$$

- *unions of k disjoint s-stars:*

$$\mu \geq \sqrt{(1-2\varepsilon)\frac{n}{2ksm}\log\left(\binom{p}{k(s+1)}-1\right)}$$

  (assume $k(s+1) \leq p$).

- *s-submatrices:* $\mu \geq \sqrt{(1-2\varepsilon)\frac{n}{2sm}\log\left(\binom{n_1}{\sqrt{s}}\binom{n_2}{\sqrt{s}}-1\right)}$.

Proof. The case of s-sets is straightforward from Proposition 2.10. The class of s-intervals is not symmetric, however, its subclass

$$\{[1,s],[s+1,2s],\ldots,[n-s+1,n]\}$$

is, therefore we can apply Proposition 2.10 for this subclass. A lower bound for any subclass is also a lower bound for the original class.

For the class of unions of intervals, we consider a similarly constructed subclass to get the bound above. In case of the unions of stars, we can consider the subclass of stars with distinct vertices. The size of this subclass is lower bounded by $\binom{p}{k(s+1)}$. For the submatrices, we consider the subclass of submatrices of size $\sqrt{s}\times\sqrt{s}$. □

Using the previous results we can state the following corollary considering the large n behavior of the non-adaptive lower bounds:

Corollary 2.7. *In order to have* $\lim_{n\to\infty}\max_{S\in\mathcal{C}}\mathbb{P}_S(\widehat{S}\neq S)=0$ *for $n\to\infty$ any non-adaptive procedure must satisfy, for some constant $c>0$, that*

- *s-sets:* $\mu \geq c\sqrt{\frac{n}{2m}\log\frac{n}{s}}$.
- *s-intervals:* $\mu \geq c\sqrt{\frac{n}{2sm}\log\frac{n}{s}}$.
- *unions of k disjoint s-intervals:* $\mu \geq c\sqrt{\frac{n}{2sm}\log\frac{n}{ks}}$.
- *unions of k disjoint s-stars:* $\mu \geq c\sqrt{\frac{n}{2m}\log\frac{\sqrt{2n}}{ks}}$.

- *s-submatrices:* $\mu \geq c\sqrt{\frac{n}{4\sqrt{s}m}\log\frac{n}{s}}$,
  when $n_1 = n_2 = \sqrt{n}$, $s_1 = s_2 = \sqrt{s}$.

The previous results shed some light on the limits of support recovery in the non-adaptive setting. When the size of the support set (s or ks) is sufficiently small relative to n, the $\log n$ factor is unavoidable for non-adaptive support recovery. On the contrary, this factor does not appear in the adaptive sensing performance bounds in any of the cases that we consider. For the class of unions of intervals, a factor of $\sqrt{1/s}$ appears in the above lower bounds, which means it might be possible to capitalize on the structure in the non-adaptive case as well (and this is indeed the case). For the unions of stars, however, this is no longer true. In fact this class is so large that non-adaptive procedures can no longer take significant advantage of the structure of the support sets. This is in stark contrast with what is possible with adaptive sensing (see Corollary 2.4). Similar remarks apply to the class of submatrices as well.

2.5.2 Adaptive Sensing

In this section we derive lower bounds for the signal strength μ in the adaptive sensing setting. We measure the performance of an estimator by the expected Hamming-distance $\mathbb{E}_S(|\widehat{S}\triangle S|)$. In some cases we are also able to prove lower bounds for the error metric $\mathbb{P}_S(\widehat{S} \neq S)$. Comparing the bounds of this section with the performance bounds of Section 2.4 shows the near optimality of the proposed procedure for sparse signals for all the classes considered.

s-sets (unstructured case)

This case is considered by Castro in [40] and lower bounds are shown for a slightly larger class, which consists of all s, $s-1$ and $s+1$ sets. However, it turns out that a similar result holds also if one considers the class of all s and $(s-1)$-sets only. Let $\mathcal{C}'$ denote this class, and suppose there is a sensing procedure and estimator $\widehat{S}$ for which

$$\max_{S\in\mathcal{C}'} \mathbb{E}_S(|\widehat{S}\triangle S|) \leq \varepsilon \ .$$

Lemma 4.1 in Castro [40] shows that it suffices to consider only *symmetric* estimators, which satisfy

$$\forall i,j \in S: \ \mathbb{P}_S(i \notin \widehat{S}) = \mathbb{P}_S(j \notin \widehat{S}) \ ,$$

and

$$\forall i,j \notin S:\ \mathbb{P}_S(i \notin \widehat{S}) = \mathbb{P}_S(j \notin \widehat{S}) \ ,$$

for any $S \in \mathcal{C}'$. This follows since any estimator $\widehat{S}$ can be symmetrized without affecting their worst case performance when the class under consideration is closed under permutations. It is easily shown that for symmetric estimators

$$\forall i,j \in S:\ \mathbb{E}_S\left(\sum_{t:A_t=i}\Gamma_t\right) = \mathbb{E}_S\left(\sum_{t:A_t=j}\Gamma_t\right) \ ,$$

and

$$\forall i,j \notin S:\ \mathbb{E}_S\left(\sum_{t:A_t=i}\Gamma_t\right) = \mathbb{E}_S\left(\sum_{t:A_t=j}\Gamma_t\right) \ ,$$

for any $S \in \mathcal{C}'$. We can then proceed as follows. Let $S \in \mathcal{C}'$ and $i \in [n]$ be arbitrary, and such that $|S| = s-1$ and $i \notin S$. Define also $S' = S \cup \{i\}$. For the event $\{i \notin \widehat{S}\}$ we have (by Theorem 2.6 and Lemma 2.6 of Tsybakov [139], see also Castro [40] for similar computations)

$$D(\mathbb{P}_S \| \mathbb{P}_{S'}) \geq -\log\left(2\mathbb{P}_S(i \notin \widehat{S}) + 2\mathbb{P}_{S'}(i \in \widehat{S})\right) \ .$$

Using the symmetry of the estimator we can easily bound the left hand side as

$$D(\mathbb{P}_S \| \mathbb{P}_{S'}) = \frac{\mu^2}{2}\mathbb{E}_S\left(\sum_{t:A_t=i}\Gamma_t\right) \leq \frac{\mu^2}{2}\frac{m}{n-s+1} \ .$$

Furthermore, also by symmetry

$$\mathbb{P}_S(i \in \widehat{S}) \leq \varepsilon/(n-s+1),\ \mathbb{P}_{S'}(i \notin \widehat{S}) \leq \varepsilon/s \ ,$$

whenever we have $\mathbb{E}_S(|\widehat{S} \triangle S|) \leq \varepsilon$. Putting everything together yields the following theorem:

Theorem 2.2. *Let $\mathcal{C}'$ denote the class of all subsets of $[n]$ with cardinality either $s-1$ or s. Suppose that $\max_{S \in \mathcal{C}'} \mathbb{E}_S(|\widehat{S} \triangle S|) \leq \varepsilon$. Necessarily*

$$\mu \geq \sqrt{\frac{2(n-s)}{m}\left(\log s + \log\frac{n-s}{n+1} + \log\frac{1}{2\varepsilon}\right)} \ .$$

In the large n regime and when considering sparse signals we have the following result:

Corollary 2.8 (s-sets)**.** *Consider the setting of Theorem 2.2, and suppose $s = o(n)$ as $n \to \infty$. If there is an adaptive sensing and estimation strategy such that $\lim_{n\to\infty} \max_{S\in\mathcal{C}'} \mathbb{E}_S(|\widehat{S}\triangle S|) = 0$ then necessarily*

$$\mu \geq \sqrt{\frac{2n}{m}(\log s + \omega_n)} \ ,$$

where $\omega_n \to \infty$.

This shows that, in the asymptotic regime, our adaptive-sensing procedure is near optimal in the unstructured case, when the signal is sparse.

Remark 2.5. *Note that the above lower bound considers the class of the union of s-sparse and $s-1$-sparse sets. Contrasting this, the procedure we considered in Section 2.4.2 for the unstructured case was designed for s-sparse sets. However, it is easy to see that the procedure has a mild adaptivity to sparsity and works well for this extended class as well.*

Several other lower bounds have a similar form, and remarks such as this apply in those cases as well.

s-intervals

In the case of s-intervals we can obtain a lower bound for the probability of error as the metric of interest. This lower bound is, however, a bit loose in the dependence on ε.

Proposition 2.11. *Let $\mathcal{C}$ be the class of s-intervals. Let $\varepsilon \in (0,1)$ and assume that $\max_{S\in\mathcal{C}} \mathbb{P}_S(\widehat{S} \neq S) \leq \varepsilon$. Then necessarily*

$$\mu \geq (1-\varepsilon)\sqrt{\frac{n}{2sm}} \ .$$

Proof. The reasoning below is inspired by a proof found in Balakrishnan et al. [17].

Assume without loss of generality that n/s is an even integer. Consider the class of disjoint intervals

$$\{[1,s],[s+1,2s],\ldots,[n-s+1,n]\} \ . \tag{2.14}$$

Since the above class is a subclass of $\mathcal{C}$ it suffices to show the lower bound for this smaller class. Now partition the class into two disjoint sets of the same size denoted by $\mathcal{C}_1$ and $\mathcal{C}_2$. To show the lower bound we consider a test of two simple hypothesis. Under H_1 assume $S \sim \text{Unif}(\mathcal{C}_1)$ and under H_2 assume $S \sim \text{Unif}(\mathcal{C}_2)$. In words, data under each hypothesis is generated by first selecting the support set S from either distribution, and then collecting data $D = \{Y_t, A_t, \Gamma_t\}_{t\geq 1}$ under the model indexed by S. Therefore this is a test between two simple hypotheses.

Given a support estimator $\widehat{S}$ one can construct a test

$$\Psi(D) = \begin{cases} 1 & \text{if } \widehat{S} \in \mathcal{C}_1, \\ 2 & \text{otherwise.} \end{cases}$$

Clearly, if $\mathbb{P}_S(\widehat{S} \neq S) \leq \varepsilon \ \forall S \in \mathcal{C}$, then $\mathbb{P}_1(\Psi(D) = 2) + \mathbb{P}_2(\Psi(D) = 1) \leq \varepsilon$, where $\mathbb{P}_i$ denotes the distribution of $D = \{Y_t, A_t, \Gamma_t\}_{t\geq 1}$ under H_i. Let $\mathbb{P}_0$ denote the distribution of D when $S = \emptyset$ and $TV(\cdot, \cdot)$ denote the total-variation distance. Assume without loss of generality that $TV(\mathbb{P}_0, \mathbb{P}_1) \geq TV(\mathbb{P}_0, \mathbb{P}_2)$. We have

$$\begin{aligned} \mathbb{P}_1(\Psi(D) = 2) + \mathbb{P}_2(\Psi(D) = 1) &\geq 1 - TV(\mathbb{P}_1, \mathbb{P}_2) \\ &\geq 1 - \Big(TV(\mathbb{P}_0, \mathbb{P}_1) + TV(\mathbb{P}_0, \mathbb{P}_2)\Big) \\ &\geq 1 - 2\ TV(\mathbb{P}_0, \mathbb{P}_1) \\ &\geq 1 - \sqrt{2\ D(\mathbb{P}_0 \| \mathbb{P}_1)}\ , \end{aligned} \tag{2.15}$$

where the second inequality follows from the triangle inequality, and the fourth inequality follows from the first Pinsker inequality, see Tsybakov [139].

The Kullback-Leibler divergence between $\mathbb{P}_0$ and $\mathbb{P}_1$ can be expressed as

$$\begin{aligned} D(\mathbb{P}_0 \| \mathbb{P}_1) &= \sum_t \mathbb{E}_0 \left(\log \frac{\mathrm{d}\mathbb{P}_0(Y_t | A_t, \Gamma_t)}{\mathrm{d}\mathbb{P}_1(Y_t | A_t, \Gamma_t)} \right) \\ &= -\sum_t \mathbb{E}_0 \left(\log \frac{\mathrm{d}\mathbb{P}_1(Y_t | A_t, \Gamma_t)}{\mathrm{d}\mathbb{P}_0(Y_t | A_t, \Gamma_t)} \right) \\ &= -\sum_t \mathbb{E}_0 \left(\log \frac{\frac{1}{|\mathcal{C}_1|} \sum_{S \in \mathcal{C}_1} \mathrm{d}\mathbb{P}_S(Y_t | A_t, \Gamma_t, S)}{\mathrm{d}\mathbb{P}_0(Y_t | A_t, \Gamma_t)} \right) \\ &= -\sum_t \mathbb{E}_0 \left(\log \frac{1}{|\mathcal{C}_1|} \sum_{S \in \mathcal{C}_1} \exp\left(-\frac{\Gamma_t}{2} \mu \mathbf{1}\{A_t \in S\}(\mu - 2Y_t) \right) \right)\ . \end{aligned}$$

Using Jensen's inequality for the final expression, we get

$$D(\mathbb{P}_0\|\mathbb{P}_1) \le -\sum_t \mathbb{E}_0\left(\frac{1}{|\mathcal{C}_1|}\sum_{S\in\mathcal{C}_1} -\frac{\Gamma_t}{2}\mu\mathbf{1}\{A_t\in S\}(\mu-2Y_t)\right) .$$

Therefore

$$\begin{aligned} D(\mathbb{P}_0\|\mathbb{P}_1) &\le \frac{1}{|\mathcal{C}_1|}\sum_{S\in\mathcal{C}_1}\mathbb{E}_0\left(\sum_t \frac{\Gamma_t}{2}\mu\mathbf{1}\{A_t\in S\}(\mu-2Y_t)\right) \\ &= \frac{1}{|\mathcal{C}_1|}\sum_{S\in\mathcal{C}_1}\mathbb{E}_0\left(\sum_{t:A_t\in S}\frac{\Gamma_t}{2}\mu^2\right) \\ &\le \frac{\mu^2}{2}\frac{1}{|\mathcal{C}_1|}\sum_{S\in\mathcal{C}_1\cup\mathcal{C}_2}\mathbb{E}_0\left(\sum_{t:A_t\in S}\Gamma_t\right) \le \frac{\mu^2}{2}\frac{2s}{n}m , \end{aligned}$$

where the second line follows from the law of total probability, by conditioning on $\{A_t,\Gamma_t\}_{t\ge1}$

From this and (2.15) we immediately get the result of the proposition. □

A closer look at the above proof gives an interesting insight. Note that in essence the previous proof claims that estimating an interval is as hard as the problem of detection, that is, deciding between the hypotheses $H_0:\ S=\emptyset$ and $H_1:\ S\in\mathcal{C}$. In fact, the method proposed in Section 2.4.2 already deals with this case, and exhibits the same performance if one "adds" the empty set to the class of s-intervals.

The following theorem gives lower bounds both when considering $\mathbb{P}_S(\widehat{S}\neq S)$ and $\mathbb{E}_S(|\widehat{S}\triangle S|)$ as the error metric, that also captures the dependence on ε.

Theorem 2.3. *Let $\mathcal{C}$ be the class of s-intervals. Let $\varepsilon\in(0,1)$.*

(i) If $\max_{S\in\mathcal{C}\cup\emptyset}\mathbb{P}_S(\widehat{S}\neq S)\le\varepsilon$, then necessarily

$$\mu \ge \sqrt{\frac{2n}{sm}\log\frac{1}{2\varepsilon}} .$$

(ii) If $\max_{S\in\mathcal{C}\cup\emptyset}\mathbb{E}_S(|\widehat{S}\triangle S|)\le\varepsilon$, then necessarily

$$\mu \ge \sqrt{\frac{2(n-s)}{sm}\left(\log\frac{n-s}{n+s}+\log\frac{s}{8\varepsilon}\right)} .$$

Proof. The assertion considering $\mathbb{P}_S(\widehat{S} \neq S)$ as the error metric immediately follows from Theorem 3.1 in Castro [40]. This theorem is directly applicable as having an estimator $\widehat{S}$ satisfying (i) implies having a test Ψ for the hypothesis testing problem

$$H_0 : \ S = \emptyset \qquad \text{versus} \qquad H_1 : \ S \in \mathcal{C}$$

with sum of type I and type II error probabilities no greater than ε.

As for the case when $\widehat{S}$ satisfies (ii), consider the following reduction of the problem. Let $\widetilde{\mathcal{C}} = \{\emptyset\} \cup \{S_i\}_{i\in[n/s]}$, where $\{S_i\}_{i\in[n/s]}$ is the class of disjoint consecutive intervals defined in (2.14). For sake of simplicity assume that n/s is an integer. It suffices to consider estimators of the form

$$\widehat{S} = \bigcup_{i\in[n/s]} S_i \ . \tag{2.16}$$

In words, the estimator can be written as a (possibly empty) union of elements from $\{S_i\}_{i\in[n/s]}$. It is not restrictive to consider such estimators since if one has an arbitrary estimator $\widehat{S}$ with expected number of errors at most ε, then we can define the estimator $\widetilde{S}$ of the form in (2.16) that has error at most 4ε. For instance let $\widetilde{S}$ be such that $S_i \subseteq \widetilde{S}$ if and only if $|\widehat{S} \cap S_i| \geq s/2$ for all $i \in [n/s]$. Then $\mathbb{E}_S(|\widetilde{S} \triangle S|) \leq 4\varepsilon$ for all $S \in \mathcal{C} \cup \{\emptyset\}$.

Considering such estimators we can write the expected number of errors as

$$\mathbb{E}_S(|\widehat{S} \triangle S|) = \sum_{i=1}^{n/s} s \ \mathbb{P}_S(\mathbf{1}\{S_i \subseteq \widehat{S}\} \neq \mathbf{1}\{S_i \subseteq S\}) \ .$$

This means that the above problem is similar to that of Theorem 2.2 with a vector of length n/s, and support set size at most 1 (but error bounded by $4\varepsilon/s$), concluding the proof. □

In the asymptotic regime for sparse signals we have the following corollary, which shows that the procedure proposed in Section 2.4.2 is nearly optimal when considering both error metrics.

Corollary 2.9 (s-intervals)**.** *Consider the setting of Theorem 2.3, and suppose* $s = o(n)$ *as* $n \to \infty$.

(i) If $\lim_{n\to\infty} \max_{S\in\mathcal{C}\cup\emptyset} \mathbb{P}_S(\widehat{S} \neq S) = 0$*, then necessarily*

$$\mu \geq \omega_n \sqrt{\frac{n}{sm}} \ ,$$

(ii) If $\lim_{n\to\infty} \max_{S\in\mathcal{C}\cup\emptyset} \mathbb{E}_S(|\widehat{S}\triangle S|) = 0$*, then necessarily*

$$\mu \geq \sqrt{\frac{2n}{sm}(\log s + \omega_n)} \ ,$$

where ω_n *is an arbitrary sequence such that* $\omega_n \to \infty$*.*

Unions of s-intervals

Consider again a slight modification of the class of interest, namely let $\mathcal{C}$ denote the class of all disjoint unions of k or $(k-1)$ s-intervals. Similarly to the previous case, we reduce the problem to look like the general s-sparse case, and apply Theorem 2.2.

Theorem 2.4. *Let* $\varepsilon > 0$ *and suppose that* $\max_{S\in\mathcal{C}} \mathbb{E}_S(|\widehat{S}\triangle S|) \leq \varepsilon$*. Then necessarily*

$$\mu \geq \sqrt{\frac{2(n-sk)}{sm}\left(\log k + \log\frac{n-sk}{n+s} + \log\frac{s}{8\varepsilon}\right)} \ .$$

Proof. Assume again for sake of simplicity that n/s is an integer and consider the class of consecutive s-intervals $\{S_i\}_{i\in[n/s]}$ defined in (2.14). Let $\widetilde{\mathcal{C}} \subset \mathcal{C}$ be the class that contains unions of k or $(k-1)$ elements of $\{S_i\}_{i\in[n/s]}$. It suffices to consider only estimators that satisfy (2.16) since if there is a general estimator $\widehat{S}$ for which $\max_{S\in\mathcal{C}} \mathbb{E}_S(|\widehat{S}\triangle S|) \leq \varepsilon$ for all $S \in \mathcal{C}$ then there is an estimator $\widetilde{S}$ of the form (2.16) satisfying $\max_{S\in\mathcal{C}} \mathbb{E}_S(|\widetilde{S}\triangle S|) \leq 4\varepsilon$. Therefore the problem can once again be viewed as the unstructured case involving a vector of length n/s and sparsity k or $(k-1)$, and requiring that the estimator has expected Hamming-distance at most $4\varepsilon/s$. Using Theorem 2.2 concludes the proof. □

Corollary 2.10 (Unions of s-intervals)**.** *Consider the setting of Theorem 2.4, and suppose* $sk = o(n)$ *as* $n \to \infty$*. If there is an adaptive sensing and estimation*

strategy such that $\lim_{n\to\infty} \max_{S\in\mathcal{C}} \mathbb{E}_S(|\widehat{S}\triangle S|) = 0$*, then necessarily*

$$\mu \geq \sqrt{\frac{2n}{sm}(\log ks + \omega_n)} \ ,$$

where ω_n *is an arbitrary sequence for which* $\omega_n \to \infty$*.*

The previous statements show the near-optimality of the procedure proposed in Section 2.4.2.

s-stars and unions of s-stars

The lower bounds for this class follow from similar arguments as the ones used for s-intervals by considering a maximal subclass of disjoint s-stars (meaning these do not share any edges). Let $\mathfrak{N}_{p,s}$ be the size of such a subclass. We have the following lemma for which we provide a short proof in Appendix 2.E:

Lemma 2.2. *Let* $\mathfrak{N}_{p,s}$ *denote the maximal number of disjoint stars of size* s *in a complete graph with* p *vertices. Then*

$$\mathfrak{N}_{p,s} \geq \frac{p(p-1-s)}{2s} \ .$$

With this in mind we can get a performance lower bound, proved in an analogous way to that of Proposition 2.11.

Proposition 2.12. *Let* $\mathcal{C}$ *be the class of* s*-stars. Assume that* $\max_{S\in\mathcal{C}} \mathbb{P}_S(\widehat{S} \neq S) \leq \varepsilon$*. Then necessarily*

$$\mu \geq (1-\varepsilon)\sqrt{\frac{\mathfrak{N}_{p,s}}{2m}} \ .$$

We can also get results analogous to Theorems 2.3 and 2.4, and Corollaries 2.9 and 2.10. We only state the results for the unions of s-stars, which shows the near-optimality of the proposed procedure.

Theorem 2.5. *Let* $\mathcal{C}'$ *denote the class of unions of* k *or* $k-1$ *disjoint* s*-stars. Let* $\varepsilon > 0$ *and suppose that* $\max_{S\in\mathcal{C}} \mathbb{E}_S(|\widehat{S}\triangle S|) \leq \varepsilon$*. Then necessarily*

$$\mu \geq \sqrt{\frac{2(\mathfrak{N}_{p,s}-k)}{m}\left(\log k + \log\frac{\mathfrak{N}_{p,s}-k}{\mathfrak{N}_{p,s}+1} + \log\frac{s}{8\varepsilon}\right)} \ .$$

Corollary 2.11 (Unions of s-stars)**.** *Consider the setting of Theorem 2.5, and suppose $ks = o(n)$ as $n \to \infty$. If there is an adaptive sensing and estimation strategy such that $\lim_{n\to\infty} \max_{S\in\mathcal{C}} \mathbb{E}_S(|\widehat{S}\triangle S|) = 0$, then necessarily*

$$\mu \geq \sqrt{\frac{2n}{sm}(\log ks + \omega_n)} \ ,$$

where ω_n is an arbitrary sequence for which $\omega_n \to \infty$.

s-submatrices

Again, akin to the s-intervals and s-stars we can get lower bounds without including the empty set to the class when considering the probability of error as the metric of interest.

Proposition 2.13. *Let $\mathcal{C}$ be the class of submatrices, and suppose that n_1, n_2, s are such that we can cover the matrix M entirely with disjoint submatrices of size s. Let $\varepsilon > 0$ and suppose $\max_{S\in\mathcal{C}} \mathbb{P}_S(\widehat{S} \neq S) \leq \varepsilon$. Then necessarily*

$$\mu \geq (1-\varepsilon)\sqrt{\frac{n}{2sm}} \ .$$

We can once more derive results when including the empty set in the class.

Theorem 2.6. *Suppose that n_1, n_2, s are such that we can cover the matrix M entirely with disjoint submatrices of size s, and let $\varepsilon > 0$.*

(i) If $\max_{S\in\mathcal{C}\cup\{\emptyset\}} \mathbb{P}_S(\widehat{S} \neq S) \leq \varepsilon$, then necessarily

$$\mu \geq \sqrt{\frac{2n}{sm}\log\frac{1}{2\varepsilon}} \ ,$$

(ii) If $\max_{S\in\mathcal{C}\cup\{\emptyset\}} \mathbb{E}_S(|\widehat{S}\triangle S|) \leq \varepsilon$, then necessarily

$$\mu \geq \sqrt{\frac{2(n-s)}{sm}\left(\log\frac{n-s}{n+s} + \log\frac{s}{8\varepsilon}\right)} \ .$$

The condition about the relation between n_1, n_2 and s in the previous results is merely to simplify presentation. This can be easily relaxed by bounding the

number of disjoint submatrices of size s in a matrix of size $n_1 \times n_2$. For instance, one such bound is $\max\{n_1 \cdot \lfloor n_2/s \rfloor, n_2 \cdot \lfloor n_1/s \rfloor\}$. This condition does not play a role when considering the behavior of the bound in the asymptotic regime for sparse signals.

Corollary 2.12 (s-submatrices)**.** *Consider the setting of Theorem 2.6, and suppose $n \to \infty$ such that $s = o(n)$.*

(i) If $\lim_{n\to\infty} \max_{S \in \mathcal{C} \cup \{\emptyset\}} \mathbb{P}_S(\widehat{S} \neq S) = 0$, *then*

$$\mu \geq \omega_n \sqrt{\frac{2n}{sm}} \ ,$$

(ii) If $\lim_{n\to\infty} \max_{S \in \mathcal{C} \cup \{\emptyset\}} \mathbb{E}_S(|\widehat{S} \triangle S|) = 0$, *then*

$$\mu \geq \sqrt{\frac{2n}{sm} (\log s + \omega_n)} \ ,$$

where ω_n is an arbitrary sequence such that $\omega_n \to \infty$.

The previous results show the near optimality of the procedure proposed in Section 2.4.2, in the case when n_1, n_2 are the same order of magnitude. When either of them is close to n the problem becomes similar to the unstructured case (however, this is not captured by the above corollary).

2.6 A Numerical Experiment

We present a short numerical experiment, to corroborate the theoretical results presented in this chapter. Note that the simulations here only serve an illustrative purpose and are by no means exhaustive.

In this simulation we gauge the performance of the adaptive sensing procedure presented in Section 2.3 for the class of s-intervals, and compare it with a reasonable non-adaptive procedure[9].

The non-adaptive procedure is as follows. We use a fixed precision budget m, and sample the signal uniformly with precision m/n. Then we pick the support es-

[9] Although we do not make a formal claim that the non-adaptive procedure implemented is indeed optimal, it is likely asymptotically optimal, as it is simly a maximum-likelihood estimator.

timator $\widehat{S}_{\text{na}}$ to be the s-interval with the highest sum of observations, and compute the Hamming distance $|S \triangle \widehat{S}_{\text{na}}|$, where S denotes the true support.

The adaptive sensing procedure is based on the one presented in Section 2.3. Note that the procedure makes a random number of measurements, and the theory we developed in Section 2.4 deals with the expected precision used. To make the comparison fair between adaptive and non-adaptive algorithms, we terminate the adaptive sensing procedure whenever the precision budget m is reached, and incur a loss of $2s$ in terms of Hamming distance. Other than that, we choose every parameter as described in Section 2.4, set $\varepsilon = 0.05$ and $\Gamma = 0.2$. Though the theory in Section 2.4 suggest we need to choose Γ close to zero, we chose a non-negligible value to see how such a choice affects the performance in practice. The performance of the procedure is evaluated by $|S \triangle \widehat{S}_{\text{a}}|$, where $\widehat{S}_{\text{a}}$ denotes the support estimator returned by the procedure.

Proposition 2.4 suggests that the adaptive sensing algorithm should satisfy $\mathbb{E}(|S \triangle \widehat{S}_{\text{a}}|) \leq \varepsilon$ when the signal strength is roughly

$$\mu_{\text{limit}} = \sqrt{\frac{8n}{s^2 m} \log \frac{4s}{\varepsilon}} \ . \tag{2.17}$$

Furthermore, according to the lower bounds of Theorem 2.3, no estimator can have small probability of error unless the signal scales as above. Note that for the SLRT at the core of Algorithm 1, we need to specify an alternative to test against, or in other words, a value for μ_{limit}. This affects the amount of precision the SLRT will use. Note that this does not mean we use knowledge of the true signal strength, as this is only a parameter of the SLRT.

We set 2 different values for this parameter: μ_{limit} defined above, and $\mu_{\text{limit}}^{(.95)}$. The latter is also defined by the formula above, but with m replaced with $0.95 \cdot m$. This means that we tune the procedure to detect a signal that is slightly larger then μ_{limit} (roughly by a factor of 1.02), but in turn this will result in a slightly smaller amount of precision used by the SLRTs (in expectation).

We run the procedures described above when the true signal strength is $t \cdot \mu_{\text{limit}}$ with the value of t varying. We set the signal dimension to be $n = 2^{15}$, the support size to be $s = 2^4$ and the precision budget to be $m = n$. We run 100 iterations for every value of the parameter t, and plot the average normalized Hamming distance of the different estimators. We also plot error bars whose total length is four times the (point-wise) standard error, which would correspond to a roughly 95% two-

sided confidence interval for normally distributed measurements. Note that the error bars are only approximate point-wise confidence bands, that are included to provide some insight about the variability of the curves.

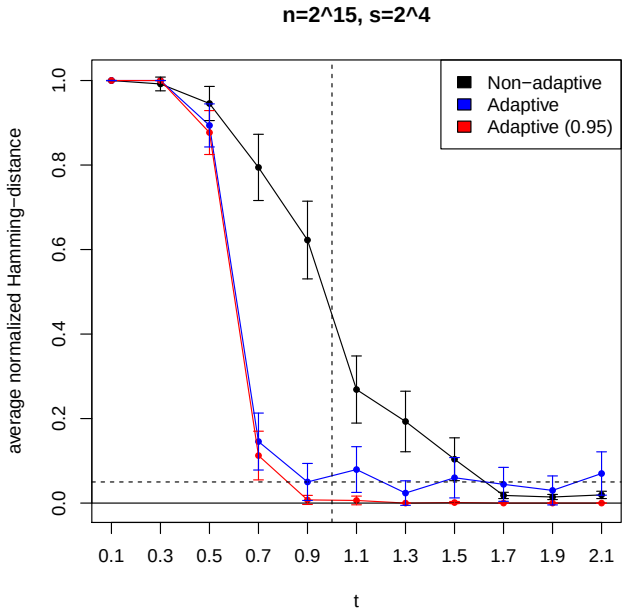

Figure 2.1: Average normalized Hamming-distance (with SE bands) for the different estimators as a function of the parameter t (the signal strength is $t \cdot \mu_{\text{limit}}$ with μ_{limit} defined in (2.17)): the non-adaptive estimator (black); adaptive sensing (AS) procedure calibrated with μ_{limit} (blue); AS calibrated with $\mu_{\text{limit}}^{(.95)}$ (red). The number of repetitions is 100 for each value of t. The vertical black dashed line is at the value $t = 1$. The horizontal black dashed line is at the value of ε (0.05).

As expected, the adaptive sensing procedures outperform the non-adaptive one. We also expect the adaptive procedures to reach the level $\varepsilon = 0.05$ at $t = 1$. In fact this level is reached somewhat earlier, which might be the result of the conservative algorithm choices, which we already highlighted in Section 2.4.

Perhaps remarkably, the adaptive sensing procedure calibrated to detect a signal strength of μ_{limit} also performs well, even though it is terminated when the precision budget is reached. Note however, that the analysis in Section 2.4 was carried out for the worst case scenario (when we need to perform all n/s tests in the search

phase). As we see above, often times the procedure does not need to perform that many tests, and thus still satisfies a strict precision budget.

Note however, that this procedure cannot have an arbitrarily small error, because regardless of the signal strength, the procedure will sometimes exhaust the precision budget in the search phase. However, as Figure 2.1 seems to illustrate, by tuning the procedure to detect a signal slightly larger then μ_{limit} boundary, this issue is circumvented (see also Appendix 2.A for more detailed comments about a strict budget constraint).

2.7 Final Remarks

In this work we have investigated the problem of support estimation of structured sparse signals under the adaptive sensing paradigm. These results broaden our understanding of the fundamental limits of adaptive sensing and also provide a method for estimating signal supports. The procedure suggested in this chapter is rather general and simple, and also turns out to be near-optimal in a variety of interesting cases.

It is important to point out that the proposed procedure requires knowledge of some parameters of the problem that might not be available in a real-life setting. Also, neither the fundamental performance limits nor the performance of the proposed procedure are yet fully understood for arbitrary classes of signal support sets. These might prove to be interesting areas for future research.

2.A Removing the expectation from the budget constraint (2.2)

We now investigate what difference it would make if we considered a more demanding budget constraint by removing the expectation from (2.2). That is, we now wish to consider algorithms that satisfy

$$\sup_{S \in \mathcal{C}} \sum_t \Gamma_t \leq m \ .$$

First, note that all the lower bounds remain valid with the latter constraint as well, since the constraint $\mathbb{E}_S(\sum_t \Gamma_t) \leq m$ is more forgiving than $\sum_t \Gamma_t \leq m$.

The basis of the procedures in Chapter 2 is the SLRT described in Section 2.4.1. To accommodate for the more strict budget constraint we need to change the bound on the expected energy for the SLRT of Proposition 2.1 to a high probability bound. We will only discuss this in detail for the s-sets as results for all other procedures follow similarly.

As a reminder, the procedure for s-sets consists of independently performing a Sequential Likelihood-ratio test (SLRT) for each component to assess whether that component is zero or not. To carry out the test for $\boldsymbol{x}_i$ we take

$$N_i = \inf \left\{ n \in \mathbb{N} : \sum_{j=1}^{n} \log \frac{\mathrm{d}\mathbb{P}_1(Y_{i,j})}{\mathrm{d}\mathbb{P}_0(Y_{i,j})} \notin (l, u) \right\}$$

measurements, where $\log \frac{\beta}{1-\alpha} = l < 0 < u = \log \frac{1-\beta}{\alpha}$ are the lower and upper stopping boundaries. Recall that all measurements are made with a fixed precision Γ. Suppose H_0 is true. In this case the upper bound on the expected number of measurements used by the test is $t_0 := \frac{2}{\Gamma \mu^2} \log \frac{2s}{\varepsilon}$ since we set $\beta = \frac{\varepsilon}{2s}$. For our purposes it is enough to show that

$$\mathbb{P}\left(\sum_{i \notin S} N_i > c(n-s)t_0 \right) \leq c'\varepsilon \ , \tag{2.18}$$

for some universal constants c, c'. If this (and a comparable result for $i \in S$) were true, then a union bound would give that the probability that the procedure uses more then cm energy is at most $2c'\varepsilon$. One then could construct a similar procedure as before with the exception that it is forced to stop once the precision budget is exhausted. By the previous result this happens with probability proportional to ε. Hence the minimum signal strength required by a procedure satisfying $\sum_t \Gamma_t \leq m$ for support recovery would still be on the same order as before, only the constants would need to be adjusted.

To show the result above we need a concentration inequality. As a start, we

show a simple tail bound for N_i under the null.

$$\begin{aligned}\mathbb{P}_0(N_i > ct_0) &\leq \mathbb{P}_0\left(\sum_{j=1}^{ct_0} \log \frac{\mathrm{d}\mathbb{P}_1(Y_{i,j})}{\mathrm{d}\mathbb{P}_0(Y_{i,j})} > l\right) \\ &\leq \mathbb{P}_0\left(\sum_{j=1}^{ct_0} \log \frac{\mathrm{d}\mathbb{P}_1(Y_{i,j})}{\mathrm{d}\mathbb{P}_0(Y_{i,j})} > \log \beta\right) \ .\end{aligned}$$

Note that $\mathrm{d}\mathbb{P}_1(Y_{i,j})/\,\mathrm{d}\mathbb{P}_0(Y_{i,j}) = \Gamma(\mu Y_{i,j} - \frac{\mu^2}{2})$, which is distributed as $N(-\frac{\Gamma\mu^2}{2}, \Gamma\mu^2)$ under the null. Using this with the Gaussian tail bound $\mathbb{P}(\xi > x) \leq e^{-x^2/2}/2$ we get

$$\mathbb{P}_0\left(\sum_{j=1}^{ct_0} \log \frac{\mathrm{d}\mathbb{P}_1(Y_{i,j})}{\mathrm{d}\mathbb{P}_0(Y_{i,j})} > \log \beta\right) \leq \frac{1}{2}\exp\left(-\frac{(\log\beta + ct_0\Gamma\mu^2/2)^2}{2ct_0\Gamma\mu^2}\right) \ .$$

Plugging in $t_0 = \frac{2}{\Gamma\mu^2}\log\frac{2s}{\varepsilon}$ and $\beta = \frac{\varepsilon}{2s}$, we get

$$\mathbb{P}_0(N_i > ct_0/\Gamma) \leq \frac{1}{2}\left(\frac{\varepsilon}{2s}\right)^{\frac{(c-1)^2}{4c}} \ ,$$

when $c > 2$. We continue by using the Craig-Bernstein inequality [53] that states that whenever the independent random variables $U_1, \ldots, U_n$ satisfy the moment condition

$$\mathbb{E}\left(|U_i - \mathbb{E}(U_i)|^k\right) \leq \frac{\mathrm{Var}(U_i)}{2}k!h^{k-2}, \ i = 1, \ldots, n \ ,$$

with some $h > 0$ then

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(U_i - \mathbb{E}(U_i)) \geq \frac{z}{n\delta} + \frac{n\delta\,\mathrm{Var}(\frac{1}{n}\sum_{i=1}^{n}U_i)}{2(1-C)}\right) \leq e^{-z} \ ,$$

for $0 < h\delta \leq C < 1$ and $z > 0$. We thus need to refine the calculations above to get a general moment bound for N_i and then we will use the inequality above with

$C = 1/2, \delta = 1/2h$ and an appropriate z. We start with the moment condition.

$$\begin{aligned}\mathbb{E}(N_i^k) &= \sum_{j=1}^{\infty} j^k \mathbb{P}(N_i = j) \\ &\leq \sum_{c=1}^{\infty} (ct_0)^k \mathbb{P}((c-1)t_0 < N_i \leq ct_0) \\ &\leq \sum_{c=1}^{\infty} (ct_0)^k \mathbb{P}((c-1)t_0 < N_i) \ .\end{aligned}$$

Plugging in the previous tail bound for N_i we get

$$\mathbb{E}(N_i^k) \leq t_0^k \left(1 + \sum_{c=2}^{\infty} c^k \frac{1}{2} \left(\frac{\varepsilon}{2s}\right)^{\frac{(c-2)^2}{4c}}\right) \leq t_0^k \left(1 + \sum_{c=2}^{\infty} c^k \frac{1}{2} \varepsilon^{\frac{(c-2)^2}{4c}}\right) \ .$$

Using $\varepsilon \leq 1$ we simply get

$$\begin{aligned}\mathbb{E}(N_i^k) &\leq t_0^k \left(1 + \frac{1}{2}(2^k + e^k + 4^k) + \sum_{c=5}^{\infty} c^k \varepsilon^{\frac{(c-2)^2}{4c}}\right) \\ &\leq t_0^k \left(1 + \frac{1}{2}(2^k + 3^k + 4^k) + \frac{1}{\varepsilon} \sum_{c=5}^{\infty} c^k \varepsilon^{c/4}\right) \ ,\end{aligned}$$

using the tail bound on N_i (also using $\varepsilon \leq 1/2$). We upper bound the sum in the last expression by an integral.

$$\begin{aligned}\sum_{c=5}^{\infty} c^k \varepsilon^{c/4} &\leq \int_4^{\infty} (x+1)^k \sqrt[4]{\varepsilon}^x \, \mathrm{d}x \\ &= \left[\frac{4(x+1)^k \sqrt[4]{\varepsilon}^x}{\log \varepsilon}\right]_4^{\infty} - \frac{4k}{\log \varepsilon} \int_4^{\infty} (x+1)^{k-1} \sqrt{\varepsilon}^x \, \mathrm{d}x \\ &= \frac{45^k \varepsilon}{\log \frac{1}{\varepsilon}} + \frac{4k}{\log \frac{1}{\varepsilon}} \int_0^{\infty} (x+1)^{k-1} \sqrt{\varepsilon}^x \, \mathrm{d}x \\ &= \dots \\ &= \varepsilon \sum_{l=0}^{k} \left(\frac{4}{\log \frac{1}{\varepsilon}}\right)^{l+1} \frac{k!}{(k-l)!} 5^{k-l} \\ &\leq \varepsilon k! 5^k \sum_{l=0}^{\infty} \left(\frac{2}{\log \frac{1}{\varepsilon}}\right)^l \leq \varepsilon k! 5^k \frac{\log \frac{1}{\varepsilon}}{\log \frac{1}{\varepsilon} - 2} \ ,\end{aligned}$$

using repeated partial integration. Plugging this back yields

$$\mathbb{E}(N_i^k) \leq t_0^k(1 + \frac{1}{2}(2^k + 3^k + 4^k) + k!5^k \frac{\log \frac{1}{\varepsilon}}{\log \frac{1}{\varepsilon} - 2}) \leq k!(Kt_0)^k \ ,$$

with some constant K. Note that the variance is on the order of t_0^2, the moment condition above is satisfied with $h = K't_0$, where K' is some constant. Hence taking $z = \log \frac{1}{\varepsilon}$ (and using $\mathrm{Var}(N_i) \leq 2kt_0^2$), the Craig-Bernstein inequality yields

$$\mathbb{P}\left(\sum_{i \notin S}(N_i - \mathbb{E}(N_i)) \geq 2K' \log \frac{1}{\varepsilon} t_0 + \frac{K}{K'}(n-s)t_0\right) \leq \varepsilon \ .$$

Unless ε is very small (less than $e^{-(n-s)}$), the expression on the left side of the inequality above is upper bounded by $c(n-s)t_0$ and thus we have shown (2.18).

2.B Proof of Proposition 2.1

To ease notation we write $N \equiv N_\Gamma$. The proof of all the statements in the proposition hinges on the derivation of upper bounds for the expected value of the stopping time N. Recall the definition of the log-likelihood ratio

$$\bar{z}_k = \sum_{i=1}^{k} \log \frac{f_1(y_i)}{f_0(y_i)} = \sum_{i=1}^{k} z_i \ ,$$

where $z_i = \frac{\Gamma}{2}\mu(2y_i - \mu)$. From Wald's identity [141] we know that

$$\mathbb{E}(\bar{z}_N) = \mathbb{E}(N)\mathbb{E}(z_1) \ .$$

Since it is easy to compute $\mathbb{E}(z_1)$ directly, in order to control $\mathbb{E}(N)$ we need to control $\mathbb{E}(\bar{z}_N)$. Note that $\mathbb{E}_0(z_1) < 0 < \mathbb{E}_1(z_1)$, thus to get an upper bound on $\mathbb{E}_0(N)$ we need to lower bound $\mathbb{E}_0(\bar{z}_N)$, and to get an upper bound on $\mathbb{E}_1(N)$ we need to upper bound $\mathbb{E}_1(\bar{z}_N)$. In what follows assume H_0 is true, as the case for H_1 is entirely analogous. Our proof hinges on the following technical lemma.

Lemma 2.3.

$$l + \mathbb{E}_0(z_1|z_1 \leq 0) \leq \mathbb{E}_0(\bar{z}_N|\bar{z}_N \leq l) \leq l \ ; \tag{2.19}$$

$$e^l \mathbb{E}_0 \left(e^{z_1} \middle| \, z_1 \leq 0 \right) \leq \mathbb{E}_0 \left(e^{\bar{z}_N} \middle| \, \bar{z}_N \leq l \right) \leq e^l \ ; \tag{2.20}$$

$$e^u \leq \mathbb{E}_0 \left(e^{\bar{z}_N} \middle| \, \bar{z}_N \geq u \right) \leq e^u \mathbb{E}_0 \left(e^{z_1} \middle| \, z_1 \geq 0 \right) \ . \tag{2.21}$$

Proof. We prove only the first statement, as the proof of the other two statements follow with essentially the same reasoning. First note that for any normal random variable $\xi \sim N(\nu, \sigma^2)$ and $c \leq 0$ we have

$$\mathbb{E}(\xi - c|\xi \leq c) \geq \mathbb{E}(\xi|\xi \leq 0) \ . \tag{2.22}$$

This can be justified by writing the conditional densities of $\xi - c|\xi \leq c$ and $\xi|\xi \leq 0$, respectively

$$f_{\xi - c|\xi \leq c}(x) = K_1 e^{-\frac{((x-\nu)+c)^2}{2\sigma^2}} \mathbf{1}\{x \leq 0\}$$
$$f_{\xi|\xi \leq 0}(x) = K_2 e^{-\frac{(x-\nu)^2}{2\sigma^2}} \mathbf{1}\{x \leq 0\} \ ,$$

where $K_2 > K_1 > 0$ are the appropriate normalization constants. It is easy to show that these densities satisfy

$$f_{\xi - c|\xi \leq c}(x) \leq f_{\xi|\xi \leq 0}(x) \quad \text{if } x \leq x_0 \ ,$$
$$f_{\xi - c|\xi \leq c}(x) \geq f_{\xi|\xi \leq 0}(x) \quad \text{if } x \geq x_0 \ ,$$

where x_0 is simply given by

$$x_0 = \frac{2\sigma^2 \log \frac{K_1}{K_2} - c^2}{2c} + \nu \ .$$

This, in turn implies (2.22), as

$$\begin{aligned}\mathbb{E}(\xi - c|\xi \le c) - \mathbb{E}(\xi|\xi \le 0) &= \int x f_{\xi - c|\xi \le c}(x)\,\mathrm{d}x - \int x f_{\xi|\xi\le 0}(x)\,\mathrm{d}x \\ &= \int x\left(f_{\xi - c|\xi\le c}(x) - f_{\xi|\xi\le 0}(x)\right)\,\mathrm{d}x \\ &\quad - x_0 \int f_{\xi - c|\xi\le c}(x) - f_{\xi|\xi\le 0}(x)\,\mathrm{d}x \\ &= \int (x - x_0)\left(f_{\xi - c|\xi\le c}(x) - f_{\xi|\xi\le 0}(x)\right)\,\mathrm{d}x \\ &\ge 0\ .\end{aligned}$$

We are now ready to prove the lemma. First note that

$$z_1 = \log\frac{f_1(y_1)}{f_0(y_1)} = \Gamma\mu y_1 - \frac{\Gamma\mu^2}{2} \overset{H_0}{\sim} N\left(-\frac{\Gamma}{2}\mu^2, \Gamma\mu^2\right)\ .$$

Therefore

$$\begin{aligned}\mathbb{E}_0\Big(\bar{z}_N\Big|\bar{z}_N \le l\Big) =& \mathbb{E}_0\Big(\mathbb{E}_0\left(\bar{z}_N\,|N, \bar{z}_{N-1}, \bar{z}_N \le l\right)\Big|\bar{z}_N \le l\Big) \\ =& l + \mathbb{E}_0\Big(\mathbb{E}_0\left(z_N - (l - \bar{z}_{N-1})\big|N, \bar{z}_{N-1}, z_N \le l - \bar{z}_{N-1}\right)\Big|\bar{z}_N \le l\Big) \\ \ge& l + \mathbb{E}_0\Big(\mathbb{E}_0\left(z_N\,|N, \bar{z}_{N-1}, z_N \le 0\right)\Big|\bar{z}_N \le l\Big) \\ =& l + \mathbb{E}_0\Big(\mathbb{E}_0\left(z_1\,|N, \bar{z}_{N-1}, z_1 \le 0\right)\Big|\bar{z}_N \le l\Big) \\ =& l + \mathbb{E}_0\Big(z_1\,|z_1 \le 0\Big)\ ,\end{aligned}$$

where the inequality follows from (2.22), concluding the proof of statement (2.19). The other two statements are shown in a similar fashion, by noting also that the exponential function is monotone increasing. □

With the lemma result at hand, note that

$$\begin{aligned}\mathbb{E}_0(\bar{z}_N) &= \alpha_\Gamma \mathbb{E}_0(\bar{z}_N|\bar{z}_N \ge u) + (1 - \alpha_\Gamma)\mathbb{E}_0(\bar{z}_N|\bar{z}_N \le l) \\ &\ge \alpha_\Gamma u + (1 - \alpha_\Gamma)l + (1 - \alpha_\Gamma)\mathbb{E}_0(z_1|z_1 \le 0)\ ,\end{aligned}$$

where the last step follows simply from the lemma. Using this together with Wald's

inequality yields

$$\begin{aligned}\Gamma\mathbb{E}_0(N) &\leq -\frac{2}{\mu^2}\Big(\alpha_\Gamma u + (1-\alpha_\Gamma)l + (1-\alpha_\Gamma)\mathbb{E}_0(z_1|z_1\leq 0)\Big)\\ &= \frac{2}{\mu^2}\Big(\alpha_\Gamma \log\frac{\alpha}{1-\beta} + (1-\alpha_\Gamma)\log\frac{1-\alpha}{\beta}\\ &\quad - (1-\alpha_\Gamma)\mathbb{E}_0(z_1|z_1\leq 0)\Big)\ . \end{aligned} \tag{2.23}$$

So, provided we can show that $\alpha_\Gamma \to \alpha$ and $\mathbb{E}_0(z_1|z_1 \leq 0) \to 0$ as $\Gamma \to 0$ the statement of the proposition follows, as we obtain the same limit as in (2.8).

Note first that $\mathbb{P}_0(z_1 \leq 0) = \Phi(\mu\sqrt{\Gamma}/2) \to 1/2$ as $\Gamma \to 0$, where Φ denotes the standard normal cumulative distribution function. Since $-|z_1| \leq z_1\mathbf{1}\{z_1 \leq 0\} \leq 0$ we conclude that $\mathbb{E}_0(z_1\mathbf{1}\{z_1 \leq 0\}) \to 0$ when $\Gamma \to 0$, since $\mathbb{E}_0(|z_1|) \leq \sqrt{\mathbb{E}_0(z_1^2)} \to 0$. Therefore $\mathbb{E}_0(z_1|z_1 \leq 0) = \mathbb{E}_0(z_1\mathbf{1}\{z_1 \leq 0\})/\mathbb{P}_0(z_1 \leq 0) \to 0$.

To conclude the proof we need to show that $\alpha_\Gamma \to \alpha$ as $\Gamma \to 0$. We can check this using the moment generating function of $\bar{z}_N$. Begin by noting that

$$\begin{aligned}1 &= \mathbb{E}_0\left(\prod_{i=1}^N \frac{f_1(y_i)}{f_0(y_i)}\right) = \mathbb{E}_0(e^{\bar{z}_N})\\ &= (1-\alpha_\Gamma)\mathbb{E}_0\left(e^{\bar{z}_N}\middle|\,\bar{z}_N \leq l\right) + \alpha_\Gamma\mathbb{E}_0\left(e^{\bar{z}_N}\middle|\,\bar{z}_N \geq u\right)\ .\end{aligned}$$

Hence

$$\alpha_\Gamma = \frac{1 - \mathbb{E}_0\left(e^{\bar{z}_N}\middle|\,\bar{z}_N \leq l\right)}{\mathbb{E}_0\left(e^{\bar{z}_N}\middle|\,\bar{z}_N \geq u\right) - \mathbb{E}_0\left(e^{\bar{z}_N}\middle|\,\bar{z}_N \leq l\right)}\ . \tag{2.24}$$

We can now use the statements (2.20) and (2.21) of Lemma 2.3.

It can be easily shown that $\mathbb{E}_0(e^{z_1}|z_1 \leq 0) \to 1$ and $\mathbb{E}_0(e^{z_1}|z_1 \geq 0) \to 1$ as $\Gamma \to 0$. Therefore from Lemma 2.3 we get that

$$\mathbb{E}_0\left(e^{\bar{z}_N}\middle|\,\bar{z}_N \leq l\right) \to e^l\ ,$$

and

$$\mathbb{E}_0\left(e^{\bar{z}_N}\middle|\,\bar{z}_N \geq u\right) \to e^u$$

as $\Gamma \to 0$. This, together with (2.24) concludes the proof of the first statement of the proposition. The proof of the second statement is entirely analogous.

2.C Fixed precision analogue of Proposition 2.1

Proposition 2.14. *Suppose the stopping boundaries of the SLRT are* $l = \log\beta$ *and* $u = \log\frac{1}{\alpha}$ *and the precision of each measurement* Γ *is fixed. Then*

$$\mathbb{E}_0(N_\Gamma) \leq \frac{2}{\Gamma\mu^2}\left(\alpha\log\alpha + (1-\alpha)\log\frac{1}{\beta} + (1-\alpha)\sqrt{\frac{\Gamma}{8\pi}}\left(\sqrt{\frac{\Gamma}{2}}\mu + e^{-\Gamma\mu^2/8}\right)\right),$$

and

$$\mathbb{E}_1(N_\Gamma) \leq \frac{2}{\Gamma\mu^2}\left((1-\beta)\log\frac{1}{\alpha} + \beta\log\beta + \sqrt{\frac{\Gamma}{8\pi}}\left(\sqrt{\frac{\Gamma}{2}}\mu + e^{-\Gamma\mu^2/8}\right)\right).$$

Proof. Note that in the SLRTs setting, the stopping boundaries are chosen as $l = \log\beta$ and $u = \log\frac{1}{\alpha}$ one has type I and II error probabilities at most α and β respectively (see [141]). Now returning to the proof of Proposition 1 if we controlled the quantities $\mathbb{E}_0(z_1|z_1 \leq l)$ and $\mathbb{E}_1(z_1|z_1 \geq u)$ in inequality (2.23) we could arrive at a fixed precision result. This can be easily done, for instance under the null $z_1 \sim N(-\Gamma\mu^2/2, \Gamma\mu^2)$ so

$$\begin{aligned}
\mathbb{E}_0(z_1|z_1 \leq 0) &= \frac{\mathbb{E}_0(z_1\mathbf{1}\{z_1 \leq 0\})}{\mathbb{P}_0(z_1 \leq 0)} \\
&\geq \frac{1}{2}\left(-\frac{\Gamma\mu^2}{2} + \int_{-\infty}^{\Gamma\mu^2/2}\frac{x}{\sqrt{2\pi\Gamma}\mu}e^{-\frac{x^2}{2\Gamma\mu^2}}\,\mathrm{d}x\right) \\
&= \frac{1}{2}\left(-\frac{\Gamma\mu^2}{2} - \sqrt{\frac{\Gamma}{2\pi}}\mu\int_{-\infty}^{\Gamma\mu^2/2}\frac{-x}{\Gamma\mu^2}e^{-\frac{x^2}{2\Gamma\mu^2}}\,\mathrm{d}x\right) \\
&= -\sqrt{\frac{\Gamma}{8\pi}}\left(\sqrt{\frac{\Gamma}{2}}\mu + e^{-\Gamma\mu^2/8}\right).
\end{aligned}$$

□

2.D Proof of Proposition 2.9

The maximum of the quantity above is attained when $\|b\|_1 = m$, so we will assume this in what follows.

For a fixed $i \in [n]$ let $\mathcal{C}_i = \{S \in \mathcal{C} : i \in S\}$. In case of symmetric classes we have that $|\mathcal{C}_i| = c$ does not depend on i. Also note that $c/|\mathcal{C}| = s/n$. To see the latter

consider a random coordinate J which is uniform on $[n]$, and a random coordinate K, which is selected sequentially: first select $S \in \mathcal{C}$ uniformly at random, then select $K \in S$ uniformly at random. When $\mathcal{C}$ is symmetric the distribution of J and K are the same, and

$$\frac{1}{n} = \mathbb{P}(J = i) = \mathbb{P}(K = i) = \frac{c}{|\mathcal{C}|}\frac{1}{s} \ .$$

With this we can write

$$\begin{aligned}
\min_{S\in\mathcal{C}} \sum_{S'\in\mathcal{C}\setminus\{S\}} \sum_{i\in S\triangle S'} b_i &= \min_{S\in\mathcal{C}} \sum_{S'\in\mathcal{C}\setminus\{S\}} \left(\sum_{i\in S\setminus S'} b_i + \sum_{i\in S'\setminus S} b_i \right) \\
&= \min_{S\in\mathcal{C}} \left(\sum_{i\in S} \sum_{S'\in\mathcal{C}\setminus\{S\}} \mathbf{1}\{i \notin S'\}\ b_i \right. \\
&\qquad \left. + \sum_{i\notin S} \sum_{S'\in\mathcal{C}\setminus\{S\}} \mathbf{1}\{i \in S'\}\ b_i \right) \\
&= \min_{S\in\mathcal{C}} \left(\sum_{i\in S} (|\mathcal{C}| - c)\ b_i + \sum_{i\notin S} c\ b_i \right) \\
&= \min_{S\in\mathcal{C}} \left((|\mathcal{C}| - c)\ b_S + c\ (m - b_S) \right) \\
&= cm + (|\mathcal{C}| - 2c)\ \min_{S\in\mathcal{C}}\ b_S \ ,
\end{aligned}$$

where $b_S = \sum_{i\in S} b_i$. However,

$$\begin{aligned}
\min_{S\in\mathcal{C}}\ b_S &\le \frac{1}{|\mathcal{C}|} \sum_{S\in\mathcal{C}}\ b_S = \frac{1}{|\mathcal{C}|} \sum_{S\in\mathcal{C}} \sum_{i\in S}\ b_i \\
&= \frac{1}{|\mathcal{C}|} \sum_{i=1}^{n} \mathbf{1}\{i \in S\} b_i = \frac{1}{|\mathcal{C}|} c \sum_{i=1}^{n} b_i \\
&= \frac{cm}{|\mathcal{C}|} \ .
\end{aligned}$$

Now note that when $b_i = m/n$ for all $i = 1, \ldots, n$, we have $b_S = sm/n = cm/|\mathcal{C}|$ for all $S \in \mathcal{C}$.

2.E Proof of Lemma 2.2

Consider the following sequence of graphs denoted by $G_0, G_1, G_2, \ldots, G_K$, where $K \in \mathbb{N}$. Let G_0 denote the graph with p vertices and no edges. The graphs $G_1, \ldots, G_K$ are obtained recursively by adding a disjoint star of s to the graph until this is no longer possible. In other words, for $k \in [K]$ the graph G_k is constructed by adding a disjoint s-star of G_{k-1}. Let $d_k(v)$ denote the degree of $v \in G_k$. Notice that for any $k \in \{0, \ldots, K\}$ if there exists $v \in G_k$ such that $d_k(v) < p - 1 - s$ we can add a star to G_k centered in v. This means that for the index K we have that $d_K(v) \geq p - 1 - s$ for all $v \in G_K$. Thus the graph G_K has at least $p(p-1-s)/2$ edges and is built entirely of disjoint stars of size s. The statement now follows.

Chapter 3

Adaptive Compressive Sensing for Structured Support Recovery

This chapter is based on joint work with Rui Castro. The results presented here can also be found in Castro & T. [43].

3.1 Introduction

Compressive sensing provides an efficient way to estimate signals that have a sparse representation in some basis or frame, see Candès & Tao [35, 34], Donoho [65], Candès & Wakin [36], Wainwright [140]. If the measurements can be chosen in a sequential and adaptive fashion, it is possible to achieve further performance gains in the sense that weaker signals can be estimated more accurately than in the non-adaptive setting, see Castro [40], Malloy & Nowak [107]. Furthermore, in some situations the signal may have additional structure that might be exploited.

In this chapter we investigate the problem of recovering the support of structured sparse signals using adaptive compressive measurements. Recall that in compressive sensing, we observe linear combinations of the signal with vectors of our choice, that are then perturbed by measurement noise. This makes the information the sample provides about the support different in nature then in the case of

coordinate-wise sampling investigated in Chapter 2.

Our focus is on the performance gains one can achieve when adaptively designing the sensing matrix compared to the situation where the sensing matrix is constructed non-adaptively. Furthermore, our aim is to highlight the way in which adaptive compressed sensing can capitalize on structural information. An appealing feature of compressed sensing is that accurate estimation can be done using only a few measurements. With this in mind we design algorithms for this problem that are sample-efficient, in the sense that they collect a number of observations that is not larger than the sample complexity of the best non-adaptive strategies.

The classes of structured support sets under consideration in this chapter are the same ones as in Chapter 2, namely

- **s-sets:** any subset of $[n]$ with size s;
- **s-intervals:** sets consisting of s consecutive elements of $[n]$;
- **unions of s-intervals:** unions of k disjoint s-intervals;
- **s-stars:** any star of size s in a complete graph (where the edges of the graph are identified with $[n]$);
- **unions of s-stars:** unions of k disjoint s-stars;
- **s-submatrices:** any submatrix of a given size $s_r \times s_c$ of an $n_r \times n_c$ matrix.

We analyze the fundamental limits of recovering support sets for the above classes under non-adaptive and adaptive sensing paradigms. We also provide adaptive sensing protocols with near-optimal performance to show the tightness of the fundamental limits mentioned before, and to illustrate how adaptive compressed sensing can capitalize on the structure of the support sets in the estimation. Finally, we provide procedures that, next to being near optimal in a statistical sense, also perform estimation using only a small number of measurements and are thus appealing from a practical point of view.

Note that, while adaptive compressive measurements might be very advantageous from a statistical and computational point of view, they also require a flexible infrastructure and hardware. In some settings, like that of the *single-pixel* camera (Duarte et al. [68]), all the necessary infrastructure is already in place. In tomography and magnetic resonance imaging the use of adaptive compressive samples is also possible, as described by Deutsch, Averbush & Dekel in [62] and Panych &

Jolesz in [118]. It is important to note that in the latter settings one has additional physical constraints that need to be accounted for. Other motivating examples include applications in sensor networks and monitoring, for instance identifying viruses in human or computer networks, or gene-expression studies, for instance when we have a group of genes co-expressed under the influence of a drug, or we have patients exhibiting similar symptoms (see Yoon et al. [147] and Moore et al. [112]). The results in this paper are foundational in nature, and aim at understanding the draws and limitations of adaptive compressive sensing in the context of structured support recovery. Furthermore, our model mostly fits the case where "compression" happens in the physical domain and before sensing takes place (e.g., the settings in Duarte et al. [68], Deutsch, Averbush & Dekel [62] and Panych & Jolesz [118]). It is important to note that if the sensing is further constrained (so that the measurement vectors cannot be arbitrary) then the performance of any algorithm will be affected. For a discussion on how such constraints can affect the performance of adaptive compressive sensing see e.g. Davenport et al. [59].

Table 3.1: Summary of scaling laws for the signal magnitude.

	Non-Adaptive Sensing (necessary)	Adaptive Sensing (necessary)	Adaptive Sensing (sufficient)
s-sets	$\mu \sim \sqrt{\frac{n}{m}\log n}$	$\mu \sim \sqrt{\frac{n}{m}\log s}$	$\mu \sim \sqrt{\frac{n}{m}\log s}$
unions of k disjoint s-intervals	$\mu \sim \frac{1}{s}\sqrt{\frac{n}{m}\log\frac{n}{ks}}$	$\mu \sim \frac{1}{s}\sqrt{\frac{n}{m}\log ks}$	$\mu \sim \frac{1}{s}\sqrt{\frac{n}{m}\log ks}$
unions of k disjoint s-stars	$\mu \sim \sqrt{\frac{n}{m}\log\frac{\sqrt{n}}{ks}}$	$\mu \sim \frac{1}{s}\sqrt{\frac{n}{m}\log ks}$	$\mu \sim \frac{1}{s}\sqrt{\frac{n}{m}\log ks}$
$\sqrt{s}\times\sqrt{s}$ submatrices of an $\sqrt{n}\times\sqrt{n}$ matrix	$\mu \sim \sqrt{\frac{n}{\sqrt{s}m}\log\frac{n}{s}}$	$\mu \sim \frac{1}{s}\sqrt{\frac{n}{m}\log s}$	$\mu \sim \frac{1}{s^{3/4}}\sqrt{\frac{n}{sm}\log s}$

Scaling laws for the signal magnitude μ (constants omitted) that are necessary/sufficient for $\max_{S\in\mathcal{C}}\mathbb{E}(\widehat{S}\triangle S)\to 0$ as $n\to\infty$, where $\mathcal{C}$ denotes the corresponding class of support sets, and m denotes the total amount of sensing energy available for our measurements in expectation (see (3.3)). The results in the last column make some sparsity assumptions, meaning $s \ll n$. For exact conditions see relevant propositions of Section 3.3.1.

Table 3.1 summarizes some of our results, showing necessary and sufficient conditions for the signal magnitude for accurate support estimation. It also highlights two different facets of the gains of adaptive sensing over non-adaptive sensing, similar to what we have seen in Chapter 2. First, note that the necessary conditions of non-adaptive sensing include a $\sqrt{\log n}$ factor for each of the classes under

consideration. This factor is replaced by the logarithm of the sparsity when considering adaptive sensing, and this is due to the fact that adaptive strategies are better able to mitigate the effects of noise. Second, for certain classes adaptive sensing can gain greater leverage from the structure of the support sets compared to non-adaptive sensing. This phenomenon is best visible considering the class of s-stars, where estimators using non-adaptive sensing gain practically nothing from the structural information whereas adaptive sensing benefits greatly from it. Note that the necessary and sufficient conditions for the class of submatrices using adaptive sensing do not match, and a full characterization of the problem in that case remains open. We also remark at this point that the results derived in this chapter are non-asymptotic in nature and also account for the constant factors in the scaling laws. The asymptotic presentation in Table 3.1 merely makes it easier to highlight the main contributions.

It is instructive to note a fundamental difference between non-adaptive sensing and adaptive sensing problems. In non-adaptive sensing support recovery methods can often be computationally demanding or even intractable, a prominent example being submatrix estimation [31, 18, 24]. Contrasting this, adaptive sensing algorithms can solve this problem using polynomial-time algorithms. Though this might seem surprising, one has to bear in mind that there is a fundamental difference between the two setups. In fact when using adaptive sensing one already shakes most of the computational burdens by tailoring the sample to facilitate inference. The bottleneck of such algorithms lies in sample collection, but given a good strategy the sample will contain much less confounders making the inference itself easier computationally. This, next to increased statistical power, can be another appealing reason for using adaptive sensing methods whenever possible.

This chapter can be seen as an extension of the previous one in that we consider the compressive sensing model, which is more general the than coordinate-wise sampling model. Although some of the techniques and insights can be used from there, changing the measurement model introduces a number of new challenges to tackle. In particular the information provided by compressive measurements is very different in nature from that provided by coordinate-wise observations. This means that structural information is captured in the observations in a different way, which influences both the theoretical limits and the way support recovery procedures need to be designed.

Related work: The results in this chapter are built on a number of recent contributions on detection and estimation of sparse signals using compressive sensing. Considering general sparse signals without structure Arias-Castro, Candès & Davenport [7] and Castro [40] provide theoretical performance limits of adaptive compressive sensing, characterizing the gains one can realize when adaptively designing the sensing matrix. Complementing these results, Haupt et al. [82] and Malloy & Nowak [108, 107] provide efficient near optimal procedures for estimation. Considering the problem of detection Arias-Castro [5] provides both theoretical limits and optimal procedures both in the non-adaptive and adaptive compressed sensing settings.

The problem of estimating structured sparse signals was examined in the past in a multitude of different settings. For an extensive literature review in the setting of coordinate-wise observations we also direct the reader to the introduction of Chapter 2.

Such problems received attention in the compressive sensing setting as well. In [20] Baraniuk et al. consider recovering tree-structured signals in the non-adaptive framework and show that structural information enhances the performance of compressive sensing methods. Recovering tree-structured signals is also investigated by Soni & Haupt [134, 135] but in these works the problem is examined in the adaptive sensing setting. The authors consider signals in which the activation pattern is a rooted subtree of a given tree, and show that one can realize further gains recovering these types of supports by adaptively designing the sensing matrix. Our work is closely related, but the structured class investigated by Soni & Haupt in [134, 135] is clearly different from the ones listed in Table 3.1.

In [97] Krishnamuthy, Sharpnack & Singh consider activation patterns that have low cut-size in an arbitrary (fixed) signal graph and also find that adaptivity enhances the statistical performance of compressive sensing. Though these types of classes seem closer to the ones investigated in this chapter, note that the structures in Table 3.1 do not have lower cut-size than an arbitrary s-sparse set, meaning that these can not be efficiently encoded with the definitions of Krishnamuthy, Sharpnack & Singh [97]. As an example, arbitrary submatrices in a 2d-lattice have typical cut-size on the order of s, the same as any s-sparse subset of the 2d-lattice. Similar comments apply to the other classes considered here as well.

Moving away from graphs, Balakrishnan et al. investigate the problem of finding block-structured activations in a signal matrix considering both non-adaptive

and adaptive measurements in [17]. They report similar findings to the previous authors, namely that both adaptivity and structural information provide gains in support recovery when dealing with block-activations in a matrix. This work extends these results by investigating more general structured activations.

As mentioned above, an appeal of compressive sensing is that accurate estimation can be done using only a small number of measurements. The number of measurements an estimation procedure uses is called the sample complexity of that procedure. The sample complexity properties of compressive sensing were studied by Aksoylar, Atia & Saligrama in [2] and [3] for the support recovery of general sparse signals in the non-adaptive and adaptive sensing settings respectively.

Organization: The chapter is structured as follows. Section 3.2 describes the problem setting in detail. In Section 3.3 we provide adaptive sensing procedures for structured support recovery and analyze the theoretical limits of the problem, both under non-adaptive and adaptive sensing paradigms. In this section we only make a restriction to the sensing power available, but not on the number of projective measurements that we are allowed to make. In Section 3.4 we further restrict the number of measurements. Finally we provide some concluding remarks in Section 3.6.

3.2 Problem Setting

In this chapter we consider the following statistical model. The signal model is the same as in the previous chapter, but we recall it here for convenience. Let $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ be a vector of the form

$$x_i = \begin{cases} \mu & \text{if } i \in S, \\ 0 & \text{if } i \notin S, \end{cases} \tag{3.1}$$

where $\mu > 0$ and S is an unknown element of a class of sets denoted by $\mathcal{C}$. We refer to $\boldsymbol{x}$ as the *signal* and to S as the *support* or *significant/active components* of the signal. The set S is our main object of interest. The signal model (3.1) may seem overly restrictive at first because of the fact that each non-zero entry has the same value μ. However, our lower bounds and the procedures of Section 3.4.1 can be generalized to signals with active components of arbitrary magnitudes and signs,

in which case the value μ would play the role of the minimal absolute value of the non-zero components. For sake of simplicity we do not discuss this extension here, but refer the reader to Arias-Castro [5], Malloy & Nowak [107] for details on how this can be done.

We are allowed to collect multiple measurements of the form

$$Y_t = \langle A_t, \boldsymbol{x} \rangle + W_t, \ t = 1, 2, \dots \ . \tag{3.2}$$

Thus each measurement is the inner product of the signal $\boldsymbol{x}$ with the vector $A_t \in \mathbb{R}^n$, contaminated by Gaussian noise. The noise terms $W_t \sim N(0,1)$ are independent and identically distributed (i.i.d.) standard normal random variables, also independent of $\{A_j\}_{j=1}^t$. Under the adaptive sensing paradigm A_t are allowed to be functions of the past observations $\{Y_j, A_j\}_{j=1}^{t-1}$. This model is only interesting if one poses some constraint on the total amount of sensing energy available. Let A denote the matrix whose row t is A_t. We require

$$\sup_{S \in \mathcal{C}} \mathbb{E}_S \left(\|A\|_F^2 \right) \ = \ \mathbb{E}_S \left(\sum_t \|A_t\|^2 \right) \ \leq m \ , \tag{3.3}$$

where $\| \cdot \|_F$ denotes the Frobenius norm, m is our total energy budget, and $\mathbb{E}_S$ denotes the expectation with respect to the joint distribution of $\{A_t, Y_t\}_{t \geq 1}$ when $S \in \mathcal{C}$ is the support set.

Remark 3.1. *As in the previous chapter, we could consider algorithms satisfying an exact energy constraint as opposed to the expected energy constraint in* (3.3)*. Similar comments apply as in Chapter 2, namely that efficient estimation can be done under the more strict constraint as well, and the results would essentially remain unchanged.*

In detail, the same comments apply for lower bounds and the procedures of Section 3.3.1 as before, thus the reader is referred to Appendix 2.A. Considering the procedures of Section 3.4.1, note that these already satisfy an energy constraint without expectations.

3.2.1 Inference Goals

This chapter aims at characterizing the difficulty of recovering structured sparse signal supports with adaptive compressive sensing. We are interested in settings where the class $\mathcal{C}$ contains sets with some sort of structure, for instance the active

components of $\boldsymbol{x}$ are consecutive. For the unstructured case, that is, when $\mathcal{C}$ contains every set of a given cardinality, there already exists a lower bound by Castro in [40], and a procedure that achieves this lower bound by Malloy & Nowak in [107]. The main goal of this chapter is to provide similar results for structured sparse sets.

We are interested in two aspects of adaptive compressive sensing: statistical power and sample complexity. First, given n, m, ε and $\mathcal{C}$ the aim is to characterize the minimal signal strength μ for which S can be reliably estimated, which means there is an algorithm and sensing strategy such that for a given $\varepsilon > 0$,

$$\max_{S \in \mathcal{C}} \mathbb{E}_S(|\widehat{S} \triangle S|) \leq \varepsilon \ , \tag{3.4}$$

where $\widehat{S} \triangle S$ is the symmetric set difference. Furthermore, we aim to construct such an adaptive sensing strategy. Although the above setting makes sense whenever $\varepsilon \in [0, |S|]$, the problem is only interesting when ε is small. Hence we will take ε as an element of $[0, 1]$.

Remark 3.2. *We remark at this point that our main interest lies in the scaling of μ in terms of the model parameters, but we do not aim to find accurate constants. With this in mind, the procedures throughout the paper could be improved with a more careful and refined analysis. However, these would only improve constant factors, and so we chose to keep technicalities to a minimum providing a smoother presentation at the price of suboptimal constants.*

Second, given n, m, μ, ε and $\mathcal{C}$ we wish to characterize the minimal number of samples needed to ensure (3.4). Considering the unstructured case, it was shown by Aksoylar, Atia & Saligrama in [2] that non-adaptive procedures need at least $O(s \log \frac{n}{s})$ measurements, and in [36] Candès & Wakin show that this bound is achievable (these results apply when the signal strength μ is close to the threshold of estimability). On the other hand, to the best knowledge of the authors, an exact characterization of the sample complexity for adaptive procedures is not yet available, though there exists work on the topic by Aksoylar & Saligrama [3]. In that work the authors present a result that states that the sample complexity of the problem essentially scales as s. However, it is not clear if that bound is tight. In Section 3.6 we provide more insight on this question.

Remark 3.3. *As mentioned in the introduction, this chapter can be seen as an extension of Chapter 2 from component-wise sampling to the more general com-*

pressive sensing, and it is instructive to briefly discuss the differences between the two setups. Component-wise observations can be viewed as restricting compressive sensing by requiring each measurement vector A_t to have exactly one non-zero entry (though the problem is set up a bit differently in Chapter 2, the two are effectively the same). In the previous chapter we have seen that the necessary conditions for support recovery for the classes considered are as follows: the condition is the same for s-sets, while for the other classes the $\frac{1}{s}$ term moves inside the square-root if one only allows component-wise observations. Also, these conditions are sufficient in the case of component-wise samples.

Lying at the heart of the difference between the rates for support recovery between the two setups is the increased detection power of compressive sensing over coordinate-wise sampling. In a nutshell, detection of a signal is the problem of differentiating between two hypotheses: the null being that all signal components are zero and the alternative being that there are s non-zero components somewhere in the signal vector. In [5] Arias-Castro shows that the necessary and sufficient conditions for detection for compressive sensing is $\frac{1}{s}\sqrt{\frac{n}{m}}$, whereas in [40] Castro shows the same for component-wise sampling to be $\sqrt{\frac{1}{s}\frac{n}{m}}$. When moving from component-wise sampling to compressive sensing, for certain structured classes it is possible to make use of this increased detection power, which in turn lowers the requirement for the signal magnitude. This also means algorithms need to be designed with a different mindset when using compressive sensing instead of coordinate-wise sampling.

In what follows we use the symbol $\mathbf{1}$ to denote both the usual indicator function (e.g., $\mathbf{1}\{i \in S\}$ takes the value 1 if $i \in S$ and zero otherwise), and to denote binary vectors with support S. For instance $\mathbf{1}_S$ denotes an element of $\{0,1\}^n$ for which the entries in S have value 1 and all the other entries have value 0. Note that to ease distinction of the two the arguments of the functions are in a different place (after the symbol in the first case and in the subscript of the symbol in the second case). Furthermore, let $\mathbb{P}_S$ denote the joint distribution of $\{A_j, Y_j\}_{1,2,\ldots}$ when $S \in \mathcal{C}$ is the support set, and let $\mathbb{E}_S$ denote the expectation with respect to $\mathbb{P}_S$.

3.3 Signal strength

We now examine the minimal signal strength required to recover structured support sets. In this setup we are allowed to make an infinite number of measurements of the

form (3.2) (provided that the budget (3.3) is satisfied). Although this might not be reasonable from a practical standpoint, it is a good place to start for understanding the fundamental performance limits of adaptive compressing sensing, and we will see in Section 3.4 that the same performance can be attained with a small number of measurements.

3.3.1 Procedures

It is instructive to briefly consider a simple support recovery algorithm for the unstructured case. When the support set can be any set of a given cardinality, and there is no restriction on the number of samples we are allowed to take, we can use essentially the same procedure as in Chapter 2. That is, we perform a Sequential Likelihood Ratio Test for each coordinate separately. More precisely, for every coordinate $i \in [n]$ collect observations of the form

$$Y_{i,j} = ax_i + W_j = \langle a\mathbf{1}_{\{i\}}, \boldsymbol{x}\rangle + W_j,\ j = 1, \dots, N_i\ ,$$

with some fixed $a > 0$, where we recall that $\mathbf{1}_{\{i\}}$ is a singleton vector. The number of observations N_i is random and is given by

$$N_i = \min\left\{n \in \mathbb{N} :\ \sum_{j=1}^{n} \log \frac{\mathrm{d}\mathbb{P}_1(Y_{i,j})}{\mathrm{d}\mathbb{P}_0(Y_{i,j})} \notin (l, u)\right\}\ ,$$

where $\mathbb{P}_0$ ($\mathbb{P}_1$) is the distribution of the observations when component i is non-active (active), and $l < 0 < u$ are the lower and upper stopping boundaries of the SLRT. Our estimator $\widehat{S}$ will be the collection of components i for which the log-likelihood process above hits the upper stopping boundary u. Considering the test of component $\boldsymbol{x}_i$ we have the following result, which is the analogue of Proposition 2.1:

Lemma 3.1. *Set $l = \log \frac{\beta}{1-\alpha}$ and $u = \log \frac{1-\beta}{\alpha}$ with $\alpha, \beta \in (0, 1/2)$, and let the type I and type II error probabilities of the SLRT described above be α_a and β_a. Then $\alpha_a \to \alpha$ and $\beta_a \to \beta$ as $a \to 0$. Furthermore*

$$a^2\mathbb{E}_0(N_i) \le \frac{2}{\mu^2}\left(\alpha \log \frac{\alpha}{1-\beta} + (1-\alpha)\log\frac{1-\alpha}{\beta}\right) \le \frac{2}{\mu^2}\log\frac{1}{\beta}$$

and

$$a^2\mathbb{E}_1(N_i) \leq \frac{2}{\mu^2}\left(\beta\log\frac{\beta}{1-\alpha} + (1-\beta)\log\frac{1-\beta}{\alpha}\right) \leq \frac{2}{\mu^2}\log\frac{1}{\alpha}$$

as $a \to 0$.

Proof. The proof is identical to that of Proposition 2.1. □

Using the previous result we can immediately analyze the above procedure. Set $\alpha = \varepsilon/2n$ and $\beta = \varepsilon/2s$ in the lemma above, and choose a to be arbitrarily small. Hence α_a and β_a will be close to the nominal error probabilities α and β and we ensure (3.4). Then using the other part of Lemma 3.1 we can upper bound the expected energy used by the tests. Summing this over all the tests and using (3.3) we arrive at the following:

Proposition 3.1. *Testing each component* $\boldsymbol{x}_i$, $i = 1, \ldots, n$ *as described above yields an estimator satisfying* (3.3) *and* (3.4) *whenever*

$$\mu \geq \sqrt{\frac{2n}{m}\log\frac{2s}{\varepsilon} + \frac{2s}{m}\log\frac{2n}{\varepsilon}}\ .$$

When the support is sparse (that is $s = o(n)$), the first term dominates the bound above. This upper bound asymptotically coincides with the lower bound of Castro [40] showing that the simple procedure above is near optimal.

Remark 3.4. *Note that the lower bound presented by Castro in [40] is valid for a slightly broader class than the class of* s*-sets, namely one also has to include* $(s-1)-$*sets into the class. However, the procedure outlined above works without any modifications for this broadened class as well, and so the result of Proposition 3.1 holds for this larger class. A similar comment applies to all the procedures presented later on: the procedures are presented for classes of a given sparsity for sake of clarity, but the analysis shows that they also work for classes containing sets of slightly different sparsity. This is important to note as, because of technical reasons, some of the lower bounds of Section 3.3.2 can only deal with such enlarged classes.*

The procedures for recovering structured support sets will be very similar in nature, but slightly modified to take advantage of the structural information. In particular we know that it is possible to detect the presence of weak signals using compressive sensing (see Arias-Castro [5]). In order to take advantage of this property our procedures consist of two phases: a *search phase* and a *refinement phase*. The aim of the search phase is to find the approximate location of the signal

using a detection-type method, that is identifying a subset of components $\boldsymbol{P} \subset [n]$ such that $|\boldsymbol{P}| \ll n$ and $S \subset \boldsymbol{P}$ with high probability. Once this is done we can focus our attention exclusively on $\boldsymbol{P}$ in the refinement phase and estimate the support in the same manner as in the unstructured case.

Unions of s-intervals

The first structured class that we consider is the unions of k disjoint s-intervals. A similar setting was considered in by Balakrishnan et al. in [17], where they considered recovering block-structured activations in a signal matrix. Hence the case $k = 1$ is a special case of their setting when the signal matrix has one row. The unions of intervals class is a good starting point to highlight the main ideas of how recovery algorithms can benefit from structural information in the adaptive compressed sensing setting, particularly because it can be viewed as a bridge between the unstructured case (with $k = s$ and intervals of length one) to the most structured class ($k = 1$).

Consider the class of sets that are unions of k disjoint intervals of length s. Formally, let $S_i = [i, i + s - 1]$ for $i \in [n - s + 1]$ and

$$\mathcal{C} = \left\{ S \subset [n] : \; S = \bigcup_{i=1}^{k} S_i, \; S_i \cap S_j = \emptyset \; \forall i \neq j \right\} .$$

In principle we could also consider overlapping intervals. Although this can still be handled in a similar fashion as done below it would result in a more cluttered presentation.

Our procedure for estimating S is as follows. Split the index set $[n]$ into consecutive bins of length $s/2$ denoted by $\boldsymbol{P}^{(1)}, \ldots, \boldsymbol{P}^{(2n/s)}$. We suppose that $2n$ is divisible by s, as it makes the presentation less cluttered. The procedure can be easily modified if this is not satisfied. Of these bins at least k (and at most $2k$) are contained entirely in S. In the search phase we aim to find the approximate location of the support by finding these bins. To do this we test the following hypotheses

$$H_0^{(i)} : \; \boldsymbol{P}^{(i)} \cap S = \emptyset \qquad \text{versus} \qquad H_1^{(i)} : \; \boldsymbol{P}^{(i)} \subset S \qquad i \in [2n/s] .$$

We use a SLRT to decide between $H_0^{(i)}$ and $H_1^{(i)}$ for each $i \in [2n/s]$, all with the same type I and type II error probabilities α and β. The choices of α and β and

the exact way of carrying out the tests will be described later. As an output of the search phase, we define the set $\boldsymbol{P}$ based on the tests above. Since some $\boldsymbol{P}^{(i)}$ may only partially intersect the support S we set $\boldsymbol{P}$ to be the union of those bins $\boldsymbol{P}^{(i)}$ for which either $H_1^{(i-1)}, H_1^{(i)}$ or $H_1^{(i+1)}$ was accepted. This way we ensure $\mathbb{P}_S(S \nsubseteq \boldsymbol{P}) \leq 2k\beta$. We also wish to ensure that $\boldsymbol{P}$ is small, and to do so we must to choose α appropriately. Once this is done we can move on to the search phase and find the support within $\boldsymbol{P}$. We can do this in a very crude way and use a similar procedure as in the unstructured case with type I and II error probabilities α', β'. The sensing energy used in this phase will be negligible due to $\boldsymbol{P}$ being small. Finally the estimator $\widehat{S}$ will be the collection of components that were deemed active at the end of the refinement phase.

We now choose $\alpha, \beta, \alpha', \beta'$ to ensure the estimator satisfies (3.4). We have

$$\begin{aligned}\mathbb{E}_S\left(|\widehat{S}\triangle S|\right) &\leq \mathbb{E}_S\left(|\widehat{S}\triangle S| \;\middle|\; S \nsubseteq \boldsymbol{P}\right)\mathbb{P}_S(S \nsubseteq \boldsymbol{P}) + \mathbb{E}_S\left(|\widehat{S}\triangle S| \;\middle|\; S \subseteq \boldsymbol{P}\right) \\ &\leq \mathbb{E}_S\left(|S \setminus \boldsymbol{P}| + \sum_{i\in\boldsymbol{P}:\; i\notin S} \alpha' + \sum_{i\in\boldsymbol{P}:\; i\in S} \beta' \;\middle|\; S \nsubseteq \boldsymbol{P}\right) 2k\beta \\ &\quad + n\alpha' + ks\beta' \; .\end{aligned}$$

Hence choosing $\alpha' = \varepsilon/4n, \beta' = \varepsilon/4ks$ and $\beta = \varepsilon/8k^2s^2$ ensures that (3.4) holds. Note that α does not influence the probability of error directly. However, it does influence the size of $\boldsymbol{P}$, and hence the total sensing energy required by the procedure.

To perform the ith test of the search phase we collect measurements using projection vectors of the form $a\mathbf{1}_{\boldsymbol{P}^{(i)}}$ with an sufficiently small a and perform a SLRT with stopping boundaries $l < 0 < u$. Let $\mathbb{E}_0$ and $\mathbb{E}_1$ denote the expectation when $H_0^{(i)}$ or $H_1^{(i)}$ is true, respectively. Similarly to the unstructured case we now have the following:

Lemma 3.2. *Set $l = \log\frac{\beta}{1-\alpha}$ and $u = \log\frac{1-\beta}{\alpha}$ with $\alpha, \beta \in (0, 1/2)$, and let the type I and type II error probabilities of the SLRT described above be α_a and β_a. Then $\alpha_a \to \alpha$ and $\beta_a \to \beta$ as $a \to 0$. Furthermore*

$$a^2\mathbb{E}_0(N_i) \leq \frac{2}{(s/2)^2\mu^2}\left(\alpha\log\frac{\alpha}{1-\beta} + (1-\alpha)\log\frac{1-\alpha}{\beta}\right) \leq \frac{2}{(s/2)^2\mu^2}\log\frac{1}{\beta} \; ,$$

and

$$a^2\mathbb{E}_1(N_i) \leq \frac{2}{(s/2)^2\mu^2}\left(\beta\log\frac{\beta}{1-\alpha}+(1-\beta)\log\frac{1-\beta}{\alpha}\right) \leq \frac{2}{(s/2)^2\mu^2}\log\frac{1}{\alpha}\ ,$$

as $a \to 0$.

Using this we can upper bound the amount of sensing energy used for the test of $\boldsymbol{P}^{(i)}$ under $H_0^{(i)}$ and $H_1^{(i)}$. However, now it is possible that neither statement in $H_0^{(i)}$ nor $H_1^{(i)}$ holds for a given bin $\boldsymbol{P}^{(i)}$. Considering a test where neither of them is true we can still carry out the the same calculations as in Lemma 3.1 and thus upper bound the expected sensing energy used for the test.

Lemma 3.3. *Set* $l = \log\frac{\beta}{1-\alpha}$ *and* $u = \log\frac{1-\beta}{\alpha}$ *with* $\alpha, \beta \in (0, 1/2)$*, and let* $\widetilde{s}$ *denote the true number of signal components in* $\boldsymbol{P}^{(i)}$*. Suppose that in the setting above neither* $H_0^{(i)}$ *nor* $H_1^{(i)}$ *is true, that is* $0 < \widetilde{s} < s/2$*. Furthermore suppose* $\widetilde{s} \neq s/4$*. Then, as* $a \to 0$*,*

$$a^2\mathbb{E}_{\widetilde{s}}(N_i) \leq \frac{2}{s\mu^2}\log\max\left\{\frac{1-\alpha}{\beta}, \frac{1-\beta}{\alpha}\right\} \leq \frac{2}{s\mu^2}\log\frac{1}{\min\{\alpha,\beta\}}\ ,$$

where $\mathbb{E}_{\widetilde{s}}$ *denotes the expectation when the number of signal components in* $\boldsymbol{P}^{(i)}$ *is* $\widetilde{s}$*.*

Proof. In what follows we drop the subscript i to ease notation. The log-likelihood ratio for an observation Y_j is

$$z_j = \log\frac{\mathrm{d}\mathbb{P}_1(Y_j)}{\mathrm{d}\mathbb{P}_0(Y_j)} = \frac{as\mu Y_j}{2} - \frac{a^2s^2\mu^2}{8},\ j = 1, \ldots, N\ .$$

Suppose first that $s/4 < \widetilde{s} < s/2$. Note that now the drift of the log-likelihood ratio process is positive. Now $z_1 \sim N\left((\widetilde{s}-\frac{s}{4})\frac{a^2s\mu^2}{2}, \frac{a^2s^2\mu^2}{4}\right)$. From normality we still have $\mathbb{E}(z_1|z_1 \geq 0) \geq \mathbb{E}(z_1 - c|z_1 \geq c),\ \forall c > 0$. Combining this with Wald's identity we get

$$\mathbb{E}(N)\mathbb{E}(z_1) = \mathbb{E}(\bar{z}_N) \leq u + \mathbb{E}(z_1|z_1 \geq 0)\ ,$$

where $\bar{z}_N = \sum_{j=1}^{N} z_j$. Denoting $\xi \sim N(0,1)$,

$$\begin{aligned}\mathbb{E}(z_1 | z_1 \geq 0) &\leq 2\mathbb{E}(z_1 \mathbf{1}\{z_1 \geq 0\}) \\ &\leq \left(\widetilde{s} - \frac{s}{4}\right) \frac{a^2 s\mu^2}{2} + 2\mathbb{E}\left(\frac{as\mu}{2}\xi \mathbf{1}\{\xi \geq -\left(\widetilde{s} - \frac{s}{4}\right)\mu\}\right) \\ &\leq as\mu\left(\left(\widetilde{s} - \frac{s}{4}\right)\frac{a\mu}{2} + 1\right) .\end{aligned}$$

Plugging this in, and using that $\mathbb{E}(z_1) \geq \frac{a^2 s\mu^2}{2}$ we get

$$a^2\mathbb{E}(N) \leq \frac{2}{s\mu^2}u + \frac{2a}{\mu}\left(\left(\widetilde{s} - \frac{s}{4}\right)\frac{a\mu}{2} + 1\right) .$$

Hence in the limit $a \to 0$ we get

$$a^2\mathbb{E}(N) \leq \frac{2}{s\mu^2}\log\frac{1-\beta}{\alpha} \leq \frac{2}{s\mu^2}\log\frac{1}{\alpha} .$$

We can treat the case $0 < \widetilde{s} < s/4$ in a similar fashion. □

Remark 3.5. *When $\widetilde{s} = s/4$ the argument breaks down, because of ties when s is divisible by 4. However this is only a technical issue that can be simply circumvented by choosing the bins to be of size $s/2 - 1$, for instance.*

Now we are ready to upper bound the expected sensing energy used by the procedure. Given α and β we can deal with the search phase and by Lemma 3.1 we can deal with the refinement phase given α', β' and $|\boldsymbol{P}|$.

Recall, the support consists of k intervals of length s. Note that we have

$$\mathbb{E}_S(|\boldsymbol{P}|) \leq 3ks + \frac{3s}{2}\sum_{i:\ \boldsymbol{P}^{(i)} \not\subseteq S} \alpha .$$

Thus choosing $\alpha = \varepsilon/6n$ we have $\mathbb{E}_S(|\boldsymbol{P}|) \leq 3ks + \varepsilon/2 \leq 4ks$.

By denoting the part of the sensing matrix A corresponding to the search and refinement phases by A_{search} and $A_{\text{refinement}}$ respectively, we have

$$\begin{aligned}\mathbb{E}_S(\|A\|_F) &\leq \mathbb{E}_S(\|A_{\text{search}}\|_F^2) + \mathbb{E}_S\left(\mathbb{E}_S(\|A_{\text{refinement}}\|_F^2 \big| |\boldsymbol{P}|)\right) \\ &\leq \frac{16n}{s^2\mu^2}\log\frac{2\sqrt{2}ks}{\varepsilon} + \frac{4k}{s\mu^2}\log\frac{6n}{\varepsilon} + \frac{2k}{\mu^2}\log\frac{6n}{\varepsilon} + \frac{8ks}{\mu^2}\log\frac{4n}{\varepsilon} . \qquad (3.5)\end{aligned}$$

When $ks \ll n$ the first term dominates the bound above. Using this and

combining the above with (3.3) we arrive at the following:

Proposition 3.2. *Consider the class of k disjoint s-intervals and let $\frac{n}{\log 4n} \geq ks^3$. Then the above estimator satisfies* (3.3) *and* (3.4) *whenever*

$$\mu \geq \sqrt{\frac{30n}{s^2 m} \log \frac{2\sqrt{2}ks}{\varepsilon}} \ .$$

Remark 3.6. *The condition on the sparsity in Proposition 3.2 is needed to ensure that the term corresponding to the search phase in* (3.5) *becomes dominant. By performing the refinement phase in a more sophisticated way one can relax that condition. For instance using k binary searches to find the left endpoint of the intervals the sparsity condition becomes $\frac{n}{\log 6n} \geq ks^2 \log s$. We expect this to be essentially the best condition one can hope for, as the lower bounds of Section 3.3.2 show that the first term in* (3.5) *is unavoidable.*

The bound of Proposition 3.2 matches the lower bound in Section 3.3.2, hence in this sparsity regime the procedure above is optimal apart from constants.

Unions of s-stars

Let the components of $\boldsymbol{x}$ be in one-to-one correspondence to edges of a complete graph $G = (V, E)$. Recall that we call a collection of s edges sharing a common vertex an s-star (for a formal definition see page 18). Let $\mathcal{C}$ be the class of unions of k disjoint s-stars. In what follows we use the notation $|V| = p$.

The procedure for support estimation is very similar to the one presented for s-intervals. We introduce the procedure when $k = 1$, but the idea can be carried through for larger k. Consider the subsets $\boldsymbol{P}^{(i)},\ i = 1, \ldots, p$, defined as follows:

$$\boldsymbol{P}^{(i)} = \{j \in [n] :\ v_i \in e_j\} \ ,$$

that is $\boldsymbol{P}^{(i)}$ contains all the components whose corresponding edges contain vertex v_i. These subsets are not a partition of $[n]$ as they are not disjoint. Nonetheless we know that

$$|\boldsymbol{P}^{(i)} \cap S| \in \{0, 1, s\} \qquad \forall i \in [p] \ ,$$

and can use this to find the approximate location of S. In the search phase we test

the hypotheses

$$H_0^{(i)}: \ |\boldsymbol{P}^{(i)} \cap S| = 1 \qquad \text{versus} \qquad H_1^{(i)}: \ |\boldsymbol{P}^{(i)} \cap S| = s \qquad i \in [p] \ .$$

In words we test whether vertex v_i is the center of the star or not for all $i \in [p]$. Note that when vertex v_i is not the center of the star we have $|\boldsymbol{P}^{(i)} \cap S| \in \{0, 1\}$. By specifying $H_0^{(i)}$ as above and using an SLRT to decide between the two hypotheses, we ensure that when $|\boldsymbol{P}^{(i)} \cap S| = 0$ both the probability of error and the expected number of steps will be smaller than when $|\boldsymbol{P}^{(i)} \cap S| = 1$. This is due to the monotonicity of the likelihood ratio.

As noted before we use independent SLRTs for the tests above with common type I and type II error probabilities α, β. The exact details of these tests will be covered later. Using these tests we can define $\boldsymbol{P}$, the output of the search phase, as the union of those $\boldsymbol{P}^{(i)}$ for which $H_1^{(i)}$ is accepted. With the appropriate choices for α and β we can ensure that with high probability $S \subset \boldsymbol{P}$ and that $|\boldsymbol{P}|$ is small. In fact we would like to accept exactly one $H_1^{(i)}$. Again the right choice for β will ensure $\mathbb{P}_S(S \nsubseteq \boldsymbol{P})$ is small whereas the right choice of α ensures that $|\boldsymbol{P}|$ is small with high probability. In the subsequent refinement phase we estimate S within $\boldsymbol{P}$. We do this using the same procedure as in the unstructured case with error probabilities α', β'. Finally the estimator $\widehat{S}$ will be the collection of those components which were deemed active in the refinement phase.

Now we choose the error probabilities for the tests such that we can ensure (3.4) for our procedure. We have

$$\begin{aligned}
\mathbb{E}_S\left(|\widehat{S} \triangle S|\right) &\leq \mathbb{E}_S\left(|\widehat{S} \triangle S| \ \middle| S \nsubseteq \boldsymbol{P}\right) \mathbb{P}_S(S \nsubseteq \boldsymbol{P}) + \mathbb{E}_S\left(|\widehat{S} \triangle S| \ \middle| S \subseteq \boldsymbol{P}\right) \\
&\leq \mathbb{E}_S\left(|S \setminus \boldsymbol{P}| + \sum_{i \in \boldsymbol{P}:\ i \notin S} \alpha' + \sum_{i \in \boldsymbol{P}:\ i \in S} \beta' \ \middle| \ S \nsubseteq \boldsymbol{P}\right) \beta \\
&\quad + n\alpha' + s\beta' \ .
\end{aligned}$$

Thus the choices $\beta = \varepsilon/4s$ and $\alpha' = \varepsilon/4n, \beta' = \varepsilon/4s$ suffice. As noted before, the choice of α will influence the size of $\boldsymbol{P}$ and will be discussed later.

To test $H_0^{(i)}$ versus $H_1^{(i)}$ we collect observations using the sensing vector $a\mathbf{1}_{\boldsymbol{P}^{(i)}}$ with an arbitrarily small a and perform a SLRT such as the one in Lemma 3.2. When there is no active component in $\boldsymbol{P}^{(i)}$ the drift of the likelihood-ratio process is smaller than if there was one active component by monotonicity of the likelihood

ratio. This results in the test terminating sooner in expectation than it would under $H_0^{(i)}$ and the probability of accepting $H_1^{(i)}$ is also smaller than the type I error probability α.

We continue by upper bounding the expected sensing energy used by the procedure. Again we have results similar to Lemma 3.2 for the tests carried out in the search phase, and we can use Lemma 3.1 to bound the energy used in the refinement phase. Hence, given $\alpha, \beta, \alpha', \beta'$ and $\boldsymbol{P}$, we can bound the total energy used by the procedure. Also note that

$$\mathbb{E}_S(|\boldsymbol{P}|) \leq p + p \sum_{i:\ \boldsymbol{P}^{(i)} \not\subseteq S} \alpha \ ,$$

thus choosing $\alpha = \varepsilon/2n$ ensures $\mathbb{E}_S(|\boldsymbol{P}|) \leq 2p$.

Using the notation A_{search} and $A_{\text{refinement}}$ as before we get

$$\begin{aligned}\mathbb{E}_S(\|A\|_F) &\leq \mathbb{E}_S(\|A_{\text{search}}\|_F^2) + \mathbb{E}_S\left(\mathbb{E}_S(\|A_{\text{refinement}}\|_F^2 \big| |\boldsymbol{P}|)\right) \\ &\leq \frac{2p(p-1)}{(s-1)^2\mu^2} \log\frac{4s}{\varepsilon} + \frac{2p}{(s-1)^2\mu^2} \log\frac{4n}{\varepsilon} + \frac{4p}{\mu^2} \log\frac{4n}{\varepsilon} \ .\end{aligned}$$

When $s \ll n$ the first term dominates the bound. Combining this with (3.3) we get the following:

Proposition 3.3. *Consider the class of s-stars and suppose that $\frac{\sqrt{n}}{\log 4n} \geq s^2$. Then the above estimator satisfies* (3.3) *and* (3.4) *whenever*

$$\mu \geq \sqrt{\frac{16n}{(s-1)^2 m} \log\frac{4s}{\varepsilon}} \ .$$

In Section 3.3.2 we show that the bound of Proposition 3.3 is near optimal in this sparsity regime. We also show that the sparsity assumption in the proposition above is needed and is not an artifact of our method.

When $k > 1$ (S consists of two or more s-stars) similar arguments hold. When $k \ll s$ it is possible to modify the procedure such that the search phase aims to find the center of the k stars. The modifications include setting $H_0^{(i)} : |\boldsymbol{P}^{(i)} \cap S| = k$, and slightly changing $\alpha, \beta, \alpha', \beta'$ to account for the fact that there is more than one star. For instance choosing α, α' to be the same as before and setting $\beta = \beta' = \varepsilon/4ks$ we get the following:

Proposition 3.4. *Consider the class of k disjoint s-stars and suppose $k < s$ and $\frac{\sqrt{n}}{\log 4n} \geq k(s-k)^2$. Then the modified estimator satisfies* (3.3) *and* (3.4) *whenever*

$$\mu \geq \sqrt{\frac{16n}{(s-k)^2 m} \log \frac{4sk}{\varepsilon}} \ .$$

We see in Section 3.3.2 that the bound above is near the optimal one when k is much smaller than s.

s_r, s_c-submatrices

Let the components of $\boldsymbol{x}$ be in one-to-one correspondence to elements of a matrix M with n_r rows and n_c columns (and let $n = n_r \cdot n_c$). Recall that a set $S \subset [n]$ an s_r, s_c-submatrix when the elements $m_i \in M$ corresponding to the components $i \in S$ form an $s_r \times s_c$ submatrix in M. Let $\mathcal{C}$ be the class of all s_r, s_c-submatrices in $\boldsymbol{x}$. Suppose without loss of generality that $s_r \geq s_c$ and recall that the number of non-zero components of $\boldsymbol{x}$ is simply $s = s_r \cdot s_c$.

One possible way to estimate S is to first find the active columns in the search phase and then focus on one or more active columns in the refinement phase to find the active rows. Let $\boldsymbol{c}^{(i)}$ denote the ith column of $\boldsymbol{x}$, for $i \in [n_c]$. In order to find the active columns we need to decide between

$$H_0^{c(i)} : \ |\boldsymbol{c}^{(i)} \cap S| = 0 \qquad \text{versus} \qquad H_1^{c(i)} : \ |\boldsymbol{c}^{(i)} \cap S| = s_r, \qquad \text{for } i \in [n_c] \ .$$

To do this we perform independent SLRTs with type I and type II error probabilities α and β respectively for every $i \in [n_c]$. At the end of the search phase we return $\boldsymbol{P}$, which is the union of columns $c^{(i)}$ for which $H_1^{c(i)}$ was accepted. Choosing α, β appropriately ensures that with high probability $\boldsymbol{P}$ contains all the active columns and only those. In the refinement phase we test whether row j of $\boldsymbol{P}$ is active or not using a similar method as above, with error probabilities α', β' for every $j \in [n_r]$. In particular the tests are formulated as

$$H_0^{r(j)} : \ |(\boldsymbol{r}^{(j)} \cap \boldsymbol{P}) \cap S| = 0 \quad \text{versus} \quad H_1^{r(j)} : \ |(\boldsymbol{r}^{(j)} \cap \boldsymbol{P}) \cap S| = s_c, \quad \text{for } j \in [n_r] \ ,$$

where $\boldsymbol{r}^{(j)}$ denotes the jth row of $\boldsymbol{x}$, $j = 1, \ldots, n_r$. Finally our estimate $\widehat{S}$ are those elements that are in a row and column that were both deemed active.

The next step is to choose the error probabilities $\alpha, \beta, \alpha', \beta'$. We can bound the

expected Hamming-distance as

$$\mathbb{E}_S(|\widehat{S}\triangle S|) \leq n\alpha + s\beta + n\alpha' + s\beta' \ ,$$

since type I error in the search phase can result in at most n_r errors in $\widehat{S}$ and there can be at most n_c type I errors in the search phase, whereas a type II error can produce at most s_r errors in the end and there are s_c possibilities to make such an error. A similar argument holds for tests in the refinement phase. Hence the choices $\alpha = \alpha' = \varepsilon/4n$ and $\beta = \beta' = \varepsilon/4s$ ensure (3.4).

We move on to bounding the expected energy used by the procedure. To test the ith hypothesis in the search phase we collect measurements using sensing vector $a\mathbf{1}_{\boldsymbol{c}^{(i)}}$ with an arbitrarily small a for all $i \in [n_c]$ and perform a SLRT similar to that described in the previous cases. To perform the jth SLRT of the refinement phase we collect measurements of the form $a\mathbf{1}_{r^{(j)}\cap \boldsymbol{P}}$ using an arbitrarily small a. For these tests we have results identical to Lemmas 3.2 and 3.3. Also for the number of columns in $\boldsymbol{P}$ denoted by $\widetilde{n}_c$ we have

$$\mathbb{E}_S(\widetilde{n}_c) \leq s_c + n_c\alpha \leq 2s_c \ .$$

Putting everything together yields

$$\begin{aligned}\mathbb{E}_S(\|A\|_F) &\leq \mathbb{E}_S(\|A_{\text{search}}\|_F^2) + \mathbb{E}_S\left(\mathbb{E}_S(\|A_{\text{refinement}}\|_F^2 \big| |\boldsymbol{P}|)\right) \\ &\leq \frac{2n}{s_r^2\mu^2}\log\frac{4s}{\varepsilon} + \frac{2n_r s_c}{s_r^2\mu^2}\log\frac{4n}{\varepsilon} + \frac{4n_r}{s_c\mu^2}\log\frac{4n}{\varepsilon} \ .\end{aligned}$$

When $s \ll n$ the first term dominates the bound above. Combining this with (3.3) yields the following result:

Proposition 3.5. *Consider the class of s_r, s_c-submatrices and suppose $\frac{n_c}{\log 4n} \geq \frac{s_r^2}{s_c}$. Then the estimator above satisfies* (3.3) *and* (3.4) *whenever*

$$\mu \geq \sqrt{\frac{8n}{s_r^2 m}\log\frac{4s}{\varepsilon}} \ .$$

Note that the condition on the sparsity in the proposition above is not very strict. Consider square submatrices within square matrices so that we have $n_r = n_c = \sqrt{n}$ and $s_r = s_c = \sqrt{s}$. Then the condition becomes $\frac{\sqrt{n}}{\log 4n} > \sqrt{s}$, which would automatically be fulfilled if there was no logarithmic term on the left. We see in

Section 3.3.2 that in some sparsity regimes the above bound matches the lower bounds we derive. Thus in those regimes this procedure is near optimal. However, in what follows we slightly modify the procedure above to improve performance for submatrices that are more sparse than the ones required in the proposition above. This combined with the results of Section 3.3.2 shows that the best performance we can hope for depends on the sparsity in a non-trivial manner in the case of submatrices.

Note that in principle it is enough to find a single active column in the search phase, as accurately estimating components within *any* active column will yield the identity of all the active rows and similarly estimating components within *any* active row yields the active columns. This motivates the following modification: return a single active column in the search phase, then focus on that column to find the active rows and finally focus on one active row to find the active columns. To do this we retain most of the algorithmic choices of the earlier approach, but choose different α and β.

Ideally we would like to accept $H_1^{c(i)}$ for exactly one active column, so our choices for α, β will be made accordingly. In the refinement phase we choose a column randomly from the ones that were deemed active and locate the active components within that column, using the same procedure as in the unstructured case. This gives us the active rows. Finally we choose a row that was deemed active, and find all the active components within that row to find the active columns. Throughout the refinement phase we set type I and type II error probabilities to be α', β'. With the right choices for the error probabilities, this procedure outperforms the previous one in certain sparsity regimes.

First we need to choose the error probabilities for the tests. We can write

$$\begin{aligned}\mathbb{E}_S(|\widehat{S} \triangle S|) &\leq 2s\mathbb{P}_S(\boldsymbol{P} = \emptyset) + (2n\alpha' + 2s)\,\mathbb{P}_S(\exists \boldsymbol{c}^{(i)} \subset \boldsymbol{P} :\ \boldsymbol{c}^{(i)} \cap S = \emptyset) \\ &\quad + (2n\alpha' + 2s\beta') \\ &\leq 2s\beta^{s_c} + (2n\alpha' + 2s)n_c\alpha + (2n\alpha' + 2s\beta')\ .\end{aligned}$$

Thus the conservative choices $\alpha = \varepsilon/16n^2, \beta = \sqrt[s_c]{\varepsilon/8s}, \alpha' = \varepsilon/8n, \beta' = \varepsilon/8s$ ensure that (3.4) holds.

Now we can move on to calculate the expected sensing energy used by the

procedure. In the same way as before,

$$\begin{aligned}\mathbb{E}_S(\|A\|_F) &\leq \mathbb{E}_S(\|A_{\text{search}}\|_F^2) + \mathbb{E}_S(\|A_{\text{refinement}}\|_F^2) \\ &\leq \frac{2n}{s_c s_r^2 \mu^2} \log \frac{8s}{\varepsilon} + \frac{4n_r s_c}{s_r^2 \mu^2} \log \frac{4n}{\varepsilon} + \frac{4 \max\{n_r, n_c\}}{\mu^2} \log \frac{8n}{\varepsilon} \ .\end{aligned}$$

Combining the above with (3.3) and using that when $s \ll n$ the first term dominates, we arrive at the following result:

Proposition 3.6. *Consider the class of s_r, s_c-submatrices and suppose that $\frac{\min\{n_r,n_c\}}{\log 8n} \geq s_c s_r^2$. Then the above estimator satisfies* (3.3) *and* (3.4) *whenever*

$$\mu \geq \sqrt{\frac{10n}{s_c s_r^2 m} \log \frac{8s}{\varepsilon}} \ .$$

The condition on the sparsity in Proposition 3.6 above is stronger than that in Proposition 3.5. On the other hand the bound for μ is smaller. This shows that in sparser regimes it is indeed possible to outperform the procedure of Proposition 3.5, hinting that the sparsity regime non-trivially influences the best possible performance of adaptive support recovery procedures in the case of sub-matrices. For instance, considering square matrices when $n_r = n_c = \sqrt{n}$ and $s_r = s_c = \sqrt{s}$, the above condition reads $\frac{\sqrt{n}}{2 \log 4n} > \sqrt{s^3}$ which is slightly stronger than the condition in Proposition 3.5.

3.3.2 Lower bounds

We turn our attention to the fundamental limits of recovering the support of structured sparse signals using compressive measurements by *any* adaptive sensing procedure. We consider both the non-adaptive sensing and adaptive sensing settings.

We use the same arguments to obtain lower bounds as in Chapter 2. We only need to adjust some the computations of the Kullback-Leibler divergences to account for the different sensing model. Nonetheless, in order to keep this chapter self contained, we provide all the steps of the arguments here as well.

Similarly to Chapter 2, some of the lower bounds presented below consider the probability of error $\mathbb{P}_S(\widehat{S} \neq S)$ as the error metric. Note that this is more forgiving than $\mathbb{E}_S(|\widehat{S} \triangle S|)$, hence lower bounds with the former metric in mind apply as lower bounds with the latter metric as well.

Non-Adaptive Sensing

First we consider the non-adaptive compressive sensing setting. Comparing these lower bounds with the performance bounds of the previous section illustrates the gains adaptivity provides in the various cases. We do not make any claim on whether these lower bounds are tight or not, as these serve mostly for comparison between adaptive and non-adaptive sensing. The lower bounds presented for the non-adaptive case consider $\mathbb{P}_S(\widehat{S} \neq S)$ as the error metric. For certain classes (s-sets, s-intervals) there exist procedures satisfying $\max_{S\in\mathcal{C}} \mathbb{E}_S(|\widehat{S}\triangle S|) \leq \varepsilon$ with performance matching the lower bounds below.

In the non-adaptive sensing setting we need to define sensing actions before any measurements are taken. That means the sensing matrix A is specified prior to taking any observations. This does not exclude the possibility that A is random, but it has to be generated before any observations are made.

Similarly to Chapter 2 all the bounds for the non-adaptive case are based on Proposition 2.3 in Tsybakov [139]. The exact statement can be found on page 49 (Lemma 2.1). In summary, the statement says that if $\mathbb{P}_0, \ldots, \mathbb{P}_M$ are probability measures, and

$$\frac{1}{M}\sum_{j=1}^{M} D(\mathbb{P}_j \| \mathbb{P}_0) \leq \alpha \ ,$$

then

$$\inf_{\Psi} \max_{j=0,\ldots,M} \mathbb{P}_j(\Psi \neq j) \geq \sup_{0<\tau<1} \left(\frac{\tau M}{1+\tau M} \left(1 + \frac{\alpha + \sqrt{\alpha/2}}{\log \tau} \right) \right) .$$

Let $\mathbb{P}_0, \ldots, \mathbb{P}_M$ be the probability measures induced by sampling $\boldsymbol{x}$ with sensing matrix A, when the support set is $S_0, \ldots, S_M$ respectively, where $S_i \in \mathcal{C}$. Now note that

$$\begin{aligned}
D(\mathbb{P}_j \| \mathbb{P}_0) &= \mathbb{E}_0 \left(\sum_t \log \frac{\mathrm{d}\mathbb{P}_0(Y_t|A_t)}{\mathrm{d}\mathbb{P}_j(Y_t|A_t)} \right) \\
&= \sum_t \mathbb{E} \left(\mathbb{E}_0 \left(-\frac{1}{2} \left((Y_t - \mu\langle A_t, \mathbf{1}_{S_0}\rangle)^2 - (Y_t - \mu\langle A_t, \mathbf{1}_{S_j}\rangle)^2 \right| \right) A \right) \right) \\
&= \sum_t \mathbb{E} \left(\mathbb{E}_0 \left(\frac{1}{2} \left(\mu^2(\langle A_t, \mathbf{1}_{S_j}\rangle^2 - \langle A_t, \mathbf{1}_{S_0}\rangle^2) - 2\mu Y_t \langle A_t, \mathbf{1}_{S_j} - \mathbf{1}_{S_0}\rangle \right) \right| A \right) \right) \ .
\end{aligned}$$

Note that $\mathbb{E}_0(Y_t|A_t) = \mu\langle A_t, \mathbf{1}_{S_0}\rangle$. We can thus continue as

$$\begin{aligned} D(\mathbb{P}_j\|\mathbb{P}_0) &= \frac{\mu^2}{2}\mathbb{E}\left(\sum_t \left(\langle A_t, \mathbf{1}_{S_j}\rangle^2 + \langle A_t, \mathbf{1}_{S_0}\rangle^2 - 2\langle A_t, \mathbf{1}_{S_j}\rangle\langle A_t, \mathbf{1}_{S_0}\rangle\right)\right) \\ &= \frac{\mu^2}{2}\mathbb{E}\left(\sum_t \langle A_t, \mathbf{1}_{S_j} - \mathbf{1}_{S_0}\rangle^2\right) \\ &\leq \frac{\mu^2}{2}\mathbb{E}\left(\sum_t |S_0 \triangle S_j| \sum_{i\in S_0\triangle S_j} A_{t,i}^2\right) \\ &= \frac{\mu^2}{2}|S_0\triangle S_j| \sum_{i\in S_0\triangle S_j} a_i^2 \ , \end{aligned} \tag{3.6}$$

where $A_{t,j}$ is the (t,j)th element of the sensing matrix A, a_i^2 denotes $\mathbb{E}(\sum_t A_{t,i}^2)$, and in the second to last step we use Jensen's inequality.

Now consider the right hand side of the lemma above and set $\tau = 1/M$. To make the bound more transparent suppose that $1 \leq (1-2\varepsilon)\log M$, which is essentially always satisfied when M is large enough and $\varepsilon \in (0, 1/2)$. This way we arrive at the inequality

$$2\alpha \geq (1-2\varepsilon)\log M \ . \tag{3.7}$$

Choosing the sets $S_0, \ldots, S_M$ and using inequality (3.6) to bound the average KL distance, we can use the above inequality to get lower bounds for μ. These choices will be specific to the different classes that we are considering.

Remark 3.7. *In the following statements we require n to be divisible by s. This condition is merely for technical convenience, and can be dropped easily at the expense of a cumbersome presentation.*

Theorem 3.1 (s-sets)**.** *Let $\mathcal{C}$ be the class of s-sets and suppose that n/s is an integer. If there is a non-adaptive estimator $\widehat{S}$ that satisfies* (3.3) *and $\mathbb{P}_S(\widehat{S} \neq S) \leq \varepsilon \ \forall S \in \mathcal{C}$, then*

$$\mu \geq \sqrt{(1-2\varepsilon)\frac{n}{4m}\log(n-s)} \ .$$

Proof. Let $S_0 \in \mathcal{C}$ be arbitrary. Partition $[n]$ into s bins of equal size denoted by $\boldsymbol{P}^{(1)}, \ldots, \boldsymbol{P}^{(s)}$ such that each bin contains exactly one element of S_0. Denote $s_i = S_0 \cap \boldsymbol{P}^{(i)}$, $i \in [s]$. Now consider the sets $S_1, \ldots, S_M$ that we get by modifying exactly one element of S_0 in the following way: pick one element of S_0 denoted by s_i and swap it with some other element in $\boldsymbol{P}^{(i)}$ thus changing the position of the

active component within $\boldsymbol{P}^{(i)}$. We can generate $M = n - s$ sets in the previous manner. From (3.6) we have that

$$\frac{1}{M}\sum_{j=1}^{M} D(\mathbb{P}_j \| \mathbb{P}_0) \leq \frac{1}{M}\mu^2 \sum_{j=1}^{M} \sum_{i \in S_0 \triangle S_j} a_i^2 = \frac{1}{n-s}\mu^2 \left(\sum_{i=1}^{n} a_i^2 + \frac{n-2s}{s} \sum_{i \in S_0} a_i^2 \right) .$$

Now note that by the total energy constraint (3.3) we have

$$\sum_{i=1}^{n} a_i^2 \leq m .$$

Also note that given A we can always choose S_0 to be the one that is the most difficult to distinguish from the other sets $S_1, \ldots, S_M$. That is we have to solve

$$\max_{A:\ \|A\|_F \leq m} \ \min_{S_0 \in \mathcal{C}} \sum_{i \in S_0} a_i^2 .$$

The expression above is clearly upper bounded by sm/n. Hence in the expression above we can take S_0 such that $\sum_{i \in S_0} a_i^2 \leq sm/n$. Combining what we have yields

$$\frac{1}{M}\sum_{j=1}^{M} D(\mathbb{P}_j \| \mathbb{P}_0) \leq \frac{1}{n-s}\left(1 + \frac{n-2s}{n}\right) m\mu^2 \leq \frac{2m}{n}\mu^2 .$$

Using this with (3.7) concludes the proof. □

Theorem 3.2 (Unions of s-intervals)**.** *Let $\mathcal{C}$ be the class of unions of k disjoint s-intervals and suppose that n/s is an integer. If there is a non-adaptive estimator $\widehat{S}$ that satisfies* (3.3) *and $\mathbb{P}_S(\widehat{S} \neq S) \leq \varepsilon \ \forall S \in \mathcal{C}$, then*

$$\mu \geq \sqrt{(1-2\varepsilon)\frac{n-(k-1)s}{4s^2 m}\log(\frac{n}{s}-k)} .$$

Proof. Partition $[n]$ into consecutive intervals of size s denoted by $S^{(1)}, \ldots, S^{(n/s)}$. Now consider the subclass whose elements are unions of the first $k-1$ intervals $S^{(1)}, \ldots, S^{(k-1)}$ and some other interval $S^{(i)}$. Formally,

$$\mathcal{C}' = \{S \in \mathcal{C} :\ S = S^{(i)} \cup \left(\bigcup_{j=1}^{k-1} S^{(j)} \right),\ i \in [k+1, n/s]\} .$$

This way we effectively reduced this problem to finding one interval in a slightly smaller vector. Let $S_0 \in \mathcal{C}'$ be arbitrary and let $S_1, \dots, S_M$ be all the other elements of $\mathcal{C}'$, so $M = n/s - k$. Let $\widetilde{S}_0 = S_0 \setminus \cup_{j=1}^{k-1} S^{(j)}$. From (3.6) we have

$$\begin{aligned}\frac{1}{M}\sum_{j=1}^{M} D(\mathbb{P}_j \| \mathbb{P}_0) &\leq s\mu^2 \frac{1}{M}\sum_{j=1}^{M}\sum_{i \in S_0 \triangle S_j} a_i^2 \\ &= \frac{s^2\mu^2}{n-ks}\left(\sum_{i=(k-1)s+1}^{n} a_i^2 + \frac{n-(k+1)s}{s}\sum_{i\in\widetilde{S}_0} a_i^2\right) .\end{aligned}$$

Again, from (3.3) and the fact that we can choose $S_0 \in \mathcal{C}'$ after the sensing strategy has been determined we have

$$\frac{1}{M}\sum_{j=1}^{M} D(\mathbb{P}_j \| \mathbb{P}_0) \leq \frac{1}{n-ks}\left(1 + \frac{n-(k+1)s}{n-(k-1)s}\right) s^2 m\mu^2 \leq \frac{2s^2 m}{n-(k-1)s}\mu^2 .$$

Using this with (3.7) concludes the proof. □

Theorem 3.3 (s-stars)**.** *Let $\mathcal{C}$ be the class of s-stars and suppose that p/s is an integer. If there is a non-adaptive estimator $\widehat{S}$ that satisfies* (3.3) *and $\mathbb{P}_S(\widehat{S} \neq S) \leq \varepsilon$ $\forall S \in \mathcal{C}$, then*

$$\mu \geq \sqrt{(1-2\varepsilon)\frac{n}{2m}\log(\sqrt{2n} - s - 1)} .$$

Proof. Consider the $p-1$ edges of the complete graph of p vertices that share a common vertex j. Denote this set of edges by E_j. The s-stars whose center is vertex j form a class of s-sets on E_j. So we can perform the same construction on this set of edges as in Theorem 3.1 to get

$$\frac{1}{M}\sum_{j=1}^{M} D(\mathbb{P}_j \| \mathbb{P}_0) \leq \frac{1}{p-1-s}\mu^2\left(\sum_{i\in E_j} a_i^2 + \frac{p-1-2s}{s}\sum_{i\in S_0} a_i^2\right) .$$

Now note that we can choose any star to be S_0 which implies $\sum_{i\in S_0} a_i^2 \leq sm/n$ and $\sum_{i\in E_j} a_i^2 \leq (p-1)m/n$ yielding

$$\frac{1}{M}\sum_{j=1}^{M} D(\mathbb{P}_j \| \mathbb{P}_0) \leq \frac{2m}{n}\mu^2 .$$

The statement now follows from (3.7) and that $p > \sqrt{2n}$. □

Considering unions of k disjoint s-stars we can get a similar lower bound by considering a subclass where $k-1$ of the s-stars are fixed and only one can change, reducing the problem to finding one s-star.

Theorem 3.4 (s-submatrices)**.** *Let $\mathcal{C}$ be the class of s-submatrices of a fixed size $s_c \times s_r$, and suppose both n_c/s_c and n_r/s_r are integers. If there is a non-adaptive estimator $\widehat{S}$ that satisfies* (3.3) *and $\mathbb{P}_S(\widehat{S} \neq S) \leq \varepsilon \ \forall S \in \mathcal{C}$, then*

$$\mu \geq \sqrt{(1-2\varepsilon)\frac{n}{4m}\max\left\{\frac{1}{s_r}\frac{n_c-s_c}{n_c}, \frac{1}{s_c}\frac{n_r-s_r}{n_r}\right\}\log\left(\max\{n_r-s_r, n_c-s_c\}\right)} \ .$$

Proof. Let $S_0 \in \mathcal{C}$ be arbitrary. Denote the indices of the rows of S_0 by $r_1, \dots, r_{s_r}$, and let $S_0^{(j)}$ denote the jth row of S_0. Consider a partition of the indexes $[n_r]$ into $\boldsymbol{r}^{(1)}, \dots, \boldsymbol{r}^{(s_r)}$ such that all of the are of the same size and $\boldsymbol{r}^{(j)}$ contains exactly one active row indexed by r_j for every $j \in [r_{s_r}]$.

Now let $S_1, \dots, S_M$ be elements of $\mathcal{C}$ that we get by replacing exactly one row index of S_0 such that if we modify r_j, then the new row index is in $\boldsymbol{r}^{(j)}$. There are $n_r - s_r$ such submatrices. The same way as for the s-sets we get

$$\frac{1}{M}\sum_{j=1}^{M} D(\mathbb{P}_j \| \mathbb{P}_0) \leq \frac{1}{n_r - s_r}\mu^2 \left(\sum_{(i,l):\ l \in C_0} a_{(i,l)}^2 + \frac{n_r - 2s_r}{s_r} \sum_{(i,l) \in S_0} a_{(i,l)}^2 \right) \ ,$$

where C_0 denotes the set of column indices of S_0. Again, the fact that we can choose an arbitrary $S_0 \in \mathcal{C}$ after the sensing strategy has been fixed results in the upper bound

$$\frac{1}{M}\sum_{j=1}^{M} D(\mathbb{P}_j \| \mathbb{P}_0) \leq \frac{2s_c m}{n}\frac{n_r}{n_r - s_r}\mu^2 \ .$$

Plugging this into (3.7) and rearranging gives a lower bound. Repeating the same arguments for columns concludes the proof. □

Adaptive Sensing

Here we provide lower bounds considering the adaptive sensing framework. Comparing these bounds with the performance bounds of Section 3.3.1 shows the near optimality of the procedures presented there.

s-sets

Adaptive sensing lower bounds for unstructured classes were proved by Castro

in [40]. In that work lower bounds are derived by slightly broadening the class, which we state here for convenience. Note that the fact that the following lower bound is valid for a slightly larger class than the class of s-sets does not cause a problem, see Remarks 3.4 and 3.8. Let $\mathcal{C}_s$ denote the class of s-sets. We have the following result:

Theorem 3.5. *Let $\mathcal{C} = \mathcal{C}_s \cup \mathcal{C}_{s-1}$, and suppose that there exists an estimator $\widehat{S}$ that satisfies* (3.3) *and* (3.4)*. Then*

$$\mu \geq \sqrt{\frac{2(n-s+1)}{m}\left(\log\frac{s}{2\varepsilon} + \log\frac{n-s+1}{n+1}\right)} \ .$$

Remark 3.8. *Note that the bound above holds for estimators for sets with sparsity s or $s-1$. The procedure presented in Section 3.3.1 works for this class of sets without any modifications. Later on for the structured classes we rely on the above proposition to derive lower bounds, hence a similar comment applies to those cases as well.*

s-intervals and unions of s-intervals

For s-intervals we have multiple ways of deriving lower bounds, just as in the case of coordinate wise sampling in Chapter 2. First we consider $\mathbb{P}_S(\widehat{S} \neq S)$ as the error metric. The following result is analogous to the lower bound derived by Balakrishnan et al. in [17], and the proof is included here for the sake of clarity.

Proposition 3.7. *Let $\mathcal{C}$ be the class of s-intervals and suppose there is an estimator $\widehat{S}$ satisfying* (3.3) *and $\max_{S\in\mathcal{C}} \mathbb{P}_S(\widehat{S} \neq S) \leq \varepsilon$. Furthermore suppose that n/s is an integer. Then*

$$\mu \geq (1-\varepsilon)\sqrt{\frac{n}{2s^2 m}} \ .$$

Proof. Consider the subclass of consecutive disjoint s-intervals

$$\{[1,s],[s+1,2s],\ldots,[n-s+1,n]\} \ .$$

Partition this subclass into two subclasses of equal size denoted by $\mathcal{C}_1$ and $\mathcal{C}_2$. Let π_i denote the uniform distribution on the subclass $\mathcal{C}_i$ for $i = 1, 2$, and consider the two hypotheses $H_i : \ S \sim \pi_i, \ i = 1, 2$. If there exists an estimator $\widehat{S}$ satisfying (3.3), then there exists a test function $\Phi : D \to \{1, 2\}$ such that

$$\mathbb{P}_1(\Phi(D) = 2) + \mathbb{P}_2(\Phi(D) = 1) \leq \varepsilon \ ,$$

where $\mathbb{P}_i$ denotes the distribution of $D = \{Y_t, A_t\}_{t\geq 1}$ when H_i is true, $i = 1, 2$. Let $\mathbb{P}_0$ denote the distribution of D when in fact $S = \emptyset$. Let $TV(\cdot, \cdot)$ be the total variation distance and $KL(\cdot, \cdot)$ be the Kullback-Leibler divergence of two distributions, and assume without loss of generality that $TV(\mathbb{P}_0, \mathbb{P}_1) \geq TV(\mathbb{P}_0, \mathbb{P}_2)$. We have

$$\begin{aligned}
\varepsilon &\geq \mathbb{P}_1(\Phi(D) = 2) + \mathbb{P}_2(\Phi(D) = 1) \geq 1 - TV(\mathbb{P}_1, \mathbb{P}_2) \\
&\geq 1 - (TV(\mathbb{P}_0, \mathbb{P}_1) + TV(\mathbb{P}_0, \mathbb{P}_2)) \geq 1 - 2TV(\mathbb{P}_0, \mathbb{P}_1) \\
&\geq 1 - \sqrt{2KL(\mathbb{P}_0, \mathbb{P}_1)} \, .
\end{aligned}$$

Now the goal is to upper bound $KL(\mathbb{P}_0, \mathbb{P}_1)$. Let Y denote the observations $\{Y_t\}_{t\geq 1}$, and let $\mathbb{P}_S$ denote the distribution of Y for a fixed support S. We have

$$\begin{aligned}
KL(\mathbb{P}_0, \mathbb{P}_1) &= \mathbb{E}_0 \left(\log \frac{\mathrm{d}\mathbb{P}_0(D)}{\mathrm{d}\mathbb{P}_1(D)} \right) = \mathbb{E}_0 \left(\log \frac{\mathrm{d}\mathbb{P}_0(Y)}{\mathrm{d}\mathbb{P}_1(Y)} \right) \\
&= -\mathbb{E}_0 \left(\log \frac{\mathrm{d}\mathbb{P}_1(Y)}{\mathrm{d}\mathbb{P}_0(Y)} \right) = -\mathbb{E}_0 \left(\log \frac{\mathbb{E}_{S\sim\pi_1} \left(\mathrm{d}\mathbb{P}_S(Y) \right)}{\mathrm{d}\mathbb{P}_0(Y)} \right) \\
&\leq -\mathbb{E}_0 \left(\mathbb{E}_{S\sim\pi_1} \left(\log \frac{\mathrm{d}\mathbb{P}_S(Y)}{\mathrm{d}\mathbb{P}_0(Y)} \right) \right) \, ,
\end{aligned}$$

by Jensen's inequality. By plugging in the densities, we get

$$\begin{aligned}
KL(\mathbb{P}_0, \mathbb{P}_1) &\leq -\mathbb{E}_0 \left(\mathbb{E}_{S\sim\pi_1} \left(-\frac{1}{2} \sum_t \left((Y_t - \mu \langle A_t, \mathbf{1}_S \rangle)^2 - Y_t^2 \right) \right) \right) \\
&= \frac{1}{2} \mathbb{E}_0 \left(\mathbb{E}_{S\sim\pi_1} \left(\sum_t \left(\mu^2 \langle A_t, \mathbf{1}_S \rangle^2 - 2\mu \langle A_t, \mathbf{1}_S \rangle Y_t \right) \right) \right) \\
&= \frac{\mu^2}{2} \mathbb{E}_0 \left(\mathbb{E}_{S\sim\pi_1} \left(\sum_t A_t^T \mathbf{1}_S \mathbf{1}_S^T A_t \right) \right) \\
&= \frac{\mu^2}{2} \mathbb{E}_0 \left(\sum_t A_t^T \mathbb{E}_{S\sim\pi_1} \left(\mathbf{1}_S \mathbf{1}_S^T \right) A_t \right) \, ,
\end{aligned}$$

where $\mathbb{E}_{S\sim\pi_1}$ is the expectation w.r.t. S when it is distributed according to π_1. Now $\mathbb{E}_{S\sim\pi_1} \left(\mathbf{1}_S \mathbf{1}_S^T \right) = \frac{2s}{n} I'$ where $I' \in \mathbb{R}^{n\times n}$ is block diagonal with $n/2s$ blocks of size $s \times s$ consisting of all ones, and the rest of the matrix consists of zeros. Thus

we can continue as

$$\begin{aligned}
KL(\mathbb{P}_0, \mathbb{P}_1) &\leq \frac{\mu^2}{2}\mathbb{E}_0\left(\sum_t A_t^T \mathbb{E}_{S\sim\pi_1}\left(\mathbf{1}_S \mathbf{1}_S^T\right) A_t\right) \\
&= \mu^2 \frac{s}{n}\mathbb{E}_0\left(\sum_t A_t^T I' A_t\right) = \mu^2 \frac{s}{n}\mathbb{E}_0\left(\sum_t \langle A_t, I' A_t\rangle\right) \\
&\leq \mu^2 \frac{s}{n}\mathbb{E}_0\left(\sum_t |\langle A_t, I' A_t\rangle|\right) \\
&\leq \mu^2 \frac{s}{n}\mathbb{E}_0\left(\sum_t \|A_t\|_2 \|I' A_t\|_2\right) \\
&\leq \mu^2 \frac{s}{n}\mathbb{E}_0\left(\sum_t \|A_t\|_2^2 \|I'\|_2\right) \\
&\leq \mu^2 \frac{ms^2}{n} \ ,
\end{aligned}$$

where $\|I'\|_2$ is the matrix norm of I' induced by the Euclidean norm and the last step follows from $\|I'\|_2 \leq s$ and (3.3). Thus we arrive at the inequality

$$\varepsilon \geq 1 - \sqrt{2\mu^2 \frac{ms^2}{n}} \ ,$$

from which the statement follows. □

In the previous bound the dependence on ε is clearly loose. When considering the Hamming distance as the error metric, we can also get lower bounds by slightly broadening the class. We cover this by considering the case of unions of k disjoint s-intervals, which as a special case contains the class of s-intervals when $k = 1$. We broaden this class by adding unions of $k-1$ disjoint s-intervals as well.

Theorem 3.6. *Let $\mathcal{C}$ be the class of unions of k or $k-1$ disjoint s-intervals with $k > 0$ fixed, and suppose that n/s is an integer. Suppose there is an estimator satisfying* (3.3) *and* $\max_{S\in\mathcal{C}} \mathbb{E}_S\big(d(\widehat{S}, S)\big) \leq \varepsilon$. *Then*

$$\mu \geq \sqrt{\frac{2\big(n - s(k-1)\big)}{s^2 m}\left(\log\frac{ks}{8\varepsilon} + \log\frac{n - s(k-1)}{n+s}\right)} \ .$$

Proof. Partition $[n]$ into consecutive disjoint s-intervals denoted by $S_1, \ldots, S_{n/s}$, that is $S_j = \{(j-1)s + 1, \ldots, js\}$, and consider the subclass $\mathcal{C}'$ of $\mathcal{C}$ consisting

of all the sets in $\mathcal{C}$ that can be written in the form $\cup S_j$. This subclass is similar to a general sparse class of sparsity k or $k-1$ with the intervals S_j playing the role of the components. This is exactly what we wish to formalize, and then use Theorem 3.5.

Clearly $\max_{S\in\mathcal{C}'} \mathbb{E}_S\big(d(\widehat{S},S)\big) \le \varepsilon$. Using $\widehat{S}$ we can construct an estimator $\widetilde{S}$ which only takes values of the form $\cup S_j$, and has the property $\max_{S\in\mathcal{C}'} \mathbb{E}_S\big(d(\widetilde{S},S)\big) \le 4\varepsilon$. For instance let $\widetilde{S}$ be such that for every $j\in[n/s]:\ S_j\subset\widetilde{S}$ if and only if $|\widehat{S}\cap S_j|\ge s/2$. The expected Hamming-distance for such estimators can be written as

$$\mathbb{E}_S\left(d(\widetilde{S},S)\right) = s\sum_{j=1}^{n/s}\mathbb{P}_S\left(\mathbf{1}\{S_j\subset\widetilde{S}\}\neq\mathbf{1}\{S_j\subset S\}\right)\ .$$

The measurements $\{Y_t\}_{t\ge1}$ can be written in the form

$$Y_t = \langle A_t,\boldsymbol{x}\rangle + W_t = \mu\sum_{i\in S}a_{i,t} + W_t = s\mu\sum_{S_j\in S}\frac{1}{s}\sum_{i\in S_j}a_{i,t} + W_t\ .$$

Also, from Jensen's inequality we have

$$\sum_t\sum_{j=1}^{n/s}\left(\frac{1}{s}\sum_{i\in S^{(d)}}a_{i,t}\right)^2 \le \sum_t\sum_{j=1}^{n/s}\frac{1}{s}\sum_{i\in S^{(d)}}a_{i,t}^2 = \frac{1}{s}\sum_t\sum_{j=1}^{n/s}\sum_{i\in S^{(d)}}a_{i,t}^2 \le \frac{m}{s}\ .$$

Therefore the problem can be viewed as estimating a general sparse support set. The sparsity is either k or $k-1$, the length of the vector is n/s, the signal strength is $s\mu$, the total sensing budget is m/s and the desired accuracy in expected Hamming-distance is $4\varepsilon/s$. By Theorem 3.5

$$s\mu \ge \sqrt{\frac{2(n/s-k+1)}{m/s}\left(\log\frac{ks}{8\varepsilon}+\log\frac{n/s-k+1}{n/s+1}\right)}\ ,$$

which concludes the proof. □

s-stars and unions of s-stars

We can use the same arguments for these classes as we did for s-intervals and unions of s-intervals. The only thing that needs to be altered is that instead of disjoint s-intervals we use disjoint s-stars. The difference this makes is that whereas before the new problem dimension became n/s, since the entire signal vector could

be covered by disjoint intervals, the same can not be said when considering s-stars.

Let $\mathfrak{N}_{p,s}$ denote the number of disjoint s-stars that can be packed in a complete graph with p vertices. We can easily check that (see Lemma 2.2 in Chapter 2)

$$\mathfrak{N}_{p,s} \geq \frac{p(p-1-s)}{2s} \ .$$

The left hand side is approximately n/s when the signal is sparse, thus essentially the same results hold as in the case of unions of intervals. Thus the analogue of Proposition 3.7 for s-stars is the following:

Proposition 3.8. *Let $\mathcal{C}$ be the class of s-stars and suppose there is an estimator $\widehat{S}$ satisfying* (3.3) *and* $\max_{S\in\mathcal{C}} \mathbb{P}_S(\widehat{S} \neq S) \leq \varepsilon$. *Then*

$$\mu \geq (1-\varepsilon)\sqrt{\frac{\mathfrak{N}_{p,s}}{2sm}} \ .$$

Remark 3.9. *When $s \ll n$, the above bound scales as $(1-\varepsilon)\sqrt{\frac{n}{s^2 m}}$.*

We also have an analogue of Theorem 3.6 for the case of multiple stars.

Theorem 3.7. *Let $\mathcal{C}$ be the class of unions of k or $k-1$ disjoint s-stars. Suppose there is an estimator satisfying* (3.3) *and* $\max_{S\in\mathcal{C}} \mathbb{E}_S\big(d(\widehat{S}, S)\big) \leq \varepsilon$. *Then*

$$\mu \geq \frac{1}{s}\sqrt{\frac{2\big(\mathfrak{N}_{p,s}-k+1\big)}{m/s}\left(\log\frac{ks}{8\varepsilon} + \log\frac{\mathfrak{N}_{p,s}-k+1}{\mathfrak{N}_{p,s}+1}\right)} \ .$$

Remark 3.10. *When $s \ll n$, the above bound scales as $\sqrt{\frac{n}{s^2 m}\log\frac{ks}{\varepsilon}}$.*

We also present another simple lower bound that illustrates that the assumption on the sparsity in Proposition 3.3 requiring approximately that $s^4 \leq n$ is needed and is not only an artifact of our method.

Consider a setting where the support set is a star of size s or $s-1$. Now consider the sub-problem of estimating the support of such a star when the center of the star is given by an oracle. This is an unstructured problem on a vector of size $p-1$. Hence we can directly apply Theorem 3.5 to get the following result:

Proposition 3.9. *Let $\mathcal{C}$ be the class of stars with sparsity s and $s-1$ and suppose*

there is an estimator $\widehat{S}$ *satisfying* (3.3) *and* (3.4)*. Then*

$$\mu \geq \sqrt{\frac{2(p-s)}{m}\left(\log\frac{s}{2\varepsilon} + \log\frac{p-s}{p}\right)} \ .$$

Remark 3.11. *When* $s \ll n$*, the above bound scales as* $\sqrt{\frac{\sqrt{n}}{m}\log\frac{s}{\varepsilon}}$*.*

Combining the results of Theorem 3.7 and Proposition 3.9 shows that considering s-stars the scaling of the signal strength needs to be at least

$$\max\left\{\frac{n}{s^2 m}\log\frac{s}{\varepsilon}, \frac{\sqrt{n}}{m}\log\frac{s}{\varepsilon}\right\} \ .$$

The first term in the maximum above dominates the second when $s^4 \leq n$. This shows that the performance of Proposition 3.3 can only be achieved in that sparsity regime.

Remark 3.12. *Note that the setting of Proposition 3.9 is slightly different than the one considered in Section 3.3.1. However, we present this result here merely to make a remark on the conditions in Proposition 3.3 and it only serves an illustrative purpose. Furthermore the procedure presented in Section 3.3.1 can be easily modified to handle classes considered in the above proposition and have similar performance guarantees to Proposition 3.3.*

s_r, s_c-submatrices

A similar setting has been studied by Balaksrishnan et al. in [17], where the authors consider block-structured activations in matrices. They provide a lower bound akin to that of Proposition 3.7 and a near-optimal procedure. Our setting is more general as we consider arbitrary sub-matrices of a given dimension. Nonetheless the same type of lower bound holds in this case as well.

Proposition 3.10. *Let* $\mathcal{C}$ *be the class of* s_r, s_c*-submatrices, and for the sake of simplicity assume that both* n_r/s_r *and* n_c/s_c *are integers. Suppose that there is an estimator satisfying* (3.3) *and* $\max_{S\in\mathcal{C}} \mathbb{E}_S\big(d(\widehat{S}, S)\big) \leq \varepsilon$*. Then*

$$\mu \geq (1-\varepsilon)\sqrt{\frac{n}{2s^2 m}} \ .$$

Proof. Since both n_r/s_r and n_c/s_c are integers, the proof goes the same way as that of Proposition 3.7 by considering any disjoint partition of the original matrix consisting of submatrices of size $s = s_r \times s_c$. □

However, our procedures do not reach this lower bound, hence the question arises whether the lower bound above is loose or the procedures are suboptimal? We partially answer this question by presenting another simple lower bound with which we illustrate that in certain sparsity regimes the procedure of Proposition 3.5 is indeed optimal. Consider the class containing all $s_r \times s_c$ and $s_r \times (s_c - 1)$ submatrices, and consider the sub-problem of estimating the support when the active rows are given. This is a problem of estimating s_c or $s_c - 1$ disjoint s_r-intervals in a signal of size $s_r \cdot n_c$. Note that the procedure of Proposition 3.5 can handle such classes without any modifications. Now we can directly apply Theorem 3.6 to get the following:

Proposition 3.11. *Let $\mathcal{C}$ be the class containing all submatrices of size $s_r \times s_c$ and $s_r \times (s_c - 1)$. Suppose there is an estimator $\widehat{S}$ satisfying* (3.3) *and* (3.4)*. Then*

$$\mu \geq \sqrt{\frac{2(n_c - s_c + 1)}{s_r m} \left(\log \frac{s}{8\varepsilon} + \log \frac{n_c - s_c + 1}{n_c + 1} \right)} .$$

When $s_r \approx n_r$ (for instance when we have linear sparsity *in the rows*: $s_r = cn_r$ with some $c \in (0, 1]$) the performance bound of Proposition 3.5 becomes essentially identical to the lower bound above. This shows that in certain regimes that procedure is optimal. Note that the condition on the number of active rows does not determine the sparsity of the signal, as there is no requirement on the number of active columns for the results to hold. Also note that by Proposition 3.6 in certain regimes it is possible to outperform the procedure of Proposition 3.5 indicating that the gains one can hope for in the case of submatrices depends on the interplay between the dimensions of the problem n_r, n_c, s_r, s_c. On a final note if we assume that the support set is such that either the active rows of active columns (but not necessary both) are consecutive then one can simply modify the procedure presented in Section 3.3.1 to even reach the lower bound of Proposition 3.10. However, the exact performance characterization of the case of submatrices with arbitrary dimensions remains an interesting open problem.

3.4 Sample complexity

In the preceding sections we have presented near-optimal procedures for structured support recovery using adaptive compressive sensing. Those procedures provide insight on how to capitalize on the structure of the support sets to achieve performance gains, but paid no regard to the number of measurements that are collected. However, an important aspect of compressive sensing is the possibility to perform estimation using only a small number of observations. Therefore we now present procedures for structured support recovery that use only a small number of observations.

3.4.1 Procedures

All the procedures presented here are based on an algorithm named Compressive Adaptive Sense and Search (CASS), introduced and analyzed by Malloy & Nowak in [107]. This procedure is designed to recover non-structured support sets. To ease presentation we briefly describe and analyze the procedure here, though the reader is referred to Malloy & Nowak [107] where this has already been done in more detail.

s-sets

The main idea of the CASS procedure is to use a binary bisection type algorithm to recover the support of the signal. In a nutshell, CASS begins by partitioning the signal into several bins and deciding if there are any significant components inside each bin. Then it continues by partitioning the bins deemed to contain signal into new bins and performing the previous step again for those. By iterating these steps the procedure is able to locate the support in a number of steps that is logarithmic in the dimension of the signal.

Assume the support set is any s-sparse set. Partition $[n]$ into $2s$ bins of equal size, denoted by $\boldsymbol{I}_1^{(1)}, \ldots, \boldsymbol{I}_{2s}^{(1)}$. For each of the $2s$ bins we wish to decide between

$$H_{i,0}^{(1)} : \ \boldsymbol{I}_i^{(1)} \cap S = \emptyset \qquad \text{versus} \qquad H_{i,1}^{(1)} : \ \boldsymbol{I}_i^{(1)} \cap S \neq \emptyset, \ \text{for } i \in [2s] \ .$$

Once having identified the non-empty bins, we split each of these into two bins of equal size denoted by $\boldsymbol{I}_1^{(2)}, \ldots, \boldsymbol{I}_{2n_1}^{(2)}$, where n_1 denotes the number of bins deemed non empty previously, and do the same as before. We know that at most s bins

can be non-empty, thus we will enforce in our procedure that $n_1 \leq s$. Hence in step j we consider bins $\boldsymbol{I}_1^{(j)}, \ldots, \boldsymbol{I}_{2n_{j-1}}^{(j)}$, where $n_{j-1} \leq s$, and test the hypotheses

$$H_{i,0}^{(j)} : \boldsymbol{I}_i^{(j)} \cap S = \emptyset \qquad \text{versus} \qquad H_{i,1}^{(j)} : \boldsymbol{I}_i^{(j)} \cap S \neq \emptyset, \text{ for } i \in [2n_{j-1}] \ .$$

When $j = \log_2 \frac{n}{2s}$ the bins consist of single components of $\boldsymbol{x}$, and the estimator of the support $\widehat{S}$ will consist of the ones deemed non-empty in this final step.

To decide between $H_{i,0}^{(j)}$ and $H_{i,1}^{(j)}$, for $j \in [\log_2 \frac{n}{2s}]$ and $i \in [2n_{j-1}]$, we collect a single measurement of the form

$$Y_i^{(j)} = \langle a\sqrt{j}\mathbf{1}_{\boldsymbol{I}_i^{(j)}}, \boldsymbol{x} \rangle + W_i^{(j)}, \text{ for } j \in [\log_2 \tfrac{n}{2s}]; i \in [2n_{j-1}] \ ,$$

where $W_i^{(j)} \sim N(0,1)$ are i.i.d., and $a > 0$. The parameter $a > 0$ needs to be chosen such that (3.3) is fulfilled. Since the length of the bins $I_i^{(j)}$ is $n/(2^j s)$ for every $i \in [2n_{j-1}]$, $n_{j-1} \leq s$ and there are $\log_2 \frac{n}{2s}$ steps we have

$$\|A\|_F^2 = \sum_{j=1}^{\log_2 \frac{n}{2s}} 2s \frac{n}{2^j s} j a^2 \leq n a^2 \sum_{j=1}^{\infty} j 2^{-(j-1)} = 4na^2 \ .$$

Combining this with (3.3) yields $a = \sqrt{\frac{m}{4n}}$. If the bin $\boldsymbol{I}_i^{(j)}$ is non-empty then $\mathbb{E}_S(Y_i^{(j)}) \geq \mu\sqrt{\frac{jm}{4n}}$. Therefore we deem the bin $\boldsymbol{I}_i^{(j)}$ to be empty when $Y_i^{(j)} \leq \frac{\mu}{2}\sqrt{\frac{jm}{4n}}$, otherwise we deem the opposite. If at any step $j \in [\log_2 \frac{n}{2s}]$ more than s bins are deemed non-empty, we select those that correspond to the s largest observations. For the method described above both the type I and type II error probabilities for the test between $H_{i,0}^{(j)}$ and $H_{i,1}^{(j)}$, for $j \in [\log_2 \frac{n}{2s}]$ and $i \in [2n_{j-1}]$ can be upper bounded using the Gaussian tail bound

$$\mathbb{P}(X > \eta) \leq \frac{1}{2} e^{-\eta^2/2} \tag{3.8}$$

by

$$\frac{1}{2} \exp\left(-\frac{jm\mu^2}{32n}\right) \ .$$

Hence the probability of error can be bounded from above as follows

$$\mathbb{P}_S(\widehat{S} \neq S) \leq \sum_{j=1}^{\log_2 \frac{n}{2s}} s \ \exp\left(-\frac{jm\mu^2}{32n}\right) \ .$$

Thus whenever $\mu^2 \geq \frac{32n}{m} \log \frac{2s}{\varepsilon}$ we have

$$\mathbb{P}_S(\widehat{S} \neq S) \leq \sum_{j=1}^{\log_2 \frac{n}{2s}} s \left(\frac{\varepsilon}{2s}\right)^j \leq \sum_{j=1}^{\log_2 \frac{n}{2s}} \left(\frac{\varepsilon}{2}\right)^j \leq \varepsilon \ .$$

When considering the expected Hamming-distance as the error metric we can use the above procedure with probability of error set to $\varepsilon/2s$. This method then yields a rate-optimal estimator for the support recovery problem described in Section 3.2 by collecting at most $2s \log_2 \frac{n}{2s}$ measurements.

Unions of s-intervals

We can modify the CASS procedure of Malloy & Nowak [107] to estimate unions of k disjoint s-intervals. Similarly to the procedure presented in Section 3.3.1 the one discussed here will consist of two phases, a search phase and a refinement phase. As before, in the search phase we wish to identify the approximate location of the support, that is, return a subset of components $\boldsymbol{P} \subset [n]$ such that $|\boldsymbol{P}| \ll n$ and $S \subset \boldsymbol{P}$ with high probability. Again we start by splitting $[n]$ into consecutive bins of size $s/2$ denoted by $\boldsymbol{P}^{(1)}, \ldots, \boldsymbol{P}^{(2n/s)}$. To ease the presentation we assume $2n/s$ is an integer since the case when this is not satisfied can be handled using simple modifications. The same holds for any divisibility issue that we encounter further on. Of these bins at least k will consist entirely of signal components. Roughly speaking we think of these bins as signal components of a vector of size $2n/s$, and use a CASS procedure to find them. Once that is done, we set $\boldsymbol{P}$ as the bins deemed active and their neighboring bins, and move on to the refinement phase. In the refinement phase we estimate the active components in $\boldsymbol{P}$ for instance by using another CASS procedure.

We now describe the method in full detail. Consider the binning $\boldsymbol{P}^{(1)}, \ldots, \boldsymbol{P}^{(2n/s)}$ described before. Partition the bins into $4k$ groups denoted by $\boldsymbol{I}_1^{(1)}, \ldots, \boldsymbol{I}_{4k}^{(1)}$. For each of these we test the hypothesis

$$H_{i,0}^{(1)} : \ \boldsymbol{I}_i^{(1)} \cap S = \emptyset \qquad \text{versus} \qquad H_{i,1}^{(1)} : \ |\boldsymbol{I}_i^{(1)} \cap S| \geq s/2, \ \text{for } i \in [4k] \ .$$

The groups for which $H_{i,1}^{(1)}$ is accepted are split into two in the middle giving us the groups $\boldsymbol{I}_1^{(2)}, \ldots, \boldsymbol{I}_{2n_1}^{(2)}$. We now test a similar hypothesis as before for these new groups. Since at most $3k$ groups can contain signal components, we will specifically enforce $n_1 \leq 3k$. Iterating this, in step j we have groups denoted by

$\boldsymbol{I}_1^{(j)}, \ldots, \boldsymbol{I}_{2n_{j-1}}^{(j)}$, where $n_{j-1} \leq 3k$, and we wish to decide between

$$H_{i,0}^{(j)} : \ \boldsymbol{I}_i^{(j)} \cap S = \emptyset \qquad \text{versus} \qquad H_{i,1}^{(j)} : \ |\boldsymbol{I}_i^{(j)} \cap S| \geq s/2, \ \text{for } i \in [2n_{j-1}] \ .$$

When $j = \log_2 n/2ks$ the groups consist of single bins. The set $\boldsymbol{P}$ will consist of the ones for which $H_{i,1}^{(1)}$ is accepted in this final step and the bins adjacent to those.

To decide between $H_{i,0}^{(j)}$ and $H_{i,1}^{(j)}$, for $j \in [\log_2 \frac{n}{2s}]$ and $i \in [2n_{j-1}]$, we collect a single measurement of the form

$$Y_i^{(j)} = \langle a\sqrt{j}\mathbf{1}_{\boldsymbol{I}_i^{(j)}}, \boldsymbol{x} \rangle + W_i^{(j)}, \ \text{for } j \in [\log_2 \tfrac{n}{2s}]; i \in [2n_{j-1}] \ ,$$

where $W_i^{(j)} \sim N(0,1)$ are i.i.d., and $a > 0$. The parameter $a > 0$ needs to be chosen such that (3.3) is fulfilled. We will use half of our energy budget for the search phase. Since the groups $I_i^{(j)}$ contain $n/(2^{j+1}k)$ components for every $i \in [2n_{j-1}]$, $n_{j-1} \leq 3k$ and there are $\log_2 \frac{n}{2ks}$ steps,

$$\|A_{\text{search}}\|_F^2 = \sum_{j=1}^{\log_2 \frac{n}{2ks}} 6k \frac{n}{2^{j+1}k} j a^2 = \frac{3}{2} n a^2 \sum_{j=1}^{\log_2 \frac{n}{2s}} j 2^{-(j-1)} = 6na^2 \ .$$

Since we use at most $m/2$ energy in the search phase we get $a = \sqrt{\frac{m}{12n}}$. When group $\boldsymbol{I}_i^{(j)}$ contains a bin which is contained in S, we have $\mathbb{E}_S(Y_i^{(j)}) \geq \frac{s\mu}{2}\sqrt{\frac{jm}{12n}}$. Therefore we declare that the group contains no signal components when $Y_i^{(j)} \leq \frac{s\mu}{4}\sqrt{\frac{jm}{12n}}$, otherwise we declare the opposite. If in step $j \in [\log_2 \frac{n}{2ks}]$ we accept $H_{i,1}^{(j)}$ for more than $3k$ groups, we choose those corresponding to the highest $3k$ observations. Considering a single test the type I and type II error probabilities can both be upper bounded using (3.8) by

$$\frac{1}{2} \exp\left(-\frac{js^2 m\mu^2}{384n}\right) \ .$$

It is also possible that neither the null nor the alternative is true, and the group contains some bins that intersect with S, but are not contained in S. However we need not pay any attention to those, as by construction $\boldsymbol{P}$ will also contain neighboring bins of those we deem non-empty. The probability of either concluding $H_{i,1}^{(j)}$ when the group $\boldsymbol{I}_i^{(j)}$ contains no signal or concluding $H_{i,0}^{(j)}$ when in fact $H_{i,1}^{(j)}$

is true can be bounded from above by

$$\sum_{j=1}^{\log_2 \frac{n}{2ks}} 3k \ \exp\left(-\frac{js^2m\mu^2}{384n}\right) \ .$$

Thus whenever $\mu \geq \sqrt{\frac{384n}{s^2m} \log \frac{9k}{\varepsilon}}$ we have that

$$\mathbb{P}_S(S \not\subseteq \boldsymbol{P}) \leq \sum_{j=1}^{\log_2 \frac{n}{2ks}} 3k \left(\frac{\varepsilon}{9k}\right)^j \leq \sum_{j=1}^{\log_2 \frac{n}{2ks}} \left(\frac{\varepsilon}{3}\right)^j \leq \varepsilon/2 \ .$$

By construction we also have that $|\boldsymbol{P}| \leq \frac{9}{2}ks$. Hence in the refinement phase we can measure each component in $\boldsymbol{P}$ separately, say, to produce $\widehat{S}$. We have $2m/18ks$ energy for each of the components in $\boldsymbol{P}$, hence it is easy to check using (3.8) that the probability of making an error in the refinement phase is at most

$$\frac{9ks}{4} \exp\left(-\frac{m\mu^2}{72ks}\right) \ .$$

Whenever $\mu \geq \sqrt{\frac{72ks}{m} \log \frac{9ks}{2\varepsilon}}$ the probability above is at most $\varepsilon/2$. Thus the procedure gives an estimator $\widehat{S}$ for which $\mathbb{P}_S(\widehat{S} \neq S) \leq \varepsilon$ whenever

$$\mu \geq \sqrt{\max\{\frac{384n}{s^2m} \log \frac{9k}{\varepsilon}, \frac{72ks}{m} \log \frac{9ks}{2\varepsilon}\}} \ .$$

When considering the expected Hamming-distance as the error metric we can use the above procedure with probability of error set to $\varepsilon/2s$ in the search phase and $\varepsilon/2ks$ in the refinement phase. This method then yields a rate-optimal estimator for the support recovery problem described in Section 3.2 by collecting at most $3k\left(\log_2 \frac{n}{2ks} + \frac{3}{2}s\right)$ measurements. We emphasize that the procedure above is only optimal in terms of rates (the way μ depends on the parameters n, m, k, s), but the constant is very large. However, as mentioned in Remark 3.2, our main interest is finding the optimal rates. Furthermore, the constants of the procedure above can be improved by better algorithm choices and a more careful analysis..

Proposition 3.12. *Consider the class of k disjoint s-intervals and suppose $n > ks^3$. Then the procedure above satisfies* (3.3) *and* (3.4) *whenever*

$$\mu \geq \sqrt{\frac{768n}{s^2 m} \log \frac{3\sqrt{2}ks}{\varepsilon}} \ .$$

Furthermore, the procedure collects at most $3k\left(\log_2 \frac{n}{2ks} + \frac{3}{2}s\right)$ observations.

Remark 3.13. *As with Proposition 3.2 the condition on the sparsity is an artifact of the simple method above and can be avoided by using a more elaborate method in the refinement phase, such as binary search.*

Unions of s-stars

Consider the class of k disjoint s-stars. To ease the discussion we focus on the case $k = 1$, but the idea can be applied to larger k. The procedure is very similar to the one used for unions of s-intervals, however due to the different nature of the structure we provide a detailed description of the procedure in Appendix 3.A.

Proposition 3.13. *Consider the class of s-stars, and suppose $\sqrt{2n} \geq s^2$. Then the procedure described in the Appendix satisfies* (3.3) *and* (3.4) *whenever*

$$\mu \geq \sqrt{\frac{392n}{s^2 m} \log \frac{9s}{\varepsilon}} \ .$$

Furthermore, the procedure collects at most $2\log_2 n + 2s\log_2 \frac{\sqrt{2n}}{s}$ observations.

Similar ideas can be used to treat the case of k disjoint s-stars when $k > 1$, but $k \ll s$.

s_r, s_c-submatrices

Consider the class of submatrices of size $s_r \times s_c$ of a matrix of size $n_r \times n_c$, and suppose $s_r \geq s_c$. The procedure we present now is very similar to the one used for unions of s-intervals, hence we only provide an outline and present performance guarantees here.

Once more we break the procedure into two phases, a search phase and a refinement phase. The aim of the search phase is to find the active columns of the signal matrix, whereas the refinement phase aims to find the active rows once the

active columns are found. If we view the columns of the signal matrix as components of a vector of dimension n_c, then finding the active columns can be viewed as estimating an unstructured s_c-sparse support set. Likewise the problem of the refinement phase can be viewed as finding an s_r-set in a signal of dimension n_r. Hence we can immediately use the CASS procedure for both sub-problems with modifications similar to those used in the case of unions of s-intervals. Thus we get the following proposition:

Proposition 3.14. *Consider the class of s_r, s_c-submatrices and suppose $n_c > s_r^2/s_c$. There exists a procedure which yields an estimator satisfying* (3.3) *and* (3.4) *whenever*

$$\mu \geq \sqrt{\frac{128n}{s_r^2 m} \log \frac{2s}{\varepsilon}} \ .$$

Furthermore the estimator takes at most $2s_c \log_2 \frac{n_c}{2s_c} + 2s_r \log_2 \frac{n_r}{2s_r}$ measurements.

The sketch of the proof of Proposition 3.14 is given in Appendix 3.B.

Note that the sample complexity of the algorithm above scales $s \log n$. Each time we collect a sample, we decide if a specific row or column is active. Therefore, the computational complexity of this adaptive sensing algorithm is also $s \log n$. This is in stark contrast to the non-adaptive sensing setting, where in order to estimate a submatrix, one needs to solve an NP-hard optimization task (see Balakrishnan et al. [18]). As mentioned in Section 3.1, this is due to a fundamental difference between adaptive and non-adaptive sensing. In particular, using adaptive sensing we can shake these computational burdens by tailoring the sample to facilitate inference.

Remark 3.14. *The result above guarantees essentially the same performance as Proposition 3.5. We remark that it is possible to formulate a CASS-type algorithm whose performance would match that in Proposition 3.6, by aiming to find only one active column in the first phase. This can be done with a binary search procedure, much like the one described by Malloy & Nowak in [108].*

3.4.2 Sample Complexity lower bounds

Necessary conditions for the sample complexity of compressive sensing have been studied both in the adaptive and the non-adaptive sensing setting by Aksoylar, Atia & Saligrama in [2] and [3]. In both works, the sample complexity was studied

for the unstructured case of s-sets. For the non-adaptive setting the authors show in Theorem 4.1 of [2] that the sample complexity can be lower bounded by an expression that scales essentially like $s \log \frac{n}{s}$. Furthermore they also show that the signal-to-noise ratio plays a role in the sample complexity of compressive sensing, and this phenomenon is also explicitly captured in their bound. Though the setting considered in their work is slightly different from that in the present work, Theorem 4.1 of Aksoylar, Atia & Saligrama [2] can be translated into our setting in the following manner.

Theorem 3.8 (Theorem 4.1 of Aksoylar, Atia & Saligrama [2])**.** *Consider the class of s-sets, and suppose there exists a non-adaptive estimator for which* (3.3) *holds and for which $\frac{1}{|\mathcal{C}|}\sum_{S\in\mathcal{C}} \mathbb{P}_S(\widehat{S} \neq S)$ is not asymptotically bounded away from zero as $n, s \to \infty$. Let $k(n,s)$ denote the number of measurements the estimator makes. Then*

$$k(n,s) \geq \frac{cs \log \frac{n}{s}}{\log\left(\mu^2 \frac{m}{n} + 1\right)} \ ,$$

with some constant c.

Remark 3.15. *For the proof the reader is referred to the paper above. However, we can show a similar, but slightly weaker result using a simple reasoning. That result is provided in Appendix 3.C for the sake of completeness, and shows that for very sparse signals, the sample complexity must scale as $s \log n$.*

This shows that the procedure presented in the previous section for s-sets performs as well in terms of sample complexity as the best non-adaptive procedure. Furthermore, when estimating structured support sets, potentially fewer samples are enough to perform accurate estimation. We now briefly discuss necessary conditions on sample complexity for non-adaptive estimators for the structured classes we have examined before.

Consider the case of unions of k disjoint s-intervals first. Without giving a rigorous formal proof we argue that the number of samples required in the non-adaptive case must scale as $k \log \frac{n}{sk}$. Let $S_1, \ldots, S_{n/s}$ be consecutive disjoint s-intervals of $[n]$ and let

$$\mathcal{C}' = \left\{ S \in \mathcal{C} : \ S = \bigcup_{j=1}^{k} S_{i_j}, \ i_1, \ldots, i_k \in [n/s] \right\} \ ,$$

that is unions of intervals that are constructed from $S_1, \ldots, S_{n/s}$. This class roughly

behaves like a class of k-sparse sets of a vector of dimension n/s, except that there is an increase in the relative sensing power arising from the fact that the building blocks of the class are s-sets instead of singletons. This results in that it is possible to detect somewhat weaker signals (see Theorem 3.2), but because of the weak dependence of the sample complexity bound of Theorem 3.8 on the signal-to-noise ratio, the scaling of the bound will still be dictated by the numerator.

The class of unions of k disjoint s-stars is even simpler to consider. Suppose $k = 1$, and that the center of the star is given by an oracle. The remaining problem is the estimation of an s-sparse set in a vector with dimension roughly $\sqrt{2n}$. Hence the sample complexity remains essentially the same as that of the unstructured case.

Finally for the class of s_r, s_c-submatrices, if an oracle provides the active columns, the problem reduces to the unions of intervals case.

This shows that the procedures presented in the previous section for structured support recovery perform as well in terms of sample complexity as the best non-adaptive procedures. However, it is plausible that adaptive procedures might outperform non-adaptive ones in terms of sample complexity. This question was investigated by Aksoylar & Saligrama in [3], where the authors provide a necessary condition for any adaptive algorithm to recover unstructured s-sets. The number of samples required depends on the signal-to-noise ratio in this case as well. Their results show that when the signal-to-noise ratio is near the boundary where accurate estimation is possible (see Theorem 3.5, and Castro [40]) the number of samples needs to scale essentially like s. It is still an open question whether this bound is achievable or not.

Although we do not yet have a rigorous proof, we conjecture that although some performance gain might be present, it is not substantial and the number of samples needs to scale essentially like $s \log \frac{n}{s}$ for adaptive estimators as well, when the signal magnitude is close to the estimation threshold. The reason behind this conjecture is roughly the following. Consider the 1-sparse case. It can be easily seen that by taking one measurement, a fraction of the n hypotheses (namely that the signal component is at coordinate $1, \ldots, n$) remains essentially indistinguishable. Focusing the next measurement on these potential signal components, again a fraction of them will remain essentially indistinguishable. With a bit of work this line of reasoning will, in principle, provide a lower bound on the sample complexity. However, formalizing this argument is challenging, because each projection does contain

some faint amount of information about these "indistinguishable" hypotheses. So one needs to show that these small amounts of information are negligible as a whole, even after collecting multiple projections. Showing this requires the proof of a sharp information-contraction bound suitable for the adaptive sensing setting. Nonetheless, the authors conjecture that because of this heuristic, a term that is logarithmic in the dimension should also be present in the sample complexity lower bounds. Foucart & Rauhut consider a different compressed sensing setting and framework in [72]. Although this setting is not directly comparable to ours, the authors show that adaptive sensing does not further reduce the sample complexity, which also leads us to believe that our conjecture is reasonable.

3.5 A Numerical Experiment

Akin to Chapter 2, we present a short numerical experiment to corroborate the theoretical results presented in this chapter. Again, the simulations here only serve an illustrative purpose.

In this simulation we gauge the performance of the adaptive sensing procedures presented in Sections 3.3.1 and 3.4.1 for the class of s-intervals, and compare it with a reasonable non-adaptive procedure[1].

For the non-adaptive procedure we randomly generate a sensing matrix of size $2 \log n \times n^2$, whose entries are independent standard normal random variables. We then re-scale the matrix so that its squared Frobenius-norm is equal to m. We use this matrix to sample the signal, as described in (3.2). Finally, we choose the s-interval that is the most closely aligned to our observations (in Euclidean sense) as our estimator. The performance of the estimator is measured in terms of Hamming distance.

The adaptive sensing procedures presented in this chapter fall into two categories: procedures based on the CASS algorithm (see Section 3.4.1) and those based on the SLRT (see Section 3.3.1). The CASS-based procedure for the detection of intervals is implemented the same way as it was described in Section 3.4.1. Note that the CASS procedure uses at most m total sensing energy by design, so it is fair to compare it with the non-adaptive procedure above.

[1] Although we do not make a formal claim that the non-adaptive procedure implemented is indeed optimal, it is likely asymptotically optimal, as it is simply a maximum-likelihood estimator.

[2] As shown in Section 3.4.2, we need at the order of $\log n$ projections for reliable support recovery. Increasing the number of rows of the sensing matrix further did not seem to have a considerable effect on the performance of the procedure.

The SLRT-based procedures use a random amount of energy, so a direct comparison with the procedures above is somewhat unfair. However, we deal with this in the same manner as in Section 2.6: when the energy budget is exhausted, we terminate the procedure and incur a loss of $2s$ in terms of Hamming distance. Other than that, we choose every parameter as described in Section 3.3.1, set $\varepsilon = 0.05$ and $a = 0.2$. The latter choice has the same motivation as the choice of Γ in Section 2.6, namely to see if such a choice has an effect on the practical performance.

Computations in Section 3.3.1 (most noatbly Lemma 3.2) suggest that the adaptive sensing algorithm based on the SLRT should satisfy $\mathbb{E}(|S \triangle \widehat{S}|) \leq \varepsilon$ when the signal strength is roughly

$$\mu_{\text{limit}} = \sqrt{\frac{8n}{s^2 m} \log \frac{4s}{\varepsilon}} \, . \tag{3.9}$$

Furthermore, according to the lower bounds of Theorem 3.6, no estimator can have small probability of error unless the signal scales as above.

The same comments apply to the calibration of the SLRT-based algorithm as in Section 2.6. In a nutshell, to implement the SLRTs we need to specify a value for μ_{limit} that we use for the alternative hypothesis in the likelihood-ratio. We emphasize once more, that this is an internal parameter of the SLRT, and does not mean that we use knowledge of the true signal strength. As in Section 2.6, we again use two versions of the SLRT-based procedure: one where we set μ_{limit} as defined above, and one where we use $\mu_{\text{limit}}^{(.95)}$, which is defined by the same expression as in (3.9), only we replace m by $0.95 \cdot m$. This way we tune the procedure to detect a signal that is slightly larger then μ_{limit} (roughly by a factor of 1.02), but in turn this will result in a slightly smaller amount of sensing energy used by the SLRTs (in expectation).

We run a similar experiment as in Section 2.6. We run the procedures described above when the true signal strength is $t \cdot \mu_{\text{limit}}$ with the value of t varying. The signal dimension is $n = 2^{15}$, the support size is $s = 2^4$ and the energy budget is $m = n$. We run 100 iterations for every value of the parameter t, and plot the average normalized Hamming distance of the different estimators. We also plot error bars whose total length is four times the (point-wise) standard error, which would correspond to a roughly 95% two-sided confidence interval for normally distributed measurements. Note that the error bars are only approximate point-wise confidence bands, that are included to provide some insight about the variability of the curves.

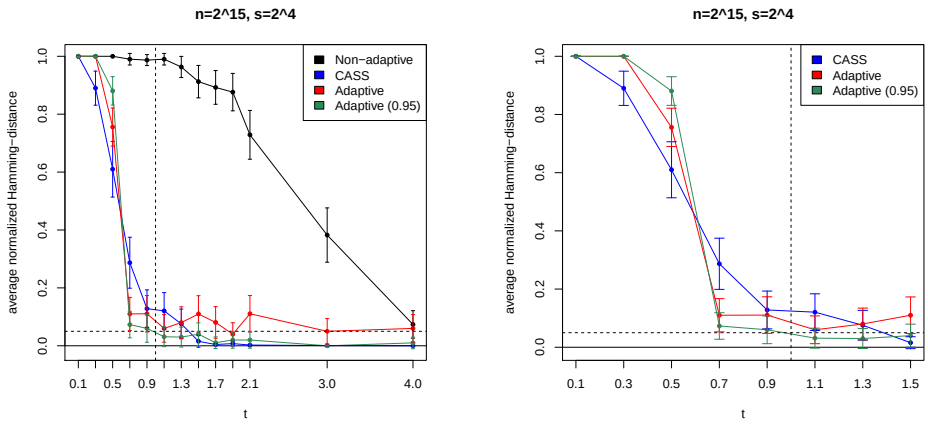

Figure 3.1: Average normalized Hamming-distance (with SE bands) for the different estimators as a function of the parameter t (the signal strength is $t \cdot \mu_{\text{limit}}$ with μ_{limit} defined in (3.9)): the non-adaptive estimator (black); the CASS-based procedure (blue); the SLRT-based procedure calibrated with μ_{limit} (red); the SLRT-based procedure calibrated with $\mu_{\text{limit}}^{(.95)}$ (green). The number of repetitions is 100 for each value of t. The vertical black dashed line is at the value $t = 1$. The horizontal black dashed line is at the value of ε (0.05).

We provide a plot on a wider range of the parameter t to be able to compare the non-adaptive estimator to the adaptive ones, and a zoom-in of the previous plot around $t = 1$ to be able to compare the adaptive sensing estimators. As expected, the adaptive sensing procedures outperform the non-adaptive one. We also expect the SLRT-based procedures to reach the level $\varepsilon = 0.05$ at $t = 1$, and the CASS-based procedure to reach this level somewhat later.

The same comments apply to the performance of the SLRT-based procedures as in Section 2.6. Note that the CASS-based procedure performs comparably to the SLRT-based estimators. This illustrates that the constants resulting from the crude analysis of Section 3.4.1 are indeed very loose.

3.6 Final remarks

In this chapter we have examined the problem of recovering structured support sets through adaptive compressive measurements. We have seen that by adaptively

designing the sensing matrix it is possible to achieve performance gains over non-adaptive protocols, and that the gains can be quite dramatic, for instance in the case of s-stars. We have also seen that these gains can be realized by simple and practically feasible estimation procedures.

However, a complete characterization of the problem for the class of submatrices is still missing. This could prove to be an interesting area for future research considering the practical relevance of that model in gene expression studies. Furthermore, it remains unclear if the sample complexity of support recovery using compressive measurements can be significantly reduced by adaptively designing the rows of the sensing matrix. Finally, the procedures of Section 3.4.1 can be modified using ideas of Arias-Castro presented in [5] to be able to handle signals with arbitrary signs and magnitudes. Working out the details could prove to be a useful extension to this work.

3.A Description of the procedure of Section 3.4.1

We begin with a search phase to find the approximate location of the support. Again we consider the subsets $\boldsymbol{P}^{(i)},\ i = 1, \ldots, p$, where $\boldsymbol{P}^{(i)}$ contains all the components whose corresponding edges lie on the vertex v_i. Our goal is to find the center of the star. We begin by forming 4 groups $\boldsymbol{I}_1^{(1)}, \ldots, \boldsymbol{I}_4^{(1)}$, where each of them is a union of $p/4$ different $\boldsymbol{P}^{(i)}$, and no subset $\boldsymbol{P}^{(i)}$ is contained in more than one group. We then take one measurement per group

$$Y_i^{(1)} = \langle a\mathbf{1}_{\boldsymbol{I}_i^{(1)}}, \boldsymbol{x}\rangle + W_i^{(1)},\ \text{for } i \in [4],$$

where $W_i^{(1)}$ are i.i.d. standard normals and $a > 0$. Large measurements should correspond to groups containing a lot of signal components, and particularly the one containing the center of the star. However, because of the structure of the support and the fact that these groups are not disjoint, large observations may also correspond to groups not containing the center of the star. Therefore instead of performing hypothesis tests we choose the two highest observations, and consider the groups corresponding to those. Once we have these groups, we split each in half in the sense that half of the $\boldsymbol{P}^{(i)}$ in a given group will form one new group, and the other half will form another new group. This way we end up with 4 groups, again not disjoint, and do the same as before. Let the groups in step j be denoted

by $\boldsymbol{I}_1^{(j)}, \ldots, \boldsymbol{I}_4^{(j)}$. The measurements we collect are

$$Y_i^{(j)} = \langle a\sqrt{j}\mathbf{1}_{\boldsymbol{I}_i^{(j)}}, \boldsymbol{x}\rangle + W_i^{(j)}, \text{ for } j \in [\log_2 \tfrac{p}{4}] \text{ and } i \in [4] \ .$$

In the final step $j = \log_2 \frac{p}{4}$ each group consists of a single $\boldsymbol{P}^{(i)}$. The output set of the search phase $\boldsymbol{P}$ will consist of the union of those two groups for which the final observation is largest.

First we specify the parameter a so as to ensure that we do not use more than half of our measurement budget. Each $\boldsymbol{I}_i^{(j)}$ contains at most $(p-1)\frac{p}{2^{j+1}} = n/2^j$ components $i \in [4]$, and $j \in [\log_2 \frac{p}{4}]$. Recalling that $n = \binom{p}{2}$, we have

$$\|A_{\text{search}}\|_F^2 \leq \sum_{j=1}^{\log_2 \frac{p}{4}} \frac{n}{2^{j-2}} j a^2 \leq 8na^2 \ .$$

Therefore $a = \sqrt{\frac{m}{16n}}$ ensures we use at most $m/2$ energy in the search phase.

Now we need to show that $S \subset \boldsymbol{P}$ with high probability. Without loss of generality suppose that $\boldsymbol{I}_1^{(j)}, \ldots, \boldsymbol{I}_4^{(j)}$ are indexed such that the center of the star is in group $\boldsymbol{I}_1^{(j)}$, and for the number of signal components in $\boldsymbol{I}_i^{(j)}$ denoted by $N_i^{(j)}$ we have $N_i^{(j)} \geq N_{i+1}^{(j)}$. Hence $\boldsymbol{I}_1^{(j)}$ contains exactly s components, and because $\sum_{i=2}^{4} N_i^{(j)} \leq s$ we know that $N_3^{(j)} \leq s/2$. Using this we conclude that in each step $j \in [\log_2 \frac{p}{4}]$ the probability that $Y_1^{(j)} < \max\{Y_3^{(j)}, Y_4^{(j)}\}$ can be bounded from above using (3.8) by

$$3 \cdot \frac{1}{2} \exp\left(-\frac{j s^2 m \mu^2}{392n}\right) \ .$$

From this we get that whenever $\mu \geq \sqrt{\frac{392n}{s^2 m} \log \frac{9}{2\varepsilon}}$ we have

$$\mathbb{P}_S(S \not\subseteq \boldsymbol{P}) \leq \sum_{j=1}^{\log_2 \frac{p}{4}} \left(\frac{\varepsilon}{3}\right)^j \leq \varepsilon/2 \ .$$

By construction we make $4 \log_2 \frac{p}{4}$ observations in this phase, and also $|\boldsymbol{P}| \leq 2(p-1)$.

In the search phase we can directly apply the CASS procedure on $\boldsymbol{P}$ to estimate the support. Since $\sqrt{2n} > p-1$ we know that whenever $\mu \geq \sqrt{\frac{64\sqrt{2n}}{m} \log \frac{4s}{\varepsilon}}$ the probability of error is at most $\varepsilon/2$, and we take at most $2s \log_2 \frac{p-1}{s}$ measurements. When considering $\mathbb{E}_S(|\widehat{S} \triangle S|)$ as the error metric one can set the probability of

error to $\varepsilon/2s$ and use the procedure above.

3.B Sketch proof of Proposition 3.14

Recall that the signal matrix is of size $n = n_r \cdot n_c$ and the support is a submatrix of size $s = s_r \cdot s_c$ with $s_r \geq s_c$.

We use half the energy for the search phase, and half for the refinement phase. In step j of the search phase the groups $\boldsymbol{I}^{(j)}$ contain $n/2^j s_c$ components and there are at most $2s_c$ components. Hence the energy used is at most

$$\sum_{j=1}^{\log_2 \frac{n_c}{2s_c}} 2s_c \frac{n}{2^j s_c} j a^2 = 4na^2 \ .$$

Thus $a = \sqrt{\frac{m}{8n}}$. This means that the probability of error is bounded by

$$\sum_{j=1}^{\log_2 \frac{n_c}{2s_c}} 2s_c \frac{1}{2} \exp\left(-\frac{s_r^2 j m \mu^2}{64n}\right) \ ,$$

sing (3.8). Thus, whenever $\mu \geq \sqrt{\frac{64n}{s_r^2 m} \log \frac{2s_c}{\varepsilon}}$ the probability of error is at most $\varepsilon/2$.

In the refinement phase the energy used is

$$\sum_{j=1}^{\log_2 \frac{n_r}{2s_r}} 2s_r \frac{n_r s_c}{2^j s_r} j a^2 = 4n_r s_c a^2 \ ,$$

hence $a = \sqrt{\frac{m}{8n_r s_c}}$. Therefore, using the same bound as before, the probability of error is at most

$$\sum_{j=1}^{\log_2 \frac{n_r}{2s_r}} 2s_r \frac{1}{2} \exp\left(-\frac{s_c j m \mu^2}{64n_r}\right) \ ,$$

which means that whenever $\mu \geq \sqrt{\frac{64n_r}{s_c m} \log \frac{2s_r}{\varepsilon}}$ the probability of error is at most $\varepsilon/2$.

Considering the expected Hamming-distance as the error metric, we can use the procedure above with probability of error set to $\varepsilon/2s$.

3.C Sample complexity lower bound for non-adaptive compressive sensing

Consider the 1-sparse case. Suppose a measurement strategy satisfying $\|A\|_F^2 \leq m$ identifies the signal support correctly, where A has r rows. If component $i \in [n]$ is active then the measurement $Y \in \mathbb{R}^r$ is distributed as $N(\mu A_i, I)$, where A_i is the ith column of A and I is the identity matrix. Recovering the support can be viewed as a multiple testing problem, namely we want to decide between $H_j : \boldsymbol{x}_j = \mu,\ j \in [n]$.

Now suppose that there is a column A_j such that $\|\mu A_i - \mu A_j\|_2^2 \leq 1$. If this were the case, we would have

$$\max\left\{\mathbb{P}_i(\widehat{S}=j), \mathbb{P}_j(\widehat{S}=i)\right\} \leq 1 - TV(\mathbb{P}_i, \mathbb{P}_j) \leq \frac{1}{2}\exp\left(-KL(\mathbb{P}_i, \mathbb{P}_j)\right) \ , \quad (3.10)$$

where the first inequality follows from Theorem 2.2 in Tsybakov [139], and the second inequality follows from Lemma 2.6 in Tsybakov [139]. Plugging in the definition of the Kullback-Leibler divergence, we continue as

$$\begin{aligned}
&\max\left\{\mathbb{P}_i(\widehat{S}=j), \mathbb{P}_j(\widehat{S}=i)\right\} \\
&\quad\leq \frac{1}{2}\exp\left(-\sum_{k=1}^{r}\mathbb{E}_i\left(\mu Y_k(A_{ik}-A_{jk}) - \frac{\mu^2}{2}(A_{ik}^2 - A_{jk}^2)\right)\right) \\
&\quad= \frac{1}{2}\exp\left(-\frac{\mu^2}{2}\|A_i - A_j\|_2^2\right) \\
&\quad\leq \frac{1}{2\sqrt{e}} \ , \qquad (3.11)
\end{aligned}$$

and so we could not have $\mathbb{P}_S(\widehat{S} \neq S) \leq \varepsilon$ for all S with $\varepsilon < 1/(2\sqrt{e})$. This means that the minimum squared Euclidean distance between the vectors $\mu A_j,\ j \in [n]$ has to be at least 1. On the other hand note that since we are considering maximal probability of error, similar arguments as in the proofs in Section 3.3.2 show that the Euclidean norm of the columns of the sensing matrix A should be identical, that is $\|A_j\|_2^2 = m/n\ \forall j \in [n]$.

This leads to the following question: at most how many unit balls can we pack into a ball with radius $\mu\sqrt{m/m}+1$ in r dimensions? Unless we are able to pack n such balls, the argument above shows that we cannot have an estimator satisfying $\max_{S\in\mathcal{C}}\mathbb{P}_S(\widehat{S}\neq S) \leq \varepsilon$ with a sensing matrix that has r rows. But the number of balls we can pack can simply be upper bounded by the ratio of the volume of the

balls and the total volume, hence we get

$$(\mu\sqrt{m/m}+1)^r < n \;\Rightarrow\; \max_{S\in\mathcal{C}} \mathbb{P}_S(\widehat{S}\neq S) > \varepsilon \;,$$

for $\varepsilon < 1/(2\sqrt{e})$, showing that we need $r \geq \log n/\log(\mu\sqrt{m/n}+1)$ measurements to recover a 1-sparse support.

We can follow the same arguments for the s-sparse case to establish

$$r \geq \log \binom{n}{s} / \log(s\mu\sqrt{m/n}+1) \;.$$

These lower bounds are somewhat weaker than the one referenced in Section 3.4.2, but are simple to obtain and give comparable results for very sparse signals.

Chapter 4

Detection of signals evolving in time

This chapter is based on joint ongoing work with Rui Castro. Because of this, some results presented in this chapter still lack rigorous proofs. In such cases the author aimed to provide heuristic arguments to show what kind of statements can be expected to hold, and wherever possible highlight the specific steps in the proofs that still require further justification.

4.1 Introduction

As outlined in Chapter 1 the detection of sparse signals is a problem that has been studied with great attention in the past. The usual setting of this problem involves a (potentially) very large number of items, of which a (typically) much smaller number *may* be exhibiting anomalous behavior. A natural question one can ask is whether it is possible to reliably detect whether there are indeed some items showing anomalous behavior?

Questions like this are encountered in a number of fields of research. Some examples include epidemiology where one wishes to quickly detect an outbreak or the environmental risk factors of a disease (see Neill & Moore [114], Kulldorff et al. [100, 86, 101]), identifying changes between multiple images (see Flenner & Hewer [71]), and microarray data studies (see Pawitan et al. [119]) to name a few.

A common point in the examples above is that even though it is not known

which items are anomalous, their identity remains fixed throughout the sampling process. However in certain situations the identity of these items may change over time. For instance consider a signal intelligence setting where one wishes to detect covert communications. Suppose that our task is to survey a signal spectrum, a small fraction of which may be used for communication, meaning that some frequencies would exhibit increased power. On one hand we do not know beforehand which frequencies are used, but also the other parties may change the frequencies they communicate through over time. This introduces a further hindrance in our ability to detect whether someone is using the surveyed signal spectrum for covert communications.

As mentioned in Chapter 1, other motivating examples for such a problem include spectrum scanning in a cognitive radio system (see Li [103], Caromi, Xin & Lai [37]), detection of hot spots of a rapidly spreading disease (see Shah & Zaman [128], Zhu & Ying [153], Luo & Tay [104], Wang et al. [144]), detection of momentary astronomical events (see Thompson et al. [138]) or intrusions into computer systems (see Gwadera, Atallah & Szpankowski [76], Phoha [121]). The main question that we aim to answer in this chapter is how the dynamical aspects of the signal affect the difficulty of the detection problem.

In the more classical framework of the signal detection problem, inference is based on observations that are collected in a non-adaptive manner. However, dealing with time-dependent signals naturally leads to a setting where measurements can be obtained in a sequential and adaptive manner, using information gleaned in the past to guide subsequent sensing actions. Furthermore, in certain situations it is impossible to monitor the entire system at once, but instead one can only partially observe the system at any given time. For instance in the above signal intelligence example, we can consider a situation where we are only able to monitor a certain frequency band at a given time, but the decision which frequencies to monitor can depend on our past observations.

Such adaptive sensing procedures can, in certain situations, outperform non-adaptive ones in signal detection and support recovery tasks, as seen in the previous two chapters. Hence a further goal is to understand the differences between adaptive and non-adaptive sensing procedures when used for detecting dynamically evolving signals, in particular in situations where the system can only be partially monitored.

In this chapter we introduce a simple framework for studying the detection prob-

lem of time-evolving signals. Our signal of interest is modeled as an n-dimensional vector. We take a hypothesis testing point of view. Under the null the signal is the zero vector, while under the alternative the signal is an s-sparse vector. At each time step $t \in \mathbb{N}$ we flip a biased coin independently for each anomalous component. The coin comes up tails with probability p, and if this happens, that component will be in a different unoccupied location of the signal vector in the next time step. Thus the parameter p encodes the speed of change of the signal support in some sense. At each time step we are allowed to select one component of the signal to observe through additive standard normal noise, and we are allowed to collect up to m measurements. Our goal is to decide whether the signal is zero or not, based on the collected observations.

We present an algorithm that solves the above task, and show its near-optimality by deriving the fundamental limits of the hypothesis testing problem above. We do this in both the adaptive sensing and non-adaptive sensing settings. It is easy to see that the above problem can not be solved reliably unless we are allowed to collect on the order of n/s measurements. When the number of measurements is of this order, we can reliably detect the presence of the signal when the minimum non-zero component scales like $\sqrt{p \log(n/s)}$ in the adaptive sensing setting. In the non-adaptive sensing setting detection is possible when the smallest non-zero component scales like $\sqrt{\log n}$. Hence, under the adaptive sensing paradigm the speed of change influences the difficulty of the detection problem, with slowly changing signals being easier to detect. Contrasting this, in the non-adaptive sensing setting, the speed of change has no effect of the problem difficulty, and in particular detection is as hard as if we were given a different s-sparse signal at each time step.

Related work: The setting where the identity of the anomalous items is fixed over time has been widely studied in the literature. Classically this problem has been studied in the non-adaptive sensing setting. In this context both the fundamental difficulties of the detection problem and the optimal tests are well understood, see for instance Ingster & Suslina [87, 88], Baraud [21], Donoho & Jin [64] and the references therein. The same problem has been investigated in the adaptive sensing setting as well. In [81], Haupt, Castro & Nowak provide an efficient adaptive sensing algorithm for identifying a few anomalous items among a large number of items. Malloy & Nowak provide algorithms for sequential testing in a similar setting, but for general distributions in [109, 106]. The algorithms outlined

in these works can in principle also be used to solve the detection problem, that is where only the presence or absence of anomalous items needs to be decided. Malloy & Nowak also provide bounds on the fundamental difficulty of the estimation problem in the adaptive sensing setting in [105], whereas Castro provides similar results for the detection problem as well in [40].

Inference problems concerning time-dependent signals have been investigated in various settings. The papers referenced below have varying degrees of connection to the problem we are considering, but we remark that we were only able to find a few instances that closely match our setting.

A closely related topic to the above is that of target search. Here, the goal is to identify anomalous processes under the constraint that we can only observe a limited number of them. These works consider a setting inspired by Chernoff [46], which deals with optimal experimental design for testing binary hypotheses, which was extended to multi-hypothesis testing by Bessler in [26]. Kadane introduced a problem called whereabouts search in [94], which was recently revisited by Zhai & Zhao [148]. Here, the aim is to find a target which is in one of several possible locations. In this setting searching at a specific location results in binary observations regarding the presence or absence of the target. Castanon [38] and Cohen & Zhao [48] consider a version of the previous problem dealing with more general observations instead of binary ones. In particular Castanon [38] considers a setting where the target process is a mean-shifted version of the baseline process, whereas Cohen & Zhao [48] deal with arbitrary distributions.

A set of related work is concerned with the spectrum scanning of multichannel cognitive radio systems. Here the aim is to quickly and accurately determine the availability of each spectrum band of a multi-band system. Alternatively one might only aim to quickly find a single band that is available. Efficient algorithms for the spectrum scanning problem are provided by Li in [103] and Caromi, Xin & Lai in [37]. A very similar problem is intrusion detection in cyber systems, investigated by Cohen, Zhao & Swami in [50, 51, 49], where the target processes correspond to unauthorized accesses or frauds in the system. We are only able to partially monitor the system, that is we can only observe a subset of the processes. Each anomalous process incurs a cost per unit time until the process is identified and fixed, thus the aim is to develop methods that detect the anomalous processes as quickly as possible in order to minimize the incurred cost.

Identification of anomalous processes among a large number of processes is

a question also encountered in the monitoring of multi-channel systems. This problem was introduced by Zigangirov in [154] and later revisited by Klimko & Yackel [95] and Dragalin [67]. In this setting each channel of a multi-channel system contains a Wiener process, a few of which are anomalous and have a deterministic drift. The observer is allowed to monitor one channel at a time with the goal to localize the anomalous channels as quickly as possible. Stone & Stanshine [136] consider a similar setting, but with an additional constraint that once the observer decides to monitor a different channel the decision about the currently monitored channel has to be made.

A more closely related problem to the one considered in this chapter is detecting the first disorder of a system involving multiple processes. In this problem, multiple sensors take observations sequentially and relay them to a so-called fusion center that determines whether the system is behaving normally or not. After some unknown time a change occurs in the statistical behavior of the observations collected by the sensors. This time can be different for different sensors, and some sensors might not exhibit any change in the behavior. The goal is to detect the first time change that occurs in any of the sensors as quickly as possible. Hadjiliadis, Zhang & Poor [77] examines one-shot schemes in the previous setup, that is, when sensors only communicate with the fusion center in case they signal an alarm. A Bayesian version of the problem was investigated by Raghavan & Veeravalli in [125]. Bayraktar & Lai [23] deal with a version of the above problem where only one of the sensors is compromised.

The detection of a change point is also a topic examined by Zhao & Ye [151]. In this work the observer is faced with multiple processes, each of which alternates between two states (called ON and OFF), but the change points of the processes are unknown. Only one process can be observed at any given time, and the goal is to catch any one process in an ON state. This problem is very similar to the one we are considering in this chapter, the main difference being that we aim to answer a different type of question, namely detecting whether there is any process in the ON state.

Organization: This chapter is organized as follows. Section 4.2 introduces the problem setup, including the signal and observation models and the inference goals. In Section 4.3 we introduce an adaptive sensing algorithm and analyze its performance. Section 4.4 is dedicated to the characterization of the difficulty of the

detection of dynamically evolving signals. In particular we show that the algorithm presented in Section 4.3 is near-optimal, and examine the difference between adaptive and non-adaptive sensing procedures. Concluding remarks and avenues for future research are provided in Section 4.6.

4.2 Problem Setup

For every $t \in [m]$, let $\boldsymbol{x}^{(t)} \in \mathbb{R}^n$ be an unknown signal of the following form

$$\boldsymbol{x}_i^{(t)} = \begin{cases} \mu & i \in S^{(t)}, \\ 0 & i \notin S^{(t)}, \end{cases}$$

where $S^{(t)} \subset [n]$ is the support of the signal at time t, and $m \in \mathbb{N}$ is our time horizon. We refer to $\{\boldsymbol{x}^{(t)}\}_{t\in[m]}$ as the *signal* and $\{\boldsymbol{x}_i^{(t)}\}_{i\in S^{(t)}}$ as the *active components* of the signal. We can collect at most m measurements of the form

$$Y_t = \boldsymbol{x}_{A_t}^{(t)} + W_t, \quad W_t \overset{iid}{\sim} N(0,1),\ t \in [m]\ , \tag{4.1}$$

where $A_t \in [n]$ is the index of the entry of the signal that we measure at time t. In an adaptive sensing scenario $\{A_t\}_{t\in\mathbb{N}}$ is a (possibly random) function of the past $\{Y_j, A_j\}_{j\in[t-1]}$. In a non-adaptive sensing scenario $\{A_t\}_{t\in\mathbb{N}}$ needs to be generated before any observations are made, or formally A_t is independent from $\{Y_j, A_j\}_{j\in[t-1]}$ for all $t \in [m]$.

This measurement model is closely related to that of Chapter 2. In particular, setting the precision of the measurements $\Gamma_t = 1$ and removing the expectation from the budget constraint in Chapter 2 leads to the model above. It is natural to assume that the precision is related to the amount of time we have for an observation, which is the reason behind the first modification. Removing the expectation from the budget constraint is done for more technical reasons, though we remark that in the previous chapters this modification did not change the results qualitatively (see Appendix 2.A), and we do not have any reason to believe that this is any different in the current setting.

Although the fact that the active components of $\boldsymbol{x}^{(t)}$ have the same value might seem overly restrictive at first glance, note that the same comment applies as in the previous chapters. That is, when the active components have different signs and magnitudes, essentially the same arguments hold throughout the chapter with μ

playing the role of the minimum absolute value of the active components. Nonetheless, we use this simple model to make discussions and results more transparent.

We consider a simple stochastic model for the evolution of the signal. In a nutshell, a support of size s is chosen uniformly at random in the first step. In subsequent steps we toss a coin which comes up tails with probability p for every active component. We then remove components for which the coin came up tails, and place them back again to a position chosen uniformly at random from available locations. Thus when $p = 1$ the signal has a new support drawn uniformly at random from all sets of size s at each time $t \in [m]$, whereas in case $p = 0$ the support is chosen randomly at the beginning and stays the same throughout the process. In general, the parameter p can be interpreted as the speed of change of the support, with larger values corresponding to faster change.

Formally, let $S^{(1)}$ be chosen uniformly at random from $\{S \subset [n] : |S| = s\}$ with some fixed $s \in [n]$. Let $\theta_i^{(t)} \sim \mathrm{Ber}(p)$ be independent for every $i \in [s],\ t \in [m]$. For a fixed $t \in [m]$, enumerate the elements of $S^{(t)}$ as $\{S_i^{(t)}\}_{i \in [s]}$. If $\theta_i^{(t)} = 0$ component $S_i^{(t)}$ will also be included in $S^{(t+1)}$, otherwise it will move. The support set $S^{(t+1)}$ is chosen uniformly at random from the set

$$\left\{ S \subset [n] :\ |S| = s,\ S \cap S^{(t)} = \{S_i^{(t)} : \theta_i^{(t)} = 0\} \right\}.$$

To help visualize the evolution of the support, we provide some simulated results in Figure 4.1.

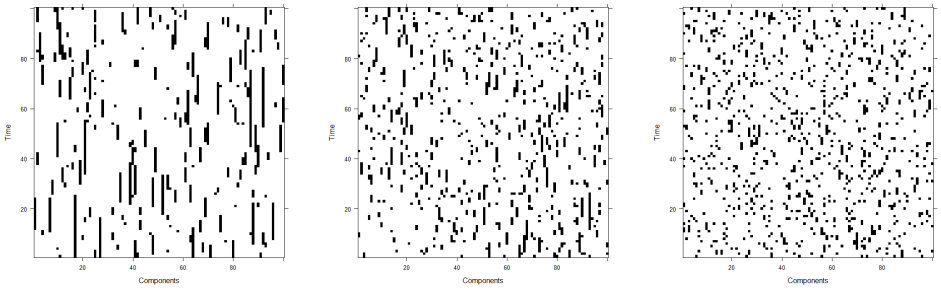

Figure 4.1: Indicator vectors of the support through time, with $p = 0,2$, $p = 0,5$ and $p = 0,8$ respectively from left to right. The signal dimension is $n = 100$, the support size is $s = 10$ and the simulations are run for 100 time steps. Time runs from bottom to top on the vertical axis, and components in the support are colored with black.

4.2.1 Inference goals

Our goal is to decide between two hypotheses. Under the null hypothesis there is no signal present, that is $S^{(t)} = \emptyset$ for every $t \in \mathbb{N}$. Under the alternative hypothesis there is a signal support evolving according to the model described above. Let $\Psi : \{A_t, Y_t\}_{t\in[m]} \to \{0, 1\}$ be a test function. We evaluate the performance of a test Ψ in terms of the maximum of the type I and type II error probabilities. We require

$$\max_{i=0,1} \mathbb{P}_i(\Psi \neq i) \leq \varepsilon \ , \tag{4.2}$$

with some fixed $\varepsilon \in (0, 1/2)$.

Note that, in contrast with Chapters 2 and 3, the alternative hypothesis is simple in the current setup. In particular, the type II error probability can be expressed by the law of total probability as

$$\mathbb{P}_1(\Psi = 0) := \mathbb{E}_1 \left(\mathbb{P} \left(\Psi = 0 \,\middle|\, \{S^{(t)}\}_{t\in[m]} \right) \right) \ . \tag{4.3}$$

Our goal is to understand how the signal strength μ needs to scale in terms of the parameters n, s, m, p and ε so that the detection problem can be solved such that (4.2) is satisfied. To this end we first propose an algorithm and evaluate its performance in Section 4.3. Then in Section 4.4 we prove the near optimality of the previous algorithm by proving necessary conditions that μ needs to satisfy so that the detection task can be solved reliably.

In the subsequent sections we will see that there is a complex interplay between the parameters n, s, m and p in how they affect the minimum signal strength required for reliable detection. Even when we restrict ourselves to the case $p = 1$, the nature of the optimal test changes radically depending on the interplay between the remaining parameters. In this case, the signal support is reset at every time $t \in [m]$, which means that regardless of the sampling strategy (the choice of A_t) we are in the situation akin to a so-called sparse mixture model.

These models are now well understood (see Ingster & Suslina [87, 88], Baraud [21], Donoho & Jin [64] and the references therein). We know that in the case of mixture models, for very sparse signals a type of scan test (which is essentially a generalized likelihood-ratio test) performs optimally, whereas for less sparse signals a global test based on the sum of all the observations is the optimal one. In our case the interplay between the parameters n, s and m determines the level of sparsity of the sample under the alternative. This in turn means that in our situation the

optimal test, and the scaling required for μ, depends on the relation between m and n/s. For more details, see also Remark 4.1.

The above phenomenon becomes even more complex with the introduction of the parameter p. Note however, that unless m is at least on the order of n/s, reliable detection is impossible. The reason behind this is that no sampling strategy will sample an active component under the alternative in fewer measurements with sufficiently large probability (this heuristic will be made rigorous in Section 4.4).

Consider the case when $p = 0$ and suppose there is no observation noise. Let the sampling strategy be arbitrary and let Ω denote the event that the algorithm does not sample an active component. When $m \leq n/s$ we have

$$\begin{aligned}\mathbb{P}_1(\Omega) &\geq \frac{\binom{n-s}{m}}{\binom{n}{m}} = \frac{(n-s)(n-s-1)\dots(n-s-m+1)}{n(n-1)\dots(n-m+1)} \\ &\geq \left(1 - \frac{s}{n-m}\right)^m \geq \left(1 - \frac{2s}{n}\right)^{n/s} .\end{aligned}$$

The expression on the right is bounded away from zero when n/s is large enough. Hence regardless of the sampling strategy, there is a strictly positive probability that no active components are sampled under the alternative, which shows that (4.2) can not hold for ε small enough. When $p \neq 0$, sampling an active component becomes even harder, hence the same result holds.

Hence, in order to take the first steps in understanding the effect of the speed of change p has on the problem difficulty, we will focus on the case when the number of measurements we are allowed to make is of the order n/s. If we are in a situation where we want to make a decision as fast as possible, then this is the regime of m we want to consider. We aim to gain understanding of the problem in full generality in the future.

4.3 A Detection Procedure

In this section we propose an algorithm for the problem described in Section 4.2 and analyze its performance. Recall that in the current setup we can select our measurement actions adaptively based on our previous observations.

The main idea of the algorithm for detection is similar to the one presented in Chapter 2. Consider a setting where there is no measurement noise, that is, when measuring a component of $\boldsymbol{x}^{(t)}$ we know for sure whether that component is zero

or not. In such a setting if we find an active component we can stop and set $\Psi = 1$. On the other hand if we look at a large number of components and only observe zeros, it becomes safe to conclude that there are no active components. Bear in mind though that in case we did not observe any active components we might have simply been unlucky - so there is always a possibility for a false negative decision regardless of how many components we observe, unless $p = 0$.

The procedure that we propose is a robustified version of the one explained above, so that it can deal with measurement noise. This is done by performing a simple sequential test to gauge the identity of the component that we are observing. Recall that in Chapter 2 this was accomplished by using a Sequential Likelihood Ratio Test (SLRT). In principle we could use the same method here, but the dynamical nature of the signal causes some difficulties. In particular since the identity of the component that we are observing might change while performing the test, the analysis of the SLRT becomes cumbersome.

Note however, that the SLRT in the case of normal observation noise is a function of the running average of the observations. Hence we will use the same statistic as the core of a more simple sequential test, that is easier to analyze and still has good performance. This test is inspired by the Distilled Sensing algorithm of Haupt, Castro & Nowak [81] and the Sequential Thresholding procedure of Malloy & Nowak [109].

We now describe the algorithm formally. We query components uniformly at random one after another and examine their identity (whether they are 0 or not) using a sequential test to be described later. Once we find a component that we deem to have non-zero mean we stop and set $\Psi = 1$. If we do not find any components with elevated mean after examining T components, or we exhaust our measurement budget m, we set $\Psi = 0$.

Formally, let $\{Q_j\}_{j\in[T]}$ denote the components we query. We choose $Q_j \sim \text{Unif}([n])$ and independent for $j \in [T]$. The number of queries $T \leq m$ will be chosen appropriately later. For each Q_j we run a sequential test to determine the identity of that component. We refer to our sequential test as Sequential Thresholding Test (STT). To gauge the identity of Q_1, STT makes multiple measurements at that coordinate. That is, we set $A_t = Q_1$ for $t \in [N_1]$, where N_1 denotes the (random) number of measurements that STT makes. If STT deems Q_1 to be an active component then we stop and set $\Psi = 1$. Otherwise we move on to the second query, and run STT for that component (that is we set $A_t = Q_2,\ t = N_1+1, \ldots, N_1+N_2$).

We continue until either we find an active component, or perform all the tests for $Q_1, \ldots, Q_T$ and find no active components, or exhaust our measurement budget. In case we stop because we have found an active component we set $\Psi = 1$, otherwise we set $\Psi = 0$.

Algorithm 2: Algorithm for detection

Parameters:

- Number of queries $T \in \mathbb{N}$
- Queries $Q_1, \ldots, Q_T \overset{iid}{\sim} \mathrm{Unif}([n])$

Set $\Psi = 0$

for $j \leftarrow 1$ **to** T **do**

  Perform STT for the component indexed by Q_j

  If STT returns **"Signal"**: **break** and set $\Psi = 1$

  If measurement budget is exhausted: **break**

end

The sequential test that we use to examine the identity of a queried component is based on the ideas of distilled sensing introduced and analyzed by Haupt, Castro & Nowak in [81]. This algorithm is designed to recover the support of a sparse signal (whose active components remain the same during the sampling process). The main idea is to use the fact that the signal is sparse and try to measure active components as often as possible, while not wasting too many measurements on components that are not part of the support. We aim to achieve the same with our procedure. On one hand we wish to quickly identify when the component that we are sampling is non-active so that we can move on to probe a different location of the signal. On the other hand in case we are sampling an active component, we wish to keep sampling as long as possible so that we collect as much evidence of an active component as possible. However, unlike in the original setting of distilled sensing, we need to be able to quickly detect that we are sampling an active component, as it might move away if we hesitate for too long.

In words, STT collects at most k measurements sequentially and keeps track of the running average until one of the stopping conditions is met. The first stopping condition says that once the running average drops below the threshold t_k we stop and declare that there is no signal present. The second says that if the running average at step j exceeds a threshold t_j, we stop and conclude that a signal component is present. Note that after each measurement the upper threshold decreases, eventually reaching t_k, hence the procedure terminates after at most k

Algorithm 3: Sequential Thresholding Test (STT)

Input:
- $k \in \mathbb{N},\ t_1 > t_2 > \cdots > t_k > 0$

for $j \leftarrow 1$ **to** k **do**
  Observe X_j and compute $\overline{X}_j = \sum_{i=1}^{j} X_i/j$
  If $\overline{X}_j \leq t_k$: **break** and declare **No signal**
  If $\overline{X}_j > t_j$: **break** and declare **Signal**
end

measurements are made.

We now use a bit of heuristics to conjecture a good choice for the parameters k and $\{t_j\}_{j\in[k]}$. The sample collected by the algorithm consists of T blocks of measurements, where each block corresponds to an application of STT. Let the block lengths be denoted by $\{N_j\}_{j\in[T]}$. Suppose for a moment that blocks entirely consist of either zero mean or non-zero mean measurements. In this case we could simply collapse each block by replacing block j with its mean multiplied by $\sqrt{N_j}$ for all $j \in [T]$.

This would reduce the problem to a detection problem in a T-dimensional vector, each component being normally distributed and having unit variance. This is a well-understood setting, and we know that in this case the signal strength needs to scale as $\sqrt{\log T}$ when there are not too many active components (see for instance Donoho & Jin [64] and the references therein). Recall that we are concerned with the case where the number of measurements we are allowed to make is of the order n/s. Hence we do not expect to encounter active components too many times.

This heuristic shows that we should calibrate STT in a way that when it encounters j consecutive measurements with elevated mean, it should be able to detect it when $\mu \approx \sqrt{\frac{1}{j}\log T}$. Furthermore, note that by the above analysis we also need $\mu \gtrsim \sqrt{\log \frac{1}{\varepsilon}}$. Recalling that $j \leq k$, we thus see that choosing k greater than $\log T$ does not buy us anything.

We choose the parameters of STT by the previous heuristics. We can prove the following result:

Lemma 4.1. *Let $\varepsilon \in (0, 1/2)$ and define the parameters of STT as*

$$k = \lfloor \log T \rfloor \ ,$$
$$t_j = \sqrt{\frac{c(\varepsilon/T)}{j} \log \frac{T}{\varepsilon}}, \ j \in [k] \ ,$$

with

$$c(x) = 2\left(1 + \frac{\log\log(1/x)}{\log(1/x)}\right).$$

Denote the observations available to the STT by $X_1, \ldots, X_k$ (note that the STT may terminate without observing all the variables). Then the following holds:

(i) If all $X_i \sim N(0,1)$ and independent for $i \in [k]$, then STT declares "Signal" with probability at most ε/T.

(ii) If the $X_i \sim N(\mu, 1)$ and independent for $i \in [j]$ with

$$\mu \geq \sqrt{\frac{c(\varepsilon/T)}{j} \log \frac{T}{\varepsilon}} + \sqrt{2 \log \frac{2}{\varepsilon}} \ ,$$

then STT declares "No Signal" with probability at most ε.

Proof. (i): Suppose we have $X_i \sim N(0,1)$ and independent for $i \in [k]$. Note that the STT declares "Signal" if at any time step $j \in [k]$ the running average $\overline{X}_j$ exceeds the threshold t_j. We can use a union bound to upper bound the probability of this event, and plug in the parameter values to verify the claim. In detail,

$$\begin{aligned}
\mathbb{P}\left(\exists j \in [k] : \overline{X}_j \geq t_j\right) &\leq \sum_{j=1}^{k} \mathbb{P}(\overline{X}_j \geq t_j) \\
&\leq \sum_{j=1}^{k} \frac{1}{2} \exp\left(-\frac{j t_j^2}{2}\right) \\
&= \sum_{j=1}^{\lfloor \log T \rfloor} \frac{1}{2} \exp\left(-\frac{c(\varepsilon/T)}{2} \log \frac{T}{\varepsilon}\right) \\
&\leq \log T \cdot \left(\frac{\varepsilon}{T}\right)^{c(\varepsilon/T)/2} \ .
\end{aligned}$$

The right hand side is at most ε/T, which can be checked by taking the logarithm:

$$\begin{aligned}\log\left(\log T\cdot\left(\frac{\varepsilon}{T}\right)^{c(\varepsilon/T)/2}\right) &= \log\log T + \left(1+\frac{\log\log(T/\varepsilon)}{\log(T/\varepsilon)}\right)\log(\varepsilon/T)\\ &= \log\log T + \log(T/\varepsilon) - \log\log(T/\varepsilon)\\ &\le \log(T/\varepsilon)\ .\end{aligned}$$

(ii): Suppose $X_i \sim N(\mu,1)$ and independent for $i\in[j]$, with μ defined above. Let

$$\Omega = \{\exists i\in[j-1] : \overline{X}_i \le t_k\}\ .$$

Note that if this event happens, we stop and declare "No signal" in one of the first $j-1$ steps. Hence, using the law of total probability, the probability of missing the signal can be upper bounded as

$$\begin{aligned}\mathbb{P}(\text{Declare "No signal"}) &= \mathbb{P}(\Omega) + \mathbb{P}(\overline{\Omega})\mathbb{P}(\text{Declare "No signal"})|\Omega)\\ &\le \mathbb{P}(\Omega) + \mathbb{P}(\overline{\Omega})\mathbb{P}(\overline{X}_j \le t_j|\Omega)\\ &\le \mathbb{P}(\Omega) + \mathbb{P}(\overline{X}_j \le t_j)\ .\end{aligned}$$

Using a union bound and the same Gaussian tail bound as before, the last expression can be upper bounded by

$$\sum_{i=1}^{j-1}\frac{1}{2}\exp\left(-\frac{i(\mu-t_k)^2}{2}\right) + \frac{1}{2}\exp\left(-\frac{j(\mu-t_j)^2}{2}\right)\ . \tag{4.4}$$

Considering he first term above, note that

$$\mu - t_k = t_j + \sqrt{2\log\frac{2}{\varepsilon}} - t_k \ge \sqrt{2\log\frac{2}{\varepsilon}}\ ,$$

since $t_j \ge t_k$ (recall that $j\le k$). Hence the first term can be upper bounded as

$$\begin{aligned}\sum_{i=1}^{j-1}\frac{1}{2}\exp\left(-\frac{i(\mu-t_k)^2}{2}\right) &\le \frac{1}{2}\sum_{i=1}^{j-1}(\varepsilon/2)^i\\ &\le \frac{1}{2}\sum_{i=1}^{\infty}(\varepsilon/2)^i\\ &= \frac{\varepsilon}{2}\frac{1}{2-\varepsilon} \le \varepsilon/2\ .\end{aligned}$$

On the other hand, when μ satisfies the inequality above, the second term is simply upper bounded by $(\varepsilon/2)^j$, and hence the claim follows. $\square$

Using Lemma 4.1, we can establish a performance guarantee for our detection algorithm. Though it is possible to derive a result for a fixed set of parameters (n, s), perhaps it is better to state an asymptotic result instead, which reads more easily, and better highlights the impact of the parameter p.

Keeping this comment in mind, note that $2 \leq c(x) \leq 2(1 + 1/e)$ and $c(x) \to 2$ as $x \to 0$. Thus, keeping ε fixed and letting $T \to \infty$, we see that if there exists a $\tau > 1$ for which

$$\mu \geq \tau \sqrt{\frac{2}{j} \log T} + \sqrt{2 \log \frac{2}{\varepsilon}} \ ,$$

then for T large enough the condition on μ in Lemma 4.1 is satisfied. Furthermore, recall that our main interest is how the algorithm performs when the time horizon is of the order n/s.

Proposition 4.1. *Fix $\varepsilon \in (0, 1/2)$ and let $n \to \infty$ and $s = o(n/(\log n)^2)$. Set $T = \frac{4n}{s} \log_2 \frac{2}{\varepsilon}$ and the parameters of STT according to Lemma 4.1. If the measurement budget is $m > 2T$ then the algorithm described above asymptotically satisfies*

$$\max_{i=0,1} \mathbb{P}_i(\Psi \neq i) \leq 2\varepsilon \ ,$$

whenever

$$\mu \geq \tau \sqrt{\frac{2}{\min\{1/(2p), \log(n/s)\}} \log(n/s)} + \sqrt{2 \log \frac{2}{\varepsilon}} \ ,$$

with $\tau > 1$ fixed, but arbitrary.

Before we move on to the proof of this result, let us discuss its message. First note that the detection algorithm is agnostic about the speed of change p and the signal strength μ, though it does require knowledge of the sparsity s to set the parameter T. On the other hand it is apparent from the proof of Proposition 4.1, that it is enough to have a lower bound on the sparsity.

The number of measurements that the Algorithm 2 requires to be effective scales like n/s, which is the minimum amount necessary to be able to solve the problem (see Section 4.4). Furthermore, when $p < 2/\log(n/s)$ the signal strength needs to scale as $\sqrt{\log(1/\varepsilon)}$, and when $p \geq 2/\log(n/s)$ it needs to scale as $\sqrt{p \log(n/s)}$. This matches our intuition that the speed of change p affects the problem difficulty

in a monotonic fashion. We will show in Section 4.4 that in the regime $m \approx n/s$ this scaling of μ is necessary to reliably solve this detection task.

Remark 4.1. *As we have mentioned in Section 4.2.1, for now we are interested in the case where the number of observations we can make is of the order n/s. In fact, in Section 4.4 we only show the optimality of the above algorithm in this regime. Note that Proposition 4.1 claims the same performance guarantee for every m that is at least of order n/s. In fact, it is not hard to see that the performance of this algorithm does not improve as m increases, hinting that it is likely suboptimal for large m. To illustrate this, we present a very simple algorithm and a back of the envelope analysis here.*

Let us sample components uniformly at random in each step $t \in [m]$. Then in each step we hit an active component with probability s/n. We then roughly have ms/n active components in our sample under the alternative. Consider the standardized sum of our observations. Under the null this follows a standard normal distribution, whereas under the alternative it is distributed as $N(\sqrt{ms}\mu/n, 1)$. Thus reliable detection using this algorithm is possible when μ is of the order $n/(\sqrt{ms})$. Hence this simple algorithm clearly outperforms the one above when m is large enough (compared to n/s).

proof of Proposition 4.1. In light of Lemma 4.1, the type I error probability is at most ε by a union bound. Hence we are left with studying the alternative.

There are two ways that our algorithm can make a type II error. Either the measurement budget is exhausted, or we fail to identify an active component in T runs of STT. We show that the probability of both events is small if n is large enough.

We start with upper bounding the probability of exhausting our measurement budget. Let N_j denote the number of measurements that STT makes when called for the jth time, for $j \in [T]$. Note that these variables are independent and identically distributed, because the components to query are selected uniformly at random independently from the past, the dynamic evolution of the model is memoryless and finally the observation noise is independent. First we upper bound $\mathbb{E}_1(N_1)$. Note that $1 \leq N_1 \leq k$, where $k = \log T$ by Lemma 4.1. Let Ω denote the event that a non-zero mean observation appears at location A_1 in any of the first k steps. By the law of total expectation we have

$$\mathbb{E}_1(N_1) \leq k\mathbb{P}_1(\Omega) + \mathbb{E}_1(N_1|\overline{\Omega}) \ .$$

Note that

$$\mathbb{P}_1(\Omega) = \mathbb{P}_1(\exists t \in [k] : A_1 \in S^{(t)}) \leq \sum_{t=1}^{k} \mathbb{P}_1(A_1 \in S^{(t)})$$
$$\leq \frac{s}{n} + (k-1)\frac{s}{n-s} \leq \frac{ks}{n-s} \ ,$$

since the choice of A_1 (and $S^{(1)}$) is random, and in each subsequent step the probability that a signal component moves to location A_1 is at most $s/(n-s)$ regardless of p. On the other hand, recalling that $t_k = \sqrt{\frac{c(\varepsilon/T}{\lfloor \log T \rfloor} \log \frac{T}{\varepsilon}} \leq \sqrt{2}$ is the lower stopping boundary of STT,

$$\mathbb{E}_1(N_1|\overline{\Omega}) = \sum_{t=1}^{k} \mathbb{P}_0(N_1 \geq t) \leq \sum_{t=1}^{k} \mathbb{P}_0(\overline{Y}_{t-1} > t_k) \leq \sum_{t=1}^{k} \mathbb{P}_0(\overline{Y}_{t-1} > \sqrt{2})$$
$$\leq 1 + \frac{1}{2}\sum_{t=1}^{k-1} e^{-t} \leq 1 + \frac{1}{2(e-1)} < 2 \ .$$

Hence

$$\mathbb{E}_1(N_1) \leq 1 + \frac{1}{2(e-1)} + \frac{k^2 s}{n-s} \leq 2 \ ,$$

eventually as $n \to \infty$ by the definition of k (and T) and since $s = o(n/(\log n)^2)$. Hence, using the notation $m = (2+c)T$ with some $c > 0$ and using Hoeffding's inequality (when n is large enough for the previous inequality to hold) we get

$$\mathbb{P}_1\left(\sum_{j=1}^{T} N_j > m\right) = \mathbb{P}_1\left(\sum_{j=1}^{T} N_j - \mathbb{E}_1\Big(\sum_{j=1}^{T} N_j\Big) > m - \mathbb{E}_1\Big(\sum_{j=1}^{T} N_j\Big)\right)$$
$$\leq \mathbb{P}_1\left(\sum_{i=1}^{T} N_i - \mathbb{E}_1\Big(\sum_{i=1}^{T} N_i\Big) > cT\right)$$
$$\leq \exp\left(-\frac{c^2 T}{k^2}\right) \to 0 \ ,$$

as $n \to \infty$ again using the definitions of T and k and the assumption about s. This shows that the probability that the measurement budget is exhausted tends to zero as $n \to \infty$.

The second step is to guarantee that the algorithm identifies an active component in one of the T tests with high probability. To show this, we first guarantee that there will be an instance in the repeated application of STT where the first

$1/(2p)$ observations that the procedure has access to have elevated mean. Then we can apply Lemma 4.1 together with a union bound to conclude the proof.

Let $T_j = \sum_{i=1}^{j-1} N_i + 1$ denote the time when STT starts for the jth time. Let $N = \sum_{j=1}^{T} \mathbf{1}\{Q_j \in S^{(T_j)}\}$ denote the number of times an active component is sampled at the start of STT. Note that $N \sim \mathrm{Bin}(T, s/n)$. Each time this happens, the first few observations STT has access to have elevated mean. Denote the number of these consecutive observations by $\{\eta_i\}_{i\in[N]}$. Note that $\eta_i \sim \mathrm{Geom}(p)$ and $\{\eta_i\}_{i\in[N]}$ are independent. We have

$$\begin{aligned}\mathbb{P}(\forall i \in [N] : \eta_i < 1/(2p)) \leq \mathbb{P}\left(\forall i \in [N] : \eta_i < 1/(2p) \mid N \geq \log_2 \tfrac{2}{\varepsilon}\right) \\ + \mathbb{P}\left(N < \log_2 \tfrac{2}{\varepsilon}\right) \ .\end{aligned}$$

On one hand, note that the median of η_i is $\lceil 1/|\log_2(1-p)| \rceil$ which is greater than $1/(2p)$. This can be easily checked by considering the cases $p \geq 1/2$ and $p < 1/2$ separately. Hence the first term above can be upper bounded as

$$\mathbb{P}\left(\forall i \in [N] : \eta_i < \lceil 1/|\log_2(1-p)| \rceil \mid N \geq \log \tfrac{4}{\varepsilon}\right) \leq 2^{-\log_2 \frac{2}{\varepsilon}} = \varepsilon/2 \ .$$

On the other hand, $N \sim \mathrm{Bin}(T, s/n)$ and so by Bernstein's inequality,

$$\mathbb{P}\left(N < (1-\delta)\frac{Ts}{n}\right) \leq \exp\left(-\frac{3\delta^2}{8(1-\delta)}\frac{Ts}{n}\right) \ ,$$

for any $\delta \in (0,1)$. However, note that plugging in the value of T together with $\delta = 2/3$ yields

$$\mathbb{P}\left(N < \log_2 \tfrac{2}{\varepsilon}\right) \leq \mathbb{P}\left(N < \frac{4}{3}\log_2 \tfrac{2}{\varepsilon}\right) \leq \exp\left(-2\log_2 \tfrac{2}{\varepsilon}\right) < \varepsilon/2 \ ,$$

since $\log_2 x > \log x$ for $x > 1$.

Union bounding concludes the proof. □

4.4 Lower bounds

In this section we identify conditions for the signal strength that are necessary for the existence of a test satisfying

$$\max_{i=0,1} \mathbb{P}_i(\Psi \neq i) \leq \varepsilon \ .$$

First, we consider the non-adaptive sensing setting for comparison purposes. Then we consider the adaptive sensing setting to show the near-optimality of the algorithm proposed in Section 4.3.

In both cases we focus on the regime $m \approx n/s$, as highlighted in Section 4.2.1.

4.4.1 Non-adaptive sensing

In the non-adaptive sensing setting, the sampling strategy $\{A_t\}_{t\in[m]}$ needs to be specified before any observations are made. Note that this does not exclude the possibility that these variables are generated randomly.

Due to technical difficulties we were so far unable to show a general lower bound for the non-adaptive sensing setting. Nonetheless, we show a lower bounding argument to illustrate the difficulty that we face compared to the more classical signal detection settings. We also show that this proof sketch can be made rigorous with two additional results that heuristically seem valid, but to which we do not have a formal proof yet.

The usual technique for deriving non-adaptive lower bounds in high-dimensional signal detection problems is the so-called second moment method. Let $\mathbb{P}_0$ and $\mathbb{P}_1$ denote the distribution of $\{Y_t\}_{t\in[m]}$ under the null and the alternative respectively. Under the null $Y_t \sim N(0,1),\ t \in [m]$ independent, regardless of the choice of $\{A_t\}_{t\in[m]}$. Under the alternative

$$Y_t \sim \mathbf{1}\{A_t \in S^{(t)}\}N(\mu,1) + \mathbf{1}\{A_t \notin S^{(t)}\}N(0,1),\ t \in [m]$$

and Y_t and $Y_{t'}$ are conditionally independent given $\{A_t\}_{t\in[m]}$ and $\{S^{(t)}\}_{t\in[m]}$ for $t \neq t'$. We use the notations $\mathbf{Y} = \{Y_t\}_{t\in[m]}, \mathbf{S} = \{S^{(t)}\}_{t\in[m]}$ and $\mathbf{A} = \{A_t\}_{t\in[m]}$.

Suppose there exists a test for which the maximum of type I and type II errors is at most ε. By Theorem 2.2 in Tsybakov [139],

$$\varepsilon \geq \inf_{\Psi} \max_{i=0,1} \mathbb{P}_i(\Psi \neq i) \geq 1 - TV(\mathbb{P}_0, \mathbb{P}_1)\ ,$$

where $TV(\cdot,\cdot)$ is the Total Variation distance. This is usually difficult to compute, hence we continue by upper bounding the Total Variation distance with the χ^2-

divergence. We have the following bound (see Section 2.4.1 in Tsybakov [139]):

$$TV(\mathbb{P}_0, \mathbb{P}_1) \leq \sqrt{\frac{1}{2}\chi^2(\mathbb{P}_0, \mathbb{P}_1)} = \sqrt{\frac{1}{2}\left(\mathbb{E}_0\left[\left(\frac{\mathrm{d}\mathbb{P}_1}{\mathrm{d}\mathbb{P}_0}\right)^2\right] - 1\right)} \, .$$

Note that there are other ways to upper bound the Total Variation distance. For instance, we could use a function of the Kullback-Leibler divergence to continue. By doing so, we would essentially get that μ needs to scale as $\sqrt{n/(sm)}$, which in the regime $m = n/s$ would only claim that the signal strength has to be above a constant. This lower bound is probably very loose. Another possibility would be to use the Hellinger-distance to continue, which would lead to similar computations and hence similar difficulties to the ones we encounter when using the χ^2-divergence.

Continuing with the bound above, we get

$$2(1-\varepsilon)^2 + 1 \leq \mathbb{E}_0\left[\left(\frac{\mathrm{d}\mathbb{P}_1}{\mathrm{d}\mathbb{P}_0}\right)^2\right] \, .$$

We need to upper bound the right hand side above. Note that the density $\mathrm{d}\mathbb{P}_1$ is a mixture. In particular, denoting the density of $N(\mu, 1)$ by f_μ, the conditional density of $\mathbf{Y}$ given $\mathbf{A}$ and $\mathbf{S}$ is

$$\mathrm{d}\mathbb{P}_1(\mathbf{Y}|\mathbf{A}, \mathbf{S}) = \prod_{t\in[m]} \left(\mathbf{1}\{A_t \in S^{(t)}\} f_\mu(Y_t) + \mathbf{1}\{A_t \notin S^{(t)}\} f_0(Y_t)\right) \, .$$

Hence we can write $\mathrm{d}\mathbb{P}_1(\mathbf{Y}) = \mathbb{E}_{\mathbf{A}}\left[\mathbb{E}_{\mathbf{S}}\left[\mathrm{d}\mathbb{P}_1(\mathbf{Y}|\mathbf{A}, \mathbf{S})\right]\right]$, where the subscripts of the expectations indicate which random variables are integrated out by that particular expectation.

Let $\mathbf{S}' = \{S'^{(t)}\}_{t\in[m]}$ denote a sequence of supports having identical distribution to that of $\mathbf{S}$, and independent of $\mathbf{S}$. Using Jensen's inequality we get

$$\begin{aligned}
\mathbb{E}_0\left[\left(\frac{\mathrm{d}\mathbb{P}_1}{\mathrm{d}\mathbb{P}_0}\right)^2\right] &= \mathbb{E}_0\left[\left(\frac{\mathbb{E}_{\mathbf{A}}\left[\mathbb{E}_{\mathbf{S}}\left[\mathrm{d}\mathbb{P}_1(\mathbf{Y}|\mathbf{A}, \mathbf{S})\right]\right]}{\mathrm{d}\mathbb{P}_0(\mathbf{Y})}\right)^2\right] \\
&\leq \mathbb{E}_0\left[\mathbb{E}_{\mathbf{A}}\left[\mathbb{E}_{\mathbf{S}}\left[\frac{\mathrm{d}\mathbb{P}_1(\mathbf{Y}|\mathbf{A}, \mathbf{S})}{\mathrm{d}\mathbb{P}_0(\mathbf{Y})}\right]^2\right]\right] \\
&= \mathbb{E}_0\left[\mathbb{E}_{\mathbf{A}}\left[\mathbb{E}_{\mathbf{S}}\left[\frac{\mathrm{d}\mathbb{P}_1(\mathbf{Y}|\mathbf{A}, \mathbf{S})}{\mathrm{d}\mathbb{P}_0(\mathbf{Y})}\right]\mathbb{E}_{\mathbf{S}'}\left[\frac{\mathrm{d}\mathbb{P}_1(\mathbf{Y}|\mathbf{A}, \mathbf{S}')}{\mathrm{d}\mathbb{P}_0(\mathbf{Y})}\right]\right]\right] \, .
\end{aligned}$$

Note that

$$\frac{\mathrm{d}\mathbb{P}_1(\mathbf{Y}|\mathbf{A},\mathbf{S}')}{\mathrm{d}\mathbb{P}_0(\mathbf{Y})} = \prod_{t\in[m]} \mathbf{1}\{A_t \in S^{(t)}\}\frac{f_\mu(Y_t)}{f_0(Y_t)} \ .$$

Hence the expression on the right above is equal to

$$\mathbb{E}_0\left[\mathbb{E}_{\mathbf{A}}\left[\mathbb{E}_{\mathbf{S},\mathbf{S}'}\left[\prod_{t\in[m]} \mathbf{1}\{A_t \in (S^{(t)} \cap S'^{(t)})\}\left(\frac{f_\mu(Y_t)}{f_0(Y_t)}\right)^2 \right.\right.\right.$$
$$\left.\left.\left. \prod_{t\in[m]} \mathbf{1}\{A_t \in (S^{(t)} \triangle S'^{(t)})\}\left(\frac{f_\mu(Y_t)}{f_0(Y_t)}\right)\right]\right]\right] .$$

By changing the order of integration we can move the outermost expectation inside, and integrate out Y_t to get

$$\mathbb{E}_0\left[\left(\frac{\mathrm{d}\mathbb{P}_1}{\mathrm{d}\mathbb{P}_0}\right)^2\right] \le \mathbb{E}_{\mathbf{A}}\left[\mathbb{E}_{\mathbf{S},\mathbf{S}'}\left[\prod_{t\in[m]} \exp\left(\mathbf{1}\{A_t \in (S^{(t)} \cap S'^{(t)})\}\mu^2\right)\right]\right] . \qquad (4.5)$$

Evaluating the expectations above in general is challenging because we can not move the expectations inside the products, since we do not have independence. Note that when $p = 1$ the supports are chosen uniformly at random at each time step, so we can evaluate the inner expectation and get

$$\mathbb{E}_0\left[\left(\frac{\mathrm{d}\mathbb{P}_1}{\mathrm{d}\mathbb{P}_0}\right)^2\right] \le \mathbb{E}_{\mathbf{A}}\left[\prod_{t\in[m]}\left(1 + \frac{s^2}{n^2}(e^{\mu^2} - 1)\right)\right]$$
$$= \left(1 + \frac{s^2}{n^2}(e^{\mu^2} - 1)\right)^m .$$

This would lead to the bound

$$\mu \ge \sqrt{\log\left(\frac{n^2}{s^2}\left(\sqrt[m]{2(1-\varepsilon)^2+1} - 1\right) + 1\right)}$$
$$\ge \sqrt{\log\left(\frac{n^2}{s^2 m}\log\left(2(1-\varepsilon)^2+1\right) + 1\right)} ,$$

using $\log x \le x - 1$ with $x = \sqrt[m]{2(1-\varepsilon)^2+1}$. In the regime $m \approx n/s$ this lower bound states that the signal strength needs to be of the order $\sqrt{\log(n/s)}$. Our conjecture is that this scaling is required regardless of the value of p (at least, in

the regime $m \approx n/s$). We illustrate the reason behind our conjecture with some heuristics.

Consider the case $p = 0$. In this case, a good strategy is to subsample the signal. This means that we select a number of locations of the signal, and sample only at the selected locations. Clearly, the number of locations we choose to sample at needs to be of the order n/s, by the same token as before. Using such a sampling scheme, it is easy to check that detection is possible whenever μ is at least of the order $\sqrt{(n/sm) \cdot \log(n/s)}$. Although we do not have a formal proof, we conjecture that such a sampling scheme is optimal in the case $p = 0$.

Note that the conjectured bound for the case $p = 0$ in the regime $m \approx n/s$ is of the order $\sqrt{\log(n/s)}$, which is the same as the bound we have for $p = 1$. Since we suspect the same result to hold (at least qualitatively) at the two extremes of the spectrum, we believe that this result should also hold for every $p \in [0, 1]$. For a rigorous argument, we need two ingredients:

(i) A lower bound for the case $p = 0$, and

(ii) A formal argument showing the monotonicity of the problem difficulty in p.

It is important to emphasize that the above conjecture is for the case when m is of the order n/s. As we have seen in Remark 4.1, this lower bound can not be valid for large m.

4.4.2 Adaptive sensing

In the adaptive sensing setting the sample collection strategy can depend on information gleaned during the sample collection process. Fundamental limits of the detection problem in this setup have been studied by Castro in [40] for the case when $p = 0$. The lower bounds presented in that work are valid for a more general observation model than the one considered in this chapter. Specifically, the setting in Castro [40] is the same as in Chapter 2, that is, we are allowed to choose the precision of the measurements as well as which coordinates of the signal to measure. Hence the lower bound presented there is also a valid lower bound for our setting.

A result in that work translated to our setting states the following:

Theorem 4.1 (Theorem 3.1 of [40]). *Let $\mathcal{C} = \{S \subset \{1, \dots, n\} : |S| = s\}$ and $\varepsilon \in (0, 1)$. If there exists an adaptive algorithm producing an test Ψ such that*

$$\max\{\mathbb{P}_0(\Psi = 1), \max_{S \in \mathcal{C}} \mathbb{P}_S(\Psi = 0)\} \leq \varepsilon$$

then necessarily

$$\mu \geq \sqrt{\frac{2n}{sm} \log \frac{1}{\varepsilon}} \ .$$

Note that the way the that performance of the test is measured in the above result is slightly different than the way we have defined it in (4.2). However, it is easy to show that these only differ by a constant multiplicative factor (this is also argued in Castro [40]), meaning that essentially the same scaling holds in our setting as well.

In the regime $m \approx n/s$ the bound states that the signal strength needs scale as $\sqrt{\log(1/\varepsilon)}$. This coincides with the bound in Proposition 4.1 when $p \leq 2/\log(n/s)$. This tells us that when the signal changes slowly enough, the problem is essentially non-dynamic in nature, which is what one would expect.

On the other extreme end of the spectrum is the case when $p = 1$. In this case the signal resets in every time instance. Let $\{A_t\}_{t\in[m]}$ be the sensing actions of an arbitrary algorithm. Under the alternative we have $\mathbf{1}\{A_t \in S^{(t)}\} \overset{iid}{\sim} \mathrm{Ber}(s/n)$, and so the observation model is a mixture model, and the same computations hold as in the non-adaptive sensing setting. In particular, when the support changes in every time step, adaptive sensing does not provide any help in detection, which is also what one would expect. Hence, when $m \approx n/s$, then for any test to have small probability of error, the signal strength has to scale like $\sqrt{\log \frac{n}{s}}$. Again, this rate coincides with the performance guarantee of our algorithm established in Proposition 4.1.

Non-extreme dynamics ($p \in (0,1)$), 1-sparse case

For general values of p we start by considering the 1-sparse case. This case is considerably simpler than the general s-sparse setting, as now whenever the active component changes, the entire signal resets. This effectively creates a number of independent static signals on the time horizon.

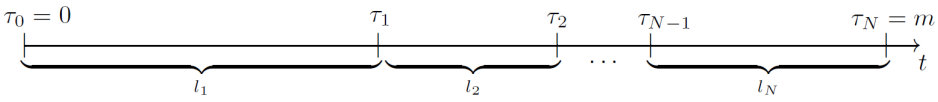

Figure 4.2: Illustration of the notations introduced for the 1-sparse case.

Let us introduce some notation, which we illustrate in Figure 4.2. Recall that the variables $\theta_i^{(t)} \overset{iid}{\sim} \mathrm{Ber}(p)$, $t \in [m]$, $i \in [s]$ identify the change points of the

signal. Since now we are dealing with the 1-sparse case we have one variable per time index, so in what follows we drop the subscript from the previous notation. Furthermore, note that our time horizon is m, so enforcing $\theta^{(m)} = 1$ does not change the model. Let the number of change points over our time horizon be $N = \sum_{t\in[m]} \mathbf{1}\{\theta^{(t)} = 1\}$. Note that $N - 1 \sim \mathrm{Bin}(m-1, p)$. Let $\tau_0 = 0$ and for $j \in N$ let $\tau_j = \min\{t > \tau_{j-1} : \theta^{(t)} = 1\}$ denote the time instances when the signal changes (so $\tau_N = m$). Note that on the time intervals $[\tau_j + 1, \tau_{j+1}]$ the signal is static. Finally, for any $t \in [m]$ let the number of change points up to time t be $N(t) = \max\{j : \tau_j \leq t\}$.

We wish to get a lower bound for all adaptive sensing algorithms described in Section 4.2. Denote the set of those algorithms by $\mathcal{A}$. Our first step is to consider a wider class of adaptive sensing algorithms $\mathcal{A} \subseteq \mathcal{A}'$ and show a lower bound for the class $\mathcal{A}'$. Let $\mathcal{A}'$ be the set of algorithms that at time t also has access to the variables $\{\theta^{(j)}\}_{j=1}^{t-1}$ next to those described in the problem setup. In words, this means the algorithm also has knowledge of when the signal changes. Due to this extra information such an algorithm has an advantage compared to those that we are interested in, or more formally $\mathcal{A} \subseteq \mathcal{A}'$. Note that since we are considering a 1-sparse problem, the entire signal resets when the active component changes. Hence, when deciding where to collect the observation at time t, all the information gathered before time $\tau_{N(t)}$ is irrelevant. More formally, an algorithm in $\mathcal{A}'$ that also uses the information in the variables $\{\theta^{(j)}\}_{j\in[t-1]}$ has the property that A_t is only a function of $\{Y_j, A_j\}_{j=\tau_{N(t)}}^{t-1}$ (instead of $\{Y_j, A_j\}_{j=1}^{t-1}$). Hence, for such algorithms, the variables $\omega_j = \sum_{t=\tau_{j-1}+1}^{\tau_j} \mathbf{1}\{A_t \in S^{(t)}\}$ are independent for $j \in [N]$.

Consider the variables $\omega_j,\ j \in N$. Under the null all those indicators are zero, since $S^{(t)} = \emptyset$ for every $t \in [m]$ and so all the observations that the algorithm collects have mean 0. Under the alternative, if an algorithm does not manage to sample an active component in the time interval $[\tau_{j-1} + 1, \tau_j]$, then all the observations are mean 0. Otherwise, some observations in that time interval have mean μ. Consider a setting where the event $\{\omega_j = 1\}$ implies that all observations in the time interval $[\tau_{j-1} + 1, \tau_j]$ have mean μ under the alternative. In such a situation the distinction between the null and the alternative is easier as we have more anomalous samples under the alternative.

Hence we aim to lower bound the following block-mixture problem. Under the null $Y_t \sim N(0, 1)$ and independent $t \in [m]$. Under the alternative

$$Y_t \sim \omega_{N(t)} N(\mu, 1) + (1 - \omega_{N(t)}) N(0, 1)$$

and independent for $t \in [m]$. In words, the model contains m measurements, consisting of N blocks (where $N-1$ is a binomial random variable, see Figure 4.2). Block j activates independently of the other blocks, for every $j \in [N]$. When $\omega_j = 1$, every measurement in the block $[\tau_{j-1}+1, \tau_j]$ has mean μ, otherwise all of them have mean zero. The lower bound for μ in this problem is also a lower bound in the original problem by the above reasoning. Let $q_j = \mathbb{P}(\omega_j = 1)$ denote the probability that the signal component is sampled in the time interval $[\tau_{j-1}+1, \tau_j]$. This probability can re-written as

$$q_j = \mathbb{P}(\cup_{t=\tau_{j-1}+1}^{\tau_j}\{A_t = S^{(t)}\}) = \mathbb{P}(S^{(t)} \in \{A_t : \ \tau_{j-1}+1 \le t \le \tau_j\}) \ .$$

However, $|\{A_t : \ \tau_{j-1}+1 \le t \le \tau_j\}| \le \tau_j - \tau_{j-1} := l_j$ thus we have $q_j \le l_j/n$.

Our strategy for deriving a lower bound for the block-mixture problem above is the following: First, we show a lemma stating that under the random signal model described above, the number of blocks N the signal consists of is at least $mp/2$, and no block is longer than cm/N with probability bounded away from zero for a well-chosen universal constant c. This means that in order for any procedure to have a small probability of error, the procedure needs to have a small probability of error on this event. Then we formulate a necessary condition for μ so that a reliable test exist, given this event.

Lemma 4.2. *Consider the event*

$$\Omega = \Big\{\{N-1 > mp/2\} \cap \{\forall j : \ l_j \le cm/N\}\Big\} \ .$$

In the model described above $\mathbb{P}(\Omega) > 1/4$ *whenever* $c \ge 6 + 3\log 2$ *and* $p \ge 8/m$.

Proof. We write

$$\begin{aligned}\mathbb{P}(\Omega) &= \mathbb{E}\left(\{N-1 > mp/2\} \cap \{\forall j : \ l_j \le cm/N\}\right) \\ &= \mathbb{E}_N\left(\mathbf{1}\{N-1 > mp/2\}\mathbb{E}_{\mathbf{l}|N}\left(\mathbf{1}\{\forall j : \ l_j \le cm/N\} \ \big|N\right)\right) \ ,\end{aligned}$$

where $\mathbf{l} = (l_1, \ldots, l_N)$ is a shorthand notation, and as before the subscripts of the expectations are reminders to which random quantities are integrated out by that particular expectation. We continue by lower bounding the expression on the right.

We first lower bound the inner conditional probability

$$\mathbb{E}_{\mathbf{l}}\left(\mathbf{1}\{\forall j : \ l_j \le cm/N\} \ \big|N\right) \ .$$

Note that if $N \leq c$ this probability is one (since $cm/N \geq m$ and $l_j \leq m$ by definition). Otherwise, we will upper bound the probability of the complementary event. This can be done by counting the number of configurations for placing the $N-1$ endpoints of the intervals in the set $[m-1]$ that results in at least one block that is longer than cm/N. To get an upper bound on this count we use the following strategy. First, place an interval of length cm/N in the set $[m-1]$. Then choose the $N-1$ endpoints of the intervals such that none of them are in the aforementioned interval. This gives a configuration in which the existence of at least one long interval is guaranteed. Furthermore, though some configurations are counted multiple times, every configuration leading to at least one long interval is counted at least once. Hence

$$\begin{aligned}
&\mathbb{P}_{\mathbf{1}}(\exists j:\ l_j > cm/N | N) \leq \\
&\quad \leq (m - cm/N)\frac{\binom{m-cm/N}{N-1}}{\binom{m-1}{N-1}} \\
&\quad = (m - cm/N)\frac{(m-cm/N)(m-cm/N-1)\dots(m-cm/N-N+2)}{(m-1)(m-2)\dots(m-N+1)} \\
&\quad \leq \frac{m}{m-1}(1-c/N)\left(\frac{m-cm/N}{m-2}\right)^{N-2} \\
&\quad < \left(\frac{m-cm/N}{m-2}\right)^{N-2} .
\end{aligned}$$

Now consider the logarithm of the expression above. Using $\log(1+x) \leq x$, we get

$$\begin{aligned}
\log \mathbb{P}_{\mathbf{1}}(\exists j:\ l_j > cm/N | N) &< (N-2)\left(\log\frac{m}{m-2} + \log(1-c/N)\right) \\
&\leq (N-2)\left(\frac{2}{m-2} - \frac{c}{N}\right) \\
&\leq -\log 2 ,
\end{aligned}$$

whenever $c \geq 6 + 3\log 2$, using the fact that $3 \leq c \leq N \leq m$.

Hence $\mathbb{P}(\Omega) \geq \mathbb{P}_N(N-1 > mp/2)/2$. All that remains is to use the fact that $N - 1 \sim \mathrm{Bin}(m-1, p)$. For instance Chebyshev's inequality yields

$$\mathbb{P}(N - 1 \leq mp/2) \leq \frac{4(m-1)p(1-p)}{(mp)^2} \leq 1/2 ,$$

when $p \geq 8/m$ and so the claim is proved. $\square$

We can use the above result to argue that in order to have a small probability of error, we need to be able to distinguish the null and the alternative under the event Ω. Under this event we can compute a suitable metric of divergence between the two (conditional) distributions, which leads to a lower bound for μ.

Theorem 4.2. *Let $\{\theta^{(t)}\}_{t=1}^m, N, \{\tau_j\}_{j=1}^N, \{l_j\}_{j=1}^N$ be defined as above and consider the problem of distinguishing $H_0 : \ Y_t \overset{iid}{\sim} N(0,1), \ t \in [m]$ and*

$$H_1 : \ Y_t \overset{iid}{\sim} \omega_{N(t)} N(\mu, 1) + (1 - \omega_{N(t)}) N(0,1), \ t \in [m] \ ,$$

where $\omega_j \overset{iid}{\sim} \mathrm{Ber}(q_j)$ (with $q_j \leq l_j/n$). Suppose

$$\inf_{\Psi} \max_{i=0,1} \mathbb{P}_i(\Psi = 1 - i) \leq \varepsilon \ ,$$

with some $\varepsilon \in (0, 1/8)$, and that $p \geq 8/m$. Then necessarily

$$\mu \geq \sqrt{\frac{p}{2c} \log\left(\frac{\log(2(1-8\varepsilon)^2 + 1) p^2 n^2}{4c^2 m} + 1 \right)} \ ,$$

with $c = 6 + 3\log 2$.

Before we prove this result, let us compare the above bound on μ with the guarantees of our detection algorithm proved in Proposition 4.1. Note that c and ε are fixed constants. Thus the bound on the signal strength in the above result scales as $\sqrt{p \log(p^2 n^2/m)}$. Recall that we are interested in the regime $m \approx n/s$ and that $s = 1$, as we are considering the 1-sparse case. Plugging this value in, the bound above scales as $\sqrt{p \log(p^2 n)}$. Also note that the scaling of the performance guarantee of Proposition 4.1 matches that of the lower bound in Theorem 4.1 when $p < 1/\log n$. Hence we only need to assess the result of Theorem 4.2 when $p \geq 1/\log n$. In this case, the scaling of that bound is at least as big as $\sqrt{p(\log n - 2\log\log n)} \approx \sqrt{p \log n}$. This shows near-optimality of the algorithm proposed in Section 4.3, in terms of its scaling in the parameters n and p.

Proof. We use similar arguments to those presented in the non-adaptive sensing setting in Section 4.4.1. First, we use the law of total probability and Theorem 2.2

in [139] to get

$$\begin{aligned}\varepsilon &\geq \inf_{\Psi} \max_{i=0,1} \mathbb{P}_i(\Psi = 1-i) \\ &\geq \frac{\mathbb{P}(\Omega)}{2} \inf_{\Psi} (\mathbb{P}_0(\Psi = 1|\Omega) + \mathbb{P}_1(\Psi = 0|\Omega)) \\ &\geq \mathbb{P}(\Omega) \frac{1 - TV(\mathbb{P}'_0, \mathbb{P}'_1)}{2} \ ,\end{aligned}$$

where $TV(\cdot,\cdot)$ is the total variation distance and $\mathbb{P}'_i$ is the conditional distribution $\mathbb{P}_i$ given Ω, $i = 0, 1$ (note that $\mathbb{P}'_0 = \mathbb{P}_0$). We continue the previous chain of inequalities by upper bounding the total variation distance with the χ^2-divergence. We have

$$TV(\mathbb{P}_0, \mathbb{P}'_1) \leq \sqrt{\frac{1}{2}\chi^2(\mathbb{P}_0, \mathbb{P}'_1)} = \sqrt{\frac{1}{2}\left(\mathbb{E}_0\left[\left(\frac{\mathrm{d}\mathbb{P}'_1}{\mathrm{d}\mathbb{P}_0}\right)^2\right] - 1\right)} \ ,$$

where $\mathrm{d}\mathbb{P}_0$ and $\mathrm{d}\mathbb{P}'_1$ are the conditional marginal densities of $\{Y_t\}_{t\in[m]}$ given Ω under the null and the alternative respectively. Hence we have

$$2(1 - 2\varepsilon/\mathbb{P}(\Omega))^2 + 1 \leq \mathbb{E}_0\left[\left(\frac{\mathrm{d}\mathbb{P}'_1}{\mathrm{d}\mathbb{P}_0}\right)^2\right] \ ,$$

and so all that remains is to get a good upper bound on the expectation on the right. We let Ω be the same event as in Lemma 4.2 with $c = 6 + 3\log 2$, that is

$$\Omega = \Big\{\{N - 1 > mp/2\} \cap \{\forall j : \ l_j \leq cm/N\}\Big\} \ .$$

By Lemma 4.2 we have $\mathbb{P}(\Omega) \geq 1/4$.

We use the shorthand notations $\mathbf{Y} = \{Y_t\}_{t\in[m]}, \mathbf{l} = \{l_j\}_{j\in[N]}$ and $\omega = \{\omega_j\}_{j\in[N]}$. As before, $\mathrm{d}\mathbb{P}'_1$ is a mixture, and we can express it in a similar fashion as we did in Section 4.4.1.

In particular, denoting the density of $N(\mu, 1)$ by f_μ, the conditional density of Y given N, l and ω is

$$\mathrm{d}\mathbb{P}_1(\mathbf{Y}|N, \mathbf{l}, \omega) = \prod_{j\in 1}^{N} \prod_{t:N(t)=j} \left(\mathbf{1}\{\omega_j = 1\} f_\mu(Y_t) + \mathbf{1}\{\omega_t = 0\} f_0(Y_t)\right) \ .$$

Hence we can write $\mathrm{d}\mathbb{P}'_1(\mathbf{Y}) = \mathbb{E}_{N,\mathbf{l}|\Omega}\left[\mathbb{E}_{\omega|N}\left[\,\mathrm{d}\mathbb{P}_1(\mathbf{Y}|N, \mathbf{l}, \omega)\right]\right]$, where the subscripts

of the expectations indicate which random variables are integrated out by that particular expectation. In detail, the inner expectation $\mathbb{E}_{\omega|N}$ is taken w.r.t. the conditional distribution $\omega|N$ (recall $\omega_j \sim \mathrm{Ber}(q_j)$ and independent for $j \in [N]$). The outer expectation $\mathbb{E}_{N,\mathbf{l}|\Omega}$ is taken w.r.t. the joint conditional distribution of $N, \mathbf{l}$ given the event Ω.

We upper bound $\mathbb{E}_0\big[(\,\mathrm{d}\mathbb{P}_1'/\,\mathrm{d}\mathbb{P}_0)^2\big]$ similarly to how it was done in Section 4.4.1. Let ω' be a vector distributed identically to ω (given N) independent of ω, and let f_μ denote the density of $N(\mu, 1)$. We have

$$\begin{aligned}
&\mathbb{E}_0\left[\left(\frac{\mathrm{d}\mathbb{P}_1'(\mathbf{Y})}{\mathrm{d}\mathbb{P}_0(\mathbf{Y})}\right)^2\right] = \\
&= \mathbb{E}_0\left[\left(\frac{\mathbb{E}_{N,\mathbf{l}|\Omega}\left[\mathbb{E}_{\omega|N}[\,\mathrm{d}\mathbb{P}_1(\mathbf{Y}|N,\mathbf{l},\omega)]\right]}{\mathrm{d}\mathbb{P}_0(\mathbf{Y})}\right)^2\right] \\
&\leq \mathbb{E}_0\left[\mathbb{E}_{N,\mathbf{l}|\Omega}\left[\mathbb{E}_{\omega|N}\left[\frac{\mathrm{d}\mathbb{P}_1(\mathbf{Y}|N,\mathbf{l},\omega)}{\mathrm{d}\mathbb{P}_0(\mathbf{Y})}\right]^2\right]\right] \\
&= \mathbb{E}_0\left[\mathbb{E}_{N,\mathbf{l}|\Omega}\left[\mathbb{E}_{\omega|N}\left[\frac{\mathrm{d}\mathbb{P}_1(\mathbf{Y}|N,\mathbf{l},\omega)}{\mathrm{d}\mathbb{P}_0(\mathbf{Y})}\right]\mathbb{E}_{\omega'|N}\left[\frac{\mathrm{d}\mathbb{P}_1(\mathbf{Y}|N,\mathbf{l},\omega')}{\mathrm{d}\mathbb{P}_0(\mathbf{Y})}\right]\right]\right]\,,
\end{aligned}$$

using Jensen's inequality. In the above expression $\mathrm{d}\mathbb{P}_1(\mathbf{Y}|N,\mathbf{l},\omega)/\,\mathrm{d}\mathbb{P}_0(\mathbf{Y})$ are products of the marginal densities of $Y_t,\ t \in [m]$, which take the value 1 when $\omega_{N(t)} = 0$ and take the value $f_\mu(Y_t)/f_0(Y_t)$ when $\omega_{N(t)} = 1$. So the expression on the right side above is equal to

$$\mathbb{E}_0\left[\mathbb{E}_{N,\mathbf{l}|\Omega}\left[\mathbb{E}_{\omega,\omega'|N}\left[\left(\prod_{j=1}^{N}\prod_{t:\ N(t)=j}\mathbf{1}\{\omega_j=\omega_j'=1\}\left(\frac{f_\mu(Y_t)}{f_0(Y_t)}\right)^2\right)\right.\right.\right.$$
$$\left.\left.\left.\left(\prod_{j=1}^{N}\prod_{t:\ N(t)=j}\mathbf{1}\{\omega_j\neq\omega_j'\}\frac{f_\mu(Y_t)}{f_0(Y_t)}\right)\right]\right]\right]\,.$$

Note that the outermost expectation $\mathbb{E}_0$ is with respect to the density of $\mathbf{Y}$ under the null. This does not depend on any of the other variables. Hence we can change

the order of integration and continue as

$$\begin{aligned}
&\mathbb{E}_0\left[\left(\frac{\mathrm{d}\mathbb{P}_1'(\mathbf{Y})}{\mathrm{d}\mathbb{P}_0(\mathbf{Y})}\right)^2\right] \leq \\
&\quad \leq \mathbb{E}_{N,\mathbf{l}|\Omega}\left[\mathbb{E}_{\omega,\omega'|N}\left[\mathbb{E}_0\left[\prod_{j=1}^{N'}\prod_{t:\ N(t)=j}\mathbf{1}\{\omega_j=\omega_j'=1\}\left(\frac{f_\mu(Y_t)}{f_0(Y_t)}\right)^2\right]\right]\right] \\
&\quad = \mathbb{E}_{N,\mathbf{l}|\Omega}\left[\mathbb{E}_{\omega,\omega'|N}\left[\prod_{j=1}^{N}\prod_{t:\ N(t)=j}\exp(\mathbf{1}\{\omega_j=\omega_j'=1\}\mu^2)\right]\right] \\
&\quad = \mathbb{E}_{N,\mathbf{l}|\Omega}\left[\prod_{j=1}^{N}(1+q_j^2(e^{l_j\mu^2}-1))\right]
\end{aligned}$$

The expectation is readily upper bounded using the fact that we are under the event Ω (recall that Ω puts an upper bound on the interval lengths $\mathbf{l}$ and a lower bound on the number of intervals N). We get

$$\mathbb{E}_0\left[\left(\frac{\mathrm{d}\mathbb{P}_1'(\mathbf{Y})}{\mathrm{d}\mathbb{P}_0(\mathbf{Y})}\right)^2\right] \leq \left(1+\frac{4c^2}{p^2n^2}\left(e^{2c\mu^2/p}-1\right)\right)^m ,$$

which when combined with the previous inequality yields

$$\mu \geq \sqrt{\frac{p}{2c}\log\left(\frac{p^2n^2}{4c^2}(\sqrt[m]{2(1-8\varepsilon)^2+1}-1)+1\right)} .$$

Using $\log x \leq x-1$ for $x = \sqrt[m]{2(1-8\varepsilon)^2+1}$, the previous inequality implies

$$\mu \geq \sqrt{\frac{p}{2c}\log\left(\frac{\log(2(1-8\varepsilon)^2+1)p^2n^2}{4c^2m}+1\right)} ,$$

which is what we wanted to show. □

Non-extreme dynamics, general sparsity

A rigorous result for the case of general sparsity levels remains out of reach at this point. However, we conjecture that in the s-sparse case, essentially the same result should hold as in Theorem 4.2 with n replaced by n/s. That is to say, we conjecture

that the signal strength needs to scale as $\sqrt{p \log(n/s)}$ in the regime $m \approx n/s$.

Although lacking a formal proof, we highlight the reason behind the above conjecture. Consider a model for the evolution of the signal support, which is the s-fold concatenation of 1-sparse problems of size n/s. Formally, let $\{B_i\}_{i \in [s]}$ be consecutive blocks of $[n]$, each having size n/s. Formally,

$$B_i = \{k : k = (i-1)n/s + j, j \in [n/s]\} \ ,$$

for $i \in [s]$. Consider a single active component that evolves according to the dynamic described in Section 4.2, restricted to the block B_1. Denote the position of this active component at time t by $S_1^{(t)}$. Define an s-sparse support for every $t \in [m]$ by setting $S^{(t)} = \{S_1^{(t)}, \ldots, S_s^{(t)}\}$ with $S_i^{(t)} = (i-1)n/s + S_1^{(t)}$ for every $i \in [s]$.

This model is essentially a 1-sparse model on a set of size n/s repeated s times. Note that in this model every measurement action A_t can be treated as if it was made inside B_1, since if $A_t \in B_i,\ i \neq 1$ then $A'_t = A_t(\mod n/s)$ yields the same type of observation (in terms of having elevated mean or not). Thus Theorem 4.2 directly applies with n replaced by n/s in this model.

Though this is a different problem then the one we would like to derive a lower bound for, heuristically this seems to be an easier problem, as the evolution of the signal support is much more restricted. Hence we believe a lower bound in this model would also be a lower bound for the one that we are considering.

4.5 A Numerical Experiment

As in Chapters 2 and 3, we present a short numerical experiment to corroborate the theoretical results presented in this chapter.

In this simulation we compare the performance of the adaptive sensing procedure of Section 4.3 with a reasonable non-adaptive procedure[1].

Recall that we are concerned with the setting when m, the time horizon, is of the order n/s. According to Proposition 4.1, we need m to be roughly $(n/s) \cdot \log_2(2/\varepsilon)$ for the adaptive sensing procedure to have error probability at most ε. However, by simply choosing $m = n/s$, the probability of not sampling an active component would be high (see the beginning of Section 4.4). Hence we use $m = (n/s) \cdot \log_2(2/\varepsilon)$

[1]Though we could not yet formally prove that this non-adaptive procedure is indeed optimal, we conjecture that it is (see Section 4.4.1).

with $\varepsilon = 0.05$.

The non-adaptive procedure simply selects components to sample at random at each time instance. Our conjecture in Section 4.4.1 is that a sub-sampling procedure should be optimal in the non-adaptive setting. Note that with this choice of m, this is in fact a sub-sampling procedure. Even if not optimal, this is a reasonable non-adaptive sampling strategy.

Once we have the sample, we simply scan through the signal and declare a detection whenever we find a large enough value. More precisely, with $\mathbf{Y} \in \mathbb{R}^m$ denoting the sample, we declare that a signal is present when

$$\max_{i \in [m]} Y_i \geq \sqrt{2 \log \frac{m}{2\varepsilon}} \ .$$

The threshold on the right side is motivated by controlling the type I error. A union bound with a Gaussian tail bound yields

$$\mathbb{P}_0 \left(\max_{i \in [m]} Y_i \geq z \right) \geq \frac{m}{2} e^{-z^2/2} \ .$$

With the above choice for the threshold, we ensure the type I error of the procedure above is at most ε.

The adaptive procedure is implemented as described in Section 4.3. In detail, we apply STT (see Algorithm 3) repeatedly to randomly selected components, until either one of them declares "Signal" in which case we stop and declare a signal is present, or reach our time horizon in which case we declare no signal is present.

According to Proposition 4.1, the type II error probability of the adaptive sensing procedure should drop below the value ε when the signal strength is about

$$\mu_{\text{limit}} = \sqrt{\frac{2}{\min\{1/(2p), \log(n/s)\}} \log(n/s)} \ . \tag{4.6}$$

Hence, set the signal strength to be $t \cdot \mu_{\text{limit}}$ and run the procedures above with different values for t. Since both procedures are designed to ensure small type I error probability, we only plot the type II errors. We choose $n = 2^{15}$ and $s = 2^4$. We run two different simulations, first when the speed of change p is equal to 0.2 and second when it is equal to 0.5. Note that with these parameters $\mu_{\text{limit}} = \sqrt{4p \log(n/s)}$.

In every instance we run 100 iterations for every value of the parameter t, and plot the average type II errors of the different tests. In order to illustrate the

variability of the curves, we also plot error bars whose total length is four times the (point-wise) standard error. These would correspond to a roughly 95% two-sided confidence interval for normally distributed measurements, and hence are approximate point-wise confidence bands.

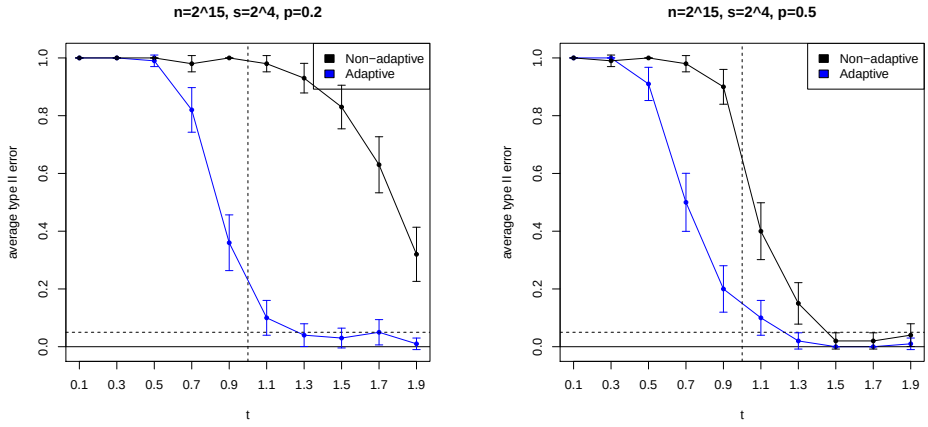

Figure 4.3: Average type II error probabilities (with SE bands) for the different estimators as a function of the parameter t (the signal strength is $t \cdot \mu_{\text{limit}}$ with μ_{limit} defined in (4.6)): the non-adaptive test (black); the adaptive sensing test based on the STT (blue). The plots from left to right correspond to $p = 0.2$ and $p = 0.5$. The number of repetitions is 100 for each value of t. The vertical black dashed line is at the value $t = 1$. The horizontal black dashed line is at the value of ε (0.05).

As expected, the adaptive sensing test outperforms the non-adaptive one, unless the speed of change p is high. The margin by which the adaptive sensing test is superior is greater as p decreases, which is what the theory suggested. Note that the type II error probability of the tests never quite reach the value zero, since with this choice of m, the probability of being able to sample an active component is not overwhelming.

4.6 Final Remarks

In this chapter we have studied the problem of the detection of signals that evolve dynamically over time. We have introduced a simple stochastic model for the

evolution of the signal support, and have analyzed the signal detection problem in this framework, with special interest in the effect the speed of the change of the support has on the problem difficulty. Though some results remain conjectures at this point, our aim as been to provide convincing heuristics as to why we believe the conjectures to be true, and hope to provide formal proofs in the near future.

Our results suggest that the speed of change has an increasing effect on the problem difficulty, when using adaptive sensing. In particular, in an adaptive sensing setting we take advantage of situations when the signal is changing slowly. This effect becomes less and less pronounced as the speed of change grows. Contrasting this, such gains can not be realized in a non-adaptive sensing context. That is, the performance of non-adaptive sensing procedures is essentially the same regardless of the speed of change, and this performance corresponds to the worst-case performance of adaptive algorithms (the case when the speed of change is the highest).

We highlight a few possibilities for future work regarding dynamically evolving signals:

Higher number of measurements: As noted before, we need at least n/s measurements to perform detection reliably, and the algorithm that we propose is only optimal when the number of measurements that we are allowed to collect is of the same order as this quantity. Understanding how the model parameters affect the problem difficulty in general remains an open question.

Restricted dynamics: In the model considered in this chapter when signal components change they can move to any unoccupied location in the signal vector. This assumption simplifies the setup, but in some applications might not be restrictive enough. For instance if signal components could only move to a location in the vicinity of where they were before, then this might make the effect of the speed of change less pronounced in the difficulty of detection. Understanding the effect of such restrictions could prove valuable in certain applications, while also being interesting from a theoretical point of view.

Structures: In certain situations the signal support can be assumed to have structure to it (as in Chapters 2, 3 and 5). For instance all anomalous items might be consecutive. In some cases the structure of the support has a huge effect on the difficulty of recovery, as illustrated in Chapters 2 and 3. In those measurement

models however, we know that structure plays (almost) no role in the difficulty of detection. How structural restrictions affect these tasks for dynamically evolving signals could be an interesting avenue of research.

Support recovery: Another common question in similar settings is how well we can estimate the support of a signal. That is, instead of deciding whether there are anomalous items or not, we need to determine which of the items are anomalous. This would also be an interesting problem to study for dynamically evolving signals, although how to cast the problem in a meaningful way is less immediate.

Chapter 5

Distribution-Free Detection of Structured Anomalies

This chapter is based on joint work with Ery Arias-Castro (UCSD), Meng Wang (UCSD), Rui Castro (TU/e). The results presented here can also be found in Arias-Castro et al. [11]. The numerical experiments of Section 5.6 were performed by Ery Arias-Castro (UCSD) and Meng Wang (UCSD).

5.1 Introduction

In the previous chapters we have investigated the properties of adaptive sensing in several signal detection and support recovery settings. This was done in the framework of clear-cut distributional models. Although we have made some remarks about possibilities to relax the distributional assumptions, a more thorough understanding of distribution-free adaptive sensing methods would be very useful in practice, as often one can not rely on such assumptions being fulfilled. A first step in this direction is to understand how such methods perform in the non-adaptive sensing setting, and this is the topic of this chapter.

We consider the problem of detecting anomalous behavior which is endowed with some structure in a high-dimensional signal. A standard way to tackle this problem in the non-adaptive sensing setting is the use of a scan statistic, which essentially inspects all possible anomalous patterns. It usually corresponds to a form of generalized likelihood ratio test (see Kulldorff [99]). Although computa-

tionally this might present a challenge, there are a number of situations where this is possible in nearly linear time (see Neill & Moore [115], Neill [113], Arias-Castro, Donoho & Huo [12] and Walther [143]).

We are interested in ways to calibrate[1] the scan statistic when the underlying distribution of the observations is unknown. For the purpose of illustration, consider the following prototypical example[2]: suppose we have event data over a certain time period and want to detect if there is a time interval with an unusually high concentration of events. To make things more concrete and move towards the setting we consider in this chapter, assume one can model this event data as a realization of a Poisson process and bin the data, so that we observe a sequence of Poisson random variables. The scan statistic in this particular case combines sums of these values over (discrete) intervals of different size and location, together with some normalization - see (5.2) further down. In this scenario we want to perform a hypotheses test, where the null hypothesis is that no anomaly is present (a homogeneous Poisson process) versus the alternative where some time intervals have an elevated rate of events (an inhomogeneous process). If the (constant) rate is known under the null, then the null distribution is completely specified and the test can be calibrated either analytically or by Monte-Carlo simulation. But what if the null event rate is unknown?

One can regard the scan statistic as a comparison between observations in one interval to those outside the interval. This point of view naturally leads to a two-sample problem for each interval, which is then followed by some form of multiple testing since we scan many intervals. Thus drawing from the classical literature on the two-sample problem, two approaches can be considered:

- *Calibration by permutation.* This amounts to using the permutation distribution of the scan statistic for inference (detection/estimation).
- *Scanning the ranks.* This amounts to replacing each observation by its rank before scanning. As any rank-based method, calibration can generally be done by Monte-Carlo simulation before the observation of data.

The perspective offered by the two-sample testing framework makes these two procedures very natural. Although less popular, as in two-sample testing, a procedure based on ranks offers some advantages over a pure calibration by permutation: it

[1] By calibration, we mean specifying the critical region of the test.

[2] In fact, this setting might have been the original motivation for the work on the scan statistic, see Wallenstein [142].

is more robust to outliers and its calibration only needs to be done once for a given sample size.[3] The latter is rather advantageous if one desires to apply the test repeatedly on several datasets of the same size. Compare with a calibration by permutation: typically, several hundred permutations are sampled at random and, for each one of them, the scan statistic is computed - and this is done each time the test is applied.

The intrinsic difficulty of the detection task depends on two things: the data distribution and the complexity of the class of anomalous sets. Regarding the data distribution, we consider the situation where the data comes from the natural exponential family. As for the class of anomalous sets, since the main motivation of our work is to develop methods and theory for the scenario when the null distribution is unknown/unspecified, we focus on the simplest and most emblematic setting, that of detecting an interval in a one-dimensional regularly sampled signal. Generalizations of our results to more complex settings (e.g., rectangles in two or more dimensions, or even blob-like subsets) are possible, and we later explain how this can be done.

In this chapter we study the performance of the two methods above and provide strong asymptotic theoretical guarantees as well as insights on the their finite-sample performance in some numerical experiments. In the context of a natural exponential family - which includes the classical normal location model and the Poisson example above - we find that the permutation scan test and the rank scan test come very close to performing as well as the oracle scan test, which we define as the scan test calibrated by Monte Carlo with (clairvoyant) knowledge of the null distribution. We perform numerical experiments on simulated data, confirming our theory, and also some experiments using a real dataset from genomics.

Related work: The permutation scan has been suggested in a number of papers and applied in a number of ways in different contexts. For example, it is suggested by Kulldorff et al. in [100, 101, 86] in the context of syndromic surveillance; by Walther in [143] in the context of sensor network monitoring with binary observations; and by Flenner & Hewer in [71] in the context of detecting a change in a sequence of images. We note that permutation tests are known to perform well in classical two-sample testing (see Lehmann & Romano [102]). However, in the context of the scan test, we are only aware of one other paper, that of Walther

[3]The latter explains why, in two-sample testing, methods based on ranks were feasible decades before methods based on permutations, which typically require access to a computer.

[143], that develops theory for the permutation scan test. This is done in the context of binary data (a Bernoulli model). Our analysis extends the theory to any natural exponential model as described in Section 5.2.1 (which also includes the binary case). This requires a different set of tools. Jung & Cho [92] also proposed a rank based method, but without any theoretical justification. That said, rank tests such as Wilcoxon's are known to perform well in classical two-sample testing (see Hettmansperger [83] and Lehmann & Romano [102].

As noted in Chapter 1, in several cases one can make some structural assumptions on the anomalous sets. For instance, grid-like networks are an important special case, arising in applications such as signal and image processing (where the signals are typically regularly sampled) and sensor networks deployed for the monitoring of some geographical area. This situation is considered in great generality and from different perspectives by several authors (see for instance Arias-Castro, Candès & Durand [8], Walther [143], Arias-Castro, Donoho & Huo [12], Desolneux, Moisan & Morel [61], Perone et al. [120], Cai & Yuan [33], Hall & Jin [78]). Also, the distribution of the corresponding scan statistic (5.2) and variants has been studied in a number of places (see Jiang [91], Boutsikas & Koutras [29], Siegmund & Venkatraman [133], Kabluchko [93], Arias-Castro & Sharpnack [14]).

In this chapter we focus exclusively on the detection of intervals, for the sake of clarity and simplicity, but our techniques and results apply naturally to more general anomaly classes. As shown by Arias-Castro, Candès & Durand in [8], similar results apply to a general (nonparametric) class $\mathcal{C}$ of blob-like ('thick') sets S when the signal corresponds to measurements from a grid-like set of arbitrary finite dimension, although the scanning is done over an appropriate approximating net for $\mathcal{C}$ (instead of the entire class $\mathcal{C}$). Furthermore, these results generalize to one-parameter exponential models, beyond the commonly assumed normal location model, as long as the sets $S \in \mathcal{C}$ are sufficiently large (poly-logarithmic in n). Other authors that develop theory for different environments include Sharpnack, Krishnamurthy & Singh [130, 131], Arias-Castro et al. [9], Addario-Berry et al. [1], Zhao & Saligrama [150]. Variants of this detection problem have been suggested, and the applied literature is quite extensive. We refer the reader to Arias-Castro, Candès & Durand [8] and references therein.

As specified below, we focus on a "static" setting, where the length of the signal being monitored is fixed a priori. Adding time is typically done by adding one 'dimension' to the framework, as done for example by Kulldorff et al. in [100].

Organization: The rest of the chapter is organized as follows. In Section 5.2 we describe the problem setting in detail. In Section 5.3 we consider the case when the null distribution is known. This section is expository, introducing the reader to the basic proof techniques that are used, for example by Arias-Castro, Candès & Durand in [8], to establish the performance of the scan statistic when calibrated with full knowledge of the null distribution - the oracle scan test, as we called it here. To keep the exposition simple, and to avoid repeating the substantially more complex arguments detailed in that paper and others, we focus on the problem of detecting an interval in a one-dimensional lattice. This allows us to set the foundation and discover what the performance bounds for the scan test in this case rely on. In Section 5.4 we consider the same setting and instead calibrate the scan statistic by permutation. In Section 5.5 we consider the same setting and instead scan the ranks. In both cases, our analysis relies on concentration inequalities for sums of random variables obtained from sampling without replacement from a finite set of reals, already established in the seminal paper of Hoeffding [84]. In Section 5.6 we perform some simulations to numerically quantify how much is lost in finite samples when calibrating by permutation or when using ranks. We also compare our methodology with that of Cai, Jeng & Li [32], on simulated data, and also on a real dataset from genomics. We conclude with a brief discussion in Section 5.7.

5.2 Problem setting

A typical framework for static anomaly detection - which includes detection in digital signals and images, sensor networks, biological data, and more - may be described in general terms as follows. We observe a set of random variables, denoted by $\mathbf{Y} = \{Y_i\}_{i \in [n]}$, which is a snapshot of the state of the environment. In this chapter we take a hypothesis testing point of view. Under the null hypothesis, corresponding to the nominal state when no anomalies are present, these random variables are independent and identically distributed (i.i.d.) with some null distribution F_0. Under the alternative, some of these random variables will have a different distribution. Let $\mathcal{C} \subset 2^{[n]}$ denote a class of possibly anomalous subsets, corresponding to the anomalous patterns that we expect to encounter (this would be a class of intervals in the example that we have used earlier). Under the alternative hypothesis, there is a subset $S \in \mathcal{C}$ such that, for each $i \in S$, $Y_i \sim F_{\theta_i}$, for

some distribution $F_{\theta_i} \neq F_0$, and independent of $\{Y_i\}_{i\in[n]\setminus S}$, which are still i.i.d. with distribution F_0. In a number of important applications, the variables are real-valued and the anomalous ones take larger-than-usual values, which can be formalized by assuming that each F_{θ_i} stochastically dominates[4] F_0. We take this to be the case throughout the paper.

Note that in the formulation above the alternative hypothesis is composite. Tackling this problem using a generalized likelihood ratio approach is popular in practice (see Kulldorff [99]) and often referred to as the scan test, as it works by scanning over the possible anomalous sets to determine if there is a set that is able to "explain" the observed data. Assuming that F_{θ_i}'s are equal for every $i \in S$ under the alternative, and that all subsets in the class $\mathcal{C}$ have the same size, some simplifications lead to considering the test that rejects for large values of the scan statistic

$$\max_{S\in\mathcal{C}} \sum_{i\in S} Y_i \ . \tag{5.1}$$

When the subsets in the class $\mathcal{C}$ may have different sizes, a more reasonable approach includes a normalization of the partial sums above, leading to the following variant of the scan statistic

$$\max_{S\in\mathcal{C}} \frac{1}{\sqrt{|S|}} \sum_{i\in S} (Y_i - \mathbb{E}_0(Y_i)) \ , \tag{5.2}$$

where $\mathbb{E}_\theta$ denotes the expectation with respect to F_θ, and for a discrete set S, $|S|$ denotes its cardinality. As argued by Arias-Castro & Grimmett in [13], this test is in a certain sense asymptotically equivalent to the generalized likelihood ratio test, but slightly simpler.

5.2.1 Exponential models

An important special case of the general framework just described is that of a one-parameter exponential model in natural form. In detail, consider a probability measure F_0 on the real line with finite moments. We assume that either F_0 is continuous (i.e., diffuse) or discrete (i.e., with discrete support). For $\theta \in (0, \theta_\star)$ define

$$f_\theta(x) = \exp(\theta x - \log \varphi_0(\theta)) \ ,$$

[4]For two distribution (functions) on the real line, F and G, we say that G stochastically dominates F if $G(t) \leq F(t)$ for all $t \in \mathbb{R}$. We denote this by $G \succeq F$.

	F_0	F_θ
Normal	$\mathrm{N}(\mu_0,\sigma_0^2)$	$\mathrm{N}(\mu_0+\theta\sigma_0^2,\sigma_0^2)$
Poisson	$\mathrm{Pois}(\lambda_0)$	$\mathrm{Pois}(\lambda_0 e^\theta)$
Bernoulli	$\mathrm{Ber}(p_0)$	$\mathrm{Ber}(p_0 e^\theta/(1+p_0(e^\theta-1)))$

Some examples of the exponential models of Section 5.2.1.

where $\varphi_0(\theta) = \int e^{\theta x}\,\mathrm{d}F_0(x)$ and $\theta_\star = \sup\{\theta > 0 : \varphi_0(\theta) < \infty\}$, assumed to be strictly positive (and possibly infinite).

In this setting, F_{θ_i} is the distribution whose density w.r.t. F_0 is f_{θ_i} as defined above, where $\theta_i > 0$. Since a natural exponential family has the monotone likelihood ratio property[5], it follows that F_θ is stochastically increasing in θ (see Lemma 3.4.2 in Lehmann & Romano [102]). In particular, we do have $F_\theta \succeq F_0$ for all $\theta > 0$.

Important special cases of such an exponential model includes the normal location model standard in many signal and image processing applications; the Poisson model popular in syndromic surveillance (see Kulldorff et al. [100]); and the Bernoulli model (see Walther [143]). The distributions that correspond to the parametrization defined above are summarized in Table 5.2.1.

5.2.2 Detection of intervals

Let $\mathcal{C}$ be the class of all discrete intervals of $[n]$, meaning

$$\mathcal{C} = \big\{\{a,\dots,b\} : 1 \le a \le b \le n\big\} \ .$$

If one assumes a normal location model then the scan test corresponding to (5.2) rejects the null for large values of

$$\max_{1\le a\le b\le n} \frac{1}{\sqrt{b-a+1}} \sum_{i=a}^{b} Y_i \ . \tag{5.3}$$

[5] A family of densities $(f_\theta : \theta \in \Theta)$, where $\Theta \subset \mathbb{R}$, has the monotone likelihood ratio property if $f_{\theta'}(x)/f_\theta(x)$ is increasing in x when $\theta' > \theta$.

Following the parameterization in Arias-Castro, Donoho & Huo [12] one assumes

$$\frac{1}{|S|}\sum_{i\in S}\theta_i \geq \tau\sqrt{\frac{2}{|S|}\log n}, \quad \text{with } \tau > 0 \text{ fixed}, \tag{5.4}$$

where S denotes the anomalous set (this ensures that all intervals are roughly equally difficult to detect). In a minimax setting it can be shown that the detection boundary is at $\tau = 1$, meaning that when $\tau < 1$ no test can simultaneously attain arbitrarily small type I and type II error probabilities in the large sample limit $n \to \infty$, while there is such a test when $\tau > 1$ (meaning such a test has risk tending to 0). In fact, the scan test (5.3) is one of them. We remark that in this model the short intervals (the most numerous in $\mathcal{C}$) drive the difficulty of the problem and a refinement is possible. See Arias-Castro, Candès & Durand [8] and Walther [143] for details.

As we will see later (Section 5.3) it is often advantageous to scan over a so-called approximating net of the class $\mathcal{C}$ instead of the entire class itself. An approximating net is simply a class such that for every set $S \in \mathcal{C}$ there is a set in the approximating net that is close to S in some sense. Analyzing a scan test restricted to an approximating net has the following advantages: the analysis is simpler as it does not require the use of chaining to achieve tight constants; it is applicable in more general settings, in particular, when the class $\mathcal{C}$ is nonparametric; and it is computationally advantageous, giving rise to potentially fast implementations.

5.2.3 Calibration by permutation

Suppose we are considering a test that rejects the null for large values of a test statistic $T(\mathbf{Y})$. Let $\mathbf{y} = \{y_i\}_{i\in[n]}$ be the observed value of $\mathbf{Y}$. If we were to know the null distribution F_0, we would return the p-value as $\mathbb{P}_0(T(\mathbf{Y}) \geq T(\mathbf{y}))$. In practice, more than knowing the null distribution, what really matters is that we can (efficiently) simulate from it, so that we can estimate this p-value by Monte-Carlo simulation.

Ignoring computational constraints for the moment, calibration by permutation amounts to computing $T(\mathbf{y}^{(\pi)})$ for all $\pi \in [n]!$, where $[n]!$ denotes the set of all permutations of $[n]$ and $\mathbf{y}^{(\pi)} = \{y_{\pi(i)}\}_{i\in[n]}$ is the permuted data. We then return the p-value

$$\frac{1}{n!}\left|\left\{\pi \in [n]! : \ T(\mathbf{y}^{(\pi)}) \geq T(\mathbf{y})\right\}\right|$$

and our decision is based on this value. Let $M = |\{T(\mathbf{y}^{(\pi)}) : \pi \in [n]!\}|$. If there are no multiplicities, meaning $M = n!$, it can be shown such tests are exact and that under the null the p-value has a (discrete) uniform distribution on $\{1/M, 2/M, \ldots, 1\}$. Otherwise the test will be conservative (see Lehmann & Romano [102]). In practice, the number of permutations is very large ($n!$) and the p-value is estimated by simulation (by uniform sampling of permutations).

In our setting, T above will be a form of a scan statistic. Assuming that T has been chosen, we define the oracle scan test as the scan test calibrated with full knowledge of the null distribution. We also define the permutation scan test as the scan test calibrated by permutation as explained above. In this chapter we characterize the performance of the permutation scan test, concluding that it has as much asymptotic power as the oracle scan test (Theorem 5.1) in most scenarios.

5.2.4 Scanning the ranks

As explained earlier, when calibrating by permutation, the scan is performed on each permutation of the original dataset. Even though this is done for only a relatively small number of permutations, that number is often chosen in the hundreds, if not thousands, meaning that the procedure requires the computation of that many scans. Even if the computation (in fact, approximation) of the scan statistic is done in linear time, this can be rather time consuming. Furthermore, for a new instantiation of the data the whole procedure must be undertaken anew. The computational burden of doing so may be prohibitive in some practical situations, for instance when monitoring a sensor network in real-time.

We propose instead a rank-based approach, which avoids the expensive calibration by permutation. The procedure amounts to simply replacing the observations with their ranks[6] before scanning, so that we end up scanning the ranks instead of the original values. As any other rank-based method, the resulting procedure is distribution-free and therefore only needs to be calibrated once for each data size n even though it can be viewed as a permutation procedure. Such a procedure is very natural given the classical literature on nonparametric tests (see Hettmansperger [83]), and from the two-sample perspective offered earlier, it is directly inspired by the rank-sum test introduced by Wilcoxon in [145].

In detail, let R_i denote the rank (in increasing order) of Y_i among $\mathbf{Y}$. If there are ties, they can be dealt with in any of the classical ways, for instance, by

[6] Throughout, the observations are ranked in increasing order of magnitude.

assigning them the average rank. For technical and also practical reasons our results are proven under the assumption that ties are broken randomly. If $T(\mathbf{Y})$ is a form of scan statistic, we then consider the rank scan, defined as $T(\mathbf{R})$, where $\mathbf{R} = \{R_i\}_{i\in[n]}$. For example, the rank variant of (5.2) is

$$\max_{S\in\mathcal{C}} \frac{1}{\sqrt{|S|}} \sum_{i\in S} (R_i - \tfrac{n+1}{2}) \ ,$$

since $\mathbb{E}_0(R_i) = \frac{n+1}{2}$. Assuming that T has been chosen, we define the rank scan test as the scan test based on the ranks. Again, the test is calibrated by permutation, since this corresponds to the null distribution once the observations are replaced by their ranks. In this chapter we establish the performance of the rank scan test, concluding that it has nearly as much asymptotic power as the oracle scan test (Theorem 5.2 and Proposition 5.2). Our results allow us to precisely quantify how much (asymptotic) power is lost when using the rank scan test versus the oracle scan test. For example, in the normal means model the rank-scan test requires a signal magnitude roughly 1.023 times larger than the regular scan test to be asymptotically powerful against anomalous sets that are not too small. In fact, in our empirical analysis of the finite sample properties of the rank-scan we actually found it is slightly more powerful than the oracle scan test.

5.3 When the null distribution is known

This section is meant to introduce the reader to the techniques underlying the performance bounds presented by Arias-Castro et al. in [8] and [12] for the scan statistic (and variants) when the null distribution is known. These provide a stepping stone for our results in regards to permutation and rank scan tests. We detail the setting of detecting an interval of unknown length in a one-dimensional lattice. Therefore, as in Section 5.2.2, consider the setting where

$$\mathcal{C} = \big\{\{a, \dots, b\} : \ 1 \le a \le b \le n\big\} \ .$$

We begin by considering the normal model - $Y_i \sim N(\theta_i, 1)$ independent - and explain later on how to generalize the arguments to an arbitrary exponential model

as described in Section 5.2.1. We are interested in testing

$$H_0 : \theta_i = 0, \forall i \in [n] \quad \text{versus} \quad H_1 : \exists S \in \mathcal{C} \text{ such that } \tfrac{1}{|S|} \sum_{i \in S} \theta_i \geq \tau \sqrt{\frac{2}{|S|} \log n} \ . \tag{5.5}$$

We consider this problem from a minimax perspective. It is shown by Arias-Castro, Donoho & Huo in [12] that, when $\tau < 1$, any test with level α has power at most $\beta(\alpha, n)$, with $\beta(\alpha, n) \to \alpha$ as $n \to \infty$. In other words, in the large-sample limit, no test can do better than random guessing, the test that rejects with probability α regardless of the data. On the other hand, if $\tau > 1$, then for any level $\alpha > 0$ there exists a test with level α and power $\beta(\alpha, n) \to 1$ as $n \to \infty$. In particular, such a test can be constructed using a form of scanning.

5.3.1 Scanning over an approximating net

Instead of considering a test that scans over all elements in $\mathcal{C}$, as in (5.3), we describe a variant that consists of scanning over an approximating net for the class $\mathcal{C}$. This brings both computational and analytical advantages over scanning all sets in $\mathcal{C}$ as discussed in Section 5.2.2. We use an approximating net similar to that of Arias-Castro, Donoho & Huo [12]; see Arias-Castro & Sharpnack [14] for an alternative construction. The underlying metric on $\mathcal{C}$ is given by

$$\rho(S, S') := \frac{|S \cap S'|}{\sqrt{|S||S'|}} \ .$$

Step 1: Construction of an approximating net. Instead of scanning over $\mathcal{C}$ we will scan over a subclass of intervals $\mathcal{C}_b$, where $0 \leq b \leq n$ is an integer to be specified later on. Such a subclass must satisfy two important properties, namely have cardinality significantly smaller than $\mathcal{C}$, and be such that any element $S \in \mathcal{C}$ can be well approximated by an element of $\mathcal{C}_b$, in terms of the metric ρ defined above. To simplify the presentation assume n is a power of 2 (that is $n = 2^q$ for some integer q). We describe the construction similar to the one given by Arias-Castro, Donoho & Huo in [12].

Let $\mathcal{D}_j$ denote the class of dyadic intervals at scale j, meaning of the form $S = [1 + k2^j, (k+1)2^j] \subset [n]$ with j and k nonnegative integers. Let $\mathcal{D}_{j,0}$ denote the class of intervals of the form $S \cup S'$ with $S, S' \in \mathcal{D}_{j-1}$. In words, $\mathcal{D}_{j,0}$ contains the dyadic intervals at scale j (that is $\mathcal{D}_j \subset \mathcal{D}_{j,0}$) and the intervals that are obtained

by shifting the previous ones by half their length. Then, for $1 \leq k < b$, let $\mathcal{D}_{j,k}$ be the class of intervals of $[n]$ of the form $S_{\text{left}} \cup S \cup S_{\text{right}}$, where $S \in \mathcal{D}_{j,k-1}$ while S_{left} (resp. S_{right}) is adjacent to S on the left (resp. right) and is either empty or in $\mathcal{D}_{j-k}$. Note that $\mathcal{D}_{j,k-1} \subset \mathcal{D}_{j,k}$ by construction. This way, $\mathcal{D}_{j,b-1}$ contains intervals of the following form: we start with an interval from $\mathcal{D}_{j,0}$ (whose length is 2^j) and we can append intervals of length $2^{j-1}, 2^{j-2}, \ldots, 2^{j-b+1}$ consecutively to either end of the interval. In the final step, $\mathcal{D}_{j,b}$ is of the same form as before, only the appended intervals S_{left} and S_{right} are either empty, or in $\mathcal{D}_{j-b+1}$. That is, in the last step we can append another interval of length 2^{j-b+1}. Finally, define $\mathcal{C}_b = \bigcup_j \mathcal{D}_{j,b}$.

We can prove the following result for this approximating net, using similar arguments to those of Arias-Castro, Donoho & Huo [12].

Lemma 5.1. *The subclass $\mathcal{C}_b \subset \mathcal{C}$ has cardinality at most $n4^{b+1}$ and is such that for any element $S \in \mathcal{C}$ there is an element $S' \in \mathcal{C}_b$ satisfying $S \subset S'$ and $\rho(S, S') \geq (1 + 2^{-b+2})^{-1/2}$.*

Remark 5.1. *It is easy to see that the subclass $\mathcal{C}_b$ can be scanned in $O(nb4^b)$ operations - this is implicit in Arias-Castro, Donoho & Huo [12]. Indeed, we start by observing that scanning all dyadic intervals can be done in $O(n)$ operations by recursion, starting from the smallest intervals and moving up (in scale) to larger intervals. We then conclude by realizing that each interval in $\mathcal{C}_b$ is the union of at most $2b + 2$ dyadic intervals.*

Step 2: Definition of the scan test. We consider a test based on scanning only the intervals in $\mathcal{C}_b$. This test rejects the null if

$$\max_{S \in \mathcal{C}_b} \mathbf{Y}_S \geq \sqrt{2(1+\eta)\log n}\ , \tag{5.6}$$

where

$$\mathbf{Y}_S = \frac{1}{\sqrt{|S|}} \sum_{i \in S} Y_i\ ,$$

and $\eta > 0$ satisfies $\eta \to 0$ and $\eta \log n \to \infty$ (the reason for these conditions will become clear shortly).

Step 3: Under the null hypothesis. By the union bound,

$$\begin{aligned}\mathbb{P}_0\left(\max_{S\in\mathcal{C}_b}\mathbf{Y}_S \geq \sqrt{2(1+\eta)\log n}\right) &\leq \sum_{S\in\mathcal{C}_b}\mathbb{P}_0\left(\mathbf{Y}_S \geq \sqrt{2(1+\eta)\log n}\right)\\ &\leq |\mathcal{C}_b|\bar{\Phi}\left(\sqrt{2(1+\eta)\log n}\right)\ ,\end{aligned}$$

where Φ denotes the standard normal distribution function and $\bar{\Phi} = 1-\Phi$ denotes the corresponding survival function. We have the well-known bound on Mill's ratio:

$$\bar{\Phi}(x) \leq e^{-x^2/2}, \quad \forall x \geq 0\ . \tag{5.7}$$

Therefore we get

$$\mathbb{P}_0\left(\max_{S\in\mathcal{C}_b}\mathbf{Y}_S \geq \sqrt{2(1+\eta)\log n}\right) \leq n4^{b+1}n^{-(1+\eta)} = n^{-\eta}4^{b+1}\ .$$

We choose $b = \frac{1}{2}\eta\log n/\log 4$. With our assumption that $\eta\log n \to \infty$, this makes the last expression tend to zero as $n\to\infty$ (it also implies that $b\to\infty$, which we use later on). We conclude the test in (5.6) has level tending to 0 as $n\to\infty$.

Step 4: Under the alternative. We now show that the power of this test tends to 1 when $\tau > 1$. Let S denote the anomalous interval. Referring to Lemma 5.1, there is a set $S'\in\mathcal{C}_b$ such that $\rho(S,S') \geq (1+2^{-b+2})^{-1/2}$, so that $\rho(S,S') = 1+o(1)$ since $b\to\infty$. Furthermore $Y_{S'}$ is normal with mean larger than $\rho(S,S')\tau\sqrt{2\log n}$, and variance 1. We thus have

$$\mathbb{P}\left(Y_{S'} \geq \sqrt{2(1+\eta)\log n}\right) \geq \bar{\Phi}(\xi)\ ,$$

where

$$\begin{aligned}\xi &:= \sqrt{2(1+\eta)\log n} - \rho(S,S')\tau\sqrt{2\log n}\\ &= \sqrt{2(1+\eta)\log n}\big(1-(1+o(1))\tau/\sqrt{1+\eta}\big)\\ &\sim -(\tau-1)\sqrt{2\log n} \to -\infty\ ,\end{aligned}$$

where we have used the fact that $\tau > 1$ is fixed and $\eta\to 0$. We conclude that the test in (5.6) has power tending to 1 as $n\to\infty$. In conclusion, we have shown the following result:

Proposition 5.1 (Arias-Castro, Donoho & Huo [12])**.** *The test defined in* (5.6)*,*

with $\eta = \eta_n \to 0, \eta_n \log n \to \infty$ *and* $b = b_n = \frac{1}{2}\eta_n \log n$*, has level converging to 0 as* $n \to \infty$*. Moreover, if* (5.4) *holds with* $\tau > 1$*, then it has power converging to 1 as* $n \to \infty$*.*

We remark that, in principle, we may choose any $b = b_n \to \infty$ such that $b_n / \log n \to 0$. From Remark 5.1 the computational complexity of the resulting scan test is of order $O(nb_n 4^{b_n})$. For example, $b_n \sim \log \log n$ is a valid choice and the resulting scan test runs in $O(n \cdot \text{polylog}(n))$ time[7].

5.3.2 Generalizations

The arguments just given for the setting of detecting an anomalous interval under a normal location model can be generalized to the problem of detecting other classes of subsets under other kinds of distributional models. We briefly explain how this is done (note that these generalizations can be combined).

Other classes of anomalous subsets: For a given detection problem, specified by a set of nodes $[n]$ and a class of subsets $\mathcal{C} \subset 2^{[n]}$, the arguments above continue to apply if one is able to construct an appropriate approximating net as in Lemma 5.1. This is done, for example, by Arias-Castro et al. in [8, 12] for a wide range of settings. We note that the construction of an approximating net is purely geometrical.

Other exponential models: To extend the result to an arbitrary (one-parameter, natural) exponential model, we require the equivalent of the tail-bound (5.7). While such a bound may not apply to a particular exponential model, it does apply asymptotically to large sums of i.i.d. variables from that model by Chernoff's bound and a Taylor development of the rate function. To ease presentation, assume without loss of generality that F_0 has mean zero, and unit variance.

Indeed, recalling the notation introduced in Section 5.2.1, let the rate function

[7] By polylog(n), we mean a polynomial of $\log n$.

of F_0 be $\psi_0(t) = \sup_{\lambda \geq 0}(\lambda t - \log \varphi_0(\lambda))$. We have, for any $\lambda \geq 0$, that

$$\begin{aligned}\mathbb{P}_0(\mathbf{Y}_S \geq x) &= \mathbb{P}_0\left(\sum_{i \in S} Y_i \geq \sqrt{|S|}x\right) \\ &= \mathbb{P}_0\left(e^{\lambda \sum_{i \in S} Y_i} \geq e^{\lambda\sqrt{|S|}x}\right) \ .\end{aligned}$$

Using Markov's inequality we can continue as

$$\begin{aligned}\mathbb{P}_0(\mathbf{Y}_S \geq x) &\leq e^{-\lambda\sqrt{|S|}x}\mathbb{E}_0\left(\prod_{i \in S} e^{\lambda Y_i}\right) \\ &= \exp\left(-\lambda\sqrt{|S|}x + |S| \log \varphi_0(\lambda)\right) \ .\end{aligned}$$

Re-writing the above in terms of the rate function,

$$\mathbb{P}_0(\mathbf{Y}_S \geq x) \leq \exp\left(-|S|\psi_0(x/\sqrt{|S|})\right) \ . \tag{5.8}$$

Assuming without loss of generality that F_0 has zero mean and unit variance, we have

$$\psi_0(t) \geq \frac{1}{2}t^2 + O(t^3) \ , \quad t \to 0 \ . \tag{5.9}$$

To see this, note that

$$\psi_0(t) = \sup_{\lambda \in [0, \theta^*)} (\lambda t - \log \phi_0(\lambda)) \geq t^2 - \log \phi_0(t) \ ,$$

with the choice $\lambda = t$. On the other hand, $\varphi_0(t)$ is infinitely many times differentiable when $t \in [0, \theta^*)$, with $\varphi_0'(0) = \mathbb{E}_0(X) = 0$ and $\varphi_0''(0) = \mathbb{E}_0(X^2) = 1$, by assumption. Therefore, $\varphi_0(t) \leq 1 + \frac{1}{2}t^2 + Kt^3$ when t is in a finite neighborhood of zero. Thus, using $\log(1 + x) \leq x$, we have

$$\psi_0(t) \geq \tfrac{1}{2}t^2 - Kt^3$$

around $t = 0$. This results in the bound

$$\mathbb{P}_0(\mathbf{Y}_S \geq x) \leq \exp\left(-\frac{1}{2}x^2 + O(x^3/\sqrt{|S|})\right) \ .$$

From this we see that our derivations for the normal model apply essentially verbatim if, for some constant $c > 0$, $|S| \geq c(\log n)^3$ for all $S \in \mathcal{C}$. Furthermore,

it can be seen the test in (5.6) is essentially optimal for exponential models, as its performance matches the lower bounds developed by Arias-Castro, Candès & Durand in [8].

5.4 Calibration by permutation

We have described in detail how a performance bound is established for the scan test variant (5.6) for the problem of detecting an interval of unknown length, and its extensions to other detection problems. Because of this, we now clearly see that the key to adapting this analysis to a calibration by permutation is a concentration of measure bound to replace (5.7) and (5.8). Since this is the same in any detection setting, we consider the problem of detecting an interval of unknown length as in Section 5.3. This time, we impose a minimum and maximum length on the intervals

$$\mathcal{C} = \big\{\{a, \dots, b\} : 1 \le a < b \le n, s_l \le b - a \le s_u\big\} \ . \tag{5.10}$$

Indeed, when calibrating the scan test by permutation, we necessarily have to assume non-trivial upper and lower bounds on the size of an anomalous interval. To see this consider intervals of length one. In this case the value of the scan for any permutation of the data is the same regardless of the underlying distributions. By symmetry the same reasoning applies to intervals of length $n - 1$.

We consider essentially the same scan statistic (5.6) as before, except for the following. We restrict the approximating net to match the class of intervals defined in (5.10) (but still call it $\mathcal{C}_b$). Specifically we only keep an element $S' \in \mathcal{C}_b$ if there is $S \in \mathcal{C}$ such that $\rho(S, S') \ge (1 + 2^{-b+2})^{-1/2}$. This ensures that the statements in Lemma 5.1 still hold, and also that $|S'| \ge s_l/(1 + 2^{-b+2})$ for all $S' \in \mathcal{C}_b$. We also do a "centering" of the statistic prior to the scan. In detail, with $\mathbf{y} = \{y_i\}_{i\in[n]}$ denoting the observed data, we define

$$\textsc{scan}(\mathbf{y}) = \max_{S\in\mathcal{C}_b} \Big(\mathbf{y}_S - \sqrt{|S|}\bar{\mathbf{y}}\Big) \ , \quad \mathbf{y}_S := \frac{1}{\sqrt{|S|}} \sum_{i\in S} y_i \ , \tag{5.11}$$

where $\bar{\mathbf{y}} = \frac{1}{n}\sum_{i\in[n]} y_i$ is the overall average. The test rejects the null when

$$\mathfrak{P}(\mathbf{y}) := \frac{1}{n!}\Big|\big\{\pi \in [n]! : \textsc{scan}(\mathbf{y}^{(\pi)}) \ge \textsc{scan}(\mathbf{y})\big\}\Big| \le \alpha \ , \tag{5.12}$$

where $\mathfrak{P}(\mathbf{y})$ is the permutation p-value, and $\alpha \in (0, 1)$ is the desired level.

Recall the definition of $\theta_\star$ in Section 5.2.1.

Theorem 5.1. *Let $0 < \alpha < 1$ and consider the test that rejects the null if $\mathfrak{P}(\mathbf{Y}) \leq \alpha$, where $\mathfrak{P}$ is defined in (5.12), with $b = b_n \to \infty$ and $b_n / \log n \to 0$. Assume that the anomalous set S belongs to $\mathcal{C}$ defined in (5.10) with $s_l/(\log n)^3 \to \infty$ and $s_u = o(n)$ as $n \to \infty$. Then the test has level at most α. When the sample distribution is from the exponential family (as defined in Section 5.2.1) with F_0 having mean zero and unit variance, the test has power converging to 1 as $n \to \infty$ whenever*

$$\bar{\theta}_S \geq \tau \sqrt{\frac{2}{|S|} \log n} \ ,$$

with $\tau > 1$, provided that either F_0 has compact support or $\max_i \theta_i \leq \widetilde{\theta} < \theta_\star$ for some fixed $\widetilde{\theta} > 0$.

The headline here is that a calibration by permutation has as much asymptotic power as a calibration by Monte-Carlo with full knowledge of the null distribution (to first-order accuracy). This is (qualitatively) in line with what is known in classical settings (see Lehmann & Romano [102]).

The conditions required here allow $\mathcal{C}$ to be any class of intervals of lengths between $(\log n)^{3+a}$ and $o(n)$, for any $a > 0$ fixed. This includes the most interesting cases of intervals not too short and also not too long. In fact, for certain families of distributions removing from consideration very small intervals is essential and cannot be avoided. For instance consider the Bernoulli model, where $Y_i \sim \text{Bernoulli}(1/2)$, for all $i \in [n]$ under the null and $Y_i \sim \text{Bernoulli}(1)$, for all $i \in S$ when S is anomalous. Even under the null we will encounter a run of ones of length $\sim \log_2 n$ (the famous Erdős-Rényi Law) with positive probability. Therefore in this case the scan test, calibrated by Monte-Carlo or permutation, is powerless for detection of intervals of length $\frac{1}{2} \log_2 n$. In fact, it can be shown that no test has any power in that case.

We place an upper bound on the nonzero θ_i's to streamline the proof arguments and also avoid special cases we were not able to rule out. For example, an open question is whether the power of this permutation test is monotone increasing in each of the θ_i when $i \in S$ and S is the anomalous set. If this is true, then obviously the upper bound (by $\widetilde{\theta}$) can be removed. We note that when F_0 does not have compact support, this can be enforced by applying a censoring. See Section 5.7.

Remark 5.2. *If we do not assume F_0 to have zero mean and unit variance, we*

can still follow the steps of the proof below. By doing so one can easily check that by replacing the condition in Theorem 5.1 with

$$\frac{1}{|S|}\sum_{i\in S}\theta_i \geq \tau\frac{1}{\sigma_0}\sqrt{\frac{2}{|S|}\log n}\ ,\ \textit{with}\ \tau > 1\ ,$$

the statement remains true, where σ_0 denotes the standard deviation of F_0.

Proof of Theorem 5.1. Consider first the null hypothesis. Note that $\mathbf{Y} = \{Y_i\}_{i\in[n]}$ are i.i.d. under the null, and therefore exchangeable. This means that, for any permutation π the marginal distributions of $\textsc{scan}(\mathbf{Y})$ and $\textsc{scan}(\mathbf{Y}^{(\pi)})$ are the same. This implies that $\textsc{scan}(\mathbf{y})$ is uniformly distributed on the set $\{\textsc{scan}(\mathbf{y}^{(\pi)}), \pi \in [n]!\}$ (with multiplicities). With this we have

$$\mathbb{P}(|\{\pi \in [n]! : \textsc{scan}(\mathbf{y}^{(\pi)}) \geq \textsc{scan}(\mathbf{y})\}| \geq \alpha n!) \leq \frac{\lfloor \alpha n! \rfloor}{n!} \leq \alpha\ ,$$

where $\lfloor z \rfloor$ denotes the integer part of z. If there were no ties, then the first inequality above would be an equality, but with ties present the test becomes more conservative. For more details on permutation tests the reader is referred to Lehmann & Romano [102].

All that remains to be done is to study the permutation test under the alternative hypothesis. This requires two main steps. First we need to control the randomness in the permutation, conditionally on the observations $\mathbf{y}$. Once this is done we remove the conditioning on the observed data.

The key to the first step is the following Bernstein's inequality for sums of variables sampled without replacement from a finite population:

Lemma 5.2 (Bernstein's inequality for sampling without replacement)**.** *Let $\{Z_i\}_{i\in[m]}$ be obtained by sampling without replacement from a given a set of real numbers $\{z_j\}_{j\in[J]} \subset \mathbb{R}$. Define $z_{\max} = \max_j z_j$, $\bar{z} = \frac{1}{J}\sum_j z_j$, and $\sigma_z^2 = \frac{1}{J}\sum_j (z_j - \bar{z})^2$. Then the sample mean $\bar{Z} = \frac{1}{m}\sum_i Z_i$ satisfies*

$$\mathbb{P}\big(\bar{Z} \geq \bar{z} + t\big) \leq \exp\left(-\frac{mt^2}{2\sigma_z^2 + \frac{2}{3}(z_{\max} - \bar{z})t}\right), \quad \forall t \geq 0.$$

This result is a consequence of Theorem 4 of Hoeffding [84] and Chernoff's bound, from which Bernstein's inequality is derived, as in Shorack & Wellner [132],

page 851[8]. See Boucheron, Lugosi, & Massart [28], Bardenet & Maillard [22] and Dembo & Zeitouni [60] for a discussion of the literature on concentration inequalities for sums of random variables sampled without replacement from a finite set.

Applying this result for a fixed (but arbitrary) set $S' \in \mathcal{C}_b$ when π is uniformly drawn from $[n]!$ and $\mathbf{y}$ is given, we get

$$\mathbb{P}\left(\mathbf{y}_{S'}^{(\pi)} - \sqrt{|S'|}\bar{\mathbf{y}} \geq t\right) \leq \exp\left(-\frac{t^2}{2\sigma_{\mathbf{y}}^2 + \frac{2}{3}(\mathbf{y}_{\max} - \bar{\mathbf{y}})t/\sqrt{|S'|}}\right) , \quad \forall t \geq 0,$$

using the same notation as in Lemma 5.2. Plugging in $t = \textsc{scan}(\mathbf{y})$, noting that $|S'| \geq s_l/(1 + 2^{-b+2}) \geq s_l/2$ eventually (because $b \to \infty$), and using this together with a union bound, we get

$$\mathfrak{P}(\mathbf{y}) \leq |\mathcal{C}_b| \exp\left(-\frac{\textsc{scan}(\mathbf{y})^2}{2\sigma_{\mathbf{y}}^2 + (\mathbf{y}_{\max} - \bar{\mathbf{y}})\textsc{scan}(\mathbf{y})/\sqrt{s_l}}\right). \tag{5.13}$$

(The $\frac{2}{3}$ in the denominator, when multiplied by $\sqrt{2}$, from $|S'| \geq s_l/2$, is still less than 1.)

Now we remove the conditioning $\mathbf{Y} = \mathbf{y}$ by plugging the random variable Y into the expression above. We then proceed by upper bounding the right-hand side, which amounts to controlling the terms $\mathbf{Y}_{\max} - \bar{\mathbf{Y}}, \sigma_{\mathbf{Y}}^2$ and $\textsc{scan}(\mathbf{Y})$ under the alternative.

Let S denote the anomalous interval under the alternative. Recall that by assumption $\theta_i \leq \widetilde{\theta}$ for all $i \in S$, and

$$\frac{1}{|S|}\sum_{i \in S} \theta_i \geq \tau\sqrt{\frac{2}{|S|}\log n} := \theta \ . \tag{5.14}$$

Note that, by the assumption on s_l, we have $\theta \to 0$ as $n \to \infty$. Also note that $\mathbb{E}_\theta(X)$ and $\mathrm{Var}_\theta(X)$ are continuous in θ (and thus bounded on the interval $[0, \widetilde{\theta}]$), $\mathbb{E}_\theta(X)$ is increasing in θ (as $\frac{\partial}{\partial\theta}\mathbb{E}_\theta(X) = \mathbb{E}_\theta(X^2) \geq 0$) and $\mathbb{E}_\theta(X) \geq \theta + O(\theta^2)$ around zero (this can be checked by noting $\mathbb{E}_\theta(X) = \int x e^{\theta x} F_0(\,\mathrm{d}x)$ and writing the Taylor expansion of $e^{\theta x}$ around zero). Also recall that F_0 has zero mean and unit variance.

[8]There is a typo in the statement of the result on page 851 in Shorack & Wellner [132], but following the proof one can find the correct result. Where the statement of the result reads $-\frac{\lambda}{2\sigma^2}$ we should have $-\frac{\lambda^2}{2\sigma^2}$ instead.

We begin by controlling $\mathbf{Y}_{\max} - \bar{\mathbf{Y}}$. We have

$$\bar{\mathbf{Y}} = \frac{1}{n}\sum_{i\in[n]} \mathbb{E}(Y_i) + \frac{1}{n}\sum_{i\in[n]} (Y_i - \mathbb{E}(Y_i)) = O(|S|/n) + o_P(1) = o_P(1) \ ,$$

as $n \to \infty$, since $|S| = o(n), \theta_i \leq \widetilde{\theta}$ for all $i \in [n]$, and using Chebyshev's inequality in the second equality. Furthermore, let $\mathbf{Y}_{\max,\mathrm{S}} = \max_{i\in S} Y_i$ be the maximum over S. A union bound together with $\mathbf{Y}_{\max} = \mathbf{Y}_{\max,\mathrm{S}} \vee \mathbf{Y}_{\max,\bar{\mathrm{S}}}$ implies

$$\mathbb{P}(\mathbf{Y}_{\max} > x) \leq \mathbb{P}(\mathbf{Y}_{\max,\mathrm{S}} > x) + \mathbb{P}(\mathbf{Y}_{\max,\bar{\mathrm{S}}} > x) \leq |S|\bar{F}_{\widetilde{\theta}}(x) + |\bar{S}|\bar{F}_0(x) \ ,$$

where $\bar{F}_\theta(x) = \mathbb{P}_\theta(X > x)$ and we used the fact that $\bar{F}_\theta(x)$ is monotone increasing in θ - see Section 5.2.1. For $c \in (0, \theta_\star - \widetilde{\theta})$, we have

$$\begin{aligned}\bar{F}_{\bar{\theta}}(x) &= \int_x^\infty e^{\bar{\theta}u - \log\varphi_0(\bar{\theta})}\,\mathrm{d}F_0(u) \\ &= \frac{1}{\varphi_0(\bar{\theta})}\int_x^\infty e^{-cu}e^{(\bar{\theta}+c)u}\,\mathrm{d}F_0(u) \leq \frac{\varphi_0(\bar{\theta}+c)}{\varphi_0(\bar{\theta})}e^{-cx} \ .\end{aligned}$$

Using this with the above union bound gives $\mathbb{P}(\mathbf{Y}_{\max} > (2/c)\log n) \to 0$ as $n \to \infty$. This and the bound on $\bar{\mathbf{Y}}$ imply that

$$\mathbb{P}(\mathbf{Y}_{\max} - \bar{\mathbf{Y}} > (3/c)\log n) \to 0 \ .$$

We now consider $\sigma^2_{\mathbf{Y}}$. Similarly as before, we have

$$\sigma^2_{\mathbf{Y}} = \frac{1}{n}\sum_{i\in[n]} (Y_i - \bar{\mathbf{Y}})^2 \leq \frac{1}{n}\sum_{i\in[n]} Y_i^2 = \frac{1}{n}\sum_{i\in[n]} \mathbb{E}(Y_i^2) + \frac{1}{n}\sum_{i\in[n]} (Y_i^2 - \mathbb{E}(Y_i^2)) \ .$$

On one hand,

$$\begin{aligned}\frac{1}{n}\sum_{i\in[n]} \mathbb{E}(Y_i^2) &= \frac{1}{n}\sum_{i\notin S} \mathrm{Var}(Y_i) + \frac{1}{n}\sum_{i\in S} (\mathrm{Var}(Y_i) + \mathbb{E}(Y_i)^2) \\ &= 1 - \frac{|S|}{n} + O\left(\frac{|S|}{n}\right) = 1 + o(1) \ ,\end{aligned}$$

using $\mathrm{Var}(Y_i) = 1$ for $i \notin S$, $\max_{i\in S}\mathrm{Var}(Y_i) < \infty$ and $\max_{i\in S}\mathbb{E}(Y_i) < \infty$ (since

$\max_{i\in S}\theta_i \le \widetilde{\theta}$), as well as our assumption that $|S| = o(n)$. On the other hand,

$$\frac{1}{n}\sum_{i\in[n]}(Y_i^2 - \mathbb{E}(Y_i^2)) = O_P(1/\sqrt{n}) \ ,$$

using the fact that $\max_{i\in[n]}\mathbb{E}(Y_i^4) < \infty$ (since $\max_{i\in S}\theta_i \le \widetilde{\theta}$) combined with Chebyshev's inequality. We may therefore conclude that

$$\mathbb{P}(\sigma^2_{\mathbf{Y}} \le 1 + \epsilon/4) \to 1 \ ,$$

with a fixed but arbitrary $\epsilon > 0$ (we will choose an appropriate value for ϵ later on).

From Lemma 5.1 (which does apply to the newly defined $\mathcal{C}_b$) there is a set $S' \in \mathcal{C}_b$ such that $S \subseteq S'$ and $\rho(S, S') \ge (1 + 2^{-b+2})^{-1/2}$. Note that $\rho(S, S') = 1 - o(1)$ by the fact that $b \to \infty$. We then have

$$\begin{aligned}\textsc{scan}(\mathbf{Y}) &\ge \mathbf{Y}_{S'} - \sqrt{|S'|}\bar{\mathbf{Y}} = \sqrt{|S'|}(\bar{\mathbf{Y}}_{S'} - \bar{\mathbf{Y}}) \\ &\ge \sqrt{|S'|}\left(\frac{|S|(n-|S'|)}{|S'|n}\bar{\mathbf{Y}}_S - \frac{n-|S|}{n}\bar{\mathbf{Y}}_{[n]\setminus S}\right) \ ,\end{aligned}$$

where $\bar{\mathbf{Y}}_S$ and $\bar{\mathbf{Y}}_{[n]\setminus S}$ are the averages of the components of $\mathbf{Y}$ over the sets S and $[n] \setminus S$ respectively. By Chebyshev's inequality,

$$\begin{aligned}\bar{\mathbf{Y}}_S &= \frac{1}{|S|}\sum_{i\in S}\mathbb{E}(Y_i) + O_P(1/\sqrt{|S|}) \ , \\ \bar{\mathbf{Y}}_{[n]\setminus S} &= O_P(1/\sqrt{|n-|S||}) \ .\end{aligned}$$

Furthermore, as noted above, $\frac{1}{|S|}\sum_{i\in S}\mathbb{E}(Y_i) \ge \mathbb{E}_\theta(X) \ge \theta + O(\theta^2)$ around zero. Recalling the expression for θ, using $\sqrt{|S'|} = (1+o(1))\sqrt{|S|}$ and $|S| = o(n)$ we get

$$\textsc{scan}(\mathbf{Y}) = (1+o(1))\tau\sqrt{2\log n} + O_P(1) - o(1) + O_P(1/\sqrt{n}) \ .$$

Hence

$$\textsc{scan}(\mathbf{Y}) \ge \sqrt{2(1+\epsilon/2)\log n} \ ,$$

with probability tending to one as $n \to \infty$, where $\tau = \sqrt{1+\epsilon}$.

Plugging this back into the upper bound on the p-value given by (5.13) and

using the condition on s_l we get

$$\begin{aligned}\log \mathfrak{P}(\mathbf{Y}) &\leq \log |\mathcal{C}_b| - \frac{2(1+\epsilon/2)\log n}{2(1+\epsilon/4) + (3/c)(\log n)\sqrt{2(1+\epsilon/2)\log n}/\sqrt{s_l}} \\ &\leq \log |\mathcal{C}_b| - \frac{(1+\epsilon/2)\log n}{1+\epsilon/4+o(1)} \ ,\end{aligned}$$

with probability going to 1. For the size of the approximating net we have

$$\log |\mathcal{C}_b| \leq \log \left(n 4^{b+1}\right) = \log n + (b+1)\log 4 = (1+o(1))\log n \ , \tag{5.15}$$

by our assumption on b. Combining these allows us to conclude that $\log \mathfrak{P}(\mathbf{Y}) \to -\infty$ (meaning $\mathfrak{P}(\mathbf{Y}) \to 0$) with probability tending to one, implying that the test has power tending to 1 as $n \to \infty$. □

5.5 Scanning the ranks

Having observed $\mathbf{y} = \{y_i\}_{i\in[n]}$, scanning the ranks amounts to replacing every observation with its rank among all the observations, and computing the scan (5.11). We call this the *rank scan*. As for all rank-based methods, the null distribution is the permutation distribution when there are no ties.

- When there are no ties with probability one we calibrate the test by permutation, and this can be done before data is observed.
- When there are ties the rank scan test is also calibrated by permutation. If one breaks ties using the average rank then the calibration must be done every time as for the permutation test. A much better alternative is to break ties randomly, so that the test can be calibrated by permutation only once (before seeing the data). The latter option is computationally superior and is the one we analyze.

See Section 5.6 for implementation issues and a computational complexity analysis.

Formally, let $\mathbf{y} = \{y_i\}_{i\in[n]}$ denote the observations as before, and for every $i \in [n]$, let r_i be the rank (in increasing order) of y_i in $\mathbf{y}$, where ties are broken randomly, and let $\mathbf{r} = \{r_i\}_{i\in[n]}$ be the vector of ranks. The rank scan test returns the p-value $\mathfrak{P}(\mathbf{r})$ defined in (5.12).

As we mentioned in Section 5.1, an important advantage of the rank scan over the permutation scan is the fact that the former only requires calibration once,

while the latter requires a new calibration with each dataset. This assumes that the extrinsic signal dimension n remains the same. An additional advantage of the rank scan is its robustness to outliers - although the permutation scan after censoring (discussed in Section 5.7) is also robust to outliers.

Because the rank scan test is a special case of the permutation scan test, we assume similarly upper and lower bounds on the size of the anomalous set as in Section 5.4. However, we will see later that it is possible to relax these conditions and prove results similar to Theorem 5.2 for small intervals as well (see Proposition 5.2).

We first prove a theorem for the rank scan test that establishes a performance bound for general distributions. This is followed by a corollary that establishes the performance of the test for the family of exponential distributions. Define

$$p_{\theta,\theta'} = \mathbb{P}(X > Y) + \tfrac{1}{2}\mathbb{P}(X = Y) \ , \tag{5.16}$$

where $X \sim F_\theta$ and $Y \sim F_{\theta'}$ are independent. Also, we use the shorthand notation $p_\theta := p_{\theta,0}$. Note that in the definition above F_θ and $F_{\theta'}$ need not be members of the exponential family.

Theorem 5.2. *Let $0 < \alpha < 1$ and consider the test that rejects the null if $\mathfrak{P}(\mathbf{R}) \leq \alpha$, where $\mathfrak{P}$ is defined in (5.12), with $b = b_n \to \infty$ and $b_n/\log n \to 0$. Assume that the anomalous set S belongs to $\mathcal{C}$ defined in (5.10) with $s_l/\log n \to \infty$ and that $s_u = o(n)$ as $n \to \infty$. Then the test has level at most α. Moreover, it has power converging to 1 as $n \to \infty$ when*

$$\frac{1}{|S|}\sum_{i\in S} p_{\theta_i} \geq \frac{1}{2} + \tau\sqrt{\frac{2}{|S|}\log n}, \quad \textit{with } \tau > \tau_0 := \frac{1}{2\sqrt{3}} \ .$$

Note that by the condition on s_l, the expression on the right in the condition above goes to $1/2$ as $n \to \infty$. This allows us to relate the parameter θ to p_θ using a Taylor expansion when the distributions are in the exponential family. Define

$$\Upsilon_0 = \mathbb{E}(X\mathbf{1}\{X > Y\}) + \tfrac{1}{2}\mathbb{E}(X\mathbf{1}\{X = Y\}) = \frac{1}{2}\mathbb{E}(\max\{X, Y\}) \ , \tag{5.17}$$

where $X, Y \sim F_0$ and independent. Note that $\Upsilon_0 \geq 0$ (with equality iff F_0 is a Dirac-measure).

Corollary 5.1. *Suppose the distributions of $\{Y_i\}_{i\in[n]}$ are in the exponential family as defined in Section 5.2.1 with F_0 having mean zero and unit variance. Under*

the conditions of Theorem 5.2, the rank scan test has level at most α and power converging to 1 as $n \to \infty$ when

$$\frac{1}{|S|}\sum_{i\in S}\theta_i \geq \tau\sqrt{\frac{2}{|S|}\log n}\ , \quad \textit{with } \tau > \frac{1}{2\sqrt{3}\Upsilon_0}\ .$$

The headline here is that rank scan requires a signal amplitude which is τ_0/Υ_0 larger than what is required of the regular scan test calibrated by Monte-Carlo with full knowledge of the null distribution. This is (qualitatively) in line with similar results in more classical settings (see Hettmansperger [83]). For the normal location model, we find that $\tau_0 = \sqrt{\pi/3} \approx 1.023$, so the detection threshold of rank scan is almost the same as that of the regular scan test. See Appendix 5.B for details.

Remark 5.3. *If we do not assume F_0 to have zero mean and unit variance, we can still follow the steps of the proof of Corollary 5.1. By doing so one can easily check that by replacing the condition in Corollary 5.1 with*

$$\frac{1}{|S|}\sum_{i\in S}\theta_i \geq \tau\sqrt{\frac{2}{|S|}\log n}\ , \quad \textit{with } \tau > \frac{1}{2\sqrt{3}(\Upsilon_0 - \frac{1}{2}\mathbb{E}_0(X))}\ ,$$

the statement remains true, where $\mathbb{E}_0(X)$ denotes the mean of F_0.

Proof of Theorem 5.2. The arguments used for the general permutation test apply verbatim under the null hypothesis, so all that remains to be done is to study the performance of the rank scan test under the alternative.

We may directly apply (5.13), to obtain

$$\mathfrak{P}(\mathbf{r}) \leq |\mathcal{C}_b| \exp\left(-\frac{\textsc{scan}(\mathbf{r})^2}{\frac{n^2}{6} + \frac{n}{2}\textsc{scan}(\mathbf{r})/\sqrt{s_l}}\right)\ , \tag{5.18}$$

where we have used that $\sigma_{\mathbf{r}}^2 = (n^2-1)/12 < n^2/12$, $\mathbf{r}_{\max} = n$ and $\bar{\mathbf{r}} = (n+1)/2$, so that $\mathbf{r}_{\max} - \bar{\mathbf{r}} < n/2$. The previous bounds can be directly computed when there are no ties in the ranks, and it is easy to verify that they also hold if ties are dealt with in any of the classical ways (assigning the average rank, randomly breaking ties, etc). As before, this is a result conditional on the observations $\mathbf{Y} = \mathbf{y}$ and hence the ranks $\mathbf{R} = \mathbf{r}$. The next step is to remove this conditioning, which now amounts to controlling the term $\textsc{scan}(\mathbf{R})$.

Let S denote the anomalous interval under the alternative. From Lemma 5.1 there is a set $S' \in \mathcal{C}_b$ such that $S \subseteq S'$ and $\rho(S, S') \geq (1 + 2^{-b+2})^{-1/2}$. Note that $\rho(S, S') = 1 - o(1)$ by the fact that $b \to \infty$. Since

$$\textsc{scan}(\mathbf{R}) \geq \mathbf{R}_{S'} - \sqrt{|S'|}\tfrac{n+1}{2}, \quad \mathbf{R}_{S'} = \frac{1}{\sqrt{|S'|}} \sum_{i \in S'} R_i \ ,$$

we focus on obtaining a lower bound on $\mathbf{R}_{S'}$ that applies with high probability.

Note that

$$\mathbb{E}(\mathbf{R}_{S'}) = \frac{1}{\sqrt{|S'|}} \sum_{i \in S'} \mathbb{E}(R_i) \ ,$$

and

$$\mathrm{Var}(\mathbf{R}_{S'}) = \frac{1}{|S'|} \left(\sum_{i \in S'} \mathrm{Var}(R_i) + \sum_{i,j \in S', i \neq j} \mathrm{Cov}(R_i, R_j) \right) \ .$$

In an analogous fashion to that in Hettmansperger [83], we can make the following claims about the first two moments of the ranks.

Lemma 5.3. *Suppose $X_i \sim F_{\theta_i}, i \in [s]$ and independent, also independent of $\{X_i\}_{i \in [s+1,n]}$ which are i.i.d. and distributed as F_0. Let R_i denote the rank (in increasing order) of X_i in the combined sample, and suppose ties are broken randomly. Recall the definition of $p_{\theta,\theta'}$ and p_θ in* (5.16)*.*

$$\mathbb{E}(R_i) = \begin{cases} (n-s)p_{\theta_i} + \sum_{j \in [s], j \neq i} p_{\theta_i,\theta_j} + 1 & \text{, when } i \in [s], \\ \frac{n+s+1}{2} - \sum_{j \in [s]} p_{\theta_j} & \text{, when } i \notin [s]. \end{cases}$$

Furthermore, as $n, s \to \infty, s = o(n)$, for $i \in [s]$

$$\mathrm{Var}(R_i) = (\lambda_{\theta_i} - p_{\theta_i}^2)n^2 + O(sn) \ ,$$

where

$$\lambda_\theta = \mathbb{P}(\{X > Y_1\} \cap \{X > Y_2\}) + \mathbb{P}(X = Y_1 > Y_2) + \tfrac{1}{3}\mathbb{P}(X = Y_1 = Y_2) \ ,$$

where $X \sim F_\theta$ independent of $Y_1, Y_2 \sim F_0$, also independent. Finally, for any $i, j \in [n]$

$$\mathrm{Cov}(R_i, R_j) = O(n) \ .$$

For the sake of completeness we sketch a proof of Lemma 5.3 in Appendix 5.A. Using the fact that for any θ, θ' we have $p_{\theta,\theta'} + p_{\theta',\theta} = 1$ we get

$$\begin{aligned}\mathbb{E}(\mathbf{R}_{S'}) &= \frac{1}{\sqrt{|S'|}}\Bigg(|S|(n-|S|)\bar{p}_S + \sum_{i\in S}\sum_{j\in S,\ j\neq i} p_{\theta_i,\theta_j} + |S| \\ &\qquad + |S' \setminus S|\tfrac{n+|S|+1}{2} - |S' \setminus S||S|\bar{p}_S\Bigg) \\ &= \frac{|S|}{\sqrt{|S'|}}(n - |S| - |S' \setminus S|)(\bar{p}_S - 1/2) + \sqrt{|S'|}\frac{n+1}{2}\ ,\end{aligned}$$

where $\bar{p}_S = \frac{1}{|S|}\sum_{i\in S} p_{\theta_i}$ is the average of p_{θ_i} over the anomalous set.

Note that for any $i \in [n]$ we trivially have $\mathrm{Var}(R_i) \leq n^2$, and by Lemma 5.3 $\mathrm{Cov}(R_i, R_j) = O(n)$, so $\mathrm{Var}(R_{S'}) = O(n^2)$. Hence, using Chebyshev's inequality we obtain

$$\begin{aligned}\mathbf{R}_{S'} - \sqrt{|S'|}\tfrac{n+1}{2} &= \frac{|S|}{\sqrt{|S'|}}(n - |S| - |S' \setminus S|)(\bar{p}_S - 1/2) + O_P(n) \\ &\geq \rho(S, S')(n - 2s_u)\tau\sqrt{2\log n} + O_P(n)\ .\end{aligned}$$

We thus have

$$\textsc{scan}(\mathbf{R}) \geq cn\sqrt{2\log n}\ ,$$

with probability going to 1 as $n \to \infty$, where $c \in (\tau_0, \tau)$, for instance $c = \frac{\tau+\tau_0}{2}$.

Plugging this back into (5.18) and using the condition on s_l we get

$$\begin{aligned}\log \mathfrak{P}(\mathbf{R}) &\leq \log|\mathcal{C}_b| - \frac{2c^2n^2\log n}{\frac{n^2}{6} + \frac{n}{2}c\sqrt{2\log n}/\sqrt{s_l}} \\ &\leq \log|\mathcal{C}_b| - \frac{(c/\tau_0)^2\log n}{1+o(1)}\ ,\end{aligned}$$

with probability going to 1. Note that the upper bound on $|\mathcal{C}_b|$ in (5.15) is still valid. This with the fact that $c/\tau_0 > 1$ yields that $\log \mathfrak{P}(\mathbf{R}) \to -\infty$ as $n \to \infty$, hence the test is asymptotically powerful. □

Proof of Corollary 5.1. We bound p_θ by using a Taylor expansion. When F_0 is discrete, we have

$$p_\theta = \int_{\mathbb{R}} \left(\bar{F}_\theta(x) + \tfrac{1}{2}f_\theta(x)F_0(x)\right) F_0(\,\mathrm{d}x).$$

We expand the integrand seen as a function of θ around $\theta = 0$ up to a second order error term. We have

$$\tfrac{\partial}{\partial\theta} f_\theta(x)\Big|_{\theta=0} = x, \quad \tfrac{\partial}{\partial\theta} \bar{F}_\theta(x)\Big|_{\theta=0} = \int_{(x,\infty)} u \ \mathrm{d}F_0(u) \ ,$$

where the second identity comes from differentiating inside the integral defining $\bar{F}_\theta$, justified by dominated convergence. Note that $\frac{\partial^2}{\partial\theta^2} f_\theta(x)$ is integrable w.r.t. F_0 when $\theta \in [0, \theta^*)$ and the same holds for $\frac{\partial^2}{\partial\theta^2} \bar{F}_\theta(x)$ as well. Hence let

$$c_0' := \int_{\mathbb{R}} \sup_{\widetilde{\theta}\in[0,\theta]} \tfrac{\partial^2}{\partial\theta^2} f_\theta(x)\Big|_{\theta=\widetilde{\theta}} \mathrm{d}F_0(x) < \infty \ , \text{ and}$$
$$c_0 := \int_{\mathbb{R}} \sup_{\widetilde{\theta}\in[0,\theta]} \tfrac{\partial^2}{\partial\theta^2} \bar{F}_\theta(x)\Big|_{\theta=\widetilde{\theta}} \mathrm{d}F_0(x) < \infty \ .$$

Hence,

$$\begin{aligned} p_\theta &\geq \int_{\mathbb{R}} \bar{F}_0(x) + \tfrac{1}{2}F_0(x) + \theta \left(\int_{(x,\infty)} u \ \mathrm{d}F_0(u) + \tfrac{1}{2}F_0(x)x \right) \mathrm{d}F_0(x) - \tfrac{\theta^2}{2}(c_0 + c_0'/2) \\ &= p_0 + \theta\big(\mathbb{E}_0(X\mathbf{1}\{X > Y\}) + \tfrac{1}{2}\mathbb{E}_0(X\mathbf{1}\{X = Y\})\big) - \tfrac{\theta^2}{2}(c_0 + c_0'/2) \\ &= \tfrac{1}{2} + \theta\Upsilon_0 - \tfrac{\theta^2}{2}(c_0 + c_0'/2) \ . \end{aligned}$$

When F_0 is continuous, we have

$$p_\theta = \int_{\mathbb{R}} \bar{F}_\theta(x) F_0(\mathrm{d}x),$$

and the same calculations lead to

$$p_\theta \geq \tfrac{1}{2} + \theta\Upsilon_0 - \tfrac{\theta^2}{2} c_0 \ .$$

In any case, $p_\theta \geq \frac{1}{2} + \theta\Upsilon_0 - O(\theta^2)$.

Following the steps of the proof of Theorem 5.2 with the above bound for p_{θ_i} the statement follows. □

Smaller intervals

The conditions of Theorem 5.2 allow for dealing with intervals of length of order (strictly) larger than $\log n$. We present results that encompass the scenario where the interval might be of smaller length. To keep the discussion simple we consider the class of intervals of a fixed size $|S| = s$ under the alternative. In this situation there is no need to consider an approximating net and we simply scan over the entire class, denoted by $\mathcal{C}$.

Recall the definition of the p-value of the permutation test (5.12) and the parameter p_θ in (5.16). We can state the following result, which we prove at the end of this section:

Proposition 5.2. *Let $0 < \alpha < 1$ and consider the test that rejects the null if $\mathfrak{P}(\mathbf{R}) \leq \alpha$. Recall that $\bar{p}_S$ is the average of p_{θ_i} over the anomalous set S. Then, in the present context, the test has level at most α and power converging to 1 as $n \to \infty$ provided*

$$(i)\ \bar{p}_S = 1 - o(n^{-2/s})\ \textit{when}\ 2 < s = o(\log n);\ \textit{or}$$
$$(ii)\ \bar{p}_S > 1 - \exp(-\tfrac{c+1}{c})\ \textit{when}\ s = c \log n\ \textit{for some}\ c > 0\ \textit{fixed.}$$

Theorem 5.2 and Proposition 5.2 together cover essentially all interval sizes that are $o(n^{2/3})$. Theorem 5.2 covers the case of larger intervals, in which case $\min_{i \in S} p_{\theta_i}$ can go to $1/2$ provided it does not converge too fast, and the test is still asymptotically powerful. In Proposition 5.2, a sufficient condition for an asymptotically powerful test is that $\min_{i \in S} p_{\theta_i}$ goes to 1 at a certain rate when the size of the anomalous interval is $o(\log n)$. If the interval size is $c \log n$ with $c > 0$ arbitrary, then the rank test is asymptotically powerful when $\min_{i \in S} p_{\theta_i}$ is greater than a constant (strictly larger than $1/2$) depending on c.

Unlike for Theorem 5.2, we can not state a general corollary for Proposition 5.2, that is a result in terms of the parameter θ instead ofp_θ. This is due to the fact that, for small intervals, the signal magnitude must necessarily be large, implying that θ is bounded away from zero. In such situations, one can only relate p_θ and θ with further knowledge about the family of distributions.

As an example, consider the normal means model when $s = o(\log n)$. In this case, we have

$$p_\theta = \Phi(-\theta/\sqrt{2}) \geq 1 - \frac{1}{2} e^{-\theta^2/4},$$

where $\Phi(x)$ is the CDF of the standard normal distribution. Hence, whenever $\frac{1}{2}e^{-{\theta_i}^2/4} = o(n^{-2/s})$ for all $i \in S$, the condition in the proposition is met. This is satisfied when

$$\min_{i \in S} \theta_i = \tau \sqrt{\frac{2}{s} \log n}, \quad \text{with } \tau > 2 \text{ fixed.} \tag{5.19}$$

This means that in this case the rank scan requires an amplitude at most two times larger than the regular scan test calibrated with full knowledge of the null distribution.

Finally note that the condition $p_\theta \to 1$ or $p_\theta > 1 - p_c$ might not be possible to meet for certain distributions of the exponential family. Recall the example of Bernoulli random variables discussed Section 5.4. In this setup $p_{\theta_i} = 3/4$ for all $i \in S$, a case that is not covered by Proposition 5.2 when the interval size is smaller than $c \log n$ and c is small enough. But this is expected since no test has any power when c is sufficiently small.

Remark 5.4. *We have considered the case when the size of the anomalous interval is known. However, we could consider the class of intervals of length greater than 2 and at most $\bar{s}$ for some given $\bar{s} = O(\log n)$. In this case we would simply scan the ranks for every fixed interval size up to $\bar{s}$ and apply a Bonferroni correction to the p-values. Following through the steps, one can see that the rank scan test would be asymptotically powerful when*

$$(i') \ \bar{p}_S = 1 - o(n \log n)^{-2/s} \ \textit{when} \ 2 < s = o(\log n) \ ;$$
$$(ii) \ \bar{p}_S > 1 - \exp(-\tfrac{c+1}{c}) \ \textit{when} \ s = c \log n \ \textit{for some} \ c > 0 \ \textit{fixed.}$$

For the normal location model and considering $\bar{s} = o(\log n)$, we can see that this is satisfied when (5.19) *holds.*

Proof of Proposition 5.2. We treat each case separately.

Condition (i). The same arguments hold as before under the null, so again we are left with studying the alternative. To deal with smaller intervals, we need a slightly different concentration inequality than before.

Lemma 5.4 (Chernoff's inequality for ranks)**.** *In the context of Lemma 5.2, assume that $z_j = j$ for all j. Then*

$$\mathbb{P}\left(\bar{Z} \geq \bar{z} + t\right) \leq \exp\left(-m \sup_{\lambda \geq 0} \psi(t, \lambda)\right) \ , \quad \forall t \geq 0 \ ,$$

where

$$\psi(t,\lambda) := \lambda t - \log\left(\frac{\sinh(\lambda n/2)}{n\sinh(\lambda/2)}\right) \ .$$

Similarly to Lemma 5.2 this result is also a consequence of Theorem 4 of Hoeffding [84] and Chernoff's bound. However, with the assumption on z_j in the lemma above we can directly compute the moment generating function of Z_j after using Chernoff's bound instead of upper bounding it, as is classically done to obtain Bernstein's inequality.

In the present context, this yields

$$\mathfrak{P}(\mathbf{r}) \leq |\mathcal{C}| \exp\left(-s\psi(\textsc{scan}(\mathbf{r})/\sqrt{s},\lambda)\right) \quad \forall \lambda > 0 \ .$$

Note that $x \leq \sinh(x) \leq e^x/2$ and $|\mathcal{C}| \leq n$, hence

$$\mathfrak{P}(\mathbf{r}) \leq n\exp\left(-\lambda\sqrt{s}\textsc{scan}(\mathbf{r}) + \frac{\lambda s n}{2} - s\log(\lambda n)\right) \quad \forall \lambda > 0 \ . \tag{5.20}$$

The next step is to remove the conditioning $\mathbf{R} = \mathbf{r}$ and bound $\textsc{scan}(\mathbf{R})$. Recall $\textsc{scan}(\mathbf{R}) \geq \mathbf{R}_S - \sqrt{s}\frac{n+1}{2}$, where S is the anomalous interval. As in the proof of Theorem 5.2 we use Lemma 5.3 to evaluate the terms $\mathbb{E}(\mathbf{R}_S)$ and $\mathrm{Var}(\mathbf{R}_S)$. We have

$$\mathbb{E}(\mathbf{R}_S) = \sqrt{s}(n-s)(\bar{p}_S - 1/2)) + \sqrt{s}\frac{n+1}{2} \ .$$

For the variance term, recalling the definition of λ_θ from Lemma 5.3, we note that $\lambda_\theta \leq p_\theta$. Hence

$$\mathrm{Var}(R_i) = (\lambda_{\theta_i} - p_{\theta_i}^2)n^2 + O(sn) \leq p_{\theta_i}(1-p_{\theta_i})n^2 + O(sn) \leq (1-p_{\theta_i})n^2 + O(sn) \ .$$

Also using $\mathrm{Cov}(R_i, R_j) = O(n)$, we get

$$\mathrm{Var}(\mathbf{R}_S) \leq (1-\bar{p}_S)n^2 + O(sn) \ .$$

According to our assumption, there exists a sequence $\omega_n \to \infty$ such that

$$\bar{p}_S \geq 1 - \omega_n^{-1} n^{-2/s} \ .$$

For reasons that become apparent at the end of the proof, we choose $\omega_n \to \infty$ not

too fast (for instance $\omega_n \le \log n$ suffices). Using Chebyshev's inequality we get

$$\begin{aligned}
&\mathbb{P}\left(\mathbf{R}_S - \sqrt{s}\tfrac{n+1}{2} \le \sqrt{s}(n-s)\left(\tfrac{1}{2} - \omega_n^{-1/4} n^{-1/s}\right)\right) \\
&\quad = \mathbb{P}\left(\mathbf{R}_S - \mathbb{E}(\mathbf{R}_S) \le \sqrt{s}(n-s)\left(1 - \omega_n^{-1/4} n^{-1/s} - \bar{p}_S\right)\right) \\
&\quad \le \mathbb{P}\left(\mathbf{R}_S - \mathbb{E}(\mathbf{R}_S) \le -\sqrt{s}(n-s)\left(\omega_n^{-1/4} n^{-1/s} - \omega_n^{-1} n^{-2/s}\right)\right) \\
&\quad \le \mathbb{P}\left(|\mathbf{R}_S - \mathbb{E}(\mathbf{R}_S)| \ge \sqrt{s}(n-s)\left(\omega_n^{-1/4} n^{-1/s} - \omega_n^{-1} n^{-2/s}\right)\right) \\
&\quad \le \frac{n^2 \omega_n^{-1} n^{-2/s} + O(sn)}{s(n-s)^2\left(\omega_n^{-1/4} n^{-1/s} - \omega_n^{-1} n^{-2/s}\right)^2} \le \frac{4n^2 \omega_n^{-1} n^{-2/s} + O(sn)}{s(n-s)^2 \omega_n^{-1/2} n^{-2/s}} \to 0 \ ,
\end{aligned}$$

where the last inequality follows because $\omega_n^{-1/4} n^{-1/s} - \omega_n^{-1} n^{-2/s} \ge \omega_n^{-1/4} n^{-1/s}/2$ eventually as $n \to \infty$. Hence,

$$\textsc{scan}(\mathbf{R}) \ge \sqrt{s}(n-s)\left(\tfrac{1}{2} - \omega_n^{-1/4} n^{-1/s}\right) \ ,$$

with probability converging to 1 as $n \to \infty$. Using this with (5.20) we get

$$\log \mathfrak{P}(\mathbf{R}) \le \log n + \frac{\lambda s^2}{2} + \lambda s(n-s)\omega_n^{-1/4} n^{-1/s} - s\log(\lambda n) \quad \forall \lambda > 0 \ ,$$

with probability tending to 1. Choosing $\lambda = \omega_n^{1/4} n^{1/s}/n$ we get

$$\log \mathfrak{P}(\mathbf{R}) \le \frac{\omega_n^{1/4} n^{1/s}}{n} s^2 + \frac{n-s}{n} s - \frac{s}{4}\log \omega_n \to -\infty \ ,$$

with probability going to 1, where we used that ω_n grows slowly enough for the first term to vanish.

Condition (ii). We can mimic the arguments above. Suppose $s = c\log n$ with arbitrary $c > 0$ and $\bar{p}_S = 1 - (1-\delta)\exp(-\frac{c+1}{c})) := 1 - (1-\delta)f(c)$ with some $\delta > 0$. As before, using Chebyshev's inequality we can show that

$$\textsc{scan}(\mathbf{R}) \ge \sqrt{s}(n-s)\left(\frac{1}{2} - \left(1 - \frac{\delta}{2}\right) f(c)\right) \ ,$$

with probability tending to 1 as $n \to \infty$. Plugging this into (5.20), choosing

$\lambda = 1/(nf(c))$ we get

$$\log \mathfrak{P}(\mathbf{R}) \leq \log n + \frac{s^2}{2f(c)n} + \frac{s(n-s)(1-\frac{\delta}{2})}{n} - s \log f(c) \ ,$$

with probability going to 1 as $n \to \infty$. Plugging in $s = c \log n$ and $f(c) = \exp(-\frac{c+1}{c})$ we see that the log of the p-value goes to $-\infty$, which is what we wanted to show. □

5.6 Numerical experiments

5.6.1 Computational complexity

We have already cited some works in Section 5.1 where fast (typically approximate) algorithms for scanning various classes of subsets are proposed (Neill & Moore [115], Neill [113], Arias-Castro, Donoho & Huo [12], Walther [143]). For example, as we saw in Lemma 5.1, Arias-Castro, Donoho & Huo [12] design an approximating net $\mathcal{C}_b$ for the class of all intervals $\mathcal{C}$ that can be scanned in $O(nb4^b)$ operations and provides an approximation in δ-metric of order $O(2^{-b/2})$. Furthermore, we saw in Proposition 5.1 that this procedure achieves the optimal asymptotic power as long as $b = b_n \to \infty$. For example, if $b_n \asymp \log \log n$, then the computational complexity is of order $(n\text{polylog}(n))$.

In any case, suppose that a scanning algorithm has been chosen and let C_n denote its computational complexity. The oracle scan test and the rank scan test are then comparable, in that they estimate the null distribution of their respective test statistic by simulation, and this is done only once for each data size n. With this preprocessing already done, the computational complexity of these two procedures is C_n, the cost of a single scan when applied to data of size n. In contrast, the permutation scan test is much more demanding, in that it requires scanning each of the permuted datasets, and this is done every time the test is applied. Assuming B permutations are sampled at random for calibration purposes, the computational complexity is BC_n, that is, B times that of the oracle or rank variants (not accounting for preprocessing). B is typically chosen in the hundreds ($B = 200$ in our experiments), if not thousands, so the computational burden can be *much* higher for the permutation test.

5.6.2 Simulations

We present the results of some basic numerical experiments that we performed to corroborate our theoretical findings in finite samples. We generated the data from the normal location model - where $F_\theta = N(\theta, 1)$ - which is arguably the most emblematic one-parameter exponential family and a popular model in signal and image processing. We used the regular scan test, calibrated with full knowledge of the null distribution, as a benchmark. The permutation scan test and rank scan test were calibrated by permutation.

The test statistic that we use in our experiments is the scan over all intervals of dyadic length. This subclass of intervals is morally similar to $\mathcal{C}_0$ (corresponding to $b = 0$) but somewhat richer. This choice allows us to both streamline the implementation and make the computations very fast via one application of the Fast Fourier Transform per dyadic length. In detail, letting $\mathcal{C}$ denote the class of all discrete intervals in $[n]$, this amounts to taking

$$\mathcal{C}_{\text{dyad}} = \Big\{ S \in \mathcal{C} : |S| = 2^j \text{ for some } j \in \mathbb{N} \Big\}$$

as an approximating set.

As explained earlier, the calibration by permutation and the rank-based approach are valid no matter what subclass of intervals is chosen, and in fact, the same mathematical results apply as long as the subclass is an appropriate approximating net. We encourage the reader to experiment with his/her favorite scanning implementation.

It is easy to see that, for each $S \in \mathcal{C}$, there is $S' \in \mathcal{C}_{\text{dyad}}$ with $S' \subset S$ and $|S'| > |S|/2$. Hence,

$$\min_{S \in \mathcal{C}} \max_{S' \in \mathcal{C}_{\text{dyad}}} \rho(S, S') \geq 1/\sqrt{2}.$$

A priori, this implies that scanning over $\mathcal{C}_{\text{dyad}}$ requires an amplitude $\sqrt{2}$ larger to achieve the same (asymptotic) performance as scanning over $\mathcal{C}$ or a finer approximating set as considered previously. To simplify things, however, in our simulations we took an anomalous interval of dyadic length, so that the detection threshold is in fact the same as before.

We set $n = 2^{15}$ and tried two different lengths for the anomalous interval

$|S| \in \{2^7, 2^{10}\}$. All the nonzero θ_i's were taken to be equal to

$$\theta_S = t\sqrt{\frac{2}{|S|}\log n} \tag{5.21}$$

with t varying. The critical values and power are based on 1000 repeats in each case. A level of significance of 0.05 was used. Also, 200 permutations were used for the permutation scan test. The results are presented in Figure 5.1. At least in these small numerical experiments, the three tests behave comparably, with the rank scan slightly dominating the others. Although the last finding is somewhat surprising, and we do not know the reason behind it, this is a finite-sample effect and is localized in the intermediate power range (around a power of 0.5) and so does not contradict the theory developed earlier. In fact, the three tests achieve power 1 at roughly the same signal amplitude, confirming the theory.

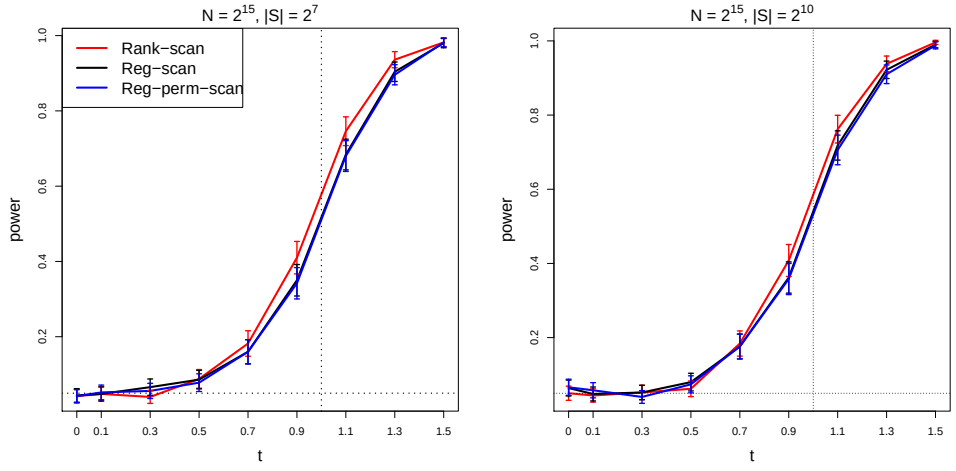

Figure 5.1: Power curves (with 95% margin of error) for the three tests (all set at level 0.05) as a function of the parameter t in (5.21): the scan test calibrated with knowledge of the null distribution (black); the permutation scan test (blue); and the rank scan test (red). On the left are the results for $|S| = 2^7$ and on the right for $|S| = 2^{10}$. $n = 2^{15}$ in both cases. Each situation was repeated 1,000 times and each time 200 permutations were drawn for calibration. The vertical black dashed line is the minimax boundary for t. The horizontal black dashed line is the significance level 0.05.

5.6.3 Comparison with RSI

Next, we compare our rank scan with the robust segment identifier (RSI) of Cai, Jeng & Li [32]. This is a recent method based taking the median over bins of a certain size (a tuning parameter of the method) and then scanning over intervals. Because the median is asymptotically normal, it allows for a calibration that only requires the value of the null density at 0. In turn, one can try to estimate this parameter. Although the method is not distribution-free proper, it appears to be the main contender in the literature. Note that the RSI was developed with a different task in mind than that investigated here, namely for identifying anomalous intervals. Hence, we first compare the two methods on simulated data in the context of detection (the problem we considered so far), and in the context of identification (a problem the RSI is designed for).

Detection: In the problem of detection, we compare the performance of the rank scan test and RSI with bin size $m \in \{10, 20\}$ in normal data. To turn RSI into a test, we reject if it detects any anomalous interval. In the simulation, we set sample size $n = 50,000$ and considered the case where there is only one signal interval with known length $|S| \in \{100, 1000\}$. The amplitude satisfies (5.21) as before. We report the empirical power curves (based on 100 repeats) in Figure 5.2.

To be fair, both methods only scan candidate signal intervals of length $|S|$. The rank scan is calibrated as before. For RSI, we set the threshold to $\sqrt{2 \log n}$ for the normalized data after localization to better control the family-wise type I error as explained in Cai, Jeng & Li [32]. From Figure 5.2, we can see that RSI is a bit more conservative. In fact, a drawback of RSI is the difficulty to calibrate it correctly.[9] In any case, the rank scan test outperforms RSI in these simulations.

Identification: In the problem of identification, we compare the rank scan and RSI. Although we focused on the problem of detection so far, a scan can be as easily used for testing as for estimation (i.e., identification). Indeed, one sets an identification threshold and extract all the intervals that exceed that threshold. Some post-processing - such as merging significant intervals that intersect or keeping the most significant among significant intervals that intersect - is often applied.

[9] Of course, it could be calibrated by permutation, but this would make the procedure much more like the permutation scan test (with the same high-computational burden), somewhat far from the intentions of Cai, Jeng & Li [32].

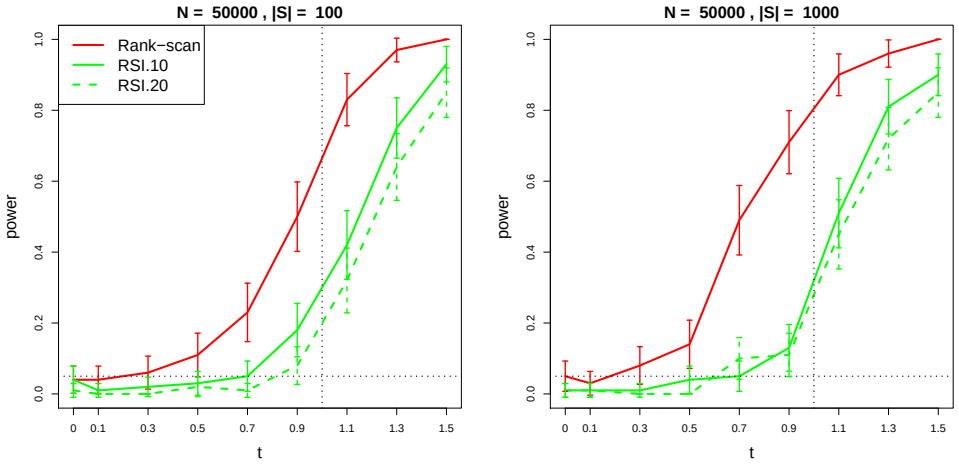

Figure 5.2: Power curves (with 95% margin of error) for the three tests as a function of the parameter t in (5.21): the rank scan test (red); RSI with bin size 10 (solid green); and RSI with bin size 20 (dashed green). The rank scan test is set at level 0.05 and its critical value is from 1000 repeats. On the left are the results for $|S| = 100$ and on the right for $|S| = 1000$. $n = 50,000$ in both cases. Each situation was repeated 100 times. The vertical black dashed line is the minimax threshold for t. The horizontal black dashed line is the significance level 0.05.

Here, in an effort to be fair, we simply took the procedure of Cai, Jeng & Li [32] - which is essentially the procedure of the same authors in [90] - but scanning ranks and calibrating as we did for testing. Note that this implies a very stringent false identification rate (at the 0.05 testing level, this means that the chances that one or more intervals are identified by mistake is 0.05).

Following Cai, Jeng & Li [32], in the simulation, we set $n = 10^4$. We consider a range of null distributions: the standard normal distribution, the t-distribution with 15 degrees of freedom and that with one degree of freedom. In each case, we set the signal mean to $\theta_S \in \{1, 1.5, 2\}$. There are three signal intervals, S_1, S_2, S_3, starting at positions 1000, 2000, 3000, and having lengths $2^4, 2^5, 2^6$, respectively. We set the threshold for the rank scan test by simulation at a significance level of 0.05. For RSI, we tried several bin sizes, $m \in \{2^3, 2^5\}$. To simplify the computation, both methods only scan dyadic intervals of length at most 2^6. As in Cai, Jeng & Li [32], we compare their performance in terms of the following dissimilarities

$$D_j = \min_{\widehat{S} \in \widehat{\mathcal{C}}} \{1 - \rho(S_j, \widehat{S})\},$$

and the number of false positives, namely

$$O = \{\widehat{S} \in \widehat{\mathcal{C}} : \widehat{S} \cap S = \emptyset, \forall S \in \mathcal{C}\},$$

where $\widehat{\mathcal{C}}$ are the estimated signal intervals.

We report the average and standard deviation (in the parenthesis in the tables below) based on 200 repeats in Tables 5.1, 5.2, and 5.3. We can see that the rank scan method performs better than RSI in when the null distribution is normal and $t(15)$, and it performs similarly to RSI with bin size $m = 2^3$ in $t(1)$. However, when the bin size of RSI is not properly chosen, RSI can perform poorly.

Table 5.1: Dissimilarity and number of over-selected intervals in $N(0, 1)$

θ_S	Method	$D_1(\lvert S_1\rvert = 2^4)$	$D_2(\lvert S_2\rvert = 2^5)$	$D_3(\lvert S_3\rvert = 2^6)$	$\#O$
1	Rank Scan	0.734 (0.421)	0.148 (0.284)	0.031 (0.049)	0.000 (0.000)
	RSI($m = 2^3$)	0.916 (0.235)	0.420 (0.406)	0.095 (0.091)	0.065 (0.267)
	RSI($m = 2^5$)	0.998 (0.029)	0.959 (0.144)	0.326 (0.278)	0.130 (0.337)
1.5	Rank Scan	0.167 (0.326)	0.019 (0.044)	0.008 (0.012)	0.000 (0.000)
	RSI($m = 2^3$)	0.593 (0.391)	0.132 (0.033)	0.069 (0.029)	0.080 (0.272)
	RSI($m = 2^5$)	0.980 (0.087)	0.729 (0.284)	0.204 (0.044)	0.025 (0.157)
2	Rank Scan	0.018 (0.051)	0.006 (0.024)	0.004 (0.008)	0.000 (0.000)
	RSI($m = 2^3$)	0.277 (0.226)	0.128 (0.021)	0.064 (0.013)	0.065 (0.247)
	RSI($m = 2^5$)	0.960 (0.122)	0.476 (0.162)	0.193 (0.032)	0.010 (0.100)

5.6.4 Application to the real data

In this section, we apply the methods to the problem of detecting the copy number variant (CNV) in the context of next generation sequencing data. We compare the rank scan method and RSI on the task of identifying short reads on chromosome 19 of a HapMap Yoruban female sample (NA19240) from the 1000 genomes project (`http://www.1000genomes.org`), which is the same data set used by Cai, Jeng & Li in [32]. Following standard protocols (see Ernst et al. [69]), we extend all the reads to 100 base pairs (BPs). We take 10^6 reads from the whole data set for comparison purposes resulting in 1,281,502 genomic locations.

We tune RSI as done by Cai, Jeng & Li in [32], setting the bin size to $m = 400$

Table 5.2: Dissimilarity and number of over-selected intervals in $t(15)$

θ_S	Method	$D_1(\lvert S_1\rvert = 2^4)$	$D_2(\lvert S_2\rvert = 2^5)$	$D_3(\lvert S_3\rvert = 2^6)$	$\#O$
1	Rank Scan	0.806 (0.369)	0.223 (0.354)	0.029 (0.048)	0.000 (0.000)
	RSI($m = 2^3$)	0.926 (0.223)	0.436 (0.406)	0.106 (0.099)	0.050 (0.218)
	RSI($m = 2^5$)	0.996 (0.041)	0.944 (0.168)	0.336 (0.278)	0.125 (0.332)
1.5	Rank Scan	0.232 (0.378)	0.026 (0.079)	0.010 (0.017)	0.000 (0.000)
	RSI($m = 2^3$)	0.554 (0.391)	0.143 (0.112)	0.069 (0.031)	0.075 (0.282)
	RSI($m = 2^5$)	0.992 (0.057)	0.732 (0.286)	0.199 (0.042)	0.020 (0.140)
2	Rank Scan	0.034 (0.097)	0.009 (0.019)	0.005 (0.014)	0.000 (0.000)
	RSI($m = 2^3$)	0.277 (0.220)	0.128 (0.022)	0.063 (0.013)	0.060 (0.238)
	RSI($m = 2^5$)	0.968 (0.107)	0.521 (0.214)	0.192 (0.030)	0.010 (0.100)

Table 5.3: Dissimilarity and number of over-selected intervals in $t(1)$

θ_S	Method	$D_1(\lvert S_1\rvert = 2^4)$	$D_2(\lvert S_2\rvert = 2^5)$	$D_3(\lvert S_3\rvert = 2^6)$	$\#O$
1	Rank Scan	0.989 (0.082)	0.878 (0.305)	0.461 (0.448)	0.000 (0.000)
	RSI($m = 2^3$)	0.950 (0.186)	0.764 (0.370)	0.332 (0.358)	4.305 (5.653)
	RSI($m = 2^5$)	0.998 (0.022)	0.982 (0.098)	0.609 (0.392)	0.520 (0.501)
1.5	Rank Scan	0.922 (0.251)	0.542 (0.455)	0.067 (0.132)	0.000 (0.000)
	RSI($m = 2^3$)	0.843 (0.307)	0.342 (0.354)	0.104 (0.080)	3.920 (2.082)
	RSI($m = 2^5$)	0.983 (0.079)	0.877 (0.236)	0.225 (0.111)	0.055 (0.229)
2	Rank Scan	0.763 (0.410)	0.206 (0.333)	0.043 (0.093)	0.000 (0.000)
	RSI($m = 2^3$)	0.619 (0.382)	0.154 (0.121)	0.089 (0.063)	3.945 (2.385)
	RSI($m = 2^5$)	0.978 (0.090)	0.667 (0.280)	0.208 (0.05)	0.060 (0.238)

and the maximum BPs in a possible CNV to $L = 2^{16}$. Note that Cai, Jeng & Li [32] took $L = 60,000$, which is a bit smaller than 2^{16} (we chose the latter because we only scan intervals of dyadic length). To save computational time, in the implementation of the rank scan we group read depths in every 200 positions and take the summation of the read depths for each bin and use that as input (meaning, we rank the sums and scan the ranks). We get the critical value for the rank scan method under the significance level 0.05 from 1000 repeats. In the

experiment, we let RSI and the rank scan method only scan dyadic intervals of lengths from 2^1 to 2^{16}.

After merging the contiguous selected segments, RSI found 30 possible CNVs and the rank scan method selected 34. Figure 5.3 shows the histograms of the read depths of the selected CNVs. We can see the read depth in the rank scan method is generally larger than that in RSI.

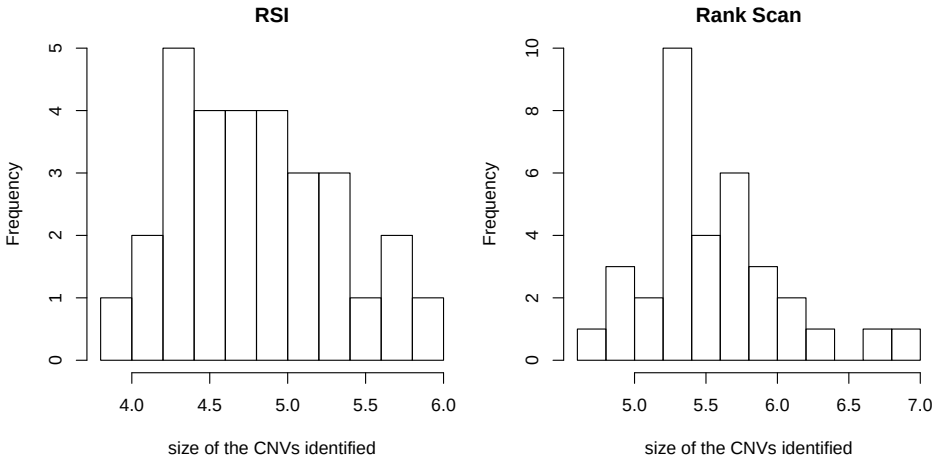

Figure 5.3: Histogram of the read depths of the selected CNVs in log scale (base 10). Both methods only scan dyadic intervals of lengths from 2^1 to 2^{16}. The RSI used a bin size $m = 400$, while the rank scan was calibrated as for testing.

5.7 Discussion

In this chapter we have considered a prototypical structured detection setting with the particularity that the null distribution is unknown. When the null distribution is known, various works have shown that a form of scan test achieves the best possible asymptotic power. When the null distribution is unknown, one can alternatively calibrate the scan test by permutation. This has been suggested a number of times in the detection literature. Theorem 5.1 implies doing this results in no loss of asymptotic power compared to a calibration by Monte Carlo with full knowledge of the null distribution. To circumvent the expense of calibrating by permutation, we propose to scan the ranks. Theorem 5.2 implies that this results in very little loss in asymptotic power. In our empirical experiments all three methods perform

comparably.

Censoring before permutation. When F_0 is not of compact support, we can enforce it by applying a censoring of the form $\widetilde{Y}_i = Y_i \mathbf{1}\{|Y_i| \le t\} + t \cdot \mathrm{sign}(Y_i)\mathbf{1}\{|Y_i| > t\}$. With a choice of threshold $t = t_n \to \infty$ slowly (e.g., $t_n = \log\log n$), Theorem 5.1 applies unchanged and without an upper bound on the θ_i's, and the proof is identical except for very minor modifications. This censoring has the added advantage of making the method more robust to possible outliers.

Other scoring functions. Although rank-sums are intuitive and classically used, any scan based on $h(\mathbf{r}_i)$, where h is increasing, is valid (recall that $\mathbf{r}_i$ is the rank of x_v in the sample). In two-sample testing, it is known that there is no uniformly best choice for the function h. See Section 6.9 in Lehmann & Romano [102] where it is shown that choosing $h(\mathbf{r}) = \mathbb{E}(Z_{(\mathbf{r})})$ - where $Z_{(1)} < \cdots < Z_{(n)}$ are the order statistics of a standard normal sample - is (in some sense) optimal in the normal location model. Our method of proof applies to a general h.

Unstructured subsets. No permutation approach (including a rank-based approach) has any power for detecting unstructured anomalies. A prototypical example is when $\mathcal{C}$ is the class of all subsets, or all subsets of given size, the latter including the class of singletons. The reason behind this is that in this case the class of possible anomalies is closed under permutations, hence the scan statistic takes the same value for every permutation of the data.

5.A Sketch proof of Lemma 5.3

First, assume that there are no ties in the ranks, with probability one. Note that we can write

$$R_i = \sum_{j\in[n],j\neq i} \mathbf{1}\{X_i > X_j\} + 1 = \sum_{j\in[s],j\neq i} \mathbf{1}\{X_i > X_j\} + \sum_{j\notin[s],j\neq i} \mathbf{1}\{X_i > X_j\} + 1\ .$$

Taking expectation yields

$$\mathbb{E}(R_i) = \begin{cases} (n-s)p_{\theta_i} + \sum_{j\in[s],j\neq i} p_{\theta_i,\theta_j} + 1 & \text{, when } i \in [s], \\ \frac{n+s+1}{2} - \sum_{j\in[s]} p_{\theta_j} & \text{, when } i \notin [s]. \end{cases}$$

since $\mathbb{P}(X_i = X_j) = 0$ for $i \neq j$ when there are no ties. The variance and covariance terms can be worked out using the same representation of the ranks as above, but we omit these straightforward computations for the sake of space.

In case of ties, to keep the presentation simple, assume that the distributions of $\{X_i\}_{i\in[n]}$ are supported on $\mathbb{Z}$. Then randomly breaking ties in the ranks amounts to using the following procedure. Let $\{\epsilon_i\}_{i\in[n]}$ be independent and uniformly distributed on $(-c, c)$ with $c \leq 1/2$, also independent from $\{X_i\}_{i\in[n]}$. Consider $X'_i = X_i + \epsilon_i,\ i \in [n]$ and let R'_i be the rank of X'_i in the combined sample $\{X'_i\}_{i\in[n]}$. Then the joint distribution of $\{R'_i\}_{i\in[n]}$ is the same as that of $\{R_i\}_{i\in[m]}$ when ties are broken randomly.

For instance, for $i \notin [s]$

$$\begin{aligned}\mathbb{E}(R'_i) &= \tfrac{n+s+1}{2} - \sum_{j\in[s]} \mathbb{P}(X'_i > X'_j) \\ &= \tfrac{n+s+1}{2} - \sum_{j\in[s]} (\mathbb{P}(X_i > X_j) + \mathbb{P}(\epsilon_i > \epsilon_j | X_i = X_j)\mathbb{P}(X_i = X_j))\end{aligned}$$

$\frac{n+s+1}{2} - \sum_{j\in[s]} p_{\theta_j}$.

The rest of the claims can be worked out similarly.

Finally, when X_i have arbitrary distributions a similar method can be applied, although it requires a bit more care and one needs to take c approaching zero.

5.B Derivation of Υ_0 in the normal location model

Assume the normal model where $F_\theta = N(\theta, 1)$. For this case we can simply compute Υ_0. Since there are no ties with probability 1, we have

$$\Upsilon_0 = \mathbb{E}(X\mathbf{1}\{X > Y\}) = \int_{-\infty}^{\infty}\int_{x}^{\infty} u f_0(u)\,\mathrm{d}u f_0(x)\,\mathrm{d}x\ .$$

Considering the inner integral we have

$$\int_x^\infty u f_0(u)\,\mathrm{d}u = \frac{1}{\sqrt{2\pi}}\int_x^\infty u e^{-u^2/2}\,\mathrm{d}u = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} = f_0(x)\ .$$

Hence

$$\Upsilon_0 = \int_{-\infty}^{\infty} f_0(x) = \int_{-\infty}^{\infty} \frac{1}{2\pi} e^{-x^2}\,\mathrm{d}x = \frac{1}{2\sqrt{\pi}}\ .$$

We conclude that, in the normal location model, $\tau_0 = \sqrt{\pi/3}$ as claimed earlier.

Chapter 6

Concluding remarks

Empirical research is a cornerstone of the scientific learning process. In most fields of science there is a stage of the scientific process where one has to verify theories or hypotheses using empirical data. Mathematical statistics is the discipline underpinning empirical research, providing methods with which the researcher can evaluate data to draw conclusions, and providing theoretical foundations to how accurately this can be done.

Considering the process of empirical learning, note that this is inherently an active mechanism in the following sense. The researcher formulates theories about the phenomenon in question, designs experiments to test these theories, and performs them to acquire data. After analyzing the data, the researcher adjusts his/her theories based on what he/she learned, and undertakes the previous sequence of actions anew. This cartoon holds for most learning processes, even ones outside the scientific world.

The focus of this work was to understand how efficient adaptive sensing is. We had the following questions in mind: Is it worth putting the effort into designing strategies to close the loop between data collection and inference, or would simply collecting data first and analyzing it afterwards be just as efficient? In case there is a benefit to active learning, what are good active learning strategies? Which are the best ones, and what is the best performance we can hope for?

We have investigated the above questions in the context of high-dimensional support recovery and signal detection. We have found that in these settings active learning is advantageous, often extremely so, compared to non-adaptive sensing. We also gained an understanding of how active learning procedures work in this

context, and what the fundamental performance limits of such procedures are. Although we have gained valuable insights about active learning, as usual, there are more questions than answers. We highlight a few below.

Robust active learning algorithms: In this work we have taken a more theoretical point of view, and have focused more on gaining fundamental understanding about active learning and less on designing practical active learning algorithms. Often the main purpose of the procedures that we have presented was to illustrate the tightness of the conditions that we have derived concerning the fundamental limits of active learning. Because of this, most of our procedures require knowledge of parameters that would perhaps be unavailable in practice, and rely on strict distributional assumptions that might be violated in real applications.

To some extent the above phenomena could be circumvented by simple modifications to the algorithms in this thesis. For instance, even though the algorithms in Chapters 2 and 3 required knowledge of the sparsity, in fact those algorithms have a mild adaptivity to sparsity. Similarly, the Sequential Likelihood Ratio Test is often at the core of those algorithms, which relies on knowledge of the underlying distributions. Replacing this with some sort of sequential thresholding procedure, one could deal with slight misspecifications of the underlying distributions.

Nonetheless, now that we see the gains that adaptive sensing can provide in this context, designing adaptive sensing procedures specifically for practical use could be of great value. We have seen in Chapter 5 that in the non-adaptive setting, distribution-free procedures perform nearly as well asymptotically as those with full knowledge of the underlying distributions, when the data distribution is a member of the exponential family. Therefore, designing similar procedures for the adaptive sensing setting, and understanding their performance could prove to be a fruitful topic of future research.

Fundamental limits of active learning in general: Although we could derive the fundamental limits of adaptive sensing in specific settings, this is a challenging task in general. The main method for understanding fundamental difficulties of statistical inference problems is to evaluate certain divergence metrics between distributions. Even when the sample is collected non-adaptively such computations might not be completely straightforward. However, such tasks become increasingly difficult in an adaptive sensing setting, because sensing actions depend on past

observations, due to which we lose the independence of the sample.

Even when we have tight bounds on these divergence metrics, obtaining sharp lower bounds for a specific class of supports often involves selecting an appropriate subclass. The subclass needs to have two properties: it needs to be simple so that we are able to compute the divergence metrics, but also the essence of the problem has to be captured in the subclass. For a given class of supports, this requires a good intuition about which subclasses account for the main difficulty of the recovery problem, and so this is difficult to generalize for arbitrary classes.

Because of the above, there are a number of open questions concerning the fundamental limits of adaptive sensing throughout this work. For instance, we are missing lower bounds for general structured classes both for the coordinate-wise sampling and the compressive sensing settings. Such lower bounds might be easier to obtain in the coordinate-wise sampling case. In particular, the procedures presented in Chapter 2 are based on a rather general idea. For a given class of supports, take the noiseless case procedure that is optimal for this class in terms of sample complexity, and then robustify it using a SLRT, which we know is also optimal in terms of the amount of precision it requires (in some sense). Heuristically, this seems to be a good candidate for an optimal procedure for general classes, and there might be a way to make such an argument formal.

Even if fully general lower bounds are temporarily out of reach, obtaining tighter upper bounds for the divergence metrics in the adaptive sensing setting would be very valuable. We illustrate this by considering the class of submatrices. To be a bit blunt, note that every lower bound in Chapters 2 and 3 essentially reduces the estimation problem to a detection problem. For certain classes (such as intervals), this reduction leads to optimal lower bounds. For submatrices it does not, at least not in the compressive sensing setting. We conjecture that the bottleneck of submatrix estimation (using adaptive sensing) is finding an active row or an active column[1], and one could quickly come up with a subclass of submatrices that captures this aspect. However, we were unable to use this idea to obtain a tight lower bound, because we lack tight bounds for the divergence metrics in the adaptive sensing setting.

Note that in the non-adaptive sensing setting, such bounds are obtained by noting certain properties of the optimal estimators. In particular, for symmetric

[1]Note that in the coordinate-wise sampling setting this is the same as the detection problem, so in that case this reduction works. However, in the compressive sensing case, this question is fundamentally different than the detection problem.

classes, our resources (precision for coordinate-wise samples and sensing energy for compressive sensing) have to be distributed evenly across the signal vector. Similar properties also hold in the adaptive sensing setting, but unfortunately these alone are not sufficient to obtain tight bounds. However, by exploring further conditions that the optimal support recovery algorithms must satisfy, we might be able to gain better control on the divergence metrics.

Another approach would be to explore different lower bounding techniques in the literature (for instance Fano's inequality). These would present a different set of challenges, but perhaps those could be overcome more easily. Such methods could also lead to deriving necessary conditions for the sample complexity of adaptive compressive sensing, and understanding the fundamental difficulty of the detection of signals evolving in time.

Combinatorial bandits: A natural generalization of the problems that we have considered in this work is to consider signals whose coordinates can take arbitrary values. If we consider settings where the signal is still sparse but the values of the active components can be arbitrary, we can obtain similar results to those presented in this thesis. This was already addressed in some comments throughout the previous chapters. However, when the signal is no longer sparse, the nature of such problems changes radically. In such settings, instead of the support of the signal, we are usually looking for the components that have the highest values.

Such problems are called multi-armed bandit problems, and were already examined in the past due to their high relevance in applications. There are multiple ways in which multi-armed bandit problems can be cast. In the classical setup the task is to minimize a cumulative regret over time as introduced by Robbins [126], whereas in more recent settings, the aim is to find the arm with the highest reward after a fixed number of pulls, which is called the best-arm problem (see Audibert & Bubeck [16]).

The latter problem is closely related to the coordinate-wise sampling setting considered earlier, and in fact can be viewed as a generalization of it. Consider the setting in Chapter 2, and recall that $\boldsymbol{x}$ denotes the signal vector, and $\mathcal{C}$ denotes the class of possible supports. Note that the support recovery problem can be formulated as finding the set in the class for which $\sum_{i \in S} \boldsymbol{x}_i$ is maximal. By letting the components of $\boldsymbol{x}$ take arbitrary values, and setting $\mathcal{C}$ to be the class of singletons, we arrive at the best-arm problem described above. In a similar fash-

ion, one could consider this problem for different classes of supports, for instance the ones considered in earlier chapters. This leads to the so-called combinatorial bandit problems.

Although the best arm problem is now well understood, knowledge about the more general combinatorial bandit problem is much more limited. In contrast to the sparse signal setting considered in this work, in multi-armed bandit problems we no longer have a baseline distribution to test against. This creates an additional difficulty, which we know how to overcome in the best arm problem, but is non-trivial to tackle in combinatorial bandit problems. The additional difficulty that combinatorial bandits pose stems from the overlaps between the different sets in the class. Whereas in the best-arm problem every set is a singleton, in combinatorial bandits, sets consist of multiple arms, which increases the uncertainty in our assessment of the values of $\sum_{i \in S} \boldsymbol{x}_i$ for a given $S \in \mathcal{C}$. Furthermore, unlike in the best-arm problem, any given arm contributes its value to many different sets. These facts make combinatorial bandit problems considerably more difficult.

Nonetheless, obtaining results in this setting would both provide valuable insights about the fundamental properties of adaptive sensing, and could also prove to be fruitful for practical applications.

Detecting/estimating correlations: The main theme of the problems investigated in this thesis is anomaly detection and estimation. In each chapter, we have considered an n-dimensional vector that represented some sort of a system, and some of its components could have different than usual values, which represented anomalous behavior in the system. However, there are different ways in modeling anomalous behavior. In particular, note that in the previous setting any given component of the system can be deemed anomalous in isolation. However, in certain situations, such a claim can not be made by examining items in isolation, but rather only when considering different components together. This is referred to as contextual anomaly detection.

A natural way to model this is to consider a model with added covariance instead of added mean. That is, anomalous behavior is modeled as certain components of the signal vector being correlated. In such a setting, one could consider a detection problem (deciding whether correlated components exist) as well as an estimation problem (identifying the correlated components). The detection problem has been investigated both in the non-adaptive sensing setting (see for instance Arias-Castro,

Bubeck & Lugosi [6]) and in the adaptive sensing setting (see Castro, Lugosi & Savalle [41]).

In the adaptive sensing setting, the picture is far from complete. In particular, Castro, Lugosi & Savalle [41] have shown that adaptive sensing can provide an improvement over non-adaptive sensing in certain settings. When the correlated components form an interval in the signal vector, the necessary and the sufficient conditions for detection nearly match. However, when the correlated components are unstructured, there is a gap between the two conditions, possibly due to the lower bounds being loose. Therefore, this problem might be another instance for which the development of general lower bounding techniques for adaptive sensing (as described above) could be extremely valuable.

Summary

On High-Dimensional Support Recovery and Signal Detection

This work investigates questions related to support recovery problems. This is a group of statistical learning problems in which the goal is to identify items with unusual/anomalous behavior among a large number of items. Typical examples of such problems arise in the fields of computer vision, genetic research or in the analysis of astronomical data. The items themselves cannot be observed directly, but rather through some sort of observation noise, which makes this a statistical learning/multiple hypotheses testing task in nature. Furthermore, in most interesting applications the number of items is very large but the number of anomalous ones among these is relatively small.

We investigate instances of the problem above where observations can be gathered in a sequential and adaptive manner. This paradigm is known as active learning/adaptive sensing as here the learner can control how data is collected while the data is being acquired. Adaptive sensing algorithms can potentially exploit the fact that only a small fraction of the items are anomalous by allocating sensing efforts more efficiently among various items. Despite their practical appeal a theoretical understanding of such procedures is challenging. Most of this thesis is aimed at providing a thorough foundational understanding of adaptive sensing algorithms.

First we consider a model in which we are allowed to sequentially collect observations of items corrupted by additive Gaussian noise, but the total precision of the measurements we make must not exceed a predetermined threshold. Our goal is to understand how much advantage do adaptive sensing procedures have in this setup compared to non-adaptive ones. To answer this question we first propose

algorithms for support recovery and analyze their performance. We then show near optimality of these algorithms, characterizing the fundamental difficulty of the support recovery task in this model. Finally we show the best possible performance of non-adaptive algorithms and contrast this to the performance of adaptive ones established previously.

Next we consider an adaptive compressive sensing model. Instead of individual measurements of the items, in compressive sensing we observe dot products between the signal vector and sensing vectors of our choice. These dot products are corrupted by additive Gaussian noise. Here the norms of the sensing vectors play the same role as the measurement precision in the previous model. Our research questions and results parallel those outlined in the previous paragraph, only now concerning the adaptive compressive sensing model outlined above.

Then we investigate the support recovery problem in a dynamical setting, when the identity of the anomalous items changes over time. Our main interest is to understand the fundamental effect the rate of change of the anomalous items has on the difficulty of the detection problem. To this end we introduce a simple model in which at each time step each signal component has a fixed probability of moving to a different location, and these moves happen independently. At each time step we are allowed to collect a noisy measurement of an item of our choice. We analyze this problem under both adaptive and non-adaptive sensing paradigms, and characterize the difficulty of signal detection in both setups.

Complementing the theoretical work above, we provide some results with a strong methodological motivation. We consider the more classical setup (non-adaptive sensing) when we are provided with one observation per item, assumed to belong to some family of distributions. Though this setup has been widely studied for a multitude of distributions, often in practice such distributional assumptions may be violated. Motivated by this we introduce distribution-free tests for the problem of detecting the presence of anomalous items. We analyze the performance of these tests and show that for distributions in the exponential family, our proposed tests have near optimal performance.

Curriculum Vitae

Ervin Tamás Tánczos was born on the 28th of August, 1985, in Szeged, Hungary. Ervin completed secondary school in 2004 at Ságvári Endre High School in Szeged, Hungary. In the same year he went on to study Applied Mathematics at the Eötvös Loránd University (ELTE). His main fields of interest were probability theory, statistics and information theory. He obtained his M.Sc. degrees from ELTE in 2010.

On the 1st of September, 2010, Ervin started working as a graduate student at the Szegedi Tudományegyetem (SZTE), in Szeged, Hungary, at the Department of Medical Informatics (DMI). There, he performed statistical data analysis for various medical research projects, contributing to several publications within the department.

On the 1st of September, 2012, Ervin started working as a graduate student at the Eindhoven University of Technology (TU/e), in Eindhoven, the Netherlands, under the supervision of dr. Rui M. Castro. As part of his PhD project, he visited to the University of California, San Diego (UCSD) and the University of Wisconsin - Madison, both for one month. Ervin defends his thesis at TU/e on the 12th of September, 2016.

On the 1st of October, 2016, Ervin starts as a postdoctoral research fellow at the University of Wisconsin - Madison, in Madison, Wisconsin, USA.

Bibliography

[1] Addario-Berry, L., Broutin, N., Devroye, L., and Lugosi, G. On combinatorial testing problems. *The Annals of Statistics 38*, 5 (2010), 3063–3092.

[2] Aksoylar, C., Atia, G., and Saligrama, V. Sparse signal processing with linear and non-linear observations: A unified shannon theoretic approach. In *Information Theory Workshop (ITW), 2013 IEEE* (2013), IEEE, pp. 1–5.

[3] Aksoylar, C., and Saligrama, V. Information-theoretic bounds for adaptive sparse recovery. *arXiv preprint arXiv:1402.5731* (2014).

[4] Amini, A. A., and Wainwright, M. J. High-dimensional analysis of semidefinite programming relaxations for sparse principal component analysis. *The Annals of Statistics 37*, 5B (2009), 2877–2921.

[5] Arias-Castro, E. Detecting a vector based on linear measurements. *Electronic Journal of Statistics 6* (2012), 547–558.

[6] Arias-Castro, E., Bubeck, S., and Lugosi, G. Detection of correlations. *The Annals of Statistics 40*, 1 (2012), 412–435.

[7] Arias-Castro, E., Candès, E. J., and Davenport, M. A. On the fundamental limits of adaptive sensing. *IEEE Transactions on Information Theory 59*, 1 (2013), 472–481.

[8] Arias-Castro, E., Candès, E. J., and Durand, A. Detection of an anomalous cluster in a network. *The Annals of Statistics 39*, 1 (2011), 278–304.

[9] Arias-Castro, E., Candès, E. J., Helgason, H., and Zeitouni, O. Searching for a trail of evidence in a maze. *The Annals of Statistics* (2008), 1726–1757.

[10] Arias-Castro, E., Candès, E. J., and Plan, Y. Global testing under sparse alternatives: Anova, multiple comparisons and the higher criticism. *The Annals of Statistics 39*, 5 (2011), 2533–2556.

[11] Arias-Castro, E., Castro, R. M., Tánczos, E., and Wang, M. Distribution-free detection of structured anomalies: Permutation and rank-based scans. *arXiv preprint arXiv:1508.03002* (2015).

[12] Arias-Castro, E., Donoho, D., and Huo, X. Near-optimal detection of geometric objects by fast multiscale methods. *IEEE Transactions on Information Theory 51*, 7 (2005), 2402–2425.

[13] Arias-Castro, E., and Grimmett, G. R. Cluster detection in networks using percolation. *Bernoulli 19*, 2 (2013), 676–719.

[14] Arias-Castro, E., and Sharpnack, J. Exact asymptotics for the scan statistic and fast alternatives. *arXiv preprint arXiv:1409.7127* (2014).

[15] Arias-Castro, E., and Verzelen, N. Community detection in random networks. *arXiv preprint arXiv:1302.7099* (2013).

[16] Audibert, J.-Y., and Bubeck, S. Best arm identification in multi-armed bandits. In *COLT-23th Conference on Learning Theory-2010* (2010), pp. 13–p.

[17] Balakrishnan, S., Kolar, M., Rinaldo, A., and Singh, A. Recovering block-structured activations using compressive measurements. *arXiv preprint arXiv:1209.3431* (2012).

[18] Balakrishnan, S., Kolar, M., Rinaldo, A., Singh, A., and Wasserman, L. Statistical and computational tradeoffs in biclustering. In *NIPS 2011 Workshop on Computational Trade-offs in Statistical Learning* (2011), vol. 4.

[19] Balcan, M.-F., Beygelzimer, A., and Langford, J. Agnostic active learning. *Journal of Computer and System Sciences 75*, 1 (2009), 78–89.

[20] Baraniuk, R. G., Cevher, V., Duarte, M. F., and Hegde, C. Model-based compressive sensing. *IEEE Transactions on Information Theory 56*, 4 (2010), 1982–2001.

[21] Baraud, Y. Non-asymptotic minimax rates of testing in signal detection. *Bernoulli 8*, 5 (2002), 577–606.

[22] Bardenet, R., and Maillard, O.-A. Concentration inequalities for sampling without replacement. *arXiv preprint arXiv:1309.4029* (2013).

[23] Bayraktar, E., and Lai, L. Byzantine fault tolerant distributed quickest change detection. *SIAM Journal on Control and Optimization 53*, 2 (2015), 575–591.

[24] Berthet, Q., and Rigollet, P. Complexity theoretic lower bounds for sparse principal component detection. In *Conference on Learning Theory* (2013), pp. 1046–1066.

[25] Berthet, Q., and Rigollet, P. Optimal detection of sparse principal components in high dimension. *The Annals of Statistics 41*, 4 (2013), 1780–1815.

[26] Bessler, S. A. Theory and applications of the sequential design of experiments, k-actions and infinitely many expreiments. part i - theory. Tech. rep., Tech. Rep. Applied Mathematics and Statistics Laboratories, Stanford University, 1960.

[27] Blanchard, G., and Geman, D. Hierarchical testing designs for pattern recognition. *Annals of Statistics* (2005), 1155–1202.

[28] Boucheron, S., Lugosi, G., and Massart, P. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

[29] Boutsikas, M. V., and Koutras, M. V. On the asymptotic distribution of the discrete scan statistic. *Journal of Applied Probability 43*, 4 (2006), 1137–1154.

[30] Brennan, S. M., Mielke, A. M., Torney, D. C., and MacCabe, A. B. Radiation detection with distributed sensor networks. *Computer 37*, 8 (2004), 57–59.

[31] Butucea, C., and Ingster, Y. I. Detection of a sparse submatrix of a high-dimensional noisy matrix. *Bernoulli 19*, 5B (2013), 2652–2688.

[32] Cai, T. T., Jeng, J. X., and Li, H. Robust detection and identification of sparse segments in ultrahigh dimensional data analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74*, 5 (2012), 773–797.

[33] Cai, T. T., and Yuan, M. Rate-optimal detection of very short signal segments. *arXiv preprint arXiv:1407.2812* (2014).

[34] Candès, E., and Tao, T. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics* (2007), 2313–2351.

[35] Candès, E. J., and Tao, T. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory 52*, 12 (2006), 5406–5425.

[36] Candès, E. J., and Wakin, M. B. An introduction to compressive sampling. *IEEE Signal Processing Magazine 25*, 2 (2008), 21–30.

[37] Caromi, R., Xin, Y., and Lai, L. Fast multiband spectrum scanning for cognitive radio systems. *IEEE Transactions on Communications 61*, 1 (2013), 63–75.

[38] Castanon, D. A. Optimal search strategies in dynamic hypothesis testing. *IEEE Transactions on Systems, Man and Cybernetics 25*, 7 (1995), 1130–1138.

[39] Castro, R. On the performance of adaptive sensing for sparse signal inference. In *10th international conference on Sampling Theory and Applications (SampTA 2013)* (Bremen, Germany, jul 2013), pp. 160–163.

[40] Castro, R. M. Adaptive sensing performance lower bounds for sparse signal estimation and testing. *Bernoulli 20*, 4 (2014), 2217–2246.

[41] Castro, R. M., Lugosi, G., and Savalle, P.-A. Detection of correlations with adaptive sensing. *IEEE Transactions on Information Theory 60*, 12 (2014), 7913–7927.

[42] Castro, R. M., and Nowak, R. D. Minimax bounds for active learning. *IEEE Transactions on Information Theory 54*, 5 (2008), 2339–2353.

[43] Castro, R. M., and Tánczos, E. Adaptive compressed sensing for estimation of structured sparse sets. *arXiv preprint arXiv:1410.4593* (2014).

[44] Castro, R. M., and Tánczos, E. Adaptive sensing for estimation of structured sparse signals. *IEEE Transactions on Information Theory 61*, 4 (2015), 2060–2080.

[45] Chamberlin, P., Pesnell, W., and Thompson, B. *The Solar Dynamics Observatory*. Springer Science & Business Media, 2012.

[46] Chernoff, H. Sequential design of experiments. *The Annals of Mathematical Statistics 30*, 3 (1959), 755–770.

[47] Cheung, Y. T. D., Spittal, M. J., Williamson, M. K., Tung, S. J., and Pirkis, J. Application of scan statistics to detect suicide clusters in australia. *PLoS ONE 8*, 1 (01 2013), e54168.

[48] Cohen, K., and Zhao, Q. Active hypothesis testing for anomaly detection. *IEEE Transactions on Information Theory 61*, 3 (2015), 1432–1450.

[49] Cohen, K., and Zhao, Q. Asymptotically optimal anomaly detection via sequential testing. *IEEE Transactions on Signal Processing 63*, 11 (2015), 2929–2941.

[50] Cohen, K., Zhao, Q., and Swami, A. Optimal index policies for quickest localization of anomaly in cyber networks. In *IEEE Global Conference on Signal and Information Processing (GlobalSIP)* (2013), IEEE, pp. 221–224.

[51] Cohen, K., Zhao, Q., and Swami, A. Optimal index policies for anomaly localization in resource-constrained cyber systems. *IEEE Transactions on Signal Processing 62*, 16 (2014), 4224–4236.

[52] Cohn, D. A., Ghahramani, Z., and Jordan, M. I. Active learning with statistical models. *Journal of Artificial Intelligence Research 4* (1996), 129–145.

[53] Craig, C. C. On the tchebychef inequality of bernstein. *The Annals of Mathematical Statistics 4*, 2 (1933), 94–102.

[54] Cui, Y., Wei, Q., Park, H., and Lieber, C. M. Nanowire nanosensors for highly sensitive and selective detection of biological and chemical species. *Science 293*, 5533 (2001), 1289–1292.

[55] Culler, D., Estrin, D., and Srivastava, M. Guest editors' introduction: Overview of sensor networks. *Computer*, 8 (2004), 41–49.

[56] Dasgupta, S. Analysis of a greedy active learning strategy. *Advances in neural information processing systems 17* (2005), 337–344.

[57] Dasgupta, S. Coarse sample complexity bounds for active learning. In *Advances in neural information processing systems* (2005), pp. 235–242.

[58] Dasgupta, S., Kalai, A. T., and Monteleoni, C. Analysis of perceptron-based active learning. In *Learning Theory*. Springer, 2005, pp. 249–263.

[59] Davenport, M. A., Massimino, A. K., Needell, D., and Woolf, T. Constrained adaptive sensing. *arXiv preprint arXiv:1506.05889* (2015).

[60] Dembo, A., and Zeitouni, O. *Large deviations techniques and applications*, vol. 38 of *Stochastic Modelling and Applied Probability*. Springer-Verlag, Berlin, 2010. Corrected reprint of the second (1998) edition.

[61] Desolneux, A., Moisan, L., and Morel, J.-M. Maximal meaningful events and applications to image analysis. *The Annals of Statistics 31*, 6 (2003), 1822–1851.

[62] Deutsch, S., Averbush, A., and Dekel, S. Adaptive compressed image sensing based on wavelet modeling and direct sampling. In *SAMPTA'09* (2009), pp. General–session.

[63] Diehl, C. P., and Hampshire, J. B. Real-time object classification and novelty detection for collaborative video surveillance. In *Proceedings of the 2002 International Joint Conference on Neural Networks* (2002), vol. 3, IEEE, pp. 2620–2625.

[64] Donoho, D., and Jin, J. Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics* (2004), 962–994.

[65] Donoho, D. L. Compressed sensing. *IEEE Transactions on Information Theory 52*, 4 (2006), 1289–1306.

[66] Dorfman, R. The detection of defective members of large populations. *The Annals of Mathematical Statistics 14*, 4 (1943), 436–440.

[67] Dragalin, V. A simple and effective scanning rule for a multi-channel system. *Metrika 43*, 1 (1996), 165–182.

[68] Duarte, M. F., Davenport, M. A., Takhar, D., Laska, J. N., Sun, T., Kelly, K. E., and Baraniuk, R. G. Single-pixel imaging via compressive sampling. *IEEE Signal Processing Magazine 25*, 2 (2008), 83–91.

[69] Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature 473*, 7345 (2011), 43–49.

[70] Fedorov, V. V. *Theory of optimal experiments*. Access Online via Elsevier, 1972.

[71] Flenner, A., and Hewer, G. A helmholtz principle approach to parameter free change detection and coherent motion using exchangeable random variables. *SIAM Journal on Imaging Sciences 4*, 1 (2011), 243–276.

[72] Foucart, S., and Rauhut, H. *A Mathematical Introduction to Compressive Sensing*. Birkhäuser, 2013.

[73] Friedman, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association 32*, 200 (1937), 675–701.

[74] Geman, D., and Jedynak, B. An active testing model for tracking roads in satellite images. *IEEE Transactions on Pattern Analysis and Machine Intelligence 18*, 1 (1996), 1–14.

[75] Guerriero, M., Willett, P., and Glaz, J. Distributed target detection in sensor networks using scan statistics. *IEEE Transactions on Signal Processing 57*, 7 (July 2009), 2629–2639.

[76] Gwadera, R., Atallah, M. J., and Szpankowski, W. Reliable detection of episodes in event sequences. *Knowledge and Information Systems 7*, 4 (2005), 415–437.

[77] Hadjiliadis, O., Zhang, H., and Poor, H. V. One shot schemes for decentralized quickest change detection. In *11th International Conference on Information Fusion* (2008), IEEE, pp. 1–8.

[78] Hall, P., and Jin, J. Innovated higher criticism for detecting sparse signals in correlated noise. *Annals of Statistics 38*, 3 (2010), 1686–1732.

[79] Hall, P., and Molchanov, I. Sequential methods for design-adaptive estimation of discontinuities in regression curves and surfaces. *The Annals of Statistics 31*, 3 (2003), 921–941.

[80] Hanneke, S. Rates of convergence in active learning. *The Annals of Statistics 39*, 1 (2011), 333–361.

[81] Haupt, J., Castro, R. M., and Nowak, R. Distilled sensing: Adaptive sampling for sparse detection and estimation. *IEEE Transactions on Information Theory 57*, 9 (2011), 6222–6235.

[82] Haupt, J. D., Baraniuk, R. G., Castro, R. M., and Nowak, R. D. Compressive distilled sensing: Sparse recovery using adaptivity in compressive measurements. In *Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers, 2009* (2009), IEEE, pp. 1551–1555.

[83] Hettmansperger, T. P. *Statistical inference based on ranks*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, 1984.

[84] Hoeffding, W. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association 58* (1963), 13–30.

[85] Holmes, A. P., Blair, R., Watson, G., and Ford, I. Nonparametric analysis of statistic images from functional mapping experiments. *Journal of Cerebral Blood Flow & Metabolism 16*, 1 (1996), 7–22.

[86] Huang, L., Kulldorff, M., and Gregorio, D. A spatial scan statistic for survival data. *Biometrics 63*, 1 (2007), 109–118.

[87] Ingster, Y. I., and Suslina, I. A. Minimax nonparametric hypothesis testing for ellipsoids and besov bodies. *ESAIM: Probability and Statistics 4* (2000), 53–135.

[88] Ingster, Y. I., and Suslina, I. A. On detection of a signal of known shape in multi-channel system. *Zapiski Nauchnykh Seminarov POMI 294* (2002), 88–112.

[89] Jang, J., Brumley, D., and Venkataraman, S. Bitshred: feature hashing malware for scalable triage and semantic analysis. In *Proceedings of the 18th ACM conference on Computer and communications security* (2011), ACM, pp. 309–320.

[90] Jeng, X. J., Cai, T. T., and Li, H. Optimal sparse segment identification with application in copy number variation analysis. *Journal of the American Statistical Association 105*, 491 (2010), 1156–1166.

[91] Jiang, T. Maxima of partial sums indexed by geometrical structures. *The Annals of Probability 30*, 4 (2002), 1854–1892.

[92] Jung, I., and Cho, H. A nonparametric spatial scan statistic for continuous data. *International Journal of Health Geographics 14*, 1 (2015), 30.

[93] Kabluchko, Z. Extremes of the standardized gaussian noise. *Stochastic Processes and their Applications 121*, 3 (2011), 515–533.

[94] Kadane, J. B. Optimal whereabouts search. *Operations Research 19*, 4 (1971), 894–904.

[95] Klimko, E., and Yackel, J. Optimal search strategies for wiener processes. *Stochastic Processes and their Applications 3*, 1 (1975), 19–33.

[96] Koltchinskii, V. Rademacher complexities and bounding the excess risk in active learning. *The Journal of Machine Learning Research 9999* (2010), 2457–2485.

[97] Krishnamuthy, A., Sharpnack, J., and Singh, A. Recovering graph-structured activations using adaptive compressive measurements. In *Asilomar Conference on Signals, Systems and Computers, 2013* (2013), IEEE, pp. 765–769.

[98] Kruskal, W. H., and Wallis, W. A. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association 47*, 260 (1952), 583–621.

[99] Kulldorff, M. A spatial scan statistic. *Communications in Statistics - Theory and Methods 26*, 6 (1997), 1481–1496.

[100] Kulldorff, M., Heffernan, R., Hartman, J., Assunçao, R., and Mostashari, F. A space-time permutation scan statistic for disease outbreak detection. *PLoS medicine 2*, 3 (2005), 216.

[101] Kulldorff, M., Huang, L., and Konty, K. A scan statistic for continuous data based on the normal probability model. *International journal of health geographics 8*, 1 (2009), 58.

[102] Lehmann, E. L., and Romano, J. P. *Testing statistical hypotheses*, third ed. Springer Texts in Statistics. Springer, New York, 2005.

[103] Li, H. Restless watchdog: Selective quickest spectrum sensing in multichannel cognitive radio systems. *EURASIP Journal on Advances in Signal Processing 2009* (2009), 6.

[104] Luo, W., and Tay, W. P. Finding an infection source under the sis model. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2013), IEEE, pp. 2930–2934.

[105] Malloy, M., and Nowak, R. On the limits of sequential testing in high dimensions. In *Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR), 2011* (2011), IEEE, pp. 1245–1249.

[106] Malloy, M., and Nowak, R. Sequential analysis in high-dimensional multiple testing and sparse recovery. In *IEEE International Symposium on Information Theory Proceedings (ISIT), 2011* (2011), IEEE, pp. 2661–2665.

[107] Malloy, M. L., and Nowak, R. D. Near-optimal adaptive compressed sensing. In *Conference Record of the Forty Sixth Asilomar Conference on Signals, Systems and Computers (ASILOMAR), 2012* (2012), IEEE, pp. 1935–1939.

[108] Malloy, M. L., and Nowak, R. D. Near-optimal compressive binary search. *arXiv preprint arXiv:1203.1804* (2012).

[109] Malloy, M. L., and Nowak, R. D. Sequential testing for sparse recovery. *IEEE Transactions on Information Theory 60*, 12 (2014), 7862–7873.

[110] Mann, H. B., and Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics 18* (1947), 50–60.

[111] Moon, N., Bullitt, E., Van Leemput, K., and Gerig, G. Automatic brain and tumor segmentation. In *Medical Image Computing and Computer-Assisted InterventionMICCAI 2002*. Springer, 2002, pp. 372–379.

[112] Moore, W. C., Meyers, D. A., Wenzel, S. E., et al. Identification of asthma phenotypes using cluster analysis in the severe asthma research program. *American journal of respiratory and critical care medicine 181*, 4 (2010), 315–323.

[113] Neill, D. B. Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74*, 2 (2012), 337–360.

[114] Neill, D. B., and Moore, A. W. A fast multi-resolution method for detection of significant spatial disease clusters. In *Advances in Neural Information Processing Systems* (2003), p. None.

[115] Neill, D. B., and Moore, A. W. Rapid detection of significant spatial clusters. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (2004), ACM, pp. 256–265.

[116] Nichols, T. E., and Holmes, A. P. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping 15*, 1 (2002), 1–25.

[117] Nitinawarat, S., Atia, G. K., and Veeravalli, V. V. Controlled sensing for multihypothesis testing. *IEEE Transactions on Automatic Control 58*, 10 (2013), 2451–2464.

[118] Panych, L. P., and Jolesz, F. A. A dynamically adaptive imaging algorithm for wavelet-encoded mri. *Magnetic resonance in medicine 32*, 6 (1994), 738–748.

[119] Pawitan, Y., Michiels, S., Koscielny, S., Gusnanto, A., and Ploner, A. False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics 21*, 13 (2005), 3017–3024.

[120] Perone Pacifico, M., Genovese, C., Verdinelli, I., and Wasserman, L. False discovery control for random fields. *Journal of the American Statistical Association 99*, 468 (2004), 1002–1014.

[121] Phoha, V. V. *Internet security dictionary*. Springer Science & Business Media, 2007.

[122] Pokrajac, D., Lazarevic, A., and Latecki, L. J. Incremental local outlier detection for data streams. In *IEEE Symposium on Computational Intelligence and Data Mining* (2007), IEEE, pp. 504–515.

[123] Qian, J., and Saligrama, V. Efficient minimax signal detection on graphs. In *Advances in Neural Information Processing Systems* (2014), pp. 2708–2716.

[124] Qian, J., Saligrama, V., and Chen, Y. Connected sub-graph detection. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics* (2014), pp. 796–804.

[125] Raghavan, V., and Veeravalli, V. V. Quickest change detection of a markov process across a sensor array. *IEEE Transactions on Information Theory 56*, 4 (2010), 1961–1981.

[126] Robbins, H. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematics Society 58* (1952), 527–535.

[127] Shabalin, A., Weigman, V., Perou, C., and Nobel, A. Finding large average submatrices in high dimensional data. *The Annals of Applied Statistics 3*, 3 (2009), 985–1012.

[128] Shah, D., and Zaman, T. Rumors in a network: who's the culprit? *IEEE Transactions on Information Theory 57*, 8 (2011), 5163–5181.

[129] Sharpnack, J., Rinaldo, A., and Singh, A. Changepoint detection over graphs with the spectral scan statistic. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS)* (2013), vol. 31 of *JMLR W&CP*, pp. 545—553.

[130] Sharpnack, J., and Singh, A. Identifying graph-structured activation patterns in networks. In *Advances in Neural Information Processing Systems* (2010), pp. 2137–2145.

[131] Sharpnack, J. L., Krishnamurthy, A., and Singh, A. Near-optimal anomaly detection in graphs using lovász extended scan statistic. In *Advances in Neural Information Processing Systems* (2013), pp. 1959–1967.

[132] Shorack, G. R., and Wellner, J. A. *Empirical processes with applications to statistics*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, 1986.

[133] Siegmund, D., and Venkatraman, E. S. Using the generalized likelihood ratio statistic for sequential detection of a change-point. *Annals of Statistics 23*, 1 (1995), 255–271.

[134] Soni, A., and Haupt, J. Efficient adaptive compressive sensing using sparse hierarchical learned dictionaries. In *Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR), 2011* (2011), IEEE, pp. 1250–1254.

[135] Soni, A., and Haupt, J. On the fundamental limits of recovering tree sparse vectors from noisy linear measurements. *IEEE Transactions on Information Theory 60*, 1 (2014), 133–149.

[136] Stone, L. D., and Stanshine, J. A. Optimal search using uninterrupted contact investigation. *SIAM Journal on Applied Mathematics 20*, 2 (1971), 241–263.

[137] Szor, P. *The art of computer virus research and defense*. Pearson Education, 2005.

[138] Thompson, D. R., Burke-Spolaor, S., Deller, A. T., et al. Real-time adaptive event detection in astronomical data streams. *IEEE Intelligent Systems 29*, 1 (2014), 48–55.

[139] Tsybakov, A. B. *Introduction to Nonparametric Estimation*, vol. 41 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer, Berlin, 2009.

[140] Wainwright, M. J. Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory 55*, 5 (2009), 2183–2202.

[141] Wald, A. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics 16*, 2 (1945), 117–186.

[142] Wallenstein, S. Joseph naus: Father of the scan statistic. In *Scan Statistics*. Springer, 2009, pp. 1–25.

[143] Walther, G. Optimal and fast detection of spatial clusters with scan statistics. *The Annals of Statistics 38*, 2 (2010), 1010–1033.

[144] Wang, H., Tang, M., Park, Y.-S., and Priebe, C. E. Locality statistics for anomaly detection in time series of graphs. *IEEE Transactions on Signal Processing 62*, 3 (2014), 703–717.

[145] Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics Bulletin 1*, 6 (1945), 80–83.

[146] Willett, R., Nowak, R., and Castro, R. M. Faster rates in regression via active learning. In *Advances in Neural Information Processing Systems* (2005), pp. 179–186.

[147] Yoon, S., Nardini, C., Benini, L., and De Micheli, G. Discovering coherent biclusters from gene expression data using zero-suppressed binary decision diagrams. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 2*, 4 (2005), 339–354.

[148] Zhai, Y., and Zhao, Q. Dynamic search under false alarms. In *IEEE Global Conference on Signal and Information Processing (GlobalSIP)* (2013), IEEE, pp. 201–204.

[149] Zhang, Z., Gao, X., Biswas, J., and Wu, J. K. Moving targets detection and localization in passive infrared sensor networks. In *Information Fusion, 2007 10th International Conference on* (2007), IEEE, pp. 1–6.

[150] Zhao, M., and Saligrama, V. Anomaly detection with score functions based on nearest neighbor graphs. In *Advances in Neural Information Processing Systems* (2009), pp. 2250–2258.

[151] Zhao, Q., and Ye, J. Quickest detection in multiple on–off processes. *IEEE Transactions on Signal Processing 58*, 12 (2010), 5994–6006.

[152] Zhong, Y., Jain, A. K., and Dubuisson-Jolly, M.-P. Object tracking using deformable templates. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 22*, 5 (2000), 544–549.

[153] Zhu, K., and Ying, L. Information source detection in the sir model: A sample path based approach. In *Information Theory and Applications Workshop (ITA)* (2013), IEEE, pp. 1–9.

[154] Zigangirov, K. S. On a problem in optimal scanning. *Theory of Probability & Its Applications 11*, 2 (1966), 294–298.