Extremes In Random Fields:
Theory and Applications
(A draft)

Benjamin Yakir, The Hebrew University

July 4, 2012
## Contents

### I Theory

1 Introduction
   1.1 Distribution of extremes in random fields 4
   1.2 Outline of the method 5
   1.3 Gaussian and asymptotically gaussian random fields 7
   1.4 Applications 9

2 Basic Examples
   2.1 Introduction 14
   2.2 A power-one sequential test 14
   2.3 A kernel-based scanning statistic 20
   2.4 Other methods 31

3 Approximation of the Local Rate
   3.1 Introduction 34
   3.2 Preliminary localization and approximation 35
      3.2.1 Localization 35
      3.2.2 A discrete approximation 40
   3.3 Measure transformation 43
   3.4 Application of the localization theorem 47
   3.5 Integration 57

4 From the Local to the Global
   4.1 Introduction 62
   4.2 Poisson approximation of probabilities 62
   4.3 Average run length to false alarm 67

5 The Localization Theorem
   5.1 Introduction 74
   5.2 A simplifies version of the localization theorem 74
   5.3 The Localization Theorem 77
   5.4 A local limit theorem 81
   5.5 Edge effects 85

### II Applications

6 Copy Number Variations
   6.1 Introduction 90
Part I

Theory
Chapter 1

Introduction
CHAPTER 1. INTRODUCTION

1.1 Distribution of extremes in random fields

The aim of this book is to present a method for analyzing the tail distribution of extreme values in random fields. A random field can be thought of as collection of random variables \( \{X_t : t \in T\} \), indexed by a set of parameters \( T \). The index set \( T \) may be quite complex. However, due to the selection bias caused by us, in the applications that we will analyze in this book it will typically turn out that \( T \) is a “nice” subset of \( \mathbb{R}^d \), the \( d \)-dimensional space of real numbers.

In some statistical applications one is interested in probabilities such as:

\[ P(\sup_{t \in T} X_t \geq x) , \]

the probability that the maximum of the random field exceeds a threshold \( x \), for large values of \( x \). There are only a few special cases in which the problem of computing such probabilities has an exact solution. In all other cases one is forced to use numerical methods, such as simulations, or to apply asymptotic approximations in order to evaluate the probability. These notes concentrate on the application of the proposed method for producing asymptotic analytical expansions of the probability. Nonetheless, some elements in the method may, and have been, applied in order to simulate numerical evaluations more efficiently. An application that illustrates the usefulness of the method in the context of simulations is presented in the second part of the book.

The notes are devoted to the task of analyzing the tail probability of extremes. We ignore completely more fundamental issues of sample-path properties of the random field and questions of measurability of the random variable \( \sup_{t \in T} X_t \). Nor, to begin with, do we even deal with the issue of the consistency of the definitions in the random fields that we consider to prove existence. Although in the bulk of the applications that we will present it turns out that the parameter set \( T \) is either finite or countable. In such a case measurability of the supremum will follow readily. In other cases one may typically rely on separability of the random field in order to establish the measurability requirement. As for us, we just ignore the issue.

The analysis of the probability that the supremum of the field exceeds the threshold \( x \) in the situation when \( T \) is such that this probability is vanishingly small will occupy a central part of the discussion. This type of analysis, the analysis of vanishingly small probabilities, is frequently referred to as large deviations. Mostly, the statement of a theorem in large deviations establishes the exponential rate by which the probability converges to zero. This first order approximation of the probability will not be sufficient for our needs. The aim of our analysis will be to produce refined approximations, approximations that include polynomial terms and associated constants. These refined expansions open the door for the production of approximations to probability of events that involve the supremum of a random field in settings where probabilities do not converge to zero an for the computation of associated expectations.

In this book we will selfishly concentrate on a specific approach for dealing with the problem at hand and thus portray the false image that the method that we present is the best method, not to say the only method, for producing asymptotic approximation of the probability that a random field obtains an extremely high value. The special situation where the index \( T \) is a subset of the real line, in which case the random field is actually a random process, has a long history and many tools for solving. Some of the alternative methods
1.2 OUTLINE OF THE METHOD

of solution in this case will be presented briefly in the next chapter. Another notable special case with a very elegant theory is the situation where the random field is gaussian with a smooth covariance structure. A more general tool may be applied in the gaussian setting that involves a continuous parameter set $T$ for cases where the covariance function does not have derivatives. There are much fewer tools available in order to deal with the even more general case where the random field is not necessarily gaussian, including the cases where the index set is discrete or the sample paths are not smooth.

The toolbox of the probabilist is not completely empty when faced with these more general problems. However, the optional method are limited in number and none is very elegant. Admittedly, one may question the elegance of the method that we will advocate. Still, the method seems to work in many different settings and thus may claim the title of generality.

The method is defined through a series of steps in which the large deviation part of the probability is separated from the refinements that are due to global and local fluctuations. These well defined steps that are recommended in order to apply the tool may help to identify the different contributors to the probability and assist in the evaluation of the relative contribution of the various sources of variability.

The demonstration of methods for approximating the extreme tail of the distribution of the maxima of a random field is initiated in the next chapter. The current introductory chapter is devoted to mental preparation. In Section 1.2 we provide a bird’s view of the proposed method. In that section we outline the role and characteristics of the different steps. All the details are left out and postpone to the subsequent chapters.

Section 1.3 we will present the type of random fields in which one may have hope to be able to apply the method as it is presented in the current notes. Essentially, we are motivated by the analysis of gaussian random fields, yet the the method is marketed as a tool that works for non-gaussian fields as well. Nonetheless, the approximation is based to some extent on the application of the Central Limit Theorem. Consequently, the type of fields that we target are those that obey the central limit theorem in an appropriate sense. We discuss such fields in Section 1.3.

In Section 1.4 we give a list of applications. Random fields associated with two relatively simple applications will serve throughout the first part of the book as demonstration of classical methods for approximating the distribution of extremes, as well as the demonstration of our method. Other, more complicated, examples will appear in the second part of the book, the part devoted to applications. For these examples we use the method described in Section 1.2.

1.2 Outline of the method

The method we propose is motivated by statistical considerations. We are inspired to think of the parameter $t$ of the field as specifying a statistical model and consider $X_t$ as a statistic that summarizes the information regarding the parameter. In many of the applications that we will consider this indeed is the context in which the field emerges. In other applications, when this is not the case, we may still consider that point of view as a motivation and a guiding principle.
CHAPTER 1. INTRODUCTION

More specifically, we propose to consider the problem of finding the tail distribution of a random field in the context of statistical hypothesis testing. In statistical hypothesis testing competing models for the distribution of the observations are grouped into two sub-collections. One sub-collection is called the null hypothesis. It reflects the absence of a scientifically significant signal in the data. The other sub-collection is composed of alternative distributions which are a reflection of the presence of such signal. A statistical test is constructed with the role of determining which of the two hypotheses is more consistent with empirical observations.

Inspired by that point of view one may regard the random variable $\sup_{t \in T} X_t$ as the test statistic, with the test itself rejecting the null hypothesis in favor of the alternative if this test statistic is above a threshold $x$. The null hypothesis itself in this context is composed of a single distribution: the actual distribution of the data. Consequently, rejecting the null hypothesis is an error. The probability that we seek is the probability of making such error, which in statistical vocabulary is called the significance level of the test.

The alternative collection of models are associated with the set $T$. Each $t \in T$ specifies a model $P_t$ of the distribution of the data. This distribution may be the actual distribution that was considered, if the problem emerged in the considered statistical context, or it may involve some artificially constructed model that fits our needs. At the heart of the method is the proposal to translate the original problem of computing the probability of being above the threshold under the null hypothesis to a problem of computing expectations under alternative models. The vehicle that carries out this translation is the likelihood ratio identity.

The likelihood ratio identity employs likelihood ratios. A likelihood is the probability of the observed data under a given probabilistic model. If the distribution of the data is continuous then the likelihood refers to the probability density. A likelihood ratio is the ratio between two likelihoods. Here we consider likelihood ratios in which the denominator is the likelihood of the data under the current distribution (the null distribution) and numerator is the likelihood of the data under the alternative distribution $P_t$. We denote this likelihood ratio by $\exp \{ \ell_t \}$, with $\ell_t$ being the log-likelihood ratio. We relate each $X_t$ to $\ell_t$, whether or not $X_t$ emerged originally as a log-likelihood ratio, and rephrase the original problem of crossing the threshold by elements from the collection $\{X_t : t \in T\}$ using instead elements from the collection $\{\ell_t : t \in T\}$, possibly with a different threshold.

In the book we present a recipe for for the application of the method that involves the likelihood ratio identity. This recipe involves a series of steps, concluding in an approximation for the tail probability that we analyze. Unlike the baking of real cakes, one need not follow the proposed steps meticulously in order to avoid disasters. On the contrary, these steps are only guidelines and are not necessarily optimal in all scenarios. Still, we find them useful.

The first step involves the identification of the large deviation rate. The method itself produces refinements to the first order approximation that is produced by this rate. Frequently, one may find the large deviation rate by the maximization of the marginal probabilities $\{ P\left( \sup_{t \in T} X_t \geq x \right) : t \in T \}$ over the collection $T$.

A large deviation rate is associated with a collection of values in the parameter set. Preparation towards the application of the likelihood ratio identity may involve the identification of a subset of parameter values that are most likely to
contain the maximizing value and restrict the analysis only to that sub-region.

In the case where the parameter space is continuous one may consider another preparation step in which the maximization is restricted to a dense, but discrete, sub-collection. Although the method can, and has been, implemented directly to a continuous parameter set there are some technical advantages to its implementation in the context of a discrete set of parameters.

After preparations one may invoke the likelihood ratio identity. The outcome of this step is a presentation of the probability under the current distribution of the maximum of the field exceeding a threshold in terms of a sum of expectations. The sum extends over the different values of the parameters. Each element in the sum is computed in the context of the probability model specified by that value. The expectation involves a deterministic term that is associated with the large deviation rate and a product of two random terms, one measuring the global behavior of the field in the context of maximization and the other measuring local fluctuations.

The localization theorem, the subsequent step, applies a local limit theorem to the global term in order to prove the asymptotic independence between the effect of the global term and the effect of local fluctuations. In the examples that are considered the local limit theorem emerges as a refinement of the central limit theorem. Consequently, the given approach is more natural in problems where the global term obeys a central limit theorem and converges to the normal distribution. The outcome of this step is an approximation of each parameter-specific expectation in the representation by the product of three factors: a factor associated with the large deviation rate, a factor associated with the density of the normal limit of the global term, and a factor that measures local fluctuations. The integrated approximation of the tail probability is obtained by the summation of these products over the collection of parameter values.

The method is employed in settings consistent with large deviation formulation. Accordingly, the probability \( P(\sup_{t \in T} X_t \geq x) \) converges to 0 when the threshold \( x \) diverges to infinity. In other applications, when the parameter set \( T \) is increasing fast enough as a function of \( x \), the probability may be converging to a positive constant. Poisson-type of approximations may be applied in order to extend the approximation obtain in large deviation settings in order to cover such cases. Convergence in distribution that emerges from poisson approximations, in conjunction with statement of uniform integrability, my be used in order to approximate functionals associated with extremes that involve expectation.

## 1.3 Gaussian and asymptotically gaussian random fields

The random field is a collection of random variables \( \{X_t : t \in T\} \) with a joint distribution. The joint distribution is uniquely specified in terms of the finite-dimensional joint distribution of the random vector \( \{X_t : t = t_1, t_2, \ldots, t_k\} \), for any finite sub-collection of parameter values \( t_1, \ldots, t_k \in T \). In the special case where these finite-dimensional joint distributions are all gaussian we say that the field is a gaussian random field.

The joint distribution of a gaussian random vector, the multi-normal dis-
CHAPTER 1. INTRODUCTION

tribution, is a function only of the vector of expectations and the matrix of variances and covariances. As a conclusion we get that the distribution of a gaussian random field is fully specified in terms of the expectation function: the expectation \( E(X_t) \), for each \( t \in T \), and the variance-covariance function: the covariance \( \text{Cov}(X_t, X_s) \), for any pair \((t, s) \in T \times T\). The distribution of the maximum of the gaussian field is influenced both by the deterministic component of the field, namely the expectation, and by the variability, which is determined by the covariance function.

The theory that deals with the investigation of extreme values in gaussian fields is highly developed. The role of the expectation, and the more delicate role of the covariance structure, in the determination of the distributions of such extremes is well understood. Extremely accurate asymptotic approximations of the distribution of these extremes exist for some sub-families of gaussian random fields. Good asymptotic approximations exist for other sub-families.

In the next chapter we will present the two main tools for analyzing the distribution of extremes in gaussian fields. One tool is based on the computation of the expectation the of Euler characteristic of the excursion set. This tool is applicable when the realizations of the random field are differentiable and is extremely accurate. The other, more general approach, is known as the double-sum method. It is not as accurate as the first tool but it may apply in situations where the realizations of the field are not smooth.

The smoothness, or lack thereof, of the realizations of a gaussian random field is determined by the covariance function in general, and the smoothness of this function in the vicinity of the diagonal \( \{(t, t) : t \in T\} \) in particular. Consequently, the answer to a question regarding which of the tools can be used is related of the ability to take partial derivatives of the covariance function at the points of the diagonal. Basically, if second order partial derivatives exist then the more accurate method of the Euler characteristic may be applied. Otherwise, one should refer to the more general method.

Our method is more similar to the double-sum method in the sense that it may be applied under more general conditions, paying for the generality of the application in coins of accuracy of the approximation. It has a further advantage that it can be applied in settings where the random field is not gaussian, although it should obey a local limit theorem that emerges from the Central Limit Theorem.

The Central Limit Theorem deals with the convergence of sums of random element to a gaussian element. If, for example, the elements are independent fields then the resulting limit is a gaussian field. A Central Limit Theorem in the context of random field relies, typically, on the convergence of the finite-dimensional distribution of the field to a gaussian limit and on a tightness property. The role of tightness is to ensure that distribution of the field, along the process of convergence, may be approximated uniformly well using a finite and fixed collection of parameter points.

A tempting approach for dealing with the distribution of the extremes in a non-gaussian setting, in the case where the field in question belongs to a sequence that converges to a gaussian limit, is to apply the approximation of the distribution of the maximum to the gaussian limit distribution. In this approach one separates between the convergence of the field to the gaussian limit, which is carried out first, to the convergence of the tail distribution of the field to zero, which is assessed after the first convergence took place. Tempting
as it may be, this approach may produce misleading outcomes. The reason for this is that the Central Limit Theorem, as the name suggests, deals with the central part of the distribution, not with the extreme tail of the distribution. There may very well exist a big difference between the tail behavior of the original field and the tail behavior of a gaussian field with the same expectation and the same covariance function as the original field. However, this difference is washed away by the central limit theorem. A better approach is to deal directly with the distribution of extreme values of the original field itself and assess its asymptotic behavior.

In the approach outlined in the previous section the probability of the maximum of the field is presented as a sum of terms. These terms are composed of a deterministic factor that relates to the large deviation rate, a factor that relates to the contribution of the global term, and a factor that measures the contribution of local perturbations. Separating out the effect of large deviation guarantees an honest assessment of the extreme tail distribution. The Central Limit Theorem plays a part in the derivation of the contribution of the global term and, in some cases, in the contribution of the local fluctuations.

The part played by the Central Limit Theorem in the assessment of the contribution of the global term is in the form of a local limit theorem. A local limit theorem deals with the probability that a statistic, typical a statistic produced by taking a sum, obtains values in an interval of fixed width. The statistic has a variance that goes to infinity. Consequently, the probability of belonging to the interval goes to 0 at a rate proportional to the standard deviation of the statistic. An accurate assessment of the rate by which the probability converges to zero may be deduced from a Central Limit Theorem that involves a higher order expansion of the approximation error. A famous theorem of that sort is the Berry-Esseen theorem. An important point to make is that the distribution of the global term is assessed in the context of an appropriate alternative distribution, not the original null distribution. Consequently, convergence may hold for the selected alternative distribution even if it does not exist for the original null distribution.

The method relies on a statement regarding the joint limit distribution of the global term and a local field that is derived by local deviations of the field. The requirement is a local limit for the global term and asymptotic independence between the global term and local deviations. The local deviations are not required to converge to a gaussian limit. This requirement is much less than the requirement that the field converges to a gaussian limit. In the particular important case where the field does converge to a gaussian limit, for example when the field is a sum of independent fields, the factor in the approximation that is associated with the local fluctuation is the same as the factor that emerges for gaussian fields (still, the factor that is associated with large deviation may be different). In other cases, the factor that is associated with local fluctuations may differ from the factors encountered in the gaussian setting.

1.4 Applications

This book targets people with interest in statistics and probability as branches of applied mathematics. As such, it will not do an honest job if it will not demonstrate the applicability of the theory in “real life situations”. “Applica-
tion of mathematical theory” is, to some extent, an oxymoron. Typically, the applications are as abstract and removed from physical reality as the theory that is being applied. To justify the relation between an application and real life the mathematician tells a story that portrays the application as something that actual practitioners, people that do real science or put their’s (or other’s) money in real risk, care about. These stories can fool outsiders but not people who actually know the details. This is true also for the applications given in this book. I am not an insider in neither of the topics that will appear in the second part of the book. In some cases, for example in genetics, I can say that I know people who know actual practitioners. In other cases I have only my imagination and what I read in Wikipedia to guide me.

After my confession let us list the applications that will be presented in the book. In the first part of the book we will use two simple applications in order to demonstrate the solutions provided by the theory, both classical solutions that will be presented in Chapter 2 and the solution provided by our method. That solution will be mainly given in Chapter 3.

The first example is an example of a sequential testing of hypothesis. In this example we will consider the case of two simple hypotheses and the application of the sequential probability ratio test for testing one hypothesis versus the other. In this example the parameter space is linear – the positive integers – and the random field is the process of a random walk. Tools that are applicable to random walks, in particular random walks stopped by a stopping time, can be used for the analysis of this example.

The second example is an example of a scanning statistic. Scanning statistics are used in order to detect rare signals in an environment contaminated by random noise. In the particular example we will consider a signal of the form of a region of elevated expectation in a linear environment with a gaussian white noise. The elevated region is parameterized by the location of its center and by its width. The resulting field of scanning statistics is two-dimensional and gaussian. We will consider cases where the field is smooth and a case where it is not. The classical tools for the analysis of gaussian fields will be applied in Chapter 2 and our alternative approach will be used in Chapter 3.

The poisson approximation for the example of a scanning statistic and for a modified version of the sequential example is discussed in Chapter 4.

More complex applications appear in the second part of the book. The applications presented are a reflection of the projects that I was involved with in recent years and in which the method was used.

The first application involves scanning for DNA copy number variations. Most of the genetic material in somatic cells comes in two copies, one originating form the mother and the other from the father. Occasionally, a segment of the DNA may be missing or may have multiple copies resulting in a copy number different than 2. A scanning statistic for detecting such intervals using genetic measurements generated from a sample can be constructed. The resulting scanning problem is not unlike the simple example that is used in the first part of the book. However, the produced two-dimensional field is not gaussian, although it is asymptotically so. This problem is analyzed in Chapter 6.

The second example combines a scanning statistic with a sequential tool for change-point detection. This example involves a story in which an image is scanned for a signal of a specific structure. The specifications of the signal are unknown nor is the time in which it will appear, if at all. The goal is to identify
the emergence of the signal as fast as possible after it appeared, but not to do so before it did. The noise involved is gaussian but the statistics that are used, and hence the associated random field, are not. The investigation of the properties of an appropriate stopping time that may be used to detect the emergence of a signal is carried out in Chapter 7.

The third application is discussed in Chapter 8. The concern in that application is the design of a buffer that is large enough to store packets waiting to be transmitted in an outgoing communication line. These packets arrive from a large number of independent sources and the outgoing line is of a fixed bandwidth. A simple model will characterize an incoming source by the distribution of the duration of time that is active, the “ON” period, and the distribution of time that it is idle, the “OFF” period. The size of the required buffer can be associated with a level and the probability of a buffer overflow can be associated with the probability that a random field associated with the sum of the ON-OFF processes exceeds the level. Apart from the fact that the field is not normally distributed, but only asymptotically so, it is also the case that the characteristic behavior of local fluctuations in this case differs from the characterizations in the other examples. This local behavior is specified by a parameter we denote by $\alpha$ that may take value between 0 and 2 in general, and between 1 and 2 in the specific case. The local behavior is characterized by constants known as Pickands’ constants that depend on this parameter $\alpha$.

The Pickands’ constants emerged as part of the development of the classical double-sum method for analyzing gaussian fields and were defined a long time ago. However, their numerical evaluation remained an illusive problem and only crude upper and lower bounds to their value existed. In Chapter 9 we will explain how the representation that emerges from the method can be used in order to evaluate the constants efficiently.

The last application considers once more the problem of quickest detection of a change in on-line sequential monitoring setting, a problem that was discussed in Chapter 7. However, the story in this example is different. Here we envision a network of distributed sensors that transmit data to a monitoring center. In order to save energy and reduce the risk of being discovered by possible adversaries a sensor transmits its data only if there is significance evidence for a meaningful signal in its vicinity. When the data is combined in the monitoring center to form a monitoring statistic a random field emerges. However, unlike previous examples, the dimension of the field increases in proportion to the number of sensors. If the number of sensors is large an asymptotic derivation should either consider a gaussian filed with a dimension that goes to infinity (under the assumption that the source of noise is gaussian) or else consider a non-gaussian field. We take the second approach in order to produce an asymptotic approximation of the working characteristics of the monitoring procedure. This analysis is presented in Chapter 10.
Chapter 2

Basic Examples
2.1 Introduction

In this chapter we give an informal description of the theory and available tools for the approximation of the probability that a random field exceeds a high threshold. The theory is described in the context of two seemingly unrelated problems: a problem of sequential testing of hypothesis and a problem of scanning for a rare signal in a noisy gaussian environment. Different tools will be applied to the different examples, tools that exploit the unique characteristics of the problems. The characteristic for the first problem is the fact that the set of parameters the defines the field is linearly ordered. The important feature for the second example is the fact that the field is gaussian.

At the end of this chapter we list some other method that appeared in the literature and have relevance to the problem at hand.

In the next chapter we will apply a different method to both problems. The similarity between the two examples will be reviled as a consequence.

2.2 A power-one sequential test

Imagine a dynamic setting where one is allowed to observe a system until enough information has been gathered in order to reach a conclusion. The issue at hand is to construct an efficient decision rule that can serve as a guide in making the decision to stop the process of observing the system and taking an appropriate action. Such dynamic settings emerge for example in clinical trials, where the concern is to make a decision regarding the relative efficacy of a new medical treatment before releasing it to the market. Ethical consideration proposes a sequential trial design in which new subjects are not treated with the new procedure once enough evidence points to the new treatment being inferior to other remedies. On the other hand, if the new treatment turns out to be superior then marketing it to the general public should not be delayed. Another situation where sequential designs are useful is in industrial quality control. The quality of the manufactured products is monitored continuously. An alarm is set if there is accumulating evidence indicating a deterioration in the quality of the products. Quality concerns should be balanced against considerations of efficiency of production. Hence, one should make sure that the monitoring procedure does not produce false alarms too frequently.

In order to understand some of the probabilistic issues involved, let us consider a simplified version of the problem. Denote by $X_1, X_2, X_3, \ldots$ the sequence of observations that one may potentially make on the system. The statistical procedure is composed of a stopping time $N$, namely an integer-valued random variable with the property that, for each integer $n$, $\{N = n\}$ is a function of the first $n$ observations $X_1, \ldots, X_n$, and a decision rule that is based on these same observations. Assume that the observations are independent and identically distribution. The exact form of the density of an observation is either $H_0 : X \sim f$ or $H_1 : X \sim g$, for two distinct density functions $f$ and $g$ that share the same support. A-priori it is unknown which of the two densities is in effect and it is the role of the decision rule to make that determination on the basis of incoming data.

Considerations of statistical theory guides one to use likelihood ratios as statistics for testing hypothesis. Equivalently, one may use the log of the likeli-
2.2. A POWER-ONE SEQUENTIAL TEST

hood ratio. In the presence of $n$ observations this statistic becomes:

$$\ell_n = \sum_{i=1}^{n} \log \left\{ \frac{g(X_i)}{f(X_i)} \right\} .$$

The increments of the sum are independent and identically distributed. If follows from Jensen’s inequality that the expectation of an increment is negative if the actual density of an observation is $f$ and it is positive if the actual density is $g$. Large values of this statistic support the alternative hypothesis $H_1$ whereas negative values of the statistic are consistent with the null hypothesis $H_0$.

A classical procedure, the sequential probability ratio test, uses these statistics as a monitoring sequence and proceeds sampling as long as the sequence is between pre-specified upper and lower boundaries. Once a boundary is crossed the stopping time is activated and a decision made. If the upper boundary was the boundary that was crossed then the decision is to reject the null hypothesis and to accept the alternative. Otherwise, if the lower boundary was crossed first then the null hypothesis is not rejected.

A special form of the sequential probability ratio test is when the lower boundary is absent. For such a procedure, if the actual density is $g$ then it is a corollary of the Law of Large Numbers that the upper bound will be eventually crossed with certainty. Consequently, the probability of correctly rejecting the null hypothesis, the statistical power of the test, is equal to 1. Our concern in the context of this procedure will be to evaluate its significance level, i.e. the probability that it ever crosses the boundary when the density of the observations is from the null distribution $f$.

The power-1 sequential probability test is associated with an upper boundary $x$. Denote the stopping time of the test by $N_x$, where:

$$N_x = \inf \{n : \ell_n \geq x\},$$

and $N_x = \infty$ if no such $n$ exist. The significance level of the test is the probability $P(N_x < \infty)$ that $N_x$ is finite, where the probability is computed under the null distribution that is governed by $f$.

We can compute this probability by splitting the event among the disjoint events associated with stopping at particular times:

$$P(N_x < \infty) = \sum_{n=1}^{\infty} P(N = n).$$

Each of the elements in the sum can be computed by changing the density that governs the probability from the null density $f$ to the alternative density $g$:

$$P(N_x = n) = \int_{\{N_x = n\}} f(x_1) \cdots f(x_n) dx_1 \cdots dx_n$$

$$= \int_{\{N_x = n\}} e^{-\ell_n} g(x_1) \cdots g(x_n) dx_1 \cdots dx_n$$

$$= \mathbb{E}_g \left[ e^{-\ell_n} ; N_x = n \right].$$

Notice that the last expectation is computed under the distribution governed by $g$. This fact is denoted by the subscript $g$ attached to the expectation functional.
Notice also that we are exploiting the fact that $N_x$ is a stopping time. As a result of this fact the computation of the probability of the event $\{N_x = n\}$ embarks only the joint distribution of the first $n$ observations and none other. Finally, observe that we are using the notation $E[Y; A]$ to indicate that the expectation involves the product between the random variable $Y$ and the indicator of the event $A$.

The “trick” of computing a probability of an event formulated in some distribution by reformulating it as an expectation in the context of an alternative distribution is called the likelihood ratio identity. We will apply this trick over and over again throughout the book. In the current sequential setting one may observe that over the event $\{N_x = n\}$ we may write the log likelihood ratio in the form

$$\ell_n = \ell_{N_x}. \quad \text{(2.1)}$$

We are interested in the probability that is given in (2.1) for large values of $x$. From the representation we may see that this probability converges to zero in an exponential rate. This rate is multiplied by a term that is given in the form of an expectation. What can be said of this expectation when $x$ goes to infinity?

The classical method for analyzing the given expectation is to apply renewal theory. Renewal theory deals with the process of partial sums of independent and positive random variables. The random variables symbolize cycle lengths. At the end of each cycle the system is restored and a new cycle is initiated. The theory deals with events that occur in a typical cycle.

The setting that led to display (2.1) also involves partial sums of independent random variables. However, although the increments have a positive expectation under the alternative distribution, still they may obtain negative values and are thus not positive random variables. A remedy for this inconsistency with the assumption of renewal theory is to consider ladder variables as substitute for the original variables. The ladder times $\tau_j$ are defined via:

$$\tau_1 = \inf \{ n : \ell_n > 0 \} \quad \text{and} \quad \sum_{i=1}^{j} \tau_i = \inf \{ n > \sum_{i=1}^{j-1} \tau_i : \ell_n > \ell_{\sum_{i=1}^{j-1} \tau_i} \} \quad \text{for} \quad j = 2, 3, \ldots.$$ 

The first ladder height is $Y_1 = \ell_{\tau_1}$ and the subsequent ladder heights are $Y_j = \ell_{\sum_{i=1}^{j} \tau_i} - \ell_{\sum_{i=1}^{j-1} \tau_i}$, considered as the periods between cycle renewals. The random variables $Y_j$ are positive, independent, and identically distributed.

The excess over the boundary $\ell_{N_x} - x$ that appears in the expectation on the righthand side of (2.1) can be reformulated in terms of ladder heights. Indeed,
since crossing an upper boundary must correspond to an increase in the level of the partial sum and hence must coincide with a ladder height we get that: 
\[ \ell_N x = \sum_{i=1}^{J_x} Y_i - x, \text{ where } J_x = \inf\{j : \sum_{i=1}^{j} Y_i \geq x\}. \]
Thereby, we may express the expectation in (2.1) in terms of the overshoot of the strictly increasing process \( S_j = \sum_{i=1}^{j} Y_i \), immediately after crossing the threshold \( x \).

To conclude, the expectation of the overshoot satisfies:
\[ E_g\left[e^{-\left(\ell_N x - x\right)}\right] = E_g\left[e^{-\left(S_{J_x} - x\right)}\right], \]
where \( S_j \) is the process of partial sums of ladder variables.

The advantage of considering partial sums of positive random variable is the fact that they form a strictly increasing sequence. Consequently, if the process is below the threshold at time \( j \) then we are assured that the process did not cross the threshold before that time. As a result, we may write the event associated with crossing at a given time in terms of the partial sum immediately before that time and the independent increment that follows:
\[ \{J_x = j\} = \{S_{j-1} < x, S_j \geq x\} = \{S_{j-1} < x, Y_j \geq x - S_{j-1}\} . \]

(In order to be consistent with the notation when \( j = 1 \) we define \( S_0 = 0 \).)

Again, we compute an expectation by partitioning it among disjoint events. Currently, the partition is with respect of the different values of the stopping time \( J_x \). Moreover, we compute the value of the integral itself by conditioning on the value of the partial sum just before stopping and obtain:
\[ E_g\left[e^{-\left(S_{J_x} - x\right)}\right] = \sum_{j=1}^{\infty} \int_0^x E_g\left[e^{-\left(Y_j - (x-u)\right)}, Y_j \geq x - u\right] P(S_{j-1} = u) du . \]

Notice that we used independence in order to replace the conditional expectation with respect to \( Y_j \), given the value of \( S_{j-1} \), by an expectation with respect to the marginal distribution of that random variable. Our notation assumes that \( S_j \) has a density. As a matter of fact, the argument does not really make such requirement. Still, to keep notation simple, we will maintain this assumption.

Finally, after the change of variable \( v = x - u \) and interchanging between summation and integration we get:
\[ = \int_0^x E_g\left[e^{-\left(Y_1 - v\right)}, Y_1 \geq v\right] \left\{\sum_{j=1}^{\infty} P(S_j = x - v)\right\} dv \]
\[ = \int_0^x E_g\left[e^{-\left(Y_1 - v\right)}, Y_1 \geq v\right] h(x - v) dv , \]
where \( H(x) = 1 + \sum_{n=1}^{\infty} P(S_n \leq x) \) is the renewal measure, defined for \( x \geq 0 \), and \( h = H' \) is the density of this measure.

The analysis up until this point was relatively straightforward. Unfortunately, the subsequent analysis becomes much more involved and complex in details, although the principles are simple. We will not give a precise derivation of this section's example in the sequel. Instead, we will describe the principles and give an outline of the proof. Details can be found in \([14]\), Chapter VIII, and the references therein.

Basically, the remainder of the derivation is an application of the renewal theorem in order to obtain the limit, as \( x \to \infty \), of the integral in (2.2).
limit is given in terms of the distribution of the first ladder variable $Y_1$. A
relation between the distribution of a ladder variables and the the distribution
of partial sums of the original increments is used in order to produce a formula
for the limit in terms of the partial sums $\ell_n$.

The renewal theorem is a type of a local limit theorem, for large values of
$x$, of the renewal measure $H(x)$. Consider an interval of the form $[x - v, x]$ of
length $v$. The key observation that is formulated by the theorem is that for
large values of $x$ the initial value of the process will have only a diminishing
effect. Therefore, one cannot identify any particular region in the interval where
renewals are more likely to occur compared to other regions. The consequence
is that in the limit, as $x \to \infty$, the density is flat over the interval.

The theory distinguishes between the case where the values of $Y_1$ reside on
an arithmetic lattice, in which case the measure $H$ must also be concentrated
on this lattice, and the non-arithmetic situation where no such lattice exists.
We will be interested here only in the non-arithmetic case. The conclusion
of the theorem in the non-arithmetic case is that the density $h(x - v)$ can be
approximated by the constant $1/E(Y_1)$, for all values of $v$ such that $x - v$ is still
large.

Consider the integral in (2.2). It depends on $x$ through the upper limit
and through the evaluation of the renewal density at $x - v$, for $v \leq x$. Under
appropriate regularity conditions on the tail of $Y_1$ one may make sure that the
expectation within the integral, taken as a function of $v$, converges to zero fast
enough for large values of $v$. Consequently, the evaluation of the integral will
depend only on value of $v$ in a bounded interval. Replacing $h(x - v)$ by its
approximation and letting the upper limit of the integral go to infinity produces
the limit of (2.2):

$$\int_0^\infty E_g [e^{-Y_1 - v}, Y_1 \geq v] \frac{dv}{E_g(Y_1)} = \frac{1 - E_g(e^{-Y_1})}{E_g(Y_1)}.$$ 

The equality follows from presenting the expectation as an integral and changing
the order of the expectation integral with the outer integral.

Recall that the random variable $Y_1$ is the first ladder height. This random
variable is associated with the first ladder time $\tau_1 = \tau_+ = \inf \{n : \ell_n > 0\}$ via
the relation $Y_1 = \ell_\tau_1$. In other words, the variable corresponds to the random
walk of log-likelihood ratios, stopped by an appropriate stopping time. In order
to compute the expectation of a stopped random walk one may apply Wald’s
identity. This identity, which is straightforward to prove, states that the expectation
of a stopped random walk is equal to the product of the expectation of an
increment times the expectation of the stopping time (provided that both expec-
tations are finite). Specifically, in our case we get that $E_g(\ell_{\tau_1}) = E_g(\ell_1)E_g(\tau_+)$. Plugging this identity in the quantity we seek to evaluate gives:

$$\frac{1 - E_g(e^{-Y_1})}{E_g(Y_1)} = \frac{1 - E_g(e^{-\ell_{\tau_+}})}{E_g(\ell_{\tau_+})} = \frac{1 - P(\tau_+ < \infty)}{E_g(\ell_1)E_g(\tau_+)}.$$ 

The value of the numerator in the last equation follows the application of the
likelihood ratio identity.

In order to give an expression for the expectation of the stopping time $\tau_+$ one may use an alternative representation of its survival function in terms the
2.2. A POWER-ONE SEQUENTIAL TEST

Let the random walk obtain a sequential minimum at a given time:

\[ P_g(\tau_+ > n) = P_g\left(\max_{0 \leq k \leq n} \ell_k \leq 0\right) = P_g\left(\max_{0 \leq k \leq n} (\ell_n - \ell_k) \leq 0\right) = P_g(\ell_n = \min_{0 \leq k \leq n} \ell_k). \]

The validity of the second equation stems from the fact that the partial sums \( \ell_n - \ell_k \) constructed in the reverse order have the same joint distribution as the original partial sums. The expectation is produced by the summation of the survival probabilities. The sum yield:

\[ E_g(\tau_+) = \sum_{n=0}^{\infty} P_g(\tau_+ > n) = E_g(\sum_{n=0}^{\infty} 1(\ell_n = \min_{0 \leq k \leq n} \ell_k)). \]

The infinite sum is a count of the number of times that the random walk obtains a sequential minimum. The same count may be constructed by the aid of a new sequence of stopping times: the times between consecutive minima. For example, the first such time is:

\[ \tau_1 = \inf\{n \geq 1 : \ell_n \leq 0\} = \tau_{(1)}. \]

Subsequent times \( \tau_j = \tau_{(j)} \) are defined in a similar fashion, in parallel to the previous definition of the stopping times \( \tau_j = \tau_{(j)} \). The count in the infinite sum terminates once a stopping time \( \tau_{(j)} \) obtains an infinite value. The stopping times are independent and identically distributed. Consequently, the distribution of the total counts of sequential minima is geometric and therefore:

\[ E_g(\tau_+) = 1/P_g(\tau_- = \infty). \]

Adding this observation to the previous data leads to the representation:

\[ \frac{1 - E_g(\ell^+_{1})}{E_g(\ell_1)} = P_g(\tau_- = \infty)P(\tau_+ = \infty)/E_g(\ell_1). \]

The final expression is obtained via representations of both probabilities in the numerator. These presentations are corollary of Theorem A.2 (obtained by letting \( s \to 1 \)) and take the form:

\[ -\log P(\tau_+ = \infty) = \sum_{n=1}^{\infty} \frac{1}{n} P(\ell_n > 0) \quad \text{and} \quad -\log P_g(\tau_- = \infty) = \sum_{n=1}^{\infty} \frac{1}{n} P_g(\ell_n \leq 0). \]

Putting all things together we get the final expression for the limit:

\[ \lim_{x \to \infty} e^x P(N_x < \infty) = \frac{\exp\left(-\sum_{n=1}^{\infty} n^{-1}[P_g(\ell_n \leq 0) + P(\ell_n > 0)]\right)}{E_g(\ell_1)} \quad \text{(2.3)} \]

Consider a specific case of testing in the normal shift model. Assume that the null density \( f \) is the standard normal density \( \phi(x) = (2\pi)^{-1/2} \exp\left(-\frac{1}{2}x^2\right) \) and the alternative density is \( g(x) = \phi(x - \mu) \), for some given \( \mu > 0 \). The log-likelihood of \( n \) observations is \( \ell_n = \sum_{i=1}^{n} \{\mu X_i - \mu^2/2\} \). This random variable has the normal distribution with variance equal to \( n\mu^2 \), both under the null distribution and under the alternative. The expectation under the alternative is \( n\mu^2/2 \) and under the null it is the negative of the alternative expectation. From the symmetry and continuity of the normal distribution we get that the limit in this case is equal to:

\[ \lim_{x \to \infty} e^x P(N_x < \infty) = \frac{\exp\left(-2\sum_{n=1}^{\infty} n^{-1}\Phi(-n^{1/2} \mu/2)\right)}{\mu^2/2} = \nu(\mu), \]

where \( \Phi(c) \) is the cumulative distribution function of the standard normal distribution.
where $\Phi(x)$ is the cumulative distribution function of the standard normal distribution.

The overshoot-correction function $\nu(\mu)$ has a reasonably nice looking representation. Unfortunately, this representation is not too convenient for numerical evaluation, especially when the value of $\mu$ is small, since in this case the convergence of the infinite sum is very slow. However, one may show that the limit value of the function when $\mu \to 0$ (in which case the Brownian motion is used for testing, is 1. Moreover, a Taylor expansion of the function near the origin is available (see, for example, [4]). For numerical purposes it was recommended in [22] to use the ad-hock formula:

$$
\nu(\mu) \approx \frac{(2/\mu)(\Phi(\mu/2) - 0.5)}{(\mu/2)\Phi(\mu/2) + \phi(\mu/2)}.
$$

### 2.3 A kernel-based scanning statistic

The second example we consider involves searching for signals in a noisy environment. The development of modern computers allows for the collection and analysis of large quantities of data. This data contains a lot of useful information. Unfortunately, the vast majority of the data is irrelevant for the specific task one seeks to carry out. Thus, a primary concern is to sort out the precisely few needles of useful data that are hidden inside the pile of useless hey.

Again, computers can come to our aid. Characteristic of needles can be specified and an algorithm for the examination of the pile in order to efficiently identify objects with the given characteristics can be used for the task. Our goal in this section will not be the development of such algorithms. Instead, we will deal with general statistical characteristics of scanning that applies such algorithms.

Noisy environment corresponds to the situation where the hey in the pile has all types of shapes. Some of the shapes may share some of the characteristics of a useful needle, but in fact correspond to useless hey. As a result, a typical application of an algorithm will produce a collection of “hits” or detections. Some of the hits may involve real needles and some not. These are, respectively, true and false detections. An important statistical concern is to make sure that the rate of hits that are unrelated to actual needles, i.e. false detections, is not too high.

In order to be concrete let us look at a simple example. Assume that we are scanning a linear region. In general, the landscape is flat at a level we denote by $0$. However, every here and there there may be elevations of a specific shape. Assume that we can characterize the elevation in the functional norm: $\beta \cdot g((x - t)/h)/h^1$, where $x$ is the location in the region of the data point we examine, $t$ is the location of the center of the elevation, $h$ is a parameter that characterizes the width of the elevated region, and $\beta$ is a measure of the overall strength of the signal. The functional form of the function $g$ is known and fixed. However, the parameters $t$ and $h$ and $\beta$ are a-priori unknown.

The noise in the environment is model as a gaussian white noise $dB_x$. If the signal is present then the observed process is $dX_x = \beta \cdot g((x - t)/h) + dB_x$. If the signal is absent then the observed process is $dX_x = dB_x$, i.e. pure noise. The model of pure noise, the null hypothesis, corresponds to the case $H_0 : \beta = 0$. The alternative hypothesis that we would like to consider is $H_1 : \beta > 0$. 

For given parameter values \((t, h, \beta)\) we get, since the signal corresponds to a shift in the expectation in a gaussian process, that the log-likelihood ratio statistic for testing these values against the null hypothesis of pure noise is:

\[
\ell_{t, h, \beta} = \int \frac{g((x - t)/h)}{h^2} dB_x - \frac{\beta^2}{2h} \int [g((x - t)/h)]^2 dx.
\]

Statistical theory suggests to use the score statistic in order to test for weak signals. The score statistic is the derivative of the log-likelihood statistic, evaluated at the null value of the parameter. Specifically, if we consider the score with respect to \(\beta\) we get:

\[
\frac{\partial}{\partial \beta} \ell_{t, h, \beta} \bigg|_{\beta = 0} = \int \frac{g((x - t)/h)}{h^2} dB_x.
\]

The expectation of this score statistic under the null distribution is 0 and the variance of the statistic under this distribution is \(\int [g((x - t)/h)]^2 h^{-1} dx\). Consequently, the standardized score statistic, for a given value of the parameter \(\theta = (t, h)\), is:

\[
Z_{\theta} = \frac{\int g((x - t)/h)/h^2 dB_x}{\left\{ \int [g((x - t)/h)]^2 h^{-1} dx \right\}^{1/2}}.
\]

The parameter \(\theta = (t, h)\) specifies the structural characteristics of a signal. Large values of the standardized score statistic for a given parameter value indicate consistency between data and the presence of a signal in the form that is characterized by that value of the parameter. Hence, considering the score statistic as a function of the parameter, one’s attention should be given to parameter values that produce high levels of the score statistic. A natural criteria is to set a threshold \(z\) and to single out all parameter values that produce statistic values at least as high.

Consider a given searching region \(T\), i.e. a given subset of parameter values, and a threshold \(z\). Typically, since true signals are rare, this region will not contain a signal. However, random fluctuation may result in relatively high levels of the score statistic and produce a false detection if they cross the threshold \(z\). Consequently, in order to avoid a high rate of false detections one should set the threshold \(z\) high enough to be above most of the null fluctuations. In that context, one should be interested in the relation between the threshold \(z\) and the probability \(P(\sup_{\theta \in T} Z_{\theta} \geq z)\), hoping to choose \(z\) such that this probability is small enough. This puts us again in a situation where the distribution of extreme values of a random field, a random field over a two-dimensional parameter space in this case, is relevant.

There are some important differences between the current random field and the one that appeared in the example of the previous section. The primary difference is that in the previous case the parameter space was one-dimensional. The one-dimensional euclidian space is ordered. The tools of sequential analysis that were heavily exploited in the analysis of the previous example are not natural in spaces with no build-in ordering such as the two-dimensional parameter space of the current example. On the other hand, the current example involves a continuum of parameters whereas the parameter space of the previous example was discrete. Thereby, there is hope to be able to use in the analysis tools of differential and integral calculus of the field as a function of the parameters, tools that could not be used in the previous case.
There is a third difference that has technical importance. The current random field is a gaussian random field, since the joint distribution of any finite collection of elements of the field is multivariate normal. The gaussian distribution has unique properties that that one can take advantage of in the development of asymptotic expansion of the tail distribution of extreme values of the random field.

In this section we will introduce two techniques that were developed in order to deal with maxima in a gaussian random field. One of the techniques investigates, using tools of differential and integral geometry, the geometrical characteristics of elevated values of the random field $Z_\theta$ in order to produce a very accurate approximation of the probability in question. However, tools of differential and integral geometric assume the smoothness of the realization of the random field as a function of the parameters. Consequently, the price one pays for the accuracy in assessing the probability is cashed in the form of a sever limitation on the range of applications on which the tool may be used. Specifically, it imposes smoothness assumptions on the realizations of the random field.

The other technique uses a more standard approach of dividing the search region into smaller subsets, assessing the probability of crossing in each such subset, and integrating the local approximations to produce a global approximation. This approach can be applied in more general gaussian field since it does not require smoothness. Albeit, the accuracy of the resulting approximation is much less than the geometrical approach. This other techniques is nicknamed the double-sum method due to the method that is used for establishing the validity of the integration of the local approximations. The technique that is being advocated in the current notes share many similar features with the double-sum method. However, the two are not the same.

The geometrical method examines the excursion set, the collection of parameter values over which the random field obtains a value at least as large as the threshold:

$$A_z = \{ \theta \in T : Z_\theta \geq z \}.$$

This set is a random set. It is not empty if, and only if, the maximal value of the field is larger than the threshold $z$:

$$P\left( \sup_{\theta \in T} Z_\theta \geq z \right) = P(A_z \neq \emptyset).$$

Instead of investigating directly the content of the set $A_z$, which may be quite complex, the idea is to use a computable characteristic of the set as a substitute. Such characteristic is the Euler characteristic. Essentially, the Euler characteristic of the set is an integer value count of the number of subsets of a simple topological structure that makeup the excursion set $A_z$, discounting double counts in the appropriate way. The approximation is based on the expectation of the given functional:

$$P\left( \sup_{\theta \in T} Z_\theta \geq z \right) \approx E(\varphi(A_z)),$$

where $\varphi(A_z)$ is the Euler characteristic of the excursion set.

The mathematics behind the Euler characteristic method have two major components. The first component is to establish the validity of the approximation, namely to identify conditions that will assure that both the expectation
of the Euler characteristic and the tail probability of the field share similar numerical values. The second component is to compute the expected Euler characteristic.

A full account of the mathematical theory behind the approach that is based on the Euler characteristic of the excursion set can be found in the book by Adler and Taylor [1]. This book covers also the required background material, 120 or so pages on the properties of gaussian fields and another 120 pages on basic and more advanced material in integral geometry that is needed for the proofs. An attempt on my side to give details of the proofs, or even to give a full description of the statement of the theorems, will be a disservice. If you are interested in learning the subject, and this subject is indeed worth learning, go read the book!

Instead of a real description of the approach we will give a superficial description of the two main theorems of the book and relate them to the discussion in the current notes.

The first main theorem gives conditions that assure the validity of the approximation of the excursion probability by the expected Euler characteristic. This theorem, Theorem 14.3.3, states that for smooth gaussian process with variance equal to 1, where the gaussian process is defined over a nice manifold $T$, one has that:

$$\left| \mathbb{P}\left( \sup_{\theta \in T} Z_\theta \geq z \right) - \mathbb{E}\left( \varphi(A_z) \right) \right| < O(e^{-\alpha^2 z^2}) ,$$

for some $\alpha > 1$. The gaussian field is required to obey some extra conditions.

This theorem makes true the praises that were expressed regarding the accuracy of the Euler characteristic approach. The leading term in the tail of the distribution of extreme values in random field is the marginal probability of exceeding the threshold $\mathbb{P}(Z_\theta \geq z)$. When the variance is equal to 1 this probability is the survival function of the standard normal distribution, $1 - \Phi(z)$, which is shares the same exponential term $e^{-\frac{1}{2}z^2}$ as the standard normal density. The actual probability is modified by a multiplicative term that is polynomial in $z$. The theorem establishes that the accuracy of the approximation of the probability by the expected Euler characteristic is super-exponential. As a corollary we get that the expected Euler characteristic captures the entire polynomial component correctly. In contrast, alternative approximations only get at the leading components of the polynomial.

The second theorem gives a computable expression for the Euler characteristic of the excursion set. The expression is given in terms of the Hermite polynomials:

$$H_n(x) = n! \sum_{j=0}^{[n/2]} \frac{(-1)^j x^{n-2j}}{j!(n-2j)!2^j} , \quad n \geq 0 , \quad x \in \mathbb{R} ,$$

and $H_{-1}(x) = \sqrt{2\pi}\Phi(x)e^{x^2/2}$, and in terms of a volume integral conducted on the manifold $T$ and its boundaries, considered themselves as manifolds of lower dimensions:

**Theorem 2.1** (Theorems 12.4.1 and 12.4.2 of [1]). Let $Z_\theta$ be a centered, unit-variance Gaussian field on a $d$-dimensional, $C^2$ manifold $T$, and satisfying ap-
propriate regularity conditions. Then

\[ E(\varphi(A_z)) = e^{-\frac{1}{2}z^2} \sum_{j=0}^{d} (2\pi)^{-(j+1)/2} L_j(T) H_{j-1}(z), \]

where \( L_j(T) \) are the Lipschitz-Killing curvatures of \( T \), calculated with respect to the metric induced by the field.

A notable property of the representation given in the theorem is the separation between the contribution of the marginal normal distribution to the polynomial part in the form of the Hermite polynomials and the spatial contribution of the field in the form of the Lipschitz-Killing curvatures \( L_j(T) \). These curvatures involve both the distribution of the random field and the geometrical properties of the manifold. The contribution of the distribution of the random fields is via the covariance matrix of the gradient field \( \dot{Z}_\theta \), where the gradient is taken with respect to the parameter \( \theta \). The determinant of the covariance matrix \( \Sigma_\theta = \text{Var}(\dot{Z}_\theta) \) enters as a density that defines the measure used in the calculation of the curvatures.

The nice formulation of the theorem disguises the fact that the curvatures \( L_j(T) \) are rather difficult to compute. If the random field is stationary, in which case the density \( |\Sigma_\theta| \) is constant, then one can produce an explicit formula for the simple rectangular manifold that is used in the example that we consider in this section. Unfortunately, the process in the example is not stationary so the computation of the Euler characteristic is more involved. On the other hand, if one gives up the attempt to compute the entire expression and is content with the computation of the highest order term only, the term the involves the highest power of \( z \), then a simpler representation emerges.

The highest power appears in the Hermite polynomial with the highest index. In the two-dimensional example that we consider \( j = d = 2 \), which corresponds to \( H_{2-1}(z) = H_1(z) = z \). The coefficient that is associated with this polynomial is \( L_j(T) = \int_T |\Sigma_\theta|^{\frac{1}{2}} d\theta \). The resulting approximation becomes

\[ P(\sup_{\theta \in T} Z_\theta \geq z) \sim ze^{-\frac{1}{2}z^2} (2\pi)^{-3/2} \int_T |\Sigma_\theta|^{\frac{1}{2}} d\theta. \tag{2.4} \]

Of course, with approximation (2.4) one may no longer claim a supper-exponential accuracy.

Return to the example of a kernel-based scanning statistic that resulted in a centered and standardized gaussian random field. The distribution of a gaussian random field is fully characterized by the expectation and the covariance functions. The expectation function in this case is the constant at level 0. The covariance between two points \( \theta = (t, h) \) and \( \vartheta = (s, w) \) is given by:

\[ \text{Cov}(Z_\theta, Z_\vartheta) = \frac{\int g((x - t)/h)g((x - s)/w) \frac{dx}{\sqrt{hw}}}{\left\{ \int [g((x - t)/h)]^2 \frac{dx}{h} \right\}^{1/2} \left\{ \int [g((x - s)/w)]^2 \frac{dx}{w} \right\}^{1/2}} = \langle g_\theta, g_\vartheta \rangle \| g_\theta \| \| g_\vartheta \|, \]

where the inner product correspond to the inner product of the Hilbert space of square-integrable functions \( \langle g, f \rangle = \int g(x)f(x)dx \) and the norm in the denominator is the norm associated with this inner product. The notation for a signal that is associated with the parameter \( \theta \) is: \( g_\theta(x) = g((x - t)/h)/h^{1/2} \).
The validity of approximation (2.4) relies on the smoothness of the random field. Smoothness of a gaussian random is tightly linked to the smoothness of the covariance function. Specifically, we would like to explore the issue of smoothness in the context of our current example that involves integration of the function $g_\theta$ with respect to the gaussian white noise. In order to simplify the discussion of smoothness, let us assume that the linear region of data points is unbounded and is composed of the entire real line. In such a case the norm of $g_\theta$ is constant as a function of $\theta$ and can be denoted $\|g\|$. The result is yet a simpler formula for the covariance function of the random field:

\[ \text{Cov}(Z_\theta, Z_\theta) = \langle g_\theta, g_\theta \rangle / \|g\|^2. \quad (2.5) \]

The smoothness of the covariance function in (2.5) is related to the smoothness of the function $g$. If the function is smooth with integrable derivatives, $g(x) = \exp\{-x^2/2\}$ is an example, then one can express the derivatives of the covariance function in terms of the derivatives of the function $g$:

\[
\frac{\partial}{\partial \theta} \text{Cov}(Z_\theta, Z_\theta) = \langle g_\theta, \dot{g}_\theta \rangle / \|g\|^2, \quad \left(\frac{\partial}{\partial \theta}\right)^2 \text{Cov}(Z_\theta, Z_\theta) = \langle g_\theta, \ddot{g}_\theta \rangle / \|g\|^2, \n\]

etc., where we use the dot notation to express partial derivatives of the function $g_\theta(x)$ with respect to $\theta$.

Let us now look more closely at the relation between the pathwise derivatives of the random field and the derivatives of the covariance function. An element of the field has the representation:

\[ Z_\theta = \int g_\theta(x) dB_x / \|g\|. \]

The gradient with respect to $\theta$ of this element, for a smooth function $g$, is:

\[ \dot{Z}_\theta = \int \dot{g}_\theta(x) dB_x / \|g\|, \]

and it is defined for each $\theta$ in the interior of $T$. Likewise, the hessian is $\ddot{Z}_\theta = \int \ddot{g}_\theta(x) dB_x / \|g\|$. Similarly, one can take higher still derivatives of the random field, pending on the smoothness of $g$. Notice that all these derivatives are integrals of the gaussian white noise, hence gaussian. Moreover, the joint distribution of the original fields and its derivatives is multivariate normal.

Clearly, the expectation of the gradient vector is the zero vector. The variance-covariance matrix of the gradient is given by:

\[ \Sigma_\theta = \text{Var}(\dot{Z}_\theta) = \int [\dot{g}_\theta(x)] [\dot{g}_\theta(x)]' dx / \|g\|^2 = \langle \dot{g}_\theta, \dot{g}_\theta \rangle / \|g\|^2, \]

where we interpret the expression $\langle \dot{g}_\theta, \dot{g}_\theta \rangle$ as the $d \times d$ matrix with components $\langle \dot{g}_\theta, \dot{g}_\theta \rangle_{ij} = \langle \dot{g}_\theta \rangle_i \langle \dot{g}_\theta \rangle_j$.

In particular, if $g(x) = \exp\{-x^2/2\}$ then $\|g\|^2 = \sqrt{\pi}/2$ and:

\[ \dot{g}_\theta(x) = \left( \frac{h^{-\frac{3}{2}}(x-t)}{h^{-\frac{3}{2}}(x-t) - \frac{1}{2}h^{-\frac{3}{2}}} \right) e^{-\frac{1}{3}h^2(x-t)^2} \]
Consequently, since:

$$\langle \hat{g}_\theta, \hat{g}_\theta \rangle_{11} = \int h^{-5}(x-t)^2 e^{-\frac{1}{h^2}(x-t)^2} dx = \frac{1}{h^3} \sqrt{\frac{\pi}{2}}$$

$$\langle \hat{g}_\theta, \hat{g}_\theta \rangle_{12} = \int h^{-6}(x-t)^2 e^{-\frac{1}{h^2}(x-t)^2} dx - \frac{1}{2} \int h^{-4}(x-t)e^{-\frac{1}{h^2}(x-t)^2} dx = \frac{1}{h^4} \sqrt{\frac{\pi}{2}}$$

$$\langle \hat{g}_\theta, \hat{g}_\theta \rangle_{22} = \int h^{-7}(x-t)^2 e^{-\frac{1}{h^2}(x-t)^2} dx + \frac{1}{4} \int h^{-3} e^{-\frac{1}{h^2}(x-t)^2} dx = \left( \frac{1}{h^5} + \frac{1}{4h} \right) \sqrt{\frac{\pi}{2}}$$

we get that

$$\Sigma_\theta = \begin{pmatrix} h^{-3} & h^{-4} \\ h^{-4} & h^{-5} + h^{-1/4} \end{pmatrix} \implies |\Sigma_\theta|^\frac{1}{2} = \frac{1}{2} h^{-2}.$$ 

In the particular case where the parameter set is the rectangle \( T = [t_0, t_1] \times [h_0, h_1] \) then approximation (2.4) specifies to:

$$P\left( \sup_{\theta \in T} Z_\theta \geq z \right) \sim z e^{-\frac{1}{2} z^2} (2\pi)^{-3/2} \cdot 0.5(t_1 - t_0)(1/h_0 - 1/h_1). \quad (2.6)$$

The entire argument falls apart if the kernel \( g \) is not a continuous function. For example, let \( g(x) \) be the indicator of the interval \([-0.5, 0.5]\). Now \( \|g\| = 1 \) and

$$\text{Cov}(Z_\theta, Z_\phi) = \langle \hat{g}_\theta, \hat{g}_\phi \rangle = \min\{t + \frac{h}{2}, s + \frac{w}{2}\} - \max\{t - \frac{h}{2}, s - \frac{w}{2}\} \sqrt{\frac{1}{haw}}$$

when the numerator is positive and \( \text{Cov}(Z_\theta, Z_\phi) = 0 \) otherwise. This function is equal to 1 when \( \theta = \phi \) and it is continuous at this point, but it does not have derivatives. Consequently, the realization of the random does not have a gradient, although it can be selected to be continuous with probability 1, and approximation (2.4) may not be applied.

An alternative method for obtaining an approximation is the double-sum method. This is a general-propose method that can be used for gaussian fields with a wide variety of covariance structures. We will use this method for the analysis of the last example involving a scanning statistic that uses a noncontinuous kernel. The same method may be used in order to obtain the first-order approximation that was specified for the case of a smooth kernel.

A good source for learning about the double-sum method and its applications is [11]. Our description of the method will diverge slightly from the more general discussion that is presented there since our main aim is to emphasis the similarities and differences between the double-sum method and our method.

The basic argument in the double-sum method involves a partitioning of the parameter space into smaller regions. In each of the regions the field may or may not cross the threshold. The probability that the global maxima of the field is above the threshold, i.e. the probability of the union of events of crossing in a subregion, is bounded from above by the sum of probabilities of crossing in a subregion and bounded from below by the same sum, minus the double sub of crossing simultaneously in a pair of distinct subregions. Gaussian arguments are used in order to obtain an approximation of the probability of crossing in a subregion and thus an approximation of the sum. On the other, it is shown
that the double-sum is of a smaller order compared to the sum, establishing the
sum as the approximation.

In the given example it is convenient to change the parametrization for
the definition of subregions. The scanning statistic in the current situation
corresponds to integration of the data process over an interval of width $h$ that
is centered at $t$. Occasionally, we would like to characterize such intervals by
their endpoints $\theta_1 = t - h/2$ and $\theta_2 = t + h/2$.

Fix $\theta = (\theta_1, \theta_2)$, the bottom-left corner of the subregion. The subregion
involves small perturbations in the “positive” direction of the endpoints about
the given parameter value:

$$ T_\theta = \{ \vartheta = (\vartheta_1, \vartheta_2) : 0 \leq \vartheta_1 - \theta_1 \leq \tau/z^2, 0 \leq \vartheta_2 - \theta_2 \leq \tau/z^2 \} . $$

In the proof we fix $\tau$ and let $z \to \infty$, resulting in a shrinkage of the size of the
subregion and convergence to a limit. Subsequently, we increase $\tau$ in order to
obtain the final approximation.

The critical computation corresponds to the probability of crossing the thresh-
hold $z$ within the local region. We carry out this computation by conditioning
on the value of $Z_\theta$, which has a standard normal distribution:

$$ P(\max_{\vartheta \in T_\theta} Z_\theta \geq z) = E\left[ P(\max_{\vartheta \in T_\theta} Z_\theta \geq z | Z_\theta) \right] $$

$$ = \frac{1}{\sqrt{2\pi}} \int e^{-\frac{1}{2}y^2} P(\max_{\vartheta \in T_\theta} Z_\theta \geq z | Z_\theta = y)dy . $$

Changing the variable to $x = -z(y-z)$ produces:

$$ = \frac{e^{-\frac{1}{2}z^2}}{\sqrt{2\pi}z} \int e^{x-\frac{1}{2}(\frac{x}{y})^2} P(\max_{\vartheta \in T_\theta} Z_\theta \geq z | Z_\theta = z - x/z)dx $$

$$ = \frac{e^{-\frac{1}{2}z^2}}{\sqrt{2\pi}z} \int e^{x-\frac{1}{2}(\frac{x}{y})^2} P(\max_{\vartheta \in T_\theta} z(\theta - Z_\theta) \geq x | Z_\theta = z - x/z)dx . $$

This is a good place to pause for a while and examine the last expression.
It involves the term $\phi(z)/z$, which is asymptotic to the survival function of
the standard normal distribution, and an integral. The first term reflects the
marginal distribution of the field elements. The integral corresponds to the
contribution of the local field about $\theta$. Elements of the local field are $z(\theta_2 - Z_\theta)$,
considered in the conditional distribution given $Z_\theta = z - x/z$.

The local field, as $z \to \infty$, is subject to two negating influences that balance
each other. On the one hand, shrinkage of the local region is shrinking the
increments $Z_\theta - Z_\theta$. On the other hand, a multiplication of these increments by
$z$ is counterbalancing the effect of shrinkage. The rate of shrinkage is selected
to assure convergence of the local field to a limit random field. In the limit, the
term that multiplies the approximation of the survival function is an appropriate
functional of the limit of the local field.

Concrete computations may help clarify the reasons for selecting the local
region. In the current parametrization we may write the covariance between
two elements of the random field in the form:

$$ \text{Cov}(Z_\theta, Z_\theta) = \frac{\theta_2/h - \theta_1/h}{\sqrt{1 + (\theta_2 - \theta_2)/h - (\theta_1 - \theta_1)/h}} $$

$$ = \frac{1 - (\theta_1 - \theta_1)/h}{\sqrt{1 + (\theta_2 - \theta_2)/h - (\theta_1 - \theta_1)/h}} . $$
When \( \vartheta = \theta \) the covariance is equal to 1. For other \( \vartheta \) in the subregion the covariance may be approximated by:

\[
1 - \operatorname{Cov}(Z_\vartheta, Z_\vartheta) \approx \frac{1}{2h} (\vartheta_1 - \vartheta_1) + \frac{1}{2h} (\vartheta_2 - \vartheta_2) = \frac{u_1}{2hz^2} + \frac{u_2}{2hz^2},
\]

where \( u_i = z^2(\vartheta_i - \vartheta_i), \ i = 1, 2 \). In general, for values of \( \vartheta \) in the vicinity of \( \vartheta \), not necessary in the first quadrant with respect to it, we will find that \( 1 - \operatorname{Cov}(Z_\vartheta, Z_\vartheta) \) is asymptotic to \([|u_1| + |u_2|]/[2hz^2]\).

The local random field \( z(Z_\vartheta - Z_\vartheta) \) in the conditional distribution is yet again a gaussian random field. As such, it is characterized by the expectation of elements and by the covariance structure. The conditional expectation is:

\[
E(z(Z_\vartheta - Z_\vartheta)|Z_\vartheta = z - x/z) = -zZ_\vartheta(1 - \operatorname{Cov}(Z_\vartheta, Z_\vartheta)) \approx -\frac{1}{2h}(u_1 + u_2).
\]

The conditional covariance does not depend on the specific value of \( Z_\vartheta \). One may obtain the conditional covariance between \( z(Z_\vartheta - Z_\vartheta) \) and \( z(Z_\eta - Z_\vartheta) \) by the examination of the conditional variance of the difference \( z(Z_\vartheta - Z_\eta) \):

\[
\operatorname{Var}(z(Z_\vartheta - Z_\eta)|Z_\vartheta) = z^2 \{ 2[1 - \operatorname{Cov}(Z_\vartheta, Z_\eta)] - [\operatorname{Cov}(Z_\vartheta, Z_\vartheta) - \operatorname{Cov}(Z_\eta, Z_\vartheta)]^2 \}.
\]

The squared term inside the curly brackets is of order \( z^{-2} \) and will make a vanishingly small contribution to the covariance. Setting \( \eta = \theta \) we get the variance of an increment:

\[
\operatorname{Var}(z(Z_\vartheta - Z_\vartheta)|Z_\vartheta) \approx 2z^2(1 - \operatorname{Cov}(Z_\vartheta, Z_\vartheta)) \approx \frac{1}{h} (u_1 + u_2).
\]

A similar expression is obtained for the conditional variance of \( z(Z_\eta - Z_\vartheta) \), with \( u_i \) replaced by \( v_i = z^2(\eta_i - \vartheta_i), \ i = 1, 2 \). Combining these facts with the general expression for \( 1 - \operatorname{Cov}(Z_\vartheta, Z_\eta) \) will give:

\[
\operatorname{Cov}(z(Z_\vartheta - Z_\vartheta), z(Z_\eta - Z_\vartheta)|Z_\vartheta) \approx \frac{1}{h} (u_1 \wedge u_1 + u_2 \wedge v_2).
\]

The covariance structure of a Brownian motion is given by \( \operatorname{Cov}(B_u, B_v) = u \wedge v \). We conclude that for each fixed \( x \), and when \( z \to \infty \), the expectation and the covariance structure of the local field converge to a limit expectation and covariance structures. The limit expectation and covariance structures, parameterized by the pair \((u_1, u_2)\), is equal to that of a sum of independent Brownian motions with negative drifts:

\[
\{ B_{u_i/h}^{(1)} \} + \{ B_{u_2/h}^{(2)} \} = \ell_i^{(1)} + \ell_i^{(2)}.
\]

Storing the information given above, let us return to the task of evaluating the probability that the field crosses the threshold within the local region. As part of this evaluation we need to assess the convergence of the integral:

\[
\int e^{-\frac{1}{2}z^2} P\big( \max_{\vartheta \in T_x} z(Z_\vartheta - Z_\vartheta) \geq x | Z_\vartheta = z - x/z \big) dx.
\]

For each fix \( x \), as \( z \to \infty \), the exponent converges to \( e^x \) and the probability converges to the probability that the maximum of the limit field is above
2.3. A KERNEL-BASED SCANNING STATISTIC

Consequently, provided that uniform integrability can be established, the integral converges to:

\[ \int e^x P \left( \max_{0 \leq u_1 \leq \tau} \ell^{(1)}_{u_1/h} + \max_{0 \leq u_2 \leq \tau} \ell^{(2)}_{u_2/h} \geq x \right) dx = \{ \mathcal{H}(\tau/h) \}^2, \]

for

\[ \mathcal{H}(t) = E \left( \max_{0 \leq u \leq t} e^{\ell_u} \right), \]

which is defined with respect to \( \ell \), a Brwonian motion with an appropriate negative drift.

The uniform integrability requires a uniform bound on the tail of the maximum of a random field. Such a bound is provided for gaussian random fields by Borel’s inequality. This inequality states that the tail of the maxima is no more than twice the survival function of the normal distribution with expectation equal to the expectation of the maximum of the random field and with variance equal to the maximal variance of components of the field. Entropy considerations can be used to bound the expectation of the maximum and finish the proof that:

\[ \sqrt{2\pi} e^{\frac{1}{2}z^2} P \left( \max_{\theta \in T_0} Z_\theta \geq z \right) \rightarrow z \rightarrow \infty \left\{ \mathcal{H}(\tau/h) \right\}^2. \] (2.7)

In order to obtain an approximation for the distribution of the global maximum over the entire region \( T = [t_0, t_1] \times [h_0, h_1] \) one splits the set into practically disjoint subsets \( T_{ij} \). The linear transformation \( A \) maps the parametrization \((t, h)\) to the new parametrization \((\theta_1, \theta_2)\), where

\[ A = \begin{pmatrix} 1 & -0.5 \\ 1 & 0.5 \end{pmatrix}. \]

In particular, \( |A| = 1 \), so this transformation preserves measure.

Consider the regular grid with span \( \tau/z^2 \) over the \((\theta_1, \theta_2)\) plane. The application of the inverse transformation \( A^{-1} \) to this grid splits the \((t, h)\) plane into rhombuses. Let \( T_{ij} \) be any of the rhombuses that cover \( T \), identified by its bottom corner. There is a total of \((t_2 - t_1)(h_2 - h_1)z^4/\tau^2\) such rhombuses of area \( \tau^2/z^4 \) each. When we apply the asymptotic approximation to the crossing probabilities in subregions we get the asymptotic upper bound:

\[ P \left( \max_{\theta \in T} Z_\theta \geq z \right) = P\left( \bigcup_{i,j} \left\{ \max_{\theta \in T_{ij}} Z_\theta \geq z \right\} \right) \leq \sum_{i,j} P\left( \max_{A\theta \in AT_{ij}} Z_{A\theta} \geq z \right) \sim \frac{z^3 e^{-\frac{1}{2}z^2}}{\sqrt{2\pi}} (t_2 - t_1) \int_{h_0}^{h_1} \frac{1}{h^2} \left\{ \frac{\mathcal{H}(\tau/h)}{\tau/h} \right\}^2 dh. \] (2.8)

The derivation of the asymptotic upper bound, which turn out to be the approximation itself, is completed by letting \( \tau \rightarrow \infty \). Towards that end a likelihood ratio identity can again be applied. Indeed, use the representation of \( \mathcal{H}(t) \) in terms of an integral of an exponent times a survival function:

\[ \mathcal{H}(t) = E \left( \max_{0 \leq u \leq t} e^{\ell_u} \right) = \int_0^\infty e^x P \left( \max_{0 \leq u \leq t} \ell_u \geq x \right) dx + 1. \]
Notice that the integral for negative values of $x$ is equal to 1 since the probability that the maximum is non-negative is equal to 1.

We would like to write the event in the probability in terms of a stopping time. Let $N_x = \inf\{u : \ell_u \geq x\}$ be the first time that the process $\ell_u$ reaches the level $x$. Stopping have occurred by time $t$ if, and only if, the maximum of the process in that interval of time is no less than $x$. Consequently, $\{\max_{0 \leq u \leq t} \ell_u \geq x\} = \{N_x \leq t\}$.

We apply the likelihood-ratio identity in order to compute the probability. Interestingly enough, the statistic $\ell_u$ is the log-likelihood ratio for testing $H_0 : \mu = 0$ versus $H_1 : \mu = 1$ on the basis of observing the Brownian motion over the interval of time $[0, u]$. Using the sequential likelihood ratio identity, which is justified by essentially the same argument as in the previous section, we get the representation:

$$\int_0^\infty e^t P(N_x \leq t) dx = \int_0^\infty e^t E_1(e^{-t\ell_u}; N_x \leq t) dx .$$

However, due to the continuity of the path of the Brownian motion, there is no overshoot involved. Upon stopping, over the event $\{N_x \leq t\}$, we have that $\ell_{N_x} = x$. Consequently, we get that the integral is equal to:

$$= \int_0^\infty P(N_x \leq t) dx = \int_0^\infty P(\max_{0 \leq u \leq t} \ell_u \geq x) dx = E_1(\max_{0 \leq u \leq t} \ell_u) .$$

The last equation results from the standard representation of the expectation in terms of the integral of the survival function.

We may conclude that:

$$\mathcal{H}(t) = E_1(\max_{0 \leq u \leq t} \ell_u) + 1 = t/2 + E_1(\max_{0 \leq u \leq t} (\ell_u - \ell_t)) + 1 ,$$

where $t/2$ is the expectation of $\ell_t$ under the alternative distribution that assigns the coefficient 1 to the expectation of the Brownian motion. This expectation is the leading term. In order to bound the second term observe that the joint distribution of the process $\ell_u - \ell_t = -(\ell_t - \ell_u)$, considered in reverse order for $u$ ranging from $t$ to 0, is the same as the null distribution of the original process. As a result:

$$0 \leq E_1(\max_{0 \leq u \leq t} (\ell_u - \ell_t)) = E(\max_{0 \leq u < \infty} \ell_u) \leq E(\max_{0 \leq u < \infty} \ell_u) = 1 .$$

The last equation follows from the fact that $P(\max_{0 \leq u < \infty} \ell_u \geq x) = e^{-x^2}$, which is the significance level of the power-one sequential probability ration test when observing a Brownian motion. Again, no overshoot is involved.

The fact that $\mathcal{H}(t)/t \to 0.5$, when it is inserted in the expression for the asymptotic upper bound $[2.9]$, gives a result for $\tau \to \infty$. The expression for the upper bound becomes:

$$P(\max_{\theta \in T} Z_\theta \geq z) \sim z^2 e^{-\frac{z^2}{2}}(2\pi)^{-\frac{1}{2}} \cdot (0.5)^2(t_1 - t_0)(1/h_0 - 1/h_1) . \quad (2.9)$$

This completes the easy part of the proof.

The more involved component of the proof is to show that the asymptotic upper bound is tight. The double-sum method establishes that fact via the
investigation of the sum of probabilities of simultaneous crossing in a pair of distinct subregions:

$$\sum_{(i,j)(k,l)\neq(i,j)} \sum P\left( \max_{A_{\theta} \in A_{ij}} Z_{A_{\theta}} \geq z, \max_{A_{\eta} \in A_{kl}} Z_{A_{\eta}} \geq z \right).$$

We will not present the proof that this double sum is of lower order. The interested reader may refer to [11]. However, we will mention the fact that proof employs Slepian’s inequality, which stochastically bounds the extreme value of a gaussian field in terms of the extreme value of a gaussian field with smaller correlation between components. This inequality, like the Borel’s inequality, is unique to the gaussian case and does not generalize to non-gaussian settings.

2.4 Other methods

TBA
Chapter 3

Approximation of the Local Rate
3.1 Introduction

In the previous chapter we presented an array of methods that can be used in order to analyze the distribution of extremes in a random field. In this chapter we concentrate on yet another method, a method that is based on a transformation of the measure. The components of this method will be demonstrated in the context of the two examples that were introduced in the previous chapter.

The first example involved a sequential testing of hypothesis. This example deals with the case where a stream of independent observations is accumulating. The aim is to test the null hypothesis that the density of the observations is $f$ versus the alternative hypothesis that the density is $g$. The statistics that are used are the log-likelihood ratio statistics:

$$\ell_n = \sum_{i=1}^{n} \log\{g(X_i)/f(X_i)\}.$$  

These statistics, in combination with a threshold $x$, define the test’s stopping time: $N_x = \inf\{n : \ell_n \geq x\}$.

The significance level of the test, the probability under the null regime that the stopping time is finite, was analyzed in the previous chapter using tools of sequential analysis. In particular, a sequential likelihood ratio identity was applied. The resulting approximation was given in (2.3) in the form:

$$\lim_{x \to \infty} e^x P(N_x < \infty) = \exp\left\{-\sum_{n=1}^{\infty} n^{-1} [P(g(\ell_n \leq 0) + P(\ell_n > 0)) \right\} \frac{E_g(\ell_1)}{\sum_{n=1}^{\infty} n^{-1}}.$$  

Currently we are interested in repeating the analysis using an alternative method. This alternative method will also apply a likelihood ratio identity, albeit a non-sequential one. The other components of the alternative method will bare no resemblance to the sequential methods that we used before.

In the second example we defined a scanning statistic for detecting a signal with the aid of the stochastic integral: \[
\int \left[g\left(\frac{x-t}{h}\right)/h\right]^2 dB_x. \]

This scanning statistic employed a kernel $g$ and was parameterized by the location and width of the signal $\theta = (t, h)$. Specifically, the components of the random field were the standardized score statistics that were associated with the parameter values:

$$Z_{\theta} = \frac{\int g((x-t)/h)/h^{1/2} dB_x}{\left\{\int [g((x-t)/h)]^2 h^{-1} dx\right\}^{1/2}},$$  

for $\theta \in T$. The goal was to approximate the probability that this two-dimensional field exceeds a high threshold $z$.

The analysis was one when the function $g$ was smooth, resulting in a smooth gaussian field, and another when the kernel was noncontinuous. In the case of a smooth kernel $g$ we were able to use a method that is based on consideration of integral geometry in order to obtain a very accurate approximation. Keeping track only of the leading term produced a formula that was presented in (2.4):

$$P\left(\sup_{\theta \in T} Z_{\theta} \geq z\right) \sim ze^{-\frac{1}{2}z^2} (2\pi)^{-3/2} \int_T \left|\Sigma_{\theta}\right|^{1/2} d\theta.$$  

The matrix in the integral is the variance-covariance matrix of the gradient of the random field, evaluated at the parameter value.
When the Gaussian random field is not smooth the geometrical method cannot be used. Instead, the double-sum approach is an option. Specifically, for \( g \) the indicator of the interval \([-0.5, 0.5]\) we saw that the double-sum method produce the approximation:

\[
P(\max_{\theta \in T} Z_\theta \geq z) \sim z^3 e^{-\frac{1}{2} z^2} (2\pi)^{-\frac{1}{2}} \cdot (0.5)^2 (t_1 - t_0)(1/h_0 - 1/h_1).
\]

This approximation appeared in (2.9).

In the following sections we will give the details of the measure-transformation technique in the context of these examples. We expect, of course, to obtain the same approximations, although they may have different representations.

For convenience we divide the application of the method to several distinct steps. The details in each step may vary depending on the specifics of the application. Occasionally, an entire step may be skipped. However, we find it useful to have this conceptual organization when analyzing a new problem that is associated with the extremes of a random field.

### 3.2 Preliminary localization and approximation

The method that we are about to use involves a likelihood ratio identity that transforms the distribution of the random field. Consequently, the central argument is conducted in the context of that transformed distribution. However, it may be useful to start by the examination of the field in its initial distribution in order to identify the order of magnitude of the approximation of the tail and to pinpoint regions of the parameter space where extreme values are more likely to occur.

In principle, the measure transformation can be carried out directly in the context of a continuous parameter space. However, it is technically more convenient to do the analysis in a discrete setting in which some boundedness conditions are naturally met. Consequently, one may consider adding a preliminary step in which maximization over the entire parameter set is replaced by a maximization over a discrete subset. This subset should be dense enough to assure an accurate approximation of the original maximum but not too dense so as to loose the advantages of discretization. As a matter of fact, the analysis that is involved in selecting the dense subset may help in understanding the local behavior of the random field and may shed light on the characteristics of the local field that is an integral part of the main tool.

In the next subsection we carry out the preliminary localization step in the context of the first example and produce in the subsequent subsection a discrete approximation of the maxima in the context of the scanning statistic example.

#### 3.2.1 Localization

A crude tool, that may be nonetheless very effective, involves the examination of the marginal tail probabilities of the random field.

Consider the first example. This example is formulated in terms of the stopping time \( N_x \). If we wish to analyze it in the context of a random field we should first reformulate the problem as such. Indeed, the stopping time is finite
if, and only if, the process (a one-dimensional field) ever crosses the threshold:

\[ \{ N_x < \infty \} = \{ \max_{1 \leq n < \infty} \ell_n \geq x \} . \]

The probability we seek can be obtained by the investigation of the extremes of the field \( \ell_n \), with \( 1 \leq n < \infty \) as the parameter set.

A lower bound on the probability in terms of the marginal distributions can be written as:

\[ P\left( \max_{1 \leq n < \infty} \ell_n \geq x \right) \geq \max_{1 \leq n < \infty} P(\ell_n \geq x) . \]

The marginal probabilities do depend on the value of the parameter and maximization provides information and sheds some light on the problem.

In the examination of the marginal probabilities \( P(\ell_n \geq x) \) it is tempting to use the Central Limit Theorem. Admittedly, the statistic \( \ell_n \) is a sum of independent and identically distributed random variables. However, some caution is advised. The Central Limit Theorem approximates the central part of the distribution. We, on the other hand, are interested in the tail of the distribution, the probability for large values of \( x \).

In order to obtain a better approximation of the tail for a sum of independent and identically distributed random variable a likelihood ratio can be used:

\[ P(\ell_n \geq x) = E_g( e^{-\ell_n}; \ell_n \geq x ) = \int_x^\infty e^{-y} g_n(y) dy , \tag{3.1} \]

where \( g_n \) is the density of \( \ell_n \) under the alternative distribution \( P_g \). At this stage we are in a better position to apply the Central Limit Theorem, with the application conducted on the distribution of \( \ell_n \) under the alternative.

However, since our goal now is to approximate the density \( g_n \), a local version of the Central Limit Theorem, a version that estimates the density, is more appropriate. Such a local version permits the replacement of \( g_n \), the density of \( \ell_n \) under the alternative distribution, by the normal density with the same expectation and variance.

In the current example we consider \( \ell_n = \sum_{i=1}^n \log\left\{ g(X_i)/f(X_i) \right\} \), a sum of independent and identically distributed components. The expectation of each of the components is the Kullback-Leibler information index:

\[ I = \int g(x) \log\left\{ f(x)/g(x) \right\} dx , \]

and the variance is \( \sigma^2 = \int g(x)(\log\left\{ f(x)/g(x) \right\} - I)^2 dx \).

If we substitute the density \( g_n \) by its approximation we get that

\[ P(\ell_n \geq x) \approx \frac{1}{\sqrt{2\pi n\sigma^2}} \int_x^\infty e^{-y} e^{-\frac{(y-x)^2}{2n\sigma^2}} dy = \frac{e^{-x}}{\sqrt{2\pi n\sigma^2}} \int_0^\infty e^{-z-\frac{(z+x-nI)^2}{2n\sigma^2}} dz , \]

with the second equality following from setting \( z = y - x \).

Consider the right-hand-side of the above approximation as a function of \( n \) and \( x \), for \( x - nI = O(\sqrt{n}) \) and for fixed \( z \). The exponent in the integral

\[ \text{We carelessly assume that } \ell_n \text{ has a density. If the distribution is discrete then integration should be replaced by a sum and densities by probabilities. Still, the formula is valid. In the general case Lebesgue-Stieltjes integration may be used.} \]
3.2. PRELIMINARY LOCALIZATION AND APPROXIMATION

is asymptotic to \(-z - (x - nI)^2/(2n\sigma^2)\) and is bounded by \(-z\). Applying the bounded convergence theorem we may be able to conclude that:

\[
P(\ell_n \geq x) \approx \frac{e^{-x}e^{-(x-nI)^2/2n\sigma^2}}{\sqrt{2\pi n\sigma^2}} \int_0^\infty e^{-z}dz = \frac{e^{-x}e^{-(x-nI)^2/2n\sigma^2}}{\sqrt{2\pi n\sigma^2}}.
\]

For values of \(n\) for which \(nI\) is still further away from \(x\) we get an approximation in which \(e^{-x}\) is multiplied by an exponentially vanishing function of the discrepancy between \(nI\) and \(x\). Maximization of the marginal probability is obtained when \(nI\) gets as close as possible to \(x\) and the maximizing probability is asymptotic to \(e^{-x}/\sqrt{2\pi\sigma^2x/I}\). Observe that the exponential part of the lower bound is correct but not the polynomial part.

It is interesting to consider an upper bound which is constructed by the summation of the marginal probabilities:

\[
P(\max_{1 \leq n < \infty} \ell_n \geq x) = P(\cup_{n=1}^\infty \{\ell_n \geq x\}) \leq \sum_{n=1}^\infty P(\ell_n \geq x).
\]

If we use again the approximations of the marginal probabilities in the critical range \(\{n : |n-x| = O(\sqrt{\pi})\} \approx \{n : |n-x/I| = O(\sqrt{x})\}\) we get that:

\[
\sum_{n=1}^\infty P(\ell_n \geq x) \approx \frac{e^{-x}}{I} \sum_{n=x/I-C\sqrt{x}}^{\infty} e^{-(n-x/I)^2/2n\sigma^2/I^2} \approx \frac{e^{-x}}{I},
\]

provided that \(C\) is large enough. Recalling (2.3) which states that the correct approximation in this case is a constant times \(e^{-x}\) we may conclude that the upper bound does produce the appropriate polynomial rate. However, the constant is still off. In summary, the crude analysis of this case proves that the correct exponential rate to use is \(e^{-x}\). This rate is produced by sequences of length \(x/I \pm O(\sqrt{x})\).

We add a little rigor to the discussion in this section by formulating a local limit theorem that can be used in order to establish (5.1). We also give more precise arguments to establish the range of parameter values in which the extremes are most likely to occur.

The specific local limit theorem is taken from [5]. The proof is essentially lifted from the chapter that deals with local limit theorems in that book. We give the proof here because in the sequel we will need a slight generalization of the given theorem. Such a generalization may be obtained by an appropriate modification to the original proof.

**Theorem 3.1** (Theorem 5.4 of Chapter 2 in [5]). Let \(X_1, X_2, \ldots \) be independent and identically distributed random variable. Assume that \(E(X_1) = 0\) and \(\text{Var}(X_1) = 1\). Let \(\varphi(\theta) = E(e^{i\theta X_1})\) be the characteristic function and assume that \(|\varphi(\theta)| < 1\) for all \(\theta \neq 0\). Then if \(\delta > 0\) is fixed and if \(x_n/\sqrt{n} \to x\), for a finite \(x\), then

\[
\lim_{n \to \infty} \sqrt{n}P(X_1 + \cdots + X_n \in (x_n, x_n + \delta)) = \delta \varphi(x).
\]

**Proof.** Let \(\delta > 0\) and set \(S_n = X_1 + \cdots + X_n\). Consider the characteristic function of the Polya’s distribution with density

\[
h_0(y) = \frac{1}{\pi} \cdot \frac{1 - \cos(\delta y)}{\delta y^2},
\]

which is of a bounded support and equal to:

\[ \hat{h}_0(u) = \begin{cases} 
1 - |u/\delta| & \text{if } |u| \leq \delta \\
0 & \text{otherwise.}
\end{cases} \]

Extend the density to a family of complex-valued functions by taking \( h_\theta(y) = e^{i\theta y}h_0(y) \) and observe that \( \hat{h}_\theta(u) = h_0(u + \theta) \).

We start by showing that for any given \( \theta \):

\[
\lim_{n \to \infty} \sqrt{n}E_h \theta(S_n - x_n) = \phi(x) \int h_\theta(y)dy ,
\]

where \( \phi \) is the density of the standard normal distribution. Indeed, from the inversion formula for characteristic functions at the presence of a density we get that:

\[
h_0(x) = \frac{1}{2\pi} \int e^{-iuu} \hat{h}_0(u)du
\]

and therefore, by the change of variable \( u = v + \theta \),

\[
h_\theta(x) = e^{i\theta x}h_0(x) = \frac{1}{2\pi} \int e^{-i(u-\theta)u} \hat{h}_0(u)du = \frac{1}{2\pi} \int e^{-ivv} \hat{h}_\theta(v)dv .
\]

Denote the distribution of \( S_n - x_n \) by \( F_n \) and apply Fubini’s theorem:

\[
E_h \theta(S_n - x_n) = \frac{1}{2\pi} \int \int e^{-iuu} \hat{h}_\theta(u)dudF_n(y) = \frac{1}{2\pi} \int \int e^{-ivu}dF_n(y) \hat{h}_\theta(u)du .
\]

The innermost integral corresponds to the characteristic function of \( S_n - x_n \), evaluated at \(-u\), hence

\[
= \frac{1}{2\pi} \int [\varphi(-u)]^n e^{ixx} \hat{h}_\theta(u)du .
\]

In order to show that the limit of the given integral is equal to the right-hand side of (3.2) we consider 3 regions. The first region is the region \([-\epsilon, \epsilon]\), for an appropriate \( \epsilon > 0 \), with the property that \( |\varphi(-u)| \leq \exp(-u^2/4) \) over the region. The second region is \([-M, M]\) \([-\epsilon, \epsilon]\), where \([-M, M]\) contains the support of \( h_\theta \). The last region is \( \mathbb{R} \setminus [-M, M] \), over which the integrand is equal to 0.

The last region does not contribute to the integral. The contribution of the second region is \( o(n^{-\frac{1}{2}}) \) small since it is bounded by \( (M/\pi)\eta^n \), for \( \eta = \sup_{\epsilon \leq |u| \leq M} |\varphi(-u)| < 1 \). For the first region we have, after multiplying by \( \sqrt{n} \) and changing the variable to \( v = u/\sqrt{n} \):

\[
\frac{\sqrt{n}}{2\pi} \int_{-\epsilon}^{\epsilon} [\varphi(-u)]^n e^{ixx} \hat{h}_\theta(u)du = \frac{1}{2\pi} \int_{-\epsilon\sqrt{n}}^{\epsilon\sqrt{n}} [\varphi(-v/\sqrt{n})]^n e^{ivv} / \sqrt{n} \hat{h}_\theta(v/\sqrt{n})dv .
\]

The integrand converges, for each fixed \( v \), to \( \exp(-v^2/2 + ivx) \hat{h}_\theta(0) \). Application of the dominated convergence theorem will give:

\[
\to n \to \infty \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-v^2/2 + ivx) \hat{h}_\theta(0)dv = \phi(x) \hat{h}_\theta(0) = \phi(x) \int h_\theta(y)dy .
\]
3.2. PRELIMINARY LOCALIZATION AND APPROXIMATION

The last equation follows from the definition of the Fourier transform. This completes the proof of (3.2), which we will use next in order to prove the statement of the theorem.

Towards that end we consider two sequences of measures on the real line. The first is the measure
\[ \mu_n(A) = \sqrt{n} P(S_n - x_n \in A), \]
which we wish to show that if converges to the measure \( \phi(x) \mu(A) \), for \( \mu \) the Lebesgue measure. The other measure is the probability measure
\[ \nu_n(A) = \frac{1}{\alpha_n} \int_A h_0(y) \mu_n(dy), \]
for \( \alpha_n = \sqrt{n} E h_0(S_n - x_n) \). Form (3.2) it follows, for \( \theta = 0 \), that \( \alpha_n \to \phi(x) \) and more generally that:
\[ \int e^{iy} \nu_n(y) \to \frac{1}{\alpha_n} \sqrt{n} E h_0(S_n - x_n) \]
Consequently, \( \nu_n \) converges in distribution to the Polya’s distribution. For the final move we apply the likelihood ratio identity:
\[ \frac{1}{\alpha_n} \nu_n(0, \delta] = \int \frac{1}{h_0(y)} h_0(y) \nu_n(dy) \to \frac{1}{\alpha_n} \sqrt{n} E h_0(S_n - x_n) \]

Consider in more carefully the range of values of \( n \) that contribute to the significance level of the stopping time \( N_x \) in the first example. Fix a large \( C \).

In order to show that \( n \) larger than \( n_1 = x/I + C \sqrt{x} \) need not be considered we may bound the probability that the stopping rule \( N_x \) obtains values larger than \( n_1 \):
\[ P(n_1 < N_x < \infty) = E_1(e^{-\ell_{N_x}}; n_1 < N_x < \infty) \leq e^{-x} P_1(n_1 < N_x) . \]

Clearly,
\[ P_1(n_1 < N_x) \leq P_1(\ell_{n_1} < x) = P_1(\ell_{n_1} - n_1 I < x - n_1 I) \leq \frac{n_1 \sigma^2}{(n_1 I - x)^2} , \]
which can be made as small as one wishes by increasing \( C \).

One may apply Kolmogorov’s maximal inequality in order to control the other end of the parameter space. Using the sequential likelihood ratio identity we get that:
\[ P(1 \leq N_x \leq n_0) \leq e^{-x} P_1(1 \leq N_x \leq n_0) = e^{-x} P_1(\max_{1 \leq n \leq n_0} \ell_n \geq x) , \]
where \( n_0 = x/I - C \sqrt{x} \). From the the fact that \( I > 0 \) it follows that
\[ P_1(\max_{1 \leq n \leq n_0} \ell_n \geq x) \leq P_1(\max_{1 \leq n \leq n_0} (\ell_n - n I) \geq x - n_0 I) \leq \frac{n_0 \sigma^2}{(n_0 I - x)^2} , \]
which is small for large \( C \). This completes the localization analysis for the first example.
Consider the second example. In this example that marginal distribution of a typical element $Z_\theta$ is the standard normal distribution. This is true across the parameter set $T$, with no subset sticking out. Consequently, the preliminary localization step can be skipped in this case. Still, it is useful to note that the survival function of the standard normal distribution provides a lower bound:

$$P(\max_{\theta \in T} Z_\theta \geq z) \geq \max_{\theta \in T} P(Z_\theta \geq z) = 1 - \Phi(z) \sim \frac{e^{-\frac{1}{2}z^2}}{z\sqrt{2\pi}}.$$ 

This lower bound does produce the correct exponential term but it fails to identify the correct polynomial modification.

### 3.2.2 A discrete approximation

The example of scanning statistic involves a continuous parameter set $T$. The asymptotic expansion of the tail probability was such that

$$P(\sup_{\theta \in T} Z_\theta \geq z) \sim z^{\alpha-1}e^{-\frac{1}{2}z^2} \times \text{const.},$$

where $\alpha = 2$ corresponded to the case where the kernel in continuous, $\alpha = 1$ corresponded to the case where the kernel was the indicator of the interval $[-0.5, 0.5]$. The constant depended on the specific selection of the kernel.

Currently, we are interested in finding a discrete subset $\hat{T} \subset T$ such that the tail probability when maximization is restricted to $\hat{T}$ is a good approximation of the tail probability of the unrestricted maxima. Specifically, we want to make sure that the choice of the subset preserves the exponential and polynomial rate. Later we will check, after the application of the measure-transformation technique, that when the subset becomes denser then the resulting constant converges to the constant that is associated with a continuous parameter set.

We produce a lower bound and an upper bound for the probability of the tail in terms of its approximation over a discrete subset. Clearly,

$$P(\max_{\theta \in \hat{T}} Z_\theta \geq z) \leq P(\max_{\theta \in T} Z_\theta \geq z).$$

On the other hand, for any given $\epsilon > 0$,

$$P(\max_{\theta \in \hat{T}} Z_\theta \geq z) \leq P\left( \max_{\theta \in T} Z_\theta \geq z - \frac{\epsilon}{z} \right) + P\left( \max_{\theta \in T} Z_\theta \leq z - \frac{\epsilon}{z}, \sup_{\theta \in T} Z_\theta \geq z \right).$$

We argue that the left-hand side of the first inequality and the first of the two terms on the right-hand side of the second inequality are approximately equal to each other. Justification follows from the fact that the expansion of these terms, as a function of the threshold, produces an exponential, a polynomial, and a converging contributions. The converging contribution is the same for both terms. The ratio of the polynomial contributions converge to one since the ration of the thresholds does. The exponential contribution is $\exp(-z^2/2)$ in the first inequality and it is $\exp\{-z^2/(2z)\}$ in the second one. The ratio between the two converges to $\exp(-\epsilon)$, which can be made as close to 1 as one wishes by an appropriate selection of $\epsilon$. Thus, in order to prove the equivalence of the discrete approximation it is sufficient to show that the error
term \( P(\max_{\theta \in T} Z_\theta \leq z - \frac{\epsilon}{z}, \sup_{\theta \in T} Z_\theta \geq z) \) is of smaller order compared to the other terms that appear in the inequalities, namely \( z^{\frac{1}{2}} e^{-\frac{1}{2}z^2} \).

Associate, with each \( \theta \in T \), a local subregion \( T_\theta \subset T \) that contains \( \theta \) and make sure that \( T \subset \bigcup_{\theta \in T} T_\theta \). The global maxima should occur in at least one of these subregions. Consequently, via the fact that \( \{\max_{\theta \in T} Z_\theta \leq z\} \subset \{Z_\theta \leq z\} \) and an application of Boole’s inequality:

\[
P(\max_{\theta \in T} Z_\theta \leq z - \frac{\epsilon}{z}, \sup_{\theta \in T} Z_\theta \geq z) \leq \sum_{\theta \in T} P(\max_{\theta \in T} Z_\theta \leq z - \frac{\epsilon}{z}, \sup_{\theta \in T_\theta} Z_\theta \geq z).
\]

The analysis proceeds by producing a bound for each of the probabilities in the sum. The bound is constructed by conditioning on the value of \( Z_\theta \) and considering the supremum of the conditional field, restricted to the local subregion.

Specifically, after dividing by the expected rate, we get that the probabilities in the sum are of the form:

\[
z^{\frac{1}{2}} e^{-\frac{1}{2}z^2} P\left(\max_{\theta \in T} Z_\theta \leq z - \frac{\epsilon}{z}, \sup_{\theta \in T_\theta} Z_\theta \geq z\right)
= z^{\frac{1}{2}} e^{-\frac{1}{2}z^2} P\left(z(Z_\theta - z) \leq -\epsilon, \sup_{\theta \in T_\theta} (Z_\theta - z) \geq 0\right)
= z^{\frac{1}{2}} e^{-\frac{1}{2}z^2} P\left(\sup_{\theta \in T_\theta} Y_\theta \geq 0\right) dy,
\]

that results from the fact that the distribution of \( Y_\theta = z(Z_\theta - z) \) is normal with mean \( -z^2 \) and variance \( z^2 \).

The random field \( \{Y_\theta : \theta \in T_\theta\} \), considered in the conditional distribution given that \( Y_\theta = y \) and restricted to the local region, is a gaussian random field. The expectation of a component of the field is:

\[
E(Y_\theta | Y_\theta = y) = y - z^2[1 - \text{Cov}(Z_\theta, Z_\theta)](1 + y/z^2).
\]

This expectation depends on \( y \). The variance of a component is:

\[
\text{Var}(Y_\theta | Y_\theta = y) = z^2[1 - \{\text{Cov}(Z_\theta, Z_\theta)\}^2]
\]

and the covariance between two components is:

\[
\text{Cov}(Y_\theta, Y_\gamma | Y_\theta = y) = z^2(\text{Cov}(Z_\theta, Z_\gamma) - \text{Cov}(Z_\theta, Z_\theta)\text{Cov}(Z_\gamma, Z_\theta)),
\]

both are independent of \( y \).

Denote by \( X_\theta = Y_\theta - E(Y_\theta | Y_\theta = y) \) the centered random field and observe that the covariance stricture of the field does not depend on \( y \). We will use the upper bound:

\[
P\left(\sup_{\theta \in T_\theta} Y_\theta \geq 0 | Y_\theta = y\right) \leq P\left(\sup_{\theta \in T_\theta} X_\theta \geq x(y, z)\right)
\]

where \( x(y, z) = -y + z^2[1 - \inf_{\theta \in T_\theta} \text{Cov}(Z_\theta, Z_\theta)](1 + y/z^2) \). We will also apply a bound on the tail distribution of a gaussian field in order to produce a upper bound for the last probability.

There are two major tools for bounding the tail of the supreme of a gaussian field. One tool in the Borel’s inequality that was mentioned in the previous
chapter. This inequality is based on bounding the expectation (or the median) of the supreme. An alternative tool, that does not require such bound, is the Fernique’s inequality. Here we will use the latter, the proof of which can be found in [9], page 164-167:

**Theorem 3.2 (Fernique’s Inequality).** Let $T = \prod_{j=1}^d [a_j, b_j]$ be a rectangle in the $d$-dimensional Euclidean space and let $\{X_\theta : \theta \in T\}$ be a centered gaussian field with a bounded variance: $0 < \sigma^2 = \sup_{\theta \in T} \text{Var}(X_\theta) < \infty$. Assume the $\mathbb{E}[(X_\theta - X_\eta)^2] \leq \varphi(||\theta - \eta||)$, for some continuous and nondecreasing function $\varphi$ that satisfies $\int_0^\infty \varphi(e^{-y})dy < \infty$. The for $\lambda > 0$, $x \geq 1$, and $C > \sqrt{2d\log 2}$ we have that:

$$
P(\sup_{\theta \in T} X_\theta \geq x\{\sigma + 2(\sqrt{2} + 1)C \int_1^\infty \varphi(\sqrt{d}\lambda 2^{-y})dy\}) \leq (2^d + B)e^{-\frac{1}{2}x^2} \prod_{j=1}^d \left( \frac{b_j - a_j}{\lambda} + \frac{1}{2} \right),$$

where $B = \sum_{n=1}^\infty \exp\{-2^{n-1}(C^2 - 2d\log 2)\}$

The critical ingredient in the application of the inequality is the relation between the Euclidian distance of points in the parameter space and the expected value of the squared difference between the elements of the random field that are associated with these points. This expected squared difference is tightly linked to the covariance function since

$$
\mathbb{E}[(X_\theta - X_\eta)^2] = \text{Var}(X_\theta) + \text{Var}(X_\eta) - 2\text{Cov}(X_\theta, X_\eta)
= 2\sigma^2 [1 - \text{Cov}(Z_\theta, Z_\eta)] - \sigma^2 [\text{Cov}(Z_\theta, \dot{Z}_\theta) - \text{Cov}(Z_\eta, Z_\theta)]^2.
$$

We consider parameter values that are close to each other. As we discovered in the previous chapter, the expansion of the covariance function about its maximal value 1 differed between the case of a smooth gaussian field and a gaussian field which is not smooth.

In the smooth case we were able to compute the gradient of the covariance function and obtain that $\frac{\partial}{\partial \theta} \text{Cov}(Z_\theta, Z_\theta) = \text{Cov}(Z_\theta, \dot{Z}_\theta)$. However, in the case under consideration we also have that $\text{Var}(Z_\theta) = \text{Cov}(Z_\theta, Z_\theta) = 1$, for all $\theta \in T$. Consequently, when we take derivatives with respect to $\theta$ we get that:

$$0 = \frac{\partial}{\partial \theta} \text{Cov}(Z_\theta, Z_\theta) = 2\text{Cov}(Z_\theta, \dot{Z}_\theta) \implies \frac{\partial}{\partial \theta} \text{Cov}(Z_\theta, Z_\theta) \bigg|_{\theta = \theta} = 0.$$

The conclusion is that in the smooth case $1 - \text{Cov}(Z_\theta, Z_\theta) \leq c||\theta - \theta||^2$, for some constant $c$. On the other hand, in the case where the kernel was an indicator of an interval we obtained that $1 - \text{Cov}(Z_\theta, Z_\theta) \leq c||\theta - \theta||$, for possibly a different constant $c$. (As a matter of fact, the relation we found was expressed in terms of the $L_1$ norm. However, all norms are equivalent to each other in the finite dimensional space.) One may combine the two cases using the statement:

$$1- \text{Cov}(Z_\theta, Z_\theta) \leq c||\theta - \theta||^\alpha,$$

with $\alpha = 2$ for the smooth case and $\alpha = 1$ for the other case.
Consider the square local subregion $T_\theta = \theta + [0, \delta z^{-\frac{2}{5}}] \times [0, \delta z^{-\frac{2}{5}}]$. For any two points $\vartheta = \theta + \gamma z^{-\frac{2}{5}}$ and $\eta = \theta + \xi z^{-\frac{2}{5}}$ that belong to this subregion we have that
$$E[(X_\vartheta - X_\eta)^2] \leq 2c(1 + o(1)) \|\gamma - \xi\|^\alpha,$$
with $\lim_{z \to \infty} o(1) = 0$, uniformly over the subregion. This means that we may use the inequality in a re-parameterized setting with $\varphi(x) = c_1 x^\alpha$, with $c_1 > 2c$.

The maximization of the variance over the subregion results in:
$$\sigma^2 = \max_{\vartheta \in T_\theta} \text{Var}(X_\vartheta) = \max_{\vartheta \in T_\theta} z^2 \{1 - \{\text{Cov}(Z_\vartheta, Z_\theta)\}^2\} \leq c_2 \delta^\alpha$$
and therefore, by taking $\lambda = \delta z^{-\frac{2}{5}}$ we obtain form the application of Fernique’s inequality that
$$P\left( \sup_{\vartheta \in T_\theta} X_\vartheta \geq C_\alpha x \delta^\frac{2}{5} \right) \leq B_\alpha e^{-\frac{1}{2} x^2},$$
for some universal constants $B_\alpha$ and $C_\alpha$ and for $x \geq 1$. For the threshold $x(y, z)$
$$x(y, z) = -y + z^2 \{1 - \inf_{\vartheta \in T_\theta} \text{Cov}(Z_\vartheta, Z_\theta)\} \{1 + y/z^2\} \sim -y(1 - \delta^\alpha/z^2) + d \delta^\alpha,$$
for yet another constant $d$. Finally, the total number of square subregions $T_\theta$ that are required in order to cover $T$ is asymptotic to $|T| \cdot \delta^{-\frac{2}{5}} z^{\frac{2}{5}}$, where $|T|$ is the area of $T$. Putting it all together gives:
$$P\left( \max_{\theta \in T} Z_\theta \leq z - \frac{\epsilon}{z}, \sup_{\theta \in T} Z_\theta \geq z \right) \leq |T| \cdot \frac{\delta^{-\frac{2}{5}}}{\sqrt{2\pi}} \int_{-\infty}^{-\frac{\epsilon}{z}} e^{-y - \frac{y^2}{2z^2}} P\left( \sup_{\theta \in T_\theta} X_\vartheta \geq x(y, z) \right) dy$$
$$\leq |T| \cdot \frac{B_\alpha \delta^{-\frac{2}{5}}}{\sqrt{2\pi}} \int_{-\epsilon}^{\infty} e^{y - \frac{y^2}{2z^2}} e^{-\frac{1}{2} \{x(y, z)\}^2/[C_\alpha^2 \delta^\alpha]} dy.$$

With the divergence of $z$ to infinity the last integrand converges to $\text{exp}\{y - (y - d \delta^\alpha)^2/[2C_\alpha^2 \delta^\alpha]\}$. Therefore, by the dominated convergence theorem we get that the upper bound converges to
$$|T| \cdot \frac{B_\alpha \delta^{-\frac{2}{5}}}{\sqrt{2\pi}} \int_{-\epsilon}^{\infty} e^{y - \frac{(y - d \delta^\alpha)^2}{2C_\alpha^2 \delta^\alpha}} dy = |T| \cdot \frac{B_\alpha C_\alpha}{\delta^{\frac{2}{5}} } e^{(d + C_\alpha^2) \delta^\alpha} \left[1 - \Phi\left( \frac{\epsilon - (d + C_\alpha^2) \delta^\alpha}{C_\alpha \delta^\frac{2}{5}} \right) \right].$$

With the decrease in delta the upper bound, which is fixed of an exponentially decreasing function of $\delta^{-\alpha}$ and a polynomially increasing function of the same quantity, is converging to 0. As a result, for each fixed $\epsilon > 0$ one may find a $\delta > 0$ so that the error that results from the approximation with a grid of span $\delta z^{-\frac{2}{5}}$ is as small as one wishes.

### 3.3 Measure transformation

After restricting the problem to the region in the parameter space that matters and substituting a continuous parameter set by a discrete approximation one can apply the measure transformation that is at the heart of the entire approach. Frequently, this step is where art is required. The idea behind it is the old tested “split and conquer”, i.e. represent the problem of computing
the probability of a maximum of the field over the set of parameters as a sum of expectations, a collection of simpler problems. Each of the elements in the sum is then approximated in the subsequent step and the approximations are summed up in the last step in order to produce the final approximation. The representation of the probability of the maximum as a sum of expectations is obtained via a likelihood-ratio identity. Each parameter value that appears in the maximization is associated with a likelihood ratio and a sum of all likelihood ratios is used to produce the identity. We illustrate this step in the two examples.

Start with the first example. Take \( T = \{n_0, \ldots, n_1\} \) to be the parameter set. The event of interest is:

\[
A = \{ \max_{n \in T} \ell_n \geq x \}.
\]

In this case, each parameter \( n \) is already associated with the likelihood \( \exp(\ell_n) \), so we might as well use these likelihoods. The sum to be used in the likelihood ratio identity can be \( \sum_{n \in T} \exp(\ell_n) \), with equal weight given to each of the likelihoods.

A careful reader should be concerned by the proposal to use \( \exp(\ell_n) \) as a likelihood ratio in the likelihood ratio identity applied to the event \( A \). The likelihood is a function of only the first \( n \) observations whereas the event is a function of all observations (or, at least the observations up to \( n_1 \)). That is true. However, it is worth noting that \( \exp(\ell_n) \) is nonetheless a likelihood ratio for the entire collection of observations. Check that it is a likelihood ratio for the alternative distribution of independent observations that assign the density \( g \) to the first \( n \) observations and the density \( f \) to the ensuing ones. Denote this alternative distribution by \( P_n \) (and the resulting expectation by \( E_n \)).

We are now in the position to apply the likelihood ratio identity:

\[
P(A) = E\left( \sum_{n \in T} \exp(\ell_n); A \right) = \sum_{n \in T} E_n \left( \frac{1}{\sum_{m \in T} \exp(\ell_m)}; A \right),
\]

where the last equality follows from changing the order of summation and expectation and the application of the likelihood ratio identity to each of the summands.

We rearrange each term and add notations:

\[
E_n \left( \frac{1}{\sum_{m \in T} \exp(\ell_m)}; A \right) = e^{-x} E_n \left( \frac{\max_{m \in T} \ell_m - \ell_n}{\sum_{m \in T} \exp(\ell_m - \ell_n)} e^{-(\ell_n - x + \max_{m \in T} (\ell_n - \ell_m))}; A \right)
\]

\[
= e^{-x} E_n \left( \frac{M_n}{S_n} e^{-(\ell_n + m_n)}; \tilde{\ell}_n + m_n \geq 0 \right),
\]

where \( S_n = \sum_{m \in T} \exp(\ell_m - \ell_n) \) is the sum of likelihood ratios, \( M_n = \max_{m \in T} \exp(\ell_m - \ell_n) \) is the maximal likelihood ratio, \( m_n = \log M_n \), and \( \tilde{\ell}_n = \ell_n - x \). In the last equation we rephrased the event \( A \) in terms of the newly defined random variables:

\[
A = \{ \max_{n \in T} \ell_n \geq x \} = \{ \tilde{\ell}_n + m_n \geq 0 \}.
\]

The essence of the final representation of a summand is the dissection to a large deviations exponential decay, given by the exponent \( e^{-x} \), and lower
order contributions that reside in the expectation. The random variables in expectation are further dissected to random variables that are influenced mainly by local perturbations and the random variable that captures the main part of the variability. The latter is the random variable $\ell_n$, which has expectation $nI - x$ under the alternative and variance $n\sigma^2$. We call this random variable the *global term*. The other random variables are $S_n, M_n$ (and its log). These random variables are functions of what we call the local field, with elements, that in the current case are of the form $\ell_m - \ell_n$ and are parameterized by $m$.

The localization theorem of the next section investigates the limiting joint distribution of the global term and the local field. If the situation is that the two components are asymptotically independent then the expectation can be dissected as a product of two terms. One is the expectation of the limit of the ratio between $M_n$ and $S_n$ and the other is similar in flavor to the expansion of the marginal probability $P(\ell_n \geq x)$ that was discussed informally in the previous section. We leave the details to the next section and to the entire chapter that is devoted to the localization theorem.

Consider next the second example of a gaussian random field. In this example the elements of the field are the statistics $Z_\theta$, for $\theta \in T$. Unlike the other example, these statistics are not log-likelihood ratios, since their exponent does not integrate to 1. Therefore, we cannot use them directly for the measure transformation as we did before. However, we can still produce for each $Z_\theta$ a likelihood ratio, for example by means of exponential tilting.

Exponential tilting, when the distribution of $Z_\theta$ is standard normal, refers to the likelihood ratio $\exp(\mu Z_\theta - \mu^2/2) = \exp(\ell_\theta)$, where $\mu$ is a number we may specify to our convenience. The alternative distribution of $Z_\theta$ after tilting is the normal distribution with expectation $\mu$ and with variance 1. Denote this distribution by $\tilde{P}_\theta$. Just as before, we should consider this distribution not only in the context the specific $Z_\theta$ but in the context of the joint distribution of the entire field. Namely, we should ask ourselves what is the alternative joint distribution of the field that produces $\ell_\theta$ as a log-likelihood ratio?

In order to answer this question consider a (finite) collection of normal random variables $Z = \{Z_\theta\}$ with zero mean and unit variance, written as a column vector. Let $\Sigma$ be the variance-covariance matrix of this collection. Consider an alternative distribution for this collection with the same variance-covariance matrix but with an expectation vector $\eta$. The resulting log-likelihood ratio is $\eta'\Sigma^{-1}Z - (\eta'\Sigma^{-1}\eta)/2$. We can obtain the form of a tilted log-likelihood ratio if we take $\eta = \mu\rho$, where $\rho$ is the column of the matrix $\Sigma$ that is associated with the variable $Z_\theta \in \{Z_\theta\}$. This is the case since $\eta'\Sigma^{-1}Z = \mu Z_\theta$ and $(\eta'\Sigma^{-1}\eta)/2 = \mu^2/2$. To conclude, the alternative distribution $\tilde{P}_\theta$ assigns a gaussian distribution to field. The covariance structure under the alternative distribution is identical to the covariance structure under the null. The expectation is shifted. Under the alternative distribution the expectation of the component $Z_\theta$ is equal to $\mu \cdot \text{Cov}(Z_\theta, Z_\theta)$, for $\theta \in T$. 

3.3. MEASURE TRANSFORMATION

45
We have still the freedom to select the numerical value of \( \mu \). We chose the value that will produce the correct large deviations exponential decay for the marginal probability \( P(Z_\theta \geq z) \). The resulting value is \( \mu = z \), which produces the log-likelihood ratio \( \ell_\theta = zZ_\theta - z^2/2 \).

Misusing the notations somewhat we denote by \( T \) the discrete approximation of the original continuous parameter set. In the examples that we follow the discrete parameter set is of the form

\[
T = \{(t, h) : t = i\delta z^{-1}, h = j\delta z^{-1}, t_0 \leq t \leq t_1, h_0 \leq h \leq h_1 \}
\]

in the case of a continuous kernel and it is of the form:

\[
T = \{(\theta_1, \theta_2) : \theta_1 = i\delta z^{-2}, \theta_2 = j\delta z^{-2}, \theta_0 \leq (\theta_1 + \theta_2)/2 \leq t_1, h_0 \leq \theta_2 - \theta_1 \leq h_1 \}
\]

when the kernel is an indicator of an interval.

In either cases, let \( A \) be the event of interest:

\[
A = \{ \max_{\theta \in T} Z_\theta \geq z \}.
\]

In the zero-mean unit-variance gaussian case with the discrete approximation of the parameter space we use the sum of the likelihood ratios \( \sum_{\theta \in T} e^{\ell_\theta} \) for the measure transformation:

\[
P(A) = E\left( \frac{1}{\sum_{\theta \in T} e^{\ell_\theta}} ; A \right) = \sum_{\theta \in T} E_\theta \left( \frac{1}{\sum_{\theta \in T} e^{\ell_\theta}} ; A \right),
\]

Each element in the sum may be represented more conveniently as:

\[
E_\theta \left( \frac{1}{\sum_{\theta \in T} e^{\ell_\theta}} ; A \right) = e^{-\frac{1}{2}z^2} E_\theta \left( \max_{s \in T} e^{\ell_\theta - \ell_s} \sum_{\theta \in T} e^{-[\theta - (\ell_s - \ell_\theta)]} ; A \right)
\]

\[
= e^{-\frac{1}{2}z^2} E_\theta \left( \max_{s \in T} e^{z(Z_\theta - Z_s)} \sum_{\theta \in T} e^{z(Z_\theta - Z_s)} e^{-[z(Z_\theta - Z_s) + \max_{s \in T} z(Z_\theta - Z_s)]} ; A \right)
\]

\[
= e^{-\frac{1}{2}z^2} E_\theta \left( \frac{M_\theta}{S_\theta} e^{-[\ell_\theta + m_\theta]} ; \ell_\theta + m_\theta \geq 0 \right),
\]

where \( S_\theta = \sum_{\theta \in T} e^{z(Z_\theta - Z_s)} \) is the term that replaces \( S_n \), \( M_\theta = \max_{s \in T} e^{z(Z_\theta - Z_s)} \) replaces \( M_n \), \( m_\theta = \log M_\theta \), and \( \ell_\theta = z(Z_\theta - z) \). Again, we rephrased the event \( A \) in terms of the newly defined random variables:

\[
A = \{ \max_{\theta \in T} Z_\theta \geq z \} = \{ \ell_\theta + m_\theta \geq 0 \}.
\]

It is worth noting that while the large deviations decay in the gaussian example is not the same as the one obtained for the example of sequential testing, still the structure of the expectation that captures lower-order contributions is the same. In the second problem, like in the first, that expectation involves a global term and a local field. The global term in the gaussian case is \( z(Z_\theta - z) \) and the local field is \( z(Z_\theta - Z_\theta) \), parameterized by \( \theta \).

The distinction between the case of a smooth gaussian field and the case where the field is continuous but not smooth is expressed in the specifications of the grid that is used in order to approximate the continuous parameter set.
In the situation of a non-smooth field a denser grid is required. However, once the grid $T$ is specified, the resulting representation of the tail probability via measure transformation is identical for both cases. Differences will emerge again when we investigate in the next section the convergence of the components in the sum and, of course, in the finite step of obtaining the overall approximation, which is the subject of the last section.

3.4 Application of the localization theorem

This step is usually the most technically involved part of the proof. It deals with the limit of terms of the form:

$$E\left[(M_\kappa/S_\kappa) \exp[-(\tilde{\ell}_\kappa + \log M_\kappa)]; \tilde{\ell}_\kappa + \log M_\kappa \geq 0\right],$$

where $\kappa$ is associated with $\eta$ in the example of sequential testing and $\kappa$ is associated with $\theta$ in the example of a scanning statistic. These expectations emerged as elements in the representation of the tail probability of the maxima of the field. The expectation corresponds to an appropriate alternative distribution, not the original null distribution. The representation is a function of a “global” term $\tilde{\ell}_\kappa$ and a “local” random field. The former is associated with the specific log-likelihood ratio and the later corresponds to the difference between other log-likelihood ratios and the specific one. The approximation identifies the asymptotic distribution of the local field and the asymptotic independence between it and the global term. The limit is given in terms of a bounded functional of the limit of the local field and in terms of the density of the global element, evaluated at the origin.

In the case where the parameter set is continuous or very refined we may choose to approximate the maximization of the random field by the maximization over a cruder subset of the parameter space. The approximating subset should be dense enough to assure accurate approximation. That was illustrated above in the context of the scanning statistic. On the other hand, the subset should not be too dense. The analysis of the local field sheds light on the last statement. The limit of the local field should be such that the resulting functional is not only bounded but also strictly positive.

The localization theorem is stated in Chapter 5 in terms of an abstract index $\kappa$. Convergence to a limit occurs as $\kappa \to \infty$. In the context of a specific application the value of the parameter can be identified by the examination of the distribution of the global term $\tilde{\ell}_\kappa$. In the formulation of the localization theorem, $\kappa_1^2$ is the multiplicative factor that balances the rate of convergence of the density of $\tilde{\ell}_\kappa$ to 0. Consequently, one may typically equate $\kappa$ with the variance of the global term or chose it to be asymptotically proportional to that variance.

The statement of the localization theorem involves also a local $\sigma$-algebra denoted by $\hat{F}_\kappa$. This local $\sigma$-algebra is asymptotically independent of $\tilde{\ell}_\kappa$ but it carries enough information to construct approximations of the random variables that summarize the contributions of the local field. These random variables, denoted $\hat{M}_\kappa$ and $\hat{S}_\kappa$, are measurable with respect to $\hat{F}_\kappa$ and serve as approximations of the original random variables $M_\kappa$ and $S_\kappa$, respectively. The expectation of the ratio of these approximating random variables converges to a limit that is used in the formulation of the approximation of the tail probability.
Specifically, in the example of sequential testing we may take $\kappa = n$. One may then choose some finite $\tau$ and use $\bar{F}_\kappa = \sigma\{X_{n-\tau}, \ldots, X_{n+\tau}\}$, for the observations in the vicinity of $n$. The local field $\{\ell_m - \ell_k\}$, restricted to the region $|m - n| \leq \tau$, is measurable with respect to the given $\sigma$-algebra. Producing $\bar{M}_\kappa$ and $\bar{S}_\kappa$ by restricting the maximization (respectively, summation) to the reduced region of parameter values will result in random variables that approximate the original maximization (or summation), as long as $\tau$ is large enough. The distribution of the ratio $\bar{M}_\kappa/\bar{S}_\kappa$ is independent of $\kappa$, for all $\kappa$ not too close to $n_0$ or $n_1$. Hence, the expectation of the ratio is unchanged in the limit. This expectation approximates a similar expectation that involves a ratio in which $\tau = \infty$, defined for a double-ended sequence of observations that extends for all $m, -\infty < m < \infty$. That last functional is the term that is denoted by $E(M/S)$ and will be used in the approximation of the significance level of the sequential test.

For the example of a scanning statistic we may use $\kappa = z^2$, but keep track of the value of $\theta$. The local $\sigma$-algebra involves the local field $z(Z_\theta - Z_0)$ and can be taken to be $\sigma\{z(Z_\theta - Z_0) : ||\theta - \theta|| \leq \tau z^{-2}\}$, for some large $\tau$. If the field is smooth we use $\alpha = 1$. For the non-continuous kernel we use $\alpha = 2$. As a consequence of the approximation of the continuous parameter set by a discrete one we obtain that the number of parameter values in $\{\theta : ||\theta - \theta|| \leq \tau z^{-2}\}$ is finite. Again, we produce $\bar{M}_\kappa$ and $\bar{S}_\kappa$ by restricting maximization and summation to the smaller subset of parameter values. In the limit, as $\kappa \to \infty$, we get the the local random field $z(Z_\theta - Z_0)$ converges to a limit gaussian field. Subsequently, the functional $E(\bar{M}_\kappa/\bar{S}_\kappa)$ converges to a limit $E[\bar{M}/\bar{S}]$ that can be expressed in terms of the limit gaussian field, restricted to the subset of parameter values. The term that appears as the outcome of the limit in the statement of the localization theorem, namely $E(M/S)$, is obtained by letting $\tau$ go to infinity.

The localization theorem is formulated as a collection of conditions and a statement of a limit. Given an application, if the conditions can be verified then the limit is validated. In Chapter III we state and prove two flavors of the theorem. For pedagogical reasons we give first a relatively simple formulation. Later, we present a more complex formulation with conditions that are usually easier to verify. In the current section we verify the conditions for the two examples that we are tracking. Both examples are simple enough and can be validated using the simpler formulation of the theorem. However, in order to practice the more useful formulation we carry out the validation using the more complex formulation.

Specifically, we need to check 5 conditions. Given $\epsilon, c_3 > 0$, for a function $g(\kappa)$ to be specified for the given application but obeying the relation $\log \kappa \leq g(\kappa) \leq c_3 \kappa^{\frac{1}{2}}$, for some $C < \infty$, and for all large $\kappa$ we should have that:

\( \textbf{I}^*: \) $M_\kappa$, $S_\kappa$, $\bar{M}_\kappa$ and $\bar{S}_\kappa$ satisfy $0 \leq M_\kappa/S_\kappa \leq C$ and $0 \leq \bar{M}_\kappa/\bar{S}_\kappa \leq C$ with probability one.

\( \textbf{II}^*: \) Denote $A_{\epsilon_1}^\delta = \{|\log M_\kappa - \log \bar{M}_\kappa| > \epsilon\} \cup \{|\bar{S}_\kappa/S_\kappa - 1| > \epsilon\}$. For some $0 < \delta$ that does not depend on $\epsilon$:

$$\max_{|x| \leq g(\kappa)} \mathbb{P}(A_{\epsilon_1}^\delta \cap \{\ell_\kappa + \hat{m}_\kappa \in x + (0, \delta]\} \cap \{|\hat{m}| \leq g(\kappa)\}) \leq c_3 \kappa^{\frac{1}{2}}.$$
Section 3.4: Application of the Localization Theorem

III*: \( \mathbb{E}[\hat{M}_\kappa/\hat{S}_\kappa] \) converges to \( \mathbb{E}[\hat{M}/\hat{S}] \) and \( |\mathbb{E}[\hat{M}/\hat{S}] - \mathbb{E}[M/S]| \leq \epsilon_3 \).

IV*: There exist \( \mu \in \mathbb{R} \) and \( \sigma \in \mathbb{R}^+ \) such that for every \( 0 < \epsilon_4, \delta \), for any event \( E \in \mathcal{F}_\kappa \) having boundary measure 0, and for all large enough \( \kappa \):

\[
\sup_{|x| \leq 3 \phi(\kappa)} |\kappa^{\frac{1}{2}} P(\tilde{m}_\kappa \in x + (0, \delta], E) - \frac{\delta}{\sigma} \phi(\frac{\mu}{\sigma}) P(E)| \leq \epsilon_4 ,
\]

and also:

\( V^* : \mathbb{P}(|\log M_\kappa| > g(\kappa)), \mathbb{P}(|\log \hat{M}_\kappa| > g(\kappa)) \) and \( \mathbb{P}(\log M_\kappa - \log \hat{M}_\kappa < -\epsilon) \) are all \( o(\kappa^{-\frac{1}{2}}) \).

The localization theorem 5.2 states that if Conditions I*-V* hold then:

\[
\lim_{\kappa \to \infty} \kappa^{\frac{1}{2}} \mathbb{E}[\hat{M}_\kappa/\hat{S}_\kappa] e^{-(\tilde{\ell}_\kappa + \log M_\kappa)}; \tilde{\ell}_\kappa + \log M_\kappa \geq 0] = \sigma^{-1} \phi(\mu/\sigma) \mathbb{E}[M/S] .
\]

We go over the list of conditions, not necessarily in the given order, and check them for the example of sequential testing and for the example of a scanning statistic.

Condition I* is straightforward. In both cases we have that the ratio is between a maximum of likelihood ratios – non-negative statistics – and the sum of the same likelihood ratios. Since the sum of non-negative terms is always larger than the largest term we get that the ratio is bounded by \( C = 1 \). This is the case when the collection of likelihood ratios is unrestricted as well as when it is restricted to belong to the local sub-collection of parameter values.

Consider next Condition V*. This condition involves three probabilities. The last probability corresponds to the event where the maximization of the likelihood ratios over the entire parameter set is less than the maximization over the local sub-collection of parameter values. Clearly, the former is larger. Therefore, the probability of the given event is 0 and that part of the condition is trivially met. Using the same argument, we may claim that the first of the two remaining probabilities is always larger than the second probability. Consequently, it is sufficient to show that the first probability meets the condition. Moreover, in both examples the trivial likelihood ratio 1 is among the likelihood ratios that participate in the maximization. Consequently, \( \hat{M}_\kappa \geq 1 \) and we need not worry about negative values of the log of this statistic.

The first probability, in the case of sequential testing, may be written as:

\[
\mathbb{P}_n(M_n > e^{g(n)}) = \mathbb{P}_n \left( \max_{n_0 \leq m \leq n_1} e^{\ell_{m-n}} > e^{g(n)} \right) \leq (n_1 - n_0) e^{-g(n)} .
\]

The last inequality follows from an application of Boole’s inequality, together with Markov’s inequality. The expectation \( \mathbb{E}_n(e^{\ell_{m-n}}) \) is equal to 1 since the random variable \( e^{\ell_{m-n}} \) is a likelihood ratio. All the values in the interval \([n_0, n_1]\) are asymptotic to \( x/I \) and the number of values in the interval is proportional to \( x^{\frac{1}{2}} \). As a result, the function \( g(x) = (1 + \epsilon) \log(x) \) will do the job.

For the case of a scanning statistic we have that the number of parameters in \( T \) is asymptotically proportional to \( z^{1/0} \). Arguing as before we get that:

\[
\mathbb{P}_\theta(M_\theta > e^{g(z^2)}) = \mathbb{P}_\theta \left( \max_{\theta \in T} e^{z(Z_\theta - Z_\theta)} > e^{g(z^2)} \right) \leq |T| e^{-g(z^2)} ,
\]
where $|T|$ is the cardinality of the set $T$. Taking $g(z^2) = (4/\alpha + 1 + \epsilon) \log z$ is sufficient.

Turn to Condition IV*, first for the scanning statistic then for sequential testing. In the former case the joint distribution of global term $z(Z_\theta - \theta)$ and the local field $(z(Z_\theta - Z_\theta))$ is multivariate normal. The expectation of the global term under the alternative distribution is $zE_\theta(Z_\theta - \theta) = 0$ and the variance is $z^2 \text{Var}_\theta(Z_\theta) = z^2$. It follows that $\mu = 0$ and $\sigma^2 = 1$. The density of the global term can be approximated by $(2\pi z^2)^{-\frac{1}{2}}$ uniformly in the range $|x| \leq 3(4/\alpha + 1 + \epsilon) \log z$. In order to verify Condition IV* it is sufficient to show that the conditional covariance structure of the local random field, given the global term, converges to the unconditioned covariance structure and that the conditional expectations of the local field converge to the unconditional ones, uniformly in the given range of values of the global term. However,}

$$z^2[\text{Cov}(Z_\theta - Z_\theta, Z_\gamma - Z_\theta|Z_\theta) - \text{Cov}(Z_\theta - Z_\theta, Z_\gamma - Z_\theta)]$$

$$= -z^2[1 - \text{Cov}(Z_\theta, Z_\theta)][1 - \text{Cov}(Z_\gamma, Z_\theta)],$$

which converges to zero, when $z \to \infty$, for all $\theta$ and $\gamma$ in the local region. Likewise,

$$z\{E_\theta[Z_\theta - Z_\theta|z(Z_\theta - \theta) = x] - E_\theta[Z_\theta - Z_\theta]\} = z^2[1 - \text{Cov}(Z_\theta, Z_\theta)]x/z^2,$$

converges to zero, uniformly in $x$, in the restricted range of $x$ values. This completes the validation of the condition for scanning statistics.

Consider sequential testing. In this case the expectation of the global term is $E_n(\ell_n - x) = nI - x$, for $I = \int g(x) \log f(x)/g(x) \, dx$, and the variance is $n\sigma^2$, for $\sigma^2 = \int g(x)(\log f(x)/g(x)) - 1)^2 \, dx$. Since $\kappa = n$ we obtain that $\sigma^2$ is the value that is used in the declaration of Condition IV* and

$$\mu = \lim_{\kappa \to \infty} \frac{E_n(\ell_n)}{\kappa} = \lim_{x \to \infty} \frac{I_n - x}{n\frac{1}{2}},$$

with values that range between $-I^\frac{1}{2}C$ and $I^\frac{1}{2}C$.

The local $\sigma$-algebra $\mathcal{F}_n = \sigma\{X_{n-\tau}, \ldots, X_{n+\tau}\}$ is generated by observations in the range $[n - \tau, n + \tau]$. One may write the global term as a sum of two independent random variables: $\ell_n - x = (\ell_{n-\tau-1} - x) + (\ell_n - \ell_{n-\tau-1})$, the first of which is independent of the local $\sigma$ algebra and the second is measurable with respect to it. We may establish Condition IV* by conditioning on the local $\sigma$-algebra and applying Theorem 5.1 to the independent component $\ell_{n-\tau-1} - x$:

$$P_n(\ell_n \in y + (0, \delta), E) = E_n[P_n(\ell_{n-\tau-1} - x \in y - (\ell_n - \ell_{n-\tau-1}) + (0, \delta)|\hat{F}_n); E],$$

where $y = \xi - \hat{m}_n$, for any $\xi$, $|\xi| \leq 3(1 + \epsilon) \log n$. The asymptotic expectation and variance of $\ell_{n-\tau-1} - x$ are the same as those of $\ell_{n-1} - x$. If the mixture of components is non-lattice then the normal density emerges as the limit, uniformly in $|y| \leq 4(1 + \epsilon) \log n$ over the event $\{\|\ell_n - \ell_{n-\tau-1}\| < \epsilon n^{\frac{1}{2}}\} \cap \{\hat{m}_n \leq (1 + \epsilon) \log n\}$. Proving that the probability of the complementary of this event is $o(n^{-\frac{1}{2}})$ completes the proof. However, for the complementary of the event for $\hat{m}_n$ we have Condition V*. For the complementary of the other event we get from Markov’s inequality that:

$$P_n(\|\ell_n - \ell_{n-\tau-1}\| \geq \epsilon n^{\frac{1}{2}}) \leq \frac{\tau \sigma^2}{(\epsilon n^{\frac{1}{2}} - \tau I)^2}.$$
3.4. APPLICATION OF THE LOCALIZATION THEOREM

as required.

Condition II* is more demanding. We analyze the condition by the identification of an event that contains $A^c_n$ and produce a bound for the event. For both cases we use the fact that $M_n \leq M_n$, $\hat{S}_n \leq S_n$. Consequently, one of the directions that is implied by the use of absolute values in the definition of $A^c_n$ is obtained for free. Still we need to bound the probability of the event

$$A^c_n = \{M_n > \hat{M}_n e^{\epsilon} \} \cup \{\hat{S}_n < (1 - \epsilon)S_n \} ,$$

intersected with an event formulated in terms of the global term and the log of the local maxima. Form the fact that $1 \leq \hat{S}_n$, $1 \leq \hat{S}_n$, and $M_n \leq M_n + S_n - \hat{S}_n$, we get that:

$$\{M_n > \hat{M}_n e^{\epsilon} \} \cup \{\hat{S}_n < (1 - \epsilon)S_n \} \subset \{S_n - \hat{S}_n > \epsilon \}$$

since $\min\{\epsilon - 1, \epsilon/(1 - \epsilon)\} \geq \epsilon$.

For both cases, the statistic $S_n - \hat{S}_n$ is a sum of likelihood ratios. In the case of sequential testing $\hat{S}_n - S_n = \sum \{m : |m - n| > \tau\} \exp(\ell_m - \ell_n)$ and in the case of the scanning statistic $\hat{S}_n - S_n = \sum \{\hat{\theta} : ||\hat{\theta} - \theta|| > z^{-\frac{2}{2}} \} \exp(z(\hat{\theta} - \theta))$. Associate with each parameter value outside the local region a non-negative quantity ($p_m$ in sequential testing and $p_\theta$ is the scanning statistic problem) such that the sum of quantities is bounded by 1. It follows that:

$$A^c_n \subset \bigcup \{\ell_m - \ell_n \geq \log(ep_m)\}$$

in the case of sequential testing and:

$$A^c_n \subset \bigcup \{z(\hat{\theta} - \theta) \geq \log(ep_\theta)\}$$

in the case of scanning for a signal hidden in gaussian white noise. The verification of Condition II* proceeds by the investigation of the probability of the events in the unions when they are intersected with the event $\{\ell_n - \hat{x} \in y + (0, \delta]\}$, $y = \xi - \hat{m}_n$, in the first case and the event $\{z(\hat{\theta} - \theta) + \hat{m}_\theta \in x + (0, \delta]\}$ in the second case.

Consider the problem of sequential testing. Up until this point we were able to get away with making the minimal assumption that the distribution of the increments of the log-likelihood ratio under the alternative distribution are non-lattice and have a finite second moment. For the current condition this will not be enough. Specifically, at this point we would like to employ the Berry-Esseen theorem in order to produce a uniform bound in the local limit approximation. This theorem requires the existence of a third moment. This is probably sufficient. However, assuming a fourth moment, which implies the existence of a third moment, simplifies the proof further. Thus, we make that assumption: $\int g(x)\log(f(x)/g(x)) - I^4 dx = \mu_4 < \infty$.

The alternative expectation of the log-likelihood statistics $\ell_m - \ell_n$ that form the local field is equal to $-I|m - n|$. Let $U_m - U_n = \ell_m - \ell_n + |m - n| \cdot I$ be the centered sum of increments and take $p_m = 0.5 \exp(\tau/2) \exp[|m - n|/I(2)]$.

Look at the case $m > n + \tau$. Intersecting with the event $\{\ell_n - x \in y + (0, \delta]\}$, that is conditionally independent of $U_m - U_n$, given the local $\sigma$-algebra $\mathcal{F}_n$, we
get that:

\[
P_n \left( \{ \ell_n - x \in y + (0, \delta) \} \right)
\leq \frac{(m-n) \mu_4 + 3(m-n)^2 \sigma^4}{\log(0.5e) + \tau/2 + (m-n)(I/2)}^2 \times cn^{-1/4},
\]

for some finite constant \(c\). The last inequality is a result of the uniform local limit produced by the Berry-Esseen theorem and applied to the distribution of \(\ell_n - \tau\). The bound on the factor on the left-hand side is produced by the application of the Markov inequality to \((U_m - U_n)^4\). The expectation of this random variable is bounded by the numerator of the ratio on the left-hand side.

In the case where \(n_0 \leq m < n - \tau\) we may not use independence directly. However, essentially the same proof still works. Denote \(A_m = \{ U_m - U_n \geq \log(\epsilon_p) + (m-n)I \}\). Condition on the value of \(U_n - U_m\) and on the local \(\sigma\)-algebra to obtain:

\[
P_n \left( \{ \ell_n - x \in y + (0, \delta) \} \right)
\leq \frac{(m-n) \mu_4 + 3(m-n)^2 \sigma^4}{\log(0.5e) + \tau/2 + (m-n)(I/2)}^2 \times (n/m)^{1/2} \times cn^{-1/4}.
\]

Summing over all parameters in the range, approximating the sum by an integral, produces the asymptotic inequality:

\[
n^2 P_n \left( A_n^c \cap \{ \ell_n - x \in y + (0, \delta) \} \right) \leq \frac{1}{\tau} \int_1^\infty \frac{2c[1 + 2C/(Ix^2)]+[y\mu_4/\tau+3y^2\sigma^2]}{[\log(0.5e)/\tau + 0.5 + y(I/2)]^4} dy.
\]

The right-hand side is \(O(1/\tau)\). Choosing \(\tau\) large enough finishes the validation of Condition II\(^*\).

Consider the same condition in the Gaussian setting. Here we may compute the probabilities involved in the bound by conditioning on the value the global term \(z(Z_0 - z)\) and using the fact that the density of the global term is bounded by its value at the origin, namely: \(P_\theta(z(Z_0 - z) = x) \leq [2\pi x^2]^{-1/2}\). However, we do intend to treat the bound on the values that \(\hat{m}_0\) may obtain more carefully than before.

Given a value of \(x\). Split the probability between the event \(\{ \hat{m}_0 \leq m \}\) and its complementary to get:

\[
P_\theta \left( A_n^c \cap \{ -m < X_\theta \leq 0, \delta \} \right) \leq P_\theta \left( A_n^c \cap \{ -m < X_\theta \leq \delta \} \right) + E_\theta \left[ P_\theta \left( \{ X_\theta + \hat{m}_0 \in (0, \delta) \} \cap \tilde{\mathcal{F}}_n \right) \right].
\]
where $X_\vartheta = z(Z_\vartheta - z) - x$. The first probability is bounded by:

$$\frac{m + \delta}{z\sqrt{2\pi}} \times \max_{x - m < y \leq x + \delta} P(A_{\bar{\eta}}^n|z(Z_\vartheta - z) = y)$$

and the second term is dominated by:

$$\left\{1 - \sum_{\theta: ||\theta - \vartheta|| \leq \tau z^{-\frac{\delta}{2}}} [1 - \operatorname{Cov}(Z_\vartheta, Z_\vartheta)]^2\right\} - \frac{1}{z\sqrt{2\pi}} P_\vartheta(\hat{m}_\vartheta > m).$$

The term in the curly brackets is the conditional variance of $Z_\vartheta$, given the local $\sigma$-algebra $\mathcal{F}_\vartheta$, and is converging to 1 with the increase in the threshold. By Condition $\text{III}^*$ we obtain that the random variable $\hat{m}_\vartheta$ converges to a finite random variable. Consequently, we can make the product of $z$ and the second term as small as we wish by the selection of a large enough $m$. The overall proof of Condition $\text{II}^*$ will be complete once we produce a uniform bound for the probabilities $P(A_{\bar{\eta}}^n|z(Z_\vartheta - z) = y)$.

The conditional probability for an event in the union that contains $A_{\bar{\eta}}^n$ is a function of the conditional expectation and conditional variance of the increment of the local field that defines the event. The conditional expectation of the local increment, given the global term, is:

$$E_\vartheta[z(Z_\vartheta - Z_\vartheta)|z(Z_\vartheta - z) = y] = -z^2[1 - \operatorname{Cov}(Z_\vartheta, Z_\vartheta)][1 + y/z^2],$$

and the conditional variance, independent of $y$, is:

$$\operatorname{Var}_\vartheta[z(Z_\vartheta - Z_\vartheta)|z(Z_\vartheta - z)] = z^2[1 - \{\operatorname{Cov}(Z_\vartheta, Z_\vartheta)\}^2] < 2z^2[1 - \operatorname{Cov}(Z_\vartheta, Z_\vartheta)].$$

We may bound the conditional probability of a typical event in the union that contains $A_{\bar{\eta}}^n$ by bounding the conditional expectation from below and bounding the conditional variance from above:

$$P_\vartheta(z(Z_\vartheta - Z_\vartheta) \geq \log(\epsilon p_\vartheta)|z(Z_\vartheta - z) = y) \leq 1 - \Phi\left(\frac{\log(\epsilon p_\vartheta) + z^2[1 - \operatorname{Cov}(Z_\vartheta, Z_\vartheta)][1 - \epsilon]}{\sqrt{2z}[1 - \operatorname{Cov}(Z_\vartheta, Z_\vartheta)]^{\frac{1}{2}}}\right),$$

for all $y > -\epsilon z^2$. For each $y$ in the range:

$$P_\vartheta(A_{\bar{\eta}}^n|z(Z_\vartheta - z) = y) \leq \sum_{\theta: ||\theta - \vartheta|| > \tau z^{-\frac{\delta}{2}}} \left\{1 - \Phi\left(\frac{\log(\epsilon p_\vartheta) + z^2[1 - \operatorname{Cov}(Z_\vartheta, Z_\vartheta)][1 - \epsilon]}{\sqrt{2z}[1 - \operatorname{Cov}(Z_\vartheta, Z_\vartheta)]^{\frac{1}{2}}}\right)\right\}.$$

We want to select positive weights $p_\vartheta$ with a sum that is bounded by 1 that permit the converges to 0, as $z \to \infty$, of the right-hand side.

In order to do so we split the bound into two parts, each of which involves a sum over a subset of parameters. One sum is over the region where $||\theta - \vartheta|| > (\log z/z)^{\frac{\delta}{2}}$ and the other is over the region where $\tau z^{-\frac{\delta}{2}} < ||\theta - \vartheta|| \leq (\log z/z)^{\frac{\delta}{2}}$

For the first region we may take $p_\vartheta = 0.5 \cdot |T|^{-1}z^{-\frac{\delta}{2}}$ to be uniform.

A total weight of $1/2$ is assigned uniformly in the first region. The cardinality of $T$ is equal to $|T|z^{\frac{\delta}{2}}$, where $|T| = (t_1 - t_0)(h_1 - h_0)$ is the area of $T$ and $\delta$ is increment used in the discrete approximation. Consequently, the magnitude
of a weight is proportional to $z^{-\frac{c}{2}}$, leading to $\log(p_\theta) \sim -(4/\alpha) \log z$. For all $\theta$ in the first region we have that

$$z^2|1 - \text{Cov}(Z_\theta, Z_\theta)| \gtrsim (\log z)^2,$$

which is more than enough to counteract the effect of the weight and to assure that the sum of the resulting normal probability is small.

Bounding the sum in a ring about $\theta$ requires a more careful examination of the asymptotic expansion of the covariance. A typical parameter value in the second region may be represented as $\theta = \theta + (i \delta z^{-\frac{c}{2}}, j \delta z^{-\frac{c}{2}})$, for $\tau < |i|$ and $\tau < |j|$. One can find a positive constant $c_0$ such that

$$z^2|1 - \text{Cov}(Z_\theta, Z_\theta)| \geq c_0 z^2 ||\theta - \theta||^\alpha = c_0 \delta^{\frac{c}{2}} (|i|^2 + |j|^2)^{\frac{c}{2}},$$

Replacing $c_0$ by a large enough finite constant $c_1$ produces an upper bound. If we choose

$$p_\theta = \epsilon^{-1} \exp\{-(c_0/2)\delta^{\frac{c}{2}} (|i|^2 + |j|^2)^{\frac{c}{2}}\}$$

we get that, after letting $z \to \infty$,

$$\frac{\log(p_\theta)}{\sqrt{2} z |1 - \text{Cov}(Z_\theta, Z_\theta)|} \leq \frac{c_0 \delta^{\frac{c}{2}}}{8c_1} (|i|^2 + |j|^2)^{\frac{c}{2}} = c \cdot (|i|^2 + |j|^2)^{\frac{c}{2}}.$$

It follows that the sum of probabilities in the second region is asymptotically bounded by

$$\sum_{|i| > \tau} \sum_{|j| > \tau} \exp\{-(c^2/2) \cdot (|i|^2 + |j|^2)^{\frac{c}{2}}\},$$

which converges to zero when $\tau \to \infty$. This completes the validation of Condition I* in the gaussian setting.

The remaining condition to check is Condition III* that states the convergence of the expected ratio of the maximum over the sum and identifies the term that this limit approximates.

The ratio is a bounded functional of the local field. Therefore, in order to identify the limit of the expectation it is sufficient to identify the limit in distribution of the local field, as the threshold goes to infinity. The term that is approximated by the limit is obtained by increasing the range of the local field. In the current situation this corresponds to letting $\tau$ go to infinity.

In the case of sequential testing the local field involves partial sums. For $m > n$, where the distribution of $X_i$ is determined by the null density $f$, the increments of the partial sum is of the form $\log \{g(X_i)/f(X_i)\}$. For $m < n$, on the other hand, the distribution of $X_i$ is determined by the density $g$ and the increments of the partial sum is of the form $\log \{g(X_i)/f(X_i)\} = \log \{f(X_i)/g(X_i)\}$.

The distribution of the local field does not changed with the change of the threshold. Therefore, the description fits the limit local field. The distribution is also independent of $n$, as long as $n_0 + \tau \leq n \leq n_1 - \tau$. Consequently, we may describe the term $E_n[M_n/S_n]$ by shifting the origin to be equal to $n$. Accordingly, the observations for $-\tau \leq i < 0$ have density $g$ and the observations for $0 < i \leq \tau$ have density $f$. The partial sum for $m = 0$ is equal to 0. Partial sums are formed in one way for negative indices and in a different way for positive indices, covering the whole range $-\tau \leq m \leq \tau$.

The resulting partial sums are exponentiated and then either summed or maximized. The ratio between the maximum and the sum is the random variable...
3.4. APPLICATION OF THE LOCALIZATION THEOREM

that enters into the expectation. The term $E[M/S]$ is the outcome of the same process for $-\infty < m < \infty$. The fact that we have convergence as a function of $\tau$ results from the application of the relevant part of the proof that was used in order to establish Condition II$^\ast$.

In the gaussian setting the bounded random variables $\tilde{M}/\tilde{S}$ is produced by a finite collection of random variables with a multi-normal distribution. This random variable converges in distribution, as $z \to \infty$, to the random variable that is produced as a function of the limit gaussian distribution of the finite collection. This limit gaussian distribution is determined by the limit expectation and the limit covariance structure. However, since

$$E_{\theta}[z(Z_{\theta} - Z_{\theta})] = -z^2[1 - \text{Cov}(Z_{\theta}, Z_{\theta})]$$

and

$$\text{Cov}[z(Z_{\theta} - Z_{\theta}), z(Z_{\eta} - Z_{\eta})] = z^2[1 + \text{Cov}(Z_{\theta}, Z_{\eta}) - \text{Cov}(Z_{\theta}, Z_{\theta}) - \text{Cov}(Z_{\eta}, Z_{\theta})]$$

it is clear that these limits emerge from expanding the covariance of the original field about $\theta$. This expansion is of one type when a smooth kernel is used and of a different type when the kernel is an indicator of an interval.

Consider the case of a smooth kernel $g$. We showed that in the case where $\text{Var}(Z_{\theta}) \equiv 1$ then the gradient of covariance, evaluated at $\vartheta = \theta$, is the zero vector. Consequently, the expansion is determined by higher order derivatives. The fact that the variance is constant as a function of the parameters implies a relation between the hessian and the gradient of the random field:

$$0 = \left( \frac{\partial}{\partial \theta} \right)^2 \text{Cov}(Z_{\theta}, Z_{\theta}) = 2 \frac{\partial}{\partial \theta} \text{Cov}(\tilde{Z}_{\theta}, Z_{\theta}) = 2[\text{Cov}(\tilde{Z}_{\theta}, Z_{\theta}) + \text{Cov}(\tilde{Z}_{\theta}, \tilde{Z}_{\theta})]$$

Consequently,

$$\left( \frac{\partial}{\partial \theta} \right)^2 \text{Cov}(Z_{\theta}, Z_{\theta}) \bigg|_{\vartheta = \theta} = \text{Cov}(\tilde{Z}_{\theta}, Z_{\theta}) = -\text{Cov}(\tilde{Z}_{\theta}, \tilde{Z}_{\theta}) = -\Sigma_{\theta}$$

where $\Sigma_{\theta}$ is the variance-covariance matrix of the gradient that was identified in Chapter 2: $\Sigma_{\theta} = (\bar{g}_{\theta}, g_{\theta})/||g||^2$. In the special case where $g(x) = \exp\{-x^2/2\}$ this matrix became:

$$\Sigma_{\theta} = \left( \begin{array}{cc} h^{-3} & h^{-4} \\ h^{-4} & h^{-5} + h^{-1}/4 \end{array} \right).$$

Consider $\vartheta = \theta + (i \delta z^{-1}, j \delta z^{-1}) = \theta + \delta z^{-1} \iota$, where $\iota$ is a point on the two-dimensional grip of integers. The limit of the expectation is given by:

$$\lim_{z \to \infty} E_{\theta}[z(Z_{\theta} - Z_{\theta})] = \lim_{z \to \infty} z^2/2 (\vartheta - \theta)^{\prime} \text{Cov}(\tilde{Z}_{\theta}, Z_{\theta})(\vartheta - \theta) = -\delta^2/2 \iota \Sigma_{\theta} \iota$$

and the limit of the covariance between two elements associated with $\vartheta$ and $\eta$:

where $\eta = \theta + \delta z^{-1} \nu$, is:

$$\lim_{z \to \infty} \text{Cov}[z(Z_{\theta} - Z_{\theta}), z(Z_{\eta} - Z_{\eta})]$$

$$= \lim_{z \to \infty} z^2/2 \langle (\vartheta - \theta)^{\prime} \text{Cov}(\tilde{Z}_{\theta}, Z_{\theta})(\vartheta - \theta) + (\eta - \theta)^{\prime} \text{Cov}(\tilde{Z}_{\theta}, Z_{\theta})(\eta - \theta)$$

$$- (\vartheta - \eta)^{\prime} \text{Cov}(\tilde{Z}_{\eta}, Z_{\eta})(\vartheta - \eta) \rangle$$

$$= \delta^2/2 [-\iota^{\prime} \Sigma_{\theta} \iota - \nu^{\prime} \Sigma_{\theta} \nu + (\iota - \nu)^{\prime} \Sigma_{\theta}(\iota - \nu)] = \delta^2/2 \iota^{\prime} \Sigma_{\theta} \iota.$$
for \( \tau \) and \( \nu \) two points on the grid of integers. The evaluation of the limit exploits the fact that \( \sum_\nu \to \sum_\theta \). Recall that the expectation of \( \bar{Z}_n \) is zero and the variance-covariance matrix is \( \Sigma_\theta \). We may conclude that the process 
\[ \{ \delta \bar{Z}_n - (\delta^2/2)\tau\Sigma_\theta \tau : \tau = (i,j), |i| \leq \tau, |j| \leq \tau \} , \]
defined over a square in the integer lattice.

The functional that measures the contribution of the local field, the expectation of the maximum of the exponentiated local field divided by its sum, converges to a functional that is defined in the same way but with respect to the limit local field:
\[
E[\hat{\mathcal{M}}/\hat{S}] = E\left[ \max_{\|i\| \leq \tau} e^{\delta \bar{Z}_n - (\delta^2/2)\tau\Sigma_\theta \tau} \right],
\]
where the norm that is being used here is the \( \|(x, y)\| = |x| \vee |y| \). However,
\[
\delta \bar{Z}_n - (\delta^2/2)\tau\Sigma_\theta \tau = \frac{\delta^2}{2}(\tau - U)^\tau\Sigma_\theta (\tau - U) + \frac{\delta}{2} \bar{Z}_n \Sigma_\theta^{-1} \bar{Z}_n ,
\]
for \( U = \Sigma_\theta^{-1} \bar{Z}_n \). Consequently,
\[
E[\hat{\mathcal{M}}/\hat{S}] = E\left[ \max_{\|i\| \leq \tau} e^{-\frac{\delta^2}{2}(\tau - U)^\tau\Sigma_\theta (\tau - U)} \right],
\]
which converges to:
\[
E[\mathcal{M}/S] = E\left[ \max_{\tau} e^{-\frac{\delta^2}{2}(\tau - U)^\tau\Sigma_\theta (\tau - U)} \right],
\]
with maximization and summation extending over the entire integer grid.

When expanding the covariance function in the case where the random field is not smooth one may see that the first order terms do not vanish. In particular, in the scanning statistic example with an indicator serving as a kernel and the parameters are the end-points of an interval, we obtained that:
\[
1 - \text{Cov}(\bar{Z}_n, Z_\theta) \approx \frac{1}{2h} |\bar{\theta}_1 - \theta_1| + \frac{1}{2h} |\bar{\theta}_2 - \theta_2| = \frac{\delta |i|}{2hz^2} + \frac{\delta |j|}{2hz^2} ,
\]
for \( \theta = \bar{\theta} + (i\delta z^{-2}, j\delta z^{-2}) \). It follows that:
\[
\lim_{z \to \infty} E_\theta [z(Z_\theta - Z_\theta)] = \frac{\delta |i|}{2h} - \frac{\delta |j|}{2h} .
\]
If we consider the asymptotic covariance between the term associated with \( \bar{\theta} \) and the term associated with \( \eta = \theta + (i\delta z^{-2}, j\delta z^{-2}) \) we get:
\[
\lim_{z \to \infty} \text{Cov}[z(Z_\theta - Z_\theta), z(Z_\eta - Z_\eta)]
= \frac{\delta}{2h} |i| + |j'| + |j'| - |i' - i| - |j - j'| + \frac{\delta}{h} |i| \wedge |i'| + |j| \wedge |j'| ,
\]
if \( i \) and \( i' \) (\( j \) and \( j' \)) share the same sign. If \( i \) and \( i' \) (\( j \) and \( j' \)) don’t share the same sign then they make zero contribution to the covariance. One may associate the limit distribution of the random field \( \{ z(Z_\theta - Z_\vartheta) \} \) with a sum of two independent processes: \( \{ W_1(i\delta/h) + W_2(j\delta/h) : |i| \leq \tau, |j| \leq \tau \} \), where \( W_1 \) and \( W_2 \) are two independent copies of the process \( W(t) = B_t - |t|/2 \). The process \( B_t \) is the double-ended Brownian motion which originates at 0 and evolves independently to the right and to the left of the origin.

The functional that measures the contribution of the local field converges in this case, due to the independence between the two processes, to:

\[
E[\hat{M}/\hat{S}] = E\left[ \max_{|i|,|j| \leq \tau} e^{W_1(i\delta/h) + W_2(j\delta/h)} \sum_{|i| \leq \tau} \sum_{|j| \leq \tau} e^{W_1(i\delta/h) + W_2(j\delta/h)} \right]^2.
\]

and when we send \( \tau \) to infinity we get:

\[
E[M/S] = \left\{ E\left[ \max_{|i| \leq \tau} e^{W(i\delta/h)} \sum_{|i| \leq \tau} e^{W(i\delta/h)} \right] \right\}^2.
\]

with maximization and summation extending to the entire collection of negative and positive integers.

### 3.5 Integration

The direction of analysis so far involved the exploration of finer and finer details until we were able to obtain an approximation of a relevant quantity associated with each given value of the parameter. Now we want to put things back together again in order to obtain the approximation of the tail probability of the extreme of the random field, which was our original motivation for the analysis.

The probability in question was presented as a sum over all parameter values of relevant quantities. The localization theorem produces an approximation for each quantity. In the final step we propose to replace the quantities in the sum by their approximated values and evaluate the resulting summation. It may be useful to think of the localization theorem as producing point-wise convergence for each value of the parameter. The integration step translates this point-wise convergence to convergence of integrals. Concepts of uniform integrability or theorems like the dominated convergence theorem may be useful in order to establish this translation.

Start with the problem of sequential testing. As a result of measure transformation we obtained the representation:

\[
P\left( \max_{\ell_n \in T} \ell_n \geq x \right) = e^{-x} \sum_{n=n_0}^{n_1} E_n \left( \frac{M_n}{S_n} e^{-[\tilde{\ell}_n + m_n]} ; \tilde{\ell}_n + m_n \geq 0 \right).
\]

The localization theorem informed us that

\[
E_n \left( \frac{M_n}{S_n} e^{-[\tilde{\ell}_n + m_n]} ; \tilde{\ell}_n + m_n \geq 0 \right) \approx \frac{1}{\sqrt{n\sigma^2}} \phi \left( \frac{n - x}{\sqrt{n\sigma^2}} \right) E[M/S],
\]

for some constant \( E[M/S] \) that did not depend on \( n \), for \( n_0 + \tau \leq n \leq n_1 - \tau \). The same type of convergence holds for parameter values in the vicinity of the
boundary points but with \(E[\mathcal{M}/S]\) replaced by a similar term that involves also the expected ratio of maximization and summation. However, the range of parameter values over which maximization and summation take place is partially truncated in either the positive or the negative direction.

If indeed the point-wise convergence implies convergence of the sum we may conclude that:

\[
P\left(\max_{n \in T} \ell_n \geq x\right) \approx e^{-x} E[\mathcal{M}/S] \sum_{n=x/1-c\sqrt{n}}^{x/1+C\sqrt{n}} \frac{1}{\sqrt{n} \sigma} \phi\left(\frac{n-x/1}{\sqrt{n} \sigma / 1}\right)
\]

\[
\approx e^{-x} \frac{E[\mathcal{M}/S]}{I} \left[2\Phi\left(C I^{3/2} \sigma^{-1}\right) - 1\right],
\]

with the second approximation resulting from the approximation of the sum by an integral, using in the process the fact that \(n/x \approx I\).

The original problem dealt with the stopping time \(N_x\) and the probability that it obtains a finite value. This probability, the significance level of the test results from allowing \(C \to \infty\), which produces:

\[
\lim_{x \to \infty} e^x P(N_x < \infty) = \lim_{C \to \infty} e^x P\left(\max_{n \in T} \ell_n \geq x\right) = \frac{1}{I} E[\mathcal{M}/S].
\]

In Chapter 2 we obtained, applying different means, an approximation of the same probability. Comparing the terms of that approximation that is given in (2.3) with the current one, recalling that \(I = E_g(\ell_1)\), teaches us that:

\[
E[\mathcal{M}/S] = \exp\left\{-\sum_{n=1}^{\infty} n^{-1}[P_g(\ell_n \leq 0) + P(\ell_n > 0)]\right\},
\]

for the ratio between the maximized likelihood ratios and the sum of these likelihood ratios that emerges in the context of the current problem. In particular, if we deal with the problem of testing hypothesis for the normal mean being equal to 0 or to \(\mu\), we get the relation: \(E[\mathcal{M}/S] = (\mu^2/2) \cdot \nu(\mu)\). Interestingly enough, the log-likelihood ratios that are used for this testing problem can be represented in the form \(\ell_n = \mu \sum_{i=1}^{n} X_i - n \mu^2/2 = B(\mu^2 n) - (\mu^2 n)/2 = W(\mu^2 n)\), where equality means equality in distribution. The conclusion we draw is that

\[
E\left[\max_{|u|} e^{W(\mu^2)} \right] = (\mu^2/2) \cdot \nu(\mu).
\]

This is nice and well, but we still need to show that point-wise convergence implies convergence of the sum. For that we may apply, for example, Berry-Esseen theorem once more and obtain a universal constant such that

\[
E_n\left(\frac{M_n}{S_n} e^{-|\ell_n+m_n|}; |\ell_n+m_n| \geq 0\right) \leq cn^{-\frac{1}{2}},
\]

which implies convergence via the dominated convergence theorem.

Now is the turn of scanning statistics. In the case of a smooth kernel we obtained that:

\[
P(\max_{\theta \in T} Z_\theta \geq z) \approx \frac{e^{-\frac{1}{2} z^2}}{\sqrt{2\pi}} \sum_{\theta \in T} \sum_{\ell \in \mathcal{T}} \left[\max_{|u|} e^{\frac{\ell - u}{2}(\ell - u)^\Sigma_u(\ell - u)} \right],
\]
and in the case of a indicator of an interval serving as a as a kernel we got:

\[
P(\max_{\theta \in T} Z_\theta \geq z) \approx \frac{e^{-\frac{1}{2}z^2}}{z\sqrt{2\pi}} \sum_{\theta \in T} \left\{ \mathbb{E} \left[ \max_{|i|} e^{W(i\delta/h)} \right] \right\}^2.
\]

Not only \( \mathbb{E}[M/S] \) but also the parametrization and the discrete collection of parameters \( T \) is different for the two cases. For the smooth case we used a grid with span \( \delta z^{-1} \) and for the other case we used a grid with span \( \delta z^{-2} \), defined for transformed parameters.

Approximating summation by integration we get for the smooth case that:

\[
\lim_{z \to \infty} \frac{z}{z^2}e^{\frac{1}{2}z^2}P(\max_{\theta \in T} Z_\theta \geq z) \approx \frac{1}{\delta^2\sqrt{2\pi}} \int_T \mathbb{E} \left[ \max_{|i|} e^{-\frac{\delta^2}{2}(i-U)'\Sigma(i-U)} \right] d\theta,
\]

where now \( T \) corresponds to the original continuous collection of parameters \( T = [t_0, t_1] \times [h_0, h_1] \). If we allow \( \delta \to 0 \) we will get that the maximum on the right-hand side converges to 1 and the sum, upon multiplication by \( \delta^2 \), converges to a gaussian integral. The result in the smooth case is the approximation:

\[
\lim_{z \to \infty} z^{-1}e^{\frac{1}{2}z^2}P(\max_{\theta \in T} Z_\theta \geq z) = (2\pi)^{-3/2} \int_T |\Sigma|^{1/2} d\theta,
\]

that is in agreement with (2.4).

In the non-smooth case we exploit the fact that the transformation of the parameters is measure-preserving and the representation of the expectation of the max over the sum in terms of the overshoot function \( \nu \) of the normal random walk. The approximation of the sum by an integral becomes:

\[
\lim_{z \to \infty} z^{-1}e^{\frac{1}{2}z^2}P(\max_{\theta \in T} Z_\theta \geq z) = (2\pi)^{-3/2} \int_T |\Sigma|^{1/2} d\theta.
\]

When \( \delta \to 0 \) the \( \nu \) function converges to 1. Therefore,

\[
\lim_{z \to \infty} z^{-3}e^{\frac{1}{2}z^2}P(\max_{\theta \in T} Z_\theta \geq z) = (2\pi)^{-\frac{1}{2}} \cdot (0.5)^2 (t_1 - t_0)(1/h_0 - 1/h_1),
\]

which is identical to the approximation in (2.9) that resulted from the application of the double-sum method.

The last point to check is uniform integrability. This follows from using the bound \( z^{-1}(2\pi)^{-\frac{1}{2}} \) on the density of \( z(Z_\theta - z) \) in both versions of gaussian scanning statistics.
Chapter 4

From the Local to the Global
4.1 Introduction

In Chapter 3 we used a method for the approximation of the tail probability of extremes in a random field. The method was applied in order to analyze two examples: an example associated with sequential testing and an example associated with scanning statistics. This method is meaningful in settings of large deviation where the probability in question converges to zero. The approximation produced by the method involved the summation, over the set of parameters, of terms. The terms were a product of a large deviation factor and two other factors that correspond to lower-order contributions. Technically, the statement of large deviation for the probability may be translated to a relation between the the large deviation factors, one for each term in the sum, which are exponentially or super-exponentially small as a function of the threshold, and the integration of such factors, which accumulates the small terms together. The setting that is assumed in the approximation is such that, even after the accumulation of all terms, the sum still converges to zero.

In the next section we deal with the case where the parameter space is large enough, in comparison to the threshold, to make the probability of crossing the threshold converge to a positive number. We approach this problem by the introduction of the poisson approximation as a vehicle to integrate small probabilities of weakly dependent events. This poisson approximation resembles in spirit the double-sum argument in the double-sum method.

Occasionally, one may be interested in settings where the probability of crossing the threshold is 1 and in functionals other than the probability. For example, in sequential change-point detection, where a change-point detection monitoring statistic is followed over time, one may be interested in the expected duration of time that elapsed until the first crossing of the detection threshold. Subject to a poisson approximation, an approximation for the expectation will emerge as a corollary of proving uniform integrability. We demonstrate this claim in Section 4.3 where we present the cusum method for change-point detection, an adaptation of the sequential probability ratio test to the setting of change-point detection.

4.2 Poisson approximation of probabilities

The example of a scanning statistic involved the examination of the region $T = [t_0, t_1] \times [h_0, h_1]$, with $[t_0, t_1]$ an interval of the real line where the signal may be located and $[h_0, h_1]$ a range of parameter values that characterizes the width of a signal. A kernel function was used in order to construct the monitoring statistic. In the case where we employed the kernel $g(x) = \exp\{-x^2/2\}$ we obtained the approximation:

$$\Pr(\sup_{\theta \in T} Z_\theta \geq z) \approx ze^{-\frac{1}{2}z^2}(2\pi)^{-3/2} \cdot 0.5 \cdot (t_1 - t_0)(1/h_0 - 1/h_1)$$

for the probability of crossing the threshold $z$ within the region $T$. If the kernel $g(x)$ is the indicator of the interval $[-0.5, 0.5]$ then the resulting approximation was:

$$\Pr(\max_{\theta \in T} Z_\theta \geq z) \approx z^3e^{-\frac{1}{2}z^2}(2\pi)^{-1/2} \cdot (0.5)^2 \cdot (t_1 - t_0)(1/h_0 - 1/h_1).$$
Both approximations were constructed having in mind the situation where the parameter set is kept fixed and the threshold \( z \) is allowed to go to infinity. If, as it may be the case, the interval \([t_0, t_1]\) is very large one may question the appropriateness of an asymptotic expansion that considers the length of the interval to a fixed quantity. More meaningful results may be obtained by imagining that the length of the interval is diverging to infinity together with the threshold. Examination of the proof that we used reviles that the approximation is still appropriate when the length of the interval is diverging to infinity in a modest rate that does not prevent an overall convergence to zero of the approximation. However, if the divergence is fast enough, leading to a non-vanishing probability of crossing the threshold, then the arguments that were used in order to establish the approximation may no longer be valid. Instead, need a new method to deal with this new problem.

In a way, the issue is not a new one and we have already encountered a similar problem when we discussed the double-sum method. Specifically, in the double-sum method, in order to analyze the probability of crossing the threshold within a fixed region \( T \) by a statistic that employs the indicator of an interval as the kernel, we subdivided the region into smaller subregions of the form:

\[
T_{ij} = T_0 = \{ \vartheta = (\vartheta_1, \vartheta_2) : 0 \leq \vartheta_1 - \theta_1 \leq \tau/z^2, 0 \leq \vartheta_2 - \theta_2 \leq \tau/z^2 \},
\]

for \( \theta = \theta_{ij} \) in a two-dimensional grid with span \( \tau/z^2 \). We were able to produce an approximation for the probability of crossing in each such subregion. This approximation was extended to the entire region via the double-sum argument that established the fact that the sum, over the grid, of the probabilities of crossing within a subinterval is an approximation to the probability of the union, i.e. the probability of crossing in \( T \).

In the current problem we also have a method for dealing with relatively small subregions and we want to extend the outcomes that are obtained for smaller subregions to a larger region. Using the same line of argument, one may propose to subdivide the increasing region \( T = [t_0, t_1] \times [h_0, h_1] \) into disjoint subregions of fixed or modestly increasing volume. The number of such subregions is increasing fast with the threshold but not the volume of each sub-region. Since we have a valid approximation for the probability associated with each subregion we may be able to use the double-sum argument in order to produce an approximation that is based on the sum of approximations of the subregions probabilities. Indeed, this approach may be put to work in the current problem. However, in order to make it work we need a slightly more sophisticated methodology, the methodology of the poisson approximation.

The poisson random variable is associated with the counting of the number of occurrences of rare events. The distribution of the count is characterized by \( \lambda \), the expected number of occurrences. Hence, if \( W \) is the total number of occurrences and is poisson then

\[
P(W = k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots
\]

In particular, \( P(W = 0) = e^{-\lambda} \) is the probability of no occurrence and \( P(W > 0) = 1 - e^{-\lambda} \) is the probability of at least one occurrence.

In order to associate the poisson approximation with the problem at hand we need to specify counts. Consider subregions of the form \( T_i = [t_0 + (i-1)m, t_0 + \]
im] × [h₀, h₁], for i = 1, 2, . . . , (t₁ − t₂)/m, and let Xᵢ be the indicator of an event that at least one crossing of the threshold occurred within the subregion Tᵢ:

\[ \{ Xᵢ = 1 \} = \left\{ \sup_{θ ∈ Tᵢ} θ ≥ z \right\}. \] (4.1)

The sum of indicators by \( \hat{W} = \sum_{i=1}^{(t₁−t₂)/m} Xᵢ \) is a count. This sum is equal to 0 if, and only if, all indicators in the sum are equal to zero. It follows that:

\[ \left\{ \sup_{θ ∈ T} θ ≥ z \right\} = \left\{ \max_{1 ≤ i ≤ (t₁−t₀)/m} Xᵢ = 1 \right\} = \{ \hat{W} > 0 \}. \]

Thus, we can assess the probability of crossing the threshold by considering the asymptotic distribution of \( \hat{W} \). We proceed by approximating this distribution.

The problem that involves an interval kernel is slightly more simple to analyze so we approach it first. Notice that the probability of the event indicated by \( Xᵢ \) is proportional to \( mz^3e^{-1/2z^2} \), thus converging to 0 when \( m \) is sub-exponential. There are \((t₁−t₂)/m\) such events. If the length of the scanning region, \( t₂ − t₁ \), is asymptotically proportional to \( z^{-3}e^{1/2z^2} \) then the sum of probabilities converges to a constant:

\[ \lambda = \lim_{z → ∞} (t₁ − t₀)z^3e^{-1/2z^2}(2π)^{-1/2} \cdot (0.5)^2 \cdot (1/h₀ − 1/h₁), \]

for \( 0 < λ < ∞ \).

The limit \( λ \) is the asymptotic expectation of \( \hat{W} \). The standard poisson approximation of the binomial distribution states that if the event that are indicated by the \( Xᵢ \)’s were independent then the distribution of \( \hat{W} \) converges to the poisson distribution with expectation \( λ \). Unfortunately, the events are not quite independent so the standard argument is not sufficient. However, extensions of the theorem to settings that involve local dependencies between events do exist and enable us to obtain the poisson limit nonetheless.

A very useful formulation of a general poisson limit theorem is given in [2]. This theorem considers a countable collection of zero-one indicator variables \( Xᵢ \) indexed by the elements of \( I \). The total variation distance between the sum of indicator variables and the poisson distribution, with the same expectation as the sum, is bounded in the theorem by a sum of three terms. These three terms are constructed in relation to “neighborhoods of dependence” \( Iᵢ \), associated with each element in the sum. The first of the three terms measures the total size of these neighborhoods of dependence and is denoted \( b₁ \). The second term measures the strength of dependence between the elements within neighborhoods and is denoted \( b₂ \). The third term measures the total dependence between each element and elements outside its neighborhood of dependence. It is denoted \( b₃ \). Specifically, the terms are:

\[ b₁ = \sum_{i ∈ I} \sum_{j ∈ I \setminus \{i\}} P(Xᵢ = 1)P(Xⱼ = 1) \]
\[ b₂ = \sum_{i ∈ I} \sum_{j ∈ I \setminus \{i\}} P(Xᵢ = 1, Xⱼ = 1) \]
\[ b₃ = \sum_{i ∈ I} E[|E(Xᵢ|σ{Xⱼ : j ∉ Iᵢ}) − E(Xᵢ)|] \]
From the statement of the theorem one may conclude that:

$$|P(W > 0) - (1 - e^{-\lambda})| \leq b_1 + b_2 + b_3 + |e^{-E(W)} - e^{-\lambda}|.$$ 

Consequently, a demonstration that the three terms \(b_1, b_2,\) and \(b_3\) can be constructed in a way that make them converge to 0, as \(z \to \infty,\) will produce the approximation:

$$P(\max_{\theta \in T} Z_\theta \geq z) \approx 1 - \exp \left\{ - z^3 e^{-\frac{1}{2} z^2} (2\pi)^{-1/2} \cdot (0.5)^2 \cdot (t_1 - t_0)(1/h_0 - 1/h_1) \right\}.$$ 

(4.2)

This approximation covers the situations where the probability does not converge to zero. If the probability does converge to zero then this approximation and the original one are asymptotically equivalent.

In order to carry out the construction let \(m > h_1/2\) and define the neighborhood of dependence of the indicator \(X_i\) to be \(I_i = \{i - 1, i, i + 1\}\) (with obvious modifications for \(i = 1\) and \(i = (t_1 - t_0)/m\)). Observe that \(X_i\) is a function of the gaussian white noise in the range \([t_0 + (i - 1)m - h_1/2, t_0 + im + h_1/2]\). It follows that \(X_i\) is independent of \(\{X_j : j \notin I_i\}\) and, as a result, \(b_3 = 0\).

All the probabilities that are involved in the computation of \(b_1\) are equal to each other. Therefore,

$$b_1 = [2(t_0 - t_1)/m - 4]P(X_1 = 1))^2 + 2P(X_1 = 1),$$

the last summand being associated with the two edge elements. It follows that \(b_1\) is asymptotic to \([2\lambda + 2]P(X_1 = 1),\) which converges to zero since \(P(X_1 = 1)\) does.

Likewise, for \(b_2\) we have:

$$b_2 = [2(t_0 - t_1)/m - 2]P(X_1 = 1, X_2 = 1).$$

Redefine parameter subregions:

$$T_1 = [t_0, t_0 + m - h_1/2] \times [h_0, h_1],$$
$$T_2 = [t_0 + m - h_1/2, t_0 + m + h_1/2] \times [h_0, h_1],$$
$$T_3 = [t_0 + m + h_1/2, t_0 + 2m] \times [h_0, h_1],$$

and let \(Y_i,\) \(i = 1, 2, 3,\) defined via \(Y_i = 1 = \{\max_{\theta \in T_i} Z_\theta \geq z\},\) be the indicators of crossing the threshold in the appropriate subregions. Notice that the indicators \(Y_1\) and \(Y_3\) are independent of each other and they share the same distribution. We use the fact that unless crossing occurs in a shared subregion it must simultaneously occur in two disjoint subregions in order to have double crossing. As a result obtain the upper bound \(X_1 \cdot X_2 \leq Y_2 + Y_1 \cdot Y_3\) and claim that:

$$P(X_2 = 1, X_3 = 1) \leq P(Y_2 = 1) + P(Y_1 = 1))^2 \leq P(Y_2 = 1) + (P(X_1 = 1))^2.$$ 

The probability \(P(Y_2 = 1)\) is proportional to \(h_1 z^3 e^{-\frac{1}{2} z^2}.\) Consequently, \(b_2\) is asymptotically bounded by \(2\lambda[h_1/m + P(X_1 = 1)].\) Hence, \(b_2\) converges to zero whenever \(m \to \infty.\) This completes the proof of the validity of approximation (4.2).
For the case of the smooth kernel \( g(x) = \exp(-x^2/2) \) we would like to establish the generalizing approximation:

\[
P(\max_{\theta \in T} Z_{\theta} \geq z) \approx 1 - \exp \left\{ -z e^{-\frac{1}{4}z^2}(2\pi)^{-3/2} \cdot 0.5 \cdot (t_1 - t_0)(1/h_0 - 1/h_1) \right\} . \quad (4.3)
\]

The difficulty is that in this case any pair of elements of the field, \( Z_\theta \) and \( Z_{\theta'} \), are correlated regardless of how far \( \theta \) is from \( \theta' \). Consequently, \( b_1 \), the ugliest of the three terms that bound the error in the poisson approximation, is not equal to zero and cannot be ignored.

Instead of analyzing the term \( b_3 \) we will use a truncation argument in order to construct a field \( \{ \tilde{Z}_\theta : \theta \in T \} \) that approximates the original field and yet has the property that remote elements are independent of each other.

Set the smooth kernel

\[
\tilde{g}(x) = \begin{cases} 
\exp(-x^2/2) & |x| \leq y , \\
\exp(-x^2/2) f(|x| - y) & |x| > y , 
\end{cases}
\]

where \( f \) is a smooth and monotone decreasing function that is equal to 1 at zero and equal to 0 for all values larger than 1. Let \( \tilde{Z}_\theta \) be the standardized statistics that is constructed with \( \tilde{g} \) as the kernel. Since the kernel is smooth the formulae \( (4.3) \) for smooth gaussian field apply. Note that for the approximating field any two elements with locations are more than \( 2h_1(y+1) \) apart are independent of each other.

The approximating field \( \tilde{Z}_t \) is related to the original field via:

\[
Z_\theta - \frac{\|g\|}{\|\tilde{g}\|} \tilde{Z}_\theta = \frac{1}{h^2 \|g\|} \int_{\{|x-t|>h\}} [1 - f(|x-t|/h-y)] e^{-\frac{(x-t)^2}{2h^2}} dB_x = \Delta_\theta .
\]

Consequently, we may bound the probability that we seek to approximate both from above and from below using the approximating field and the associated field of discrepancies:

\[
P(\max_{\theta \in T} Z_{\theta} \geq z) \leq P(\max_{\theta \in T} \tilde{Z}_{\theta} \geq \frac{\|g\|}{\|\tilde{g}\|} \frac{z - \epsilon}{z}) + P(\max_{\theta \in T} |\Delta_\theta| \geq \frac{\epsilon}{z}) ,
\]

\[
P(\max_{\theta \in T} Z_{\theta} \geq z) \geq P(\max_{\theta \in T} \tilde{Z}_{\theta} \geq \frac{\|g\|}{\|\tilde{g}\|} \frac{z + \epsilon}{z}) - P(\max_{\theta \in T} |\Delta_\theta| \geq \frac{\epsilon}{z}) .
\]

We proceed by showing that the probabilities associated with \( \tilde{Z}_\theta \) may be approximated using \( (4.3) \) and the probabilities associated with the discrepancies \( \Delta_\theta \) are vanishingly small.

For the probabilities that are associated with \( \tilde{Z}_\theta \) we may use the poisson approximation. In particular, if \( m \) and \( y \) are selected so that \( y/m \to 0 \) then the term \( b_3 \) may be eliminated. The approximations will (almost) coincide with \( (4.3) \) if \( \|g\|/\|\tilde{g}\| = 1 + o(z^{-1}) \) and if the matrix \( |\Sigma_\theta| \), the determinant of the matrix of inner products of the components of the gradient of \( \tilde{g} \), converges to \( |\Sigma_\theta| = h^{-2}/4 \). However,

\[
0 \leq \|g\|^2 - \|\tilde{g}\|^2 = 2 \int_y^{\infty} \left[ 1 - f(x-y) \right]^2 e^{-x^2} dx \leq \frac{\sqrt{2}}{y} e^{-y^2} .
\]

Therefore, provided that \( (\log z)^{1/2}/y \to 0 \) we have the required rate of convergence. A similar argument, applied to the gradient with respect to \( \theta \) of \( \tilde{g}_\theta \), will show the convergence of the determinant whenever \( y \to \infty \).
Let

$$\sigma^2 = \max_{\theta \in T} \text{Var}(\Delta_\theta) = \frac{2}{\|g\|^2} \int_y^\infty (1 - f(x - y))^2 e^{-x^2} \, dx \leq \frac{2}{\sqrt{\pi}y} e^{-y^2}.$$ 

We use Fernique’s Inequality (Theorem 3.2) with \(\varphi(x) = c_1 x^2\), for an appropriate \(c_1\), and choose \(\lambda = \sqrt{\sigma}\). The inequality produces the bound

$$P(\max_{\theta \in T} |\Delta_\theta| \geq \frac{\epsilon}{z}) \leq B\sigma^{-1} (t_2 - t_1)(h_1 - h_0) e^{-C\epsilon^2/(z^2\sigma^2)},$$

for some finite constants \(B\) and \(C\). The order of magnitude of the right-hand side is \((y/z) \exp\{y^2 + z^2/2 - \sqrt{\pi}C\epsilon^2 z^{-2} ye^{y^2/2}\}\), which converges to zero for \(y = \log z\). Thus, selecting \(y\) to be this large will be sufficient in order to validate (4.3).

### 4.3 Average run length to false alarm

We would like to discuss the poisson approximation in the sequential setting. However, the probability that we computed in the context of the sequential probability ratio test converged to 0 for a maximization that already involved the entire parameter space. Consequently, a poisson approximation will not add new information. Instead, we will consider a closely related problem of sequential monitoring that involves statistics that resemble the statistics that are used for the sequential probability ratio test.

The problem that we would like to consider emerged originally in the context of industrial production quality control. In the third decade of the previous century Shewhart invented the control charts that are named after him. In these charts a quality index is plotted sequentially. As long as the index is below a pre-specified threshold the production is judged to be in control. If the threshold is crossed then the presumption is that something went wrong in the production line prompting action to restore quality. 30 years later Page suggested to use the cusum, short for cumulative sum, statistic as the quality index. Our goal will be to investigate the statistical properties of Page’s cusum statistic, formulated in the same formulation that was used for the sequential probability ratio test.

Consider the construction of the quality index based on the accumulation of \(n\) independent observations \(X_1, X_2, \ldots, X_n\). The distribution of observation \(X_i\) is \(f\) if production was in control at the \(i\)-th production period and it is \(g\) if it went out of control. It production was in control for the entire duration then the density of all observations is \(f\). On the other hand, if production was in control prior to the \(k\)-th production period, but then something went wrong causing the \(k\)-th and subsequent observations to have the alternative density \(g\) then we will get a likelihood which involved \(k - 1\) products of \(f\) followed by \(n - k + 1\) products of \(g\). The log-likelihood ratio statistic to compare the null in-control distribution to an alternative that involves a change at time \(k\) is given by:

$$\ell_{k,n} = \sum_{i=k}^{n} \log\{g(X_i)/f(X_i)\}.$$ 

The change can occur at any production period and is treated as an unknown parameter. Estimating this unknown parameter from the data, using the maximum-
CHAPTER 4. FROM THE LOCAL TO THE GLOBAL

likelihood estimator, produces the cusum statistic:

\[ q_n = \max_{1 \leq k \leq n} \ell_{k,n} \]

that serves as the quality index. The stopping time that is associated with this index corresponds to the first time that the quality index crosses the threshold:

\[ N_x = \inf \{ n : q_n \geq x \} . \]

We are interested in the distribution of the stopping rule \( N_x \) when the process is in control.

We start with the investigation of the probability \( P(N_x \leq m) \), which is carried out by a reformulation of the event in terms of a random field:

\[ \{ N_x \leq m \} = \{ \max_{1 \leq n \leq m} q_n \geq x \} = \{ \max_{1 \leq n \leq m} \max_{1 \leq k \leq n} \ell_{k,n} \geq x \} . \]

This formulation is similar to the formulation that was used for the sequential probability ratio test, but there are differences. The main difference is that in the current example we have a two dimensional random field, parameterized by the pair \((k, n)\) in the triangle \(1 \leq k \leq n \leq m\), and not a one-dimensional process. This will introduce some modifications in the analysis and will produce a different asymptotic behavior. However, most of the details that are involved in the application of the advocated method are more or less the same. Consequently, we will allow ourselves the freedom to skip the proof of some of the details. This freedom is not granted to the active reader. Please check to see that you agree with all the statements that are made and you know how to complete the proofs, especially the conditions of Theorem 5.2.

The first step is localization. Observe that the marginal probabilities of crossing the threshold \( x \) are a function of the difference \( n - k + 1 \) and are maximized when \( n - k + 1 = x/I \). We can restrict the maximization to the strip:

\[ T = \{ (k, n) : x/I - C\sqrt{x} \leq n - k + 1 \leq x/I + C\sqrt{x} \} , \]

for a large \( C \), without compromising the probability by much. You may use the fact that the probability that we seek to approximate is of the order \( m e^{-x} \).

Consequently, the argument that was used for localization in the case of the probability ratio test can be carried out for each \( k, 1 \leq k \leq m \). The resulting errors can be summed up in order to produce a bound on the overall error. The bound on the error will be of a smaller order compared to the size of the probability.

The second step is the transformation of the measure via the likelihood ratio identity. Here we use the likelihood ratio \( \sum_{(k, n) \in T} \exp \{ \ell_{k,n} \} \) to get the representation:

\[ P \left( \max_{(k, n) \in T} \ell_{k,n} \geq x \right) = e^{-x} \sum_{(k, n) \in T} E_{k,n} \frac{M_{k,n} e^{-\ell_{k,n} + m_{k,n}}}{S_{k,n}} \left[ \ell_{k,n} + m_{k,n} \geq 0 \right] , \]

where \( S_{k,n} = \sum_{(j, m) \in T} e^{\ell_{j,m} - \ell_{k,n}} \) and \( M_{k,n} = \max_{(j, m) \in T} e^{\ell_{j,m} - \ell_{k,n}} \) are the sum, respectively the maximal, of likelihood ratios, \( m_{k,n} = \log M_{k,n} \), and \( \ell_{k,n} = \ell_{k,n} - x \).
4.3. AVERAGE RUN LENGTH TO FALSE ALARM

For the application of the localization theorem we use $\kappa = n - k + 1$ and take $
abla_k = \sigma \{ X_{k-\tau}, \ldots, X_{k+\tau}, X_{n-\tau}, \ldots, X_{n+\tau} \}$, for the observations in the vicinity of either $k$ or $n$. In particular, this $\sigma$-field is independent of the observations $X_{k+\tau+1}, \ldots, X_{n-\tau-1}$. The local field $\{ \ell_{j,m} - \ell_{k,n} \}$, restricted to the region $|m-n| \leq \tau$ and $|j-k| \leq \tau$, is measurable with respect to the given $\sigma$-algebra. We produce $\bar{M}_n$ and $\bar{S}_n$ by restricting the maximization (respectively, summation) to the reduced region of parameter values. The distribution of the ratio $\bar{M}_n / \bar{S}_n$ is independent of $\kappa$, for all pairs $(k,n)$ not too close to the boundary of the strip $T$. Moreover, this ratio can be written as a product of two ratios of maximum divided by the sum. One ratio is associated with the observations in the vicinity of $k$ and the other is associated with the observations in the vicinity of $n$. The ratios are independent of each other and each of them is of the same structure as the ratio obtained in the analysis of the sequential probability ratio test. It follows that

$$
\lim_{\tau \to \infty} E_{k,n} [ \bar{M}_{k,n} / \bar{S}_{k,n} ] = \left( E[M/S] \right)^2,
$$

where $E[M/S]$ is the limit obtained for the probability ratio test.

The conclusion from the application of the localization theorem is that

$$
P \left( \max_{(k,n) \in T} \ell_{k,n} \geq x \right) \approx e^{-x} \sum_{(k,n) \in T} \frac{1}{\sqrt{(n-k+1)\sigma^2}} \phi \left( \frac{I(n-k+1) - x}{\sqrt{(n-k+1)\sigma^2}} \right) \left[ E[M/S] \right]^2,
$$

where $\sigma^2$ is the variance of a log-likelihood ratio of an observation under the alternative distribution and $\phi$ is the standard normal density. Finally, the integration step, the approximation becomes:

$$
\lim_{x \to \infty} (e^x/m)P \left( \max_{(k,n) \in T} \ell_{k,n} \geq x \right) = \frac{1}{7} \left( E[M/S] \right)^2 \left[ 2\Phi(C1^{\frac{1}{2}}\sigma^{-1}) - 1 \right].
$$

The extra factor of $m$ results from the fact that each value of $n - k + 1$ appears approximately $m$ times in the strip $T$. Letting $C \to \infty$ results in the statement:

$$
\lim_{x \to \infty} (e^x/m)P(N_x \leq m) = \frac{1}{7} \left( E[M/S] \right)^2.
$$

In the special case of monitoring for a change in a normal mean from 0 to $\mu$ the constant on the right-hand side of (4.4) becomes $(\mu^2/2) \cdot \{ \nu(\mu) \}^2$, for $\nu$ the function associated with an overshoot in a stopped normal random walk.

The approximation in (4.4) is valid as long as $x \ll m \ll e^x$, which makes the probability converge to zero. Approximation of the entire distribution of $N_x$ calls for the use of a poisson approximation.

An attempt to apply the poisson approximation directly on $N_x$ will result in a nonzero $b_4$ term in the bound on the approximation. In order to avoid that we use a truncation argument. The original stopping time is defined with the aid of the quality index $q_n = \max_{1 \leq k \leq n} \ell_{k,n}$. We consider instead an approximation of the quality index that maximizes the log-likelihood ratios only in the range that matters. Fixing a large $C$, let

$$
\hat{q}_n = \max_{\{k; |n-k|/\sqrt{T} \leq C\}} \ell_{k,n}
$$
and define $\tilde{N}_x = \inf\{n : q_n \geq x \}$. We will start by analyzing the distribution of $\tilde{N}_x$ and then show that for a large value of $C$ the distribution of $\tilde{N}_x$ and the distribution of $N_x$ are about the same.

Consider the time interval $[1, ye^x]$ and define $T = \{(k, n) : 0 < n \leq ye^x, |n - k - I/x| \leq C\sqrt{x} \}$. Clearly,

$$P(\tilde{N}_x \leq ye^x) = P(\max_{(k, n) \in T} \ell_{k,n} \geq x).$$

In order to analyze the probability on the right-hand side we divide $T$ into disjoint regions:

$$T_i = \{(k, n) : (i - 1)m < n \leq im, |n - k - I/x| \leq C\sqrt{x} \}, \quad i = 1, 2, \ldots, ye^x/m,$$

and define indicators $X_i$ via the relation $X_i = 1 = \{\max_{(k,n) \in T_i} \ell_{k,n} \geq x\}$. The sum of these indicators, $W = \sum_{i=1}^{ye^x/m} X_i$ is the random variable that we want to approximate by the poisson distribution. The expectation of the limit distribution is the limit of the sum of probabilities of the indicators:

$$\lambda = \lim_{x \to \infty} ye^x/m P(X_1 = 1) = \frac{y}{I} \{E[M/S]\}^2 [2\Phi(CI^2\sigma^{-1}) - 1].$$

The justification for the limit is that, modulo minor edge effects for $X_1$, the distribution of all the indicators is the same. In any case, the rate in which their probability converge to zero when $x \to \infty$, and for $x \ll m \ll e^x$ is the same and is given by the rate that was computed for $X_1$.

In order to validate the poisson convergence of $W$ we need to define neighborhoods of dependence and show that the terms $b_1$ and $b_2$ converge to zero. As before we set $I_i = \{i - 1, i, i + 1\}, 1 < i < ye^x/m$ and let the neighborhoods on the edges contain two indecies only. By construction, $X_i$ is independent of $\{X_j : j \notin I_i\}$ so indeed $b_3 = 0$.

An argument no different than the argument that was used for the scanning statistic will produce:

$$b_1 = [2ye^x/m - 4]P(X_1 = 1)^2 + 2P(X_1 = 1),$$

which is asymptotic to $[2\lambda + 2]P(X_1 = 1)$ and converges to 0. For $b_2$ we have:

$$b_2 = [2ye^x/m - 2]P(X_1 = 1, X_2 = 1).$$

Let $w$ be such that $x \ll w$ but $w/m = o(1)$. Splitting the union $T_1 \cup T_2$ into 3 disjoint subsets, according to $n < m - w, m - w \leq n \leq m + w$ and $n > m + w$, and defining $Y_1, Y_2$ and $Y_3$ to be indicators associated with the new split will produce the upper bound:

$$P(X_2 = 1, X_3 = 1) \leq P(Y_2 = 1) + [P(Y_1 = 1)]^2 \leq P(Y_2 = 1) + \{P(X_1 = 1)\}^2.$$

The probability $P(Y_2 = 1)$ is proportional to $2we^{-x}$. Consequently, $b_3$ is asymptotically bounded by $2\lambda[2w/m + P(X_1 = 1)]$ and it converges to zero when the threshold diverges to infinity. The conclusion is an expression for the asymptotic distribution of $\tilde{N}_x$:

$$\lim_{x \to \infty} P(\tilde{N}_x \leq ye^x) = \lim_{x \to \infty} P(W > 0) = 1 - e^{-y(E[M/S])^2[2\Phi(CI^2\sigma^{-1}) - 1]/I}.$$
4.3. AVERAGE RUN LENGTH TO FALSE ALARM

An alternative statement of the result is to say that $e^x \hat{N}_x$ has an asymptotic exponential distribution with a rate parameter $\{E[M/S]\}^2 [2 \Phi(CI^2 \sigma^{-1}) - 1]/I$.

Our main goal is to obtain the asymptotic distribution of the cusum stopping time $N_x$. For that we show that for large $C$ the distribution of $N_x$ and $\hat{N}_x$ almost coincide. Fixing $y$, we want to compare the probabilities $P(N_x \leq ye^x)$ and $P(\hat{N}_x \leq ye^x)$ to each other. Luckily, one direction is automatically taken care of. By the definitions we have that $q_n \geq \hat{q}_n$, since the latter involves maximization over a smaller subset. Consequently, the process $q_n$ must cross the threshold before (or at the same time) the process $\hat{q}_n$ does. Hence, $N_x \leq \hat{N}_x$ and we get that $P(N_x \leq ye^x) \geq P(\hat{N}_x \leq ye^x)$. What is left to do is to show that the latter probability, for a large $C$, is not much smaller than the former probability.

Fix $y > 0$. Observe that

$$P(\hat{N}_x > ye^x) \leq P(N_x > ye^x) + P(\hat{N}_x > ye^x, N_x \leq ye^x).$$

The event $\{\hat{N}_x > ye^x, N_x \leq ye^x\}$, in reference to the collection of parameters $\{(k, n) : 1 \leq k \leq n \leq ye^x\}$, corresponds to the statement that there exists a log-likelihood $\ell_{k, n}$ that goes above the threshold $x$, but that any such log-likelihood does not belong to the strip $T$ that defines $\hat{N}_x$. One may conclude that the event in question is a subset of:

$$\bigcup_{k=1}^{ye^x} \{ \max_{k \leq n < k + 1/x - C \sqrt{x}} \ell_{k, n} \geq x \} \cup \{ \max_{k+1/x + C \sqrt{x} < n} \ell_{k, n} \geq x \}$$

The probabilities of the event in the big brackets are identical to each other and are equivalent to similar probabilities defined for the sequential probability ratio test (in which $k = 1$). Using exactly the same bounds that were used in the localization argument for the sequential probability ratio test we get that

$$P(\hat{N}_x > ye^x, N_x \leq ye^x) \leq \sum_{k=1}^{ye^x} \left( \frac{\sigma(x/I - C \sqrt{x})}{e^x C^2 I^2 x} + \frac{(x/I + C \sqrt{x}) \sigma^2}{e^x C^2 I^2 x} \right) = \frac{2y \sigma^2}{C^2 I^3},$$

which converges to zero when $C \to \infty$. This concludes the proof of the approximation:

$$\lim_{x \to \infty} P(N_x \leq ye^x) = 1 - e^{-y[E[M/S]]^2/I}, \quad (4.5)$$

that generalizes (4.4) and validates the limit exponential distribution of $e^{-x}N_x$, with a rate given by $\{E[M/S]\}^2/I$.

Change-point detection rules in sequential quality control are eventually activated in probability 1 even if the production process is in control throughout. Such activation is considered a false detection and is to be delayed as much as possible. The principle characteristic of the stopping rule which is used in order to bound from below the delay in false detection is the expectation of the stopping rule. This expectation is called the average run length to false alarm. In particular, there is interest in the asymptotic evaluation of the average run length to false alarm of the cusum change-point detection rule $N_x$, which allows setting the threshold in order to obey a specified average run length to false alarm.

We obtained in (4.5) the asymptotic distribution of the stopping rule. The natural conjecture that the limit of expectations is equal to the expectation of
the limit distribution,

$$\lim_{x \to \infty} e^{-x}E(N_x) = I/\{E[M/S]\}^2,$$

(4.6)
is true but requires proof. The proof follows from the validation of the fact
that the collection of random variables \( \{e^{-x}N_x\} \), indexed by \( x \), is uniformly
integrable.

All random variables involved are non-negative. Therefore, in order to estab-
lish uniform integrability it is sufficient to give a uniform bound on the positive
tail of the distributions of the random variables, for all random variables asso-
ciated with a large enough \( x \).

Let us consider again the sequence of indicators \( \{X_i\} \) that are associated
with exceeding the threshold by likelihood ratio in a strip of length \( m \) and
width of \( 2C\sqrt{x} \). Define the geometric stopping time \( U \) that identifies the index
of the first even element in the sequence that obtains the value one:

\[
U = \inf \{k : X_{2k} = 1 \}.
\]
The distribution of \( U \) is geometric since the even indicators are independent of
each other. Moreover, since \( N_x \leq \hat{N}_x \leq 2mU \) we get that

\[
P(e^{-x}N_x > y) \leq P(U > ye^x/(2m)) = \{1 - P(X_2 = 1)\}^{ye^x/(2m)}.
\]
The sequence \( (e^x/m)P(X_2 = 1) \) converges to a positive limit. Thus, for some
positive number \( \delta \) the quantity is larger than \( \delta \) for all \( x > x_\delta \). Consequently,

\[
P(e^{-x}N_x > y) \leq \{1 - \delta me^{-x}\}^{\{xe^x/(m\delta)\}[\delta y/2]} \leq e^{-\delta y/2},
\]
which completes the proof of uniform integrability and establishes the approxi-
mation in (4.6).
Chapter 5

The Localization Theorem
5.1 Introduction

In this section we present proofs of technical elements that are used in the derivation of the approximation of the distribution of extremes in a random field. The element of the analysis that is specific to the problem at hand is a limit argument that expresses the asymptotic value of expectations that emerge from measure transformation. The argument is summarized in a theorem that we denote the Localization Theorem. The full formulation is given in Theorem 5.2. A simplified version, where the main concepts are easier to grasp, is formulated in Theorem 5.1.

The localization theorem applies a special version of a local limit theorem for distributions. This limit theorem deals with multivariate distributions. Specifically, the joint distribution of the global term and the local field. Unlike standard formulations of limit theorems in distribution, in this version there is a distinction between the limit requirements in one direction, the direction of the global term, and the requirements for the other directions that are specified by the local field. The global term is required to obey a rate of convergence that is consistent with a local limit whereas the demand for the local field is the parallel of standard convergence in distribution.

In general, the limit of the local field need not be gaussian. However, if the local field is a sum of independent fields it will typically be gaussian in the limit. In several applications that we will encounter the global term and the local field are produced from the summation of independent and identically distributed fields. In Theorem 5.3 we show that the minimal conditions of a one-dimensional local limit theorem and a finite-dimension Central Limit Theorem, when combined with a vanishing correlation between the global term and the local field, is sufficient in order to establish the local limit theorem that is required by Theorem 5.2.

5.2 A simplified version of the localization theorem

Consider a triangular array, in which $\kappa$ is the primary index. Let $\ell_{\kappa}$ be a random variable that is closely related to the global log-likelihood ratio. Typically, it will be the difference between the log-likelihood ratio and a threshold. This difference is expected to converge after rescaling (dividing, say, by $\kappa^{1/2}$) to a normal distribution with a possibly non-zero mean. Let $M_{\kappa}$ and $S_{\kappa}$ be a pair of random variables, measurable with respect to the collection of observations. Typically, $\log M_{\kappa}$ will correspond to the maximization of some random field, and $S_{\kappa}$ will be the sum of the exponentiated random field. In many cases $M_{\kappa}$ will be completely analogous to $S_{\kappa}$, but with the summation replaced by maximization. In such cases the ratio $M_{\kappa}/S_{\kappa}$ will have values between zero and one.

After a change of measure along the lines discussed in preceding chapters, we are interested in the limit, as $\kappa \to \infty$, of the term

$$\kappa^{1/2} \mathbb{E}[(M_{\kappa}/S_{\kappa}) \exp[-(\ell_{\kappa} + \log M_{\kappa})]; \ell_{\kappa} + \log M_{\kappa} \geq 0].$$

We first apply a localization step that replaces the quantities $M_{\kappa}$ and $S_{\kappa}$ by local versions that are almost independent of $\ell_{\kappa}$. A local central limit theorem is
applied to the conditional distribution of $\tilde{\ell}_\kappa$, and distributional approximations are applied to the localized version of $M_\kappa/S_\kappa$. Together these yield the desired limit.

Consider a sequence of $\sigma$-fields $\mathcal{F}_\kappa$, $\kappa = 1, 2, \ldots$. Let $\tilde{M}_\kappa$ and $\tilde{S}_\kappa$ be approximations of $M_\kappa$ and $S_\kappa$, respectively, which are measurable with respect to $\mathcal{F}_\kappa$. Given $\epsilon > 0$, we will assume that for all large $\kappa$:

I: $M_\kappa$ and $S_\kappa$ satisfy $0 < M_\kappa \leq S_\kappa < \infty$ with probability one.

II: There exist $\tilde{M}_\kappa$ and $\tilde{S}_\kappa$ measurable with respect to $\mathcal{F}_\kappa$ such that $|\tilde{M}_\kappa/M_\kappa - 1| \leq \epsilon$ and $|\tilde{S}_\kappa/S_\kappa - 1| \leq \epsilon$, with probability at least $1 - \epsilon \kappa^{-\frac{\epsilon}{2}}$.


IV: There exist $\mu_\kappa \in \mathbb{R}$ and $\sigma_\kappa \in \mathbb{R}^+$ such that for every $0 < \epsilon, \delta$ and for all large enough $\kappa$ the probability of the event

$$B_\kappa = \left\{ \sup_{|x| \leq \log \kappa} \kappa^2 P(\tilde{\ell}_\kappa \in x - \log \tilde{M}_\kappa + (0, \delta) | \tilde{\mathcal{F}}_\kappa) - \frac{\delta}{\sigma_\kappa} \phi\left(\frac{\mu_\kappa}{\sigma_\kappa}\right) \leq \epsilon \right\}$$

is bounded from below by $1 - \epsilon \kappa^{-\frac{\epsilon}{2}}$.

**Theorem 5.1.** If Conditions I-IV hold and $\mu_\kappa \to \mu$, $\sigma_\kappa \to \sigma > 0$, then

$$\lim_{\kappa \to \infty} \kappa^2 E\left[(M_\kappa/S_\kappa)e^{-(\tilde{\ell}_\kappa + \log M_\kappa)}; \tilde{\ell}_\kappa + \log M_\kappa \geq 0\right] = \sigma^{-1} \phi(\mu/\sigma) E\left[M/S\right],$$

where $\phi$ is the density of the standard normal distribution.

Before giving a proof of the theorem, we consider the conditions. The random variable $M_\kappa$ typically corresponds to the maximization of a finite collection of non-negative random variables, and $S_\kappa$ corresponds to summation over the same collection. Condition I is satisfied in such a case. The $\sigma$-field $\mathcal{F}_\kappa$ contains the essential information for the determination of the local behavior of the likelihood ratios. Conditions II and III deal with approximation of the functionals that summarize the effect of the local process by the use of terms that are functions of the local field.

The global component $\tilde{\ell}_\kappa$ is approximately normal. Condition IV provides a local central limit statement and establishes a suitable relation between the local components and the global one.

**Proof of Theorem 5.1** Denote by $A_\kappa$ the event over which Condition II holds. We will produce both an upper bound and a lower bound, which together prove the theorem.

Start with an upper bound. On the event $A_\kappa$ we replace $S_\kappa$ by the lower bound $\tilde{S}_\kappa/(1 + \epsilon)$ and replace $M_\kappa$ by an upper bound given in terms of $\tilde{M}_\kappa$. Specifically, we replace $\log M_\kappa$ on the event $\{\tilde{\ell}_\kappa + \log(M_\kappa) \geq 0\}$ by the upper bound $\log \tilde{M}_\kappa - \log(1 - \epsilon)$ in order to obtain:

$$E\left[(M_\kappa/S_\kappa)e^{-(\tilde{\ell}_\kappa + \log M_\kappa)}; \tilde{\ell}_\kappa + \log M_\kappa \geq 0\right] \leq P(A_\kappa) + \frac{1 + \epsilon}{1 - \epsilon} E\left[(\tilde{M}_\kappa/\tilde{S}_\kappa)e^{-(\tilde{\ell}_\kappa + \log M_\kappa - \log(1 - \epsilon))}; \tilde{\ell}_\kappa + \log M_\kappa - \log(1 - \epsilon) \geq 0, A_\kappa\right].$$
In this inequality we exploit the fact that
\[ 0 < (M_\kappa/S_\kappa)e^{-(\ell_\kappa + \log M_\kappa)}I_{(\hat{\ell}_\kappa + \log M_\kappa)} \leq 1, \]
thus replacing the expectation over the complementary of the event \( A_\kappa \) by the probability of the complementary event. Due to the assumption in Condition II we can ignore \( P(A_\kappa^c) \).

Denote \( \hat{m}_\kappa = \log M_\kappa - \log(1 - \epsilon) \). By conditioning on the \( \sigma \)-field \( \hat{F}_\kappa \) and the fact that \( \bar{M}_\kappa \) and \( \bar{S}_\kappa \) are measurable with respect to \( \hat{F}_\kappa \) one finds that:

\[
E\left[ \frac{\bar{M}_\kappa}{\bar{S}_\kappa} e^{-(\ell_\kappa + \hat{m}_\kappa)}; \ell_\kappa + \hat{m}_\kappa \geq 0 \right] = E\left[ \frac{\bar{M}_\kappa}{\bar{S}_\kappa} E\left( e^{-(\ell_\kappa + \hat{m}_\kappa)}; \ell_\kappa + \hat{m}_\kappa \geq 0 | \hat{F}_\kappa \right) \right].
\]

We proceed the construction of an upper bound by analyzing the conditional expectation. The exponential function is a monotone function. Hence:

\[
E\left( e^{-(\ell_\kappa + \hat{m}_\kappa)}; \ell_\kappa + \hat{m}_\kappa \geq 0 | \hat{F}_\kappa \right) \leq \sum_{j=0}^{\log(\kappa)/\delta} e^{-\delta j} P(\ell_\kappa \in [\delta j - \hat{m}_\kappa + (0, \delta) | \hat{F}_\kappa) + \frac{1}{\kappa(1 - e^{-\delta})}.}
\]

After multiplication by \( \kappa^\frac{1}{2} \) we get, directly from Condition IV, that over the event \( B_\kappa \) the sum is bounded from above by
\[
\left[ \delta/(1 - e^{-\delta}) \right] \left[ \sigma_\kappa^{-1} \phi(\mu_\kappa/\sigma_\kappa) + \epsilon \right]
\]
and the remainder term still converges to zero. Taking expectation with respect to \( \hat{F}_\kappa \) and using Condition III and the fact that the ratio \( M_\kappa/\bar{S}_\kappa \) is bounded, eliminating in the process the vanishingly small error caused by the complementary of the event \( B_\kappa \), complete the upper bound.

The lower bound is constructed in a similar way. This time we replace \( S_\kappa \) by an upper bound and replace \( M_\kappa \) by a lower bound. This produces:

\[
E\left[ \left( M_\kappa/S_\kappa \right)e^{-(\ell_\kappa + \log M_\kappa)}; \ell_\kappa + \log M_\kappa \geq 0 \right] \geq \frac{1 - \epsilon}{1 + \epsilon} E\left[ \left( M_\kappa/S_\kappa \right)e^{-(\ell_\kappa + \log M_\kappa - \log(1 + \epsilon))}; \ell_\kappa + \log M_\kappa - \log(1 + \epsilon) \geq 0, A_\kappa \right].
\]

Denote now \( \hat{m}_\kappa = \log M_\kappa - \log(1 + \epsilon) \) and consider the expectation on the righthand side:

\[
E\left[ \left( M_\kappa/S_\kappa \right)e^{-(\ell_\kappa + \hat{m}_\kappa)}; \ell_\kappa + \hat{m}_\kappa \geq 0, A_\kappa \right] \geq E\left[ \left( M_\kappa/S_\kappa \right)e^{-(\ell_\kappa + \hat{m}_\kappa)}; \ell_\kappa + \hat{m}_\kappa \geq 0, B_\kappa \right] - P(A_\kappa^c).
\]

As before, we will ignore the error resulting from the complementary of the event \( A_\kappa \) and use the monotonicity of the exponential function:

\[
E\left( e^{-(\ell_\kappa + \hat{m}_\kappa)}; \ell_\kappa + \hat{m}_\kappa \geq 0 | \hat{F}_\kappa \right) \geq \sum_{j=0}^{\left[ \log(\kappa)/\delta \right] - 1} e^{-\delta(j+1)} P(\ell_\kappa \in [\delta j - \hat{m}_\kappa + (0, \delta) | \hat{F}_\kappa) .
\]
5.3. THE LOCALIZATION THEOREM

Observe that we are using the fact that $B_\kappa$ is measurable with respect to $\mathcal{F}_\kappa$.

After multiplication by $\frac{1}{2} \kappa$ we get, directly from Condition IV, that over the event $B_\kappa$ the sum is bounded from below by:

$$e^{-\delta}(1 - 1/\kappa)\left[\delta/(1 - e^{-\delta})\right] \left[\sigma^{-1}\phi(\mu/\sigma) - \epsilon\right].$$

Taking expectation with respect to $\hat{\mathcal{F}}_\kappa$, using again Condition III and the fact that the ratio $\hat{M}_\kappa/\hat{S}_\kappa$ is bounded, complete the lower bound and the proof of Theorem 5.1.

5.3 The Localization Theorem

The conditions for Theorem 5.1 produce a simple proof. In simple cases, for example when the random field is gaussian to begin with, one may be able to verify the conditions. However, in more complex scenarios the verification is more difficult. In particular, the most problematic condition is Condition IV, which is a form of a conditional local central limit theorem of the global term, given the $\sigma$-algebra generated by the local field.

There are two difficulties with the condition as stated. Firstly, the localization argument that is used in the proof of Theorem 5.1 requires that the approximation of $M_\kappa$ and $S_\kappa$ by $\hat{M}_\kappa$ and $\hat{S}_\kappa$ improves as $\kappa$ increases. This, in turn, calls for an increase in the number of elements of the local field that are involved in the approximation. Consequently, the consideration of a conditional local central limit theorem for $\tilde{\ell}_\kappa$, given only a bounded number of elements of the local field, will not be sufficient.

The other difficulty is that a conditional local-central limit theorem is stronger than what is required. Actually, considering the joint distribution of the global term and the local field, only the part that involves the global term demands a local limit theorem. For the part that involves the local field a weak convergence is sufficient. Consideration of conditional distribution essentially calls for the investigation of ratios of densities, thereby enforcing an approximation of the density in the denominator, namely the density of the local field.

As a remedy for the shortcomings of Theorem 5.1 we propose a modified version of the theorem. The conditions in this modified version are more difficult to state but are easier to validate. The price that we pay is in the form of a more complex proof.

Given $\epsilon, \epsilon_3 > 0$, we assume that for some function $g(\kappa)$, $\log \kappa \leq g(\kappa) \leq \epsilon k^{\frac{1}{2}}$, for some $C < \infty$, and for all large $\kappa$ the following conditions:

$\Gamma$: $M_\kappa$, $S_\kappa$, $\hat{M}_\kappa$ and $\hat{S}_\kappa$ satisfy $0 \leq M_\kappa/S_\kappa \leq C$ and $0 \leq \hat{M}_\kappa/\hat{S}_\kappa \leq C$ with probability one.

$\Pi$: Denote $A^\epsilon_{\text{II}} = \{|\log M_\kappa - \log \hat{M}_\kappa| > \epsilon\} \cup \{|\hat{S}_\kappa/S_\kappa - 1| > \epsilon\}$. For some $0 < \delta$ that does not depend on $\epsilon$:

$$\max_{|x| \leq 3g(\kappa)} P\left(A^\epsilon_{\text{II}} \cap \{\tilde{\ell}_\kappa + \log \hat{M}_\kappa \in x + (0, \delta]\} \cap \{|\tilde{m}| \leq g(\kappa)\}\right) \leq \epsilon k^{-\frac{1}{2}}. \quad (5.1)$$

$\Pi^*$: $E[\hat{M}_\kappa/\hat{S}_\kappa]$ converges to $E[\hat{M}/\hat{S}]$ and $|E[\hat{M}/\hat{S}] - E[\hat{M}/\hat{S}]| \leq \epsilon_3$. 

IV*: There exist $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}^+$ such that for every $0 < \epsilon_1, \delta$, for any event $E \in \mathcal{F}_\kappa$ having boundary measure 0, and for all large enough $\kappa$:

$$\sup_{|x| \leq \log(\kappa)} |\kappa^{\frac{1}{2}} P(\bar{\ell}_\kappa + \log \hat{M}_\kappa \in x + (0, \delta], E) - \frac{\delta}{\sigma} \phi\left(\frac{\mu}{\sigma}\right) P(E)| \leq \epsilon_4,$$

and also:

$$V*: P\left(\left| \log M_\kappa \right| > g(\kappa) \right), P\left(\left| \log \hat{M}_\kappa \right| > g(\kappa) \right) \text{ and } P\left(\log M_\kappa - \log \hat{M}_\kappa < -\epsilon\right) \text{ are all } o(\kappa^{-\frac{3}{4}}).$$

Theorem 5.2. If Conditions I*-V* hold then:

$$\lim_{\kappa \to \infty} \kappa^\frac{1}{2} E[(M_\kappa/S_\kappa)e^{-\bar{\ell}_\kappa + \log M_\kappa}; \bar{\ell}_\kappa + \log M_\kappa \geq 0] = \sigma^{-1} \phi(\mu/\sigma) E[M/S].$$

Conditions I* and III* are essentially the same as Conditions I and III of Theorem 5.1. Condition II* is less restrictive than the parallel condition of Theorem 5.1. In the previous condition the probability of the event where the approximations of the maximum and the sum is not valid was required to go to zero in a rate faster than $\kappa^{-\frac{3}{2}}$. In the new condition that rate should hold when intersecting the event with an event that already has a rate of convergence of $O(\kappa^{-\frac{3}{4}})$. Condition IV* is the statement of the local limit theorem. The condition is formulated in terms of the joint distribution of the local field and $\bar{\ell}_\kappa$ and is substantially weaker than the corresponding requirement in Condition IV, which involves conditional distribution of the latter given the former. Condition V* is a new condition. It is a natural condition when localization involves both a reduction in the set of indices that enter into the maximization/summation and an approximation of the local process. Note that the sum and the maximum of likelihood ratios are both decreasing if the set of indices is reduced. Hence, the actual requirement is for an improving approximation of the elements of local field.

Notice that there is another discrepancy between Condition IV and Condition IV*. In the former the maximization of the error is taken with respect to $x$ in the region $|x| \leq \log \kappa$. In the latter the maximization is over a larger region $|x| \leq g(\kappa)$. There is a tradeoff between Conditions II* and IV* and Condition V*. The better we can control the magnitude of the local field the less stringent the bound in the local limit theorem that we need to validate for the global term.

Proof of Theorem 5.2. The proof is similar to that of Theorem 5.1 but the details are more involved.

Consider first an upper bound. Denote by $A^c_\nu$ the union of the events entering into Condition V*. As before, on $A_{ni}$, the complementarity of $A^c_\nu$ from Condition II*, we decrease $S_\kappa$ and increase $\log M_\kappa$ in the event $\{\bar{\ell}_\kappa + \log \hat{M}_\kappa \geq 0\}$ to obtain:

$$E[(M_\kappa/S_\kappa)e^{-(\bar{\ell}_\kappa + \log M_\kappa)}; \bar{\ell}_\kappa + \log M_\kappa \geq 0] \leq$$

$$\frac{1 + \epsilon}{1 - \epsilon} E[(\bar{M}_\kappa/S_\kappa)e^{-(\bar{\ell}_\kappa + \log M_\kappa - \log(1 - \epsilon))}; \bar{\ell}_\kappa + \log \hat{M}_\kappa - \log(1 - \epsilon) \geq 0, A^c_\nu] + CE[e^{-(\bar{\ell}_\kappa + \log M_\kappa)}; \bar{\ell}_\kappa + \log M_\kappa \geq 0, A_{ni}^\nu \cap A_\nu] + CP(A_\nu^c).$$
5.3. THE LOCALIZATION THEOREM

We used the fact that \(M_\kappa/S_\kappa \leq C\) in the last two terms. By Conditions I* and V* we can ignore the last of these terms.

Denote \(\tilde{m}_\kappa = \min\{\log \tilde{M}_\kappa, g(\kappa)\} = \log(1 - \epsilon)\) and consider the first expectation. Notice that \(M_\kappa/S_\kappa\) is a bounded function of the local process. Let \(\sum_{n=1}^n e_u I_{E_n}\) be a simple function that bounds the function from above and approximates it up to \(\epsilon\) in the maximal norm. The partition \(\{E_n\}\) may be formed with sets of zero boundary that belong to \(\tilde{F}_\kappa\). The number of events in the partition \(n = n(\epsilon)\) is finite for any positive \(\epsilon\). By the monotonicity of the exponential function:

\[
E[(\tilde{M}_\kappa/S_\kappa)e^{-(\ell_\kappa + \tilde{m}_\kappa)}; \tilde{\ell}_\kappa + \tilde{m}_\kappa \geq 0] \leq \sum_{j=0}^{\log(\kappa)/\delta} \sum_u e_u e^{-j\delta} P(\tilde{\ell}_\kappa + \tilde{m}_\kappa \in j\delta + (0, \delta], E_u) + \frac{1}{\kappa(1 - e^{-\delta})}.
\]

After multiplication by \(\kappa^2\) we find by Conditions IV* and V* that the sum is bounded over the event defined in these conditions by:

\[
\frac{\delta}{1 - e^{-\delta}} \frac{1}{\sigma} \phi(\frac{\mu}{\sigma}) \sum_u e_u P(E_u) - \epsilon_4 \sum_u e_u.
\]

Since \(\sum_u e_u P(E_u)\) approximates the expectation of \(\tilde{M}_\kappa/S_\kappa\) and since \(\sum_u e_u\) is bounded and \(\epsilon_4\) can be made arbitrarily small relative to both \(\epsilon\) and \(\delta\), we get the upper bound associated with the first expectation.

In order to complete the upper bound we need to show that

\[
\kappa^2 E[e^{-(\ell_\kappa + \log M_\kappa)}; \tilde{\ell}_\kappa + \log M_\kappa \geq 0, A_{m_i}^c \cap A_v]
\]

is negligible.

We examine the expectation over each of the events \(\{|\log M_\kappa| > g(\kappa)\}, \{|\log M_\kappa| \leq g(\kappa), |\tilde{\ell}_\kappa| > 2g(\kappa)\}, \) and \(\{|\tilde{\ell}_\kappa| \leq 2g(\kappa)\}\). The expectation over the first event is negligible by Condition V*. The expectation over the second event is negligible because the event implies that \(\tilde{\ell}_\kappa + \log(M_\kappa) > g(\kappa)\). Consequently, we get a vanishingly small bound of order \(\kappa^2 \exp\{-g(\kappa)\}\). Hence, we only need to deal with the expectation over the third event.

Let \(m_\kappa = \log M_\kappa, \tilde{m}_\kappa = \log \tilde{M}_\kappa\). Set \(B_0 = \{|m_\kappa - \tilde{m}_\kappa| < \epsilon\}\), and for \(0 < i < 2g(\kappa)/\delta\), let \(B_i = \{m_\kappa - \tilde{m}_\kappa \in \epsilon + (i - 1)\delta + (0, \delta]\}\). Observe that \(A_{m_i}^c \cap A_v \subset \bigcup_{i>0} (B_i \cap A_{m_i}^c) \subset A_{m_i}^c\).

Let \(C_j = \{\ell_\kappa + \tilde{m}_\kappa \in j\delta + (0, \delta]\} \cap \{|\tilde{m}_\kappa| \leq g(\kappa)\}\). This event is a function of \(\tilde{\ell}_\kappa\) and the local \(\sigma\)-algebra \(\tilde{F}_\kappa\). Conditioning on \(\tilde{F}_\kappa\) and on \(\tilde{\ell}_\kappa\) we get

\[
E[e^{-(\ell_\kappa + m_\kappa)}; \ell_\kappa + m_\kappa \geq 0, A_{m_i}^c \cap A_v \cap B_i \cap C_{j-1} \cap \tilde{F}_\kappa, \tilde{\ell}_\kappa] \leq e^{\epsilon} e^{-j\delta} I_{(j \geq 0)} I_{C_{j-1}} P(A_{m_i}^c \cap B_i, \tilde{F}_\kappa, \tilde{\ell}_\kappa) \leq e^{\epsilon} e^{-j\delta} I_{(j \geq 0)} I_{C_{j-1}} P(A_{m_i}^c, \tilde{F}_\kappa, \tilde{\ell}_\kappa).
\]

The value of \(\ell_\kappa + m_\kappa\) is essentially equal to the quantity \(j\delta\). The largest value of \(j\) that needs to be considered is \(j = 3g(\kappa)/\delta\). The smallest value is \(j = 0\). The fact that only \(j \geq 0\) needs to be consider follows from the fact that \(\ell_\kappa + m_\kappa \geq 0\).
Taking expectations with respect to the conditional distribution, given \( \hat{F}_\kappa \), produces:

\[
E[e^{-(\bar{\ell}_\kappa + m_\kappa)}; \bar{\ell}_\kappa + m_\kappa \geq 0, A_{11} \cap A_\nu, |\bar{\ell}_\kappa| \leq 2g(\kappa)|\hat{F}_\kappa|] \\
\leq e^\delta \sum_{j=0}^{\lceil 3g(\kappa)/\delta \rceil} \sum_{i=0}^{\lceil 3g(\kappa)/\delta \rceil} e^{-j\delta} E[P(A_{11}^c | \hat{F}_\kappa, \bar{\ell}_\kappa); C_{j-i} | \hat{F}_\kappa] \\
= e^\delta \sum_{j=0}^{\lceil 3g(\kappa)/\delta \rceil} \sum_{i=0}^{\lceil 3g(\kappa)/\delta \rceil} e^{-j\delta} P(A_{11}^c \cap \{\bar{\ell}_\kappa + \hat{m}_\kappa \in (j-i)\delta + (0, \delta]\} | \hat{F}_\kappa).
\]

When we integrate with respect to the local random field and over the event \( D = \{|\hat{m}| \leq g(\kappa)\} \), we get that:

\[
P(A_{11}^c \cap \{\bar{\ell}_\kappa + \hat{m}_\kappa \in (j-i)\delta + (0, \delta]\} \cap D) \leq \max_{|x| \leq 3g(\kappa)} P(A_{11}^c \cap \{\bar{\ell}_\kappa + \hat{m}_\kappa \in x + (0, \delta]\} \cap D)
\]

that does not depend any longer on \( i \) or \( j \). Multiplying by \( \kappa^{2} \) and applying Condition II* leads to the bound:

\[
e^\delta (1 + \epsilon) \frac{(1 - e^{-\delta^2})}{1 - e^{-\delta}} \times \epsilon.
\]

This completes the proof of the upper bound.

Now we consider the lower bound. This time we increase \( M_\kappa \) in the event \( \{\bar{\ell}_\kappa + \log M_\kappa \geq 0\} \) to obtain

\[
E[ \frac{M_\kappa}{S_\kappa} e^{-(\bar{\ell}_\kappa + \log M_\kappa)}; \bar{\ell}_\kappa + \log M_\kappa \geq 0] \\
\geq \frac{1 - \epsilon}{1 + \epsilon} E[ \frac{\hat{M}_\kappa}{\hat{S}_\kappa} e^{-(\bar{\ell}_\kappa + \log \hat{M}_\kappa - \log(1 + \epsilon))}; \bar{\ell}_\kappa + \log \hat{M}_\kappa - \log(1 + \epsilon) \geq 0, A_{11} \cap A_\nu].
\]

Denote now \( \hat{m}_\kappa = \log \hat{M}_\kappa - \log(1 + \epsilon) \) and consider the expectation on the righthand side:

\[
E[(\hat{M}_\kappa/\hat{S}_\kappa)e^{-(\bar{\ell}_\kappa + \hat{m}_\kappa)}; \bar{\ell}_\kappa + \hat{m}_\kappa \geq 0, A_{11} \cap A_\nu] = \\
E[(\hat{M}_\kappa/\hat{S}_\kappa)e^{-(\bar{\ell}_\kappa + \hat{m}_\kappa)}; \bar{\ell}_\kappa + \hat{m}_\kappa \geq 0] \\
- E[(\hat{M}_\kappa/\hat{S}_\kappa)e^{-(\bar{\ell}_\kappa + \hat{m}_\kappa)}; \bar{\ell}_\kappa + \hat{m}_\kappa \geq 0, A_{11}^c, A_\nu] = \text{CP}(A_\nu^c).
\]

We may ignore the last probability. For the first term on the righthand side of the equation we get by the monotonicity of the exponential function:

\[
E[(\hat{M}_\kappa/\hat{S}_\kappa)e^{-(\bar{\ell}_\kappa + \hat{m}_\kappa)}; \bar{\ell}_\kappa + \hat{m}_\kappa \geq 0] \geq \\
\sum_{j=0}^{\lceil \log(\kappa)/\delta \rceil} \sum_{u} e_u e^{-j\delta} P(\bar{\ell}_\kappa + \hat{m}_\kappa \in j\delta + (0, \delta], E_u) I(\hat{m}_\kappa < g(\kappa)),
\]

where this time \( \sum_u e_u E_u \) is a simple and non-negative function that approximates \( \hat{M}_\kappa/\hat{S}_\kappa \) from below. After multiplication by \( \kappa^{2} \) and the application of Condition IV* we get (over the event \( \{\hat{m}_\kappa < g(\kappa)\} \)) the lower bound:

\[
\frac{e^{-\delta(1 - 1/\kappa)}}{1 - e^{-\delta}} \frac{1}{\sigma^2} \phi\left(\frac{\mu}{\sigma}\right) \sum_u e_u P(E_u) - \frac{\epsilon_4}{1 - e^{-\delta}} \sum_u e_u.
\]
Again, ignoring the negligible reduction that is caused by the truncation \( \{ \hat{m}_\kappa < g(\kappa) \} \), we get a first term which is almost equal to the leading term that appears in the upper bound.

In order to complete the proof we need an upper bound on the term

\[
E[\hat{M}_\kappa/\hat{S}_\kappa e^{-(\tilde{\ell}_\kappa + \hat{m}_\kappa)}; \tilde{\ell}_\kappa + \hat{m}_\kappa \geq 0, A^*_{ii_i}, A_{\nu}].
\]

But this can be obtained by a repeat of the proof that was used in order to bound (5.2), the parallel term that emerged in the context of constructing an upper bound.

5.4 A local limit theorem

Condition IV* of Theorem 5.2 is an interesting conditions and it is formulated the way it does for a reason. The context of the condition are local limit theorems and higher order approximation in the central limit theorem. The two issues are related. Consider a distribution that converges to the normal distribution. If an expansion with an error of order \( o(n^{-\frac{1}{2}}) \) is given to the distribution function then that expansion can be used in order to produce a local limit type of approximation.

In the one-dimensional i.i.d. setting such an approximation exists under very mild conditions. An evidence is Theorem 3.1 which requires only a second moment and the minimal assumption of a non-lattice distribution. Adding a third moment will guarantee a uniform approximation.

However, in the multivariate case the picture is not as simple. An attempt to obtain an accurate approximation of the joint distribution is not trivial at all and typically requires conditions that applications don’t share. The best result we know of that may be used in our context is Theorem 20.1 in [3] that produces an expansion of the distribution of a standardized sum of i.i.d. random vectors with an accuracy that depends on the highest moment. In particular, if a third moment exists then the expansion seems to be good enough for our needs.

There is a catch. The joint characteristic function in the theorem is required to satisfy the Cramér’s condition. This condition states that the absolute value of the characteristic function must not converge to 1 along any subsequence that diverges to infinity. In comparison, the non-lattice assumption in the one-dimensional case simply states that the value is 1 only at the origin. The non-lattice condition does not exclude cases where the absolute value converges to 1 for a sequence that converges to infinity. The Cramér’s condition does not allow such convergence.

This condition would not have been a serious concern. Alas, it excludes discrete distributions, preventing us from using for example functions of the binomial or poisson distributions. Consequently, it is worthwhile to look for proofs that are less restrictive.

An important point in Condition IV* is the difference between the global term, which requires an accurate assessment of the distribution, and the local field. For the local field, the multivariate part, only weak converges is called for. The rate of that convergence is not critical. Consequently, one may hope, that the situation is closer to the situation encountered in the one-dimensional case and more general proofs can be obtained.
Theorem 5.3 is an attempt in that direction. It deals with the situation when the field is a sum of independent and identically distributed finite dimensional fields. One coordinate in each of the fields, the coordinate associated with the global term, has variance 1. The other coordinates share a vanishing covariance structure that converges to a limit when summed. The correlation between the (standardized) global term and the local field is vanishing in the limit, which implies independence between the limit gaussian local field and the global term. It should be noted that this theorem does not cover situations, such as the case of the sequential probability ratio test, where the limit of the local field is not gaussian.

Theorem 5.3. Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random vectors of dimension $d + 1$. Assume that $E(X_1) = (0, \mu_n)'$ and

$$\text{Var}(X_1) = \left( \begin{array}{cc} 1 & \rho_n \\
 & \Sigma_n \end{array} \right),$$

where $\rho_n$ is a correlation vector and $\Sigma_n$ is a $d$-dimensional variance-covariance matrix. Assume that the marginal distribution of the first coordinate is non-lattice and identical for all $n$. Consider the sum $S_n = X_1 + \cdots + X_n$ and let $m = m_n : \mathbb{R}^d \to \mathbb{R}$ be a bounded function. The term $m(S_n)$ corresponds to the application of the function $m$ to the last $d$ coordinates of $S_n$. Then if $\delta > 0$ and a set $A$ with a zero measure boundary are fixed and if $x_n/\sqrt{n} \to x$, for a finite $x$, $n\mu_n \to \mu$, $n\Sigma_n \to \Sigma$, $n^2 \rho_n \to 0$, and $\max |m| = o(\sqrt{n})$ then

$$\lim_{n \to \infty} \sqrt{n} P(S_n + m(S_n) \in (x_n, x_n + \delta) \times A) = \delta \phi(x) P(Z \in A),$$

for $Z \sim N(\mu, \Sigma)$.

Proof. Let $\delta > 0$, set $S_n = X_1 + \cdots + X_n$, and assume initially that $\mu_n = 0$ and $m \equiv 0$. Use a kernel produced of the joint density function of independent Polya’s random variables:

$$H_0(y, z) = h(y)k(z) = \frac{1}{\pi} \frac{1 - \cos(\epsilon_0y)}{\epsilon_0y^2} \times \prod_{j=1}^{d} \left[ \frac{1}{\pi} \frac{1 - \cos(\epsilon_j z_j)}{\epsilon_j z_j^2} \right],$$

for $\epsilon_j > 0$, $0 \leq j \leq d$, $y$ a real number and $z = (z_1, \ldots, z_d)'$ a vector.

Consider the characteristic function of this Polya kernel, a product of the marginal characteristic functions, which is of a bounded support and is equal to:

$$\hat{H}_0(u_0, u) = \hat{h}(u_0)\hat{k}(u) = \left\{ \begin{array}{ll} \prod_{j=0}^{d} (1 - |u_j/\epsilon_j|) & \text{if, for all } j, |u_j| \leq \epsilon_j, \\
0 & \text{otherwise.} \end{array} \right.$$  

Extend the density to a family of complex-valued functions by taking

$$H_0(y, z) = h_0(y)k_0(z) = e^{i\eta y}h(y)e^{i(\eta z)}k(z) = e^{i(\eta y + (\eta z))}H_0(y, z),$$

for $\theta = (\vartheta, \eta)$, and observe that $\hat{H}_0(u_0, u) = \hat{H}_0(u_0 + \vartheta, u + \eta)$.

We start by showing that for any given $\theta$:

$$\lim_{n \to \infty} \sqrt{n} E H_\theta(S_n - (x_n, 0)') = \phi(x) \int h_\theta(y)dy \cdot E_{\eta}(Z),$$  \hspace{1cm} (5.3)
where $\phi$ is the density of the standard normal distribution and $Z \sim N(0, \Sigma)$. For that we use the inversion formula for characteristic functions and get that:

$$H_0(y, z) = (2\pi)^{-d+1} \iint e^{- \imath u_0 y - \imath u z} \hat{H}_0(u_0, u) du_0 du$$

and therefore, by the change of variable $u_0 = v + \vartheta$, $u = w + \eta$

$$H_\theta(y, z) = e^{\imath \vartheta y + \imath \eta z} H_0(y, z) = \frac{1}{(2\pi)^d+1} \iint e^{- \imath (u_0 - \vartheta) y - \imath (u - \eta) z} \hat{H}_0(u_0, u) du_0 du = \frac{1}{(2\pi)^d+1} \iint e^{- \imath ivx - \imath (w, z)} \hat{H}_\theta(v, w) dv dw .$$

Denote the distribution of $S_n - (x_n, 0)'$ by $F_n$ and apply Fubini’s theorem to write:

$$EH_\theta(S_n - (x_n, 0)') = \frac{1}{(2\pi)^d+1} \iint \iint e^{- \imath ivy - \imath (w, z)} \hat{H}_\theta(v, w) dv dw dF_n(y, z) = \frac{1}{(2\pi)^d+1} \iint \iint e^{- \imath ivy - \imath (w, z)} dF_n(y, z) \hat{H}_\theta(v, w) dv dw .$$

The innermost integral corresponds to the characteristic function of $S_n - (x_n, 0)'$, evaluated at $(-v, -w)'$, hence

$$E \hat{H}_\theta(v, w) = \frac{1}{(2\pi)^d+1} \iint [\varphi_n(-v, -w)] n e^{\imath ivx} \hat{H}_\theta(v, w) dv dw .$$

Notice that

$$|\varphi_n(-v, -w) - \varphi_n(-v, 0)| \leq E|e^{- \imath ((0, w)'X_1)} - 1| = o(w\Sigma n) ,$$

so the joint characteristic function may be bounded by the marginal characteristic function of the first component with an error of order smaller than $1/n$, uniformly over bounded regions of $w$. Moreover, the joint characteristic function is bounded by

$$|\varphi_n(-v, -w)| \leq \exp\{-(1/4)|v|^2 + 2v(w, w') + w\Sigma n\} ,$$

whenever the term in the square bracket is small enough.

In order to show that the limit of the given integral, multiplied by $\sqrt{n}$, is equal to the right-hand side of (5.3) we consider 4 regions. The first region is the region $[-\epsilon_n, \epsilon_n] \times B$, where $B = [-M, M]^d$ and where $\epsilon_n = \log(n)/n^{\frac{1}{4}}$. The second region is $([-\epsilon, -\epsilon_n) \cup (\epsilon_n, \epsilon)] \times B$, for $\epsilon > 0$ small enough to assure the validity, for all $n \geq n_\epsilon$, of the bound (5.5). The third region is $([-M, M] \times B) \setminus ([\epsilon, \epsilon] \times B)$, for $[-M, M] \times B$ that contains the support of $H_\theta$. The last region is $\mathbb{R}^{d+1} \setminus ([-M, M] \times B)$, over which the integrand is equal to 0.

The last region does not contribute to the integral. The contribution of the third region is bounded by $(M/\pi)^{d+1} [1 + o(1/n)]^n$, for $n = \sup|\varphi_n(-v, -w)| < 1$, where the suprema is taken over the third region and it holds for all large enough $n$ in light of (5.4). Such contribution is $o(n^{-\frac{1}{2}})$ small. In the second region we apply (5.5). The dominating contribution results from $v$ and produces a bound
of order \( n^{-0.5 \log n} \), enough to eliminate the \( \sqrt{n} \) factor. For the first region we have, after multiplying by \( \sqrt{n} \) and changing the variable to \((\xi, w) = (\nu n^{-\frac{1}{2}}, w)\):

\[
\frac{\sqrt{n}}{(2\pi)^{d+1}} \int_{-\epsilon}^{\epsilon} \int_{-\epsilon}^{\epsilon} [\varphi_n(-v, -w)]^n e^{i\xi x_n} \hat{H}_\theta(v, w) dv dw = \frac{1}{(2\pi)^{d+1}} \int_{-\log n}^{\log n} \int_{B} [\varphi_n(-\xi n^{-\frac{1}{2}}, w)]^n e^{i\xi x_n} n^{-\frac{1}{2}} \hat{H}_\theta(\xi n^{-\frac{1}{2}}, w) d\xi dw .
\]

The integrand converges, for each fixed \((\xi, w),\) to:

\[ e^{-\frac{1}{2}((\xi^2 + w^2)\Sigma + i\xi x)\hat{H}_\theta(0, w)}. \]

Application of the dominated convergence theorem, justified by (5.5), will give:

\[
\lim_{n \to \infty} \frac{1}{(2\pi)^{d+1}} \int_{-\log n}^{\log n} \int_{B} e^{-\frac{1}{2}((\xi^2 + w^2)\Sigma + i\xi x)\hat{H}_\theta(0, w)} d\xi dw = \phi(x) \hat{h}_\theta(0) \int e^{-\frac{1}{2}w^2\Sigma \hat{k}_\eta(w)} dw = \phi(x) \int h_\eta(y) dy \cdot E k_\eta(Z) .
\]

The last equation follows from the definition of the Fourier transform. This completes the proof of (5.3), which we will use next in order to prove the statement of the theorem.

For that proof we consider two sequences of measures on \( \mathbb{R}^{d+1} \). The first is the measure that is generated by

\[ \mu_n([a, b] \times A) = \sqrt{n} P(S_n - (x_n, 0)^\top \in [a, b] \times A) , \]

which we wish to show that if converges to the measure generated by \( \phi(x) P(Z \in A) \mu([a, b]), \) for \( \mu \) the Lebesgue measure. The other measure is the probability measure generated by:

\[ \nu_n([a, b] \times A) = \frac{1}{\alpha_n} \int_a^b \int_A H_\theta(y, z) \mu_n(dy, dz) , \]

for \( \alpha_n = \sqrt{n} E H_\theta(S_n - (x_n, 0)^\top) \). Form (5.3) it follows, for \( \theta = 0 \), that \( \alpha_n \to \phi(x) E k(Z) \) and more generally that:

\[
\int \int e^{i\varphi y + (y, z)} d\nu_n(y, z) = \frac{1}{\alpha_n} \sqrt{n} E H_\theta(S_n - (x_n, 0)^\top) \\
\lim_{n \to \infty} \int e^{i\varphi y} h(y) dy \to \frac{E [e^{i\varphi y} k(Z)]}{E k(Z)} .
\]

Consequently, \( \nu_n \) converges in distribution to a distribution with a density that is the product of a one-dimensional Polya’s distribution times a density proportional to the product of of independent Polya’s densities and a multivariate normal density. For the final move we apply the likelihood ratio identity, making sure that the \( \epsilon \)’s are selected so that the support of the of the numerator in contained in the part of the support of the denominator over which the denominator is strictly positive:

\[
\frac{1}{\alpha_n} \mu_n([0, \delta] \times A) = \int \int \frac{1_{[0, \delta]}(y) \cdot 1_A(z)}{h(y)k(z)} d\nu_n(y) \\
\lim_{n \to \infty} \int \int \frac{1_{[0, \delta]}(y)}{h(y)} h(y) dy \cdot \int \frac{1_A(z)}{k(z)} k(z) f(z) dz = \delta \cdot P(Z \in A) ,
\]

which completes the proof.
where $f$ is the density of $Z$. This completes the proof of the theorem for bounded events $A$. Unbounded events may be intersected with a bounded event of $Z$ probability $1 - \epsilon$. The error in the limit in assessing the probability that involves the original set is no more than $\epsilon \phi(x)$. Letting $\epsilon \to 0$ deals with this general case.

The effect of having non-zero values for the sequence of vectors $\mu_n$ is the addition of the term $i\langle \mu_n, w \rangle$ to the joint characteristic function, which is then reflected in the limit gaussian density. The bound on the approximation of the joint characteristic function for all components of $X_1$ but the first has an addition of $o(\langle \mu_n, w \rangle)$ term. These changes do not change the argument.

Finally, we would like to deal with the case of a non-zero bounded function $m$. Choose $\epsilon > 0$ and approximate $m$ from above by a simple function $\hat{m} = \sum_j (j\epsilon)1_{B_j}$ such that $\hat{m} \leq m + \epsilon$. Using the partition $\{B_j\}$ we get:

$$P(S_n + m(S_n) \times A) = \sum_j P(S_n + m(S_n) \times (x_n, x_n + \delta) \times A \cap B_j) .$$

Each probability in the sum can be bounded over the event $B_j$ by:

$$P(S_n + m(S_n) \times A \cap B_j) \leq P(S_n \in (x_n - (j+1)\epsilon, x_n - j\epsilon + \delta) \times A \cap B_j) .$$

The event $A \cap B_j$ is a function of the last $d$ coordinates of the sum. Without loss of generality, this event has a zero boundary. It follows from the proof for a zero $m$ function that:

$$\lim_{n \to \infty} \sqrt{n}P(S_n \in (x_n - (j+1)\epsilon, x_n - j\epsilon + \delta) \times A \cap B_j) = (\delta + \epsilon)\phi(x)P(Z \in A \cap B_j) .$$

Taking a sum over $j$ and moving to a limit establishes the upper limit:

$$\limsup_{n \to \infty} \sqrt{n}P(S_n + m(S_n) \times (x_n, x_n + \delta) \times A) \leq (\delta + \epsilon)\phi(x)P(Z \in A) .$$

A lower limit can be obtained in a similar way by replacing $\delta + \epsilon$ by $\delta - \epsilon$. The theorem follows from letting $\epsilon$ converge to 0.

5.5 A remark on edge effects and higher order approximations

TBA
CHAPTER 5. THE LOCALIZATION THEOREM
Part II

Applications
Chapter 6

Copy Number Variations
6.1 Introduction

In the past decade or so, the story of biology in general, and genetics in particular, was the story of the development of high throughput technology for reading biological sequences, primarily DNA and RNA sequences. At least, this has been the impression I got from my exposure to the subject. Immediately after a new platform for RNA expression was announced everybody hurried to use that platform and publish a list of genes that show elevated expression under some specified conditions. Likewise, a purchase of a new genotyping machine resulted in a rush of publications that associated that gene or the other with the phenotype close to the researcher's heart. I am not complaining. They all need statisticians to help them analyze their data and this is a way for people like myself to make a living.

And statisticians, or more generally data-management and data-analysis people, are definitely needed. The data that is generated by these technologies is characterized by two features: quantity and error. The statement about quantity is clear. The number of measurements that are produced by a single application of a platform, a number that was in the order of several hundreds only a decade ago and was considered revolutionary, is now reaching the order of several millions and the technologies continue to develop. Handling this multitude of data is a profession.

But the issue of systematic and random error is as important and should be recognized and dealt with. What enabled the revolution was the development of technologies for performing experiments in parallel and on a very small space. In principle, the same type of experiments that were originally conducted individually, each in its own testing tube, is now done on tiny bids or on the surface of a chip, all at once and using the same chemical solution. Sometimes my kids have difficulties not interfering with each other's affairs when they are together at home. They are only 4 and the apartment is of a reasonable size. Now condense it and multiply it by a million... Still, you are reading this text not in order to learn about my domestic affairs but in order to relate the theory of random fields to real life problems. In this chapter we will deal with a specific application that arises in the context of biological sequence data and see the relevance of random fields for handling the issue of random noise.

The problem is the detection of inherited DNA copy number variants (CNV). CNV are gains and losses of (usually very small) segments of the DNA molecule in a chromosome, and comprise an important class of genetic variation in human populations. The modern approach for detecting CNV involves measurements of genetic markers as probes.

Genetic markers are mapped loci on the genome where genetic measurements can be made. In the simplest form, the measurement involves an intensity that reflects the copy number at that locus. The standard copy number of 2 is associated with some average intensity level. Variation of this number, say to 0, 1, 3, or any other number is associated with a change in that average intensity: a decrease if the copy number is below 2 and an increase if it is more than 2. A huge number of such probes, each reflecting a distinct position on the genome, are measured in parallel for the same individual.

As promised, these measurements involve errors. Consequently, it is difficult to determine on the basis of a single measurement the presence of a CNV at
a given locus or the absence thereof. A partial solution is to use repeated measurements. However, applying the platform is costly. Applying it more than once on each subject, especially when the experiment involves a sample of subjects, is currently still prohibitively expensive. A more economical solution is to average measurements in neighboring loci in order to produce a typical intensity for a genomic interval. This approach has the advantage of increasing the strength of a signal if the same CNV extends over several markers. On the other hand, the signal from a very short CNV may be diluted by including in the average intensities from markers that do not belong to the CNV. Nonetheless, this is the method of choice.

The resulting scanning statistic, the statistic that is produced by a standardized average of the intensities from markers in a given genomic interval, has approximately a normal distribution. This is the case, in particular, if data analyst did a proper job in the preliminary preparation and normalization of the data. This preliminary work is essential, and may involve the identification and removal of systematic errors, the handling of unwanted correlations, and the application of normalizing transformations to the data to make its distribution closer to the normal. Failing to do this step may produce results that reflect flaws in the technology instead of biological phenomena of interest.

Assuming the normality and the independence of the per-marker intensities, an assumption we henceforth make, we get that basic scanning statistic has a normal distribution. Moreover, if we consider the specifications of the interval, say the location of the center $t$ and the length of the interval $h$, as parameters we obtain that the resulting random field is gaussian. Moreover, under the null assumption of the absent of CNV in an entire section of the chromosome, we get that the distribution of the field is practically identical to the distribution that we obtained in Chapter 2 in the context of a scanning statistic that uses the interval $[-0.5, 0.5]$ as a kernel. In particular, we may use (4.2) to approximate the probability of a false detection in the interval. Alternatively, we may use the discrete version of this approximation, a version that uses the normal random walk overshoot function $\nu$, if we feel that markers are not dense enough to justify the continuous field approximation.

The field in this case of scanning for CNV in a single individual is gaussian and as such may be handled by traditional techniques such as the double-sum method. This example hardly serves as a justification for our method. Instead, we will present an analyze an alternative that is designed to use the data from a sample of subjects in order to identify inherited CNV that are present in the population. The approach that we consider was proposed in [30] and the statistical properties of the approach were discussed in [23, 24].

The detection of CNV is relevant in diagnostics of cancer. Part of the process that produces abnormal cancer cells results from pathogenic variations in the genetic material of somatic cells, an important part of which is in the form of CNV. Such variations are very irregular and are specific to the individual tumor. For such irregular behavior a method that concentrates on the data from that individual is more appropriate.

However, CNVs are also a relatively benign part of the normal genetic variation and can be passed on by gamete cells from parents to offsprings. Such variations may spread in the population and be present in a noticeable fraction of a sample from that population. Consequently, one may propose to accumulate data on the CNV in a given genomic interval from the sub-sample that
carries that CNV as yet another method to overcome random noise and enhance the chances of successful detection.

In the next section we formulate a statistical model for the data and describe the proposed method for accumulating data from sub-samples. The result is a new form of scanning statistic that is again parameterized by \( t \) and \( h \) but is not a gaussian field. In Section 6.3 we analyze the statistical properties of the scanning statistic via the application of the method of measure-transformation.

### 6.2 The statistical model

We assume that data from different subjects are independent of each other. We further assume that the data from each subject was normalized to produce measurements which, at the absence of a signal, are independent standard normal random variables. The observed data may be arranged as a two dimensional array \( \{ X_{it} : 1 \leq i \leq n, t_{0} < t \leq t_{1} \} \), where \( X_{it} \) is the data point for the \( i \)-th subject at location \( t \). The number \( n \) is the total number of subjects, and \( t_{1} - t_{0} \) is the total number of markers. In current days genome-wide scanning studies, \( n \) may range from several scores to thousands and the number of markers is several hundreds of thousands to about one or two million.

Given a genomic interval \([ t - h/2, t + h/2 ]\), parameterized by \( \theta = (t, h) \), we compute a summary statistic for the evidence from subject \( i \) regarding the presence of a CNV in the interval:

\[
Z_{i,\theta} = \frac{1}{\sqrt{h + 1}} \sum_{j=t-h/2}^{t+h/2} X_{ij}.
\]

This statistic would have been the scanning statistic had we considered only the data from subject \( i \). The statistic tends to obtain values in an interval about the origin if the subject does not carry a CNV, it has the standard normal distribution in this case, or else it tends to have a more extreme value – positive or negative – if it does carry a CNV.

The question is how to accumulate the data from a sample of subjects of which only part, if at all, are carriers of a CNV? Taking a sum may not be a good idea, especially if the carriers of the CNV are a small minority in the sample, since that will dilute the signal and make it harder to detect. Even worse, since a CNV can be expressed for some in the form of an increase in the intensity and for others in the form of a decrease in the intensity, the signal may be canceled out by the summation. This problem of opposite-sign cancelation may be solved by taking the absolute value of \( Z_{i,\theta} \) or squaring it before the summation, still leaving the dilution of the signal as a concern.

An alternative would be to sum the statistics, or their absolute values, only for the subjects in the sub-sample that carries a CNV. Unfortunately, we do not know who these subjects are. Yet, this is not a silly proposal since we do have data on each subject in the form of \( Z_{i,\theta} \). Thus we can identify an ad-hoc sub-group of patients with relatively elevated values of \( |Z_{i,\theta}| \) and form the sum for these subjects, and for these subjects only. A generalization of the same idea may apply a function \( g(z) \), a function that produces relatively small values when the absolute value \( z \) is not sufficiently large but produces high values when the
6.2. **THE STATISTICAL MODEL**

absolute value of $z$ is large. Such a function can be used in order to construct a scanning statistic:

$$Y_\theta = \sum_{i=1}^{n} g(Z_{i\theta}).$$  \hspace{1cm} (6.1)

If that scanning statistic is applied and detections declared whenever a threshold $y$ is crossed then the probability $P(\max_{\theta \in \mathcal{T}} Y_\theta \geq y)$ reflects the rate of false detections.

There is flexibility in the selection of $g$. Taking $g$ to be the identity function will produce the sum of the subject-specific scanning statistics. This approach was rejected as a bad one. Still, for this selection of $g$ we get that the sum has a gaussian distribution and we already have an approximation for the rate of false alarms for this case.

Taking the sum of squares of the subject-specific scanning statistics corresponds to choosing $g(z) = z^2$. The distribution of the scanning statistic will no longer be normal. Instead it will be the chi-square distribution on $n$ degrees of freedom. It requires work, but one may squeeze the theory of gaussian fields to allow results for this chi-square field. However, for other selections of $g$ the field will not be gaussian nor will it be linked to a gaussian field other than the general statement that with the increase in $n$ the field converges to a gaussian field.

For example, the idea of preselecting a sub-sample for producing the statistic may be obtained with a threshold function that is equal to zero for values less than the threshold and equal to the identity for values larger than the threshold. Such a function can be applied to the absolute value of the square of $Z_{i\theta}$. A similar approach, which is motivated by ideas stemming from mixture of distributions, is to use the function $g(z) = \log(1 - p + p \cdot \exp\{z^2/2\})$, where $p$ is a parameter that represents the frequency of the CNV in the population. This function is a smooth function, a fact that facilitates the analysis that we will carry below, although the analysis can be conducted with slightly more effort for the non-continuous threshold function.

The resulting random field $Y_\theta$, for both the threshold function or the mixture-type function, is not gaussian nor is it directly related to a random field in a way that allows application of traditional methods for gaussian fields. Admittedly, the limit distribution of the field when $n$ diverges to infinity is gaussian. However, we may remind the reader of the dangers in applying gaussian approximations to the extreme tail of a non-gaussian field. In the current setting, if the threshold in the threshold function is relatively large or if the parameter $p$ in the mixture-type function is small then the distribution of $Y_\theta$ will be severely skewed. A normal approximation, especially for values of $n$ not extremely large, will be questionable.

The inability to apply the standard theory of maxima in gaussian fields is exactly the property that we cherish. The standard gaussian techniques do not work but the measure-transformation approach does. In the next section we apply this approach to the random field $Y_\theta$ in order to obtain approximations of the rate of false discoveries. For convenience, we will make the assumption that the function $g$ is differentiable and leave the case where it is not as one of the projects/exercises that the reader may do if the reader wishes.
6.3 Analysis of statistical properties

We derive the asymptotic expansion of the probability $P(\max_{\theta \in \mathcal{T}} Y_\theta \geq y)$, for $Y_\theta = \sum_{i=1}^{n} g(Z_i, \theta)$, as a function of the threshold $y$ and the sample size $n$. The parameter set is $\mathcal{T} = [t_0, t_1] \times [h_0, h_1]$. In the expansion we will let both $y$ and $n$ grow to infinity. As you may read between the lines below, the analysis can be made rigorous also in the case where $n$ is kept fixed. However, the large sample asymptotic formulae are simpler and the proof, in light of Theorem 5.3, is more streamlined.

The first decision to make is which alternative distribution to use. There is no natural built-in alternative. Consequently, we choose to use the omnibus technique of exponential tilting. This technique can also help us identify the large deviation factor.

It is helpful to start by tilting the distribution of a statistic for a single subject and then extend to the entire sample. Define the log-moment generating function $\psi(\xi) = \log \mathbb{E} \exp\{\xi g(Z)\}$, where $Z$ is a standard normal random variable. Observe that $Z_i, \theta$ has the standard normal distribution for all $i$ and $\theta$. It follows that $\ell_i, \theta = \xi g(Z_i, \theta) - \psi(\xi)$ is a log-likelihood that is associated with the $i$-th sample. One may extend this basic construction in order to produce a tilted distribution for the entire sample by taking the sum and considering the likelihood ratio $\ell_\theta = \sum_{i=1}^{n} \ell_i, \theta$. Under the tilted distribution the samples are still independent of each other, each of them tilted according to the parameter $\xi$. The statistic $Y_\theta$, a sum of independent elements from the tilted distribution, inherits this distribution.

The large deviation properties of the marginal probability $P(Y_\theta \geq y)$, and as a byproduct, the value of the parameter $\xi$ that will be used may be obtained from the examination of the log-likelihood function of $Y_\theta$, which is $n\psi(\xi)$. The expectation of $Y_\theta$ under the tilted distribution is $n\psi'(\xi)$, where $\psi'$ is the derivative of $\psi$. The value of $\xi$ that is associated with the large deviation rate is the value that equates the expectation with the threshold:

$$n \psi'(\xi) = y \implies \xi = \xi(y, n) = [\psi']^{-1}(y/n),$$

which we use henceforth. With this selection of parameter $\xi$ the large deviation factor that will be produced is $\exp\{n[\psi(\xi) - \xi \psi'(\xi)]\}$. Since the distribution of the statistic $Y_\theta$ is the same for all $\theta$ we will get the same factor all across the parameter set.

A preparation step that precedes the application of the likelihood ratio identity is the initial localization and the approximation of the original parameter set by a discrete subset. In the current example this preliminary step is not needed. The marginal distributions of the random field are identical to each other for all values of the parameter, hence there is no subset of parameters that is associated with a larger than other large deviation factor. Likewise, the parameter space is discrete to begin with, no further action is required. However, if it is judged that the density of genetic markers, though discrete, is not consistent with the asymptotic derivation we intend to perform then our current decision not to apply this step may be revised. Furthermore, as a premonition we do want to maintain the flexibility to set the appropriate range of values of parameter set, namely the selection of $t_0$, $t_1$, $h_0$, and $h_1$ as a function of $y$ of $n$, to assure that the asymptotic derivation that we produce is meaningful. This
freedom of choice and the decision regarding the approximation of the parameter set by a subset are closely related and reflect the larger issue of determining appropriate scalings.

We turn to the transformation of the measure. We produced above an alternative distribution associated with a specific parameter value \( \theta \) and the random field element \( Y_{\theta} \) that is connected to it. This distribution, defined for a specific element, needs to be extended to the entire random field. We do so by defining the alternative distribution for the field in a way that retains \( \ell_{\theta} \) as the likelihood ratio. Hence, if \( dP \) is the joint density of all the elements in the field then we define the alternative distribution using the density \( dP_{\theta} = e^{\ell_{\theta}}dP \).

It is worthwhile to pay attention to the fact that the property of independence between subject-specific fields that was in the original distribution is still intact under the tilted distribution. Specifically, if the joint null distribution of the fields is characterized by the product \( \prod_{i=1}^{n} [dP_{i}] \), with \( dP \) the \( n \)-fold convolution of the \( dP_{i} \) densities, then the joint alternative distribution of the subject-specific fields is \( \prod_{i=1}^{n} [e^{\ell_{i,\theta}}dP_{i}] \). The distribution \( dP_{\theta} = e^{\ell_{\theta}}dP \) is the \( n \)-fold convolution of the densities in the square brackets.

With the given alternative distributions the likelihood ratio identity produces:

\[
P(\max_{\theta \in T} Y_{\theta} \geq y) = \sum_{\theta \in T} E_{\theta}\left( \frac{1}{\sum_{\theta \in T} e^{\ell_{\theta}}} ; \max_{\theta \in T} Y_{\theta} \geq y \right).
\]

The log-moment generating function \( n\psi(\xi) \) that appears in each of the log-likelihoods is the same for all values of \( \theta \). Consequently, it follows that the local field, the collection \( \{\ell_{\theta} - \ell_{\tilde{\theta}}\} \), is composed of elements of the form

\[
\xi(Y_{\theta} - Y_{\tilde{\theta}}) = \sum_{i=1}^{n} \xi[g(Z_{i,\theta}) - g(Z_{i,\tilde{\theta}})].
\]

After the usual rearrangement of terms in each of the expectations that were produced by the likelihood ratio identity these expectations become:

\[
E_{\theta}\left( \frac{1}{\sum_{\theta \in T} e^{\ell_{\theta}}} ; \max_{\theta \in T} Y_{\theta} \geq y \right) = e^{n[\psi(\xi) - \psi^{'}(\xi)]} E_{\theta}\left( \frac{M_{\theta}}{S_{\theta}} e^{-[\tilde{\ell}_{\theta} + m_{\theta}]} ; \tilde{\ell}_{\theta} + m_{\theta} \geq 0 \right),
\]

where

\[
\tilde{\ell}_{\theta} = \sum_{i=1}^{n} \xi[g(Z_{i,\theta}) - g(Z_{i,\tilde{\theta}})],
\]

\[
S_{\theta} = \exp \left\{ \sum_{\theta \in T} \xi[g(Z_{i,\theta}) - g(Z_{i,\theta})] \right\},
\]

and

\[
M_{\theta} = \max_{\theta \in T} \exp \left\{ \sum_{i=1}^{n} \xi[g(Z_{i,\theta}) - g(Z_{i,\theta})] \right\}, \quad m_{\theta} = \log M_{\theta}.
\]

We intend to apply the localization theorem, Theorem 5.2, to the expectations that are produced by the likelihood ratio identity with \( \kappa = n \), a parameter proportional to the variance of \( \tilde{\ell}_{\theta} \). The application involves determining a local \( \sigma \)-algebra \( \tilde{F}_{n} \) and checking the 5 conditions of the theorem. For the determination of the local field it is convenient to temporarily change the parametrization.
to \( \theta = (\theta_1, \theta_2) \), with \( \theta_1 = t - h/2 \) and \( \theta_2 = t + h/2 \). With this in mind, we set

\[
\hat{F}_n = \sigma^2 \left( \sum_{i=1}^n g[Z_i, \theta] g[Z_i, \theta] \right): |g_j - g_{ij}| \leq \tau, j = 1, 2, \text{ for some } \tau. \]

Of the 5 conditions, Condition I is trivial, Condition IV follows directly from the local limit theorem and Condition III*, in light of the local limit theorem, is an immediate corollary of the identification of the limit expectation and covariance structure of the Gaussian field that results from taking the limit of the local field, jointly with the global term.

In order to understand the issues involved in the derivation of the limit distribution let us consider in detail the expectation of a component of the local field of the form: \( \sum_{i=1}^n \xi[g(Z_i, \theta) - g(Z_i, \theta)] \).

The expectation of the sum is the sum of expectations, each taken under the alternative distribution for the component in the sum. The statistic \( Z_{i, \theta} \) is a sufficient statistic for that alternative distribution. Namely, the conditional distribution, given the statistic, is the same as the conditional distribution under the alternative distribution for the component in the sum. The statistic \( Z_{i, \theta} \) is the maximum of a Gaussian field to the limit Gaussian distribution of the standardized statistics \( \xi g[Z_i, \theta] - g[Z_i, \theta] \).

The asymptotic approximation produced by a naïve application of the formulas for the formation of the statistic \( Z_{i, \theta} \) is the same as the asymptotic approximation produced by a naïve application of the formulas for the maximum of a Gaussian field to the limit Gaussian distribution of the standardized statistics \( Y_\theta - n \psi'(0) / [n \psi''(0)]^{1/2} \).

A second justification can be made on a more scientific basis. The level of the required threshold \( y \) is a reflection of
the affective number of elements in the field, which in our case is of the order of magnitude of the product of the number of genetic markers times the sample size, since we will take \( h \) to be proportional to \( n \). The number of markers is several orders of magnitude larger than the sample size \( n \), and that should be reflected in a very high threshold \( y \) if an effective control on the rate of false detection of CNVs is intended.

With the choice we made, let us return to the expectation of the element of the local field that is given in (6.3). We are in a situation where \( r \) is converging to 1 and \( g \) is a smooth function. Define \( V_r = V = \{1 - r^2\}^{1/2} W + (r - 1) Z \) and take a second order Taylor expansion of the function \( g \) about \( Z \) to get:

\[
\begin{align*}
n \xi E_\theta \left( [g(Z_{1, \theta}) - g(Z_{1, \theta})] \right) &= n \xi E \left( [g(Z + V) - g(Z)]e^{\xi g(Z) - \psi(\xi)} \right) \\
&= n \xi E \left( [g'(Z)V + \frac{1}{2} g''(Z) V^2] e^{\xi g(Z) - \psi(\xi)} \right) + O([1 - r]^3 n) \\
&= -(1 - r) n \xi E \left( [Z g'(Z) - g''(Z)] e^{\xi g(Z) - \psi(\xi)} \right) + O([1 - r]^2 n),
\end{align*}
\]

that follows from the independence between the standard normal variables \( W \) and \( Z \) and the fact that \( 1 - r^2 \approx 2(1 - r) \). Integrating by parts we get that

\[
\int g''(z) e^{\xi g(z) - \frac{1}{2} z^2} dz = -\int \left[ \xi g'(z) \right]^2 - z g'(z) e^{\xi g(z) - \frac{1}{2} z^2} dz,
\]

which implies that:

\[
n \xi E_\theta \left( [g(Z_{1, \theta}) - g(Z_{1, \theta})] \right) \approx -n[1 - \text{Cov}(Z_{1, \theta}, Z_{1, \theta})]\xi^2 E \left( [g'(Z)]^2 e^{\xi g(Z) - \psi(\xi)} \right).
\]

(6.4)

For the variance of an increment of the local field we may use independence of the random fields to conclude that the variance of the sum is the \( n \) times the variance of an single random local field. Here it is sufficient to take a first order Taylor expansion of the function \( g \):

\[
\begin{align*}
 n \text{Var}_\theta \left( \xi [g(Z_{1, \theta}) - g(Z_{1, \theta})] \right) &= n \xi^2 \text{Var} \left( [g(Z + V) - g(Z)]e^{\xi g(Z) - \psi(\xi)} \right) \\
&= n \xi^2 \text{Var} \left( [g'(Z)V] e^{\xi g(Z) - \psi(\xi)} \right) + O([1 - r]^2 n) \\
&= (1 - r^2) n \xi^2 E \left( [g'(Z)]^2 e^{\xi g(Z) - \psi(\xi)} \right) + O([1 - r]^1 n) \\
&\approx -2n E_\theta \left( \xi [g(Z_{1, \theta}) - g(Z_{1, \theta})] \right),
\end{align*}
\]

with the passage from the second to the third line being partially justified by the computation of the variance via the conditioning on the value of the random variable \( Z \). We obtained that the asymptotic variance of an increment is twice the absolute value of the asymptotic expectation of the increment, a characteristic feature of a gaussian log-likelihood ratio when testing for a non-zero mean vector. Specifically in this case, an asymptotic expansion of the covariance between two increments of the local field, which can be conducted by considering again a first-order Taylor expression, demonstrates that the limit distribution of the local random field is that of a two-sided random walk of gaussian likelihood ratios for testing the hypothesis that the expectation of the increments is some fixed value. We omit the details.

For the asymptotic expression in Condition III* we may use the fact that:

\[
1 - \text{Cov}(Z_{\theta}, Z_\theta) \approx \frac{1}{2h} (\theta_1 - \theta_1) + \frac{1}{2h} (\theta_2 - \theta_2),
\]
to obtain expressions that are similar to the expressions that emerged in the similar situation of a scanning statistic with an indicator of an interval as a kernel, an example that was among the basic examples of Chapter 2. Specifically, in the current case and in the context of Condition III*, we get that 

\[ E[M/S] = \left\{ I(\theta) \nu\left( [2I(\theta)]^{1/2} \right) \right\}^2 \]

where \( \nu \) is the overshoot function of a gaussian likelihood ratio random walk and:

\[ I(\theta) = \frac{n \xi^2}{h} E_0\left\{ g'(Z)^2 \right\} = \frac{n \xi^2}{h} \int [g'(z)]^2 e^{\xi g(z) - \psi(\xi)} \phi(z) dz = \frac{n}{h} \cdot \iota(\xi) \quad (6.5) \]

is the expectation of an increment of the random walk.

The expectation of the global term \( \tilde{\ell}_n \) is \( \mu = 0 \) by construction. The variance of the global term is:

\[ \text{Var}_\theta(\tilde{\ell}_n) = n \xi^2 \text{Var}_\theta(g(Z_{1,\theta})) = n \xi^2 \psi''(\xi) = n \sigma^2. \]

Jumping our guns, we may predict that as a result of the integration step, if we take \( h_0 = n \eta_0 \) and \( h_1 = n \eta_1 \), and if we change to the variable \( w = h/n \), then we will obtain the approximation:

\[ P\left( \max_{\theta \in \Theta} Y_\theta \geq y \right) \approx \frac{e^{n[\psi(\xi) - \xi\psi'(\xi)](t_1 - t_2)/2}}{(2\pi n \xi^2 \psi''(\xi))^{1/2}} \int_{\eta_0}^{\eta_1} \left\{ (\iota(\xi)/w) \nu\left( [2\iota(\xi)/w]^{1/2} \right) \right\}^2 dw. \]

(6.6)

However, before reaching this conclusion we still need to finish validating the appropriateness of Theorem 5.3 for the current problem and to check the two remaining conditions of the Localization Theorem 5.2.

In order to complete the checking of the conditions for Theorem 5.3 we want to show that the global term and an increment of the random field are asymptotically uncorrelated. Taking a first order Taylor expansion we obtain:

\[ n \xi^2 \text{Cov}_\theta(g(Z_{1,\theta}), g(Z_{1,\theta}) - g(Z_{1,\theta})) \approx n \xi^2 E\left[ \{g(Z) - \psi'(\xi)\} [g'(Z) V] e^{\xi g(Z) - \psi(\xi)} \right]. \]

The covariance between two random variables is the expectation of the product of the centered random variables. It is enough to center one of the two variables. We arrived at the approximation by considering the centered global term and a first order approximation of the increment of the random field. We used once more independence between subject-specific fields to represent the covariance as a sum of covariances.

The resulting approximation of the covariance is of the order of a constant for \( V = \sqrt{1 - r^2} W - (1 - r) Z \), since the term associated with \( W \) makes a zero contribution and \( n(1 - r) \) converges to a constant. Consequently, the correlation between the global term and the local field is of order \( n^{-1/2} \) and converges to zero.

Regarding the rest of the unchecked conditions of Theorem 5.2 they can be verified using methods that are not unlike the methods that were used in the context of gaussian scanning statistics.

For example let us look at Conditions V*. For the same reasons that were given for scanning statistics it is sufficient to bound the right tail of \( M_\theta \).
6.3. ANALYSIS OF STATISTICAL PROPERTIES

trivial bound can be provided by taking:

\[
P_\theta(M_\theta > e^x) = P_\theta \left( \max_{\theta \in T} \exp \left\{ \sum_{i=1}^{n} \xi_i [g(Z_{i, \theta}) - g(Z_{i, \theta})] \right\} > e^x \right)
\]

\[
\leq |T| P_\theta \left( \exp \left\{ \sum_{i=1}^{n} \xi_i [g(Z_{i, \theta}) - g(Z_{i, \theta})] \right\} > e^x \right) \leq |T| e^{-x},
\]

by the fact that \( \sum_{i=1}^{n} \xi_i [g(Z_{i, \theta}) - g(Z_{i, \theta})] = \ell_\theta - \ell_\theta \) is a log-likelihood ratio under the \( P_\theta \) distribution and the application of the Markov inequality.

The derivation of (6.6) requires that \( t_1 - t_2 \gg h_1 - h_0 = (\eta_1 - \eta_2)n \). Consequently, \( |T| \) should grow faster than \( n^2 \) but, otherwise, any polynomial rate will do. Consequently, \( x = (c+0.5) \log n \) is sufficient for \( c \) larger than the polynomial order of \( |T| \).

Condition II is the last on the list. This condition deals with the probability \( P_\theta(A_{ii}^c \cap \{ \ell_\theta + \hat{m}_\theta \in x + (0, \delta) \}) \), for a truncated \( \hat{m}_\theta \) and for \( x \) in a relatively narrow range. We may apply the same approach that was used for a gaussian scanning statistic:

\[
P_\theta(A_{ii}^c \cap \{ \ell_\theta + \hat{m}_\theta \in x + (0, \delta) \}) \leq P_\theta(A_{ii} \cap \{ -m < X_\theta \leq \delta \}) + P_\theta(\{ X_\theta + \hat{m}_\theta \in (0, \delta) \} \cap \{ \hat{m}_\theta > m \})
\]

for \( X_\theta = \ell_\theta - x \) and a finite \( m \).

If \( m \) is large enough to make the probability of the event \( A = \{ \hat{m}_\theta > m \} \) small then we will have by Theorem 5.3 that the last probability, multiplied by \( \sqrt{n} \), converges to a quantity no larger than \( P(A)/\sqrt{2\pi} \).

For the first of the two probabilities in the bound we consider the sum of probabilities:

\[
\sum_{\tau < \| \theta - \theta \| \leq \log n} P_\theta(\{ \ell_\theta - \ell_\theta \geq \log(\epsilon p_\theta) \} \cap \{ -m < X_\theta \leq \delta \})
\]

\[
+ \sum_{\| \theta - \theta \| > \log n} P_\theta(\ell_\theta - \ell_\theta \geq \log(\epsilon p_\theta))
\]

This statement follows from the fact that the event \( A_{ii} \) is included in a union. For elements in the sum that are associated with remote parameters we ignore the intersection with the event that involves the global term.

We use log-moment generating functions in order to bound the probabilities, either for small values of \( \| \theta - \theta \| \) or for larger values. In both cases we consider the conditional log-moment generating function, given the values of \( \{ Z_{i, \theta} \} \), evaluated at \( 1/2 \). Specifically we are interested in terms of the form:

\[
\psi_r(\xi/2|Z) = \log E \left( \exp \left\{ (\xi/2)[g(V_r + Z) - g(Z)] \right\} | Z \right)
\]

for \( V_r = (1 - r^2)^{1/2} W - (1 - r)Z \).

The main concern is for values of \( \theta \) for which \( 1 - r \) is small. For such values we expand the function \( g \), taking a two-term Taylor approximation:

\[
\psi_r(\xi/2|Z) = \log E \left( \exp \left\{ (\xi/2)[g'(Z)V_r + 0.5g''(Z)V_r^2] \right\} | Z \right) + R .
\]

If the function \( g \) has a locally bounded third derivative then the error term \( R \) is \( O_p((1 - r)^{3/2}) \) and may be ignored, even after multiplication by \( n \), for \( r \) such that
(1 - r)n(\log n)^{-1} \to 0. For larger r it may not produce a vanishing term after multiplication by n, but it cannot significantly change the effect of the leading term.

Keeping track only of terms that make a non-vanishing contribution to the conditional moment generating function yields:

\[
g'(Z)V_r + \frac{1}{2}g''(Z)V_r^2 \approx (1 - r^2)\frac{1}{2}g'(Z)W + (1 - r)g''(Z)W^2 - (1 - r)Zg''(Z).
\]

This term should be multiplied by \(\xi/2\), exponentiated, and integrate with respect to the density of \(W\). Notice that the term associated with \(W^2\) that appears in the exponent of the density can be combined, changing in effect the variance, which is 1 in the original density, to \(1 - \xi(1 - r)g''(Z)\)\(^{-1}\). The expectation is still equal to zero. The outcome from the integration of the approximation is:

\[
\psi_r(\xi/2)Z \approx -(1 - r)(\xi/2)Zg'(Z) - \frac{1}{2}\log(1 - \xi(1 - r)g''(Z))
\]

\[
+ \frac{\xi^2}{8}(1 - r^2)[g'(Z)]^2(1 - \xi(1 - r)g''(Z))
\]

\[
\approx -(1 - r)(\xi/2)\{Zg'(Z) - g''(Z) - (\xi/2)[g'(Z)]^2\}.
\]

Consider the approximation of the conditional log-moment generating function as a function of \(Z\). Recall that \(E_\theta[Zg'(Z) - g''(Z)] = \xi E_\theta[(g'(Z))^2]\). Consequently, the expectation of the approximation is \(-(1 - r)(\xi/2)E_\theta[(g'(Z))^2]\). The variance is proportional to \((1 - r)^2\), i.e. of a much smaller order.

Define \(U_\theta = \sum_{i=1}^n\{\psi_r(\xi/2)Z_{i,\theta} - E_\theta\psi_r(\xi/2)Z_{i,\theta}\}\), with \(r = \text{Cov}(Z_{i,\theta}, Z_{i,\theta})\). The bound on the probability will follow from the relation:

\[
P(\ell_\theta - \ell_\theta \geq \log(ep_\theta)\{Z_{i,\theta}\}) \leq \exp\{nE_\theta\psi_r(\xi/2)Z_{1,\theta} - (1/2)\log(ep_\theta) + U_\theta\}.
\]

Consequently, Condition II* is a result of the examination of the bound:

\[
\frac{c(m + \delta)}{\sqrt{n}} \times \sum_{\tau < ||\theta - \theta|| < \log n} e^{-\left(n/2\right)E_\theta\psi_r(\xi/2)Z_{1,\theta}} + \sum_{\log n < ||\theta - \theta||} \frac{e^{-\left(n/2\right)E_\theta\psi_r(\xi/2)Z_{1,\theta}}}{\sqrt{ep_\theta}}
\]

\[
+ \sum_{\tau < ||\theta - \theta||} P_\theta(U_\theta > \left(n/2\right)E_\theta\psi_r(\xi/2)Z_{1,\theta})\).
\]

The constant \(c\) that appears in the term that multiplies the first sum is produced by the application of the Berry-Esseen theorem to \(X_\theta\), assuming that a third moment for the increments exist. Selecting \(p_\theta \propto \exp\{-\varepsilon(\xi/2)E_\theta\psi_r(\xi/2)Z_{1,\theta}\}\), for a small enough \(\varepsilon > 0\), will assure that the first two sums are \(o(n^{-\delta/2})\).

In order to complete the story we need to deal with the probabilities. Restricting ourselves to the case where \(E_\theta\{\psi_r(\xi/2)Z_{i,\theta} - E_\theta\psi_r(\xi/2)Z_{i,\theta}\}\)\(^4\), the centered forth moment, is finite and asymptotic to \((||\theta - \theta||/\sqrt{n})^4\) we may use the bound:

\[
P_\theta(U_\theta > \left(n/2\right)E_\theta\psi_r(\xi/2)Z_{1,\theta}) \leq \frac{E_\theta(U_{\theta}^4)}{\{\left(n/2\right)E_\theta\psi_r(\xi/2)Z_{1,\theta}\}^4} \approx \frac{c}{n^2||\theta - \theta||^4},
\]

for some limit constant \(c\). But, \(\sum_{\theta \in \mathcal{R}}||\theta - \theta||^{-4} < \infty\), uniformly in \(n\). Therefore, the contribution of the sum of probabilities is negligible.
This completes the validation of the conditions for Theorem 5.2 and establishes the approximation \( (6.6) \). It goes almost without saying that this approximation can be generalized to the case where the probability does not converge to zero with the aid of a poisson approximation. The generalization applies the function \( f(\lambda) = 1 - \exp\{-\lambda\} \) to the term that is given on the right-hand side of \( (6.6) \). Interestingly enough, the poisson approximation has further implications in this example which we would like to discuss.

### 6.4 The False Discovery Rate (FDR)

The context in which this book is written is controlling the error rate. The computations are conducted under the null distribution where signals are completely absent. The probability that is being computed is associated with wrongly declaring the presence of a signal when we should have not done so. The motivation for wanting to verify that the probability of falsely rejecting the null hypothesis is small is the worry that the consequences for such error may be severe and should be avoided as much as reasonable.

There are other situations, among them the detection of CNVs, where making an error as mentioned is not too catastrophic. Usually, one may apply other laboratory techniques, that are albeit more costly per locus, in order to validate that an identified CNV is genuine. Hence, one may consider the application of the high throughput technology as a screening step which is used in order to eliminate the parts of the genome that do not carry such variations in order to concentrate subsequent efforts on the parts that most likely express the variations. Still, one may not what the remaining regions to be polluted by too many false positive. A reasonable approach to address this concern is to use a statistical method that controls the rate of false detection.

In the statistical literature such methods fall under the acronym FDR, that stands for false discovery rate. The standard technique deals with a setting that involve independent random variables and do not fit the situation of a random field with dependence between elements. However, there are methods that can be used in the given context.

One such method will call for a more careful definition of what constitutes a discovery. In the case of screening for CNVs, for example, one may relate a region in which the the threshold was crossed to a discovery, which may be true or false. Still, the exact meaning of the notion of a region may need to be specified but we will not worry about it here. The false discovery rate is the expected ratio of false discoveries to the total of discoveries. The method makes sure that this expected ratio is not above some pre-specified level.

Notice that we no longer deal with a distribution which is generated purely from the null distribution. Currently, we are in a situation where some parts of the parameter set are associated with a genuine signals were as in other parts there is only random noise. The computations that we carried out were not designed to work in such a setting. However, if we restrict the computations to the part of the parameter space devoid of true signals, presumably the greater part of the parameter set, then the results that we obtained are still valid.

In order for the above-mentioned method for controlling the false discovery rate to work two conditions should be met. The first is that the number of true discoveries is asymptotically independent of the number of false discoveries.
The other condition is that the asymptotic distribution of the number of false discoveries is poisson. There is no other restriction on the distributions involved. The results that we obtained are relevant for the justification of the second condition.

Indeed, under a reasonable association of the count of discoveries with the number of connected excursion sets that are associated with a given threshold, we will get that this count follows, approximately, the poisson distribution. The poisson probability of the count $k$ is given by the application of the function $f_k(\lambda) = e^{-\lambda} \lambda^k / k!$ to the expression on the right-hand side of (6.6). This computed value of $\lambda$ is used for the application of the method.
Chapter 7

Sequential Monitoring of an Image

TBA
Chapter 8

Buffer Overflow

TBA
Chapter 9

Computing Pickands’ Constants

TBA
Chapter 10

Detecting a Change with a Network of Sensors

TBA
Bibliography


Appendix A

Mathematical Background
A.1 Transforms

Given a random variable $X$, the moment generating function that is associated with its distribution is

$$ M_X(\theta) = E[e^{\theta X}], $$

for all $\theta$ for which the function is finite. The function is non-negative and convex. The moments of the distribution coincide with the derivatives of the function evaluated at the origin. The moment generating function determines the distribution uniquely, provided that the origin is in the interior of the range of the function. Moment generating function of a sum of independent random variables is the product of their moment generating functions.

The cumulant generating function is the log of the moment generating function:

$$ \psi(\theta) = \log M(\theta). $$

We will typically call it the log-moment generating function. The cumulants are the derivative of the cumulant generating function, evaluated at the origin. The first cumulant is the expectation and the second cumulant is the variance. The log-moment generating function of the normal distribution with mean $\mu$ and variance $\sigma^2$ is

$$ \theta \mu + \theta^2 \sigma^2 / 2. $$

A family of distribution measure is of an exponential class in the natural form if the density, with respect to an appropriate measure, can be written in the form

$$ f_\theta(x) = h(x) \exp\{\theta \cdot t(x) - \psi(\theta)\}. $$

The sufficient statistic $t(x)$ and the natural parameter $\theta$ can be both vectors, in which case the product becomes an inner product. The expectation of $t(X)$ is given by the derivative $\psi'(\theta)$ and the variance is the second derivative $\psi''(\theta)$. A sum of independent and identically distributed random variable from an exponential family form an exponential family. Many, but not all, models of statistical interest are from an exponential family.

One may produce a family of distributions of an exponential class using the log-moment generating function $\psi(\theta)$. The log-likelihood ratio with respect to the baseline density, namely the log of the ratio between the density of the distribution associated with $\theta$ and the density associated with 0, is $\theta \cdot t(x) - \psi(\theta)$.

The large deviation rate for a sum of i.i.d. random variables may be obtained with the aid of the log-moment generating function:

$$ \lim_{n \to \infty} n^{-1} \log P(S_n \geq xn) = -I(x), $$

where $S_n$ is a sum of the random variables, $I(x) = \sup\{x \theta - \psi(\theta) : \theta \in \mathbb{R}\}$, and $x > E(X_1)$.

The characteristic function of a random variable $X$ is the moment generating function evaluated at $\theta = it$. It is defined via:

$$ \hat{\psi}(t) = Ee^{itX} = \int e^{itx}dF_X(x). $$

Thus, it is the Fourier transform of the distribution of $X$. The definition extends to a random vector $X$ by taking $t$ to be a vector and replacing the product $tX$ by the inner product $\langle t, X \rangle$. The characteristic function uniquely determines the distribution of $X$ as is guaranteed by the inversion theorem:

$$ \lim_{T \to \infty} \int_{-T}^{T} \frac{e^{-ita} - e^{-ith}}{it} \hat{\psi}(t) dt = P(a < X < B) + \frac{P(X = a) + P(X = b)}{2}. $$
A.2. APPROXIMATIONS OF SUM OF INDEPENDENT RANDOM ELEMENTS

When the distribution has a density with respect to the Lebesgue measure and the characteristic function is integrable then the inverse transform is

\[ f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{\psi}(t) e^{-itx} dt \]

The characteristic function of the normal distribution with mean \( \mu \) and variance \( \sigma^2 \) is \( \exp\{ -t\mu - t^2\sigma^2/2 \} \).

The error in the approximation of a characteristic function by a polynomial constructed with the moments of the random variable is bounded by:

\[ \left| \psi(t) - \sum_{m=0}^{n} \frac{(it)^m E(X^m)}{m!} \right| \leq E \min \left( \frac{|tX|^n+1}{(n+1)!}, \frac{2|tX|^n}{n!} \right). \]

In particular, if \( E(X) = \mu \) and \( \text{Var}(X) = \sigma^2 < \infty \) then \( \psi(t) = 1 + it\mu - t^2\sigma^2/2 + o(t^2) \), and the error term is bounded by \( 6t^2\sigma^2 \).

A.2 Approximations of sum of independent random elements

The classical method for dealing with the distribution of the sum of independent random variable uses approximations of the characteristic function of the sum. The Central Limit Theorem in its general form was originally proved this way as well as the Berry-Esseen theorem. This theorem states that for a standardized sum of i.i.d. random variables \( Z_n \):

\[ \sup_{x \in \mathbb{R}} |P(Z_n \leq x) - \Phi(x)| \leq C \frac{\mu_3}{\sigma^3 \sqrt{n}} , \]

for \( \mu_3 \) the centered third moment of an increment in the sum, \( \sigma^2 \) the variance, and \( C \) some universal constant. With more conditions the result may be refined.

For example, consider a generalization to higher dimensions. A distribution \( Q \) on \( \mathbb{R}^d \) is said to satisfy the Cramér’s condition if:

\[ \limsup_{\|t\| \to \infty} |\psi(t)| < 1 , \]

where \( \psi(t) \) is the characteristic function of \( Q \). Let \( \sum_{r=0}^{k-2} n^{-r} P(-\Phi_{\Sigma}, \{\chi_r\}) \) a higher order approximation of a distribution with respect to the Gaussian distribution, given in terms of the cumulants of the distribution. Define

\[ M_s(f) = \begin{cases} \sup_{x \in \mathbb{R}^d} (1 + \|x\|^s)^{-1} |f(x)| & s > 0 , \\ \sup_{x, y \in \mathbb{R}^d} |f(x) - f(y)| & s = 0 . \end{cases} \]

and let

\[ \tilde{\omega}_f(2e^{-cn}, \Phi_{\Sigma}) = \int_{\{y : \|y-x\| \leq 2e^{-cn}\}} |f(y) - f(x)| \Phi_{\Sigma}(dx) . \]

Theorem A.1 (Theorem 20.1 in [3]). Let \( X_n \) be i.i.d. sequence of random vectors in \( \mathbb{R}^d \), whose common distribution \( Q_1 \) satisfies Cramér’s condition. Assume that \( Q_1 \) has mean zero and a finite kth absolute moment for some integer
APPENDIX A. MATHEMATICAL BACKGROUND

Let $\Sigma$ denote the covariance matrix of $Q_1$ and $\chi_\nu$ its $\nu$-th cumulant ($3 \leq |\nu| \leq k$). Then for every real-valued, Borel-measurable function $f : \mathbb{R}^d \to \mathbb{R}$ satisfying $M_s(f) < \infty$, for some $0 \leq s \leq k$, one has that:

$$\left| \int f \left( Q_n - \sum_{r=0}^{k-2} n^{-\frac{r}{2}} \mathbb{P}(-\Phi_{\Sigma}, \{\chi_\nu\}) \right) \right| \leq M_s(f) \delta(n) + c(k,d) \bar{\omega} f(2e^{-cn}, \Phi_{\Sigma}),$$

where $Q_n$ is the distribution of $n^{-\frac{1}{2}} \sum_{i=1}^{n} X_i$, $\Phi_{\Sigma}$ is the multivariate normal distribution with 0 mean and variance-covariance matrix $\Sigma$, $c$ is an absolute constant, $c(k,d)$ constants that depend on the dimension of the space and the level of approximation, and $\delta(n) = o(n^{-(k-2)/2})$. All terms that are not otherwise specified do not depend on $f$.

A.3 Random walks

A random walk is a process produced by partial sums of i.i.d. random variables $S_n$. Statements about the distribution associated with the process are of interest. For example, the Law of Large Numbers states that if the expectation of an increment is finite then $(1/n)S_n \to E(X_1)$, almost surely.

Results may be obtained in relation to a random walk stopped by a stopping time. For example, Wald’s identity states that if the expectation of an increment is finite and if the expectation of the stopping rule $N$ is finite then $E(S_N) = E(X_1)E(N)$.

A useful inequality that can be proved using stopping times is Kolmogorov’s maximal inequality that states that if the increments are centered and the variance of an increment $\sigma_i^2$ is finite then

$$\mathbb{P}\left( \max_{1 \leq k \leq n} |S_k| \geq x \right) \leq \frac{1}{x^2} \sum_{i=1}^{n} \sigma_i^2.$$

In the book, when we discuss the asymptotic distribution of an overshoot for a random walk stopped by a bound we refer to a theorem that appears in Feller’s book. The theorem is:

**Theorem A.2** (Theorem 4 in [1], Vol. II Chapter XII.7). Consider the process $S_n$ of partial sums of independent random variables. Let $\tau_n = \mathbb{P}(S_1 \leq 0, \ldots, S_{n-1} \leq 0, S_n > 0)$ and let $p_n = \mathbb{P}(S_1 > 0, \ldots, S_{n-1} > 0, S_n > 0)$. Define the generating functions $\tau(s) = \sum_{n=1}^{\infty} \tau_n s^n$ and $p(s) = \sum_{n=1}^{\infty} p_n s^n$. Then

$$\log p(s) = -\log(1 - \tau(s)) = \sum_{n=1}^{\infty} \frac{s^n}{n} \mathbb{P}(S_n > 0).$$

A.4 The gaussian distribution

The standard normal distribution has the density $\phi(x) = (2\pi)^{-\frac{1}{2}} \exp\{-x^2/2\}$ over the real line and the cumulative distribution function $\Phi(x) = \int_{-\infty}^{x} \phi(z)dz$. The cumulative distribution function does not have a closed form, but a good
asymptotic approximation, for large values of \( x \), can be obtained with the aid of Mill’s ratio:

\[
\frac{x}{x^2 + 1} \phi(x) \leq \Phi(x) \leq \frac{1}{x} \phi(x).
\]

Modifications can be used to obtain better evaluations near the origin.

A random variable \( X \) has normal distribution if it is a linear transformation of a standard normal random variable: \( X = aZ + b \), for \( Z \) standard normal. In general, a random element is gaussian if any linear functional applied to it is normally distributed.

A Gaussian vector is characterized by the vector of expectations of the components and the matrix of covariances between components. Decompose a gaussian random vector \( X = (X_1, X_2)' \) into sub-vectors. Also, decompose accordingly the vector of expectations and matrix of covariances:

\[
\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.
\]

The conditional distribution of \( X_1 \), given the value of \( X_2 \) – denoted the regression of \( X_1 \) on \( X_2 \) – is gaussian. If \( \Sigma_{22} \) is of full rank then

\[
E(X_1 | X_2) = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2), \quad \text{Var}(X_1 | X_2) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.
\]

In particular, the conditional covariance structure of \( X_1 \) does not depend on the values of \( X_2 \).

Two central results that are used in the analysis of extremes in a centered gaussian random field are the Borell's inequality and the Slepian's inequality. The former gives a bound on the extreme tail and the latter makes a comparison between two random fields with one being more inter-correlated than the other.

The Borell’s inequality is similar to the Fernique’s inequality (Theorem 3.2):

**Theorem A.3 (Borell’s inequality).** Let \( \{X_t : t \in T\} \) be a centered gaussian field with almost surely bounded realizations. Let \( \|X\| = \sup_{t \in T} X_t \) and let \( \sigma^2 = \sup_{t \in T} \text{Var}(X_t) < \infty \). Then \( E\|X\| < \infty \) and for all \( x > 0 \):

\[
P(\|X\| - E\|X\| > x) \leq 2e^{-\frac{x^2}{2\sigma^2}}.
\]

Bounds on the term \( E\|X\| \) may be obtained via considerations of entropy.

The Slepian’s inequality introduces another centered random field \( \{Y_t : t \in T\} \) with the same variance: \( \text{Var}(Y_t) = \text{Var}(X_t) \), for all \( t \in T \). Then:

**Theorem A.4 (Slepian’s inequality).** If \( \text{Cov}(X_t, X_s) \geq \text{Cov}(Y_t, Y_s) \), for all \( (s,t) \in T \times T \), then for all \( x \):

\[
P(\|X\| > x) \leq P(\|Y\| > x).
\]

### A.5 Integration

Some facts regarding integration and convergence:

**Theorem A.5 (Fubini’s theorem).** Suppose \( X \times Y \) is a complete measure space equipped with the product measure \( d(x,y) = dx \times dy \) and suppose that \( f(x,y) \) is a measurable and integrable function then:

\[
\int_X \left( \int_Y f(x,y)dy \right) dx = \int_Y \left( \int_X f(x,y)dx \right) dy.
\]
Let \( \{f_n\} \) be a sequence of real valued measurable functions over a measure space. Assume that \( f_n(x) \to_{n \to \infty} f(x) \), for almost all \( x \). Then:

**Theorem A.6** (Dominated convergence theorem). If \( |f_n(x)| \leq g(x) \) and \( g \) is a measurable function with \( \int g(x)dx < \infty \) then \( \int f(x)dx < \infty \). Moreover, we have that \( \lim_{n \to \infty} \int f_n(x)dx = \int f(x)dx \).

A translation of this theorem to the language of random variables will imply from the point-wise convergence \( X_n(\omega) \to X(\omega) \) and the relation \( |X_n| \leq Y_i \) with \( EY < \infty \), the convergence of the expectations: \( \lim_{n \to \infty} E(X_n) = E(X) \). For the convergence of the expectations to hold it is sufficient that \( X \) is uniformly integrable.

A more general concept is the concept of uniform integrability: We say that a collection of random variables \( \{X_n\} \) is uniformly integrable if for all \( \epsilon > 0 \) one can find a universal \( x \) such that:

\[
E(|X_n|; |X_n| > x) \leq \epsilon , \quad \text{for all } X_n \text{ in the collection}.
\]

This concept is tightly linked to the concept of converges of expectations. Given a sequence of random variables \( \{X_n\} \) that converges in probability to a random variable \( X \) then \( \lim_{n \to \infty} E|X_n - X| = 0 \) if, and only if, the collection \( \{X_n\} \) is uniformly integrable.

### A.6 Poisson approximation

The poisson distribution serves as an approximation to the number of occurrences of independent rare events. Consider, for example, a binomial random variable \( X \) associated with the count of the number of successes among \( n \) trials with probability of success \( p = \lambda/n \) in each. It is easy to show that if \( np_n \to \lambda \) then \( \lim_{n \to \infty} P(X_n = x) = e^{-\lambda}\lambda^x/x! \), for any non-negative integer \( x \).

More sophisticated theorems will establish a similar type of convergence when the events do not have identical probabilities and/or are weakly dependent. For example, in the book we use the theorem:

**Theorem A.7** (Theorem 1 in [2]). Let \( \{X_i : i \in I\} \) be Bernoulli random variables. Set \( \hat{W} = \sum_{i} X_i \) and \( \lambda = \sum_{i} P(X_i = 1) \in (0, \infty) \). Associate with each \( i \) a neighborhood of dependence \( I_i \subset I \). Define:

\[
\begin{align*}
  b_1 &= \sum_{i \in I} \sum_{j \in I_i \setminus \{i\}} P(X_i = 1)P(X_j = 1) \\
  b_2 &= \sum_{i \in I} \sum_{j \in I_i \setminus \{i\}} P(X_i = 1, X_j = 1) \\
  b_3 &= \sum_{i \in I} E\left[ E(X_i | \sigma\{X_j : j \notin I_i\}) - E(X_i) \right]
\end{align*}
\]

Then

\[
2 \sup_{A} |P(\hat{W} \in A) - P(W \in A)| \leq 2(b_1 + b_2 + b_3) ,
\]

where \( W \) is a poisson random variable with expectation \( \lambda \). Also,

\[
|P(\hat{W} = 0) - e^{-\lambda}| \leq (1 - \lambda^{-1})(b_1 + b_2 + b_3) .
\]
Appendix B

Proposed Projects

TBA