学位論文
Doctoral Dissertation

Unsupervised Anomaly Detection within Non-Numerical Sequence Data
(非数値系列データにおける教師無し異常検出)

シテファン　ヤン　スクドラレク
Stefan Jan Skudlarek

Contents

Acknowledgement

I would like to thank first and foremost my supervisor, Professor Hirosuke Yamamoto, who accepted me as his student, helped me with the application for the scholarship financing my stay in Japan, and supported me in every aspect of my research, on the same hand urging me to work independently.

Secondly, I became greatly indebted to one of my master thesis supervisors, Dr. Nascimento, for encouraging me to pursue a Ph.D. degree and referring me to Professor Yamamoto for supervision.

I would like to thank for the recommendations of two of my former teachers, Professor Hagenauer and Professor Eberspaecher, and of my former employer, Mr. Matthias Meyer, which ensured the grant of the scholarship mentioned above. In addition, Mr. Meyer also supported me by allowing the premature cancellation of the employment contract and my absence from work for attending the interview crucial for scholarship application.

For emotional support during my stay in Japan, I have to thank my family and friends.

Chapter 1

Introduction

1.1 Review of Anomaly Detection

The term anomaly detection refers either to the problem of detecting data not fitting an expected or normal behavior defined previously or to the problem of detecting data diverging from the majority of a set of data given. These two descriptions mark the partition of the research regarding the problem into supervised and unsupervised approaches.

- (Semi)Supervised Anomaly Detection: Based on a labeled set of normal training data describing the expected behavior, we compute a score (binary or continuous) expressing the likelihood of a newly presented test data point to conform with the expected behavior.
- Unsupervised Anomaly Detection[1]: We are given a set of unlabeled data, and are presented with the task of evaluating every data point based on the assumption that the majority of the data is normal.

While the fundamental problem was first defined within the statistics research community during the late 19^{th} century, more widespread interest in the topic did not occur until the rise of computer networks and large scale automated plants during the second half of the 20^{th} century. The technological progress enforced the development of methods for automatically judging the state of a complex system - for example a power plant, an aircraft engine, or a production line - based on the output data - pressure or altitude real

[1]This term is sometimes also used to refer to supervised anomaly detection as defined above when discriminating it from an anomaly detection approach using training data to deduce a model of both the normal and the erroneous data.

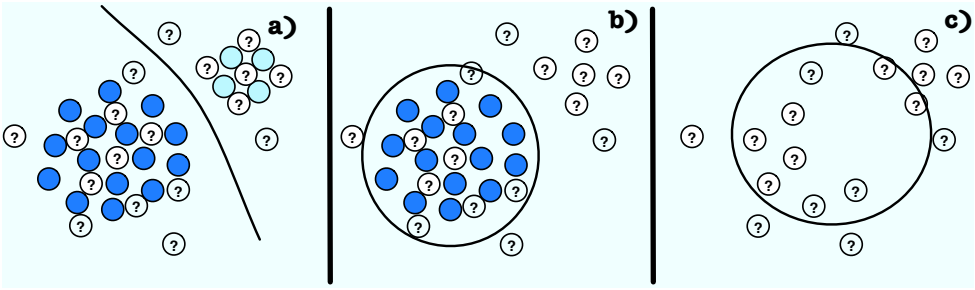

Figure 1.1: a) Conventional two-class classifier built from sample data of both the normal class (dark) and the anomalous class (light) to classify test data (question marks) b) Supervised anomaly detection: one-class classifier built from sample data of the normal class (dark) c) Unsupervised anomaly detection: classifier built from unlabeled test data

values, power down binary values, access rate etc. - generated by the system. The system could be in a normal or in an erroneous state. The generic approach to the problem had been to properly define both erroneous and normal output, turning the situation into a two-class discrimination problem as depicted by Fig. 1.1 a). But while the definition of the output characteristic of the normal behavior of a complex system is fairly easy, defining the output characteristic of every possible erroneous system state may be quite difficult.

Thus, in order to circumvent the problem, the concept of supervised anomaly detection was applied, defining any observation deviating from the expected normal output, any abnormal data, as indicative of an erroneous system state. Figure 1.1 b) shows a classifier based on this concept. While early applications where mostly related to industrial production [1], the advance of computer, multimedia, and network technology during the second half of the eighties created new application fields like computer network intrusion detection [2], financial fraud detection [3], satellite image analysis [4], and medical imaging [5].

A new chapter was opened with the rise of complex dynamic systems like wireless networks, the internet, or mobile robots, during the last decade of the 20^{th} century. Earlier applications of supervised anomaly detection had mostly been set in fixed environments, for example a computer pool with a limited number of terminals and registered users, making it easy to collect the normal training data necessary for supervised anomaly detection. In a dynamic and anonymous environment like the internet, on the other hand, clean training data becomes difficult to acquire because the desired application may change and evolve rapidly. Therefore, the scenario of unsupervised anomaly detection is more realistic, raising the problem of building a classifier from unlabeled data as shown by Fig. 1.1 c).

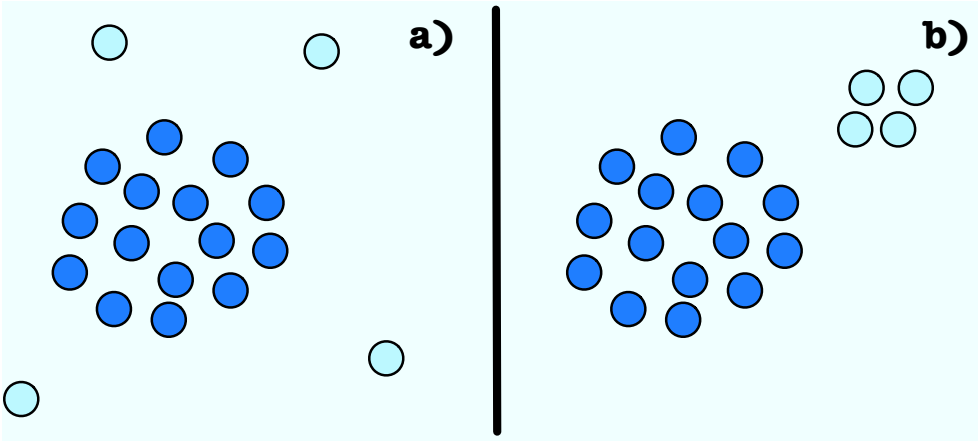

Figure 1.2: a) Heterogenous anomalies (light) spread around the bulk of the normal data (dark) b) Homogenous cluster of anomalous data

While the concept of anomaly detection is convenient and intuitive, there is an aspect easily overlooked when defining anomalous data points (which are also referred to as outliers) within a data set of unlabeled data, as in the unsupervised scenario. In his frequently quoted definition of outliers, Grubbs [6] states that

"An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs."

This definition echoes the statistical roots of anomaly detection, but leaves one important point ambiguous: If several anomalous data points or outliers are present in a set of data, how are they related to each other? If the anomalous data is a heterogenous group, the only common feature is the deviation from the normal data, which forms the majority of the set. Detection is easy compared to the case of of a homogeneous anomalous data subset, because it is sufficient to search for data deviating from all the other data points. Figure 1.2 illustrates the point, using a geometric example. The first subfigure shows a situation reminiscent of a two dimensional distribution, with the outliers simply being extreme values scattered around the bulk of the observations, distorting the estimation of parameters such as mean and covariance. Contrary, the second subfigure shows a homogeneous cluster of abnormal data, so Hawkin's definition of an outlier [7] as an

"observation that deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism."

is more befitting. The second situation might arise upon permanent malfunction of a measurement device, or mixing data from different sources (such as gene sequences from different organisms). Hereafter, we will refer to the first situation as *point-based* anomaly and to the second situation as *cluster-based* anomaly, a denomination introduced by Lian et al. [8].

The numerous anomaly detection techniques (supervised and unsupervised) proposed so far (for an exhaustive list of publications see Chandola et al. [9]) may roughly be divided into the subsequent four groups. Note, however, that those techniques are not mutually exclusive but often used in combination. For example, principal component analysis may be used to preprocess the vectors used for training a support vector machine.

- Anomaly Detection Based on Classification
  A classifier is built using the available training data, judging the test data as either normal or anomalous. Techniques used include one-class support vector machines, neural networks, principal component analysis, as well as rule based approaches.

  One-class support vector machines [10] are a derivative of the original support vector machine algorithm [11], which had been designed to minimize the structural risk when separating two classes by a linear classifier, first mapping the data to a suitable high-dimensional space via a so called Kernel function [12] if necessary to solve non-linear separation problems. Instead of a function separating two classes, the one-class support vector algorithm optimizes a function enclosing the majority of the training data, judging any data outside this boundary as anomalous. While showing good generalization capabilities, the computational expense of the quadratic optimization procedure involved has spawned several simplifications, which exploit the fact that after a suitable mapping, the training data will form a hypersphere (or hyperellipsoid) [13] [14]. This turns the problem into a minimum-enclosing-ball problem. As for unsupervised anomaly detection, the above approaches are robust with respect to point-based anomalies, but cluster-based anomalies might seriously undermine detection capability.

  Artificial Neural networks [15] use a network consisting of units modeled after biological neurons to build a nonlinear classifier function, using techniques likes radial basis functions, hopfield networks, or self-organizing maps. While training is easy, there is a tendency for overfitting, and for the unsupervised case, even a small percentage of anomalies might result in serious distortion.

  Another family of methods very suitable to the supervised anomaly detection scenario are rule-based approaches [16]. In contrast to the cohesive approach of statistical modeling discussed below, rule-based approaches generate a set of rules characterizing the normal training data, assigning a confidence value dependent on the frequency of occurrence. A test data point fitting only rules with low confidence is likely to

be anomalous. This approach is especially suitable for sparse categorical training datasets [17] [18] featuring a large alphabet, when statistical modeling is likely to yield unstable results. The unsupervised application of rule-based approaches is difficult, since even a minor share of anomalous data may generate faulty rules or lower the confidence of correct rules.

Classification by principal component analysis [19] belongs to a larger family of methods called spectral methods, which aim at reducing the data complexity (number of dimensions, mapping of time series to frequency) while improving the discrimination of anomalous and normal data. Principal component analysis performs a linear transformation of the coordinate system, using an eigenvector decomposition according to the covariance matrix of the training data. Within the new system, the dimensions are uncorrelated, and the share of variance carried by one dimension decreases with rising order of dimension (and is identical to the eigenvalue of the dimension). This enables the exclusion of dimensions carrying little information. Using normalization by the eigenvalues of the remaining dimensions, the probability distribution of the normal data can be approximated using the chi-square distribution if the normal data obeys a multivariate normal distribution. While originally proposed for the supervised scenario [20] (with some filtering of outliers of the normal data), recently also applicability to unsupervised anomaly detection has been explored [21].

A common strong point of classifier systems is that while the training may take a considerable amount of time, classification of the testing data is usually fast. The main common drawback is the dependence on clean training data of most methods, especially regarding rule based approaches and neural networks. Also, often no meaningful score regarding the reliability of the judgement is generated.

- Anomaly Detection Based on Data Distance
  After defining a suitable distance measure for the data, these approaches compute the matrix of pairwise distances of the data points in the set. Basically unsupervised, some algorithms may be modified to process labeled data for parameter estimation. The techniques may be divided into nearest-neighbor-based techniques and clustering techniques.

  k-nearest-neighbor-based techniques compute a numerical measure for any point based on the average distance to the adjoining k data points, thus estimating the local density. The most straightforward algorithm simply classifies the points of lowest nearest-neighbor distance within the set as anomalous [22]. It detects both point-based and cluster-based anomalies (sparse clusters are supposed to be anomalous). In order to account for density variations of the normal data, the local outlier factor [23] was introduced. Supposing only point-based anomalies, this approach calculates the ratio of the k-nearest-neighbor distance of the point in question, and the average of the k-nearest-neighbor distances of the k nearest neighbors. Points of

small ratio are points located outside a cluster above a certain density.

In contrast to the local estimation of the nearest-neighbor techniques, clustering techniques establish a global model [24]. Crisp partition-based techniques like the k-means algorithm [25] group the data into several non-overlapping groups or clusters and then judge a data point as anomalous if it is either member of a sparse cluster or located exceptionally far from its assigned cluster center. Fuzzy partition-based methods, on the other hand, assign to any point in the the data set a degree of membership in each of the clusters [26]. Data points with a small degree of membership in any cluster or high membership in sparse clusters are considered anomalous. Contrary to partition-based techniques, hierarchical techniques either gradually agglomerate the points of the data set into a single cluster [27] (bottom-up), or break down the global cluster (top-down), creating a hierarchy of the data points allowing for the detection of both point-based and cluster-based anomalies.

There have also been some attempts to combine elements of the two above approaches [28] [8], which are sometimes referred to as density-based clustering.

While distance-based approaches are unsupervised in nature and mostly independent of the data distribution, the quality of results depends on the setting of parameters, which often have to be found out by trial-and error. Another point is the quadratic computational complexity.

- Anomaly Detection Based on Statistics
  These techniques generate a statistical model based on the data. A data point is judged as anomalous if the likelihood of the data point within the estimated model drops below a certain threshold. The techniques may be divided into parametric and non-parametric techniques.

  Parametric approaches assume that the normal data has been generated by a particular class of distributions (for example Gaussian) or a mixture of several distributions [29], and try to infer the parameters of the distribution from the data given, using maximum-likelihood techniques like the expectation-maximization algorithm [30].

  Non-parametric approaches do not assume any particular distribution but adapt to the data. The most popular technique of this kind creates a histogram by counting the occurrences of data points within regions of equal size or bins of the data space, thus estimating the density [31]. If a test data point is located in a bin of low density, it is judged anomalous. The other frequently employed non-parametric technique places a standard conditional distribution above every data point, estimating the overall distribution by combining the conditional distributions [32].

  A special subgroup of supervised statistical approaches deals with the problem of online anomaly detection: Given a stream of data, evaluate the present data point based on the past data points judged to be correct [33] [34].

The advantage of statistical methods, which are mostly supervised, is a reliable score associated with a confidence interval if the distribution has been estimated correctly, detecting both point-based and cluster-based anomalies. However, many parameter estimation algorithms are very susceptible to noisy or sparse training data.

- Anomaly Detection based on Information Theory
  Information theoretic techniques are based on the observation that outliers within a set of data instances cause a notable inflation of information theoretic measures like entropy or complexity. For this purpose, the data is usually represented as a set of strings. The complexity of a string, as calculated by parsing algorithms of the type frequently used in compression algorithms, is an approximation of the theoretic Kolmogorov complexity [35]. The Kolmogorov complexity $C(x)$ of a string x is defined as the length of the shortest program that will output the string x. Extending the original thought, the conditional Kolmogorov complexity $C(x|y)$ is defined as the minimum program length for generating x given y. Most strings of a given length generated by the same source are supposed to show similar Kolmogorov complexities.

  Research regarding the application of information theory to anomaly detection has mostly focused on the supervised scenario. The basic assumption is that the source of the normal data is stationary or does at least produce similar output over time. Thus, normal sequences share more patterns with the training data than anomalous sequences do, and so the conditional Kolmogorov complexity with respect to the training data should be higher for anomalous sequences [36] [37]. While most supervised approaches are static, estimating the conditional Kolmogorov complexity based entirely on the training data, adaptive approaches start off with the training data and progressively add test data to the training data, in some cases dropping older training data. Besides standard information measures, also the use of model selection criteria (see 3.3.3) has been proposed [38]. In this adaptive test, one looks for a significant increase in the complexity of the optimum statistical model, which hints at the inclusion of anomalous data.

  For the unsupervised scenario, the standard approach searches for the minimum set of data points which, when removed from the calculation of the information theoretic measure with respect to the whole of the set, will result in the maximum decrease of the measure [39]. Another class of methods defines a mutual dissimilarity score (or pseudo-distance) of two sequences, based on the estimated single and concatenated complexity of the two strings [40] [41]. A suitable clustering algorithm may then be used to divide normal and abnormal data. However, mutual dissimilarity measures based on conventional universal compression algorithms not may not meet the conditions of a metric, and tend to become unstable for short sequence lengths.

1.2 Research Problem Definition

At about the same time unsupervised anomaly detection became interesting, the advancement of technology spawned numerous anomaly detection problems involving non-numerical sequence data, like fraud detection or protein classification. Due to the application-specific properties of real-world non-numerical sequence data, unsupervised anomaly detection for such data is especially difficult. Previous methods are hampered by difficulties of parameter selection and computational expense. Therefore, we chose this problem as the topic of our research. We define our scenario as follows:

1. We are given a set $\mathcal{S}$ of n sequences or blocks $x^{b_1}, x^{b_2}, \ldots, x^{b_{n-1}}, x^{b_n}$, with $x \in \mathcal{X} = \{a_1, a_2, \ldots, a_{Z-1}, a_Z\}$, with the alphabet size Z not fixed in advance. We assume that each b_i is a multiple of a minimum block length b. The index $i \in \{1, 2, \ldots, n-1, n\}$ of the sequences may either be assigned at random or indicate the temporal order of sequence generation.

2. We suppose that the majority of $1 - \rho$ ($0 \leq \rho \leq \rho_{\max} = 0.33$) of the sequences was generated by one stationary normal source N (one-class scenario), while the remaining share ρ of the sequences was generated by one or more stationary abnormal sources A. Note that this scenario includes the case of $\rho = 0$ (No anomalous data). The indices of the anomalous sequences may be random with respect to the overall index range, or cover a limited section of the range.

3. The task is to derive a measure for the normality of each sequence.

We developed two approaches dealing with this scenario, which may be categorized as distance based.

The first approach strings together the sequences of the set $\mathcal{S}$ into one global sequence, using the index difference or distance of identical symbols within this global sequence for anomaly detection.

The second approach computes the distance of any pair of sequences via a suitable kernel, retrieving a sequence representative of the normal data, the distance from which is then used for anomaly detection.

1.3 Non-Numerical Sequence Data and Unsupervised Anomaly Detection

Supervised anomaly detection for symbolic sequences [42] has been proposed for application in various fields, such as proteomics [43], flight safety [44], and computer intrusion detection [45] [46]. While also data distance based methods have been proposed, statistical methods and classification methods prevail.

Unsupervised anomaly detection within symbols sequences, on the other hand, has received less attention. Contrary to the case of supervised detection, the majority of methods is based on mutual data distance or dissimilarity. How to define the distance of two sequences of non-numerical data of probably different length is a fundamental problem of machine learning [47] [12]. In contrast to numerical data, no ready ordinal space exists.

For evaluation of our results, we implemented two unsupervised approaches representative of previous research: An unsupervised probabilistic suffix tree algorithm [48] and a clustering technique called fixed width clustering [22]. While the suffix tree algorithm is an information theoretic method, with encoding based on a statistical model of variable memory length [49], the clustering approach is based on the mutual distance of sequences.

The suffix tree algorithm first creates a probabilistic suffix tree based on the whole of the data $\mathcal{S}$ given. Using the tree, a dissimilarity measure for judging any sequence $x^b \in \mathcal{S}$ of b symbols is calculated as follows:

$$\begin{aligned} \mathrm{DSIM}\left(x^b\right) &= \frac{-\log P\left(x_1,\ldots,x_b\right)}{b} \\ &= \frac{-\log\left(P\left(x_1\right)\cdot P\left(x_2|x_1\right)\cdot\ldots P\left(x_b|x_1,\ldots,x_{b-1}\right)\right)}{b} \\ &= \frac{-\log\left(P\left(x_1\right)\cdot P\left(x_2|x_1\right)\cdot\ldots P\left(x_b|x_{b-t},\ldots,x_{b-1}\right)\right)}{b} \end{aligned} \tag{1.1}$$

For $b \to \infty$, this measure will converge towards the entropy rate of the t^{th} Markov source. Note that this measure has no upper limit, the lower limit being zero. In contrast to a criterion based on the frequency of occurrence of the respective subsequences used by Agrawal et al. [50] for supervised methods, Kam et al. [48] used the Corrected Akaike Information Criterion for setting the tree depth t i.e. the maximum memory length of the source description for calculating the dissimilarity measure. While the method performs well in case of small fixed symbol alphabet size Z, for a combination of long memory, large variable alphabet size, and small size of the data set $\mathcal{S}$, numerous sparse nodes are created, forcing the tree to suboptimal depth because of inflation of the parameter term within the Corrected Akaike Information Criterion. In order to suppress this effect, we have to fix a

minimum number $S_{\min}$ of occurrences of a preceding subsequence $x_{b-\gamma}, \ldots, x_{b-1}$ in order to become a node of the tree at level $\gamma \leq t$. Besides evaluating the performance of our algorithm, we used the optimum tree depth for deducing certain parameter values. We explain this in detail in Section 3.3.4.

The clustering algorithm defines a distance measure for any two symbol sequences in $\mathcal{S}$ using the normalized spectrum kernel function [51] controlled by the dimension parameter k. Having computed the distance of any two blocks in our data set, we may estimate the density of each block by counting the number of neighboring blocks within a certain radius w. Blocks of high density are likely to belong to normal data, while blocks of low density are likely to be outliers. In order to speed up calculation, the algorithm defines the first data block as the center of a cluster. If one of the following blocks falls within the range w of one or more blocks previously designated as cluster centers, it is assigned to those clusters and the respective count of the clusters is increased by one. If not, the block forms the center of a new cluster. After all sequences have been assigned, the local density of a block is defined as the density of the cluster it is assigned to. Multiple cluster assignment is resolved by taking the average. While the algorithm is easy to handle and many extensions have been proposed to speed up the calculation, the setting of the parameters w and k poses a difficult problem, and in practice is often done based on experience or trial-and-error.

1.4 Thesis Organization

This thesis is organized as follows:

The first chapter contains the introduction to the topic, problem definition and explanation of the notation used throughout the thesis.

The second chapter introduces the first approach, which was inspired by the work of Kennel [52] and Rieke et al. [53] [54], and is based on the idea of stringing together all sequences within $\mathcal{S}$ into one global sequence[2]. We use a function called the average index difference to create a numerical value for every symbol based on the distance of identical symbols within the global sequence. After introducing the function and explaining its properties, the first algorithm, which assumes the anomalous data to be clustered within a subsection of the global sequence, is presented, including the setting of the parameters and the global cost. Next, the second algorithm, which removes several drawbacks of the first one, most important allowing for arbitrary location of the anomalous blocks at the cost of increased complexity, is explained, again including parameter setting and computational cost. The chapter closes with an evaluation of the experimental results of both algorithms using both artificial and real world data.

[2]The contents of the chapter were presented to some extent at the ADMA2009 conference [55].

The third chapter introduces the second approach, which is based on selecting a sequence representative of the normal data and classifying the sequences in the set according to the distance from this representative, after initially using a suitable kernel for mapping the sequence data to a numerical space[3]. After stating the algorithm and explaining the parameter setting, we discuss the usage of kernels for mapping sequence data to a suitable numerical space, as well as the parameter setting of the kernel used for our experiments. After stating the computational cost, we close the chapter by discussing the results of experiments using both artificial and real-world data.

The last chapter contains the conclusion, summing up the work presented and pointing out possible topics of further research.

1.5 Notation

In this section, we present an overview of the notation used in the thesis in the order of first appearance in the text.

- $\mathcal{S}$: set of non-numerical sequences
- $\mathcal{X}$: alphabet which is the set of symbols appearing in $\mathcal{S}$
- Z: the size of the alphabet $\mathcal{X}$, i.e. $\mathcal{X} = \{a_1, a_2, \ldots, a_Z\}$
- n: number of sequences within $\mathcal{S}$
- b_i: length of the i^{th} sequence or block within $\mathcal{S}$ with $i \in \{1, \ldots, n\}$
- b: minimum block length. All b_i are considered multiples of b.
- ρ: share of anomalous sequences within $\mathcal{S}$
- $\rho_{\max}$: upper limit of ρ
- g: length of the global sequence created by concatenating the sequences of $\mathcal{S}$
- DSIM $\left(x^{b_i}\right)$: Dissimilarity measure for judging the irregularity of a sequence based on a given distribution
- $S_{\min}$: Minimum number of occurrences of a subsequence within $\mathcal{S}$ in order to create a node within a probabilistic suffix tree describing $\mathcal{S}$
- t: depth of the suffix tree structure

[3]The contents of the chapter were presented to some extent at the ICMLC2010 conference [56].

- w: radius parameter of the fixed width clustering algorithm
- k: dimensional parameter of the spectrum kernel deciding the length of the subsequences, the respective occurrences of which are then counted in order to create a numerical vector describing the sequence
- j_{A}: start index of anomalous data within the global sequence incase of concentrated anomalous data
- l_{A}: total length of anomalous data within the global sequence
- j: index within the global sequence of length g
- $P_X^{\mathrm{N}}(x)$: generation probability of symbol $x \in \mathcal{X}$ within normal data.
- $P_X^{\mathrm{A}}(x)$: generation probability of symbol $x \in \mathcal{X}$ within abnormal data.
- $h(x)$: ratio of $P_X^{\mathrm{N}}(x)$ and $P_X^{\mathrm{A}}(x)$ of $x \in \mathcal{X}$, i.e. $h(x) \stackrel{\mathrm{def}}{=} \frac{P_X^{\mathrm{N}}(x)}{P_X^{\mathrm{A}}(x)}$ for $x \in \mathcal{X}$ such that $P_X^{\mathrm{A}}(x) > 0$
- C_{b}: occurrences of symbols identical to the one at index j below j
- C_{a}: occurrences of symbols identical to the one at index j above j
- $T_j(x)$: average index difference of the symbol $x \in \mathcal{X}$ located at index j
- $\Delta(x)$: index difference of consecutive occurrences of $x \in \mathcal{X}$
- $\Delta_i(x)$: i^{th} index difference between consecutive occurrences of a particular $x \in \mathcal{X}$ counting from the start of the global sequence
- $\overline{\Delta}(x)$: mean value of index differences between consecutive occurrences of $x \in \mathcal{X}$
- j_i^{a}: i^{th} occurrence of the symbol $x \in \mathcal{X}$ found at index j counting up from j
- j_i^{b}: i^{th} occurrence of the symbol $x \in \mathcal{X}$ found at index j counting down from j
- j_i: i^{th} occurrence of the symbol $x \in \mathcal{X}$ counting from the start of the sequence
- τ_{th}: threshold for processing the average index difference values
- c_{th} threshold for processing the scalar value assigned to a block
- M: memory variable when searching for typical subsequences of symbols
- $M_{\max}$: upper threshold of the memory variable M

- $h\left(x^{M+1}\right)$: ratio of $P_X^{\mathrm{N}}\left(x^{M+1}\right)$ and $P_X^{\mathrm{A}}\left(x^{M+1}\right)$ of $x^{M+1} \in \mathcal{X}^{M+1}$, i.e. $h\left(x^{M+1}\right) \stackrel{\mathrm{def}}{=} \frac{P_X^{\mathrm{N}}\left(x^{M+1}\right)}{P_X^{\mathrm{A}}\left(x^{M+1}\right)}$ for $x^{M+1} \in \mathcal{X}^{M+1}$ such that $P_X^{\mathrm{A}}\left(x^{M+1}\right) > 0$
- h_{th}: threshold value of $h\left(x^{M+1}\right)$ for defining the $x \in \mathcal{X}$ considered representative of the anomalous data by the algorithm.
- $\Delta\left(x^{M+1}\right)$: index difference of consecutive occurrences of $x^{M+1} \in \mathcal{X}^{M+1}$
- $\Delta_i\left(x^{M+1}\right)$: i^{th} index difference between consecutive occurrences of a particular $x^{M+1} \in \mathcal{X}^{M+1}$ counting from the start of the global sequence
- $\overline{\Delta}\left(x^{M+1}\right)$: mean value of index differences between consecutive occurrences of $x^{M+1} \in \mathcal{X}^{M+1}$
- $C_{\text{b block}}$: occurrences of symbols identical to the one at index j below j within the same block
- $C_{\text{a block}}$: occurrences of symbols identical to the one at index j above j within the same block
- $\nu_{\min}$: lower threshold of the number of identical subsequences within the same block, $C_{\text{b block}} + C_{\text{a block}}$
- $S_j\left(x^{M+1}\right)$: scaled average index difference using the expected value of $T_j\left(x\right)$ in case of $l_{\mathrm{A}} = 0$ for normalization
- $a + 1$: dimension of the vector used for vector processing of $S_j\left(x^{M+1}\right)$
- β: parameter used to create a threshold via multiplication with $\overline{\Delta}\left(x^{M+1}\right)$ for judging the $\Delta_{C_{\mathrm{b}}}\left(x^{M+1}\right)$ and $\Delta_{C_{\mathrm{b}}+1}\left(x^{M+1}\right)$ in order to decide whether j is located inside anomalous data
- ξ: parameter used to create a threshold via multiplication with $\overline{\Delta}\left(x^{M+1}\right)$ in order to decide whether to exclude a particular $\Delta_i\left(x^{M+1}\right)$ from calculation
- $\Upsilon\left(k, x^{M+1}, l_{\mathrm{A}}\right)$: Parameter describing the share of the sequence length g covered by $\Delta\left(x^{M+1}\right)$ surpassing the threshold k
- m: mean of the scalar values of the normal blocks estimated via the median
- σ: deviation of the scalar values of the normal blocks estimated via the median absolute deviation.
- $\phi\left(x^{b_1}\right)$: Transforms the sequence x^{b_1} to an inner product space, calculating a vector of non-negative numerical entries.

- $K\left(x^{b_1}, x^{b_2}\right)$: calculates a numerical value describing the similarity of two sequences x^{b_1}, x^{b_2} as a dot product of the numerical vectors $\phi\left(x^{b_1}\right), \phi\left(x^{b_2}\right)$
- $K_{\text{norm}}\left(x^{b_1}, x^{b_2}\right)$: calculates a numerical value describing the similarity of two sequences x^{b_1}, x^{b_2}, which is normalized to the interval $[0, 1]$ by
  $K_{\text{norm}}\left(x^{b_1}, x^{b_2}\right) = \frac{K\left(x^{b_1}, x^{b_2}\right)}{\sqrt{K\left(x^{b_1}, x^{b_1}\right)} \cdot \sqrt{K\left(x^{b_2}, x^{b_2}\right)}}$
- $d_K\left(x^{b_1}, x^{b_2}\right)$: pseudo-distance or dissimilarity of two sequences x^{b_1}, x^{b_2} based on the kernel $K\left(x^{b_1}, x^{b_2}\right)$ by
  $d_K\left(x^{b_1}, x^{b_2}\right) = \sqrt{K\left(x^{b_1}, x^{b_1}\right) - 2K\left(x^{b_1}, x^{b_2}\right) + K\left(x^{b_2}, x^{b_2}\right)}$
- D: matrix of mutual distances of the n sequences of $\mathcal{S}$
- $\hat{D}$: matrix created from D by rearranging the elements of each row from smallest to largest
- θ: Parameter regulating the number of distances to be enclosed by the minimum radius of the algorithm.
- $C\left(x^{b_1}\right)$: Kolmogorov complexity of the sequence x^{b_1}, defined as the length of the shortest program able to generate x^{b_1}.
- $C\left(x^{b_1} | x^{b_2}\right)$: Conditional Kolmogorov complexity of the sequence x^{b_1} given x^{b_2}, defined as the length of the shortest program able to generate x^{b_1} given x^{b_2}.
- $N_{x^{b_1}}\left(x^k\right)$: A function outputting the number of occurrences of the subsequence x^k within the sequence x^{b_1}
- $B_{x^{b_1}}\left(x^k\right)$: A function outputting a binary result indicating the occurrence of the subsequence x^k within the sequence x^{b_1}
- $P_{\mathcal{X}}^{\mathcal{S}}\left(x_{t+1} | x^t\right)$: conditional probability of a symbol $x \in \mathcal{X}$ to appear in any of the sequences of the set $\mathcal{S}$, given a certain sequence x^t has appeared immediately before

Chapter 2

Unsupervised Anomaly Detection based on Average Index Difference

In this chapter we present the first approach for unsupervised anomaly detection within non-numerical sequence data. The approach concatenates the sequences of the set into one global sequence. It exploits the fact that in case of stationary ergodic symbol generation, the expected value of two consecutive occurrences of a symbol is the inverse of the generation probability.

We introduce a function called the average index difference, which performs a weighted comparison of the index differences of neighboring identical symbols to assign a numerical value to the symbol at index j within the global sequence. We show that the average index difference function converges to an expected value dependent only on the global index j, but not on the symbol generation probability, for the case of stationary ergodic symbol generation and no anomalies present. In case of anomalies, the average index difference values both within normal and abnormal data will deviate form the expected value in a way that enables the detection of anomalous data.

Two algorithms employing the deviation of the average function for anomaly detection are presented. The first algorithm supposes a local concentration of anomalous data. This requirement is lifted by the second algorithm, at the cost of increased complexity. Nonetheless, the parameters of both algorithms are shown to be theoretically deducible. Finally, we demonstrate the correctness of the theoretical predictions by experiments using both artificial and real-world data.

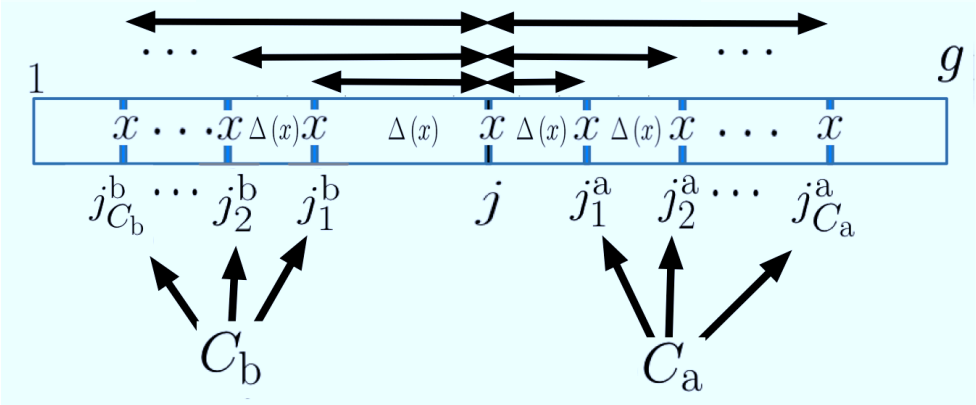

Figure 2.1: Notation of average index difference function $T_j(x)$. C_{b} represents the occurrences of x *below* j, while C_{a} represents the occurrences *above* j.

2.1 Average Index Difference Function

2.1.1 Definition

Consider a sequence of overall length g generated by a stationary ergodic normal data source N and a stationary ergodic anomalous data source A. It is partitioned into subsequences or blocks of length b, with

$$b \ll g. \tag{2.1}$$

A share of l_{A} symbols is anomalous data. A particular block contains either anomalous or normal data. One may also imagine the situation as the generating source switching between the normal and the anomalous state. l_{A} is assumed to be bounded by

$$0 \leq l_{\mathrm{A}} \leq \frac{g}{3} \tag{2.2}$$

and is a multiple of the block length b. For a particular symbol $x \in \mathcal{X}$, the expected value of the number of occurrences and the expected value of the index difference of two consecutive occurrences of x inside both the anomalous and the normal data are determined by the pair of stationary ergodic generation probabilities $P_X^{\mathrm{N}}(x), P_X^{\mathrm{A}}(x)$, related by

$$h(x) \stackrel{\mathrm{def}}{=} \frac{P_X^{\mathrm{N}}(x)}{P_X^{\mathrm{A}}(x)}. \tag{2.3}$$

In case of $P_X^{\mathrm{N}}(x) > 0, P_X^{\mathrm{A}}(x) = 0$, $h(x)$ is undefined.

In Section 2.2, we suppose that the l_{A} abnormal symbols are generated as a closed subsequence of symbols (a consecutive sequence of anomalous blocks) located somewhere within the the overall sequence g. Note that there may as well be no abnormal data at all.

$\Delta(x)$ represents the interval between two consecutive occurrences of $x \in \mathcal{X}$. The expected value of $\Delta(x)$ in either abnormal or normal state of the source for stationary ergodic symbol generation is given by

$$\mathrm{E}\left(\Delta(x)\,|\mathrm{A}\right) = \frac{1}{P_X^{\mathrm{A}}(x)} \tag{2.4}$$

or

$$\mathrm{E}\left(\Delta(x)\,|\mathrm{N}\right) = \frac{1}{P_X^{\mathrm{N}}(x)} \tag{2.5}$$

from Kac's Lemma [57](for a more formal statement see Lemma A.1 in the appendix), where E represents the expectation. The validity of the subsequent approach rests on these two expressions.

The average index difference $T_j(x)$ of the symbol x found at index j is defined as the average over the respective index differences of the symbol x found at index j and all symbols of identical value x within the whole of the sequence of length g. Using the notation depicted by Fig. 2.1, $T_j(x)$ is written as follows:

$$T_j(x) \stackrel{\mathrm{def}}{=} \frac{\sum_{o=1}^{C_\mathrm{b}} \left(j - j_o^{\mathrm{b}}\right) + \sum_{q=1}^{C_\mathrm{a}} \left(j_q^{\mathrm{a}} - j\right)}{C_\mathrm{b} + C_\mathrm{a}}. \tag{2.6}$$

In case of $C_\mathrm{b} + C_\mathrm{a} = 0$, the average index difference is defined to be zero. (2.6) may be rewritten using a notation of the index differences $\Delta(x)$ of neighboring identical sequences or symbols, indexing the $\Delta(x)$ from 1 to $C_\mathrm{a} + C_\mathrm{b}$, starting from the beginning of the sequence.

$$T_j(x) \;=\; \sum_{o=1}^{C_\mathrm{a}} \left(\frac{C_\mathrm{b} - o + 1}{C_\mathrm{b} + C_\mathrm{a}}\right) \cdot \Delta_{C_\mathrm{b}-o+1}(x) + \sum_{q=1}^{C_\mathrm{a}} \left(\frac{C_\mathrm{a} - q + 1}{C_\mathrm{b} + C_\mathrm{a}}\right) \cdot \Delta_{C_\mathrm{b}+q}(x) \tag{2.7}$$

Figure 2.2 shows a visualization of (2.7) for an example featuring $C_\mathrm{b} = 2$ and $C_\mathrm{a} = 6$. The average index difference function processes the $\Delta(x)$ alike to an asymmetric weighting window.

We define $\overline{\Delta}(x)$ as the empirical mean value of the index difference of neighboring identical subsequences or symbols observed within the overall sequence of length g.

$$\overline{\Delta} = \frac{\sum_{q=1}^{C_\mathrm{b}+C_\mathrm{a}} \Delta_q(x)}{C_\mathrm{b} + C_\mathrm{a}} \tag{2.8}$$

2.1.2 Properties of Average Index Difference Function in Case of No Anomaly

We only have to consider $P_X^{\mathrm{N}}(x)\ \forall x \in \mathcal{X},\ \forall j \in \{1, \ldots, g\}$. In order to calculate the expected value of the average index difference of a symbol x at index j within a symbol

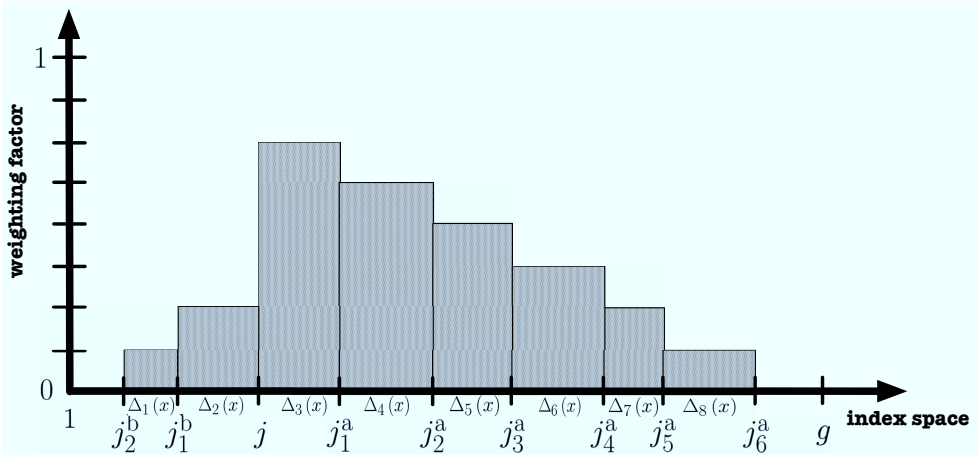

Figure 2.2: Visual example of the calculation of $T_j(x)$ based on $\Delta(x)$ for $C_{\mathrm{b}} = 2$ and $C_{\mathrm{a}} = 6$.

sequence of length g, we introduce the expected values of C_{b} and C_{a}:

$$\mathrm{E}(C_{\mathrm{b}}) = j \cdot P_X^{\mathrm{N}}(x) \tag{2.9}$$

$$\mathrm{E}(C_{\mathrm{a}}) = (g-j) \cdot P_X^{\mathrm{N}}(x) \tag{2.10}$$

Inserting (2.9), (2.10) and (2.5) into (2.6), the expected average index difference may be approximated by

$$\begin{aligned}
\mathrm{E}(T_j(x)) &\approx \frac{1}{g \cdot P_X^{\mathrm{N}}(x)} \left(\sum_{o=1}^{j \cdot P_X^{\mathrm{N}}(x)} \frac{o}{P_X^{\mathrm{N}}(x)} + \sum_{q=1}^{(g-j) \cdot P_X^{\mathrm{N}}(x)} \frac{q}{P_X^{\mathrm{N}}(x)} \right) \\
&= \frac{1}{g \cdot P_X^{\mathrm{N}}(x)^2} \left(\sum_{o=1}^{j \cdot P_X^{\mathrm{N}}(x)} o + \sum_{q=1}^{(g-j) \cdot P_X^{\mathrm{N}}(x)} q \right) \\
&= \frac{1}{g \cdot P_X^{\mathrm{N}}(x)^2} \left(\frac{\left(j \cdot P_X^{\mathrm{N}}(x)\right)^2}{2} + \frac{\left((g-j) \cdot P_X^{\mathrm{N}}(x)\right)^2}{2} + \frac{\left(g \cdot P_X^{\mathrm{N}}(x)\right)}{2} \right) \\
&= \frac{j^2 + (g-j)^2}{2g} + \frac{1}{2 \cdot P_X^{\mathrm{N}}(x)} \qquad (2.11) \\
&\approx \frac{j^2 + (g-j)^2}{2g} \qquad (2.12) \\
&\geq \frac{g}{4} \qquad (2.13)
\end{aligned}$$

The first term of expression 2.11 show the effect of taking the average of the index differences. The expected value of the average index difference mainly consists of the weighted

sum of the differences between the index j and centroid indices of closed areas of constant probability distribution *below and above* the index j. The weights consist of the average percentage of symbols contributed by the respective area.

The second term of expression (2.11) reminds us that in case of very small generation probability compared to sequence length g, above approximation formula becomes futile. Together with the lower bound of the remaining terms given by (2.13), this yields the subsequent condition:

$$\frac{1}{2 \cdot P_X^{\mathrm{N}}(x)} \ll \frac{g}{4} \tag{2.14}$$

Formula (2.12) is symmetric with respect to the index $\frac{g}{2}$. The symmetry is achieved because the average index *difference* is calculated. While the approximation presented here is informal, we formally deduced the formula by the subsequent theorems, the proofs of which are given in the appendix:

Theorem 2.1
For a sequence of g symbols $x^g \in \mathcal{X}^g$ generated by a single stationary ergodic source, for the expected value of the average index difference $\mathrm{E}(\mathrm{T_j}(\mathrm{x}))$ of the symbol found at index $j = \lceil g \cdot y \rceil$ $(0 < y < 1)$, the subsequent convergence holds

$$\lim_{g \to \infty} \frac{1}{g} \cdot \left| \mathrm{E}\left(T_j(x) \,|l_{\mathrm{A}} = 0\right) - \frac{j^2 + (g-j)^2}{2g} \right| = 0 \tag{2.15}$$

Theorem 2.2
For a sequence of g symbols $x^g \in \mathcal{X}^g$ generated by an i.i.d. source, the expected value of the average index difference $\mathrm{E}(T_j(x)\,|l_{\mathrm{A}} = 0)$ of the symbol found at index $j \in \{1, \ldots, g\}$ is given by

$$\mathrm{E}\left(T_j(x)|l_{\mathrm{A}} = 0\right) \;=\; \left(1 - \left(1 - P_X^{\mathrm{N}}(x)\right)^{g-1}\right) \cdot \left(\frac{j^2 + (g-j)^2}{2 \cdot (g-1)} + \frac{g - 2j}{2 \cdot (g-1)}\right). \tag{2.16}$$

Corollary 2.1
For a sequence of g symbols $x^g \in \mathcal{X}^g$ generated by an i.i.d. source, the expected value of the average index difference $\mathrm{E}(T_j(x)\,|l_{\mathrm{A}} = 0)$ of the symbol found at index $j \in \{1, \ldots, g\}$ obeys the relative bound

$$\left| \frac{\mathrm{E}\left(T_j(x)|l_{\mathrm{A}} = 0\right) - \frac{j^2 + (g-j)^2}{2g}}{\frac{j^2 + (g-j)^2}{2g}} \right| \;\leq\; \frac{4}{g-1} + \left(1 - P_X^{\mathrm{N}}(x)\right)^{g-1} \cdot \left(\frac{g+3}{g-1}\right). \tag{2.17}$$

Theorem 2.3
For a sequence of g symbols $x^g \in \mathcal{X}^g$ generated by i.i.d. source, the variance of the average index difference $\mathrm{E}(T_j(x))$ of the symbol found at index $j \in \{1, \ldots, g\}$ obeys the upper bound

$$Var(T_j(x)|l_\mathrm{A} = 0) \quad \leq \quad \frac{g + 39 + \frac{2}{g}}{12 \cdot P_X^\mathrm{N}(x)} \tag{2.18}$$

if $P_X^\mathrm{N}(x) \geq \frac{5}{g}$ holds.

2.2 Algorithm Supposing Local Concentration of Anomalous Blocks

2.2.1 Algorithm Statement

The average index difference values are used for block classification according to the subsequent algorithm:

A subsequence or block is classified as an anomaly if the percentage of symbols showing an average index difference below a certain limit τ_th surpasses a certain percentage threshold c_th. The approach is based on the assumption that the overall sequence of g symbols exhibits a single symbol subsequence of a certain length l_A generated by intrusion. Thus, the symbols characteristic of the anomaly (i.e. symbols frequently occurring inside the anomalous but not the normal data) will show a smaller average index difference, because the majority of the identical symbols will be located close to the symbol compared to the overall sequence length g.

A subsequence or block of length b is classified using the subsequent algorithm:

1. average index difference and percentage calculation:

   Calculate the percentage c of the b symbols featuring an average index difference of $T_j(x) \leq \tau_\mathrm{th}$ $(0 \leq \tau_\mathrm{th} \leq g)$. The average index difference of symbols only occurring once within the overall sequence of length g is defined to be zero.

2. percentage evaluation:

   If $c \geq c_\mathrm{th}$ $(0 \leq c_\mathrm{th} \leq 1)$, the block is classified as anomalous.

The performance of the algorithm is obviously determined by the setting of the parameter τ_th. The task of the step employing τ_th is to search for symbols which are characteristic

of the anomaly ($P_X^{\mathrm{N}}(x) \ll P_X^{\mathrm{A}}(x)$) when compared to the normal output *and* are located within the anomalous data. A pseudocode transcription of the algorithm is shown by Algorithm 1.

Algorithm 1 Block Classification Algorithm

Ensure: index difference threshold $\tau_{\mathrm{th}} \leq g$
Ensure: percentage threshold $c_{\mathrm{th}} \leq 1$
Ensure: block start index $j_{\mathrm{S}} \leq g - b$
  block variable $c \Leftarrow 0$
  index variable $j \Leftarrow j_{\mathrm{S}}$
  while $j < j_{\mathrm{S}} + b$ **do**
    if $T_j(x) \leq \tau_{\mathrm{th}}$ **then**
      $c = c + \frac{1}{b}$
    end if
    $j++$
  end while
  if $c \geq c_{\mathrm{th}}$ **then**
    classify block as anomalous
  else
    classify block as normal
  end if

2.2.2 Parameter Setting

As mentioned before, we suppose a single anomaly of length l_{A} starting at index j_{A}. Due to symmetry with regard to $\frac{g}{2}$, we may limit our consideration to $j \leq j_{\mathrm{A}} + l_{\mathrm{A}}$. We deduce bounds of the expected value of the average index difference, discriminating three cases for any $x \in \mathcal{X}$:

- Case 1: $P_X^{\mathrm{N}}(x) \ll P_X^{\mathrm{A}}(x) \leftrightarrow h(x) \ll 1$
  The symbol is characteristic of the anomalous data when compared to the normal data.

- Case 2: $P_X^{\mathrm{N}}(x) \approx P_X^{\mathrm{A}}(x) \leftrightarrow h(x) \approx 1$
  The symbol is characteristic of neither the anomalous nor the normal data.

- Case 3: $P_X^{\mathrm{N}}(x) \gg P_X^{\mathrm{A}}(x) \leftrightarrow h(x) \gg 1$
  The symbol is characteristic of the normal data when compared to the anomalous data.

Bounds of $\mathrm{E}\left(T_j\left(x\right)\right)$ for $j > j_\mathrm{A}$ First we examine the case of the reference symbol being located inside the anomaly. Figure 2.3 shows the centroid indices of both normal and anomalous areas R_1, R_2, R_3 and R_4. We define C_{R_1}, C_{R_2}, C_{R_3} and C_{R_4} as the

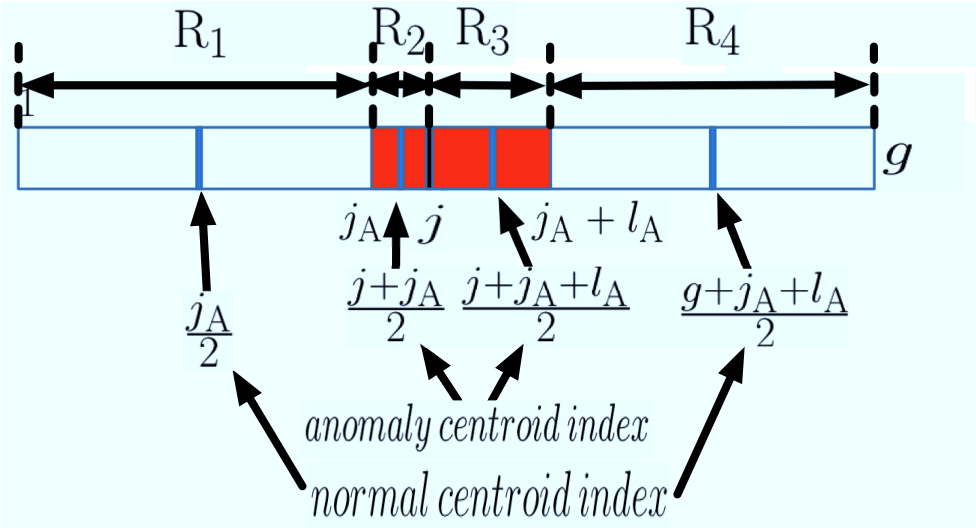

Figure 2.3: Reference symbol located inside anomalous data

occurrence numbers of x in the respective areas. Supposing $P_X^{\mathrm{N}}\left(x\right)$ and/or $P_X^{\mathrm{A}}\left(x\right)$ suitably large, the subsequent approximations hold:

$$C_{\mathrm{R}_1} \approx j_\mathrm{A} \cdot P_X^{\mathrm{N}}\left(x\right) \tag{2.19}$$
$$C_{\mathrm{R}_2} \approx \left(j - j_\mathrm{A}\right) \cdot P_X^{\mathrm{A}}\left(x\right) \tag{2.20}$$
$$C_{\mathrm{R}_3} \approx \left(j_\mathrm{A} + l_\mathrm{A} - j\right) \cdot P_X^{\mathrm{A}}\left(x\right) \tag{2.21}$$
$$C_{\mathrm{R}_4} \approx \left(g - j_\mathrm{A} - l_\mathrm{A}\right) \cdot P_X^{\mathrm{N}}\left(x\right) \tag{2.22}$$

$$\mathrm{E}_{\mathrm{R}_1} = \mathrm{E}\left(\left|j - j_i^{\mathrm{b}}\right| \left| j_i^{\mathrm{b}} \in \mathrm{R}_1\right.\right) \approx j - \frac{j_\mathrm{A}}{2} \;;\; i \in \{1, \ldots, C_\mathrm{b}\} \tag{2.23}$$
$$\mathrm{E}_{\mathrm{R}_2} = \mathrm{E}\left(\left|j - j_i^{\mathrm{b}}\right| \left| j_i^{\mathrm{b}} \in \mathrm{R}_2\right.\right) \approx \frac{j - j_\mathrm{A}}{2} \;;\; i \in \{1, \ldots, C_\mathrm{b}\} \tag{2.24}$$
$$\mathrm{E}_{\mathrm{R}_3} = \mathrm{E}\left(\left|j - j_i^{\mathrm{a}}\right| \left| j_i^{\mathrm{a}} \in \mathrm{R}_3\right.\right) \approx \frac{j_\mathrm{A} + l_\mathrm{A} - j}{2} \;;\; i \in \{1, \ldots, C_\mathrm{a}\} \tag{2.25}$$
$$\mathrm{E}_{\mathrm{R}_4} = \mathrm{E}\left(\left|j - j_i^{\mathrm{a}}\right| \left| j_i^{\mathrm{a}} \in \mathrm{R}_4\right.\right) \approx \frac{g + j_\mathrm{A} + l_\mathrm{A}}{2} - j \;;\; i \in \{1, \ldots, C_\mathrm{a}\} \tag{2.26}$$
$$\tag{2.27}$$

The approximation of expected average index distance goes:

$$\mathrm{E}\left(T_j\left(x\right)\right) \approx \frac{C_{\mathrm{R}_1} \cdot \mathrm{E}_{\mathrm{R}_1} + C_{\mathrm{R}_2} \cdot \mathrm{E}_{\mathrm{R}_2} + C_{\mathrm{R}_3} \cdot \mathrm{E}_{\mathrm{R}_3} + C_{\mathrm{R}_4} \cdot \mathrm{E}_{\mathrm{R}_4}}{C_{\mathrm{R}_1} + C_{\mathrm{R}_2} + C_{\mathrm{R}_3} + C_{\mathrm{R}_4}} \tag{2.28}$$

inserting above approximations

$$\begin{aligned}\mathrm{E}\left(T_j\left(x\right)\right) \approx & \frac{1}{l_{\mathrm{A}} \cdot P_X^{\mathrm{A}}\left(x\right) + \left(g - l_{\mathrm{A}}\right) \cdot P_X^{\mathrm{N}}\left(x\right)} \cdot \left[\left(j - \frac{j_{\mathrm{A}}}{2}\right) \cdot j_{\mathrm{A}} \cdot P_X^{\mathrm{N}}\left(x\right)\right. \\ & + \left(\frac{j - j_{\mathrm{A}}}{2}\right) \cdot \left(j - j_{\mathrm{A}}\right) \cdot P_X^{\mathrm{A}}\left(x\right) \\ & + \left(\frac{j_{\mathrm{A}} + l_{\mathrm{A}} - j}{2}\right) \cdot \left(j_{\mathrm{A}} + l_{\mathrm{A}} - j\right) \cdot P_X^{\mathrm{A}}\left(x\right) \\ & \left. + \left(\frac{g + j_{\mathrm{A}} + l_{\mathrm{A}}}{2} - j\right) \cdot \left(g - j_{\mathrm{A}} - l_{\mathrm{A}}\right) \cdot P_X^{\mathrm{N}}\left(x\right)\right] \end{aligned} \tag{2.29}$$

For the three cases introduced above we respectively get

- Case 1: $P_X^{\mathrm{N}}\left(x\right) \ll P_X^{\mathrm{A}}\left(x\right)$
  The terms containing $P_X^{\mathrm{A}}\left(x\right)$ will determine both the enumerator and the denominator of (2.29), thus making

$$\mathrm{E}\left(T_j\left(x\right)\right) \approx \frac{1}{l_{\mathrm{A}}} \cdot \left[\left(\frac{j - j_{\mathrm{A}}}{2}\right) \cdot \left(j - j_{\mathrm{A}}\right) + \left(\frac{j_{\mathrm{A}} + l_{\mathrm{A}} - j}{2}\right) \cdot \left(j_{\mathrm{A}} + l_{\mathrm{A}} - j\right)\right] \tag{2.30}$$

  a valid approximation. (2.30) may be rewritten and upper bounded as

$$\mathrm{E}\left(T_j\left(x\right)\right) \approx \frac{1}{2j_{\mathrm{A}}}\left(\left(j - j_{\mathrm{A}}\right)^2 + \left(l_{\mathrm{A}} - \left(j - j_{\mathrm{A}}\right)\right)^2\right) \leq \frac{j_{\mathrm{A}}}{2} \tag{2.31}$$

- Case 2: $P_X^{\mathrm{N}}\left(x\right) \approx P_X^{\mathrm{A}}\left(x\right)$
  Since the probability of generation is approximately equal for the whole of the sequence, we get a result identical to the case of no anomaly, and (2.29) will converge to:

$$\begin{aligned}\mathrm{E}\left(T_j\left(x\right)\right) &\approx \frac{j^2 + \left(g - j\right)^2}{2g} \\ &\geq \frac{g}{4}\end{aligned} \tag{2.32}$$

- Case 3: $P_X^{\mathrm{N}}\left(x\right) \gg P_X^{\mathrm{A}}\left(x\right)$
  The terms containing $P_X^{\mathrm{N}}\left(x\right)$ will determine both the enumerator and the denominator of (2.29), so we may approximate

$$\mathrm{E}\left(T_j\left(x\right)\right) \approx \frac{1}{g - l_{\mathrm{A}}} \cdot \left[\left(j - \frac{j_{\mathrm{A}}}{2}\right) \cdot j_{\mathrm{A}} + \left(\frac{g + j_{\mathrm{A}} + l_{\mathrm{A}}}{2} - j\right) \cdot \left(g - j_{\mathrm{A}} - l_{\mathrm{A}}\right)\right] \tag{2.33}$$

For the subsequent derivation of the lower bound of (2.33), we first replace j by $j_\mathrm{A} + zl_\mathrm{A}$, with $z \in [0,1]$. We may do so because j is located inside the anomaly. Insertion yields

$$\begin{aligned}\mathrm{E}\left(T_{j_\mathrm{A}+zl_\mathrm{A}}(x)\right) &\approx \frac{1}{g-l_\mathrm{A}} \cdot \left[\left(j_\mathrm{A} + zl_\mathrm{A} - \frac{j_\mathrm{A}}{2}\right) \cdot j_\mathrm{A} \right. \\ &\quad \left. + \left(\frac{g + j_\mathrm{A} + l_\mathrm{A}}{2} - j_\mathrm{A} - zl_\mathrm{A}\right) \cdot (g - j_\mathrm{A} - l_\mathrm{A})\right] \end{aligned} \tag{2.34}$$

Taking the first derivative of (2.34) with respect to z we get

$$\begin{aligned}\frac{d\mathrm{E}\left(T_{j_\mathrm{A}+zl_\mathrm{A}}(x)\right)}{dz} &\approx \frac{1}{g-l_\mathrm{A}} \cdot [l_\mathrm{A} j_\mathrm{A} - l_\mathrm{A} \cdot (g - j_\mathrm{A} - l_\mathrm{A})] \\ &= \frac{l_\mathrm{A}}{g-l_\mathrm{A}} \cdot [2j_\mathrm{A} + l_\mathrm{A} - g] \end{aligned} \tag{2.35}$$

The derivative is independent of z. For $j_\mathrm{A} < \frac{g-l_\mathrm{A}}{2}$, the derivative is negative, while for $j_\mathrm{A} > \frac{g-l_\mathrm{A}}{2}$, the derivative is positive for any $z \in [0,1]$. This means that for any possible setting of $j_\mathrm{A}, l_\mathrm{A}$, the minimum and the maximum value of (2.34) with respect to z are located at $z = 0$ and $z = 1$, with the locations switching at $j_\mathrm{A} = \frac{g-l_\mathrm{A}}{2}$. Thus, because the behavior of (2.34) for $z = 0$ is symmetric to the behavior for $z = 1$ with respect to $\frac{g-l_\mathrm{A}}{2}$, we may simply set z to zero and take the derivative with respect to j_A searching for the minimum value for $j_\mathrm{A} \in [1, g - l_\mathrm{A}]$. We get

$$\mathrm{E}\left(T_{j_\mathrm{A}+zl_\mathrm{A}|z=0}(x)\right) \approx \frac{1}{g-l_\mathrm{A}} \cdot \frac{1}{2} \cdot \left[j_\mathrm{A}^2 + (g - j_\mathrm{A} + l_\mathrm{A}) \cdot (g - j_\mathrm{A} - l_\mathrm{A})\right] \tag{2.36}$$

and

$$\frac{d\mathrm{E}\left(T_{j_\mathrm{A}+zl_\mathrm{A}|z=0}(x)\right)}{dj_\mathrm{A}} = \frac{1}{g-l_\mathrm{A}} \cdot [2j_\mathrm{A} - g] \tag{2.37}$$

Because (2.37) is strictly monotonic decreasing for $j_\mathrm{A} < \frac{g}{2}$ and strictly monotonic increasing for $j_\mathrm{A} > \frac{g}{2}$, we may deduce that the lower bound of the expression (2.34) is found at $j_\mathrm{A} = \frac{g}{2}$ with $z = 0$ i.e. $j = j_\mathrm{A}$, which yields,

$$\begin{aligned}\mathrm{E}\left(T_j(x)\right) &\approx \frac{1}{g-l_\mathrm{A}} \cdot \left[\left(j - \frac{j_\mathrm{A}}{2}\right) \cdot j_\mathrm{A} + \left(\frac{g + j_\mathrm{A} + l_\mathrm{A}}{2} - j\right) \cdot (g - j_\mathrm{A} - l_\mathrm{A})\right] \\ &\geq \frac{\frac{g^2}{2} - l_\mathrm{A}^2}{2 \cdot (g - l_\mathrm{A})} \end{aligned} \tag{2.38}$$

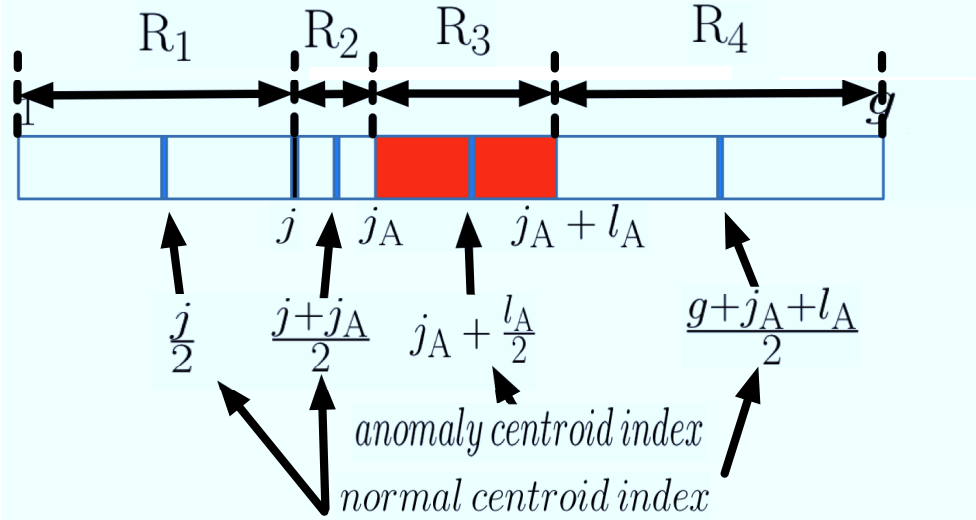

Figure 2.4: Reference symbol located outside anomalous data

Bounds of $\mathrm{E}\left(T_j\left(x\right)\right)$ **for** $j < j_{\mathrm{A}}$ In the case of $j < j_{\mathrm{A}}$, the reference symbol is located outside the anomaly. Figure 2.4 shows the centroid indices of both normal and anomalous areas. Supposing $P_X^{\mathrm{N}}\left(x\right)$ and/or $P_X^{\mathrm{A}}\left(x\right)$ suitably large, the subsequent approximations hold:

$$\begin{aligned} C_{\mathrm{R}_1} &\approx j \cdot P_X^{\mathrm{N}}\left(x\right) && (2.39)\\ C_{\mathrm{R}_2} &\approx \left(j_{\mathrm{A}} - j\right) \cdot P_X^{\mathrm{N}}\left(x\right) && (2.40)\\ C_{\mathrm{R}_3} &\approx l_{\mathrm{A}} \cdot P_X^{\mathrm{A}}\left(x\right) && (2.41)\\ C_{\mathrm{R}_4} &\approx \left(g - j_{\mathrm{A}} - l_{\mathrm{A}}\right) \cdot P_X^{\mathrm{N}}\left(x\right) && (2.42) \end{aligned}$$

$$\begin{aligned} \mathrm{E}_{\mathrm{R}_1} = \mathrm{E}\left(\left|j - j_i^{\mathrm{b}}\right| \middle| j_i^{\mathrm{b}} \in \mathrm{R}_1\right) &\approx \frac{j}{2} \;\; ; \; i \in \{1, \ldots, C_{\mathrm{b}}\} && (2.43)\\ \mathrm{E}_{\mathrm{R}_2} = \mathrm{E}\left(\left|j - j_i^{\mathrm{a}}\right| \left|j_i^{\mathrm{a}} \in \mathrm{R}_2\right.\right) &\approx \frac{j_{\mathrm{A}} - j}{2} \;\; ; \; i \in \{1, \ldots, C_{\mathrm{b}}\} && (2.44)\\ \mathrm{E}_{\mathrm{R}_3} = \mathrm{E}\left(\left|j - j_i^{\mathrm{a}}\right| \left|j_i^{\mathrm{a}} \in \mathrm{R}_3\right.\right) &\approx j_{\mathrm{A}} - j + \frac{l_{\mathrm{A}}}{2} \;\; ; \; i \in \{1, \ldots, C_{\mathrm{a}}\} && (2.45)\\ \mathrm{E}_{\mathrm{R}_4} = \mathrm{E}\left(\left|j - j_i^{\mathrm{a}}\right| \left|j_i^{\mathrm{a}} \in \mathrm{R}_4\right.\right) &\approx \frac{g + j_{\mathrm{A}} + l_{\mathrm{A}}}{2} - j \;\; ; \; i \in \{1, \ldots, C_{\mathrm{a}}\} && (2.46)\\ & && (2.47) \end{aligned}$$

The expected average index distance may be written as follows:

$$\mathrm{E}\left(T_j\left(x\right)\right) \approx \frac{C_{\mathrm{R}_1} \cdot \mathrm{E}_{\mathrm{R}_1} + C_{\mathrm{R}_2} \cdot \mathrm{E}_{\mathrm{R}_2} + C_{\mathrm{R}_3} \cdot \mathrm{E}_{\mathrm{R}_3} + C_{\mathrm{R}_4} \cdot \mathrm{E}_{\mathrm{R}_4}}{C_{\mathrm{R}_1} + C_{\mathrm{R}_2} + C_{\mathrm{R}_3} + C_{\mathrm{R}_4}} \qquad (2.48)$$

Supposing $P_X^{\mathrm{N}}(x)$ and/or $P_X^{\mathrm{A}}(x)$ suitably large, the expected average index distance may be approximated as follows:

$$\begin{aligned}\mathrm{E}(T_j(x)) \approx\ & \frac{1}{l_{\mathrm{A}} \cdot P_X^{\mathrm{A}}(x) + (g - l_{\mathrm{A}}) \cdot P_X^{\mathrm{N}}(x)} \left[\left(\frac{j}{2} \right) \cdot j \cdot P_X^{\mathrm{N}}(x) \right. \\ & + \left(\frac{j_{\mathrm{A}} - j}{2} \right) \cdot (j_{\mathrm{A}} - j) \cdot P_X^{\mathrm{N}}(x) \\ & + \left(j_{\mathrm{A}} + \frac{l_{\mathrm{A}}}{2} - j \right) \cdot l_{\mathrm{A}} \cdot P_X^{\mathrm{A}}(x) \\ & \left. + \left(\frac{g + j_{\mathrm{A}} + l_{\mathrm{A}}}{2} - j \right) \cdot (g - j_{\mathrm{A}} - l_{\mathrm{A}}) \cdot P_X^{\mathrm{N}}(x) \right]\end{aligned} \tag{2.49}$$

For the three cases introduced above we respectively get

- Case 1: $P_X^{\mathrm{N}}(x) \ll P_X^{\mathrm{A}}(x)$
  The terms containing $P_X^{\mathrm{A}}(x)$ will determine both the enumerator and the denominator of (2.49), so we may approximate

$$\begin{aligned}\mathrm{E}(T_j(x)) &\approx \frac{1}{l_{\mathrm{A}}} \cdot \left[\left(j_{\mathrm{A}} + \frac{l_{\mathrm{A}}}{2} - j \right) \cdot l_{\mathrm{A}} \right] \\ &= j_{\mathrm{A}} + \frac{l_{\mathrm{A}}}{2} - j\end{aligned} \tag{2.50}$$

  We replace j by zj_{A}, with $z \in [0, 1]$, getting

$$\mathrm{E}(T_j(x)) \approx j_{\mathrm{A}}(1 - z) + \frac{l_{\mathrm{A}}}{2}, \tag{2.51}$$

  finally deriving the lower bound

$$\begin{aligned}\mathrm{E}(T_j(x)) &\approx \frac{1}{l_{\mathrm{A}}} \cdot \left[\left(j_{\mathrm{A}} + \frac{l_{\mathrm{A}}}{2} - j \right) \cdot l_{\mathrm{A}} \right] \\ &\geq \frac{l_{\mathrm{A}}}{2}\end{aligned} \tag{2.52}$$

- Case 2: $P_X^{\mathrm{N}}(x) \approx P_X^{\mathrm{A}}(x)$
  Since the probability of generation is approximately equal for the whole of the sequence, we get a result identical to the case of no anomaly, and (2.49) will converge to:

$$\mathrm{E}(T_j(x)) \approx \frac{j^2 + (g - j)^2}{2g} \tag{2.53}$$

- Case 3: $P_X^{\mathrm{N}}(x) \gg P_X^{\mathrm{A}}(x)$
  The terms containing $P_X^{\mathrm{N}}(x)$ will determine both the enumerator and the denominator of (2.49), so we may approximate

$$\begin{aligned}\mathrm{E}(T_j(x)) &\approx \frac{1}{g-l_{\mathrm{A}}}\cdot\left[\left(\frac{j}{2}\right)\cdot j+\right.\\ &\quad+\left(\frac{j_{\mathrm{A}}-j}{2}\right)\cdot(j_{\mathrm{A}}-j)\\ &\quad+\left.\left(\frac{g+j_{\mathrm{A}}+l_{\mathrm{A}}}{2}-j\right)\cdot(g-j_{\mathrm{A}}-l_{\mathrm{A}})\right] \qquad (2.54)\end{aligned}$$

  For deducing a lower bound, we first take the derivative of (2.54) according to j_{A} while fixing j:

$$\frac{d\mathrm{E}(T_j(x))}{dj_{\mathrm{A}}} \approx -\frac{l_{\mathrm{A}}}{g-l_{\mathrm{A}}} \qquad (2.55)$$

  The result is negative for any $j_{\mathrm{A}} \geq j$. As mentioned before, because of symmetry, we may limit our consideration to $j \leq j_{\mathrm{A}}$. This means that no matter where in the interval $[1, j_{\mathrm{A}}]$ j is placed, (2.54) can be diminished by raising j_{A}. Thus, after setting $j_{\mathrm{A}} = g - l_{\mathrm{A}}$ we may move on to find a lower bound for (2.54).

$$\begin{aligned}\mathrm{E}(T_j(x)) &\approx \frac{1}{g-l_{\mathrm{A}}}\cdot\left[\left(\frac{j}{2}\right)\cdot j+\left(\frac{g-l_{\mathrm{A}}-j}{2}\right)\cdot(g-l_{\mathrm{A}}-j)\right]\\ &= \frac{1}{2(g-l_{\mathrm{A}})}\cdot\left[j^2+(g-l_{\mathrm{A}}-j)^2\right]\\ &\geq \frac{g-l_{\mathrm{A}}}{4} \qquad (2.56)\end{aligned}$$

Table 2.1 sums up the upper and lower bounds of (2.29) and (2.49) deduced for a particular l_{A}, while Table 2.2 generalizes the bounds to the range of l_{A} defined by (2.2).

In order to identify anomalous blocks, we would like to deduce a threshold τ_{th} for separating average index differences generated inside the anomaly by symbols typical of the anomaly (i.e. $P_X^{\mathrm{A}}(x) \gg P_X^{\mathrm{N}}(x)$) from all other index differences. If we know l_{A}, we may use $\tau_{\mathrm{th}} = \frac{l_{\mathrm{A}}}{2}$ because for any particular $l_{\mathrm{A}} \leq \frac{g}{3}$ as defined by (2.2), the upper bound given by (2.57) falls below all the other bounds in Table 2.1. But since we only know the range of (2.2), we have to consult Table 2.2. We see that the upper bound of (2.64) overlaps with the lower bound of (2.68), causing some misclassification if we want to detect all the index differences falling below (2.64). However, only few of those symbols are generated outside the anomaly

Table 2.1: Upper and lower bounds of the expected value of the average index difference for a single anomalous subsequence of a particular anomaly length l_{A}

	$P_X^{\mathrm{N}}(x) \ll P_X^{\mathrm{A}}(x)$	$P_X^{\mathrm{N}}(x) \approx P_X^{\mathrm{A}}(x)$	$P_X^{\mathrm{N}}(x) \gg P_X^{\mathrm{A}}(x)$
j inside	$\mathrm{E}(T_j(x)) \leq \frac{l_{\mathrm{A}}}{2}$ (2.57)	$\mathrm{E}(T_j(x)) \geq \frac{g}{4}$ (2.58)	$\mathrm{E}(T_j(x)) \geq \frac{\frac{g^2}{2} - l_{\mathrm{A}}^2}{2 \cdot (g - l_{\mathrm{A}})}$ (2.59)
j outside	$\mathrm{E}(T_j(x)) \geq j_{\mathrm{A}}(1-z) + \frac{l_{\mathrm{A}}}{2}$ (2.60) $\geq \frac{l_{\mathrm{A}}}{2} \quad z \in [0,1]$ (2.61)	$\mathrm{E}(T_j(x)) \geq \frac{g}{4}$ (2.62)	$\mathrm{E}(T_j(x)) \geq \frac{g - l_{\mathrm{A}}}{4}$ (2.63)

Table 2.2: Upper and lower bounds of the expected value of the average index difference for a single anomalous subsequence of length l_{A} within the range defined by (2.2).

	$P_X^{\mathrm{N}}(x) \ll P_X^{\mathrm{A}}(x)$	$P_X^{\mathrm{N}}(x) \approx P_X^{\mathrm{A}}(x)$	$P_X^{\mathrm{N}}(x) \gg P_X^{\mathrm{A}}(x)$
j inside	$\mathrm{E}(T_j(x)) \leq \frac{g}{6}$ (2.64)	$\mathrm{E}(T_j(x)) \geq \frac{g}{4}$ (2.65)	$\mathrm{E}(T_j(x)) \geq \frac{g}{4}$ (2.66)
j outside	$\mathrm{E}(T_j(x)) \geq j_{\mathrm{A}}(1-z) + \frac{b}{2}$ (2.67) $\geq \frac{b}{2} \quad z \in [0,1]$ (2.68)	$\mathrm{E}(T_j(x)) \geq \frac{g}{4}$ (2.69)	$\mathrm{E}(T_j(x)) \geq \frac{g}{6}$ (2.70)

(i.e. $P_X^{\mathrm{A}}(x) \gg P_X^{\mathrm{N}}(x)$), and the normal block containing them may not necessarily adjoin anomalous data, a fact expressed by (2.64). Hence, the threshold setting

$$\tau_{\mathrm{th}} = \frac{g}{6} \tag{2.71}$$

is used.

2.2.3 Computational Cost

The computational cost of calculating the average index difference of all symbols for a sequence of length g is upper bounded by a function of order

$$O(Z \cdot g) \tag{2.72}$$

For every $x \in \mathcal{X}$ within the sequence of length g, the number of occurrences $C_{\mathrm{b}}+C_{\mathrm{a}}+1$, the indices of the occurrences, and the respective index differences to neighboring occurrence(s) of identical value, as well as the sum of the index differences between the first occurrence and the subsequent occurrences have been retrieved, with the computational cost obeying above bound. What is left to show is that for a particular x, the enumerator of (2.6) of the occurrences may be calculated starting from the enumerator of the first occurrence, which has been calculated during the previous step. Changing the notation to index differences of neighboring symbols of identical value, the average index difference of the 1^{st} occurrence of x counting from the head of the sequence, the index of which is noted as j_1 is rewritten as

$$T_{j_1}(x) = \frac{\sum_{q=1}^{C_b+C_a} q \cdot \Delta_{C_{\mathrm{a}}+C_{\mathrm{b}}+1-q}(x)}{C_b + C_a}, \tag{2.73}$$

while the index difference of the i^{th} occurrence at j_i is expresses

$$T_{j_i}(x) = \frac{\sum_{o=1}^{i-1} o \cdot \Delta_o(x) + \sum_{q=1}^{C_b+C_a-i+1} q \cdot \Delta_{C_b+C_a+1-q}(x)}{C_b + C_a} \tag{2.74}$$

$\Delta_i(x)$ represents the index difference between the i^{th} and the $i+1^{th}$ occurrence of x. The change of the enumerator between the average index difference of the ith and the average index difference of the $i+1$th occurrence consists of subtracting $\Delta_i(x)$ $C_b + C_a - i + 1$ times and adding it i times. In other words, the upper limit of the index of the second sum of the enumerator of (2.74) is decreased by one, while the upper limit of the index of the first sum is increased by one.

$$\begin{aligned} T_j(x, i+1) - T_j(x, i) &= \frac{i \cdot \Delta_i(x) - (C_b + C_a - i + 1) \cdot \Delta_i(x)}{C_b + C_a} \\ &= \frac{(2 \cdot i - C_b - C_a - 1) \cdot \Delta_i(x)}{C_b + C_a} \end{aligned} \tag{2.75}$$

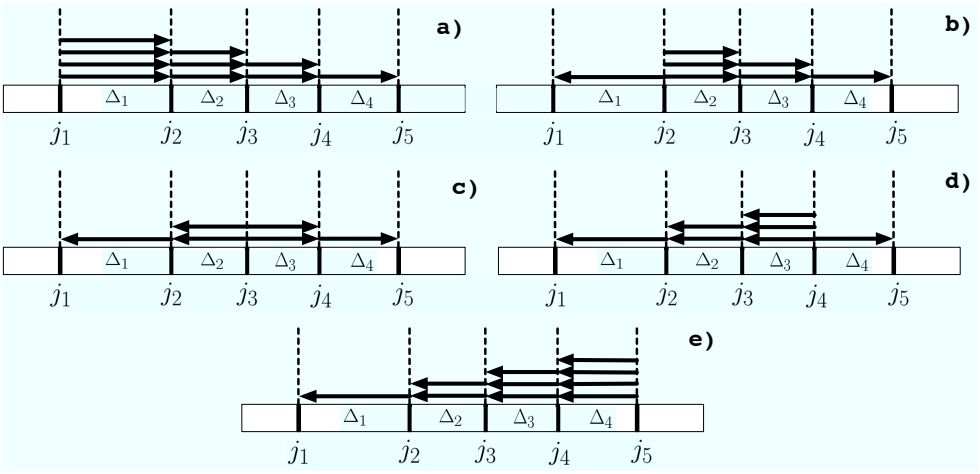

Figure 2.5: Consecutive computation of average index difference based on $\Delta_i(x)$

The number of additions and multiplications required to calculate the difference given by (2.75) is a constant independent of the number of occurrences. Thus, the computational effort required to calculate the average index differences of the occurrences of a particular symbol is a linear function of the number of occurrences, and because the number of occurrences for any $x \in \mathcal{X}$ is upper bounded by g, the bound of the computational cost is proven to be correct.

Figure 2.5 shows a graphic example of the process for $C_\mathrm{a} + C_\mathrm{b} + 1 = 5$, with the number of arrows representing the multiplier of the respective $\Delta_i(x)$ within the sums in the enumerator of (2.74) and the direction indicating to which of the two sums the $\Delta_i(x)$ contributes. While for $i = 1$ all $\Delta_i(x)$ contribute to the second sum, for every $i > 1$ only the arrows representing the $\Delta_{i-1^{th}}(x)$ are turned and their number is altered. For all other $\Delta(x)$, the number and direction of the arrows remains the same.

2.2.4 Processing of Subsequences (Grams)

While the previous sections introduced and analyzed an algorithm processing symbols, it is important to note that the processing of subsequences of symbols or grams might sometimes be more effective for sources with memory. Figure 2.6 shows an example for an alphabet of size $Z = 3$. The stationary distribution of symbols is identical for anomalous and normal data. Thus computing the average index difference for the symbol b found at index j cannot detect the anomaly. The average index difference of the symbol subsequence of length 2 found at index j, ba, on the other hand, is able to discriminate normal and anomalous data.

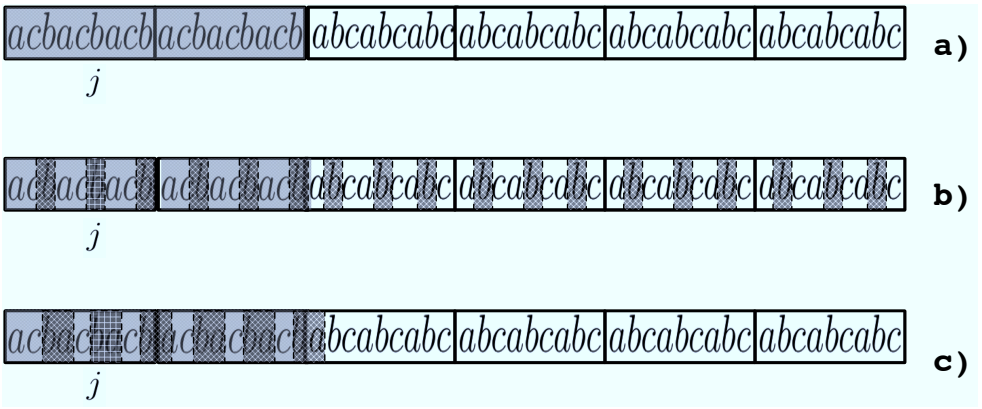

Figure 2.6: a) Original sequence headed by anomalous blocks b) Symbols identical to the one found at index j c) Symbol subsequences of length 2 identical to the one found at index j

A suitable search algorithm for typical subsequences of length $M + 1$ is shown by Algorithm 2. It repeatedly increases the memory parameter M starting from $M = 0$ (symbols) until either the maximum memory $M_{\max}$ is reached or the occurrence number $C_{\text{a block}} + C_{\text{b block}}$ of identical symbols/subsequences within the same block drops below a threshold $\nu_{\min}$. The retrieved symbol sequences are then used within the subsequent stages of the algorithm. Figure 2.7 shows the notation used. The setting of the parameter $\nu_{\min}$ is guided by the consideration that $\nu_{\min}$ has to be chosen small enough to make the generation of $\nu_{\min}$ grams within a single block b (an anomaly of minimum length) highly probable even for small generation probabilities. On the other hand, a large setting increases stability of the algorithm if no anomaly was generated or the retrieved symbol sequences (grams) are typical of the normal data. Regarding $M_{\max}$ on the other hand, it is sufficient to select a setting just small enough to keep effects at the edge of the anomalous data to a minimum. Thus a choice below 5% of the block length is suitable. Another concern is the increase of computational cost in case of large $M_{\max}$. A suitable setting for $M_{\max}$ can be derived as the depth of a probabilistic suffix tree optimized according to the Akaike Information Criterion.

2.2.5 Drawbacks of Algorithm

The supposition of a consecutive sequence of anomalous data blocks, while convenient, may not be met by the data. Figure 2.8 shows an example: the anomalous block sequence is split up and moved to opposite ends of the sequence. Here, although the symbol c found at index j is obviously typical of the anomalous data, calculation of the average

Algorithm 2 Search Algorithm for Typical Subsequences

Ensure: nonnegative maximum memory length $M_{\max} \geq 0$

Ensure: positive minimum number of identical sequences $\nu_{\min} > 0$ within the block containing j

memory length variable $M \Leftarrow 0$

identical sequence number variable $\nu \Leftarrow 0$

Calculate the number of symbols $\nu \Leftarrow C_{\text{a block}} + C_{\text{b block}}$ within the block containing j identical to the one found at index j

if $\nu \geq \nu_{\min}$ **then**

  while $M < M_{\max}$ and $M < (j-1)$ **do**

    $M++$

    Calculate the number of $M+1$-grams $\nu \Leftarrow C_{\text{a block}} + C_{\text{b block}}$ within the block containing j identical to the one found at index j

    if $\nu < \nu_{\min}$ **then**

      $M--$

      Calculate the number of $M+1$-grams $\nu \Leftarrow C_{\text{a block}} + C_{\text{b block}}$ within the block containing j identical to the one found at index j

      break the while loop

    end if

  end while

end if

Output the $C_{\text{a}} + C_{\text{b}}$ $M+1$-grams within the overall sequence g identical to the one found at index j

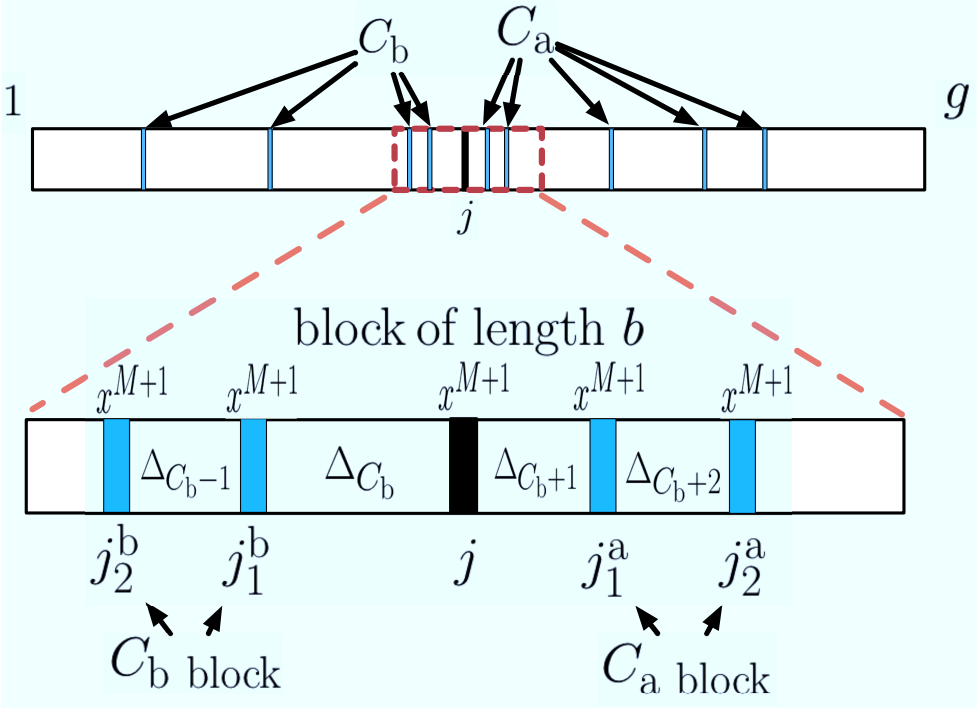

Figure 2.7: Notation extension for occurrences within the same block

index difference is futile because the consecutive index difference $\Delta_3(x)$ creates a bias. We observe that in order to suppress the influence of the gap, we have to compare every $\Delta(x)$ to the mean value defined by (2.8), and eliminate it from the calculation if it surpasses a threshold defined as a multiple of the mean.

Another point is the implicit supposition of the existence of symbols $x \in \mathcal{X}$ featuring $P_X^{\mathrm{N}}(x) \ll P_X^{\mathrm{A}}(x)$ i.e. $h(x) \ll 1$. While allowing for a convenient derivation of a threshold setting τ_{th} yielding a low error rate, the supposition may not be met by the data.

2.3 Algorithm Allowing for Arbitrary Distribution of Anomalous Blocks

2.3.1 Algorithm Statement

Addressing the main drawback of the algorithm presented in the previous section, we present an algorithm allowing for arbitrary distribution of anomalous blocks consisting of

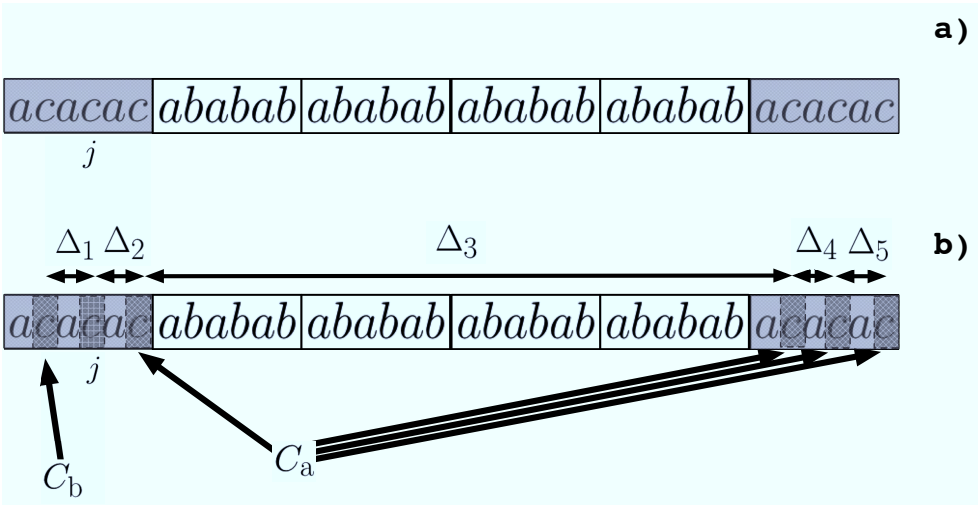

Figure 2.8: Example sequence featuring a distributed anomaly: (a) anomalous data, b) symbols identical to the one found at index j

the following steps:

1. Search for typical symbol subsequences.
2. Close the gaps between anomalous blocks by comparing the index differences $\Delta_i(x)$ to the mean value $\overline{\Delta}\left(x^{M+1}\right)$.
3. Calculate the average index differences.
4. Process the average index differences of each block to generate a scalar.

In order to broaden applicability, instead of the demanding definition of the ratio of generation probabilities marking symbol subsequences x^{M+1} typical of the anomalous data given by $h\left(x^{M+1}\right) \ll 1$, this algorithm uses a threshold $h_{\mathrm{th}} < 1$ to define the range of generation probability ratios considered anomalous.

$$\frac{P_X^{\mathrm{N}}\left(x^{M+1}\right)}{P_X^{\mathrm{A}}\left(x^{M+1}\right)} = h\left(x^{M+1}\right) \quad \leq \quad h_{\mathrm{th}} \tag{2.76}$$

Here, M is a nonnegative integer expressing the memory of the present average index difference calculation. A suitable setting of h_{th} will be determined as $h_{\mathrm{th}} = \frac{1}{4}$ in Section 2.3.2.

In order for the detection to work, the algorithm supposes the existence of symbol subsequences x^{M+1} meeting the conditions

$$\frac{P_X^{\mathrm{N}}\left(x^{M+1}\right)}{P_X^{\mathrm{A}}\left(x^{M+1}\right)} \leq h_{\mathrm{th}} \tag{2.77}$$

and

$$\mathrm{E}\left(C_{\text{b block}} + C_{\text{a block}} + 1 \,|\mathrm{A}\right) \;\; \geq \;\; 3, \tag{2.78}$$

i.e.

$$\frac{1}{P_X^{\mathrm{A}}\left(x^{M+1}\right)} \;\; \leq \;\; \frac{b}{3}, \tag{2.79}$$

with condition (2.79) being caused by the arbitrary arrangement of blocks allowed by the algorithm.

The fundamental properties of the average index difference function introduced previously are also valid for subsequences of symbols, because Kac's Lemma holds for subsequences of symbols in case of stationary ergodic symbol generation.

After preprocessing the symbol sequence in order to remove gaps between anomalous blocks and detect typical subsequences, we calculate the average index difference according to (2.6) and scale the returned value using the expected value given by (2.12).

$$S_j\left(x^{M+1}\right) \stackrel{\text{def}}{=} \frac{T_j(x^{M+1}, \text{Algorithm 2})}{\mathrm{E}\left(T_j\left(x^{M+1}\right)|l_{\mathrm{A}}=0\right)} \cdot \frac{g}{2} = \frac{T_j(x^{M+1}, \text{Algorithm 2})}{\frac{j^2+(g-j)^2}{2g}} \cdot \frac{g}{2} \tag{2.80}$$

Thus, a constant expected value for any j in case of no anomaly is assured.

Vector Processing of Average Index Difference

As mentioned before, the original approach of calculating the scalar value representing a block consisted of simply calculating the percentage of symbols featuring an average index $T_j\left(x\right)$ difference falling below a threshold τ_{th}. This was feasible because of the strong requirement of $h\left(x\right) \ll 1$. But the increase of the applicability by introduction of the threshold h_{th}, as well as the scaling, increase the noisiness in the average index differences $S_j\left(x^{M+1}\right)$ returned by the second algorithm. Thus, we split up the range defined by τ_{th}, $[0, \tau_{\mathrm{th}}]$, into a sections or bins of equal size, creating a vector of dimension $a+1$ with entries according to the average index differences observed in the block, and finally calculating the

Euclidean norm of the vector, thus creating the scalar used for calculating the receiver operating characteristic curve[1]. Since the length of a block is limited, we choose a small setting of a to avoid sparse vectors.

Algorithm 3 shows the procedure. In order to smooth the returned values, we not only increase the counter of the bin an average index difference value is falling into, but also those of adjoining bins. This is also the reason the vector has $a+1$ instead of a entries. This detail is inspired by distribution estimation from sparse samples using kernel functions.

Algorithm 3 Calculation of Vector Entries Based on the Average Index Differences of the Block

Ensure: the vector of average index differences $S_j\left(x^{M+1}\right)^b$ within the block b has been calculated before
reset the entries of the output vector $V^{a+1} \Leftarrow 0$
for $com_1 = 1$ to $com_1 = b$ **do**
  for $com_2 = 1$ to $com_2 = a$ **do**
    if $com_2 == 1$ **then**
      if $T_{com_1} \leq \frac{\tau_{\text{th}}}{a}$ **then**
        $V_1 \Leftarrow V_1 + 3$
        $V_2 \Leftarrow V_2 + 1$
      end if
    else
      if $T_{com_1} \leq \frac{\tau_{\text{th}}}{a} \cdot com_2$ and $T_{com_1} \geq \frac{\tau_{\text{th}}}{a} \cdot (com_2 - 1)$ **then**
        $V_{com_2} \Leftarrow V_{com_2} + 2$
        $V_{com_2-1} \Leftarrow V_{com_2-1} + 1$
        $V_{com_2+1} \Leftarrow V_{com_2+1} + 1$
      end if
    end if
  end for
end for
calculate and output the Euclidean norm of the vector V^{a+1}

[1]The receiver operating characteristic curve or simply ROC curve is a method used for visualization of the performance of an algorithm on a classification problem. For anomaly detection, the unit of the x-coordinate is the rate of normal blocks misclassified as anomalous blocks (false positive rate), while the y-coordinate is the rate of correctly detected anomalous blocks (true positive rate). A point is the combination of true positive rate and false positive rate returned by a particular parameter setting. The curve of those points thus marks the tradeoff between true positive and false positive rate achievable by parameter variation.

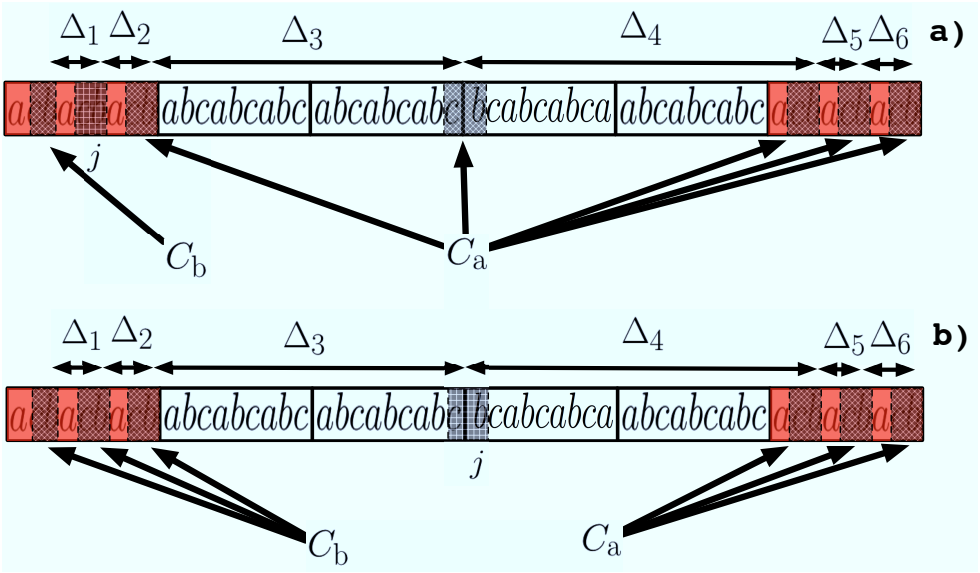

Figure 2.9: a) Index j located inside anomalous data b) Index j located inside normal data

Gap Closing

This processing step was designed to close the gaps between anomalous blocks prior to the calculation of the average index difference, if the symbol was generated inside an anomaly and is characteristic of the anomaly. In order to illustrate the rationale of this approach, we present an example sequence with the index j located inside and outside the anomaly data (Fig. 2.9). While the 2-grams returned are typical of the anomaly, an occasional 2-gram is located inside the normal data. We are interested in retaining index differences $\Delta\left(x^{M+1}\right)$ inside the anomaly, while eliminating the index differences spanning the gap, *if* the index j is actually located inside the anomaly. Therefore, we check the index differences[2] $\Delta_{C_\mathrm{b}}(x)$ and $\Delta_{C_\mathrm{b}+1}(x)$ adjourning the index j (Fig. 2.7). If both fall below a suitable fraction of the mean $\beta \cdot \overline{\Delta}\left(x^{M+1}\right)$, it is very likely that j is located inside the anomaly. On the other hand, if one or both of the adjourning index differences equals or surpasses the mean, j is more likely to be located inside the normal data, in which case no elimination is desirable. We also must avoid eliminating index differences in case of no anomaly as much as possible. Another point illustrated by the example is that in order to allow for effective gap elimination, the anomalous data blocks have to be distributed uniformly within the whole of the index sequence. Thus we randomly rearrange the order of the blocks prior to gap elimination, performing an urn experiment without replacement.

[2]These are the index differences between the symbol sequence found at index j and the neighboring identical symbol sequences.

A more formal description reads as follows

1. Check if both $\Delta_{C_\mathrm{b}}(x)$ and $\Delta_{C_\mathrm{b}+1}(x)$, the index differences adjoining j within the same block, fall below a suitable fraction of the empirical mean of the index difference $\overline{\Delta}\left(x^{M+1}\right)$ defined as $\beta \cdot \overline{\Delta}\left(x^{M+1}\right)$, with the parameter $\beta \leq 1$[3]. If not, skip the next step.

2. Check the sequence of index differences of neighboring symbol sequences of identical value, $\Delta^{C_\mathrm{a}+C_\mathrm{b}}$, for index differences exceeding a multiple of the empirical mean $\overline{\Delta}\left(x^{M+1}\right)$, $\xi \cdot \overline{\Delta}\left(x^{M+1}\right)$, with the parameter $\xi \geq 1$. Exclude those sequences from the calculation of the average index difference for this j.

3. Proceed to calculate the average index difference based on the remaining $\Delta\left(x^{M+1}\right)$.

A pseudocode transcription of the processing is shown as Algorithm 4.

Algorithm 4 Gap Elimination Algorithm

Require: $C_\mathrm{a} + C_\mathrm{b} \geq 2$
Ensure: $0 \leq \beta \leq 1$
Ensure: $\xi \geq 1$
$\overline{\Delta} \Leftarrow \frac{\sum_{q=1}^{C_\mathrm{a}+C_\mathrm{b}} \Delta_q}{C_\mathrm{a}+C_\mathrm{b}}$
$z \Leftarrow 1$
if $\Delta_{C_\mathrm{b}} \leq \beta \cdot \overline{\Delta}$ and $\Delta_{C_\mathrm{b}+1} \leq \beta \cdot \overline{\Delta}$ and $\Delta_{C_\mathrm{b}} + \Delta_{C_\mathrm{b}+1} \leq b$ hold **then**
  while $z \leq C_\mathrm{a} + C_\mathrm{b}$ **do**
    if $\Delta_z \geq \xi \cdot \overline{\Delta}$ **then**
      exclude Δ_z from calculation
    end if
    $z++$
  end while
end if
use the remaining Δ for calculating the average index difference.

[3]The existence of subsequences x^{M+1} falling below the threshold and featuring adjoining occurrences within the same block is demanded by (2.77) and (2.79).

2.3.2 Parameter Setting: Stationary Ergodic Source

Setting of β

As mentioned before, the choice of β regulates which subsequences x^{M+1} are judged to be located inside anomalous data blocks.

In case of anomaly, the expected value of the mean of index differences of neighboring occurrences given by (2.8) may be approximated by

$$\begin{aligned} \mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right)\right) &= \mathrm{E}\left[\frac{\sum_{q=1}^{C_\mathrm{a}+C_\mathrm{b}} \Delta_q\left(x^{M+1}\right)}{C_\mathrm{a}+C_\mathrm{b}}\right] \\ &\approx \mathrm{E}\left[\frac{g}{C_\mathrm{a}+C_\mathrm{b}}\right] \\ &= \frac{g}{\mathrm{E}\left(C_\mathrm{a}+C_\mathrm{b}\right)} \\ &= \frac{g}{l_\mathrm{A}\cdot P_X^\mathrm{A}\left(x^{M+1}\right)+\left(g-l_\mathrm{A}\right)\cdot P_X^\mathrm{N}\left(x^{M+1}\right)} \end{aligned} \tag{2.81}$$

for a suitably large $g \gg \frac{1}{P_X^\mathrm{N}(x^{M+1})}, \frac{1}{P_X^\mathrm{A}(x^{M+1})}$. We desire a setting of β which is able to discriminate the $\Delta_\mathrm{A}(x)$ and $\Delta_\mathrm{N}(x)$.

If we define the symbol sequences x^{M+1} typical of the anomalous data compared to the normal data by

$$P_X^\mathrm{A}\left(x^{M+1}\right) \geq \frac{1}{h_\mathrm{th}}\cdot P_X^\mathrm{N}\left(x^{M+1}\right) \tag{2.82}$$

$$\tag{2.83}$$

i.e.

$$h\left(x^{M+1}\right) \leq h_\mathrm{th} \leq 1, \tag{2.84}$$

the subsequent inequalities may be derived by combining (2.81) and (2.82), with the last inequality based on the assumption of the maximum anomaly length (2.2).

$$\mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right)\,\middle|\, h\left(x^{M+1}\right)\leq h_\mathrm{th}\right) \geq \frac{g}{l_\mathrm{A}+\left(g-l_\mathrm{A}\right)\cdot h_\mathrm{th}}\cdot\frac{1}{P_X^\mathrm{A}\left(x^{M+1}\right)} \geq \frac{3}{1+2h_\mathrm{th}}\cdot\frac{1}{P_X^\mathrm{A}\left(x^{M+1}\right)}, \tag{2.85}$$

yielding

$$\mathrm{E}\left(\Delta(x^{M+1})|\mathrm{A}\right) = \frac{1}{P_X^\mathrm{A}\left(x^{M+1}\right)} \leq \frac{1+2h_\mathrm{th}}{3}\cdot\mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right)|h\left(x^{M+1}\right)\leq h_\mathrm{th}\right). \tag{2.86}$$

Moreover, in case of $h\left(x^{M+1}\right) \approx 0$, we get

$$\mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right)|h\left(x^{M+1}\right) \approx 0\right) \approx \frac{g}{l_{\mathrm{A}} \cdot P_X^{\mathrm{A}}\left(x^{M+1}\right)} \geq \frac{3}{P_X^{\mathrm{A}}\left(x^{M+1}\right)}. \tag{2.87}$$

Examining the relation of the expected index difference within normal data and the mean of index differences, we get

$$\mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right)|h\left(x^{M+1}\right) \leq h_{\mathrm{th}}\right) \leq \frac{g}{\frac{l_{\mathrm{A}}}{h_{\mathrm{th}}} + g - l_{\mathrm{A}}} \cdot \frac{1}{P_X^{\mathrm{N}}\left(x^{M+1}\right)} \leq \frac{g}{\frac{b}{h_{\mathrm{th}}} + g - b} \cdot \frac{1}{P_X^{\mathrm{N}}\left(x^{M+1}\right)}, \tag{2.88}$$

$$\begin{aligned}
\mathrm{E}\left(\Delta\left(x^{M+1}\right))|\mathrm{N}\right) &= \frac{1}{P_X^{\mathrm{N}}\left(x^{M+1}\right)} \\
&\geq \left(\left(\frac{1}{h_{\mathrm{th}}} - 1\right) \cdot \frac{l_{\mathrm{A}}}{g} + 1\right) \cdot \mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right)|h\left(x^{M+1}\right) \leq h_{\mathrm{th}}\right) \qquad (2.89) \\
&\geq \left(\left(\frac{1}{h_{\mathrm{th}}} - 1\right) \cdot \frac{b}{g} + 1\right) \cdot \mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right)|h\left(x^{M+1}\right) \leq h_{\mathrm{th}}\right). \qquad (2.90)
\end{aligned}$$

On the other hand, if we define the symbol sequences x^{M+1} typical of the normal data *and* the neutral sequences by

$$P_X^{\mathrm{A}}\left(x^{M+1}\right) \leq \frac{1}{h_{\mathrm{th}}} \cdot P_X^{\mathrm{N}}\left(x^{M+1}\right), \tag{2.91}$$

using (2.2) we get

$$\mathrm{E}\left(\overline{\Delta}|h\left(x^{M+1}\right) \geq h_{\mathrm{th}}\right) \leq \frac{g}{(g - l_{\mathrm{A}})} \cdot \frac{1}{P_X^{\mathrm{N}}\left(x^{M+1}\right)} \leq \frac{3}{2} \cdot \frac{1}{P_X^{\mathrm{N}}\left(x^{M+1}\right)}. \tag{2.92}$$

Hence

$$\mathrm{E}\left(\Delta(x^{M+1})|\mathrm{N}\right) = \frac{1}{P_X^{\mathrm{N}}\left(x^{M+1}\right)} \geq \frac{2}{3} \cdot \mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right)|h\left(x^{M+1}\right) \geq h_{\mathrm{th}}\right). \tag{2.93}$$

Since the lower bounds of (2.90) and (2.89) are above (2.93), for separating anomalous and normal index differences $\Delta\left(x^{M+1}\right)$, using (2.86), we deduce the condition

$$\frac{2}{3} > \frac{1 + 2h_{\mathrm{th}}}{3} \tag{2.94}$$

i.e.

$$0 < h_{\mathrm{th}} < \frac{1}{2} \tag{2.95}$$

In order to increase protection against noise, we set

$$h_{\text{th}} = \frac{1}{4}. \tag{2.96}$$

Then, according to (2.86), the setting of β must satisfy

$$\beta \leq \frac{1 + 2h_{\text{th}}}{3} = \frac{1}{2} \tag{2.97}$$

Note that while the upper bound of (2.86) depends both on the maximum anomaly length and the maximum ratio of stationary probabilities defining sequences typical of the anomaly, the lower bound (2.93) depends only on the maximum anomaly length. Having thus deduced an upper limit for the setting of β, we are also interested in a lower limit. It is obvious that we may choose a very small setting for β in case of $h_{\text{th}} \approx 0$. Examining (2.87) we find a lower bound with

$$\begin{aligned} \mathrm{E}\left(\Delta_{\mathrm{A}}\left(x\right)\right) &= \frac{1}{P_X^{\mathrm{A}}\left(x^{M+1}\right)} \\ &\approx \frac{l_{\mathrm{A}}}{g} \cdot \mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right) | h\left(x^{M+1}\right) \approx 0\right) \geq \frac{2b}{g} \cdot \mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right) | h\left(x^{M+1}\right) \approx 0\right), \end{aligned} \tag{2.98}$$

with $2b$ being the minimum length l_{A} of a partitioned anomaly. Thus a reasonable range of β is given by by

$$\frac{2b}{g} \leq \beta \leq \frac{1}{2} \tag{2.99}$$

Because we want to keep our algorithm as general as possible, from now on we use the upper bound of the possible range as a setting for β.

$$\beta = \frac{1 + 2h_{\text{th}}}{3} = \frac{1}{2} \tag{2.100}$$

Summing up, we get

$$\mathrm{E}\left(\Delta\left(x^{M+1}\right) \left| \mathrm{A}, h\left(x^{M+1}\right) \leq h_{\text{th}}\right.\right) \leq \beta \cdot \mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right) | h\left(x^{M+1}\right) \leq h_{\text{th}}\right) \tag{2.101}$$

$$= \frac{1 + 2h_{\text{th}}}{3} \cdot \mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right) | h\left(x^{M+1}\right) \leq h_{\text{th}}\right). \tag{2.102}$$

One might now be tempted to simplify the algorithm by only looking at the number of symbol sequences within a block featuring adjourning index differences below the threshold $\beta \cdot \overline{\Delta}\left(x^{M+1}\right)$. However, experiments show the performance is degraded for large l_{A} and by adjourning symbols. Moreover, it is easy to construct counter-examples of two index difference distributions which, despite featuring different expected values, show cumulative probabilities of similar range at the threshold.

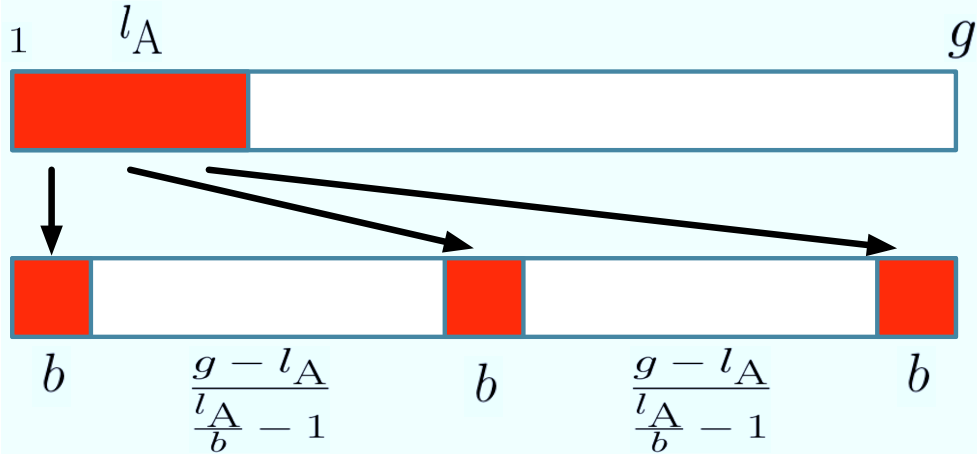

Figure 2.10: Gaps between the anomalous blocks

Setting of ξ

The parameter ξ is used for filtering index differences unlikely to have been generated inside anomalous data *and* a limited share of index differences in case of no anomaly.

In case of $l_{\mathrm{A}} \neq 0$, even when supposing $h\left(x^{M+1}\right) \ll 1$, we must avoid choosing an overly high setting of ξ. It holds from (2.81) that

$$\mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right) \left| h\left(x^{M+1}\right) \ll 1\right.\right) \approx \frac{1}{P_X^{\mathrm{N}}\left(x^{M+1}\right)} \cdot \frac{g}{l_{\mathrm{A}}\left(\frac{1}{h_{\mathrm{th}}}-1\right)+g} \tag{2.103}$$

$$\approx \frac{1}{P_X^{\mathrm{N}}\left(x^{M+1}\right)} \cdot \frac{g}{\frac{l_{\mathrm{A}}}{h_{\mathrm{th}}}+g} \tag{2.104}$$

$$\ll \frac{1}{P_X^{\mathrm{N}}\left(x^{M+1}\right)}. \tag{2.105}$$

But the observed index differences covering normal data will be dominated by the index differences between symbols located in different anomalous blocks in case of a partitioned anomaly. The worst case is shown in Figure 2.10, where the gap between anomalous blocks is given by

$$G\left(l_{\mathrm{A}}\right) = \frac{g-l_{\mathrm{A}}}{\frac{l_{\mathrm{A}}}{b}-1} \tag{2.106}$$

For $2b \leq l_{\mathrm{A}} \leq \frac{g}{3}$ we derive the subsequent bounds of (2.106):

$$G\left(l_{\mathrm{A}}\right) \geq G\left(\frac{g}{3}\right) = 2b \tag{2.107}$$

$$G\left(l_{\mathrm{A}}\right) \leq G\left(2b\right) = g-2b \tag{2.108}$$

Even a low multiple of the empirical mean of the average index difference may surpass above limit.

As an example, suppose $h\left(x^{M+1}\right) \ll 1$ and $P_X^{\mathrm{A}}\left(x^{M+1}\right) \approx \frac{3}{b}$ in accordance with (2.79). We approximate the expected value of $\overline{\Delta}\left(x^{M+1}\right)$ for the worst case of Fig. 2.10, we get

$$\mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right)\right) \approx \frac{g}{P_X^{\mathrm{A}}\left(x^{M+1}\right) \cdot l_{\mathrm{A}} - 1} \tag{2.109}$$

$$\approx \frac{b}{3} \cdot \frac{g}{l_{\mathrm{A}} - \frac{b}{3}} \tag{2.110}$$

we derive the subsequent expected values:

$$\mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right) | l_{\mathrm{A}} = 2b\right) \approx \frac{g}{5} \tag{2.111}$$

and

$$\mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right) \left| l_{\mathrm{A}} = \frac{g}{3}\right.\right) \approx b \tag{2.112}$$

Comparing (2.112) and (2.107), we see that a setting of

$$\xi \approx 2 \tag{2.113}$$

is needed to cover the worst case.

In case of no anomaly, the possibility of the index difference of neighboring occurrences exceeding a certain multiple of the expected value is bounded by

$$P\left(\Delta\left(x^{M+1}\right) \geq \xi \cdot \mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right)\right)\right) = P\left(\Delta\left(x^{M+1}\right) \geq \xi \cdot \mathrm{E}\left(\Delta\left(x^{M+1}\right)\right)\right) \leq \frac{1}{\xi} \tag{2.114}$$

Setting of τ_{th} for $\xi = 2$

Having deduced the settings of $\xi = 2$ and $h_{\mathrm{th}} = \frac{1}{4}$, the final step is the determination of the threshold parameter τ_{th}, which is used for evaluation of the scaled average index difference as given by (2.80). We have to determine how much the gap closing routine - if applied - using the setting $\xi = 2$ shrinks the overall sequence of length g, and thus the average index difference calculated from index differences within this sequence, both for $h\left(x^{M+1}\right) > h_{\mathrm{th}}$ and $h\left(x^{M+1}\right) \leq h_{\mathrm{th}}$.

For the subsequent calculations, we define $\Upsilon\left(k, x^{M+1}, l_{\mathrm{A}}\right)$ as the expected share of a sequence of suitable length $g \gg \frac{1}{P_X^{\mathrm{N}}(x^{M+1})}, \frac{1}{P_X^{\mathrm{A}}(x^{M+1})}$ including an anomaly of overall length

l_{A} covered by index differences $\Delta\left(x^{M+1}\right)$ of length above a positive integer k:

$$\Upsilon\left(k, x^{M+1}, l_{\mathrm{A}}\right) \stackrel{\mathrm{def}}{=} \mathrm{E}\left(\frac{\sum_{i=1}^{C_{\mathrm{a}}+C_{\mathrm{b}}} f(\Delta_i\left(x^{M+1}\right),\ k) \cdot \Delta_i\left(x^{M+1}\right)}{g}\right)$$
$$f(\Delta_i\left(x^{M+1}\right), k) = \begin{cases} 0 & : \Delta_i\left(x^{M+1}\right) < k \\ 1 & : \Delta_i\left(x^{M+1}\right) \geq k \end{cases} \quad \forall k \geq 0 \quad (2.115)$$

We will supplement above definition by additional conditions when necessary. We subsequently approximate $\Upsilon\left(k = \xi \cdot \overline{\Delta}\left(x^{M+1}\right), x^{M+1}, l_{\mathrm{A}}\right)$ by

$$\begin{aligned}
&\Upsilon\left(k = \xi \cdot \overline{\Delta}\left(x^{M+1}\right), x^{M+1}, l_{\mathrm{A}}\right) \\
&\approx \left(\frac{l_{\mathrm{A}} \cdot P_X^{\mathrm{A}}\left(x^{M+1}\right) \cdot \mathrm{E}\left(\Delta_{\mathrm{A}}\left(x^{M+1}\right) | \Delta_{\mathrm{A}}\left(x^{M+1}\right) \geq \xi \cdot \overline{\Delta}\left(x^{M+1}\right)\right)}{g}\right) \\
&\cdot P\left(\Delta_{\mathrm{A}}\left(x^{M+1}\right) \geq \xi \cdot \overline{\Delta}\left(x^{M+1}\right)\right) \\
&+ \left(\frac{(g - l_{\mathrm{A}}) \cdot P_X^{\mathrm{N}}\left(x^{M+1}\right) \cdot \mathrm{E}\left(\Delta_{\mathrm{N}}\left(x^{M+1}\right) | \Delta_{\mathrm{N}}\left(x^{M+1}\right) \geq \xi \cdot \overline{\Delta}\left(x^{M+1}\right)\right)}{g}\right) \\
&\cdot P\left(\Delta_{\mathrm{N}}\left(x^{M+1}\right) \geq \xi \cdot \overline{\Delta}\left(x^{M+1}\right)\right).
\end{aligned} \quad (2.116)$$

The two sum terms respectively represent the share of the overall sequence length g covered by index differences $\Delta_{\mathrm{N}}\left(x^{M+1}\right)$ and $\Delta_{\mathrm{A}}\left(x^{M+1}\right)$ above the threshold $k = \xi \cdot \overline{\Delta}\left(x^{M+1}\right)$. The respective first terms of the products are the expected values of C_{b} and C_{a}.

Using the approximation of (2.81), we derive the subsequent relations between the expected value $\overline{\Delta}\left(x^{M+1}\right)$ and the expected values of the index difference of consecutive occurrences within N and A:

$$\mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right)\right) = \frac{g}{\frac{l_{\mathrm{A}}}{h(x^{M+1})} + (g - l_{\mathrm{A}})} \cdot \mathrm{E}\left(\Delta_{\mathrm{N}}\left(x^{M+1}\right)\right) \quad (2.117)$$

$$\mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right)\right) = \frac{g}{l_{\mathrm{A}} + (g - l_{\mathrm{A}}) \cdot h\left(x^{M+1}\right)} \cdot \mathrm{E}\left(\Delta_{\mathrm{A}}\left(x^{M+1}\right)\right) \quad (2.118)$$

The distribution of the $\Delta\left(x^{M+1}\right)$ of symbols of high stationary probability is fairly concentrated around the mean value given by

$$\mathrm{E}\left(\Delta\left(x^{M+1}\right)\right) = \frac{1}{P_X\left(x^{M+1}\right)}, \quad (2.119)$$

and may be approximated by a simple geometric distribution. In case of rare events however, a more sophisticated approach is necessary. Hirata et al. [58] and Abadi [59] showed that for rare events within a fairly general stationary ergodic process, the distribution of

$\Delta\left(x^{M+1}\right)$ is best approximated by a mixture of a dirac pulse carrying the probability for immediate repetition of the symbol subsequence, such that $\Delta\left(x^{M+1}\right) \approx 0$, and a geometric distribution of mean $\frac{1}{P(\Delta(x^{M+1})>0)\cdot P_X(x^{M+1})}$, such that

$$P\left(\Delta\left(x^{M+1}\right) > k\right) \approx P\left(\Delta\left(x^{M+1}\right) > 0\right) \cdot \exp^{-P\left(\Delta\left(x^{M+1}\right)>0\right)\cdot P_X\left(x^{M+1}\right)\cdot k} \quad \forall k > 0, \tag{2.120}$$

which yields an expected value alike to the one given by (2.119). For the subsequent considerations we define

$$P_{\min}^{X^{M+1},\mathrm{N}} \stackrel{\text{def}}{=} \min_{x^{M+1}\in\mathcal{X}^{M+1}} P\left(\Delta_{\mathrm{N}}\left(x^{M+1}\right) > 0\right) \tag{2.121}$$

$$P_{\min}^{X^{M+1},\mathrm{A}} \stackrel{\text{def}}{=} \min_{x^{M+1}\in\mathcal{X}^{M+1}} P\left(\Delta_{\mathrm{A}}\left(x^{M+1}\right) > 0\right) \tag{2.122}$$

$$P_{\min}^{X^{M+1}} \stackrel{\text{def}}{=} \min\left(P_{\min}^{X^{M+1},\mathrm{N}}, P_{\min}^{X^{M+1},\mathrm{A}}\right). \tag{2.123}$$

We examine the behavior of (2.116) for several values of $h\left(x^{M+1}\right)$ and $\xi = 2$. The deduction of the expressions for $\mathrm{E}\left(\Delta_{\mathrm{N(A)}}\left(x^{M+1}\right)|\Delta_{\mathrm{N(A)}}\left(x^{M+1}\right) \geq k\right)$ and $P\left(\Delta_{\mathrm{N(A)}}\left(x^{M+1}\right) \geq k\right)$ is delayed until Section 2.3.3.

- $l_{\mathrm{A}} = 0$
  Using the equivalence

$$\mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right)\right) = \mathrm{E}\left(\Delta_{\mathrm{N}}\left(x^{M+1}\right)\right) = \frac{1}{P_X^{\mathrm{N}}\left(x^{M+1}\right)}, \tag{2.124}$$

  the expected shrinking ratio may be calculated by

$$\begin{aligned}
&\Upsilon\left(k = \frac{2}{P_X^{\mathrm{N}}\left(x^{M+1}\right)}, x^{M+1}, l_{\mathrm{A}} = 0\right) \\
&\quad\approx\; P_X^{\mathrm{N}}\left(x^{M+1}\right) \cdot \mathrm{E}\left(\Delta_{\mathrm{N}}\left(x^{M+1}\right)|\Delta_{\mathrm{N}}\left(x^{M+1}\right) \geq \frac{2}{P_X^{\mathrm{N}}\left(x^{M+1}\right)}\right) \\
&\qquad\cdot P\left(\Delta_{\mathrm{N}}\left(x^{M+1}\right) \geq \frac{2}{P_X^{\mathrm{N}}\left(x^{M+1}\right)}\right) \\
&\quad\approx\; P_X^{\mathrm{N}}\left(x^{M+1}\right) \cdot \left(\frac{2}{P_X^{\mathrm{N}}\left(x^{M+1}\right)} + \frac{1}{P\left(\Delta_{\mathrm{N}}\left(x^{M+1}\right) > 0\right) \cdot P_X^{\mathrm{N}}\left(x^{M+1}\right)}\right) \\
&\qquad\cdot P\left(\Delta_{\mathrm{N}}\left(x^{M+1}\right) > 0\right) \cdot \exp^{-P\left(\Delta_{\mathrm{N}}\left(x^{M+1}\right)>0\right)\cdot P_X^{\mathrm{N}}\left(x^{M+1}\right)\cdot\frac{2}{P_X^{\mathrm{N}}\left(x^{M+1}\right)}} \\
&\quad=\; \left(2\cdot P\left(\Delta_{\mathrm{N}}\left(x^{M+1}\right) > 0\right) + 1\right) \cdot \exp^{-2\cdot P\left(\Delta_{\mathrm{N}}\left(x^{M+1}\right)>0\right)}
\end{aligned} \tag{2.125}$$

  (2.125) is monotonically decreasing with $P\left(\Delta_{\mathrm{N}}\left(x^{M+1}\right) > 0\right)$. Thus, the bound

$$\begin{aligned}
&\Upsilon\left(k = \frac{2}{P_X^{\mathrm{N}}\left(x^{M+1}\right)}, x^{M+1}, l_{\mathrm{A}} = 0\right) \\
&\quad\leq\; \left(2\cdot P_{\min}^{X^{M+1},\mathrm{N}} + 1\right) \cdot \exp^{-2\cdot P_{\min}^{X^{M+1},\mathrm{N}}}
\end{aligned} \tag{2.126}$$

holds.

- $h\left(x^{M+1}\right) \approx 1$
  With a deduction similar to the case of $l_{\mathrm{A}} = 0$, the shrinking ratio may be calculated by

$$\begin{aligned}
&\Upsilon\left(k = 2 \cdot \overline{\Delta}\left(x^{M+1}\right), x^{M+1}, l_{\mathrm{A}} \neq 0\right) \\
&\approx \frac{l_{\mathrm{A}}}{g} \cdot \left(2 \cdot P\left(\Delta_{\mathrm{A}}\left(x^{M+1}\right) > 0\right) + 1\right) \cdot \exp^{-2 \cdot P\left(\Delta_{\mathrm{A}}\left(x^{M+1}\right) > 0\right)} \\
&\quad + \frac{g - l_{\mathrm{A}}}{g} \cdot \left(2 \cdot P\left(\Delta_{\mathrm{N}}\left(x^{M+1}\right) > 0\right) + 1\right) \cdot \exp^{-2 \cdot P\left(\Delta_{\mathrm{N}}\left(x^{M+1}\right) > 0\right)}
\end{aligned} \tag{2.127}$$

  Thus, the bound

$$\begin{aligned}
&\Upsilon\left(k = 2 \cdot \overline{\Delta}\left(x^{M+1}\right), x^{M+1}, l_{\mathrm{A}} \neq 0\right) \\
&\leq \left(\frac{l_{\mathrm{A}}}{g}\left(2 \cdot P_{\min}^{X^{M+1},\mathrm{A}} + 1\right) \cdot \exp^{-2 \cdot P_{\min}^{X^{M+1},\mathrm{A}}}\right) \\
&\quad + \left(\frac{g - l_{\mathrm{A}}}{g}\left(2 \cdot P_{\min}^{X^{M+1},\mathrm{N}} + 1\right) \cdot \exp^{-2 \cdot P_{\min}^{X^{M+1},\mathrm{N}}}\right) \\
&\leq \left(2 \cdot P_{\min}^{X^{M+1}} + 1\right) \cdot \exp^{-2 \cdot P_{\min}^{X^{M+1}}}
\end{aligned} \tag{2.128}$$

  holds.

- $h\left(x^{M+1}\right) \neq 1$
  Using (2.117) and (2.118), we write

$$\begin{aligned}
&\Upsilon\left(2 \cdot \mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right)\right), x^{M+1}, l_{\mathrm{A}} \neq 0\right) \\
&\approx \left(\frac{l_{\mathrm{A}}}{g}\right) \cdot \left(\frac{2 \cdot P\left(\Delta_{\mathrm{A}}\left(x^{M+1}\right) > 0\right)}{\frac{l_{\mathrm{A}}}{g} + h\left(x^{M+1}\right) \cdot \frac{g - l_{\mathrm{A}}}{g}} + 1\right) \cdot \exp^{-\frac{2 \cdot P\left(\Delta_{\mathrm{A}}\left(x^{M+1}\right) > 0\right)}{\frac{l_{\mathrm{A}}}{g} + h\left(x^{M+1}\right) \cdot \frac{g - l_{\mathrm{A}}}{g}}} \\
&\quad + \left(\frac{g - l_{\mathrm{A}}}{g}\right) \cdot \left(\frac{2 \cdot P\left(\Delta_{\mathrm{N}}\left(x^{M+1}\right) > 0\right)}{\frac{l_{\mathrm{A}}}{g} \cdot \left(\frac{1}{h\left(x^{M+1}\right)} - 1\right) + 1} + 1\right) \cdot \exp^{-\frac{2 \cdot P\left(\Delta_{\mathrm{N}}\left(x^{M+1}\right) > 0\right)}{\frac{l_{\mathrm{A}}}{g} \cdot \left(\frac{1}{h\left(x^{M+1}\right)} - 1\right) + 1}} \\
&= \left(\frac{l_{\mathrm{A}}}{g}\right) \cdot \left(\frac{2 \cdot P\left(\Delta_{\mathrm{A}}\left(x^{M+1}\right) > 0\right)}{\frac{l_{\mathrm{A}}}{g} + h\left(x^{M+1}\right) \cdot \frac{g - l_{\mathrm{A}}}{g}} + 1\right) \cdot \exp^{-\frac{2 \cdot P\left(\Delta_{\mathrm{A}}\left(x^{M+1}\right) > 0\right)}{\frac{l_{\mathrm{A}}}{g} + h\left(x^{M+1}\right) \cdot \frac{g - l_{\mathrm{A}}}{g}}} \\
&\quad + \left(\frac{g - l_{\mathrm{A}}}{g}\right) \cdot \left(\frac{2 \cdot P\left(\Delta_{\mathrm{N}}\left(x^{M+1}\right) > 0\right) \cdot h\left(x^{M+1}\right)}{\frac{l_{\mathrm{A}}}{g} + h\left(x^{M+1}\right) \cdot \frac{g - l_{\mathrm{A}}}{g}} + 1\right) \\
&\quad \cdot \exp^{-\frac{2 \cdot P\left(\Delta_{\mathrm{N}}\left(x^{M+1}\right) > 0\right) \cdot h\left(x^{M+1}\right)}{\frac{l_{\mathrm{A}}}{g} + h\left(x^{M+1}\right) \cdot \frac{g - l_{\mathrm{A}}}{g}}}.
\end{aligned} \tag{2.129}$$

For $h\left(x^{M+1}\right) > 1$ (symbols typical of the normal data), we would like to derive an upper bound of $\Upsilon\left(2 \cdot \mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right)\right), x^{M+1}, l_{\mathrm{A}} \neq 0\right)$. As we will show below, using a suitable approximation of (2.129) given by (2.133), $\Upsilon\left(2 \cdot \mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right)\right), x^{M+1}, l_{\mathrm{A}} \neq 0\right)$ is approximately strictly monotonic increasing with respect to $h\left(x^{M+1}\right)$ for all $l_{\mathrm{A}} \in \left[0, \frac{g}{3}\right]$. Assuming $h\left(x^{M+1}\right) \gg 1$, $\Upsilon\left(2 \cdot \mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right)\right), x^{M+1}, l_{\mathrm{A}} \neq 0\right)$ is approximately strictly monotonic increasing with respect to l_{A}. The upper bound of $\Upsilon\left(2 \cdot \mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right)\right), x^{M+1}, l_{\mathrm{A}} \neq 0\right)$ is thus derived from (2.129) as follows:

$$\begin{aligned}
&\Upsilon\left(2 \cdot \mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right)\right), x^{M+1}, l_{\mathrm{A}} \neq 0\right) \\
&\leq \Upsilon\left(2 \cdot \mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right)\right), x^{M+1}, l_{\mathrm{A}} = \frac{g}{3} \,\middle|\, h\left(x^{M+1}\right) \gg 1\right) \\
&\approx \frac{1}{3} + \frac{2}{3} \cdot \left(3 \cdot P\left(\Delta_{\mathrm{N}}\left(x^{M+1}\right) > 0\right) + 1\right) \cdot \exp^{-3 \cdot P\left(\Delta_{\mathrm{N}}\left(x^{M+1}\right) > 0\right)} \\
&\leq \frac{1}{3} + \frac{2}{3} \cdot \left(3 \cdot P_{\min}^{X^{M+1},\mathrm{N}} + 1\right) \cdot \exp^{-3 \cdot P_{\min}^{X^{M+1},\mathrm{N}}}
\end{aligned} \tag{2.130}$$

For symbols featuring $h\left(x^{M+1}\right) \leq h_{\mathrm{th}}$ (symbols typical of the anomaly), we derive a lower bound of $\Upsilon\left(2 \cdot \mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right)\right), x^{M+1}, l_{\mathrm{A}} \neq 0\right)$ using the second sum term of (2.129):

$$\begin{aligned}
&\Upsilon\left(2 \cdot \mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right)\right), x^{M+1}, l_{\mathrm{A}} \neq 0\right) \\
&\geq \left(\frac{g - l_{\mathrm{A}}}{g}\right) \cdot \left(\frac{2 \cdot P\left(\Delta_{\mathrm{N}}\left(x^{M+1}\right) > 0\right) \cdot h\left(x^{M+1}\right)}{\frac{l_{\mathrm{A}}}{g} + h\left(x^{M+1}\right) \cdot \frac{g - l_{\mathrm{A}}}{g}} + 1\right) \\
&\quad \cdot \exp^{-\frac{2 \cdot P\left(\Delta_{\mathrm{N}}\left(x^{M+1}\right) > 0\right) \cdot h\left(x^{M+1}\right)}{\frac{l_{\mathrm{A}}}{g} + h\left(x^{M+1}\right) \cdot \frac{g - l_{\mathrm{A}}}{g}}}
\end{aligned} \tag{2.131}$$

(2.131) is strictly monotonic decreasing with respect to $h\left(x^{M+1}\right)$ and $P\left(\Delta_{\mathrm{N}}\left(x^{M+1}\right) > 0\right)$. Thus the lower bound uses $h\left(x^{M+1}\right) = h_{\mathrm{th}} < 1$ and $P\left(\Delta_{\mathrm{N}}\left(x^{M+1}\right) > 0\right) = 1$. (2.131) has a single maximum within the range $l_{\mathrm{A}} \in \left[0, \frac{g}{3}\right]$. For $h_{\mathrm{th}} = \frac{1}{4}$, the minimum is found at $l_{\mathrm{A}} = 0$. Thus, the lower bound

$$\begin{aligned}
&\Upsilon\left(2 \cdot \mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right)\right), x^{M+1}, l_{\mathrm{A}} \neq 0\right) \\
&\geq \Upsilon\left(2 \cdot \mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right)\right), x^{M+1}, l_{\mathrm{A}} \to 0\right) \\
&\geq 3 \cdot \exp^{-2} \approx 0.4
\end{aligned} \tag{2.132}$$

holds.

We now explain the validity of the deduction for the upper bound of $\Upsilon\left(2 \cdot \mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right)\right), x^{M+1}, l_{\mathrm{A}} \neq 0\right)$ for $h\left(x^{M+1}\right) > 1$ given by (2.130):

Making use of the fact that for $h\left(x^{M+1}\right) > 1$ the terms $\frac{2 \cdot P\left(\Delta_{\mathrm{N}}\left(x^{M+1}\right) > 0\right) \cdot h\left(x^{M+1}\right)}{\frac{l_{\mathrm{A}}}{g} + h\left(x^{M+1}\right) \cdot \frac{g - l_{\mathrm{A}}}{g}}$ and $\frac{2 \cdot P\left(\Delta_{\mathrm{A}}\left(x^{M+1}\right) > 0\right)}{\frac{l_{\mathrm{A}}}{g} + h\left(x^{M+1}\right) \cdot \frac{g - l_{\mathrm{A}}}{g}}$ are respectively restricted to the ranges $\left[2P\left(\Delta_{\mathrm{N}}\left(x^{M+1}\right) > 0\right), 3\right]$

and $[0,2]$, we create a linear approximation of (2.129):

$$\begin{aligned}
&\Upsilon\left(2\cdot \mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right)\right), x^{M+1}, l_{\mathrm{A}} \neq 0\right) \\
&\approx \left(\frac{l_{\mathrm{A}}}{g}\right)\cdot\left(1-0.3\cdot\frac{2\cdot P\left(\Delta_{\mathrm{A}}\left(x^{M+1}\right)>0\right)}{\frac{l_{\mathrm{A}}}{g}+h\left(x^{M+1}\right)\cdot\frac{g-l_{\mathrm{A}}}{g}}\right) \\
&\quad +\left(\frac{g-l_{\mathrm{A}}}{g}\right)\cdot\left(0.83-0.21\cdot\frac{2\cdot P\left(\Delta_{\mathrm{N}}\left(x^{M+1}\right)>0\right)\cdot h\left(x^{M+1}\right)}{\frac{l_{\mathrm{A}}}{g}+h\left(x^{M+1}\right)\cdot\frac{g-l_{\mathrm{A}}}{g}}\right) \qquad (2.133)
\end{aligned}$$

The numerical parameters of (2.133) can be deduced a follows:

For the range $z\in[0,3]$, the function $f(z)=(z+1)\cdot e^{-z}$ is a strictly monotonic decreasing function with $f(z)\in\left[1,4\cdot e^{-3}\right]$. Thus we approximate the behavior of the first term of (2.129) by

$$\begin{aligned}
(z+1)\cdot e^{-z} &\approx f(z=0)-\frac{f(z=0)-f(z=2)}{2}\cdot z \\
&\approx 1-0.3z \\
z &= \frac{2\cdot P\left(\Delta_{\mathrm{A}}\left(x^{M+1}\right)>0\right)}{\frac{l_{\mathrm{A}}}{g}+h\left(x^{M+1}\right)\cdot\frac{g-l_{\mathrm{A}}}{g}},\quad z\in[0,2]\,. \qquad (2.134)
\end{aligned}$$

For the second term of (2.129), we derive

$$\begin{aligned}
&(z+1)\cdot e^{-z} \\
&\approx f\left(z=2P\left(\Delta_{\mathrm{N}}\left(x^{M+1}\right)>0\right)\right) \\
&\quad -\frac{f\left(z=2P\left(\Delta_{\mathrm{N}}\left(x^{M+1}\right)>0\right)\right)-f(z=3)}{3-2P\left(\Delta_{\mathrm{N}}\left(x^{M+1}\right)>0\right)}\cdot\left(z-2P\left(\Delta_{\mathrm{N}}\left(x^{M+1}\right)>0\right)\right) \\
&\approx 0.83-0.21z \\
&\quad z=\frac{2\cdot P\left(\Delta_{\mathrm{N}}\left(x^{M+1}\right)>0\right)\cdot h\left(x^{M+1}\right)}{\frac{l_{\mathrm{A}}}{g}+h\left(x^{M+1}\right)\cdot\frac{g-l_{\mathrm{A}}}{g}},\quad z\in\left[2P\left(\Delta_{\mathrm{N}}\left(x^{M+1}\right)>0\right),3\right], \\
&\qquad (2.135)
\end{aligned}$$

supposing $P\left(\Delta_{\mathrm{N}}\left(x^{M+1}\right)>0\right)\geq 0.9$. (2.135) forms an upper bound of the actual term for all $P\left(\Delta_{\mathrm{N}}\left(x^{M+1}\right)>0\right)\in[0.5,1]$ because of the falling inflection point of $f(z)$ being located at $z=1$.

Comparing (2.132), (2.130), (2.128) and (2.126), it is evident that a parameter setting for avoiding false positives depends on $P_{\min}^{X^{M+1}}$ and $P_{\min}^{X^{M+1},\mathrm{N}}$. We supposed

$$P_{\min}^{X^{M+1}},\ P_{\min}^{X^{M+1},\mathrm{N}}>0.9. \qquad (2.136)$$

Thus,

$$\Upsilon\left(2 \cdot \mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right)\right), x^{M+1}, l_{\mathrm{A}} \neq 0 \mid h\left(x^{M+1}\right) \geq 1\right) < 0.5 \tag{2.137}$$

holds. Using the notation introduced above and remembering that the anomalous blocks are uniformly distributed within the sequence, the expected value of the unscaled average index difference after eliminating the $\Delta\left(x^{M+1}\right)$ according to the setting $\xi = 2$ and h_{th} may be expressed as

$$\begin{aligned}
&\mathrm{E}\left(T_j(x^{M+1}, \text{Algorithm 2}) \mid l_{\mathrm{A}} \neq 0, \xi = 2\right) \\
&\quad \approx \quad \left(1 - \Upsilon\left(2 \cdot \mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right)\right), x^{M+1}, l_{\mathrm{A}} \neq 0\right)\right) \cdot \mathrm{E}\left(T_j\left(x^{M+1}\right) \mid l_{\mathrm{A}} = 0, \xi = \infty\right) \\
&\quad = \quad \left(1 - \Upsilon\left(2 \cdot \mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right)\right), x^{M+1}, l_{\mathrm{A}} \neq 0\right)\right) \cdot \frac{j^2 + (g-j)^2}{2g}.
\end{aligned} \tag{2.138}$$

Combining (2.138) and the scaling definition of (2.80), we get

$$\mathrm{E}\left(S_j\left(x^{M+1}\right) \mid l_{\mathrm{A}} \neq 0, \xi = 2\right) \quad \approx \quad \left(1 - \Upsilon\left(2 \cdot \mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right)\right), x^{M+1}, l_{\mathrm{A}} \neq 0\right)\right) \cdot \frac{g}{2}. \tag{2.139}$$

Therefore, a setting τ_{th} avoiding false positives has to be chosen according to

$$\tau_{\mathrm{th}} \quad \leq \quad \left(1 - \Upsilon\left(\frac{2}{P_X^{\mathrm{N}}\left(x^{M+1}\right)}, x^{M+1}, l_{\mathrm{A}} \neq 0 \mid h\left(x^{M+1}\right) \geq 1\right)\right) \cdot \frac{g}{2}. \tag{2.140}$$

Inserting (2.137), we get

$$\tau_{\mathrm{th}} \quad \leq \quad 0.25 \cdot g. \tag{2.141}$$

2.3.3 Parameter Setting: i.i.d. Source

Setting of ξ

In case of i.i.d. generation, the index difference of neighboring occurrences of identical symbol subsequences forms an independent geometrically distributed variable, a fact we may use to deduce the setting of $\xi > 1$. A convenient approximation of the probability that a $\Delta\left(x^{M+1}\right)$ within a certain range expressed as a product of the expected value of the index difference

$$\mathrm{E}\left(\Delta\left(x^{M+1}\right)\right) = \frac{1}{P_X^{\mathrm{N}}\left(x^{M+1}\right)} \tag{2.142}$$

and ξ is observed, may be deduced as follows. We first express the desired probability using the geometric distribution:

$$\begin{aligned}
P\left(\Delta\left(x^{M+1}\right) \leq \xi \cdot \mathrm{E}\left(\Delta\left(x^{M+1}\right)\right)\right) &= 1 - \left(1 - P_X^{\mathrm{N}}\left(x^{M+1}\right)\right)^{\xi \cdot \mathrm{E}\left(\Delta\left(\mathrm{x}^{\mathrm{M+1}}\right)\right)} \\
&= 1 - \left(1 - P_X^{\mathrm{N}}\left(x^{M+1}\right)\right)^{\frac{\xi}{P_X^{\mathrm{N}}\left(x^{M+1}\right)}}
\end{aligned} \tag{2.143}$$

By using the approximation $(1-p)^{\frac{1}{p}} \approx \exp^{-1}$, we get

$$P\left(\Delta\left(x^{M+1}\right) \leq \xi \cdot \mathrm{E}\left(\Delta\left(x^{M+1}\right)\right)\right) \approx 1 - \exp^{-\xi}, \tag{2.144}$$

and hence

$$P\left(\Delta\left(x^{M+1}\right) \geq \xi \cdot \mathrm{E}\left(\Delta\left(x^{M+1}\right)\right)\right) \approx \exp^{-\xi}. \tag{2.145}$$

For the subsequent calculations, we use $\Upsilon\left(k, x^{M+1}, l_{\mathrm{A}}\right)$ as defined by (2.115) for representation of the expected share of a sequence of suitable length $g \gg \frac{1}{P_X^{\mathrm{N}}(x^{M+1})}, \frac{1}{P_X^{\mathrm{A}}(x^{M+1})}$ including an anomaly of overall length l_{A} covered by index differences $\Delta\left(x^{M+1}\right)$ of length above k. For $l_{\mathrm{A}} = 0$, $\Upsilon\left(k, x^{M+1}, l_{\mathrm{A}} = 0\right)$ may be approximated by

$$\begin{aligned}
&\Upsilon\left(k, x^{M+1}, l_{\mathrm{A}} = 0\right) \\
&\approx \frac{\left(g \cdot P_X^{\mathrm{N}}\left(x^{M+1}\right) - 1\right) \cdot P\left(\Delta\left(x^{M+1}\right) > k\right) \cdot E\left(\Delta\left(x^{M+1}\right) | \Delta\left(x^{M+1}\right) > k\right)}{g}
\end{aligned} \tag{2.146}$$

The first term within the enumerator represents the expected number $C_{\mathrm{b}} + C_{\mathrm{a}}$ of $\Delta\left(x^{M+1}\right)$. The second term represents the probability that a $\Delta\left(x^{M+1}\right)$ will surpass k. The third term represents the expected length of a $\Delta\left(x^{M+1}\right)$, if $\Delta\left(x^{M+1}\right)$ is longer than k. Keeping in mind the condition of $\frac{1}{P_X^{\mathrm{N}}(x^{M+1})} \ll g$, the complete transcription of 2.146 reads

$$\begin{aligned}
\Upsilon\left(k, x^{M+1}, l_{\mathrm{A}} = 0\right) &\approx \frac{\left(g \cdot P_X^{\mathrm{N}}\left(x^{M+1}\right) - 1\right) \cdot \left(1 - P_X^{\mathrm{N}}\left(x^{M+1}\right)\right)^k \cdot \left(k + \frac{1}{P_X^{\mathrm{N}}(x^{M+1})}\right)}{g} \\
&\approx P_X^{\mathrm{N}}\left(x^{M+1}\right) \cdot \left(1 - P_X^{\mathrm{N}}\left(x^{M+1}\right)\right)^k \cdot \left(k + \frac{1}{P_X^{\mathrm{N}}\left(x^{M+1}\right)}\right)
\end{aligned} \tag{2.147}$$

The third term of (2.147) is derived by the fact that for all $\Delta\left(x^{M+1}\right) > k$, the difference $\Delta\left(x^{M+1}\right) - k$ is again a geometrically distributed variable of expected value $\frac{1}{P_X^{\mathrm{N}}(x^{M+1})}$, and hence

$$\begin{aligned}
\mathrm{E}\left(\Delta\left(\mathrm{x}^{\mathrm{M+1}}\right) | \Delta\left(\mathrm{x}^{\mathrm{M+1}}\right) > \mathrm{k}\right) &= k + \mathrm{E}\left(\Delta\left(\mathrm{x}^{\mathrm{M+1}}\right)\right) && (2.148) \\
&= k + \frac{1}{P_X^{\mathrm{N}}\left(x^{M+1}\right)} && (2.149)
\end{aligned}$$

holds. Replacing k by $\xi \cdot \mathrm{E}\left(\Delta\left(x^{M+1}\right)\right)$ and using (2.145), (2.147) may be simplified to

$$\begin{aligned}
&\Upsilon\left(\xi \cdot \mathrm{E}\left(\Delta\left(x^{M+1}\right)\right), x^{M+1}, l_{\mathrm{A}} = 0\right) \\
&\approx P_X^{\mathrm{N}}\left(x^{M+1}\right) \cdot \left(1 - P_X^{\mathrm{N}}\left(x^{M+1}\right)\right)^{\xi \cdot \mathrm{E}\left(\Delta\left(x^{M+1}\right)\right)} \cdot \left(\xi \cdot \mathrm{E}\left(\Delta\left(x^{M+1}\right)\right) + \frac{1}{P_X^{\mathrm{N}}\left(x^{M+1}\right)}\right) \\
&= P_X^{\mathrm{N}}\left(x^{M+1}\right) \cdot \left(1 - P_X^{\mathrm{N}}\left(x^{M+1}\right)\right)^{\frac{\xi}{P_X^{\mathrm{N}}(x^{M+1})}} \cdot \left(\xi \frac{1}{P_X^{\mathrm{N}}\left(x^{M+1}\right)} + \frac{1}{P_X^{\mathrm{N}}\left(x^{M+1}\right)}\right) \\
&\approx \exp^{-\xi} \cdot (1 + \xi),
\end{aligned} \tag{2.150}$$

again dependent on the condition $\frac{1}{P_X^{\mathrm{N}}(x^{M+1})} \ll g$. Using (2.150), we may set ξ to a value leaving the majority of index differences of neighboring identical sequences in case of no anomaly untouched. Because in case of no anomaly the thus deleted index differences are uniformly distributed within the overall sequence, the expected average index difference in case of no anomaly after deletion may be calculated using (2.12):

$$\begin{aligned} &\mathrm{E}\left(S_j\left(x^{M+1}\right)\,|\xi, \Delta_{C_\mathrm{b}} \leq \beta \cdot \overline{\Delta}\left(x^{M+1}\right), \Delta_{C_\mathrm{b}+1} \leq \beta \cdot \overline{\Delta}\left(x^{M+1}\right)\right) \\ &\quad \approx \quad \mathrm{E}\left(S_j\left(x^{M+1}\right)\,|l_\mathrm{A} = 0, \xi = \infty\right) \cdot \left(1 - \Upsilon\left(\xi \cdot \mathrm{E}\left(\Delta\left(x^{M+1}\right)\right), x^{M+1}, l_\mathrm{A} = 0\right)\right) \end{aligned} \tag{2.151}$$

Note that the expected value given by (2.151) is attained if and only if both of the index differences adjourning the symbol sequence found at index j fall below the threshold $\overline{\Delta}\left(x^{M+1}\right)$. Otherwise, the expected value is given by (2.12). Because the anomalous blocks are roughly uniformly distributed within the sequence by the initial block rearrangement, (2.151) holds for the case of anomaly as well.

In order to derive a suitable setting of $\xi > 1$, we are interested in the behavior of $\Upsilon\left(\xi \cdot \mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right)\right), x^{M+1}, l_\mathrm{A} \neq 0\right)$ for both $h\left(x^{M+1}\right) < h_\mathrm{th}$ and $h\left(x^{M+1}\right) \geq 2$. The limit of 2 was chosen in order to create a set of neutral sequences. This examination splits the possible range of $h\left(x^{M+1}\right)$ into three parts.

- $h\left(x^{M+1}\right) < h_\mathrm{th}$: Sequences considered representative of the anomalous data.
- $h_\mathrm{th} \leq h\left(x^{M+1}\right) \leq 2$: Neutral sequences.
- $h\left(x^{M+1}\right) \geq 2$: Sequences considered representative of the normal data.

We express $\mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right)\right)$ as a function of $P_X^{\mathrm{N}}\left(x^{M+1}\right)/P_X^{\mathrm{A}}\left(x^{M+1}\right)$ using (2.81) and $h\left(x^{M+1}\right)$:

$$\begin{aligned} \mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right)\right) \quad &\approx \quad \frac{g}{l_\mathrm{A} + (g - l_\mathrm{A}) \cdot h\left(x^{M+1}\right)} \cdot \frac{1}{P_X^{\mathrm{A}}\left(x^{M+1}\right)} \\ &= \quad \frac{1}{\frac{l_\mathrm{A}}{g} + h\left(x^{M+1}\right) \cdot \frac{g - l_\mathrm{A}}{g}} \cdot \frac{1}{P_X^{\mathrm{A}}\left(x^{M+1}\right)} \end{aligned} \tag{2.152}$$

and

$$\begin{aligned} \mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right)\right) \quad &\approx \quad \frac{g}{\frac{l_\mathrm{A}}{h(x^{M+1})} + g - l_\mathrm{A}} \cdot \frac{1}{P_X^{\mathrm{N}}\left(x^{M+1}\right)} \\ &= \quad \frac{1}{\frac{l_\mathrm{A}}{g} \cdot \left(\frac{1}{h(x^{M+1})} - 1\right) + 1} \cdot \frac{1}{P_X^{\mathrm{N}}\left(x^{M+1}\right)} \end{aligned} \tag{2.153}$$

Respectively inserting into (2.150), we derive:

$$\begin{aligned}
&\Upsilon\left(\xi \cdot \mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right)\right), x^{M+1}, l_{\mathrm{A}} \neq 0\right) \\
&\approx \left(\frac{g-l_{\mathrm{A}}}{g}\right) \cdot \left(\frac{\xi}{\frac{l_{\mathrm{A}}}{g} \cdot \left(\frac{1}{h\left(x^{M+1}\right)}-1\right)+1}+1\right) \cdot \exp^{-\frac{\xi}{\frac{l_{\mathrm{A}}}{g} \cdot \left(\frac{1}{h\left(x^{M+1}\right)}-1\right)+1}} \\
&\quad + \left(\frac{l_{\mathrm{A}}}{g}\right) \cdot \left(\frac{\xi}{\frac{l_{\mathrm{A}}}{g}+h\left(x^{M+1}\right) \cdot \frac{g-l_{\mathrm{A}}}{g}}+1\right) \cdot \exp^{-\frac{\xi}{\frac{l_{\mathrm{A}}}{g}+h\left(x^{M+1}\right) \cdot \frac{g-l_{\mathrm{A}}}{g}}}
\end{aligned} \tag{2.154}$$

The first term of the sum of (2.154) represents the share of
$\Upsilon\left(\xi \cdot \mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right)\right), x^{M+1}, l_{\mathrm{A}} \neq 0\right)$ within the normal data, while the second term represents the share within anomalous data. The subsequent bounds hold because $(\xi+1) \cdot \exp^{-\xi}$ of (2.150) is strictly monotonic decreasing with respect to $\xi > 0$.

$$\begin{aligned}
&\Upsilon\left(\xi \cdot \mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right)\right), x^{M+1}, l_{\mathrm{A}} \neq 0 \mid h\left(x^{M+1}\right) \leq h_{\mathrm{th}}\right) \\
&\geq \left(\frac{g-l_{\mathrm{A}}}{g}\right) \cdot \left(\frac{\xi}{\frac{l_{\mathrm{A}}}{g} \cdot \left(\frac{1}{h_{\mathrm{th}}}-1\right)+1}+1\right) \cdot \exp^{-\frac{\xi}{\frac{l_{\mathrm{A}}}{g} \cdot \left(\frac{1}{h_{\mathrm{th}}}-1\right)+1}}
\end{aligned} \tag{2.155}$$

$$\begin{aligned}
&\Upsilon\left(\xi \cdot \mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right)\right), x^{M+1}, l_{\mathrm{A}} \neq 0, h\left(x^{M+1}\right) \geq 2\right) \\
&\leq \left(\frac{g-l_{\mathrm{A}}}{g}\right) \cdot \left(\frac{\xi}{\frac{l_{\mathrm{A}}}{g} \cdot \left(\frac{1}{2}-1\right)+1}+1\right) \cdot \exp^{-\frac{\xi}{\frac{l_{\mathrm{A}}}{g} \cdot \left(\frac{1}{2}-1\right)+1}} + \left(\frac{l_{\mathrm{A}}}{g}\right)
\end{aligned} \tag{2.156}$$

Plotting the two bounds (2.155) (2.156) and (2.150) for $l_{\mathrm{A}} = 0.05g$, $l_{\mathrm{A}} = 0.15g$, $l_{\mathrm{A}} = 0.25g$ and $l_{\mathrm{A}} = 0.35g$ (Figure 2.11), we observe that a setting of $\xi \approx 2$ keeps (2.155) stable for various l_{A} while still retaining a reasonable distance to the case of no anomaly. Thus, we subsequently use a setting of $\xi = 2$.

Figure 2.12 shows the plotting of (2.154) for various l_{A} and $\xi = 2$. The plots show there is an almost linear increase of $\Upsilon\left(\xi \cdot \mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right)\right), x^{M+1}, l_{\mathrm{A}} \neq 0 \mid \xi = 2\right)$ for $h\left(x^{M+1}\right) > 1$ with respect to l_{A}. Contrary, for $h\left(x^{M+1}\right) < 0.25$, there will be no significant increase of $\Upsilon\left(\xi \cdot \mathrm{E}\left(\overline{\Delta}\left(x^{M+1}\right)\right), x^{M+1}, l_{\mathrm{A}} \neq 0 \mid \xi = 2\right)$ with respect to l_{A} for $l_{\mathrm{A}} \geq 0.15 \cdot g$. Moreover, for $h\left(x^{M+1}\right) \leq 0.25$, Fig. 2.12 shows a distinctive decrease of (2.154) for $l_{\mathrm{A}} = 0.35g$, indicating the strong degradation of performance for $l_{\mathrm{A}} \geq \frac{g}{3}$.

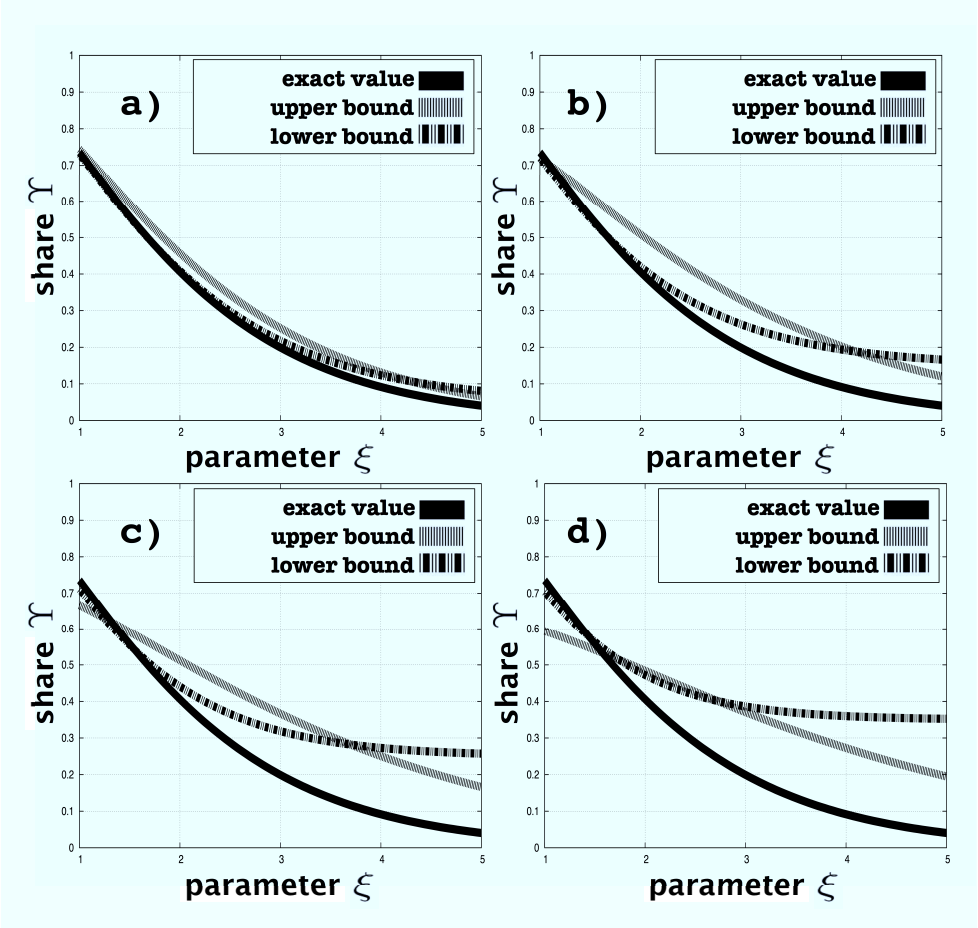

Figure 2.11: $a) l_{\mathrm{A}} = 0.05g$, b)$l_{\mathrm{A}} = 0.15g$, c)$l_{\mathrm{A}} = 0.25g$, d)$l_{\mathrm{A}} = 0.35g$

Setting of τ_{th} for $\xi = 2$

In case of i.i.d. generation, an improved setting of the threshold τ_{th} guided by the theory developed above is possible. Using (2.150), (2.151), (2.154), and (2.80), we deduce that for $\xi = 2$, the threshold has to be selected from a range

$$\begin{aligned} \tau_{\mathrm{th}} &\leq \left(1 - \Upsilon\left(\frac{2}{P_X^{\mathrm{N}}\left(x^{M+1}\right)}, x^{M+1}, l_{\mathrm{A}} = 0\right)\right) \cdot \mathrm{E}\left(S_j\left(x^{M+1}\right) | l_{\mathrm{A}} = 0, \xi = \infty\right) \\ &\approx 0.6 \cdot \mathrm{E}\left(S_j\left(x^{M+1}\right) | l_{\mathrm{A}} = 0, \xi = \infty\right) \qquad (2.157) \\ &= 0.6 \cdot 0.5 \cdot g. \qquad (2.158) \end{aligned}$$

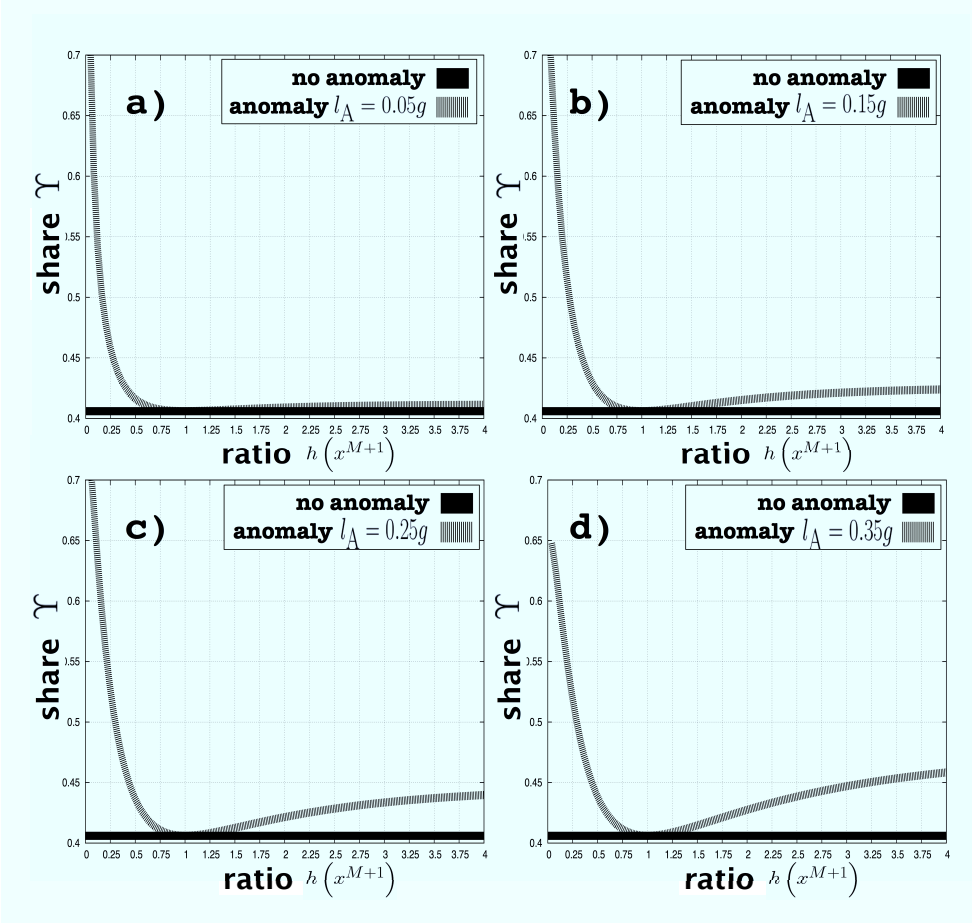

Figure 2.12: Plot of (2.154) for various anomaly ratios.

If we may be sure that a substantial portion of the generated symbols of the anomaly data shows a $h\left(x^{M+1}\right) \ll 0.25$, we may set $\tau_{\mathrm{th}} = 0.45 \cdot \mathrm{E}\left(S_j\left(x^{M+1}\right) | l_{\mathrm{A}} = 0, \xi = \infty\right)$, which will create an almost perfect detector result. On the other hand, a relaxed condition supposing that the symbols typical of the anomaly are $h\left(x^{M+1}\right) \approx 0.25$ demands a higher threshold $\tau_{\mathrm{th}} = 0.5 \cdot \mathrm{E}\left(S_j\left(x^{M+1}\right) | l_{\mathrm{A}} = 0, \xi = \infty\right)$, increasing the false positive rate.

Table 2.3: Possible combinations of true class and classification result.

	Classification Result: Anomaly	Classification Result: Normal
True Class: Anomaly	True Positive TP	False Negative FN
True Class: Normal	False Positive FP	True Negative TN

2.3.4 Computational Cost

With a line of argument similar to the one used in Section 2.2.3, we can show that the computational cost is bounded by a function of order

$$O\left(g^2\right) \tag{2.159}$$

2.4 Experimental Results

2.4.1 Detection Quality Evaluation: Receiver Operating Characteristic

After classification, for every member of the set $\mathcal{S}$, there are four possible combinations of true class and classification result, as shown in Table 2.3. Here, TP, FN, FP and TN are variables representing the occurrence numbers of the respective combinations within $\mathcal{S}$, such that

$$n = TP + FN + FP + TN. \tag{2.160}$$

These four variables can be used to calculate various coefficients, which show the tradeoff between the number of correctly detected anomalous data points and the number of normal data points mistakenly classified anomalous. These include

- True Positive Rate or Recall: percentage of correctly detected anomalous data

$$TPR = \frac{TP}{TP + FN} \tag{2.161}$$

- False Positive Rate: percentage of normal data mistakenly classified anomalous

$$FPR = \frac{FP}{FP + TN} \tag{2.162}$$

- Precision: percentage of true anomalies within the data classified anomalous

$$PR = \frac{TP}{TP + FP} \tag{2.163}$$

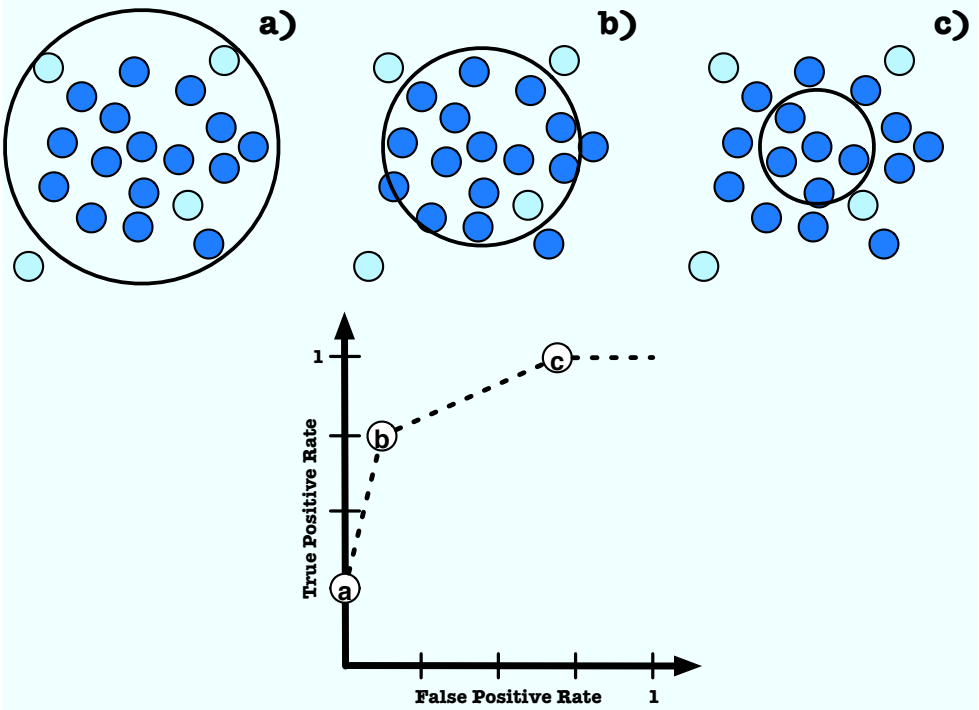

Figure 2.13: a) Overly low true positive rate b) Optimum combination of true positive rate and false positive rate c) Overly high false positive rate

- Accuracy: percentage of overall correct classifications

$$ACC = \frac{TP + TN}{TP + FN + FP + TN} \tag{2.164}$$

One of the most frequently used measures of classification quality is the combination of true positive rate and false positive rate. In order to show the effect of parameter tuning on the result, a so-called Receiver Operating Characteristic (ROC) graph [60] can be used. The name was coined by radar engineers during the second world war, who where interested in finding the optimum tradeoff point of correct detection of enemy aircraft and false alarms. They plotted the points representing the combination of true positive rate and false positive rate for certain classifier settings. Figure 2.13 shows an example with a primitive classifier. Every data point inside the circle is classified as normal, while everything outside the circle is classified as anomalous. Fixing the center, we alter the radius. While an overly large radius yields zero false positive rate but insufficient true positive rate, an overly small radius yields an unacceptable false positive rate. The optimum radius yields a good tradeoff of true positive and false positive rate.

2.4.2 ROC Parameter Calculation

In order to enhance the comparability of ROC curves generated from different data samples featuring different anomaly lengths, we conceived a new scheme for ROC calculation. We estimated the mean m and the deviation σ of the normal values within the retrieved sequence of scalars (similarity measures, densities, average index difference vector lengths, percentages etc.) by calculating the median and the median absolute deviation (MAD)[4] [61]. The threshold c_{th} used for ROC calculation is then chosen from the interval $[m, m + 4\sigma]$.

2.4.3 Artificial Data

We used i.i.d. symbol generation, with both the normal distribution and the anomalous distribution generated according to the subsequent methods. The expected value of the generated probabilities is equal to the inverse of the alphabet size Z.

$$\mathrm{E}\left(P\left(x\right)\right) = \frac{1}{Z} \tag{2.165}$$

1. Uniform Generation:
   Z random values uniformly distributed within a fixed interval with non-negative limits, $[u_1, u_2]$ are generated independently and assigned to the respective symbols, followed by normalization using the sum of the generated values. The range of percentage values observed with high probability depends on $[u_1, u_2]$

   $$z \in \left[\frac{u_1}{Z \cdot (\frac{u_2+u_1}{2})} > 0, \frac{u_2}{Z \cdot (\frac{u_2+u_1}{2})} < \frac{2}{Z}\right] \tag{2.166}$$

   The generated percentage values are approximately uniformly distributed.

2. Exponential Generation:
   While above model is easy to handle, certain real world data displays an exponential distribution of percentages. Thus we used an exponential distribution to generate the Z independent samples.

The overall sequence length g was set to $4,000$. Because we set the alphabet size $Z = 100$ to the same range as the block size $b = 200 = 0.05 \cdot g$ and because of i.i.d. generation, the subsequent simulations use zero maximum memory length $M_{\max}$.

[4] In order to avoid zero median absolute deviation, we only consider non-zero entries of the sequence of differences between the scalars and the median.

Setting of the Algorithm Supposing Local Concentration of Anomalous Blocks

Because the length of the anomaly l_{A} is supposed to be unknown, we choose a setting $\tau_{\mathrm{th}} = \frac{g}{6} = 666$.

Settings of the Algorithm Allowing for Arbitrary Distribution of Anomalous Blocks

A setting $\xi = 2$ leaves approximately 60% of the covered area in case of no anomaly untouched, while at least 45% are eliminated in case of anomaly and the symbol sequence being typical of the anomaly. The parameter β is set to 0.5. The threshold τ_{th} is set to $1,000$. The dimension of the vector calculated from the average index differences is set to $a = 3$.The minimum number of identical symbols within a block is set to $\nu_{\min} = 4$.

Settings of the Previous Approaches

The cluster algorithm is set up with $k = 1$ and $w = 0.9$, with the latter setting retrieved via trial and error. The optimum depth of the suffix tree turned out to be $t = 0$, regardless of the setting of $S_{\min}$.

We simulate anomaly lengths l_{A} of $0.05 \cdot g$, $0.15 \cdot g$, and $0.25 \cdot g$ respectively, using uniform distribution (interval $[0.1, 1]$) and exponential generation for symbol distribution generation. The uniform generation limits the range of $h(x)$ to $\frac{1}{10} \leq h(x) \leq 10$ ($P(x) \in \left[\frac{1}{5.5 \cdot Z}, \frac{10}{5.5 \cdot Z}\right]$), thus posing are more difficult task than a symbol distribution generated by the exponential approach. Each simulation consists of $10,000$ repetitions of sequence generation, with the symbol distributions generated anew after every 100th run. The anomalous blocks are randomly rearranged after initial generation in case of the algorithm allowing for arbitrary distribution of anomalous blocks.

The Figures 2.14 and 2.15 show the ROC curves returned by our algorithms and previous approaches, as well as the decrease of the false positive rate in case of no anomalous data within the interval $[m + 2\sigma, m + 4\sigma]$. The main difference between our two algorithms is the extraordinary low false positive rate of the algorithm supposing local concentration of anomalous blocks. The reason is found in the fact that the algorithm will only detect symbols inside anomalous data in case of long l_{A} or $h(x) \ll 1$, while the settings of $\xi = 2$ and $\beta = 0.5$ of the algorithm allowing for arbitrary distribution of anomalous blocks will also shrink a certain percentage of normal average index differences, causing a higher percentage of the average index differences of normal symbols to be misclassified. Because no generation probability pairs featuring $h(x) \ll 1$ for any $x \in X$ exist in

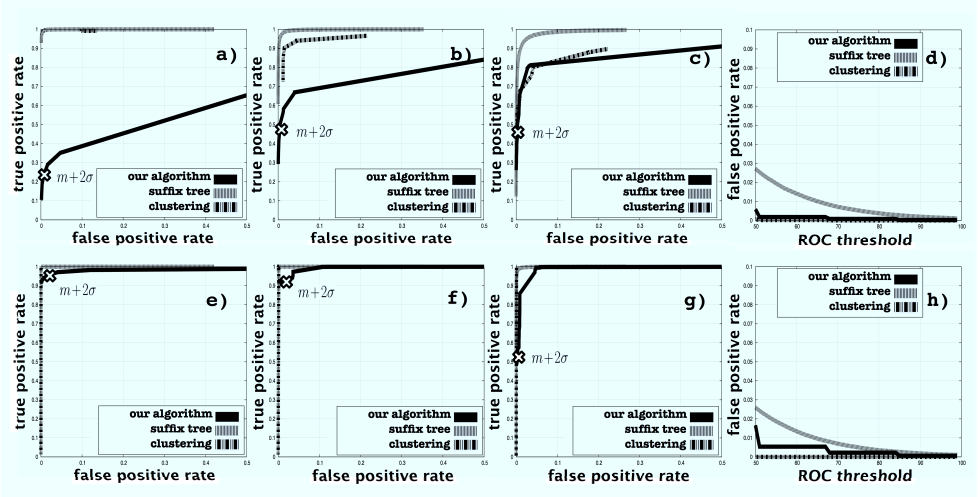

Figure 2.14: Algorithm supposing local concentration of anomalous blocks: Uniform distribution generation: a) 5% anomalous data b) 15% anomalous data c) 25% anomalous data d) Decrease of false positive rate; Exponential distribution generation: a) 5% anomalous data b) 15% anomalous data c) 25% anomalous data d) Decrease of false positive rate

case of uniform distribution generation, the algorithm allowing for arbitrary distribution of anomalous blocks is outperformed by the previous approaches. Contrary, the algorithm supposing local concentration of anomalous blocks yields results comparable to those of previous algorithms if l_{A} is suitably large, the higher share of anomalous data compensating for the lesser difference of generation probabilities. While the suffix tree algorithm delivers perfect results for all l_{rmA}, the cluster algorithm is slightly affected. In case of exponential distribution generation, all the algorithms produce very good results. An ROC threshold of $c_{\mathrm{th}} = m + 2\sigma$ (marked by an x) returns low false positive rates for both exponential and uniform distribution data.

2.4.4 Computer Security Data

One possible application of our algorithm is network masquerade attack detection. A masquerade attack consists of an attacker somehow stealing the password and login of a regular user. Because he is able to perform a regular login to the computer network, there will be no noticeable anomalies within the network traffic before or during the attack. Thus, the only possibility to immediately detect the attack is the analysis of the user input during the session. This input often solely consists of the command line data input by the user, which may be modeled as a time series of symbol sequences, one sequence representing a single session. Ever since the Schonlau et al. published their groundbreaking paper

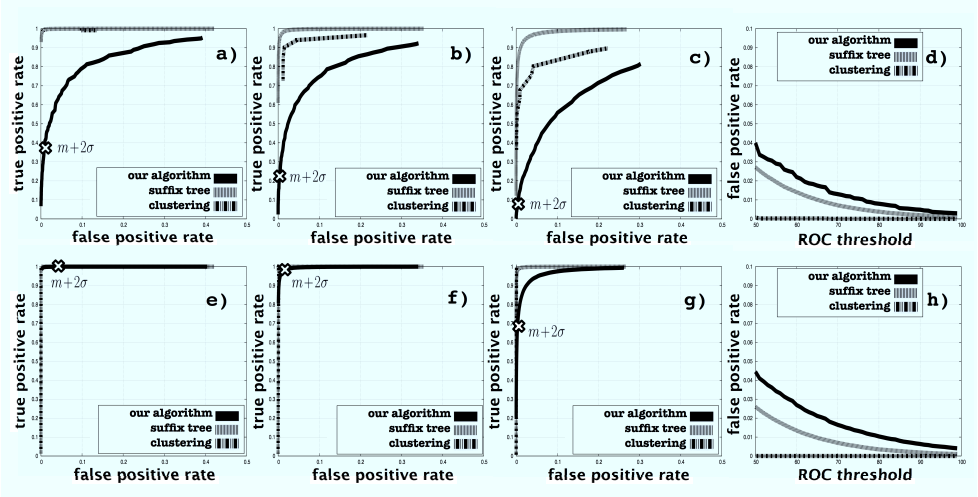

Figure 2.15: Algorithm allowing for arbitrary distribution of anomalous blocks: Uniform distribution generation: a) 5% anomalous data b) 15% anomalous data c) 25% anomalous data d) Decrease of false positive rate; Exponential distribution generation: a) 5% anomalous data b) 15% anomalous data c) 25% anomalous data d) Decrease of false positive rate

comparing the performance of several supervised statistical classifiers [62], a plethora of supervised anomaly detection approaches to the problem has been proposed [63]. The dataset created by them for their experiments quickly become the standard data set for evaluating new detection algorithms, and is commonly referred to as the SEA dataset.

The data was captured using the UNIX acct auditing mechanism. Any parameters and time stamps were removed, leaving a truncated command dataset, (i.e. a sequence of commands). Examples of commands are: sed, eqn, troff dpost, echo, sh, cat, netstat, tbl, sed, eqn, sh and so forth.

$15,000$ sequential commands of 70 users were originally recorded. Among those 70 users, 50 were randomly chosen as victims and the remaining 20 as intruders. The first $5,000$ commands for each victim do not contain any commands generated by masqueraders and are commonly used as classifier training data. The next $10,000$ commands can be thought of as 100 blocks of 100 commands each, and command data generated by the group of users used as masqueraders was randomly inserted for testing purposes. Note that 20 users feature no anomaly data at all($l_{\mathrm{A}} = 0$). The ROC curves shown below were created by applying the respective methods separately to the data of every user, finally calculating the average of true positive and false positive rate. Note that in contrast to the uniform distribution of generation probabilities in case of the artificial data above, the real world

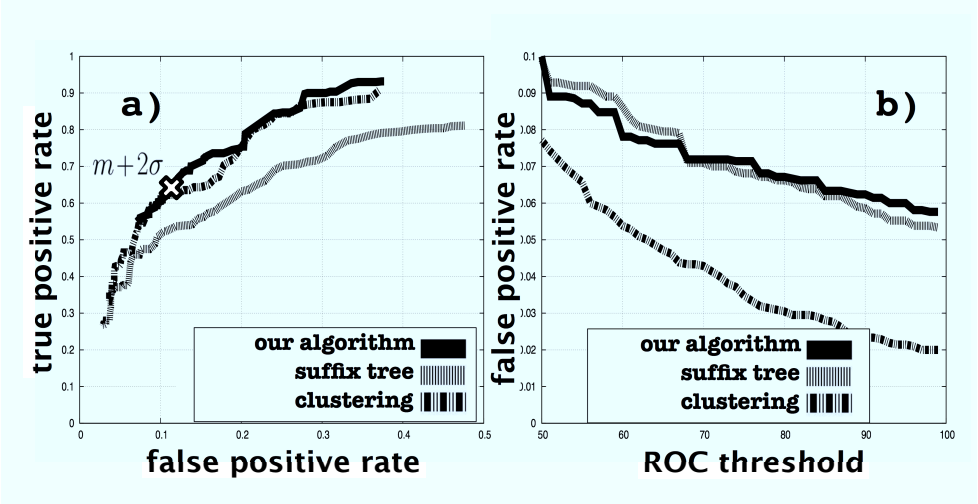

Figure 2.16: Real world masquerade data and $S_{\min} = 1$: a) Average ROC for algorithm supposing local concentration of anomalous blocks. b) Decrease of false positive rate

data features an exponential distribution of command probabilities. While a few commands are used very often, the majority features very low stationary generation probabilities.

Because our topic is unsupervised anomaly detection, we discarded the designated training data. The blocks are randomly rearranged in case of the algorithm allowing for arbitrary distribution of anomalous blocks.

The data set features $g = 10,000$ and $b = 100$. The parameters $M_{\max}$ and k are set using the depth t of the tree returned by the probabilistic suffix tree algorithm. In order to show the impact of $S_{\min}$ for data generated by a source with memory, we use the settings $S_{\min} = 1$ and $S_{\min} = Z$.

Setting of the Algorithm Supposing Local Concentration of Anomalous Blocks

Because the length of the anomaly l_{A} is supposed to be unknown, we choose a setting $\tau_{\mathrm{th}} = \frac{g}{6} = 1,666$.

Settings of the Algorithm Allowing for Arbitrary Distribution of Anomalous Blocks

The setting $\xi = 2$ leaves roughly 50% of the covered area in case of no anomaly untouched. The parameter β is set to 0.5. The threshold τ_{th} is set to $2,500$. The dimension of the vector calculated from the average index differences is set to $a = 3$. The minimum number of identical symbols within a block was set to $\nu_{\min} = 4$.

Settings of the Previous Approaches

The cluster algorithm uses a radius $w = 1.3$, again set using trial and error. For the suffix tree algorithm, $S_{\min} = 1$ and $S_{\min} = Z$ yields $t = 0 \rightarrow \ M_{\max} = 0,\ k = 1$ and $t = 1 \rightarrow \ M_{\max} = 1,\ k = 2$ respectively.

The ROC curves returned by the algorithms introduced for the settings $S_{\min} = 1$ and $S_{\min} = Z$, as well as the decrease of the respective false positive rates, are shown by Fig. 2.16 and Fig. 2.17. Although the probabilistic suffix tree algorithm provides suitable estimates for the parameter k of the clustering algorithm, the actual performance of the algorithm is inferior to both our algorithms and the cluster algorithm. The rather poor performance of the suffix tree algorithm may be explained by the fact that the combination of small g and large Z causes an overly rough estimation of the probabilities in case of exponential probability distribution. Opposite to the case of artificial data, the algorithm supposing local concentration of anomalous blocks features a high false positive rate, which hampers its performance despite high true positive rate. The reason for this behavior is found in the fact that within the masquerade data, local bursts of single commands may occur, causing a small share of blocks to display percentage values above the range given by m and σ, because the share of those blocks is just small enough not to affect the calculation of median and median absolute deviation. The algorithm allowing for arbitrary distribution of anomalous blocks, on the other hand, outperforms the suffix tree algorithm and equals the fixed-width clustering algorithm for both $S_{\min} = 1$ and $S_{\min} = Z$ within the most interesting range of false positive rates below ten percent. An ROC threshold of $m + 2\sigma$ returns a suitable combination of false positive/true positive rate. The main difference between$S_{\min} = 1$ and $S_{\min} = Z$ is a faster decrease of the false positive rate for the algorithm allowing for arbitrary distribution of anomalous blocks, and a significantly improved true positive rate of the clustering algorithm.

2.5 Concluding Remarks

In this section we presented two unsupervised anomaly detection algorithms for non-numerical sequence data based on the average index difference function. Besides a suitable detection performance comparable to those of previous approaches, both algorithms share the advantage of theoretically deducible parameter settings. While the first algorithm features low computational cost, it supposes the local concentration of anomalous blocks, a requirement not posed by the second algorithm.

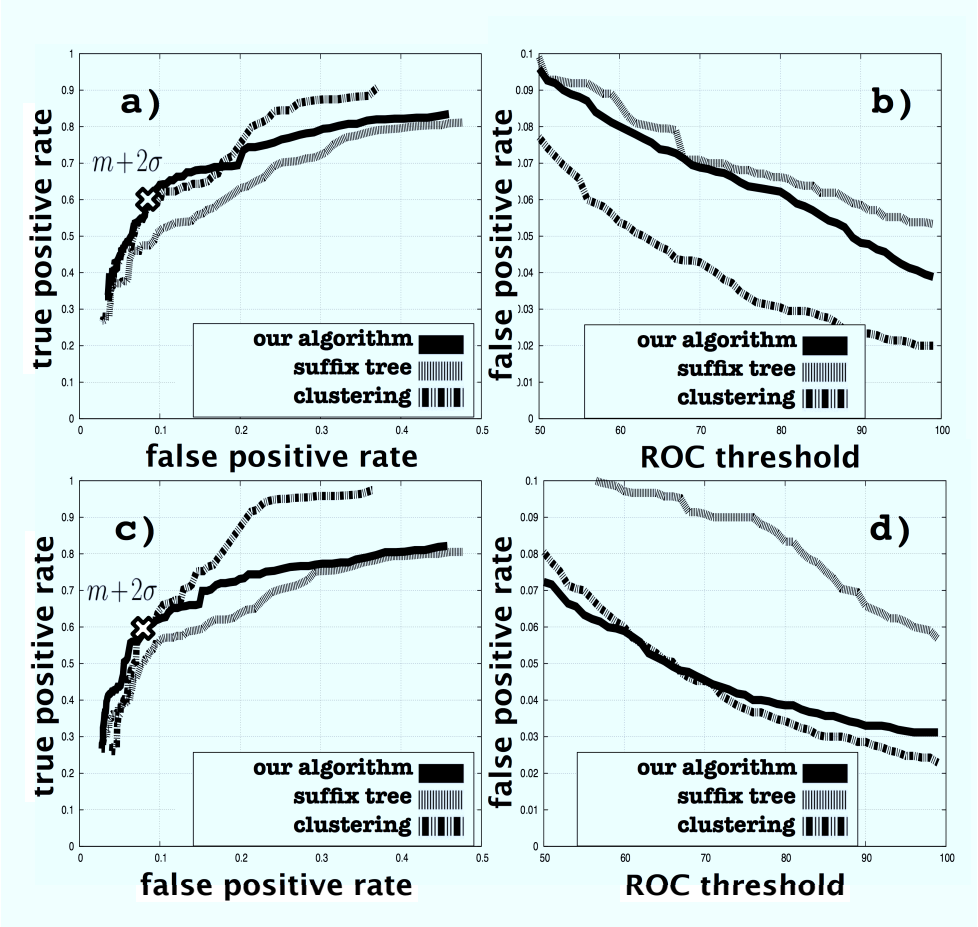

Figure 2.17: Real world masquerade data and $S_{\min} = 1$: a) Average ROC for algorithm allowing for arbitrary distribution of anomalous blocks b) Decrease of false positive rate ; Real world masquerade data and $S_{\min} = Z$: c) Average ROC for algorithm allowing for arbitrary distribution of anomalous blocks d) Decrease of false positive rate