# Coding of Non-Stationary Sources as a Foundation for Detecting Change Points and Outliers in Binary Time-Series

**Peter Sunehag** **Wen Shao** **Marcus Hutter**

Research School of Computer Science
Australian National University
ACT 0200 Australia
Email: {peter.sunehag, wen.shao, marcus.hutter}@anu.edu.au

**Abstract**

An interesting scheme for estimating and adapting distributions in real-time for non-stationary data has recently been the focus of study for several different tasks relating to time series and data mining, namely change point detection, outlier detection and online compression/ sequence prediction. An appealing feature is that unlike more sophisticated procedures, it is as fast as the related stationary procedures which are simply modified through discounting or windowing. The discount scheme makes older observations lose their influence on new predictions. The authors of this article recently used a discount scheme for introducing an adaptive version of the Context Tree Weighting compression algorithm. The mentioned change point and outlier detection methods rely on the changing compression ratio of an online compression algorithm. Here we are beginning to provide theoretical foundations for the use of these adaptive estimation procedures that have already shown practical promise.

*Keywords:* Non-stationary sources, time-series, compression, detection, change point, outlier

## 1 Introduction

Data mining in time series data is an active and vast area of research with many applications Fu (2011) relating to various tasks like change detection Guralnik and Srivastava (1999), Kawahara and Sugiyama (2012) and outlier detection Fawcett and Provost (1999), Zhang et al. (2009). A unifying framework for these two tasks were developed in Yamanishi and Takeuchi (2002), Takeuchi and Yamanishi (2006) based on online learning in non-stationary environments using probabilistic modeling which discounts experiences over time so as to focus on recent observations. Recently in Kawahara and Sugiyama (2012), this was further developed into a real-time change detection method based on sequential discounting normalized maximum likelihood coding that was applied to security applications, in particular malware detection. In the framework of Yamanishi and Takeuchi (2002), Takeuchi and Yamanishi (2006), Kawahara and Sugiyama (2012), a scoring function based on log loss, or in other words on arithmetic code length, was used to decide if recent observations were anomalous. If the average score over a number of consecutive time steps is sufficiently much higher than before, then a change has been detected. In compression terminology, the compressed size of those observations is higher than those before. This basic idea is also underlying the classical works E.S. (1955), Lorden (1971) on detecting change in a distribution.

Encoding a data source into a more compact representation is a long standing problem. In this paper, we are only concerned with the task of lossless data compression, which requires reproducing the exact original data from the compressed encoding. A number of different techniques for lossless data compression have been developed, for example Ziv and Lempel (1977, 1978), Cleary and Witten (1984), Cormack and Horspool (1987), Burrows and Wheeler (1994) to name a few. Many data compressors make use of a concept called arithmetic coding Rissanen (1976), Rissanen and Langdon (1979), which when provided with a probability distribution for the next symbol can be used for lossless compression of the data. In general, however, the true distribution for the next symbol is unknown and must be estimated. For stationary distributions, this estimation task is in many situations a solved problem and arithmetic coding based on the estimated distribution is optimal. For non-stationary distributions, estimating the true distribution is a much harder task. The Bayesian approaches Zacks (1983), Barry and Hartigan (1993) are attractive in that they are principled and automatically optimal but they are usually much more computationally expensive in their full form and, therefore, require approximation, in particular if they are going to run online R.P and D.J. (2007), Turner et al. (2009). If one is only interested in sequence prediction in the presence of change points and not in the change points themselves, the Bayesian approach Willems (1996) offers the possibility of using a mixture over all possible segmentations into piece-wise stationary intervals. Instead of using segmentation for sequence prediction, the framework by Yamanishi and Takeuchi (2002), Takeuchi and Yamanishi (2006), Kawahara and Sugiyama (2012) uses sequence prediction for segmentation.

Our interest here lies in methods that are as fast as their counterpart for the stationary case. Kawahara and Sugiyama (2012) achieves this using a simple discounting scheme and a similar technique is used by the authors of this article in O'Neill et al. (2012) to create a sequence prediction and compression algo-

This work was supported by ARC grant DP120100950.

rithm for non-stationary environments based on the Context Tree Weighting (CTW) algorithm Willems et al. (1995), which relies upon the Krichevsky-Trofimov (KT) estimator. In O'Neill et al. (2012), we introduce an adaptive version of the KT estimator and use this to define the adaptive CTW algorithm. In the case of non-stationary binary sequences, the algorithm of Kawahara and Sugiyama (2012) would also naturally be based on this estimator. By proving redundancy bounds for the adaptive KT estimator for interesting classes of environments, we automatically get a bound for adaptive CTW as well as a theoretical foundation for the empirically successful change detection algorithm from Kawahara and Sugiyama (2012). An alternative approach to discounting for dealing with non-stationarity is to use a moving window. The discounting version can be viewed as an approximation of this approach. The windowed KT and the resulting windowed CTW was studied in Kawabata and Rashid (2003) and redundancy bounds was proved for stationary ($d$:th order)Markov sources. We are instead first going to consider a source whose Bernoulli parameter moves within an interval that is small in the Kullback-Leibler sense and then consider drifting sources as well as sources when the parameters (or interval of parameters) can jump significantly but rarely.

**Related work.** Stationary sources have been extensively studied, Krichevsky and Trofimov (1981) provides a good survey. For example, Krichevsky (1968) provides an asymptotic lower bound for redundancy of block to variable universal code for Bernoulli sources; Krichevsky (1970) provides a corresponding upper bound. Trofimov (1974) provides a finite bound for stationary $d$:th order Markov sources which is also studied by Kawabata and Rashid (2003). Krichevsky (1998) looks at asymptotic one step redundancy bound for stationary Bernoulli sources.

## 2 Windowed Krichevsky-Trofimov Estimation for Non-Stationary Sources

The KT estimator Krichevsky and Trofimov (1981), in this article often referred to as the regular KT estimator, is obtained using a Bayesian approach by assuming a $(\frac{1}{2},\frac{1}{2})$-Beta prior on the parameter of a Bernoulli distribution. Let $y_{1:t}$ be a binary string containing $a$ zeros and $b$ ones. We write $P_{kt}(a,b)$ to denote $P_{kt}(y_{1:t})$. The KT estimator can be incrementally calculated by: $P_{kt}(a+1,b)=\frac{a+1/2}{a+b+1}P_{kt}(a,b)$ and $P_{kt}(a,b+1)=\frac{b+1/2}{a+b+1}P_{kt}(a,b)$ with $P_{kt}(0,0)=1$.

Allowing changes in the underlying sources suggests that 'outdated' histories do not necessarily provide useful and accurate information for predicting the next bit as it does in the stationary case. The regular KT estimator is very slow to update once many samples have been collected, so it cannot quickly adapt to a change in the source. Therefore, we will in this section look at a scheme where we estimate the probability of the next bit using the KT estimator, however, as opposed to counting the number of zeros and ones in the entire history, we only take the latest $n$ bits into account. We call this moving window KT or windowed KT.

**Redundancy bounds for windowed KT.** We are interested in one-step prediction. Assuming a stationary Bernoulli source $\theta$, an estimation for the probability of the next bit $x$ when given the latest $n$ bits, as a string $w$, yields a code length $-\ln\hat{p}(x|w)$. We then take an expectation over all possible $x$ and history $w$ to define the (expected) redundancy by

$$R_\theta(n)=\sum_{|w|=n}p_\theta(w)\sum_{x\in\mathbb{B}}p_\theta(x)(-\ln\hat{p}(x|w))-H(\theta)$$

where $H(\theta)$ is the entropy of source $\theta$. $p_\theta(w)$ and $p_\theta(x)$ are the probabilities of observing string $w$ and $x$ under $\theta$ respectively. $\hat{p}(x|w)$ is given by the KT-estimator

$$\hat{p}(x|w)=\frac{r_x(w)+1/2}{n+1} \tag{1}$$

where $r_x(w)$ is the number of $x$ that appears in $w$. For a non-stationary Bernoulli source, the one step redundancy is defined accordingly. Suppose $x_{1:m}$ is generated by a non-stationary Bernoulli process, with $x_i$ being sampled according to $\theta_i$, the one step redundancy $R_m(n)$ at step $m$ given a window size $n$ is

$$\sum_{|w|=n}p_{\theta_{m-n+1:m}}(w)\sum_{x\in\mathbb{B}}p_{\theta_{m+1}}(x)(-\ln\hat{p}(x|w))-H(\theta_{m+1}).$$

**Theorem 1.** *Suppose that a binary sequence is generated by a non-stationary Bernoulli process, with parameters $\theta_i$ where $\theta_i=\theta^1$ when $i\leq n$ and $\theta_i=\theta^2$ when $i>n$. We estimate the probability of the $(n+1)$:th letter by the KT-estimator. If $\theta^1,\theta^2\in(0,1)$, $\theta^1\leq\theta^2$, then*

$$R(n)\leq KL(\theta^2||\theta^1)+\frac{1+o(1)}{n}+\frac{\theta^2(3-4\theta^1)}{2n\theta^1}+\frac{(1-\theta^2)(4\theta^1-1)}{2n(1-\theta^1)}$$

The proof technique used is largely borrowed from Krichevsky (1998) where the following Lemma was proven.

**Lemma 2.** *Krichevsky (1998). Let*

$$b(n,k,\theta)=\binom{n}{k}\theta^k(1-\theta)^{n-k}.$$

*There is a constant $C$ such that the inequality*

$$\sum_{k=0}^{\lambda-\delta}b(n,k,\theta)<C\lambda e^{-(\delta^2/\lambda)+((\delta+1)^2/2(\lambda-\delta))}$$

*holds for $n>2$, $0<\theta<1$, $\lambda=np$, $1<\delta<\lambda$.*

*Proof for Theorem 1.* The one step redundancy we want to bound is $R(n)=$

$$\sum_{|w|=n}p_{\theta_{m-n+1:m}}(w)\sum_{x\in\mathbb{B}}p_{\theta_{m+1}}(x)(-\ln\hat{p}(x|w))-H(\theta_{m+1}) \tag{2}$$

where $\hat{p}(x|w))$ is given by the classic KT-estimator

$$\hat{p}(x|w)=\frac{r_x(w)+1/2}{n+1} \tag{3}$$

with $r_x(w)$ being the number of $x$ in string $w$. More specifically, the redundancy for the special case of this

Theorem, $R_{\theta^1,\theta^2}(n)$, can be rewritten as

$$\sum_{|w|=n} p_{\theta^1}(w) \sum_{x\in\mathcal{B}} p_{\theta^2}(x)(-\ln\hat{p}(x|w))$$

$$-H(\theta^2) \tag{4}$$

We can rewrite $H(\theta^2)$ as

$$H(\theta_R)=\ln n-\frac{1}{n}(\lambda_{R,1}\ln\lambda_{R,1}+\lambda_{R,0}\ln\lambda_{R,0}) \tag{5}$$

where $\lambda_{R,x}$ is the number of expected $x$ that appear in $n$, i.e. $\lambda_{R,x}=n\theta_R^x(1-\theta_R)^{(1-x)}$. Noticing that $-\ln\hat{p}(x|w)$ in equation (2) contains $\ln(n+1)$ while $H(\theta_R)$ contains an $\ln n$ term, we Taylor expand the function $\ln(n+1)$ at the origin and get that

$$\ln(n+1)<\ln n+\frac{1}{n}-\frac{1}{2n^2} \tag{6}$$

Plugging equation (3,5,6) into (4) yields

$$nR_{\theta^1,\theta^2}(n)\leq 1+\frac{1}{n}-\frac{1}{2n^2}$$
$$+\lambda_{R,1}\ln\lambda_{R,1}-\lambda_{R,1}\sum_{k=0}^{n} b(n,k,\theta^1)\ln(k+\frac{1}{2})$$

$$+\lambda_{R,0}(\ln\lambda_{R,0}-\sum_{k=0}^{n} b(n,k,1-\theta^1)\ln(k+\frac{1}{2})) \tag{7}$$

where $b(n,k,\theta)=\begin{pmatrix} n \\ k \end{pmatrix}\theta^k(1-\theta)^{n-k}$. Letting

$$F(n,\theta,\theta')=$$
$$\frac{1}{2}+\lambda\ln\lambda-\lambda\sum_{k=0}^{n} b(n,k,\theta')\ln(k+\frac{1}{2}) \tag{8}$$

where $\lambda=n\theta$, we can rewrite equation (7) as

$$nR_{\theta^1,\theta^2}(n)\leq\frac{1}{n}-\frac{1}{2n^2}+$$
$$F(n,\theta^2,\theta^1)+F(n,1-\theta^2,1-\theta^1) \tag{9}$$

and to bound this we are going to show that

$$F(n,\theta,\theta')\leq\frac{1}{2}+n\theta\ln\frac{\theta}{\theta'}+C''n$$

for some constant $C''$. Next we Taylor expand $\ln(k+\frac{1}{2})$ at $\lambda'=n\theta'$

$$\ln(k+\frac{1}{2})=\ln\lambda'+\frac{k+\frac{1}{2}-\lambda'}{\lambda'}+\mathcal{R}(k) \tag{10}$$

The remainder $\mathcal{R}(k)$ is

$$\mathcal{R}(k)=-\frac{(k+\frac{1}{2}-\lambda')^2}{2\xi(k)^2} \tag{11}$$

where $\xi(k)$ lies between $\lambda'$ and $k+\frac{1}{2}$. By plugging equation (10) into (8), we get

$$F(n,\theta,\theta')= \tag{12}$$
$$\frac{1}{2}+n\theta\ln\frac{\theta}{\theta'}-\frac{\theta}{2\theta'}-\lambda\sum_{k=0}^{n} b(n,k,\theta')\mathcal{R}(k) \tag{13}$$

Take $n>\frac{1}{\theta'}+1$ and choose a natural number $\delta$ with $1<\delta<\lambda'$. We split the summation in the last term into two parts: $0\leq k\leq\lambda-\delta$ and $\lambda-\delta<k\leq n$. To bound the first part, we use that $\xi(k)>\frac{1}{2}$. Putting $\delta=\lambda^{3/4}$ and using the previous lemma it follows that $\mathcal{R}(k)\geq-2(k-\lambda+\frac{1}{2})^2$ and therefore

$$-\lambda\sum_{k=0}^{\lambda-\delta} b(n,k,\theta')\mathcal{R}(k)\leq$$
$$2\lambda(\lambda'+\frac{1}{2})^2\sum_{k=0}^{\lambda-\delta} b(n,k,\theta')<C'\lambda(\lambda'+\frac{1}{2})^2e^{-\sqrt{\lambda'}}$$

for some constant $C'$. To deal with the second part, we choose $n$ large enough such that $\xi(k)>\frac{\lambda'}{2}$ and then we have

$$\mathcal{R}(k)\geq-\frac{2(k-\lambda'+\frac{1}{2})^2}{\lambda'^2}$$

Therefore,

$$-\lambda\sum_{k=\lambda-\delta}^{n} b(n,k,\theta')\mathcal{R}(k)\leq$$

$$\frac{2\lambda}{\lambda'^2}\sum_{k=0}^{n} b(n,k,\theta')(k-\lambda'+\frac{1}{2})^2$$

Using the second central moment of binomial distribution $m_2=\lambda'(1-\theta')$ together with the first central moments, we have

$$-\lambda\sum_{k=\lambda-\delta+1}^{n} b(n,k,\theta')\mathcal{R}(k)\leq\frac{2\theta(1-\theta')}{\theta'}+\frac{\theta}{n\theta'^2}$$

Thus, we conclude that for large enough $n$

$$F(n,\theta,\theta')\leq\frac{1}{2}+n\theta\ln\frac{\theta}{\theta'} \tag{14}$$
$$+\frac{\theta(3-4\theta')}{2\theta'}+C'\lambda(\lambda'+\frac{1}{2})^2e^{-\sqrt{\lambda'}} \tag{15}$$

The last term decrease exponentially as $n\to\infty$ and can be replaced by $o(1)$, and write

$$F(n,\theta,\theta')<\frac{1}{2}+n\theta\ln\frac{\theta}{\theta'}+\frac{\theta(3-4\theta')}{2\theta'}+o(1)$$

Therefore, through Equation 9 we have

$$R_{\theta^1,\theta^2}(n)\leq KL(\theta^2||\theta^1)+\frac{1+o(1)}{n}$$

$$+\frac{\theta^2(3-4\theta^1)}{2n\theta^1}+\frac{(1-\theta^2)(4\theta^1-1)}{2n(1-\theta^1)}$$

□

If we allow the parameters to move within an in-

terval $[\theta_L,\theta_R]$, then Theorem 1 above deals with the worst case situation, namely when $\theta_i$ is at one end point for $m$ steps and then jumps to the other.

**Corollary 3.** *Suppose $x_{1:m}$ is generated by a non-stationary Bernoulli process, with $x_i$ being sampled according to $\theta_i$. We estimate the probability of the $i$:th letter by the KT-estimator with a moving window of size $n<m$. If $i$ is such that $m-n+1\leq i\leq m+1$, $\theta_i\in[\theta_L,\theta_R]$,$\theta_L,\theta_R\in(0,1)$ and $\theta_L\leq\theta_R$, then the redundancy for this prediction is bounded by*

$$\begin{aligned}&R(n)\leq\frac{1+o(1)}{n}+\\&\max\{KL(\theta_L||\theta_R),KL(\theta_R||\theta_L)\}\\&+\frac{1}{n}\max\{\frac{\theta_R(3-4\theta_L)}{2\theta_L}+\frac{(1-\theta_R)(4\theta_L-1)}{2(1-\theta_L)},\\&\frac{\theta_L(3-4\theta_R)}{2\theta_R}+\frac{(1-\theta_L)(4\theta_R-1)}{2(1-\theta_R)}\}\end{aligned}$$

**Example 4.** *In the above bounds we notice that the constant factor in the $O(1/n)$ term grows unboundedly when the parameters tend to 0 or 1. This is not just a problem with the bounds but a genuine phenomenon. Suppose that $\theta_i=1$ for $n$ time steps and then switch to $\theta<1$. The redundancy for the next time step is then $O(log(1+n))$. We conclude that if we want a uniform constant for the $O(1/n)$ term we need to assume that we are a minimum distance away from the end points.*

**Corollary 5.** *Suppose $x_{1:m}$ is generated by a non-stationary Bernoulli process, with $x_i$ being sampled according to $\theta_i\in[L,R]$ where $0<L\leq R<1$. We estimate the probability of the $i$:th letter by the KT-estimator with a moving window of size $n<m$. If $i$ is such that $m-n+1\leq i\leq m+1$, $\theta_i\in[\theta_L,\theta_R]$,$\theta_L,\theta_R\in[L,R]$ and $\theta_L\leq\theta_R$, then, the redundancy for this prediction is bounded by*

$$R(n)\leq\max\{KL(\theta_L||\theta_R),KL(\theta_R||\theta_L)\}+C/n$$

*where $C$ does depend on $L$ and $R$ but not on $\theta_L$ or $\theta_R$.*

**Remark 6.** *For the case when $\theta_L=\theta_R$ we do not have a problem at the end points. Consider $\theta_i=1\ \forall i$ which means that we will almost surely have a constant sequence. Then the redundancy is $-\log\frac{1/2+n}{n+1}=\log(1+\frac{1}{2(n+1)})\leq\frac{1}{2(n+1)}$. Corollary 5 holds for $L=0$ and $R=1$ as long as $\theta_R=\theta_L$.*

**Geometrically drifting sources.** Suppose $x_{1:m}$ is generated by a non-stationary Bernoulli process, with $x_i$ being sampled according to $\theta_i$. If the source is such that for all $i$, $KL(\theta_{max(i,n)},\theta_{min(i,n)})\leq g(n)$, where

$$\theta_{min(i,n)}=\min_{i\leq j\leq i+n}\{\theta_j\}$$

$$\theta_{max(i,n)}=\max_{i\leq j\leq i+n}\{\theta_j\}$$

we can for any fixed $i$, apply Theorem 1. We next define a class of drifting sources for which there is a simple function $g$ of this sort.

**Definition 7** (Geometrically drifting source)**.** *Suppose a sequence $\{x_i\}_{i=1}^{\infty}$ is generated by a non-stationary Bernoulli process, identified by $\{\theta_i\}_{i=1}^{\infty}$ (with $\theta_1\in(0,1)$) with each $x_i$ sampled according to $\theta_i$. We say that the source is geometrically drifting if and only if $1\leq\max\{\frac{\theta_i}{\theta_{i+1}},\frac{1-\theta_i}{1-\theta_{i+1}}\}\leq c$ for all $i$ and some constant $c\geq1$.*

The idea behind this definition is that the source can only drift, i.e. increase or decrease by a certain ratio $c$. This notion of drift allows us to bound the KL divergence of the maximum and minimum $\theta$ during $n$ consecutive steps.

$$KL(\theta_i||\theta_{i+1})=\theta_i\ln\frac{\theta_i}{\theta_{i+1}}+(1-\theta_i)\ln\frac{1-\theta_i}{1-\theta_{i+1}}<\ln c$$

for all $i$ and it holds that

$$\begin{cases}1\leq\frac{\theta_{max}}{\theta_{min}}\leq c^n\\1\leq\frac{1-\theta_{min}}{1-\theta_{max}}\leq c^n\end{cases}$$

which results in a bound for $KL(\theta_{max}||\theta_{min})$ (and the same for $KL(\theta_{min}||\theta_{max})$), namely

$$\begin{aligned}&KL(\theta_{max}||\theta_{min})=\\&\theta_{max}\ln\frac{\theta_{max}}{\theta_{min}}+(1-\theta_{max})\ln\frac{1-\theta_{max}}{1-\theta_{min}}\\&\leq n\ln c\end{aligned}$$

## 3 Discounted Estimation

When dealing with non-stationary sources, it is natural that one wants to weight recent history higher. We define an adaptive KT estimator, which we call discounting KT based on replacing $a_n$ and $b_n$ in the definition of the KT estimator with discounted counts. These counts are defined by applying the following discounting operation after adding a new zero ($a_n=a_n+1$) or a new one ($b_n=b_n+1$),

$$a_{n+1}:=(1-\gamma)\,a_n\qquad b_{n+1}:=(1-\gamma)\,b_n$$

where $\gamma\in[0,1)$ denotes the discount rate. For discounting KT with $\gamma>0$, we have an effective horizon of length $\frac{1}{1-\gamma}$. The windowed estimator from the previous section can be viewed as a hard version of this scheme.

Consider a situation where we have a stationary source ($\theta_i=\theta\ \forall i$) where we use a windowed KT with window length $n=\frac{1}{1-\gamma}$. Compare the distribution for the coefficient $a$ (the number of zeroes in the window) with the distribution for the $a$ coefficient defined from discounting from an infinite history. Both distributions are symmetric around the same mean but the one arising from the discounting has more mass close to the mean. Hence the discounting method will have a lower redundancy. This is not surprising in this situation because the discounting estimate gets to use an infinite history of observations and if we use the full history KT we have zero redundancy. This is, however, not the situation that we want to use discounting KT in. Discounting KT effectively only depends on a small number of observations. The reason we let it depend at all on things further back is for convenience, it yields a very simple update formula where nothing has to be stored. This is very convenient when, as in the CTW algorithm, a KT estimator is created for every node in a tree, which might be deep. The conclusion is that the upper bounds for the redundancy of windowed KT should at least approximately also hold for discounting KT. Furthermore, when the

source has been close to stationary for longer then the window length, one should expect marginally better from the discounting algorithm. Another case when one expect better from the discounting algorithm is for slowly drifting sources. We will below provide a class of drifting sources that is such that for any window, we are going to satisfy the assumption of Corollary 5 and one can conclude a redundancy bound for moving window which does tell us what we should at least expect from discounting KT.

## 4 Implications for Compression and Detection of Change Points and Outliers

We have showed that if the parameters stay within a small interval and we have a large enough (though not too large) window, the expected redundancy for windowed KT is small. This also applies if instead the parameter drift is small. We argue that this is not only true for windowed KT with a suitable window length (for the amount of drift) but for discounting KT with an appropriate discount factor. In this section we discuss the implications for the motivating applications.

**Compression.** Since expectation is a linear operation, the total redundancy is the sum of the per step redundancies. In the case of a stationary source, a source constrained to a small interval or a slowly drifting source within a larger interval our per bit redundancy bounds are simply multiplied by the file length to get the total redundancy. When we have a source with a small number of jumps and otherwise only slow (geometric) drift, Theorem 1 tells us that we have to add the sum of the KL-divergences times the window length for the jumps to the estimate. Note that with a window length $n$, the worst code length for a step is less than $\log(1+n)$. If we have $h(N)$ jumps during the first $N$ time steps, the total accumulated redundancy is less than $h(N)\log(n+1)+O(N/n)$. Hence, if jumps are rare, they will not affect the total compressed length significantly. Regular KT has unbounded one step redundancy and the regular KT estimator also continues to be affected by all old data as much as newer observations. Hence, if there is substantial change in the middle of the file, the first part will adversely affect the rest of the file. In O'Neill et al. (2012), an empirical advantage of adaptive CTW over regular CTW was demonstrated on files, which were created by concatenating two different shorter files.

Here we have so far only proved bounds for the adaptive KT estimators and not the CTW algorithm which is more relevant for practical compressions since it takes context into account. However, bounds for the adaptive KT estimator is all that is needed to prove bounds for adaptive CTW. This can be done easily because there are three parts contributing to the redundancy bounds for CTW Willems et al. (1995): (1) redundancy used to find the 'right' tree (2) parameter redundancy given the 'right' tree (3) arithmetic coding redundancy (always bounded by 2). The first term has nothing to do with the underlying estimator but with the code length of the 'right' tree. The problem is thus reduced to the second term (the main term), which is the redundancy of the underlying KT estimator. For regular KT in the stationary case, this is $O(1/m)$ for the $m$:th time step. This adds up to a logarithmic term for the first $N$ time steps and this is the only non-constant term. The $O(1/n)$ terms is for windowed KT replaced by an $O(1/n)$ term (Theorem 1) in the stationary case (where $n$ is the size of the window) which accumulates to $O(N/n)$ if $N$ is the total length of the file. In the interval case the term is replaced by $KL+O(1/n)$, again according to Theorem 1. In these cases the total redundancy of adaptive CTW is $O(N/n)$. If we have $h(N)$ jump point of KL-divergence at most $D$, the redundancy is $h(N)D+O(N/n)$.

**Detecting change points and outliers.** The algorithm for detecting change points and outliers in Yamanishi and Takeuchi (2002), Takeuchi and Yamanishi (2006), Kawahara and Sugiyama (2012) relies on $\log Pr(x_{t+1} \mid x_{1:t})$ as a score. The idea is that this score is large in expectation if the distributions have changed significantly and smaller if it has not. They average over a number of time steps to get a good estimate of this expectation. That it will be large if the distribution change by much is clear but for it to be a well founded method one also needs to be able to say that it will be small if the distribution has at most changed a little. This is what we provide a theory for in this article.

## 5 Conclusion

Some recent advances in real-time online change point detection, outlier detection and compression have relied on discounting when estimating parameters of a distribution that is changing over time. A closely related alternative for dealing with non-stationarity in a computationally efficient manner is to only use the last few observations for the estimation. In this article we have provided a theoretical analysis of how these estimators behave for important classes of non-stationary environments and outlined the implication of the results for the application of the mentioned algorithms.

## References


Barry, D. & Hartigan, J. A. (1993), 'A bayesian analysis for change point problems', *Journal of the American Statistical Association* **88**(421), 309–319.

Burrows, M. & Wheeler, D. (1994), 'A block-sorting lossless data compression algorithm', *Digital SRC Research Report* .

Cleary, J. G. & Witten, I. H. (1984), 'Data compression using adaptive coding and partial string matching', *IEEE Transactions on Communications* **32**, 396–402.

Cormack, G. V. & Horspool, R. N. S. (1987), 'Data compression using dynamic Markov modelling', *The Computer Journal* **30**(6), 541–550.

E.S., P. (1955), 'A test for change in a parameter occurring at an unknown point', *Biometrika* **42**.

Fawcett, T. & Provost, F. (1999), Activity monitoring: noticing interesting changes in behavior, *in* 'Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining', KDD '99, ACM, New York, NY, USA, pp. 53–62.

Fu, T. (2011), 'A review on time series data mining', *Eng. Appl. Artif. Intell.* **24**(1), 164–181.

Guralnik, V. & Srivastava, J. (1999), Event detection from time series data, *in* 'KDD', pp. 33–42.

Kawabata, T. & Rashid, M. M. (2003), 'A zero redundancy estimator for the context tree weighting method with a finite window', *IEEE International Symposium on Information Theory* p. 114.

Kawahara, Y. & Sugiyama, M. (2012), 'Sequential change-point detection based on direct density-ratio estimation', *Stat. Anal. Data Min.* **5**(2), 114–127.

Krichevsky, R. E. (1968), 'The connection between the redundancy and reliability of information about the sources', *Problems of Information Transmission* **4**(3), 48–57.

Krichevsky, R. E. (1970), Lectures on information theory, Technical report, Novosibirsk State University.

Krichevsky, R. E. (1998), 'Laplace's law of succession and universal encoding', *IEEE Transactions on Information Theory* **44**, 296–303.

Krichevsky, R. E. & Trofimov, V. K. (1981), 'The performance of universal encoding', *IEEE Transactions on Information Theory* **27**(2), 199–207.

Lorden, G. (1971), 'Procedures for reacting to a change in distribution', *Annals of Mathematical Statistics* **42(6)**, 1897–1908.

O'Neill, A., Hutter, M., Shao, W. & Sunehag, P. (2012), Adaptive context tree weighting, *in* '2012 Data Compression Conference, Snowbird, UT, USA, April 10-12, 2012', IEEE Computer Society, pp. 317–326.

Rissanen, J. J. (1976), 'Generalized Kraft inequality and arithmetic coding', *IBM Journal of Research and Devlopment* **20**(3), 198–203.

Rissanen, J. J. & Langdon, G. G. (1979), 'Arithmetic coding', *IBM Journal of Research and Devlopment* **23**, 149–162.

R.P, A. & D.J., M. (2007), 'Bayesian online changepoint detection'.

Takeuchi, J. & Yamanishi, K. (2006), 'A unifying framework for detecting outliers and change points from time series', *IEEE Trans. Knowl. Data Eng.* **18**(4), 482–492.

Trofimov, V. K. (1974), 'The redundancy of markov source encoding', *Problems of Information Transmission* **10**(4), 16–24.

Turner, R., Saatci, Y. & Rasmussen, C. E. (2009), Adaptive sequential Bayesian change point detection, *in* 'Advances in Neural Information Processing Systems (NIPS): Temporal Segmentation Workshop'.

Willems, F. M. J. (1996), 'Coding for a binary independent piecewise-identically-distributed source', *IEEE Transactions on Information Theory* **42**(6), 2210–2217.

Willems, F. M. J., Shtarkov, Y. M. & Tjalkens, T. (1995), 'The context tree weighting method: Basic properties', *IEEE Transactions on Information Theory* **41**, 653–664.

Yamanishi, K. & Takeuchi, J. (2002), A unifying framework for detecting outliers and change points from non-stationary time series data, *in* 'Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23-26, 2002, Edmonton, Alberta, Canada', ACM, pp. 676–681.

Zacks, S. (1983), *Survey of classical and Bayesian approaches to the change-point problem: Fixed sample and sequential procedures in testing and estimation*, Academic Press.

Zhang, K., Hutter, M. & Jin, W. (2009), A new local distance-based outlier detection approach for scattered real-world data, *in* 'Proc. 13th Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD'09)', Vol. 5467 of *LNAI*, Springer, Bangkok, pp. 813–822.

Ziv, J. & Lempel, A. (1977), 'A universal algorithm for sequential data compression', *IEEE Transactions on Information Theory* **23**, 337–342.

Ziv, J. & Lempel, A. (1978), 'Compression of individual sequences via variable-rate coding', *IEEE Transactions on Information Theory* **24**, 530–536.