Symbolic Aggregate ApproXimation (SAX) under Interval Uncertainty

Chrysostomos D. Stylios
Laboratory of Knowledge and Intelligent Computing
Department of Computer Engineering
Technological Educational Institute of Epirus
47100 Kostakioi, Arta, Greece
E-mail: stylios@teiep.gr

Vladik Kreinovich
Department of Computer Science
University of Texas at El Paso
500 W. University
El Paso, TX 79968, USA
E-mail: vladik@utep.edu

***Abstract*—In many practical situations, we monitor a system by continuously measuring the corresponding quantities, to make sure that any abnormal deviation is detected as early as possible. Often, we do not have readily available algorithms to detect abnormality, so we need to use machine learning techniques. For these techniques to be efficient, we first need to compress the data. One of the most successful methods of data compression is the technique of Symbolic Aggregate approXimation (SAX); see, e.g., [10]. While this technique is motivated by measurement uncertainty, it does not explicitly take this uncertainty into account. In this paper, we show that we can further improve upon this techniques if we explicitly take measurement uncertainty into account.**

I. Formulation of the Problem

Need for diagnostics. In many practical situations, we are monitoring a certain process for possible problems:

- We may be monitoring a mechanical system to check if there are mechanical problems that require correction: e.g., whether the observed vibrations indicate some abnormality. Detecting such problem is known as *diagnostics*.
- We may be monitoring the vital signs of a patient to see if an urgent medical intervention is needed.

Need for machine learning. In some cases, we have an algorithm that, based on the observed time series, tells us whether the intervention is necessary – and what kind of intervention is needed. However, such situations are rare. In most practical applications – especially in medicine – no such algorithm is readily available.

What we have instead is numerous past data series corresponding both to the cases when situation turned out to be normal and situations in which further developments revealed abnormality. We thus need to extract such an algorithm from all these examples, i.e., use one of the *machine learning* algorithms; see, e.g., [1].

Need for data compression. Most machine learning algorithms work well if we have up to dozens of inputs. However, as a result of monitoring, we get hundreds and thousands of values $x(t)$ corresponding to different moments of time.

So, to efficiently apply machine learning algorithms, we first need to compress the input data; see, e.g., [1].

Symbolic Aggregate approXimation (SAX): main idea. The main reason why we have a large number of inputs is that we keep track of the values $x(t)$ for many different moments of time. The main objective of monitoring is to catch deviations from the normal regimes as early as possible. As a result, monitoring is performed at a high rate, so that we will be able to catch a deviation while this deviation is small, way before this deviation may lead to catastrophic consequences. Thus, when the monitoring is arranged properly, with the time between two consequent measurements small enough to capture even small changes, values change very little from one moment to the next.

Hence, during the periods of normal functioning, the function is almost constant for a significant number of values, i.e., we have a sequence of values $x(t)$ which are practically equal to each other. So, instead of keeping all these almost-equal values, we can simply keep an average and a duration of this almost-constant period. In mathematical terms, we replace the original function $x(t)$ with a piece-wise constant approximation.

By definition, a piece-wise function attains the same value for several consequent moments of time from a certain time interval. It is therefore not necessary to store its value at each of these moments of time, it is sufficient to store only the corresponding time interval and the value of the function in this interval. This representation indeed leads to a drastic reduction in data size; see, e.g., [3].

Often, the resulting sequence of time intervals and corresponding values x still requires too many bits, so it needs to be compressed further. Such further compression is definitely possible since each average value is a computer-represented real number, and such numbers require dozens of bits to store, corresponding to potential accuracy of up to ten decimal digits. In some computations, we need these many digits, but for usual monitoring measurements, the accuracy usually ranges from 1% to 10%, so two decimal digits is more than enough. Symbolic Aggregate approXimation (SAX) is a technique for such a reduction.

In this technique, in the interval $[\underline{x}, \overline{x}]$ of possible values of $x(t)$, we select thresholds $x_0 = \underline{x}, x_1, x_2, \ldots, x_m$. These

978-1-4673-7248-0/15/$31.00 ©2015 IEEE

thresholds divide the interval $[\underline{x}, \overline{x}]$ into $m+1$ subintervals $[x_i, x_{i+1}]$, $i = 0, 1, \ldots, m$. Then, for each moment of time t, instead of keeping the original value $x(t)$, we simply keep the index i of the subinterval that contains this value [10]. At present, SAX is the most efficient data compression technique.

SAX: details and successes. To maximize the amount of information after compression, the current implementation of SAX methods take into account that the maximum amount of Shannon's information $-\sum\limits_{i=0}^{m} p_i \cdot \log_2(p_i)$, where p_i is the frequency with which we get the i-th subinterval, is attained when all the probabilities p_i are equal to each other – and is, thus, equal to $p_i = \dfrac{1}{m+1}$; see, e.g., [7]. Thus, the thresholds x_i are selected in such a way that for each i, the proportion of moments of time for which $x(t)$ is between x_i and x_{i+1} is equal to $\dfrac{1}{m+1}$.

SAX techniques led to many practical applications ranging from engineering [5], [6], [9] to medicine [4].

Problem. The problem is that while measurement errors were a motivation for SAX techniques, the actual implementation does not take measurement errors into account. As a result, for an almost-constant signal $x(t)$, when most of the observed values are concentrated on a very small interval, we may get thresholds x_i and x_{i+1} which are much closer to each other than the measurement accuracy: e.g., differ by 5% while the measurement accuracy is 10%. In this case, because of the measurement inaccuracy, we cannot really tell whether the actual value $x(t)$ was in the i-th interval or in the next interval.

This problem was noticed, e.g., in [2]. It is therefore desirable to explicitly take measurement uncertainty into account when applying SAX techniques.

What we do in this paper. In this paper, we describe how to take measurement uncertainty into account when selecting the thresholds x_i.

Structure of the paper. In Section 2, we recall the motivations for the usual SAX selection of thresholds – and for similar threshold selection techniques. In Section 3, we show how the corresponding optimization problems can be modified to take into account measurement uncertainty, and we show how to solve the corresponding optimization problems.

II. How to Optimize Threshold Selection: Case When Measurement Inaccuracy Can Be Ignored (Reminder)

Towards a formal description of the problem. In this section, we describe how we can select optimal thresholds in the usual setting, when the measurement inaccuracy is ignored.

To formulate the corresponding optimization problem, we need to describe:

- what is known,
- what we want, and
- how we decide which threshold selection is better.

Let us answer these questions one by one.

What we know. We have a large number of observed values $x(t)$ corresponding to different moments of time t. Based on these values, we can find the frequencies (probabilities) with which different values of x occur. These probabilities can be naturally described by a probability density function $\rho(x)$.

Here, the probabilities should add up to one, i.e., we should have $\int \rho(x)\,dx = 1$.

Comment. In many practical situations, the observed signal is a joint effect of many different independent processes. In such situations, the Central Limit Theorem implies that the resulting distribution should be close to Gaussian; see, e.g., [12]. An indeed, in many practical situations, the empirical distribution is close to Gaussian, with appropriate mean μ and standard deviation σ; see, e.g., [4], [5], [6], [9].

What we want. We want to select the appropriate thresholds x_1, x_2, ... Since there are many thresholds, a reasonable way to describe them is to describe the *distribution* of thresholds, i.e., to describe, for every value x, the density $\rho_t(x)$ of the thresholds – how many thresholds we have per unit length.

Here, the overall number of selected thresholds is m, so we should have $\int \rho_t(x)\,dx = m$.

How to decide which threshold selection is better: first idea. After the data compression, the only information that we have about each value $x(t)$ in the index i of the subinterval that contains this value. So, if we want to reconstruct the value $x(t)$ based on this information, the best we can do is to select a midpoint $\widetilde{x}(t)$ of this subinterval. This reconstruction is approximate, there is an approximation error $\varepsilon(t) \stackrel{\text{def}}{=} \widetilde{x}(t) - x(t) \neq 0$.

Ideally, we would like to have all these errors to be as close to 0 as possible. In other words, we would like to have the vector $\varepsilon = (\varepsilon(t_1), \varepsilon(t_2), \ldots)$ consisting of all these errors to be as close to the zero vector $\vec{0} = (0, 0, \ldots)$ as possible.

It is natural to use the usual Euclidean distance between these two vectors to estimate how close they are. In this Euclidean case, the distance has the form

$$d(\varepsilon, \vec{0}) = \sqrt{\sum_k (\varepsilon(t_k))^2}.$$

Since the square root is a monotonic function, minimizing this sum is equivalent to minimizing the sum of the squares $\sum\limits_k (\varepsilon(t_k))^2$. In the continuous approximation, this is equivalent to minimizing the corresponding integral $\int (\varepsilon(t))^2\,dt$.

Limitations of the first idea. The above least-squares approach is indeed ubiquitous in data processing, but it has a known problem: it is very vulnerable to outliers. For example, if we simply estimate a constant value a based on several repeated measurements $a_1, \ldots, a_n$, then the least squares methods means minimizing the sum $\sum\limits_{i=1}^{n} (a_i - a)^2$. If we

differentiate this expression with respect to a and equate the resulting derivative to 0, we get the usual arithmetic average

$$a = \frac{a_1 + \ldots + a_n}{n}.$$

This estimate works well if all the measured values a_i are close to a, but sometimes, we get an outlier, a value that – due, e.g., to some malfunction, is drastically different from a. For example, if the actual value of the measured quantity is 10, and we have nine very exact measurements $a_i = 10$ and one outlier $a_i = 1000$, then the arithmetic average formula leads to

$$\frac{10 + \ldots + 10 \ (9 \text{ times}) + 1000}{10} = \frac{1090}{10} = 109 \gg 10.$$

How to decide which threshold selection is better: second idea. An alternative idea is to use estimates that try to avoid the above sensitivity. Such estimates are known as *robust*; see, e.g., [8]. Among the most widely used robust estimates are ℓ^p-estimates, when instead of minimizing the sum of the squares, we minimize the sum of the p-th degrees, for some p from the interval $[1, 2)$. In other words, we minimize the sum $\sum\limits_k |\varepsilon(t_k)|^p$ or the corresponding integral $\int |\varepsilon(t)|^p \, dt$.

These methods are indeed more robust: for example, for $p = 1$, minimizing the sum $\sum\limits_{i=1}^{n} |a_i - a|$ leads to the median, and the median is clearly much more robust than the arithmetic mean – for example, in the above 10s and 1000 example, the outlier does not affect the median at all.

Third idea: decreasing the number of bits. Since our objective is data compression, another natural idea is to decrease (and, ideally, minimize) the number of bits needed to describe all the thresholds. The number of bits that we need to transmit each threshold depends on how close it is to the next threshold.

For example, if the two neighboring thresholds differ by 0.1, then it does not make much sense to describe the second decimal digit of the corresponding interval: using a subinterval $[1.201, 1.302]$ leads, in effect, to the same results as the subinterval $[1.2, 1.3]$, the difference is miniscule.

Indeed:

- the values $x(t)$ for moment t which are close to the threshold t_i and smaller than t_i are replaced by the "average" value of the signal on the previous time interval, while
- the values $x(t)$ corresponding to the moment t which are close to the threshold t_i and for which $t > t_i$ are replaced by the average value of the signal in the next time interval.

For a smooth function $x(t)$, the average value on an interval is approximately equal to its value at the midpoint of this interval. Thus:

- for $t < t_i$, the inaccuracy of the resulting approximation is approximately equal to the difference $|x(t) - x(t_-)|$, where $t_- = \dfrac{x_{i-1} + x_i}{2}$ is the midpoint of the previous time interval, while
- for $t > t_i$, the inaccuracy of the resulting approximation is approximately equal to the difference $|x(t) - x(t_+)|$, where $t_+ = \dfrac{x_i + x_{i+1}}{2}$ is the midpoint of the following time interval.

For smooth signals, $|x(t) - x(t_-)| \approx |x'(t_i)| \cdot (t - t_-)$ and $|x(t) - x(t_+)| \approx |x'(t_i)| \cdot (t_+ - t)$.

For the threshold point itself, we have $t_i - t_- \approx t_+ - t$, so these two inaccuracies are approximately the same. If the increase the threshold a little bit, this means that for this threshold and for a few neighboring points $t = t_i + \delta t$ (with $\delta t \ll t_i - t_-$) we replace the inaccuracy proportional to $t_+ - t = t_+ - t_i - \delta t$ with the inaccuracy proportional to $t - t_- = t + \delta t - t_-$. The relative increase in inaccuracy can thus be estimated as the ratio

$$\frac{(t + \delta t - t_-) - (t_+ - t_i - \delta t)}{t_+ - t_i} = \frac{2 \cdot \delta t}{t_+ - t_i} = \frac{4 \cdot \delta t}{t_{i+1} - t_i}.$$

So, when $\delta t \ll t_{i+1} - t_i$, this relative accuracy is indeed small. And this is accuracy with which we estimate approximation accuracy – so this small difference can be indeed safely ignored: computing something with an accuracy 9.5% is, in effect, the same as computing something with an accuracy 10%.

In general, if $x_{i+1} - x_i \approx 2^{-b}$, then it is sufficient to describe the first b binary digits of the corresponding interval. This, the number of bits needed to store each threshold is approximately equal to $b \approx -\log_2(x_{i+1} - x_i)$. In this arrangement, we minimize the average number of bits, i.e., the sum $-\sum\limits_k \log_2(x_{i+1} - x_i)$ or the corresponding integral.

Towards formulating the corresponding optimization problems in precise terms. Since subintervals are small, the probability density functions do not change much over this subinterval, so we can safely assume that the corresponding distributions are uniform on this subinterval.

On the unit interval around a value x, there are $\rho_t(x)$ thresholds. Thus, the unit interval is divided into $\rho_t(x)$ subintervals. Hence, the width $w = x_{i+1} - x_i$ of each subinterval can be estimated as the ratio

$$w = \frac{1}{\rho_t(x)}.$$

On this interval, as one can easily check, the absolute value $a \stackrel{\text{def}}{=} |\varepsilon|$ of the difference ε between the midpoint and the actual value is uniformly distributed on the interval

$$\left[0, \frac{w}{2}\right] = \left[0, \frac{1}{2\rho_t(x)}\right].$$

This uniform distribution has a probability density

$$\rho_0(a) = \frac{1}{w/2} = \frac{2}{w}.$$

Thus, the average value of $|\varepsilon|^2$ on this interval is equal to

$$\int_0^{w/2} a^2 \cdot \rho_0(a) \, da = \frac{2}{w} \cdot \int_0^{w/2} a^2 \, da = \frac{2}{w} \cdot \left.\frac{a^3}{3}\right|_0^{w/2} =$$

$$\frac{2}{w}\cdot\frac{1}{3}\cdot\frac{w^3}{3}=\text{const}\cdot w^2=\text{const}\cdot\frac{1}{(\rho_t(x))^2}.$$

Each value x occurs with probability density $\rho(x)$, so minimizing the integral $\int(\varepsilon(t))^2\,dt$ is equivalent to minimizing the integral

$$\int\rho(x)\cdot\frac{1}{(\rho_t(x))^2}\,dx.$$

Similarly, for every $p\in[1,2)$, the average value on $|\varepsilon|^p$ of this interval is equal to

$$\int_0^{w/2}a^p\cdot\rho_0(a)\,da=\frac{2}{w}\cdot\int_0^{w/2}a^p\,da=\frac{2}{w}\cdot\frac{a^p}{p+1}\Big|_0^{w/2}=$$
$$\frac{2}{w}\cdot\frac{1}{p+1}\cdot\frac{w^{p+1}}{p+1}=\text{const}\cdot w^p=\text{const}\cdot\frac{1}{(\rho_t(x))^p}.$$

Each value x occurs with probability density $\rho(x)$, so minimizing the integral $\int|\varepsilon(t)|^p\,dt$ is equivalent to minimizing the integral

$$\int\rho(x)\cdot\frac{1}{(\rho_t(x))^p}\,dx.$$

For minimizing the number of bits, for each interval,

$$x_{i+1}-x_i\approx\frac{1}{\rho_t(x)},$$

so

$$-\log_2(x_{i+1}-x_i)=-\text{const}\cdot\ln(\rho_t(x)),$$

so the corresponding minimization is equivalent to minimizing the integral

$$-\int\rho(x)\cdot\ln(\rho_t(x))\,dx.$$

Let us solve the corresponding optimization problems. For the least squares optimization, we need to minimize the integral

$$\int\rho(x)\cdot\frac{1}{(\rho_t(x))^2}\,dx$$

under the constraint

$$\int\rho_t(x)\,dx=m.$$

This is a constrained optimization problem, and such problems can be solved by using the Lagrange multiplier method. For this particular problem, the Lagrange multiplier method mean optimizing the following objective function:

$$\int\rho(x)\cdot\frac{1}{(\rho_t(x))^2}\,dx+\lambda\cdot\int\rho_t(x)\,dx.$$

Differentiating this objective function with respect to each unknown $\rho_t(x)$ and equating the resulting derivative to 0, we conclude that

$$-2\cdot\frac{\rho(x)}{(\rho_t(x))^3}+\lambda=0,$$

i.e., that $(\rho_t(x))^3=\text{const}\cdot\rho(x)$ and $\rho_t(x)=\text{const}\cdot(\rho(x))^{1/3}$. The corresponding constant can be found from the condition that $\int\rho_t(x)\,dx=m$, thus,

$$\rho_t(x)=\frac{(\rho(x))^{1/3}}{\int(\rho(y))^{1/3}\,dy}.\qquad(1)$$

In particular, when $\rho(x)$ is a normal distribution with means μ and variance σ^2, the threshold distribution $\rho_t(x)$ is proportional to the normal distribution with the same mean and the variance $\frac{\sigma^2}{3}$.

For the ℓ^p-optimization, we need to minimize the integral

$$\int\rho(x)\cdot\frac{1}{(\rho_t(x))^p}\,dx$$

under the same constraint

$$\int\rho_t(x)\,dx=m.$$

For this problem, the Lagrange multiplier method means optimizing the following objective function:

$$\int\rho(x)\cdot\frac{1}{(\rho_t(x))^p}\,dx+\lambda\cdot\int\rho_t(x)\,dx.$$

Differentiating this objective function with respect to each unknown $\rho_t(x)$ and equating the resulting derivative to 0, we conclude that

$$-p\cdot\frac{\rho(x)}{(\rho_t(x))^{p+1}}+\lambda=0,$$

i.e., that $(\rho_t(x))^{p+1}=\text{const}\cdot\rho(x)$ and

$$\rho_t(x)=\text{const}\cdot(\rho(x))^{1/(p+1)}.$$

The corresponding constant can be found from the condition that $\int\rho_t(x)\,dx=m$, thus,

$$\rho_t(x)=\frac{(\rho(x))^{1/(p+1)}}{\int(\rho(y))^{1/(p+1)}\,dy}.\qquad(2)$$

In particular, when $\rho(x)$ is a normal distribution with means μ and variance σ^2, the threshold distribution $\rho_t(x)$ is proportional to the normal distribution with the same mean and the variance $\frac{\sigma^2}{p+1}$.

For the bit minimization, we need to minimize the integral

$$-\int\rho(x)\cdot\ln(\rho_t(x))\,dx$$

under the constraint

$$\int\rho_t(x)\,dx=m.$$

For this problem, the Lagrange multiplier method means optimizing the following objective function:

$$-\int\rho(x)\cdot\ln(\rho_t(x))\,dx+\lambda\cdot\int\rho_t(x)\,dx.$$

Differentiating this objective function with respect to each unknown $\rho_t(x)$ and equating the resulting derivative to 0, we conclude that

$$-\frac{\rho(x)}{\rho_t(x)}+\lambda=0,$$

i.e., that $\rho_t(x)=\text{const}\cdot\rho(x)$. The corresponding constant can be found from the condition that $\int\rho_t(x)\,dx=m$, thus,

$$\rho_t(x)=m\cdot\rho(x).\qquad(3)$$

In particular, when $\rho(x)$ is a normal distribution with means μ and variance σ^2, the threshold distribution $\rho_t(x)$ is proportional to the this normal distribution, with the same mean and the variance σ^2.

Comment. By definition of the threshold density, on each subinterval, we have

$$\int_{x_i}^{x_{i+1}} \rho_t(x)\,dx = 1,$$

thus, since $\rho(x) = \dfrac{1}{m} \cdot \rho_t(x)$, we conclude that the probability

$$p_i = \int_{x_i}^{x_{i+1}} \rho(x)\,dx$$

of being in the i-th subinterval is equal to

$$p_i = \int_{x_i}^{x_{i+1}} \rho(x)\,dx = \int_{x_i}^{x_{i+1}} \frac{1}{m} \cdot \rho_t(x)\,dx = \frac{1}{m} \cdot \int_{x_i}^{x_{i+1}} \rho_t(x)\,dx = \frac{1}{m}.$$

Thus, indeed, we have an "equiprobable" division into subintervals, when the probability p_i of being in a subinterval is the same for all the subintervals i.

III. How to Optimize Threshold Selection When We Take Measurement Inaccuracy into Account: Case of Interval Uncertainty

Case of interval uncertainty. In the ideal world, for each measuring instrument, we should know the probability distribution of measurement errors. This distribution can be determined if we compare the results of the given measuring instrument with the results of a super-precise "standard" measuring instrument.

This "calibration" process is possible, but it is usually very costly: sensors are cheap nowadays, but super-precise measuring instruments are not. As a result, in many cases, all we know is the upper bound Δ on the absolute measurement error; see, e.g., [11].

How measurement errors affect threshold selection. In the above analysis, the approximation error was equal to the difference $\varepsilon(t) = \widetilde{x}(t) - x(t)$ between the midpoint $\widetilde{x}(t)$ and the measured value $x(t)$. In the ideal-measurement case, any deviation from $x(t)$ is an inaccuracy.

However, if we take measurement uncertainty into account, then deviations not exceeding Δ are OK: the (unknown) actual value of the measured quantity can be anywhere within the interval $[x(t) - \Delta, x(t) + \Delta]$, so if the midpoint $\widetilde{x}(t)$ is within this interval, it can still be exactly equal to the actual value.

Only when $|\varepsilon(t)| > \Delta$, we know that there is an approximation error. This error can be gauged as the distance

$$d(\widetilde{x}(t), [x(t) - \Delta, x(t) + \Delta]) = \min\{d(\widetilde{x}(t), x) : x \in [x(t) - \Delta, x(t) + \Delta]\}$$

between the midpoint $\widetilde{x}(t)$ and the corresponding interval. One can check that this distance is equal to

$$d(\widetilde{x}(t), [x(t) - \Delta, x(t) + \Delta]) = \max(|\varepsilon(t)| - \Delta, 0).$$

This distance is the value that we should take into account (instead of $|\varepsilon(t)|$) when we select the optimal thresholds. Specifically, we should minimize either the sum of the squares of these distances, or, if we want a robust approach, the sum of their p-th powers.

Let us describe the average value of the corresponding power. For the square, as we have mentioned earlier, the value $a = |\varepsilon(t)|$ is uniformly distributed on the interval $\left[0, \dfrac{w}{2}\right]$. On this interval, the distance is equal to $\max(a - \Delta, 0))$, so the average value of the square of the distance is equal to

$$\frac{2}{w} \cdot \int_0^{w/2} (\max(a - \Delta, 0))^2\,da.$$

The value of $\max$ is non-zero only when $a \geq \Delta$. So when $w/2 \leq \Delta$, the integral is simply equal to 0. When $w/2 > \Delta$, then the integral takes the form

$$\frac{2}{w} \cdot \int_\Delta^{w/2} (a - \Delta, 0))^2\,da.$$

By introducing a new variable $a' = a - \Delta$, we get

$$\frac{2}{w} \cdot \int_0^{w/2-\Delta} (a')^2\,da' = \text{const} \cdot \frac{1}{w} \cdot \left(\frac{w}{2} - \Delta\right)^3 = \text{const} \cdot \frac{(w - 2\Delta)^3}{w}.$$

Here, $w = \dfrac{1}{\rho_t(x)}$, so, in terms of the threshold density, the integral takes the form

$$\text{const} \cdot \left(\frac{1}{\rho_t(x)} - 2\Delta\right)^3 \cdot \rho_t(x).$$

This expression corresponds to a given value x. To get an overall average, we need to multiply this expression by the probability density $\rho(x)$ of different x-values, and integrate over all possible x-values. Thus, we get

$$\int \left(\frac{1}{\rho_t(x)} - 2\Delta\right)^3 \cdot \rho_t(x) \cdot \rho(x)\,dx.$$

For the p-th powers, we similarly get

$$\frac{2}{w} \cdot \int_0^{w/2} (\max(a - \Delta, 0))^p\,da,$$

hence

$$\frac{2}{w} \cdot \int_\Delta^{w/2} (a - \Delta, 0))^p\,da$$

and

$$\frac{2}{w} \cdot \int_0^{w/2-\Delta} (a')^p\,da' = \text{const} \cdot \frac{1}{w} \cdot \left(\frac{w}{2} - \Delta\right)^{p+1} = \text{const} \cdot \frac{(w - 2\Delta)^{p+1}}{w}.$$

Substituting $w = \dfrac{1}{\rho_t(x)}$, we get

$$\text{const} \cdot \left(\frac{1}{\rho_t(x)} - 2\Delta\right)^{p+1} \cdot \rho_t(x).$$

So, we need to optimize the integral

$$\int \left(\frac{1}{\rho_t(x)} - 2\Delta\right)^{p+1} \cdot \rho_t(x) \cdot \rho(x)\, dx.$$

Let us solve the corresponding optimization problems. For the least squares case, the Lagrange multiplier methods leads to optimizing the expression

$$\int \left(\frac{1}{\rho_t(x)} - 2\Delta\right)^{3} \cdot \rho_t(x) \cdot \rho(x)\, dx + \lambda \cdot \int \rho_t(x)\, dx.$$

Differentiating this objective function with respect to each unknown $\rho_t(x)$ and equating the resulting derivative to 0, we conclude that

$$-3\left(\frac{1}{\rho_t(x)} - 2\Delta\right)^{2} \cdot \frac{\rho_t(x)}{(\rho_t(x))^2} \cdot \rho(x) + \left(\frac{1}{\rho_t(x)} - 2\Delta\right)^{3} \cdot \rho(x) + \lambda = 0.$$

Moving the first two terms into the right-hand side and taking into account that they have a common factor

$$\left(\frac{1}{\rho_t(x)} - 2\Delta\right)^{2} \cdot \rho(x),$$

we conclude that

$$\left(\frac{1}{\rho_t(x)} - 2\Delta\right)^{2} \cdot \left(\frac{3}{\rho_t(x)} - \frac{1}{\rho_t(x)} + 2\Delta\right) \cdot \rho(x) = \lambda.$$

Dividing both sides by $\rho(x)$, we get

$$\left(\frac{1}{\rho_t(x)} - 2\Delta\right)^{2} \cdot \left(\frac{2}{\rho_t(x)} + 2\Delta\right) = \frac{\lambda}{\rho(x)}. \qquad (4)$$

In contrast to the previous case, where we had an explicit solution, this is a generic cubic equation in terms of the unknown $\frac{1}{\rho_t(x)}$; we can still solve it, but no longer with a simple formula. The parameter λ needs to be determined from the condition that the overall number of thresholds is equal to m: $\int \rho_t(x)\, dx = m$.

For the ℓ^p-case, the Lagrange multiplier methods leads to optimizing the expression

$$\int \left(\frac{1}{\rho_t(x)} - 2\Delta\right)^{p+1} \cdot \rho_t(x) \cdot \rho(x)\, dx + \lambda \cdot \int \rho_t(x)\, dx.$$

Differentiating this objective function with respect to each unknown $\rho_t(x)$ and equating the resulting derivative to 0, we conclude that

$$-(p+1) \cdot \left(\frac{1}{\rho_t(x)} - 2\Delta\right)^{p} \cdot \frac{\rho_t(x)}{(\rho_t(x))^2} \cdot \rho(x) + \left(\frac{1}{\rho_t(x)} - 2\Delta\right)^{p+1} \cdot \rho(x) + \lambda = 0.$$

Moving the first two terms into the right-hand side and taking into account that they have a common factor

$$\left(\frac{1}{\rho_t(x)} - 2\Delta\right)^{p} \cdot \rho(x),$$

we conclude that

$$\left(\frac{1}{\rho_t(x)} - 2\Delta\right)^{p} \cdot \left(\frac{p+1}{\rho_t(x)} - \frac{1}{\rho_t(x)} + 2\Delta\right) \cdot \rho(x) = \lambda.$$

Dividing both sides by $\rho(x)$, we get

$$\left(\frac{1}{\rho_t(x)} - 2\Delta\right)^{p} \cdot \left(\frac{p}{\rho_t(x)} + 2\Delta\right) = \frac{\lambda}{\rho(x)}. \qquad (5)$$

The parameter λ also needs to be determined from the condition that the overall number of thresholds is equal to m: $\int \rho_t(x)\, dx = m$.

What if we minimize the number of bits. So far, we have described what will happen if we minimize the sum of the squares or the sum of the p-th powers. What if we minimize the number of bits?

In this case, the only restriction is that the width $w = \frac{1}{\rho_t(x)}$ cannot be smaller than 2Δ, and thus, the threshold density $\rho_t(x)$ cannot be larger than $\frac{1}{2\Delta}$. Minimizing the number of bits under this constraint leads to

$$\rho_t(x) = C \cdot \min\left(\rho(x), \frac{1}{2\Delta}\right). \qquad (6)$$

The constant C must also be determined from the condition that $\int \rho_t(x)\, dx = m$.

IV. Conclusions and Future Work

Conclusions. In this paper, we explore and propose enhancements to the use of Symbolic Aggregate Approximations (SAX) for data compression. While the intent of SAX is to take uncertainty into account, the current implementations of SAX do not account for all the uncertainty. So, we propose to extend the current SAX methodology to taking interval uncertainty into account.

Specifically, we propose to take interval uncertainty into account when selecting the thresholds. In this paper, we propose theoretical foundations and the resulting asymptotically optimal algorithms.

Future work. It is desirable to test the new algorithms on several real-life examples.

While the new algorithms lead to an asymptotically better data compression, which will hopefully lead to faster computations, implementing these algorithms requires an additional computational overhead. We know that asymptotically, the advantages outweigh this overhead. Testing on real-life examples would help us:

- to check whether the new algorithm is still beneficial for real-size data,
- and if this is not always the case, to find out what is the minimal size after which the new algorithm should be recommended.

ACKNOWLEDGMENT

This work was supported in part by the National Science Foundation grants HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and DUE-0926721. This work was performed when C. Stylios was a Visiting Researcher at the University of Texas at El Paso.

The authors are thankful to Martine Ceberio and to the anonymous referees for valuable suggestions.

REFERENCES

[1] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 2006.

[2] M. Butler and D. Kazakov, "SAX discretization does not guarante equiprobable symbols", *IEEE Transactions on Knowledge and Data Engineering*, 2015, Vol. 27, No. 4, pp. 1162–1166.

[3] K. Chakrabarti, E. Keogh, S. Mhrotha, and M. Pazzani, "Locally adaptive dimensionality reduction for indexing large time series databases", *ACM Transactions on Database Systems*, 2002, Vol. 27, No. 2, pp. 188–228.

[4] J. Fairley, P. Karvelis, G. Georgoulas, C. Stylios, D. Rye, and D. Bliwise, "Symbolic representation of human electromyograms for automated detection of phasic activity during sleep", *Proceedings of 2014 IEEE Statistical Signal Processing Workshop SSP'14*, Gold Coast, Queensland, Australia, June 29 – July 2, 2014.

[5] G. Georgoulas, P. Karvelis, T. Loutas, C. D. Stylios, "Rolling element bearings diagnostics using the Symbolic Aggregate approXimation", *Mechanical Systems and Signal Processing*, 2015, Vol. 60–61, pp. 229–242.

[6] G. Georgoulas, P. Karvelis, C. D. Stylios, I. P. Tsoumas, J. A. Antonino-Daviu, and V. Climente-Alarcon, "Automatizing the broken bar detection process via short time Fourier transform and two-dimensional piecewise aggregate approximation representation", *Proceedings of the IEEE Energy Conversion Congress and Exposition ECCE'2014*, Pittsburgh, Pennsylvania, September 14–18, 2014, pp. 3104–3110.

[7] R. W. Hamming, *Coding and Information Theory*, Ptentice Hall, Englewood Cliffs, New Jersey, 1986.

[8] P. J. Huber, *Robust Statistics*, Wiley, Hoboken, New Jersey, 2004.

[9] P. Karvelis, G. Georgoulas, I. Tsoumas, C. D. Stylios, J. Antonino-Daviu, and V. Climente-Alarcon, "An Intelligent icons approach for rotor bar fault detection", *Proceedings of the 39th Annual conference of the IEEE Industrial Electronics Society IECON'2013*, Vienna, Austria, November 10–13, 2013, pp. 5526–5531.

[10] J. Lin, E. Keogh, L. Weu, and S. Lonardi, "Experiencing SAX: a novel symbolic representation of time series", *Data Mining and Knowledge Discovery*, 2007, Vol. 15, No. 2, pp. 107–144.

[11] S. Rabinovich, *Measurement Errors and Uncertainties: Theory and Practice*, Springer Verlag, New York, 2005.

[12] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman & Hall/CRC, Boca Raton, Florida, 2011.