Computations Based on Delta-Modulation Representation of Measurement Results

George Djuro Zrilic

I. Direct dynamic measurement with interval uncertainty

In many real-life situations, we want to *monitor* the value of a physical quantity x for all moments of time t (e.g., to check that the object of measurement is performing in the right manner).

We want to make as many measurements as possible, so we have lots of information to send. However, the capacity of the communication channel is limited (and in many situations, e.g., in space exploration, the cost of adding an extra communication channel can be enormous). The smaller the interval between consequent measurements, the more information we need to send. So, the limit on the capacity of the communication channel restricts the interval between the measurements. Let us denote the smallest interval between the measurements (that the communication channel can still support) by Δt. If we denote the starting moment for our monitoring by t_0, then, since we want to measure as many values of the quantity $x(t)$ as possible, we will measure the value of $x(t)$ of the quantity y in the moments t_0, $t_1 = t_0 + \Delta t$, ..., $t_k = t_0 + k\Delta t$, etc.

Measurements are never absolutely precise [3]; therefore, the measurement result $\tilde{x}$ can differ from the actual value x of the corresponding quantity by the *measurement error* $\Delta x = \tilde{x} - x$. For a measuring instrument or a sensor to make sense, the manufacturer must provide us with the guaranteed error (if there is no guaranteed error, then we can conclude nothing from the fact that, e.g., the measured value was $\tilde{x} = 10.0$: the actual value can be 9.9, can be 2,000). In some cases, we know the probabilities of possible errors Δx. In many cases, however, the guaranteed upper bound Δ is the only information about the errors Δx that the manufacturer provides. In this cases, the only information that we know about the actual value of x is that this actual value belongs to an interval $\mathbf{x} = [\tilde{x} - \Delta, \tilde{x} + \Delta]$.

The author is with the Department of Engineering, New Mexico Highlands University, Las Vegas, NM 87701, email djuro@edison.nmhu.edu.

II. The main idea behind delta-modulation

A possibility to decrease the number of transmitted bits (and thus, to make more frequent monitoring measurements) comes from the fact that the measured quantities are usually changing continually, and we usually know the upper estimate M on the rate with which the measured quantity $x(t)$ changes (if we do not have any limits on M, then we have no information about the intermediate values $x(t)$, and our monitoring is of limited usage).

In this case, if we know the value $x(t_k)$ in the moment of time t_k, then the next value $x(t_{k+1})$ cannot deviate from $x(t_k)$ by more than $M \cdot \Delta t$.

Let us give an example of why this idea can indeed decrease number of bits that is necessary to carry a single measurement.

Example. Let us assume that we are measuring the temperature every millisecond, with an accuracy of 1 degree: $|\tilde{x}(t_k) - x(t_k)| \le \Delta = 1$. Let us also assume that the measured value $\tilde{x}(t_k)$ of the temperature x at some moment t_k is equal to $1,826$ degrees, and we know that during the interval between the two consequent measurement (i.e., during 1 millisecond) the temperature can change by no more than 2 degrees, i.e., that $|x(t_{k+1}) - x(t_k)| \le 2$. Therefore, the difference between the *measured* values of temperature cannot exceed 4 degrees:

$$|\tilde{x}(x_{k+1}) - \tilde{x}(t_k)| \le |\tilde{x}(t_{k+1}) - x(t_{k+1})| +$$

$$|x(t_{k+1}) - x(t_k)| + |x(t_k) - \tilde{x}(t_k)| \le 1 + 2 + 2 = 4.$$

According to the traditional approach, in the next moment of time t_{k+1}, we must send the numerical value of the measured temperature $\tilde{x}(t_{k+1})$. This value is an integer between $1,822 = 1,826 - 4$ and $1,830 = 1,826 + 4$. Therefore, it is an integer between $1,024 = 2^{10}$ and $2,048 = 2^{11}$, so, we need 11 binary digits to describe this measurement result.

On the other hand, instead of sending the value $\tilde{x}(t_{k+1})$, we can simply send the difference between $\tilde{x}(t_{k+1})$ and $\tilde{x}(t_k)$. This difference is an integer between -4 and 4, so it has only 9 possible values $(-4, -3, ..., 0, 1, ..., 4)$. So, we only need 4 bits (1 bit for sending a sign, and 3 bits for sending the absolute value of the difference) as opposed to 11 in the traditional approach. Because we need fewer bits to send the results of the measurements, we can hold measurements 11/4 (> 2) times more frequently than before.

We have already mentioned that ideally, we should be monitoring the value of $x(t)$ for every moment t, but in reality, we only get the values in the moments $t_1, ..., t_k, ...$. So, if we are interested in the value of $x(t)$ for some intermediate moment of time t, i.e., in a moment of time that lies in between t_k and t_{k+1} for some k, then as an estimate for $x(t)$, we take the latest available measured value, i.e., $\tilde{x}(t_k)$. Even if we measured $x(t_k)$ precisely, this difference in times between t and t_k would still to an error in this estimate, an error $x(t) - x(t_k)$ that is limited by $M \cdot (t - t_k) \le M \cdot \Delta t$. The additional measurement error $\tilde{x}(t_k) - x(t_k)$ may increase the total error $x(t) - \tilde{x}(t_k)$ of using $\tilde{x}(t_k)$ as an estimate for $x(t)$.

Since we already have an error component of size $M \cdot \Delta t$, it makes no big sense to measure the values $x(t_k)$ with accuracy that is much better than $M \cdot \Delta t$ (i.e., there is no sense in trying to achieve measurement errors that are much smaller than $M \cdot \Delta t$): such super-accurate measurements would mean using very expensive sensors, but their usage will not seriously improve the resulting error, because this error will still be of order $M \cdot \Delta t$.

So, the measurement accuracy Δ is usually chosen to be smaller than $M \cdot \Delta t$, but approximately of the same order ($\Delta < M \cdot \Delta t$, $\Delta \approx M \cdot \Delta t$). With this choice, the difference $x(t_{k+1}) - x(t_k)$ (that is $\le M \cdot \Delta t$) is measured with an error that is close to the value of this difference. With such a huge measurement error, we can basically distinguish between only two cases:

- the case when this difference is positive, and
- the case when this difference is negative.

So, the sensor gets the measurement results $\tilde{x}(t_1), ...$, but it sends for processing only one bit per moment of time. This bit actually represents a sign of the difference between the two consequent values of the signal (i.e., whether x increased with respect to the previous moment of time or not), so it is natural to represent this bit not as 0 or 1, but as a *sign*, i.e., as $+1$ or -1. Let us denote the sign bit that comes out of the sensor in the moment t_k by $s(k)$. Then, at the receiving end of the communication channel, all we have is a sequence of sign bits $s(1), ..., s(k), ...$. In order to be able to reconstruct the signal from the this sequence, we must know the initial value of the signal $\tilde{x}(t_0)$. How can we reconstruct the signal from this sequence? There is not much that we can do but follow the following natural algorithm:

- As the initial value $r(0)$ of the reconstructed signal r, we simply take $r(0) = \tilde{x}(t_0)$.
- To get further reconstructed values $r(k)$, we proceed as follows:
  - if we have already computed $r(k)$, and the next bit that comes out of the communication channel is $+1$, we add $\alpha = M \cdot \Delta t$ (i.e., take $r(k+1) = r(k) + M \cdot \Delta t$);
  - if we have already computed $r(k)$, and the next bit that comes out of the communication channel is -1, we subtract $M \cdot \Delta t$ (i.e., take $r(k+1) = r(k) + M \cdot \Delta t$).

This natural reconstruction algorithm leads to the following natural idea of selecting a signal $s(k)$ that would go through the communication channel:

- At every moment of time, after we generate the communication bit, we also simulate the reconstruction procedure at the sensor's end of the communication channel (thus getting $r(0)$, $r(1)$, ...).
- After a new measured value $\tilde{x}(t_k)$ arrives, we compare it with the previously reconstructed signal $r(k-1)$. Now, we have two options:
  - If we choose $s = +1$, then the next reconstructed signal $r(k)$ will be greater than $r(k-1)$.
  - If we choose $s = -1$, then the next reconstructed signal $r(k)$ will be greater than $r(k-1)$.

  So, to get $r(k)$ as close to $\tilde{x}(t_k)$ as possible, we will choose:
  - $s(k) = +1$ if the value of $r(k-1)$ is smaller than $\tilde{x}(t_k)$ (and thus needs to be increased), and
  - $s(k) = -1$ if the value of $r(k-1)$ is greater than $\tilde{x}(t_k)$ (and thus needs to be decreased).

The resulting algorithm is called *delta-modulation.*

Comment. For a recent survey of delta-modulation techniques, see, e.g., [2] and references therein. These methods and results, however, are mainly developed for the statistical case.

III. Direct dynamic measurement and its error estimate

Definition 1. By a *dynamic measuring instrument*, we mean a pair $(\Delta, \Delta t)$ of two positive numbers:

- a number $\Delta > 0$ will be called the *measurement accuracy;*
- a number $\Delta t > 0$ will be called a *time quantum.*

Definition 2. By a *dynamic measurement situation*, we mean a tuple $(I, M, t_0, x, \{t_k\}, \{\tilde{x}(t_k)\})$, where:

- I is a dynamic measuring instrument.
- M is a positive real number called the *(prior) bound on the rate of change of the signal.* We will assume that $\Delta \leq M \cdot \Delta t$.
- t_0 is a real number called the *initial moment of time.*
- x is a function from real numbers into real numbers that is an $M-$Lipschitz function (i.e., $|x(t) - x(s)| \leq M|t - s|$ for all t and s).
- $\{t_k\}$, $0 \leq k$, is a sequence of real numbers defined as $t_k = t_0 + k \cdot \Delta t$. A number t_k will be called *$k-th$ measurement moment.*
- $\{\tilde{x}(t_k)\}$ is a sequence of real numbers for which for every k, $|\tilde{x}(t_k) - x(t_k)| \leq \Delta$. The element $\tilde{x}(t_k)$ will be called the *result of $k-th$ measurement.*

Comment. Let us first consider the case when we do not use delta-modulation.

Definition 3. For every dynamic measurement situation, and for every moment of time $t \geq t_0$, by a *monitoring error*, we mean the difference $\tilde{x}(t_k) - x(t)$, where k is the largest value for which $t_k \leq t$.

PROPOSITION 1.

- *For every dynamic measurement situation, and for every moment of time t, the absolute value of the monitoring error does not exceed $M \cdot \Delta t + \Delta$.*
- *For every $M > 0$, t_0, $\Delta t > 0$, $\Delta > 0$, and $\delta > 0$, there exists a dynamic measurement situation and a moment of time $t \geq t_0$ for which the monitoring error is not smaller than $M \cdot \Delta t + \Delta - \delta$.*

Comments.

1. These two statements mean that:
   - $M \cdot \Delta t + \Delta$ is the error bound for monitoring error, and
   - that no better bound is possible.
2. The fact that the error bound is $M \cdot \Delta t + \Delta$ can be easily explained by the fact that we have two sources of error:
   - the measurement error, whose bound is Δ, and
   - the error caused by the difference between t and t_k whose upper bound is $M \cdot \Delta t$.
3. The results from this paper has been proven by V. Kreinovich and me.

A. Proof of Proposition 1

Let us first prove that the monitoring error is always bounded by $M \cdot \Delta t + \Delta$. Indeed, if $t_k \leq t < t_{k+1}$, then $0 \leq t - t_k < t_{k+1} - t_k = \Delta t$ and therefore, by the definition of a measurement situation, $|x(t) - x(t_k)| \leq M \cdot |t - t_k| \leq M \cdot \Delta t$. From the same definition, it follows that $|\tilde{x}(t_k) - x(t_k)| \leq \Delta$. Therefore, $|x(t) - \tilde{x}(t_k)| \leq |x(t) - x(t_k)| + |x(t_k) - \tilde{x}(t_k)| \leq M \cdot \Delta t + \Delta$. So, the monitoring error is indeed always bounded by $M \cdot \Delta t + \Delta$.

Let us now show that a smaller bound for a monitoring error is impossible. Indeed, assume that $\delta > 0$, t_0, M, Δt are fixed. As a measured signal, let us take the function $x(t) = M(t - t_0)$. As measurement results, we will take $\tilde{x}(t_k) = x(t_k) - \Delta$. Then, for $t = t_1 - \delta/M$, the monitoring error is equal to

$$x(t) - \tilde{x}(t_0) = M(t_1 - \delta/M - t_0) - (M(t_0 - t_0) - \Delta) =$$

$$M \cdot \Delta t - \delta + \Delta.$$

Q.E.D.

IV. Delta-modulation: formal definition, and why it really helps

Let us now define delta-modulation and show that its usage (while saving on communication) does not decrease the resulting monitoring error. Namely, we will show that with delta-modulation, we can achieve the same monitoring error if we double the measuring rate. At this rate, we need to transmit twice as many measurement results. However, since we only need one bit to transmit a single delta-modulated measurement result, and we need several bits to transmit the actual measurement result $\tilde{x}(t_k)$, the resulting total amount of bits per second that needs to be transmitted is smaller when we use delta-modulation.

Definition 4.

- For every dynamic measurement situation, by the *result of delta-modulation* (applied to the sequence $\tilde{x}(t_k)$), we mean the sequence

  $$s(1), ..., s(k), ...$$

  whose elements are determined by the formula

  $$s(k) = \text{sign}\big[\tilde{x}(t_k) - \big(\tilde{x}(t_0) + M \cdot \Delta t \cdot \sum_{j=1}^{k-1} s(j)\big)\big], \tag{1}$$

  where:

  - $\text{sign}(a) = +1$ if $a \geq 0$, and
  - $\text{sign}(a) = -1$ if $a < 0$.

- For every dynamic measurement situation, by the *reconstructed* or *delta-demodulated* signal, we mean a sequence

$$r(k) = \tilde{x}(t_0) + M \cdot \Delta t \cdot \sum_{j=1}^{k} s(j). \qquad (2)$$

Definition 5. For every dynamic measurement situation, and for every moment of time $t \geq t_0$, by a *monitoring error after (delta-)demodulation*, we mean the difference $r(k) - x(t)$, where k is the largest value for which $t_k \leq t$.

PROPOSITION 2.

- *For every dynamic measurement situation, and for every moment of time t, the absolute value of the monitoring error after delta-demodulation does not exceed $2M \cdot \Delta t + \Delta$.*
- *For every $M > 0$, t_0, $\Delta t > 0$, $\Delta > 0$, and $\delta > 0$, there exists a dynamic measurement situation and a moment of time $t \geq t_0$ for which the monitoring error after delta-demodulation is not smaller than $2M \cdot \Delta t + \Delta - \delta$.*

Comments.

1. These two statements mean that:
   - $2M \cdot \Delta t + \Delta$ is the error bound for monitoring error after delta-demodulation, and
   - that no better bound is possible.
2. Because of Proposition 2, if for measurements with delta-modulation, we take the time quantum Δt that is twice smaller than the one that was used for regular measurements, we will get exactly the same monitoring error as for measurements without modulation. Let us give two examples:
   - Suppose that we measure $x(t_k)$ with an accuracy of 1%. This means that possible measurement results run from -100 to 100 quanta. The binary representation of 100 takes 7 bits, so, with an extra bit for sign, we need 8 bits to transmit the result of a single measurement. If we use delta-modulation, then we only need one bit per measurement, but these measurements must be two times more frequent. So, when initially we needed 8 bits, we now need only two. Therefore, if we use delta-modulation, we can keep the same total error and reduce the information flow by the factor of four.
   - Suppose now that we measure $x(t_k)$ with an accuracy of 0.1%. In this case, possible measurement results run from -1000 to 1000. The binary representation of 1000 takes 10 bits, so we need 11 bits to transmit the result of a single measurement. If we use delta-modulation, and aim at the same accuracy of the final result, we thus need two bits during the same time quantum Δt. So, we decrease the information flow by the factor of 5.5.

   In general, the more accurate the measurements, the more we save by using delta-modulation.

A. Proof of Proposition 2

Let us first start with proving the inequality, and then produce an example that proves the second part of this proposition.

To prove the inequality, we will first prove (by induction over k) that $|x(t_k) - r(k)| \leq M \cdot \Delta t + \Delta$ for all k.

Induction base. The initial reconstructed value $r(0)$ is defined as $r(0) = \tilde{x}(t_0)$. But by definition of a measurement situation, we have $|\tilde{x}(t_0) - x(t_0)| \leq \Delta$. Therefore,

$$|r(0) - x(t_0)| \leq \Delta < M \cdot \Delta t + \Delta.$$

Induction step. Assume that we have already proved the desired inequality for k, i.e., that

$$|x(t_k) - r(k)| \leq M \cdot \Delta t + \Delta. \qquad (3)$$

We must prove that a similar inequality holds for $k+1$, i.e., that

$$|x(t_{k+1}) - r(k+1)| \leq M \cdot \Delta t + \Delta. \qquad (4)$$

To prove that, we will consider two cases:

- the case when $\tilde{x}(t_{k+1}) \geq r(k)$, and therefore, $s(k+1) = 1$.
- the case when $\tilde{x}(t_{k+1}) < r(k)$, and therefore, $s(k+1) = -1$.

In the first case, $r(k+1) = r(k) + M \cdot \Delta t$. From (3), we conclude that

$$x(t_k) \leq r(k) + M \cdot \Delta t + \Delta. \qquad (5)$$

From the definition of a measuring situation, we conclude that $x(t_{k+1}) - x(t_k) \leq M \cdot (t_{k+1} - t_k) = M \cdot \Delta t$, so $x(t_{k+1}) \leq x(t_k) + M \cdot \Delta t$. Replacing $x(t_k)$ by its upper bound taken from (5), we conclude that

$$x(t_{k+1}) \leq r(k) + M \cdot \Delta t + \Delta + M \cdot \Delta t.$$

Since $r(k) + M \cdot \Delta t = r(k+1)$, we conclude that $x(t_{k+1}) \leq r(k+1) + \Delta + M \cdot \Delta t$, i.e., that $x(t_{k+1}) - r(k+1) \leq M \cdot \Delta t + \Delta$.

This is a half of the desired inequality (4). To complete the proof for this case, it is thus necessary to prove the other half of this inequality, i.e., to prove that $x(t_{k+1}) \geq r(k+1) - (M \cdot \Delta t + \Delta)$. Since in the case under consideration, $r(k+1) = r(k) + M \cdot \Delta t$, this inequality is equivalent to $x(t_{k+1}) \geq r(k) - \Delta$. This follows from the following sequence of inequalities:

- in this case, $\tilde{x}(t_{k+1}) \geq r(k)$;
- by definition of a measurement situation,
$$x(t_{k+1}) \geq \tilde{x}(t_{k+1}) - \Delta;$$
- therefore, $x(t_{k+1}) \geq \tilde{x}(t_{k+1}) - \Delta \geq r(k) - \Delta$.

This inequality is thus proved, and so (4) is true in the first case.

The second case is proved similarly. Now, the desired inequality follows in a manner similar to the proof of Proposition 1: if $t_k \leq t < t_{k+1}$, then

$$|x(t) - r(k)| \leq |x(t) - x(t_k)| + |x(t_k) - r(k)| \leq$$
$$M \cdot |t - t_k| + M \cdot \Delta t + \Delta \leq$$
$$M \cdot \Delta t + M \cdot \Delta t + \Delta.$$

Let us now give an example of the measurement situation in which the monitoring error is not smaller than $2M \cdot \Delta t + \Delta - \delta$:

x: we choose the following function $x(t)$:
  $x(t_2) = -M\Delta t$;
  $x(t_k) = 0$ for $k \neq 2$:
  $x(t)$ is linear on each of the intervals $[t_k, t_{k+1}]$.
$\tilde{x}$: we take $\tilde{x}(t_k) = x(t_k) + \Delta$ for all k.
t: we take $t = t_2 - \delta/M$.

In this case:

- $r(0) = \tilde{x}(t_0) = \Delta$.
- Since $\tilde{x}(t_1) = \Delta \geq r(0)$, we have $s(1) = +1$ and $r(1) = r(0) + M \cdot \Delta t = \Delta + M \cdot \Delta t$.
- We have $t_1 \leq t < t_2$. Since x is linear on $[t_1, t_2]$, we have
$$x(t) = x(t_1) \cdot \frac{t - t_1}{t_2 - t_1} + x(t_2) \cdot \frac{t_2 - t}{t_2 - t_1} =$$
$$(-M\Delta t) \cdot \frac{\Delta t - \delta/M}{\Delta t} = -M \cdot \Delta t + \delta.$$

So, here,

$$|r(1) - x(t)| = |(\Delta + M \cdot \Delta t) - (-M \cdot \Delta t + \delta)| =$$
$$2M \cdot \Delta t + \Delta - \delta.$$

Q.E.D.

V. Error estimation of the result of delta-demodulation

When the signals arrive in a delta-modulated form, the following problem arises: modulation and demodulation (reconstruction) take some extra computation time. So, if we want to process the measured values, and we simply demodulate and process, then we lose extra time. Can we somehow modify data processing algorithms so that they can deal directly with a delta-modulated form and produce the desired result without spending extra time on demodulation? It will be shown that in many real-life cases, we can process the delta-modulated sequence directly, and this usage will not only *not* increase the computation time, but often drastically *decrease* this time.

References

[1] M. Freedman, D. G. Zrilic, *Nonlinear arithmetic operations on the delta sigma pulse stream*, Signal Processing, 1990, Vol. 21, pp. 25–35.

[2] I. Galton, *Granular quantization noise in a class of delta-sigma modulators*, IEEE Transactions on Information Theory, 1994, Vol. 40, No. 3, pp. 848–859.

[3] S. Rabinovich, *Measurement errors: theory and practice*, American Institute of Physics, N.Y., 1993.