Optimized Sampling Frequencies for Weld Reliability Assessments of Long Pipeline Segments

Cesar J. Carrasco
Department of Civil Engineering
University of Texas at El Paso
500 W. University
El Paso, TX 79968, USA

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

***Abstract*— In this paper, we describe new faster algorithms that design an optimal testing strategy for long pipeline segments.**

I. Introduction

Inspecting Pipelines is Important: Modern technology depends on oil, gas, and other substances which are often transported by long buried pipelines. High-pressure corrosive substances transported within a pipeline and often severe soil interaction make pipelines vulnerable. An especially vulnerable part of the pipeline is a weld where different pipes forming a pipeline are welded together.

A pipeline disruption can lead to serious environmental problem and sometimes, when the pipeline disruption occurs in populated areas, to human disasters. It is therefore necessary to periodically inspect pipelines.

This problem is especially important for older pipelines particular those designed in the 1950s and earlier, when the quality of welding was much lower than at present.

Inspecting Pipelines is Possible: In principle, each pipeline segment can be inspected by using mechanized ultrasonic inspection techniques.

Inspecting Pipelines is Expensive: For long pipelines, it is prohibitively expensive to inspect all the welds, especially if we take into account the high cost of excavation.

Solution: Statistical Sampling: As a result, instead of inspecting all parts of the pipeline, practitioners sample several locations, and make statistical estimates based on the results of the sampling.

Result of Statistical Sampling: Failure Assessment: Based on the measurements, we estimate the probability distributions describing different pipeline flaws – i.e., to be more precise, we estimate the parameters of the corresponding distributions such as their mean and standard deviation.

There exist accurate mathematical models that translate the values of these parameters into the probability of the pipeline failure p. If this probability exceeds a certain regulated (small) threshold p_0, the pipeline must be repaired.

Our Estimates of p are Also Approximate: Because of the limited sampling, we can only determine these parameters with uncertainty; thus, the value $\widetilde{p}$ calculated based on these measurement result is also only an approximation to the desired probability p – in other words, we have an interval $[\underline{p},\overline{p}]$ of possible values of p. To be on the safe side, regulations require that a pipeline be repaired when it is possible that $p \ge p_0$, i.e., when $\overline{p} \ge p_0$.

Need for Optimal Resource Allocation for Pipeline Assessment: Pipeline repairs are extremely expensive. To avoid unnecessary repairs, it is therefore important to generate estimates for p which are as accurate as possible.

So, given the amount of resources available for the weld reliability assessment, we must allocate these resources to different possible measurements so as to provide the most accurate estimation of the probability failure.

Where Uncertainty Comes into Picture: Due to the uncertainty with which we know many of the factors, we need to provide an optimal solution under uncertainty.

What Was Known and What We Propose: At present, the only way to find the optimal design is, in effect, to exhaustively search all possible combinations of numbers of different measurements $n_1, \ldots, n_k$. In this paper, we provide an analytical (fast) algorithm that find an almost optimal design (and we show that finding the exactly optimal design is NP-hard).

II. Statistical Approach to Pipeline Reliability Assessment: Motivations

Traditionally, a statistical approach is used to assess reliability of a pipeline. Let us describe the motivations behind this approach.

Let $x_1, \ldots, x_k$ be measurable parameters that describe the reliability of a pipeline – such as the pipe's thickness, parameters describing pipe deformation, different degrees of corrosion, etc.

The value of each of these parameters randomly varies from point to point. It is therefore reasonable to consider each of these parameters as a random variable.

The actual value of each of these parameters x_i is caused by a large number of different independent factors. In principle, there are some major factors that affect the state of the pipeline; to extend the pipeline's service, pipelines are designed in such a way that the effect of these major factors is minimized. For example, a proper insulation is placed around the pipeline, special anti-corrosion layers are added inside the pipe, etc. After we exclude these major factors, the state of the pipe is affected by the large number of relatively small difficult-to-exclude processes. By itself, each of these processes has a rather small influence on the pipeline, but together (and especially in the long run), these processes can lead to a drastic decrease in the pipeline's reliability.

It is known that for large n, the sum of n independent identically distributed random variables is almost normally distributed. This *Central Limit Theorem* is one of the main reasons why Gaussian distribution is so frequent in practice; see, e.g., [10]. We can therefore conclude that each of the parameters x_i can be characterized as a normally distributed random variable.

It is well known (see, e.g., [10]) that a normally distributed random variable is uniquely determined by its mean and its variance. Thus, to fully characterize the state of the pipeline, we must know the means $a_1, \ldots, a_k$ and the standard deviations $\sigma_1, \ldots, \sigma_k$ of the parameters $x_1, \ldots, x_k$.

Based on this information, we need to assess the reliability of the pipeline – measured, e.g., by the probability p of the pipeline's failure. For different types of pipelines, there exist models f that estimate the probability p of the pipeline's failure as a function of the values $a_1, \ldots, a_k, \sigma_1, \ldots, \sigma_k$:

$$p = f(a_1, \ldots, a_k, \sigma_1, \ldots, \sigma_k). \qquad (1)$$

III. To Make a Meaningful Decision About the Pipeline, We Must Also Know How Accurate Is the Reliability Estimate

Our objective is to make sure that the probability of failure p does not exceed the established small threshold p_0.

The models used in assessing pipeline reliability are reasonably sophisticated. So, in the ideal case when we know the exact values of the statistical characteristics a_i and σ_i, these models provide a very good estimate of the pipeline's reliability. Therefore, in this ideal case, it is easy to check whether the pipeline can still be exploited or needs immediate maintenance:

- if the resulted value p is smaller than the threshold value p_0, this means that we can continue to safely exploit this pipeline;
- on the other hand, if the resulted value p exceeds the desired threshold, this means that the pipeline needs to be serviced before it can be further exploited.

In practice, we do not know the exact values of these statistical characteristics. To find their values, we measure the values of the corresponding quantities x_i at different locations, and then use standard statistical techniques to estimate these characteristics. Specifically, for each quantity x_i, we perform n_i measurements at different places along the pipeline. Based on the results $x_k^{(1)}, \ldots, x_k^{(n_k)}$ of these measurements, we then compute the following estimates $\widetilde{a}_k$ and $\widetilde{\sigma}_k$ for the desired values a_k and σ_k:

$$\widetilde{a}_k = \frac{x_k^{(1)} + \ldots + x_k^{(n_k)}}{n_k};$$

$$\widetilde{\sigma}_k = \sqrt{\frac{(x_k^{(1)} - \widetilde{a}_k)^2 + \ldots + (x_k^{(n_k)} - \widetilde{a}_k)^2}{n_k - 1}}.$$

We then compute an estimate $\widetilde{p}$ for the pipeline's reliability – by substituting these estimates into the reliability model f:

$$\widetilde{p} = f(\widetilde{a}_1, \ldots, \widetilde{a}_k, \widetilde{\sigma}_1, \ldots, \widetilde{\sigma}_k). \qquad (2)$$

How can we use this estimate to gauge the pipeline reliability? In the ideal case, we can simply compare the probability p with the desired threshold p_0. However, since the estimates $\widetilde{a}_i$ and $\widetilde{\sigma}_i$ are only approximate, the resulting estimate $\widetilde{p}$ for p is also only approximate. So, we cannot simply conclude that the pipeline can be exploited by simply comparing the estimate $\widetilde{p}$ with the desired threshold p_0: even if $\widetilde{p} < p_0$, it is still possible that the actual probability p is larger than the estimate $\widetilde{p}$ and larger than the threshold p_0.

So, to make a correct decision on the pipeline's state, we must know not only the estimate $\widetilde{p}$ for the pipeline's probability of failure, we must also know how accurate is this estimate. In other words, we would like to have some information about the estimation error $\Delta p \stackrel{\text{def}}{=} \widetilde{p} - p$.

IV. Formulas for the Accuracy of the Reliability Estimate

Based on the measurements, we get reasonable approximations $\widetilde{a}_i$ and $\widetilde{\sigma}_i$ to the actual values a_i and σ_i of the corresponding statistical characteristics. In other words, the corresponding estimation errors $\Delta a_i \stackrel{\text{def}}{=} \widetilde{a}_i - a_i$ and $\Delta \sigma_i \stackrel{\text{def}}{=} \widetilde{\sigma}_i - \sigma_i$ are small – and thus, in our computations, we can safely ignore terms which are quadratic and higher order in terms of Δa_i and $\Delta \sigma_i$.

In particular, if we substitute the expressions $a_i = \widetilde{a}_i - \Delta a_i$ and $\sigma_i = \widetilde{\sigma}_i - \Delta \sigma_i$ into the formula (1), expand the result in Taylor series in terms of small quantities Δa_i and $\Delta \sigma_i$, and then ignore quadratic and higher order terms in this expansion, we conclude that

$$p = \widetilde{p} - \frac{\partial f}{\partial a_1} \cdot \Delta a_1 - \ldots - \frac{\partial f}{\partial a_k} \cdot \Delta a_k -$$
$$\frac{\partial f}{\partial \sigma_1} \cdot \Delta \sigma_1 - \ldots - \frac{\partial f}{\partial \sigma_k} \cdot \Delta \sigma_k.$$

Thus, for the desired estimation error $\Delta p = \widetilde{p} - p$, we get the following formula:

$$\Delta p = \sum_{i=1}^{k} \frac{\partial f}{\partial a_i} \cdot \Delta a_i + \sum_{i=1}^{k} \frac{\partial f}{\partial \sigma_i} \cdot \Delta \sigma_i. \qquad (3)$$

For a reasonably large number of measurements, the estimation errors Δa_i and $\Delta \sigma_i$ are independent and (almost) normally distributed. It is known that the standard statistical estimates are un-biased – so the mean values of the estimation errors is 0, that the standard deviation of the estimation error Δa_i decreases with the number of measurements as $\dfrac{\sigma_i}{\sqrt{n_i}}$, and that the standard deviation of the estimation error $\Delta \sigma_i$ decreases with the number of measurements as $\dfrac{\sigma_i}{\sqrt{2n_i}}$; see, e.g., [9], [10]. So, the desired estimation error Δp is a linear combination of independent normally distributed random variables with 0 means and known standard deviations.

It is known that, in general, a linear combination $\sum\limits_{i=1}^{k} \alpha_i \cdot \xi_i$ of independent normally distributed random variables ξ_i with 0 mean and standard deviations σ_i is also normally distributed, with 0 mean and the standard deviation σ for which $\sigma^2 = \sum\limits_{i=1}^{n} \alpha_i^2 \cdot \sigma_i^2$. By applying this known formula to the expression (3), we conclude that the reliability estimation error Δp is normally distributed, with 0 mean and standard deviation σ for which

$$\sigma^2 = \sum_{i=1}^{k} \left(\frac{\partial f}{\partial a_i}\right)^2 \cdot \frac{\sigma_i^2}{n_i} + \sum_{i=1}^{k} \left(\frac{\partial f}{\partial \sigma_i}\right)^2 \cdot \frac{\sigma_i^2}{2n_i}. \tag{4}$$

V. How to Access the Pipeline Reliability? Assessment Process and a Possible Need for Additional Measurements

Suppose that we have performed a few measurements, and came up with the estimates $\widetilde{a}_i$, $\widetilde{\sigma}_i$, and $\widetilde{p}$. By using the formula (4) and the known properties of the normal distribution, we can now estimate the accuracy of the estimate $\widetilde{p}$:

- For example, it is known that with probability 95%, a normally distributed random variable with a mean a and standard deviation σ is within a “two sigma” interval
$$[a - 2\sigma, a + 2\sigma].$$
In our case, this means that with probability 95%, the actual values p of the probability of failure does not exceed $\widetilde{p} + 2\sigma$.
- Similarly, from the fact that with probability 99.9%, a normally distributed random variable with a mean a and standard deviation σ is within a “three sigma” interval $[a - 3\sigma, a + 3\sigma]$. In our case, this means that with probability 99.9%, the actual values p of the probability of failure does not exceed $\widetilde{p} + 3\sigma$, etc.

If the resulting upper bound for p is smaller than p_0, this means that the pipeline is still operational.

However, if $\widetilde{p} < p_0$, but the resulting upper bound $\widetilde{p} + k_0 \cdot \sigma$ exceeds p_0, the problem may be that we have performed too few measurements and, as a result, our estimate $\widetilde{p}$ is too crude. In this situation, to check the pipeline's reliability, we must perform additional measurements.

VI. Towards Optimal Way of Accessing the Pipeline Reliability: Formulation of the Problem

How many measurements do we need? Once we know the current reliability estimate $\widetilde{p} < p_0$, and have selected the value k_0 (corresponding to the desired certainty), we will need as many measurements to be able to guarantee, with the selected degree of certainty, that $p \le p_0$. In other words, we want to make sure that the resulting standard deviation σ satisfies the inequality $\widetilde{p} + k_0 \cdot \sigma \le p_0$ – i.e., equivalently, the inequality $\sigma \le \sigma_0 \stackrel{\text{def}}{=} \dfrac{p_0 - \widetilde{p}}{k_0}$. So, we must select the number of different measurements n_i in such a way that the standard deviation σ, as described by the formula (4), does not exceed the given value σ_0. This condition can be described as follows:

$$\sum_{i=1}^{k} \frac{b_i}{n_i} \le \varepsilon_0, \tag{5}$$

where we denoted

$$b_i \stackrel{\text{def}}{=} \left(\frac{\partial f}{\partial a_i}\right)^2 \cdot \sigma_i^2 + \left(\frac{\partial f}{\partial \sigma_i}\right)^2 \cdot \frac{\sigma_i^2}{2};$$

$$\varepsilon_0 \stackrel{\text{def}}{=} \sigma_0^2.$$

We have one inequality condition to find k different characteristics $n_1, \ldots, n_k$ – sampling frequencies for k different quantities $x_1, \ldots, x_k$. Therefore, there exist several possible designs that satisfy this condition. Which of these designs should we select?

We have already mentioned that measurements are expensive – this is one of the main reasons why we did not perform a large number of measurements in the first place. Thus, when selecting an experiment design, it is extremely important to minimize the overall measurement cost.

This cost consists of two major parts: the cost of measurements themselves, and the cost of excavation that is needed to gain access to the pipeline. We know the cost c_i of a single measurement of the quantity x_i, and we know the cost c_0 of a single excavation. To perform all the measurements, we need $\max_i n_i$ excavations. Thus, the overall costs can be described as

$$\sum_{i=1}^{k} c_i \cdot n_i + c_0 \cdot \max_i n_i. \tag{6}$$

Thus, we arrive at the following exact formulation of the problem of optimization of sampling frequencies for assessing weld reliability of long pipeline segments:

- we are given positive values $b_1, \ldots, b_k$, ε_0, $c_1, \ldots, c_k$, and c_0;
- among all integer arrays $\vec{n} = (n_1, \ldots, n_k)$ that satisfy the inequality (5), we must find a one that minimizes the overall cost (6).

In this paper, we prove that if we want to solve this problem exactly, then this problem is computationally intractable (NP-hard). We also produce a reasonable efficient algorithm that provides an asymptotically optimal solution to this problem.

VII. First Result: Finding the Exact Optimum is Computationally Intractable (NP-Hard)

Let us prove that the exact optimization problem is NP-hard; for exact definitions of NP-hardness, see, e.g., [4], [5]. Specifically, we will prove that even a simplified version of our original optimization problem is NP-hard, a version in which instead of choosing arbitrary values n_i, we only have two choices n and n'. In other words, for each i, we either select to perform a small number n of measurements, or a large number $n' > n$ of measurements.

This selection can be described by a single Boolean variable y_i that is equal:

- to 1 if, for x_i, we select a small number of measurements, and
- to 0 if, for x_i, we select a large number of measurements.

If we did not have a restriction on accuracy, then, of course, the smallest cost would mean that for every i, we perform the small number of measurements $n_i = n$. To avoid this trivial solution, we assume that to achieve the desired accuracy, it is not sufficient to perform a small number of measurements for all i – i.e., that $\sum\limits_{i=1}^{k} \frac{b_i}{n} > \varepsilon_0$. In this case, for at least one of the quantities x_i, we must have $n_i = n'$ (i.e., $y_i = 0$). Thus, $\max(n_i) = n'$, and the objective function takes the form $\sum\limits_{i=1}^{k} c_i \cdot n_i + c_0 \cdot n'$.

To prove NP-hardness of our problem, we will reduce a known NP-hard problem to the problem whose NP-hardness we try to prove: namely, to the inverse problem for piecewise smooth velocity distributions.

Specifically, we will reduce, to our problem, the following *subset sum* problem [4], [5] that is known to be NP-hard:

- Given:
  - k positive integers $s_1, \ldots, s_k$ and
  - an integer $s > 0$,
- check whether it is possible to find a subset of this set of integers whose sum is equal to exactly s.

For each i, we can take $y_i = 0$ if we do not include the i-th integer in the subset, and $y_i = 1$ if we do. Then the subset problem takes the following form: check whether there exist values $y_i \in \{0, 1\}$ for which $\sum s_i \cdot y_i = s$.

We will reduce each instance of this problem to an instance of our optimization problem. Indeed, for each i, we can write that $\frac{b_i}{n_i} = \frac{b_i}{n'} + s_i \cdot y_i$, where we denoted

$$s_i \stackrel{\text{def}}{=} b_i \cdot \left(\frac{1}{n} - \frac{1}{n'}\right). \qquad (7)$$

Thus, if we are given the values s_i, we must choose

$$b_i = \frac{s_i \cdot n \cdot n'}{n' - n}. \qquad (8)$$

For this choice of s_i, we get

$$\sum_{i=1}^{k} \frac{b_i}{n_i} = \sum_{i=1}^{k} \frac{b_i}{n'} + \sum_{i=1}^{k} s_i \cdot y_i.$$

So, for any number s_0, the condition $\sum\limits_{i=1}^{k} s_i \cdot y_i \le s_0$ is equivalent to $\sum\limits_{i=1}^{k} \frac{b_i}{n_i} \le \sum\limits_{i=1}^{k} \frac{b_i}{n'} + s_0$. Therefore, if we take

$$\varepsilon_i \stackrel{\text{def}}{=} s_0 + \sum_{i=1}^{k} \frac{b_i}{n'} \qquad (9)$$

we will be able to conclude that $\sum\limits_{i=1}^{k} s_i \cdot y_i \le s_0$ if and only if $\sum\limits_{i=1}^{k} \frac{b_i}{n_i} \le \varepsilon_0$. So, with our choice of b_i and ε_0, the constraint in the resulting optimization problem becomes very similar to the condition in the subset problem – with the only difference that we have an inequality $\sum s_i \cdot y_i \le s_0$, while in the subset problem, we need equality.

To reduce inequality to equality, let us select appropriate coefficients c_i in the objective function. Indeed, in general, $n_i = n' - (n' - n) \cdot y_i$, hence for arbitrary c_i, we get

$$\sum_{i=1}^{k} c_i \cdot n_i = n' \cdot \sum_{i=1}^{k} c_i - \sum_{i=1}^{k} (n' - n) \cdot y_i.$$

So, if we select c_i in such a way that $c_i \cdot (n' - n) = s_i$, i.e., as

$$c_i \stackrel{\text{def}}{=} \frac{s_i}{n' - n}, \qquad (10)$$

then we conclude that

$$\sum_{i=1}^{k} c_i \cdot n_i = n' \cdot \sum_{i=1}^{k} c_i - \sum_{i=1}^{k} s_i \cdot y_i. \qquad (11)$$

Thus, the cost $\sum\limits_{i=1}^{k} c_i \cdot n_i + c_0 \cdot n'$ attains its smallest possible value if and only if the linear combination $\sum\limits_{i=1}^{k} s_i \cdot y_i$ attains its largest value. Since $\sum\limits_{i=1}^{k} s_i \cdot y_i \le s_0$, this largest possible value cannot exceed s_0, and the only possibility for it to attain the value s_0 is when the subset problem has a solution.

For simplicity, we can choose $c_0 = 0$. In this case, due to the formula (11), $\sum\limits_{i=1}^{k} s_i \cdot y_i = s_0$ if and only if the smallest possible value of the cost $\sum c_i \cdot n_i$ is equal to $n' \cdot \sum c_i - s_0$, i.e., if equal to

$$C_0 \stackrel{\text{def}}{=} n' \cdot \sum_{i=1}^{k} \frac{s_i}{n' - n} - s_0. \qquad (12)$$

Thus, to check whether an instance of the subset problem has a solution, it is sufficient to form the corresponding instance of our optimization problem and check whether its minimal cost is equal to C_0. If the minimal cost is indeed equal to C_0, this means that the original subset problem has a solution – actually the same values y_i that correspond to this cost provide the solution to the original instance of the subset problem. If the minimal cost is $> C_0$, this means that the given instance of the subset problem has no solution.

This reduction proves that our optimization problem is indeed NP-hard.

VIII. How This Problem Is Solved Now

In the previous section, we have shown that in general, the problem of finding the optimal sampling frequencies for weld reliability assessments of long pipeline segments is NP-hard. Crudely speaking, NP-hardness means that, in general, any algorithm that exactly solves all the instances of this problem requires, in some cases, computation time that grows exponentially with the number of inputs k. In other words, if we want to solve the problem exactly, then, most probably, we cannot find the optimal frequencies faster than by using exhaustive (or almost exhaustive) search.

At present, exhaustive (or almost exhaustive) search is example how this problem is solved in practice – by trying all possible combinations of n_i. This is very time-consuming, so we need faster algorithms for finding n_i.

IX. Uncertainty in the Input Makes Faster Algorithms Possible

In practice, the parameters b_i, c_i, and c_0 are only approximately known. Since we only know these parameters *approximately*, it does not make much sense to waste computational resources on solving the *exact* optimization problem – because the solution that is optimal for the given values of these parameters may be not exactly optimal for the (unknown) actual values of these quantities.

It is therefore reasonable, instead of looking for an optimal solution, to only look for an asymptotically optimal one.

X. New Algorithm: Motivations

One of the main reasons why our optimization problem is computationally difficult is because the desired values n_i are integers. It is well known (see, e.g., [5]) that such discrete optimization problems are much more difficult to solve than the similar continuous optimization problems, i.e., problems in which the values n_i can take arbitrary real values.

In view of this comment, a reasonable idea is to treat the original problem as a continuous optimization problem, find the optimal values n_i as real numbers, and then round off these values to the nearest integers to get an implementable testing schedule.

The resulting algorithm is not exact – it replaces each value n_i with the nearest integer which may differ from n_i by 0.5. Thus, the relative accuracy with which this algorithm returns the values is $0.5/n_i$. Hence, the relative difference between the cost corresponding to this selection and the optimal cost is also of the order $\sim 1/n_i$. Thus, when n_i increases, this relative accuracy decreases – in this sense, this algorithm is asymptotically optimal.

Let us therefore find real value n_i that optimize the objective function (6) under the constraint (5). The constraint is described by a function that smoothly (differentiably) depends on the unknowns n_i. If the objective function was also smooth, then we could use the Lagrange multiplier method to find the maximum. However, the objective function contains a non-smooth term $\max\limits_i n_i$, so we cannot directly use the Lagrange multiplier technique. Instead, we will use the following argument.

In the optimal solution, some values n_i are equal to the maximum $\max_i n_i$ and some are not. Let us first consider two unknowns n_j and n_l whose values are smaller than the maximum. Then, if we select sufficiently small changes Δn_j and Δn_l, the resulting values $n_j + \Delta n_j$ and $n_l + \Delta n_l$ are still smaller than $\max_i n_i$. We want to make changes after which the condition is still satisfied – i.e., that do not affect the left-hand side of the condition (5). For that, we need to make sure that

$$\frac{b_j}{n_j + \Delta n_j} + \frac{b_l}{n_l + \Delta n_l} = \frac{b_j}{n_j} + \frac{b_l}{n_l}.$$

For small Δn_j and Δn_l, this means that

$$-\frac{b_j}{n_j^2} \cdot \Delta n_j - \frac{b_l}{n_l^2} \cdot \Delta n_l + o(\Delta n_j) = 0,$$

so we can take an arbitrary small Δn_j and

$$\Delta n_l = -\Delta n_j \cdot \frac{b_j}{n_j^2} \cdot \frac{n_l^2}{b_l} + o(\Delta n_j).$$

Substituting the changed values of n_j and n_l into the objective function (6), we thus add, to the resulting cost, the value

$$c_j \cdot \Delta n_j + c_l \cdot \Delta n_l =$$

$$\Delta n_j \cdot \left(c_j - c_l \cdot \frac{b_j}{n_j^2} \cdot \frac{n_l^2}{b_l}\right) + o(\Delta n_j). \qquad (13)$$

The original vector $(n_1, \ldots, n_k)$ was optimal, so this change in the cost must be non-negative for all possible values Δn_j. Since the value Δn_j can be both (small) positive and (small) negative, the only way for the cost change to be always non-negative is when the coefficient at Δn_j is equal to 0 – otherwise,

- if this coefficient is positive, the change will be negative for $\Delta n_j < 0$, and
- if this coefficient is negative, the change will be negative for $\Delta n_j > 0$.

When this coefficient is equal to 0, it means that $c_j = c_l \cdot \dfrac{b_j}{n_j^2} \cdot \dfrac{n_l^2}{b_l}$. Moving all terms related to x_j to the left side and all other terms to the right, we conclude that $\dfrac{c_j \cdot n_j^2}{b_j} = \dfrac{c_l \cdot n_l^2}{b_l}$. This equality must be true for every two indices j and l for which $n_j, n_l < \max_i n_i$. Thus, for all such indices, the expression $\dfrac{c_j \cdot n_j^2}{b_j}$ attains the same value. Let us denote this common value by λ. Then, from the equation $\dfrac{c_j \cdot n_j^2}{b_j} = \lambda$, we conclude that

$$n_j = \sqrt{\lambda} \cdot \sqrt{\frac{b_j}{c_j}}. \qquad (14)$$

If $n_j < \max_i n_i$ and $n_l = \max_i n_i$, and if there are at least two different indices l for which $n_l = \max_i n_i$, then the condition that $\max_i n_i$ remains the same is only preserved when we decrease n_l, i.e., when $\Delta n_l < 0$. This requirement corresponds to $\Delta n_j > 0$. So, in this case, from the fact that the original cost was the smallest, we can only conclude that the change in cost is non-negative for $\Delta n_j > 0$. This means that the coefficient at Δn_j at the expression (13) must be non-negative, i.e., that $c_j \ge c_l \cdot \dfrac{b_j}{n_j} \cdot \dfrac{n_l^2}{b_l}$. Moving terms related to j into the left and all other terms into the right and taking into account that $\dfrac{c_j \cdot n_j^2}{b_j} = \lambda$, we conclude that $\dfrac{c_l \cdot n_l^2}{b_l} \le \lambda$, i.e., that $n_l \le \sqrt{\lambda} \cdot \sqrt{\dfrac{b_l}{c_l}}$.

In other words, for j for which $n_j < \max_i n_i$, we have $n_j = \sqrt{\lambda} \cdot \sqrt{\dfrac{b_j}{c_j}} < \max_i n_i$, and for l for which $n_l = \max n_i$, we get $n_l = \max_i n_i \le \sqrt{\lambda} \cdot \sqrt{\dfrac{b_l}{c_l}}$. Thus, for every two indices j and l for which $n_j < \max_i n_i$ and $n_l = \max_i n_i$, we have $\dfrac{b_j}{c_j} \le \dfrac{b_l}{c_l}$. Thus, there is a threshold t such that if $\dfrac{b_j}{c_j} \le t$, we get $n_j < \max_i n_i$ and if $\dfrac{b_l}{c_l} \ge t$, we get $n_l = \max_i n_i$.

So, as the first step of our algorithm we sort all the indices j in the increasing order of the ratio b_j/c_j. For notation simplicity, let us assume that these indices are already sorted in this manner. Then, we must select a threshold value t so that we will have $n_1, \ldots, n_{t-1}$ smaller than $\max_i n_i$ and $n_t = \ldots = n_k = \max_i n_i$.

Replacing n_j ($j < t$) with $n_j + \Delta n_j$ and all the values $n_t, \ldots$ with $n_l + \Delta n$, we can similarly conclude that

$$n_l = \sqrt{\lambda} \cdot \sqrt{\frac{\sum\limits_{l=t}^{k} c_l}{\sum\limits_{l=t}^{k} c_l + c_0}}. \tag{15}$$

Let us determine the value λ. Clearly, if we guarantee more accuracy than necessary, we can perform fewer measurements and still get the desired accuracy, Thus, the minimal cost is attained when $\sum\limits_i \dfrac{b_i}{n_i} = \varepsilon_0$. Substituting the above expressions (14) and (15) for n_j and n_l into this formula, we conclude that $\dfrac{1}{\sqrt{\lambda}} \cdot Z_t = \varepsilon_0$, where

$$Z_t \stackrel{\rm def}{=} \sum_{j-1}^{t-1} \sqrt{b_j \cdot c_j} + \sqrt{\sum_{l=t}^{k} b_l} \cdot \sqrt{\sum_{l=t}^{k} c_l + c_0}. \tag{16}$$

Hence, $\sqrt{\lambda} = Z_t/\varepsilon_0$. For these values n_i, the overall cost $\sum c_j \cdot n_j + (c_0 + \sum c_l) \cdot n_l$ takes the form $\sqrt{\lambda} \cdot Z_t$, i.e., the form Z_t^2/ε_0. Thus, to find the smallest possible cost, we must find t for which Z_t is the smallest.

One can check that the same formulas work also when we only have one value n_l for which $n_l = \max_i n_i$. Thus, we arrive at the following algorithm:

XI. Resulting Algorithm

- First, we sort all the indices in the increasing order of the ratio b_i/c_i. Sorting requires time $O(k \cdot \log(k))$.
- Then, for every t from 1 to k, we compute Z_t by using the formula (16). When we move from Z_t to Z_{t+1}, each sum changes by only one term, so we only need a constant number of terms to find each of k values Z_t – to the total of $O(k)$.
- We find t for which Z_t is the smallest. For this t, we compute $\sqrt{\lambda} = Z_t/\varepsilon_0$, and the find the optimal n_i by using the formula (14) for $i < t$ and the formula (15) for $i \ge t$.

This algorithm requires computation time $O(k \cdot \log(k)) + O(k) = O(k \cdot \log(k))$.

Acknowledgments

This work was supported in part by NASA under cooperative agreement NCC5-209, by NSF grants EAR-0225670 and DMS-0532645, Star Award from the University of Texas System, and by Texas Department of Transportation grant No. 0-5453.

References

[1] A. H.-S. Ang and W. H. Tang, *Probability Concepts in Engineering Planning and Design*, Vol. II, J. Wiley and Sons, New York, 1990.

[2] J. Brown and T. Cruse, "Efficiently Reducing Statistical Uncertainty Through a Sample Sensitivity Index," *Proc. 43rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference*, AIAA-2002-1275, Denver, CO, April 22–25, 2002.

[3] B. Gross, T. S. Connelly, H. Van Dijk, and A. Gilroy-Scott, "Flaw Sizing Using Mechanized Ultrasonic Inspection on Pipeline Girth Welds," *Proc. ICAWT'99: Pipeline Welding and Technology*, EWI, October 1999.

[4] V. Kreinovich, A. Lakeyev, J. Rohn, and P. Kahl, *Computational complexity and feasibility of data processing and interval computations*, Kluwer, Dordrecht, 1997.

[5] C. H. Papadimitriou and K. Steiglitz, *Combinatorial Optimization: Algorithms and Complexity*, Dover Publications, Inc., Mineola, New York, 1998.

[6] R. Rackwitz and B. Fiessler, "Structural Reliability Under Combined Random Load Sequences," *Computers and Structures*, 1978, Vol. 9, pp. 489–494.

[7] B. H. Thacker, C. F. Popelar, and R. C. McClung, "A New Probabilistic Method for Predicting the Long Life Reliability of Pipelines," In: R. W. Revie and K. C. Wang (eds.) *Proc. Int. Conf. on Pipeline Reliability*, Calgary, Alberta, Canada, June 2–5, 1992.

[8] T. Y. Torng and B. H. Thacker, "An Efficient Probabilistic Scheme for Constructing Structural Reliability Confidence Bounds," *Proc. 34th AIAA/ASME/ASCE/AHS/ASC SDM Structures, Structural Dynamics, and Materials Conf.*, Paper No. AIAA-93-1627, La Jolla, CA, April 19–21, 1993.

[9] R. Trejo and V. Kreinovich, "Error Estimations for Indirect Measurements: Randomized vs. Deterministic Algorithms For 'Black-Box' Programs", In: S. Rajasekaran, P. Pardalos, J. Reif, and J. Rolim (eds.), *Handbook on Randomized Computing*, Kluwer, 2001, pp. 673–729.

[10] H. M. Wadsworth Jr., *Handbook of Statistical Methods for Engineers and Scientists*. McGraw-Hill, N.Y., 1990.

[11] M. J. Wagner and B. M. Patchett, "Girth Weld Defects in Mechanized GMA Field-Welded Pipelines," *Welding Journal*, 1991, Vol. 70, No. 6.

[12] R. W. Warke, K. C. Koppenhoefer, and W. E. Amend, "Case Study in Probabilistic Assessment: Seismic Integrity of Girth Welds in a Pre-World War II Pipeline," *Proc. ICAWT'99: Pipeline Welding and Technology*, EWI, October 1999.

[13] R. W. Warke, Y.-Y. Wang, C. M. Ferregut, C. J. Carrasco, and D. J. Horseley, "A FAD-based method for probabilistic flaw assessment of strength-mismatched girth welds", *Proc. 1999 Pressure Vessels and Piping Conference PVP'99 "Probabilistic and Environmental Aspects of Fracture and Fatigue*, American Sociery of Mechanical Engineers (ASME), Vol. PVP-386, 1999.

[14] P. H. Wirsching, H. P. Nguyen, D. Osage, and A. E. Mansour, "Probability Based fitness for Service for Pressure Vessels and Piping," *Proc. 8th ASCE Specialty Conf. On Probabilistic Mechanics and Structural Reliability*, PMC2000-048, 2000.