A Method to Determine Strict Gaussian Bounds of a Sample Distribution

Juan Blanch, Todd Walter, Per Enge
Stanford University

ABSTRACT

The integrity analysis of systems like Satellite-Based Augmentation Systems (SBAS), Ground-based Augmentation Systems (SBAS), Receiver Autonomous Integrity Monitoring (RAIM), or Advanced RAIM (ARAIM) requires proving that the actual distribution of errors can be replaced by a simpler distribution, often a gaussian distribution. Two key results have been used to generate these overbounding distributions: cdf bounding and paired overbounding. Although these two results are very powerful, each of them has a weakness that limits their direct application.

The goal of this paper is to overcome some of these weaknesses. We will first present a method to determine gaussian overbounding distributions that combines cdf bounding and paired overbounding. This method is based on the determination of an intermediate overbounding distribution that is symmetric and unimodal. Then, we present a modification of the paired overbounding theorem that relaxes the cdf paired overbounding requirement mentioned above. These techniques are the basis of a MATLAB toolset that computes rigorous gaussian overbounding distributions for any sample distribution.

INTRODUCTION

The proof of safety of GNSS integrity systems like Satellite-Based Augmentation Systems (SBAS), Ground-based Augmentation Systems (GBAS), Receiver Autonomous Integrity Monitoring (RAIM), or Advanced RAIM (ARAIM) requires proving that an empirical distribution of errors, which is assumed to be the expected distribution, can be replaced in the analysis by a simpler distribution. These errors are for example the contributors to the pseudorange error: code noise and multipath, clock and ephemeris errors, or residual tropospheric error. This simpler distribution, sometimes called the overbounding distribution, must be such that the computed user integrity risk is larger than if it had been computed using the expected distribution.

Replacing the expected distribution by the overbounding distribution is a necessary step for at least two reasons. First, an arbitrary distribution would require an amount of data too large to be sent through the typically low bandwidth channels available for augmentation systems. Second, and perhaps more critically, it would be very difficult at the user level to compute the integrity risk using a set of empirical distribution for each of the error components. When the position solution is computed by the user, all the pseudorange errors are combined linearly to form the position error, which is what the user is interested in. The user therefore needs to characterize the convolution of the different error sources. Computing the convolution of up to a hundred error sources characterized by arbitrary empirical distributions is likely to be prohibitive, even with the potentially increased computational power available to users. That is why

replacing these distributions by simpler distributions is essential. Because the only finite variance distribution that is stable through convolution is the normal distribution, it is almost unavoidable to use it as the basis for the overbounding distribution. In addition, several of the empirical distributions occurring in augmentation systems are well approximated by a gaussian, at least at the core of the distribution.

The problem described above is not new and several techniques have been developed to treat it, especially for SBAS and GBAS. There are in particular two key results: cdf bounding and paired overbounding [1,2]. These two results are very powerful and have been key in several safety of life systems. However, each of these results has a weakness that limits their direct application. The first one only applies to symmetric unimodal distributions. This is a weakness because empirical distributions are never symmetric nor unimodal: an empirical distribution is a series of delta impulses set at each of the samples. For the second one, it is required that the overbounding cdf bounds the empirical distribution for all values. It turns out that this requirement is very stringent.

After stating the two above results and their limitations, we first describe a new method to determine gaussian overbounding distributions that combines both of them. Second, we present a modification of the paired overbounding theorem [2] that relaxes the cdf paired overbounding requirements mentioned above. Finally, we describe the MATLAB set of scripts that implements these two techniques and apply these scripts to the determination of an overbounding gaussian distribution of the GPS clock and ephemeris error distribution (to support the determination of the ARAIM Integrity Support Message).

PREVIOUS WORK

Cdf overbounding for unimodal and symmetric distributions (DeCleene, 2000)

The first result, [1], applies to symmetric and unimodal distributions. It states that if f , f_{ob}, and g are zero mean and unimodal distributions, and if:

$$\int_{x}^{+\infty} f \le \int_{x}^{+\infty} f_{ob} \quad \text{for} \quad x \ge 0 \tag{1}$$

Then:

$$\int_{L}^{+\infty} f * g \le \int_{L}^{+\infty} f_{ob} * g \quad \text{for} \quad L \ge 0 \tag{2}$$

Paired overbounding (Rife, 2006)

The second one, first described in [2], states that if:

$$\int_{x}^{+\infty} f \le \int_{x}^{+\infty} f_{ob} \quad \text{for any } x \tag{3}$$

Then:

$$\int_{x}^{+\infty} f * g \le \int_{x}^{+\infty} f_{ob} * g \quad \text{for any } x \tag{4}$$

The second result applies to any distribution.

Limitations

These two results are very powerful and have been key for the proof of integrity of several safety of life systems. However, each of these results has a weakness that limits its direct application.

The first one only applies to symmetric unimodal distributions. This is a weakness because empirical distributions are never symmetric nor unimodal (an empirical distribution is a series of delta impulses set at each of the samples). This limitation can be bypassed by asserting that the underlying true distribution is symmetric and unimodal. It is however not always obvious to justify this assertion.

For the second one, it turns out that the requirement (3) is very stringent. For example, using this criterion, if $\sigma_1 < \sigma_2$, $N(0,\sigma_1)$ is not bounded by $N(0,\sigma_2)$. Also, it is strictly impossible to fulfil requirement (3) if f is an empirical distribution and f_{ob} is a gaussian. This is due to the fact that the tails of the empirical distribution are finite, such that there exists x_0 such that:

$$\int_{x_0}^{+\infty} f = 1 \tag{5}$$

But for a gaussian, we always have:

$$\int_{x_0}^{+\infty} f_{ob} < 1 \tag{6}$$

The two contributions proposed below mitigate these weaknesses.

COMBINING CDF BOUNDING AND PAIRED OVERBOUNDING

We first describe a method that combines the two previous results to produce a gaussian overbound of the expected distribution. This method does not require the sample distribution to be symmetric or unimodal, and it does not require the final gaussian overbound to fulfil condition (3). The idea consists in introducing an intermediate overbounding distribution. There are two steps: determining a unimodal and symmetric (about its mean) distribution F_s that is a cdf overbound of the empirical distribution, and determining a gaussian overbound $N(\beta,\sigma)$ of the distribution F_s and apply DeCleene's result on overbounding. The method is applied independently to each side.

Determining a unimodal and symmetric overbounding distribution

In this first step, we determine a unimodal and symmetric distribution F_s such that:

$$F(x) \le F_s(x) \text{ for any } x \tag{7}$$

There are many ways of generating such a distribution. Here we used the following process:

1) bin the data of F in equal sized bins.
2) modify the distribution defined by the binned data by forcing unimodality for the bins above the median. This is done by modifying the bins one by one from the right hand side by making sure that each is equal or larger than the previous one.
3) stop when the top half of the distribution is defined. Each side of the distribution is defined by forcing the symmetry around the mean. The width of the middle bin is chosen to make sure that the distribution is unimodal
4) within each bin, assume a uniform distribution. We found it was practical to define a "regularized" set of samples that approximates the uniform distribution within each bin using the same number of samples. We label these points x_i from i = 0 to K-*1*, where K is the number of samples.
5) modify F_s so that that (3) holds for $x \geq M$, where M is a tunable parameter. The default is the m, mean of F_s. This modification can be done either by shifting F_s to the right or shrinking it, or a combination of both.

Figure 1 shows as an example the result of this process for a random variable drawn from the pdf formed of two gaussians with mean 2 and -2.

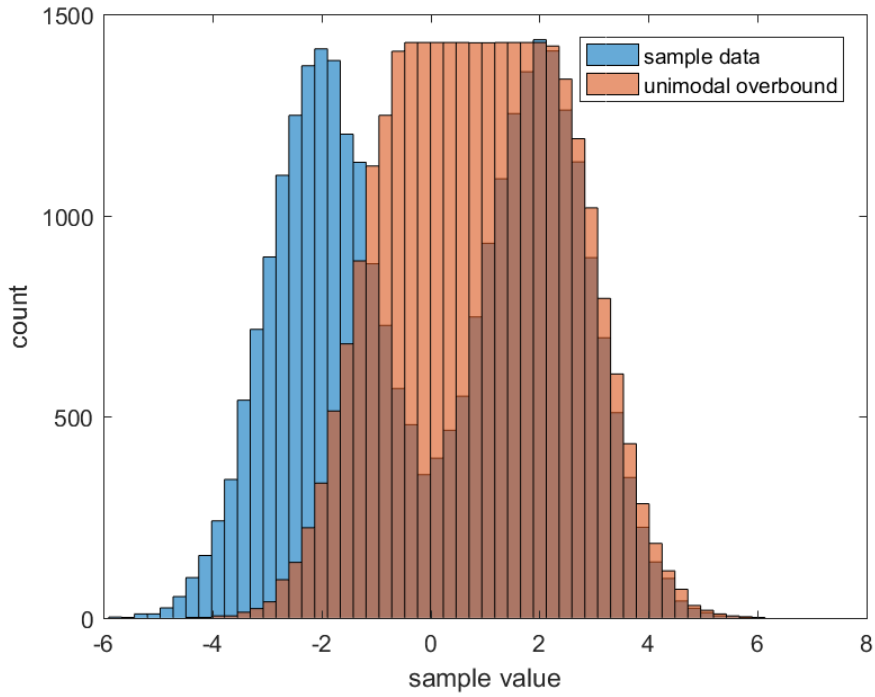

Figure 1. Symmetric and unimodal right hand side overbound of a bimodal distribution

Determining a gaussian overbounding distribution

The second step consists on finding σ such that:

$$F_s(x) \le \frac{1}{\sqrt{2\pi}} \int_{x-m}^{+\infty} e^{-\frac{t^2}{2\sigma^2}} dt \quad \text{for any } x \ge m \tag{8}$$

If the two conditions (7) and (8) are met, the gaussian distribution *N(m, σ)* will be an overbound for any interval of the form [*x*,+∞[where *x* is above the mean. The distribution F_s, as determined in the first part of the process, is piecewise uniform. Appendix B describes how to rigorously obtain *σ* using a half interval search.

PAIRED OVERBOUNDING RELAXATION

In some cases, it may be the case that condition (7) is only achievable by considerably modifying F_s, in which case the overbounding distribution could end up being too conservative. This section describes a modification of the paired overbounding theorem that relaxes this requirement.

We consider two independent random variables *X* and *Y*, with probability density functions *f* an *g*. We note:

$$F(x) = \int_x^{+\infty} f(t)\,dt \tag{9}$$

$$G(x) = \int_x^{+\infty} g(t)\,dt \tag{10}$$

Result 1

For any density function f_{ob} such that:

$$\forall x \ge 0 \quad F(x) \le F_{ob}(x) \tag{11}$$

and

$$\forall x \le 0 \quad F(x) - F_{ob}(x) \le \varepsilon \tag{12}$$

where

$$F_{ob}(x) = \int_x^{+\infty} f_{ob}(u)\,du \tag{13}$$

We have:

$$H(x) = \int_x^{+\infty} f * g = \int_{-\infty}^{+\infty} g(x-u)F(u)\,du \le \int_{-\infty}^{+\infty} g(x-u)F_{ob}(u)\,du + \varepsilon G(x) \tag{14}$$

Proof:

We write:

$$H(x) = \int_{-\infty}^{+\infty} g(x-u)F(u)du$$
$$= \int_{-\infty}^{+\infty} g(x-u)F_{ob}(u)du + \int_{-\infty}^{0} g(x-u)\left(F(u)-F_{ob}(u)\right)du + \int_{0}^{+\infty} g(x-u)\left(F(u)-F_{ob}(u)\right)du \quad (15)$$

The last term is negative because of (11). In the second term, we use (12) to show:

$$\int_{-\infty}^{0} g(x-u)\left(F(u)-F_{ob}(u)\right)du \le \varepsilon G(x)$$

Result 2:

Let us consider n distributions with pdf f_i, and n distributions with pdf $f_{i,ob}$, and their respective right hand cdf, F_i,and $F_{i,ob}$, such that:

$$\forall x \ge 0 \quad F_i(x) \le F_{i,ob}(x)$$
$$\forall x \le 0 \quad F_i(x) - F_{i,ob}(x) \le \varepsilon_i \quad (16)$$

Then we have:

$$\forall x \ge 0 \quad \int_{x}^{+\infty} f_1 * \cdots * f_n \le \prod_{i=1}^{n}(1+\varepsilon_i) \int_{x}^{+\infty} f_{1,ob} * \cdots * f_{n,ob} \quad (17)$$

As long as:

$$\forall i_1, \cdots, i_k \quad i_k \ne i_{k'}$$
$$\int_{x}^{+\infty} f_{i_1,ob} * \cdots * f_{i_k,ob} \le \int_{x}^{+\infty} f_{1,ob} * \cdots * f_{n,ob} \quad (18)$$

This last condition is trivial in the case of a gaussian for x larger than the mean of $f_{1,ob} * \cdots * f_{n,ob}$.

The inflation in probability resulting from (17) can either be neglected (if it is considered small enough), or taken into account by inflating the final error bound.

Proof: the result is obtained by applying result 1 repeatedly and using the last condition

MATLAB SCRIPTS AND AN EXAMPLE OF APPLICATION TO ARAIM INTEGRITY SUPPORT MESSAGE PARAMETERS

List of functions

gaussian_overbound.m: main function which takes as an argument the samples to be bounded, and the bin size used for the symmetrization and "unimodalization". This function outputs the mean and the sigma of the gaussian distribution that overbounds the sample distribution, as well as the bound on the difference between the cdf for negative values. A positive value means that the final probability needs to be multiplied by $(1+\varepsilon)$.

symmetric_overbound.m: function that determines the intermediate unimodal and symmetric distribution described above. It implements the process described between Equations (7) and (8). It takes as argument the sample distribution, the bin size used for the symmetrization and unimodalization, and the allowable probability inflation. It is used within gaussian_overbound.m

evaluate_sigma.m: determines the standard deviation of the overbounding sigma. It is based on the conditions described in Appendix B. It is used within gaussian_overbound.m.

These scripts will be made available at: gps.stanford.edu

Application to GPS clock and ephemeris errors

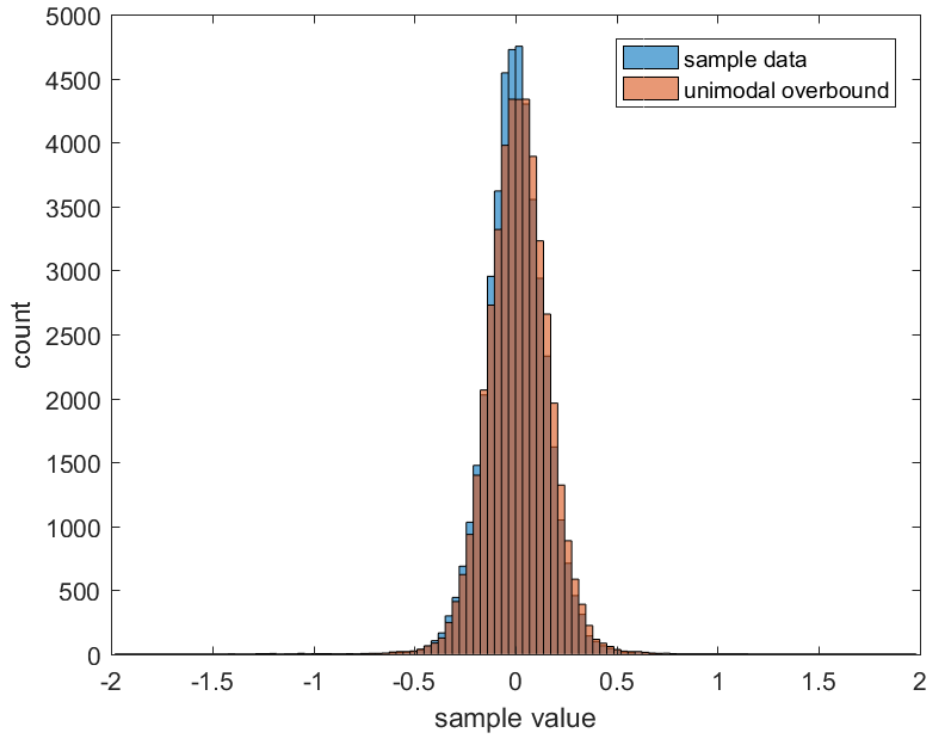

Figure 2. *Histogram of sample data and unimodal and symmetric right hand overbound*

We now apply the function gaussian_overbound.m to the clock and ephemeris GPS residuals obtained using the process described in [3]. This process produces a sample distribution for each satellite and each user (from a set of users covering the globe). Figure 2 shows a histogram of the clock and ephemeris GPS residuals normalized by the URA for SVN 41 from January 2008 to December 2016 for a user located at 39° N, 45° W (blue histogram). This is the distribution that needs to be bounded by a gaussian distribution. Figure 2 also shows the result of the unimodalization and symmetrization step for the right hand side (after step 4) above).

Figure 3 shows the corresponding cdfs. The difference at the tails between the unimodal overbound and the sample distribution is due to the irregularity of the sample distribution at the tails.

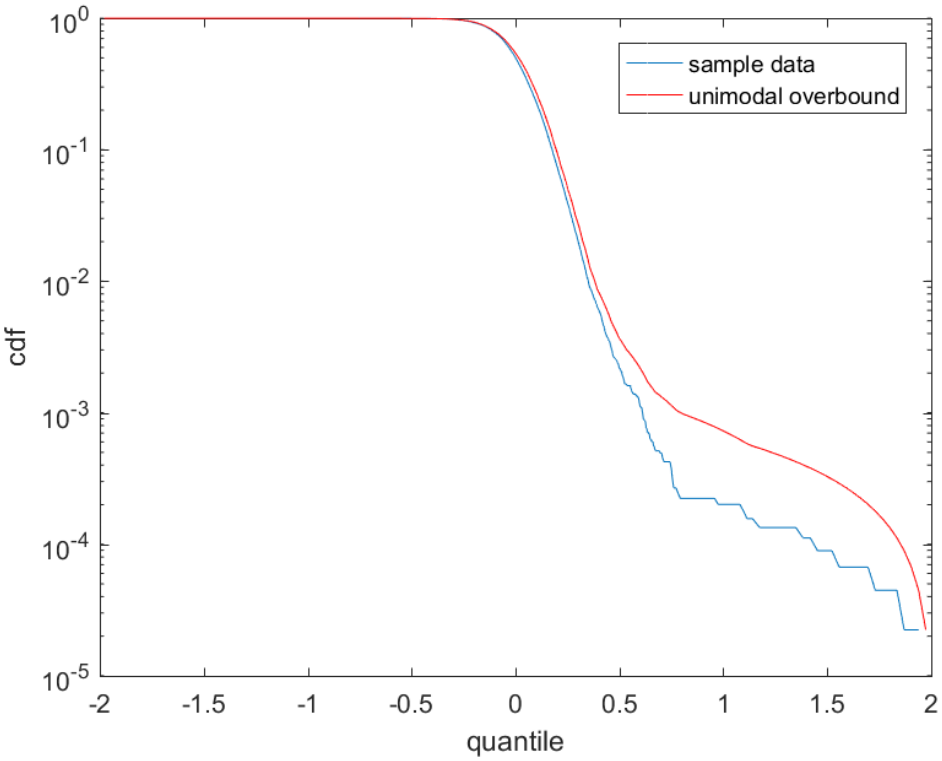

Figure 3. *Cdfs of sample distribution and symmetric and unimodal overbound*

The mean of the final overbounding distribution will be the mean (and median) of the symmetric and unimodal overbound F_s. For the data sample shown in Figure 2, the mean is equal to 0.015. In the last step, we determine the minimum standard deviation of a gaussian overbound, which is equal to 0.49. As mentioned above, the algorithm also outputs the inflation of the final probability, which is in this case is equal to 0.0009, and therefore negligible.

Note: The size of this inflation can be changed through the parameter *M*. For the above results, we used the default.

SUMMARY

There are three contributions in this paper. First, we have presented a modification of the paired overbounding theorem that relaxes the cdf paired overbounding requirement. Second, we have presented a method to determine gaussian overbounding distributions that combines cdf bounding and paired overbounding. This combination removes important limitations in the two main results used in overbounding. It is based on the determination of an intermediate overbounding distribution that is symmetric and unimodal, but not gaussian. Finally, these techniques are the basis of a MATLAB set of scripts and functions that computes rigorous gaussian overbounding distributions for any sample distribution.

REFERENCES

[1] DeCleene, Bruce, "Defining Pseudorange Integrity - Overbounding," *Proceedings of the 13th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GPS 2000)*, Salt Lake City, UT, September 2000, pp. 1916-1924.

[2] Rife, J., Pullen, S., Pervan, B., and Enge, P. Paired Overbounding for Nonideal LAAS and WAAS Error Distributions. IEEE Transactions on Aerospace and Electronic Systems, 2006, 42, 4, 1386 -1395.

[3] Walter, Todd, Blanch, Juan, "Characterization of GNSS Clock and Ephemeris Errors to Support ARAIM," *Proceedings of the ION 2015 Pacific PNT Meeting*, Honolulu, Hawaii, April 2015, pp. 920-931.

APPENDIX A: PROOF FOR THE COMBINED CDF AND PAIR BOUNDING

We consider n probability distributions defined by their pdf f_k from k=1 to n. For each k we consider a pdf $f_{k,su}$ that is a right side cdf overbound of f_k,:

$$\int_x^{+\infty} f_k \le \int_x^{+\infty} f_{k,su} \quad \text{for any } x \tag{19}$$

Using the results on paired bounding in [2], we have:

$$\int_x^{+\infty} f_1 * \cdots * f_n \le \int_x^{+\infty} f_{1,su} * \cdots * f_{n,su} \quad \text{for any } x \tag{20}$$

Now, let us further assume that each pdf f_k is unimodal and symmetric about its mean m_k.

In the next step, for each k, we consider a second probability distribution $f_{k,ob}$ with mean m_k that is also unimodal and symmetric about its mean. Let us suppose that:

$$\int_x^{+\infty} f_{k,su} \le \int_x^{+\infty} f_{k,ob} \text{ for } x \ge m_k \tag{21}$$

Using the results on cdf bounding from [1], we have:

$$\int_x^{+\infty} f_{1,su} * \cdots * f_{n,su} \le \int_x^{+\infty} f_{1,ob} * \cdots * f_{n,ob} \text{ for any } x \ge \sum_{k=1}^{n} m_k \tag{22}$$

Typically, $f_{k,ob}$ will be a gaussian distribution. Combining the results (20) and (22), we get:

$$\int_x^{+\infty} f_1 * \cdots * f_n \le \int_x^{+\infty} f_{1,ob} * \cdots * f_{n,ob} \text{ for any } x \ge \sum_{k=1}^{n} m_k \tag{23}$$

The second step might seem unnecessary, as we have replaced a symmetric and unimodal overbound by another one. The key is that the requirements on the second one are significantly weaker: the overbound only needs to apply for one side of the distribution.

APPENDIX B: GAUSSIAN OVERBOUND OF A PIECEWISE LINEAR CDF

In what follows we have defined:

$$P(x) = F_s(x - m) \tag{24}$$

The distribution defined by the cdf *P(x)* is zero mean, symmetric, unimodal, and piecewise uniform.

Gaussian cdf overbound of a piecewise uniform pdf

We consider the interval $[x_1,x_2]$. The cdf of a piececewise uniform distribution over each uniform interval is given by:

$$P(x) = P(x_1) + \frac{P(x_2) - P(x_1)}{x_2 - x_1}(x - x_1) \tag{25}$$

The right hand cdf of $N(0,\sigma)$ is given by:

$$Q\left(\frac{x}{\sigma}\right) = \int_x^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{u^2}{2\sigma^2}} du \tag{26}$$

This function is convex for *x* positive. Therefore, it is bounded by its tangent. The equation for the tangent at point x_0 is:

$$y = Q\left(\frac{x_0}{\sigma}\right) - \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x_0^2}{2\sigma^2}} \left(x - x_0\right) \tag{27}$$

We therefore have:

$$Q\left(\frac{x_0}{\sigma}\right) - \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x_0^2}{2\sigma^2}} \left(x - x_0\right) \le Q\left(\frac{x}{\sigma}\right) \text{ for } x \ge 0 \tag{28}$$

For $x_0 = \frac{x_1 + x_2}{2}$ this gives:

$$Q\left(\frac{x_1 + x_2}{2\sigma}\right) - \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\left(\frac{x_1+x_2}{2}\right)^2}{2\sigma^2}} \left(x - \frac{x_1 + x_2}{2}\right) \le Q\left(\frac{x}{\sigma}\right) \tag{29}$$

A sufficient condition to have:

$$P(x) \le Q\left(\frac{x}{\sigma}\right) \text{ for } x \in [x_1, x_2] \tag{30}$$

is therefore given by:

$$P(x_1) \le Q\left(\frac{x_1 + x_2}{2\sigma}\right) - \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\left(\frac{x_1+x_2}{2}\right)^2}{2\sigma^2}} \left(\frac{x_1 - x_2}{2}\right) \tag{31}$$

And

$$P(x_2) \le Q\left(\frac{x_1 + x_2}{2\sigma}\right) - \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\left(\frac{x_1+x_2}{2}\right)^2}{2\sigma^2}} \left(\frac{x_2 - x_1}{2}\right) \tag{32}$$

Initial points for the search of the bounding gaussian sigma

The search for the bounding sigma is done using a half-interval search. Such a search requires an upper bound and a lower bound on the solution. An upper bound is given by:

$$\sigma_{\min} = \frac{x}{Q^{-1}\left(P(x)\right)} \text{ for any } x > 0 \tag{33}$$

In the code, we take x to be the lower bound of the last interval where $P(x)$ is non zero. An upper bound is obtained be computing the overbound of a uniform pdf between $-x_{max}$ and x_{max} where x_{max} is the upper bound of the last interval where $P(x)$ is non zero:

$$\sigma_{\max} = \frac{1}{\sqrt{2}p_u} \text{ where } p_u = \frac{1}{2x_{\max}} \tag{34}$$