Signal Processing 127 (2016) 185–190

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Short communication

Mean and variance of round off error

Rui Li, Saralees Nadarajah *

School of Mathematics, University of Manchester, Manchester M13 9PL, UK

A R T I C L E I N F O

Article history:
Received 6 November 2015
Received in revised form
25 February 2016
Accepted 7 March 2016
Available online 14 March 2016

Keywords:
Continuous random variable
Finite interval
Floor
Real line

A B S T R A C T

Gadzhiev [4] derived expressions for round off error mean and round off error variance when the rounded variable follows the centered uniform and centered Gaussian distributions. Here, we derive general expressions for round off error mean and round off error variance when the rounded variable is any continuous random variable on the real line or any continuous random variable over a finite interval. Numerical studies are given.

1. Introduction

Round off errors arise in many areas of signal processing: wave digital-filters (Ullrich [14]); fast state-space decimator (Zeman and Lindgren [19]); polynomial FIR predictors and predictive FIR differentiators (Tanskanen and Dimitrov [13]); state-space digital filters (Lu and Hinamoto [10], Hinamoto et al. [6]); block floating-point treatment to the LMS algorithm (Mitra et al. [11]); efficient digital filter structures (Zhao and Li [20]); normal realizations of digital filters (He et al. [5]); recursive digital frequency synthesis (Kountouris [8]); digital filters (Li et al. [9]); and so on.

Two important measures of round off errors are their mean and variance. Gadzhiev [4] derived expressions for these measures by assuming that the rounded variable say X follows the centered uniform and centered Gaussian distributions.

The aim of this note is to extend the work of Gadzhiev [4] for any continuous random variable X defined on either the real line or a finite interval. The derived expressions for mean and variance are general and yet simple. They involve only $E(X)$, $\mathrm{Var}(X)$, the cumulative distribution function of X and another functional. Simple computer programs have been written by the authors that implement the derived expressions for any continuous random variable X. The programs can be obtained from the corresponding author.

Various distributions have been used to model roundoff errors in the signal processing area. Some evidence of experiments based on the Gaussian, uniform, triangular, and trapezoidal distributions are:

- Press [12] considers “the accumulation of round-off error in a floating-point digital filter”. He assumes that “the error committed at each arithmetic operation is an independent random variable uniformly distributed in $(-2-t, 2t)$ where t is the length of the mantissa”.
- Barnes et al. [2] consider “roundoff error after fixed-point multiplication”. They show that “if the multiplier coefficient is expressed as $a = N/M$, where M is a positive integral power of two and N is an odd integer, then the errors generated by roundoff after multiplication can generally be modeled as uniformly distributed, white, and uncorrelated with the signal, if the signal has sufficiently wide bandwidth and has a dynamic range that extends over approximately M quantum steps”.
- Ardalan and Alexander [1] present a “fixed-point roundoff error analysis of the exponentially windowed RLS algorithm” by assuming that “the input signal is a white Gaussian random process”.
- Kawarai and Murakami [7] while studying the “roundoff errors in floating-point arithmetic” and presenting “an optimization procedure for cascade floating-point digital filters”, apply “the isosceles trapezoidal distribution to the error analysis of cascade floating-point digital filters” and obtain “results that are in extremely close agreement with computer simulation values”.
- Wong [17] considers “the characteristics of the error resulting when a continuous amplitude signal x_n is quantized and then multiplied by a constant multiplier under fixed-point roundoff arithmetic”. It is shown that “regardless of the probability distribution of the input signal x_n, it is always possible to add a suitable dither signal to the input of the system so that both the quantization error and the roundoff error are uniformly distributed, white, and mutually uncorrelated”.
- Vladimirov and Diamond [15] show the validity of the uniform

* Corresponding author.

http://dx.doi.org/10.1016/j.sigpro.2016.03.007
0165-1684/

distribution for fixed-point roundoff noise “in digital implementation of linear systems arising due to overflow, quantization of coefficients and input signals, and arithmetical errors”.
- Csordas et al. [3] consider dithering “a known method for increasing the precision of analog-to-digital conversion”. They investigate and compare the effects “on bias and variance using uniform and triangularly distributed digital dithers”.
- Yu and Lim [18] present “an analysis for the roundoff noise of signal represented using a limited number of signed power-of-two terms”. Their analysis estimates “roundoff errors for Gaussian distributed inputs”.

In each of these experiments, the derived formulas for the mean and variance can be used to provide basic measures of roundoff error.

Furthermore, the excellent book by Widrow and Kollar [16] mentions the following: the Gaussian distribution, uniform distribution, the triangular distribution, the sinusoidal distribution, the convolution of triangular and uniform distributions and the convolution of triangular and triangular distributions can be used as models for dithers (see pages 497, 500, 501, 502, 503); the Gaussian, uniform and triangular distributions can be used as models for digital dithers (see pages 689–694); the Gaussian distribution can be used as models for quantizer input (see pages 84, 85, 106, 107, 123, 124, 125, 206, 207, 225–254, 411). In these cases too, the derived formulas can be used to provide basic measures of the associated roundoff error.

The contents of this note are organized as follows: four theorems deriving expressions for the mean and variance of $X - \text{floor}(X)$ and $X - \text{floor}\left(X + \frac{1}{2}\right)$ are given in Section 2; their proofs are given in the appendix; two numerical studies showing the use of the theorems and checking correctness of their derivations are given in Section 3.

2. Main results

Our main results are Theorems 1–4. Theorems 1 and 2 derive the mean and variance of $X - \text{floor}(X)$ and $X - \text{floor}\left(X + \frac{1}{2}\right)$ when X is a continuous random variable on the domain $(-\infty, \infty)$. Theorems 3 and 4 derive the mean and variance of $X - \text{floor}(X)$ and $X - \text{floor}\left(X + \frac{1}{2}\right)$ when X is a continuous random variable on the domain (a, b) for $-\infty < a < b < \infty$.

Throughout, we use the following notations: for a random variable X with probability density function (pdf) f and cumulative distribution function (cdf) F, define

$$F^*(y) = F\left(y - \frac{1}{2}\right),$$

$$M(x) = \int_{-\infty}^{x} z f(z)\, dz,$$

$$M^*(y) = M\left(y - \frac{1}{2}\right) + \frac{1}{2} F\left(y - \frac{1}{2}\right).$$

Note that $F^*(y)$ is the cdf of $Y = X + \frac{1}{2}$ at y and that $M^*(y)$ is the M function of $Y = X + \frac{1}{2}$ at y.

Theorem 1. *Let X be a continuous random variable on the domain $(-\infty, \infty)$ with pdf f and cdf F. Then*

$$E\left[X - \text{floor}(X)\right] = E[X] - \sum_{k=-\infty}^{\infty} k\left[F(k+1) - F(k)\right]$$

and

$$\text{Var}\left[X - \text{floor}(X)\right] = E\left[X^2\right] - 2\sum_{k=-\infty}^{\infty} k\left[M(k+1) - M(k)\right] + \sum_{k=-\infty}^{\infty} k^2\left[F(k+1) - F(k)\right] - \{E\left[X - \text{floor}(X)\right]\}^2.$$

Theorem 2. *Let X be a continuous random variable on the domain $(-\infty, \infty)$ with pdf f and cdf F. Then*

$$E\left[X - \text{floor}\left(X + \frac{1}{2}\right)\right] = E[X] - \sum_{k=-\infty}^{\infty} k\left[F^*(k+1) - F^*(k)\right]$$

and

$$\text{Var}\left[X - \text{floor}\left(X + \frac{1}{2}\right)\right] = E\left[X^2\right] + E\left[X\right] + \frac{1}{4} - 2\sum_{k=-\infty}^{\infty} k\left[M^*(k+1) - M^*(k)\right] + \sum_{k=-\infty}^{\infty} k^2\left[F^*(k+1) - F^*(k)\right] - \{E\left[X - \text{floor}\left(X + \frac{1}{2}\right)\right] + \frac{1}{2}\}^2.$$

Theorem 3. *Let X be a continuous random variable on the domain (a, b) for $-\infty < a < b < \infty$ with pdf f and cdf F. Then*

$$E\left[X - \text{floor}(X)\right] = E\left[X\right] - q + \sum_{k=p+1}^{q} F(k)$$

and

$$\text{Var}\left[X - \text{floor}(X)\right] = E\left[X^2\right] - 2qE\left[X\right] + 2\sum_{k=p+1}^{q} M(k) + q^2 - 2\sum_{k=p+1}^{q} kF(k) + \sum_{k=p+1}^{q} F(k) - \{E\{X - \text{floor}(X)\}\}^2,$$

where $p = \text{floor}(a)$ *and* $q = \text{ceiling}(b)$.

Theorem 4. *Let X be a continuous random variable on the domain (a, b) for $-\infty < a < b < \infty$ with pdf f and cdf F. Then*

$$E\left[X - \text{floor}\left(X + \frac{1}{2}\right)\right] = E\left[X\right] - q + \sum_{k=p+1}^{q} F^*(k)$$

and

$$\text{Var}\left[X - \text{floor}\left(X + \frac{1}{2}\right)\right] = E\left[X^2\right] + (1 - 2q)E\left[X\right] + \frac{(1-2q)^2}{4} + 2\sum_{k=p+1}^{q} M^*(k) - 2\sum_{k=p+1}^{q} kF^*(k) + \sum_{k=p+1}^{q} F^*(k) - \left\{E[X - \text{floor}\left(X + \frac{1}{2}\right)] + \frac{1}{2}\right\}^2,$$