A PIECEWISE LINEAR APPROXIMATION FOR PROBABILITY DISTRIBUTION FUNCTIONS USING FEW MOMENTS

E.C.G. WILLE

Federal University of Technology-Paraná (UTFPR)
Av. Sete de Setembro 3165, 80230-901, Curitiba (PR), Brazil.
ewille@utfpr.edu.br

***Abstract*— A procedure is presented for approximating a given probability distribution function or statistical data considering a subset of their moments. This is done by a method of fitting moments of a piecewise linear function to the moments of the known data. The approach has many advantages over popular approximation approaches. The procedure is demonstrated with commonly used cdfs (Exponential, Gamma, Log-Normal, Normal) and more difficult problems involving sum and product of random variables, obtaining good agreement between the theoretical/simulation curves and the piecewise linear approximations.**

***Keywords*— Moment problem, Sum of RVs, Product of RVs, Piecewise linear functions, Five-parameter logistic function.**

I. INTRODUCTION

Approximating a given cumulative (or probability) distribution function (cdf) or statistical data considering a subset of their moments is a known problem for which there does not exist a satisfactory solution.

The literature offers varying strategies to tackle the problem, each one with its own drawbacks. For instance, it is possible to obtain approximation by means of series expansions, e.g. Gram-Charlier series and Edgeworth series (Das Gupta, 2008). These expansions have a serious drawback, the approximation could result in negative values. Lagrange interpolants can be used to fit the inverse cdf. Because of the oscillatory behavior exhibited by a high-order interpolant, this method does not preserve densities' monotonicity (Gallier, 1999). Estimating the generalized lambda distribution (gld) parameters to match the first moments of the empirical data is another popular method. The gld is defined by a percentile function. From this it is easy to derive the probability density function for the gld using differentiation by parts, however the cumulative distribution function needs to be calculated numerically. Unfortunately, in this approach, more than one set of parameter values can exist for a set of moments. These different sets of parameters will produce a different set of approximations (Au-Yeung, 2003). Other conventional approach is the saddlepoint approximation. It is derived using saddlepoint integration techniques applied to the moment generation function (mgf) inversion. This approximation has a very simple form, but often not calculable explicitly, since an analytic expression for the saddlepoint value is not always available (Kolassa, 2006). Aside from those, there have been several studies dealing with the numerical computation of densities and distributions given the characteristic function (chf). For example, Beaulieu's series expressions are very useful because the chf is readily available in closed form, but it is difficult or impossible to carry out the inverse Fourier transform to get the pdf in closed form. Nevertheless, difficulties in selecting the sampling rate parameter, numerical instability, and truncation errors can hinder the process (Tellambura and Annamalai, 2000).

In practice high-order moments are hard to find and/or very inaccurate, thereby an alternative to evaluate (even in approximate form) probability amounts based on few information is helpful. The present paper proposes a procedure for approximating a distribution function by setting up a piecewise linear function based only on N moments of the cdf. The procedure has the advantage over popular approximation approaches in that:

(i) it approximates a probability distribution function based on few moments;
(ii) it produces a single normalized, monotonically non-decreasing, and valid cdf;
(iii) it does not need to set any control parameter; and
(iv) it does not suffer from numerical instability problems.

The rest of this paper is organized as follows: In Section II, a procedure is provided for approximating a distribution function based on piecewise linear functions. In Section III, we present examples of the procedure applied to commonly used probability distributions, and also more difficult problems involving sum and product of random variables. The fitting of a logistic function is also considered. Finally in Section IV, concluding remarks are made.

II. THE CDF APPROXIMATION

Consider a positive random variable X with cumulative probability function (cdf) $F(x)$. Then the statistical moment of order n is defined by:

$$E[X^n] = n\int_0^\infty x^{n-1}\,(1-F(x))dx. \qquad (1)$$

Instead of $F(x)$ it will be convenient to work with the complementary cumulative probability function (ccdf) $\bar{F}(x) = 1 - F(x)$. In order to approximate $\bar{F}(x)$, $x \in [a,b]$ we choose a partition $a = x_0 < x_1 < \cdots < x_{N+1} = b$ to divide the domain of $\bar{F}$ into $N+1$ sub-intervals. This *piecewise linear approximation* is defined by $N+1$ line segments as follows:

$$\bar{F}_i(x) = y_i + \frac{(y_{i+1}-y_i)}{(x_{i+1}-x_i)} \cdot (x - x_i), \quad (2)$$

$$x \in [x_i, x_{i+1}], \;\; i = 0, \cdots, N.$$

Geometrically, this can be seen as a curve being represented as a set of connected straight line segments as shown in Fig. 1.

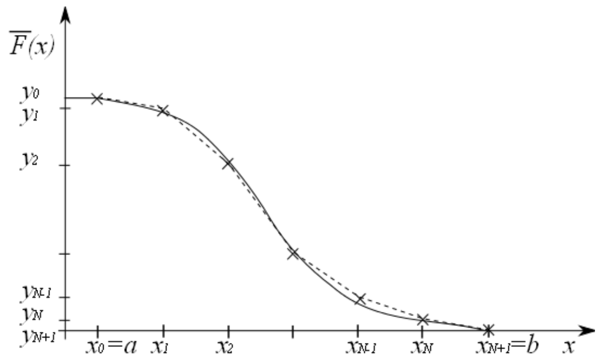

Figure 1: Piecewise linear approximation of a complementary cumulative probability function.

Then it is possible to obtain the moment of order *n*, $E_N[X^n]$, related to the approximate ccdf as follows:

$$E_N[X^n] = n\sum_{i=0}^{N}\left(\int_{x_i}^{x_{i+1}} x^{n-1}\bar{F}_i(x)dx\right). \quad (3)$$

We assume that the intersection points x_i, $i = 0,\ldots,N+1$, are known. Clearly, $y_0 = 1$ and $y_{N+1} = 0$. By substituting Eq. 2 into Eq. 3, the last equation can be put in function of the unknowns y_i as follows:

$$E_N[X^n] = \alpha(n, x_0, x_1) + \sum_{i=1}^{N}\beta(n, x_{i-1}, x_i, x_{i+1}) \cdot y_i\,, \quad (4)$$

where:

$$\alpha(k,q,r) = \frac{kq^{k+1} - (k+1)q^k r + r^{k+1}}{(k+1)(r-q)}, \quad (5)$$

and

$$\beta(k,q,r,s) = \alpha(k,r,s) - \alpha(k,r,q)$$
$$= \frac{q^{k+1}(s-r) - rs(r^k - s^k) + q(r^{k+1} - s^{k+1})}{(k+1)(q-r)(r-s)}. \quad (6)$$

The method of moments estimates parameters of a predefined distribution by equating moments of sample values and moments of the distribution. In our case, we need to determine the unknowns y_i, which parametrize $\bar{F}(x)$ and provide an optimal fit in the sense of minimizing the following problem:

$$\min_{\mathbf{y}} G(\mathbf{y}) = \sum_{i=1}^{N}(E_N[X^n] - E[X^n])^2 \quad (7)$$

subject to: $\quad 1 \geq y_1 \geq y_2 \geq \cdots \geq y_N \geq 0. \quad$ (8)

The constraints (8) ensure the monotonicity of the distribution. This *bounded-variable least squares problem* can be put in matrix form (Stark and Parker, 1995):

$$\min_{\mathbf{y}} G(\mathbf{y}) = \|\mathbf{B} \cdot \mathbf{y} - \mathbf{d}\|^2 \quad (9)$$

subject to: $\quad 1 \geq y_1 \geq y_2 \geq \cdots \geq y_N \geq 0,$

where:

$$\mathbf{B} = \begin{bmatrix} \beta(1,x_0,x_1,x_2) & \beta(1,x_1,x_2,x_3) & \cdots & \beta(1,x_{N-1},x_N,x_{N+1}) \\ \beta(2,x_0,x_1,x_2) & \beta(2,x_1,x_2,x_3) & \cdots & \beta(2,x_{N-1},x_N,x_{N+1}) \\ \vdots & \vdots & \ddots & \vdots \\ \beta(N,x_0,x_1,x_2) & \beta(N,x_1,x_2,x_3) & \cdots & \beta(N,x_{N-1},x_N,x_{N+1}) \end{bmatrix} \quad (10)$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \;\; \mathbf{d} = \begin{bmatrix} E[X^1] - \alpha(1,x_0,x_1) \\ E[X^2] - \alpha(2,x_0,x_1) \\ \vdots \\ E[X^N] - \alpha(N,x_0,x_1) \end{bmatrix}, \quad (11)$$

and $\|\cdot\|$ is the Euclidean norm and the moments $E[X^n]$ are given.

This optimization problem is convex, since $G(\mathbf{y})$ is quadratic with a positive definite Hessian matrix and the non-negativity constraints are linear, and hence has a unique global minimizer. Then we found a valid, monotone non-decreasing, approximation for the unknown cumulative probability function. With the set $\{x_i, y_i\}$, it is in turn possible to evaluate analytically the probability $P[X \leq x] \approx 1 - \bar{F}(x)$ using Eq. 2. It is worth mentioning that the pdf can be obtained by analytical differentiation of the proposed piecewise linear cdf.

III. APPLICATIONS

In this section we present a set of numerical experiments with the procedure presented above. We consider two ways to set up the distribution of the points x_i in the interval $[a, b]$. For symmetrical probability distributions, we use equidistant points, i.e., $x_i = a + \frac{(b-a)}{(N+1)} \cdot i$, $i = 0, \cdots, N+1$. For asymmetrical ones, we use logarithmically-spaced points, then $x_i = (1+a) \cdot \left(\frac{1+b}{1+a}\right)^{\frac{i}{N+1}} - 1$, $i = 0, \cdots, N+1$.

According to our experiments, good results are obtained if $\sqrt{G(\mathbf{y})} \leq 10^{-6}$. In some cases it is necessary to increase (or decrease) N or change $[a, b]$ in order to find a better approximation.

A. Approximations of Known CDFs

In Fig. 2, we present examples of the procedure applied to commonly used cdfs (Exponential, Gamma, Log-Normal, Normal). In this case we set $N = 6$. Solid curves denote the theoretical cdf and dashed curves the cdf obtained with the proposed procedure. The moments $E[X^N]$ were directly calculated from explicit expressions for each distribution. By visual inspection of the plots, we observe a good agreement between the theoretical curves and the piecewise linear approximations.

The approximation behavior with increasing number of moments is illustrated with the example of Fig. 3. In this case a Normal distribution with $\mu = 2.5$ and $\sigma = 0.5$ was chosen, and the values $N = 4, 8, 12$ were considered.

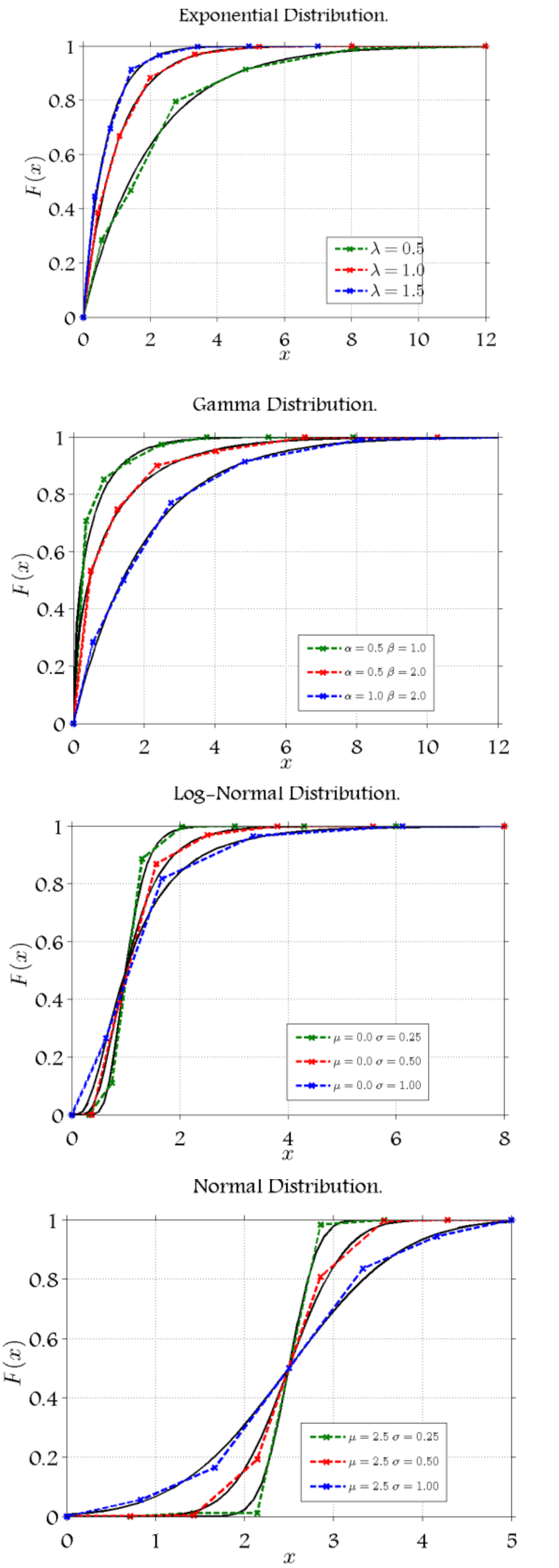

Figure 2: Piecewise linear approximations of known probability distributions (Exponential, Gamma, Log-Normal, Normal) considering $N = 6$. Markers: approximate cdfs, solid lines: theoretical cdfs.

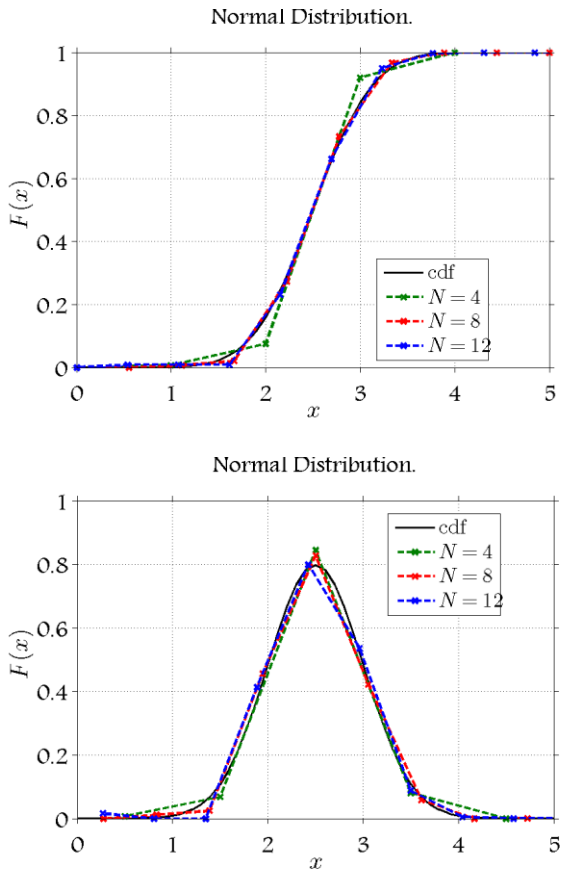

Figure 3: Piecewise linear approximation of Normal distribution ($N = 4$; 8; 12). The densities were recovered from the approximate distributions. Markers: approximate, solid lines: theoretical.

The densities were recovered from the approximate distributions. We observe that, using only a few moments, a good approximation for $P[X \leq x]$ can be obtained. Moreover, as N increases the better is the agreement between the curves.

B. Approximations for Sum and Product of RVs.

In this section, we deal with more difficult cases in which only moments are known.

B.1. Compound distributions

The distribution of the random sum $Y = X_1 + X_2 + \cdots + X_K$, where X_i are independent and identically distributed positive random variables and K is a counting random variable, is called a *compound distribution* (Grubbström and Tang, 2006). In Fig. 4 (top plot), we present approximations of a compound Geometric-Gamma distribution, in which the gamma distribution has *shape parameter* α and *scale parameter* θ, and the geometric distribution[1] has parameter p. The mgf of Y (with no known inversion function) is as follows:

$$M_Y(t) = P_K[M_X(t)] = \frac{p}{(1-\theta t)^{\alpha} + p - 1}, \qquad (12)$$

where $P_K[\cdot]$ is the probability generating function (pgf)

[1] We mean: $p[K = k] = (1-p)^{k-1} \cdot p, \quad k = 1,2, \ldots..$

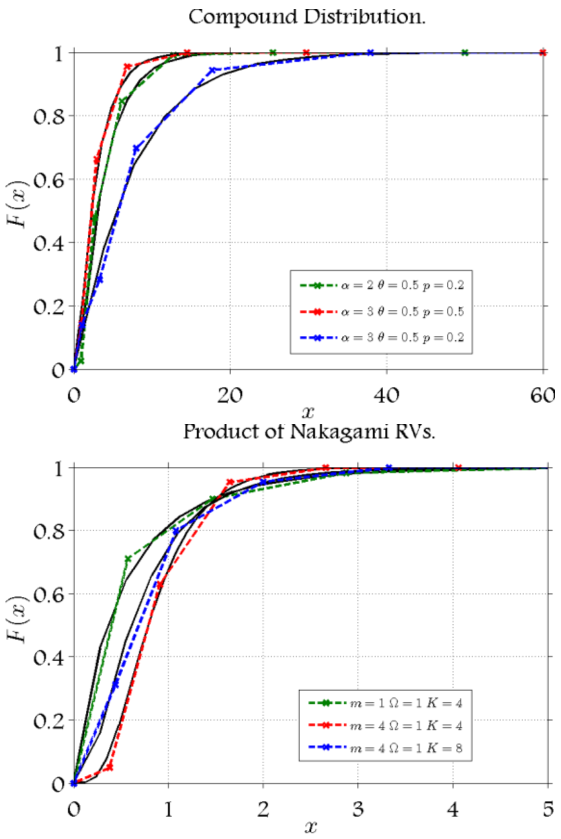

Figure 4: Compound (top plot) and *K*-Nakagami (bottom plot) distributions ($N = 5$). Markers: approximate cdfs, solid lines: simulated cdfs.

of K, and $M_X(t)$ is the mgf of X. The moments were obtained based on: $E[Y^n] = \frac{d^n}{dt^n} M_Y(t)|_{t=0}$. The value $N = 5$ was considered. Monte Carlo simulations, considering 2×10^5 samples of the random variable Y, were carried out to estimate $F_Y(\cdot)$. We observe a good agreement between the curves.

It is worth mentioning that Y is *moment-determinate*, i.e., there is only one pdf with the moment sequence $E[Y^i]$, $i \geq 0$. This fact follows from the existence of the mgf of Y (Lin, 2017).

B.2. Product of Nakagami-m RVs

The Nakagami-m distribution has gained considerable research interest in the area of wireless communications thanks to its good fit to measured fading data (Crepeau, 1992). The Nakagami-m pdf of the signal's envelope is given by the formula:

$$f(x) = \frac{2}{\Gamma(m)} \left(\frac{m}{\Omega}\right)^m x^{2m-1} e^{-mx^2/\Omega}, \ x \geq 0, m \geq 0.5, \quad (13)$$

where $\Gamma(m)$ expresses the gamma function, $m = E^2(X^2)/Var(X^2)$ is the shape parameter, determining the severity of fading, and $\Omega = E(X^2)$. The pdf of the product of K independent Nakagami-m RVs, $R = X_1 \times X_2 \times \cdots \times X_K$, is difficult to evaluate (requiring intensive numerical calculations), however its moments are given by:

$$E[R^n] = \prod_{i=1}^{K} \frac{\Gamma(m_i + \frac{n}{2})}{\Gamma(m_i)} \left(\frac{m_i}{\Omega_i}\right)^{n/2}. \quad (14)$$

In Fig. 4 (bottom plot), we present approximations based in our approach. Solid lines show Monte Carlo simulations. We observe a good agreement between the curves.

It should be noted, however, that R is moment-indeterminate (for $K \geq 3$), i.e., there are pdfs which are different from the pdf of R but with the same moments. This can be proved by considering a Nakagami-m pdf as a special case of a generalized Gamma distribution (which is moment-indeterminate for $K \geq 3$; Lin and Stoyanov (2015)).

C. Fitting of a logistic function

An alternative to $1 - \bar{F}(x)$, for the evaluation of probabilities, corresponds to the fitting of a single continuous function to the points $\{x_i, y_i\}$ obtained with our proposed procedure. Following this idea, we consider the five-parameter logistic function (5PL), reparameterized by Liao and Rong (2009), which accounts for curve asymmetry and has improved statistical performance compared to current 5PL. In this case, approximate cdfs and pdfs are given by:

$$F_L(x) = c_1 + \frac{c_2 - c_1}{\left(1 + \left(2^{\frac{1}{c_5}} - 1\right)\left(\frac{x}{c_4}\right)^{c_3}\right)^{c_5}}. \quad (15)$$

$$f_L(x) = \frac{dF_L}{dx} = \frac{(c_2 - c_1) c_3 c_5 \left(2^{\frac{1}{c_5}} - 1\right)\left(\frac{x}{c_4}\right)^{c_3 - 1}}{c_4 \left(1 + \left(2^{\frac{1}{c_5}} - 1\right)\left(\frac{x}{c_4}\right)^{c_3}\right)^{c_5 + 1}}. \quad (16)$$

In order to find the coefficients c_j, it is necessary to solve the following *non-linear least squares problem:*

$$\min_{\mathbf{c}} W(\mathbf{c}) = \sum_{i=0}^{N+1} (y_i - F_L(x_i))^2. \quad (17)$$

Fits for the re-parameterized 5PL function, considering some cases of Section III B, are displayed in Fig. 5. Dashed lines show cdfs and the corresponding pdfs estimated by Eq. 15 and Eq. 16. Solid lines show Monte Carlo simulations. Good agreement between the curves are observed.

IV. CONCLUSION

In this paper, we have presented a new procedure to approximate probability distribution functions based on few moments. This is done by a method of fitting moments of a piecewise linear function to the moments of the known data. The procedure has the advantage over popular approximation approaches in that it approximates a probability distribution function based on few moments; it produces a single normalized, monotonically non-decreasing, and valid cdf; it does not need to set any control parameter; and it does not suffer from numerical instability problems. We have demonstrated the efficiency of the proposal by applying it to several study cases. Considering commonly used cdfs (Exponential,

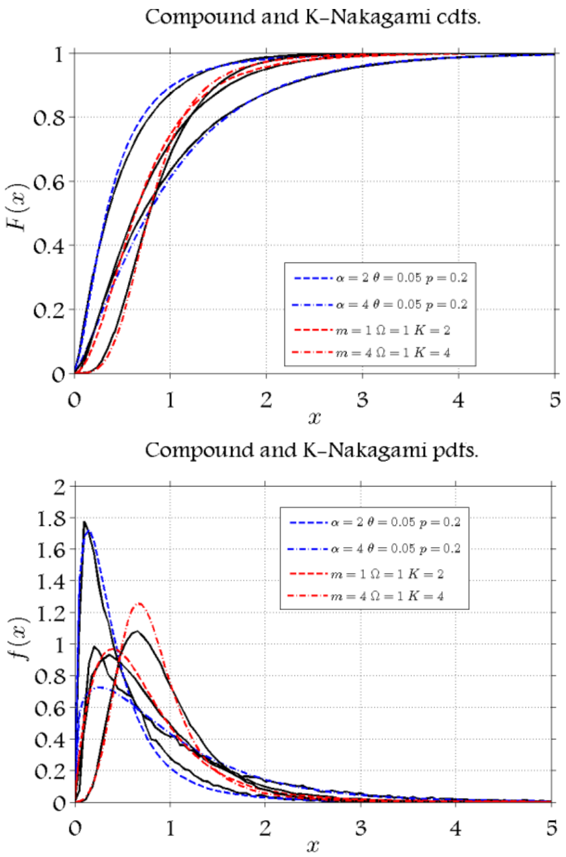

Figure 5: Approximations using fitted logistic functions. Dashed lines: 5PL, solid lines: simulation.

Gamma, Log-Normal, Normal), we obtained good agreement between the theoretical curves and the piecewise linear approximations. Good results were also obtained for more difficult problems involving a compound Geometric-Gamma distribution and the product of K independent Nakagami-m random variables.

Future work will consider the generalization of the method to higher dimension. For instance, in the case of bivariate probability functions, the idea is simply to replace the piecewise linear function $\bar{F}_i(x)$ with a joint one $\bar{F}_i(x,y)$ and the moments $E[X^n]$ with joint moments $E[X^n.Y^k]$. Another idea is to adapt the proposed approach to consider a more elaborated optimization criterion such as the Cumulative Residual Entropy subject to the moments.

ACKNOWLEDGEMENTS

The author is grateful to the anonymous referees for their valuable suggestions.

REFERENCES

Au-Yeung, S.W.M. (2003). *Finding probability distributions from moments*. MSc. thesis, University of London.

Crepeau, P.J. (1992). "Uncoded and coded performance of MFSK and DPSK in Nakagami fading channels," *IEEE Transactions on Communications*, **40**, 487-493.

DasGupta, A. (2008). *Asymptotic theory of statistics and probability*. Springer Texts in Statistics.

Gallier, J. (1999*) Curves and surfaces in geometric modeling: Theory and algorithms*.

Grubbström, R.W. and Tang, O. (2006). "The moments and central moments of a compound distribution," *European Journal of Operational Research*, **170**, 106-119.

Kolassa, J.E. (2006). *Series approximation methods in statistics*, 3rd ed. Lecture Notes in Statistics. Springer, Dordrecht.

Liao, J.J.Z. and Rong, L. (2009) "Re-parameterization of five-parameter logistic function," *Journal of Chemometrics*, **23,** 248-253.

Lin, G.D. (2017), "Recent developments on the moment problem," *Journal of Statistical Distributions and Applications*, **4**, 1-17.

Lin, G.D. and Stoyanov, J. (2015). "Moment determinacy of powers and products of nonnegative random variables," *Journal of Theoretical Probability*, **28**, 1337-1353.

Stark, P.B. and Parker, R.L. (1995). "Bounded variable least-squares: an algorithm and applications," *Computational Statistics*.

Tellambura, C. and Annamalai, A. (2000). "Further results on the Beaulieu series," *IEEE Transactions on Communications*, **48**, 1774-1777.

Received March 3, 2019
Sent to Subject Editor September 7, 2019
Accepted November 7, 2019
Recommended by Subject Editor José Guivant