MATHEMATICS OF COMPUTATION, VOLUME 31, NUMBER 137
JANUARY 1977, PAGES 214–225

Approximations for Hand Calculators Using Small Integer Coefficients

By Stephen E. Derenzo

Abstract. Methods are presented for deriving approximations containing small integer coefficients. This approach is useful for electronic hand calculators and programmable calculators, where it is important to minimize the number of keystrokes necessary to evaluate the function. For example, the probability $P(x)$ of exceeding x standard deviations of either sign (Gaussian probability integral) is approximated by

$$P(x) \approx \mathrm{EXP}\left[-\frac{(83x + 351)x + 562}{703/x + 165}\right]$$

with a relative error less than 0.042% over the range $0 < x < 5.5$ $(1 > P(x) > 4 \times 10^{-8})$. Other examples presented are the functional inverse of $P(x)$; the Klein-Nishina cross section for Compton scattering; photoelectric cross sections in H_2O, Bone, Fe, NaI, and Pb; and the pair production cross section in Pb.

1. Introduction. By the use of suitable approximations most functions can be conveniently evaluated on automatic digital computers [1], [2]. However, these approximations (usually polynomials or rational functions) are often inconvenient for hand calculators because many keystrokes are required to enter the coefficients. In this paper we describe methods for deriving approximations containing small integer coefficients, which substantially reduce the number of keystrokes required. This approach is also important for programmable calculators, where the stored programs are usually limited to a certain number of keystrokes.

In many cases such approximations can be evaluated as rapidly and will generally be as accurate as interpolation from tables, eliminating the need for tables in those cases. Although graphical representation permits ready interpolation it has limited accuracy, especially when the function spans many decades.

The method consists of four parts: (1) selecting a suitable form for the approximation, (2) fitting the approximation to the function, (3) eliminating unnecessary terms in the approximation and (4) determining small integer coefficients that give a fit not substantially worse than the best fit of (2).

2. Method.

2.1. *Selecting a Suitable Form.* This part of the method rests heavily on the existing body of approximation theory [1], [2], but a few comments seem appropriate. (1) Many electronic calculators are equipped with $\sqrt{x}$, $\sin x$, $\cos x$, $\tan x$, e^x, $\ln x$, x^y, etc., keys and these should be considered if the function to be approximated resembles one of them. (2) There is much merit however, in the polynomial, which (when arranged according to Horner's rule) has a repetitive pattern that lends itself to a rapid keystroke

Received August 19, 1975; revised April 12, 1976.
AMS (MOS) subject classifications (1970). Primary 41A20.

pace. Moreover, electronic calculators can evaluate polynomials as rapidly as the keys are depressed but this is not so for the transcendental functions. (3) Asymptotic limits are important in the selection of a form. For example, the Klein-Nishina formula (discussed in Section 3.3) approaches a constant value at low photon energy E and decreases as E^{-1} at large values of E. This suggests a form such as

$$g(E) = \frac{a_1 + a_2E + a_3E^2}{a_4 + a_5E + a_6E^2 + a_7E^3}.$$

Note that no simple power series can satisfy these limits.

2.2. *Fitting the Approximation to the Function.* After a form $g(x)$ has been chosen, its (unknown) coefficients a_1 to a_N must be selected so that $g(x)$ fits the function $f(x)$ to be approximated. The usual criterion is the minimax (or least maximum) error criterion, requiring that the largest deviation of $|d(x)|$ be minimized, where

$$d(x) = w(x)[g(x) - f(x)] \tag{1}$$

and $w(x)$ is a weighting function [3]. Under this criterion, the function $d(x)$ oscillates about zero with equal positive and negative excursions. For rational approximating forms Chebyshev's Theorem gives the minimum number of excursions that are necessary and sufficient for a best approximation [2].

Unfortunately, the minimax criterion does not lend itself to the minimization code used in this work [4], as the code assumes that the function to be minimized is locally quadratic in the coefficients a_i. It was found, however that the code could minimize D given by:

$$D = \sqrt{\sum_{j=1}^{M} d(x_j)^4}, \tag{2}$$

where the base points x_j were chosen with sufficient density that $d(x_j)$ was a reasonable representation of $d(x)$. Moreover, the resulting deviations $d(x_j)$ oscillated about zero with very nearly equal positive and negative excursions and had the necessary minimum number of excursions for a best fit (see Figure 1 and [5]). Thus, while the best fit coefficients given in Section 3 may not be unique, no other values can yield a significantly better fit [6].

In the event that the deviations $d(x_j)$ are larger then the required accuracy, it is necessary to go back and improve the form of the approximation. This usually means increasing the number of terms and consequently increasing the number of coefficients.

2.3. *Eliminating Unnecessary Terms.* As a rule, we started with a form that contained a sufficient number of terms to give a good fit. Then the computer code set each coefficient in turn to zero while all others were varied to minimize D. If the best of these fits was acceptable, the related coefficient was set permanently to zero and the process was automatically repeated to try to eliminate other terms. Although this procedure usually resulted in the elimination of the highest order terms, it was applied equally to all coefficients.

2.4. *Determining Small Integer Coefficients.* The methods described in this section assume that the approximation remains numerically unchanged when all coefficients

are multiplied by a common factor. For example, this assumption is valid for power series (provided they are divided by a single coefficient) and more generally for the rational approximations, but not for expansions in transcendental functions.

We now define a scale factor b_1 that is allowed to take on the integer values 1, 2, 3, ... Renaming the best fit coefficients a_1 to a_N so that the coefficient closest to zero is a_1, the scaled best fit coefficient values are given by:

$$b_i = a_i \cdot b_1/a_1 \quad \text{and} \quad |b_i| \geq b_1.$$

Clearly, in the limit of large integer values of b_1 it is possible to round all the other coefficients to their nearest integer values and still remain very close to the best fit approximations. This suggests a straightforward integer search algorithm that consists of tabulating D (and the deviations $d(x_j)$) for $b_1 = 1, 2, 3, \ldots$ where in each case the best fit values of b_2 to b_N are rounded to the nearest integer. Usually, the resulting values of D are far from monotonic; and it is possible to stop the search at a downward fluctuation in D that corresponds to an acceptable fit.

The above search method was not used in this paper because, for a given b_1, the integer values of b_2 to b_N closest to the best fit are usually not the best integer values (i.e., those that result in the lowest value of D).

By searching the space of b_2 to b_N it is often possible to find a set of integer values that result in a lower value of D, because the variation in each coefficient from its best fit value has been nearly compensated by the variations in the other coefficients.

As an example of the need for such a search, consider Eq. (10) of Section 3.2. The best fit coefficients are $(b_1 \equiv 1)$, $b_2 = 260.40 \ldots$, $b_3 = 503.60 \ldots$, $b_4 = 134.16 \ldots$, $b_5 = 543.36 \ldots$ $(D = 0.412)$ and varying these coefficients by less than 0.15% to their nearest integer values $b_2 = 260$, $b_3 = 504$, $b_4 = 134$, $b_5 = 543$ yields a rather poor fit $(D = 6.8)$. A complete search of the integer space of b_2 to b_5 for the lowest value of D results in coefficients that are far from their best fit values: $b_2 = 280$, $b_3 = 572$, $b_4 = 144$, $b_5 = 603$ but $D = 0.520$, not much larger than the best fit value.

Unfortunately, a straightforward search over a wide range of integer coefficient values requires calculating D an unacceptably large number of times (typically 10^7). Moreover, it is not obvious from the values of D how far each coefficient should be stepped.

A more efficient algorithm was therefore devised that restricted the search (as much as possible) to the volume V' within which $D < D'$ where D' is the lowest value of D yet achieved during the integer coefficient search. The appendix is an example of this algorthim as used in this work to search V' and determine the best integer values of b_2 to b_N for each successive integer value b_1. Although it only covers the case $N = 4$, it is clear from its structure how it may be modified to handle any other value of N. It is hoped that the way in which it was written is self-explanatory.

This procedure permits a complete search for the smallest integer coefficients that result in an acceptable fit, subject to the condition that all subspaces of b_1 to b_N have a single minimum value of D. It can search a deep, narrow valley while avoiding regions too far from the valley to be fruitful. The examples below required typically 10^4 to 10^6 evaluations of D, depending on the number of coefficients.

3. **Examples.** The examples that follow were chosen largely on the basis of their usefulness to physicists and engineers. No claim is made that the approximating forms are the best that could be chosen, only that no smaller integer coefficients can be used in those forms to give a significantly better fit. In each example the approximation with integer coefficients has deviations that are within a factor of two of those that result from using the best fit coefficients. (For details of how the fits were performed and for additional graphs of the deviation functions $d(x)$, see [5].)

In our experience an integer coefficient fit that approaches the best fit involves fewer keystrokes than a fit using smaller integers but one or two additional terms.

The approximations make considerable use of polynomial forms and these have been arranged according to Horner's rule to minimize the number of keystrokes and the need for intermediate storage.

3.1. *Gaussian Probability Integral.* The probability $P(x)$ of exceeding x standard deviations of either sign is given by:

$$P(x) = \sqrt{2/\pi} \int_x^\infty e^{-x^2/2}\, dx, \qquad x \geqslant 0. \tag{3}$$

Approximation 1:

$$\tilde{P}_1(x) = \mathrm{EXP}\left[-\frac{(83x + 351)x + 562}{703/x + 165}\right]. \tag{4}$$

Error: $|(\tilde{P}_1(x) - P(x))/P(x)| < 0.042\%$.
Range: $0 < x \leqslant 5.5$ $(1 > P(x) > 3.8 \times 10^{-8})$.
Approx. number of keystrokes = 26.
Approximating form:

$$\ln[\tilde{P}_1(x)] = -\frac{((b_1x + b_2)x + b_3)x}{b_4x + b_5}. \tag{5}$$

Best fit values [7]: $(b_1 \equiv 1)$, $b_2 = 4.20075 \pm 0.00020$, $b_3 = 6.72175 \pm 0.00083$, $b_4 = 1.988778 \pm 0.000075$, $b_5 = 8.39964 \pm 0.00036$. In Figure 1 we compare the deviation functions $d(x)$ for the best fit and integer approximations.

Approximation 2:

$$\tilde{P}_2(x) = \sqrt{2/\pi}\left(\frac{1}{x}\right)\mathrm{EXP}(-x^2/2 - 0.94/x^2). \tag{6}$$

Error: $|(\tilde{P}_2(x) - P(x))/P(x)| < 0.040\%$.
Range: $x \geqslant 5.5$ $(P(x) < 3.8 \times 10^{-8})$.
Approx. number of keystrokes = 20.

3.2. *Inverse of the Gaussian Probability Integral.* Defining $P(x)$ by Eq. (3):
Approximation 1:

$$\tilde{x}_1 = \sqrt{\frac{((4y + 100)y + 205)y^2}{((2y + 56)y + 192)y + 131}}, \tag{7}$$

where $y = -\ln(P)$.

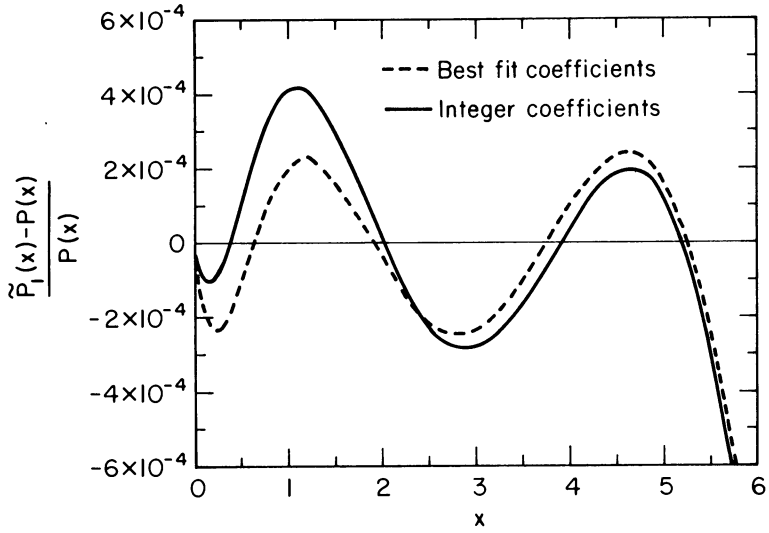

FIGURE 1

Solid curve—relative deviation between approximation $\tilde{P}_1(x)$ with integer coefficients (Eq. (4)) and the Gaussian probability $P(x)$ of exceeding x standard deviations of either sign (Eq. (3)). Dashed curve—same but with best fit coefficients (Eq. (5)).

Error: $|x_1 - x| < 1.3 \times 10^{-4}$.

Range: $1 > P(x) > 2 \times 10^{-7}$ $(0 < x < 5.2)$.

Approx. number of keystrokes = 38.

Approximation 2:

$$\tilde{x}_2 = \sqrt{\frac{((2y + 280)y + 572)y}{(y + 144)y + 603}}, \tag{8}$$

where $y = -\ln(P)$.

Error: $|x_2 - x| < 4 \times 10^{-4}$.

Range: $2 \times 10^{-7} > P > 10^{-112}$ $(5.2 < x < 22.6)$.

Approx. number of keystrokes = 30.

Approximating forms:

$$\tilde{x}_1 = \sqrt{\frac{((2b_1 y + b_2)y + b_3)y^2}{((b_1 y + b_4)y + b_5)y + b_6}} \tag{9}$$

with best fit coefficients [7] $(b_1 \equiv 1)$, $b_2 = 48.8740 \pm 0.0024$, $b_3 = 95.976 \pm 0.015$, $b_4 = 27.4283 \pm 0.0016$, $b_5 = 91.446 \pm 0.012$, $b_6 = 61.231 \pm 0.025$, and

$$\tilde{x}_2 = \sqrt{\frac{((2b_1 y + b_2)y + b_3)y}{(b_1 y + b_4)y + b_5}} \tag{10}$$

with best fit coefficients [7] $(b_1 \equiv 1)$, $b_2 = 260.403 \pm 0.016$, $b_3 = 503.60 \pm 0.62$, $b_4 = 134.1596 \pm 0.0087$, $b_5 = 543.36 \pm 0.37$.

Material	Z	K = 0.6Z/A (cm^2/gm)	K = 0.9964Z (barns/atom)	Density (gm/cm^3)
H_2O	–	0.3330	–	1
Al	13	0.2891	12.95	2.692
Si	14	0.2991	13.95	2.4
Fe	26	0.2793	25.90	7.86
Ge	32	0.2645	31.88	5.4
NaI	–	0.2563	–	3.67
W	74	0.2415	73.7	19.3
Pb	82	0.2375	81.7	11.35
U	92	0.2319	91.7	18.7

TABLE I. *K values for the Klein-Nishina approximation (Eq. (12))*

3.3. *Klein-Nishina Formula.* The Klein-Nishina formula describes the narrow beam attenuation of photons by Compton scattering on free electrons [8]. Its exact expression is given by:

$$\sigma_{KN} = 2\pi r_0^2 \left[\frac{((\alpha + 9)\alpha + 8)\alpha + 2}{\alpha^2(1 + 2\alpha)^2} + \frac{(\alpha + 2)\alpha - 2}{2\alpha^3} \ln(1 + 2\alpha) \right], \tag{11}$$

where $\alpha = E/m_e c^2$ and r_0 is the classical electron radius (2.8179×10^{-13}cm). Although this expression may be evaluated directly, approximately 60 keystrokes are required.

Approximation:

$$\tilde{\sigma} = K \frac{(E + 28)E + 16}{((E + 54)E + 134)E + 24} \tag{12}$$

E is the photon energy in MeV

$K = 0.6000\ Z/A$ (cm^2/gm)

$K = 0.9964\ Z$ (barns/atom)

$K = 0.9964$ (barns/electron)

Error: $|(\tilde{\sigma} - \sigma_{KN})/\sigma_{KN}| < 0.56\%$.

Range: $0 < E < 100$ MeV.

Approx. number of keystrokes = 31, using the K values given in Table I.

Approximating form:

$$K \frac{(E + b_1)E + b_2}{((E + b_3)E + b_4)E + b_5}. \tag{13}$$

Best fit coefficients [7]: $K = 0.9667 \pm 0.0020$ barns/electron, $b_1 = 23.718 \pm 0.098$, $b_2 = 13.186 \pm 0.050$, $b_3 = 45.62 \pm 0.22$, $b_4 = 108.08 \pm 0.64$, $b_5 = 19.206 \pm 0.097$.

Material	Density (g/cm^3)	Energy range (MeV)	Photoelectric cross section (cm^2/gm)[c]	Error[d] e_1 (%)	Error[d] e_2 (cm^2/gm)	Approx. No. Keystrokes
H_2O	1	> 0.01	(((X/7+35)X-97)X+196)X/10^7	1.1	1×10^{-5}	23
Compact bone	∿ 2	> 0.01	((20X-31)X+52)X/10^6	2.7	1.5×10^{-5}	18
Fe	7.87	> 0.01	(((-X/23+22)X-11)X+25)X/10^5	2.9	2×10^{-5}	24
NaI	3.67	> 0.0332[a]	(((-X+107)X+62)X+82)X/7×10^4	2.8	2×10^{-5}	23
Pb	11.35	0.0159-0.088[b]	(-X/26+8)X+32)X^2/10^4	1.4	0	19
Pb	11.35	> 0.088[a]	(((-X+52)X+93)X+42)X/10^4	1.4	1×10^{-5}	21

[a] above K edge.

[b] between L1 and K edges.

[c] X = 1/E in units MeV^{-1}.

[d] Error is the greater of e_1 and e_2.

TABLE II. *Photoelectric cross sections*

3.4. *Photoelectric Cross Sections in* H_2O, Bone, Fe, NaI, and Pb. It is well known that the photoelectric cross sections may be approximated by expansions in inverse powers of the photon energy [9], and we have used the same form for our approximations (Table II). Each of these approximations was fit to typically 25 data points from [9]. The deviations $d(x_i)$ were not smooth functions of x_j because the data are partially based on experimental measurements. Moreover, as stated in [9] these cross sections have not been established with accuracies much better than 5%.

3.5. *Pair Production Cross Section in* Pb.

Approximation:

$$\tilde{\sigma}_p(\mathrm{cm}^2/\mathrm{gm}) \approx \frac{(x^2+9)x-1}{((4x+55)x-168)x+358}, \tag{14}$$

where $x = \log_{10}(E)$ and E is the photon energy in MeV.

Error: $|\tilde{\sigma}_p - \sigma_p| < 8 \times 10^{-4}\mathrm{cm}^2/\mathrm{gm}$.

Range: 1.5 MeV $< E <$ 10^5 MeV.

Approx. number of keystrokes = 28.

The approximation $\tilde{\sigma}_p$ was fit to 26 data points σ_p from 1.5 MeV to 10^5 MeV [9]. The deviations $d(x_j)$ were not smooth functions of x, because the data were only given to three significant figures. Also, the lower energy data points are quite sparse; and the error bound given above is only an estimate.

Acknowledgements. I am grateful to P. Concus and B. Pardoe for helpful discussions. Work was supported by the U. S. National Institute of Health. Facilities were provided by Lawrence Berkeley Laboratory through the U. S. Energy Research and Development Administration.

APPENDIX

EXAMPLE OF INTEGER COEFFICIENT SEARCH PROCEDURE FOR 4 COEFFICIENTS

```
A1 = 1
DETERMINE BEST FIT VALUES OF A2,A3,A4 BY MINIMIZING D

RENAME A1,A2,A3,A4 SO THAT A1 IS CLOSEST TO ZERO


LOOP B1=1,2,3 . . . .
B2=A2·B1/A1
B3=A3·B1/A1
B4=A4·B1/A1
D'=D EVALUATED AT B1, [B2], [B3], [B4]

LOOP J2=0,1
LOOP I2=1,2,3,. . . .
E1=B1
*E2 = [B2] + J2 + (-1)^J2 · I2
**HOLDING E1 AND E2 FIXED, DETERMINE BEST FIT VALUES OF E3 AND E4
    BY MINIMIZING D
EVALUATE D AT E1, E2, E3, E4
***IF D > 1.2 D', EXIT I2 LOOP AND TAKE NEXT J2

LOOP J3=0,1
LOOP I3=1,2,3,. . . .
F1=E1
F2=E2
F3 = [E3] + J3 + (-1)^J3 · I3
HOLDING F1, F2, AND F3 FIXED DETERMINE BEST FIT VALUE OF F4 BY
    MINIMIZING D
EVALUATE D AT F1, F2, F3, F4
IF D > 1.2 D', EXIT I3 LOOP AND TAKE NEXT J3

LOOP J4=0,1
LOOP I4=1,2,3,....
G1=F1
G2=F2
G3=F3
G4 = [F4] + J4 + (-1)^J4 · I4
EVALUATE D AT G1, G2, G3, G4
IF D > 1.2 D', EXIT I4 LOOP AND TAKE NEXT J4
IF D ≤ 1.2 D', PRINT OUT ALL RELEVENT QUANTITIES FOR THIS FIT
IF D ≤ D', SET D' = D

NEXT I4
NEXT J4
NEXT I3
NEXT J3
NEXT I2
NEXT J2

STOP PROCEDURE IF D' IS SUFFICIENTLY SMALL
NEXT B1
```

*For J2=0, the I2 loop sets E2 to successive values [B2] + 1, [B2] + 2, . . . where [B2] is the integer part of B2. For J2 = 1 the I2 loop sets E2 to successive values [B2], [B2]-1, [B2]-2, ...

**For efficiency, the starting values E3 and E4 are determined whenever possible by a linear extrapolation of previous best fit values of E3 and E4 (obtained at this point in the code) as a function of E2. Moreover, this minimization can be skipped if the D value associated with E1, E2 and the extrapolated values of E3 and E4 is less than D'. Similar efficiencies are employed for the minimizations in all the other loops.

***This test assumes that the preceding step has found the true minimum rather than a local minimum. If the minimum D is greater than 1.2 D', then this value of E2 (and all subsequent values) define a subspace within which D > 1.2 D'; and the I2 loop is ended.

Lawrence Berkeley Laboratory
University of California
Berkeley, California 94720

1. C. HASTINGS, JR., *Approximations for Digital Computers*, Princeton Univ. Press, Princeton, N. J., 1955. MR 16, 963.

2. J. F. HART ET AL., *Computer Approximations*, Wiley, New York, 1968.

3. For cases where the absolute error $g(x) - f(x)$ is important, $w(x)$ is equal to a constant. For cases where the relative error $(g(x) - f(x))/f(x)$ is important, $w(x)$ is proportional to $1/f(x)$.

4. STEPHEN E. DERENZO, Lawrence Radiation Laboratory Group A Programming Note P-190, Berkeley, Calif., 1969. Most other minimizing codes can also be used.

5. STEPHEN E. DERENZO, Lawrence Berkeley Laboratory Report No. LBL-3804, Berkeley, Calif., March, 1975. (Available from the author.)

6. Henceforth we use the term "best fit coefficients" to mean those that result from the minimization of D (Eq. (2)).

7. The number given after the ± symbol is the amount that the corresponding coefficient must be varied from its best fit value to double the value of D, holding all other coefficients at their best fit values.

8. O. KLEIN AND Y. NISHINA, *Nature*, v. 122, 1928, p. 398. Formulas are also available in the *American Institute of Physics Handbook*, 3rd ed. (D. E. Gray, Editor), McGraw-Hill, New York, 1972, pp. 8–197.

9. J. H. HUBBELL, *Photon Cross Sections, Attenuation Coefficients, and Energy Absorption Coefficients from* 10 keV *to* 100 GeV, Report No. NSRDS-NBS 29, U. S. Nat. Bur. of Standards, 1969.