PHYS-4007/5007: Computational Physics
Course Lecture Notes
Section IV

Dr. Donald G. Luttermoser
East Tennessee State University

Version 7.1

Abstract

These class notes are designed for use of the instructor and students of the course PHYS-4007/5007: Computational Physics taught by Dr. Donald Luttermoser at East Tennessee State University.

IV. Error Analysis and Uncertainties

A. Errors and Uncertainties.

1. Types of Errors.

a) **Blunders:** (Can control these with sleep! → look for inconsistencies in data points.)

i) Typographical errors in program or data.

ii) Running wrong program.

iii) Using wrong data file.

iv) Using wrong equations for analysis.

b) **Random Errors:** (Can't control these → harder to detect.)

i) Electronic fluctuations due to power surges.

ii) Cosmic ray damage on detectors and/or chips.

iii) Somebody pulled the plug!

c) **Systematic Errors:** (Very hard to detect.)

i) Faulty calibration of equipment.

ii) Bias from observer or experimenter.

iii) These must be estimated from analysis of experimental conditions and techniques.

iv) In some cases, corrections can be made to data to compensate for these errors where type and extent of error is known.

v) In other cases, the uncertainties resulting from these errors must be estimated and combined with uncertainties from statistical fluctuations.

d) Approximation Errors:

i) These arise from simplifying the mathematics so that the problem can be solved on a computer:

→ infinite series being replaced by a finite sum

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} \qquad \text{(IV-1)}$$

$$\approx \sum_{n=0}^{N} \frac{x^n}{n!} = e^x + \mathcal{E}(x, N) \; . \qquad \text{(IV-2)}$$

ii) In Eq. (IV-2), $\mathcal{E}(x, N)$ is the **total absolute error**.

iii) This type of error also is called *algorithmic error*, the *remainder*, or *truncation error*.

iv) To have a small approximation error, $N \gg x$ in Eq. (IV-2).

e) Roundoff Errors:

i) Inaccuracies in stated numbers on a computer due to a finite (typically small) number of bits being used on a computer to represent numbers.

ii) The total number of possible "machine" numbers is much less than the total number of "real" numbers (which of course is infinite).

iii) The more calculations a computer does, the larger the roundoff error.

iv) This error can cause some algorithms to become unstable.

v) If the roundoff error value > value of the number being represented $\Longrightarrow$ the result is referred to as *garbage*.

vi) For instance, a computer may compute the following as

$$2\left(\frac{1}{3}\right)-\frac{2}{3} = 0.6666666-0.6666667 = -0.0000001 \neq 0\,.$$

vii) We need to worry about **significant figures** on a computer:

- On a 32-bit CPU, *single precision* is stored in <u>one</u> word $\rightarrow$ 4 bytes (REAL$*$4 or REAL).

- On a 32-bit CPU, *double precision* is stored in <u>two</u> words $\rightarrow$ 8 bytes (REAL$*$8).

- Trailing digits beyond these "bytes" are lost (if you ask a code to print more digits out, the printed digits should be considered as *garbage*).

f) What kind of error do we make in representing a real number in a floating point number system?

i) Absolute Error:

$$\text{true value - approximate value.}$$

ii) Relative Error:

$$\frac{\text{true value - approximate value}}{\text{true value}}.$$

Relative error is not defined if the true value is zero.

2. Accuracy versus Precision.

a) **Accuracy** is how close an experiment comes to the "true" value.

i) It is a measure of the correctness of the result.

ii) For an experimenter, it is a measure of how skilled the experimenter is.

iii) For a programmer, it is a measure on how good they are at programming and the assumptions used in the algorithm.

b) **Precision** of an experiment is a measure of how exactly the result is determined without reference to what the results means.

i) It is a measure of the precision of the instruments being used in the experiment.

ii) The precision of an experiment is dependent on how well we can overcome or analyze random errors.

iii) In programming, it is a measure of how many bits are used to store numbers and perform calculations.

3. Uncertainties.

a) The term *error* signifies a deviation of the result from some "true" value.

i) However, since we often cannot know the "true" value of a measurement prior to the experiment, we can only determine *estimates* of the errors inherent to the experiment.

ii) The difference between two measurements is called the **discrepancy** between the results.

iii) The discrepancy arises due to the fact that we can only determine the results to a certain **uncertainty**.

b) There are two classes of uncertainties:

i) The most prominent type: Those which result from fluctuations in repeated measurements of data from which the results are calculated.

ii) The secondary type: Those which result from the fact that we may not always know the appropriate theoretical formula for expressing the result.

c) **Probable Error:** The magnitude of the error which we estimate we have made in our determination of the results.

i) This does not mean that we expect our results to be wrong by this amount.

ii) Instead, it means that if we are wrong in our results, it *probably* won't be wrong by more than the probable error.

iii) As such, probable error will be synonymous with uncertainty in a measurement or calculation.

4. Implications for Numerical Computing.

a) Most numbers that we use in floating point computations must be presumed to be somewhat in error. They may be *rounded* from the values we have in mind by:

i) Input conversion errors.

ii) Inexact arithmetic from the limited significance of the input numbers due to limited RAM sizes of the computer.

b) How is the value of function $f(x)$ affected by errors in the argument x?

i) Suppose x has a relative error of ϵ so that we actually use the value $x(1+\epsilon)$.

ii) Then the value of the function is $f(x(1+\epsilon))$.

iii) If f is differentiable, the absolute error in the value of f caused by the error in x can be approximated by

$$f(x+\epsilon x) - f(x) \approx \epsilon x f'(x) \ . \qquad \text{(IV-3)}$$

iv) The relative error is then

$$\frac{f(x+\epsilon x) - f(x)}{f(x)} \approx \epsilon x \frac{f'(x)}{f(x)} \ . \qquad \text{(IV-4)}$$

v) As we had above, suppose $f(x) = e^x$, then the absolute error is approximately $\epsilon x e^x$ and the relative error is ϵe^x.

vi) These errors can be serious if x is large. Thus, even if the routine EXP were perfect, there could be serious errors in using EXP(X) for e^x when x is large.

vii) Another significant example is cosine x when x is near $\pi/2$. The absolute error is approximately

$$-\epsilon x \sin x \approx -\epsilon \cdot \frac{\pi}{2} \cdot 1 \ .$$

viii) This is not troublesome, but the relative error is very much so:

$$-\epsilon x \frac{\sin x}{\cos x} \approx -\epsilon \cdot \frac{\pi}{2} \cdot \frac{1}{0} \ .$$

Very small changes in x near $\pi/2$ cause very large relative changes in $\cos x \Longrightarrow$ we say the evaluation is *unstable* there. The accurate values

$$\begin{aligned} \cos 1.57079 &= 0.63267949 \times 10^{-5} \\ \cos 1.57078 &= 1.63267949 \times 10^{-5} \end{aligned}$$

demonstrate how a small change in the argument can have a profound effect on the function value.

c) An important part of computational physics is deriving solutions to problems in terms of series. When one is working to high accuracies or if the series converges slowly, it is necessary to add up a great many terms to sum the series.

B. Useful Theorems in Computational Physics.

1. **Intermediate Value Theorem.** Let $f(x)$ be a continuous function on the closed interval $[a, b]$. If for some number α and for some $x_1, x_2 \in [a, b]$ we have $f(x_1) \leq \alpha \leq f(x_2)$, then there is some point $c \in [a, b]$ such that

$$\alpha = f(c) \ . \tag{IV-5}$$

Here the notation $[a, b]$ means the interval consisting of the real numbers x such that $a \leq x \leq b$.

2. **Rolle's Theorem.** Let $f(x)$ be continuous on the closed, finite interval $[a, b]$ and differentiable on (a, b). If $f(a) = f(b) = 0$, there is a point $c \in (a, b)$ such that

$$f'(c) = 0 \ . \tag{IV-6}$$

Here the notation (a, b) means the interval consisting of the real numbers x such that $a < x < b$.

3. **Mean-Value Theorem for Integrals.** Let $g(x)$ be a non-negative function integrable on the interval $[a, b]$. If $f(x)$ is continuous on $[a, b]$, then there is a point $c \in [a, b]$ such that

$$\int_a^b f(x)g(x)\,dx = f(c) \int_a^b g(x)\,dx \tag{IV-7}$$

(more to come in §VI of the notes).

4. **Mean-Value Theorem for Derivatives.** Let $f(x)$ be continuous on the finite, closed interval $[a, b]$ and differentiable on (a, b). Then there is a point $c \in (a, b)$ such that

$$\frac{f(b) - f(a)}{b - a} = f'(c) \tag{IV-8}$$

(more to come in §VI of the notes).

5. Taylor's Theorem (with Remainder). Let $f(x)$ have the continuous derivative of order $n+1$ on some interval (a,b) containing the points x and $x_\circ$. Set

$$\begin{aligned} f(x) &= f(x_\circ) + \frac{f'(x_\circ)}{1!}(x-x_\circ) + \frac{f''(x_\circ)}{2!}(x-x_\circ)^2 + \\ &\quad + \cdots + \frac{f^{(n)}(x_\circ)}{n!}(x-x_\circ)^n + R_{n+1}(x) \ . \end{aligned} \tag{IV-9}$$

Then there is a number c between x and $x_\circ$ such that

$$R_{n+1}(x) = \frac{f^{(n+1)}(c)}{(n+1)!}(x-x_\circ)^{n+1} \ . \tag{IV-10}$$

6. Let $f(x)$ be a continuous function on the finite, closed interval $[a,b]$. Then $f(x)$ assumes its maximum and minimum values on $[a,b]$; *i.e.*, there are points $x_1, x_2 \in [a,b]$ such that

$$f(x_1) \le f(x) \le f(x_2) \tag{IV-11}$$

for all $x \in [a,b]$.

7. Integration by Parts. Let $f(x)$ and $g(x)$ be real values functions with derivatives continuous on $[a,b]$. Then

$$\int_a^b f'(t)g(t)\,dt = f(t)g(t)\big|_{t=a}^{t=b} - \int_a^b f(t)g'(t)\,dt \ . \tag{IV-12}$$

8. Fundamental Theorem of Integral Calculus. Let $f(x)$ be continuous on the interval $[a,b]$, and let

$$F(x) = \int_a^x f(t)\,dt \quad \text{for all} \quad x \in [a,b] \ . \tag{IV-13}$$

Then $F(x)$ is differentiable on (a,b) and

$$F'(x) = f(x) \ . \tag{IV-14}$$

C. The Mathematics of Errors and Uncertainties.

1. Subtractive Cancellation.

a) For the following, let the "actual" numbers be represented by *unmarked* variables and those on the computer be designated with a 'c' subscript.

b) The representation of a simple subtraction is then

$$a = b - c \implies a_c = b_c - c_c \ , \tag{IV-15}$$

$$a_c = b(1+\epsilon_b) - c(1+\epsilon_c) \ , \tag{IV-16}$$

$$\implies \frac{a_c}{a} = 1 + \epsilon_b \frac{b}{a} - \frac{c}{a}\epsilon_c \ , \tag{IV-17}$$

where the error of the number/variable is given by

$$\epsilon = \frac{\text{computer number - actual number}}{\text{actual number}} \ . \tag{IV-18}$$

c) From Eq. (IV-17), the average error in a is a weighted average of the errors in b and c.

d) We can have some cases, however, when the error in a increases when $b \approx c$ because we subtract off (and thereby lose) the most significant parts of both numbers $\implies$ this leaves the least significant parts.

e) **If you subtract two large numbers, and end up with a small one, there will be less significance in the small one.**

i) For example, if a is small, it must mean that $b \approx c$ and so

$$\frac{a_c}{a} = 1 + \epsilon_a \ , \tag{IV-19}$$

$$\epsilon_a \approx \frac{b}{a}(\epsilon_b - \epsilon_c) \ . \tag{IV-20}$$

ii) This shows that even for small relative errors in b and c, the uncertainty in a_c will be large since it is multiplied by b/a (remember $b \approx c \gg a$).

iii) This subtraction cancellation pokes its ugly head in the series for e^{-x}.

2. Multiplicative Errors.

a) Error in computer multiplication arises in the following way:

$$a = b \times c \implies a_c = b_c \times c_c \ , \tag{IV-21}$$

$$\implies \frac{a_c}{a} = \frac{(1+\epsilon_b)(1+\epsilon_c)}{(1+\epsilon_a)} \approx 1 + \epsilon_b + \epsilon_c \ . \tag{IV-22}$$

Since ϵ_b and ϵ_c can have opposite signs, the error in a_c is sometimes larger and sometimes smaller than the individual errors in b_c and c_c.

b) Often, we can estimate an average roundoff error for a series of multiplications by assuming that the computer's representation of a number differs *randomly* from the actual number.

c) In these cases, we can use the **random walk** technique:

i) Let R be the average total distance covered in N steps each of length r, then

$$R \approx \sqrt{N}\, r \ . \tag{IV-23}$$

ii) Each step of a multiplication has a roundoff error of length ϵ_m, the machine's precision.

iii) In the random walk analogy, the average relative error ϵ_{ro} arising after a large number N of multiplicative steps is then

$$\boxed{\epsilon_{ro} \simeq \sqrt{N}\,\epsilon_m\ .} \qquad \text{(IV-24)}$$

We will be making use of Eq. (IV-24) many times in this course.

d) When roundoff errors do not occur randomly, careful analysis is needed to predict the dependence of the error on the number of steps N.

i) If there are no cancellations of error, the relative error may increase like $N\,\epsilon_m$.

ii) Some recursive algorithms where the production of errors is coherent (*e.g.*, upward recursive Bessel functions), the error increases like $N!\,\epsilon_m$.

e) This is something to keep in mind when you hear about computer calculations requiring hours to complete.

i) A fast computer may complete 10^{10} floating-point operations per second. Hence, a program running for 3 CPU hours performs about 10^{14} operations.

ii) Then, after 3 hours, we can expect roundoff errors to have accumulated to a relative importance of $10^7\,\epsilon_m$.

iii) For the error to be smaller than the answer, this demands that $\epsilon_m < 10^{-7}$.

iv) As such, this long of a calculation with 32-bit arithmetic (hence inherently possesses only 6 to 7 places of precision) probably contains much noise.

3. Definitions from Statistics and Probability Theory.

a) The **mean**, μ, of the parent population (*i.e.*, measurements) is defined as the limit of the sum N determinations x_i of the quantity x divided by the number N determinations.

$$\mu \equiv \lim_{N\to\infty} \left(\frac{1}{N} \sum_{i=1}^{N} x_i \right) \ . \qquad \text{(IV-25)}$$

The mean is therefore equivalent to the centroid or **average** value of the quantity x.

b) The **median**, $\mu_{1/2}$, of the parent population is defined as that value for which, in the limit of an infinite number of determinations x_i of the quantity x, half of the observations will be less than the median and half will be greater than the median.

i) In terms of the parent distribution, this means that the probability, P, is 50% that any measurement x_i will be large or smaller than the median:

$$P(x_i \le \mu_{1/2}) = P(x_i \ge \mu_{1/2}) = 50\% \ . \qquad \text{(IV-26)}$$

ii) Much computer time is wasted by figuring out median values. As such, fast sorting routines have been developed in many programming languages (*e.g.*, SORT in IDL). Then, the median is just the element in an array that is at the midway point.

c) The **most probable value**, $\mu_{\max}$, of the parent population is that value for which the parent distribution has greatest value:

$$P(\mu_{\max}) \geq P(x \neq \mu_{\max}) \ . \qquad \text{(IV-27)}$$

d) The **deviation** d_i of any measurement x_i from the mean μ of the parent distribution is defined as

$$d_i = x_i - \mu \ . \qquad \text{(IV-28)}$$

i) Deviations are generally defined with respect to the mean for computational purposes.

ii) If μ is the true value of the quantity, then d_i is the true error in x_i.

iii) The average deviation $\overline{d}$ for an infinite number of observations must vanish by virtue of the definition of the mean (see Eq. IV-25):

$$\begin{aligned} \lim_{N\to\infty} \overline{d} &= \lim_{N\to\infty} \left[\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu) \right] \\ &= \lim_{N\to\infty} \left(\frac{1}{N} \sum_{n=1}^{N} x_i \right) - \mu = \mu - \mu = 0 \ . \end{aligned}$$

iv) The **average deviation** α, therefore, is defined as the average of the magnitudes of the deviations, which are given by the absolute values of the deviations:

$$\alpha \equiv \lim_{N\to\infty} \left(\frac{1}{N} \sum_{n=1}^{N} |x_i - \mu| \right) \ . \qquad \text{(IV-29)}$$

v) The average deviation is a measure of the dispersion of the expected observations about the mean.

e) The **variance** σ^2 is defined as the limit of the average of the squares of the deviations of the mean:

$$\sigma^2 \equiv \lim_{N\to\infty}\left(\frac{1}{N}\sum_{n=1}^{N}(x_i-\mu)^2\right) = \lim_{N\to\infty}\left(\frac{1}{N}\sum_{n=1}^{N}x_i^2\right) - \mu^2 \ . \tag{IV-30}$$

i) The **standard deviation** σ is the square root of the variance.

ii) The standard deviation is thus the root mean square of the deviations, where we define the **root mean square** to be the square root of the mean or average of the square of an argument.

iii) In computational work, the standard deviation σ is considered an appropriate measure of the uncertainty of a measurement or a calculation.

4. Errors in Algorithms. An algorithm is often characterized by its step size h or by the number of steps N it takes to reach its goal. If the algorithm is "good," it should give an exact answer in the limit $h \to 0$ or $N \to \infty$. Here, we present methods for determining the error in your code.

a) Let's assume that an algorithm takes a large number of N steps to get a good answer and that the approximation error approaches zero like

$$\epsilon_{\text{apprx}} \simeq \frac{\alpha}{N^\beta} \ . \tag{IV-31}$$

b) In Eq. (IV-31), α and β are empirical constants that would change for different algorithms, and may be "constant" only for $N \to \infty$.

c) Meanwhile, the roundoff error keeps accumulating as you take more steps in a random fashion following

$$\epsilon_{ro} \simeq \sqrt{N}\,\epsilon_m \ , \tag{IV-32}$$

where ϵ_m is the machine precision (see Eq. IV-24).

d) The **total error** is just the sum of these two errors:

$$\epsilon_{\rm tot} = \epsilon_{\rm apprx} + \epsilon_{ro} \ , \tag{IV-33}$$

$$\simeq \frac{\alpha}{N^\beta} + \sqrt{N}\,\epsilon_m \ . \tag{IV-34}$$

e) Say we have a test problem where an analytic solution exists. By comparing a computed solution with the analytic solution, we can deduce the total error $\epsilon_{\rm tot}$.

i) If we then plot $\log(\epsilon_{\rm tot})$ against $\log(N)$, we can use the slope of this graph (*i.e.*, the power of N) to deduce which term in Eq. (IV-34) dominates the total error $\Longrightarrow$ if the slope $= 1/2$, roundoff error dominates, if not, the slope $= -\beta$ and the approximation error dominates.

ii) Alternatively, we can start at very large values of N where there should be no approximation error, and continuously lower N to deduce the slope in the approximation error.

5. Optimizing with Known Error Behavior. For this section, we will assume that the approximation error has $\alpha = 1, \beta = 2$, so

$$\epsilon_{\rm apprx} \simeq \frac{1}{N^2} \ . \tag{IV-35}$$

a) The extremum occurs when $d\epsilon_{\rm tot}/dN = 0$, which gives

$$-2N^{-3} + \frac{1}{2}N^{-1/2}\epsilon_m = 0 \ ,$$

$$\begin{aligned} N^{-1/2}\epsilon_m &= 4N^{-3} \ , \\ N^{5/2} &= \frac{4}{\epsilon_m} \ . \end{aligned} \qquad \text{(IV-36)}$$

b) For single precision on a 32-bit chip processor, $\epsilon_m \simeq 10^{-7}$, so the minimum total error is reached when

$$\begin{aligned} N^{5/2} &\simeq \frac{4}{10^{-7}} \\ N &\simeq 1099 \end{aligned}$$

and

$$\begin{aligned} \epsilon_{\rm tot} &\simeq \frac{1}{N^2} + \sqrt{N}\,\epsilon_m \\ &= (8 \times 10^{-7}) + (33 \times 10^{-7}) \simeq 4 \times 10^{-6} \ . \end{aligned}$$

c) This shows for a typical algorithm, most of the error is due to roundoff. Also, even though that this is the minimum error, the best we can do is to get some 40 times machine precision (in single precision).

d) As such, to reduce roundoff error, we should come up with a code that requires less steps (N) to achieve convergence.

6. Sections 3.10 and 3.11 of your textbook go into further details on optimizing error behavior and setting up an empirical error analysis. You should look these sections over in detail.

D. Propagation of Uncertainties.

1. We will demonstrate the propagation of errors with a simple example. Suppose we wish to find the volume V of a box of length L, width W, and height H.

a) We measure $L_\circ$, $W_\circ$, and $H_\circ$ and determine

$$V_\circ = L_\circ\, W_\circ\, H_\circ \ . \qquad \text{(IV-37)}$$

b) How do the uncertainties in $L_\circ$, $W_\circ$, and $H_\circ$ affect the uncertainty in $V_\circ$? Let L, W, H, and V be the actual (*i.e.*, 'true') value, then $\Delta L = L_\circ - L$, $\Delta W = W_\circ - W$, and $\Delta H = H_\circ - H$.

c) The error in V is approximately the sum of the products of the errors in each dimension times the effect that dimension has on the final value of V:

$$\Delta V \simeq \Delta L \left(\frac{\partial V}{\partial L}\right) + \Delta W \left(\frac{\partial V}{\partial W}\right) + \Delta H \left(\frac{\partial V}{\partial H}\right) \ , \quad \text{(IV-38)}$$

of for our example,

$$\Delta V \simeq W_\circ H_\circ \Delta L + L_\circ H_\circ \Delta W + L_\circ W_\circ \Delta H \ . \quad \text{(IV-39)}$$

Dividing this equation by Eq. (IV-37) gives

$$\frac{\Delta V}{V_\circ} \simeq \frac{\Delta L}{L_\circ} + \frac{\Delta W}{W_\circ} + \frac{\Delta H}{H_\circ} \ . \quad \text{(IV-40)}$$

2. In general, we do not know the actual errors in the determination of any of the parameters. The following describes how we estimate the uncertainties.

a) Let's define

$$x = f(u, v, \ldots) \ . \quad \text{(IV-41)}$$

b) Assume the most probable value for x is given by

$$\overline{x} = f(\overline{u}, \overline{v}, \ldots) \ . \quad \text{(IV-42)}$$

c) Individual results for x (x_i) are found by individual measurements of other parameters, that is u_i, v_i, ..., giving

$$x_i = f(u_i, v_i, \ldots) \ . \quad \text{(IV-43)}$$

d) Using Eq. (IV-30), the variance in x is then

$$\sigma_x^2 = \lim_{N\to\infty} \frac{1}{N} \sum (x_i - \overline{x})^2 \ . \quad \text{(IV-44)}$$

e) Following Eq. (IV-38), the deviation for measurement in x is

$$x_i - \overline{x} \simeq (u_i - \overline{u}) \left(\frac{\partial x}{\partial u}\right) + (v_i - \overline{v}) \left(\frac{\partial x}{\partial v}\right) + \cdots \quad \text{(IV-45)}$$

f) Combining these two equations gives

$$\begin{aligned}
\sigma_x^2 &= \lim_{N\to\infty} \frac{1}{N} \sum \left[(u_i - \overline{u}) \left(\frac{\partial x}{\partial u}\right) + (v_i - \overline{v}) \left(\frac{\partial x}{\partial v}\right) + \cdots \right]^2 \\
&= \lim_{N\to\infty} \frac{1}{N} \sum \left[(u_i - \overline{u})^2 \left(\frac{\partial x}{\partial u}\right)^2 + (v_i - \overline{v})^2 \left(\frac{\partial x}{\partial v}\right)^2 + \right. \\
&\qquad \left. 2(u_i - \overline{u})(v_i - \overline{v}) \left(\frac{\partial x}{\partial u}\right) \left(\frac{\partial x}{\partial v}\right) + \cdots \right] \\
&= \sigma_u^2 \left(\frac{\partial x}{\partial u}\right)^2 + \sigma_v^2 \left(\frac{\partial x}{\partial v}\right)^2 + \\
&\qquad 2\sigma_{uv}^2 \left(\frac{\partial x}{\partial u}\right) \left(\frac{\partial x}{\partial v}\right) + \cdots \ , \qquad \text{(IV-46)}
\end{aligned}$$

where

$$\sigma_{uv}^2 \equiv \lim_{N\to\infty} \frac{1}{N} \sum \left[(u_i - \overline{u})(v_i - \overline{v}) \right] \ . \qquad \text{(IV-47)}$$

g) If measurements in u and v are uncorrelated, one should get as many negative values as positive values for the terms in the series of Eq. (IV-47). As such, this summation in Eq. (IV-47) $\to 0$. So

$$\sigma_x^2 = \sigma_u^2 \left(\frac{\partial x}{\partial u}\right)^2 + \sigma_v^2 \left(\frac{\partial x}{\partial v}\right)^2 + \cdots \qquad \text{(IV-48)}$$

3. Specific Formulas.

a) **Addition and Subtraction.** Let x be the weighted sum of u and v:

$$x = au \pm bv \ . \qquad \text{(IV-49)}$$

i) Then, the partial derivatives are simply the weighting coefficients (which are constant):

$$\left(\frac{\partial x}{\partial u}\right) = a \ , \quad \left(\frac{\partial x}{\partial v}\right) = \pm b \ . \qquad \text{(IV-50)}$$

ii) Eq. (IV-46) then becomes

$$\sigma_x^2 = a^2\sigma_u^2 + b^2\sigma_v^2 + 2ab\sigma_{uv}^2 \ . \qquad \text{(IV-51)}$$

b) Multiplication and Division. Now let x be the weighted product of u and v:

$$x = \pm auv \ . \qquad \text{(IV-52)}$$

i) The partial derivatives of each variable contain the values of the other variable:

$$\left(\frac{\partial x}{\partial u}\right) = \pm av \ , \quad \left(\frac{\partial x}{\partial v}\right) = \pm au \ . \qquad \text{(IV-53)}$$

ii) Eq. (IV-46) yields

$$\sigma_x^2 = a^2v^2\sigma_u^2 + a^2u^2\sigma_v^2 + 2a^2uv\sigma_{uv}^2 \ , \qquad \text{(IV-54)}$$

which can be expressed more symmetrically as

$$\frac{\sigma_x^2}{x^2} = \frac{\sigma_u^2}{u^2} + \frac{\sigma_v^2}{v^2} + 2\frac{\sigma_{uv}^2}{uv} \ . \qquad \text{(IV-55)}$$

iii) Similarly, if x is obtained through division,

$$x = \pm\frac{au}{v} \ , \qquad \text{(IV-56)}$$

the variance for x is given by

$$\frac{\sigma_x^2}{x^2} = \frac{\sigma_u^2}{u^2} + \frac{\sigma_v^2}{v^2} - 2\frac{\sigma_{uv}^2}{uv} \ . \qquad \text{(IV-57)}$$

c) Powers. Let x be obtained by raising the variable u to a power:

$$x = au^{\pm b} \ . \qquad \text{(IV-58)}$$

i) The partial derivative of x with respect to u is

$$\left(\frac{\partial x}{\partial u}\right) = \pm abu^{\pm b-1} = \pm\frac{bx}{u} \ . \qquad \text{(IV-59)}$$

ii) Eq. (IV-46) becomes

$$\frac{\sigma_x}{x} = b\frac{\sigma_u}{u} \ . \qquad \text{(IV-60)}$$

d) Exponentials. Let x be obtained by raising the natural base to a power proportional to u:

$$x = ae^{\pm bu} \ . \qquad \text{(IV-61)}$$

i) The partial derivative of x with respect to u is

$$\left(\frac{\partial x}{\partial u}\right) = \pm abe^{\pm bu} = \pm bx \ . \qquad \text{(IV-62)}$$

ii) Eq. (IV-46) becomes

$$\frac{\sigma_x}{x} = b\sigma_u \ . \qquad \text{(IV-63)}$$

e) Logarithms. Let x be obtained by taking a log of u:

$$x = a\ln(\pm bu) \ . \qquad \text{(IV-64)}$$

i) The partial derivative of x with respect to u is

$$\left(\frac{\partial x}{\partial u}\right) = \frac{a}{u} \ . \qquad \text{(IV-65)}$$

ii) Eq. (IV-46) becomes

$$\sigma_x = a\frac{\sigma_u}{u} \ , \qquad \text{(IV-66)}$$

which is essentially the inverse of Eq. (IV-63).