Some comments on Differentials and Errors

John McCuan

December 1, 2019

In the main, section 14.6 of Thomas' Calculus (fourteenth edition) is well-written, and I'm not going to give anything like a full exposition here. On the other hand, problems 62 and 63 of the Chapter 14 Practice Problems bring out some ambiguities which should perhaps be addressed. We'll stick to two variables. The first order Taylor approximation in that case can be written as

$$f(x,y) \sim f(x_0,y_0) + \frac{\partial f}{\partial x}(x_0,y_0)(x-x_0) + \frac{\partial f}{\partial y}(x_0,y_0)(y-y_0).$$

The function on the right

$$L(x,y) = f(x_0,y_0) + \frac{\partial f}{\partial x}(x_0,y_0)(x-x_0) + \frac{\partial f}{\partial y}(x_0,y_0)(y-y_0), \tag{1}$$

is identified as the **linearization** on page 857, and on the next page the **error** is implicitly defined as the difference

$$E(x,y) = f(x,y) - L(x,y).$$

Finally, the **differential** is defined at the bottom of page 858 by

$$df = \frac{\partial f}{\partial x}(x_0,y_0)\,dx + \frac{\partial f}{\partial y}(x_0,y_0)\,dy. \tag{2}$$

This is all somewhat standard, though I would present it somewhat differently. Let me mention some aspects of my perspective, which especially make sense if one is familiar with linear algebra.

It will be noted that the *linearization* appearing in (1) is not a linear function. I would call it the **affine approximation** of f and perhaps write

$$\alpha(x,y) = f(x_0,y_0) + \frac{\partial f}{\partial x}(x_0,y_0)(x-x_0) + \frac{\partial f}{\partial y}(x_0,y_0)(y-y_0). \tag{3}$$

The differential, on the other hand, is really an honest linear function $df : \mathbb{R}^2 \to \mathbb{R}^1$ given by a dot product. It's just that the variables have been given sort of weird names. To make it look more familiar, we can write it like this:

$$df(u,v) = Df(x_0,y_0)\cdot(u,v) = \frac{\partial f}{\partial x}(x_0,y_0)\,u + \frac{\partial f}{\partial y}(x_0,y_0)\,v.$$

Note that $df(0,0) = 0$ while $L(0,0) = f(x_0,y_0)$ does not necessarily vanish.

Exercise 1 *(Recall that) a linear function $L : \mathbb{R}^n \to \mathbb{R}^m$ is one for which $L(a\mathbf{v} + b\mathbf{w}) = aL(\mathbf{v}) + bL(\mathbf{w})$ for all $a, b \in \mathbb{R}$ and all $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$. Show that any such linear function L is given by matrix multiplication*

$$L(\mathbf{v}) = A\mathbf{v}$$

where A is an $m \times n$ matrix and in the special case when $m = 1$ this reduces to a dot product. Conclude that the image under L of the zero vector in $\mathbb{R}^n$ is the zero vector in $\mathbb{R}^m$. In particular, when $m = 1$, $L(0,0,\ldots,0) = 0$, so the function L called the linearization above *is not linear.*

An affine function is a linear function plus a shift: $\alpha(\mathbf{v}) = L(\mathbf{v}) + \mathbf{w}_0$; if $L : \mathbb{R}^n \to \mathbb{R}^m$, then $\mathbf{v} \in \mathbf{r}^n$ and $\mathbf{w}_0 \in \mathbf{r}^m$ in this formula.

In any case, one is typically interested in using the values of the affine approximation (linearization) to approximate the values of the function, but problems 14.6.51-56 implicitly introduce a different aspect/kind of approximation and the Practice Problems 59-64, and 62 and 63 in particular, on page 893 make this explicit. Problem 62 mentions the **error in estimating the area of an ellipse**. This suggests, on the one hand, that one estimates the error using the affine approximation based on estimate near the top of page 858 involving the second partials, and that is sort of fine. Let us carry out the problem from this point of view.

The formula for the area of the ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

is $f(a,b) = \pi ab$. Therefore, the affine approximation is given by

$$\alpha(a,b) = \pi a_0 b_0 + Df(a_0,b_0)\cdot(a-a_0, b-b_0) = \pi[a_0b_0 + b_0(a-a_0) + a_0(b-b_0)].$$

We are told that $a = 10$ cm and $b = 16$ cm to the nearest millimeter. A natural interpretation of this information is that there are some **actual values** a_0 and b_0, and

these are **measured values** associated with some **potential error in measurement** for which

$$|a - a_0| < 0.1 \text{ cm} \qquad \text{and} \qquad |b - b_0| < 0.1 \text{ cm}.$$

From this, we have all the information needed to compute the error in the **affine approximation/estimation** of the area of the ellipse from the estimate on page 858: Note that the second partials of $f = \pi ab$ are all bounded by $M = \pi$, thus[1]

$$\begin{aligned} |f(a,b) - f(a_0,b_0) - Df(a_0,b_0)\cdot(a-a_0, b-b_0)| &\leq \frac{M}{2}[(a-a_0)^2 + (b-b_0)^2] \\ &= \frac{\pi}{2}[(0.1)^2 + (0.1)^2] \\ &\leq (0.01)\pi \text{ cm}^2. \end{aligned}$$

To get the percentage error we should then consider

$$\frac{|f(a,b) - f(a_0,b_0) - Df(a_0,b_0)\cdot(a-a_0, b-b_0)|}{f(a_0,b_0)}.$$

This is the ratio of the error to the actual size of the approximated quantity. Of course, in practice—and in this problem—we do not know the actual value of the area $f(a_0, b_0)$. There are a couple things we can do. One is that we can simply take the lowest possible value for $f(a_0, b_0) = f(9.9, 15.9) = 157.41\pi$. This gives an error ratio of

$$\frac{0.01}{157.41} \sim 0.000635$$

or about 0.06 %. Alternatively, we might also, in practice, be a bit more sloppy and just consider the percentage error with respect to the size of our measured (approximate) values:

$$\frac{|f(a,b) - f(a_0,b_0) - Df(a_0,b_0)\cdot(a-a_0, b-b_0)|}{f(a,b)} \leq \frac{0.01}{160} \sim 0.000625.$$

[1]The book does not go into the proof of this error estimate at this point, but it is easy to obtain from the first order Taylor expansion formuala **with remainder** which reads

$$f(x,y) = f(x_0,y_0) + Df(x_0,y_0)\cdot(x-x_0, y-y_0) + \frac{1}{2}D^2f(x^*,y^*)(x-x_0, y-y_0)\cdot(x-x_0, y-y_0)$$

where (x^*, y^*) is some point in a convex domain containing (x, y) and (x_0, y_0) and D^2f is the usual Hessian matrix of second partials. A justification for the error estimate on page 858 is given in section 14.9.

So you see, we get about the same answer either way.

On the other hand, the problem itself asks about the percentage error in the **calculated area**, which is a different thing. In fact, in practice, the error in the linear/affine approximation we have calculated above is of no use to us since we do not know the zero order term $f(a_0, b_0)$. Thus, the **error** that makes sense in this problem is that which is pretty clearly asked for at the end of the statement—though the introductory title **Maximum error in estimating the area of an ellipse** is ambiguous.[2]

What makes sense is the maximum value of

$$|f(a, b) - f(a_0, b_0)|$$

alone rather than $|f(a, b) - [f(a_0, b_0) + Df(a_0, b_0) \cdot (a - a_0, b - b_0)]|$. For this, we need estimates on first derivatives rather than second derivatives. In fact, the **zero order Taylor expansion with remainder** reads

$$f(x, y) = f(x_0, y_0) + Df(x^*, y^*) \cdot (x - x_0, y - y_0)$$

where (x^*, y^*) is a point on the line segment joining (x, y) and (x_0, y_0). (This line segment is assumed to lie in the domain of f.) Applying this in our case, we see

$$|f(a, b) - f(a_0, b_0)| \leq \pi |(b^*, a^*) \cdot (a - a_0, b - b_0)| \leq \pi[(10.1)(0.1) + (16.1)(0.1)] = 2.62\pi.$$

Therefore, the desired percentage error in estimation is given by

$$\frac{|f(a, b) - f(a_0, b_0)|}{f(a_0, b_0)} \leq \frac{2.62}{157.41} \sim 0.0166$$

or about 1.7 %. If, as before, we substitute the exact measured values as the reference calculation, we get

$$\frac{|f(a, b) - f(a_0, b_0)|}{f(a, b)} \leq \frac{2.62}{160} \sim 0.0164$$

so about 1.6 %. Almost the same thing.

Now we can move on to problem number 63, the next problem, where things really go off the rails. The heading says **Error in estimating a product**. As we know from the previous problem, what is presumably intended is **Error in calculating a product**. That is, we have some error in measuring values u and v, so that the

[2] A better heading might be **Maximum error in *calculating* the area of an ellipse**.

tolerances, or maximum increments, with respect to some (unknown) exact values are known:

$$|u - u_0| < \epsilon_1 \qquad \text{and} \qquad |v - v_0| < \epsilon_2. \tag{4}$$

And we are asked to determine the largest possible error in the calculated product $y = uv$, as compared to the (unknown) exact product $u_0 v_0$. It is also mentioned that we are restricting attention to only **positive** values of u and v. In practice, we are not given increments as in (4) but rather **percentage error** in measurement. This means something like this:

$$\frac{|u - u_0|}{u_0} < p_1 = 0.02 \qquad \text{and} \qquad \frac{|v - v_0|}{v_0} < p_2 = 0.03. \tag{5}$$

That is, u is known/measured within a 2 % error and v, likewise, is measured within a 3 % error. One might also, under the assumption that u and u_0 are relatively close together, interpret the given measurement tolerances as

$$\frac{|u - u_0|}{u} < q_1 = 0.02 \qquad \text{and} \qquad \frac{|v - v_0|}{v} < q_2 = 0.03. \tag{6}$$

Exercise 2 *Given (5), what values are implied for (6)? Show for example that*

$$\frac{|u - u_0|}{u} < q_1 = \frac{p_1}{1 - p_1} \qquad \textit{and if } p_0 = 0.02, \textit{ then} \qquad \frac{p_1}{1 - p_1} = 0.98 p_1.$$

Hint: If $u < u_0$, *then* $|u - u_0| = u_0 - u < u_0 p_1$ *and*

$$\frac{|u - u_0|}{u} < \frac{|u - u_0|}{(1 - p_1) u_0}.$$

In any case, we now have at least four different notions of **error** floating about. There is the **error of first order approximation** discussed in sections 14.6 and 14.9 associated with the **first order Taylor expansion formula with remainder**:

$$f(\mathbf{x}) = f(\mathbf{x}_0) + Df(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + \left\langle D^2 f(\mathbf{x}^*)(\mathbf{x} - \mathbf{x}_0), \mathbf{x} - \mathbf{x}_0 \right\rangle$$

where

$$\begin{aligned} \left\langle D^2 f(\mathbf{x}^*)(\mathbf{x} - \mathbf{x}_0), \mathbf{x} - \mathbf{x}_0 \right\rangle &= \sum_{\mathbf{e}_{ij}} \frac{1}{\mathbf{e}_{ij}!} \frac{\partial^2 f}{\partial x_i, \partial x_j}(\mathbf{x}^*) \, (\mathbf{x} - \mathbf{x}_0)^{\mathbf{e}_{ij}} \\ &= \frac{1}{2} \left(\sum_i \frac{\partial^2 f}{\partial x_i^2}(\mathbf{x}^*) \, (x_i - x_{0,i})^2 \right. \\ &\qquad \left. + 2 \sum_{i<j} \frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}^*) \, (x_i - x_{0,i})(x_i - x_{0,j}) \right), \end{aligned}$$

and $\mathbf{x}^*$ is some point on the segment between $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{x}_0 = (x_{0,1}, \ldots, x_{0,n})$. We recall the multi-index conventions for $\mathbf{e}_{ij} = \mathbf{e}_i + \mathbf{e}_j \in \mathbb{N}^n = \{(i_1, \ldots, i_n) : i_k \in \{0, 1, 2, \ldots\} \text{ for } k = 1, 2, \ldots, n\}$:

$$\mathbf{e}_{ij}! = i!j! \qquad \text{and} \qquad \mathbf{v}^{\mathbf{e}_{ij}} = v_i v_j.$$

Notice that the actual error here (or remainder) involves second order derivatives. **In order to bound the error of a first order Taylor approximation, you use second derivatives.** A justification of the Taylor formula with remainder is as follows:[3]

Let $\mathbf{x}_0$ and $\mathbf{x}$ be fixed with all the points $(1-t)\mathbf{x}_0 + t\mathbf{x}$ in the domain of f. Notice that these are the points on the line segment from $\mathbf{x}_0$ to $\mathbf{x}$. In fact, $\ell(t) = (1-t)\mathbf{x}_0 + t\mathbf{x}$ parameterizes this segment on the interval $0 \leq t \leq 1$ with $\ell(0) = \mathbf{x}_0$ and $\ell(1) = \mathbf{x}$. Consider the function

$$g(t) = f(\mathbf{x}_0) + Df(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0)t + [f(\mathbf{x}) - f(\mathbf{x}_0) - Df(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0)]t^2,$$

and the difference $\delta(t) = f((1-t)\mathbf{x}_0 + t\mathbf{x}) - g(t)$. Notice that $g(0) = f(\mathbf{x}_0)$, so $\delta(0) = 0$. Also, $g(1) = f(\mathbf{x})$, so $\delta(1) = 0$. By the mean value theorem, there is some t_1 such that $\delta'(t_1) = 0$. This means

$$g'(t_1) = Df(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + 2[f(\mathbf{x}) - f(\mathbf{x}_0) - Df(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0)]t_1 = 0,$$

and $\delta'(0) = Df(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0) - g'(0) = 0 = \delta'(t_1)$. Therefore, considering the function $\delta' : [0, t_1] \to \mathbb{R}$, we can apply the mean value thereom again to conclude there is some t_* with $0 < t_* < t_1$ such that $\delta''(t_*) = 0$. Computing $\delta''(t_*)$ we conclude

$$\begin{aligned}\langle D^2 f((1-t_*)\mathbf{x}_0 + t_*\mathbf{x})(\mathbf{x} - \mathbf{x}_0), \mathbf{x} - \mathbf{x}_0\rangle - g''(t_*) = \\ \langle D^2 f((1-t_*)\mathbf{x}_0 + t_*\mathbf{x})(\mathbf{x} - \mathbf{x}_0), \mathbf{x} - \mathbf{x}_0\rangle \\ - 2[f(\mathbf{x}) - f(\mathbf{x}_0) - Df(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0)] \\ = 0.\end{aligned}$$

This means there is a point $\mathbf{x}^* = (1-t_*)\mathbf{x}_0 + t_*\mathbf{x}$ on the segment connecting $\mathbf{x}_0$ and $\mathbf{x}$ for which

$$f(\mathbf{x}) - f(\mathbf{x}_0) - Df(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0) = \frac{1}{2}\langle D^2 f(\mathbf{x}^*)(\mathbf{x} - \mathbf{x}_0), \mathbf{x} - \mathbf{x}_0\rangle,$$

[3]There is some version of this in section 14.9, and we will also use/review an argument from the end of section 10.9 in Chapter 10.

and this is the first order Taylor formula with remainder.

So that's one kind of error estimation. In addition, we are interested in the **zero order Taylor approximation formula with remainder**. That is,

$$f(\mathbf{x}) = f(\mathbf{x}_0) + Df(\mathbf{x}^*) \cdot (\mathbf{x} - \mathbf{x}_0).$$

Exercise 3 *Use the mean value theorem, as we did above in the first order case, to prove the zero order Taylor formula with remainder.*

Finally, there is **error in measurement** and the resulting **error in calculation** of a formula like $y = uv$. The error in measurement has already been interpreted in terms of inequalities involving increments and as percentages of the value measured. We can apply a similar interpretation for the values of a function like $y = uv$. The error in calculation would be

$$|uv - u_0v_0|$$

where as above u_0 and v_0 are some (unknown) actual exact values and u_0v_0 is the (unknown) exact value of the calculation. Also, the quantities

$$\frac{|uv - u_0v_0|}{u_0v_0} \quad \text{or} \quad \frac{|uv - u_0v_0|}{uv}$$

may be used to represent the percentage error in such a calculation. I think it's pretty clear what is required/expected in part (a) of this problem. We are given (5), and and it is desired for us to use the zero order approximation formula to estimate the error in the calculation of y like this:

$$\begin{aligned}
\frac{|uv - u_0v_0|}{u_0v_0} &= \frac{|Dy(u^*, v^*) \cdot (u - u_0, v - v_0)|}{u_0v_0} \\
&= \frac{|(v^*, u^*) \cdot (u - u_0, v - v_0)|}{u_0v_0} \\
&= \frac{|(v^*(u - u_0) + u^*(v - v_0)|}{u_0v_0} \\
&\leq \frac{v^*}{v_0}\frac{|u - u_0|}{u_0} + \frac{u^*}{u_0}\frac{|v - v_0|}{v_0} \\
&\leq \frac{v^*}{v_0}p_1 + \frac{u^*}{u_0}p_2. \qquad (7)
\end{aligned}$$

At this point, we are supposed to assume $u^*/u_0 \sim 1$ and $v^*/v_0 \sim 1$ so that the error is

$$\frac{|uv - u_0v_0|}{u_0v_0} \sim p_1 + p_2 = 0.02 + 0.03 = 0.05$$

where "$\sim$" here means "roughly no more than." Thus, we get the 5 % answer in the back of the book. Notice that this last step is a bit sloppy.

Exercise 4 *Start with (7) and use the approach of Exercise 2 involving estimates for the quotients u^*/u_0 and v^*/v_0 carefully, and see if you can still get* 5 *% to the nearest tenth.*

Let's move on to part (b). We are asked to consider a second function $z = u + v$, which we note is linear in u and v, and compare the percentage errors in the calculated values of the functions y and z. In particular, we are supposed to show the percentage error in z is less than the percentage error in y. This assertion seems to simply be false. Nevertheless, it is more or less clear what one is expected to do. That is, the incorrect reasoning of—and expected by—the person who wrote the problem can be easily guessed. Remember the percentage error is represented by

$$\begin{aligned}\frac{|uv - u_0v_0|}{u_0v_0} &= \frac{|(v^*(u-u_0) + u^*(v-v_0)|}{u_0v_0} \\ &\leq \frac{v^*}{v_0}\frac{|u-u_0|}{u_0} + \frac{u^*}{u_0}\frac{|v-v_0|}{v_0} \\ &< \frac{v^*}{v_0}\, p_1 + \frac{u^*}{u_0}\, p_2.\end{aligned}$$

The corresponding percentage error for the sum $z = u + v$ is

$$\begin{aligned}\frac{|u+v-u_0-v_0|}{u_0+v_0} &= \frac{|u-u_0+v-v_0|}{u_0+v_0} \\ &\leq \frac{|u-u_0|}{u_0+v_0} + \frac{|v-v_0|}{u_0+v_0} \\ &= \frac{u_0}{u_0+v_0}\,\frac{|u-u_0|}{u_0} + \frac{v_0}{u_0+v_0}\,\frac{|v-v_0|}{v_0} \\ &< \frac{u_0}{u_0+v_0}\, p_1 + \frac{v_0}{u_0+v_0}\, p_2.\end{aligned}$$

It is clear, moreover, that we can continue the last approximation/estimation to obtain

$$\frac{|u+v-u_0-v_0|}{u_0+v_0} < p_1 + p_2.$$

Therefore, if we (by whatever means) interpret the "percentage error" in the measurement of y to be $p_1 + p_2$, then we have "shown" what the problem asks us to show.

There are several problems with this. At the heart of those problems is basically this:

> *If you show $A < B$ and $C < D$, then the fact that $B < D$ does* **not** *imply $A < C$.*

To be very explicit, let us imagine that u_0 and v_0 are both 1. But our measurements are $u = 2$ and $v = 1/2$. Then the percentage error in the product is

$$\frac{|uv - u_0v_0|}{u_0v_0} = 0. \tag{8}$$

The supposedly "smaller" percentage error in the sum $z = u + v$ is

$$\frac{|u + v - u_0 - v_0|}{u_0 + v_0} = \frac{1/2}{2} = 0.25$$

or 25 %!

There is (yet!) a fifth notion of error mentioned in the text. This appears in the discussion of differentials on page 858. That is the **increment of the linearization**. Thus, in groping for a way to devine what the author of this problem had in mind, we might imagine taking as representative of the percentage error, that produced by the error in the affine approximation. This would mean considering for the product $y = uv$ the function $\alpha(u, v) = u_0v_0 + (v_0, u_0) \cdot (u - u_0, v - v_0)$ and the quantity

$$\frac{|\alpha(u, v) - \alpha(u_0, v_0)|}{u_0v_0} = \frac{|v_0(u - u_0) + u_0(v - v_0)|}{u_0v_0} \tag{9}$$

in comparison to the corresponding quantity for the sum, which since the sum is linear, is just

$$\frac{|u + v - u_0 - v_0|}{u_0 + v_0}$$

as above. Writing these as

$$\left|\frac{u - u_0}{u_0} + \frac{v - v_0}{v_0}\right| \qquad \text{and} \qquad \left|\frac{u - u_0}{u_0 + v_0} + \frac{v - v_0}{u_0 + v_0}\right|$$

we see that putting $u_0 = 1 = v_0$ and $u = 2$, $v = 1/2$ as before, the two quantities are $1/2$ and $1/4$ respectively, so this "affine error" for the sum is less than the "affine error" for the product, and it looks like there may be some hope to show

$$\left|\frac{u - u_0}{u_0 + v_0} + \frac{v - v_0}{u_0 + v_0}\right| \leq \left|\frac{u - u_0}{u_0} + \frac{v - v_0}{v_0}\right|.$$

Putting $u_0 = 2$ and $v_0 = 1$ with $u = 3$ and $v = 1/2$, however, gives

$$\left|\frac{u-u_0}{u_0} + \frac{v-v_0}{v_0}\right| = |1/2 - 1/2| = 0$$

for the "affine error" for the product and

$$\left|\frac{u-u_0}{u_0+v_0} + \frac{v-v_0}{u_0+v_0}\right| = |1/3 - 1/6| = 1/6 > 0$$

for the "affine error" for the sum. The bottom line is that there does not seem any reasonable way to make sense of part (b) of this problem. No solution is given for this part in the back of the book. I haven't checked the solutions manual.

Okay, I checked the solutions manual, and the error there is even worse. Basically, all increments are assumed positive. So the problem in the offered solution of 63 part (b) is that it is always assumed $u - u_0 \geq 0$ and $v - v_0 \geq 0$. Based on this, the error for the sum is expressed as

$$\frac{u-u_0}{u_0+v_0} + \frac{v-v_0}{u_0+v_0}$$

(without the absolute values). And the "error" for the product is

$$\frac{u-u_0}{u_0} + \frac{v-v_0}{v_0}$$

(again without the absolute values). Then—if you incorrectly assume $u - u_0$ and $v - v_0$ are positive, ie., that all measurements are higher than the exact values—then indeed

$$\frac{u-u_0}{u_0+v_0} + \frac{v-v_0}{u_0+v_0} \leq \frac{u-u_0}{u_0} + \frac{v-v_0}{v_0}.$$

(Of course, it is mentioned as justification that the assumption of positivity for u and v is used, but more than that is needed.)

Incidentally, the explanation given in the solution manual for part (a) uses what we've called the affine error for the product, and then the solution is very clean:

$$\left|\frac{u-u_0}{u_0} + \frac{v-v_0}{v_0}\right| \leq \left|\frac{u-u_0}{u_0}\right| + \left|\frac{v-v_0}{v_0}\right| \leq p_1 + p_2. \tag{10}$$

Here, the application of the triangle inequality is correct, so there is no problem with assuming the increments are positive.

On the other hand, I see that technically the argument given is somewhat different (though more or less equivalent to) what we've done in (10) . Let's think carefully about what is done there because something like it will come up naturally later. Starting with the notation directly from the solutions manual:

$$dy = v\,du + u\,dv \qquad \text{and} \qquad \frac{dy}{y} = \frac{du}{u} + \frac{dv}{v}.$$

Translating to our notation the first equation would be

$$dy(u - u_0, v - v_0) = v(u - u_0) + u(v - v_0).$$

This is not the correct formula for the linear part of the product $y = uv$ at the actual values (u_0, v_0) which would be (what we have used above, namely)

$$dy(u - u_0, v - v_0) = v_0(u - u_0) + u_0(v - v_0).$$

So what has been done here? Apparently, the gradient was evaluated at the known (measured) values which are overall assumed "close" to the unknown actual values:

$$dy = dy(u - u_0, v - v_0) \sim Dy(u, v) \cdot (u - u_0, v - v_0) = v(u - u_0) + u(v - v_0).$$

Then a comparison is made to the known (measured) product uv for the percentage:

$$\left|\frac{dy}{y}\right| \sim \left|\frac{v(u - u_0) + u(v - v_0)}{uv}\right| < 0.05.$$

Of course, nothing is mentioned about the (double) approximation

$$\frac{dy}{y} = \frac{v_0(u - u_0) + u_0(v - v_0)}{u_0 v_0} \sim \frac{v_0(u - u_0) + u_0(v - v_0)}{uv} \sim \frac{v(u - u_0) + u(v - v_0)}{uv}.$$

Again, one could give precise estimates for the errors in these approximations.

Exercise 5 *Taking careful account of both of these approximations, what can you say about the percentage of affine error calculated by the method in the solutions manual?*

Just to be clear, the percentage of affine error we have in mind here is represented by

$$\frac{|v_0(u - u_0) + u_0(v - v_0)|}{u_0 v_0},$$

which we can estimate directly under the assumptions $|u-u_0|/u_0 < 0.02$ and $|v-v_0|/v_0 < 0.03$. The quantity estimated in the solutions manual,

$$\frac{|v(u-u_0)+u(v-v_0)|}{uv},$$

is an approximation of the percentage of affine error which is approximated there under the assumptions $|u-u_0|/u < 0.02$ and $|v-v_0|/v < 0.03$. The question of whether to evaluate the gradient at the exact values (u_0, v_0), which happen to be unknown, or to consider some kind of strange affine approximation obtained by evaluating the gradient at the approximate measured values (u, v), which have the virtue of being known, will come up below.

Saying more

While we have shown some shortcomings in the composition of Problem 63, especially part (b), when looking at the expressions

$$\frac{|uv-u_0v_0|}{u_0v_0} = \frac{|(v^*(u-u_0)+u^*(v-v_0)|}{u_0v_0} = \left|\frac{v^*}{v_0}\ \frac{u-u_0}{u_0} + \frac{u^*}{u_0}\ \frac{v-v_0}{v_0}\right| \tag{11}$$

and

$$\left|\frac{u-u_0}{u_0+v_0} + \frac{v-v_0}{u_0+v_0}\right| \tag{12}$$

for the percentage error for the product $y = uv$ and the sum $z = u+v$ respectively, we have the feeling that under some (or most) circumstances the second (percentage) error should be smaller. This is based on the larger denominators appearing in the second (percentage) error. Of course, we know it is not *always* true that the (percentage) error in the sum calcuation is less than the (percentage) error in the product calculation, but still we can ask:

> *When or how much (for how many values both exact and measured) does the desired inequality fail?*

On the face of it, this is a pretty hard question, and here is why: If we call the percentage error in the product P_p and the percentage error in the sum P_s, then these represent two functions of **four** variables:

$$P_p : (0,\infty)^4 \to \mathbb{R} \quad \text{with } P_p = P_p(u, v, u_0, v_0) \text{ by the formula (11).}$$

And we have a similar representation for the sum with $P_s = P_s(u, v, u_0, v_0)$ defined on the same region of $\mathbb{R}^4$ using (12). One expects it is very difficult to compare two complicated functions on a big open set in $\mathbb{R}^4$. Nevertheless, perhaps something more can be said, so let's try to say something more.

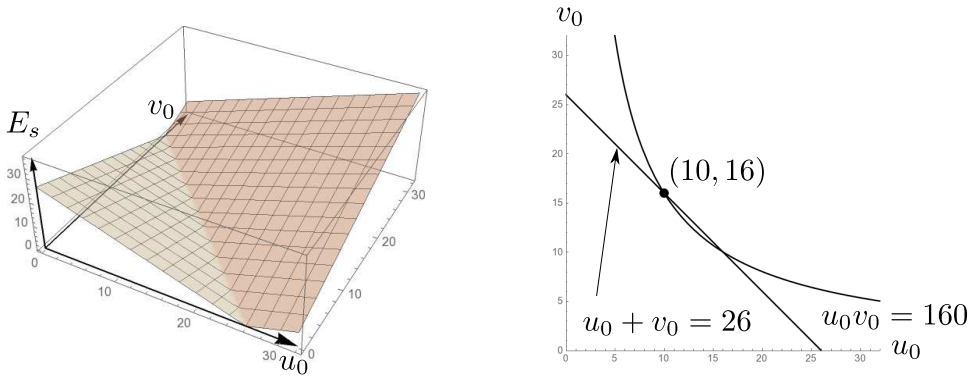

Figure 1: The graph of the error E_s in calcuation of a sum $u_0 + v_0$ when one takes/measures $u = 10$ and $v = 16$ (left). Zero level sets for E_s and E_p (right).

First of all, we are given no specific values for u and v (like $a = 10$ and $b = 16$ in Prolem 62), but we could take some specific values and reduce our consideration to two functions of two variables u_0 and v_0. Moreover, we may simplify the functions under consideration by using different measurements for error, like the absolute error, rather than the percentage error. Similarly, we could have a look at the affine percentage error. Let's try to look at all of these at least in some special cases. Let us start with the simplest, which is the actual value of the error. Setting

$$E_p = |uv - u_0 v_0| \qquad \text{and} \qquad E_s = |u + v - u_0 - v_0|,$$

if we set $u = 10$ and $v = 16$ these quantities become $E_p = |160 - u_0 v_0|$ and $E_s = |26 - u_0 - v_0|$. Both quantities are zero at $(10, 16)$ and non-negative elsewhere. Let us understand this error $E_s = E_s(u_0, v_0)$ for the sum first; it is the easier/simpler one. The graph is piecewise planar as indicated on the left in Figure 1. Obviously, the zero level set Z_s is the line $v_0 = 26 - u_0$. Let us note, finally, that over the region $u_0 + v_0 < 26$ in the first quadrant, the error E_s for the sum is given by $E_s = 26 - u_0 - v_0$, and in the complementary region $26 < u_0 + v_0$ we have $E_s = u_0 + v_0 - 26$.

The graph of E_p for the error in the product is somewhat more complicated. The zero level set Z_p for E_p is not a straight line, but the curve $u_0 v_0 = 160$. This is a

convex curve, and we know it intersects the line $u_0 + v_0 = 26$ in the points $(10, 16)$ and $(16, 10)$. By convexity, these are the only points of intersection of Z_p with Z_s as indicated on the right in Figure 1. Furthermore, $E_p = 160 - u_0 v_0$ in the region where $u_0 v_0 < 160$ and $E_p = u_0 v_0 - 160$ in the complementary region where $u_0 v_0 > 160$. Taking these considerations into account as illustrated in Figure 1 (right), we can try to understand the set of points (u_0, v_0) where $E_p = E_s$. From that, we should be able to understand the entire situation.

First of all, if we look for such points in $R_1 = \{(u_0, v_0) : u_0 + v_0 < 26 \text{ and } u_0 v_0 < 160\}$, then we want

$$26 - u_0 - v_0 = 160 - u_0 v_0 \qquad \text{or} \qquad u_0 v_0 - u_0 - v_0 = 134 \qquad \text{or} \qquad v_0 = \frac{u_0 + 134}{u_0 - 1}.$$

This represents a convex graph in the first quadrant which we know also passes through $(10, 16)$ and $(16, 10)$. For $u_0 < 10$ we find

$$v_1 = \frac{u_0 + 134}{u_0 - 1} > \frac{160}{u_0}$$

since $u_0^2 + 134 u_0 > 160 u_0 - 160$. (This is because $u_0^2 - 26 u_0 + 160 = (u_0 - 10)(u_0 - 16)$ has zeros at $u_0 = 10$ and $u_0 = 16$.) For essentially the same reason, we obtain a convex curve connecting $(10, 16)$ to $(16, 10)$ in the region R_1. This separates R_1 into two regions, a large one

$$R_{1a} = \{(u_0, v_0) : u_0 + v_0 < 26 \text{ and } v_0 < v_1(u_0)\}$$

containing $(u_0, v_0) = (1, 1)$ in which $E_s < E_p$, and a second very thin region

$$R_{1b} = \{(u_0, v_0) : 10 < u_0 < 16 \text{ and } v_1(u_0) < v_0 < 160/v_0\}$$

close to the portion of ∂R_1 consisting of Z_p and on which $E_p < E_s$. These regions are indicated in Figure 2 though region R_{1b} is so small it is difficult to see.

Moving to the region(s) between the line Z_s and the curve Z_p, one of the expressions for E_s or E_p changes sign, so for equality we get

$$-(26 - u_0 - v_0) = 160 - u_0 v_0 \qquad \text{or} \qquad u_0 v_0 + u_0 + v_0 = 186 \qquad \text{or} \qquad v_0 = \frac{186 - u_0}{u_0 + 1}.$$

Finally, there is another large region $\{(u_0, v_0) : v_0 > \max\{v_1(u_0), v_2(u_0)\},\ 0 < u_0\}$ where

$$v_2 = \frac{186 - u_0}{u_0 + 1},$$

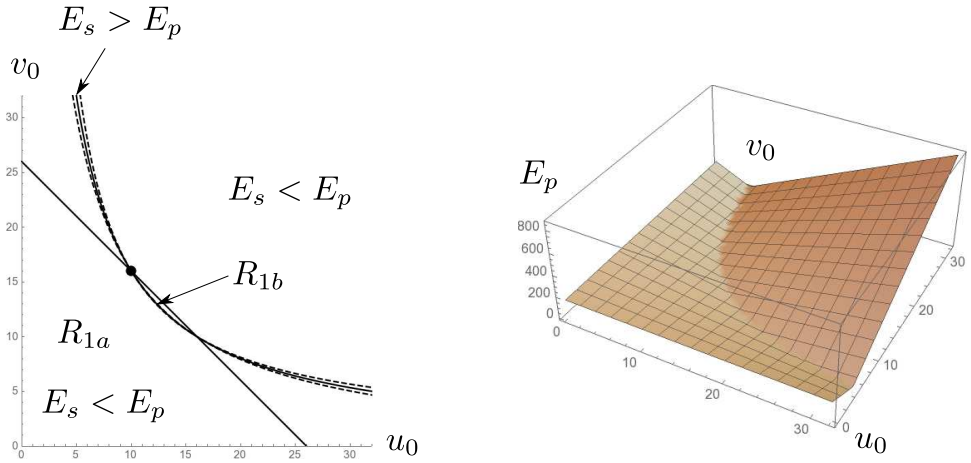

Figure 2: The dashed curves in the figure on the left bound a narrow disconnected region around the branch of the hyperbola $Z_p = \{(u_0, v_0) : v_0 = 160/u_0\}$ on which the error E_s in the calcuation of the sum $u_0 + v_0$ is **greater** than the error E_p in the calculation of the product $u_0 v_0$. On the right is the graph of the error E_p for the product; note the scale on the E_p axis is very different from that used in the plot of E_s in Figure 1.

and $E_s < E_p$ on this region.

I would like to pause and point out that my use of the words "large" and "small" here, in the sense that R_{1a} is "large" and R_{1b} is "small," while appearing convincing, say from the figures, is rather **inexact**. It would be very nice to have some **quantitative measure** to compare precisely how much more likely it is to have $E_s < E_p$. One might consider measuring the area of the regions, though there is some difficulty with this due to the fact that the areas of both regions are infinite. Another reasonable approach would be to determine the angle θ at which the dashed curves meet, and then consider the relative sizes of θ/π and $(\pi - \theta)/\pi$.

Exercise 6 *How much of our discussion of the sets Z_s and Z_p, the graphs of the errors E_s and E_p, and the regions where $E_s < E_p$ above can be extended to arbitrary (fixed) measured values u and v? In particular, what happens in cases where $u = v$ that is different when $u \neq v$? Can you compute the angles between the dashed curves as described above as functions of general u and v?*

Note that partially due to the particular measured values $(u, v) = (10, 16)$ involved, the gradient of the function appearing in the absolute values of the product error $E_p = |uv - u_0v_0|$ at $(10, 16)$ is rather large compared to the gradient of the corresponding function in the sum error $E_s = |u + v - u_0 - v_0|$:

$$\nabla(uv - u_0v_0) = (-v_0, -u_0) = (-16, -10) \qquad \text{while} \qquad \nabla(u+v-u_0-v_0) = (-1, -1).$$

In fact, at each point along the curve Z_p, the absolute value of the affine approximation

$$|(-v_0, -u_0) \cdot (\xi - u_0, \eta - v_0)|$$

associated with the product E_p is a piecewise affine function like E_s whose graph may be visualized as two planes meeting along the tangent line to Z_p and tangent to the graph of E_p at $(u_0, v_0, 0)$. Comparing the relative slopes of these affine approximations suggests something interesting: *The region(s) on which $E_p < E_s$ should become larger, in some sense, when u and v are smaller.* Let's see if we can realize this suggestion.

Instead of $(u, v) = (10, 16)$ which was obtained in a rather arbitrary way simply by taking the semi-axes of the ellipse considered in the nominally unrelated previous problem (Problem 62), let us take for measured values $(u, v) = (1/2, 1/4)$. In this case,

$$E_p = |1/8 - u_0v_0| \qquad \text{and} \qquad E_s = |3/4 - u_0 - v_0|.$$

The curves Z_s and Z_p are given by

$$v_0 = 3/4 - u_0 \qquad \text{and} \qquad v_0 = \frac{1}{8u_0}$$

respectively. The bounding curves are given by

$$v_1 = \frac{u_0 - 5/8}{u_0 - 1} \qquad \text{and} \qquad v_2 = \frac{7/8 - u_0}{u_0 + 1}.$$

Not only is the "smaller" or bounded region somewhat more substantial and easier to see—our choice of measured values has caused it to "fatten up"—but there are two other dramatic changes. First, this smaller disconnected region is now a region along Z_s rather than along Z_p. Second, the large and unbounded region is the region where the error in the sum calculation is **larger** than the error in the product calculation. There are two immediate conclusions to draw from this. First, it should not be expected that the actual error in calculating the product is larger than the actual error in calculating the sum, though this says nothing about the relative (percentage) errors which have denominators to seemingly magnify the (percentage) error of the product calculation over that of the sum calculation. Second, some kind of major transition has taken place in moving the measured values from $(10, 16)$ to $(1/2, 1/4)$.

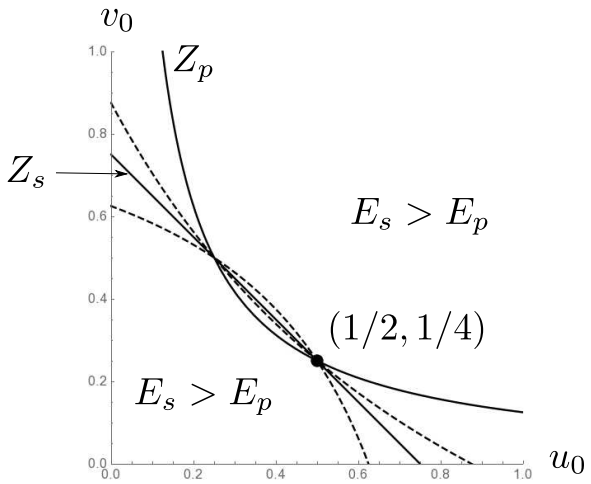

Figure 3: The dashed curves in the figure bound a narrow disconnected region around the line $Z_s = \{(u_0, v_0) : v_0 = 3/4 - u_0\}$ on which the error E_s in the calcuation of the sum $u_0 + v_0$ is **less** than the error E_p in the calculation of the product $u_0 v_0$.

Exercise 7 *Explain the transition in terms of a continuous family of comparisons depending on the measurements* (u, v).

Let us turn now to the percentages

$$P_p = \frac{|uv - u_0 v_0|}{u_0 v_0} \qquad \text{and} \qquad P_s = \frac{|u + v - u_0 - v_0|}{u_0 + v_0}.$$

As we have set these up, we are multiplying the actual errors by functions which are singular along the coordinate axes $u_0 = 0$ and $v_0 = 0$. To be precise,

$$P_p = \left(\frac{1}{u_0 v_0}\right) E_p \qquad \text{and} \qquad P_s = \left(\frac{1}{u_0 + v_0}\right) E_s.$$

This, at least potentially, complicates matters substantially. The level sets

$$Z_s = \{(u_0, v_0) : P_s(u_0, v_0) = 0\} \qquad \text{and} \qquad Z_p = \{(u_0, v_0) : P_p(u_0, v_0) = 0\}$$

are the same as the sets of the same name above. This tells us that there will always be some region around Z_p on which $P_s > P_p$, but the graphs to be considered are quite different, so maybe now this region will always be, in some sense, small. The

removal of the absolute values proceeds much as before, so that in the unbounded regions determined by Z_s and Z_p, the equality $P_s = P_p$ is represented by

$$\frac{uv - u_0v_0}{u_0v_0} = \frac{u + v - u_0 - v_0}{u_0 + v_0}. \tag{13}$$

This leads to the relation

$$uv(u_0 + v_0) = (u + v)u_0v_0 \qquad \text{or} \qquad v_0 = v_1(u_0) = \frac{uvu_0}{(u + v)u_0 - uv} = \frac{u_0}{\frac{u+v}{uv}u_0 - 1}.$$

Similarly, the other curve is given by

$$-\frac{uv - u_0v_0}{u_0v_0} = \frac{u + v - u_0 - v_0}{u_0 + v_0} \tag{14}$$

which becomes

$$-uv(u_0 + v_0) + 2u_0v_0(u_0 + v_0) = (u + v)u_0v_0$$

or

$$2u_0v_0^2 + [2u_0^2 - uv - (u + v)u_0]v_0 - uvu_0 = 0$$

which is quadratic in v_0. Thus, we are led to

$$v_0 = v_2(u_0) = \frac{-[2u_0^2 - uv - (u + v)u_0] \pm \sqrt{[2u_0^2 - uv - (u + v)u_0]^2 + 8uvu_0^2}}{4u_0}.$$

We have plotted these curves for the choice $(u, v) = (1/2, 1/4)$ and $(u, v) = (2, 1/2)$ in Figure 4. One presumes that for the percentages, there is always a (small) banded region about Z_p corresponding to the inequality $P_p < P_s$. We have not shown this. Presumably, the following is correct:

Conjecture/Exercise 1 *The functions $v_1 = v_1(u_0)$ and $v_2 = v_2(u_0)$ defined above have natural domains containing both the minimum $m = \min\{u, v\}$ and the maximum $M = \max\{u, v\}$ of the known measured values u and v, and satisfy*

$$v_s(u_0) = u+v-u_0 < v_2(u_0) < v_p(u_0) = \frac{uv}{u_0} < v_1(u_0) \qquad \text{for} \qquad u_0 < m \quad \text{and} \quad u_0 > M$$

$$v_2(u_0) < v_p(u_0) < v_1(u_0) < v_s(u_0) \qquad \text{for} \qquad m < u_0 < M.$$

And $P_p(u_0, v_0) < P_s(u_0, v_0)$ in the region between the graphs of v_1 and v_2 containing the graph of v_p (except for the points (u, v) and (v, u) which are a single point when $u = v$) while $P_s(u_0, v_0) < P_p(u_0, v_0)$ elsewhere in the open first quadrant.

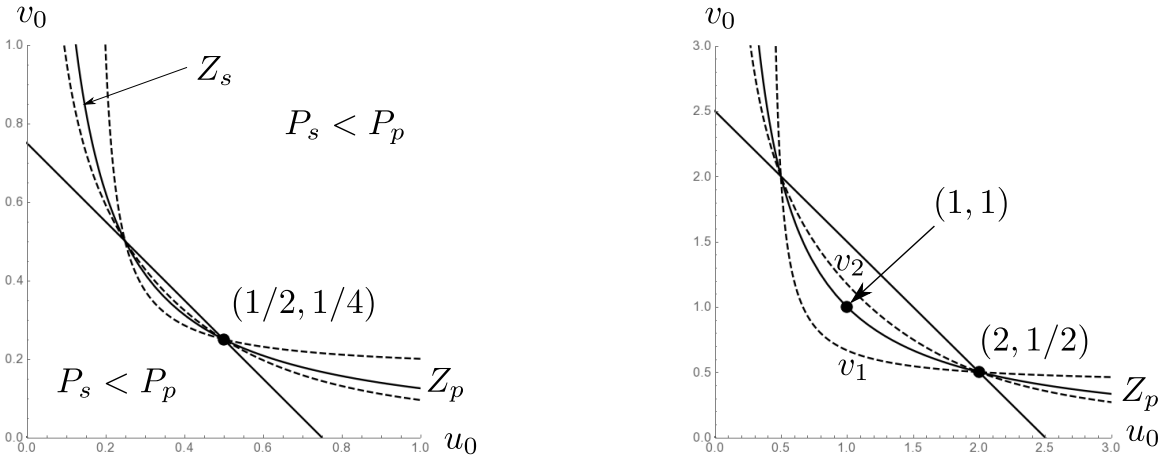

Figure 4: Percentages of error. The first counterexample $(u_0, v_0) = (1, 1) \in Z_p$ is illustrated on the right.

Considering the graphs of v_1 and v_2 as level sets, we can compute the angle between them as the angle between the respective normals. Writing (13) as

$$uv(u_0 + v_0) - (u + v)u_0v_0 = 0,$$

the gradient of the function on the left (with respect to (u_0, v_0)) is

$$(uv - v_0(u + v), uv - u_0(u + v))$$

which evaluted at $(u_0, v_0) = (u, v)$ becomes $(-v^2, -u^2)$. This is a downward normal, so let us set $\mathbf{v} = (v^2, u^2)$.

Similarly, (14) can be written as

$$-uv(u_0 + v_0) + 2u_0v_0(u_0 + v_0) - (u + v)u_0v_0 = 0,$$

and the gradient here is

$$(-uv + 4u_0v_0 + 2v_0^2 - v_0(u + v), -uv + 4u_0v_0 + 2u_0^2 - u_0(u + v))$$

with evaluated value $\mathbf{w} = (2uv + v^2, 2uv + u^2)$. (This is an upward normal.) Thus, the angle between $\mathbf{v}$ and $\mathbf{w}$ satisfies

$$\cos\theta = \frac{2uv^3 + v^4 + 2u^3v + u^4}{\sqrt{u^4 + v^4}\ \sqrt{4u^2v^2 + 4uv^3 + v^4 + 4u^2v^2 + 4u^3v + u^4}}.$$

It's not exactly clear how much we can say about this angle. We can say, however, that both normal vectors point strictly into the first quadrant. This means the expression for $\cos\theta$ is strictly between 0 and 1, and this means the angle corresponding to $P_p < P_s$ is greater than zero (so the region near the measured value (u, v) in which the percentage error for the product is smaller than the percentage error for the sum always corresponds to a smaller pair of sectors) and less than 1 (so there is no chance of showing the percentage error for the sum is always smaller than the percentage error for the product—as required by the problem). The sectors corresponding of size θ and $\pi - \theta$ are illustrated in Figure 5.

Let's go back and realize the counter-examples given above in the framework we have worked out here. The first one was when $(u, v) = (2, 1/2)$, $v = 1/2$ and we have simply picked the actual exact values $(u_0, v_0) = (1, 1)$ to be on the curve Z_p which gives zero error for the product (Figure 4 (right)). This is, on the one hand, difficult for the sum to beat and on the other hand clearly in the region where the percent error for the product should beat the sum.

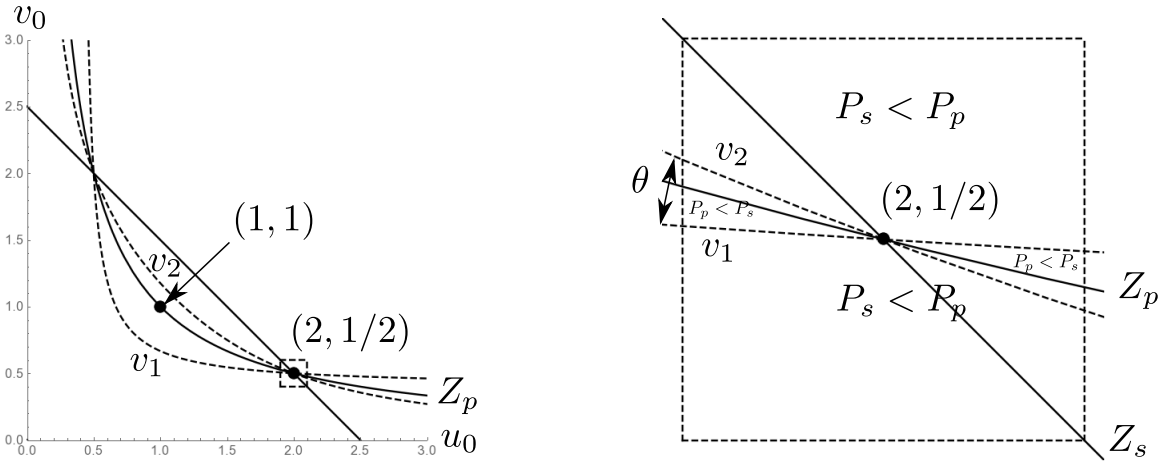

Figure 5: Example where the percentage of error in the calculation of the product is less than the percentage of error in the calculation of the sum. In this case $(u, v) = (2, 1/2)$ and $(u_0, v_0) = (1, 1)$. The illustration on the left is the same as Figure 4 (right) with a region of interest around $(2, 1/2)$ added. On the right we have zoomed in to the region of interest.

The second example is rather more interesting. This is where the question of where to linearize comes up. We have argued that to approximate the product uv or

the product error $uv - u_0v_0$ it is most natural to take the affine approximation with respect to u and v which involves the gradient (v, u) with respect to u and v followed by an evaluation at the actual values (u_0, v_0). The approach we have taken in which (u, v) is fixed, and then we consider all values (u_0, v_0) which the actual values may take, makes it natural to linearize with respect to (u_0, v_0), and then plug in (u, v) for (u_0, v_0). We did this, for example, in the discussion following Exercise 6. Now we will contemplate both processes in the same discussion. Here our counterexample is given by $(u, v) = (3, 1/2)$ and $(u_0, v_0) = (2, 1)$. And we are using the affine error obtained by linearizing $uv - u_0v_0$ as a function of u and v at (u_0, v_0). This expression does not make a natural geometric appearance in our framework. In particular, we should consider the sets where

$$Q_p = \frac{|v_0(u - u_0) + u_0(v - v_0)|}{u_0v_0}$$

is greater than, less than, or equal to P_s. First of all, the zero set Z_q for Q_p is given by the relation $v_0(u - u_0) + u_0(v - v_0) = 0$ or

$$v_0 = v_q(u_0) = \frac{vu_0}{2u_0 - u}.$$

Associated with this zero curve, we obtain also two transition curves

$$\pm\frac{v_0(u - u_0) + u_0(v - v_0)}{u_0v_0} = \frac{u + v - u_0 - v_0}{u_0 + v_0}.$$

These transition curves are (apparently) not always graphs, but presumably they have properties rather similar to v_1 and v_2 above. In fact, one should be careful with regard to the conjecture/exercise above; maybe the transition curves there are not always graphs either. Nevertheless, we expect that each such curve may be represented by at most two graphs each corresponding to signs "$\pm$", and we have that $Q_p = P_s$ along the graphs of

$$v_0 = v_3(u_0) = u_0\,\frac{u_0 \pm \sqrt{u_0^2 + 4vu_0 - 4uv}}{2(u - u_0)}$$

and

$$v_0 = v_4(u_0) = u_0\,\frac{3u_0 - 2(u + v) \pm \sqrt{9u_0^2 - 12uu_0 + 4(u^2 + uv + v^2)}}{2(u - 3u_0)}.$$

Presumably, conclusions somewhat similar to those concerning P_s and P_p and the "graphs" of v_1 and v_2 given in Conjecture/Exercise 1 hold for Q_p, P_s and the curves

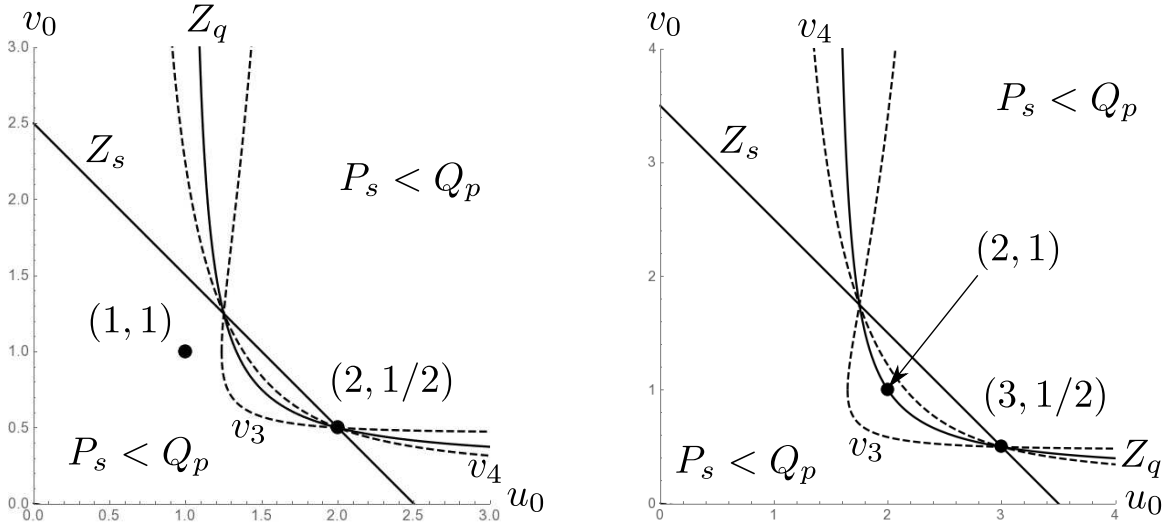

Figure 6: Affine error. In these figures, we compare the percentage error in the calculation of the affine approximation of the product with the percentage error in the calculation of the sum (which is the same as the affine approximation of the calculation of the sum). On the left we illustrate that when the exact values are $(u_0, v_0) = (1, 1)$ and the measured values are $(u, v) = (2, 1/2)$ (which give the exact product exactly) then the percentage error in the affine approximation of the product is still more than the percentage error in the calculation of the sum. Thus, these values do not give a counterexample to the assertion of the text (Problem 63 part (b)) when one uses the affine approximation to represent the product. On the right, however, we illustrate a counterexample in which the percentage of error in the affine approximation of the calculation of the product is less than the percentage of error in the calculation of the sum. In this case $(u, v) = (3, 1/2)$ and $(u_0, v_0) = (2, 1)$.

associated with v_3 and v_4 as illustrated in the special case of our second counterexample in Figure 6. As mentioned above, the linearization of the product error at the point (u_0, v_0) giving the (unknown) exact values is not easily represented in our figures. This is because this notion of "affine error" dependends on the position (u_0, v_0) in a nonlinear way. Nevertheless, this is the most natural quantity to consider.

Finally, if one is willing to use the sloppy approximations $u_0\tilde{u}$ and $v_0\tilde{v}$ as justification for it, one might consider the alternative affine error determined by linearizing with respect to (u_0, v_0), and then evaluating at the measured values (u, v). That is,

one may compare P_s to

$$W_p(u_0, v_0) = \frac{|v(u - u_0) + u(v - v_0)|}{uv}.$$

Notice that W_p depends on u_0 and v_0 linearly. Thus, the graph of W_p is the "wedge" determined by two half planes meeting along the line $Z_w = \{(u_0, v_0) : W_p(u_0, v_0) = 0\}$ as discussed just after Exercise 6. This greatly simplifies the discussion, but still does not lead to any possibility to justify the conclusion of Problem 63 part (b).

Exercise 8 *Reproduce the illustration on the right in Figure 6. Apply the discussion concerning the comparison of P_p to P_s and insert also the curve Z_p along with the graphs of v_1 and v_2 on the same illustration. Determine the transition curves (say v_5 and v_6) where $P_s(u_0, v_0) = W_p(u_0, v_0)$ and sketch them on the same figure. You should then have three distinct wedge shaped regions determined near $(u, v) = (3, 1/2)$:*

1. *The region where $P_p < P_s$ determined by v_1 and v_2.*
2. *The region where $Q_p < P_s$ determined by v_3 and v_4.*
3. *The region where $W_p < P_s$ determined by v_5 and v_6.*

What is the relation between these three wedge shaped regions? Can you find a point (u_0, v_0) in all three of them? Have we already found such a point?

Final Remarks

One may become irritated when there are errors in textbooks, but many of the best and most famous (advanced level) textbooks in mathematics are especially known and famous for the errors and lack of clarity within them. One does not necessarily strive for lack of clarity, but when it occurs, it can provide an opportunity for learning—which people who love to learn often appreciate. It took me quite a long time to understand and explain clearly what was going on in Practice Problems 62 and 63 of Chapter 14 in Thomas' Calculus—and there aren't many serious or interesting errors in the book—but I thought about things in a different way than I had before, and I learned some things.