The UMAP Journal

Vol. 32, No. 3

Subscription Rates for 2011 Calendar Year: Volume 32

Institutional Web Membership (Web Only)

Institutional Web Memberships do not provide print materials. Web memberships allow members to search our online catalog, download COMAP print materials, and reproduce them for classroom use.

(Domestic) #3130 $467 (Outside U.S.) #3130 $467

Institutional Membership (Print Only)

Institutional Memberships receive print copies of *The UMAP Journal* quarterly, our annual CD collection UMAP Modules, *Tools for Teaching*, and our organizational newsletter *Consortium*.

(Domestic) #3140 $312 (Outside U.S.) #3141 $351

Institutional Plus Membership (Print Plus Web)

Institutional Plus Memberships receive print copies of the quarterly issues of *The UMAP Journal*, our annual CD collection UMAP Modules, *Tools for Teaching*, our organizational newsletter *Consortium*, and online membership that allows members to search our online catalog, download COMAP print materials, and reproduce them for classroom use.

(Domestic) #3170 $615 (Outside U.S.) #3171 $659

For individual membership options visit
www.comap.com for more information.

To order, send a check or money order to COMAP, or call toll-free
1-800-77-COMAP (1-800-772-6627).

The UMAP Journal is published quarterly by the Consortium for Mathematics and Its Applications (COMAP), Inc., Suite 3B, 175 Middlesex Tpke., Bedford, MA, 01730, in cooperation with the American Mathematical Association of Two-Year Colleges (AMATYC), the Mathematical Association of America (MAA), the National Council of Teachers of Mathematics (NCTM), the American Statistical Association (ASA), the Society for Industrial and Applied Mathematics (SIAM), and The Institute for Operations Research and the Management Sciences (INFORMS). The Journal acquaints readers with a wide variety of professional applications of the mathematical sciences and provides a forum for the discussion of new directions in mathematical education (ISSN 0197-3622).

Periodical rate postage paid at Bedford, MA and at additional mailing offices.

Send address changes to: info@comap.com
COMAP, Inc., Suite 3B, 175 Middlesex Tpke., Bedford, MA, 01730

Vol. 32, No. 3 2011

Table of Contents

Publisher's Editorial

Minimodule

UMAP Modules

Reviews

Publisher's Editorial

I'm Mad as Hell

Solomon A. Garfunkel
Executive Director
COMAP, Inc.
175 Middlesex Turnpike, Suite 3B
Bedford , MA 01730–1459
s.garfunkel@comap.com

Here are excerpts from a recent report on the status of the National Science Foundation's merit review criteria:

> The National Science Foundation's (NSF) merit review criteria and process lie at the core of NSF funding decisions. . . . [C]riteria consist of both intellectual merit and broader impacts, and . . . proposed revisions attempt to provide clearer guidance on broader impacts. . . . The proposed principles and changes . . . align with language included in the America COMPETES Reauthorization Act . . . [which] states eight goals for broader impacts activities:
> (1) Increased economic competitiveness of the United States
> (2) Development of a globally competitive STEM workforce
> (3) Increased participation of women and underrepresented minorities in STEM
> (4) Increased partnerships between academia and industry
> (5) Improved pre-K–12 STEM education and teacher development
> (6) Improved undergraduate STEM education
> (7) Increased public scientific literacy
> (8) Increased national security
>
> [The] Task Force on Merit Review . . . received approximately 280 comments, with the majority coming from individuals rather than organizations. . . . feedback was diverse, ranging from very positive to very negative, with some suggesting that NSF do away entirely with the broader impacts criteria while others were appreciative of the more descriptive language. Overall two issues stand out to the Task Force:
>
> 1. There is a perception that broader impacts detract from NSF's basic research goal; and

The UMAP Journal 32 (3) (2011) 185–186.

2. Other broader impacts have been diminished by the listing of the national goals congressionally authorized in the America COMPETES Act.

[T]he Task Force will . . . discuss how to move forward, giving full consideration to the stakeholder feedback while also being responsive to the congressional language.

Here is what I wrote in response to the call for public comment:

> My reaction to the laundry list of "important national goals" that sit in a portion of the proposed new review criteria is that they represent a much too narrow view of scientific research. Of course, NSF is the "national" organization that supports scientific research and education. But science (and I would argue science education) is an international and cooperative effort. I for one do not care if a Chinese scientist finds a way to reverse global warming or a Russian invents cold fusion (we can no longer visit the International Space Station without the help of our Russian colleagues). The entire tone of this list feels like something that was written in the 1950s. Of course NSF should improve the state of U.S. scientific research and education. But *much* more importantly, NSF should improve the state of science and of science education everywhere. This document cries out for such recognition. For shame.

Certainly, some of these goals are motherhood and apple pie, but the overall nationalistic thrust, in my opinion, is simply wrong. Let me be clear: I am far from politically naïve. I am well aware that the NSF may be held hostage to the current political winds. I am equally aware that this is a defensive move on NSF's part, i.e., shift a bit to the America-first position before you are forced to shift further, or have your appropriations cut. There was a good deal of this that went on during the recent Republican administration.

The trouble is—it doesn't work! Look at what has gone on in the recent debt ceiling debate. Someone has to stand up for pure science and mathematics research. Someone has to call on the best of us to work in education. And if the leadership of NSF is too timid, then those of us in the field have to speak with a loud clear voice. Otherwise the shame will be on us.

While the period of public review of these criteria has past, it is still not too late to write tothe House Science, Space and Technology Subcommittee on Research and Science Education, Rep. Mo Brooks, chair.

About the Author

Solomon Garfunkel is the founder and Executive Director of COMAP and Executive Publisher of this *Journal*.

Minimodule
Profit in a Mutual Fund

Floyd Vest
1103 Brightwood
Denton, TX 76209
wyonav@verizon.net

Table 1 is a simulation of an annual statement from a mutual fund that invests primarily in stocks. On 6/25, a dividend of \$0.20 per share was awarded. Since the investor owned 779.834 shares, the dollar amount of the transaction was (779.834)(\$0.20) = \$155.97. At \$32.46 a share, this dollar amount of the transaction was converted into 155.97/32.46 = 4.805 additional shares, for a total of 4.805 + 779.834 = 784.639 shares. Similarly, on 12/15 a dividend was awarded yielding 82.847 additional shares, so the total shares owned at the end of the year was 867.486. (For a more complete introduction to a mutual fund statement, see Vest [2011].)

Table 1.
Annual stock mutual fund statement.

Trade date	Dividend per share	Dollar amount of transaction	Share price	Shares transacted	Total shares owned	Account value
12/31/2009			\$28.61		779.834	\$22,311.05
6/24			\$32.66		779.834	
6/25	\$0.20	\$155.97	\$32.46	4.805	784.639	\$25,469.39
12/14			\$32.46		784.639	
12/15	\$3.11	\$2440.22	\$29.36	82.847	867.486	\$25,469.38
12/31/2010			\$29.84		867.486	\$25,885.78

(Roundings are responsible for Account values on 6/25 and 12/15 differing by \$0.01.)

You Try It #1

Do the calculations for the 12/15 row in **Table 1**. Explain each number.

The UMAP Journal 32 (3) (2011) 187–198.

Profit

The profit to the investor comes from appreciation of share prices and from additional shares awarded from dividends. From **Table 1**, based on share price at the beginning, the gain from additional shares awarded was

$$28.61(4.805 + 82.847) = \$2507.72.$$

This gain was 70.15% of the total profit for the year:

$$\text{Total profit} = \$25{,}885.78 - \$22{,}311.05 = \$3{,}574.73.$$

These calculations illustrate that the number of shares awarded from dividends can play a major role in overall annual gain to the investor. Even if share price doesn't increase, there can still be a profit.

You Try It #2

Do the calculations based on **Table 1** to show that if share price at the end of the year equals share price at the beginning, then account value at the end of the year is greater than account value at the beginning of the year. Explain why this is true. Explain the values as you calculate.

Generalization

To investigate in general terms the role of additional shares awarded, we build **Table 2** based on **Table 1**.

Table 2.
The algebra of the mutual fund statement.

Trade date	Dividend per share	Dollar amount of transaction	Share price	Shares transacted	Total shares owned	Account value
12/31/2009			P_B		T	TP_B
Day before dividend 1			P_1			
Day of dividend 1 (Day 1)	D_1	TD_1	$P_1 - D_1$			
Day before dividend 2			P_2			
Day of dividend 2 (Day 2)	D_2		$P_2 - D_2$			
12/31/2010			P_3			

The meanings of the terms and the symbols in the table are:

- P_B is the share price at the beginning of the year.
- T is the total shares owned at the beginning.
- P_1 is the share price on the day before dividend Day 1.
- D_1 is the dividend per share awarded and $P_1 - D_1$ is the new share price.

Similarly for P_2, D_2, and $P_2 - D_2$, while P_3 indicates the share price at the end of the year.

In **Table 2**, the dollar amount of the transaction is the number of shares owned times the dividend per share. The shares transacted is the dollar amount of the transaction divided by the new share price on the dividend day. The algebraic expressions required to extend the table are as follows:

- Dollar amount of transaction on Day 1: TD_1.
- Shares transacted on Day 1: $\dfrac{TD_1}{P_1 - D_1}$. **(1)**
- Total shares owned on Day 1: $T + \dfrac{TD_1}{P_1 - D_1}$.
- Account value on Day 1: $\left[T + \dfrac{TD_1}{P_1 - D_1}\right](P_1 - D_1)$.

You Try It #3

Apply the above approach to **Table 1** to calculate account value on Day 1. Explain the values as you calculate.

- Dollar amount of transaction on Day 2: $\left[T + \dfrac{TD_1}{P_1 - D_1}\right] D_2$.
- Shares transacted on Day 2: $\dfrac{\left[T + \dfrac{TD_1}{P_1 - D_1}\right] D_2}{P_2 - D_2}$. **(2)**
- Total shares owned on Day 2: $T + \dfrac{TD_1}{P_1 - D_1} + \dfrac{\left[T + \dfrac{TD_1}{P_1 - D_1}\right] D_2}{P_2 - D_2}$. **(3)**
- Total account value on Day 2:

$$(P_2 - D_2)T\left[\left(1 + \frac{D_1}{P_1 - D_1}\right) + \frac{\left(1 + \dfrac{D_1}{P_1 - D_1}\right) D_2}{P_2 - D_2}\right].$$

You Try It #4

Apply the above approach to **Table 1** to calculate account value on Day 2. Explain the values as you calculate.

General Formulas for the End of the Year

- Total account value at end of year:

$$(P_3)(T)\left[\left(1+\frac{D_1}{P_1-D_1}\right)+\frac{\left(1+\frac{D_1}{P_1-D_1}\right)D_2}{P_2-D_2}\right] \tag{4}$$

- Total profit at end of year:

$$(P_3)(T)\left[\left(1+\frac{D_1}{P_1-D_1}\right)+\frac{\left(1+\frac{D_1}{P_1-D_1}\right)D_2}{P_2-D_2}\right]-P_BT. \tag{5}$$

Formulas 1–4 indicate that the final account value at the end of the year is

(Ending Share Price)

$\times$ (T + Shares transacted on Day 1 + Shares transacted on Day 2).

You Try It #5

Do the calculations in **Table 1** to illustrate formula **(4)** for total account value and formula **(5)** for total profit. Explain the values as you calculate.

When Is There a Profit for the Year?

Examination of **(5)** for total profit at the end of the year indicates that if $P_3 = P_B$, the investor still makes a profit if $P_1 - D_1 > 0$ and $P_2 - D_2 > 0$. This is because of the increase in number of shares from dividends.

This observation suggests the question of how low can P_3 be and still make a profit. A general investigation would involve five variables: D_1, D_2, and three share prices. For now, we look just at an example from **Table 1**. Considering **Table 1** and **Table 2**,

$$\text{Account value at end of the year} = (P_3)(867.486), \quad \text{and}$$

$$\text{Account value at beginning of the year} = \$22{,}311.05.$$

We solve the inequality $(P_3)(867.486) \geq 22{,}311.05$ to get a share price $P_3 \geq \$25.72$. This value is less than $P_B = \$28.61$, so we could have in **Table 1** a depreciation of share price to as low as \$25.72 and still not lose money.

Notice that in **Table 1**, $P_3 = \$29.84 > \$28.61 = P_B$, so there has been an appreciation in share prices. We will examine the role of appreciation of share prices by using contiguous intervals on share prices.

- Let appreciation rate b_1 be such that $P_1 - D_1 = P_B(1 + b_1)$.
- Let b_2 be such that $P_2 - D_2 = (P_1 - D_1)(1 + b_2)$.
- Let b_3 be such that $P_3 = (P_2 - D_2)(1 + b_3)$.

Doing so gives

$$P_3 = P_B(1 + b_1)(1 + b_2)(1 + b_3). \quad \textbf{(6)}$$

The reader can verify that in **Table 1**,

$$b_1 = 0.1345683, \qquad b_2 = -0.0955022, \qquad b_3 = 0.0163488.$$

You Try It #6

Do the calculations in **Table 1** to verify each of the above rates of appreciation of share prices. Explain the values as you calculate.

Further Results

Entering $(1 + b_1)$, $(1 + b_2)$, and $(1 + b_3)$ in the above formulas in **Table 2** gives various new and informative formulas. For example,

- Total shares owned on Day 2:

$$T\left[1 + \frac{D_1}{(P_B)(1 + b_1)}\right]\left[1 + \frac{D_2}{\big((P_B)(1 + b_1)\big)(1 + b_2)}\right]. \quad \textbf{(7)}$$

This formula gives two rates of increase in total shares owned. The second factor gives the rate of increase in shares owned resulting from shares awarded on D_1, and the third gives the rate of additional increase from the shares awarded at D_2.

You Try It #7

Do the calculations in **Table 1** to illustrate that the second factor in **(7)** gives the rate of increase in number of shares owned resulting from shares awarded on Day 1, and the third factor gives the additional rate of increase on Day 2. Calculate those rates and explain the values as you calculate.

- Total profit from shares awarded on Day 1 and on Day 2:

$$T(1+b_3)\left[(1+b_2)(D_1)+\left(1+\frac{D_1}{P_B(1+b_1)}\right)(D_2)\right]. \tag{8}$$

- Total account value at the end of the year:

$$= T(1+b_3)[P_B(1+b_1)(1+b_2)+D_2]\left[1+\frac{D_1}{P_B(1+b_1)}\right] \tag{9}$$

$$= T(P_B)(1+b_3)\left[(1+b_1)+\frac{D_1}{P_B}\right]\left[(1+b_2)+\frac{D_2}{P_B(1+b_1)}\right] \tag{10}$$

$$= T(1+b_1)(1+b_2)(1+b_3)\left[1+\frac{D_1}{P_B(1+b_1)}\right] \times\left[1+\frac{D_2}{P_B(1+b_1)(1+b_2)}\right] \tag{11}$$

$$= TP_3\left[1+\frac{D_1}{P_B(1+b_1)}\right]\left[1+\frac{D_2}{P_B(1+b_1)(1+b_2)}\right] \tag{12}$$

All of the formulas **(9)–(12)** are significant in that they express total account value in terms of basic variables. The last two factors in **(12)** give the rate of increase of total shares owned, the first at D_1, and the second at D_2. Notice that the presence P_3 in **(12)** indicates that the increased number P_3 of shares gives greater participation in appreciation.

Criteria for End-of-Year Profit

To determine if there is profit for the year, we can use **(5)** and set

- Total profit at the end of the year:

$$(P_3)(T)\left[\left(1+\frac{D_1}{P_1-D_1}\right)+\frac{\left(1+\frac{D_1}{P_1-D_1}\right)D_2}{P_2-D_2}\right]-P_BT>0.$$

We can express this inequality as

$$\left[1+\frac{D_1}{P_B(1+b_1)}\right]\left[1+\frac{D_2}{P_B(1+b_1)(1+b_2)}\right]>\frac{P_B}{P_3}.$$

There are two cases:

- The share price is the same or higher at the end of the year than at the beginning, i.e., $P_3 \geq P_B$, so that $P_B/P_3 \leq 1$. As long as $b_1 > -1$ and $b_2 > -1$ (meaning that the share price does not become negative!), then each of the factors is greater than 1 and there is profit for the year.
- The share price is lower at the end of the year than at the beginning, i.e., $P_3 < P_B$, so that $P_B/P_3 > 1$. For $b_1 > -1$ and $b_2 > -1$, each of the factors is greater than 1; for an end-of-year profit, their product must be greater than P_B/P_3.

We can analyze both cases together. Using the fact from **(6)** that

$$P_3 = (1 + b_1)(1 + b_2)(1 + b_3)P_B,$$

we can render the condition for profit as

$$\left[1 + \frac{D_1}{P_B(1 + b_1)}\right]\left[1 + \frac{D_2}{P_B(1 + b_1)(1 + b_2)}\right] > \frac{1}{(1 + b_1)(1 + b_2)(1 + b_3)}.$$

Let the rate a be such that

$$(1 + a) = (1 + b_1)(1 + b_2)(1 + b_3),$$

so that

$$P_3 = (1 + a)P_B \quad \text{and} \quad P_3/P_B = 1 + a.$$

If $P_3 \geq P_B$, then $a \geq 0$ and $1 + a \geq 1$; if $P_3 < P_B$, then $a < 0$ and $1 + a < 1$.

The requirement for end-of-year profit can then be formulated as

$$\left[1 + \frac{D_1}{P_B(1 + b_1)}\right]\left[1 + \frac{D_2}{P_B(1 + b_1)(1 + b_2)}\right] > \frac{1}{1 + a}.$$

In Vest [2011], separate appreciation rates a_1, a_2, and a_3 are used so that

$$P_1 = P_B(1 + a_1), \quad P_2 = (P_1 - D_1)(1 + a_2),$$

$$P_3 = (P_2 - D_2)(1 + a_3).$$

One could ask about the relationships between a_1, a_2, a_3, and b_1, b_2, b_3. Obviously, $a_3 = b_3$. It turns out that

$$1 + a_1 = (1 + b_1) + \frac{D_1}{P_B}, \qquad 1 + a_2 = (1 + b_2) + \frac{D_2}{P_B(1 + b_1)}. \qquad \textbf{(13)}$$

More simply, from Theorem 9 in Vest [2011], a requirement for end-of-year profit is

$$T(P_B)[(1 + a_1)(1 + a_2)(1 + a_3) - 1] > 0,$$

that is,

$$(1 + a_1)(1 + a_2)(1 + a_3) > 1.$$

Substituting from **(13)** gives a requirement for end-of-year profit of

$$\left[(1 + b_1) + \frac{D_1}{P_B}\right]\left[(1 + b_2) + \frac{D_2}{P_B(1 + b_1)}\right](1 + b_3) > 1.$$

What is the requirement for a certain rate of return (total return TR) for the year? From Theorem 10 in Vest [2011], we have

$$\text{TR} = \text{Rate of return} = (1 + a_1)(1 + a_2)(1 + a_3) - 1.$$

To acquire a rate of return TR requires

$$\left[(1 + b_1) + \frac{D_1}{P_B}\right]\left[(1 + b_2) + \frac{D_2}{P_B(1 + b_1)}\right](1 + b_3) - 1 = \text{TR}.$$

Exercises

1. Express total shares transacted on Day 2 in terms of T, P_B, D_1, D_2, b_1, and b_2.
2. Prove in general that total account value on Day 1 is the same as total account value on the day before Day 1.
3. Use the general formulas to prove that total account value on Day 2 is the same as total account value on the day before Day 2.
4. Do the calculations in **Table 1** to verify **(7)** for total shares owned on Day 2.

For the following exercises, give general derivations.

5. Derive **(7)** for total shares owned on Day 2 from previous formulas.
6. Do a derivation to prove **(9)**.
7. Do derivations to prove both parts of **(13)**.
8. Prove that if $b_1 > -1$, $D_1 > 0$, and $P_B > 0$, then

$$1 + \frac{D_1}{P_B(1 + b_1)} > 1.$$

9. Prove that if $b_1 > -1$, $b_2 > -1$, $D_2 > 0$, and $P_B > 0$, then

$$1 + \frac{D_2}{P_B(1+b_1)(1+b_2)} > 1.$$

10. Prove in general that expressions **(10)** and **(11)** are equal.

11. Prove that

Total account value at end of year

$$= (1 + b_3)(\text{Total account value on Day 2}).$$

Sidebar Notes

Investing in Stocks Can Be Discouraging. From June 2007 through February 2009, the S&P 500 major stock index fell 49.7%. Over the 10 years ending December 31, 2010, the S&P 500 averaged a piddling annual 1.32% profit. The worst 20-year period on stocks was 1962 to 1981—stocks averaged only 0.8% per year return (`Scottburns.com`, *Money Magazine*, June 2011).

Investing in Stocks Requires Taking the Long View. From 1900 to 2010, the annualized return on U.S. stocks after inflation was 6.3%. The return on stocks after inflation outside the U.S. was 5%. Using an estimate of average inflation rate of 3.2%, and the formula for inflation-adjusted rate of return

$$\frac{r - I}{1 + I} = \frac{r - 0.032}{1 + 0.032} = 0.063$$

gives $r = 9.7\%$ as the before-inflation long-term return on stocks (*Money Magazine*, June 2011). (See Vest [2012] for a derivation of inflation-adjusted rate of return.) (See also `http://www.usinflationcalculator.com`.)

References

Vest, Floyd. 2011. The advanced arithmetic and theorems of mutual fund statements. *The UMAP Journal*. To appear.

_____. 2012. Living and investing with inflation, Fisher's effect. *The UMAP Journal*. To appear.

Answers to Selected Exercises

5. From equation **(3)**, total shares owned on Day 2 is

$$= T\left[1 + \frac{D_1}{P_B(1+b_1)}\right]\left[1 + \frac{D_1}{P_B(1+b_1)(1+b_2)}\right]$$

$$= T + \frac{T(D_1)}{P_1 - D_1} + \frac{\left(T + \dfrac{T(D_1)}{P_1 - D_1}\right) D_2}{P_2 - D_2}$$

$$= T\left[\left[1 + \frac{D_1}{P_1 - D_1}\right] + \frac{\left[1 + \dfrac{D_1}{P_1 - D_1}\right] D_2}{P_2 - D_2}\right]$$

$$= T\left[1 + \frac{D_1}{P_1 - D_1}\right]\left[1 + \frac{D_2}{P_2 - D_2}\right]$$

$$= T\left[1 + \frac{D_1}{P_B(1+b_1)}\right]\left[1 + \frac{D_2}{P_B(1+b_1)(1+b_2)}\right].$$

6. Total account value at the end of the year is

$$= P_3[\text{Total shares owned on Day 2}]$$

$$= [P_2 - D_2](1+b_3)[\text{Total shares owned on Day 2}]$$

$$\text{since } P_3 = (P_2 - D_2)(1+b_3)$$

$$= (1+b_3)[\text{Total account value on Day 2}]$$

$$\text{since total account value on Day 2 is}$$

$$(P_2 - D_2)(\text{Total shares owned on Day 2})$$

$$= (1+b_3)(\text{Total account value on day before on Day 2})$$

$$= (1+b_3)(P_2)(\text{Total shares owned on Day 1})$$

$$= (1+b_3)(P_2)\left[T\left(1 + \frac{D_1}{P_1 - D_1}\right)\right]$$

$$= T(1+b_3)[P_B(1+b_1)(1+b_2) + D_2]\left[1 + \frac{D_1}{P_1 - D_1}\right]$$

$$\text{since } P_2 = P_B(1+b_1)(1+b_2) + D_2.$$

7. A common way to arrive at the idea of a proof is to start with the equality and simplify, then reverse the steps to make sure every step can be reversed. We first simplify. Multiplying through by $P_B(1+b_1)$ gives

$$\begin{aligned}
P_B(1+b_1)(1+a_2) &= P_B(1+b_1)(1+b_2)+D_2 \\
P_B(1+b_1)(1+a_2) &= P_2 \\
&\qquad \text{since } P_2 = P_B(1+b_1)(1+b_2)+D_2 \\
P_B(1+b_1)(1+a_2) &= (1+a_2)(P_1-D_1) \\
&\qquad \text{since } P_2 = (1+a_2)(P_1-D_1) \\
P_B(1+b_1) &= SP - D_1 \qquad \text{(by canceling)} \\
P_1 - D_1 &= P_1 - D_1 \\
&\qquad \text{since } P_B(1+b_1) = P_1 - D_1.
\end{aligned}$$

We leave it to the reader to see that these steps can be reversed for a proof.

8. $b_1 > -1$, therefore $1+b_1 > 0$, so

$$\begin{aligned}
P_B(1+b_1) &> 0, \\
\frac{D_1}{P_B(1+b_1)} &> 0, \\
1 + \frac{D_1}{P_B(1+b_1)} &> 1.
\end{aligned}$$

11. See the proof of **Exercise 6**.

Notes for the Instructor

This article is a companion to Vest [2011]. Both are about the same annual stock mutual fund statement, but each from a different point of view.

Have your students set up a lifetime file on personal finance and include this article.

Students can study the history of earnings by stock market indexes at `http://www.wikipedia.org` and `http://www.investopedia.com` (type in "Dow Jones Industrial Average" and "S&P 500 Index of Stocks"). The history of index funds that invest in major stock indexes can be found at `http://www.vanguard.com`.

About the Author

Floyd Vest is retired Professor of Mathematics and Education, Mathematics Dept., University of North Texas.

UMAP

Modules in Undergraduate Mathematics and Its Applications

Published in cooperation with

The Society for Industrial and Applied Mathematics,

The Mathematical Association of America,

The National Council of Teachers of Mathematics,

The American Mathematical Association of Two-Year Colleges,

The Institute for Operations Research and the Management Sciences, and

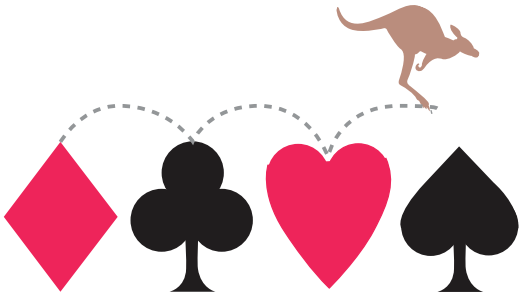
The American Statistical Association.

COMAP

Module 808

Cards, Codes, and Kangaroos

Lindsey R. Bosko-Dunbar



Numerical Analysis

COMAP, Inc., Suite 3B, 175 Middlesex Tpk., Bedford, MA 01730 (781) 862-7878

INTERMODULAR DESCRIPTION SHEET: UMAP Unit 808

TITLE: Cards, Codes, and Kangaroos

AUTHOR: Lindsey R. Bosko-Dunbar
Dept. of Natural Sciences and Mathematics
West Liberty University
West Liberty, WV 26074
lindsey.bosko@westliberty.edu

MATHEMATICAL FIELD: Numerical analysis

APPLICATION FIELD: Games

TARGET AUDIENCE: Undergraduate or early graduate student with one semester each of abstract algebra and linear algebra

ABSTRACT: The Kruskal Count is a card trick invented by mathematician (not magician) Martin Kruskal. The mathematics of the trick illustrates Pollard's kangaroo method, which was designed to solve the discrete logarithm problem: Given a finite cyclic group, $G = \langle g \rangle$, and $X \in G$, find $x \in \mathbb{Z}$ such that $g^x = X$. In this Module, we demonstrate the card trick and in revealing its secret, we uncover connections to the discrete logarithm problem, cryptography, and Markov chains.

PREREQUISITES: Cyclic groups, generators, modular arithmetic, matrix inverses, modular arithmetic and factoring functions for a computer algebra system (e.g., for Maple, the functions **mod** and **ifactor**, if-then statements and for-loops in programming in such a system, expected value, and standard results about Markov chains (transient and absorbing states, canonical form of the transition matrix, and the fundamental matrix).

The UMAP Journal 32 (3): 199–236.

COMAP, Inc., Suite 3B, 175 Middlesex Tpke., Bedford, MA 01730
(800) 77-COMAP = (800) 772-6627, or (781) 862-7878; http://www.comap.com

Cards, Codes, and Kangaroos

Lindsey R. Bosko-Dunbar

Dept. of Natural Sciences and Mathematics
West Liberty University
West Liberty, WV 26074
lindsey.bosko@westliberty.edu

Table of Contents

Modules and Monographs in Undergraduate Mathematics and its Applications (UMAP) Project

Paul J. Campbell	Editor
Solomon Garfunkel	Executive Director, COMAP

The goal of UMAP is to develop, through a community of users and developers, a system of instructional modules in undergraduate mathematics and its applications, to be used to supplement existing courses and from which complete courses may eventually be built.

The Project was guided by a National Advisory Board of mathematicians, scientists, and educators. UMAP was funded by a grant from the National Science Foundation and now is supported by the Consortium for Mathematics and Its Applications (COMAP), Inc., a nonprofit corporation engaged in research and development in mathematics education.

1. Introduction

This Module begins with a card trick explained in **Section 2**, followed by an example of the trick succeeding. Before revealing the secret behind the trick's success, we describe several cryptosystems that rely on what is known as the discrete logarithm problem (DLP) (**Section 3**). The link between these topics is Pollard's kangaroo method, which is an attack on the DLP as well as the basis for the trick. In **Section 4**, we reveal the secret behind the trick, together with many of the links among cryptography, discrete logarithms, cards, and kangaroos. Before further analysis on the card trick, we set out results about Markov chains **Section 5**. We then model the card trick as a Markov process **Section 6**. The results of the Markov chain analysis confirm that the card trick will "probably" work. "Probably" was the precise word used by Martin Gardner [1978], who first published the trick; his intent was to convey its difference from a typical magician's trick. We also perform another probabilistic analysis, on a modified version of the card trick, in **Section 7**, which provides an upper bound on the probability of the trick failing.

2. A Card Trick

A group of mathematicians gathers at a party. To entertain her peers, one brings a standard deck of 52 cards (i.e., a poker/bridge deck) and throughout the evening performs the following trick:

> As dealer, she invites a person to be the player and instructs the player to choose secretly a number between 1 and 10. She informs all watching that each card has a value: Aces 1, face cards 5, and all other cards the number on the card. She deals the cards, one at a time face up. When the number of cards dealt equals the player's secretly chosen number, the player is to note (silently) the value of the corresponding card, which we refer to as the player's first *key card*. With the first key card's value in mind, the player silently counts until the number of cards dealt from his key card on equals its value; the corresponding card is the player's second key card.
>
> For instance, suppose that the player initially chose 4 as the secret number and the first four cards dealt are $A, 10, 2, 8$; then the player's first key card is the 8. The player next counts to the eighth card dealt after the 8 to arrive at the player's second key card.
>
> Play continues in this manner, from one key card to the next, until all 52 cards have been dealt. The dealer then announces what she believes to be the player's last secret card. To the player's astonishment, the dealer is correct (well, maybe—as we will see).

We give an example, with the player's key cards in red (light gray):

Example 1.

$$A, 10, 2, 8, 7, 4, 7, J, 6, 5, K, 8, 2, 3, 9, 8, 5, Q, J, 5, 6, 8, K, 10, K, 3,$$

$$9, 5, 2, K, 4, Q, 9, Q, 3, 7, 6, A, J, 10, J, A, 6, 4, 9, 4, 7, A, 2, 3, \mathbf{Q}, 10$$

The initial secret number was 4 and the last key card is a Queen. (The trick does not distinguish suits or colors.)

The dealer does not always correctly guess the player's last card, but most of the time she is successful. As she performs the trick with other players, her fellow mathematicians begin to inquire how the trick works. They note that the deck is random, neither shuffled nor stacked in any particular order. The dealer, not being a true magician, is happy to discuss the secret, but first she must give some background on a seemingly unrelated topic—cryptography.

For Websites where you can play the card trick, see Haga and Robins [1997] or Montenegro [2009].

Exercises

1. **a)** Perform the card trick from **Example 1** as if you were the player, using a different initial secret number. What is the last card that you land on (your last key card)?
   b) Is there an initial secret number between 1 and 10 that will lead to something other than the Queen being the last key card? If so, what number(s) will do so?

2. What are the possible last key cards if the card trick is played on the ordered deck

$$A, A, A, A, 2, 2, 2, 2, 3, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 7,$$
$$8, 8, 8, 8, 9, 9, 9, 9, 10, 10, 10, 10, J, J, J, J, Q, Q, Q, Q, K, K, K, K?$$

3. Cryptography

This section gives realistic scenarios in which the mathematics behind the card trick's secret can be used. We will divulge the secret of the trick in **Section 4**. For now, we say only that the trick is related to an attack on the *discrete logarithm problem* (DLP), which also has ties to cryptography and will be discussed later in this section.

3.1 Ciphertext

Cryptography's goal is to send messages across an insecure channel to the intended recipient without eavesdroppers reading them. To achieve this goal, an encryption algorithm turns the original plaintext message into ciphertext, which in turn is sent to the intended recipient. The recipient uses a decryption algorithm to convert the message back into its original plaintext. Figuring out a cryptoquote is an example of deciphering ciphertext (**Figure 1**).

XFTBPWANWPL
is
CRYPTOQUOTE
Each letter stands for another letter. In the sample above, W
represents O and P represents T.

Cryptoquote: C TCBYHTCXEXAQ XW C FHDXEH MAU
BZUQXQJ EAMMHH XQBA BYHAUHTW
—NCZK HUFAW

Figure 1. A cryptoquote that encrypts a quotation from a famous mathematician. The correspondences in the sample (XFTBPWANWPL→CRYPTOQUOTE, with W→O and P→ T) are not necessarily the same as the correspondence in the cryptoquote to be solved. **Exercise 3** asks you to solve the cryptoquote. (Inspired by the daily cryptoquote in the *Arkansas Democrat Gazette*.)

The cryptoquote puzzle uses a very naïve encryption algorithm, in which a bijective function f maps each letter to another letter. The decryption algorithm applies the inverse of f to the ciphertext and uncovers the original plaintext.

A stronger and more useful encryption algorithm would be a function that receives plaintext and a secret key as inputs, and outputs ciphertext. The decryption algorithm would input ciphertext and a secret key and output plaintext. The key is not known to eavesdroppers. Thus, the security would rely less on the algorithm and more on the security of the keys.

3.2 Diffie-Hellman Key Exchange Protocol

How do two people agree on a key while keeping it secret? Certainly, two people can meet in person and decide on a key, but such a meeting is not always feasible. Instead, we consider a method by which two parties can agree on a secret key in communication across a possibly insecure channel, the Diffie-Hellman key exchange protocol, devised by Whitfield Diffie and Martin Hellman [1976]. An earlier version was discovered independently in 1974 by Malcolm J. Williamson of British Government Communications Headquarters but kept classified. Singh [1999, Ch. 6, 243–292] tells the stories of the discoveries.

The Diffie-Hellman key exchange protocol uses the group of nonzero integers modulo p, denoted $\mathbb{Z}_p^*$. So we remind the reader of a relevant definition:

> An element g is a *generator* of a group G if for all $x \in G$, there exists an $n \in \mathbb{N}$ such that $g^n = x$. If so, we write $\langle g \rangle = G$.

Suppose that Alice and Bob wish to exchange private information. They can arrive at a shared key through the Diffie-Hellman protocol as detailed in **Figure 2**.

1. Alice and Bob agree on a prime p and an integer g such that $1 < g < p$ and $\langle g \rangle = Z_p^*$.
2. Now, Alice and Bob secretly choose integers x and y, respectively, with $1 < x, y < p$.
3. Alice sends $X = g^x \pmod p$ to Bob and Bob sends $Y = g^y \pmod p$ to Alice.
4. Alice computes $Y^x = g^{xy} \pmod p$ while Bob computes $X^y = g^{yx} \pmod p$. They arrive at the same number, which is the *key*.

Figure 2. The Diffie-Hellman key exchange protocol.

Example 2. In practice, the value of p is often large (with length 1024 or 2048 bits) and the computations are done with a computer algebra system. We use a small p, $p = 23$ (with 5 bits, since $23 = 10111_2$), to exemplify the steps of the protocol. We first find a generator g for the group $G = \mathbb{Z}_{23}^* = \{1, 2, \ldots, 22\}$. Any such generator will have the property that

$$g^{p-1} = 1 \pmod p, \tag{1}$$

but no integer smaller than $p - 1 = 22$ satisfies **(1)**. Since

$$2^{11} = 3^{11} = 1 \pmod{23},$$

neither 2 nor 3 generates the group. Because 2 is not a generator, $4 = 2^2$ is not a generator either. After verification that $5^{22} = 1 \pmod{23}$ and no other power less than 22 satisfies $5^n = 1 \pmod{23}$, we know that 5 is a generator. (Verification can be done with Maple using the command **mod** or with Mathematica using its `Mod[m,n]`.)

Now we arbitrarily pick two numbers, say $x = 6$ and $y = 15$, and compute modulo 23:

$$X = 5^6 = (5^2)^3 = 2^3 = 8,$$
$$Y = 5^{15} = (5^2)^7 5 = (2^7)5 = (13)5 = 19.$$

Thus in G,

$$Y^x = 19^6 = (19^2)^3 = 16^3 = (3)16 = 2,$$
$$X^y = 8^{15} = (8^3)^5 = 6^5 = (6^2)^2 6 = (13^2)6 = (8)6 = 2,$$

so the key is 2. This key is then used to encrypt messages to be sent between the two parties. We have used laws of exponents to complete these computations in detail, though the use of a computer algebra system can simplify the work.

There are many ways in which a key can encrypt a message. A type of substitution cipher known as the *Caesar cipher* is credited to Julius Caesar. Applied to modern-day English letters, it would convert the letters A. . . Z to the numbers $0 \ldots 25$, respectively, and use a key k to shift each number x in the message to

$$y = x + k \pmod{26}.$$

The message could be decrypted by computing $y - k \pmod{26}$. If an eavesdropper knew the key and the algorithm, the message could be deciphered. For this reason, the security of the key needs high priority.

In the Diffie-Hellman key exchange, p, g, X, and Y are considered *public*, meaning that an eavesdropper may know the values of these variables—and can know them without compromising the security of transmissions. But if in addition an eavesdropper can determine x and y, then he or she can calculate the key k—and with k, any encrypted message can be deciphered. Thus, solving $X = g^x$ for x and then $Y = g^y$ for y is the eavesdropper's goal, since the key is $k = g^{xy}$.

If this computation were with real numbers, one could use logarithms to determine x and y. However, the Diffie-Hellman protocol uses a finite group, in which the analogue to the logarithm defined over the real numbers is not easily computed. The smallest nonnegative integer x that solves $g^x = X \pmod{n}$ is called the *discrete logarithm*, or the *index*, of X with respect to the base g modulo n; calculating it is the *discrete logarithm problem* (DLP). This problem is difficult and no efficient general solution is known. There are, however, methods faster than the brute-force guess-and-check method of taking higher powers of g until $g^x = X$ for some x.

We measure efficiency in terms of the number of computational steps needed to complete the algorithm. A *polynomial-time algorithm* is one that can be run in no more than cn^k steps where c and k are positive constants and n represents the length of the input of the algorithm. The length of the input is the number of bits needed to represent the input, not the value itself of the input. Examples of polynomial-time algorithms include addition, subtraction, multiplication, and division. There is no known method to solve the discrete logarithm problem in steps bounded by a polynomial in the length of the input. In the next section, we describe a method that runs in time $\mathcal{O}(w^{1/2}) = \mathcal{O}(e^{(\log w)/2})$, where w is a bound on the length of the inputs and $\mathcal{O}$ is big-oh notation for the order of magnitude.

But first we introduce a cryptosystem similar to the Diffie-Hellman key exchange protocol. Our intent is to show that the inherent difficulty in solving the DLP is useful in cryptography. An attack on the DLP will connect these cryptography concepts to the card trick.

3.3 ElGamal Cryptosystem

Bob selects a prime p and generator g of $\mathbb{Z}_p^*$. He also picks $a \in \mathbb{Z}_p^*$ such that $a < p-1$. He then makes his key (p, g, g^a) public. Now, suppose that Alice then wants to send Bob a message. She encrypts the message through the following steps:

1. Look up Bob's public key (p, g, g^a).
2. Convert the message into integers $m_1, m_2, m_3, \ldots, m_n$, with $m_i \in \mathbb{Z}_p$.
3. Choose a random number $b \in \mathbb{Z}_p^*$ such that $b < p-1$.
4. Compute $B = g^b \pmod{p}$ and $C_i = m_i(g^a)^b \pmod{p}$ for each i.
5. Send the ciphertext $(B, C_1, C_2, \ldots, C_n)$ to Bob.

To decode the ciphertext, Bob uses the following algorithm.

1. Use the private key a to compute $B^{p-a} \pmod{p}$, the inverse of g^{ab}.
2. Compute, for each i,

$$B^{-a}C_i = (g^b)^{-a}m_i(g^a)^b = g^{-ab}m_i g^{ab} = m_i \pmod{p}.$$

Exercises

3. Determine f and the plaintext message of the ciphertext from **Figure 1**. Hint: M→F.

In **Example 2**, we showed an example of the Diffie-Hellman key exchange protocol using 5 as a generator. In the following exercises, we explore how to find other generators of this cyclic group.

4. Lagrange's Theorem states that any subgroup H of a finite group G has the property that $|H|$ divides $|G|$:

$$|H| \,\big|\, |G|.$$

Use this theorem to determine the possible orders of $\langle x \rangle$ where $x \in \mathbb{Z}_{23}^*$.

5. Determine the other generators for $\mathbb{Z}_{23}^*$ using the assertion above and **(1)**.

6. Use a computer algebra system, plus **(1)**, to determine the smallest generator for Z_{102877}^*. The **mod**(e, m) command in Maple calculates $e \pmod{m}$, while `Mod[e,m]` does the same in Mathematica.

7. Identify the public and private information in **Example 2**. How might you determine the private numbers from the public ones?

8. In the ElGamal cryptosystem, prove the assertion that B^{p-a} is the inverse of g^{ab}.

9. Suppose that $p = 1777$, $g = 6$, and $a = 1009$. Use the ElGamal cryptosystem and either Maple or Mathematica to encipher $m = 1341$, assuming that $b = 701$.

10. Use the same values for p, g, a, and b as given in **Exercise 9** to decipher $C = 1031$ with $B = 1664$, using either Maple or Mathematica.

11. Identify the public and private components of the ElGamal cryptosystem. How do they compare to the public and private information in the Diffie-Hellman key exchange?

4. Pollard's Kangaroo Method for the DLP

We now know of cryptosystems whose security relies on the difficulty of the DLP. Even though there is no general efficient way to solve the DLP, we discuss one way to attack the DLP, which foreshadows the card trick.

4.1 Jumping Kangaroos

Let G be a finite multiplicative cyclic group generated by g, and let $X \in G$. We can think of G as the group of units in some $\mathbb{Z}_k$.

We wish to find the discrete logarithm (index) of X with respect to g, that is, to solve $g^x = X$. We begin by constructing two sequences:

- one sequence has what we consider "tame" behavior, because we know its starting position; and
- the other sequence has what we consider "wild" behavior, because it starts at a position unknown to us.

We can visualize these sequences as kangaroos jumping through the cyclic group, landing on elements of G (**Figure 3**).

To control the jumps of the kangaroos, we use a *hash function*

$$h : G \to J \subset \mathbb{Z}_k.$$

A hash function converts a large amount of data to a smaller amount; an example would be the function that converts your name to just your initials. We let our hash function h map the $|G|$ elements of G to the smaller set

$$J = \left\{1, 2, 3, 4, \ldots, \left\lfloor \sqrt{|G|} \right\rfloor\right\}$$

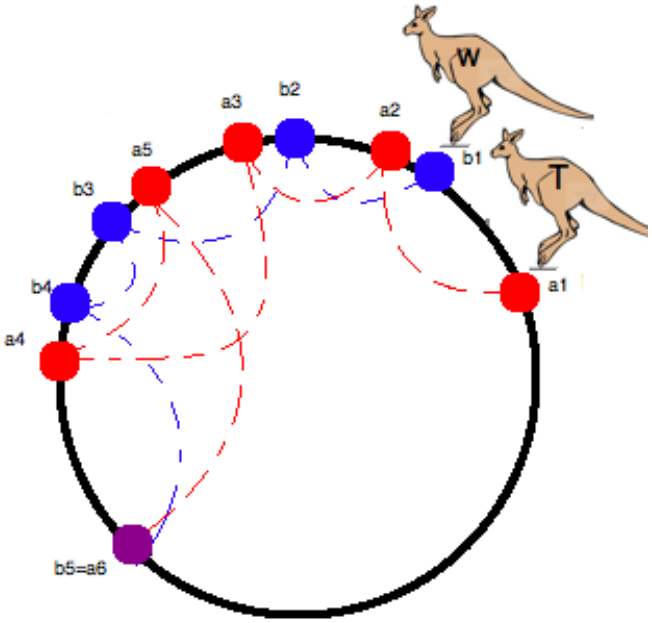


Figure 3. Conceptual diagram of the kangaroo method, with T and W denoting the tame and wild kangaroos.

of size $\lfloor\sqrt{|G|}\rfloor$, where $\lfloor x\rfloor$ is the floor function that outputs the largest integer less than or equal to x. Further explanation of the choice of this hash function (and other possibilities) will follow after this section.

Now, the tame kangaroo starts at a known value $a_0 = g$ and proceeds by jumping a distance $h(a_0)$, so that at the end of its first jump it is at $a_1 = a_0 g^{h(a_0)}$. In general, with subsequent jumps, $a_{m+1} = a_m g^{h(a_m)}$. With recursion and simplification, we have

$$a_{m+1} = g^{1+h(a_0)+h(a_1)+\ldots+h(a_m)}.$$

After $m+1$ jumps, the kangaroo has traveled a distance

$$d_{m+1} = h(a_0) + h(a_1) + \ldots + h(a_m).$$

More importantly, the path of this tame kangaroo is comprised of powers of g, the generator of G.

The wild kangaroo travels on a path with unknown starting value $b_0 = X = g^x$. It uses the same hash function to determine jump size, so that $b_{n+1} = b_n g^{h(b_n)}$. We have

$$b_{n+1} = X g^{h(b_0)+h(b_1)+\ldots+h(b_n)}.$$

Here, the distance traveled by the wild kangaroo after $n+1$ jumps is

$$d'_{n+1} = h(b_0) + h(b_1) + \ldots + h(b_n).$$

The path of the wild kangaroo is composed of X times powers of g.

Now, if at some point the wild kangaroo lands at a site visited also by the tame kangaroo, we have the equation

$$g^{1+h(a_0)+h(a_1)+\ldots+h(a_m)} = Xg^{h(b_0)+h(b_1)+\ldots+h(b_n)}. \quad \textbf{(1)}$$

Thus, $X = g^{1+h(a_0)+h(a_1)+\ldots+h(a_m)-[h(b_0)+h(b_1)+\ldots+h(b_n)]}$ and we now know the index of X. Explicitly, we have

$$x = 1 + h(a_0) + h(a_1) + \ldots + h(a_m) - [h(b_0) + h(b_1) + \ldots + h(b_n)].$$

This attack is known as *Pollard's kangaroo method*, a set of steps that can lead to a solution of the DLP (which has connection to the Diffie-Hellman key exchange and to the ElGamal cryptosystem from **Section 3**). We will use the kangaroo method to explain the success of card trick. For now, we remark that both Pollard's kangaroos and the card trick's player (and the dealer) perform jumps, the former pair through a cyclic group and the latter pair through a deck of cards. We give an example of the kangaroo method.

Example 3. Let $G = \mathbb{Z}_{29}^*$ and $g = 3$, $X = 2$. We wish to find x such that $3^x = 2 \pmod{29}$. We have j = $\lfloor\sqrt{|G|}\rfloor = 5$, so let us define the hash function $h : G \to J = \{1, 2, 3, 4, 5\}$ as follows, where h repeats with period $8 = 2s - 2$ for $s = \lfloor\sqrt{|G|}\rfloor = \lfloor\sqrt{29}\rfloor = 5$:

a	1	2	3	4	5	6	7	8	9	10	11	$\cdots$	28
$h(a)$	1	2	3	4	5	4	3	2	1	2	3	$\cdots$	4

Then, modulo 29 we have

$$\begin{aligned}
a_0 &= 3,\\
a_1 &= 3 \cdot 3^{h(3)} = 3 \cdot 3^3 = 3^4 = 23,\\
a_2 &= 3^4 \cdot 3^{h(23)} = 3^4 \cdot 3^3 = 3^7 = 12,\\
a_3 &= 3^7 \cdot 3^{h(12)} = 3^7 \cdot 3^4 = 3^{11} = 15,\\
a_4 &= 3^{11} \cdot 3^{h(15)} = 3^{11} \cdot 3^3 = 3^{14} = 28,\\
a_5 &= 3^{14} \cdot 3^{h(28)} = 3^{14} \cdot 3^4 = 3^{18} = 6,\\
a_6 &= 3^{18} \cdot 3^{h(6)} = 3^{18} \cdot 3^4 = 3^{22} = 22,
\end{aligned}$$

and

$$\begin{aligned}
b_0 &= 2,\\
b_1 &= 2 \cdot 3^{h(2)} = 2 \cdot 3^2 = 18,\\
b_2 &= 2 \cdot 3^2 \cdot 3^{h(18)} = 2 \cdot 3^4 = 17,\\
b_3 &= 2 \cdot 3^4 \cdot 3^{h(17)} = 2 \cdot 3^5 = 22.
\end{aligned}$$

Since $a_6 = b_3$, we have $3^{22} = 2 \cdot 3^5 \pmod{29}$. Solving for 2 results in $3^{17} = 2 \pmod{29}$. Thus, in only 9 calculations, we were able to

determine the index of 2 in this group relative to the generator 3. You can verify the answer by using the function **mod** in Maple or the function `Mod` in Mathematica.

4.2 Analysis of Pollard's Kangaroo Method

To analyze the method, we assume that:

The landing positions of the kangaroos are independent random samples from a uniform distribution of the elements of G.

Now, suppose that the tame kangaroo jumps a total of $c\sqrt{|G|}$ times for some constant c. Then, at every jump made by the wild kangaroo, there is a chance

$$\frac{c\sqrt{|G|}}{|G|} = \frac{c}{\sqrt{|G|}}$$

that the wild kangaroo lands on the tame kangaroo's path.

Suppose instead that the wild kangaroo misses the tame kangaroo's path at each of $\sqrt{|G|}$ jumps. This occurs with probability:

$$\left(1 - \frac{c}{\sqrt{|G|}}\right)^{\sqrt{|G|}} = \left(1 + \frac{1}{\frac{-\sqrt{|G|}}{c}}\right)^{\frac{-\sqrt{|G|}}{c}(-c)} \approx e^{-c},$$

using the facts from calculus that

$$\left(1 + \frac{1}{n}\right)^{n} \approx e, \qquad \left(1 - \frac{1}{n}\right)^{-n} \approx e$$

for large n. We want c to be small compared to $\sqrt{|G|}$, since the larger $\sqrt{|G|}/c$, the closer the approximation to e^{-c}. In addition, we do not want to calculate an infeasible number of jumps. So, the probability of failure is small when c is sufficiently large, since e^{-c} will be small. For instance,

c	1	2	3	4	5
e^{-c}	.3679	.1353	.0498	.0183	.0067

We credit the above analysis to Lacey [2002]; Pollard himself computes a similar analysis [1978].

4.3 Hash Functions

As promised, we discuss the choice of the hash function h, which establishes the jump sizes. It may seem a bit arbitrary. Pollard [2000a, 438] states

> . . . the methods work well when the jumps of the kangaroos are powers of two. . . or powers of another number. We are not claiming that these are the only good choices. Possibly most sufficiently large sets are good choices.

He cites Oorschot and Wiener [1999] along with his own work [1978].

In the exercises, you will explore a hash function using powers of 2. But first, we provide an example of a poorly chosen hash function that results in the method not succeeding.

Example 4. Given $G = \mathbb{Z}_7^*$, which is generated by 3, suppose that we want to know the value of x for which $3^x = 6$. If we (foolishly) take $h(y) = 6$ for all $y \in G$, then $a_0 = g = 3$ and

$$a_{i+1} = a_i 3^{h(a_i)} = a_i 3^6 = a_i.$$

Since $a_i = 3$ for all i, the tame kangaroo's sequence is $\{3, 3, 3, \ldots\}$. Now, $b_0 = 6$ and

$$b_{i+1} = b_i 3^{h(b_i)} = b_i 3^6 = b_i.$$

Thus, the wild kangaroo's sequence is $\{6, 6, 6, \ldots\}$, and a collision will never occur.

The DLP was not solved, and the blame lies solely with the hash function. We need a hash function that inserts some variety into each kangaroo's sequence; a kangaroo that jumps in place is not of much use.

A hash function that guarantees success is the constant function $h(y) = 1$ for all $y \in G = \langle g \rangle$. The tame kangaroo's sequence is $\{g, g^2, g^3, \ldots\}$, so this kangaroo eventually jumps to each element in the group and hence must collide with the wild kangaroo at some point. But this hash function is no better than a guess-and-check method.

4.4 The Secret

Returning to the card trick, we relate it to the kangaroo method and use a similar analysis. Recall that the player's sequence is initialized by choosing a number between 1 and 10, followed by counting until the number of cards dealt is precisely the chosen number. From there, the first key card, the player secretly notes the value of the card (aces worth 1, face cards worth 5, and all others worth face value), with the player continuing to count from key card to key card. While the player is constructing a sequence through

the deck, the dealer constructs her own sequence in the same manner. The prediction that the dealer makes of the player's last secret card is actually the last card in the dealer's sequence. Simply put, this is the trick's secret.

Most of the time, the trick ends with the dealer correctly guessing the player's last secret card. So, at some point, either at the last card or before, the two sequences met at the same card. Just as in the kangaroo method, we have two sequences running through a set. We can view the dealer's and player's sequences as tame and wild kangaroos, respectively. Visualizing the two paths becoming one resembles the lowercase Greek letter λ. For this reason, Pollard's kangaroo method is sometimes referred to as the λ-*method*. **Figure 4** shows an example, using the sequences from **Example 3**.

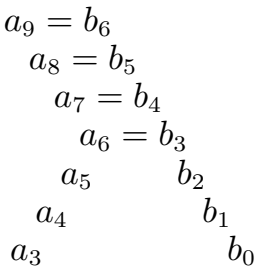


Figure 4. Graphical illustration of why the Pollard kangaroo method is sometimes called the λ-method.

Exercises

12. Use $G = \mathbb{Z}_{13}^*$ with $g = 6$ and $X = 3$ to determine $x \in \mathbb{Z}_{>0}$ such that $g^x = 6^x = X = 3 \pmod{13}$—that is, find the index of 4 in $\mathbb{Z}_{13}^*$ with respect to the generator 6. Define $h : G \to J = \{1, 2, 3\}$ by the table below, where h repeats modulo $4 = 2s - 2$ for $s = \lfloor \sqrt{|G|} \rfloor = \lfloor \sqrt{13} \rfloor = 3$. It may be useful to note that $6^{12} = 1$ and $6^{-1} = 6^{11}$.

a	1	2	3	4	5	6	7	8	9	10	11	12
$h(a)$	1	2	3	2	1	2	3	2	1	2	3	2

13. Use $G = \mathbb{Z}_{102877}^*$ and the generator g found in **Exercise 6** to apply Pollard's kangaroo method to determine x for which $g^x = 7 \pmod{102877}$. Use either the Maple or the Mathematica code in **Appendix A**, which take as inputs a prime p, a generator g, and an integer $X \in Z_p^*$. The output is the index of X in Z_p^* with respect to the generator g, arrived at by Pollard's kangaroo method. (The result can be confirmed using the command `MultiplicativeOrder[g,p,{X}]` in Mathematica.)

14. In the computer code in **Appendix A**, we have $s = \lfloor \sqrt{|G|} \rfloor$ and the hash function h:

Hash function in code of Appendix A

a	1	2	3	$\cdots$	s	$s+1$	$s+2$	$\cdots$	$2s-2$	$2s-1$	$2s$	$2s+1$	$\cdots$
$h(a)$	1	2	3	$\cdots$	s	$s-1$	$s-2$	$\cdots$	2	1	2	3	$\cdots$

Rewrite either the Maple or the Mathematica code using the following different hash function h, where t is chosen such that $2^t < s$:

New hash function

a	1	2	3	$\cdots$	$t+1$	$t+2$	$t+3$	$\cdots$	$2t$	$2t+1$	$2t+2$	$2t+3$	$\cdots$
$h(a)$	2^0	2^1	2^2	$\cdots$	2^t	2^{t-1}	2^{t-2}	$\cdots$	2^1	2^0	2^1	2^2	$\cdots$

15. Run both the computer code from **Appendix A** and the revised code from **Exercise 14** using the numbers from **Exercise 13**. Does either hash function lead to a quicker collision of the sequences? What did you compare to come to this conclusion? Rerun the code using a few different inputs to determine if one hash function leads to quicker collisions.

16. Prove that if $a_n = b_m$ in Pollard's kangaroo method, then $a_{n+k} = b_{m+k}$ for all $k \geq 0$.

5. Markov Chains

Thus far, this Module has explored cryptography, the discrete logarithm problem, and Pollard's kangaroo method, all of which relate to the card trick described in the Introduction. In the card trick, the "jump" to the player's next key card depends on only the current key card. For instance, suppose that the player's first six key cards are $8, 8, 5, K, K, 3$ (as in **Example 1**). The next key card would be three cards from the 3 card. The important point is that the next "jump" only depends on the current card, not on the previous cards in the sequence. This is similar to Pollard's kangaroo method, in which the next position depends only on the current position.

Both the card trick and the kangaroo method can be represented in terms of Markov chains, a useful tool to analyze long-term behavior of models using probability and matrix theory. Below we summarize results about Markov chains from Grinstead [1997], Kemeny and Snell [1960], and Roberts [1976], to which we refer the reader for further details and proofs of the results.

5.1 Results about Markov Chains

A *Markov process* (or *Markov chain*) is a system of states in which the probability of moving from one state to another depends only on the current state, not on states visited in the past.

We can display the probabilities involved in a *transition matrix*, with the previous states corresponding to rows and the next state corresponding to columns. A simple example for a system with three states is:

$$T = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} .5 & 0 & 0.5 \\ 0 & 0.5 & 0.5 \\ 0.25 & 0.25 & 0.5 \end{pmatrix} \end{array}.$$

Theorem. *For a Markov process with transition matrix T and states $1, 2, \ldots, n$, the probability that the process will be in state j after k steps given that it started in state i is $T^k_{i,j}$, for $1 \le i, j \le n$.*

For a Markov chain with transition matrix, T, suppose that we are given the probabilities that the process begins in each of the n states as a vector $v = \{v_1, v_2, \ldots, v_n\}$. Then vT^k is the vector whose jth component gives the probability that the process is in state j after k steps.

A set of states is a *transient set* if there is a way in which to leave the set, and once the process leaves the set, there is zero probability that the process will return to any state in the set. A state is *transient* if it is in a transient set.

An *absorbing state* is a state from which one cannot leave. An *absorbing Markov chain* is a Markov chain that contains at least one absorbing state and a path from each nonabsorbing state to an absorbing state.

An absorbing Markov chain can be simplified to a *canonical form*, as shown in **Figure 5**.

$$\begin{array}{l} \\ \text{absorbing states} \\ \text{nonabsorbing states} \end{array} \begin{array}{c} \begin{array}{cc} \text{absorbing states} & \text{nonabsorbing states} \end{array} \\ \begin{pmatrix} I & 0 \\ R & Q \end{pmatrix} \end{array}$$

Figure 5. Canonical form of a transition matrix for an absorbing Markov chain.

Let there be m absorbing states and $n - m$ nonabsorbing states. Then

- I is the $m \times m$ identity matrix composed of absorbing states,
- 0 is the $m \times (n - m)$ zero matrix,
- R is an $(n - m) \times m$ matrix, and
- Q is an $(n - m) \times (n - m)$ matrix composed of probabilities that the process moves from a transient state to a transient state.

The *fundamental matrix* for an absorbing Markov chain with transition matrix in canonical form is $N = (I - Q)^{-1}$.

The fundamental matrix N has useful properties:

- The entries of N give the expected value of the number of times the process is in each nonabsorbing state for each possible starting state.
- The sum of a row i of N is the expected value of the number of steps before the process is absorbed, assuming that it started in state i.

- NR is the matrix whose entries are the probabilities that the process ends in a particular absorbing state for each possible nonabsorbing starting state.

Exercises

17. Explain why each row of a transition matrix T must sum to 1.

18. An equivalence relation $\sim$ can be placed on a Markov chain: Two states, s_i and s_j, are considered equivalent and we write $s_i \sim s_j$ if either $i = j$ or there is a path from s_i to s_j and a path from s_j to s_i. Verify that $\sim$ is an equivalence relation. This equivalence relation partitions the states into two equivalence classes.

19. Explain why an absorbing chain cannot have a transition matrix with the property that $T_{ij}^n > 0$ for some n and all i, j.

20. For an absorbing Markov chain with transition matrix T, where we have $\lim_{n\to\infty} Q^n = 0$, confirm that $I - Q$ has an inverse by finding an explicit form for it.

21. Describe a scenario in which $NR = [1]$, the 1×1 identity matrix.

6. A Simplified Kruskal Count as a Markov Process

To view the card trick as a Markov process, we must first modify the deck to lead to more-feasible probability calculations. As with the original trick, we distinguish only the face value of the card, not the suit or color. We follow the treatment in Haga and Robins [1997]: To give every value equal probability, we toss out the face cards; and we assume an infinite random deck, with each value 1 through 10 having an equal and independent chance of being at any position in the deck.

We define the states of the Markov process as corresponding to the *distances*—the numbers of cards—between the dealer's current key card and the player's key card that is immediately ahead of it, or zero if they are the same card. With cards valued 1 though 10, this distance can be 0 through 9, so we have 10 states in the Markov chain. The state corresponding to the distance 0 is absorbing, since when the player's card and the dealer's card are the same, their paths have converged and the distance between each successive card remains 0. We define entry $M_{i,j}$ in the transition matrix to be the probability that the distance goes from i to j.

We illustrate with an example using a deck of cards whose values are only 1 and 2.

Example 5. Suppose that we have an infinite deck of cards with values 1 and 2, with each value equally likely at each position and independent of values at all other positions. For the card trick, the states of the Markov chain correspond to the number of cards between the dealer's current card and the closest player's card that is even with or ahead of the dealer's card, so the states are 0 and 1. The card trick is initialized by the player secretly choosing a number from $\{1, 2\}$ and then counting that number of cards to the next key card. We compute each entry of the transition matrix:

- $M_{0,0}$: the probability that the distance between the sequences starts at 0 and stays at 0 after one turn. If the distance is 0, then the player and dealer are on the same card and will move the same number of spaces to again be on the same card after one turn. Therefore, $M_{0,0} = 1$.
- $M_{0,1}$: the probability that the distance between the sequences starts at 0 and then becomes 1 after a turn. By the preceding explanation, this is not possible. So, $M_{0,1} = 0$.
- $M_{1,1}$: the probability that the distance between the sequences starts at 1 and stays 1 after a turn. If the distance is 1, then the player is on the card directly after the dealer's card. The equally probable options for two successive cards are: 1 1, 1 2, 2 1, 2 2. The first three cases result in the player and dealer's sequences colliding on the next turn. When the dealer and player are on cards 2 2, they are 1 apart and will stay 1 apart after completing the turn. See **Figure 6**, in which d_i and p_j represent the dealer's ith card and player's jth card. The cases in the figure confirm that $M_{1,1} = \frac{1}{4}$.

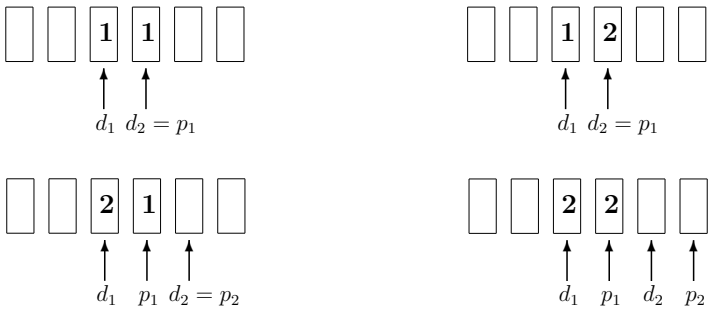


Figure 6. Cases for **Example 5** relating to $M(1, 1)$.

- $M_{1,0}$: the probability that the distance between the sequences starts at 1 and then becomes 0 after a turn. By the explanation above, $M_{1,0} = \frac{3}{4}$ and

$$M = \begin{bmatrix} 1 & 0 \\ 3/4 & 1/4 \end{bmatrix}.$$

This matrix is already in canonical form with $N = (I - Q)^{-1} = [4/3]$. According to the results about Markov chains, the expected number of turns until the process is absorbed is $4/3$.

Now we explore what happens for a deck of cards with 10 independently and identically distributed values. The transition matrix M is a 10×10 matrix indexed from 0 to 9. From Haga and Robins [1997], we have the following two theorems.

Theorem. *The transition matrix M satisfies*

$$(a)\ M_{0,0} = 1 \ \text{ and } \ M_{0,j} = 0, \qquad \text{for } \ 0 \le j \le 9;$$

$$(b)\ M_{9,9} = \frac{1}{100} \ \text{ and } \ M_{9,j} = \frac{1}{10}\left(1 + \frac{1}{10}\right), \qquad \text{for } 0 \le j \le 8;$$

$$(c)\ M_{i,j} = \left(1 + \frac{1}{10}\right) M_{i+1,j}, \qquad \text{for } 0 < i \ne j < 9;$$

$$(d)\ M_{i,i} = \left(1 + \frac{1}{10}\right) M_{i+1,i} - \frac{1}{10}, \qquad \text{for } 0 < i < 9.$$

Proof: (a) Since a distance of 0 is the absorbing state, once the distance between the cards is 0, it will remain 0. Thus, $M_{0,0} = 1$ and $M_{0,j} = 0$ for $1 \le j \le 9$.

(b) The only way in which the distance between the cards is 9 and stays 9 is if both the dealer and the player have cards of value 10. This occurs with probability $\left(\frac{1}{10}\right)\left(\frac{1}{10}\right)$. Thus, $M_{9,9} = \frac{1}{100}$.

(c), (d) The remainder of the proof is in **Appendix B**. □

By induction, Haga and Robins [1997] get a closed form for M:

Theorem.

$$M_{i,i} = \frac{1}{10}\left[\left(1 + \frac{1}{10}\right)^{10-i} - 1\right], \qquad \text{for } 1 \le i \le 8;$$

$$M_{i,j} = \frac{1}{10}\left(1 + \frac{1}{10}\right)^{10-i}, \qquad \text{for } 0 < j < i < 9;$$

$$M_{i,j} = \frac{1}{10}\left(1 + \frac{1}{10}\right)^{j-i}\left[\left(1 + \frac{1}{10}\right)^{10-j} - 1\right], \quad \text{for } 0 < i < j \le 9.$$

Example 6. We verify the entry for $M_{8,9}$ by examining all possible paths that might be taken for the distance between the dealer and player to go from 8 to 9.

First, we suppose that the player's key card is 8 cards ahead of the dealer. The distance will become 9 if either of the following occurs:

- The dealer's key card is a 9 and the player's key card is a 10. Since all card values have equal and independent probability $1/10$, this occurs with probability $1/10^2$.
- The dealer's key card is a 10, the player's current key card is a 1, and the player's next key card is a 10. This case occurs with probability $1/10^3$.

The sum of these two probabilities is

$$M_{8,9} = \frac{1}{10^2} + \frac{1}{10^3} = \frac{1}{10}\left(1 + \frac{1}{10}\right)\frac{1}{10},$$

as stated in the theorem.

A more detailed construction of this matrix is done in Haga and Robins [1997], where there is also a general form for the transition matrix M for a deck with values 1 through m.

The closed form of the matrix M is already in canonical form, so we can calculate the matrices Q, R, and N. Since N provides the most important output, we explicitly calculate it:

$$\frac{1}{11}\begin{bmatrix}
20 & 8 & 7 & 6 & 5 & 4 & 3 & 2 & 1 \\
10 & 19 & 7 & 6 & 5 & 4 & 3 & 2 & 1 \\
10 & 9 & 18 & 6 & 5 & 4 & 3 & 2 & 1 \\
10 & 9 & 8 & 17 & 5 & 4 & 3 & 2 & 1 \\
10 & 9 & 8 & 7 & 16 & 4 & 3 & 2 & 1 \\
10 & 9 & 8 & 7 & 6 & 15 & 3 & 2 & 1 \\
10 & 9 & 8 & 7 & 6 & 5 & 14 & 2 & 1 \\
10 & 9 & 8 & 7 & 6 & 5 & 4 & 13 & 1 \\
10 & 9 & 8 & 7 & 6 & 5 & 4 & 3 & 12
\end{bmatrix}$$

Thus, summing the rows of N yields a 9×1 matrix:

$$\frac{1}{11}\begin{bmatrix}56 & 57 & 58 & 59 & 60 & 61 & 62 & 63 & 64\end{bmatrix}^T.$$

Recall that the summed rows of matrix N yield the expected values of the number of steps before the process is absorbed for each nonabsorbing starting state. For instance, if we start in state 1, meaning that the distance between the dealer's card is 1 away from the player's card, we expect $56/11 = 5.\overline{09}$ turns before the distance between the cards becomes 0. Over the nine nonabsorbing states, the largest such value is $64/11 = 5.\overline{81}$ turns, for a distance of 9. Though the deck is potentially infinite, we expect success after only a few turns! While this is not a concrete argument for why the card trick will work a majority of the time, it provides evidence for that

conclusion. For a more mathematically convincing rationale that the trick works approximately $5/6$ of the time, see Lagarias et al. [2009] and Pollard [2000a].

Exercise

22. Verify that the entry for $M_{7,9} = \frac{1}{10}\left(1+\frac{1}{10}\right)^2\left(\frac{1}{10}\right)$ by determining all possible paths that might be taken for the distance between the dealer and player to go from 7 to 9.

7. The Kruskal Count

The trick that we have been discussing is credited to mathematician Martin D. Kruskal and is sometimes called the *Kruskal Count*.

7.1 How to Increase the Chance of the Trick's Success

Further analysis has shown that the dealer can increase the chance of correctly guessing the player's last card by

- **using two decks of cards instead of one:** Using more cards lengthens the dealer's and player's paths, thus increasing the likelihood that the two paths will eventually meet.
- **decreasing the value of the face cards:** Decreasing the value of the face cards increases the average number of cards on the dealer's and player's paths, making the probability of collision greater. This would be the result of a version of the trick where each face card has value equal to the number of letters in its name: A jack is worth 4, queen worth 5, and king worth 4.
- **starting the dealer's sequence with the first card rather than randomly picking a digit 1 to 10:** By starting on the first card (i.e., choosing a secret number to be 1), the dealer potentially lengthens the dealer's sequence of cards, which increases the probability that the player will land on one of them.

In each of these modifications, we increase the number of jumps by the kangaroos, thereby making the chance of landing on the same card more probable. For a rigorous analysis of alterations that improve the probability of success, see Lagarias et al. [2009].

7.2 Estimating the Chance of Success

The Markov chain analysis in **Section 6** provides an expected value until success, but not a probability of success. So we consider a second Markov

chain for the card trick, one that allows us to bound the probability of failure of the trick. The technique used to examine two independent lines of chains is common within the theory of Markov chains. Lagarias et al. [2009] produce a "reduced" version; we provide details of an unreduced version of the chain.

Define the Markov chain as follows. We deal from two independent decks of cards with independent and equally distributed values from the set $S = 1, 2, \ldots, L$. The dealer plays the card trick on one deck and the player uses the other deck. Both choose a number from S; and starting from the top of their decks, they simultaneously count down the cards until their number is reached. The card on which each lands is the first key card for each.

We consider one card to be *behind* another if it is closer to the top of its deck.

- Whichever card is behind the other moves again until it is ahead of the other deck's key card. [This is similar to the order of turns in the game of golf.] As before, the value of the key card is the number of cards counted until reaching the next key card.
- If both cards are an equal distance from the top of the deck, both the player and the dealer move onto the next key card in their deck.

We define the *distance* as the difference between the player's current key card and the top of the deck less the difference between the dealer's current key card and the top of the deck. The distance can range from $-(L-1)$ to $L-1$.

The states of the Markov chain are the $2L-1$ distances

$$-(L-1), -(L-2), \ldots, -1, 0, 1, \ldots, L-2, L-1.$$

The time until the distance becomes 0 is called the *coupling time*, which we denote by τ. We perform the actions described above until the Nth card in one of the decks is passed. We calculate the probability of failure of the decks to couple by the Nth card, i.e., $P(\tau > N)$.

First, we define the random variable $Z_{L,N}$ to be the total number of key cards in both decks up to and including the Nth card. Thus, our probability space is $\Omega = \bigcup_{k=0}^{\infty} \{Z_{L,N} = k\}$. So,

$$\{\omega \in \Omega \mid \tau(\omega) > N\} = \{\omega \in \Omega \mid \tau(\omega) > N\} \cap \left(\bigcup_{k=0}^{\infty} \{Z_{L,N} = k\} \right)$$

$$\implies \{\tau > N\} = \bigcup_{k=0}^{\infty} \left(\{\tau > N\} \cap \{Z_{L,N} = k\} \right)$$

$$\implies P(\tau > N) = \sum_{k=0}^{\infty} P(\tau > N,\ Z_{L,N} = k)$$

$$= P(\tau > N \mid Z_{L,N} = k)\, P(Z_{L,N} = k)$$

$$= \sum_{k=0}^{\infty} \left(\frac{L-1}{L}\right)^k \cdot P\left(Z_{L,N} = k\right)$$
$$= E\left[\left(1 - \frac{1}{L}\right)^{Z_{L,N}}\right].$$

Since $Z_{L,N} \geq 2N/L$, we have

$$\left(1 - \frac{1}{L}\right)^{Z_{L,N}} \leq \left(1 - \frac{1}{L}\right)^{2N/L}.$$

Thus,

$$E\left[\left(1 - \frac{1}{L}\right)^{Z_{L,N}}\right] \leq E\left[\left(1 - \frac{1}{L}\right)^{2N/L}\right] = \left(1 - \frac{1}{L}\right)^{2N/L}.$$

Therefore, $P(\tau > N) \leq \left(1 - \frac{1}{L}\right)^{2N/L}$.

Exercises

23. Suppose $L = 5$. This means that the two decks are composed of a uniform distribution of five cards: Ace, 2, 3, 4, and 5. Determine the transition matrix T of size 5×5 for this Markov chain. Why can't we use the analysis performed in **Section 6** on this Markov chain?

24. Assuming $L = 10$ (because a standard deck of cards has card values 1, 2, 3, ..., 10) and $N = 52$, compute the probability that the trick will succeed.

8. Other Results and Open Problems

The DLP's difficulty gives rise to several algorithms in cryptography. Apart from the Diffie-Hellman key exchange and the ElGamal cryptosystem (whose security relies on the DLP), other cryptosystems using the DLP include the U.S. Government's Digital Signature Algorithm and its elliptic curve analogue [Teske 2001].

Montenegro and Tetali [2009] provide bounds on the probability of success that supplement work by Pollard [2000a].

As is common practice, in our analysis we took a uniformly-distributed deck of cards. However, a standard deck of 52 cards with the original assignment of card values does not have equally likely card values; so our calculations did not give an exact calculation of the probability trick's success. Pollard himself admits that "An exact calculation seems difficult"

[Pollard 2000b]. He gives an approximate calculation to show that we expect the trick to succeed with probability 89.3%, together with a simulation that admits 85.4% success.

Grime [n.d.] uses a geometric distribution and gives an approximate probability of success to be 83.88%. Grime also includes several other interesting results, including probabilities on the placement of the final card in the sequence and how many possible different last cards can occur. In his simulation, 58.39% of the decks had the following property: Independent of the 10 possible starting positions, the final card reached was the same. Moreover, taking all 10 starting positions into account, 97.9% of decks will have no more than two different possible final cards. We end this discussion with proof of a result of Pollard that any deck of cards cannot have more than six different final cards.

The goal is to show that seven final cards or more is impossible. To do so, we examine the length of a path through the deck. We show that the total of the minimum possible lengths of seven distinct paths is longer than the total value of all the cards in the entire deck. For a standard 52-card deck and the usual card values, that total value is

$$4 \cdot (1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9 + 10 + 5 + 5 + 5) = 280.$$

We look at the 10 sequences formed by the 10 starting positions. Suppose that these result in seven different final cards. Now, take one additional turn with each of these seven cards; doing so will lead beyond the last of the 52 cards, but we just imagine the deck to extend a bit further. The total length of each card's extended path is minimized if the path ends right on the imaginary 53rd card (let it have value 0). The lengths of the extended paths that end at the 53rd card and start at cards 4 through 10 (which are the shortest seven such paths) are

$$\begin{aligned} 53 - 4 &= 49 \\ 53 - 5 &= 48 \\ 53 - 6 &= 47 \\ 53 - 7 &= 46 \\ 53 - 8 &= 45 \\ 53 - 9 &= 44 \\ 53 - 10 &= 43 \end{aligned}$$

for a total length of 322, which is greater than 280. This makes seven paths impossible.

Six paths have a minimum total length of $322 - 49 = 273$, so six is the upper bound on the number of final cards possible at the end of the trick. However, no example of an ordered deck with six final cards appears in the literature. The minimum total length discussed would involve starting the six paths on cards 5 through 10 and finishing on cards with values

$6, 5, 4, 3, 2, 1$ where the value 6 would appear on the 47th card in the deck, value 5 on the 48th, etc.

9. Conclusion

The purpose of this Module was to introduce Pollard's kangaroo method via an interesting hook (the card trick). The Diffie-Hellman key exchange and the ElGamal cryptosystem both rely on the difficulty of solving the DLP, for which the kangaroo method provides an attack. We used Markov chains to achieve results about the card trick, concluding that it is successful most of the time. We encourage the reader to attempt the role of dealer and perform the trick on an audience!

10. Answers to Exercises

1. Regardless of which secret number $(1, 2, \ldots, 10)$ is chosen, the last card will be the 51st card in the deck, a Queen.

2.

Secret Number	Last Card
$1, 2, 3, 4, 5, 7, 9$	fourth Queen
$6, 8, 10$	first King

3. Not all letters were encrypted, so we skip several in the definition of f in **Table S1**. Quote: "A mathematician is a device for turning coffee into theorems." —Paul Erdos (Erdős).

x	A	C	D	E	F	G	H	I	L	M	N	O	P	R	S	T	U	V
$f(x)$	C	E	F	H	M	J	Y	X	K	T	Q	A	N	U	W	B	Z	D

Table S1. Solution for Exercise 3.

4. $|\langle x \rangle| \in \{1, 2, 11, 22\}$.

5. We need to find x such that $x^1 \neq 1$, $x^2 \neq 1$, $x^{11} \neq 1$, but $x^{22} = 1$. The acceptable x are 5, 7, 10, 11, 14, 15, 17, 19, 20, and 21. The Mathematica command `Solve[x^11==22,{x},Modulus->23]` gives these values, plus 22, which does not satisfy $x^2 \neq 1$.

6. Since $|\mathbb{Z}^*_{102877}| = 102876 = 2^2 \cdot 3 \cdot 8573$, we need to find x such that $x \neq 1$, $x^2 \neq 1$, $x^3 \neq 1$, $x^4 \neq 1$, $x^6 \neq 1$, $x^{12} \neq 1$, $x^{8573} \neq 1$, $x^{17146} \neq 1$, $x^{25719} \neq 1$, $x^{34292} \neq 1$, and $x^{51428} \neq 1$, but $x^{102876} = 1$. Using Maple or Mathematica, it can be verified that $x = 2$ qualifies (and of course is the smallest such).

7. The public information is $X = 8, Y = 19, p = 23$, and $g = 5$. The private information is $Y^x = 2 = X^y$. This can be calculated if one can solve $19^x = 2$ for x or $8^y = 2$ for y with computations in Z^*_{23}.

8. $B^{p-a} g^{ab} = (g^b)^{p-a} g^{ab} = g^{bp-ab+ab} = (g^p)^b = 1^b = 1$.

9. $B = 1664$ and $C = 1103$. The ciphertext is $(1664, 1103)$.

10. $m = 108$.

11. The public information is p, g, g^a, $B = g^b$, and $C = mg^{ab}$. The private information is a, b, and m. Since we know g, g^b, and p, if we can solve the DLP, we can compute a and b, which will lead to m. This is the same as the Diffie-Hellman key exchange in which an eavesdropper tries to solve for x if Y and Y^x are known.

12.
$$\begin{aligned} a_0 &= 6, & b_0 &= 4, \\ a_1 &= 6 \cdot 6^2 = 6^3 = 8, & b_1 &= 4 \cdot 6^2 = 1, \\ a_2 &= 6^3 \cdot 6^2 = 6^5 = 2, & b_2 &= 4 \cdot 6^2 \cdot 6 = 4 \cdot 6^3 = 6. \end{aligned}$$
The collision occurs when $a_0 = 6 = b_2$. Thus, $6 = 4 \cdot 6^3$ and $6^{-2} = 4$. Since $6^{-1} = 6^{11}$, we have $4 = 6^{22} = 6^{12+10} = 6^{12}6^{10} = 6^{10}$.

13. 86843.

14. The new **Index2** code with changes is shown in **Figure S1** (Maple) and in **Figure S2** (Mathematica) on the following pages.

15. Running **Index2** yields the same answer of 86843 that **Index** provided. The sequences a_i and b_j are longer for **Index2**. Running each program on the numbers from **Example 3** results in sequences of equal length.

As another example, it can be verified that 1987 is prime with generator 2. Suppose that we want to find the index of 1000 for that generator. Both programs yield the answer $2^{1356} = 1000$, but again **Index2** shows longer sequences for a_i and b_j.

16. Suppose that $a_n = b_m$; then $a_{n+1} = a_n g^{h(a_n)} = b_m g^{h(b_m)} = b_{m+1}$. Now, suppose that $a_{n+k} = b_{m+k}$ for some $k \geq 1$. Then

$$a_{n+k+1} = a_{n+k} g^{h(a_{n+k})} = b_{m+k} g^{h(b_{m+k})} = b_{m+k+1}.$$

```
> Index2 := proc(p, g, X)
  local s, L1, L2, p1, p2, i, k, m, n, h, f, z, y, t;
  s := floor(sqrt(p)); L1 := [g]; p1 := [1];
  t := floor( ln(floor(sqrt(p))) / ln(2) );
   print t;
  L2 := [X]; p2 := [0]; i := 1; k := 0;
  while k = 0 do
      for z in L1 do
          for y in L2 do
              if (k = 0) and (z = y) then k := 1;
              print(z, y); member(z, L1,'u'); member(y, L2,'v');
             break; end if;
           end do
      end do;
      m := mod(op(i, L1) - 1, 2 ·t); n := mod(op(i, L2) - 1, 2·t );
      if m ≤ t then j := 2^m; else h := 2^(2t - m); end if;
      L1 := [op(L1), mod(op(i, L1)·g^h, p)];
      p1 := [op(p1), op(i, p1) + h];
      if n ≤ t then f := 2^n; else f := 2^(2t - m); end if;
      L2 := [op(L2), mod(op(i, L2)·g^f, p)];
      p2 := [op(p2), op(i, p2) + f];
      i := i + 1;
  end do;
  print(L1); print(L2); print(p1); print(p2);
  if op(u, p1) - op(v, p2) > 0 then print(op(u, p1) - op(v, p2));
  else print(p - 1 + (op(u, p1) - op(v, p2)));end if;
  end proc;
```

Figure S1. Revised Maple code for solution to **Exercise 14**, with changes in red (light gray).

17. The columns of a transition matrix T represent all possible states of the process. Given that the process is currently in row i, it either stays in state i or moves to one of the other states by the next move. As such, $\sum_j T_{i,j} = 1$.

18. Symmetric: If $s_i \sim s_j$ then $s_j \sim s_i$ by definition. Reflexive: Additionally, $s_i \sim s_i$ by definition. Transitive: For $s_i \sim s_j$ and $s_j \sim s_k$, there is a path from s_i to s_j to s_k and a path from s_k to s_j to s_i, so $s_i \sim s_k$.

19. In an absorbing Markov chain's transition matrix, there must be an entry of 1 somewhere and all other entries in that row must be 0. For any power of this matrix, the same argument will hold.

```
Index2[p_, g_, X_](* prime, generator, element*) := Module[
  {s, Collision, L1, L2, p1, p2, i, u, v, Jump}, (*local vars*)
  s = Floor[Sqrt[p]]; (*size of hash table*)
  Collision = False;
  L1 = {g}; (*tame kangaroos positions*)
  L2 = {X}; (*wild kangaroos positions*)
  p1 = {1}; (*power of g for tame kangaroo*)
  p2 = {0}; (*power of g for wild kangaroo*)
  u = 0;    (*jump number for tame kangaroo at collision*)
  v = 0;    (*jump number for wild kangaroo at collision*)
  i = 1;    (*counter for jumps*)
  Jump[L_, plist_] := Module[{m, h, t},  (*local vars*)
    t = Floor[Log[2, s]];                (*log to base 2*)
    m = Mod[L[[i]] - 1, 2 * t];          (*hash to find jump size*)
    If[m ≤ t,                            (*then*)
     h = 2^m,                            (*else*)
     h = 2^(2 * t - m)];
    {Append[L, Mod[(Last[L] * g^h), p]], (*add new jump location*)
     Append[plist, Last[plist] + h]}];   (*record power of g for new location*)
  While[(Collision == False && i < p),   (*no collision and not exhausted jumps*)
   If[Intersection[L1, L2] ≠ {},         (*if jump location in common*)
                                         (*then*)
    (Collision = True;                   (*note fact of collision*)
     u = Flatten[Position[L1, Intersection[L1, L2][[1]]]];
     v = Flatten[Position[L2, Intersection[L1, L2][[1]]]],
                                         (*record jump numbers*)
                                         (*else*)
    Collision = False];                  (*ready to jump again*)
   {L1, p1} = Jump[L1, p1];              (*tame kangaroo jumps*)
   {L2, p2} = Jump[L2, p2];              (*wild kangaroo jumps*)
   i = i + 1];                           (*increment jump counter*)
  Print["tame jump locations: ", L1];
  Print["wild jump locations: ", L2];
  Print["tame powers of p: ", p1];
  Print["wild powers of p: ", p2];
  Print["tame intersection location: ", u];
  Print["wild intersection location: ", v];
  If [i ≠ p,                             (*not tried all jumps*)
                                         (*then*)
   (diff = p1[[u]][[1]] - p2[[v]][[1]]; (*difference in powers of g*)
    If [(diff > 0),                      (*then*)
     (Print["index of X for generator", g ": ", diff]),
     (Print["index of X for generator", g ": ", p - 1 + diff])]),
                                         (*else*)
   Print["there is no such index"];]]
```

Figure S2. Revised Mathematica code for solution to **Exercise 14**, with changes in red (light gray).

20. $(I-Q)^{-1} = \sum_{n=0}^{\infty} Q^n$. Since $\lim_{n\to\infty} Q^n = 0$ (because the sum converges),

$$(I-Q)(I+Q+Q^2+Q^3+\cdots) = I.$$

21. Since NR is a matrix whose entries represent the probabilities that the process ends in a particular absorbing state for each nonabsorbing state, $NR = [1]$ means that there is one nonabsorbing state in the process.

22. Begin by supposing that the player's card is 7 cards ahead of the dealer's card. For the distance to go from 7 to 9, one of the following must occur:

- The dealer's current card is an 8 and the player's current card is a 10. The probability of this occurring is $(1/10)^2$.
- The dealer's current card is a 9 and the player's current card is 1. Since distance is defined by a player's card being ahead of the dealer's card, this turn is not complete until the player moves again. For the distance to be 9, the player's card must be a 10. The probability is $(1/10)\,(1/10)^2$.
- The dealer's current card is a 10 and either
  - the player's next cards are 1, 1, and 10; or
  - the player's next cards are 2 and 10.

  This probability is $\frac{1}{10}\left[\left(\frac{1}{10}\right)^3 + \left(\frac{1}{10}\right)^2\right]$.

  In total, $M_{7,9} = \frac{1}{100}\left(1+\frac{1}{10}\right)^2$.

23.

$$T = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 5 & 8 & 6 & 4 & 2 \\ 5 & 5 & 5 & 5 & 5 \\ 5 & 10 & 5 & 5 & 0 \\ 5 & 10 & 5 & 0 & 0 \\ 5 & 10 & 5 & 5 & 0 \end{pmatrix} \end{array} \times \frac{1}{25}.$$

This is not an absorbing Markov chain.

24. $P(\{\tau > N\}) \le \left(1-\frac{1}{10}\right)^{104/10} \approx 0.334$.
So the probability of success is at least $1 - 0.334 = 0.666$.

11. Appendix A: Computer Code

The code in **Figures A1** (Maple) and **A2** (Mathematica) is for use with **Exercises 13–15**.

These and the other programs in this Module are available at `http://www.comap.com/product/periodicals/supplements.html`.

```
> Index := proc(p, g, X)
    # input prime p, generator g, element X
  local s, L1, L2, p1, p2, i, k, m, n, h, f, z, y;
    # identify local variables
  s := floor(sqrt(p));
    # smallest integer <= sqrt of order of G
  L1 := [g]; L2 := [X];
    # lists terms of sequences (a_i) and (b_i)
  p1 := [1]; p2 := [0];
    # lists powers of g for a_i and b_i
  i := 1; k := 0;
    # loop variables

  while k = 0 do
    for z in L1 do
      for y in L2 do
        if (k = 0) and (z = y) then k := 1;
          # there is a collision
        print(z, y); member(z, L1,'u'); member(y, L2,'v');
        break; end if;
      end do
    end do;
  m := mod(op(i, L1), 2·s - 2);
    # determine g(a_i)
  if m = 0 then h := 2;
    elif m ≤ s then h := m;
    else h := 2·s - m; end if;
  L1 := [op(L1), mod(op(i, L1)·g^h, p)];
  p1 := [op(p1), op(i, p1) + h];

  n := mod(op(i, L2), 2·s - 2);
    # determine g(b_i)
  if n = 0 then f := 2;
    elif n ≤ s then f := n;
    else f := 2·s - n; end if;
  L2 := [op(L2), mod(op(i, L2)·g^f, p)];
  p2 := [op(p2), op(i, p2) + f];
  i := i + 1;
  end do;
  print(L1); print(L2); print(p1); print(p2);
  if op(u, p1) - op(v, p2) > 0 then print(op(u, p1) - op(v, p2));
  else print(p - 1 + (op(u, p1) - op(v, p2)));end if;
  end proc;
```

Figure A1. Maple code for Pollard's kangaroo method.

```
Index[p_, g_, X_](* prime, generator, element*) := Module[
  {s, Collision, L1, L2, p1, p2, i, u, v, Jump}, (*local vars*)
  s = Floor[Sqrt[p]]; (*size of hash table*)
  Collision = False;
  L1 = {g}; (*tame kangaroos positions*)
  L2 = {X}; (*wild kangaroos positions*)
  p1 = {1}; (*power of g for tame kangaroo*)
  p2 = {0}; (*power of g for wild kangaroo*)
  u = 0;    (*jump number for tame kangaroo at collision*)
  v = 0;    (*jump number for wild kangaroo at collision*)
  i = 1;    (*counter for jumps*)
  Jump[L_, plist_] := Module[{m, h},      (*local vars*)
    m = Mod[L[[i]], 2 * s - 2];           (*hash to find jump size*)
    If[m == 0,                            (*then*)
     h = 2,                               (*else*)
     If[m ≤ s,                            (*then*)
      h = m,                              (*else*)
      h = 2 * s - m]];
    {Append[L, Mod[(Last[L] * g^h), p]], (*add new location*)
     Append[plist, Last[plist] + h]}];   (*record power of g for new location*)
  While[(Collision == False && i < p),   (*no collision and not exhausted jumps*)
   If[Intersection[L1, L2] ≠ {},         (*if jump location in common*)
                                         (*then*)
    (Collision = True;                   (*note fact of collision*)
     u = Flatten[Position[L1, Intersection[L1, L2][[1]]]];
     v = Flatten[Position[L2, Intersection[L1, L2][[1]]]]),
                                         (*record jump numbers*)
                                         (*else*)
    Collision = False];                  (*ready to jump again*)
   {L1, p1} = Jump[L1, p1];              (*tame kangaroo jumps*)
   {L2, p2} = Jump[L2, p2];              (*wild kangaroo jumps*)
   i = i + 1];                           (*increment jump counter*)
  Print["tame jump locations: ", L1];
  Print["wild jump locations: ", L2];
  Print["tame powers of p: ", p1];
  Print["wild powers of p: ", p2];
  Print["tame intersection location: ", u];
  Print["wild intersection location: ", v];
  If [i ≠ p,                            (*not tried all jumps*)
                                         (*then*)
   (diff = p1[[u]][[1]] - p2[[v]][[1]]; (*difference in powers of g*)
    If [(diff > 0),                      (*then*)
     (Print["index of X for generator", g ": ", diff]),
     (Print["index of X for generator", g ": ", p - 1 + diff])]),
                                         (*else*)
   Print["there is no such index"];]]
```

Figure A2. Mathematica code for Pollard's kangaroo method.

12. Appendix B: Continuation of Proof

We continue expositing the proof from Haga and Robins [1997] of the first of their Theorems in **Section 6**.

First, we introduce some notation. Define sequences (y_i) and (x_i) to be the dealer's and player's secret cards, respectively. Let d_k be the distance between the dealer's current card y_i and the player's next-closest card appearing after the dealer's card in the deck, x_j.

To prove part (b), we assume that $d_k = 9$ and $d_{k+1} = j < 9$. There are two ways in which this can happen:

- Either the dealer's card has face value $9 - j$, with probability $1/10$; or
- the dealer's card has face value 10 and the player's card has face value $j + 1$, with probability $1/100$.

Thus, $M_{9,j} = \frac{1}{10}\left(1 + \frac{1}{10}\right)$.

For $M_{i,j}$ we will compare the **Figures B1 and B2**:

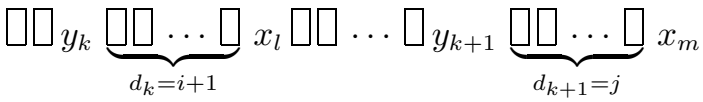


Figure B1.

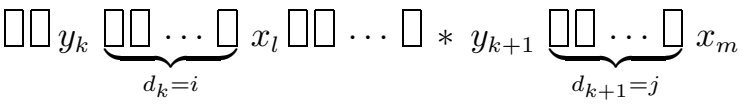


Figure B2.

Now, we claim that the difference between computing $M_{i+1,j}$ depicted in the **Figure B1** and $M_{i,j}$ depicted in **Figure B2** is the card just before y_{k+1}, which Haga and Robins call "star" and we denote by $*$.

If $*$ in **Figure B2** is vacant, then the player's cards can occupy any of the empty boxes between x_l and y_{k+1} in the same number of ways as **Figure B1**. Thus, when $*$ is vacant, $M_{i,j} = M_{i+1,j}$.

When a player's card occupies $*$, the player's sequence of cards would have followed a path described when $d_k = i + 1$ and $d_{k+1} = j$, but with one additional turn, since some x_n lands on $*$.

When $*$ is occupied, $M_{i,j} = \frac{1}{10} M_{i+1,j}$.

In total,

$$M_{i,j} = \left(1 + \frac{1}{10}\right) M_{i+1,j}, \qquad \textit{for } i \neq j \textit{ and } 0, 9 \neq i.$$

Last, we examine part (d) from the theorem. The paths in which $d_k = i + 1$ and $d_{k+1} = i$ include the cases:

- y_k has face value 1, which occurs with probability $1/10$.

- Any other path can be modified to work in the case where $d_k = d_{k+1} = i$ by forcing y_k to have face value one less. This accounts for all cases in which $y_k \neq 10$ and we have $M_{i,i} = M_{i+1,i} - \frac{1}{10}$.
- y_k has face value 10; we again examine a path traveled by the player's cards in which $d_k = i + 1$ and $d_{k+1} = i$. Say this path has $y_k = p$. Then we modify the path of the player's cards by inserting a card of face value $10 - p + 1$ as the first card in the player's path and all other cards having an unchanged face value (see example). So, this means when $y_k = 10$, $M_{i,i} = \frac{1}{10} M_{i+1,i}$. In total, $M_{i,i} = \left(1 + \frac{1}{10}\right) M_{i+1,i} - \frac{1}{10}$.

  Example for Proof: We want to show explicitly

$$M_{7,7} = \left(1 + \frac{1}{10}\right) M_{8,7} - \frac{1}{10}.$$

  First, we label the paths in which $d_k = 8$ and $d_{k+1} = 7$:

$$y_k \underbrace{\square\square\cdots\square}_{d_k=8} x_l \square\square\cdots\square\, y_{k+1} \underbrace{\square\square\cdots\square}_{d_{k+1}=7} x_{l+m}$$

  Path 1: $y_k = 1$ and $x_l = x_{l+m}$
  Path 2: $y_k = 9$ and $x_l = 8$
  Path 3: $y_k = 10$ and $x_l = 1,\ x_l + 1 = 8$
  Path 4: $y_k = 10$ and $x_l = 9$

  Now, we modify these paths in the two ways described in the preceding proof.

  (a) Decrease y_k by 1 for $y_k \neq 1$. So, we have the paths given below.
  Path 2a: $y_k = 8$ and $x_l = 8$
  Path 3a: $y_k = 9$ and $x_l = 1,\ x_{l+1} = 8$
  Path 4a: $y_k = 9$ and $x_l = 9$
  Since the values of the cards are equally distributed throughout the deck, the probabilities of paths 2a, 3a, and 4a are equal to the probabilities of paths 2, 3, and 4, respectively.

  (b) Change y_k to 10 and insert extra x_j. The paths are then
  Path 1b: $y_k = 10$ and $x_l = 10$
  Path 2b: $y_k = 10$ and $x_l = 2,\ x_{l+1} = 8$
  Path 3b: $y_k = 10$ and $x_l = 1,\ x_{l+1} = 1,\ x_{l+2} = 8$
  Path 4b: $y_k = 10$ and $x_l = 1,\ x_{l+1} = 9$
  Since we have inserted an additional card into the path, the probabilities of paths 1b, 2b, 3b, and 4b are each $\frac{1}{10}$ times the probability of paths 1, 2, 3, and 4, respectively.

  In total, we have $M_{7,7} = \left(1 + \frac{1}{10}\right) M_{8,7} - \frac{1}{10}$.

References

Diffie, W., and M. Hellman. 1976. New directions in cryptography. *IEEE Transactions on Information Theory* IT-22 (6): 644–654.

Gardner, Martin. 1978. Mathematical Games: On checker jumping, the Amazon game, weird dice, card tricks and other playful pastimes. *Scientific American*, 238 (2) (February): 19–32. 1989. Reprinted under the title: Sicherman dice, the Kruskal count and other curiosities. Chapter 19 in *Penrose Tiles to Trapdoor Ciphers... and the Return of Dr. Matrix*. New York: W.H. Freeman. 1997. Rev. ed., with addendum, 265–280. Washington, DC: Mathematical Association of America. 2005. Reproduced in *Martin Gardner's Mathematical Games*. CD-ROM. Washington, DC: Mathematical Association of America.

Grime, J. n.d. Kruskal's count. `http://www.singingbanana.com/Kruskal.pdf` .

Grinstead, C.M. 1997. *Introduction to Probability*. Providence, RI: American Mathematical Society.

Haga, Wayne, and Sinai Robins. 1997. On Kruskal's principle. *Organic Mathematics; Proceedings of the Organic Mathematics Workshop*. 20: 407–412. `http://oldweb.cecm.sfu.ca/cgi-bin/organics/doc_vault/name=paper+access_path=vault/robins` . A demonstration of the card trick—with face cards having value 1—is at `http://oldweb.cecm.sfu.ca/cgi-bin/organics/carddemo.pl` .

Kemeny, J., and J. Snell. 1960. *Finite Markov Chains*. Princeton, NJ: D. Van Nostrand Company, Inc.

_____, and G. Thompson. 1957. *Introduction to Finite Mathematics.* Englewood Cliffs, NJ: Prentice-Hall.

Klima, R. 1999. Applying the Diffie-Hellman key exchange to RSA. *The UMAP Journal* 20 (1): 21–27.

Lacey, Michael T. 2002. Cryptography, card tricks, and kangaroos. `http://people.math.gatech.edu/~lacey/talks/roos.pdf` .

Lagarias, Jeffrey C., Eric Rains, and Robert J. Vanderbei. 2001. The Kruskal count. `http://arxiv.org/abs/math/0110143` . 2009. In *The Mathematics of Preference, Choice and Order: Essays in Honor of Peter J. Fishburn*, edited by S. Brams, W.V. Gehrlein, and F.S. Roberts, 371–391. New York: Springer-Verlag.

Montenegro, Ravi. 2009. Kruskal count and kangaroo method. `http://faculty.uml.edu/rmontenegro/research/kruskal_count/index.html` . Includes links to one-deck and two-deck demonstrations of the card trick (with face cards having value 5) at `http://faculty.uml.`

edu/rmontenegro/research/kruskal_count/kruskal.html, with a demonstration of variations on the kangaroo method to solve the DSP at http://faculty.uml.edu/rmontenegro/research/kruskal_count/kangaroo.html .

_____, and Prasad Tetali. 2009. How long does it take to catch a wild kangaroo? *Proceedings of 41st ACM Symposium on Theory of Computing (STOC 2009)*, 553–559. 2010. Rev. version (v2). http://arxiv.org/abs/0812.0789 .

van Oorschot, P.C., and M.J. Wiener. 1999 Parallel collision search with cryptanalytic applications. *Journal of Cryptology* 12: 1–28. http://homes.chass.utoronto.ca/~krybakov/links_files/papers/parallel%20colision%20search.pdf .

Pollard, J.M. 1978. Monte Carlo methods for index computation $(\mathrm{mod}\ p)$. *Mathematics of Computation* 32 (No. 143, July 1978): 918–924. http://www.ams.org/journals/mcom/1978-32-143/S0025-5718-1978-0491431-9/S0025-5718-1978-0491431-9.pdf .

_____. 2000a. Kangaroos, Monopoly and discrete logarithms. *Journal of Cryptology* 13: 437–447. http://www4.ncsu.edu/~singer/437/Kangaroos.pdf .

_____. 2000b. 84.29 Kruskal's card trick. *Mathematical Gazette* 84 (No. 500, July 2000): 265–267. http://www4.ncsu.edu/~singer/437/Kruskal_card.pdf .

Roberts, F.S. 1976. *Discrete Mathematical Models*. Englewood Cliffs, NJ: Prentice-Hall.

Singh, Simon. 1999. *The Code Book: The Science of Secrecy from Ancient Egypt to Quantum Cryptography*. New York: Doubleday.

Teske, E. 2001. Square-root algorithms for the discrete logarithm problem (a survey). In *Public-Key Cryptography and Computational Number Theory*, edited by Kazimierz Alster, Jerzy Urbanowicz, and Hugh C. Williams, 283–301. New York: Walter de Gruyter.

Acknowledgments

Work on this Module was done at North Carolina State University under the direction of Dr. Michael Singer with assistance from Dr. Min Kang and Dr. Ernie Stitzinger, and I gratefully acknowledge their help, with a special thank you to Dr. Min Kang, whose knowledge of probability and patience allowed me to take the project further. I also thank the referee for the very helpful comments, references, and Mathematica codes.

The author was partially supported by NSF Grants CCR-0634123 and CCF-1017217.

About the Author

Lindsey Bosko-Dunbar grew up in southeastern Wisconsin and graduated with a degree in mathematics education from Elizabethtown College in Pennsylvania. From there, she went on to study Lie algebras at North Carolina State University, earning her doctorate in 2011. Lindsey and her husband now reside in northern West Virginia working as mathematics professors at West Liberty University. In her free time, Lindsey enjoys playing soccer and cooking.

UMAP

Modules in Undergraduate Mathematics and Its Applications

Published in cooperation with

The Society for Industrial and Applied Mathematics,

The Mathematical Association of America,

The National Council of Teachers of Mathematics,

The American Mathematical Association of Two-Year Colleges,

The Institute for Operations Research and the Management Sciences, and

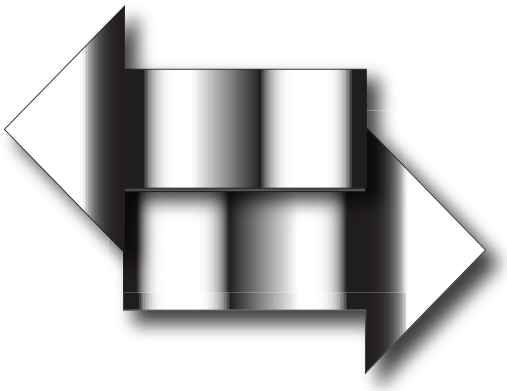
The American Statistical Association.

COMAP

Module 805

Forward and Backward Analyses of Quadratic Equations

Yves Nievergelt



Numerical Analysis

COMAP, Inc., Suite 3B, 175 Middlesex Tpk., Bedford, MA 01730 (781) 862-7878

INTERMODULAR DESCRIPTION SHEET:	UMAP Unit 805
TITLE:	Forward and Backward Analyses of Quadratic Equations
AUTHOR:	Yves Nievergelt Dept. of Mathematics Eastern Washington University 216 Kingston Hall Cheney, Washington 99004–2418 ynievergelt@ewu.edu
MATHEMATICAL FIELD:	Numerical analysis
APPLICATION FIELD:	Accounting, chemistry, finance, taxation
TARGET AUDIENCE:	Students of numerical analysis and of mathematics applied in science or finance
ABSTRACT:	This Module demonstrates the methods of forward and backward analysis to prove the degree of accuracy of the computed real solutions of real quadratic equations. This Module focuses on the methods of proof but also features specific results. Applications include the computation of acidity in chemistry, yield-to-maturity (also called the internal rate of return) of one-year U.S. Treasury bills, and the Excess Related Person Indebtedness according to the U.S. corporate tax code.
PREREQUISITES:	None
RELATED UNITS:	Unit 806: *Forward and Backward Analyses of Cubic Equations* by Yves Nievergelt. *The UMAP Journal*. To appear. Unit 807: *Applications of Cubic Equations* by Yves Nievergelt. *The UMAP Journal* 32 (1) (2011): 5–37.

The UMAP Journal 32 (3): 237–266.

COMAP, Inc., Suite 3B, 175 Middlesex Tpke., Bedford, MA 01730
(800) 77-COMAP = (800) 772-6627, or (781) 862-7878; http://www.comap.com

Forward and Backward Analyses of Quadratic Equations

Yves Nievergelt

Dept. of Mathematics
Eastern Washington University
216 Kingston Hall
Cheney, Washington 99004–2418
ynievergelt@ewu.edu

Table of Contents

Modules and Monographs in Undergraduate Mathematics and its Applications (UMAP) Project

Paul J. Campbell — Editor
Solomon Garfunkel — Executive Director, COMAP

The goal of UMAP is to develop, through a community of users and developers, a system of instructional modules in undergraduate mathematics and its applications, to be used to supplement existing courses and from which complete courses may eventually be built.

The Project was guided by a National Advisory Board of mathematicians, scientists, and educators. UMAP was funded by a grant from the National Science Foundation and now is supported by the Consortium for Mathematics and Its Applications (COMAP), Inc., a nonprofit corporation engaged in research and development in mathematics education.

1. Introduction

This Module explains *backward analysis*, a method to *prove* mathematically that specific algorithms deliver accurate results despite all intermediate rounding errors. Called "a landmark in the history of error analysis" [Wilkinson 1971, § 4, 552], backward analysis was developed and publicized for linear systems and eigenvalues by James Wallace Givens, Jr. [1954] and James Hardy Wilkinson [1960; 1961; 1963; 1971; 1972; 1988]. Backward analysis now shines in proving the accuracy of solutions of partial differential equations [Isaacson and Keller 1994, § 5.1].

Inspired by an example from William Kahan [1972, 1215–1220], we demonstrate the method in a simpler context: We apply backward analysis to the solutions of quadratic equations, requiring only algebra.

To find the zeros of a quadratic polynomial f defined by

$$f(x) = a\,x^2 + b\,x + c, \tag{1}$$

calculate (a multiple of) the discriminant d and the zeros $r_\pm$ of f by the following implementation of the quadratic formula:

$$h = b/2, \tag{2}$$

$$d = h^2 - ac, \tag{3}$$

$$r_\pm = \frac{\pm\sqrt{d} - h}{a}. \tag{4}$$

We assume $a \neq 0 \neq c$. Rationalizing the numerator gives another formula,

$$r_\pm = \frac{-c}{\pm\sqrt{d} + h}. \tag{5}$$

We use backward analysis to prove that despite rounding errors, **(4)** is always accurate for *one zero* whereas **(5)** is always accurate for *the other zero*.

We present applications from chemistry and finance; for the latter, we find multimillion-dollar discrepancies in results from various financial computing hardware and software.

2. Rounding Errors

Computers and calculators store a number x in a format called *floating-point*, like scientific notation, with a sign $(\pm)$ and an integer exponent $\mathrm{E}(x)$ but a fixed number m of significant digits $x_0, \ldots, x_{m-1}$:

$$\begin{aligned} x &= \pm R^{\mathrm{E}(x)} \times x_0.x_1 \ldots x_{m-1} \\ &= \pm R^{\mathrm{E}(x)} \times \left(x_0 + x_1 \cdot R^{-1} + \cdots + x_{m-1} \cdot R^{1-m}\right) \end{aligned} \tag{6}$$

relative to a radix R. For instance, $R = 10$ for a decimal calculator, while $R = 2$ for a binary computer. Because of the fixed number m of digits, computers and calculators must round all transcendental numbers (such as e and π), all irrational numbers (such as $\sqrt{2}$), and all rational numbers with a period longer than m, thereby causing rounding errors and propagating them through subsequent computations. To avoid tracking every error through the entire computation, the *standard model* provides upper bounds for rounding errors.

2.1 The Standard Model of Floating-Point Arithmetic

To fit results within the fixed number m of digits, computers and calculators approximate each arithmetic operation $\circ \in \{+, -, *, /\}$ by a rounded operation $\circledcirc \in \{\oplus, \ominus, \circledast, \oslash\}$ so that for all floating-point numbers p and q, there exists a *relative rounding error* $\delta_{\circledcirc,p,q}$ such that

$$p \circledcirc q = (p \circ q) \cdot (1 + \delta_{\circledcirc,p,q}). \tag{7}$$

The subscripts p and q are usually omitted. Thus, if $p \circ q \neq 0$, then **(7)** is equivalent to

$$\frac{(p \circledcirc q) - (p \circ q)}{p \circ q} = \delta_{\circledcirc}.$$

The square root is also approximated by a rounded operation `sqrt` with rounding error $\delta_{\texttt{sqrt}}$ such that

$$\texttt{sqrt}(p) = \sqrt{p} \cdot (1 + \delta_{\texttt{sqrt}}).$$

For $p \circledcirc q$ and $\texttt{sqrt}(p)$ to fit in the floating-point format **(6)**, the machine rounds each nonzero result r to the nearest m-digit number $\tilde{r}$, so that r and $\tilde{r}$ agree to m digits, which means that

the rounding error does not exceed one-half of one unit in the mth digit:

In the floating-point format **(6)**, we have $r_0 \geq 1$ because $r \neq 0$ and

$$\begin{aligned}
r &= \pm R^{\mathrm{E}(r)} \times r_0 . r_1 \ldots r_{m-2} \quad r_{m-1} \quad r_m \ldots, \\
\tilde{r} &= \pm R^{\mathrm{E}(\tilde{r})} \times \tilde{r}_0 . \tilde{r}_1 \ldots \tilde{r}_{m-2} \quad \tilde{r}_{m-1}, \\
|\tilde{r} - r| &\leq R^{\mathrm{E}(r)} \times 0 . 0 \ldots 0 \quad \tfrac{1}{2} \\
&= R^{\mathrm{E}(r)} \times \quad \tfrac{R^{-(m-1)}}{2} \\
&= R^{\mathrm{E}(r)+1} \times \tfrac{1}{2} \cdot R^{-m} \\
&< |r| \times \tfrac{1}{2} \cdot R^{1-m}.
\end{aligned}$$

Therefore, the relative rounding error $\delta = (\tilde{r} - r)/r$ satisfies the inequality

$$|\delta| = \frac{|\tilde{r} - r|}{|r|} < \mu = \frac{R^{1-m}}{2}. \qquad \textbf{(8)}$$

The number μ is the *unit roundoff*. The inequality **(8)** is strict, because if $|r| = R^{\mathrm{E}(r)}$, then no rounding occurs; rounding occurs only if $|r| > R^{\mathrm{E}(r)}$. For the same reason [Kincaid and Cheney 2002, 52, exercise 6],

$$|\delta| = \frac{|\tilde{r} - r|}{|r|} \leq \Delta = \frac{R^{1-m}/2}{1 + R^{1-m}/2} = \frac{\mu}{1+\mu}. \qquad \textbf{(9)}$$

Exercise 1 shows that if

$$|\delta| = \frac{|\tilde{r} - r|}{|r|} \leq \frac{R^{n-m}}{2},$$

then $\tilde{r}$ agrees with r to at least $m - n$ significant digits.

Example 1 For a two-digit decimal calculator, we have $R = 10$, $m = 2$, $\mu = 10^{1-2}/2 = 0.05$, and $\Delta = 0.\overline{047619}$. With also $a = .48$, $h = -.47$, and $c = .45$, **Table 1** gives examples of two-digit decimal floating-point arithmetic. A single rounding error in each of the multiplications drowns the fact that **(1)** has two distinct zeros.

Table 1.
Examples of two-digit decimal floating-point arithmetic.

Exact arithmetic			Two-digit floating-point decimal arithmetic				
$h * h$	$=$	$.2209$	$h \circledast h$	$=$	$.22$	$=$	$.2209 \cdot (1 - .00407\ldots)$
$a * c$	$=$	$.216$	$a \circledast c$	$=$	$.22$	$=$	$.216 \cdot (1 + .0\overline{185})$
$(h * h) - (a * c)$	$=$	$.0049$	$(h \circledast h) \ominus (a \circledast c)$	$=$	$.00$	$\neq$	$.0049 \cdot (1 + \delta)$

2.2 Simplifications of Compounding Factors

Error analyses and proofs of accuracy make extensive use of inequalities to simplify lengthy formulae and estimate the effect of compounding rounding errors. Since this Module focuses not on the results but on the methods of proof, this subsection presents inequalities relevant to quadratic equations. **Lemma 1** allows for the replacement of a product of different factors $(1 + \delta_k) \cdots (1 + \delta_N)$ by a power of a single factor $(1 + \delta)^N$.

Lemma 1 *For all $\delta_1, \ldots, \delta_N$, and Γ with $|\delta_1|, \ldots, |\delta_N| \leq \Gamma \leq 1$, there exists a δ such that $|\delta| \leq \Gamma$ and*

$$(1 + \delta_1) \cdots (1 + \delta_N) \quad = \quad (1 + \delta)^N. \qquad \textbf{(10)}$$

Proof: By hypothesis, $1+\delta_1 \geq 0, \ldots, 1+\delta_N \geq 0$. Hence solving **(10)** for δ gives

$$\delta = \sqrt[N]{(1+\delta_1)\cdots(1+\delta_N)} - 1. \tag{11}$$

Also by hypothesis,

$$1-\Gamma \leq 1+\delta_1, \ldots, 1+\delta_N \leq 1+\Gamma,$$

whence $-\Gamma \leq \delta \leq \Gamma$ by formula (11). The existence of the Nth root results from the Intermediate Value Theorem applied to the function f with

$$f(\delta) = (1+\delta)^N$$

and the inequalities

$$f(-\Gamma) = (1-\Gamma)^N \leq (1+\delta_1)\cdots(1+\delta_N) \leq (1+\Gamma)^N = f(\Gamma). \qquad \square$$

Similarly, **Lemma 2** provides a means to factor out a sum of rounding factors $1+\delta_1$ and $1+\delta_2$.

Lemma 2 *If p_1 and p_2 have the* same *sign and $|\delta_1|, |\delta_2| \leq \Gamma \leq 1$, then there is a δ such that $|\delta| \leq \Gamma$ and*

$$[p_1 \cdot (1+\delta_1)] + [p_2 \cdot (1+\delta_2)] = (p_1+p_2)\cdot(1+\delta). \tag{12}$$

If also $\delta_1 = 0$ and $|p_2| \leq |p_1|$, then $|\delta| \leq \Gamma/2$.

Proof: If $p_1 = 0 = p_2$, then both sides of **(12)** vanish, so that **(12)** holds with $\delta = 0$. If $p_1 \neq 0$ or $p_2 \neq 0$, then solving **(12)** for $1+\delta$ by dividing both sides by $p_1+p_2 \neq 0$ gives

$$1+\delta = \frac{p_1\cdot(1+\delta_1) + p_2\cdot(1+\delta_2)}{p_1+p_2} = 1 + \frac{p_1\cdot\delta_1 + p_2\cdot\delta_2}{p_1+p_2}. \tag{13}$$

Formula **(13)** shows that δ is a weighted average of δ_1 and δ_2, whence δ lies between δ_1 and δ_2.

If also $\delta_1 = 0$ and $|p_2| \leq |p_1|$, then **(13)** becomes

$$1+\delta = 1 + \delta_2 \cdot \frac{p_2}{p_1+p_2}, \tag{14}$$

where $|p_2/(p_1+p_2)| \leq 1/2$ by the hypothesis that $|p_2| \leq |p_1|$. $\square$

Lemma 3 establishes an upper bound for the product and quotient of rounding factors $1+\delta_k$.

Lemma 3 *For all $\delta_0, \ldots, \delta_N$ and Γ, if $|\delta_0|, \ldots, |\delta_N| \leq \Gamma \leq 1$, then*

$$\begin{aligned} |(1+\delta_1)\cdots(1+\delta_N) - 1| &\leq (1+\Gamma)^N - 1, & \textbf{(15)} \\ \left|\frac{1+\delta_0}{(1+\delta_1)\cdots(1+\delta_N)} - 1\right| &\leq \frac{1+\Gamma}{(1-\Gamma)^N} - 1. & \textbf{(16)} \end{aligned}$$

Proof: From the hypotheses that $|\delta_0|, \ldots, |\delta_N| \leq \Gamma \leq 1$ follow the inequalities

$$0 \leq (1-\Gamma)^N \leq (1+\delta_1)\cdots(1+\delta_N) \leq (1+\Gamma)^N, \quad \textbf{(17)}$$

$$0 \leq \frac{1-\Gamma}{(1+\Gamma)^N} \leq \frac{1+\delta_0}{(1+\delta_1)\cdots(1+\delta_N)} \leq \frac{1+\Gamma}{(1-\Gamma)^N}. \quad \textbf{(18)}$$

Applying the identity

$$A^N - B^N = (A-B)\cdot(A^{N-1} + A^{N-2}B + \cdots + AB^{N-2} + B^{N-1})$$

first to $A = 1$ with $B = 1-\Gamma$ and second to $A = 1+\Gamma$ with $B = 1$ gives the inequalities

$$0 \leq 1 - (1-\Gamma)^N = \Gamma \cdot \sum_{k=0}^{N-1}(1-\Gamma)^k \leq \Gamma \cdot \sum_{k=0}^{N-1}(1+\Gamma)^k = (1+\Gamma)^N - 1. \quad \textbf{(19)}$$

Inequalities **(17)** and **(19)** yield inequality **(15)**. Hence from $(1-\Gamma^2) \leq 1$ follow $1-\Gamma \leq 1/(1+\Gamma)$ and

$$0 \leq 1 - \frac{1}{(1+\Gamma)^N} \leq 1 - (1-\Gamma)^N \leq (1+\Gamma)^N - 1 \leq \frac{1}{(1-\Gamma)^N} - 1 \quad \textbf{(20)}$$

by inequality **(19)**. Consequently, we have

$$\begin{aligned} 0 \leq 1 - \frac{1-\Gamma}{(1+\Gamma)^N} &= \frac{\Gamma}{(1+\Gamma)^N} + 1 - \frac{1}{(1+\Gamma)^N} \\ &\leq \frac{\Gamma}{(1+\Gamma)^N} + \frac{1}{(1-\Gamma)^N} - 1 \leq \frac{1+\Gamma}{(1-\Gamma)^N} - 1. \end{aligned} \quad \textbf{(21)}$$

Inequalities **(18)** and **(21)** yield inequality **(16)**. □

Lemmas 4, 5, 6, 7 take advantage of formula **(9)** to simplify quotients and products of rounding factors.

Lemma 4 *If $|\delta_1|, \ldots, |\delta_N|, |\delta'_1|, \ldots, |\delta'_M| \leq \Delta$, then there exists δ such that $|\delta| \leq \mu$ and*

$$\frac{(1+\delta_1)\cdots(1+\delta_N)}{(1+\delta'_1)\cdots(1+\delta'_M)} = (1+\delta)^{N+M}. \quad \textbf{(22)}$$

Proof: By **Lemma 3**, it suffices to prove **Lemma 5** with $\delta_1 = \cdots = \delta_N = \Delta$ and $\delta'_1 = \cdots = \delta'_M = -\Delta$. If $M \le N$, then

$$\frac{(1+\Delta)^N}{(1-\Delta)^M} = \frac{[1+\mu/(1+\mu)]^N}{[1-\mu/(1+\mu)]^M} = \frac{(1+2\cdot\mu)^N}{(1+\mu)^{N-M}}$$

$$\le \frac{[(1+\mu)^2]^N}{(1+\mu)^{N-M}} = (1+\mu)^{N+M}. \tag{23}$$

If $M \ge N$, then multiplying the denominator and numerator by $(1+\mu)^M$ gives

$$\frac{[1+\mu/(1+\mu)]^N}{[1-\mu/(1+\mu)]^M} = \frac{(1+\mu)^{M-N}\cdot(1+2\cdot\mu)^N}{1}$$

$$\le \frac{(1+\mu)^{M-N}\cdot[(1+\mu)^2]^N}{1} = (1+\mu)^{N+M}. \tag{24}$$

Thus,

$$\delta = \sqrt[N+M]{(1+\delta_1)\cdots(1+\delta_N)/(1+\delta'_1)\cdots(1+\delta'_M)} - 1$$

solves equation **(22)**, with $|\delta| \le \mu$ by inequality **(23)** or **(24)**. □

Lemma 5 *If $|\delta_0|, \ldots, |\delta_N| \le \Delta$, then there exists δ such that $|\delta| \le \mu$ and*

$$\frac{1+\delta_0}{(1+\delta_1)\cdots(1+\delta_N)} = (1+\delta)^{N-1}\cdot(1+2\cdot\delta). \tag{25}$$

Proof: By **Lemma 3**, it suffices to prove **Lemma 5** with $\delta_0 = \Delta$ and with $\delta_1 = \cdots = \delta_N = -\Delta$, for which

$$\frac{1+\Delta}{(1-\Delta)^N} = \frac{[1+\mu/(1+\mu)]}{[1-\mu/(1+\mu)]^N} = (1+\mu)^N \cdot \frac{1+\mu/(1+\mu)}{1^N}$$

$$= (1+\mu)^{N-1}\cdot(1+2\cdot\mu). \tag{26}$$

□

For at most four rounding factors, **Lemmas 6** and **7** simplify the upper bounds from **Lemmas 1** and **5**.

Lemma 6 *If $|\delta_1|, \ldots, |\delta_N| \le \Delta$ and $N \le 4$, then there exists a δ such that $|\delta| \le \mu$ and*

$$(1+\delta_1)\cdots(1+\delta_N) = 1 + N\cdot\delta. \tag{27}$$

Proof: By **Lemma 3**, it suffices to prove **Lemma 6** with $\delta_1 = \cdots = \delta_N = \Delta$. Substituting $\Delta = \mu/(1+\mu)$ from formula **(9)** in $(1+\Delta)^N - 1 \le N \cdot \mu$, multiplying out the left-hand side, clearing denominators, and collecting similar powers of μ as in the proof of **Lemma 5** proves **Lemma 6**. □

Lemma 7 *If* $|\delta_0|, \ldots, |\delta_N| \le \Delta$ *and* $N \le 3$ *with* $\mu \le 1/6$, *then there exists* δ *with* $|\delta| \le \mu$ *and*

$$\frac{1+\delta_0}{(1+\delta_1)\cdots(1+\delta_N)} = 1 + (N+2)\cdot\delta. \tag{28}$$

Proof: Equation **(26)** in **Lemma 5** gives

$$\frac{1+\Delta}{1-\Delta} = 1 + 2\cdot\mu$$

for $N = 1$, while for $N \in \{2, 3\}$,

$$\frac{1+\Delta}{(1-\Delta)^2} = (1+\mu)\cdot(1+2\cdot\mu) = 1 + 3\mu + 2\mu^2 \le 1 + 4\cdot\mu, \tag{29}$$

$$\frac{1+\Delta}{(1-\Delta)^3} = (1+\mu)^2\cdot(1+2\cdot\mu) = 1 + 4\mu + 5\mu^2 + 2\mu^3 \le 1 + 5\cdot\mu, \tag{30}$$

provided that $5\cdot\mu + 2\cdot\mu^2 \le 1$, which holds for every $\mu \le 1/6$, in particular, for $R \ge 2$ and $m \ge 3$. □

2.3 Exercises on Rounding Errors

1. Prove that if

$$|\delta| = \frac{|\tilde{r} - r|}{|r|} \le \frac{R^{n-m}}{2},$$

then $\tilde{r}$ agrees with r to at least $m - n$ significant digits.

2. Prove inequality **(9)**.

3. Prove that if $r \ne 0 \ne \tilde{r}$, then

$$\frac{|(1/\tilde{r}) - (1/r)|}{|1/\tilde{r}|} = \frac{|\tilde{r} - r|}{|r|}.$$

4. Prove that if $r \ne 0 \ne c$, then

$$\frac{|(c\cdot\tilde{r}) - (c\cdot r)|}{|c\cdot r|} = \frac{|\tilde{r} - r|}{|r|}.$$

3. An Accurate Algorithm

Using one formula instead of another leads through different intermediate rounding errors [Nievergelt 2003]. Hence, *specifying exactly* the formula or algorithm in use is necessary to design a proof that it yields an accurate result. To this end, the literature demonstrates by heuristics, examples, and counterexamples—but usually without proof!— that formula **(4)** is more accurate for one zero whereas formula **(5)** is more accurate for the other zero [Kahan 1972; Thompson 1987]. Both cases lead to **Algorithm 1**, where $\texttt{sign}(h) = 1$ for $h \geq 0$ while $\texttt{sign}(h) = -1$ for $h < 0$.

Algorithm 1: Real Roots of Quadratics
For all nonzero reals a, b, c, compute the discriminant d:

$$h = b/2, \tag{31}$$

$$d = h^2 - (a \cdot c). \tag{32}$$

Thereafter, if $d \geq 0$ (and a, h, and c are all nonzero real numbers), then compute the real zeros r_1 and r_2:

$$s = \sqrt{d}, \tag{33}$$

$$t = -\left[h + \texttt{sign}(h)\, s\right], \tag{34}$$

$$r_1 = c/t, \tag{35}$$

$$r_2 = t/a. \tag{36}$$

Special cases handle the situations where $a = 0$, $b = 0$, $c = 0$, or $d < 0$. □

Remark If $a > 0$, then $r_- \leq r_+$; if also $ac < 0 < a$, then $d = h^2 - ac > h^2$, whence $r_- < 0 < r_+$. And if $a > 0$ but $ac > 0$, then $d = h^2 - ac < h^2$, whence r_- and r_+ have the same sign as h.

If $d > 0$ and $b < 0 < a$, then $r_1 = r_- < r_+ = r_2$, but if $0 < a, b$, then $r_2 = r_- < r_+ = r_1$.

However, calculators and computers can introduce rounding errors at each step, especially from taking a square root, which need not be a rational number, so that its expansion with any radix R must be rounded for the machine to continue with the algorithm.

One exception is the `sign` function, which does not cause any rounding error. For instance, with $\texttt{S} \geq 0$, the Fortran command `SIGN(S,H)` [American Standard Association 1966, 23] and the `C` command `copysign(S,H)` [American National Standards Institute 1999, 235] merely impart the sign of `H` onto `S` without necessarily performing any multiplication.

To simplify the exposition, we assume either that b is even or that the machine is binary, so that $b \oslash 2 = b/2$ exactly, because a binary division by 2 in floating-point **(6)** amounts to decreasing the exponent by 1, just as a

division by 10 also amounts to decreasing the exponent by 1 on a decimal calculator with scientific notation. The general case with

$$b \oslash 2 = (b/2) \cdot (1 + \delta_{\oslash})$$

merely carries along the additional multiplicative factor $1 + \delta_{\oslash}$.

Thus, with b even or $R = 2$, a computer or calculator starts **Algorithm 1** by computing the discriminant:

$$\begin{aligned} h &= b/2, \\ h \circledast h &= h^2 \cdot (1 + \delta_{\circledast}), \\ a \odot c &= (a \cdot c) \cdot (1 + \delta_{\odot}), \\ \tilde{d} &= (h \circledast h) \ominus (a \odot c) \\ &= \left[h^2 \cdot (1 + \delta_{\circledast}) - (a \cdot c) \cdot (1 + \delta_{\odot})\right] \cdot (1 + \delta_{\ominus}). \end{aligned} \tag{37}$$

Henceforth, simplifications apply to (the one-half of all) quadratic equations with $a \cdot c \leq 0$, thanks to **Lemma 2**.

4. Forward Analyses, for $a \cdot c \leq 0$

Forward analysis derives upper bounds for the rounding errors as they occur chronologically during the computations. For quadratic equations with $a \cdot c \leq 0$, forward analysis alone, without backward analysis, suffices to guarantee the accuracy of the computed discriminant and roots.

4.1 The Computed Discriminant

If $a \cdot c \leq 0$, then $p_1 = h^2 \geq 0$ and $p_2 = -(a \cdot c) \geq 0$ have the same sign, whence there are δ, $\delta_{\tilde{d}}$, and $\delta'_{\tilde{d}}$ with $|\delta|, |\delta_{\tilde{d}}| \leq \Delta$ by **Lemmas 2** and **1**, and $|\delta'_{\tilde{d}}| \leq \mu$ by **Lemma 6**, so that formula **(37)** becomes

$$\begin{aligned} \tilde{d} &= \left[h^2 - (a \cdot c)\right] \cdot (1 + \delta) \cdot (1 + \delta_{\ominus}) \\ &= \left[h^2 - (a \cdot c)\right] \cdot (1 + \delta_{\tilde{d}})^2 \qquad \textbf{(38)} \\ &= \left[h^2 - (a \cdot c)\right] \cdot (1 + 2 \cdot \delta'_{\tilde{d}}). \qquad \textbf{(39)} \end{aligned}$$

Thus, if $a\,c \leq 0$, then by formula **(38)** the computed discriminant $\tilde{d}$ is a perturbation by $(1 + \delta_{\tilde{d}})^2 > 0$ of the exact discriminant $d = h^2 - ac$. In particular, if $d > 0$, then also $\tilde{d} > 0$; similarly, if $d = 0$, then also $\tilde{d} = 0$. In other words, if $ac \leq 0$, then, despite the rounding errors, computations do not change the number of zeros, in contrast to **Example 1**, where $ac > 0$. Moreover, by formula **(39)**, we have $|2 \cdot \delta'_{\tilde{d}}| \leq 2 \cdot R^{1-m}/2 \leq R^{2-m}/2$ for every $R \geq 2$, which guarantees at least $m - 2$ correct significant digits in the computed discriminant $\tilde{d}$.

4.2 The Computed Zeros

Computing hardware approximates the square root of the discriminant $\tilde{s} = \texttt{sqrt}(\tilde{d})$ with bounds by **(38)**:

$$\begin{aligned}\tilde{s} &= \texttt{sqrt}(\tilde{d}) = (1+\delta_{\texttt{sqrt}})\sqrt{\tilde{d}} = (1+\delta_{\texttt{sqrt}})(1+\delta_{\tilde{d}})\sqrt{h^2-ac} \\ &= (1+\delta_{\tilde{s}})^2\sqrt{h^2-a\,c},\end{aligned}$$

for some $\delta_{\tilde{s}}$ such that $|\delta_{\tilde{s}}| \le \Delta$ by **Lemma 2**. The machine then continues **Algorithm 1** with

$$\begin{aligned}\tilde{t} &= -[h \oplus \texttt{sign}(h)\,\tilde{s}] = -[h + \texttt{sign}(h)\,\tilde{s}]\,(1+\delta_{\oplus}) \\ &= -\left[h + \texttt{sign}(h)\sqrt{h^2-a\,c}\right](1+\delta_{\tilde{t}})^2\,(1+\delta_{\oplus}) = t\cdot(1+\delta_{\tilde{t}})^2\,(1+\delta_{\oplus}),\end{aligned}$$

for some $\delta_{\tilde{t}}$ such that $|\delta_{\tilde{t}}| \le \Delta$ by **Lemma 2** with $p_1 = \texttt{sign}(h)\sqrt{h^2-ac}$ and $\delta_1 = (1+\delta_{\tilde{s}})^2 - 1$, and $p_2 = h$ with $\delta_2 = 0$. Hence, the machine completes **Algorithm 1** with

$$\begin{aligned}\tilde{\tilde{r}}_1 = c \oslash \tilde{t} = (1+\delta_{\oslash})\,c/\tilde{t} &= (1+\delta_{\oslash})/[(1+\delta_{\tilde{t}})^2(1+\delta_{\oplus})]\cdot c/t && \textbf{(40)}\\ &= (1+\delta_{\tilde{\tilde{r}}_1})^4\cdot r_1 && \textbf{(41)}\\ &= (1+5\cdot\delta'_{\tilde{\tilde{r}}_1})\cdot r_1, && \textbf{(42)}\\ \tilde{\tilde{r}}_2 = \tilde{t} \oslash a = (1+\delta_{\oslash})\,\tilde{t}/a &= (1+\delta_{\tilde{t}})^2(1+\delta_{\oplus})(1+\delta_{\oslash})\cdot t/a && \textbf{(43)}\\ &= (1+\delta_{\tilde{\tilde{r}}_2})^4\cdot r_2 && \textbf{(44)}\\ &= (1+4\cdot\delta'_{\tilde{\tilde{r}}_2})\cdot r_2, && \textbf{(45)}\end{aligned}$$

with $|\delta_{\tilde{\tilde{r}}_2}| \le \Delta$ by **Lemma 1**, $|\delta_{\tilde{\tilde{r}}_1}| \le R^{1-m}/2$ by **Lemma 5**, and finally $|\delta'_{\tilde{\tilde{r}}_1}|, |\delta'_{\tilde{\tilde{r}}_2}| \le R^{1-m}/2$ by **Lemma 6**. Thus,

$$\tilde{\tilde{r}}\cdot(1-5\cdot\mu) \le \frac{\tilde{\tilde{r}}}{(1-\mu)^4} \le r \le \frac{\tilde{\tilde{r}}}{(1+\mu)^4} \le \tilde{\tilde{r}}\cdot(1+5\cdot\mu). \qquad \textbf{(46)}$$

□

Remark These results show that the computed zeros $\tilde{\tilde{r}}_1$ and $\tilde{\tilde{r}}_2$ are perturbations by factors $(1+\delta_{\tilde{\tilde{r}}})^4$ of the exact zeros r_1 and r_2 of the initial quadratic equation with coefficients a, b, and c.

Example 2 For a decimal calculator, we have $4 \le 5 = 10/2 = R/2$, whence $|4\cdot\delta'_{\tilde{\tilde{r}}_2}| \le 4\cdot R^{1-m}/2 \le R^{2-m}/2$, guaranteeing $m-2$ digits. Thus, a 10-digit decimal calculator delivers at least 8 correct decimal digits.

For a binary computer, $4 = 2^2 = R^2$, whence $|4 \cdot \delta'_{\tilde{r}_2}| \le 4 \cdot R^{1-m}/2 = R^{3-m}/2$, guaranteeing $m - 3$ digits. For example, a single-precision 24-digit binary computer delivers at least 21 correct binary digits.

4.3 Computation of Acidity (pH)

In chemistry, the computation of acidity (pH) leads to quadratic equations with $ac < 0$, as in **Example 3**.

Example 3 For a solution of 0.00100 mole [M] per liter [L] of sodium chloroacetate [$C_2H_2ClO_2Na$] and 0.0100 [M/L] of chloroacetic acid [$ClCH_2CO_2H$], whose ionization constant is $K = 1.40 \cdot 10^{-3}$, Thompson [1987] shows how to compute the hydrogen concentration (pH) as the positive root of

$$x^2 + (0.0100 + 0.00140) \cdot x - (1.40 \cdot 10^{-3}) \cdot (0.00100) = 0. \quad \textbf{(47)}$$

Table 2 shows the computations with $a = 1$, $c = -1.4 \times 10^{-6}$, and $h = 0.0100/2 + 0.00140/2 = 5.7 \times 10^{-3}$. With $R = 10$, $m = 2$, and $\mu = 10^{1-2}/2 = 0.05$ from **Example 1**, **Table 2** and inequalities **(46)** *prove* that

$$1.0695 \times 10^{-4} \le \frac{1.3 \times 10^{-4}}{(1 + .05)^4} \le r_1 \le \frac{1.3 \times 10^{-4}}{(1 - .05)^4} \le 1.6 \times 10^{-4}. \quad \textbf{(48)}$$

Thus, although $r_1 = r_+$, computations of $\tilde{\tilde{r}}_1$ with two digits yield one accurate significant decimal digit, and hence two accurate digits for the acidity $\mathrm{pH} = -\log(\tilde{\tilde{r}}_1)$, whereas the computation of $-\log(\tilde{\tilde{r}}_+)$ fails.

Table 2.
Computations of acidity with a fictitious two-digit decimal calculator.

	Exact arithmetic			Two-digit decimal floating-point		
			Algorithm 1			
	$h * h$	$=$	3.249×10^{-5}	$h \circledast h$	$=$	3.2×10^{-5}
	$a * c$	$=$	-1.4×10^{-6}	$a \circledast c$	$=$	-1.4×10^{-6}
$d =$	$(h * h) - (a * c)$	$=$	3.389×10^{-5}	$\tilde{d} = (h \circledast h) \ominus (a \circledast c)$	$=$	3.3×10^{-5}
$s =$	$\sqrt{d}$	$=$	$5.821\ldots \times 10^{-3}$	$\tilde{s} =$ $\mathtt{sqrt}(\tilde{d})$	$=$	5.7×10^{-3}
$t =$	$-[h + \mathtt{sign}(h)\, s]$	$=$	$-1.152\ldots \times 10^{-2}$	$\tilde{t} = -[h \oplus \mathtt{sign}(h)\, \tilde{s}]$	$=$	-1.1×10^{-2}
$r_1 =$	c/t	$=$	$1.215\ldots \times 10^{-4}$	$\tilde{\tilde{r}}_1 = c \oslash \tilde{t}$	$=$	1.3×10^{-4}
$\mathrm{pH} =$	$-\log(r_1)$	$=$	$3.915\ldots$	$\widetilde{\mathrm{pH}} = -\log(\tilde{\tilde{r}}_1)$	$=$	3.9
			Quadratic formula **(4)**			
	$s - h$	$=$	$1.215\ldots \times 10^{-4}$	$\tilde{s} \ominus h$	$=$	0
$r_+ =$	$(s - h * h)/a$	$=$	$1.215\ldots \times 10^{-4}$	$\tilde{\tilde{r}}_+ = [\tilde{s} \ominus (h \circledast h)] \oslash a$	$=$	0
$\mathrm{pH} =$	$-\log(r_1)$	$=$	$3.915\ldots$	$\widetilde{\mathrm{pH}} = -\log(\tilde{\tilde{r}}_+)$	$=$	undefined

4.4 Yield of Discounted Treasury Bills

To save or invest your money for a year or less, you can buy U.S. Treasury bills for a price $\mathcal{P}$ on the date of issue and later cash them in at a *face value* $\mathcal{F}$ on the date of maturity. Thus, you get $\mathcal{F} - \mathcal{P}$ in interest. Different buyers may bid and pay different prices, and get different interest rates, which can be computed. The rate is compounded semiannually.

Example 4 The Federal Reserve Bank of New York [Trainer 1982, 18] shows how to compute the yield rate r of a U.S. Treasury bill issued on 2 January 1981 and maturing on 31 December 1981, after $D = 363$ days, in a calendar year with $Y = 365$ days, that has a face-value $\mathcal{F} = \$100$ and an average purchase price $\mathcal{P} = \$87.825$ depending on the buyer [Wall Street Journal 1980b]. As a convenience, the face value is stated as 100 [\$], so that the price coincides with a percentage of the face value: In this example, $\mathcal{P} = 87.825[\$] = 87.825\%$ of $\mathcal{F} = 100$ [\$]. Yet in reality this bill was sold only in integer multiples of $\mathcal{F} = \$10\,000$ [Wall Street Journal 1980b], for which $\mathcal{P} = \$8\,782.5 = 878\,250$ [¢]. Thus the data consist of integers: $D = 363$ and $Y = 365$ [days], with $\mathcal{F} = 1\,000\,000$ and $\mathcal{P} = 878\,250$ [¢]. Although this bill lasts for less than a year, to approximate semiannual compounding, the yield rate r is *defined* as the non-negative solution of the quadratic equation with coefficients

$$\begin{aligned}
b = D/Y &= 363/365 \\
&= 0.994\,520\,547\,945\ldots, \\
a = (b/2) - (1/4) &= [363/(2 * 365)] - (1/4) \\
&= 0.247\,260\,273\,973\ldots, \\
c = (\mathcal{P} - \mathcal{F})/\mathcal{P} &= (878\,250 - 1\,000\,000)/878\,250 \\
&= -0.138\,627\,953\,316\ldots.
\end{aligned}$$

Because the decimal expansions of the coefficients do not terminate, rounding them would already cause rounding errors. Instead, multiplying all the coefficients by $4 * \mathcal{P} * Y$ gives integer coefficients. Optionally, dividing throughout by the greatest common divisor of all three integer coefficients, in this example dividing by 25, gives smaller integers. Thus the yield rate r is also the nonnegative solution of

$$1268193\,x^2 + 5100876\,x - 711020 = 0. \qquad \textbf{(49)}$$

The Bank finds 13.49% for the yield rate r [Trainer 1982, 18]. **Table 3** lists other computed values $\tilde{\tilde{r}}$ of the rate r.

Conclusions from this example. Real life does not have solutions at the back of the book. The only basis available here to assess the accuracy of the computed rate is the foregoing proof that a 10-digit

Table 3.

Numerical computations of the internal rate of return $r = r_1$.

SYSTEM[a]	METHOD	COMPUTED ROOT $\tilde{\tilde{r}}_1$
$1268193\,x^2 + 5100876\,x - 711020 = 0$		
HP-12C[b]	Algorithm 1	0.134 869 362 2
`C` program	Algorithm 1, `single precision`	0.134 869 351 . . . ($\tilde{\tilde{r}}_1$ = `3e0a1b31` hex)
`C` program	Algorithm 1, `double precision`	0.134 869 362 221 160 749 015 780
Financial Hardware and Software		
HP-12C [c]	Yield-to-Maturity function `YTM`	`Error 8` (dates rejected)
HP-12C [d]	Yield-to-Maturity function `YTM`	0.134 899 977 6
Matlab 7[e]	Financial Toolbox `yldtbill`	0.137 482 267 751 657
Matlab 7[f]	Financial Toolbox `xirr`	0.139 442 686 656 321

[a] `C` and Matlab computations on a Power Mac G4 with OS 9.2 and gcc-937.2 compiler

[b] A financial programmable 10-digit decimal calculator

[c] Following the user's guide [HP 2005, 67]: `87.825 PV 0 PMT g D.MY 02.011981 ENTER 31.121981 f YTM`

[d] Delayed by one day to mature in 1982: `87.825 PV 0 PMT g D.MY 03.011981 ENTER 01.011982 f YTM`

[e] `yldtbill('02-Jan-1981','31-Dec-1981', 100, 87.825)`

[f] `xirr( [ -87825; 100000 ], [ '02-Jan-1981';'31-Dec-1981' ])`

calculator with **Algorithm 1** delivers at least $m - 2 = 10 - 2 = 8$ correct digits. Therefore, with an error smaller than 5×10^{-9}, **Table 3** guarantees that $0.134\,869\,357\,2 < r_1 < 0.134\,869\,367\,2$. With $R = 10$, $m = 10$, and $\mu = 10^{1-10}/2 = 10^{-9}/2$, **Table 3** and inequalities **(46)** also *prove* that $\tilde{\tilde{r}}_1$ is accurate to 9 digits:

$$0.134\,869\,361\,7 \le \frac{0.134\,869\,362\,2}{[1 + (10^{-9}/2)]^4} \le r_1 \le \frac{0.134\,869\,362\,2}{[1 - (10^{-9}/2)]^4} \le 0.134\,869\,362\,7. \quad \textbf{(50)}$$

Table 3 also guarantees that the financial hardware/software in various calculators and computers use methods different from the Federal Reserve Bank of New York's [Trainer 1982]. The smaller discrepancies between the computed yield rates, in the fifth digit for the financial calculator's Yield-To-Maturity `YTM` function, and in the third digit for the financial software's `yldtbill` function, translate into much larger discrepancies for the Treasury: \$137 769.30 and \$11 million respectively, because the total of all bills sold amounts to \$4.5 billion [Wall Street Journal 1980a]:

$$(0.134\,899\ldots - 0.134\,869\ldots) \times (4.5 \times 10^9)$$
$$= 3.06\ldots \times 10^{-5} \times (4.5 \times 10^9) = 137\,769.30[\$],$$
$$(0.137\,482\ldots - 0.134\,869\ldots) \times (4.5 \times 10^9)$$
$$= 2.61\ldots \times 10^{-3} \times (4.5 \times 10^9) = 11\,758\,074.98[\$],$$

enough to pay an accountant's yearly salary in 1981. Is your financial advisor aware of such discrepancies?

4.5 Exercises on Forward Analysis

5. For each even radix R and each number of digits $m \geq 3$, derive an upper bound for the rounding errors from **Algorithm 1** in computing the *larger* root, called the *golden ratio* [Falbo 2005], of the equation

$$G^2 - G = 1.$$

6. For each even radix R and each number of digits $m \geq 3$, derive an upper bound for the rounding errors from **Algorithm 1** in computing the *smaller* root, called the *golden ratio* [Falbo 2005], of the equation

$$g^2 + g = 1.$$

7. Derive a formula for the yield rate of Treasury bills maturing exactly one year past their issue.

8. For each even radix R and each number of digits $m \geq 10$, derive an upper bound for the rounding errors from **Algorithm 1** or from **Exercise 7** in computing the internal rate of return (yield rate) r for an investor who bought a one-year U.S. Treasury bill issued on 15 January 1960 at the price $\mathcal{P} = \$9{,}484.90$, and received a balance (face value) $\mathcal{F} = \$10{,}000$ on 15 January 1961 [Goldstein 1962, 449, table 1].

5. Backward Analyses, for $a \cdot c > 0$

If $a \cdot c > 0$, then **Lemma 2** does not apply, because $h^2 \geq 0$ and $-ac < 0$ have different signs.

5.1 The Computed Discriminant

A machine starts **Algorithm 1** by computing the discriminant with formula **(37)**, copied here for convenience:

$$\tilde{d} = (h \circledast h) \ominus (a \odot c) = \left[h^2 \cdot (1 + \delta_{\circledast}) - (a \cdot c) \cdot (1 + \delta_{\odot})\right] \cdot (1 + \delta_{\ominus}). \quad \textbf{(37)}$$

The rounding errors in the multiplications can be associated arbitrarily to any coefficient, for instance, c:

$$\tilde{c} = c \cdot \frac{1+\delta_\odot}{1+\delta_\circledast} = c \cdot (1 + 2 \cdot \delta_{\tilde{c}}), \tag{51}$$

$$\tilde{d} = \left[h^2 - a\,\tilde{c}\right]\,(1+\delta_\circledast)\,(1+\delta_\ominus) = \left[h^2 - a\,\tilde{c}\right]\,(1+\delta_{\tilde{d}})^2, \tag{52}$$

for some $\delta_{\tilde{c}}$ and $\delta_{\tilde{d}}$ such that $|\delta_{\tilde{c}}| \le \mu$ by **Lemma 5** and $|\delta_{\tilde{d}}| \le \Delta$ by **Lemma 1**.

Remark Formula **(52)** shows that the computed discriminant $\tilde{d}$ is a perturbation by $(1+\delta_{\tilde{d}})^2$ of the exact discriminant $h^2 - a\tilde{c}$ of a "perturbed" quadratic equation with coefficients a and b but $\tilde{c}$ instead of c.

Through distributivity and associativity, such simplifications associate some of the effects of intermediate rounding errors *backward* to a perturbation $\tilde{c}$ of the initial data c, whence the name *backward analysis*.

5.2 The Computed Zeros

Backward analysis continues the same as does forward analysis, but with $\tilde{c}$ instead of c and other modifications:

$$\begin{aligned}\tilde{s} &= \mathtt{sqrt}(\tilde{d}) = (1+\delta_{\mathtt{sqrt}})\,\sqrt{\tilde{d}} = (1+\delta_{\tilde{d}})(1+\delta_{\mathtt{sqrt}})\sqrt{h^2 - a\,\tilde{c}} \\ &= \sqrt{h^2 - a\,\tilde{c}}(1+\delta_{\tilde{s}})^2 = \sqrt{h^2 - a\,\tilde{c}}(1 + 2\cdot\delta'_{\tilde{s}}),\end{aligned}$$

with $|\delta_{\tilde{s}}| \le \Delta$ by **Lemma 2**, and $|\delta'_{\tilde{s}}| \le R^{1-m}/2$ by **Lemma 6**. The machine then continues **Algorithm 1** with

$$\begin{aligned}\tilde{t} &= -[h \oplus \mathtt{sign}(h)\,\tilde{s}] = -[h + \mathtt{sign}(h)\,\tilde{s}]\,(1+\delta_\oplus) \\ &= -\left[h + \mathtt{sign}(h)\,\sqrt{h^2 - a\,\tilde{c}}\right]\,(1+\delta_{\tilde{t}})\,(1+\delta_\oplus),\end{aligned}$$

for some $\delta_{\tilde{t}}$ such that $|\delta_{\tilde{t}}| \le \delta'_{\tilde{s}}$ by the second part of **Lemma 2**, for $p_1 = h$ with $\delta_1 = 0$, and $p_2 = \mathtt{sign}(h)\,\tilde{s}$ with $\delta_2 = 2\cdot\delta'_{\tilde{s}}$, because $h^2 - a\,\tilde{c} \le h^2$ for $a \cdot c \ge 0$. Hence the machine completes **Algorithm 1** with

$$\tilde{\tilde{r}}_1 = c \oslash \tilde{t} = (1+\delta_\oslash)\,c/\tilde{t} \tag{53}$$

$$= -(1+\delta_\oslash)/[(1+\delta_{\tilde{t}})(1+\delta_\oplus)]c/\left[h + \mathtt{sign}(h)\,\sqrt{h^2 - a\,\tilde{c}}\right] \tag{54}$$

$$= (1+\delta_{\tilde{r}_1})^3 \cdot \tilde{r}_1 = (1 + 4\cdot\delta_{\tilde{r}'_1})\cdot\tilde{r}_1, \tag{55}$$

$$\tilde{\tilde{r}}_2 = \tilde{t} \oslash a = (1+\delta_\oslash)\,\tilde{t}/a \tag{56}$$

$$= -(1+\delta_{\tilde{t}})(1+\delta_\oplus)(1+\delta_\oslash)\left[h + \mathtt{sign}(h)\,\sqrt{h^2 - a\,\tilde{c}}\right]/a \tag{57}$$

$$= (1+\delta_{\tilde{r}_2})^3 \cdot \tilde{r}_2 = (1 + 4\cdot\delta'_{\tilde{r}_2})\cdot\tilde{r}_2, \tag{58}$$

with $|\delta_{\tilde{r}_1}| \le \mu$ by **Lemma 5**, $|\delta_{\tilde{r}_1'}| \le \mu$ by **Lemma 7**, $|\delta_{\tilde{r}_2}| \le \Delta$ by **Lemma 1**, and $|\delta'_{\tilde{r}_2}| \le \mu$ by **Lemma 6**.

Remark Formulae **(53)–(58)** show that the computed zeros $\tilde{\tilde{r}}_1$ and $\tilde{\tilde{r}}_2$ are perturbations by factors $(1+\delta)^3$ of the exact zeros $\tilde{r}_1$ and $\tilde{r}_2$ of a perturbed quadratic equation with coefficients a, b, and $\tilde{c}$ from **(51)**:

$$\tilde{f}(x) = a\,x^2 + b\,x + \tilde{c}. \tag{59}$$

The task at hand now consists of estimating $\tilde{r} - r$ in terms of a, b, and $\tilde{c} - c$.

5.3 The Relative Rate of Change

Subtracting the quadratic equation **(1)** with $x = r$ from the perturbed equation **(59)** with $x = \tilde{r}$ gives

$$0 = 0 - 0 = f(r) - \tilde{f}(\tilde{r}) = a\,(\tilde{r} + r)\cdot(\tilde{r} - r) + b\,(\tilde{r} - r) + (\tilde{c} - c), \tag{60}$$

for either zero $r \in \{r_+, r_-\}$. Yet we also have $\tilde{r} + r = 2\,r + (\tilde{r} - r)$ and $2a\,r_\pm + b = \pm\sqrt{d}$ by **(2)** and **(4)**. Hence,

$$0 = a\,(\tilde{r}_\pm - r_\pm)^2 \pm \sqrt{d}\,(\tilde{r}_\pm - r_\pm) + (\tilde{c} - c), \tag{61}$$

which is a quadratic equation for $\tilde{r}_\pm - r_\pm$. Solving **(61)** using **(5)** with $b = \pm\sqrt{d}$ and $\tilde{c} - c$ instead of c gives

$$\frac{(\tilde{r}_\pm - r_\pm)/r_\pm}{(\tilde{c} - c)/c} = \frac{\sqrt{d} \pm h}{\sqrt{d} + \sqrt{d - a\,c\,(\tilde{c} - c)/c}}. \tag{62}$$

Alternatively, using **(4)** with $\tilde{c}$ for $\tilde{r}_\pm$ and c for $r_\pm$ on the numerator, but using **(5)** for $r_\pm$ on the denominator, also yields equation **(62)**. The left-hand side of equation **(62)** may be called the *relative* rate of change of $r_\pm$ with respect to c, because it is the ratio of the *relative* change in the zero $r_\pm$ to the *relative* change in the constant coefficient c. To avoid computing this ratio, which is more cumbersome than computing $r_\pm$, let $\delta_c = (\tilde{c} - c)/c$: With $d - |a\,c|\,|\delta_c| \ge 0$ to ensure real zeros, from **(62)** the triangle inequality and algebra give [Kahan 1972, 1219–1220]:

$$\left|\frac{\tilde{r}_\pm - r_\pm}{r_\pm}\right| \le \sqrt{|\delta_c|}\,\frac{\sqrt{|\delta_c|}\,\sqrt{d} + \sqrt{|\delta_c|}\,|h|}{\sqrt{d} + \sqrt{\tilde{d}}} \le \sqrt{|\delta_c|}\,\left(\sqrt{|\delta_c|} + \sqrt{1 + |\delta_c|}\right), \tag{63}$$

with equality if and only if $d = |a\,c|\,|\delta_c| = a\,c\,\delta_c$. Specifically,

$$|h| - \sqrt{\tilde{d}} = \sqrt{h^2} - \sqrt{h^2 - a\,c\,(1 + \delta_c)} = \frac{a\,c\,(1 + \delta_c)}{\sqrt{h^2} + \sqrt{h^2 - a\,c\,(1 + \delta_c)}} \tag{64}$$

$$= \frac{a\,c\,(1+\delta_c)}{|h| + \sqrt{\tilde{d}}}. \qquad \textbf{(65)}$$

Also, if $\delta_c \geq 0$, then $0 \leq \tilde{d} = d - a\,c\,\delta_c$, whence $0 \leq a\,c\,\delta_c \leq d$; whereas if $\delta_c \leq 0$, then $0 \leq \tilde{d} = d + |a\,c\,\delta_c|$, whence $0 \leq |a\,c\,\delta_c| \leq \tilde{d}$; thus, in either case, we have $0 \leq |a\,c\,\delta_c| \leq \max\{d, \tilde{d}\}$. Hence, the upper bound **(63)** becomes

$$\frac{\sqrt{|\delta_c|}\,\sqrt{d} + \sqrt{|\delta_c|}\,\sqrt{\tilde{d}} + \sqrt{|\delta_c|}\,\left(|h| - \sqrt{\tilde{d}}\right)}{\sqrt{d} + \sqrt{\tilde{d}}}$$

$$= \sqrt{|\delta_c|} + \frac{\sqrt{|a\,c\,\delta_c|}\,\sqrt{a\,c\,(1+\delta_c)}\,\sqrt{1+\delta_c}}{\left(\sqrt{d} + \sqrt{\tilde{d}}\right)\cdot\left(|h| + \sqrt{\tilde{d}}\right)}$$

$$\leq \sqrt{|\delta_c|} + \sqrt{1+\delta_c} \qquad \textbf{(66)}$$

from $0 \leq \tilde{d} = h^2 - a\,c\,(1+\delta_c)$ and $\sqrt{a\,c\,(1+\delta_c)} \leq |h|$. Thus, without knowledge of the zero $r_\pm$, the foregoing backward analysis gives an upper bound **(63)** on the relative discrepancy in the zero $\tilde{r}_\pm$ of the perturbed quadratic equation **(59)** in terms of the relative discrepancy δ_c in the datum $\tilde{c}$. The previous results yield

$$\tilde{\tilde{r}}_\pm = r_\pm \cdot (1 + 4\cdot\delta'_\pm)\cdot\left[1 + \sqrt{|\delta_c|}\,\left(\sqrt{|\delta_c|} + \sqrt{1+|\delta_c|}\right)\right]. \qquad \textbf{(67)}$$

Example 5 In the context of rounding errors, from formula **(51)** with $\tilde{c} = c\cdot(1 + 2\cdot\delta_{\tilde{c}})$ and $|\delta_{\tilde{c}}| \leq R^{1-m}/2$, it follows that

$$|\delta_c| = |\tilde{c} - c|/|c| \leq 2\cdot\delta_{\tilde{c}} \leq R^{1-m}.$$

If $R \geq 2$ and $m \geq 5$, then $R^{1-m} \leq 2^{1-5} = 2^{-4}$, whence $\sqrt{|\delta_c|} \leq 1/4$ and $\sqrt{1+|\delta_c|} \leq \sqrt{1+2^{-4}} < 5/4$, so that $\sqrt{|\delta_c|} + \sqrt{1+|\delta_c|} < 3/2$. Hence, $|\tilde{r} - r|/|r| \leq (3/2)\cdot\sqrt{|\delta_c|} \leq (3/2)\cdot R^{(1-m)/2}$ by inequality **(63)**. Equivalently, $\tilde{r}_\pm = r_\pm\cdot(1+\delta_{\tilde{r}_\pm})$ with

$$|\delta_{\tilde{r}_\pm}| \leq (3/2)\cdot R^{(1-m)/2} = 3\cdot R^{(1-m)/2}/2.$$

For decimal calculators, if m is even and $R = 10$, then $3 < \sqrt{10} = \sqrt{R}$, whence $3\cdot R^{(1-m)/2}/2 < R^{(1+1-m)/2}/2 = R^{(2-m)/2}/2$, and hence $\tilde{r}_\pm$ and $r_\pm$ agree to at least $(m/2) - 1$ significant decimal digits.

For binary computers, if m is odd and $R = 2$, then

$$4\cdot R^{(1-m)/2}/2 = R^2\cdot R^{(1-m)/2}/2 = R^{2+(1-m)/2}/2,$$

and hence $\tilde{r}_\pm$ and $r_\pm$ agree to at least $[(m-1)/2] - 2$ significant binary digits.

Exercises 9 and **10** lead to similar upper bounds relative to perturbations of a or b instead of c.

Example 6 (Example 1 continued) For a two-digit decimal calculator, we have $R = 10$, $m = 2$, $\mu = 10^{1-2}/2 = 0.05$, and $\Delta = 0.\overline{047619}$. For the values $a = .48$, $h = -.47$, and $c = .45$, **Table 4** demonstrates the computation of the zeros r_1 and r_2 of the quadratic polynomial **(1)** with two-digit decimal floating-point arithmetic.

Table 4.
Examples of two-digit decimal floating-point arithmetic.

Exact arithmetic	Two-digit floating-point decimal arithmetic	
$h * h = .2209$	$h \circledast h = .22$	$= .2209 \cdot (1 - .00407\ldots)$
$a * c = .216$	$a \circledast c = .22$	$= .216 \cdot (1 + .0\overline{185})$
$d = (h * h) - (a * c) = .0049$	$\tilde{d} = (h \circledast h) \ominus (a \circledast c) = .00$	$\neq .0049 \cdot (1 + \delta)$
$s = \sqrt{d} = .07$	$\tilde{s} = \texttt{sqrt}(\tilde{d}) = .00$	$= .00 \cdot (1 + 0)$
$t = -[-.47 - .7] = .54$	$\tilde{t} = -[-.47 \ominus .00] = .47$	$= .47 \cdot (1 + 0)$
$r_1 = c/t = .45/.54 = 5/6 = .8\overline{3}$	$\tilde{r}_1 = c/\tilde{t} = .45/.47 = .957\ldots$	
	$\tilde{\tilde{r}}_1 = c \oslash \tilde{t} = .45 \oslash .47 = .96$	
$r_2 = t/a = .54/.48 = 9/8 = 1.125$	$\tilde{\tilde{r}}_2 = \tilde{t} \oslash a = .47 \oslash .48 = .98$	

We have

$$\frac{\tilde{r}_1 - r_1}{r_1} = \frac{.45/.47 - .45/.54}{.45/.54} = \frac{6 \times 45 - 5 \times 47}{5 \times 47} = \frac{35}{235} = 0.148\,936\ldots \tag{68}$$

$$\begin{aligned} 0 = \tilde{d} &= (h \circledast h) - (a \circledast c) = h^2 - a \cdot c \cdot (1 + \delta_c) \\ &= .2209 - .48 \times .45 \cdot (1 + \delta_c), \end{aligned} \tag{69}$$

$$1 + \delta_c = \frac{.2209}{.48 \times .45} = 1.022\,685\ldots, \tag{70}$$

$$\sqrt{\delta_c} = \sqrt{0.022\,685\ldots} = 0.150\,616\ldots \tag{71}$$

Similarly,

$$\frac{\tilde{\tilde{r}}_1 - r_1}{r_1} = \frac{6 \times .96 - 5}{5} = .152$$

with $\sqrt{\delta_c} = 0.151 < .152 < 1.75 = \sqrt{\delta_c} \cdot \left(\sqrt{\delta_c} + \sqrt{1 + \delta_c}\right)$.

5.4 Excess Related Person Indebtedness

The U.S. corporate tax code calls for the smaller nonnegative solution of a quadratic equation [Internal Revenue Service 1990; Schmedel 1988] whose coefficients satisfy the inequalities $0 \le ac \le h^2$.

Example 7 The U.S. Internal Revenue Service (IRS) documentation [Internal Revenue Service 1990, §1.861-12T, 203] shows an example of a U.S. corporation X with total assets A and total debts D, which also controls a foreign subsidiary X' with total assets A' and total debts D'. If the allowed percentage p (80% after 1989) of the debt-to-asset ratio D/A of the U.S. corporation exceeds that of the foreign subsidiary D'/A', then the IRS code defines the "excess related person indebtedness" [Schmedel 1988] as the smaller nonnegative solution Z of the equation

$$\frac{D-Z}{A-Z}\cdot p = \frac{D'+Z}{A'}. \tag{72}$$

The IRS equation **(72)** means that the parent corporation X could adjust its debt-to-asset ratio by transferring an amount Z to its subsidiary X' to meet the allowed percentage p. Equation **(72)** is equivalent to

$$Z^2 + [d-(A+pA')]Z + (pDA' - AD') = 0. \tag{73}$$

Thus, $a = 1$, and the condition that $pD/A > D'/A'$ is equivalent to $c > 0$, so that $0 < ac$. If both debt-to-asset ratios remain sufficiently small, then also $ac \le h^2$. The IRS code gives a numerical example with $A = 2\times10^6$, $D = 10^6$, $A' = 5\times10^5$, $D' = 10^5$ [$], and $p = 0.8$, which produces the coefficients $a = 1$, $b = D' - (A + pA') = -2\,300\,000$, $c = pDA' - AD' = 2\times10^{11}$, and hence gives

$$z^2 - 2\,300\,000z + 2\times10^{11} = 0. \tag{74}$$

Thus, the inequality $ac = 2\times10^{11} < 6.6125\times10^{11} = h^2/2$ holds, from which $0 \le r_1 = r_- < r_+ = r_2$ because $b < 0 < a$, by the **Remark** in **Section 3** (p. 8). The IRS example gives the value 90 519 for the smaller solution z_1. The calculator produces results accurate to all displayed digits, as do other systems, as shown in **Table 5**.

Table 5.
Numerical computations of the excess related person indebtedness.

$z^2 - 2\,300\,000 * z + 2 * 10^{11} = 0$		
System[a]	Formulae	Computed smaller zero $\tilde{\tilde{z}}_1 \ge 0$
HP-12C	Algorithm 1	90 518.994 98
`C` program	Algorithm 1, `single precision`	90 519.000 . . . ($\tilde{\tilde{z}}_1$ = `47b0cb80` hex)
`C` program	Algorithm 1, `double precision`	90 518.994 979 145 471 006 631 851 196

[a] `C` computations on a Power Mac G4 with OS 9.2 and gcc-937.2 compiler.

Accuracy analysis for this example. From **Example 5**, we have

$$|\tilde{r} - r|/|r| \le (3/2) \cdot \sqrt{|\delta_c|} \le (3/2) \cdot R^{(1-m)/2}$$

by formula **(51)** and inequality **(63)**. Equivalently, $\tilde{r}_\pm = r_\pm \cdot (1 + \delta_{\tilde{r}_\pm})$ with $|\delta_{\tilde{r}_\pm}| \le (3/2) \cdot \sqrt{|\delta_c|}$. Consequently,

$$\tilde{\tilde{r}}_1 = (1 + 4 \cdot \delta_{\tilde{\tilde{r}}_1}) \cdot \tilde{r}_1 = (1 + 4 \cdot \delta_{\tilde{\tilde{r}}_1}) \cdot (1 + \delta_{\tilde{r}_\pm}) \cdot r_1 = (1 + \delta'_{\tilde{\tilde{r}}_1}) \cdot r_1, \tag{75}$$

with $|\delta_{\tilde{\tilde{r}}_1}| \le R^{(1-m)/2}$. For decimal calculators, if $m \ge 8$ is even and $R = 10$, then

$$\begin{aligned}
0 &\le (1 + 4 \cdot \delta_{\tilde{\tilde{r}}_1}) \cdot (1 + \delta_{\tilde{r}_\pm}) \\
&\le (1 + 2 \cdot 10^{1-m}) \cdot [1 + (3/2) \cdot 10^{(1-m)/2}] \\
&= 1 + \frac{1}{2} \cdot 10^{(1-m)/2} \cdot \left(3 + 4 \cdot 10^{(1-m)/2} + 6 \cdot 10^{1-m}\right) \\
&\le 1 + \frac{1}{2} \cdot 10^{(1-m)/2} \cdot \sqrt{10} \\
&= 1 + 10^{(2-m)/2}/2,
\end{aligned}$$

which still guarantees at least $(m/2) - 1$ correct significant decimal digits. For instance, a 10-digit calculator delivers at least 4 correct significant decimal digits.

For binary computers, if $m \ge 7$ is odd and $R = 2$, then

$$\begin{aligned}
0 &\le (1 + 4 \cdot \delta_{\tilde{\tilde{r}}_1}) \cdot (1 + \delta_{\tilde{r}_\pm}) \\
&\le (1 + 2 \cdot 2^{1-m}) \cdot [1 + (3/2) \cdot 2^{(1-m)/2}] \\
&= 1 + \frac{1}{2} \cdot 2^{(1-m)/2} \cdot \left(3 + 4 \cdot 2^{(1-m)/2} + 6 \cdot 2^{1-m}\right) \\
&\le 1 + \frac{1}{2} \cdot 2^{(1-m)/2} \cdot 4 \\
&= 1 + \frac{1}{2} \cdot 2^{2+(1-m)/2},
\end{aligned}$$

which still guarantees at least $[(m-1)/2] - 2$ correct significant binary digits, for instance, $[(53-1)/2] - 2 = 24$ with double-precision IEEE arithmetic, or at least 7 decimal digits. In particular, **Table 5** shows that the `double precision` result *proves* that the IRS's value 90 519 is correct to all five displayed digits.

5.5 Exercises on Backward Analysis

9. For a nonzero zero $r \in \{r_+, r_-\}$ of f as in equation **(1)** and the corresponding zero $\hat{r}$ of $\hat{f}$ with

$$\hat{f}(x) = a\,x^2 + \hat{b}\,x + c,$$

derive an upper bound for $|\hat{r} - r|/|r|$ in terms of $|\hat{b} - b|/|b|$ with $b \neq 0$.

10. For a nonzero zero $r \in \{r_+, r_-\}$ of f as in equation **(1)** and the corresponding zero $\breve{r}$ of $\breve{f}$ with

$$\breve{f}(x) = \breve{a}\,x^2 + b\,x + c,$$

derive an upper bound for $|\breve{r} - r|/|r|$ in terms of $|\breve{a} - a|/|a|$ with $a \neq 0$.

6. Acknowledgments

I thank Dr. John Douglas, retired professor of chemistry, for his insight into the chemical examples. This work was completed in part thanks to a professional leave granted by Eastern Washington University.

7. Solutions to Odd-Numbered Exercises

1. In the floating-point format **(6)**,

$$\begin{aligned} r &= \pm R^{\mathrm{E}(r)} \times r_0.r_1 \ldots r_{m-1} r_m \ldots \\ |r| &\leq \quad R^{\mathrm{E}(r)+1}, \\ \tilde{r} &= \pm R^{\mathrm{E}(\tilde{r})} \times \tilde{r}_0.\tilde{r}_1 \ldots \tilde{r}_{m-1}, \\ \frac{|\tilde{r}-r|}{R^{\mathrm{E}(r)+1}} \leq \frac{|\tilde{r}-r|}{|r|} &\leq \frac{1}{2} \cdot R^{n-m}, \\ |\tilde{r}-r| &\leq R^{\mathrm{E}(r)+1} \times \frac{1}{2} \cdot R^{n-m} \\ &= R^{\mathrm{E}(r)} \times 0.0 \ldots 0 \frac{1}{2} \cdot R^{-m}, \end{aligned}$$

with the coefficient $\frac{1}{2}$ in the position of the digit number $(m-n)-1$.

3. Multiply the denominator and the numerator by $r \cdot \tilde{r}$.

5. If $R = 2 \cdot N$ is even, then $N^2 < R^2$, whence there are at most two not necessarily distinct digits $K, L \in \{0, \ldots, R-1\}$ such that $N^2 = KL_R$. Hence one-half $= N/R = 0.N_R$, whence one-quarter

$$(N/R)^2 = N^2/R^2 = 0.KL_R,$$

and five-quarters $= 1.KL_R$ all fit exactly as m-digit floating point numbers. Consequently, $\tilde{h} = h$ and $\tilde{d} = d$. Thus,

$$\tilde{s} = \mathtt{sqrt}(d) = s \cdot (1 + \delta_{\mathtt{sqrt}})$$

with $|\delta_{\mathtt{sqrt}}| \leq \Delta$. Also, $s = \sqrt{1.25}$, whence $1 \leq \tilde{s} \leq 1.25$ and hence the addition $1.5 \leq |h + \mathtt{sign}(h)\, \tilde{s} \leq 1.75$ does not entail any carry, so that $h \oplus \mathtt{sign}(h)\, \tilde{s} = h + \mathtt{sign}(h)\, \tilde{s}$. Moreover,

$$\begin{aligned} \tilde{t} &= -[h \oplus \mathtt{sign}(h)\, \tilde{s}] = -[h + \mathtt{sign}(h)\, \tilde{s}] = -[h + \mathtt{sign}(h)\, s] \cdot (1 + \delta_t) \\ &= t \cdot (1 + \delta_t), \end{aligned}$$

whence

$$\tilde{r}_2 = \tilde{t} \oslash a = \tilde{t} \oslash 1 = \tilde{t} = t \cdot (1 + \delta_t) = r_2 \cdot (1 + \delta_t),$$

with $|\delta_t| \leq \Delta$. Furthermore, $1 < |r_2| = 1/2 + \sqrt{5/4} < 2$, whence

$$r_2 \cdot (1 + \delta_t) = r_2 + r_2 \cdot \delta_t,$$

and hence $|\tilde{r}_2 - r_2| = |r_2 \cdot \delta_t| < R^{1-m}$. Therefore, the magnitude of the final rounding error does is smaller than one unit in the last significant digit. For instance, the 10-digit decimal HP-12C calculator gives $\tilde{r}_2 = 1.618\,033\,989$, and the foregoing argument guarantees that $1.618\,033\,988 < r_2 < 1.618\,033\,990$.

7. For Treasury bills maturing exactly one year past issue, the number of days to maturity D equals the number of days in a year Y, so $D = Y$, whence $b = D/Y = 1$, and hence $a = (b/2) - (1/4) = 1/4$. Thus,

$$(1/4)r^2 + r + c = 0,$$

with $c = (\mathcal{P} - \mathcal{F})/\mathcal{P} = 1 - (\mathcal{F}/\mathcal{P})$. Hence,

$$r^2 + 4 \cdot r + 4 + 4 \cdot (c - 1) = 0,$$

and thence $r = 2 \cdot (\sqrt{\mathcal{F}/\mathcal{P}} - 1)$ [Goldstein 1962, 450].

9. To avoid divisions by 2 throughout the analysis, let $h = b/2$. Similarly, to avoid the function `sign` throughout the analysis, multiply quadratic equations by ± 1 so that $h \geq 0$. Let r and $\tilde{r}$ be real zeros of the quadratic equations

$$a\,x^2 + 2hx + c = 0, \tag{76}$$

$$a\,x^2 + 2\tilde{h}x + c = 0. \tag{77}$$

The goals are bounds for $|\tilde{r} - r|/r$ in terms of $|\tilde{h} - h|/h$. If the zeros r and $\tilde{r}$ and real, then

$$d = h^2 - a\,c \geq 0, \tag{78}$$

$$\tilde{d} = \tilde{h}^2 - a\,c \geq 0. \tag{79}$$

Rearranging inequalities **(78)** and **(79)** gives bounds for the difference of the discriminants

$$0 \leq d = \tilde{d} - (\tilde{h}^2 - h^2), \tag{80}$$

$$0 \leq \tilde{d} = d + (\tilde{h}^2 - h^2). \tag{81}$$

Rearranging inequalities **(80)** and **(81)** gives bounds for the difference of the squared coefficients

$$\tilde{h}^2 - h^2 \leq d, \tag{82}$$

$$h^2 - \tilde{h}^2 \leq \tilde{d}, \tag{83}$$

$$|\tilde{h}^2 - h^2| \leq \max\{d, \tilde{d}\} \leq d + \tilde{d} \leq \left(\sqrt{d} + \sqrt{\tilde{d}}\right)^2. \tag{84}$$

Rearranging inequalities **(82)** and **(83)** gives bounds for sums of products of the coefficients

$$h^2 = \tilde{h}^2 + (h^2 - \tilde{h}^2) \le \tilde{h}^2 + \tilde{d} \le \left(\tilde{h} + \sqrt{\tilde{d}}\right)^2, \tag{85}$$

$$\tilde{h}^2 = h^2 + (\tilde{h}^2 - h^2) \le h^2 + d, \tag{86}$$

$$\tilde{h} \cdot h \le \tilde{h} \cdot \sqrt{\tilde{h}^2 + \tilde{d}} \le \sqrt{\tilde{h}^2 + \tilde{d}} \cdot \sqrt{\tilde{h}^2 + \tilde{d}}, \tag{87}$$

$$h^2 + \tilde{h} \cdot h \le 2 \cdot \left(\tilde{h}^2 + \tilde{d}\right) \le 2 \cdot \left(\tilde{h} + \sqrt{\tilde{d}}\right)^2. \tag{88}$$

The zero of smaller magnitude is $r_+ = (\sqrt{d} - h)/a$. If $r_+ \neq 0$, then h and d cannot both vanish, whence $h + \sqrt{d} > 0$, under the current hypothesis that $h \ge 0$. Also, from $r_+ \cdot r_- = c/a = \tilde{r}_+ \cdot \tilde{r}_-$ follows $\tilde{r}_+/r_+ = r_-/\tilde{r}_-$, whence

$$\frac{\tilde{r}_+}{r_+} = \frac{\sqrt{\tilde{d}} - \tilde{h}}{\sqrt{d} - h} = \frac{\sqrt{d} + h}{\sqrt{\tilde{d}} + \tilde{h}} = \frac{r_-}{\tilde{r}_-}. \tag{89}$$

The right-hand formula **(89)** gives

$$\frac{\tilde{r}_+ - r_+}{r_+} = \frac{\sqrt{d} + h}{\sqrt{\tilde{d}} + \tilde{h}} - 1 = \frac{(\sqrt{d} + h) - (\sqrt{\tilde{d}} + \tilde{h})}{\sqrt{\tilde{d}} + \tilde{h}} = \frac{\frac{h^2 - \tilde{h}^2}{\sqrt{d} + \sqrt{\tilde{d}}} + \frac{h - \tilde{h}}{h} \cdot h}{\sqrt{\tilde{d}} + \tilde{h}}. \tag{90}$$

Substituting inequalities **(84)**, **(85)**, and **(88)** in equations **(90)** gives

$$\frac{|\tilde{r}_+ - r_+|}{|r_+|} \le \frac{\frac{\sqrt{|h^2 - \tilde{h}^2|}}{\sqrt{d} + \sqrt{\tilde{d}}} \cdot \frac{\sqrt{|h - \tilde{h}|}}{\sqrt{h}} \cdot \sqrt{|h + \tilde{h}|} \cdot \sqrt{h} + \frac{|h - \tilde{h}|}{h} \cdot h}{\sqrt{\tilde{d}} + \tilde{h}} \tag{91}$$

$$\le \sqrt{\frac{|\tilde{h} - h|}{|h|}} \cdot \left(\sqrt{2} + \sqrt{\frac{|\tilde{h} - h|}{|h|}}\right). \tag{92}$$

A bound similar to the bound **(92)** also holds for the zero of larger magnitude, $r_- = -(\sqrt{d} + h)/a$, because $r_- = (c/a)/r_+$, and relative changes in scalar multiples of reciprocals have similar magnitudes.

References

American National Standards Institute. 1999. *International Standard ISO/IEC 9899:1999(E)*. 2nd ed. New York: American National Standards Institute.

American Standard Association. 1966. *USA Standard FORTRAN (USAS X3.9-1966)*. New York: United States of America Standards Institute (formerly American Standard Association).

Falbo, Clement. 2005. The golden ratio: A contrary viewpoint. *The College Mathematics Journal* 36 (2): 1223–1234.

Givens, J. Wallace. 1954. Numerical computation of the characteristic values of a real symmetric matrix. Technical Report ORNL-1574. Oak Ridge, TN: Oak Ridge National Laboratory.

Goldstein, Henry N. 1962. Should the Treasury auction long-term securities? *Journal of Finance* 17 (3): 444–464.

Hewlett-Packard Co. 2005. *HP-12C Financial Calculator User's Guide*. 4th ed. HP Part Number 0012C-90001. San Diego, CA: Hewlett-Packard Co.

Internal Revenue Service. 1990. A codification of documents of general applicability and future effect. In *Code of Federal Regulations*, vol. 26. Internal Revenue Service. National Archives and Records Administration: Office of the Federal Register. 45,454–45,462 and 200–203. Chapter 1, §1.861-10T and §1.861-12T.

Isaacson, Eugene, and Herbert Bishop Keller. 1994. *Analysis of Numerical Methods*. Mineola, NY: Dover.

Kahan, W. 1972. A survey of error analysis. In *Information Processing 71: Proceedings of IFIP Congress 71*, vol. 2: *Applications*, edited by C.V. Freiman. Amsterdam: North-Holland. 1214–1239.

Kincaid, David R., and E. Ward Cheney. 2002. *Numerical Analysis: The Mathematics of Scientific Computing*. 3rd ed. Providence, RI: American Mathematical Society.

Nievergelt, Yves. 2003. How (not) to solve quadratic equations. *College Mathematics Journal* 34 (2): 90–104.

Schmedel, Scott R. 1988. A math footnote. *Wall Street Journal* 119 (62): 1.

Thompson, H. Bradford. 1987. Good numerical technique in chemistry: The quadratic equation. *Journal of Chemical Education* 64 (12): 1009–1010.

Trainer, Richard D.C. 1982. *The Arithmetic of Interest Rates*. Public Information Department, 33 Liberty Street, New York, NY 10045: Federal Reserve Bank of New York.

Wall Street Journal. 1980a. Treasury sale brings lower 12.074% yield on its 52-week bills. *Wall Street Journal* (24 December 1980): 13.

_____. 1980b. Treasury will offer $4.5 billion of bills. *Wall Street Journal* (18 December 1980): 44.

Wilkinson, James Hardy. 1960. Error analysis of floating-point computation. *Numerische Mathematik* 2 (2): 319–340. *Mathematical Reviews* 0116477 (22 #7264).

_____. 1961. Error analysis of direct methods of matrix inversion. *Journal of the Association for Computing Machinery* 8 (3): 281–330.

_____. 1963. *Rounding Errors in Algebraic Processes*. Englewood Cliffs, NJ: Prentice-Hall.

_____. 1971. Modern error analysis. *SIAM Review* 13 (4): 548–568.

_____. 1972. The perfidious polynomial. In *Studies in Numerical Analysis*, edited by Gene H. Golub, 1–28. Washington, DC: Mathematical Association of America.

_____. 1988. *The Algebraic Eigenvalue Problem*. Oxford, UK: Oxford University Press.

About the Author

Yves Nievergelt completed his diploma in mathematics from the École Polytechnique Fédérale de Lausanne, Switzerland, in 1976, and then earned a Ph.D. in Several Complex Variables under the guidance of James R. King at the University of Washington, Seattle, in 1984. Since 1985 he has been teaching complex and numerical analysis at Eastern Washington University.

Reviews

Levi, Mark. 2009. *The Mathematical Mechanic: Using Physical Reasoning to Solve Problems*; 196 pp, $19.95. Princeton, NJ: Princeton University Press. ISBN 978–0–691–14020–9.

Anyone educated in the American calculus curriculum is inured to the idea that *mathematics can explain physics*. Kepler's laws, the brachistochrone, and the rising fastball in baseball are but three dramatic examples of this phenomenon; there are myriad others. This perspective, of course, was Newton's view of physics. It is a world view that he created, and it is a large part of his legacy. We should cherish it.

We all are dimly aware—though one would be hard-pressed to produce examples—that *physics can explain mathematics*. To be sure, contemplation of the potential energy on concentric spheres virtually forces Newton's inverse square law; the inverse square law and some physical analysis give rise to the elliptical orbits of planets.

Nevertheless, it is charming, and a new pleasure, to encounter a book that *systematically* presents purely mathematical results with genuine physical justifications. Author Levi has really done his homework and given us a treasure-trove of nifty results. Among these are

- a proof of the Pythagorean theorem, by examining the torques of pressure forces in a fish tank;
- a proof that the arithmetic mean dominates the geometric mean, by way of circuit analysis;
- a derivation of the cylinder of given volume and least area, using an analysis of potential energy;
- a proof of Pappus's volume theorem, by way of an analysis of work;
- a proof of Ceva's theorem (from projective geometry), by analysis of center of mass;
- a proof of Green's theorem, by way of fluid analysis;
- a proof of the Gauss-Bonnet theorem, by way of mechanics;
- a derivation of the Euler-Lagrange equations, via stretched springs;

and the list can go on at length.

The UMAP Journal 32 (3) (2011) 267–276.

I should stress that these "physical proofs" are quite rigorous and consistent with the types of arguments that we typically give in calculus. They are presented cogently and clearly and with considerable finesse. Levi is a gifted writer, and he has no trouble keeping his readers fascinated and motivated. He also, where appropriate, provides enough background to make the reading easy. Readers will find the book to be generally self-contained, and therefore a pleasure to read.

The book is organized around concepts, and each chapter conveys a cogent and well-organized body of material. One could use this book for enrichment material for students—for example, for capstone projects, or as seeds for summer research work.

The figures in the book tend to be simple line drawings, but they are well done and very clear. They certainly add considerably to the quality of the exposition, and there are plenty of them.

Levi has considerable physical intuition—much more than myself. So I found myself *really thinking* about some of his arguments. But they always worked out, and they always made me happy. Newton himself would have been charmed by this book.

The book has some nice (if brief) problem sets at the ends of some of the chapters. These tend to illustrate the ideas presented, and to give the reader more to think about.

A teacher of calculus or analysis could have a lot of fun with this book. It certainly gives a means of presenting the ideas to the students in a genuinely new way, and will be a device for keeping lectures lively and innovative. Anyone using this book to good effect in teaching calculus will not often be confronted with the question, "What is this stuff good for?"

I generally feel that I have not done my job as reviewer unless I can find at least one little thing to criticize, and it would be this: The index for this book is far too terse. This is a book where one wants to jump around and find things; and a detailed index would have helped a lot in that task. Today, with LaTeX, it is really quite straightforward to produce a very impressive index that offers considerable insight and depth. That is what is needed here. Perhaps a future edition could correct this shortcoming.

In sum, this is a delightful and unusual book that is a welcome addition to the literature. Certainly, any calculus teacher and many others of us as well will want to have it on the shelf for ready reference. It not only will enhance our teaching experience but will also teach us (the instructors) something in the process.

Steven G. Krantz, Professor of Mathematics, Washington University in St. Louis, One Brookings Drive, St. Louis, MO 63130; `sk@math.wustl.edu`.

Klymchuk, Sergiy. 2010. *Counterexamples in Calculus*. Washington DC: Mathematical Association of America; ix + 101 pp, $45.95 (P) ($35.95 to MAA members). ISBN 978–0–88385–765–6.

The toolbox of functions that we give to students prior to calculus consists of polynomials, exponentials and logarithms, rational (polynomial) expressions, trigonometric functions, and a further miscellany such as root functions, the absolute value function, and the floor function. Students are also taught to combine functions, by multiplying or adding them together, by composition, or by piecewise definitions. The calculus course adds the hyperbolic functions but primarily helps the student to analyze functions in new ways.

The student taking the first year of calculus is usually pretty busy. Furthermore, the student is on the cusp of advanced mathematics, but there is very little available that they can actually handle. To a mathematician, the archetype counterexample (or "pathological function") would be the function due to Weierstrass that is continuous everywhere but differentiable nowhere. To the calculus student, the absolute value function is an interesting counterexample because it is continuous at a single point but not differentiable there. The absolute value function is also a worthwhile example because it is defined piecewise and because it can be written as the product of the identity function and the sign function. This last observation also shows that the product of a continuous function and a noncontinuous function can be continuous.

The book reviewed here appears to be aimed at the instructor as much as to the student. However, I do not imagine many of these examples coming up in class. The counterexamples are organized by topic: Functions (15), limits(7), continuity (17), differential calculus (34), and integral calculus (16). The book is superb for a student in first-year calculus and for undergraduate math majors in general. A big virtue is that students do not have to read it in any given order but can put it down and come back to it as they like.

The book is a republication of Klymchuk [2004], which is cited (p ii, but without the hyphen in the original title). This new version has a foreword by John Mason (p. vii), which serves to cite an earlier book of Mason's. What is not cited is the more-recent book by Mason and Klymchuk [2009] that contains all of the examples of the new version with only cosmetic differences in language, plus a single additional example in the section on differential calculus (the Weierstrass function, which is merely mentioned in passing in the new version (p. 8)). The Mason and Klymchuk volume also has more material at the beginning, an extra 20 pp on both creating and using counterexamples. I prefer the new version: It is compact, I prefer its formatting, and I am not interested in the explication at the beginning.

Curiously, Mason and Klymchuk [2009] is available in paperback for $32 at Amazon, while Klymchuk [2010] costs $45.95 there, and Klymchuk

[2004] is not listed. Hint to the Mathematical Association of America: The price of Klymchuk [2010] is too high for a 100-page book, and in any case a reprint should cost less than an imported original.

References

Klymchuk, Sergiy. 2004. *Counter-Examples in Calculus*. Auckland, NZ: Maths Press.

Mason, John, and Sergiy Klymchuk. 2009. *Using Counter-Examples in Calculus*. London, UK: Imperial College Press.

James M. Cargal, Mathematics Department, Troy University—Montgomery Campus, 231 Montgomery St., Montgomery, AL 36104; jmcargal@sprintmail.com .

Thomson, Brian S., Judith B. Bruckner, and Andrew M. Bruckner. *Elementary Real Analysis*. 2nd ed. ClassicalRealAnalysis.Com, 2008; xvi + 667 pp, $33.95. ISBN 978–143484–367–8.

Bruckner, Andrew M., Judith B. Bruckner and Brian S. Thomson. *Real Analysis*. 2nd ed. ClassicalRealAnalysis.Com, 2008; xvi + 642 pp, $31.25. ISBN 978–143484–412–5.

The Crime of the Century

It is claimed by some that everyone has a secret shame, ranging from a minor faux pas to the tragic. Mine is at the second extreme: Some 35 years ago, *I did poorly in my analysis courses*. This is not something that I can escape! I know that if I run into anyone who knew me then, they will feel it necessary to point out that the funny squiggly symbol is called an "integral sign"; and in any case, transcripts live forever.

However, as I completed my master's degree, my behavior and attitudes changed and I went from being a bad student to a good one. Some of these changes were pathetically overdue (e.g., the recognition that attendance is critical), and other changes might seem trite (I moved from the back of the class to the front). However, my transformation was more qualitative that quantitative. The more-demanding professors now liked me instead of merely tolerating my presence, while the less-demanding ones now tolerated my presence. My grades did not improve much.

Nevertheless, my transformation caused me to believe that each student deserves a second chance and (if needed) a third one. Still, looking back at more than 20 years of full-time teaching, I remember a then-A+ student who reminded me that her first grade from me had been a C in Abstract Algebra I. That was, she said, "a wake-up call." I have given lots of wake-up calls, but she has been the only one to wake up.

You Can Go Home Again

I did only one more semester of pure mathematics after my master's degree, an introductory course in analysis taught very loosely out of Berberian [1970]. The professor gave a small homework assignment at the end of the first lecture; but it was not discussed subsequently, and he never again assigned homework. I alone passed the midterm; there was no final, and we were all given Bs. The lectures never reached the point of defining the Lebesgue integral. The students (and faculty) all knew that this professor could not (or would not) teach, and so did the administration. I believe that his Faustian bargain, presumably to get tenure, was that he would stop flunking students. However, I did not sign up for that course, pay tuition, and attend all of the lectures simply to get a (minimally) passing grade. I was 25, I was now serious, and—since the course seemed representative of that department—I left. That was 35 years ago, but it still makes me angry. You can go home again, but in general you do not get second chances.

When I went into industry, I needed to study probability; and I used that need as an excuse to study analysis. It did not take a long time to acquire most of the analysis that I had failed to learn in my previous courses. I did my Ph.D. in industrial engineering, and it so happens that there are individuals in that field who know analysis. My experience, though, was that professors mentioned "measure" or "Lebesgue" as a shibboleth to show that they were mathematically sophisticated. One professor insisted that a random variable is not a function from a sample space to the reals but is a "measurable" function. This single use of a form of the word "measure," as well as knowledge of such theoretical esoterica as the St. Petersburg paradox, were enough to distinguish him as one of the great theoretical geniuses of the 20th century. Now, I have taught analysis countless times during the last 20 years—but that means nothing. The fact is, that with most classes, if I were to plug a hole in some proof by invoking astrology or the Magna Carta, the students would not object.

Two Books

The two books reviewed here, *Elementary Real Analysis*, 2nd ed. (hereafter, ERA) by Thomson et al., and *Real Analysis*, 2nd ed. (hereafter, RA) by Bruckner et al., are my new favorite books in analysis. They are large inexpensive paperbacks, very well-written and organized—the sort of book that is easy to pick up and start reading in the middle. Also, as texts go, they are fairly complete. The combination of factors makes them great references, but it is also a weakness in them as texts. The authors use a scissors icon to indicate text that can be skipped; however, since most instructors will have to skip more material than the passages marked, skipping large swaths of text can make the going tougher for the student.

These books are friendly, as opposed to dry, but they are too strong for the weakest students. Then again, one can argue that the whole idea of analysis

courses is to filter out the weakest students. For example, Bryant [1990] is at the most elementary level. It is well written, sometimes quite ingenious, and makes for a decent text; but it goes to such lengths in helping the student to navigate delta-epsilon arguments that sometimes there is little for the instructor to do. But the more serious problem that recurs in any text that is so friendly is that if the student needs so much hand-holding, then maybe the student is in the wrong field. Bryant's text is ideal for self-study by the sophomore or junior trying to get a head start on analysis, but it may be too elementary for a course text.

The Question of Applications

At *The UMAP Journal*, we are interested in applied mathematics and especially in modeling. Analysis undergirds much applied mathematics. There are texts that try to combine analysis and applications; Cooper [2004] is one such book and it received a rave review from Steven Krantz [2005]. Estep [2002] is a strictly undergraduate book that is also very strong on applications, and I like it a great deal. Estep spends a great deal of time on Lipschitz-continuous functions, which leads to a nice analysis of fixed-point iteration. Uniform continuity then appears as a generalization of Lipschitz continuity; this is pedagogically a very nice way of teaching analysis. However, the book is 621 pp long. In a pure analysis course, I simply do not have time for topics such as fixed-point iteration. The question then is, what course do you use this book or Cooper's book for?

A Model Text

Understanding Analysis by Abbott [2001] is one of the most-used analysis texts in 2010. It is the text that I am teaching from, and I consider it a model undergraduate text in analysis. At 257 pp, it is less than half the length of ERA but at the same depth. ERA simply covers more material in more detail than Abbot.

ERA (and RA) are pedagogically strong; they have good examples; their exercises are excellent, and the writing is lucid and informative. They simply are more complete than many other texts. I would recommend RA to any student studying for a Ph.D. qualifying exam in analysis. (The other book that I would recommend would be *Lebesgue Integration on Euclidean Space* by Frank Jones [2001]. Despite its title, it might be a great first resource for students encountering Lebesgue integration. It has 588 pp but is quite readable.) I am not much impressed by most of the short books intended to be quick introductions to Lebesgue theory. Lebesgue integration has many details, and these details cannot be skipped—which is a reason to go to the generalized Riemann integral.

Riemann, Lebesgue, or Riemann Part II (the Revenge)?

ERA, in my view, is virtually a complete course in undergraduate real analysis; and similarly, RA is is a complete course in graduate real analysis. As a result, ERA covers the Riemann integral and RA covers the Lebesgue integral.

Jean Dieudonné suggested scrapping the Riemann integral in favor of other integrals (I myself heard him say this). In fact, there is much discussion on discarding the Riemann integral by the authors of the books reviewed here, at their Website `http://www.classicalrealanalysis.com`.

The standard higher integral is the Lebesgue integral. Its advantages over the Riemann integral are as follows:

- The Lebesgue integral applies to every function that the Riemann integral handles and then some more. The simplest example of a function that the Lebesgue integral will handle that the Riemann integral will not handle is probably the indicator function of the rational numbers, a historically important function attributed to Dirichlet. However, this function holds no interest for applied mathematicians, engineers, and physicists. It does have an interesting interpretation in probability, viz., if one does an infinite number of tosses of a fair coin, the probability that the sequence of heads and tails will eventually fall into a repeating pattern is zero. But this result too is outside of applied mathematics (to the extent that applied mathematics is about the real world). This particular view of the Lebesgue measure of subsets of $[0, 1)$—in terms of coin tossing sequences—is in fact the opening motivation for Lebesgue theory in Adams and Guillemin [1996]. I consider this book along with Capiński and Kopp [1999] as among the best short introductions to the Lebesgue integral. (However, one should go into both with a prior knowledge of probability). The standard reference for Lebesgue theory and probability is Billingsley [1995]. Rosenthal [2006] may be a more elementary treatment.

- More importantly, the Lebesgue integral leads to limit theorems that do not hold for the Riemann integral. For example, one can prove sharper versions of central theorems in probability such as the Law of Large Numbers and the Central Limit Theorem. However, proving these theorems in their most abstract forms is of little interest to mathematicians in industry or to engineers.

- Rosenthal [2006, 1] motivates the Lebesgue integral by considering a mixed random variable (one with both discrete and continuous components). Specifically, he considers a Poisson variate and a normal variate and chooses one or the other based on a coin flip. The problem with this example is that it is easy to analyze this mixed variate without using measure theory.

A fellow I knew with a Ph.D. in analysis insisted that the Riemann

integral was all one ever needs in industry. I find it hard to argue against that position. My contention, though, is that the Riemann integral is not taught to undergraduates until they take analysis. What they learn in calculus is that if $F(t)$ is the antiderivative of $f(t)$, then

$$\int_a^b f(t)\,dt = F(b) - F(a).$$

This result is what Yee and Výborný [2000, p. 1] and RA (p. 40) call *Newton's integral*. Calculus students do exercises related to the fundamental theorem of calculus, but that does not mean that they understand it. Most first-semester calculus students have difficulty viewing an expression such as $\int_a^x f(t)\,dt$ as a function of x. In any case, when do engineers and physicists or mathematicians working in industry pull out the definition of the Riemann integral? They generally use just Newton's integral, and the closest they get to the Riemann integral is numerical integration. From their vantage point, the Lebesgue integral is more abstract than the Riemann integral, is a great deal more complex, and has little utility.

The integral of Denjoy and Perron is more general than the Lebesgue integral and definitely more abstract. In the 1950s, Kurzweil and Henstock came up with a generalization of the Riemann integral that turns out to be equivalent to that integral. This integral is known variously as the *Kurzweil integral*, the *Henstock integral*, the *gauge integral*, or the *K–H integral*; I will refer to it as the *generalized Riemann integral*.

Yee and Výborný [2000] offer a worthwhile introduction to the generalized Riemann integral, as does Swartz [2001]; but the classic work by McLeod [1980] is the gold standard.

Amazingly, the generalized Riemann integral is barely more abstract conceptually than the Riemann integral and can be defined with almost exactly the same definition. That is, you can take a definition of the Riemann integral and slightly augment the wording to get a definition of the generalized Riemann integral. A short and clear statement by the late Robert Bartle and five other mathematicians defines the generalized Riemann integral and discusses bringing this definition into basic calculus texts [Bartle et al. 1996]. Although I am skeptical about that goal, I am impressed by their argument. Certainly, the generalized Riemann integral can be brought into the undergraduate real analysis course; the graduate course can then be built around it. In so doing, we do not lose measure theory as such, but the measure of a set S is now defined as $\int_S 1$. (If the integral does not exist, then S is a nonmeasurable set.)

A Little History

Whether it is a good idea to take a historical approach to first learning a subject has two answers:

- A student should do whatever he or she finds helpful regardless of what others think.
- Some disciplines have nice historically oriented introductory texts, while other disciplines do not. For example, in number theory the text by Ore [1988] works very well as an introduction, whereas the text by Goldman [1997] does the same thing but at a higher level of mathematical maturity and covers a great deal more material. In analysis, David Bressoud has written two very well-reviewed historical texts on analysis [2006; 2008], with the second devoted to the history of the Lebesgue integral; Steven Krantz gave it a rave review in this journal [2008]. I like Dunham [2008] a great deal for a superb introduction to the Lebesgue integral. Both Bressoud and Dunham owe something to Hawkins [2001]. Both ERA and RA do a good job of integrating history into the text, although not on the scale of these books.

I believe that both ERA and RA are great additions to the literature on analysis. The first is a good investment for both undergraduates and graduate students studying analysis, the second is worthwhile for students studying the Lebesgue integral. They are well-written and rich in content.

Nevertheless, it is time for the mathematics community to switch to the generalized Riemann integral—and to develop the Lebesgue integral as needed from there.

References

Abbott, Stephen. 2001. *Understanding Analysis*. New York: Springer.

Adams, Malcolm and Victor Guillemin. 1996. *Measure Theory and Probability*. Boston, MA: Birkhäuser.

Bartle, Robert, Ralph Henstock, Jaroslav Kurzweil, Eric Schechter, Stefan Schwabik, and Rudolf Výborný. 1996. An open letter to authors of calculus textbooks. `http://www.math.vanderbilt.edu/~schectex/ccc/gauge/letter/` .

Berberian, Sterling K. 1970. *Measure and Integration*. New York: Chelsea.

Billingsley, Patrick. 1995. *Probability and Measure*. 3rd ed. New York: Wiley.

Bressoud David. 2006. *A Radical Approach to Real Analysis*. 2nd ed. Washington, DC: Mathematical Association of America.

_____. 2008. *A Radical Approach to Lebesgue's Theory of Integration*. Washington, DC: Mathematical Association of America, and New York: Cambridge University Press.

Bryant, Victor. 1990. *Yet Another Introduction to Analysis*. New York: Cambridge University Press.

Capiński, Marek, and Ekkehard Kopp. 1999. *Measure, Integral and Probability*. New York: Springer. 2007. 2nd ed.

Cooper, Jeffery. 2004. *Working Analysis*. Boston, MA: Academic Press.

Dunham, William. 2008. *The Calculus Gallery: Masterpieces from Newton to Lebesgue*. Princeton, NJ: Princeton University Press.

Estep, Donald. 2002. *Practical Analysis in One Variable*. New York: Springer.

Goldman, Jay. 1997. *The Queen of Mathematics: An Historically Motivated Guide to Number Theory*. Natick, MA: A K Peters

Hawkins, Thomas. 2001. *Lebesgue's Theory of Integration: Its Origins and Development*. 2nd ed. Providence, RI: American Mathematical Society.

Jones, Frank. 2001. *Lebesgue Integration on Euclidean Space*. Revised ed. Sudbury, MA: Jones and Bartlett.

Krantz, Steven. 2005. Review of Cooper [2004]. *The UMAP Journal* 26(4): 471–473.

_____. 2008. Review of Bressoud [2008]. *The UMAP Journal* 29(1): 85–87.

McLeod, Robert F. 1980. *The Generalized Riemann Integral*. Washington, DC: Mathematical Association of America.

Ore, Øystein. 1988. *Number Theory and Its History*. Mineola, NY: Dover.

Rosenthal, Jeffrey S. 2006. *A First Look at Rigorous Probability Theory*. Hackensack, NJ: World Scientific Press.

Swartz, Charles. 2001. *Introduction to Gauge Integrals*. Hackensack, NJ: World Scientific Press.

Yee, Lee Peng, and Rudolph Výborný. 2000. *The Integral: An Easy Approach after Kurzweil and Henstock*. New York: Cambridge University Press.

James M. Cargal, Mathematics Department, Troy University—Montgomery Campus, 231 Montgomery St., Montgomery, AL 36104; `jmcargal@sprintmail.com`.