# Interesting Examples in Undergraduate Mathematics

Matthew McGonagle

June 30, 2021

## Contents

# 1 Introduction

The purpose of these notes is examples in undergraduate mathematics that the author considers to be interesting; this could be from applications or pure mathematical interest.

# 2 Precalculus

## 2.1 Descarte's Method of Tangents

### The Setup

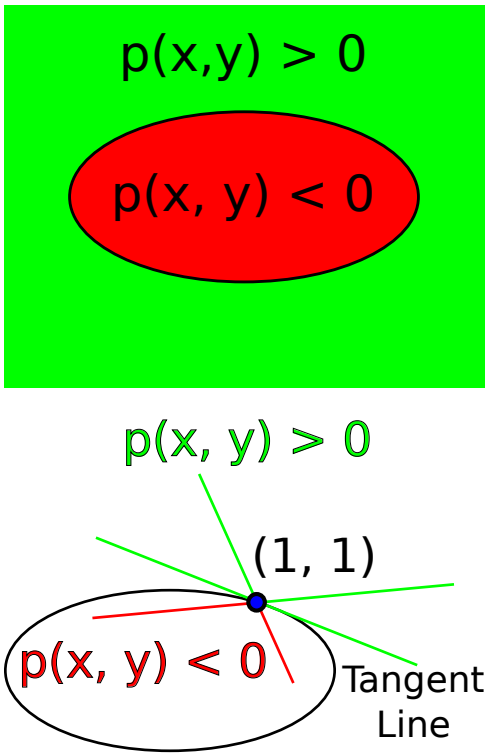
Prior to the differential calculus created by Leibniz and Newton, Descarte invented several methods of finding tangent lines to curves that are described by algebraic equations. These methods are purely algebraic; they don't use either the concept of limits or infinitesimals. A nice reference on these methods is Suzuki [16]. Here we will discuss one of Descarte's simpler methods. As an example, we will use Descarte's method to find the tangent line to the ellipse $\{x^2 + 2y^2 = 3\}$ at the point $(1, 1)$.

To begin, let us define the two variable polynomail $p(x, y) = x^2 + 2y^2 - 3$; we see that our ellipse is the zero set $\{p(x, y) = 0\}$.

Next, let us describe all of the non-vertical lines (not necessarily tangent lines) that pass through the point $(1, 1)$. Consider any line $y = mx + b$. If this line passes through $(1, 1)$, then we have $1 = m + b$. So all of these lines are described by $y = mx + 1 - m$ for different constants $m$.

Next, let us observe that the interior of the ellipse is the set $\{p(x, y) < 0\}$ and the exterior is the set $\{p(x, y) > 0\}$.

Next, let us see what the sign of $p(x, y)$ looks like along different lines passing through $(1, 1)$. When a line passing through $(1, 1)$ is not a tangent line, close to $(1, 1)$ the line passes through both the interior and exterior of the ellipse. Therefore, for a non-tangent line, the sign of $p(x, y)$ changes. However, a tangent line will remain in the exterior of the ellipse (except at the point $(1, 1)$), so $p(x, y)$ will not change sign along the line. Consider the figure below.




So, given a line $y = mx + 1 - m$ passing through $(1, 1)$, let us compute the value of $p(x, y$ along the line. We parameterize this value by $x$, so we have

$$\begin{aligned} q(x) &\coloneqq p(x, mx + 1 - m), && (1)\\ &= x^2 + 2(mx + 1 - m)^2 - 3, && (2)\\ &= (1 + 2m^2)x^2 + 4(m - m^2)x + 2(1 - m)^2 - 3. && (3) \end{aligned}$$

Now, we know that $q(1) = 0$ since the line must intersect the ellipse at $(1, 1)$. Futhermore, from our argument above, we know that for non-tangent lines $q(x)$ changes sign at $x = 1$ and that the tangent line must NOT change sign at $q(x) = 1$.

Since the tangent line doesn't change sign, we know that for the correct value of the slope $m$ for the tangent line, the polynomial $q(x)$ has a double root at $x = 1$. Therefore, for the tangent line, the polynomial is divisible by $(x-1)^2 = x^2 - 2x + 1$; that is, after computing the polynomial division $\frac{q(x)}{x^2-2x+1}$ we find a remainder of zero.

If we find that there is only one value of $m$ that gives us a remainder of zero for this polynomial division, then it must be the slope of the tangent line. Note that for non-tangent lines, we could also have that $q(x)$ is divisible by $(x - 1)^2$, since the sign will also change if there is a triple root.

**The Problem**

Compute the slope of the tangent line $y = mx + 1 - m$ passing through $(1, 1)$ for the ellipse $\{x^2 + 2y^2 = 3\}$ by finding the value of $m$ for which the polynomial division

$$\frac{(1 + 2m^2)x^2 + 4(m - m^2)x + 2(1 - m)^2 - 3}{x^2 - 2x + 1}, \tag{4}$$

has vanishing remainder.

**The Solution**

Let us perform the polynomial division.

$$\frac{(1 + 2m^2)x^2 + 4(m - m^2)x + 2(1 - m)^2 - 3}{x^2 - 2x + 1}, \tag{5}$$

$$= 1 + 2m^2 + \frac{(2 + 4m^2 + 4m - 4m^2)x - 1 - 2m^2 + 2(1 - m)^2 - 3}{x^2 - 2x + 1}, \tag{6}$$

$$= 1 + 2m^2 + \frac{(2 + 4m)x - 4m - 2}{x^2 - 2x + 1}. \tag{7}$$

The degree of the numerator is less than the denominator, so the numerator is the remainder. Therefore, we must have that $(2 + 4m)x - 4m - 2 = 0$. So the coefficients must vanish and we get

$$\begin{cases} 2 + 4m = 0, \\ -4m - 2 = 0. \end{cases} \tag{8}$$

The system has the unique solution $m = -1/2$; therefore the slope of the tangent line must be $m = -1/2$. So we find the tangent line to be

$$y = -\frac{1}{2}x + \frac{3}{2}. \tag{9}$$

## 2.2 Cardano's Method of Solving Cubic Equations

In this section we will look at Cardano's method for solving cubic equations using a specific example. We will motivate the algebraic manipulations using Cardano's geometric manipulations. For a reference on Cardano's point of view, see [1].

**The Setup**

Cardano's method is based on geometry. Consider the equation

$$x^3 = 6x^2 + x + 2. \tag{10}$$

Cardano considers this as a geometric problem of matching volumes. The left hand side represents the volume of a cube of side lengths $x$. Each term on the right hand side represents a specific volume. So we have

| Left | Right |
|---|---|
| • Cube with side lengths $x$. | • Box of dimensions $6 \times x \times x$.<br>• Box of dimensions $1 \times 1 \times x$.<br>• Cube of side lengths $2^{1/3}$. |

Now, Cardano's idea is to break up the length $x$ into two lengths $a$ and $b$. That is we, look at $x = a + b$.

When we do so, we get

| Left | Right |
|---|---|
| • Cube of side lengths $a$.<br>• Cube of side lengths $b$.<br>• Box of dimensions $3a \times b \times b$.<br>• Box of dimensions $3b \times a \times a$. | • Box of dimensions $6 \times a \times a$.<br>• Box of dimensions $6 \times b \times b$.<br>• Box of dimensions $12 \times a \times b$.<br>• Box of dimensions $1 \times 1 \times a$.<br>• Box of dimensions $1 \times 1 \times b$.<br>• Cube of side lengths $2^{1/3}$. |

Cardano's idea is to use some or all of the non-cube volumes on the left to cancel problematic terms on the right. This is done in two stages:

1. (Depress the Cubic) Divide $x = a + b$, and use $a, b$ to remove the second order terms from the equation.
2. Say the new free variable is $a$, then divide $a = c + d$ and use $c, d$ to remove the first order terms.

We will look at following these ideas using algebraic manipulation.

**The Problem**

Use Cardano's ideas to solve

$$x^3 = 6x^2 + 3x + 2. \tag{11}$$

**The Solution**

We follow the two stages of Cardano.

1. Fist we depress the cubic. We split $x = a + b$. So we get

$$a^3 + 3a^2b + 3ab^2 + b^3 = 6a^2 + 12ab + 6b^2 + 3a + 3b + 2. \tag{12}$$

Now, note that if we fix $b$ and let $a$ be a new free variable, then it is possible to eliminate all terms for $a^2$. That is we need to find $b$ such that $3a^2b = 6a^2$. So we choose $b = 2$. Then we get

$$a^3 + 12a + 8 = 24a + 24 + 3a + 8. \tag{13}$$

Collecting all non-third order terms to the right, we get

$$a^3 = 15a + 24. \tag{14}$$

2. Now we split $a = c + d$ to get

$$c^3 + 3c^2d + 3cd^2 + d^3 = 15c + 15d + 24. \tag{15}$$

Now, note that to get rid of first order terms and not introduce second order terms, we must use the entirety of the non-cube terms on the left $3c^2d + 3cd^2$ and eliminate all of the linear terms on the right $15c + 15d$. So we look at the possibilities of matching

$$3c^2d + 3cd^2 = 15(c + d). \tag{16}$$

Now, note that $3c^2d + 3cd^2 = 3cd(c + d)$. So, we are lead to $3cd = 15$. Now after eliminating the matching, we have the following system

$$\begin{cases} c^3 + d^3 = 24, \\ cd = 5. \end{cases} \tag{17}$$

From this, we get that

$$125 + d^6 = 24d^3. \tag{18}$$

This is quadratic in $d^3$, so we may solve for $d^3$ using the quadratic formula. We get

$$d^3 = \frac{24 \pm \sqrt{24^2 - 4 * 125}}{2}, \tag{19}$$

$$= 12 \pm \sqrt{144 - 125}, \tag{20}$$

$$= 12 \pm \sqrt{19}. \tag{21}$$

Before, we continue, note that

$$\frac{1}{12 + \sqrt{19}} = \frac{12 - \sqrt{19}}{144 - 19}, \tag{22}$$

$$= \frac{12 - \sqrt{19}}{125}. \tag{23}$$

Therefore, from $cd = 5$, we see that when $c = \left(12 + \sqrt{19}\right)^{1/3}$, we have that $d = \left(12 - \sqrt{19}\right)^{1/3}$. The opposite situation holds, but without a loss

of generality we may suppose that the magnitudes $|c|$ and $|d|$ are given by the above.

Finally, we deal with the fact that there are three cube roots in the complex numbers. Let $\zeta = e^{i2\pi/3}$. From $cd = 5$, we have

$$c = \left(12 + \sqrt{19}\right)^{1/3} \zeta^k, d = \left(12 - \sqrt{19}\right)^{1/3} \zeta^{-k}, \tag{24}$$

for $k = 0, 1, 2$.

Finally, we use that $a = c + d$ to get that

$$a = \left(12 + \sqrt{19}\right)^{1/3} \zeta^k + \left(12 - \sqrt{19}\right)^{1/3} \zeta^{-k}, \tag{25}$$

for $k = 0, 1, 2$.

Then we use $x = 2 + a$ to get

$$x = 2 + \left(12 + \sqrt{19}\right)^{1/3} \zeta^k + \left(12 - \sqrt{19}\right)^{1/3} \zeta^{-k}, \tag{26}$$

for $k = 0, 1, 2$. We have found the three roots of the equation.

# 3 One Variable Differential Calculus

## 3.1 Gauss and the Gauss Distribution

### History

A good reference on the history of the gaussian distribution is [13].

The history of how to deal with errors is intimately tied to astronomy; astronomical predictions involve quantities that need to be measured to high precision. Practical limits force astronomers to deal with the errors of predictions or measurements never being in complete agreement.

In the 18th century and early 19th century, there was some confusion as to how to deal with these errors in measurement. As an example, there was some dispute as to whether to use the average or the median of measurements. One of the problems was a theoretical foundation for understanding error was in its infancy. For example, Laplace created a model of typical error that is far from the typical gaussian distribution considered today.

So how did Gauss arrive at his distribution? First it should be noted that he worked on modeling error while solving a problem in astronomy. On January 1, 1801, Giuseppe Piazzi observed the Ceres asteriod. He was interested in whether Ceres was a new planet, but he could only take a small number of observations of its position before it disappered behind the sun. Ceres was estimated to be visible again after about a year, which left many astronomers with the question of where to find it in the sky.

Gauss greatly increased his reputation by correctly solving this problem; in fact, his correct answer was actually in disagreement with most reputable

astronomers. Aside from his masterful use of geometry, part of his solution is how to deal with the errors in measurements that were made. It is this problem that lead him to the gaussian distribution as a model for the error.

His approach to modeling the error is the following.

He considers the errors to be random described by a differentiable probability density $p(x)$. The distribution of the errors should satisfy the following:

1. Smaller errors are more probable, i.e. the density $p(x)$ should have a maximum at $x = 0$.

2. The distribution of errors should symmetric, i.e. $p(-x) = p(x)$.

3. Consider any observed quanitity $X$ with true value $X_0$ and errors modeled by our distribution, i.e. $X = X_0 + G$ where $P(G = x) = p(x)$.

   Given any set of observations $\{x_1, x_2, ..., x_n\}$, then the likelihood $P(x_1, x_2, ..., x_n|X_0)$ (i.e. the probability of observing $x_1$, $x_2$, ... $x_n$ given the true value is $X_0$) is maximized by $X_0$ being the average of $\{x_1, x_2, ..., x_n\}$ (i.e. $X_0 = \frac{x_1+x_2+...+x_n}{n}$). Let us explain this in a little more detail.

   We are assuming that the errors $\{x_1, x_2, ..., x_n\}$ are independent. So

   $$P(x_1, x_2, ..., x_n|X_0) = p(x_1 - X_0)p(x_2 - X_0)...p(x_n - X_0). \tag{27}$$

   When we speak of maximizing the likelihood, we think of all of the observations $x_i$ being fixed. So the above is considered to a function of only the one variable $X_0$. That is, we are considering the likelihood functions

   $$L(X_0) = p(x_1 - X_0)p(x_2 - X_0)...p(x_n - X_0). \tag{28}$$

   Then our assumption is that the maximum of $L(X_0)$ occurs at the average of our observations $X_0 = \frac{x_1+x_2+...+x_n}{n}$.

   This amounts to Gauss's justification of using averages over median. He is purposefully choosing a model of error where the average of the observations is the most likely explanation of the true value.

**The Problem**

Show that Gauss' requirements on $p(x)$ force $p(x)$ to be a Gaussian distribution.

**The Solution**

First, let us consider condition (3) and the consequences of maximizing the likelihood. First note that $L(X_0) \geq 0$, so maximizing $L(X_0)$ is equivalent to maximizing $f(X_0) = \log(L(X_0))$. Using the logarithm will be more convenient as it will turn the product of the $p(x_i - X_0)$ into a sum of logarithms; so we have

$$h(X_0) = \log(p(x_1 - X_0)) + \log(p(x_2 - X_0)) + ... + \log(p(x_n - X_0)). \tag{29}$$

To find the maximum, let's set the derivative to be zero:

$$0 = h'(X_0) = -\left(\frac{p'(x_1 - X_0)}{p(x_1 - X_0)} + \frac{p'(x_2 - X_0)}{p(x_2 - X_0)} + ... + \frac{p'(x_n - X_0)}{p(x_n - X_0)}\right). \quad (30)$$

Now the key is that condition (3) applies to any possible set of observations, no matter how unprobable. Since $p(x)$ is continuous with maximum at $x = 0$, we know that there exists an interval $[-\delta, \delta]$ around $x = 0$ such that $p(x) > 0$ for all $x \in [-\delta, \delta]$. In particular, we know that observations in $[X_0 - \delta, X_0 + \delta]$ are all possible. So now consider any real number $r \in [-\delta, \delta]$ and the observations $\{x_1 = X_0\}$ and $\{x_2 = ... = x_n = X_0 + r\}$.

Since we have already fixed $X_0$ to represent our true value, let us now use $y$ as the independent variable for our likelihood. So we seek to maximize

$$L(y) = p(x_1 - y)p(x_2 - y)...p(x_n - y). \quad (31)$$

Condition (3) says this maximum is at $y = \frac{x_1 + x_2 + ...x_n}{n} = X_0 + \frac{n-1}{n}r$. To simplify notation, let $f(x) = \frac{p'(x)}{p(x)}$. So we get

$$0 = f\left(-\frac{n-1}{n}r\right) + (n-1)f\left(\frac{1}{n}r\right). \quad (32)$$

Now, note that $p(x)$ symmetric implies that $p'(x)$ is anti-symmetric. Therefore $f(x)$ is anti-symmetric. So we get that

$$f\left(\frac{n-1}{n}r\right) = (n-1)f\left(\frac{1}{n}r\right). \quad (33)$$

What are the consequences of this equation? Fix any $r_0$ small enough such that $2r_0$ is in the interval $[-\delta, \delta]$. Now note that $\frac{n}{n-1}r_0$ is also in the interval for any $n > 1$, and consider $r = \frac{n}{n-1}r_0$. Then we have that

$$\frac{1}{n-1}f(r_0) = f\left(\frac{r_0}{n-1}\right). \quad (34)$$

Now consider $0 < k \leq n + 1$. Note that $\frac{k}{n}r$ is in the interval too, and now apply 32 for $n- > k$ and $r- > \frac{k}{n}r$, we get

$$f\left(\frac{k-1}{n}r\right) = (k-1)f\left(\frac{1}{n}r\right). \quad (35)$$

So for any fraction of the form $fracmn$ where $0 < m \leq n$, we have that

$$f\left(\frac{m}{n}r_0\right) = \frac{m}{n}f(r_0). \quad (36)$$

Consider the function $g(x) = f(r_0)x$. We have that $g(x) - f(x) = 0$ for any $x$ that is a rational multiple of $r_0$ and $0 < |x| \leq r_0$. Hence, by the continuity of $f(x)$, we have that $f(x) = g(x) = f(r_0)x$ for all $|x| \leq r_0$.

Therefore, we may write that $f(x) = kx$ for some constant $k$ and $x$ on some interval around zero. This gives us the differential equation

$$\frac{p'(x)}{p(x)} = k, \tag{37}$$

locally around $x = 0$. Integrating we get

$$\log(p(x)) = \frac{k}{2}x^2 + C. \tag{38}$$

This can be written in the form $p(x) = Ae^{-Bx^2}$. So we see the probability distribution extends to be non-zero on all $x$. Furthermore, the constants $A$ and $B$ can be related by the fact that $p(x)$ is a probability density so that $\int_{-\infty}^{\infty} Ae^{-Bx^2}\,dx = 1$. It is standard to solve for $A$ in terms of $B$.

To solve $\int_{-\infty}^{\infty} e^{-Bx^2}\,dx$, we square the integral and switch to polar coordinates to get

$$\int_{-\infty}^{\infty} e^{-Bx^2}\,dx = \sqrt{\frac{\pi}{B}}. \tag{39}$$

So we get that $A = \sqrt{\frac{B}{\pi}}$, and then

$$p(x) = \sqrt{\frac{B}{\pi}}e^{-Bx^2}. \tag{40}$$

## 3.2 The Schwarzian Derivative

In this example we will derive the form of the Schwarzian derivative $Su$ for a smooth function $u : \mathbb{R} \to \mathbb{R}$ given by

$$Su = \frac{u'''}{u'} - \frac{3}{2}\left(\frac{u''}{u'}\right)^2. \tag{41}$$

What is special about the Schwarzian derivative is that you get the same result if you apply it to a function $w$ that is a linear fractional transformation of $f$, i.e. for

$$w(x) = \frac{A + Bu(x)}{C + Du(x)}, \tag{42}$$

for constants $A, B, C$, and $D$, we have that $Su = Sw$. We will first briefly discuss linear fractional transformations and then how we aim to derive the Schwarzian derivative from the above invariance property.

A nice introductory reference to the Schwarzian derivative and its relationship to projective differential geometry is [10].

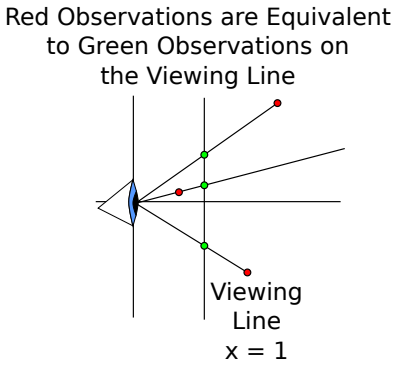
**The Setup: Linear Fractional Transformations**

A linear fractional transformation for one real variable is a function of the form

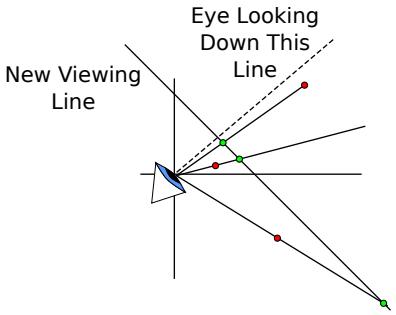
$$T(y) = \frac{A + By}{C + Dy}, \tag{43}$$

for constants $A, B, C$ and $D$. Note that the function is not defined for the single point $C + Dy = 0$ when $D \neq 0$.

An important motivation for linear fractional transformations is perspective transformations. First imagine an eye sitting at the origin in $\mathbb{R}^2$, and the eye is looking down the x-axis towards $+\infty$. Everything the eye observes is equivalent to seeing something sitting in its viewing line (or plane for 3-dimensions). We will take the viewing line of the eye to be $\{x = 1\}$. See the figure below.




Now imagine that the eye rotates $\pi/4$ radians counter-clockwise; it is now looking down a line that makes an angle of $\pi/4$ radians with the x-axis. Furthermore, its viewing line has rotated to be the line $\left\{y + x = 2\sqrt{\frac{1}{2}}\right\}$. Let $\tilde{y}$ denote the position along this new viewing line. See the figure below




Let $\tilde{y}$ measure position on the viewing line relative the line $\{y = x\}$, i.e. the line that eye is looking down. If an object appears to be at a point $y$ on the original viewing line, what is its new position $\tilde{y}$ on the new viewing line? The

answer turns out to be a linear fractional transformation, i.e.

$$\tilde{y} = \frac{A + By}{C + Dy}, \tag{44}$$

for some constants $A, B, C$, and $D$ depending on the angle of rotation.

Linear fractional transformations are also a classical topic in complex analysis, projective geometry, and conformal geometry.

### Finding the Schwarzian Derivative

Now we will discuss our method of finding the Schwarzian derivative $Su$ of a function $u : \mathbb{R} \to \mathbb{R}$. We will demand that $Su$ has the following properties:

1. The Schwarzian derivative is defined by

$$Su = F(u, u', u'', u'''), \tag{45}$$

   for some function $F(a, b, c, d)$.

2. Given any function $u(x)$ and linear fractional transformation $T(y) = \frac{A+By}{C+Dy}$, we have that the function defined by

$$w(x) = T(u(x)) = \frac{A + Bu(x)}{C + Du(x)}, \tag{46}$$

   satisfies $Sw = Su$.

You can interpret the Schwarz derivative via perpective transformations. Suppose there is a particle moving in a 2-dimensional plane and the eye from our first set-up in the previous section observes it at position $u(t)$ in its viewing line. Next, suppose another observer is at 45 degrees relative to the first observer (so the second set-up) and observes the particle at position $w(t)$ on their viewing line. Furthermore, suppose neither observer knows the distance of their viewing plane and doesn't know the angle difference of their perspectives. Is there some measure of the rate of change in the particle's position that they can agree on?

Yes, the Schwarzian derivative. Their observations differ by some linear fractional transformation; since they are missing data on their perspectives, they can't reconstruct the actual position of the particle and they also can not reconstruct the details of the linear fractional transformation that relates their observations. However, the Schwarzian derivative will always be the same despite no knowing the exact transformation. It is enough to just know that it is a linear fractional transformation.

### The Problem

Given the two properties of the Schwarzian derivative listed above, show that

$$F(u, u', u'', u''') = J\left(\frac{u'''}{u'} - \frac{3}{2}\left(\frac{u''}{u'}\right)^2\right), \tag{47}$$

for some function $J$. Thus, deduce that a reasonable definition of the Schwarzian derivative is

$$Su = \frac{u'''}{u'} - \frac{3}{2}\left(\frac{u''}{u'}\right)^2. \tag{48}$$

**The Solution**

First, we start by looking at the consequences of having an invariant for translations $T(y) = A + y$ which are fractional transformations for $B = 1, C = 1$, and $D = 0$. Note that the derivatives of $u$ and $w$ are equal. So we get that for any function $u$ that

$$F(u + A, u', u'', u''') = F(u, u', u'', u'''). \tag{49}$$

Now, note that for any point $(a, b, c, d)$, we can find a function $u$ such that $(u, u', u'', u''') = (a, b, c, d)$ for some point $x$. So we get that

$$F(a + A, b, c, d) = F(a, b, c, d), \tag{50}$$

for any $A, a, b, c$, and $d$. In particular, since we can vary $A$ without changing the right hand side, we see that the value $F(a, b, c, d)$ is actually independent of $a$. Therefore, we may find a function $G(b, c, d)$ such that

$$F(a, b, c, d) = G(b, c, d). \tag{51}$$

Note that $G(u', u'', u''')$ will be invariant for linear fractional transformations of $u$ and that $G(u', u'', u''')$ depends only on the derivatives of $u$.

Next, we consider the effect of scalings $T(y) = By$, these are also a specific class of linear fractional transformations. We have that

$$G(Bu', Bu'', Bu''') = G(u', u'', u'''), \tag{52}$$

for any $B$. Similar to before, we get that

$$G(Bb, Bc, Bd) = G(b, c, d), \tag{53}$$

for any $B, b, c$, and $d$. Therefore, we see that $G$ must be homogeneous of degree 0, i.e.

$$G(\lambda b, \lambda c, \lambda d) = G(b, c, d), \tag{54}$$

for any constant scaling $\lambda$.

Next, we consider the specific linear fractional transformation $T(y) = \frac{1}{y}$. We then have that

$$w = \frac{1}{u}, \tag{55}$$

$$w' = -\frac{u'}{u^2}, \tag{56}$$

$$w'' = \frac{2u'^2}{u^3} - \frac{u''}{u^2}, \tag{57}$$

$$w''' = \frac{6u'u''}{u^3} - \frac{6u'^3}{u^4} - \frac{u'''}{u^2}. \tag{58}$$

So from the invariance of $G$, we get that

$$G(u', u'', u''') = G\left(-\frac{u'}{u^2}, \frac{2u'^2}{u^3} - \frac{u''}{u^2}, \frac{6u'u''}{u^3} - \frac{6u'^3}{u^4} - \frac{u'''}{u^2}\right). \tag{59}$$

Now use the 0-homoegeneity of $G$ to get

$$G(u', u'', u''') = G\left(-1, \frac{2u'}{u} - \frac{u''}{u'}, \frac{6u''}{u} - \frac{6u'^2}{u^2} - \frac{u'''}{u'}\right) \tag{60}$$

Note that the expression on the right hand side contains terms involving $u$ without any derivatives, which isn't directly supplied as an argument to $G$ on the left hand side. So when like before we pass to general coordinates, the $u$ terms turn into general constants $A$. So we get

$$G(b, c, d) = G\left(-1, \frac{2b}{A} - \frac{c}{b}, \frac{6c}{A} - \frac{6b^2}{A^2} - \frac{d}{b}\right), \tag{61}$$

for any constant $A$. So there exists a function $H$ such that

$$G(b, c, d) = H\left(\frac{2b}{A} - \frac{c}{b}, \frac{6c}{A} - \frac{6b^2}{A^2} - \frac{d}{b}\right), \tag{62}$$

for any constant $A$.

The fact that $A$ is any general constant that appears on the right hand side but is not on the left means that there is more information to draw from $H$. Let $\theta = \frac{2b}{A} - \frac{c}{b}$. So

$$\frac{6c}{A} - \frac{6b^2}{A^2} - \frac{d}{b} = \frac{3c}{b}\left(\theta + \frac{c}{b}\right) - \frac{3}{2}\left(\theta + \frac{c}{b}\right)^2 - \frac{d}{b}, \tag{63}$$

$$= 3\left(\theta + \frac{c}{b}\right)\left(\frac{c}{2b} - \frac{\theta}{2}\right) - \frac{d}{b}, \tag{64}$$

$$= \frac{3}{2}\left(\frac{c^2}{b^2} - \theta^2\right) - \frac{d}{b}. \tag{65}$$

So, we get that

$$G(b, c, d) = H\left(\theta, \frac{3}{2}\left(\frac{c^2}{b^2} - \theta^2\right) - \frac{d}{b}\right). \tag{66}$$

Now, freeze $b, c$ and $d$ while varying $A$. The left hand side is constant, but the value of $\theta$ on the right hand side is varying. So we see that the right hand side is constant on $\{\theta < -\frac{c}{b}\}$ and it is also constant on $\{\theta > -\frac{c}{b}\}$. Note as long as $c \neq 0$ and $b \neq 0$, we can find a value of $A$ such that $\theta = 0$. So we have

$$G(b, c, d) = H\left(0, \frac{3}{2}\left(\frac{c}{b}\right)^2 - \frac{d}{b}\right), \tag{67}$$

as long as $c \neq 0$ and $b \neq 0$.

So for $b, c \neq 0$, we can write

$$G(b,c,d) = J\left(\frac{d}{b} - \frac{3}{2}\left(\frac{c}{b}\right)^2\right). \tag{68}$$

As long as $J$ is continuous, it extends to $b \neq 0$. So putting it all together, we get

$$F(u,u',u'',u''') = J\left(\frac{u'''}{u'} - \frac{3}{2}\left(\frac{u''}{u'}\right)^2\right). \tag{69}$$

We may take the expression inside parenthesese on the right hand side to be the Schwarzian derivative, i.e.

$$Su = \frac{u'''}{u'} - \frac{3}{2}\left(\frac{u''}{u'}\right)^2. \tag{70}$$

You can now directly verify that if $w(x) = T(u(x))$ for a linear fractional transformation $T$, then $Sw = Su$.

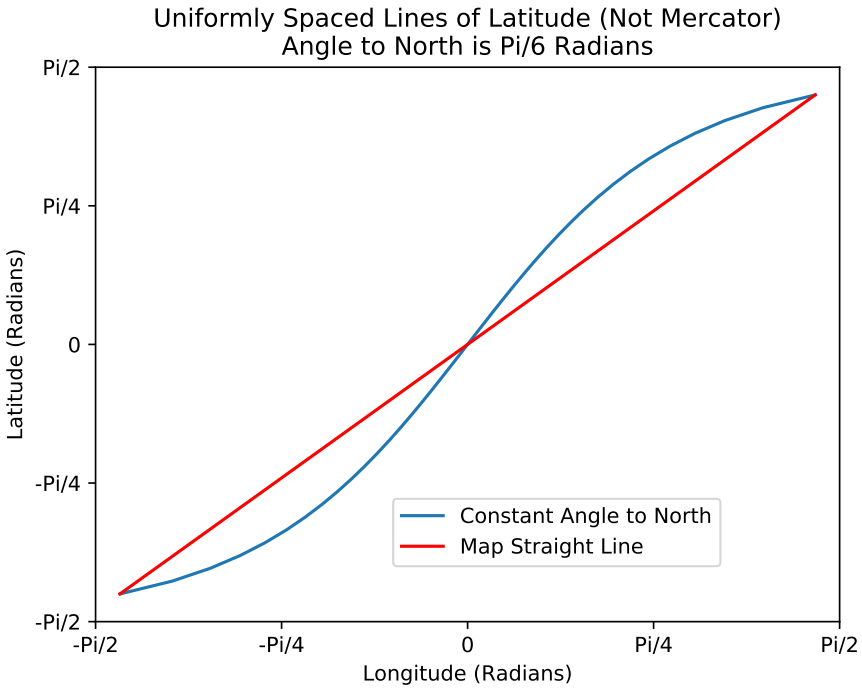
# 4 One Variable Integral Calculus

## 4.1 The Mercator Map and the Integral of Secant

### Historical Motivation

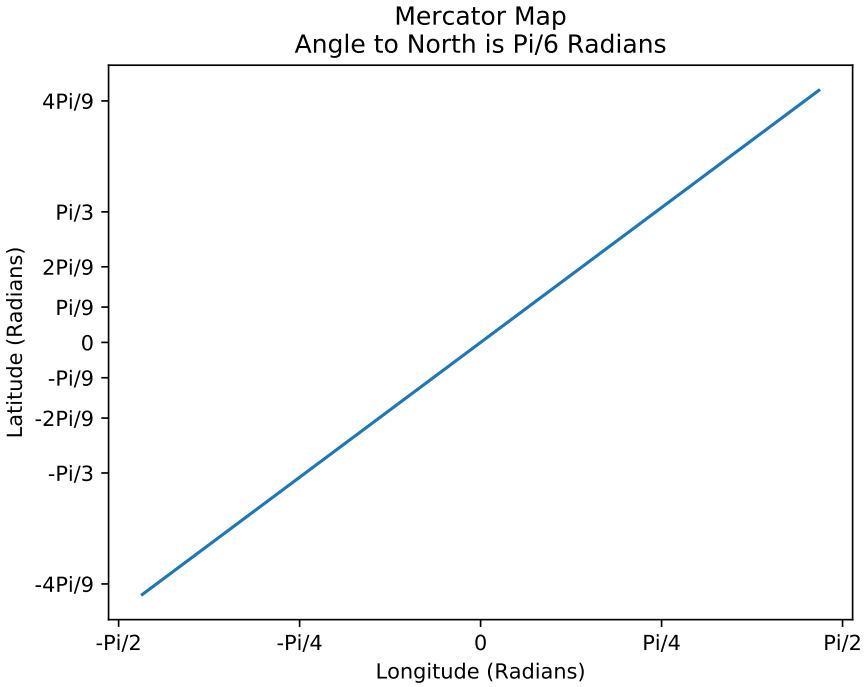
A great reference for the history of the integral of the secant function and its relation to the Mercator map is [12]. We will give a brief overview here.

The Mercator Map of the world spaces out the lines of latitude in a particular way inorder to solve a problem in naval navigation. The problem is that ships would navigate by sailing with a fixed angle to due north (e.g. as seen on a compass). This creates an issue for making map. Consider a map where the lines of latitude are spaced out evenly in the vertical direction (so NOT the Mercator map); for such a map, a course with fixed angle to magnetic north is NOT a straight line on the map. The figure below shows the path of a course with constant angle to magnetic north on a map with uniformly spaced lines of latitude.




The problem is that the lines of latitude get represent shorter and shorter distances as you move from the equator towards either of the poles. This means that there is a complicated relationship between the angle measured on this map and the true angle to magnetic north it represents.

In 1569, Mercator had the idea that he could create a map where the lines of latitude are NOT spaced evenly; if you choose the variation in spacing in the correct manner, then a course with fixed angle to magnetic north will be a straight line on this new map. Furthermore, the angle measured on the map will match the true angle to magnetic north. Consider the following figure that shows a course with constant angle to magnetic north on a Mercator map, note that the lines of latitude are not evenly spaced (the values marked are multiples of $\pi/9$ radians).




Unfortunately, Mercator didn't give a clear formula to precisely describe how to space out the lines of latitude. However, in 1599, Edward Wright found a precise mathematical description of how to space out the lines; he found that the spacing depended on the area under the secant function. He didn't know how to precisely compute this area, but he was able to approximate it.

Later in the 1640's, Henry Bond looked at a table of these approximate areas and a table of logarithms of trigonometic functions. He noticed a similarity in the two tables, and he was able to conjecture a precise formula for the area under the secant function. We now know that his conjecture was correct, but at the time there was no proof beyond numerical tables.

A proof was later given by Isaac Barrow; this proof is the earliest known publication of the use of integration by partial fractions.

### The Problem

Compute the integral

$$\int_0^x \sec(u)\,du. \tag{71}$$

**The Solution**

Recall that $\sec(u) = \frac{1}{\cos(u)}$. First, let's use algebraic manipulation combined with the trigonometric formula $\cos^2(u) + \sin^2(u) = 1$.

$$\int_0^x \frac{1}{\cos(u)}\,du = \int_0^x \frac{\cos(u)}{\cos^2(u)}\,du, \tag{72}$$

$$= \int_0^x \frac{\cos(u)}{1-\sin^2(u)}\,du. \tag{73}$$

Now, we do a $u$-substitution. However, we are already use the variable $u$, so let's make it a "$w$-substitution". We use $w = \sin(u)$, and so $dw = \cos(u)\,du$. Then we have that our integral is:

$$\int_0^{\sin(x)} \frac{1}{1-w^2}\,dw. \tag{74}$$

Now, we use partial fractions:

$$\frac{1}{1-w^2} = \frac{1}{(1-w)(1+w)}, \tag{75}$$

$$= \frac{A}{1-w} + \frac{B}{1+w}. \tag{76}$$

Combining terms and comparing numerators, we get $A+B+(A-B)w = 1$. So we have

$$\begin{cases} A+B=1, \\ A-B=0. \end{cases} \tag{77}$$

Solving we get $A = B = \frac{1}{2}$.

Therefore, our integral becomes

$$\int_0^{\sin(x)} \frac{1}{2(1-w)} + \frac{1}{2(1+w)}\,dw = \frac{1}{2}\log\left(\frac{1+w}{1-w}\right)\Bigg|_0^{\sin(x)}, \tag{78}$$

$$= \frac{1}{2}\log\left(\frac{1+\sin(x)}{1-\sin(x)}\right). \tag{79}$$

To simplify things, we can now use some trigonometric identities.

$$\frac{1}{2}\log\left(\frac{1+\sin(x)}{1-\sin(x)}\right) = \log\sqrt{\frac{1+\sin(x)}{1-\sin(x)}}, \tag{80}$$

$$= \log\sqrt{\frac{(1+\sin(x))^2}{1-\sin^2(x)}}, \tag{81}$$

$$= \log\left(\frac{1+\sin(x)}{\cos(x)}\right), \tag{82}$$

$$= \log(\sec(x)+\tan(x)). \tag{83}$$

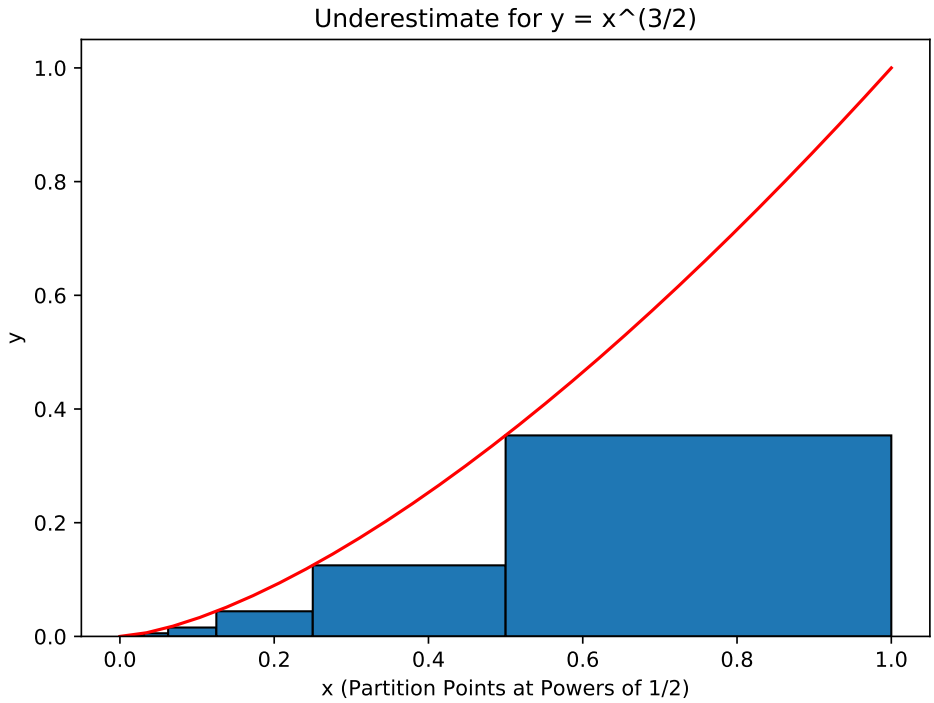
## 4.2 Fermat's Method of Integrating Powers of $x$

### Motivation

Consider the problem of finding the area underneath the curve for a particular power of $x$; here we will concentrate on the particular case of $y = x^{3/2}$ and $0 \leq x \leq 1$. Note that the restriction of $x$ to $x \geq 0$ keeps everything well-defined within the realm of real numbers.

Before Leibniz and Newton developed the use of integral calculus to find the area under curves, Fermat had already developed a method to solve this particular problem. Let us put his idea into modern terms. His idea is to use a method of exhaustion to give lower bounds and upper bounds for the area. In particular, the key to his idea is that we will use rectangles whose widths are not uniform. In fact, their widths decrease geometrically.

For example, on the interval $0 \leq x \leq 1$, one can bound the area above and below by the area of an infinite number of rectangles whose widths are the powers of $1/2$. To see this, consider the graph below for the lowerbound.




The behavior of the rectangles is dictated by the following pattern:

- The rectangle of width $1/2$ is from $1/2 \leq x \leq 1$.
- The rectangle of width $1/4$ is from $1/4 \leq x \leq 1/2$.
- The rectangle of width $1/8$ is from $1/8 \leq x \leq 1/4$.
- Etc...

The amazing consequence of Fermat's idea is that we can actually compute the area of these infinite number of rectangles since their areas turn out to be a geometric series. Recall that a geometric sum is of the form $b + b^2 + ... + b^n$ for some base $b$ and power $n$; this can be expressed more explicitly as

$$b + b^2 + ... + b^n = \frac{b - b^{n+1}}{1 - b}. \tag{84}$$

So for bases $b$ satisfying $|b| < 1$, we get an expression for the infinite series:

$$b + b^2 + b^3 + ... = \frac{b}{1 - b}. \tag{85}$$

Let us see how this is related to the area of the rectangles described above.

The area of the rectangles are:

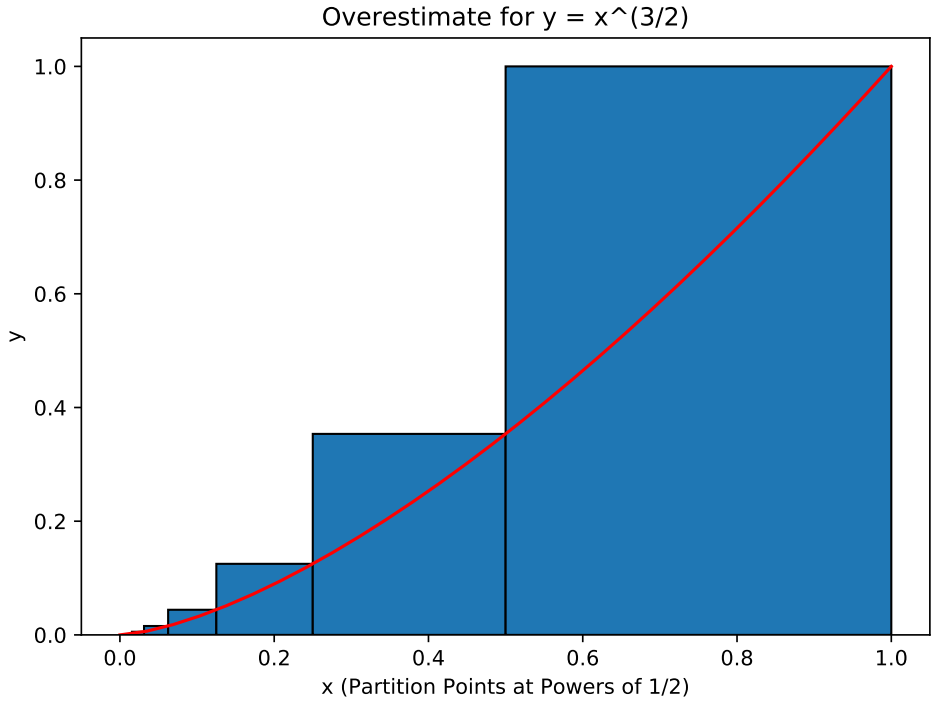
- The area of the rectangle on $1/2 \leq x \leq 1$ is $(1/2)^{3/2}(1/2) = (1/2)^{5/2}$.
- The area of the rectangle on $1/4 \leq x \leq 1/2$ is $(1/4)^{3/2}(1/4) = (1/4)^{5/2}$.
- The area of the rectangle on $1/8 \leq x \leq 1/4$ is $(1/8)^{3/2}(1/8) = (1/8)^{5/2}$.
- Etc.

So we have that the total area is

$$(1/2)^{5/2} + (1/4)^{5/2} + (1/8)^{5/2} + ... = (1/2)^{\frac{5}{2}} + \left((1/2)^{\frac{5}{2}}\right)^2 + \left((1/2)^{\frac{5}{2}}\right)^3 + ... \quad (86)$$

So we have an infinite geometric series with base $b = (1/2)^{\frac{5}{2}}$. Therefore, this set of an infinite rectangles gives us a lowerbound of $\frac{(1/2)^{5/2}}{1-(1/2)^{5/2}}$ for the area under $y = x^{3/2}$ and $0 \leq x \leq 1$.

A similar argument can be applied to another set of rectangles whose widths are powers of $1/2$, as pictured below:




The final part of Fermat's idea is to do the above for general base $0 < b < 1$. Then we consider the limit as $b \to 1$.

### The Problem

Use Fermat's method to find the area under the curve $y = x^{\frac{3}{2}}$ and $0 \leq x \leq 1$.

**The Solution**

Let us first construct the lower bound using rectangles whose widths are powers of a fixed base $b$ satisfying $0 < b < 1$. Similar to the case of base $1/2$ we discussed before, we have that the rectangles satisfy

- There is a rectangle on $b \leq x \leq 1$ of height $b^{\frac{3}{2}}$.
- There is a rectangle on $b^2 \leq x \leq b$ of height $\left(b^2\right)^{\frac{3}{2}}$.
- There is a rectangle on $b^3 \leq x \leq b^2$ of height $\left(b^3\right)^{\frac{3}{2}}$.
- Etc...

These rectangles have areas (repectively) $(1-b)b^{\frac{3}{2}}$, $(1-b)b\left(b^{\frac{3}{2}}\right)^2$, $(1-b)b^2\left(b^{\frac{5}{2}}\right)^3$, ... So we see that the total area of these rectangles (and hence our lowerbound) is:

$$(1-b)b^{\frac{3}{2}} + (1-b)b\left(b^{\frac{3}{2}}\right)^2 + (1-b)b^2\left(b^{\frac{5}{2}}\right)^3 + ... = \frac{1-b}{b}\left(b^{\frac{5}{2}} + \left(b^{\frac{5}{2}}\right)^2 + \left(b^{\frac{5}{2}}\right)^2 + ...\right), \tag{87}$$

$$= \frac{1-b}{b}\frac{b^{\frac{5}{2}}}{1-b^{\frac{5}{2}}}, \tag{88}$$

$$= b^{\frac{3}{2}}\frac{1-b}{1-b^{\frac{5}{2}}}. \tag{89}$$

Now we use the algebraic fact that we can factor $1-y^2 = (1-y)(1+y)$ and $1-y^5 = (1-y)(1+y+y^2+y^3+y^4)$ for $y = b^{\frac{1}{2}}$ to get that the lower bound is given by

$$b^{\frac{3}{2}}\frac{1+b^{\frac{1}{2}}}{1+b^{\frac{1}{2}}+b^{\frac{2}{2}}+b^{\frac{3}{2}}+b^{\frac{4}{2}}}. \tag{90}$$

This lower bound is valid for all $0 < b < 1$. So taking the limit as $b \to 1$ we get a lower bound of

$$\frac{1+1}{1+1+1+1+1} = \frac{2}{5}. \tag{91}$$

We can use a similiar process to find an upper bound of $\frac{2}{5}$. Since the upper bound and lower bound are the same, we must have that they are exaclty equal to the area.

Therefore, the area under $y = x^{3/2}$ from $0 \leq x \leq 1$ is $\frac{2}{5}$.

## 4.3 Cavalieri's Quadrature of $x^n$

In this example we consider Cavalieri's method of finding the area under $x^n$ for positive integers $n$. In particular, we will use a couple simple integral properties,

which may be simply proven using a method of exhaustion, to create a system of linear equations that will allow us to compute

$$\int_0^a x^n dx, \tag{92}$$

for $a > 0$ and any positive integer $n$. Actually we will create an algorithm for finding the value of such an integral for a given value of $n$ ($a$ will still be general). The author is unaware of how mathematical induction may be applied to find the general formula for general $n$ and general $a$; the equations seem to be too complicated to be immediately tractable.

**The Setup**

First, an introduction to Cavalieri's idea. Struik [14] gives a nice discussion of Cavalieri's methods with more modern notation for $n = 2, 3$. Cavalieri uses a notion of "summing over all lines", "summing over all squares,", "summing over all cubes", etc; for example, he is comfortable with writing in words the equivalent of $\sum x^2$, where the sum is considered to be over all x-values for $0 \leq x \leq a$. He does not consider details of convergence or whether such a sum is well-defined. This method is very non-rigorous by modern standards, and we will update his arguments; when necessary, we will prove the integral properties we need using an exhaustion by Riemann sums.

Furthermore, the argument in Struik [14] doesn't precisely match ours, but the author feels our argument amounts to an induction step that is easier to analyze. The difference being that we will induct over set of integrals depending on $a$ and $n$, instead of just the values of $\int_0^a x^n dx$. Next we introduce the set of integrals we need for our induction.

We introduce another variable $y$ such that $x + y = a$; here Cavalieri is imagining breaking the line segment of length $a$ into two parts. Then the variables $x$ and $y$ represent their lengths, which must add up to $a$. We will be interested in the following integrals

$$I_{p,q}^n \coloneqq \int_0^a x^p y^q dx, \tag{93}$$

where $p, q$ are non-negative integers such that $p+q = n$. Note that our integrals may be rewritten entirely in terms of $x$ as $I_{p,q}^n = \int_0^a x^p (a-x)^q dx$, but we will find it more convenient to express them in terms of $x$ and $y$.

Next, we describe our induction step. We will assume that we know that values of $I_{p,q}^m$ for all $m < n$ and $p, q$ non-negative integers with $p+q = m$. Then we will show that we can compute $I_{p,q}^n$ for all $p, q$ non-negative integers with $p + q = n$.

Cavalieri's idea can be broken into several parts:

1. There is a symmetry coming from interchanging the roles of $x$ and $y$, i.e. $I_{p,q}^n = I_{q,p}^n$. This implies that there are actually less than $n+1$ unknowns.

They can be expressed in the form $I^n_{p,q}$ for $p, q$ non-negative integers with $p + q = n$.

In the case of even $n = 2k$, we only have the $k+1$ unknowns $I^{2k}_{0,2k}, ..., I^{2k}_{k,k}$. In the case of odd $n = 2k + 1$, we have only have the $k + 1$ unknowns $I^{2k+1}_{0,2k+1}, ..., I^{2k+1}_{k,k+1}$.

The symmetry may be proven using a method of exhaustion. Take a uniform partition of $[0, a]$ with an even number of intervals $2P$ of length $h = a/2P$. Let $x_i$ be the center of each partition interval. We approximate the integral using the sum

$$J^n_{p,q,P} \coloneqq \sum_1^{2P} x_i^p y_i^q h, \tag{94}$$

and we have that $I^n_{p,q} = \lim_{P\to\infty} J^n_{p,q,2P}$. By the symmetry of our partition and choice of $x_i$, we have the symmetry $J^n_{p,q,P} = J^n_{q,p,P}$. Taking the limit we get that $I^n_{p,q} = I^n_{q,p}$.

2. The next part is to create recurrence relations for the $I^n_{p,q}$ for the induction step using

$$aI^{n-1}_{p,q} = \int_0^a (y + x)x^p y^q dx \tag{95}$$

$$= I^n_{p,q+1} + I^n_{p+1,q}, \tag{96}$$

where $p + q = n - 1$.

We see that for the case of even $n = 2k$, we have the $k$ equations in the $k + 1$ unknowns:

$$aI^{2k-1}_{0,2k-1} = I^{2k}_{0,2k} + I^{2k}_{1,2k-1}, \tag{97}$$

$$... \tag{98}$$

$$aI^{2k-1}_{k-2,k+1} = I^{2k}_{k-2,k+2} + I^{2k}_{k-1,k+1}, \tag{99}$$

$$aI^{2k-1}_{k-1,k} = I^{2k}_{k-1,k+1} + I^{2k}_{k,k}. \tag{100}$$

Note that this is an underdetermined system, and we will need another equation. In the next part we will obtain another equation for this case.

For the case of odd $n = 2k + 1$, we have the $k + 1$ equations in the $k + 1$ unknowns:

$$aI^{2k}_{0,2k} = I^{2k+1}_{0,2k+1} + I^{2k+1}_{1,2k}, \tag{101}$$

$$... \tag{102}$$

$$aI^{2k}_{k-1,k+1} = I^{2k+1}_{k-1,k+2} + I^{2k+1}_{k,k+1}, \tag{103}$$

$$aI^{2k}_{k,k} = 2I^{2k+1}_{k,k+1}. \tag{104}$$

We can see that this system is actually solvable; the last equation allows us to solve for $I_{k,k+1}^{2k+1}$, and then we work backward substitute in equations to solve for the rest of the quantities one by one.

3. For the final part we need to find another equation for the case of even $n = 2k$. Cavalieri's idea involves a scaling argument to find a relation between $I_{0,2k}^{2k}$ and $I_{k,k}^{2k}$.

   First, introduce a variable $z$ such that $x = a/2 + z$ and $y = a/2 - z$; that is, $z$ represents the offset of $x$ from $a/2$.

   Then we have that

$$I_{k,k}^{2k} = \int_0^a x^k y^k dx = \int_0^a (a^2/4 - z^2)^k dx. \tag{105}$$

   An exhaustion argument can be used to easily show $\int_0^a (f \pm g)dx = \int_0^a f dx \pm \int_0^a g dx$, especially in the case that $f$ is non-negative and $g$ is non-positive.

   Then, we have that

$$I_{k,k}^{2k} = \sum_{l=0}^{k} \binom{k}{l} (-1)^l (a/2)^{2l} \int_0^a z^{2(k-l)} dx. \tag{106}$$

   Now we claim that $J^{2p} \coloneqq \int_0^a z^{2p} dx = (1/2)^{2p} \int_0^a x^{2p} dx$; this can be proven by breaking the integral into pieces and applying a scaling argument.

   First note that

$$J^{2p} = \int_0^{a/2} z^{2p} dx + \int_{a/2}^a z^{2p} dx, \tag{107}$$

$$\coloneqq J_1^{2p} + J_2^{2p} \tag{108}$$

   One can use an exhaustion argument to easily show that an integral is translation invariant. Next, not that $z = x - a/2$, so that $z$ is the translation of $x$ by $a/2$. Therefore, we have that $J_2^{2p} = \int_0^{a/2} x^{2p} dx$.

   Similarly we have that $J_1^{2p} = \int_{-a/2}^0 x^{2p} dx$; since $x^{2p}$ is an even power, then it is also an even function. Therefore $J_1^{2p} = \int_0^{a/2} x^{2p} dx$.

   Therefore we have that $J^{2p} = 2 \int_0^{a/2} x^{2p} dx$.

   Next we claim that $\int_0^{a/2} x^{2p} dx = (1/2)^{2p+1} \int_0^a x^p dx$. To prove the we apply a scaling argument to an exhaustion of $\int_0^{a/2} x^{2p} dx$. We partition the interval $[0, a/2]$ into $Q$ uniform intervals of length $h_Q = a/2Q$ and $x_i$ be the left endpoints of each interval.

Then we have that $J_Q^{2p} := \sum_i x_i^{2p} h_Q$. Next note that $x_i = (1/2)\tilde{x}_i$ for points $\tilde{x}_i$ of a partition of $[0, a]$. Similarly $h_{2Q} = (1/2)\tilde{h}_Q$, where $\tilde{h}_Q$ is the length of the corresponding uniform partition of $[0, a]$. Therefore, we have that $J_Q^{2p} = (1/2)^{2p+1} \sum_i \tilde{x}_i^{2p} \tilde{h}_Q$, a scalar multiple of an exhaustion for $\int_0^a x^{2p} dx$. Therefore we get that

$$\int_0^a z^{2p} dx = (1/2)^{2p} \int_0^a x^{2p} dx, \tag{109}$$

$$= (1/2)^{2p} I_{2p,0}^{2p}. \tag{110}$$

So we have that

$$I_{k,k}^{2k} = \sum_{l=0}^{k} \binom{k}{l} (-1)^l (a/2)^{2l} (1/2)^{2(k-l)} I_{2(k-l),0}^{2(k-l)}, \tag{111}$$

and we get the equation

$$(1/2)^{2k} \sum_{l=1}^{k} \binom{k}{l} (-1)^{l+1} a^{2l} I_{2(k-l),0}^{2(k-l)} = (1/2)^{2k} I_{2k,0}^{2k} - I_{k,k}^{2k}, \tag{112}$$

where the left hand side is known. It is possible to see that after including this system, we get $k+1$ equations in the $k+1$ unknowns system for $I_{0,2k}^{2k}, ..., I_{k,k}^{2k}$, and this system is in fact solvable (it is easier this last fact using the matrix structure).

Note, you may be tempted to introduce an equation for $I_{k,k}^{2k}$ by using that $a^{2k+1} = \int_0^a (x+y)^{2k} dx$ and binomial expanding. However, this equation is a consequence of the $k$ equations we already found. Namely we have that $a^{2k+1} = a \int_0^a (x+y)^{2k-1} dx = a \sum_{p=0}^{2k-1} \binom{2k-1}{p} I_{p,2k-1-p}^{2k-1}$, and then we can use that $\binom{2k}{p} = \binom{2k-1}{p} + \binom{2k-1}{p-1}$ to relate the two binomial expansions in each equation. So trying to use the binomial expansion will not give us anything new.

It is noteworthy that amazingly we may solve for the case of odd $n$ by just solving a linear system that it created from very simple knowledge of integrals (i.e. no need to do anything complicated like the $z$ trick used in the even case).

Let us demonstrate how to use Cavalieri's ideas to compute the cases of $n = 2, 3$.

**Example for the Case of $n = 2$**

First note that $\int_0^a x dx = \int_0^a y dy = a^2/2$ follows from basic geometry. So we already know the base case of $n = 1$. Our unknowns are $I_{0,2}^2$ and $I_{1,1}^2$.

We start by writing down the recurrence relation

$$a^3/2 = I_{0,2}^2 + I_{1,1}^2. \tag{113}$$

Next we expand

$$I^2_{1,1} = \int_0^a (a^2/4 - z^2)dx, \tag{114}$$

$$= a^3/4 - (1/4)I^2_{0,2}. \tag{115}$$

We can solve these equations to get

$$I^2_{0,2} = a^3/3, \tag{116}$$

$$I^2_{1,1} = a^3/6. \tag{117}$$

**Example for the Case of $n = 3$**

We start by writing the recurrence relations

$$aI^2_{0,2} = a^4/3 = I^3_{0,3} + I^3_{1,2}, \tag{118}$$

$$aI^2_{1,1} = a^4/6 = 2I^3_{1,2}. \tag{119}$$

We see that we can directly solve the last equation for $I^3_{1,2} = a^4/12$, and then we can solve the first equation for $I^3_{0,3} = a^4/4$.

**The Problem**

Use Cavalieri's ideas to compute the next two cases of $n = 4, 5$.

**The Solution**

Let us first compute the case of $n = 4$. We need to compute the integrals $I^4_{0,4}, I^4_{1,3}, I^4_{2,2}$. We start by writing down the recurrence relations

$$aI^3_{0,3} = a^5/4 = I^4_{0,4} + I^4_{1,3}, \tag{120}$$

$$aI^3_{1,2} = a^5/12 = I^4_{1,3} + I^4_{2,2}. \tag{121}$$

Next we expand

$$I^4_{2,2} = \int_0^a (a^2/4 - z^2)^2 dx, \tag{122}$$

$$= a^5/16 - (a^2/2)\int_0^a z^2 dx + \int_0^a z^4 dx, \tag{123}$$

$$= a^5/16 - a^5/24 + (1/16)I^4_{0,4}, \tag{124}$$

$$= a^5/48 + (1/16)I^4_{0,4}. \tag{125}$$

So we have three equations in three unknowns, and this system is in fact solvable. We have $I^4_{0,4} = a^5/5$, $I^4_{1,3} = a^5/20$, and $I^4_{2,2} = a^5/30$.

Next we consider the case of $n = 5$. We still have only three unknowns: $I^5_{0,5}, I^5_{1,4}$, and $I^5_{2,3}$. Let us write down the recurrence relations:

$$aI^4_{0,4} = a^6/5 = I^5_{0,5} + I^5_{1,4}, \tag{126}$$

$$aI^4_{1,3} = a^6/20 = I^5_{1,4} + I^5_{2,3}, \tag{127}$$

$$aI^4_{2,2} = a^6/30 = 2I^5 2, 3. \tag{128}$$

We can solve this system for our unknown integrals (no use of $z$ necesssary). We obtain $I^5_{0,5} = a^6/6$, $I^5_{1,4} = a^6/30$, and $I^5_{2,3} = a^6/60$.

## 4.4 Quadrature of the Hyperbola and Logarithms

Now we investigate the relationship between the area under the graph of the hyperbola $y = 1/x$ and logarithms.

### The Setup

John Napier created the logarithm with the purpose of aiding in computations of multiplication and division. [2] He noted that comparing an arithmetic progression to a geometric progression would allow one to, e.g., switch from division to subtraction. By having a logarithm table, one can make division easier by converting to the logarithm using the table, performing subtraction, and then using the logarithm table to convert back.

So the original intention of logarithms is to take advantage of the algebraic fact that

$$\log(ab) = \log(a) + \log(b). \tag{129}$$

The relationship between logarithms and the area under the hyperbola $y = 1/x$ was established by the work of Gregory St. Vincent in 1647 and Alfons Anton de Sarasa in 1649 (note that their work is before the invention of calculus of Newton and Leibniz). Using modern terminology, we seek to show that

$$f(x) \equiv \int_1^x \frac{1}{t} dt, \tag{130}$$

satisfies the algebraic rule of algorithms $f(ab) = f(a) + f(b)$. Since this work predates the invention of calculus, we will prove this result directly using methods of exhaustion, a technique that also predates the invention of calculus.

### The Problem

1. Use a method of exhaustion to show that for any $a, b, c > 0$, one has that

$$\int_{ac}^{bc} \frac{1}{t} dt = \int_a^b \frac{1}{t} dt. \tag{131}$$

2. Prove that
$$\int_1^{ab} \frac{1}{t} dt = \int_1^{a} \frac{1}{t} dt + \int_1^{b} \frac{1}{t} dt. \tag{132}$$

**The Solution**

1. First we partition $[ac, bc]$ using the points $t_i = ac + \frac{bc-ac}{N} i$, for $0 \leq i \leq N$. Next, note that the points $s_i = a + \frac{b-a}{N} i$ for $0 \leq i \leq N$ give a partition of the interval $[a, b]$.

   Let $U_N$ be the upper sum for the partition of $[ac, bc]$. We have that
$$U_N = \sum_{i=0}^{N-1} \frac{1}{t_i} \frac{bc-ac}{N}, \tag{133}$$
$$= \sum_{i=0}^{N-1} \frac{1}{a + \frac{b-a}{N} i} \frac{b-a}{N}, \tag{134}$$
$$= \sum_{i=0}^{N-1} \frac{1}{s_i} \frac{b-a}{N}, \tag{135}$$
$$\tag{136}$$

   Let us relate this to $\int_a^b 1/t dt$. Now, we see that $U_N$ isn't quite a lower bound for this integral. However, we do have that it gives a lower bound for a slightly different interval. Let $\delta_N = \frac{b-a}{N}$, we have that
$$U_N < \int_{a-\delta_N}^{b-\delta_N} \frac{1}{t} dt. \tag{137}$$

   Similarly let the lower sum be $L_N$; we have that
$$L_N = \sum_{i=0}^{N-1} \frac{1}{t_{i+1}} \frac{bc-ac}{N}, \tag{138}$$
$$= \sum_{i=0}^{N-1} \frac{1}{a + \frac{b-a}{N}(i+1)} \frac{b-a}{N}, \tag{139}$$
$$= \sum_{i=0}^{N-1} \frac{1}{s_{i+1}} \frac{b-a}{N}, \tag{140}$$

   We also see that
$$L_N > \int_{a+\delta_N}^{b+\delta_N} \frac{1}{t} dt. \tag{141}$$

So we have that

$$\int_{a+\delta_N}^{b+\delta_N} \frac{1}{t}dt < \int_{ac}^{bc} \frac{1}{t}dt < \int_{a-\delta_N}^{b-\delta_N} \frac{1}{t}dt. \tag{142}$$

As we let $N \to \infty$, we have that $\delta_N \to 0$. Therefore, we get that

$$\int_{a}^{b} \frac{1}{t}dt \leq \int_{ac}^{bc} \frac{1}{t}dt \leq \int_{a}^{b} \frac{1}{t}dt. \tag{143}$$

Therefore,

$$\int_{ac}^{bc} \frac{1}{t}dt = \int_{a}^{b} \frac{1}{t}dt. \tag{144}$$

2. Next, consider the area

$$\int_{1}^{ab} \frac{1}{t}dt. \tag{145}$$

From our previous result, we have that

$$\int_{a}^{ab} \frac{1}{t}dt = \int_{1}^{} b\frac{1}{t}dt. \tag{146}$$

Next, note that

$$\int_{a}^{ab} \frac{1}{t}dt = \int_{1}^{ab} \frac{1}{t}dt - \int_{1}^{a} \frac{1}{t}dt. \tag{147}$$

Therefore, we get that

$$\int_{1}^{ab} \frac{1}{t}dt = \int_{1}^{a} \frac{1}{t}dt + \int_{1}^{b} \frac{1}{t}dt. \tag{148}$$

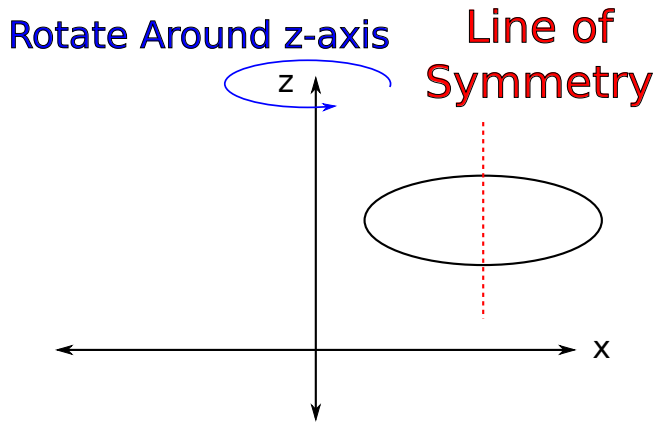
## 4.5 Volume of Particular Tori by Exhaustion

In this section we look at using a method of exhaustion/rearrangement to evaluate a torus created from rotating a 2-dimensional cross-section that has a vertical line of symmetry.

### The Setup

The volume of a solid of revolution is provided by the classic Pappus' Centroid Theorem. The theorem is mainly attributed to Pappus of Alexandria and Paul Guldin. Now, Pappus is a mathematician of ancient Greece, and Guldin wrote his manuscript containing the theorem, *De centro gravitatis trium specierum quanitatis continuae* in 1640 (two years before Newton was born). Typically this theorem is presented using integral calculus. However, given that the creators of the theorem predate the invention of integral calculus, we are lead to wonder if we can prove the theorem more directly using a method of exhaustion.

Here we will prove the theorem in the case that the cross-section of the revolution has a vertical line of symmetry, much like in the figure below.




### The Problem

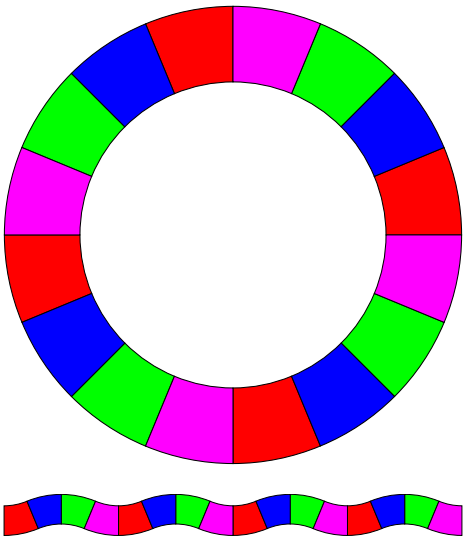
Given that the cross-section of the body of revolution has a vertical line of symmetry, find a method of exhaustion/rearrangement that will compute its volume.

### The Solution

We divide the body of revolution into $4n$ pieces by the angle of revolution, i.e. a number of pieces that is a multiple of four. For example, in the figure below is a picture of such a partition from above.

Since the cross-section has a vertical line of symmetry, if we take any piece and flip it to rotate in the opposite direction, the ends of that piece can still be made to line-up with the previous piece. Doing so, we can rearrange the pieces to "snake" back and forth. We use a number of pieces that are a multiple of four, because this way the endpoints of the "snake" will be aligned. See the figure below,

Now, intuitively, as we make the number of pieces larger, the "snake" gets closer to a straight cylinder. We see that each side of the cylinder is made of an



equal number of pieces of arcs coming from the "outer" circle of revolution as the "inner" circle of revolution. Let these radii be $r_1$ and $r_2$. Then, intuitively we should have that sides of the cylinder are $2\pi(r_1 + r_2)/2$.

Let us prove this more rigorously.

Now, for each arc coming from the outer radius $r_2$, from elementary geometry we have that the width of the arc (NOT the length) is exactly

$$w_{\text{piece}} = 2r_2 \sin\left(\frac{\pi}{4n}\right). \tag{149}$$

Similarly for $r_1$.

So we get that the total width of the snake is

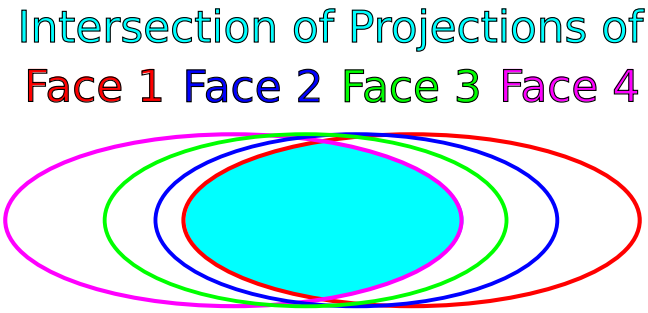
$$w_n = 4nr_1 \sin\left(\frac{\pi}{4n}\right) + 4nr_2 \sin\left(\frac{\pi}{4n}\right). \tag{150}$$

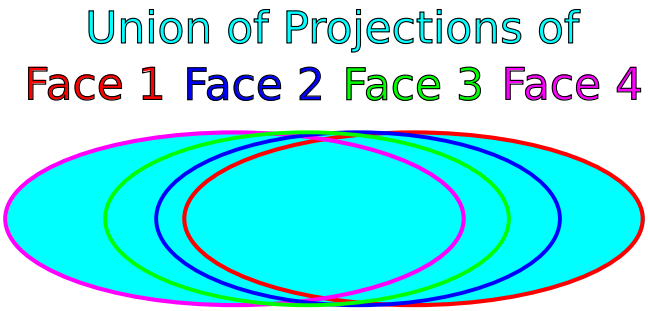
So $w_n \to \pi r_1 + \pi r_2$. This is equal to the circumference of the circle formed by the centroid of the cross-section (the symmetry of the cross-section gives us that the centroid has radius $(r_1 + r_2)/2$).

Technically, this last limit used that $\frac{\sin(\theta)}{\theta} \to 1$ as $\theta \to 0$. This doesn't need to the full force of calculus to be shown as it follows from the classical argument of comparing areas of triangles to areas of circular regions.

Now, we need lower and upper bounds on the volume of each rearrangement depending on $n$. Look at four consecutive segments of the "snake" and project their faces onto the plane of the base of the "snake". We can use the areas of the intersections and the unions of these projections to give us lower and upper bounds on the volume of the "snake" (which is equal to the volume of the revolution). This is due to the face that snake is between the cylinder formed by the intersection of the proejections and the cylinder formed by the union of the projections.

See the figures below for the pictures of the intersections and unions of the projections.




Let the area of the intersection be $I_n$ and the area of the union by $U_n$. So we see that the volume $V$ of the body of revolution is bounded by

$$I_n w_n \leq V \leq U_n w_n. \tag{151}$$

How it should be clear that are the number of pieces gets larger, the areas $I_n$ and $U_n$ get closer to the area of the original cross section $A$. Furthermore, we already showed that $w_n \to 2\pi r_{\text{centroid}}$. So from the pinching theorem, we get that

$$V = 2\pi r_{\text{centroid}} A. \tag{152}$$

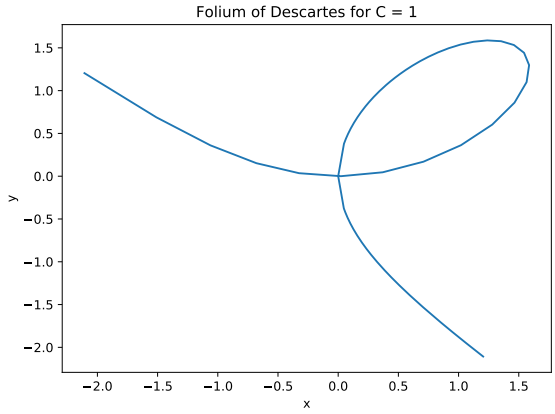
## 4.6 Fermat's Quadrature of the Folium of Descartes

### The Setup

In this section we investigate how Fermat was able to use integration by parts to find the area of the loop in the Folium of Descartes. A reference for Fermat's treatment of this problem is given by [11].

The Folium of Descartes is the curve given by

$$x^3 + y^3 = 3Cxy, \tag{153}$$




where $C$ is a constant. A graph of the curve for $C = 1$ is given below. The curve crosses itself at the origin.

Let $\bar{x}$ be the x-coordinate of the right-most part of the loop; so the loop can be written as the graph of two functions $y_1 \leq y_2$ on $[0, \bar{x}]$.

Fermat's idea comes in two parts:

1. If we can make a substitution for a new variable $z$ to replace $y$ given by $y = C^{1-m-n}x^m z^n$ such that the equation now becomes

$$x^p = \sum_i D_i^{p-p_i} z^{p_i}, \tag{154}$$

then we know how to compute the integral

$$\int x^p dz. \tag{155}$$

A note here about the choice of $C^{1-m-n}$, Fermat likes to follow the rules of homogeneity, which is probably best explained in terms of units. If $x$ and $y$ were distances in say meters, then the left hand side of $x^3 + y^3 = 3Cxy$ has units meters$^3$. Fermat chooses a constant $C$ that is also in meters, and so the right hand side also has units meters$^3$. Similarly, the left and right hand sides of $y = C^{1-m-n}x^m z^n$ have units of just meters.

2. Use integration by parts to write the original area integral in terms of the integral we now know how to compute. Let $\bar{z}$ be the largest value of the $z$-coordinate obtained by the loop (we will in fact see that $\bar{z} = \infty$). So we wish to try to find $p$ and $p_i$ to ensure that

$$\int_0^{\bar{x}} (y_2 - y_1)dx = \int_0^{\bar{z}} x^p dz. \tag{156}$$

It may be that $\bar{z} = \pm\infty$, and so to be rigorous we will need to take an improper integral.

**The Problem**

Do the two steps of Fermat's plan to evaluate the area of the loop in the Folium of Descartes.

**The Solution**

1. Let's do step one; let's make the substitution $y = C^{1-m-n}x^m z^n$ into $x^3 + y^3 = 3Cxy$. We get

$$x^3 + C^{3-3m-3n}x^{3m}z^{3n} = 3C^{2-m-n}x^{1+m}z^n. \tag{157}$$

Let's us now try to consolidate the $x$-terms into one term $x^p$; to do so, two of the terms need to have the same power of $x$. Let's look at the three cases:

(a) If we try equating $x^3 = x^{3m}$, then we must have $m = 1$. So we get

$$1 + C^{-3n}z^{3n} = 3C^{1-n}x^{-1}z^n. \tag{158}$$

Consolidating the z-terms, this becomes

$$3C^{1-n}x^{-1} = z^{-n} + C^{-3n}z^{2n}. \tag{159}$$

The problem here is that there is no way to produce the power $x^{-1}$ using integration by parts without using a logarithm. So this case isn't really any help.

(b) Now let's try equating $x^{3m} = x^{1+m}$. We then have that $m = \frac{1}{2}$. So we get

$$x^{3/2} = -C^{3/2-3n}z^{3n} + 3C^{3/2-n}z^n. \tag{160}$$

Again, we can't directly get $x^{3/2}$ without introducing some radical power of $x$ that isn't already present. We could just square the equation, but we will see that it is better to work with the last case.

(c) Now we equate $x^3 = x^{1+m}$ and we get $m = 2$. So then we have that

$$1 + C^{-3-3n}x^3z^{3n} = 3C^{-n}z^n, \tag{161}$$

which gives us

$$x^3 = 3C^{3+2n}z^{-2n} - C^{3+3n}z^{-3n}. \tag{162}$$

The power $x^3$ is a nice integer power for us to work with.

So we take $y = C^{-1-n}x^2z^n$ where $n$ is still left to be determined. Our equation defining the folium is now

$$x^3 = 3C^{3+2n}z^{-2n} - C^{3+3n}z^{-3n}. \tag{163}$$

2. Let us now consider the area integral

$$\int_0^a (y_2 - y_1)dx = C^{-1-n} \int_0^a (z_2^n - z_1^n)x^2 dx \tag{164}$$

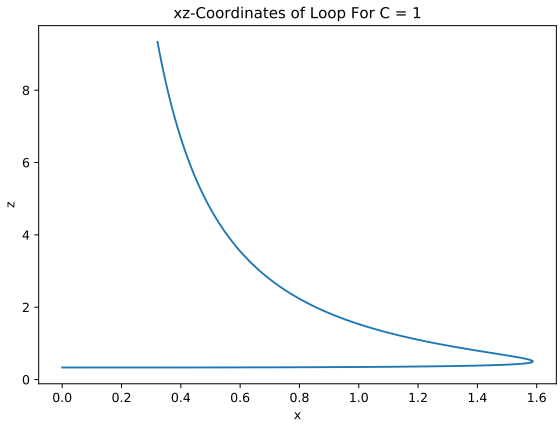
When we perform integration by parts, we will get an integral of the form (ignoring the boundary terms for now)

$$\frac{n}{3} \int z^{n-1} x^3 dz. \tag{165}$$

This is most simple when $n = 1$. So we will take $y = C^{-2}x^2 z$ for our variable transformation. The equation of the folium is now

$$x^3 = 3C^5 z^{-2} - C^6 z^{-3}. \tag{166}$$

Below is a picture of the Folium of Descartes in $xz$-coordinates.




Now we need to be more rigorous with our boundary terms.

First note that the loop of the folium of descartes is in the first quadrant, i.e. $\{x, y \geq 0\}$. Furthermore, the only non-negative solutions to $3C^5 z^{-2} - C^6 z^{-3} = 0$ are $z = \frac{C}{3}$ and $z = \infty$.

Next, note that for $x^3 + y^3 = 3Cxy$ and $x, y$ close to the origin, the $3Cxy$ term dominates. So we expect the loop to be asymptotically like $3xy = 0$. That is, one direction of approach to the origin is horizontal while the other is vertical. We need to investigate this asymptoticness more closely.

For any point $(x, y)$ on the folium and close to the origin, let $\lambda = \|(x, y)\|$ and let $(\tilde{x}, \tilde{y}) = \frac{1}{\lambda}(x, y)$. Using the equation for the folium, we have that

$$0 \leq \tilde{x}\tilde{y} \leq \frac{\lambda}{3C}. \tag{167}$$

Note that $(\tilde{x}, \tilde{y})$ occur on the unit circle. Consider the function $f(x, y) = xy$ and its parameterization on the unit circle given by $g(\theta) = f(\cos\theta, \sin\theta) = \cos\theta \sin\theta$. Now note that where $g$ vanishes for the first quadrant is $\theta = 0$ or $\theta = \pi/2$. Furthermore, $g'(0) \neq 0$ and $g'(\pi/2) \neq 0$.

We can use this and some basic trignometry to show that $\tilde{x}\tilde{y} \leq \frac{\lambda}{3C}$ gives us that $\|(\tilde{x}, \tilde{y}) - (1, 0)\| \leq K\lambda$ or $\|(\tilde{x}, \tilde{y}) - (0, 1)\| \leq K\lambda$ for some constant $K$ depending on $C$.

The first case is the horizontal part of the loop approaching the origin. In this case, we have that $z = C^2 y/x^2 = C^2 \frac{\tilde{y}}{\lambda \tilde{x}^2}$. Now, as we approach the origin, $\lambda \to 0$, $\tilde{x} \to 1$ and $\tilde{y} \leq K\lambda$. So we get that $z$ is bounded as we approach the origin along the horizontal portion of the loop. So it must approach $z = \frac{C}{3}$.

The second case is the vertical portion of the loop approacing the origin. In this case, we write $z = C^2 \frac{\tilde{y}}{x\tilde{x}}$. We have that $\tilde{x} \to 0$, $x \to 0$, and $\tilde{y} \to 1$. Therefore $z \to \infty$ along the vertical portion of the loop as we approach the origin.

From the $xz$-equation of the folium, we see that in either case that $zx^3 \to 0$ as $x \to 0$. Now let us rigorously apply integration by parts. Let $z_a$ be the z-coordinate of the right-most part of the loop. We have that

$$\int_0^{\bar{x}} (z_2 - z_1)x^2 dx = \lim_{b \to 0+} \int_b^{\bar{x}} (z_2 - z_1) d\left(\frac{x^3}{3}\right). \tag{168}$$

Let us study each integral involving $z_i$ individually. Let $z_a$ be the z-coordinate of the rightmost part of the loop. For the bottom part of the loop, we see that

$$\lim_{b \to 0+} \int_b^{\bar{x}} z_1 d\left(\frac{x^3}{3}\right) = \lim_{b \to 0+} z_1 \frac{x^3}{3}\Big|_b^a - \frac{1}{3} \int_{C/3}^{z_a} x^3 dz, \tag{169}$$

$$= \frac{z_b a^3}{3} - 0 - \frac{1}{3} \int_{C/3}^{z_a} x^3 dz. \tag{170}$$

For the top part of the loop, we see that

$$\lim_{b \to 0+} \int_b^{\bar{x}} z_2 d\left(\frac{x^3}{3}\right) = \lim_{b \to 0+} z_2 \frac{x^3}{3}\Big|_b^a - \frac{1}{3} \int_{\infty}^{z_a} x^3 dz, \tag{171}$$

$$= \frac{z_a a^3}{3} + \frac{1}{3} \int_{z_a}^{\infty} x^3 dz. \tag{172}$$

Putting them together, we get

$$\lim_{b\to 0+}\int_b^{\bar{x}}(z_2-z_1)d\left(\frac{x^3}{3}\right)=\frac{1}{3}\int_{z_a}^{\infty}x^3dz+\frac{1}{3}\int_{C/3}^{z_a}x^3dz, \tag{173}$$

$$=\frac{1}{3}\int_{C/3}^{\infty}x^3dz. \tag{174}$$

Now, we can use the $xz$-coordinate equation for the folium to get that

$$\frac{1}{3}\int_{C/3}^{\infty}x^3dz=\int_{C/3}^{\infty}\left(C^5z^{-2}-\frac{C^6z^{-3}}{3}\right)dz, \tag{175}$$

$$=C^5\frac{3}{C}-C^6\frac{9}{6C^2}, \tag{176}$$

$$=3C^4-\frac{3C^4}{2}, \tag{177}$$

$$=\frac{3C^4}{2}. \tag{178}$$

Combining with the area integral, we get

$$\int_0^{\bar{x}}(y_2-y_1)dx=\frac{3}{2}C^2. \tag{179}$$

# 5 Multivariable Differential Calculus

Here are examples related to differential calculus in more than one variable.

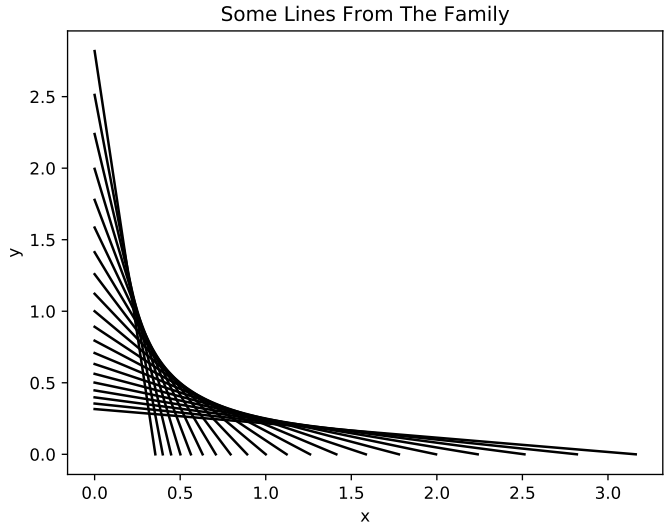
## 5.1 Envelopes

In the simplist cases, the **envelope** of a family $\mathfrak{F}$ of curves is a curve $\gamma$ that is in some sense extremal to the entire family of curves. What is often the case, is that every point of the curve $\gamma$ touches exaclty one curve from the the family $\mathfrak{F}$, and furthermore this touching is only tangential (i.e. they cross at an angle of zero). This is best illustrated with examples.

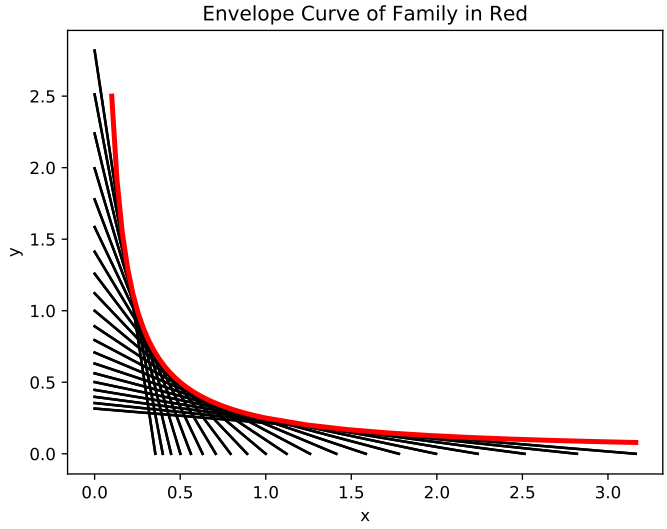
### 5.1.1 The Hyperbola as an Envelope

#### The Set Up

Consider the family $\mathfrak{F}$ of straight lines in $\mathbb{R}^2$, where each line crosses the x-axis and y-axis at pairs of points of the form $(s,0)$ and $(0,1/s)$ for some $s>0$. So we see that each line is of the form $\frac{1}{s}x+sy=1$ for some $s>0$. Some of the lines from the family are pictured in the following figure.




We can see that extemal to the family of lines is a curve concave up in the first quadrant $\{x, y > 0\}$. In the following figure, you can see the curve superimposed with some of the lines from the family.




### The Problem

Let us consider computing the envelope curve $\gamma(x)$ of the family $\mathfrak{F}$.

**The Solution**

To compute the envelope curve $\gamma(x)$, let us consider the auxilliary function $g(x,y,s) = \frac{1}{s}x + sy - 1$. Let us see how the Implicit Function Theorem of vector calculus let's us use $g(x,y,s)$ to find the exremal envelope curve $\gamma(x)$. As we discuss this, please consider the similarities to the ordinary first derivative test.

First, consider any point $(x_0, y_0)$ NOT on the extremal envelope curve $\gamma(x)$, but is touched by some line in $\mathfrak{F}$. So there is some $s_0 > 0$ such that $\frac{1}{s_0}x_0 + s_0 y_0 = 1$; note that this is equivalent to $g(x_0, y_0, s_0) = 0$. Since $(x_0, y_0)$ isn't on the boundary of the region of points touched by lines in $\mathfrak{F}$, we know that for any other points $(x_1, y_1)$ close to $(x_0, y_0)$ we may find another line in $\mathfrak{F}$ touching $(x_1, y_1)$. That is, for every $(x_1, y_1)$ close to $(x_0, y_0)$, we may find $s_1 > 0$ such that $g(x_1, y_1, s_1) = 0$.

This can be summarized as saying that for all points $(x_0, y_0)$ that are touched by a line in $\mathfrak{F}$ and also isn't on the envelope $\gamma$, we can locally solve $s = S(x,y)$ such that $g(x, y, S(x,y)) = 0$. Now, you may begin to see the connection to the Implicit Function Theorem.

Recall that the Implicit Function Theorem can only confirm that we CAN locally solve $s = S(x,y)$ such that $g(x, y, S(x,y)) = 0$. However, we seek for the extremal points where we CAN'T locally solve. This is similar to the first derivative test of ordinary calculus. Technically, the first derivate test only says when a point is NOT an extemum of a function; then the candidate points for extrema are reduced to some finite list by solving for the vanishing of the derivative.

Here, we are in a similar situation. We solve for a set of candidate points that must contain our extremal curve $\gamma$. It will happen to be the case that our candidate set will allow only one curve and so this must be the envelope. However, we are being a little reckless here as we haven't proven the curve must exist; we will consider the picture to be very convincing and ignore this technical detail.

So we seek for when we can't locally solve $s = S(x,y)$ such that $g(x, y, S(x,y)) = 0$. The Implicit Function Theorem tells us this will only be possible for those $(x,y,s)$ with $g(x,y,s) = 0$ and $\frac{\partial g}{\partial s}(x,y,s) = 0$.

So we look for

$$0 = \frac{\partial g}{\partial s}, \tag{180}$$

$$= -\frac{x}{s^2} + y \tag{181}$$

We wish to find an equation restricting $x$ and $y$; so it is most efficient to solve the above for $s$. Also, from the picture it is clear that we should restrict to $x, y > 0$. Therefore, for $x, y > 0$, we have $s = \sqrt{\frac{x}{y}}$. Plugging this into the equation for $g(x,y,s) = 0$, we get

$$\sqrt{xy} + \sqrt{xy} - 1 = 0. \tag{182}$$

Therefore, we find that the envelope curve must lie inside the set $S = \{xy = \frac{1}{4}\}$. However, one will recognize that for each point $x > 0$, there is only one $y$ such that $y \in S$. Therefore, the envelope must be this curve.

So the envelope $\gamma(x)$ is the curve $y = \frac{1}{4x}$ for $x > 0$.

**Final Remark**

Note that the set $S$ is actually a hyperbola. Therefore, the hyperbola can be realized as the envelope of a simple family of straight lines. For this reason, hyperbolas are (approximately) reproducible in "string art": art formed from straight line segments where each segment is made by tightened string.

## 5.2 Differentiable Function With Bounded Non-Continuous Derivatives

**Setup**

Functions that are differentiable everywhere do not necessarily have continuous derivatives, even if the derivatives of the function are bounded. For the case of one variable, a classic example is

$$f(x) = \begin{cases} x^2 \sin(\frac{1}{x}) & x \neq 0, \\ 0 & x = 0. \end{cases} \tag{183}$$

Here, the altering of the amplitude by the factor of $x^2$ forces the function to be differentiable at $x = 0$ and $f'(0) = 0$. This can be checked by directly applying the definition of differentiability. However, the derivative $f'(x)$ alternates infinitely between values close to $f' = -1$ and $f' = 1$ as $x \to 0$. Therefore, the function does not have continuous derivatives.

However, one may wonder if this "infinite frequency oscillation" is necessary. In this example, we show that this is unnecessary in two-dimensions. That is, we construct a function $u(x, y)$ that is differentiable everywhere, has bounded derivatives, and does not have continuous derivatives at $(0, 0)$.

**The Problem**

Find a simple function $u(x, y)$ such that $u$ is differentiable on $\mathbb{R}^2$, the derivatives of $u$ are bounded, and at least one of the derivatives of $u$ is NOT continuous at $(0, 0)$.

**The Solution**

First let us consider a phenomenon in one-variable calculus. If $g(x)$ is a differentiable function with bounded derivatives: $|g'| \leq M$, then for any constant $\lambda > 0$ the function $h(x) = \lambda g\left(\frac{x}{\lambda}\right)$ has the same bound for its derivatives, i.e.

$|h'| \leq M$. This is easily proved using the chain rule:

$$h'(x) = \lambda \frac{d}{dx}\left(g\left(\frac{x}{\lambda}\right)\right), \tag{184}$$

$$= \lambda g'\left(\frac{x}{\lambda}\right)\frac{1}{\lambda}, \tag{185}$$

$$= g'\left(\frac{x}{\lambda}\right). \tag{186}$$

Therefore, if $|g'| \leq M$, then $|h'| \leq M$ too.

So the idea is that we can construct a function $u(x,y)$ by setting $u(x,y) = (x^2+y^2)h\left(\frac{x}{x^2+y^2}\right)$ for an appropriate function $h(x)$. What properties should we require of $h(x)$? First, let's check what is necessary for bounded derivates. Although we were guided by our idea in one-variable differentiation, we must now explicitly check that everything works out okay since our factor $x^2+y^2$ isn't actually constant.

First, let's calculate the $x$-derivative when $(x,y) \neq (0,0)$:

$$\frac{\partial u}{\partial x} = 2xh + (x^2+y^2)h'\left(\frac{1}{x^2+y^2} - \frac{2x^2}{(x^2+y^2)^2}\right), \tag{187}$$

$$= 2xh + h' - h'\frac{x^2}{x^2+y^2}. \tag{188}$$

Now, let's calculate the $y$-derivative when $(x,y) \neq (0,0)$:

$$\frac{\partial u}{\partial y} = 2yh + (x^2+y^2)h'\frac{2xy}{(x^2+y^2)^2}, \tag{189}$$

$$= 2yh + h'\frac{2xy}{x^2+y^2}. \tag{190}$$

Therefore, we see that the derivatives $u_x$ and $u_y$ will be bounded for $(x,y) \neq (0,0)$ when the function $h(x)$ and its derivative $h'(x)$ are both bounded; note that expressions like $\frac{x^2}{x^2+y^2}$ and $\frac{2xy}{x^2+y^2}$ are bounded for points away from the origin because they are invariant under scaling (i.e. homogeneous of order 0).

Finally, when $h(x)$ is bounded, the fact that we mulitply $h$ by $x^2+y^2$ to construct $u$ implies that $u$ will be differentiable at the origin and that $u_x(0,0) = u_y(0,0) = 0$. Now, focus on the expression $h'\frac{2x^2}{x^2+y^2}$ in the expression for $u_x$. If we approach the origin along $y^4 = x$, then

$$h'\left(\frac{x}{x^2+y^2}\right) = h'\left(\frac{\sqrt{x}}{x^{3/2}+1}\right), \tag{191}$$

$$\to h'(0). \tag{192}$$

Furthermore as we approach the origin along $y^4 = x$,

$$\frac{x^2}{x^2+y^2} = \frac{x^{3/2}}{x^{3/2}+1}, \tag{193}$$

$$\to 0. \tag{194}$$

Therefore, as we approach the origin along $y^4 = x$, we have that $u_x \to h'(0)$. So we will want $h'(0) \neq 0$, e.g. $h'(0) = 1$.

Let us recap our requirements for $h(x)$.

- The function $h(x)$ is differentiable with continuous derivatives on $\mathbb{R}$
- The values of the function $h(x)$ and its derivative $h'(x)$ are both bounded.
- At $x = 0$, we have $h'(0) = 1$.

A function that meets all of these requirements is $h(x) = \frac{x}{1+x^2}$. So our function $u(x, y)$ is

$$u(x,y) = (x^2+y^2)h\left(\frac{x}{x^2+y^2}\right), \tag{195}$$

$$= (x^2+y^2)\frac{x}{(x^2+y^2)(1+x^2(x^2+y^2)^{-2})}, \tag{196}$$

$$= \frac{x(x^2+y^2)^2}{(x^2+y^2)^2+x^2}. \tag{197}$$

## 5.3 Maximizing Likelihood for a Three Step Markov Process

In this example we will look at deriving statistical estimates for a three step Markov process. For more background information, see chapter one of [15]; however [15] is missing a derivation of the final formula.

### The Setup : The Constraints

We have three random variables $X_1$, $X_2$, and $X_3$ where each $X_i = 0$ or $X_i = 1$. The order of the variables matter, and we think of them as randomly being chosen in sequence according to their indices one, two, or three.

So there are eight possible outcomes according to the two possible values for each $X_i$; we label these outcomes as $X_1X_2X_3$, e.g. 000 or 110. We label the probabilities of the outcomes according to these possibilities, e.g. $p_{000}$ or $p_{110}$. We think of the probabilities forming a vector $\vec{p} \in \mathbb{R}^8$, i.e. the vector $\vec{p} = (p_{000}, p_{001}, ..., p_{111})$.

Finally, one more notation we will use. We will use $\bar{i}$ to denote the other value of 0 or 1 that is not $i$; i.e. if $i = 1$ then $\bar{i} = 0$.

To be a three step Markov process, the transition from $X_2$ to $X_3$ needs to depend only on $X_2$ and not on the entire history, i.e. not depend on $X_1$ and $X_2$. In terms of probabilities, this is expressed as

$$P(X_3 = k | X_1 = i, X_2 = j) = P(X_3 = k | X_2 = j). \tag{198}$$

Applying Bayes' formula to both sides of this equation, we get

$$\frac{p_{ijk}}{\sum_\gamma p_{ij\gamma}} = \frac{\sum_\alpha p_{\alpha jk}}{\sum_{\alpha,\gamma} p_{\alpha j\gamma}}. \tag{199}$$

We can rewrite this as

$$p_{ijk} \sum_{\alpha,\gamma} p_{\alpha j \gamma} = \left( \sum_{\alpha} p_{\alpha j k} \right) \left( \sum_{\gamma} p_{ij\gamma} \right). \tag{200}$$

Next, let's expand the sums over $\gamma$ as sums over the values $k$ and $\overline{k}$; we get

$$p_{ijk} \sum_{\alpha} p_{\alpha j k} + p_{ijk} \sum_{\alpha} p_{\alpha j \overline{k}} = p_{ijk} \sum_{\alpha} p_{\alpha j k} + p_{ij\overline{k}} \sum_{\alpha} p_{\alpha j k}. \tag{201}$$

Cancelling terms we get

$$p_{ijk} \sum_{\alpha} p_{\alpha j \overline{k}} = p_{ij\overline{k}} \sum_{\alpha} p_{\alpha j k}. \tag{202}$$

Now expand the sum over $\alpha$ as a sum over the values $i$ and $\overline{i}$, we get

$$p_{ijk} p_{ij\overline{k}} + p_{ijk} p_{\overline{i}j\overline{k}} = p_{ij\overline{k}} p_{ijk} + p_{ij\overline{k}} p_{\overline{i}jk}. \tag{203}$$

Again, cancelling terms we get

$$p_{ijk} p_{\overline{i}j\overline{k}} = p_{ij\overline{k}} p_{\overline{i}jk}. \tag{204}$$

Now, at first this appears to be eight different equations, one for each possible choice of $(i, j, k)$. However, we will now show that it is actually just two different equations without doing a brute force plug and check.

First, notice that the equation is exactly the same if we make the substitution $i \to \overline{i}$; the effect is to merely switch the left and right hand side of the equation. Therefore, the equation is the same no matter which value of $i$ we choose. So let us choose $i = 0$.

Similarly, using the substitution $k \to \overline{k}$, we can choose $k = 0$. Both of these choices give

$$p_{0j0} p_{1j1} = p_{0j1} p_{1j0}, \tag{205}$$

for either $j = 0$ or $j = 1$. It is not very hard to see that we get different equations for each different value of $j$.

Therefore, we find that the constraints for $\vec{p}$ to be a three step Markov process are exactly the four following constraints:

$$\begin{cases} p_{ijk} \geq 0, \\ \sum_{\alpha,\beta,\gamma} p_{\alpha\beta\gamma} = 1, \\ p_{000} p_{101} = p_{001} p_{100}, \\ p_{010} p_{111} = p_{011} p_{110}. \end{cases} \tag{206}$$

All probability vectors $\vec{p} \in \mathbb{R}^8$ that belong to three step Markov processes are exactly the probability vectors $\vec{p} \in \mathbb{R}^8$ that satisfy all of the above constraints.

## The Setup: Maximizing Likelihood

We are interested in creating statistical estimates for the different $p_{ijk}$ based on data recording sample counts $n_{ijk}$; that is, we run $N$ independent trials, and $n_{ijk}$ is the number of times we see outcome $X_1 = i$, $X_2 = j$, and $X_3 = k$. We denote the collection of all the $n_{ijk}$ as a vector $\vec{n}$ similarly to how we used $\vec{p}$ above.

First, let us briefly discuss the notion of likelihood. For the notion of likelihood, you consider the data $\vec{n}$ to be fixed, and we consider varying the probabilities of our model $\vec{p}$. The likelihood is defined as the probabilitiy $l(\vec{p}) = P(\vec{n}|\vec{p})$ for our three step Markov model. Assuming the trials are independent, we have

$$l(\vec{p}) = \prod_{\alpha,\beta,\gamma} (p_{\alpha\beta\gamma})^{n_{\alpha\beta\gamma}}. \tag{207}$$

The term "likelihood" is used instead of probability, because $l(\vec{p})$ does not in general represent a probability distribution on $\vec{p}$.

The idea is that a good estimate of the true probabilities should come from finding $\vec{p}$ that maximizes the likelihood $l(\vec{p}$.

Next note that maximizing likelihood is equivalent to maiximizing the logarithm of likelihood; however, the latter has a nicer form. So let $L(\vec{p}) = \log(l(\vec{p}))$. We see that

$$L(\vec{p}) = \sum_{\alpha,\beta,\gamma} n_{\alpha\beta\gamma} \log(p_{\alpha\beta\gamma}). \tag{208}$$

Now, recall that those $\vec{p}$ that represent three step Markov processes are exactly those $\vec{p}$ satisfying four constraints. So we are lead to a constrained maximization problem. We will assume that the maximum occurs at the interior of the constraints, i.e. $p_{ijk} > 0$ for all $(i, j, k)$.

### The Problem

Let the data $n_{ijk}$ be fixed. Assume that the maximum of the following constrained problem occurs at $p_{ijk} > 0$ for all $(i, j, k)$:

$$\begin{cases} \text{maximize } L(\vec{p}) = \sum_{\alpha,\beta,\gamma} n_{\alpha\beta\gamma} \log p_{\alpha\beta\gamma}, & \\ \sum_{\alpha,\beta,\gamma} p_{\alpha\beta\gamma} = 1, & \\ p_{0j0}p_{1j1} - p_{0j1}p_{1j0} = 0, & \text{for } j \in \{0, 1\}. \end{cases} \tag{209}$$

Find the $p_{ijk}$ where the maximum occurs in terms of the data $n_{ijk}$.

### The Solution

For convenience of notation, let us make the following definitions

$$f(\vec{p}) \coloneqq \sum_{\alpha,\beta,\gamma} p_{\alpha\beta\gamma}, \tag{210}$$

$$g_j(\vec{p}) \coloneqq p_{0j0}p_{1j1} - p_{0j1}p_{1j0}. \tag{211}$$

Let us look the Lagrangian condition for finding the constrained critical points of $L(\vec{p})$. Now, note that

$$\frac{\partial L}{\partial p_{ijk}} = \frac{n_{ijk}}{p_{ijk}}, \tag{212}$$

$$\frac{\partial f}{\partial p_{ijk}} = 1, \tag{213}$$

for all $(i, j, k)$. Next, let us consider the derivatives of $g_j(\vec{p})$. We see that

$$\frac{\partial g_j}{\partial p_{ijk}} = (-1)^{i+k} p_{\bar{i}j\bar{k}}, \tag{214}$$

$$\frac{\partial g_j}{\partial p_{i\bar{i}jk}} = 0. \tag{215}$$

Now, let $\lambda$ be the Lagrangian coefficient for $f(\vec{p})$ and let the two coefficients $\tau_j$ be the Lagrangian coefficients for $g_j(\vec{p})$. The Lagrangian condition gives us that

$$\frac{n_{ijk}}{p_{ijk}} = \lambda + \tau_j (-1)^{i+k} p_{\bar{i}j\bar{k}}, \tag{216}$$

for each $(i, j, k)$. Note how coefficient $\tau_0$ and variables $p_{i0k}$ are decoupled from $\tau_1$ and $p_{i1k}$; that is no equation has some variable or coefficient from both sets. However, $\lambda$ is coupled to all of them.

Now, multiply through to get

$$n_{ijk} = \lambda p_{ijk} + \tau_j (-1)^{i+k} p_{ijk} p_{\bar{i}j\bar{k}}. \tag{217}$$

Next consider this equation for $(i, j, \bar{k})$ and note that $(-1)^{i+k} = -(-1)^{i+\bar{k}}$. So we have

$$n_{ij\bar{k}} = \lambda p_{ij\bar{k}} - \tau_j (-1)^{i+k} p_{ij\bar{k}} p_{\bar{i}jk}. \tag{218}$$

Next, use the constraint $p_{ijk} p_{\bar{i}j\bar{k}} = p_{ij\bar{k}} p_{\bar{i}jk}$ and add together the two equations to get

$$n_{ijk} + n_{ij\bar{k}} = \lambda (p_{ijk} + p_{ij\bar{k}}). \tag{219}$$

Similarly do the same for $(\bar{i}, j, k)$, and we obtain

$$\frac{n_{ijk} + n_{\bar{i}jk}}{p_{ijk} + p_{\bar{i}jk}} = \frac{n_{ijk} + n_{ij\bar{k}}}{p_{ijk} + p_{ij\bar{k}}} = \lambda. \tag{220}$$

Let us concentrate on the following constraints:

$$\begin{cases} \sum_{\alpha,\beta,\gamma} p_{\alpha\beta\gamma} = 1, \\ \frac{n_{0jk} + n_{1jk}}{p_{0jk} + p_{1jk}} = \frac{n_{ij0} + n_{ij1}}{p_{ij0} + p_{ij1}} = \lambda, \\ p_{0j0} p_{1j1} = p_{0j1} p_{1j0}. \end{cases} \tag{221}$$

Note that these constraints are invariant under the substitution $i \to \bar{i}$ (appropriately interpreted for the quadratic constraint); that is the substitution $i \to \bar{i}$ is a symmetry for these constraints. Similarly $j \to \bar{j}$ is a symmetry as well.

We will want to use these symmetries to help us solve these system of constraints. Before we do so, let us show how to formalize these symmetries.

Let $S(\vec{p})_{ijk} = p_{\bar{i}jk}$ be the linear transformation that involves switching components according to the substitution $i \to \bar{i}$. Similarly let $T(\vec{p})_{ijk} = p_{ij\bar{k}}$ be that corresponding to the substitution $j \to \bar{j}$.

A symmetry in the substitution $i \to \bar{i}$ means that $\vec{p}$ satisfies these constraints if and only if $S(\vec{p})$ does too; similarly for a symmetry in the substitution $j \to \bar{j}$ and the transformation $T$.

The key point is that these constraints are easier to understand if we change to coordinates that are special for the transformations $S$ and $T$. Note that $ST = TS$ and they are both orthogonal transformations, and so we can decompose $\mathbb{R}^8$ into an eigenbasis for both $S$ and $T$; if you are uncomfortable with these theoretical details, then it is enough to know that we want to try and we will see that it will work.

First, let us decompose into an eigenbasis of $S$. Note that $S$ just switches the components of $\vec{p}$ in pairs, and so $S^2 = 1$. Therefore the eigenvalues of $S$ are at most $\pm 1$. In fact, it is simple to correctly guess the eigenvectors of $S$. This leads us to an intial change of variables

$$q_{+jk} \coloneqq p_{0jk} + p_{1jk}, \tag{222}$$

$$q_{-jk} \coloneqq p_{0jk} - p_{1jk}. \tag{223}$$

Note that we have used subscripts of $\pm$ to indicate whether the variable corresponds to eigenvalue $\pm 1$. Also take note of the symmetry for the substitution $p_{ijk} \to p_{\bar{i}jk}$ in the definition of $q_{+jk}$; furthermore, such a substitution in the definition of $q_{-jk}$ is an anti-symmetry, i.e. it changes the sign.

Next, we wish to further decompose for $T$. Again, it is simple enough to guess at the right answer. The coordinates are

$$r_{+j+} \coloneqq p_{0j0} + p_{1j0} + p_{0j1} + p_{1j1}, \tag{224}$$

$$r_{+j-} \coloneqq p_{0j0} + p_{1j0} - p_{0j1} - p_{1j1}, \tag{225}$$

$$r_{-j+} \coloneqq p_{0j0} - p_{1j0} + p_{0j1} - p_{1j1}, \tag{226}$$

$$r_{-j-} \coloneqq p_{0j0} - p_{1j0} - p_{0j1} + p_{1j1}, \tag{227}$$

$$\tag{228}$$

Again, note the symmetries and anti-symmetries for the substitution $p_{ijk} \to p_{\bar{i}jk}$ and for the substitution $p_{ijk} \to p_{ij\bar{k}}$.

It will be convenient if we do a similar change of coordinates to $n_{ijk}$ to make coordinates $\{m_{+j+}, m_{+j-}, m_{-j+}, m_{-j-}\}$.

Next, we rewrite our constraints in terms of these new coordinates.

First, we note that $\sum_{\alpha,\beta,\gamma} p_{\alpha\beta\gamma} = \sum_j r_{+j+}$, so we get the constraint

$$r_{+0+} + r_{+1+} = 1. \tag{229}$$

Next, let's look at the linear constraints $n_{ij0} + n_{ij1} = \lambda(p_{ij0} + p_{ij1})$ for each value of $i$. We note that both sides are invariant under the substitution $k \to \bar{k}$.

So we seek expressing the right side in terms of $r_{+j+}$ and $r_{-j+}$, and similar expressions for the left hand side.

We immediately see that $r_{+j+}+r_{-j+} = 2(p_{0j0}+p_{0j1})$ and that $r_{+j+}-r_{-j+} = 2(p_{1j0}+p_{1j1})$. We get a similar result for $m_{+j+}+m_{-j+}$ and $m_{+j+}-m_{-j+}$. Therefore we have

$$m_{+j+} + m_{-j+} = \lambda(r_{+j+} + r_{-j+}), \tag{230}$$
$$m_{+j+} - m_{-j+} = \lambda(r_{+j+} - r_{-j+}), \tag{231}$$
$$\tag{232}$$

So we get

$$m_{+j+} = \lambda r_{+j+}, \tag{233}$$
$$m_{-j+} = \lambda r_{-j+}. \tag{234}$$

Similarly, using the linear constraints $n_{0jk} + n_{1jk} = \lambda(p_{0jk} + p_{1jk})$ for each value of $k$, we get another equation (but only one new equation):

$$m_{+j-} = \lambda r_{+j-}. \tag{235}$$

Now, it is easy to see that $N = m_{+0+} + m_{+1+}$. So using our new equations we get

$$N = \lambda(r_{+0+} + r_{+1+}), \tag{236}$$
$$= \lambda. \tag{237}$$

So we get

$$r_{+j+} = \frac{m_{+j+}}{N}, \tag{238}$$
$$r_{-j+} = \frac{m_{-j+}}{N}, \tag{239}$$
$$r_{+j-} = \frac{m_{+j-}}{N}. \tag{240}$$

Note that the right hand sides only depend on the data $n_{ijk}$, which is fixed. In order to obtain our original $p_{ijk}$, we still need to find $r_{-j-}$ in terms of the data.

To do so, we will have to use the quadratic equations $p_{0j0}p_{1j1} - p_{0j1}p_{1j0} = 0$. We need to rewrite this quadratic polynomial in $p_{ijk}$ as a quadratic polynomial in $r$-coordinates. For now, forget the other constraints and just consider these quadratic equations for any $\vec{p} \in \mathbb{R}^8$.

Recall that these equations are described using the quadratic forms $g_j(\vec{p})$. Next, note that the quadratic forms $g_j$ are anti-symmetric for the substitution $p_{ijk} \to p_{\bar{i}jk}$ and for the substitution $p_{ijk} \to p_{ij\bar{k}}$.

These substitutions correspond to our transformations $S$ and $T$. So when we find how $g_j$ depends on the coordinates $r_{+j+}$, $r_{+j-}$, $r_{-j+}$, and $r_{-j-}$, we can make our work much shorter by paying attention to the anti-symmetries of the substitutions.

For example, we know that $g_j$ will be a quadratic polynomial in the $r$-coordinates. Let us consider what the coefficient of $r_{+j+}r_{+j+}$ can be. So consider

$$g_j(\vec{p}) = Ar_{+j+}r_{+j+} + ... \tag{241}$$

From the above discussion we know that $g_j(S\vec{p}) = g_j(s\vec{p})$. However, the $r_{+j+}$ coordinates of $S\vec{p}$ is the same as $\vec{p}$. Therefore we get

$$\begin{aligned} -(Ar_{+j+}r_{+j+} + ...) &= -g_j(\vec{p}), &(242)\\ &= g_j(S\vec{p}), &(243)\\ &= Ar_{+j+}r_{+j+} \pm .... &(244)\end{aligned}$$

Now, we haven't discussed what happens to the other $r$-coordinates, but it doesn't matter as none of them transform into having any $r_{+j+}$ coordinate. Therefore, we have $-A = A$, and so we get $A = 0$. Hence the coefficient of $r_{+j+}r_{+j+}$ is zero. The key for why this worked out so nicely is that the $r$-coordinates are constructed to from the eigenvectors of both $S$ and $T$.

In fact, we can use the anti-symmetries to narrow down our list of which terms can be non-zero. They need to match the anti-symmetry of the substitutions. Consider first the substitution $p_{ijk} \to p_{\bar{i}jk}$. This results in an anti-symmetry means that we can only have non-zero coefficients for those terms that have exactly one minus in their $i$ spot, i.e. $r_{+jx}r_{-jy}$ where $x, y \in \{-, +\}$.

Similarly, considering the anti-symmetry of the substitution $p_{ijk} \to p_{ij\bar{k}}$, we narrow down the terms to

$$g_j(\vec{p}) = Br_{+j+}r_{-j-} + Cr_{+j-}r_{-j+}, \tag{245}$$

for some unknown constants $B, C$.

To find $B$ and $C$, we simply substitute in some choice of $\vec{p}^0$ (recall that we our now considering all $\vec{p} \in \mathbb{R}^8$). Let us use $p^0_{0j0} = s$, $p^0_{1j1} = t$, and all other values of $p^0_{ijk} = 0$. We have

$$\begin{aligned} st = p^0_{0j0}p^0_{1j1} - p^0_{0j1}p^0_{1j0}, &(246)\\ r^0_{+j+} = s + t, &(247)\\ r^0_{+j-} = s - t, &(248)\\ r^0_{-j+} = s - t, &(249)\\ r^0_{-j-} = s + t. &(250)\end{aligned}$$

So we get

$$\begin{aligned} Ar^0_{+j+}r^0_{-j-} + Br^0_{+j-}r^0_{-j+} &= A(s+t)^2 + B(s-t)^2, &(251)\\ &= (A+B)s^2 + 2(A-B)st + (A+B)t^2. &(252)\end{aligned}$$

So we get

$$\begin{aligned} A + B = 0, &(253)\\ 2(A - B) = 1 &(254)\end{aligned}$$

Therefore, we get $A = 1/4$ and $B = -1/4$. So

$$g_j = \frac{r_{+j+}r_{-j-} - r_{+j-}r_{-j+}}{4}. \tag{255}$$

So the condition that $p_{0j0}p_{1j1} = p_{0j1}p_{1j0}$ becomes $r_{+j+}r_{-j-} = r_{+j-}r_{-j+}$. Therefore, we can solve for $r_{-j-}$ to get

$$r_{-j-} = \frac{m_{+j-}m_{-j+}}{Nm_{+j+}}. \tag{256}$$

Now that we have solved the $r$-coordinates in terms of the data, we can solve for the $p_{ijk}$. Again, we can pay attention to the symmetries to minimize the amount of work this entails.

First, let us solve for $p_{0j0}$. We note that

$$\begin{aligned}
p_{0j0} &= \frac{r_{+j+} + r_{+j-} + r_{-j+} + r_{-j-}}{4}, && (257)\\
&= \frac{m_{+j+}(m_{+j+} + m_{+j-} + m_{-j+}) + m_{+j-}m_{-j+}}{4Nm_{+j+}}, && (258)\\
&= \frac{m_{+j+}(m_{+j+} + m_{+j-}) + m_{-j+}(m_{+j+} + m_{+j-})}{4Nm_{+j+}}, && (259)\\
&= \frac{(m_{+j+} + m_{-j+})(m_{+j+} + m_{+j-})}{4Nm_{+j+}}, && (260)\\
&= \frac{4(n_{0j0} + n_{0j1})(n_{0j0} + n_{1j0})}{4Nm_{+j+}}, && (261)\\
&= \frac{\left(\sum_\gamma n_{0j\gamma}\right)\left(\sum_\alpha n_{\alpha j0}\right)}{N\sum_{\alpha,\gamma} n_{\alpha j\gamma}}. && (262)
\end{aligned}$$

Now that we have solved for $p_{0j0}$, we can use symmetries to solve for the rest of the $p_{ijk}$. Let $\tilde{n}_{ijk} = n_{\bar{i}jk}$. Note that this amounts to relabeling the outcomes of $X_1$. Next, let $\tilde{p}_{ijk}$ be the optimal probabilities for $\tilde{n}_{ijk}$. Since we are really only relabeling $X_1$, we have that $\tilde{p}_{ijk} = p_{\bar{i}jk}$.

Now we use the formula for $\tilde{p}_{0j0}$. We get

$$\begin{aligned}
p_{1j0} &= \tilde{p}_{0j0}, && (263)\\
&= \frac{\left(\sum_\gamma \tilde{n}_{0j\gamma}\right)\left(\sum_\alpha \tilde{n}_{\alpha j0}\right)}{N\sum_{\alpha,\gamma} \tilde{n}_{\alpha j\gamma}}, && (264)\\
&= \frac{\left(\sum_\gamma n_{1j\gamma}\right)\left(\sum_\alpha n_{\alpha j0}\right)}{N\sum_{\alpha,\gamma} n_{\alpha j\gamma}}. && (265)
\end{aligned}$$

Similarly we can compute every $p_{ijk}$ as

$$p_{ijk} = \frac{\left(\sum_\gamma n_{ij\gamma}\right)\left(\sum_\alpha n_{\alpha jk}\right)}{N\sum_{\alpha,\gamma} n_{\alpha j\gamma}}. \tag{266}$$

# 6 Multivariable Integral Calculus

## 6.1 Function Not Satisfying Fubini's Theorem

### Set Up

Here we consider Fubini's theorem in two-dimensions.

Fubini's theorem tells us two things:

- When we know we can compute a two-dimensional integral as a repeated application of one-dimensional integration over the variables $x$ and $y$.
- When we know the order of the repeated one-dimensional integration over $x$ and $y$ doesn't depend on the order of integrating over $x$ and $y$.

We will consider the problem of finding a simple function that doesn't satisfy Fubini's theorem. In particular, the result of applying the repeated one-dimensional integration will depend on the order of $x$ and $y$. We will aim to find a simple elementary function, and we will keep the domain of integration simple, i.e. the square $S = \{0 \leq x \leq 1 \text{ and } 0 \leq y \leq 1\}$.

As a hint as to how this process will work, let us recall the fact that for a convergent series $\sum_i a_i$, the limit of the partial sums is independent of the order of the sum when the series is absolutely convergent, i.e. $\sum_i |a_i| < \infty$. For our problem of finding an appropriate function $u(x, y)$, we are then lead to consider finding a function $u(x, y)$ that satisfies the following:

- The function $u(x, y)$ takes positive and negative values.
- The integral of the aboslute value of $u$ is not convergent, i.e. $\iint_s |u| \, dA = \infty$.
- The integrals of the positive and negative parts of the function must cancel out in some way such that the repeated integration gives nice finite values despite the fact that $\iint_s |u| \, dA = \infty$.

### The Problem

Find a nice elementary function $u(x, y)$ defined on the square $s = \{0 \leq x \leq 1 \text{ and } 0 \leq y \leq 1\}$ such that $u(x, y)$ doesn't satisfy Fubini's theorem in the following sense:

- Both of the repeated integrals

$$\int_0^1 \int_0^1 u(x, y) \, dx \, dy, \tag{267}$$

and

$$\int_0^1 \int_0^1 u(x, y) \, dy \, dx, \tag{268}$$

exist and are finite.

- However, the repeated integrals mentioned above are NOT equal.

**Solution**

To keep things simple, we find a function $u(x, y)$ that blows up to both $\pm\infty$ at the corner $(0, 0)$. First let us observe that the function

$$f(x, y) = \frac{-1}{(x + y)^2} \tag{269}$$

has $\iint_S |f|\, dA = \infty$; you can quickly see that convergence of this integral is suspect because $|f|$ of order $r^{-2}$ as the radius $r \to 0$. This is the edge case of convergence for two-dimensions (recall that the area element for polar coordinates in two-dimensions includes an extra $r$, i.e. $dA = rd\theta dr$).

However, $f(x, y)$ is always negative inside the square $S$; so we won't get the cancellation of positive and negative parts that we desire. To fix this we set up $u(x, y)$ to be a difference of $f(x, y)$ and a similar function. First, let $A, B$ be constants that we will determine later. Then we use

$$u(x, y) = \frac{1}{(Ax + By)^2} - \frac{1}{(x + y)^2}. \tag{270}$$

We need that $u$ takes positive and negative values in $S$. To make sure $u$ is always defined in $S$ we will restrict to considering $A, B > 0$.

Next, consider the values of $u$ along the line $\{x + y = 1\}$. This line joins two corners of $S$, i.e. $(0, 1)$ and $(1, 0)$. We will design $A$ and $B$ to make sure $u$ has opposite signs at these two corners. To do so, we need to compare the sizes of $Ax + By$ and $x + y$ at these two corners.

To get opposite signs, we need that $Ax + By$ is above 1 at one of these two corners and below 1 at the other corner. At $(0, 1)$, $Ax + By = B$ and at $(1, 0)$, $Ax + By = A$. So we can choose $A$ is above 1 and $B$ is below 1. A convenient choice is $A = 2$ and $B = 1/2$ (if you dive deeper into the construction, you will find that the reciprocal nature of $A$ and $B$ is also necessary, but we won't go into detail on this).

So we have that

$$u(x, y) = \frac{1}{(2x + y/2)^2} - \frac{1}{(x + y)^2}. \tag{271}$$

Let us verify that this function $u(x, y)$ satisfies the conditions on the repeated integrals that we are looking for.

First, note that for any $y > 0$, we have

$$\int_0^1 \frac{1}{(2x + y/2)^2} - \frac{1}{(x + y)^2}\, dx = \frac{1}{x + y} - \frac{1}{2(2x + y/2)}\bigg|_{x=0}^{1}, \tag{272}$$

$$= \frac{1}{1 + y} - \frac{1}{y} - \frac{1}{4 + y} + \frac{1}{y}, \tag{273}$$

$$= \frac{1}{1 + y} - \frac{1}{4 + y}. \tag{274}$$

So we get that

$$\int_0^1 \left( \int_0^1 \frac{1}{(2x+y/2)^2} - \frac{1}{(x+y)^2}\, dx \right) dy = \int_0^1 \frac{1}{1+y} - \frac{1}{4+y}\, dy, \tag{275}$$

$$= \log(1+y) - \log(4+y)\big|_{y=0}^{1}, \tag{276}$$

$$= \log(2) - \log(1) - \log(5) + \log(4), \tag{277}$$

$$= 3\log(2) - \log(5). \tag{278}$$

Now, let us consider the other iterated integral. First, for any $x > 0$, we have that

$$\int_0^1 \frac{1}{(2x+y/2)^2} - \frac{1}{(x+y)^2}\, dy = \frac{1}{x+y} - \frac{2}{2x+y/2}\Big|_{y=0}^{1}, \tag{279}$$

$$= \frac{1}{x+1} - \frac{1}{x} - \frac{2}{2x+1/2} + \frac{2}{2x}, \tag{280}$$

$$= \frac{1}{x+1} - \frac{2}{2x+1/2}. \tag{281}$$

So we have that

$$\int_0^1 \left( \int_0^1 \frac{1}{(2x+y/2)^2} - \frac{1}{(x+y)^2}\, dy \right) dx = \int_0^1 \frac{1}{x+1} - \frac{2}{2x+1/2}\, dx, \tag{282}$$

$$= \log(x+1) - \log(2x+1/2)\big|_{x=0}^{1}, \tag{283}$$

$$= \log(2) - \log(1) - \log(5/2) + \log(1/2), \tag{284}$$

$$= \log(2) - \log(5). \tag{285}$$

And so we see the repeated integrals are finite, but do NOT match.

## 6.2 Interesting Property of Significands Under Multiplication

### The Setup

Before we get started, let's explain the concept of the significand of a number(also sometimes referred to as the mantissa). Essentially, the significand of a number is the collection of significant digits in scientific notation, which we will take to be normalized to be between 1 and 10. For example, the significand of 1059 is 1.059, because $1059 = 1.059 \times 10^3$ in scientific notation. So the significand is found by simply ignoring the exponent in scientific notation. Note that the significand isn't defined for the number zero.

Let us denote the signficand of a number $x$ by $s(x)$.

Now we consider the case that we have a distribution of random significands $X$ such that their probability distribution is given by an inverse distribution. That is, the probability density is $\frac{1}{x\log(10)}$, i.e. for any $1 \leq x < 10$,

$$P\left(X \leq x\right) = \int_1^x \frac{1}{s\log(10)} ds. \tag{286}$$

We also assume we have another distribution of random significands $Y$ such that their probability distribution is given by an arbitrary density $f(y)$. That is, for any $1 \leq y < 10$, we have

$$P\left(Y \leq y\right) = \int_0^y f(s) ds. \tag{287}$$

We will consider the distribution of the significands for the product of $X$ and $Y$; that is $P(s(XY) \leq z)$. Let $h(z)$ be the density of the probability distribution of $s(XY)$. We will show that

$$h(z) = \frac{1}{z\log(10)}. \tag{288}$$

That is, the significands of the products also have an inverse distribution, no matter what the density $f(y)$ is. For a more detailed discussion, please see [6].

Since we seek the distribution of a random variable that is real valued, it is helpful to consider the appropriate cumulative distribution. Since we denote the density of the signficands $s(XY)$ by $h(z)$, we will then let $H(z)$ be the cumulative distribution, i.e.

$$H(z) = P(1 \leq s(XY) \leq z) \tag{289}$$

$$= \int_1^z h(z) dz, \tag{290}$$

for any $1 \leq z < 10$. Recall that the significand is always between 1 and 10.

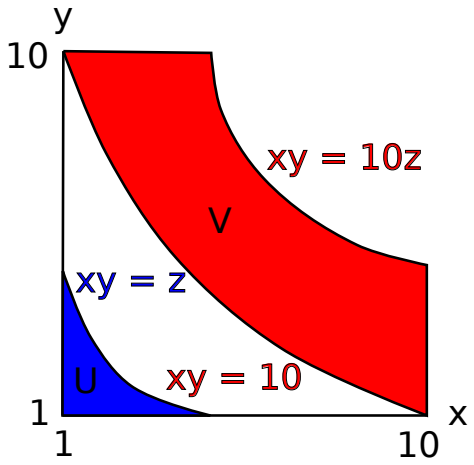
Similary, we let $F(y)$ be the cumulative distribution for $f(y)$.

Next, let us describe which significands $X$ and $Y$ satisfy $1 \leq s(XY) \leq z$. First, we have the case that the product satisfies $XY < 10$. In this case, we have that the significand $s(XY) = XY$. However, in the case that $XY \geq 10$, we have that $s(XY) = XY/10$; we need to divide by 10 to move the decimal point to the right. So we see that

$$\{1 \leq s(XY) \leq z\} = \{1 \leq XY \leq z\} \cup \{10 \leq XY \leq 10z\}. \tag{291}$$

Let $U = \{(x,y) | 1 \leq xy \leq z \text{ and } 1 \leq x, y < 10\}$ and $V = \{(x,y) | 10 \leq xy \leq 10z \text{ and } 1 \leq x, y < 10\}$. Note that $U$ and $V$ both depend on $z$.

Let's take a look at these regions.




Let $p(x,y)$ be the density for $(X,Y)$; we have that $p(x,y) = \frac{f(y)}{x\log(10)}$. Then we have that our cumulative distribution is given by

$$H(z) = \iint\limits_{U\cup V} p(x,y)\,dA_{xy}, \tag{292}$$

$$= \frac{1}{\log(10)} \iint\limits_{U\cup V} \frac{f(y)}{x} dA_{xy}. \tag{293}$$

We can use this expression for $H(z)$ to compute the density $h(z) = H'(z)$.

**The Problem**

For $1 \le z < 10$, consider the regions $U = \{(x,y)|1 \le xy \le z \text{ and } 1 \le x,y < 10\}$ and $V = \{(x,y)|10 \le xy \le 10z \text{ and } 1 \le x,y < 10\}$. Use that the cumulative distribution $H(z)$ satisfies

$$H(z) = \frac{1}{\log(10)} \iint\limits_{U\cup V} \frac{f(y)}{x}\,dA_{xy}, \tag{294}$$

to show that the density $h(z) = H'(z)$ satisfies

$$h(z) = \frac{1}{z\log(10)}, \tag{295}$$

for $1 \le z \le 10$.

**The Solution**

Let us first compute $H'_U(z)$ for $H_U(z)$ defined by

$$H_U(z) \coloneqq \frac{1}{\log(10)} \iint\limits_{U} \frac{f(y)}{x}\,dA_{xy}. \tag{296}$$

The region $U$ is simple enough that we can compute this using iterated integrals. We have that

$$H_U(z) = \frac{1}{\log(10)} \int_1^z \left( \int_1^{z/x} \frac{f(y)}{x} dy \right) dx. \tag{297}$$

Now, we have that

$$\int_1^{x/z} \frac{f(y)}{x} dy = \frac{1}{x} F\left(\frac{z}{x}\right). \tag{298}$$

So,

$$H_U = \frac{1}{\log(10)} \int_1^z \frac{1}{x} F\left(\frac{z}{x}\right) dx. \tag{299}$$

Therefore, using that $F(1) = 0$ and a $u$-substitution, we have that

$$H_U'(z) = 0 + \frac{1}{\log(10)} \int_1^z \frac{1}{x^2} f\left(\frac{z}{x}\right) dx, \tag{300}$$

$$= \frac{1}{z \log(10)} \int_1^z f(u) du, \tag{301}$$

$$= \frac{F(z)}{z \log(10)}. \tag{302}$$

Next, we compute $H_V'(z)$ for $H_V(z)$ defined by

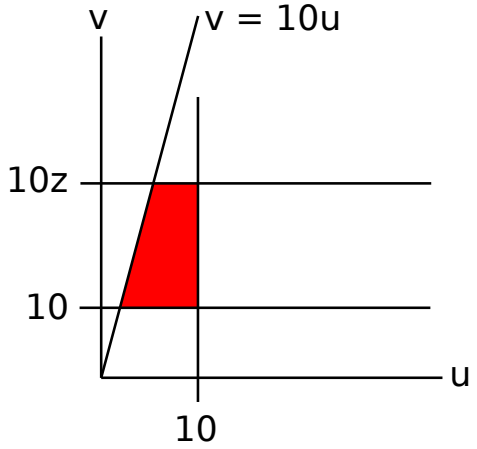
$$H_V(z) \coloneqq \frac{1}{\log(10)} \iint_V \frac{f(y)}{x} \, dA_{xy}. \tag{303}$$

We can either compute this by breaking up the integral into regions where we can apply double iterated integrals or we can make a change of coordinates. We will change our coordinates.

Define coordinates $u = x$ and $v = xy$; note that these are valid cooridinates for the square $1 \leq x, y \leq 10$. In $xy$-coordinates, the region $V$ is specified by its four boundary pieces: $\{x = 10\}$, $\{y = 10\}$, $\{xy = 10\}$, and $\{xy = 10z\}$. In $uv$-coordinates, these become $\{u = 10\}$, $\{v = 10u\}$, $\{v = 10\}$, $\{v = 10z\}$. Denote this tranformed region by $\overline{V}$; below is a picture of $\overline{V}$.

We see that we can integrate over the region $\overline{V}$ in the $uv$-plane with one application of iterated integrals. However, first we need to compute the Jacobian factor for the transformation of coordinates. We have

$$\frac{\partial(u, v)}{\partial(x, y)} = x. \tag{304}$$




So we get that

$$\frac{\partial(x,y)}{\partial(u,v)} = \frac{1}{x}, \tag{305}$$

$$= \frac{1}{u}. \tag{306}$$

So we see that

$$H_V(z) = \frac{1}{\log(10)} \iint\limits_{\overline{V}} \frac{1}{u^2} f\left(\frac{v}{u}\right) dA_{uv}, \tag{307}$$

$$= \frac{1}{\log(10)} \int\limits_{10}^{10z} \left( \int\limits_{v/10}^{10} \frac{1}{u^2} f\left(\frac{v}{u}\right) du \right) dv. \tag{308}$$

For the inner integral we get

$$\int\limits_{v/10}^{10} \frac{1}{u^2} f\left(\frac{v}{u}\right) du = \frac{1}{v} \int\limits_{v/10}^{10} f(w) dw, \tag{309}$$

$$= \frac{1}{v}\left(F(10) - F\left(\frac{v}{10}\right)\right), \tag{310}$$

$$= \frac{1}{v}\left(1 - F\left(\frac{v}{10}\right)\right). \tag{311}$$

Here we have used that $F(10) = 1$.

So then

$$H_V(z) = \frac{1}{\log(10)} \int\limits_{10}^{10z} \frac{1}{v}\left(1 - F\left(\frac{v}{10}\right)\right) dv, \tag{312}$$

and therefore

$$H'_V(z) = \frac{10}{10z\log(10)}\left(1 - F(z)\right), \tag{313}$$

$$= \frac{1}{z\log(10)} - \frac{1}{z\log(10)}F(z). \tag{314}$$

Since $H(z) = H_U(z) + H_V(z)$, we have that $h(z) = H'_U(z) + H'_V(z)$ and so we get

$$h(z) = \frac{1}{z\log(10)}. \tag{315}$$

# 7 Ordinary Differential Equations

## 7.1 Lie Symmetries

In this section we will briefly investigate the idea of studying the Lie symmetries of a differential equation and how they can sometimes be used to transform a trivial solution into the general solution.

A nice reference for this is [4].

### The Setup: The Differential Equation

Consider the differential equation

$$-1 + y^2 + xy\frac{dy}{dx} = 0. \tag{316}$$

We will be making use of partial derivatives and derivatives of quantities that really depend on only one variable. So we will be explicit as to which derivative we are using and whether it should be thought of as a partial derivative or a complete derivative; this will be made more clear later.

Furthermore note that the above differential equation isn't exact since

$$\frac{\partial}{\partial y}\left(-1 + y^2\right) = 2y, \tag{317}$$

$$\neq y = \frac{\partial}{\partial x}\left(xy\right). \tag{318}$$

It should also be clear that the extra constant term $-1$ keeps this differential equation from being separable; i.e. we can't separate the variables through division or multiplication.

So how are we to solve this equation? One way (which is usually taught in a differential equations class) is to search for an appropriate integration factor. However, here we will take a different approach. First note that this equation has a trivial constant solution $y(x) = 1$. We will then investigate symmetries of the differential equation that will then allow us to transform this particular solution into the general solution.

**The Setup: Lie Symmetries**

We will be studying the so called Lie Group symmetries of the differential equation. First define

$$F(x, y, z) = -1 + y^2 + xyz. \tag{319}$$

Note that any solution $y(x)$ to our differential equation satisfies $F\left(x, y, \frac{dy}{dx}\right) = 0$. Our ultimate goal is to find change of variables $(\tilde{x}, \tilde{y}, \tilde{z}) = \phi(x, y, z)$ that takes solutions of our differential equation to other solutions of our differential equation. Ensuring this happens is a sort of complicated process, so let us tackle it in stages. First, for clarity, we will denote the components of $\phi$ by $\phi_{\tilde{x}}, \phi_{\tilde{y}}$ and $\phi_{\tilde{z}}$.

First, we want to ensure that for a function $y = f(x)$, if $\phi$ transforms $\left(x, f(x), \frac{df}{dx}\right)$ into a curve such that $\tilde{y} = g(\tilde{x})$, i.e. $\tilde{y}$ is a function of $\tilde{x}$, then the point $\left(x, f(x), \frac{df}{dx}\right)$ transforms into $\left(\tilde{x}, g(\tilde{x}), \frac{dg}{d\tilde{x}}\right)$. That is, we have that $\tilde{z} = \frac{dg}{d\tilde{x}}$.

The key to making sure this will work is to use a chain rule for derivatives, and to consider $\tilde{y}$ and $\tilde{x}$ to both be functions of only $x$ (which is possible because we are assuming $y$ and $z$ are functions of only $x$, i.e. $y = f(x)$ and $z = f'(x)$). From the single variable chain rule we have that

$$\frac{dg}{d\tilde{x}} = \frac{\frac{d\tilde{y}}{dx}}{\frac{d\tilde{x}}{dx}}. \tag{320}$$

Let us investigate the numerator and denominator. From the fact that we are looking at the transformation of the curve $\left(x, f(x), \frac{df}{dx}\right)$, we have that

$$\tilde{y}(x) = \phi_{\tilde{y}}\left(x, f(x), \frac{df}{dx}\right). \tag{321}$$

Therefore, using the multi-variable chain rule we have that

$$\frac{d\tilde{y}}{dx} = \frac{\partial \phi_{\tilde{y}}}{\partial x}\left(x, f(x), \frac{df}{dx}\right) + \frac{df}{dx}\frac{\partial \phi_{\tilde{y}}}{\partial y}\left(x, f(x), \frac{df}{dx}\right) + \frac{d^2 f}{dx^2}\frac{\partial \phi_{\tilde{y}}}{\partial z}\left(x, f(x), \frac{df}{dx}\right) \tag{322}$$

We would like to relate the terms on the right hand side to the original coordinates $(x, y, z)$, but the term $\frac{d^2 f}{dx^2}$ has no relation to these coordinates. It depends on the original curve in a way that can't be expressed in terms of the coordinates $(x, y, z)$ themselves.

To eliminate this problem, we make the choice that $\frac{\partial \phi_{\tilde{y}}}{\partial z} = 0$; that is we require $\phi_{\tilde{y}}(x, y)$ to be only a function of $x$ and $y$. Then using that we get

$$\frac{d\tilde{y}}{dx} = \frac{\partial \phi_{\tilde{y}}}{\partial x}(x, y) + z\frac{\partial \phi_{\tilde{y}}}{\partial y}(x, y), \tag{323}$$

where $(x, y, z) = \left(x, f(x), \frac{df}{dx}\right)$.

Similarly, we require that $\phi_{\tilde{x}}(x,y)$ be only a function of $x$ and $y$, and we find

$$\frac{d\tilde{x}}{dx} = \frac{\partial \phi_{\tilde{x}}}{\partial x}(x,y) + z\frac{\partial \phi_{\tilde{x}}}{\partial y}(x,y), \tag{324}$$

where $(x,y,z) = \left(x, f(x), \frac{df}{dx}\right)$.

Now using that we require that $\phi_{\tilde{z}} = \tilde{z} = \frac{dg}{d\tilde{x}}$, we get

$$\phi_{\tilde{z}}(x,y,z) = \frac{\frac{\partial \phi_{\tilde{y}}}{\partial x}(x,y) + z\frac{\partial \phi_{\tilde{y}}}{\partial y}(x,y)}{\frac{\partial \phi_{\tilde{x}}}{\partial x}(x,y) + z\frac{\partial \phi_{\tilde{x}}}{\partial y}(x,y)}. \tag{325}$$

When $\phi_{\tilde{z}}$ satisfies the above relation, we have that when the $z$-variable acts as a derivative for a function, then the transformed $\tilde{z}$-variable will act as a derivative for the transformed curve.

Next, we change our viewpoint slightly. Instead of considering a single transformation $\phi(x,y,z)$, we consider a family of transformations $\phi^t(x,y,z)$ paramaterized by time $t$ and such that $\phi^0$ is the identitiy transformation, i.e. $\phi^0(x,y,z) = (x,y,z)$. Furthermore, we require that $\frac{\partial \phi^t}{\partial t}(x,y,z) = \vec{V}(x,y,z)$, a vector field independent of time. This restriction will allow us to solve for possible $\vec{V}$ and then integrate the symmetry forward in time.

Note that to calculate $\vec{V}$ it suffices to calculate $\frac{\partial \phi^0}{\partial t}$, the time derivative at time $t = 0$. This is more tractable as we know that $\phi^0$ is the identity transformation.

From our restrictions above we have that $\phi^t_{\tilde{x}}(x,y)$ doesn't depend on $z$. So we have that the component $V_x$ satisfies

$$V_x = \frac{\partial \phi^0_{\tilde{x}}}{\partial t}(x,y). \tag{326}$$

That is, $V_x(x,y)$ doesn't depend on $z$. Similarly $V_y(x,y)$ doesn't depend on $z$.

Next, let us derive a restriction on $V_z$. We take a time derivative $\frac{\partial}{\partial t}$ of the equation 325 and use that $\phi^0(x,y,z) = (x,y,z)$ to get

$$V_z(x,y,z) = \frac{1}{1+0z}\left(\frac{\partial V_y}{\partial x} + z\frac{\partial V_y}{\partial y}\right) - \frac{0+1z}{(1+0z)^2}\left(\frac{\partial V_x}{\partial x} + z\frac{\partial V_x}{\partial y}\right), \tag{327}$$

$$= \frac{\partial V_y}{\partial x} + z\frac{\partial V_y}{\partial y} - z\left(\frac{\partial V_x}{\partial x} + z\frac{\partial V_x}{\partial y}\right). \tag{328}$$

If you wish, this may be expressed more succinctly as

$$V_z = \left(\frac{\partial}{\partial x} + z\frac{\partial}{\partial y}\right)(V_y - zV_x). \tag{329}$$

So far we haven't actually used the differential equation; we have only found conditions on our vector field $\vec{V}$ such that its flow $\frac{\partial \phi^t}{\partial t} = \vec{V}(\phi^t)$ will properly preserve using the $z$-coordinate to represent a derivative of a function $y = f(x)$.

Now we find conditions on our vector field $\vec{V}$ such that $\phi^t$ takes a solution of the differential equation to a solution of the differential equation.

That is, suppose $y = f(x)$ is a solution to our differential equation, i.e. $F\left(x, f, \frac{df}{dx}\right) = 0$. We require that the transformed function $\tilde{y} = g(\tilde{x})$ is also a solution, i.e. $F\left(\tilde{x}, g, \frac{dg}{d\tilde{x}}\right) = 0$. However, from our previous restrictions, we can directly express this using the coordinates of the transformation $\phi^t$. We have that $F(\phi^t_{\tilde{x}}, \phi^t_{\tilde{y}}, \phi^t_{\tilde{z}}) = 0$ for all time $t$.

Taking a time derivative at $t = 0$ and using the multi-variable chain rule, we then get that

$$\frac{\partial F}{\partial x}V_x + \frac{\partial F}{\partial y}V_y + \frac{\partial F}{\partial z}V_z = 0, \tag{330}$$

at all point $(x, y, z)$. This is our final condition on the vector field $\vec{V}$.

**The Problem : Part 1**

For the differential equation

$$-1 + y^2 + xy\frac{dy}{dx} = 0, \tag{331}$$

use the function $F(x, y, z) = -1 + y^2 + xyz$ and the following restrictions on the vector field $\vec{V}(x, y, z)$ to find all possible components $V_x(x, y)$ and $V_y(x, y)$ generating the symmetries of the differential equation. The restrictions on $\vec{V}$ are:

1. The components $V_x(x, y)$ and $V_y(x, y)$ do not depend on $z$.
2. The component $V_z(x, y, z)$ satisfies

$$V_z = \frac{\partial V_y}{\partial x} + z\frac{\partial V_y}{\partial y} - z\left(\frac{\partial V_x}{\partial x} + z\frac{\partial V_x}{\partial y}\right). \tag{332}$$

3. The components of $\vec{V}$ are related to the differential equation by the following constraint:

$$\frac{\partial F}{\partial x}V_x + \frac{\partial F}{\partial y}V_y + \frac{\partial F}{\partial z}V_z = 0. \tag{333}$$

**The Problem : Part 2**

After finding all possible $\vec{V}$ generating symmetries, use one particular $\vec{V}$ to find a particular family of symmetries $\phi^t(x, y, z)$ to turn the constant trivial solution $y_0(x) = 1$ into the general solution of the differential equation.

**The Solution to Part 1**

We first compute that

$$\frac{\partial F}{\partial x} = yz, \tag{334}$$

$$\frac{\partial F}{\partial y} = 2y + xz, \tag{335}$$

$$\frac{\partial F}{\partial z} = xy. \tag{336}$$

So we get the following constraint on the components of $\vec{V}$:

$$yzV_x + (2y + xz)V_y + xyV_z = 0. \tag{337}$$

Now we use the constraint giving $V_z$ in terms of $V_x$ and $V_y$; we also group like powers of $z$.

$$0 = 2yV_y + xy\frac{\partial V_y}{\partial x} \tag{338}$$

$$+ z\left(yV_x + xV_y + xy\left(\frac{\partial V_y}{\partial y} - \frac{\partial V_x}{\partial x}\right)\right) \tag{339}$$

$$- z^2\left(xy\frac{\partial V_x}{\partial y}\right). \tag{340}$$

Since $V_x(x, y)$ and $V_y(x, y)$ don't depend on $z$, we see that the only parts of the above equation that depend on $z$ are $z$ and $z^2$. Therefore, we see that their coefficients must vanish (and so must the term not depending on $z$). So we get three equations

$$0 = 2yV_y + xy\frac{\partial V_y}{\partial x}, \tag{341}$$

$$0 = yV_x + xV_y + xy\left(\frac{\partial V_y}{\partial y} - \frac{\partial V_x}{\partial x}\right), \tag{342}$$

$$0 = xy\frac{\partial V_x}{\partial y} \tag{343}$$

The last equation tells us that $\frac{\partial V_x}{\partial y} = 0$ and so $V_x(x)$ is independent of both $y$ and $z$.

Next, we rewrite the first equation as

$$0 = 2xV_y + x^2\frac{\partial V_y}{\partial x}. \tag{344}$$

This is a linear ODE, and we may readily solve it to get

$$0 = \frac{\partial(x^2V_y)}{\partial x}, \tag{345}$$

which gives

$$V_y = \frac{g(y)}{x^2}, \tag{346}$$

for some function $g(y)$.

Then we plug into the second equation to get

$$0 = yV_x + \frac{g(y)}{x} + \frac{y}{x}\frac{dg}{dy} - xy\frac{\partial V_x}{\partial x}. \tag{347}$$

We can then separate the variables of this equation:

$$x^2\frac{\partial V_x}{\partial x} - xV_x = \frac{g(y)}{y} + \frac{dg}{dy}. \tag{348}$$

Since the left hand side depends only on $x$ and the right hand side depends only on $y$, we have that each side is equal to the same constant $C$. So

$$C = \frac{g(y)}{y} + \frac{dg}{dy}, \tag{349}$$

$$C = x^2\frac{\partial V_x}{\partial x} - xV_x. \tag{350}$$

The first equation for $g$ may be rewritten as

$$Cy = \frac{d(yg)}{dy}. \tag{351}$$

So we get that

$$g(y) = \frac{C}{2}y + \frac{D}{y}, \tag{352}$$

for some constand $D$.

Now, let us use our equation for $V_y$ in terms of $g(y)$ to get

$$V_y = \frac{Cy}{2x^2} + \frac{D}{x^2y}. \tag{353}$$

Then we rewrite the equation for $V_x$ as

$$\frac{C}{x^3} = \frac{d(x^{-1}V_x)}{dx}. \tag{354}$$

Then we may integrate to get

$$V_x = \frac{C}{2x^2} + Ex, \tag{355}$$

for another constant $E$.

Therefore, all components $V_x$ and $V_y$ are given by

$$V_x = \frac{C}{2x^2} + Ex, \tag{356}$$

$$V_y = \frac{Cy}{2x^2} + \frac{D}{x^2y}. \tag{357}$$

**Solution to Part 2**

Now we choose a simple version of $V_x$ and $V_y$ that will actually change the $y$-values of our initial solution. So we need a simple version of both such that $V_y \neq 0$. A nice choice is given by $C = E = 0$ and $D = 1$. We then have

$$V_x = 0, \tag{358}$$

$$V_y = \frac{1}{x^2 y}. \tag{359}$$

So now let us integrate this vector field for initial points starting on the graph of the trivial solution $y(x) = 1$. Let's start with an intial point of $(x_0, 1)$ on the graph of the trivial solution. So we seek a solution to the following differential system in time:

$$\frac{dx}{dt} = 0, \tag{360}$$

$$\frac{dy}{dt} = \frac{1}{x^2 y}, x(0) = x_0, \tag{361}$$

$$y(0) = 1. \tag{362}$$

Since $\frac{dx}{dt} = 0$ for all time $t$, we have that $x$ is the constant function $x(t) = x_0$ for all time. Then we may integrate

$$\frac{dy}{dt} = \frac{1}{x_0^2 y}, \tag{363}$$

to get

$$\frac{d(y^2)}{dt} = \frac{2}{x_0^2}. \tag{364}$$

Then we get

$$y(t)^2 - 1 = \frac{2t}{x_0^2}. \tag{365}$$

Fixing a value of $t = t_0$ and varying $x_0 = x$ gives us a new solution:

$$y^2 - 1 = \frac{2t_0}{x^2}. \tag{366}$$

We can rewrite this as

$$x^2 y^2 - x^2 = 2t_0. \tag{367}$$

As we change $t_0$ the right hand side just varies over different constants. So we see that the solutions are given implicilty by the family of curves

$$x^2 y^2 - x^2 = C, \tag{368}$$

for different constants $C$.

Okay, there is one catch. We have shown that these are solutions, but we haven't shown they are all solutions. This can be seen by studying the curves $x^2 y^2 - x^2 = C$ and seeing which points in the $xy$-plane are part of this family.

# 8 Topology

## 8.1 No Metric For Pointwise Convergence

In this section we look at a motivating example for why metric spaces aren't enough to capture all notions of convergence; a more abstract concept such as point-set topology is necessary to cover a simple example of convergence.

We will in particular consider pointwise convergence of a sequence of functions $f_n(x) : [0,1] \to \mathbb{R}$. We will show that there is no metric $d$ on these functions such that pointwise convergence of $f_n(x)$ is equivalent to $d(f_n, f) \to 0$. In particular, we will show that given a metric $d$ on these functions, there exists a sequence of functions $f_n$ pointwise converges to the zero function, but $d(f_n, 0) \not\to 0$.

### The Setup

Historically, topology arose from a desire to have a universal framework to discuss different notions of convergence. Since Weierstrass, analysts recognized that there are different notions of convergence for functions, e.g. uniform convergence vs point-wise convergence. It was desirable to put all of these ideas on even footing.

In a sense, Frechet began the introduction of more abstract spaces with the creation of his so-called L-spaces in 1904. Later, in his doctoral dissertation in 1906, Frechet introduced the concept of a metric space. Convergence of sequences was defined by abstract definitions of distance.

In 1914, Hausdorff took a more set-theoretic approach; he introduced the precursor to the modern topology. The abstract concept of distance to a point is replaced with an abstract idea of neighborhoods of a point. For a nice historical discussion of the creation of point-set topology, see [8].

Now, let $X$ be the set of $\mathbb{R}$-valued functions defined on $[0,1] \subset \mathbb{R}$. We will show that there is no metrix $d$ on $X$ such that point-wise convergence is equivalent to convergence in the metric. Now, at first you may think the issue is that the functions in $X$ are unbounded, but this is not the issue.

For example, the metric $d_0(f,g) = \sup_{x\in[0,1]} \min\{|f(x) - g(x)|, 1\}$ is a bounded metric on $X$. Furthermore, if $d_0(f_n, g) \to 0$ then it isn't too hard to see that $f_n$ must converg pointwise to $g$ (in fact the convergence must be uniform).

We will show that pointwise convergence for functions in $X$ is not equivalent to convergence in any metric, but is pointwise convergence equivalent to convergence in a topology? Yes, if we think of the space of functions $X$ as a product of copies of $\mathbb{R}$, $X = \prod_{x\in[0,1]} \mathbb{R}_x$ where the $x^{\text{th}}$ component of $f$ is $f(x)$, then the product topology will work.

### The Problem

Let $X$ be the $\mathbb{R}$-valued functions defined on $[0,1] \subset \mathbb{R}$. Show that there is no metric such that a sequence of functions $f_n \in X$ converges point-wise to $g$ if and only if $d(f_n, g) \to 0$.

### The Solution

We have already seen that there exists metrics $d$ on $X$ such that $d(f_n, g) \to 0$ implies that $f_n$ converges point-wise to $g$. So we must show that the other direction of implication does not hold.

We will prove by contradiction; so for the purpose of creating a contradiction, assume that $d$ is a metric on $X$ such that $f_n$ converges point-wise to $g$ if and only if $d(f_n, g) \to 0$. We will show that there is a sequence of functions $f_n$ converging point-wise to the zero function $g(x) = 0$, but $d(f_n, 0) \not\to 0$.

So, let $d$ be a metric on $X$. We prove the following claim: for every $x \in [0,1]$, there is a $\delta_x > 0$ such that if $|f(x)| \geq 1$, then $d(f(x), 0) > \delta_x$.

Fix $x \in [0,1]$. Let $\delta_x = \inf\{d(f,0) || f(x)| \geq 1\}$. If $\delta_x = 0$, then there is a sequence of $f_n$ such that $d(f_n, 0) \to 0$ and $|f_n(x)| \geq 1$. However, our assumptions on $d$ then imply that $f_n$ point-wise converges to 0, in particular $f_n(x) \to 0$. So we have a contradiction. Therefore, $\delta_x > 0$.

Now we partition the $x \in [0,1]$ according to which of the following intervals $\delta_x$ lies in: we use a countable collection of intervals given by $[1, \infty)$, $[1/2, 1)$, $[1/3, 1/2)$, and so on. Note that we have used the positivity of $\delta_x$.

Since this is a countable partition of the uncountable set $[0,1]$, there must be a partition that contains an infinite number of $x$. Therefore, there is a fixed $N$ such that there are an infinite number of $x \in [0,1]$ such that $\delta_x > 1/N$; Now let $x_n$ be a sequence of distinct such $x$.

Next, define the sequence of functions

$$f_n(x) \coloneqq \begin{cases} \frac{1}{n-k}, & x = x_k \text{ for some } k < n, \\ 1, & x = x_k \text{ for some } k \geq n, \\ 0, & \text{otherwise.} \end{cases} \tag{369}$$

We then have that $f_n$ converges pointwise to the zero function. However, since $f_n(x_n) = 1$, we have that $d(f_n, 0) \geq \delta_{x_n} > 1/N$.

So $f_n$ gives us a sequence of functions that converge point-wise to zero, but doesn't converge in the metric. Hence we have our contradiction.

## 8.2 No Metric For Convergence From Above

### The Setup

Here we look at another notion of convergence for which there is no metric $d(x,y)$ that describes it. In particular we will look at a notion of convergence for sequences in $\mathbb{R}$ such that there is no metric $d(x,y)$ on $\mathbb{R}$ where the convergence can be described using the metric space convergence of $d$.

First let us define that a sequence $a_n \in \mathbb{R}$ *converges from above* to $L\}$ when for every $\epsilon > 0$, there exists an $N$ such that if $n \geq N$ then $L \leq a_n < L + \epsilon$.

Roughly, the above definition means that in addition to converging to $L$ (in the normal sense), eventually the sequence must also be above $L$. Note that we do not assume that $a_n$ satisfies any sort of monotonicity (or eventual monotonicity).

The fact that this is notion of convergence depends on a direction depends on a direction (i.e. from above) gives one the intuition that this convergence must depend on more than distance. That is, it is no surprise that a metric $d(x, y)$ will not be enough to describe it.

However, there must be some care here. Note that there are subsets $U \subset \mathbb{R}$ such that *converges from above* for $a_n \in U$ can be described using the standard metric on $\mathbb{R}$. For example, when $U = \{0\} \cup \{1/m | m \in \mathbb{Z}^+\}$, a sequence $a_n \in U$ converges with respect to the standard sub-space metric if and only if it converges from above. This can easily be seen, because any sequence $a_n \in U$ converging to some $1/m \in U$ must eventually be constant. Finally, if $a_n$ converges to 0, then by the defintion of $U$ every point $a_n$ already satisfies $a_n \geq 0$.

From our above example of a sub-set $U$, we see that non-existence of such a metric for $\mathbb{R}$ will be a little more deep than just the directionality. We will need to use the uncountable nature of $\mathbb{R}$.

**The Problem**

Prove that there does not exists a metric $d(x, y)$ on $\mathbb{R}$ such that a sequence $a_n \in \mathbb{R}$ *converges from above* to $L$ if and only if the sequence converges to $L$ in the sense of metric-space convergence for $d$.

**The Solution**

We prove by contradiction. Assume first that such a metric $d$ exists.

We claim that every point $x \in \mathbb{R}$ has a $\delta_x > 0$ such that all points $y$ in the metric ball $B_{\delta_x}(x)$ satisfy $x \leq y$. If this was not the case, then we would have a sequence of points $a_n$ converging to $x$ with respect to the metric convergence of $d$ but satisfying $a_n < x$. This is a sequence that can not converge to $x$ from above; by our assumption on $d$ this is impossible.

We will now find a sequence that converges from above but does not converge with respect to the metric. Since the open interval $(0, 1)$ is uncountably infinite, we may find a strictly decreasing sub-sequence of points $0 < a_n < 1$ such that $a_n$ converges to $0 \leq x_0 \leq 1$ and $\delta_{a_n} > 1/K$ for some $K \in \mathbb{Z}^+$. Note that $a_n$ converges from above to $x_0$. So $d(x_n, x_0) \to 0$. However $a_n < a_m$ for $m < n$, so $a_n \notin B_{1/K}(a_m)$.

Since $d(x_n, x_0) \to 0$, there exists $a_m, a_n \in B_{1/3K}(x_0)$ and $m < n$. However, we then have that $d(x_m, x_n) < 2/3K < 1/K$ which contradicts that $a_n \notin B_{1/K}(a_m)$.

## 8.3 Example of Weak Convergence that is not Metrizable

In this example we explore a normed vector space where weak convergence of its continuous linear functionals is not given by a metric.

### The Setup

Let $V$ be the space of $\mathbb{R}$ valued sequences $v_i$ such that $v_i \neq 0$ for at most a finite number of $i$; we give $V$ the $l^\infty$ norm where $\|v\|_{l^\infty} = \sup_i |v_i|$. For example, $(1, 1, 0, 0, 0, ...) \in V$, but $(1, 1, 1, 1, ...) \notin V$.

Let $V^*$ be the space of continuous linear functionals on $V$ with respect to the $l^\infty$ norm; i.e. the linear functions $\omega : V \to \mathbb{R}$ that satisfy the usual continuity condition $|\omega(v)| \leq C_\omega \|v\|_{l^\infty}$.

A sequence $\omega_n \in V^*$ is said to converge weakly to $\psi \in V^*$ when for any fixed vector $v \in V$ we have that $\omega_n(v) \to \psi(v)$.

We wish to show that there is no metric $d$ on $V^*$ such that weak convergence on $V^*$ is equivalent to convergence with respect to the metric $d$; therefore, weak convergence properly belongs to the notion of convergence with respect to topology and not metric spaces. Note that weak convergence is covered by the topology generated by sets $U_{\psi,v,\delta}$ of the form $U_{\psi,v,\delta} :- \{\omega \in V^* \mid |\omega(v) - \psi(v)| < \delta\}$. That is, it is covered by the topology generated by a family of semi-norms.

To show our desired conclusion, let $d$ be any metric on $V^*$ such that if $d(\omega_n, \psi) \to 0$ then $\omega_n$ weakly converges to $\psi$. We will show that there must exist a sequence $\omega_n$ weakly converging to 0, but $\omega_n$ does not converge to 0 with respect to the metric $d$.

### The Problem

Given a metric $d$ on $V^*$ such that if $d(\omega_n, \psi) \to 0$ then $\omega_n$ weakly converges to $\psi$. Show that there must exist a sequence $\omega_n$ weakly converging to 0, but $\omega_n$ does not converge to 0 with respect to $d$.

### The Solution

For any index $i \in \mathcal{N}$, let $1_i :- v \in V$ such that $v_i = 1$ and $v_j = 0$ for $j \neq i$.

Consider any finite subset $S \subset \mathcal{N}$. We let $\delta(S)$ be the infimum of all $d(\omega, 0)$ such that $\omega(1_i) \geq 1$ for some $i \in S$.

We claim that $\delta(S) > 0$. Since $S$ is finite, there exists a fixed $i \in S$ and a sequence $\omega_n$ such that $d(\omega_n, 0) \to \delta(S)$ and $\omega_n(1_i) \geq 1$. Hence we have that $\omega_n$ does not weakly converge to 0. Therefore by our assumption on $d$, we must have that $\omega_n$ also does not converge to 0 with respect to the metric $d$. Hence, $\delta(S) > 0$.

Now partition $(0, \infty)$ into a countably infinite set of disjoint intervals, each of which is bounded away from zero: $[1, \infty)$, $[1/2, 1)$, $[1/4, 1/2)$, and so on. Since the set of all finite subsets $S \subset \mathcal{N}$ is uncountable infinite, we must have that one of the above intervals contains $\delta(S)$ for an infinite collection of $S$. Therefore,

we may find an infinite collection $\mathcal{C}$ of $S$ such that $\delta(S) \geq \delta > 0$ for some fixed $\delta > 0$ independent of $S$ in $\mathcal{C}$.

Since $\mathcal{C}$ is infinite, we can find a strictly increasing sequence of integers $n_i$ such that $n_i \in S$ for some $S \in \mathcal{C}$. Then consider the sequence of continuous linear functionals $\omega_i(v) = v_{n_i}$. Note that $\omega(1_{n_i}) = 1$. Therefore we have that $d(\omega_i, 0) \geq \delta$; hence $\omega_i$ does not converge to 0 with respect to the metric $d$.

However, for any fixed $v \in V$, $v_i$ is non-zero for only finitely many $i$. Therefore, for large enough $i$ we have that $v_{n_i} = 0$. Hence $\omega_i(v) \to 0$. So $\omega_i$ converges weakly to 0 despite the fact that it does not converge to 0 with respect to $d$.

## 8.4 A Nonmetrizable Topology for Compactly Supported Continuous Functions

In this section we look at a topology on the compactly supported continuous functions on $\mathbb{R}$, i.e. $C_0(\mathbb{R}) = \{f \in C(\mathbb{R}) | \text{suppt}(f) \text{ compact}\}$.

### The Setup

We will define the topology using an exhaustion of $\mathbb{R}$ by compact intervals. So first define

$$U_n = \{f \in C_0(\mathbb{R}) | \text{suppt}(f) \subset [-n, n]\}. \tag{370}$$

Note that $C_0(\mathbb{R}) = \bigcup_n U_n$. Each $U_n$ is given the topology of uniform convergence, i.e. the open sets of $U_n$ are generated by the balls $B_\epsilon(f) = \{g \in U_n | \sup_{x \in [-n,n]} |f(x) - g(x)| < \epsilon\}$ for $f \in U_n$.

We define a topology $\tau$ on $C_0(\mathbb{R})$ using the final topology given by the inclusions $U_n \to C_0(\mathbb{R})$. In particular, $U \subset C_0(\mathbb{R})$ is open if and only if $U \bigcap U_n$ is open for all $n$.

This topology is a useful topology for the theory of distributions [7](page 180). In particular, it has many continuous linear functionals. For example, the functional $T : C_0(\mathbb{R}) \to \mathbb{R}$ defined by $T(f) = \int_{\mathbb{R}} x^2 f(x) dx$ is continous for the topology $\tau$, but it is not continuous for the topology of the supremum norm (i.e. the topology induced by $\|f\| = \sup |f(x)|$).

The topology $\tau$ can also be described as the largest topology on $C_0(\mathbb{R})$ that allows the inclusions $U_n \to C_0(\mathbb{R})$ to be continuous.

There are alternative ways to realize this topology, see counterexample 1.1.8 of [5]. In particular, it can be seen as coming from an uncountable collection of semi-norms.

### The Problem

Give a direct proof that the above topology on $C_0(\mathbb{R})$ is non-metrizable.

## The Solution

Let $d$ be any metric on $C_0(\mathbb{R})$ such that the topology $\tau_d$ generated by the metric $d$ is contained in $\tau$. We will show that there exists a sequence $f_n \to 0$ in $\tau_d$, but $f_n \not\to 0$ in $\tau$.

First note that for any $n > 0$, we can find a sequence of non-zero functions $f_k$ with support in $[n-1, n+1]$ such $f_k \to 0$ in $\tau$ (simply take a sequence uniformly converging to 0 in $U_{n+1}$ with the appropriate support restrictions). In particular, we may take that $f_k(n) > 0$ for all $k$. Therefore, given a ball $B_{1/n}(0)$ for the metric $d$, we can find a function $f_n$ such that $f_n(n) > 0$ for all $n$ and $f_n \in B_{1/n}(0)$. We see that $f_n \to 0$ in $\tau_d$.

However, it is then easy to construct a function $\rho$ such that $0 < \rho(n) < f_n(n)/2$ for all $n$. We see that $V = \{|f(x)| < \rho(x)\}$ is open in $\tau$, but $f_n \notin V$ for all $n$. Therefore $f_n \not\to 0$ in $\tau$, despite $f_n \to 0$ in $\tau_d$.

## 8.5 A Compact Subspace of Sequences of Non-negative Integers

Here we consider a compact sub-space of a metric-space topology defined on the space of sequences of non-negative integers. Both this metric space and sub-space are considered in [3] in relation to Suslin sets.

## The Setup

Let $d_0(i,j)$ be the following bounded metric on the set of non-negative integers $\mathbb{Z}_{\geq 0} = \{0, 1, 2, ...\}$:

$$d_0(i,j) \coloneqq \frac{|i-j|}{1+|i-j|}. \tag{371}$$

Note that $d_0$ satisfies all of the conditions of being a metric, and that $d(i,j) < 1$.

We let $\mathcal{N}$ be the metric space of sequences of non-negative integers with following metric:

$$d(m,n) = \sum_{i=0}^{\infty} \frac{d_0(m_i, n_i)}{2^i}, \tag{372}$$

for $m, n \in \mathcal{N}$.

The importance of the space $\mathcal{N}$ is that it is used to create the notion of Suslin sets which are useful for studying the behavior of Borel sets under continuous maps [3]; these are topics that are well beyond these notes and are not necessary for this example. We only mention them to provide some context for $\mathcal{N}$.

Next, given a fixed sequence $M \in \mathcal{N}$, we will define the sub-space

$$K \coloneqq \{n \in \mathcal{N} \mid n_i \leq M_i \text{ for all } i\}. \tag{373}$$

We will show that $K$ is a compact sub-space of $\mathcal{N}$.

Before we do so, let us discuss the topology of $\mathcal{N}$. Note that $d_0(i,j) \geq d_0(0,1) = 1/2$ when $i \neq j$. Therefore, if $d(m,n) < 2^{-j}$ for some $j > 1$, then

we know that $m_i = n_i$ for all $i < j$. Therefore for every point $m$ in the ball $B_{2^{-j}}(n)$, we have that $m_i = n_i$ for all $i < j$.

You may be attempted to interpret members $n \in \mathcal{N}$ as numeral expansions of numbers in $\mathbb{R}$ with an "infinite base". However, such an interpretation is not completely correct; let us look at such a relationship and show that it isn't a homeomorphism.

Consider the interval $[0, 1) \subset \mathbb{R}$. We construct a function $g : \mathcal{N} \to [0, 1)$ that is onto and continuous. First partition the interval $[0, 1)$ into an infinite set of intervals of length $1/2, 1/4, 1/8, 1/16, ...$; explicitly the intervals are $[0, 1/2), [1/2, 3/4), [3/4, 7/8), ....$ Then for $n \in \mathcal{N}$, we let $n_0$ choose one of the preceding intervals. Next, we take our previously chosen interval, break it up into intervals of geometrically decreasing sizes like we have done before, and let $n_1$ choose the next interval. And we inductively continue with all $n_i$. This gives a continuous function $g : \mathcal{N} \to [0, 1)$ that is onto. Explicitly we have

$$g(n) = (1 - 2^{-n_0}) + 2^{-1-n_0}(1 - 2^{-n_1}) + 2^{-2-n_0-n_1}(1 - 2^{-n_2}) + ... \quad (374)$$

Note that $g(n) < (1 - 2^{-n_0}) + 2^{-1-n_0} + 2^{-2-n_0} + ...$ and so we have that $g(n) < 1 - 2^{-n_0} + 2^{-n_0} = 1$. Similarly, by taking $n_i \to \infty$, we see that $\sum_{n \in \mathcal{N}} g(n) = 1$. Finally, $g(0) = 0$ and $g(n) > 0$ for $n \neq 0$. Therefore, we have that $g : \mathcal{N} \to [0, 1)$ is well defined.

Furthermore, $g$ is continuous. If $d(m, n) < 2^{-j}$, then $m_i = n_i$ for $i < j$. So we have that

$$|g(m) - g(n)| < 2^{-j} + 2^{-j-1} + ... = 2^{1-j}. \quad (375)$$

Therefore, we have that $g$ is continuous.

From the desciption of the choice of intervals that are used to define $g$, we see that if $m < n$ for $m, n \in \mathcal{N}$, then $g(m) < g(n)$; continuing the analogy to, e.g., a decimal expansion, this where there is a difference. Two decimal expansions may represent the same number, e.g. $0.9999... = 1.0$. However, since there is no upper limit on the "numerals" used to construct $g : \mathcal{N} \to [0, 1)$, such a situation is impossible for $g$. Furthermore, the construction of $g$ may be used to see that $g$ is onto $[0, 1)$.

We have shown that $g : \mathcal{N} \to [0, 1)$ is a one-to-one, onto, and continuous. However, $g^{-1}$ is not continuous and so $g$ fails to be a homeomorphism. This may be seen in the following way. Let $n^j = (0, j, 0, 0, ....) \in \mathcal{N}$ and $m = (1, 0, 0, ...) \in \mathcal{N}$. Then we have that $g(n^j) \to g(m)$, but $d(n^j, m) > 1/2$. Therefore $g^{-1}$ is not continuous. Despite $g$ failing to completely characterize the topology of $\mathcal{N}$, it can stil be useful to add to your understanding of the nature of $\mathcal{N}$.

## The Problem

Given a fixed sequence $m \in \mathcal{N}$, we define the sub-space $K \coloneqq \{n \in \mathcal{N} \mid n_i \leq m_i \text{ for all } i\}$. Show that $K$ is compact.

## The Solution

It is very clear if $\tilde{K}$ is the space corresponding to another fixed sequence $\tilde{m}_i$ such that $m_i \leq \tilde{m}_i$ for all $i$, then $K$ is a closed subspace of $\tilde{K}$. Since a closed subspace of a compact space must also be compact, without any loss in generality we may consider the case that $m_i \geq 1$ for all $i$.

The key is to again interpret $K$ as a numeral expansions as we did for $g$ and $\mathcal{N}$; however this time, instead of having an "infinite base", we have a varying finite base given by the $m_i$ (the base will actually be $1 + m_i$). We will use this to construct a function $f : K \to [0, 1]$ that is continuous and onto. Note that the function $f$ will not be one to one, but the inverse image of any point will be finite; also note that it is onto a closed interval. With $f$ constucted, we can then imitate the proof that a closed and bounded interval in $\mathbb{R}$ is compact by looking at the set $S \coloneqq \{x \in [0, 1] \mid f^{-1}([0, x]) \text{ is compact } \}$.

Interpreting $K$ as numeral expansions for numbers in $[0, 1]$ with a variable base $m_i$, we are lead to the following construction for $f(n)$ for $n \in \mathcal{N}$. First, we divide $[0, 1]$ into $m_0 + 1$ intervals and choose the $n_0^{\text{th}}$ interval. Next we divide that interval into $m_1 + 1$ intervals and choose the $n_1^{\text{th}}$ sub-interval. We continue this for every $n_i$. We are lead to the following formula:

$$f(n) \coloneqq \frac{n_0}{1 + m_0} + \frac{n_1}{(1 + m_0)(1 + m_1)} + ... \tag{376}$$

Note that

$$f(n) \leq \frac{m_0}{1 + m_0} + \frac{m_1}{(1 + m_0)(1 + m_1)} + ... \tag{377}$$

$$= 1 - \frac{1}{1 + m_0} + \frac{m_1}{(1 + m_0)(1 + m_1)} + ... \tag{378}$$

$$= 1 - \frac{1}{(1 + m_0)(1 + m_1)} + ... \tag{379}$$

and the pattern continues. So we see that the sum of the first $j + 1$ terms is equal to $1 - \frac{1}{(1+m_0)...(1+m_j)}$. Therefore, $f(n)$ is well-defined by a convergent series and $f(n) \leq 1$. Furthermore, in the case that $n_i = m_i$ for all $i$, we have equality; so $f(m) = 1$.

From the interval construction of $f$, it is clear that $f : K \to [0, 1]$ is also onto. Next note that when $d(x, y) < 2^{-j}$ for $x, y \in K$, then $x_i = y_i$ for $i < j$. Hence

$$|f(x) - f(y)| \leq \frac{|x_j - y_j|}{(1 + m_0)...(1 + m_j)} + ... \tag{380}$$

Since $|x_j - y_j| \leq m_j$, the analysis we used above will apply here as well. So we have that $|f(x) - f(y)| \leq \frac{1}{(1+m_0)...(1+m_j)} \leq 2^{-j-1}$, since $m_i \geq 1$. Therefore, $f$ is continuous.

Finally, note that although $f$ is not one to one, we do have the fact that $f^{-1}(x)$ consists of at most two points in $K$ for any $x \in [0, 1]$. Furthermore, if $f^{-1}(x)$ consists of two points, then they both must be of the form

$(n_0, ..., n_j, m_{j+1}, m_{j+2}, ...)$ and $(n_0, ..., n_j + 1, 0, 0, ...)$; you may find it helpful to compare to how the decimal expansions 0.999... and 1.000... represent the same real number.

With the above facts we have what we need to prove that $K$ is compact. Consider the set $S \coloneqq \{x \in [0,1] \mid f^{-1}(x) \text{ is compact }\}$, and let $\mu = \sup S$. Note that $f^{-1}(0)$ consists of the single point $(0, 0, ...) \in K$; therefore $0 \in S$.

Next, we show that $\mu \in S$. Let $\mathfrak{F}$ be a collection of open sets covering $f^{-1}([0, \mu])$.

First, consider the case that $f^{-1}(\mu)$ consists of a single point $n \in K$. So we know that $n_i > 0$ for infinitely many $i$. There exists a $j$ such that the open ball $B_{2^{-j-1}}(n)$ is contained in a single open set $U_0 \in \mathcal{F}$ and $n_j > 0$. So we know that the set $V \coloneqq \{x \in K | x_i = n_i \text{ for all } i \leq j\} \subset U_0$. Now, since $n_j > 0$, we must have that over $x \in f^{-1}([0, \mu]) \setminus V$, $f(x)$ is maximized by $x_0 = (n_0, ..., n_j - 1, m_{j+1}, m_{j+2}, ...)$. Now, for some $k > j$, we have also have that $n_k > 0$. It is clear then clear that $f(n) \geq f(x_0) + \frac{1}{(1+m_0)...(1+m_k)}$. Therefore, there is a $\nu$ such that $f(x) \leq \nu < \mu$ for all $x \in f^{-1}([0, \mu]) \setminus V$. Since $[0, \nu]$ is compact, cover it with a finite collection of open sets $\mathfrak{G}$ from $\mathfrak{F}$. Then we have that $\{U_0\} \cup \mathfrak{G}$ is a finite collection of open sets from $\mathfrak{F}$ covering $f^{-1}([0, \mu])$, and so $\mu \in S$.

Next, consider the case that $f^{-1}(\mu)$ consists of two points. They must be of the form $n = (n_0, n_1, ..., n_j, m_{j+1}, m_{j+2}, ...)$ and $p = (n_0, n_1, ..., n_j + 1, 0, 0, ...)$ for some $j$. We may cover both $n$ and $p$ with at most two open sets from $\mathfrak{F}$. Then for any $x$ satisfying $f(x) < \mu$, we have that $x < n$. Using a ball $B_{2^{-j-1}}(n)$ and the fact that $m_i > 0$, we may apply the same argument from the previous case. Therefore, we again have that $\mu \in S$. We have shown that in all cases $\mu \in S$.

Finally, we show that if $\mu < 1$ then $\mu + \delta \in S$ for some $\delta > 0$. Let $n$ be the largest element of $f^{-1}(\mu)$. We know that there exists $j$ such that the set $W \coloneqq \{x \in K | x_i = n_i \text{ for all } i < j\}$ is covered by a single open set $U_0$ from $\mathcal{F}$. Furthermore, there is a $k > j$ such that $n_k < m_k$; let us show that this must be true. If not then $n = (n_0, ..., n_j, m_{j+1}, m_{j+2}, ...)$. Since $f(n) < 1$, we have that for some $i \leq j$ that $n_i < m_i$. Let $i$ the largest $i \leq j$ satisfying this. Then we have that $p = (n_0, ..., n_i + 1, 0, 0, ...)$ is a larger element of $f^{-1}(\mu)$, but this can not be the case. Thus, by contradiction we have show that there exists a $k > j$ such that $n_k < m_k$.

Now let $q = (n_0, ..., n_{k-1}, m_k, m_{k+1}, ...)$. We have that $q > n$, $q \in U_0$, and $f(q) \geq f(n) + \frac{1}{(1+m_0)...(1+m_k)} > \mu + 2\delta$ for some $\delta > 0$. So if $f(x) \leq \mu + \delta$, then $x < q$. If $x < q$, then either the first $j$ terms of $x$ agree with $n$ and $x \in U_0$, or $x_i < n_i$ for some $i \leq j$. In the latter case, $f(x) \leq \mu$. Since $f^{-1}([0, \mu]$ is compact, we then have that combining $\{U_0\}$ with a finite sub-convering from $f^{-1}([0, \mu])$ gives us that $f^{-1}([0, \mu + \delta])$ is compact. Therefore, when $\mu < 1$, we have that $\mu + \delta \in S$ for some $\delta > 0$.

Combining the above, we must have that $\sup S = 1$, and therefore $K = f^{-1}([0, 1])$ is compact.

## 8.6 Global Analysis

### The Setup

It is sometimes possible for us to use topology to extend an argument that holds in a limit to a more global case.

Here we look at an example from Osserman [9] where we can use winding number arguments to show that a map is in fact a diffeomorphism. In particular, consider the domain $D = \{0 < |z| < 1\} \subset \mathbb{C}\}$, and consider the map

$$f(z) = \frac{1}{\bar{z}} + \frac{z^3}{3}, \tag{381}$$

defined on $D$. Let $C = \{|z| = 1\}$ be the outermost component of the boundary of $D$.

Let us next show that $f$ is a local diffeomorphism. Recall that we by the inverse function theorem, we need to only show that on all points of $D$ we have that $Df$ is of rank two. This is equivalent to only having trivial solutions $r_i \in \mathbb{R}$ to $r_1\partial_x f + r_2\partial_y f$; this is equivalent to only a trivial solution $c \in \mathbb{C}$ of $c\partial_z f + \bar{c}\partial_{\bar{z}} f = 0$.

We have that

$$\partial_z f = z^2, \tag{382}$$

$$\partial_{\bar{z}} f = -\frac{1}{\bar{z}^2}. \tag{383}$$

Hence, we have that $\frac{c^2}{|c|^2}|z|^4 = 1$. Note that we only need to show that such an equation has no solutions with $|c| = 1$ and $z \in D$. So we consider $c^2|z|^4 = 1$. Note that every term other than $c$ is in $\mathbb{R}$, and $|z|^4 \geq 0$. Therefore we must have that $c = \pm 1$, and therefore $|z|^4 = 1$. Therefore there are no solutions in $D$, and $Df$ must be rank two throughout $D$. Hence, $f$ is a local diffeomorphism.

We will show that $f(C)$ is a Jordan curve, and we will use a topological argument to show that $f$ is in fact a global diffeomorphism of $D$ onto the exterior of $f(C)$. To do so, we will employ a winding number argument.

### The Problem

1. Show that $f$ is a one-to-one map of the boundary component $C = \{|z| = 1\}$.
2. For any $w_0$ in the exterior of the Jordan curve $f(C)$, use a winding number argument to show that $w_0$ is the image of exactly one point in $D$.
3. Similarly show that no points $w_0$ in the interior of $f(C)$ are in the image of $D$.
4. Finally, argue that every point $w_0 \in f(C)$ is not in the image $f(D)$.

Therefore we have that $f$ is a global diffeomorphism of $D$ onto the exterior of $f(C)$.

**The Solution**

1. First let us show that $f$ is a one-to-one map when restricted to $C = \{|z| = 1\}$. On $C$, we have that $\bar{z} = 1/z$, so we have that $f(z) = z + z^3/3$.

   So if $|z| = |\zeta| = 1$ and $f(z) = f(\zeta)$, then

   $$z - \zeta + \frac{z^3 - \zeta^3}{3} = 0. \tag{384}$$

   We factor this equation by $z - \zeta$ to get either $z = \zeta$ or

   $$3 + z^2 + z\zeta + \zeta^2 = 0. \tag{385}$$

   Now, $|z^2| = |z\zeta| = |\zeta^2| = 1$; so $|z^2 + z\zeta + \zeta^2| \leq 3$ with equality only when all three terms are the same. So from the above we must have that all three terms are the same, and hence $z = \zeta$.

   In either case $z = \zeta$, and so we have that $f$ is one-to-one when restricted to $C$.

2. If $w_0$ is in the exterior of the Jordan curve $f(C)$, then notice that $f(C)$ is a simply connected curve in $\mathbb{C} \setminus w_0$.

   However, for $r$ small, the curve $\theta \to f(re^{i\theta})$ is a curve with winding number 1 about $w_0$.

   Next, note that since $f$ is a local diffeomorphism, the points in $f^{-1}(w_0)$ must be separated. Since $|f| \to \infty$ as $z \to 0$, we must then have that the image $f^{-1}(w_0)$ consists of at most finitely many $k$ points, where $k \geq 0$.

   We may homotope a counter-clockwise route $\gamma$ on the circe $C$ to a counter-clockwise circle around $z = 0$, counter-clockwise circles around each point of $f^{-1}(w_0)$, and overlapping pairs of curves (running in opposite directions) that connect each of the former circles.

   Since $w_0$ is in the exterior of $f(C)$, we have that the winding number of $f(\gamma)$ is 0. Furthermore, we know that $f$ is orientation reversing near $z = 0$ and so it must be orientation reversing on $D$ since it is a local diffeomorphism on $D$. Therefore, we have that winding numbers of each of the above curves are $-1$.

   From the above homotopy, we then have that $0 = k - 1$, and therefore we must have that $k = 1$. So $f^{-1}(w_0)$ consists of a single point, and so f is one to one onto the part of the image in the exterior of $f(C)$.

3. Similarly, when $w_0$ in the interior of $f(C)$, we may apply the above the argument. However, now the winding number of $f(\gamma)$ is 1. Therefore, we get $1 = k - 1$, and so $k = 0$. Hence, the image of $f$ does not intersect the interior of $f(C)$.

4. Finally note that there can't be any points $z \in D$ such that $f(z) \in f(C)$, because this would imply that there are points in a neighborhood of $f(C)$ that has more than one point in their pre-image. However, this impossible by our work above.

# 9 Algebra

## 9.1 Solving Cubic Equations Using Galois Theory

### The Setup

In this example we explore using Galois Theory to explicitly solve cubic equations. Note, plan to go beyond just showing that the cubic is solvable; we will actually solve for the roots.

Consider a general cubic polynomial

$$p(x) = x^3 + Ax^2 + Bx + C. \tag{386}$$

We are considering this as a polynomial over the field $K = \mathbb{Q}(A, B, C)$, and $A, B, C$ to be general non-algebraic elements. Let the roots of $p$ be $a, b, c$, no particular relation to the capital versions that make up the coefficients of $p$.

We will use without proof that $\operatorname{Gal} K(a, b, c)/K = S_3$, the symmetric group of three elements.

Recall that the classical method of solving cubic polynomials involves solving an auxilliary quadratic equation. This quadratic equation could have roots that are complex even if the roots of our particular cubic equation are real. This is infact the original historical motivation for the construction of the imaginary number $i = \sqrt{-1}$.

Galois' idea for approaching the solutions of polynomials is manifold:

- Our aim is to reduce Galois group to the trivial group by extending the base field. It is exactly in this case that the roots must be contained in the extended base field.
- Adjoining the roots of an auxilliary equation does not by itself change the Galois group. Consider $\tilde{K} = K(\alpha_1, \alpha_2, ..., \alpha_n)$ where the $\alpha_i$ are roots of an auxialliary equation with no factors in common with the original polynomial $p(x)$. We expect that $\operatorname{Gal} \tilde{K}(a, b, c)/\tilde{K} = \operatorname{Gal} K(a, b, c)/K$.
- We adjoin roots of auxilliary equations, because they allow us to write down expressions that:
  - Are expressed in terms of the roots.
  - Are invariant under a normal sub-group of the current Galois group.
  - Are not invariant under the complete Galois Group.
  - The images of the quantity under the Galois group can be combined with primitive roots of unity to form a new quanitity such a simple power is invariant under the entire current Galois Group. This allows us to express the new quantity as a root of terms involving only the current base field $\tilde{K}$.

    This will in turn allow us to express the new quantity in terms of other known quantities: coefficients of the original polynomial $p(x)$ and

roots of our auxialliary polynomials. Recall that we don't actually know the values of $a, b, c$ yet, and so writing the expression just in terms of $a, b, c$ doesn't give us a quantity that we actually know.

For example, consider we have an extension $\tilde{K}$ of $K$ such that $\operatorname{Gal}\tilde{K}(a,b,c)/\operatorname{Gal}\tilde{K}$ is generated by the cyclic permutations of the roots $a, b, c$, and $\tilde{K}$ contains the primitive third-roots of unity. Let $\zeta_3$ be a primitive third-root of unity. Consider the quanitity $\beta = a$; this satisfies that:

- $\beta$ is NOT preserved by $\operatorname{Gal}\tilde{K}(a,b,c)/\operatorname{Gal}\tilde{K}$;
- $\beta$ IS trivially preserved by the trivial normal subgroup that only consists of the identity.
- Note that the images of $\beta$ under the Galois group are all of the roots $\{a, b, c\}$. So we form the new quantity $\alpha = a + b\zeta_3 + c\zeta_3^2$. The quantity $\alpha$ is also NOT preserved by the Galois Group; however, since the Galois Group is generated by cyclic permutations, we see that the images of $\alpha$ under the Galois Group are $\{\alpha, \alpha\zeta_3, \alpha\zeta_3^2\}$.

  So, the power $\alpha^3$ IS preserved by $\operatorname{Gal}\tilde{K}(a,b,c)/\tilde{K}$. So we may write $\alpha^3$ as an expression in $\tilde{K}$. Later we will see that in practicality, this amounts to writing $\alpha^3$ as a polynomial in quantities that are known, e.g. quantities like $C = -abc$.

**The Problem**

Use Galois theory to find the roots of the general cubic

$$p(x) = x^3 + Ax^2 + Bx + C, \tag{387}$$

where the coefficients are in $\mathbb{Q}$.

**The Solution**

Recall that we treat the coefficients as independent non-algebraic elements. So our initial base field is $K = \mathbb{Q}(A, B, C)$. We use without proof that $\operatorname{Gal} K(a,b,c)/K = S_3$, and is generated by all of the permutations of $a, b, c$.

The first normal sub-group of $S_3$ is the alternating group $A_3$, which also happens to be sub-group of cyclic permutations of $a, b, c$. To reduce the Galois group to $A_3$ we seek to extend $K$ by a quantity $\Delta$ such that $\Delta$ is preserved by $A_3$, but is NOT preserved by all of $S_3$.

We make use of the classical quantity associated with the alternating group:

$$\Delta = (a-b)(a-c)(b-c). \tag{388}$$

For any permutation $\sigma \in \S_3$ we have that $\sigma\Delta = \Delta$ exactly when $\sigma \in A_3$ and $\sigma\Delta = -\Delta$ otherwise. So $\Delta$ satisfies the properties we are interested in.

However, note that $\Delta$ is expressed in terms of the roots $a, b, c$, which are currently unknown. How can we find a quantity that is actually known to us?

Let $\zeta_2$ be the primitive square-root of unity, i.e. $\zeta_2 = -1$. We can combine powers of $\Delta$ with the images of $\Delta$ under $S_3$ to get a value that is expressed as a root of a quantity that is invariant under $A_3$. Note that the images of $\Delta$ are $\{\Delta, -\Delta\}$. So we consider the quantity

$$\alpha = \Delta + (-\Delta)\zeta_2, \tag{389}$$
$$= \Delta + (-1)^2\Delta, \tag{390}$$
$$= 2\Delta. \tag{391}$$

We have that $\alpha^2$ is preserved by $\operatorname{Gal} K(a,b,c)/\operatorname{Gal} K$, and similarly $\Delta^2$ is also preseved by the Galois Group. So we extend by the roots of the auxilliary polynomial $q(x) = x^2 - \Delta^2$. This amounts to just extending by $\Delta$ itself.

So our first extension $K(\Delta)$.

Now, we need to express $\Delta^2$ in terms of our only known quantities, the coefficients of the polynomial $A, B, C$. First, we will need how $A, B, C$ are related to the roots (despite the fact that we don't know what the roots are):

$$A = -(a+b+c), \tag{392}$$
$$B = ab + ac + bc, \tag{393}$$
$$C = -abc. \tag{394}$$

You may recognize these as being, within a change of sign, classical fundamental symmetric polynomials. In fact, our method is related to the classical fact a symmetric polynomial in $a, b, c$ can be expressed as a polynomial in the classic symmetric polynomials:

$$\sigma_1(a,b,c) = a + b + c, \tag{395}$$
$$\sigma_2(a,b,c) = ab + ac + bc, \tag{396}$$
$$\sigma_3(a,b,c) = abc. \tag{397}$$

So $A = -\sigma_1, B = \sigma_2, C = -\sigma_3$.

Since $\Delta^2 \in K$, we have that $\Delta^2$ is a rational expression of $A, B, C$. This implies that $\Delta^2$ is also a rational expression of $\sigma_1, \sigma_2, \sigma_3$. However, $\Delta^2$ is a polynomial in $a, b, c$, and $\sigma_1, \sigma_2$, and $\sigma_3$ are also polynomials in $a, b, c$. The only way this is possible for general $a, b, c$ is that $\Delta^2$ is actually a polynomial expression of $\sigma_1, \sigma_2$, and $\sigma_3$.

Now, the fact that $\Delta^2$ is a six-degree polynomial in $a, b, c$ will put restrictions on the possible polynomial expressions of $\sigma_1, \sigma_2, \sigma_3$. We have that

$$\Delta^2 = d_1\sigma_3^2 + d_2\sigma_3\sigma_2\sigma_1 + d_3\sigma_3\sigma_1^3 + d_4\sigma_2^3 + d_5\sigma_2^2\sigma_1^2 + d_6\sigma_2\sigma_1^4 + d_7\sigma_1^6, \tag{398}$$

for constants $d_i \in \mathbb{Q}$.

Now, initially this seems like a huge mess and a complete headache to solve. However, we will see that by looking at the right powers of the resulting equations, it isn't as bad as it initially looks. Consider the largest power of $c$ on the left hand side: $(a-b)^2c^4$. Compare this to the right hand side (each $\sigma_i$ can only

contribute at most one power of $c$). The largest power of $c$ for the right hand side is seen to be $d_7c^6$. So we have $d_7 = 0$.

Now find the new largest power of $c$ for the right hand side, $d_6(a+b)c^5$. Again, we must then have that $d_6 = 0$.

Repeat to get a largest power $d_3abc^4 + d_5(a+b)^2c^4$. Comparing to the left hand side, we then have that $(a-b)^2 = d_3ab + d_5(a+b)^2$. Then we get $d_5 = 1$ and $d_3 = -4$. So we have

$$\Delta^2 = d_1\sigma_3^2 + d_2\sigma_3\sigma_2\sigma_1 - 4\sigma_3\sigma_1^3 + d_4\sigma_2^3 + \sigma_2^2\sigma_1^2. \tag{399}$$

Next, look at the smallest powers of $c$. The left hand side has lowest power $(a-b)^2a^2b^2$, note that the term is independent of $c$. The right hand side has $d_4a^3b^3 + a^2b^2(a+b)^2$; note that terms like $d_1\sigma_3^2$ never contribute a term independent of $c$. So we must have that $(a-b)^2a^2b^2 = d_4a^3b^3 + a^2b^2(a+b)^2$. Writing it out, we get

$$a^4b^2 - 2a^3b^3 + a^2b^4 = d_4a^3b^3 + a^4b^2 + 2a^3b^3 + a^2b^4. \tag{400}$$

So we see that $-2 = d_4 + 2$ and so $d_4 = -4$. Hence we have that

$$\Delta^2 = d_1\sigma_3^2 + d_2\sigma_3\sigma_2\sigma_1 - 4\sigma_3\sigma_1^3 - 4\sigma_2^3 + \sigma_2^2\sigma_1^2. \tag{401}$$

Finally, let us plug in $a = b = 1$ and $c = -2$. So we have that $\Delta = 0$ and $\sigma_1 = 0$. We get that

$$0 = 4d_1 - 4(1-2-2)^3, \tag{402}$$

$$= 4d_1 + 4(27) \tag{403}$$

So we have that $d_1 = -27$. Therefore,

$$\Delta^2 = -27\sigma_3^2 + d_2\sigma_3\sigma_2\sigma_1 - 4\sigma_3\sigma_1^3 - 4\sigma_2^3 + \sigma_2^2\sigma_1^2. \tag{404}$$

Finally plug in $a = b = c = 1$ to get that $0 = -27 + 9d_2 - 4(27) - 4(27) + 3^4$. So $d_2 = 3 + 12 + 12 - 9 = 18$. So we finally get that

$$\Delta^2 = -27\sigma_3^2 + 18\sigma_3\sigma_2\sigma_1 - 4\sigma_3\sigma_1^3 - 4\sigma_2^3 + \sigma_2^2\sigma_1^2. \tag{405}$$

Then, rewriting in term of the original coefficients, we have that

$$\Delta^2 = -27C^2 + 18ABC - 4A^3C - 4B^3 + A^2B^2. \tag{406}$$

Now, we know $\Delta$ as a radical expression of the original equation and we have that $\operatorname{Gal} K(a,b,c)/K(\Delta) = A_3$, the cyclic permutations. The next normal subgroup is simply the trivial group. An expression of the roots that is invariant under the trivial group is simply one of the roots, say $\beta = a$. This has images $\{a, b, c\}$ under the cyclic permutations.

Now, let $\zeta_3$ be a primitive third-root of unity. Then we have that the quantity $\alpha = a + b\zeta_3 + c\zeta_3^2$ satisfies that $\alpha^3$ is invariant under the cyclic permutations

while $\alpha$ itself is not. The problem is that we can't use $\alpha$ yet in our extension as $\zeta_3 \notin K(a,b,c)$.

So, we must extend $K$ itself so that we are able to use $\zeta_3$. So let $\tilde{K} = K(i\sqrt{3})$; note that $\tilde{K}$ contains $\zeta_3$. We then have that $\operatorname{Gal}\tilde{K}(a,b,c)/\tilde{K}(\Delta) = \operatorname{Gal}K(a,b,c)/K(\Delta)$, generated by the same permutations of $a,b,c$.

So, we have that $\alpha^3 \in \tilde{K}(\Delta)$. However, we can save ourselves some work by going a little further in narrowing down which field it resides in. Consider the automorphism of $\tau \in \operatorname{Gal}\tilde{K}(a,b,c)/\tilde{K}$ generated by

$$\begin{cases} \tau a = a, \\ \tau b = c, \\ \tau c = b, \\ \tau i\sqrt{3} = -i\sqrt{3}. \end{cases} \tag{407}$$

Note that $\tau$ restricts to an automorphism of $\tilde{K}(\Delta)$. Furthermore, we have that $\tau\alpha = \alpha$. Therefore, $\alpha^3$ is in the fixed field of the group generated by $\tau$, i.e. $\{\mathrm{id}, \tau\}$. What is the fixed sub-field of $\tilde{K}(\Delta)$?

The automorphism $tau$ takes an element $k_1 + k_2 i\sqrt{3} + k_3\Delta + k_4\Delta i\sqrt{3}$ to $k_1 - k_2 i\sqrt{3} - k_3\Delta + k_4\Delta i\sqrt{3}$. Therefore we see that the fixed sub-field is $K(\Delta i\sqrt{3})$. Hence $\alpha^3 \in K(\Delta i\sqrt{3})$.

Since $\alpha^3$ is a polynomial in $a,b,c$ and $i\sqrt{3}$, we may express it using polynomial functions of $\sigma_1, \sigma_2$, and $\sigma_3$ instead of more general rational expressions; furthermore, we need to only use the standard expansion for $\Delta i\sqrt{3}$. Now, also note that $\alpha^3$ is a homogeneous polynomial in $a,b,c$ of degree 3, and that $\Delta = (a-b)(a-c)(b-c)$ is also homogeneous of degree 3. So we have that

$$\alpha^3 = e_1\sigma_3 + e_2\sigma_2\sigma_1 + e_3\sigma_1^3 + e_4\Delta i\sqrt{3}, \tag{408}$$

where the $e_i \in \mathbb{Q}$.

We play a similar game as before to find the $e_i$. Find the largest powers of $a$ on the left and right hand sides. For the left hand side, we have that the largest power of $a$ is $a^3$. For the right hand side, we get $e_3a^3$. Therefore, $a^3 = e_3a^3$, and so $e_3 = 1$.

Now look at the lowest powers of $a$. The left hand side gives us $(b\zeta_3 + c\zeta_3^2)^3 = b^3 + 3\zeta_3 b^2c + 3\zeta_3^2 bc^2 + c^3$. The right hand side gives us $e_2bc(b+c) + (b+c)^3 + e_4bc(b-c)i\sqrt{3} = b^3 + (e_2 + 3 + e_4 i\sqrt{3})b^2c + (e_2 + 3 - e_4 i\sqrt{3})bc^2 + c^3$.

So we get that

$$\begin{cases} 3\zeta_3 = e_2 + 3 + e_4 i\sqrt{3}, \\ 3\zeta_3^2 = e_2 + 3 - e_4 i\sqrt{3}. \end{cases} \tag{409}$$

So we get that $2e_2 + 6 = 3(\zeta_3 + \zeta_3^2) = -3$. Therefore $e_2 = -9/2$.

We also get that $e_4 2i\sqrt{3} = 3(\zeta_3 - \zeta_3^2) = 3i\sqrt{3}$. So $e_4 = 3/2$.

So we have that

$$\alpha^3 = e_1\sigma_3 - \frac{9}{2}\sigma_2\sigma_1 + \sigma_1^3 + \frac{3}{2}\Delta i\sqrt{3}. \tag{410}$$

Now, plug in $b = c = 1$ and $a = -2$. We have that $\sigma_1 = \Delta = 0$. We get

$$(-2 + \zeta_3 + \zeta_3^2)^3 = (-2 - 1)^3 = -27, \tag{411}$$

and so

$$-27 = -2e_1. \tag{412}$$

So $e_1 = 27/2$. So we finally have that

$$\alpha^3 = \frac{27}{2}\sigma_3 - \frac{9}{2}\sigma_2\sigma_1 + \sigma_1^3 + \frac{3}{2}\Delta i\sqrt{3}. \tag{413}$$

Rewriting in terms of the original coefficients, we have that

$$\alpha^3 = -\frac{27}{2}C + \frac{9}{2}AB - A^3 + \frac{3}{2}\Delta i\sqrt{3}. \tag{414}$$

So we have that $\tilde{K}(\Delta, \alpha)$ is a splitting field for our original polynomial $p(x)$. That is, $\tilde{K}(a, b, c) = \tilde{K}(\Delta, \alpha)$. However, to minimize our work in identifying the roots, we want to find $K(a) \subset \tilde{K}(a, b, c)$. Note that $\operatorname{Gal} \tilde{K}(a, b, c)/K(a)$ is of order four and generated by two automorphisms: the first sending $i\sqrt{3} \to -i\sqrt{3}$ and the second sending $b \to c$. To find the sub-field $K(a)$ we need to describe the fixed field in terms of the original polynomial coefficients.

We note that the images of $\alpha$ under this group of autormorphisms consists of $\alpha$ itself and the element $\beta = a + b\zeta_3^2 + c\zeta_3$. It is immediately clear that $\alpha + \beta = 2a - b - c$ is fixed by this group of automorphisms, but it is NOT fixed by the automorphisms generated by the cyclic permutations. So we then have that $K(a) = K(\alpha + \beta)$.

How can we compute $\beta$? Notice that the automorphisms in $\operatorname{Gal} \tilde{K}(\Delta, \alpha)/\tilde{K}(\Delta)$, i.e. those generated by the cyclic permutations, preserve the product $\alpha\beta$. Therefore, $\alpha\beta \in \tilde{K}(\Delta)$. Furthermore, it is preserved by the automorphism generated by the transposition switching $b$ and $c$. From the structure of $S_3$, this is enough to give us that $\alpha\beta$ is preserved by the entirity of $\operatorname{Gal} K(a, b, c)/K$. Therefore, $\alpha\beta \in K$.

Since $\alpha\beta$ is a homogenous polynomial in $a, b, c$ of degree 2, we have that

$$\alpha\beta = f_1\sigma_2 + f_2\sigma_1^2, \tag{415}$$

where $f_i \in \mathbb{Q}$.

For the highest power of $a$, on the left hand side we get $a^2$ and on the right hand side we get $f_2a^2$. Therefore $f_2 = 1$.

Now, look at the lowest power of $a$. On the left hand side we get $b^2 + (\zeta_3 + \zeta_3^2)bc + c^2$. On the right hand side we get $f_1bc + (b + c)^2$. So we get that $-1 = f_1 + 2$, and that $f_1 = -3$. Therefore,

$$\alpha\beta = -3\sigma_2 + \sigma_1^2. \tag{416}$$

In terms of the original coefficients,

$$\alpha\beta = -3B + A^2. \tag{417}$$

Now, we wish to find one the root $a \in K(a) = K(\alpha + \beta)$. Note that $a$ is itself a homogeneous linear polynomial of degree one in the roots, and $\alpha + \beta$ is homogeneous of degree 1. So we are left with the following possibilities:

$$a = g_1\sigma_1 + g_2(\alpha + \beta), \tag{418}$$

where $g_i \in \mathbb{Q}$.

Next, realize that $\alpha + \beta = 2a - b - c$. So we see that

$$a = \frac{1}{3}\sigma_1 + \frac{1}{3}(\alpha + \beta). \tag{419}$$

Rewriting in terms of the coefficients of the original polynomial we get

$$a = -\frac{1}{3}A + \frac{1}{3}(\alpha + \beta). \tag{420}$$

To get the rest of the roots, simply apply the automorphism of $K(a, b, c)$ generated by cyclic permutations. We get that

$$\begin{cases} a = -\frac{1}{3}A + \frac{1}{3}(\alpha + \beta), \\ b = -\frac{1}{3}A + \frac{1}{3}(\zeta_3^{-1}\alpha + \zeta_3\beta), \\ c = -\frac{1}{3}A + \frac{1}{3}(\zeta_3\alpha + \zeta_3^{-1}\beta). \end{cases} \tag{421}$$

Now, in practice we can't be sure we have selected the precise $\Delta$ and $\alpha$ we have used above. For example, we can only by sure that we select $\hat{\Delta} \in \{-\Delta, \Delta\}$. Similarly, depending on our choice of $\hat{\Delta}$, we select $\hat{\alpha} \in \{\alpha, \zeta_3\alpha, \zeta_3^2\alpha, \beta, \zeta_3\beta, \zeta_3^2\beta\}$. However, everything will be okay as long as we enforce the following:

$$\hat{\Delta}^2 = -27C^2 + 18ABC - 4A^3C - 4B^3 + A^2B^2, \tag{422}$$

$$\hat{\alpha}^3 = -\frac{27}{2}C + \frac{9}{2}AB - A^3 + \frac{3}{2}\hat{\Delta}i\sqrt{3}, \tag{423}$$

$$\hat{\alpha}\hat{\beta} = -3B + A^2, \tag{424}$$

$$a = -\frac{1}{3}A + \frac{1}{3}(\hat{\alpha} + \hat{\beta}), \tag{425}$$

$$b = -\frac{1}{3}A + \frac{1}{3}(\zeta_3^{-1}\hat{\alpha} + \zeta_3\hat{\beta}), \tag{426}$$

$$c = -\frac{1}{3}A + \frac{1}{3}(\zeta_3\hat{\alpha} + \zeta_3^{-1}\hat{\beta}). \tag{427}$$

# References


[1] William Branson. Solving the cubic with cardano. *Convergence*, 2013.

[2] Florian Cajori. History of the exponential and logarithmic concepts. *The American Mathematical Monthly*, 1913.

[3] Herbert Federer. *Geometric Measure Theory*.

[4] Robert Gilmore. *Lie Groups, Physics, and Geometry.*

[5] Richard Hamilton. The inverse function theorem of nash and moser. *Bulletin of the American Mathematical Society*, 1982.

[6] Richard Hamming. *Numerical Methods for Scientists and Engineers.* Dover, 1987.

[7] Steven Krantz and Harold Parks. *Geometric Integration Theory.*

[8] Gregory Moore. The emergence of open sets, closed sets, and limit points in analysis and topology. 2008.

[9] Robert Osserman. *A Survey of Minimal Surfaces.*

[10] Valentin Ovsienko and Sergei Tabachnikov. What is... the schwarzian derivative? *Bulletin of the AMS*, 2009.

[11] Jaume Paradis; Josep Pla; and Pelegri Viader. Fermat's treatise on quadrature: A new reading. 2004.

[12] V. Frederick Rickey and Philip M. Tuchinsky. An application of geography to mathematics: History of the integral of the secant. *Mathematics Magazine*, 1980.

[13] Saul Stahl. The evolution of the normal distribution. *Mathematics Magazine*, 2006.

[14] Dirk Jan Struik. *Source Book in Mathematics.*

[15] Seth Sullivant. *Algebraic Statisitcs.*

[16] Jeff Suzuki. The lost calculus (1637 - 1670): Tangency and optimization without limits. *Mathematics Magazine*, 2005.