Interesting Examples in Undergraduate Mathematics

Matthew McGonagle

June 27, 2018

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 One Variable Differential Calculus

2.1 Gauss and the Gauss Distribution

History

A good reference on the history of the gaussian distribution is [3].

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

$$P(x_1, x_2, ..., x_n|X_0) = p(x_1 - X_0)p(x_2 - X_0)...p(x_n - X_0). \qquad (1)$$

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

$$L(X_0) = p(x_1 - X_0)p(x_2 - X_0)...p(x_n - X_0). \tag{2}$$

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

$$h(X_0) = \log(p(x_1 - X_0)) + \log(p(x_2 - X_0)) + ... + \log(p(x_n - X_0)). \tag{3}$$

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

$$0 = h'(X_0) = -\left(\frac{p'(x_1 - X_0)}{p(x_1 - X_0)} + \frac{p'(x_2 - X_0)}{p(x_2 - X_0)} + ... + \frac{p'(x_n - X_0)}{p(x_n - X_0)}\right). \tag{4}$$

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

$$L(y) = p(x_1 - y)p(x_2 - y)...p(x_n - y). \tag{5}$$

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

$$0 = f\left(-\frac{n-1}{n}r\right) + (n-1)f\left(\frac{1}{n}r\right). \tag{6}$$

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

$$f\left(\frac{n-1}{n}r\right) = (n-1)f\left(\frac{1}{n}r\right). \tag{7}$$

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

$$\frac{1}{n-1}f(r_0) = f\left(\frac{r_0}{n-1}\right). \tag{8}$$

Now consider $0 < k \leq n+1$. Note that $\frac{k}{n}r$ is in the interval too, and now apply 6 for $n- > k$ and $r- > \frac{k}{n}r$, we get

$$f\left(\frac{k-1}{n}r\right) = (k-1)f\left(\frac{1}{n}r\right). \tag{9}$$

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

$$f\left(\frac{m}{n}r_0\right) = \frac{m}{n}f(r_0). \tag{10}$$

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

$$\frac{p'(x)}{p(x)} = k, \tag{11}$$

locally around $x = 0$. Integrating we get

$$\log(p(x)) = \frac{k}{2}x^2 + C. \tag{12}$$

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

$$\int\limits_{-\infty}^{\infty} e^{-Bx^2}\,dx = \sqrt{\frac{\pi}{B}}. \tag{13}$$

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

$$p(x) = \sqrt{\frac{B}{\pi}}e^{-Bx^2}. \tag{14}$$

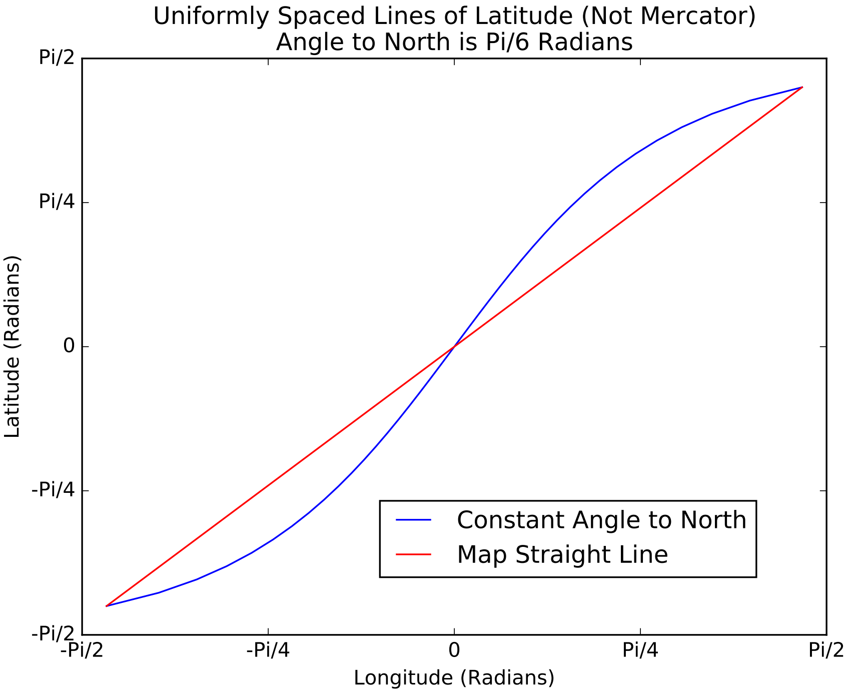
3 One Variable Integral Calculus

3.1 The Mercator Map and the Integral of Secant

Historical Motivation

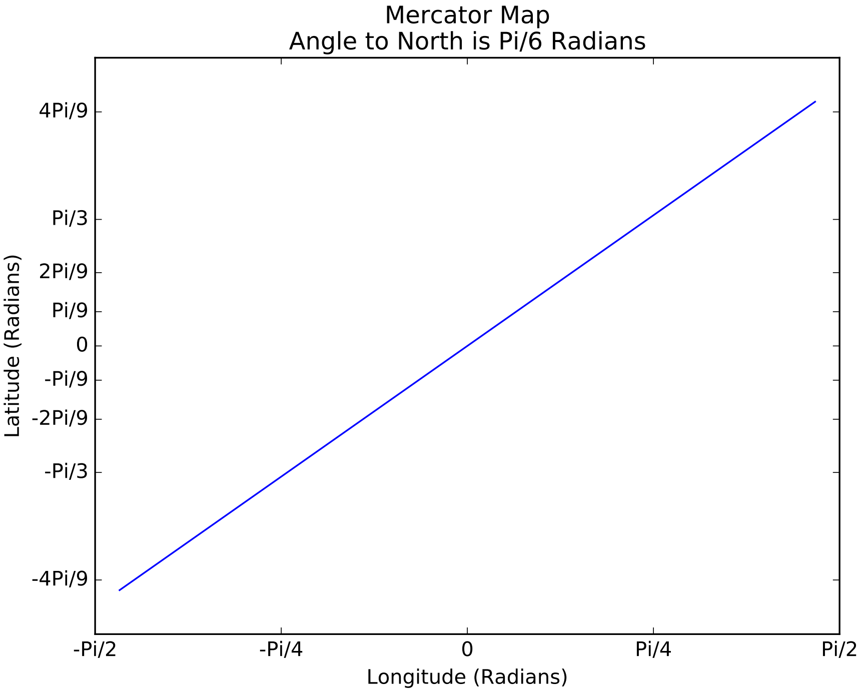
A great reference for the history of the integral of the secant function and its relation to the Mercator map is [2]. We will give a brief overview here.

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

$$\int\limits_0^x \sec(u)\,du. \tag{15}$$

The Solution

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

$$\begin{aligned}\int\limits_0^x \frac{1}{\cos(u)}\,du &= \int\limits_0^x \frac{\cos(u)}{\cos^2(u)}\,du, &(16)\\ &= \int\limits_0^x \frac{\cos(u)}{1-\sin^2(u)}\,du. &(17)\end{aligned}$$

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:

$$\int\limits_0^{\sin(x)} \frac{1}{1-w^2}\,dw. \tag{18}$$

Now, we use partial fractions:

$$\begin{aligned}\frac{1}{1-w^2} &= \frac{1}{(1-w)(1+w)}, &(19)\\ &= \frac{A}{1-w} + \frac{B}{1+w}. &(20)\end{aligned}$$

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

$$\begin{cases} A + B = 1, \\ A - B = 0. \end{cases} \tag{21}$$

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

Therefore, our integral becomes

$$\begin{aligned}\int\limits_0^{\sin(x)} \frac{1}{2(1-w)} + \frac{1}{2(1+w)}\,dw &= \frac{1}{2}\log\left(\frac{1+w}{1-w}\right)\Bigg|_0^{\sin(x)}, &(22)\\ &= \frac{1}{2}\log\left(\frac{1+\sin(x)}{1-\sin(x)}\right). &(23)\end{aligned}$$

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

$$\begin{aligned}
\frac{1}{2}\log\left(\frac{1+\sin(x)}{1-\sin(x)}\right) &= \log\sqrt{\frac{1+\sin(x)}{1-\sin(x)}}, && (24)\\
&= \log\sqrt{\frac{(1+\sin(x))^2}{1-\sin^2(x)}}, && (25)\\
&= \log\left(\frac{1+\sin(x)}{\cos(x)}\right), && (26)\\
&= \log(\sec(x)+\tan(x)). && (27)
\end{aligned}$$

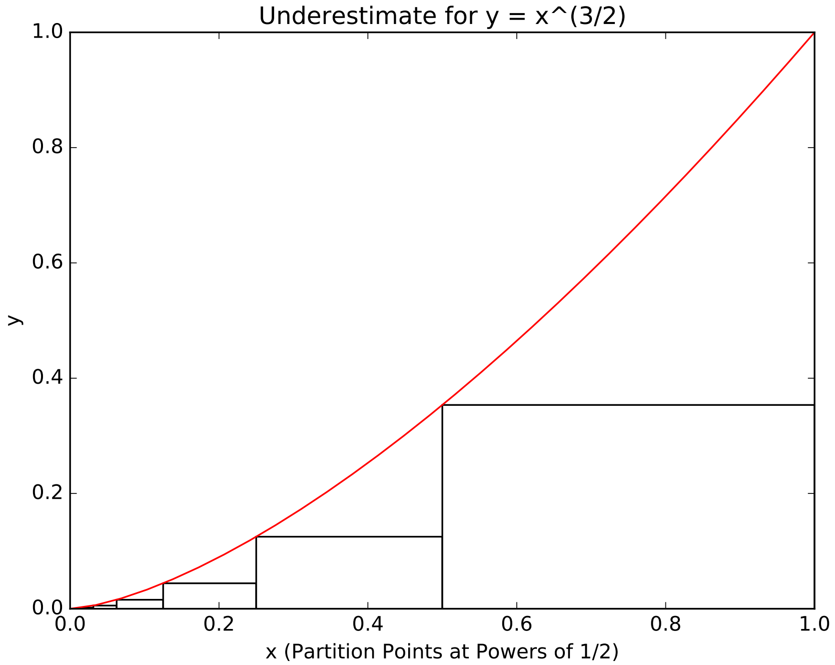
3.2 Fermat's Method of Integrating Powers of x

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

$$b + b^2 + ... + b^n = \frac{b - b^{n+1}}{1 - b}. \tag{28}$$

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

$$b + b^2 + b^3 + ... = \frac{b}{1 - b}. \tag{29}$$

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

The area of the rectangles are:

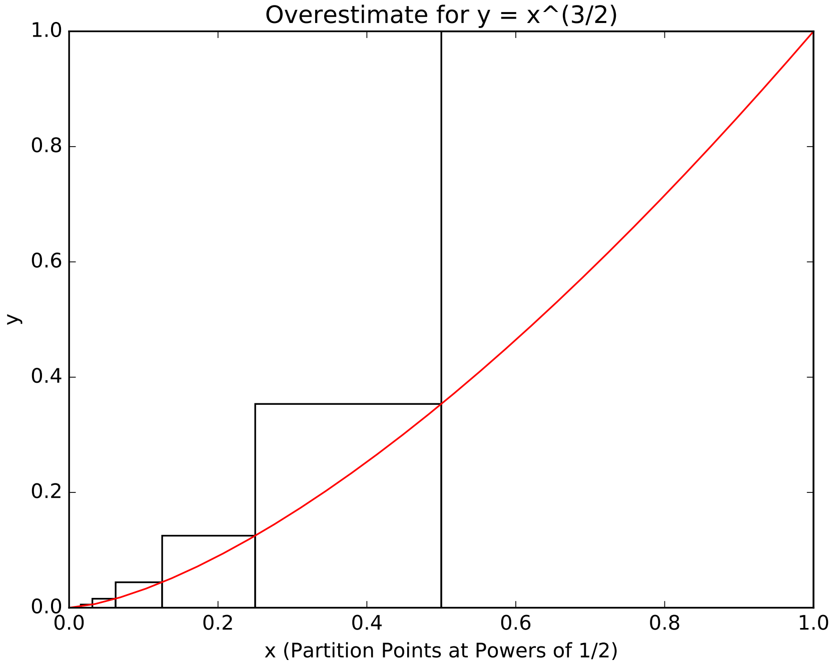
- The area of the rectangle on $1/2 \leq x \leq 1$ is $(1/2)^{3/2}(1/2) = (1/2)^{5/2}$.
- The area of the rectangle on $1/4 \leq x \leq 1/2$ is $(1/4)^{3/2}(1/4) = (1/4)^{5/2}$.
- The area of the rectangle on $1/8 \leq x \leq 1/4$ is $(1/8)^{3/2}(1/8) = (1/8)^{5/2}$.
- Etc.

So we have that the total area is

$$(1/2)^{5/2} + (1/4)^{5/2} + (1/8)^{5/2} + ... = (1/2)^{\frac{5}{2}} + \left((1/2)^{\frac{5}{2}}\right)^2 + \left((1/2)^{\frac{5}{2}}\right)^3 + ... \quad (30)$$

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:

$$(1-b)b^{\frac{3}{2}} + (1-b)b\left(b^{\frac{3}{2}}\right)^2 + (1-b)b^2\left(b^{\frac{5}{2}}\right)^3 + ... = \frac{1-b}{b}\left(b^{\frac{5}{2}} + \left(b^{\frac{5}{2}}\right)^2 + \left(b^{\frac{5}{2}}\right)^2 + ...\right), \tag{31}$$

$$= \frac{1-b}{b}\frac{b^{\frac{5}{2}}}{1-b^{\frac{5}{2}}}, \tag{32}$$

$$= b^{\frac{3}{2}}\frac{1-b}{1-b^{\frac{5}{2}}}. \tag{33}$$

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

$$b^{\frac{3}{2}}\frac{1+b^{\frac{1}{2}}}{1+b^{\frac{1}{2}}+b^{\frac{2}{2}}+b^{\frac{3}{2}}+b^{\frac{4}{2}}}. \tag{34}$$

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

$$\frac{1+1}{1+1+1+1+1} = \frac{2}{5}. \tag{35}$$

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 Multivariable Differential Calculus

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

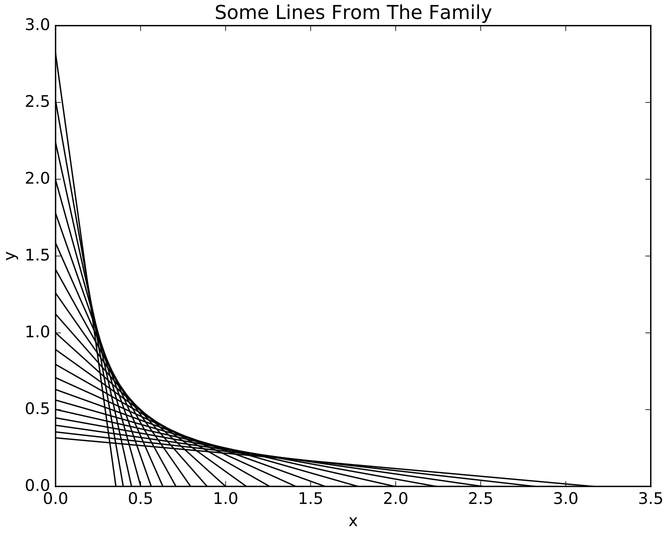
4.1 Envelopes

In the simplist cases, the **envelope** of a family $\mathfrak{F}$ of curves is a curve γ that is in some sense extremal to the entire family of curves. What is often the case, is that every point of the curve γ 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.

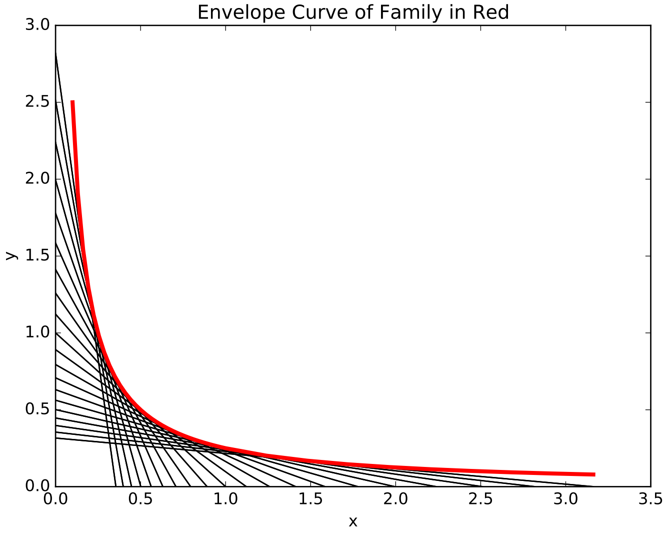
4.1.1 The Hyperbola as an Envelope

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 γ, 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 γ. 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

$$\begin{aligned} 0 &= \frac{\partial g}{\partial s}, && (36) \\ &= -\frac{x}{s^2} + y && (37) \end{aligned}$$

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

$$\sqrt{xy} + \sqrt{xy} - 1 = 0. \tag{38}$$

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.

4.2 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

$$f(x) = \begin{cases} x^2 \sin(\frac{1}{x}) & x \neq 0, \\ 0 & x = 0. \end{cases} \tag{39}$$

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:

$$h'(x) = \lambda \frac{d}{dx}\left(g\left(\frac{x}{\lambda}\right)\right), \tag{40}$$

$$= \lambda g'\left(\frac{x}{\lambda}\right)\frac{1}{\lambda}, \tag{41}$$

$$= g'\left(\frac{x}{\lambda}\right). \tag{42}$$

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 $(x, y) \neq (0, 0)$:

$$\frac{\partial u}{\partial x} = 2xh + (x^2 + y^2)h'\left(\frac{1}{x^2 + y^2} - \frac{2x^2}{(x^2 + y^2)^2}\right), \tag{43}$$

$$= 2xh + h' - h'\frac{x^2}{x^2 + y^2}. \tag{44}$$

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

$$\frac{\partial u}{\partial y} = 2yh + (x^2 + y^2)h' \frac{2xy}{(x^2 + y^2)^2}, \tag{45}$$

$$= 2yh + h' \frac{2xy}{x^2 + y^2}. \tag{46}$$

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

$$h' \left(\frac{x}{x^2 + y^2} \right) = h' \left(\frac{\sqrt{x}}{x^{3/2} + 1} \right), \tag{47}$$

$$\to h'(0). \tag{48}$$

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

$$\frac{x^2}{x^2 + y^2} = \frac{x^{3/2}}{x^{3/2} + 1}, \tag{49}$$

$$\to 0. \tag{50}$$

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

$$u(x, y) = (x^2 + y^2)h \left(\frac{x}{x^2 + y^2} \right), \tag{51}$$

$$= (x^2 + y^2) \frac{x}{(x^2 + y^2)(1 + x^2(x^2 + y^2)^{-2})}, \tag{52}$$

$$= \frac{x(x^2 + y^2)^2}{(x^2 + y^2)^2 + x^2}. \tag{53}$$

4.3 Maximizing Likelihood for a Three Step Markov Process

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

$$P(X_3 = k|X_1 = i, X_2 = j) = P(X_3 = k|X_2 = j). \tag{54}$$

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

$$\frac{p_{ijk}}{\sum_\gamma p_{ij\gamma}} = \frac{\sum_\alpha p_{\alpha jk}}{\sum_{\alpha,\gamma} p_{\alpha j\gamma}}. \tag{55}$$

We can rewrite this as

$$p_{ijk} \sum_{\alpha,\gamma} p_{\alpha j\gamma} = \left(\sum_\alpha p_{\alpha jk} \right) \left(\sum_\gamma p_{ij\gamma} \right). \tag{56}$$

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

$$p_{ijk} \sum_\alpha p_{\alpha jk} + p_{ijk} \sum_\alpha p_{\alpha j\bar{k}} = p_{ijk} \sum_\alpha p_{\alpha jk} + p_{ij\bar{k}} \sum_\alpha p_{\alpha jk}. \tag{57}$$

Cancelling terms we get

$$p_{ijk} \sum_\alpha p_{\alpha j\bar{k}} = p_{ij\bar{k}} \sum_\alpha p_{\alpha jk}. \tag{58}$$

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

$$p_{ijk}p_{ij\bar{k}} + p_{ijk}p_{\bar{i}j\bar{k}} = p_{ij\bar{k}}p_{ijk} + p_{ij\bar{k}}p_{\bar{i}jk}. \tag{59}$$

Again, cancelling terms we get

$$p_{ijk}p_{\bar{i}j\bar{k}} = p_{ij\bar{k}}p_{\bar{i}jk}. \tag{60}$$

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 $i \to \bar{i}$; 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 $k \to \bar{k}$, we can choose $k = 0$. Both of these choices give

$$p_{0j0}p_{1j1} = p_{0j1}p_{1j0}, \tag{61}$$

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:

$$\begin{cases} p_{ijk} \geq 0, \\ \sum_{\alpha,\beta,\gamma} p_{\alpha\beta\gamma} = 1, \\ p_{000}p_{101} = p_{001}p_{100}, \\ p_{010}p_{111} = p_{011}p_{110}. \end{cases} \tag{62}$$

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

$$l(\vec{p}) = \prod_{\alpha,\beta,\gamma} (p_{\alpha\beta\gamma})^{n_{\alpha\beta\gamma}}. \tag{63}$$

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

$$L(\vec{p}) = \sum_{\alpha,\beta,\gamma} n_{\alpha\beta\gamma} \log(p_{\alpha\beta\gamma}). \tag{64}$$

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):

$$\begin{cases} \text{maximize } L(\vec{p}) = \sum_{\alpha,\beta,\gamma} n_{\alpha\beta\gamma} \log p_{\alpha\beta\gamma}, & \\ \sum_{\alpha,\beta,\gamma} p_{\alpha\beta\gamma} = 1, & \\ p_{0j0}p_{1j1} - p_{0j1}p_{1j0} = 0, & \text{for } j \in \{0, 1\}. \end{cases} \tag{65}$$

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

$$f(\vec{p}) := \sum_{\alpha,\beta,\gamma} p_{\alpha\beta\gamma}, \tag{66}$$

$$g_j(\vec{p}) := p_{0j0}p_{1j1} - p_{0j1}p_{1j0}. \tag{67}$$

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

$$\frac{\partial L}{\partial p_{ijk}} = \frac{n_{ijk}}{p_{ijk}}, \tag{68}$$

$$\frac{\partial f}{\partial p_{ijk}} = 1, \tag{69}$$

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

$$\frac{\partial g_j}{\partial p_{ijk}} = (-1)^{i+k} p_{\bar{i}j\bar{k}}, \tag{70}$$

$$\frac{\partial g_j}{\partial p_{i\bar{j}k}} = 0. \tag{71}$$

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

$$\frac{n_{ijk}}{p_{ijk}} = \lambda + \tau_j (-1)^{i+k} p_{\bar{i}j\bar{k}}, \tag{72}$$

for each (i, j, k). Note how coefficient τ_0 and variables p_{i0k} are decoupled from τ_1 and p_{i1k}; that is no equation has some variable or coefficient from both sets. However, λ is coupled to all of them.

Now, multiply through to get

$$n_{ijk} = \lambda p_{ijk} + \tau_j(-1)^{i+k} p_{ijk} p_{\bar{i}j\bar{k}}. \tag{73}$$

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

$$n_{ij\bar{k}} = \lambda p_{ij\bar{k}} - \tau_j(-1)^{i+k} p_{ij\bar{k}} p_{\bar{i}jk}. \tag{74}$$

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

$$n_{ijk} + n_{ij\bar{k}} = \lambda(p_{ijk} + p_{ij\bar{k}}). \tag{75}$$

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

$$\frac{n_{ijk} + n_{\bar{i}jk}}{p_{ijk} + p_{\bar{i}jk}} = \frac{n_{ijk} + n_{ij\bar{k}}}{p_{ijk} + p_{ij\bar{k}}} = \lambda. \tag{76}$$

Let us concentrate on the following constraints:

$$\begin{cases} \sum_{\alpha,\beta,\gamma} p_{\alpha\beta\gamma} = 1, \\ \frac{n_{0jk}+n_{1jk}}{p_{0jk}+p_{1jk}} = \frac{n_{ij0}+n_{ij1}}{p_{ij0}+p_{ij1}} = \lambda, \\ p_{0j0}p_{1j1} = p_{0j1}p_{1j0}. \end{cases} \tag{77}$$

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 ±1. In fact, it is simple to correctly guess the eigenvectors of S. This leads us to an intial change of variables

$$q_{+jk} := p_{0jk} + p_{1jk}, \tag{78}$$

$$q_{-jk} := p_{0jk} - p_{1jk}. \tag{79}$$

Note that we have used subscripts of ± to indicate whether the variable corresponds to eigenvalue ±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

$$r_{+j+} := p_{0j0} + p_{1j0} + p_{0j1} + p_{1j1}, \tag{80}$$

$$r_{+j-} := p_{0j0} + p_{1j0} - p_{0j1} - p_{1j1}, \tag{81}$$

$$r_{-j+} := p_{0j0} - p_{1j0} + p_{0j1} - p_{1j1}, \tag{82}$$

$$r_{-j-} := p_{0j0} - p_{1j0} - p_{0j1} + p_{1j1}, \tag{83}$$

$$\tag{84}$$

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

$$r_{+0+} + r_{+1+} = 1. \tag{85}$$

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

$$m_{+j+} + m_{-j+} = \lambda(r_{+j+} + r_{-j+}), \tag{86}$$

$$m_{+j+} - m_{-j+} = \lambda(r_{+j+} - r_{-j+}), \tag{87}$$

$$\tag{88}$$

So we get

$$m_{+j+} = \lambda r_{+j+}, \tag{89}$$

$$m_{-j+} = \lambda r_{-j+}. \tag{90}$$

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):

$$m_{+j-} = \lambda r_{+j-}. \tag{91}$$

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

$$N = \lambda(r_{+0+} + r_{+1+}), \tag{92}$$

$$= \lambda. \tag{93}$$

So we get

$$\begin{aligned} r_{+j+} &= \frac{m_{+j+}}{N}, && (94)\\ r_{-j+} &= \frac{m_{-j+}}{N}, && (95)\\ r_{+j-} &= \frac{m_{+j-}}{N}. && (96) \end{aligned}$$

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

$$g_j(\vec{p}) = Ar_{+j+}r_{+j+} + ... \tag{97}$$

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

$$\begin{aligned} -(Ar_{+j+}r_{+j+} + ...) &= -g_j(\vec{p}), && (98)\\ &= g_j(S\vec{p}), && (99)\\ &= Ar_{+j+}r_{+j+} \pm .... && (100) \end{aligned}$$

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

$$g_j(\vec{p}) = Br_{+j+}r_{-j-} + Cr_{+j-}r_{-j+}, \tag{101}$$

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

$$\begin{aligned} st &= p^0_{0j0}p^0_{1j1} - p^0_{0j1}p^0_{1j0}, && (102)\\ r^0_{+j+} &= s + t, && (103)\\ r^0_{+j-} &= s - t, && (104)\\ r^0_{-j+} &= s - t, && (105)\\ r^0_{-j-} &= s + t. && (106) \end{aligned}$$

So we get

$$\begin{aligned} Ar^0_{+j+}r^0_{-j-} + Br^0_{+j-}r^0_{-j+} &= A(s+t)^2 + B(s-t)^2, && (107)\\ &= (A+B)s^2 + 2(A-B)st + (A+B)t^2. && (108) \end{aligned}$$

So we get

$$\begin{aligned} A + B &= 0, && (109)\\ 2(A - B) &= 1 && (110) \end{aligned}$$

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

$$g_j = \frac{r_{+j+}r_{-j-} - r_{+j-}r_{-j+}}{4}. \qquad (111)$$

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

$$r_{-j-} = \frac{m_{+j-}m_{-j+}}{Nm_{+j+}}. \qquad (112)$$

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

$$\begin{align} p_{0j0} &= \frac{r_{+j+} + r_{+j-} + r_{-j+} + r_{-j-}}{4}, \tag{113} \\ &= \frac{m_{+j+}(m_{+j+} + m_{+j-} + m_{-j+}) + m_{+j-}m_{-j+}}{4Nm_{+j+}}, \tag{114} \\ &= \frac{m_{+j+}(m_{+j+} + m_{+j-}) + m_{-j+}(m_{+j+} + m_{+j-})}{4Nm_{+j+}}, \tag{115} \\ &= \frac{(m_{+j+} + m_{-j+})(m_{+j+} + m_{+j-})}{4Nm_{+j+}}, \tag{116} \\ &= \frac{4(n_{0j0} + n_{0j1})(n_{0j0} + n_{1j0})}{4Nm_{+j+}}, \tag{117} \\ &= \frac{\left(\sum_\gamma n_{0j\gamma}\right)\left(\sum_\alpha n_{\alpha j0}\right)}{N\sum_{\alpha,\gamma} n_{\alpha j\gamma}}. \tag{118} \end{align}$$

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

$$\begin{align} p_{1j0} &= \tilde{p}_{0j0}, \tag{119} \\ &= \frac{\left(\sum_\gamma \tilde{n}_{0j\gamma}\right)\left(\sum_\alpha \tilde{n}_{\alpha j0}\right)}{N\sum_{\alpha,\gamma} \tilde{n}_{\alpha j\gamma}}, \tag{120} \\ &= \frac{\left(\sum_\gamma n_{1j\gamma}\right)\left(\sum_\alpha n_{\alpha j0}\right)}{N\sum_{\alpha,\gamma} n_{\alpha j\gamma}}. \tag{121} \end{align}$$

Similarly we can compute every p_{ijk} as

$$p_{ijk} = \frac{\left(\sum_\gamma n_{ij\gamma}\right)\left(\sum_\alpha n_{\alpha jk}\right)}{N\sum_{\alpha,\gamma} n_{\alpha j\gamma}}. \tag{122}$$

5 Multivariable Integral Calculus

5.1 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

  $$\int_0^1 \int_0^1 u(x, y) \, dx \, dy, \tag{123}$$

  and

  $$\int_0^1 \int_0^1 u(x, y) \, dy \, dx, \tag{124}$$

  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

$$f(x, y) = \frac{-1}{(x + y)^2} \tag{125}$$

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

$$u(x,y) = \frac{1}{(Ax+By)^2} - \frac{1}{(x+y)^2}. \tag{126}$$

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

$$u(x,y) = \frac{1}{(2x+y/2)^2} - \frac{1}{(x+y)^2}. \tag{127}$$

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

$$\int\limits_0^1 \frac{1}{(2x+y/2)^2} - \frac{1}{(x+y)^2}\,dx = \left.\frac{1}{x+y} - \frac{1}{2(2x+y/2)}\right|_{x=0}^{1}, \tag{128}$$

$$= \frac{1}{1+y} - \frac{1}{y} - \frac{1}{4+y} + \frac{1}{y}, \tag{129}$$

$$= \frac{1}{1+y} - \frac{1}{4+y}. \tag{130}$$

So we get that

$$\int_0^1 \left(\int_0^1 \frac{1}{(2x+y/2)^2} - \frac{1}{(x+y)^2}\, dx \right) dy = \int_0^1 \frac{1}{1+y} - \frac{1}{4+y}\, dy, \tag{131}$$

$$= \log(1+y) - \log(4+y)\big|_{y=0}^{1}, \tag{132}$$

$$= \log(2) - \log(1) - \log(5) + \log(4), \tag{133}$$

$$= 3\log(2) - \log(5). \tag{134}$$

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

$$\int_0^1 \frac{1}{(2x+y/2)^2} - \frac{1}{(x+y)^2}\, dy = \frac{1}{x+y} - \frac{2}{2x+y/2}\bigg|_{y=0}^{1}, \tag{135}$$

$$= \frac{1}{x+1} - \frac{1}{x} - \frac{2}{2x+1/2} + \frac{2}{2x}, \tag{136}$$

$$= \frac{1}{x+1} - \frac{2}{2x+1/2}. \tag{137}$$

So we have that

$$\int_0^1 \left(\int_0^1 \frac{1}{(2x+y/2)^2} - \frac{1}{(x+y)^2}\, dy \right) dx = \int_0^1 \frac{1}{x+1} - \frac{2}{2x+1/2}\, dx, \tag{138}$$

$$= \log(x+1) - \log(2x+1/2)\big|_{x=0}^{1}, \tag{139}$$

$$= \log(2) - \log(1) - \log(5/2) + \log(1/2), \tag{140}$$

$$= \log(2) - \log(5). \tag{141}$$

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

5.2 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 $1 \leq x < 10$,

$$P(X \leq x) = \int\limits_1^x \frac{1}{s \log(10)} ds. \tag{142}$$

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

$$P(Y \leq y) = \int\limits_0^y f(s) ds. \tag{143}$$

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

$$h(z) = \frac{1}{z \log(10)}. \tag{144}$$

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 [1].

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.

$$\begin{aligned} H(z) &= P(1 \leq s(XY) \leq z) && (145) \\ &= \int_1^z h(z) dz, && (146) \end{aligned}$$

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

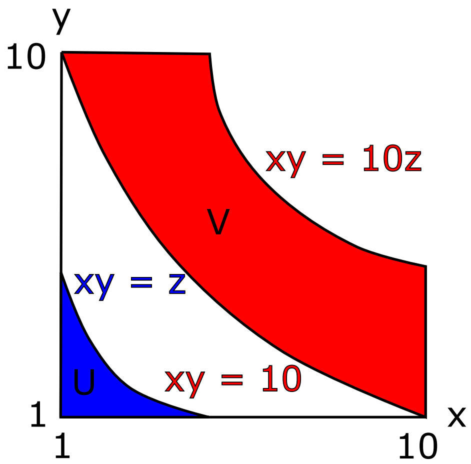
Similary, we let $F(y)$ be the cumulative distribution for $f(y)$.

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

$$\{1 \leq s(XY) \leq z\} = \{1 \leq XY \leq z\} \cup \{10 \leq XY \leq 10z\}. \tag{147}$$

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

$$H(z)=\iint\limits_{U\cup V} p(x,y)\,dA_{xy}, \tag{148}$$

$$=\frac{1}{\log(10)}\iint\limits_{U\cup V}\frac{f(y)}{x}dA_{xy}. \tag{149}$$

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

$$H(z)=\frac{1}{\log(10)}\iint\limits_{U\cup V}\frac{f(y)}{x}\,dA_{xy}, \tag{150}$$

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

$$h(z)=\frac{1}{z\log(10)}, \tag{151}$$

for $1\le z\le 10$.

The Solution

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

$$H_U(z)\coloneqq\frac{1}{\log(10)}\iint\limits_{U}\frac{f(y)}{x}\,dA_{xy}. \tag{152}$$

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

$$H_U(z) = \frac{1}{\log(10)} \int_1^z \left(\int_1^{z/x} \frac{f(y)}{x} dy \right) dx. \tag{153}$$

Now, we have that

$$\int_1^{x/z} \frac{f(y)}{x} dy = \frac{1}{x} F\left(\frac{z}{x}\right). \tag{154}$$

So,

$$H_U = \frac{1}{\log(10)} \int_1^z \frac{1}{x} F\left(\frac{z}{x}\right) dx. \tag{155}$$

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

$$H_U'(z) = 0 + \frac{1}{\log(10)} \int_1^z \frac{1}{x^2} f\left(\frac{z}{x}\right) dx, \tag{156}$$

$$= \frac{1}{z\log(10)} \int_1^z f(u) du, \tag{157}$$

$$= \frac{F(z)}{z\log(10)}. \tag{158}$$

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

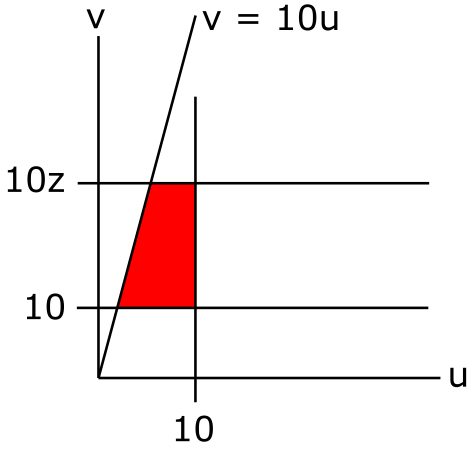
$$H_V(z) \coloneqq \frac{1}{\log(10)} \iint_V \frac{f(y)}{x} \, dA_{xy}. \tag{159}$$

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

$$\frac{\partial(u,v)}{\partial(x,y)} = x. \tag{160}$$



So we get that

$$\frac{\partial(x,y)}{\partial(u,v)} = \frac{1}{x}, \tag{161}$$

$$= \frac{1}{u}. \tag{162}$$

So we see that

$$H_V(z) = \frac{1}{\log(10)} \iint\limits_{V} \frac{1}{u^2} f\left(\frac{v}{u}\right) dA_{uv}, \tag{163}$$

$$= \frac{1}{\log(10)} \int\limits_{10}^{10z} \left(\int\limits_{v/10}^{10} \frac{1}{u^2} f\left(\frac{v}{u}\right) du \right) dv. \tag{164}$$

For the inner integral we get

$$\int\limits_{v/10}^{10} \frac{1}{u^2} f\left(\frac{v}{u}\right) du = \frac{1}{v} \int\limits_{v/10}^{10} f(w) dw, \tag{165}$$

$$= \frac{1}{v}\left(F(10) - F\left(\frac{v}{10}\right)\right), \tag{166}$$

$$= \frac{1}{v}\left(1 - F\left(\frac{v}{10}\right)\right). \tag{167}$$

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

So then

$$H_V(z) = \frac{1}{\log(10)} \int\limits_{10}^{10z} \frac{1}{v}\left(1 - F\left(\frac{v}{10}\right)\right) dv, \tag{168}$$

and therefore

$$H'_V(z) = \frac{10}{10z\log(10)}\left(1 - F(z)\right), \tag{169}$$

$$= \frac{1}{z\log(10)} - \frac{1}{z\log(10)}F(z). \tag{170}$$

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

$$h(z) = \frac{1}{z\log(10)}. \tag{171}$$

References

[1] Richard Hamming. *Numerical Methods for Scientists and Engineers*. Dover, 1987.

[2] V. Frederick Rickey and Philip M. Tuchinsky. An application of geography to mathematics: History of the integral of the secant. *Mathematics Magazine*, 1980.

[3] Saul Stahl. The evolution of the normal distribution. *Mathematics Magazine*, 2006.