Authentic Applications

Editors: Kseniya Garaschuk and Anton Mosunov

August 2022

Contents

1   Introduction

We often tell the students that math is everywhere, but they can rarely recognize it because they do not know what to look for: there are no polynomials hanging out in mid-air and no ladders sliding down random walls. Students do not put themselves in the shoes of an observer, a biologist collecting data, a mathematician set out to prove a theorem. Their math always came from books with carefully constructed examples and mostly integer solutions, with applications that are never personally relevant, relatable or significant enough. So students learn to ignore the make-up of the problems or really any words surrounding the formulas and practice choosing (or guessing) the correct algorithm to apply.

We want to challenge this view of the subject by discussing mathematics in the context of current events, important phenomena and interesting information relevant to students' lives. The applications you find in this volume are authentic in the sense that they are based in the "real-world" first, with math derived out of the context, rather than the other way around. Notable exceptions include Section 10 that involves an imaginary but charming application for a math majors course and Section 11 that describes various open education technologies, such as CoCalc and PreTeXt, that can be used to teach applied content to students.

This document is a result of Authentic Applications session organized by Kseniya Garaschuk and Anton Mosunov at the 2022 Canadian Mathematical Society Summer Meeting and the consequent call for contributions. We present to you this collection of applications suitable for a wide variety for math and stats courses.

<div align="right">Kseniya and Anton</div>

2 Mono-Mono Twins: Modelling the Optimal Length of Pregnancy

Eric Cytrynbaum
cytryn@math.ubc.ca
University of British Columbia

Audience. This problem is versatile and can be used in first-semester calculus, though will require proper scaffolding. Given the context, it is particularly suitable for life-sciences calculus. Comments on the subparts of the problem, including necessary student background, follow the problem statement.

Problem Statement. If a human egg splits between day 8 and day 13 post fertilization, the pregnancy will result in a particular type of genetically identical twins referred to as mono-mono twins. The first mono is short for monochorionic (a single shared placenta) and the second mono is short for monoamniotic (a single shared sac enclosing the fetuses and amniotic fluid). This occurs in about 1 out of 50,000 pregnancies and about 1% of twins amounting to approximately 8 such pregnancies per year in Canada and about 10 times as many in the US. Sharing a sac is dangerous because ambilical cords can entangle or get compressed. Sharing a placenta puts such twins at risk of twin-twin-transfusion syndrome in which one twins gets a substantially smaller fraction of the blood flow, potentially resulting in dramatically different developmental progress and other problems. As a result, mono-mono twins have a declining probability of survival the longer they remain in the uterus. The standard method for mediating that risk is a scheduled c-section but premature delivery is also dangerous. Suppose that the probability of both mono-mono twins surviving *in utero* up until week w of the pregnancy in light of the mono-mono risks is given by $f(w)$ and that the probability of survival past infancy for a premature infant delivered at week w is $g(w)$. From probability theory, assuming independence of the following events, we know that the probability of both twins surviving until delivery at w and both twins surviving past infancy from delivery at week w is $s(w) = f(w)g(w)^2$. The goal is to choose w so as to maximize the survival probability $s(w)$.

a) During pregnancy, a fraction of mono-mono twins are at risk of death due to entanglement and compression of their umbilical cords. Out of 25 such pregnancies that were identified at 24 weeks, 16 of them were still viable at 32 weeks. Use this information to determine the function $f(w)$, assuming that $f(w)$ is of the form $f(w) = Ae^{-rt}$.

b) The probability that an infant delivered at week w survives past infancy is given for several values of w in the table below. Plot the data and consider possible forms for $g(w)$. Using (linear or nonlinear) least squares, use the data to determine the best parameters for your proposed functional form.

w	$s(w)$
26	80%
27	90%
30	95%
34	98%

c) Once you have calculated the optimal w for scheduling delivery, you can estimate the loss associated with deviating from the optimum. Assume that across Canada there are 8 mono-mono pregnancies per year. What is the expected number of surviving babies if they are all delivered at the optimal time? And what is it if they are all delivered a week after the optimal time?

Guidance and Directions for Students.

Comments on the subparts:

a) This problem as stated should be manageable for students who have worked with exponential function before as this is equivalent to a radioactive decay problem. If this problem is used in a modelling course of the appropriate type, the students could be expected to come up with the form of f, only being asked to assume the time of pregnancy loss is exponentially distributed. In that case, they could be asked to do a bit of reading and encouraged to critique this assumption (there is reason to suspect that the risk of cord compression, if not entanglement also, gets worse as the fetuses get big enough to fill space in the uterus, violating the assumption of an exponential distribution). More sophisticated functions can also be used.

b) This problem as stated is targeting a more sophisticated student but I have used it with more scaffolding in a first term calculus course where students fit a Hill function using a [Lineweaver-Burk analysis]. Approximate fitting of any two-parameter function can be done using a spreadsheet instead (making a table with rows for one parameter and columns for the other and the residuals filling the table) if such tools are discussed in the course.

c) Using the functions from the previous parts, find the optimal value of w at which to schedule the C-section. Depending on the functions used, this could be done analytically or numerically. I once asked a highly scaffolded version of this problem on a life science calculus midterm using exponential times Hill function (omitting the square on $g(w)$) so this ought to be a fairly versatile example.

General comments:

The numbers in this example were gathered in 2012 and are likely to be different now (specifically premie survival probabilities). Updating these numbers could be a good challenge for students. For example, [1] has some useful data and [2] has some other references.

To mediate the risk of complications, women pregnant with mono-mono twins usually

undergo twice daily ultrasound monitoring from around 26-28 weeks until delivery. If there is any indication of cord entanglement or other serious complications, an emergency delivery is performed. This monitoring is not accounted for in the simple optimization problem posed here but provides a much harder modelling challenge for the ambitious student.

References

[1] S. E. Juul et al., *Deaths in a modern cohort of extremely preterm infants from the preterm erythropoietin neuroprotection trial*, JAMA Netw. Open 5(2), 2022, available at https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2788752.

[2] *Preterm birth*, Wikipedia, available at https://en.wikipedia.org/wiki/Preterm_birth#Outcomes_and_prognosis.

3 Buying, Renting, Investing: a Financial Project

Lauren DeDieu
lauren.dedieu@ucalgary.ca
University of Calgary

Audience. This project was used in a mathematics appreciation course at the University of Calgary (MATH 205 — Mathematical Explorations). The enrolment was 100 students, and almost all planned on becoming elementary school teachers or junior/high school math teachers. The course is a terminal course with no pre-requisites.

Description of a Project. In this project, you and your partner will act as financial advisors. You will create a professional report to present to your client, Donna, and will meet with your client for 10 minutes to report your findings and answer their questions. (Our TA will be playing the role of Donna.)

Problem Statement. Your client, Donna, is currently renting an apartment near downtown Calgary. She is considering buying a house or condo but is not sure what she can afford. The whole process of potentially buying a property is overwhelming for Donna, and she wants your help.

Some facts about Donna:

- Donna currently lives alone and would be purchasing the property alone.

- Donna currently pays about $1,800 in rent each month.

- Donna wants to continue living near downtown Calgary.

- Donna has excellent credit and no debt.

- Donna makes about $85,000 per year. After taxes and other deductions, that amounts to about $4,500 each month.

- Donna currently has about $60,000 in liquid savings (money in her savings account that is not invested).

- Donna currently has about $90,000 in her TFSA invested in broad market index ETFs. She could potentially use some of this money as a down payment but would prefer not to.

- Donna is a fairly numerate person; in particular, she is familiar with the terminology and formulas we discussed in this course.

This is all we know about Donna; we cannot assume other things about her.

Donna has the following questions (which your report should address):

Question 1: If I were to buy a house/condo, what can I afford? In particular,

- how much would my mortgage be?

- how much would my down payment be?

- how much would I be paying each month for this house/condo?

<u>Question 2</u>: Suppose I don't buy a house/condo, and instead continue renting for the next 25 years and invest my savings (and any money I would potentially save each month by renting) in ETFs. Would I have more money this way?

Note that Donna is not expecting you to tell her exactly what she should do. She wants you to paint a picture of her possible options so that she can make an informed decision.

Guidance and Directions for Students.

<u>Guidelines for Question 1:</u>

- You should offer Donna a range of possible mortgages (at least three, where one is the maximum amount she could possibly afford, one is the lowest amount that appears possible based on house/condo prices near downtown Calgary, and one is somewhere in between). This website may be helpful: https://www.realtor.ca/

- When calculating the monthly payments, assume a 5-year fixed rate mortgage and a 25-year amortization period. This website may be helpful: https://tools.td.com/mortgage-payment-calculator/

- Note that the down payment she makes will affect her monthly payments. It would be good to paint a picture of how her monthly payments would change based on her down payment by presenting a few possible options.

- Normally, down payments in Canada must be at least 5%. If a down payment is less than 20%, then you need to purchase insurance (https://www.rateshop.ca/cmhc-insurance). This insurance is added to the total mortgage price.

<u>Guidelines for Question 2:</u>

- **Investment Returns:** There is no way to predict the market. In particular, there is no way to know for sure what annual rate of return Donna would have on her investments. Since Donna is invested in broad market index ETFs, they generally track the market's performance. An annual rate of return of 5% would be a conservative estimate. Anywhere between 5-9% annual rate of return would be reasonable (although, of course, there's no way to know for sure and predict the future).

- **House Appreciation:** I would recommend estimating how much money Donna would make from renting and investing her savings. Then consider the scenario where Donna buys a condo/house. How much would the condo/house need to appreciate by for this option to be better? This website may be helpful: https://www.calculator.net/investment-calculator.html

- **Inflation:** To make our calculations easier, please ignore inflation and assume Donna's monthly pay and rent remain constant. (Although inflation will happen each year, presumably Donna's pay will increase to keep up with this, so we are making the assumption that these things will cancel out.)

Submission Instructions:

- Create an infographic to summarize your findings for Donna.
  - Your infographic should contain at least one chart (following the best practices for representing data we discussed in class).
  - Donna should be able to quickly and fully understand what your recommendation is by reading your infographic.
- Create a report (at most 3 pages) to accompany your infographic(s), which contains full details explaining how you arrived at your recommendations.
  - Remember that your audience is Donna, so your report should be written in a way that is clear and easy for Donna to understand.
  - All calculations should be included and explained, except for the monthly mortgage payments and the interest acquired by monthly investments.
  - Be sure to tell Donna what assumptions are being made (e.g., 5% annual rate of return, mortgage interest rate, etc.).

Meeting with Client:

- You will meet with your client, Donna (played by our TA), for 10-minutes to discuss your recommendation.
- Donna will have read your report in advance. She will most likely ask you to summarize your findings and will ask you any questions she has about the report.
- Remember that your audience for this is Donna (not our TA), so your responses and explanations should be tailored to Donna.

Rubric:

This project is worth 25% of the course grade and will be graded according to the rubric attached at the end.

Additional Notes. In class, we spent two weeks on financial things (e.g., percentage change, interest rates, investing, inflation) but did not discuss mortgages. This project gave students the opportunity to apply their learning from class to a 'real world' scenario. The outcomes of this project, and the course more generally, was for students to develop their quantitative reasoning skills, cultivate positive feelings/attitudes about math, and to be able to identify math in the 'real world'. By the end of the course, the hope was that students developed a new appreciation for

the beauty and importance of mathematics and that they felt inspired to share this passion with their future students.

Student chose their partner three weeks before the report was due, and booked their client meeting with our TA via https://youcanbook.me/ three days before the report was due. The client meetings occurred during the last four weeks of classes.

Student feedback about this project was overwhelmingly positive. In an end-of-term reflection, almost all students reported that the financial project and the financial topics were the most meaningful part of the course for them. They found it super applicable to their personal lives and wish that it was taught in high school. Several students reported that the client meeting was a very memorable part of the course and reported that it helped to boost their self-confidence, because they were able to explain a topic in which they very recently knew very little about confidently to someone else. Several students reported that the Financial Project and related topics inspired them to open a TFSA and learn more about investing.

Attribute	Criteria	Exceeds Expectations (10 points)	Meets Expectations (8.5 points)	Partially Meets Expectations (7 points)	Needs Improvement (5 points)
Content: Question 1	Question 1 is fully and accurately answered, and contains an appropriate amount of detail.	Question 1 is **insightfully** answered, and **no** pieces require additional clarification.	Question 1 is **accurately** answered, and **some** pieces require additional clarification.	Question 1 is **partially** answered, and **some** pieces require additional clarification.	Question 1 is **weakly** answered, and **many** pieces require additional clarification.
Content: Question 2	Question 2 is fully and accurately answered, and contains an appropriate amount of detail.	Question 2 is **insightfully** answered, and **no** pieces require additional clarification.	Question 2 is **accurately** answered, and **some** pieces require additional clarification.	Question 2 is **partially** answered, and **some** pieces require additional clarification.	Question 2 is **weakly** answered, and **many** pieces require additional clarification.
Infographic	The infographic clearly, accurately, and creatively summarizes the recommendations for Questions 1 and 2. It contains graphs and tables where appropriate.	The clarity, accuracy, and creatively of the infographic *exceeds all expectations*.	The clarity, accuracy, and creatively of the infographic *meets expectations*.	The clarity, accuracy, and creatively of the infographic **partially meets** expectations.	The clarity, accuracy, and creatively of the infographic **does not meet** expectations.
Clarity & Delivery	The report is organized and presented in a way that makes the content cohesive and easy to understand. (e.g. content is sequenced in a thoughtful way, content is communicated at a level appropriate for the intended audience)	The organization and presentation of the report are *exemplary* and are *incredibly* effective in making report easy to understand.	The organization and presentation of the report are *very good* and are *very* effective in making report easy to understand.	The organization and presentation of the report are *satisfactory* and are *reasonably* effective in making report easy to understand.	The organization and presentation of the report **need improvement** and are *partially* effective in making report easy to understand.

4   Using Calculus to Model UV Index

Miroslav Lovrić
lovric@mcmaster.ca
McMaster University

Audience. This topic is suitable for a lecture in a level 1 course in calculus. Although the theme will likely be new for many students, the mathematical content (such as: piecewise functions, exponential functions, log scale, approximate integration) should be familiar to them. The value of this lecture is that it shows an authentic application, and demonstrates the use of functions and concepts covered in calculus. Introducing the background (UV radiation) takes very little time.

Description. UV index is a measure of the strength of ultraviolet (UV) radiation coming from the Sun. In moderation, exposure to UV radiation has benefits for our health; however, overexposure could cause sunburn, skin ageing, as well as serious health problems, including DNA damage, immunosuppression and skin cancer. In [2], we read "After sporadic attempts by various meteorologists to define a 'sunburn index' and growing concern about ozone depletion, Environment Canada scientists James B. Kerr, C. Thomas McElroy, and David I. Wardle invented the modern UV index in Toronto, Ontario. Environment Canada launched it as part of the weather forecast on May 27, 1992, making Canada the first country in the world to issue official predictions of UV levels for the next day."

UV radiation is an electromagnetic radiation with wavelengths in the range from 100 to 400 nanometres (nm). (Recall that 1 nanometre $= 10^{-9}$ metre.) It comes in three variants, called UVA, UVB, and UVC, defined by the wavelength ranges in this way: UVC: 100nm-290 nm, UVB: 290nm-320nm, and UVA: 320nm-400nm.

How dangerous is UV radiation? A common way to quantify the danger to humans is to use the *erythemal action spectrum function* (also referred to as the *erythema*). The erythema is a piecewise function of wavelength, given by:

$$w(\lambda) = \begin{cases} 1 & 100 \leq \lambda \leq 298 \\ 10^{0.094(298-\lambda)} & 298 < \lambda \leq 328 \\ 10^{0.015(139-\lambda)} & 328 < \lambda \leq 400 \\ 0 & \lambda > 400 \end{cases}$$

The larger the number, the more dangerous UV radiation is.

It is a good exercise to sketch $w(\lambda)$ by hand (see Figure 1): it consists if two constant functions, and two decreasing exponential functions. Note that there is a discontinuity at $\lambda = 328$; calculating, we find that the left and the right limits at $\lambda = 328$ are 0.0015 and 0.0014 respectively.

This diagram shows that all UVC, and the lowest UVB wavelengths are the most dangerous. As the wavelength increases, UVB radiation danger decreases exponen-

tially, and that trend continues throughout the UVA spectrum. Checking for some specific values, we find, for instance, $w(290) = 1$, $w(305) = 0.22$, and $w(340) = 0.001$.
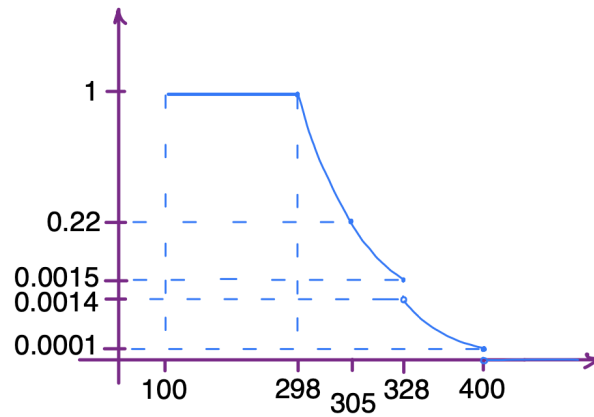
Figure 1: Erythema function

In Figure 2, taken from [2], erythema (Erythema action spectrum in the figure) is plotted in green. Note that the scale on the vertical axis is logarithmic, and hence the exponentially decreasing pieces are plotted as straight lines with negative slope.
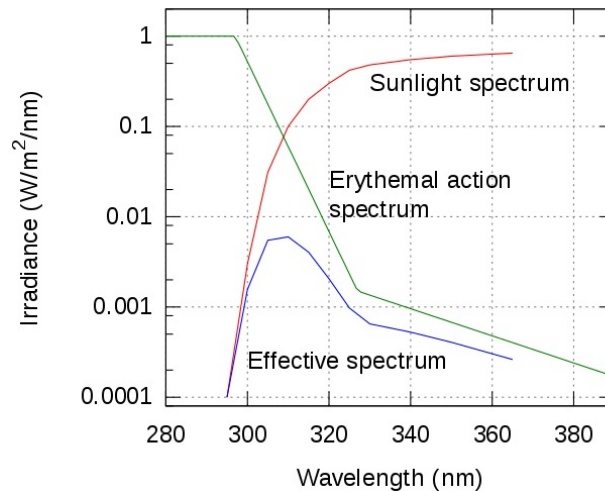
Figure 2: Erythema function, log scale

The erythema function tells us how dangerous each wavelength is to humans. To calculate the impact (severity) of UV radiation, we need to know how much of each UV wavelength reaches Earth.

The severity of UV radiation depends on numerous factors, including the time of day (the UV index is usually calculated at solar noon, i.e., when the angle of Sun rays with respect to the horizontal is the largest), duration of exposure, skin tone, and reflection from water, snow, sand or other materials. The good news is that ozone blocks UV radiation (and that's one reason why we are able to live on Earth):

12

it blocks all of UVC (thus, the most dangerous wavelengths do not reach us) and about 95% of UVB wavelengths. In other words, about 5% of UVB and all UVA wavelengths reach the surface of our planet.

The quantity of each wavelength that reaches Earth is described by the *spectral irradiance* or *spectral intensity* function $I(\lambda)$. $I(\lambda)$ is a discrete function, whose domain is a range of integer and half-integer wavelengths:

$$\{286.5, 287, 287.5, 288, \ldots, 400\}$$

It is measured up to certain wavelengths, and estimated for others (we will skip technical details here). In Figure 3, (taken from [1]) the spectral irradiance is shown in red.
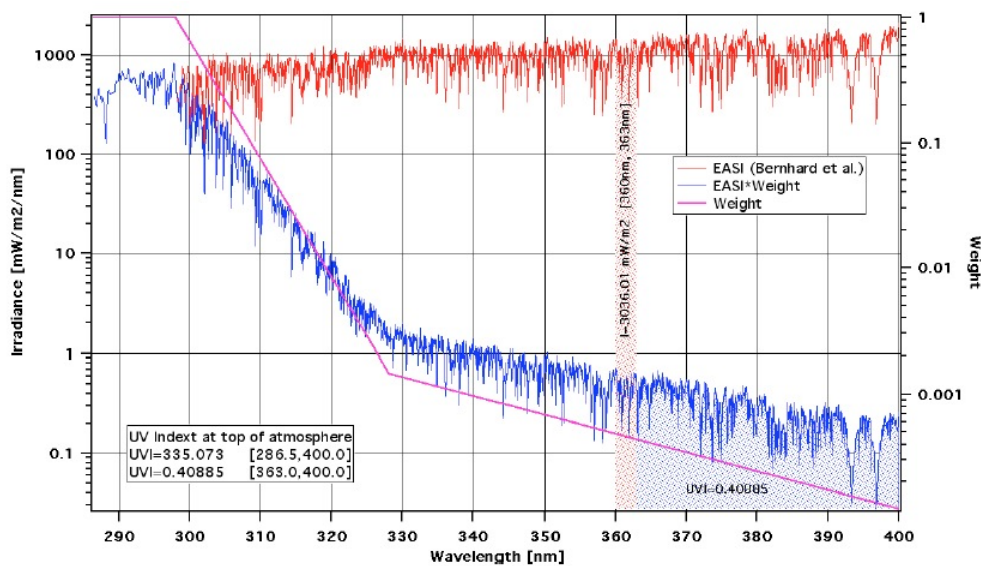

Figure 3: Erythema function, log scale

We are now ready to define the UV index.

Definition. The UV index is given by the integral

$$\text{UVI} = \frac{1}{25} \int_{286.5}^{400} w(\lambda) I(\lambda) \, d\lambda$$

Note that this definition makes sense: $w(\lambda)I(\lambda)$ is the product of the level of danger and the irradiance a given wavelength (in Figure 3, this product is drawn in blue); thus, it represents the severity of radiation from this particular wavelength. We then add these severity terms for all wavelengths from 286.5nm to 400nm, namely those wavelengths of UV radiation that reach Earth. The integral is divided by 25 for normalization, as its inventors wanted UVI to be 10 for a hot, sunny summer day in Toronto.

13

Note that this is an integral of a discrete function, and thus we need to apply numeric integration to calculate it (in practice, Trapezoid rule is used).

As mentioned, weather forecasts include the UVI values, in some cases updated on an hourly basis (otherwise, the highest UVI daily value is reported). On a typical (not hottest) summer day, UVI values are around 8 for Toronto, Montreal and Calgary; a bit lover (around 7) for Vancouver, St. John's and Winnipeg, and about 5 for Yellowknife. World-wide, UVI could be higher. Due to the thinness of the ozone layer over New Zealand, UVI can reach 14 there. As well, the UVI in Grand Canyon (Arizona) is measured to be 14, in Mauna Loa (Hawaii) it is 18, and in Antarctica it reaches 16. Keep in mind that the UVI produced by tanning beds routinely goes as high 11 or 12!

Guidance and Directions for Students. In my view, the best way to guide students through this topic would be to lecture, and stop frequently to ask questions. For instance, after defining the erythema function, an instructor can ask students to interpret its values, and sketch the graph by hand. As students attempt to do that, they will inevitably come across the fact that that the function is discontinuous, and that the range of its values is large (covering four orders of magnitude). Hence, a more convenient way to graph is to use semi-log plots (i.e., the vertical axis is a log scale). The instructor can ask students to describe how each of the four pieces of $w(\lambda)$ looks like in a semi-log coordinate system.

Next, looking at Figure 3, there could be a discussion about how the purple and red graphs generate the blue graph (How do we graph the product of two graphs?). After the definition of UVI, a course instructor could ask for suggestions about how the integral involved can be calculated. As illustration, the instructor could assume that there are only 4 or 5 wavelengths involved, make up the values for $w(\lambda)$ and $I(\lambda)$ and ask students to approximate the integral using a Riemann sum, and then using Trapezoid rule.

As mentioned this is a rich math problem, as in involves a number of concepts: a piecewise function, exponential functions, log scale and log graph; understanding of magnitudes; discrete functions; graphs and multiplying graphs; numeric integration (approximation methods).

Depending on students' background, one could attempt to code the functions involved here (which would include finding real data for the intensity function $I(\lambda)$.

References

[1] P. Kiedron, S. Stierle and K. Lantz, *Instantaneous UV Index and Daily UV Dose Calculation*, NOAA-EPA Brewer Network, pp. 1–5, 2007, available at https://bit.ly/3utTred.

[2] *Ultraviolet Index*, Wikipedia, available at https://en.wikipedia.org/wiki/Ultraviolet_index.

5   Counting Cards in Blackjack

Diana Skrzydlo
dkchisho@uwaterloo.ca
University of Waterloo

Audience: Introductory Probability class for first or second year students. All activities were specifically targeted at Computer Science students, although no specific computer knowledge is assumed. Calculus 1 knowledge is assumed for the course, but it is not used in the activity described below.

Description of Activities: No specific classroom setup is required, and these examples can be integrated into class slides or given out separately. This course had 24 activities in total, but I will discuss only one of them (more activities are available by request). You can find video explanation of this activity, as well as of other activities on my YouTube channel in this playlist.

Guidance and Directions for Students: I present the example of counting cards in Blackjack by going through it and asking for student input before revealing each step of the answer. If you are giving out material in advance you don't need to include the solution. It would probably also use it as an in-class activity where students work on it in small groups, but then you may want to include some checkpoints rather than letting them loose on the whole problem.

The rules of Blackjack are as follows:

- You get 2 cards, and so does the "dealer";

- The cards 10, J, Q and K are worth 10, A is worth 1 or 11;

- You can ask for as many more cards as you want, one at a time, and you want to have a higher total than the dealer, without going over 21;

- A combination of one A and one card worth 10 (10, J, Q and K) is called a *Blackjack*;

- There can be multiple players, but they are all playing against the dealer, not against each other, so we will just consider one player.

Suppose you are playing with one standard deck (52 cards). Out of 52 cards, 20 are high cards (10, J, Q, K, A) which you want to get. A combination of any two of those would be good for you. Assume they are randomly spread throughout the deck. At the start, the probability you get two of those cards is $\binom{20}{2}\binom{32}{0}/\binom{52}{2} = 190/1326 \approx 0.143$. Why? Because if we let X to be the number of high cards you get, then X follows the hypergeometric distribution with population size $N = 52$, number of success states $K = 20$, and number of draws $n = 2$:

$$X \sim \text{Hypergeometric}(N = 52, \ K = 20, \ n = 2).$$

We determine the probability of observing $x = 2$ successes as follows:

$$P(X = x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}} = \frac{\binom{20}{2}\binom{32}{0}}{\binom{52}{2}} = \frac{190}{1326} \approx 0.143.$$

As you continue to play more rounds, you get better information about what cards are remaining in the deck. If you keep track of this information, this is called "card counting". Suppose 26 cards have been played, so you know there are 26 cards left: r high cards and $26 - r$ low cards, i.e., you have been counting how many high cards that have been played so far. Now the number of high cards you get is

$$X \sim \text{Hypergeometric}(N = 26, \ K = r, \ n = 2),$$

and the probability you get two high cards is $\binom{r}{2}\binom{26-r}{0}/\binom{26}{2}$. We can evaluate this probability for different values of r:

High cards played	11	10	9	8	7	6	5	4	3	2
r	9	10	11	12	13	14	15	16	17	18
$P(X = 2)$	0.11	0.14	0.17	0.20	0.24	0.28	0.32	0.37	0.42	0.47

Note that r could be as low as 0 or as high as 20, but it is extremely unlikely, so I have left them off. Having that information gives you a lot of power: if there are lots of high cards left (relative to the remaining cards in the deck), you are much more likely to get a good hand, and so you can bet more money accordingly. **That's why BlackJack is not played with one deck. It's played with at least 4 decks. But why does that matter?**

Suppose you are playing with 4 decks. Then there are 80 high cards in the 208 combined cards. Assuming they are equally scattered in the deck, the probability you get 2 high cards at the start is $\binom{80}{2}\binom{128}{0}/\binom{208}{2} = 3160/21528 \approx 0.147$, very close to what it was before with one deck.

Now again assume that 26 cards have been played, and that there are r high cards of the $208 - 26 = 182$ cards left. So the number of high cards we get is

$$X \sim \text{Hypergeometric}(N = 182, \ K = r, \ n = 2).$$

If we have been counting the number of high cards played so far, we know r, so we get the following for $P(X = 2)$:

High cards played	11	10	9	8	7	6	5	4	3	2
r	69	70	71	72	73	74	75	76	77	78
$P(X = 2)$	0.14	0.15	0.15	0.16	0.16	0.16	0.17	0.17	0.18	0.18

The exact same information (knowing how many high cards have been played out of the first 26) gives us hardly any difference in the chance our hand has two high cards! So it's harder to use that information to adjust your betting strategy. Of course,

as we get closer to the end of the 4 decks of cards, we have more information and we can use it if we count cards. That's why the dealer usually stops and reshuffles before getting too close to the end of the 4 decks, before the information becomes too powerful.

What if we could play with an infinite number of decks? If we have an infinite number of standard decks, then the probability that any card is a high card is just $20/52 \approx 0.3846$. And importantly, the cards are all independent, so we don't need a Hypergeometric distribution anymore, we can use the Binomial distribution. If X is the number of your cards that are high cards and

$$X \sim \text{Binomial}(n = 2, \ p = 0.3846),$$

then

$$P(X = 2) = \binom{2}{2}(0.3846)^2(1 - 0.3846)^0 \approx 0.148.$$

This is pretty close to the 4-deck case. Why is it so close? When N is large and n is small, we can approximate a Hypergeometric with a Binomial pretty well. The parameters $N = 208$ and $n = 2$ fit this observation nicely. But if we have an infinite number of decks, then NO MATTER WHAT happens in the first 26 cards (even if all of them or none of them are high cards), the probability of you having two high cards never changes: it's always 0.148, so you never have any information to adjust your strategy.

6 Introductory Statistics Authentic Final Project

Emily Kozlowski
eikozlowski@uwaterloo.ca
University of Waterloo

Audience. This project is appropriate for first or second year students in an introductory statistics course. In such a course students have already taken a pre-requisite probability course. The introductory statistics course covers content such as numerical and graphical data summaries, maximum likelihood estimation, sampling distributions, confidence intervals and hypothesis tests, simple linear regression, and causation.

Description of Project. This is a final project for the course that focuses on the applications of goodness of fit tests for an assumed model, simple linear regression, and the idea of causal relationships. Students are given a real/contextualized dataset containing real-estate prices in Taiwan with potential explanatory variates such as house age, number of stores within walking distance, and distance to nearest metro station [1].

Problem Statement and Its Solution. Students are asked to answer a number of questions, which involve conducting analysis using the programming language R for statistical computing and then interpreting and summarizing results. The questions guide them through:

- Data summaries and doing a chi-squared goodness of fit test for an exponential model for distance to nearest metro station;

- Creating a two-way table for the price of the house and one of the explanatory variates to formally test the hypothesis of independence;

- Fitting two different simple linear regression models for the price of the house as a function of potential explanatory variates.

After completing all the analysis, students were to create a report that was written in a style similar to a report that one would write for an employer. That is, it was to be well organized, well-written in complete sentences, and not containing R code. The final conclusion section of their report asked them to summarize the key results from their analysis that address the main purpose of this study: whether the price of a house is associated with factors such as the age of the house, the distance to the nearest public transit, or the number of convenience stores within walking distance, and whether a simple linear regression model is appropriate for these relationships.

Secondly, they were asked to discuss and explain whether or not one can conclude there are casual relationships between the explanatory variates and the response of house price.

Guidance and Directions for Students. Students were given the base dataset and were instructed to each create their own unique dataset as a subset from the full

dataset.

To guide in the completion of the R analysis, students were provided R Tutorials throughout the course which contained sample code that was similar to that which was needed for the assessment, but applied to other problems. An R Tutorial for the final project was provided to give guidance for the goodness of fit tests. Simple linear regression was covered in Assignment 5 and the R tutorial for that assignment was sufficient for that section of the project.

To aid in consistent reports for marking (this course had approximately 600 students), students were provided a Word document template that they were to populate with their written responses and appropriate R output.

Additional Notes. Included in the associated files are the following:

- PDF of the project file presented to students;

- PDF of the project solutions;

- Word document template that students were to populate;

- R Tutorials for the Final Project and Assignment 5 (which covered simple linear regression);

- The dataset provided to the students to be used for analysis (some minor modifications to the one downloaded from the following repository):

  https://archive.ics.uci.edu/ml/datasets/Real+estate+valuation+data+set.

References

[1] I. C. Yeh and T. K. Hsu, *Building real estate valuation models with comparative approach through case-based reasoning*, Applied Soft Computing 65, pp. 260–271, 2018.

7 Shannon's Entropy and Applications to Cyber

David Thomson
dthomson@math.carleton.ca
Carleton University and The Tutte Institute for Mathematics and Computing

Audience. The Cyber Branch was introduced to the United States Army on 1 September 2014, and the Army Cyber Institute was established in part to support the development of the Cyber discipline throughout the Army and its officers. The Army Cyber Institute is housed at the United States Military Academy at West Point, and its civilian faculty are cross-appointed with a home department the Academy. Part of the goal of this cross-appointment is the integration of Cyber topics across the entire West Point curriculum.

At West Point, section sizes are capped at 18 cadets, and there is a heavy emphasis on active cadet participation (each classroom has blackboards on every wall). In the Department of Mathematical Sciences, instructors have the freedom to pick their teaching method, but read-ahead is expected from the onset and online homework on the day's material is due before class, which encourages a "flipped"-style of teaching.

Description of the Project. This lecture was developed as part of MA104: Calculus I, a 4.5 credit-hour 17-week course, and the second semester of the mathematics core curriculum taken by every West Point cadet. The course covers single variable differential and integral calculus, and a significant portion of the middle of the course focuses on modeling and applications. Integration in particular gets a modern treatment, where little time is spent hammering home mechanical integration techniques, and instead cadets employ computational methods by modeling examples in Mathematica and relying on software to do the dirty work. Since courses are highly coordinated by necessity (in 2016 there were approximately 20 instructors covering over 2 dozen sections of MA104) creating a new topic includes creating a read-ahead for the students as well as a lesson plan and pre-lesson presentation for the instructors. This cyber lesson was placed at approximately 2/3 mark of the course (lesson 46). One of the previous lessons introduced the notion of a probability distribution as evidenced by its occurrence in the lecture material.

Problem Statement. Two academic papers by Claude Shannon launched the revolution of digital communications with the realization that an analog signal could be encoded as a series of *discrete* numbers (e.g., bits) with no practical loss in information. Moreover, motivated by statistical mechanics, Shannon shows that the secrecy of these communications could be assured given any amount of computational power.

Today, digital communications are ubiquitous in our lives: our computers, cellular phones, even our homes and cars are networked together, and Hollywood has ingrained green 0s and 1s falling down a black screen for nearly two decades now.

Before Shannon, transmitting a message, a picture or an audio clip was understood only as propagating an electromagnetic signal across a wire. With Shannon's work, we are able to discretize communications systems, effectively encoding information

as a series of *binary digits* or *bits*. Shannon's key observation was to relate messages passing through a wire to probabilities. When a message arrives at a receiver, the decoder selects it from a set of possible messages with some probability. More specifically, the decoder selects a message x from a sample space with probability $p(x)$. Considering messages in this way, selecting a given message gives *information* about the distribution.

Definition. *Shannon defines the amount of information $I(x)$ of the message x to be*

$$I(x) = -\log_2(p(x)),$$

where the logarithm is in base 2 in order to measure the information in bits.

In this lesson, we will focus on digital secrecy. A modern secrecy system can be considered in an abstract way as the application of a set of transformations to a message. If two parties agree upon the transformations to apply in advance, the agreement is called a *key* and the secrecy system is called *symmetric*. This is the general framework we will be discussing.

Before we jump into the precise definition of entropy, let's look at what entropy means intuitively. Entropy is a measure of the uncertainty in a system: if we view a probability density function $p(x)$ as the probability of a certain occurrence x (such as a message being encrypted with a given key), then the entropy of the system will be high if the outcome of any given experiment reveals little information about the remainder of the system, other keys or messages.

For one example, your web browser may negotiate a secure connection to a website using "128-bit encryption", which means there is approximately 128 bits of entropy in the system used to encrypt your information. One way of looking at this is that a brute-force attack would require approximately 2^{128} operations to find your key. A modern processor may operate at 2 GHz, which means its clock cycles 2 billion (roughly 2^{31}) times per second. If that processor has 16 (2^4) cores, and a supercomputer has 1024 (2^{10}) processors, then this computer can perform roughly 2^{64} clock ticks per week (there are roughly 2^{19} seconds per week). Remember: $2^{128} = 2^{64} \cdot 2^{64}$, so this means that it should take around 2^{64} weeks to brute-force your key. The current age of the universe is estimated to be about 14 billion years (that's about 2^{40} weeks), and heat death of the universe (the ultimate uniform distribution) is estimated to be approximately 2^{300} years away, so you have plenty of time to break that secret before things get too cold!

Definition. *The entropy of a continuous density function $p(x)$ is*

$$H(x) = -\int_{-\infty}^{\infty} p(x) \log_2(p(x)) \ dx.$$

The entropy is the expected value *(think weighted average) of the information of $p(x)$.*

Concretely, when deciding to use a particular secrecy system, the provider (your bank, iTunes, etc.) effectively chooses which distribution $p(x)$ to encipher messages

into. There are practical considerations in picking any system, but think about some distributions you would like to encode your message (or your password) into. let's look at a a distribution for which every outcome is equally likely.

Example. *Consider the uniform distribution on the closed interval $[0, n]$, that is*

$$u(x) = \begin{cases} 0 & \text{if } x < 0, \\ 1/n & \text{if } 0 \leq x \leq n, \\ 0 & \text{if } x > n. \end{cases}$$

1. Verify that $u(x)$ is a probability distribution.

2. Calculate the entropy of $u(x)$.

Solution. *To verify that $u(x)$ is actually a probability, we need to ensure that*

1. $u(x)$ is non-negative for all x.

2. $\int_{-\infty}^{\infty} u(x)\mathrm{d}x = 1$.

The first point is clear from the definition of u. To ensure the second condition is met, observe that

$$\int_{-\infty}^{\infty} u(x)\mathrm{d}x = \int_{0}^{n} u(x)\mathrm{d}x = \int_{0}^{n} \frac{1}{n}\mathrm{d}x,$$

Caution: *The number n here is just a constant, like 2 or 5, whereas the variable we are integrating over is x, so the uniform distribution is just the constant function.*

We could integrate the function, or we could use the fact that the integral is equal to the area (of the rectangle) under the curve. Thus $\int_{0}^{n} \frac{1}{n} \, dx = n \cdot \frac{1}{n} = 1$ for any positive n, and the uniform distribution is indeed a probability.

To calculate the entropy of $u(x)$, use the formula

$$H(x) = -\int_{-\infty}^{\infty} u(x) \log_2(u(x))\mathrm{d}x.$$

As before, $-\int_{-\infty}^{\infty} u(x) \log_2(u(x))\mathrm{d}x = -\int_{0}^{n}(1/n) \log_2(1/n)\mathrm{d}x = \log_2(n)$.

Caution: Again, n is a constant, while x is the variable we are integrating over.

We have just shown that the uniform distribution on an interval of width n provides exactly $\log_2(n)$ bits of entropy. The uniform distribution is precisely the case where all outcomes are equally likely. Drawing any enciphered message does not give additional information about other messages from the distribution, because all messages are equally likely.

Intuitively, the uniform distribution *should* be the distribution with the maximum amount of entropy.

In modern communications, both parties do not need to share their private keys ahead of time. Public key cryptography allows a user to encrypt information for a

recipient using only public information provided to that recipient. For example, this is how you can send your iTunes password to Apple and assume that only Apple is able to read it and authenticate you. It took 30 years after Shannon's groundbreaking work on secrecy systems for public key cryptography to be devised.

Suggested exercises and considerations for in-class discussion

1. Given that one of the supercomputers available through West Point can perform approximately 64 bits of work in 1 week, estimate how many bits of entropy a secrecy system requires to hold a key secure for up to 1 year.

2. Let
$$p(x) = \begin{cases} 2\ln(2) \cdot 2^{-x} & \text{for } x \geq 1, \\ 0 & \text{otherwise.} \end{cases}$$
Use integration by parts to calculate the entropy H of p.

3. Suppose Bob and Sue want to communicate securely over a radio channel. Alice and Bob have a shared key over the internet that they use to hide their messages from Eve. Should Bob use the same key to communicate with Sue?

4. Watch the Khan Academy videos together and discuss: https://www.khanacademy.org/computing/computer-science/cryptography/crypt/v/intro-to-cryptography

5. Let $f(x) = 2x$ for $0 \leq x \leq 1$ and 0 elsewhere. Follow the Exercise from the notes above and show it is a probability and remind them how to integrate by parts in order to compute the entropy.

6. Open Mathematica and define $f(x)$ and $H = -\int f(x)\log_2(f(x))\mathrm{d}x$. Re-do the previous example using $f(x) = \frac{1}{\sqrt{\pi}}e^{-x^2}$, which is the $(0,1)$-normal distribution.

References

[1] W. Diffie and M. Hellman, *New directions in cryptography*, IEEE Transactions in Information Theory 22, pp. 644–654, 1976.

[2] A. J. G. Hey and P. Walters, *The New Quantum Universe*, Cambridge University Press, Cambridge UK, 2003.

[3] D. Kahn, *The Codebreakers: The Comprehensive History of Secret Communication from Ancient Times to the Internet*, Simon and Schuster, 1996.

[4] J. J. O'Connor and E. F. Robertson, *Claude Ellwood Shannon*, 2003, available at https://mathshistory.st-andrews.ac.uk/Biographies/Shannon/.

[5] J. J. O'Connor and E. F. Robertson, *John Wilder Tukey*, 2004, available at https://mathshistory.st-andrews.ac.uk/Biographies/Tukey/.

[6] R. Rivest, A. Shamir and L. Adleman, *Cryptographic Communications System and Method*, US Patent no. 4405829 A, 1977.

[7] C. E. Shannon, *A mathematical theory of communication*, The Bell System Technical Journal 27, pp. 379–423, 1948.

[8] C. E. Shannon, *Communication theory of secrecy systems*, The Bell System Technical Journal 28, pp. 656–715, 1949.

8 PageRank Algorithm: Linear Algebra of Google Search

Shirin Boroushaki
sboroushaki@tru.ca
Thompson Rivers University

Audience. This material can be covered in a first year linear algebra course either as a motivating example or in more details when the notion of eigenvectors is introduced.

Description of the Project. This module gives a brief description of how the Google search engine provides a relevant search result by using concepts from linear algebra.

Problem Statement. Back in early 90's, the first generation of search engines would use a *text-based ranking system* to determine whether a page is a relevant search result. Basically, what they would do was to count the number of occurrences of the key word in a given search query in their inventory of all web pages. However, the result wouldn't be necessary the most relevant. For example, suppose you search the word "Mars", and there's a website whose only content is the word "Mars" repeated a million times, so this website will be the top search result which is obviously not what you were looking for. So how does Google do it?

Two graduate students at Stanford Larry Page and Sergey Brin who later became co-founders of Google, first introduced an algorithm that revolutionized the world of search engines. The algorithm is called the *PageRank algorithm*. What the PageRank algorithm does is to assign a score[1] to a web page based on the number of the links to that page called *backlinks*, and the quality of the pages where the links come from (authority).

Basic Idea of the Solution. Consider a simple case where there are four web pages in a network with a collection of links among them. First we turn the network into a *directed graph*[2]. Each web page is a node in the graph and links between two web pages are the edges. Figure 4 below shows the directed graph corresponding to a web of four pages and their links. Let's denote the number of backlinks to page i by x_i. So to find x_i for each node we need to count the number of edges that are pointing to x_i. In the example below we have $x_1 = 2,\ x_2 = 1,\ x_3 = 3,\ x_4 = 2$.

The backlinks to a page count as votes of support for that page, so in this example page #3 must be the most important because it has most of the votes. Moreover, pages #1 and #4 tie for second, but page #1 has a backlink from the relatively important page #3 while page #4 has a backlink from the not-very-important page #1. Hence, to better capture the importance of pages, we modify our scoring by assigning a weight to each link. To do so, we make the assumption that each page

---

[1] Sometimes referred to as *importance score*

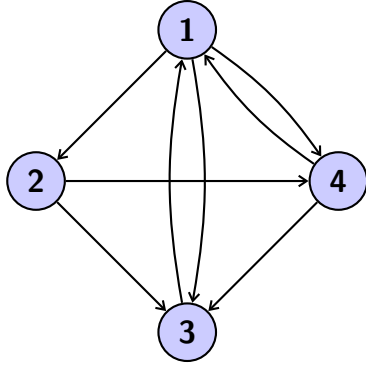[2] An introduction or a review of directed graphs may be needed.
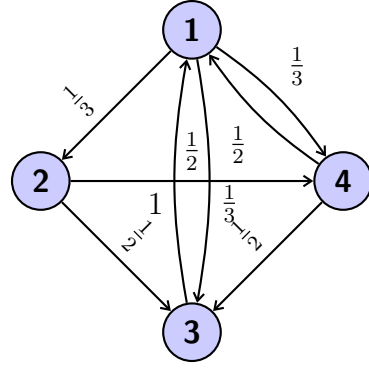
Figure 4: Directed graph   Figure 5: Weighted directed graph

should transfer evenly its importance to the pages that it links to. Node 1 links to three other pages corresponding to the three outgoing edges, so its importance is divided by 3; meaning that each of the three nodes with backlinks from node 1 receives $1/3$ of node 1's importance. Node 3 has only one outgoing edge to node 1, so all its weight is passed on to node 1. In fact, a node with m outgoing edges will pass on $\frac{1}{m}$ of its importance to each of the nodes it links to. This way we have the directed graph with weights for each link given in Figure 5 above.

Now we re-evaluate the importance score of each web page by incorporating the weight of each backlink when we add them up. This means that the total score of page #1 with backlinks from page #3 and page #4, is the whole score of page #3 and half of the score of page #4 given the weight of the backlinks.. With x_i notation this amounts to the equation $x_1 = 1 \cdot x_3 + \frac{1}{2} \cdot x_4$. Doing this for the other three nodes we obtain the following system of linear equations

$$
\begin{cases}
x_1 & = x_3 + \frac{x_4}{2} \\
x_2 & = \frac{x_1}{3} \\
x_3 & = \frac{x_1}{3} + \frac{x_2}{2} + \frac{x_4}{2} \\
x_4 & = \frac{x_1}{3} + \frac{x_2}{2}
\end{cases}
\iff \quad A\mathbf{x} = \mathbf{x} \quad \text{where } A =
\begin{bmatrix}
0 & 0 & 1 & \frac{1}{2} \\
\frac{1}{3} & 0 & 0 & 0 \\
\frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2} \\
\frac{1}{3} & \frac{1}{2} & 0 & 0
\end{bmatrix}
$$

Matrix A is called the *transition* matrix. Note that the sum of entries on each column of A is 1. A matrix with such a property is a *column stochastic* matrix. The solution vector to the system above gives the score of each webpage. But we observe that the solution is in fact the eigenvector of the matrix A corresponding to the eigenvalue $\lambda = 1$. Straightforward calculations show that these eigenvectors are multiples of the vector $\mathbf{v}$ given below which can be rescaled to the vector $\mathbf{u}$ below whose entries add up to 1.

$$
\mathbf{v} =
\begin{bmatrix} 12 \\ 4 \\ 9 \\ 6 \end{bmatrix}
\xrightarrow{\text{rescale}}
\mathbf{u} = \frac{1}{31}
\begin{bmatrix} 12 \\ 4 \\ 9 \\ 6 \end{bmatrix}
\approx
\begin{bmatrix} 0.38 \\ 0.12 \\ 0.29 \\ 0.19 \end{bmatrix}
$$

Entries of vector $\mathbf{u}$ are in fact scores of each web page in our graph. This is called the

PageRank vector. Unlike what we calculated before by only counting the backlinks, we observe that the score of page #1 is higher than page #3.

In reality a graph of web pages contains at least 8 billion nodes, meaning that the matrix A is 8 billion by 8 billion. Finding an eigenvector for such a huge matrix even with mathematical software is almost impossible. Alternatively, we can use an approach similar to discrete dynamical systems to converge to an eigenvector. The technique in this approach is called the *Power Method.* The goal is to find the PageRank vector $\mathbf{u}$ whose entries represent the importance of each page in our web graph. We start with an initial guess for the vector $\mathbf{u}$, say $\mathbf{u_0}$, and assume that at the beginning the importance is equally distributed. Multiplying A by $\mathbf{u_0}$ returns an updated importance vector as the new PageRank vector $\mathbf{u_1}$, where $\mathbf{u_1} = A\mathbf{u_0}$. Iterating this process, we obtain a sequence of importance vectors $\mathbf{u_0}, \mathbf{u_1}, \mathbf{u_2}, ...$ where $\mathbf{u_n} = A\mathbf{u_{n-1}}$.

$$\mathbf{u_0} = \begin{bmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{bmatrix} \Rightarrow \mathbf{u_1} = A\mathbf{u_0} = \begin{bmatrix} 0.37 \\ 0.08 \\ 0.33 \\ 0.20 \end{bmatrix} \Rightarrow \mathbf{u_2} = A\mathbf{u_1} = A^2\mathbf{u_0} \Rightarrow \ ... \ \Rightarrow \mathbf{u_n} = A^n\mathbf{u_0}$$

In this example, the process shows that after 8 iterations the sequence of importance vectors converges to the PageRank vector

$$\mathbf{u} = \begin{bmatrix} 0.38 \\ 0.12 \\ 0.29 \\ 0.19 \end{bmatrix}.$$

This result coincides with what we obtained from the eigenvector method. In terms of efficiency of computations, it is much easier for computer software to compute powers of A and multiply them by an initial vector $\mathbf{u_0}$ until they converge than computing eigenvectors of a huge A, especially in the case of the internet web graph where A is sparse.

Exercise. Consider the following graph of three pages.

(a) Write the number of backlinks to each page. Can you guess which page has the highest score given the number of backlinks?
(b) Draw the weighted graph corresponding to this graph. Given the weights, now can you make a better guess of the highest-scored page?
(c) Write the linear system of equation whose solution is the importance vector of the web pages.
(d) Use the eigenvector method and find the importance vector $\mathbf{u}$. Is this vector consistent with your guess in part (b)?
(e) Use the Power method (and a mathematical software) to find the PageRank vector starting with a reasonable initial vector.
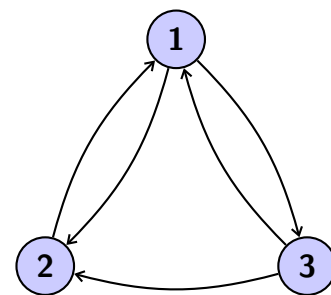

Figure 6

The discussion above only works for simple cases where the graph is connected and all pages have backlinks, which is not true in reality. Using the eigenvector as the PageRank vector in the following exercises results in a score of 0 for the webpages or multiple eigenvectors corresponding to $\lambda = 1$, which are not desirable.

Exercise. For each of the following web graphs find the PageRank vector.
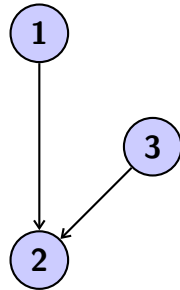

Figure 7: (a) Node 2 is a dangling node because it has no outgoing link (zero row in A, $\lambda = 1$ is not an eigenvalue.)
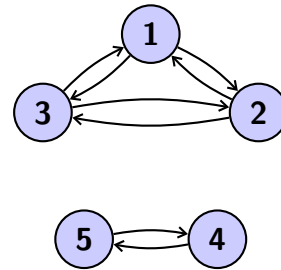
Figure 8: (b) Disconnected components. (Two eigenvectors corresponding to $\lambda = 1$.)

The exercise above can be used to further the discussion to the general formula for the PageRank matrix involving a damping factor. The idea is to construct a matrix M (called PageRank or Google matrix) that is *positive*[3] and *column stochastic* and use its probability eigenvector as the PageRank vector. The matrix M defined by

$$M = pA + (1 - p)H_n \qquad \text{where} \quad 0 \le p \le 1, \ H_n = \left[\frac{1}{n} \right]_{n \times n}$$

is called the Google matrix, and p is the damping factor. It can be proved that M is a column stochastic matrix and positive. Then the following theorem guarantees the existence of a PageRank vector. Forming the matrix M and using this theorem for the exercise above, we can find a meaningful PageRank vector for each graph.

Perron-Frobenius Theorem. If M is a positive and column stochastic matrix, then $\lambda = 1$ is an eigenvalue with multiplicity 1 and all other eigenvalues are less than 1. Moreover, the eigenvector corresponding to $\lambda = 1$ is unique and the sequence $\mathbf{u_0}, M\mathbf{u_0}, \ldots, M^k\mathbf{u_0}$ converges to the PageRank vector for some initial vector $\mathbf{u_0}$.

---

[3]A matrix A is positive if either A or some of its powers has all positive entries. For example the identity matrix is not positive.

References

[1] K. Bryan, T. Leise, *The \$25,000,000,000 Eigenvector: The Linear Algebra behind Google*, SIAM Rev. 48.3, pp. 569–581, 2006.

[2] R. Tanase, R. Radu, *PageRank Algorithm — The Mathematics of Google Search*, available at `https://pi.math.cornell.edu/~mec/Winter2009/RalucaRemus/Lecture3/lecture3.html`.

[3] D. Austin, *Understanding Linear Algebra*, Department of Mathematics, Grand Valley State University, 2021, available at `https://davidaustinm.github.io/ula/ula.html`.

9 Infectious Disease Spread

Asmita Sodhi
acsodhi@dal.ca
Dalhousie University

Audience. This topic is suitable for a course in calculus in which Euler's method and early topics in integration (Riemann sums, integration of polynomials) are covered. If Euler's method is not a standard part of the course curriculum, it would not be difficult to add an introduction to the process as a preliminary section to the topic. It is assumed that Euler's method has only been seen applied to a single first-order differential equation (not a system of first-order differential equations).

Description. Epidemiology is (for obvious reasons) a topic which the general public has a much greater interest in than they did three years ago. For me personally, having an understanding of the SIR model and the mathematics of diseases spread helped me stay calm and objective during the spring of 2020 when so much was unknown. The amount of talk about disease spread and data makes it a very natural topic of interest to students for a calculus assignment, particularly in a calculus for life sciences course (which is the context where I discussed this topic).

Part 1: SIR Model

This section was inspired by [3].

Epidemiologists use **compartmental models** to model infectious diseases. These models assign a population to a single "compartment" associated with a disease at a given time – such as susceptible, infected, and recovered – and an individual can move between these compartments. Often these compartmental models are simplified by ignoring natural birth and death rates (though more complex models do include these).

One fairly simple, but very useful, model is the SIR model for disease, which splits a population into three distinct categories:

- the fraction who are **susceptible** to a disease, represented by $S(t)$,

- the fraction who are **infected** with a disease, represented by $I(t)$,

- the fraction who have **recovered** from a disease, represented by $R(t)$.

The SIR model involves a coupled system of differential equations:

$$\frac{dS}{dt} = -b\,S(t)\,I(t)$$
$$\frac{dI}{dt} = b\,S(t)\,I(t) - k\,I(t)$$
$$\frac{dR}{dt} = k\,I(t)$$

where $b > 0$ represents the rate of infection and $k > 0$ represents the rate of recovery. This model assumes that once a person recovers from a disease, they cannot become sick again (such as with chickenpox).

Most first-year calculus courses do not introduce methods to solve these differential equations, but they do cover Euler's method to approximate solutions to a differential equation given an initial condition, and we can extend the understanding of using Euler's method on one differential equation to applying it to a system.

The equations for using Euler's Method in the SIR model are given by

$$
\begin{aligned}
S_{t+1} &= S_t - b\,S_t\,I_t\,\Delta t \\
I_{t+1} &= I_t + (b\,S_t\,I_t - k\,I_t)\,\Delta t \\
R_{t+1} &= R_t + k\,I_t\,\Delta t
\end{aligned}
$$

This is now something that students who have seen Euler's method can work with, given values for b, k, Δt, and initial conditions (S_0, I_0, R_0). A spreadsheet can be used to determine the population who is susceptible, infected, and recovered after a given amount of time. There is also room for students to play with the parameters b and k if so desired.

Part 2: Interpreting COVID-19 data

There is a wealth of COVID-19 data out there available for use (such as [1], [2], and [4]). Depending on the mathematical skill you want students to practice, you may want to use different data than what I chose. Because I wanted students to practice integration, I chose my data set to be the number of new recorded COVID-19 cases per day in my province over a certain period of time – this data was extracted from [2]. I applied a degree-four polynomial fit to this data set (degree four was chosen as a "Goldilocks" fit of being a high enough degree to actually model the data reasonably well and low enough degree that it's not frustrating for students to work with – just right!), and this is what students used in their assignment problems.

Since polynomials are easy for students to work with, they can practice integration techniques such as Riemann sums and using the Fundamental Theorem of Calculus without having to worry about special functions. Using real data also gives students an opportunity to interpret what the integral actually represents in the context of the data – something we do not get to emphasize when asking students to integrate functions that do not have any meaning assigned to them, and something some students find a bit more challenging.

Guidance and Directions for Students. I used this topic in a standalone assignment, but there is no reason it could not be covered in class instead (my assignment was used in an online asynchronous course, and my approach may have been different in an in-person class).
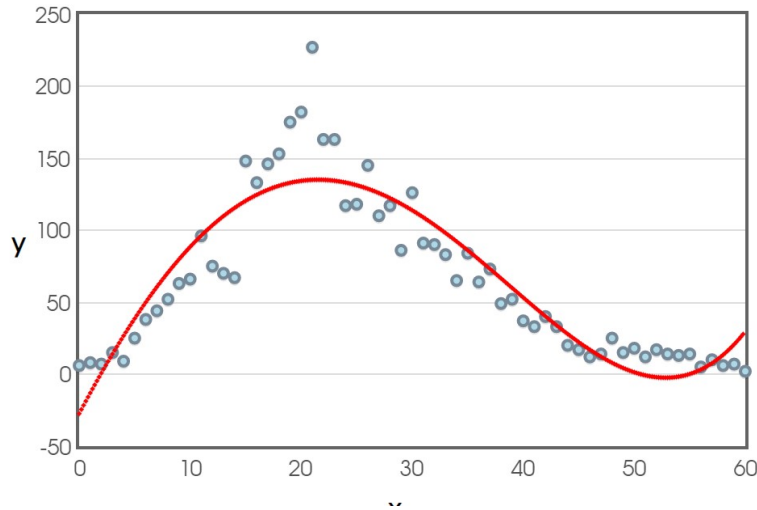
Figure 9: A scatterplot of the number of number of new recorded COVID-19 cases per day in Nova Scotia from April 16, 2021 to June 15, 2021 (a 60-day period) with polynomial fit $f(x) = 0.0001x^4 - 0.0050x^3 - 0.2643x^2 + 14.6366x - 27.4120$.

Students may have difficulty with their SIR model spreadsheets and the interacting data, and doubt that their work is correct. Some good checks are reminding students that all values in their spreadsheet should be between 0 and 1 inclusive, and that each row of their spreadsheet (i.e. $S_t + I_t + R_t$ for any time t) should add to 1.

Another challenge students faced was in interpreting the meaning of the results of their integrals when working with COVID-19 data. Students were sometimes not confident in what their integral represented. Reminding students that the integral is related to the area under the curve and suggesting they think about what the axes of the graph represented was usually a sufficient hint to get students on the right track.

Additional Notes. The assignment included here was part of a series of assignments I did in which I interviewed people who work in the life sciences and designed an assignment related to the work that they do. My favourite assignment out of the six was this one on disease spread, which was inspired by a friend who is an epidemiologist.

Creating calculus problems embedded in a scientific area without needing to link it to the work of a specific person would likely have been a lot easier, but it was important to me to show students that these are actual topics that come up in the work of actual people, and not just an "application" created by a mathematician for the sake of doing calculus. It also was really fun for me to get to interview so many interesting people and talk to them about the work they do!

References

[1] Centers for Disease Control and Prevention, *COVID Data Tracker*, available at https://covid.cdc.gov/covid-data-tracker/.

[2] Government of Canada, *COVID-19 epidemiology update*, available at https://health-infobase.canada.ca/covid-19/#a2.

[3] D. Smith and L. Moore, *The SIR Model for Spread of Disease*, 2004, available at https://www.maa.org/press/periodicals/loci/joma/the-sir-model-for-spread-of-disease-introduction.

[4] H. Ritchie, E. Mathieu, L. Rodés-Guirao, C. Appel, C. Giattino, E. Ortiz-Ospina, J. Hasell, B. Macdonald, D. Beltekian, and M. Roser, *Coronavirus Pandemic (COVID-19)*, 2020, available at https://ourworldindata.org/coronavirus.

10 Visiting Hilbert's Hotel

Karyn McLellan
karyn.mclellan@msvu.ca
Mount Saint Vincent University

Audience. Appropriate for any "mathematical concepts" type survey course, an introduction to proofs course or an introductory number theory course. Prerequisites include even and odd numbers, primes and prime factorization, the Fundamental Theorem of Arithmetic; possibly sets, bijections, countable and uncountable sets of numbers.

Description of a Project. In 1924, German mathematician David Hilbert created a thought experiment about an "infinite hotel" to illustrate the strange and seemingly paradoxical properties of infinity [3]. We start with a hotel containing infinitely many rooms and no vacancy and move through a series of questions about how we can fit in more guests. No special materials are required. Instructors may wish to number students to act out the various situations, if space permits.

Problem Statement and Its Solution. Suppose a hotel has an infinite number of rooms, numbered $1, 2, 3, 4, 5, \ldots$ and a "No Vacancy" sign in front. We will attempt to fit in new guests that arrive and to do this we must assign each of them a room number.

1. What if 1 new guest arrives?

Solution: Put the new guest in Room 1. But there's already someone in there! Put that person in Room 2. And put the person in Room 2 into Room 3, and so on. In general we put the person in Room n into Room $n + 1$.
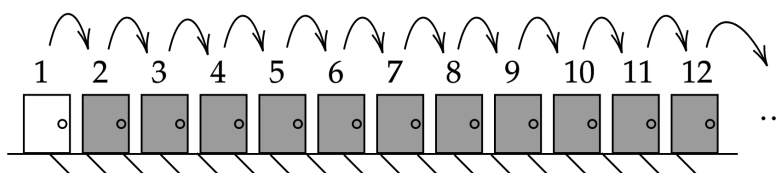

Figure 10: Taken from [4].

2. What if 2 new guests arrive?

Solution: We can use the same technique, this time putting the two new guests in Rooms 1 and 2. The current occupants go to Rooms 3 and 4, respectively. In general we put the person in Room n into Room $n + 2$.

3. What if k new guests arrive?

Solution: Following the logic above, we can do this for any finite number, k, of new guests, if each current occupant shuffles down k rooms. This frees up k rooms for

the new guests, and in general we put the person in Room n into Room $n + k$.

4. What if a bus containing infinitely many guests on seats $1, 2, 3, 4, 5, \ldots$ arrives?

Solution: The current method no longer works! We can't tell a guest to shuffle down an infinite number of rooms — what would their new room number be? We need a different tactic. We ask each current guest to move to double their room number, i.e., the guest in Room n moves to Room $2n$. This frees up all the odd-numbered rooms, of which there are still infinitely many! We now put person in Seat n of the bus into Room $2n - 1$. Everyone gets a room and all of the hotel rooms are now filled.
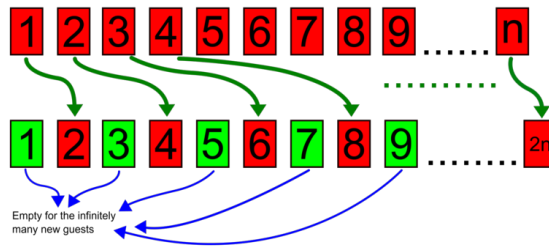

Figure 11: Taken from [1].

5. What if infinitely many busses, numbered $1, 2, 3, 4, 5, \ldots$, each containing infinitely many guests, arrive?

Solution: We need to be a bit more clever now. Suppose each person who arrives is on Bus number b and Seat number s. We can then put each person into Room $2^s 3^b$. Everybody gets a room, and because prime factorization is unique by the Fundamental Theorem of Arithmetic, each room number is different. But where do the current guests go? Which room numbers have the new guests NOT taken? We could put them in the prime-numbered rooms. There are infinitely many primes, and because the room number of each new guest has at least two factors, there is no overlap. It may be a bit difficult to tell the current occupants which room to move to, however (e.g., "go to the 47th prime number" may not be entirely clear!) Alternately we could have the current occupant of Room n move to Room 2^n. Again there is no overlap, as the room numbers of all new guests have a factor of 3. (We are essentially looking at the $b = 0$ case of Room $2^s 3^b$.) Note that although all old and new guests now have a room, not all rooms in the hotel are filled! For example, no one is in Room 5. This is called the Prime Factorization Method [6].

6. What if infinitely many ferries, numbered $1, 2, 3, 4, 5, \ldots$, each containing infinitely many busses, arrive?

Solution: The above method extends nicely here. First suppose the new guests arrive on Ferry number f, Bus number b and Seat number s. Each guest is then assigned to Room $2^s 3^b 5^f$. The current occupants can again move to Rooms 2^n.

7. What if we wish to extend the above idea to any number of layers of infinite

guests? (What type of vehicle might be able to carry infinitely many ferries each containing infinitely many busses?) How can we do this?

Solution: Given another "layer of infinity", say an infinite number of jumbo jets, each containing an infinite number of ferries, and so on, we can put person on Jet j, Ferry f, Bus b and Seat s into Room $2^s 3^b 5^f 7^j$. We can continue process this for any finite number of layers of infinity.

8. What if we have infinitely many such infinite layers? Does your method hold?

Solution: With infinitely many layers we need infinitely many prime numbers in the factorization to determine room numbers. At this point we cannot explicitly tell someone what their new room number is. The method fails to work, and in fact, we may not actually be able to fit everyone into the hotel [6]! The reason is that there are different sizes of infinity! The hotel rooms are numbered $1, 2, 3, 4, 5, \ldots$, and we say that this set of numbers, i.e, the *natural numbers*, is *countably infinite*, or just *countable*. The integers and the rational numbers are also countably infinite, because there is a systematic way to count them. In contrast, the set of real numbers is *uncountable*. For example, if a bus contained infinitely many people, but the people were numbered using every real number (including all possible decimals), we could not fit them into the hotel.

9. Can you think of any other ways to answer Questions 4, 5 and 6?

Solution: For Question 4, we can choose any two infinite and disjoint subsets of the integers. For example, we could have all new guests in Seat n of the bus move to room $3n + 1$ and all current guests in Room n move to room $3n + 2$. For Question 5, we can put the person in Bus b and Seat s into room p^s, where p is the $(b+1)^{\text{st}}$ prime number. E.g., the person in Bus 4 and Seat 5 will go to Room 11^5. We can put the existing guests in Rooms 2^n and these room numbers are all distinct. We will have the additional issue of no exact formula for the room number, however. It is possible to generalize this method using powers of powers. This is called the Prime Powers Method [6], [2]. For Question 6 (and 5) there exist other creative approaches that do not involve primes. For example, if a guest arrives on Ferry f, Bus b and Seat s, let their room number, in binary, be 1 followed by f zeroes, then 1 followed by b zeroes, then 1 followed by s zeroes [6]. For example, a person arriving on Ferry 3, Bus 1 and Seat 2 would be assigned Room number 100010100, which in decimal, is Room 276. Each number, and its corresponding decimal representation, is unique. These numbers are all even and so the current occupant of Room n can be assigned to Room $2n - 1$.

Guidance and Directions for Students.

- Have students think about why we simply can't put the new guest(s) "at the end".

- Have students think about the different ways we can divide up the natural numbers into sets that are infinite and disjoint.

- Emphasize that all new and old guests must get a room and all these rooms must be distinct. Further, we should be able to assign each of the guests a room number. Also, not every room need be filled, as long as everyone fits in.

- Emphasize the idea of using variables. If each guest has a bus number b, seat number s, etc., how can we map these variables to a room number?

- We can imagine giving one set of directions and then all the guests move simultaneously. This avoids problems with the time it may take for guests to move one at a time.

- Students may need some guidance on Questions 5–7. A hint like "think about using prime numbers" may help.

- Encourage students to be creative as there are many possible solutions to these questions!

Additional Notes.

- Instructors may wish to expand on certain topics such as paradoxes, sets and bijections and countability.

- Discuss that for infinite sets, the idea that "the part is smaller than the whole" no longer applies. I.e., when there are infinitely many rooms, the total number of even-numbered rooms is not smaller than the total number of rooms. Similarly, the statements "every room contains a guest" and "no more guests can fit in" are not equivalent, as they are for finitely many rooms. This was Hilbert's takeaway — infinity behaves in strange ways.

- Can prove that two sets have the same size using a bijection — a 1:1 and onto mapping between sets. We can create a mapping between the natural numbers and an infinite set of objects, for example the even numbers, to prove it has the same size as the natural numbers, i.e., it is countable. It can similarly be proven that the set of even numbers has the same size as the set of odd numbers.

- Can discuss Georg Cantor, who proved in 1874 that the set of real numbers is uncountable [5]. His proof was very controversial. It not only implied different sizes of infinity, but that there are an infinite number of different infinities! It encountered resistance from his mathematical contemporaries. Mathematician Kronecker called him a "scientific charlatan" and a "corruptor of youth". Some Christian theologians saw Cantor's work as challenging the unique infinite nature of God!

References

[1] J. Beránek, CC BY-SA 4.0 https://creativecommons.org/licenses/by-sa/4.0, via Wikimedia Commons, https://commons.wikimedia.org/wiki/File:Hilbert%27s_Hotel.png.

[2] M. Freiburger, *Hilbert's Hotel*, Plus: Bringing Mathematics to Life, 2017, available at https://plus.maths.org/content/hilberts-hotel#:~:text=Welcome%20to%20Hilbert's%20hotel!,%2C%202%2C%203%2C%20etc.

[3] M. Hallett, U. Majer, and D. Schlimm, *David Hilbert's Lectures on the Foundations of Arithmetic and Logic 1917–1933 (Vol. 3)*, Springer-Verlag, 2013.

[4] J.P. MacManus, *Hilbert's hotel but the guests are mere mortals*, 2020, available at https://jpmacmanus.me/2020/07/26/hilbertshotel.html.

[5] *Georg Cantor*, Wikipedia, available at https://en.wikipedia.org/wiki/Georg_Cantor.

[6] *Hilbert's paradox of the grand hotel*, Wikipedia, available at https://en.wikipedia.org/wiki/Hilbert%27s_paradox_of_the_Grand_Hotel.

11 Supporting Applied Content with Open Education Technologies

Sean Fitzpatrick
sean.fitzpatrick@uleth.ca
University of Lethbridge

Audience. Faculty considering applications requiring computational tools, who would like to provide their students with no-cost access to software, while ensuring a consistent experience for all students, regardless of hardware access or operating system choices.

Students (and faculty) are assumed to have basic familiarity with operating a computer, including downloading, saving, and opening files, and using a web browser to access a URL.

For faculty, some familiarity with GitHub is an asset, as is the ability to use Google to figure out the correct way to format snippets of code. Experience with coding, in particular Python and R, is an asset, but not required.

Description of Tools. Here we describe a set of tools that can be used to teach applied content, and empower students to explore a project using technology.

Python and R are both excellent programming languages for undergraduate teaching, and both are ubiquitous in industry, especially in fields such as data science. One of the reasons we prefer these languages is that they are open source, and freely available on a variety of platforms and operating systems. Both can also be run in the Jupyter notebook environment, which allows a mixture of code and rich text to be combined in a single document.

For students with no programming background, or who don't feel comfortable with computers, a Jupyter notebook provides instructors with a way to present and explain code to students, while letting them edit and execute that code in the same document.

Another major advantage is that a Jupyter notebook is designed to run in a web-browser, and several cloud-based options exist that do not require students to install any software:

- Many Canadian universities, including PIMS member institutions, have free access to a Syzygy Jupyter hub service (see [2] for more details);

- A paid (but relatively inexpensive) option is CoCalc, which is designed to run Sage code in a Jupyter notebook, but languages such as Python and R are also supported;

- A relatively new organization, known as 2i2c, provides cloud-based Jupyter solutions to institutions without the resources to set up their own.

Running Jupyter in the cloud means that students can use these tools on any computer capable of running a modern web browser, including low cost Chromebooks.

It also gives instructors the option of using tools like "nbgitpuller" links to distribute notebooks to students by simply providing a URL.

Students in an introductory calculus or linear algebra class probably can't be expected to know how to code. But with Python we can access several libraries designed for doing mathematics, including NumPy, SciPy, and SymPy. The SymPy library is designed to provide functionality similar to what one would find in expensive computer algebra systems such as Maple or Mathematica. When used in a Jupyter notebook, it has the nice feature that output can be formatted using MathJax to appear as it would in a LaTeXdocument.

None of the following qualify as "authentic" applications, but they do give examples of the types of worksheets one can create using Jupyter and Python:

- Partial fraction decomposition;

- Graphical and numerical methods for ODEs;

- Numerical integration;

- Singular value decomposition. [4]

Another way to provide students with universal access to live code is to embed it directly into an HTML document. One way to do this is with the SageMathCell environment. It can be embedded in any webpage, and it is also possible to embed several linked cells, so that subsequent cells are aware of code that has already been run in other cells on the same page. As the name suggests, SageMathCell is designed primarily for running Sage, but it does well with pure Python and R code as well.

A great way to author documents (or even entire textbooks [1]) is to use the PreTeXt language. PreTeXt has native support for Sage cells, and a book written in PreTeXt can be produced in a variety of formats, including HTML, PDF, and Jupyter. Recently, a lot of work has taken place to integrate PreTeXt with Runestone Academy. PreTeXt can now be used to author all of the interactive question types available in Runestone.

Examples of linear algebra topics that can be presented with a mix of live code (via Sage cells) include

- Gram-Schmidt ortho-normalization,

- Matrix factorization, and

- Jordan Canonical Form.

The examples above link to a work-in-progress for an edition of a linear algebra textbook written in PreTeXt that includes interactive Runestone exercises (so far these appear in the first three chapters). Once this edition is complete, it will be able to rune on a Runestone server (either Runestone academy, or a local server),

---

[4]The SymPy library includes an algorithm for the singular value decomposition, but the worksheet was designed to make students follow the steps.

allowing students to log in, and earn credit by answering the problems as they go. Long answer exercises can be designated as "reading problems", which will accept long answer responses from students that can be reviewed by an instructor.

Tools such as Jupyter and PreTeXt solve two problems in the classroom:

- They allow computational tools such as Python to be introduced in a "just in time" manner that does not require students to have prior knowledge of software or a programming language;

- They give the instructor freedom to explore more complicated examples, or assign problems that are more computationally intensive, because access to technology means that computational difficulty is less of a barrier when working with interesting problems.

This allows students (and the instructor) in a math class to explore authentic problems that might otherwise be inaccessible, because the calculations are too difficult or time-consuming to do by hand. Moreover, the tools presented here are all completely free for students to access, and can be made available in a way that requires nothing more than a web browser on the part of the students.

References

[1] S. Fitzpatrick, *Lecture Notes for Math 3410, with Computational Examples*, 2022, available at https://opentext.uleth.ca/Math3410/.

[2] M. Lamoureux, *Teaching with Syzygy*, 2022, available at https://intro.syzygy.ca/teaching/.