MCB137L/237L: Physical Biology of the Cell
Spring 2024
Homework 5
(Due 2/27/24 at 2:00pm)

Hernan G. Garcia

1 Dynamics of A → B reactions.

One of the most interesting topics in science is how we have learned to probe deep time. Surprisingly, DNA sequence has permitted us to explore deep time in the biological setting. Of course, biology and the dynamics of the Earth are not independent phenomena and the point of the rest of this problem is to better understand the details of how scientists figure out how old the Earth is as well as how old various fossil-bearing strata are. To that end, we will first consider a simple model of the radioactive decay process for potassium-argon dating methods, recognizing that there are many other dating methods that complement the one considered here.

Potassium-Argon dating

Potassium-argon dating is based upon the decay of ^{40}K into ^{40}Ar. To a first approximation, this method can be thought of as a simple stopwatch in which at t = 0 (i.e. when the rocks crystallize), the amount of ^{40}Ar is zero, since it is presumed that all of the inert argon has escaped. We can write an equation for the number of potassium nuclei at time $t + \Delta t$ as

$$N_{\mathrm{K}}(t + \Delta t) = N_{\mathrm{K}}(t) - (\lambda \Delta t) N_{\mathrm{K}}(t). \tag{1}$$

Stated simply, this means that in every small time increment Δt, every nucleus has a probability $\lambda \Delta t$ of decaying, where λ is the decay rate of ^{40}K into ^{40}Ar. We also employ the important constraint that the number of total nuclei in the system must remain constant, so that

$$N_{\mathrm{K}}(0) = N_{\mathrm{K}}(t) + N_{\mathrm{Ar}}(t), \tag{2}$$

where $N_{\mathrm{K}}(0)$ is the number of ^{40}K nuclei present when the rock is formed, $N_{\mathrm{K}}(t)$ is the number of ^{40}K nuclei present in the rock at time t, and $N_{\mathrm{Ar}}(t)$ is likewise the number of ^{40}Ar nuclei present in the rock at time t. In this part of the problem you will use equations 1 and 2 to construct differential equations to find the relationship between $N_{\mathrm{K}}(t)$, $N_{\mathrm{Ar}}(t)$, and t.

(a) Using equations 1 and 2 as a guide, write differential equations for $N_{\mathrm{K}}(t)$ and $N_{\mathrm{Ar}}(t)$. How do these two expressions relate to one another?

(b) Next, we note that the solution for a linear differential equation of the form $\frac{dx}{dt} = kx$ is given by $x(t) = x(0)e^{kt}$. Use this result to solve for $N_{\mathrm{K}}(t)$.

(c) Use the constraint encapsulated by equation 2 to write an equation for the lifetime of the rock, t, in terms of the ratio $\frac{N_{\mathrm{Ar}}}{N_{\mathrm{K}}}$.

Age of the Galapagos Islands

The potassium-argon dating method described above has been used in several contexts central to some of the most important evolutionary questions in biology. As we go from West to East in the Galapagos Archipelago, the ages of the islands increase, with Santa Cruz older than Isabella, for example. But how are these numbers known and what evidence substantiates these claims when naturalist guides make them? In a beautiful article from Science Magazine in 1976 (Science, New Series, Vol. 192, No. 4238 (Apr. 30, 1976), pp. 465-467), Kimberly Bailey tells us of her efforts to determine the ages of the islands of Santa Cruz, San Cristobal and Espanola. We will now use her data to find out the K-Ar ages of several of these islands ourselves.

(d) Read Bailey's short paper and give a brief synopsis (1 paragraph) of her approach and findings.

(e) Use the results from Sample H70-130 and JD1088 of Table 1 from Bailey's paper to determine ages for Santa Cruz Island and Santa Fe Island. To do this, you will need to navigate a few subtleties. First, note that the amount of Argon is presented in moles, and so you can use those numbers directly. To determine the number of moles of ^{40}K, you will need to use the weight percentage that is K_2O and use that in combination with the mass of the sample to figure out how much K is present. Note that not all of the potassium in the sample will be the isotope ^{40}K, so you will need to use the ratio of ^{40}K to total potassium, $\frac{^{40}\mathrm{K}}{\mathrm{K_{total}}} \approx 1.2 \times 10^{-4}$. Additionally, use the decay constant $\lambda \approx 5.8 \times 10^{-11}\ \mathrm{yr}^{-1}$.

Determining Lucy's age

In 1974, a fossil of *Australopithecus afarensis* (shown in Figure 1) was discovered in Ethiopia. This specimen, which was dubbed "Lucy," marks an important step in understanding human evolution because at the time of its discovery, it was the earliest known species to show evidence of bipedal locomotion. Because Lucy was found in an area that was rich in volcanic rock, potassium-argon dating was an ideal method for determining Lucy's age (Aronsen 1977).

Unfortunately for us, real-world K-Ar dating data are generally not neatly presented in the form of N_{Ar} and N_{K}. Instead, geologists will measure a concentration of ^{40}Ar in mol/g and a weight percent of $\mathrm{K_2O}$. These data must be used to identify the number of ^{40}Ar and ^{40}K nuclei in the sample. In this part of the problem, we will look at such measurements from an actual paleontological specimen as reported in Aronsen (1977) in order to determine its age.

Figure 1: The remains of Lucy, a specimen of *Australopithecus afarensis*.

(f) Using the table of ^{40}Ar and K_2O measurements below (Aronsen 1977), obtain an estimate for Lucy's age. Be sure to explain the steps you take to obtain your answer. Since each sample is taken from the area in which Lucy was found, we expect each sample to give you roughly the same answer; you will need to take the mean of the ages of each sample to obtain an estimate for Lucy's age.

Assume that each sample has a total mass of 1 g. Also, note that not all of the potassium in the sample will be the isotope ^{40}K, so you will need to use the ratio of ^{40}K to total potassium, $\frac{^{40}\mathrm{K}}{\mathrm{K_{total}}} \approx 1.2 \times 10^{-4}$. Additionally, use the decay constant $\lambda \approx 5.8 \times 10^{-11}$ yr^{-1}.

Sample Number	**^{40}Ar $\times 10^{-12}$ mol/g**	**wt.%K_2O**
1	2.91	0.657
2	3.18	0.755
3	3.08	0.680

Table 1: Outcome of measurements of potassium and argon for dating the rocks in the vicinity of Lucy.

2 Synthesizing a Transcriptome: Big Data in Transcription

In class, we briefly discussed the myriad of different ways to measure gene expression. Writ large, we can either find ways to count the mRNA transcripts or the protein products that result from these transcripts. For example, when properly calibrated, the green fluorescent protein (GFP) in conjunction with fluorescence microscopy is a favorite approach for measuring protein copy numbers. Recently, a different way to engage in the dialogue between theory and experiment has been afforded by the advent of technologies that make it possible to take a census of the full complement of transcripts inside individual cells.

One of the key applications of single-cell mRNA sequencing has been its use to identify "transcriptional fingerprints" that define discrete cell types within a population containing cells that have committed to multiple possible fates. One of the best examples of this application of single-cell transcriptome-wide sequencing comes from projects such as the *Tabula muris*. This project measured RNA counts for tens of thousands of genes within tens of thousands of individual cells in the mouse, derived from tens of distinct organs and tissues. Each single cell transcriptome is a giant $\approx$ 10,000 dimensional vector with the i^{th} entry corresponding to the mRNA count of the i^{th} gene.

One widespread approach to visualizing the results from these types of experiments is shown in Figure 2. In the figure, each point corresponds to an individual cell whose transcriptome was sequenced. Here, the extremely high dimensional data resulting from single-cell RNA sequencing (i.e., the number of mRNA molecules corresponding to each of $\approx$ 10,000 genes in each cell) was projected onto two dimensions using methods we will later explore. Further, once this projection is performed, cells are grouped in clusters. The idea is that cells within a cluster share much of their gene expression profile and are therefore identified as unique cell types corresponding to different tissues within the mouse. In this problem, we will attempt to build some intuition for how this identification of unique cell types is achieved by working with a synthetic transcriptome that we build ourselves using our understanding of the constitutive promoter. Obviously this is a caricature of the real situation where most genes are *not* constitutively expressed.

(a) Let's start by creating a mental picture of the high dimensionality of single-cell sequencing data by picturing how this data is stored. Specifically, think of a matrix $\mathbf{G}$ where you store the RNA counts for 10,000 genes measured in 1,000 cells where each row of the matrix corresponds to a given cell. How many rows and columns would this matrix have? Draw this matrix schematically, clearly indicating what each dimension of the matrix represents. Further, identify the gene expression vector that corresponds to the number of mRNA molecules detected for all species in cell number 1.

(b) To begin to get a feeling for this kind of data, we imagine an experiment on cells containing only two genes. These cells can adopt three different fates based on the expression state of these genes (i.e., low/low, low/high and high/high). Further, let's assume that these two

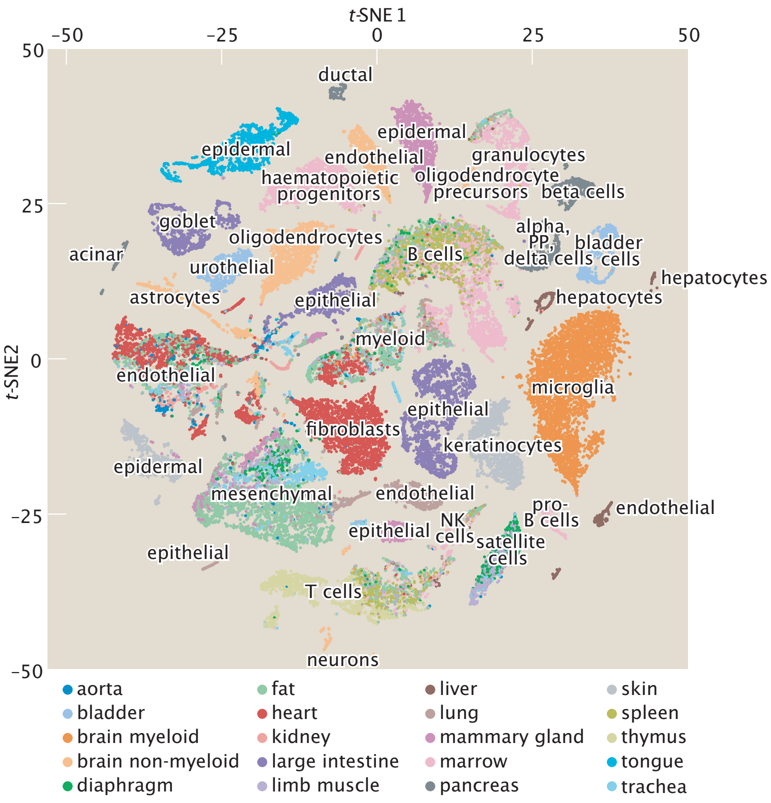


Figure 2: Graphical representation of the *Tabula muris* single-cell sequencing data. Individual cells of different organs in the mouse were subjected to single-cell transcriptome sequencing. Each dot represents a single cell, with its high-dimensional gene expression vector reduced to a two t-SNE lower dimensional representation. Clustering and manual annotation reveal different tissues and cell types. Adapted from The Tabula Muris Consortium et al., *Nature* 562:367-372, 2018.

genes are constitutively expressed, and that low and high gene expression levels correspond to an average of 10 and 35 mRNA molecules per cell, respectively. To remind ourselves of what the null hypothesis for constitutive promoters looks like, write the chemical master equation for a constitutive promoter and show that solving this equation in steady state results in a Poisson distribution. In the case of the low and high expression levels, give the formula for the specific Poisson distribution for those two cases.

(c) Plot histograms of the number of mRNA molecules of gene 1 and gene 2 for each cell type, assuming 1,000 cells of each type. This means that you will invoke the Poisson distribution you derived in the previous part of the problem and use it to describe the distribution of mRNA counts for the different cell types.

(d) Generate a synthetic transcriptome matrix **G** with 1,000 cells of each type (for a total of 3,000 cells in your dataset) by sampling from the Poisson distributions that you derived above. Make a plot of this low-dimensional synthetic transcriptome data set consisting of number of mRNA molecules of gene 2 vs. number of mRNA molecules of gene 1, where each dot within the plot corresponds to an individual cell.

Now, we will imagine that we are given this transcriptome data without any more information than the fact that there should be three cell types within it. Note that in reality we will rarely have information about number of cell types within a sample a priori. However, this is a good first step toward building intuition about the challenges of analyzing single-cell sequencing data.

In order to find cell types in our synthetic transcriptome, we will resort to so-called k-means clustering. The steps of this algorithm are illustrated in figure 3 and can be enumerated as follows:

1. The transcriptome data is plotted. In this case, because we only have two genes, this corresponds to a two dimensional plot of the number of mRNA molecules of gene 2 as a function of the number of mRNA molecules of gene 1 for each single cell. A set of N random points within this data set are then selected, with N being the number of clusters we are trying to identify. These N points will be called the centroids.

2. The distance of every data point to the centroids is calculated. Each data point is assigned to its closest centroid. This is our first approximation to the assignment of cells to our three clusters.

3. Based on the categorization of data points, new centroids are calculated. For each cluster, calculate their corresponding centroids by taking the average values of expression for the two genes.

4. Data points are reassigned to their closest centroid. This means that we now need to take every data point and compute the distance to all three updated centroids and then to assign them to the centroid they are closest to.

5. Steps (3) and (4) are repeated until convergence is achieved.

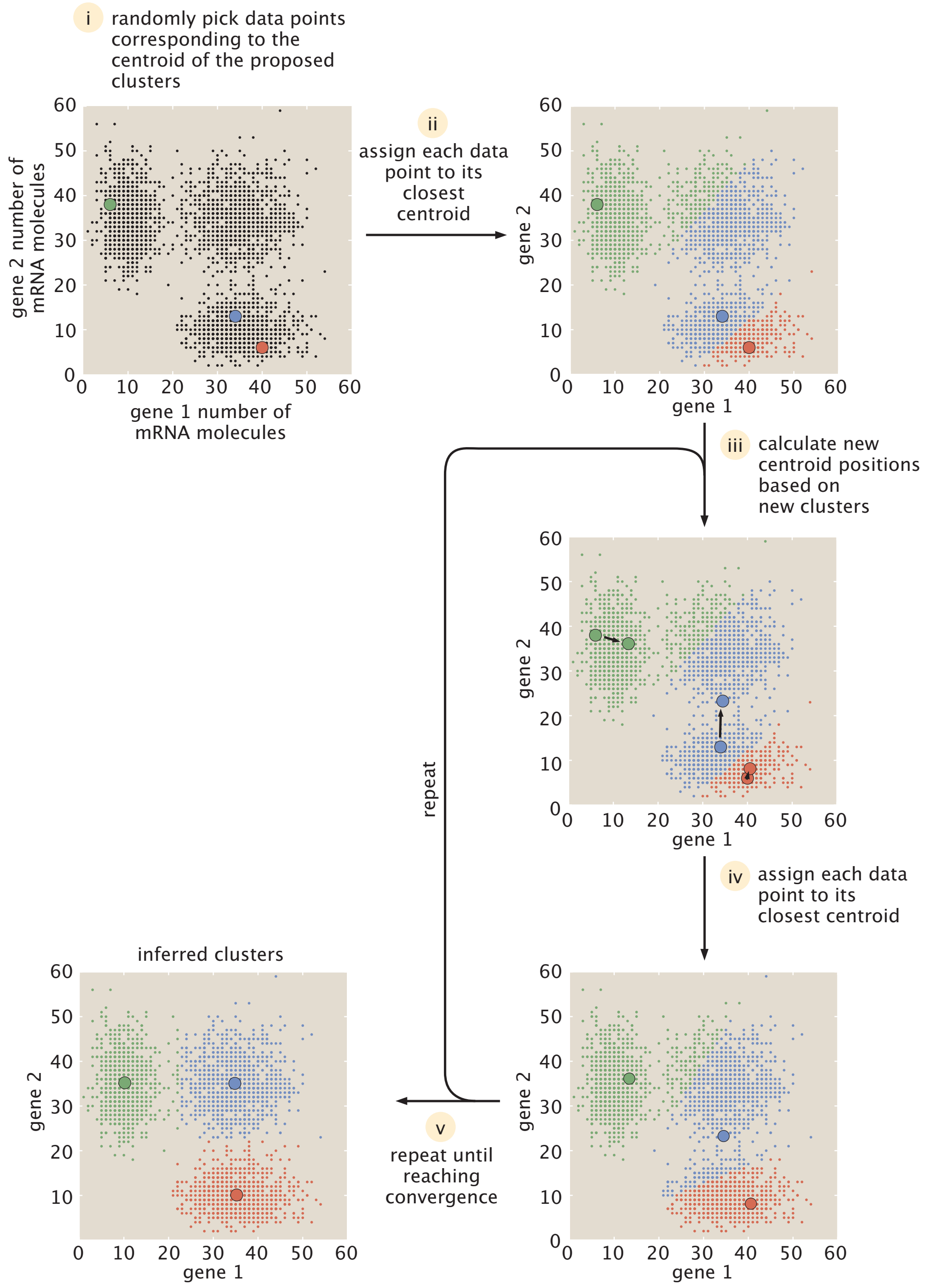


Figure 3: The k-means clustering algorithm. (i) A set of N points are chosen randomly from the dataset to become the centroids of the N clusters to identify. (ii) Each data point is assigned to its closest centroid. (iii) New centroids are calculated for each new cluster. (iv) Data points are reassigned to their new centroids. By iteratively repeating steps (iii) and (iv) convergence can be ultimately reached.

(e) Write a k-means algorithm to find 3 clusters in your synthetic transcriptome data set. In doing so, generate intermediate plots for the iterations of the algorithm such as those shown in Figure 3.

(f) One of the biggest drawbacks of k-means clustering is that we need to commit to a given number of clusters in advance. Explore what happens if you tell your algorithm to look for two and four clusters instead of three. Document some of the final answers from the algorithm and comment on why it converged to that answer. Comment on how all of these answers correspond to what you actually know about the system given that you generated the transcriptomes!

Finally, it is important to note that all algorithms are limited in the sense that they require commitments by specifying parameters. In k-means, we had to commit to a number of clusters. However, there are other approaches to finding clusters that do not require specifying cluster number a priori such as DBSCAN.

(g) Read about DBSCAN and explain how it works by drawing a graphical example (this can be in cartoon form). For this algorithm, what are the parameters we need to commit to?

3 Taylor series

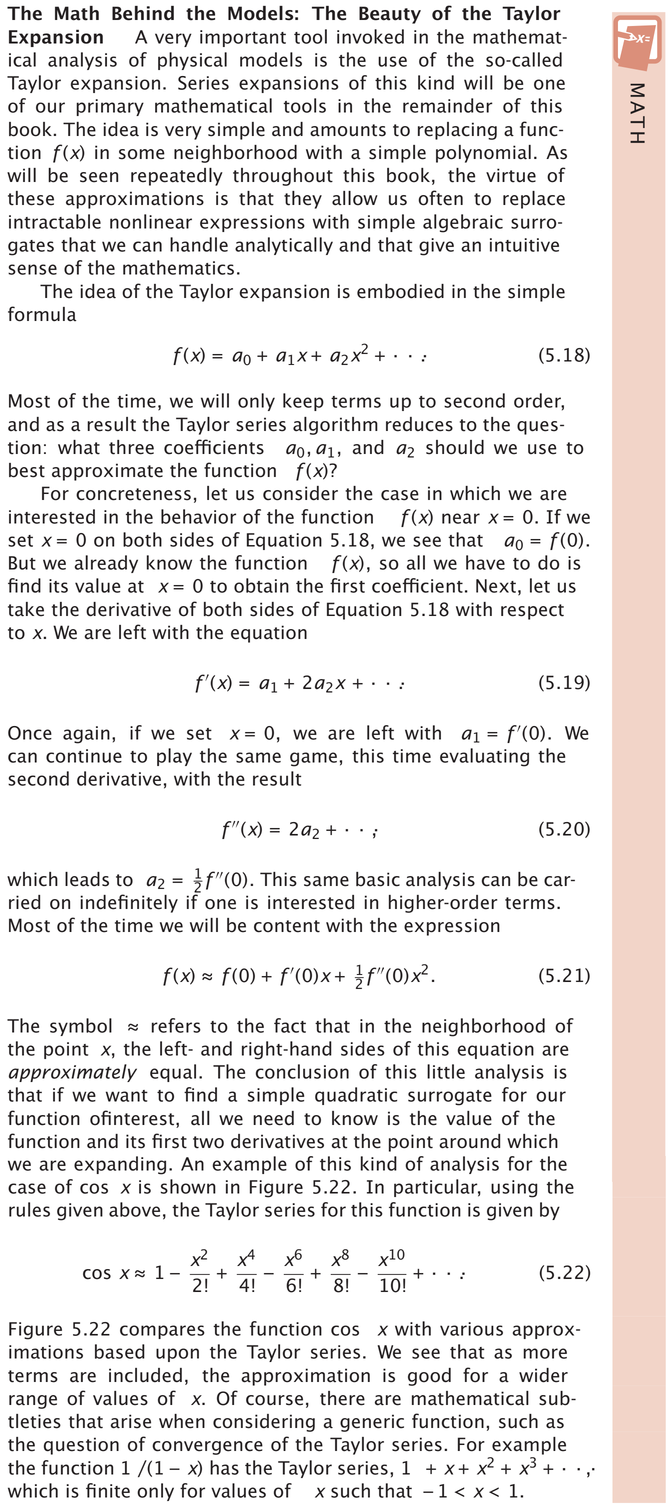
In class, we solved the master equation for mRNA production and concluded that the mRNA production in steady state can be described by a Poisson distribution. To make this possible, we had to invoke the result that

$$\sum_{m=0}^{+\infty} \frac{1}{m!} x^m = e^x = 1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3 + \dots \tag{3}$$

In this problem, we introduce the Taylor expansion to prove that the equation above is correct. This expansion is perhaps one of the most important tools used in the mathematical analysis of physical models.

(a) Read the section “The Math Behind the Models: The Beauty of the Taylor Expansion” on page 215 of PBoC2 shown below in Figure 4.

(b) The idea behind Equation 3 is that, as we sum more of the terms in the equation, our summation will converge to the function e^x. Here, we check this assertion using your favorite programming language. Make a plot like that shown in Figure 5.22 of PBoC (shown below in Figure 5), but for the function e^x. Specifically, plot the function e^x as well as the sum in the equation up until different powers. This means that you will plot e^x, together with 1, $1 + x$, $1 + x + \frac{1}{2}x^2$, etc. Go until the fourth order for a total of five lines on your plot.



MATH

The Math Behind the Models: The Beauty of the Taylor Expansion A very important tool invoked in the mathematical analysis of physical models is the use of the so-called Taylor expansion. Series expansions of this kind will be one of our primary mathematical tools in the remainder of this book. The idea is very simple and amounts to replacing a function $f(x)$ in some neighborhood with a simple polynomial. As will be seen repeatedly throughout this book, the virtue of these approximations is that they allow us often to replace intractable nonlinear expressions with simple algebraic surrogates that we can handle analytically and that give an intuitive sense of the mathematics.

The idea of the Taylor expansion is embodied in the simple formula

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots ; \tag{5.18}$$

Most of the time, we will only keep terms up to second order, and as a result the Taylor series algorithm reduces to the question: what three coefficients a_0, a_1, and a_2 should we use to best approximate the function $f(x)$?

For concreteness, let us consider the case in which we are interested in the behavior of the function $f(x)$ near $x = 0$. If we set $x = 0$ on both sides of Equation 5.18, we see that $a_0 = f(0)$. But we already know the function $f(x)$, so all we have to do is find its value at $x = 0$ to obtain the first coefficient. Next, let us take the derivative of both sides of Equation 5.18 with respect to x. We are left with the equation

$$f'(x) = a_1 + 2a_2 x + \cdots ; \tag{5.19}$$

Once again, if we set $x = 0$, we are left with $a_1 = f'(0)$. We can continue to play the same game, this time evaluating the second derivative, with the result

$$f''(x) = 2a_2 + \cdots ; \tag{5.20}$$

which leads to $a_2 = \frac{1}{2} f''(0)$. This same basic analysis can be carried on indefinitely if one is interested in higher-order terms. Most of the time we will be content with the expression

$$f(x) \approx f(0) + f'(0)x + \tfrac{1}{2} f''(0)x^2 . \tag{5.21}$$

The symbol $\approx$ refers to the fact that in the neighborhood of the point x, the left- and right-hand sides of this equation are *approximately* equal. The conclusion of this little analysis is that if we want to find a simple quadratic surrogate for our function ofinterest, all we need to know is the value of the function and its first two derivatives at the point around which we are expanding. An example of this kind of analysis for the case of cos x is shown in Figure 5.22. In particular, using the rules given above, the Taylor series for this function is given by

$$\cos x \approx 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \frac{x^8}{8!} - \frac{x^{10}}{10!} + \cdots ; \tag{5.22}$$

Figure 5.22 compares the function cos x with various approximations based upon the Taylor series. We see that as more terms are included, the approximation is good for a wider range of values of x. Of course, there are mathematical subtleties that arise when considering a generic function, such as the question of convergence of the Taylor series. For example the function $1/(1-x)$ has the Taylor series, $1 + x + x^2 + x^3 + \cdots ,$ which is finite only for values of x such that $-1 < x < 1$.

Figure 4: Math Behind the Models: The Beauty of the Taylor Expansion. From PBoC2, page 215.

Figure 5.22: Comparison of the function cos x and its Taylor expansion. The curves are labeled by the order of the highest term kept in the Taylor series. For example, $n = 2$ means that the series goes to quadratic order, etc. The cosine function we are approximating is shown in bold for comparison with the approximate expressions.

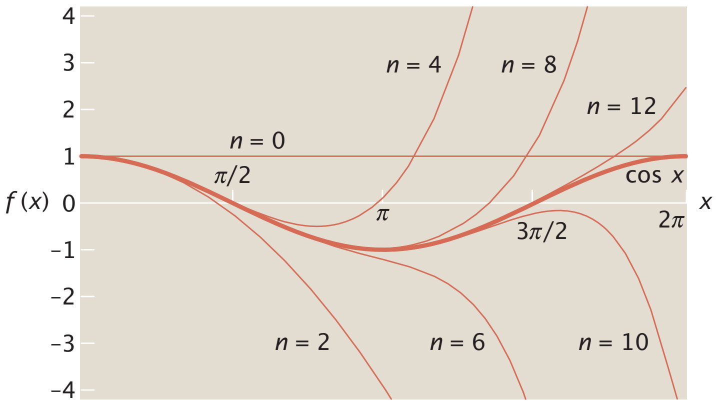


Figure 5: Figure 5.22 from PBoC2.