Chapter 2

Mathematical Preliminaries and Notation

A ***quantitative*** understanding of evolution will require the language of mathematics, so we'll have to spend a little time reviewing the relevant concepts and establishing a common notation. We'll assume you have a general comfort with manipulating equations, calculus, and basic differential equations (ODEs).[1]

However, there is one set of concepts that you might not have seen in previous math or physics courses, but will be extremely useful for studying evolutionary problems. These methods fall under the general headings of ***series expansions*** / ***asymptotic approximations*** / ***self-consistency arguments***.

2.1 Series expansions / asymptotic approximations

We can illustrate these concepts with a simple example which we already know how to solve. Suppose we want to find the positive root of the quadratic equa-

[1] See the *Mathematical Background* document for more details: `https://bgoodlab.github.io/courses/apphys237/math_background.pdf`.

tion

$$\epsilon \cdot x^2 + x - 1 = 0 \tag{2.1}$$

Using the quadratic formula, we can see that this equation has one positive root:

$$x = \frac{-1 + \sqrt{1 + 4\epsilon}}{2\epsilon} \equiv F(\epsilon) \tag{2.2}$$

which we can write as some arbitrary function of ϵ.

As anticipated by our choice of variables, we will often want to understand the behavior of this function in certain limits, e.g. as $\epsilon \to 0$. We can do this by performing a Taylor series of $F(\epsilon)$ around $\epsilon = 0$:

$$x = F(0) + F'(0) \cdot \epsilon + \dots \tag{2.3}$$

In this case, we find[2] that:

$$x = \underbrace{(1)}_{\text{leading order}} + \underbrace{(-\epsilon)}_{\text{next order}} + \underbrace{\dots}_{\text{(h.o.t.)}} \tag{2.4}$$

We'll call the first (non-zero) term in an expansion like this the ***leading-order term***, and we'll call next one the ***next-order term*** (and so on for the higher-order terms, or h.o.t.).

When ϵ is small, each term in this expansion will often be smaller than the one before it, and this suggests a natural approximation scheme for x. The most extreme version is to take the leading-order term and forget the rest. This is how we'll often look at these expansions: the leading-order term will tell us how to approximate x, and the next-order term will tell us how "good" that approximation is.

[2] WolframAlpha can be really useful for calculating series expansions, so feel free to use it if you want to save yourself some time. You can use a query like "expand (-1+sqrt(1+4*epsilon))/(2*epsilon) around epsilon=0"

In other words, we expect that $F(0)$ will be a good approximation for x if the next-order term is much smaller:

$$F'(0)\epsilon \ll F(0)\,. \tag{2.5}$$

We can also write this in terms of ϵ as

$$\epsilon \ll \epsilon^* \equiv \frac{F(0)}{F'(0)}\,. \tag{2.6}$$

In this example, we have $F(0) = 1$ and $\epsilon^* = 1$, so we will write this as

$$x \approx 1; \quad (\epsilon \ll 1) \tag{2.7}$$

and read it as "x is approximately equal to 1 when ϵ is much smaller than 1".

2.1.1 Dominant balance

The use of a Taylor expansion to approximate a function will likely be familiar to many readers. What may be slightly less familiar is that one can do this whole process directly from Eq. (2.1) using a technique known as the ***method of dominant balance***.

This technique gets its name from the observation that, for a random equation involving multiple terms, it will often be the case that two of them are much larger than the others, and are therefore providing the ***dominant*** contribution to ***balance***-ing the two sides of the equation. This leads to an approximation method that involves two steps:

- **Step 1.** We first ***guess*** that one of the terms (in this case, ϵx^2) is much smaller than the others (i.e., we guess that x and -1 are providing the dominant balance in the equation). This yields a much simpler equation,

  $$x - 1 \approx 0\,, \tag{2.8}$$

  which allows us to obtain the ***leading-order approximation***:

  $$x \approx 1\,. \tag{2.9}$$

- **Step 2.** We can then substitute this solution to the original equation to ***check*** whether the approximation is ***self-consistent***. In this case, the neglected term would be equal to

$$\epsilon x^2 \approx \epsilon(1)^2 \approx \epsilon \tag{2.10}$$

so the assumption that ϵx^2 is smaller than the other two terms (x and -1) will be self-consistent if

$$\epsilon \ll 1\,. \tag{2.11}$$

This self-consistency argument will naturally tell us if we had guessed wrong at the beginning. E.g. if we had assumed that the dominant balance was between the ϵx^2 and -1 terms (and that the x term was negligible) we would have found that

$$\epsilon x^2 - 1 \approx 0 \implies x \approx \frac{1}{\sqrt{\epsilon}} \gg 1\,, \tag{2.12}$$

which would have contradicted our assumption that $x \ll 1$. Thus, a good (but rather brute force) is can be to try *all* the possible dominant balances, and see which ones are self-consistent.

Another nice feature of this approach is that it tells us *when* our leading-order approximation will break down. E.g., if we had started from the slightly different quadratic equation,

$$100 \cdot \epsilon \cdot x^2 + x - 1 = 0\,, \tag{2.13}$$

we would have found the same leading-order approximation, but with a slightly different condition of validity:

$$\epsilon \ll \frac{1}{100}\,, \tag{2.14}$$

which is much smaller than above. We can compare this to the more traditional math notation, $\lim_{\epsilon \to 0} x = 1$, which gives us little indication of the region of validity. This feature can be really important when we start making connections to data and experiments. A big theme of this course will be figuring out leading order approximations ($x \approx 1$) as well as their regions of validity ($\epsilon \ll 1$) for a variety of evolutionary problems.

Higher-order corrections. We can use an extension of this dominant balance approach calculate the next-order correction:

- **Step 1.** We ***guess*** that x is approximately equal to the leading-order term, ***plus a small correction***:

$$x \approx 1 + \delta \leftarrow \text{correction term} \tag{2.15}$$

- **Step 2.** We then substitute this guess into Eq. (2.1), expand to lowest order in δ, and solve:

$$\begin{gathered} \epsilon(1+\delta)^2 + (1+\delta) - 1 \approx 0 \\ \downarrow \\ \epsilon(1+2\delta) + (1+\delta) - 1 \approx 0 \\ \downarrow \\ \delta = \frac{-\epsilon}{1+2\epsilon} \approx -\epsilon \end{gathered} \tag{2.16}$$

  where we have kept only the leading-order contribution in ϵ. We can ***verify*** that this answer is ***self-consistent*** by checking that δ is small compared to the leading order term (as we assumed in step 1):

$$|\delta| \ll 1 \implies \epsilon \ll 1 \tag{2.17}$$

  We notice that this is the same condition required for the leading-order approximation to be valid.

Approximations for other regimes. We could have used this same approach to understand the behavior in the opposite limit where $\epsilon \to \infty$. From the Taylor expansion of the quadratic formula in Eq. (2.2), we have

$$x \approx \frac{1}{\sqrt{\epsilon}} - \frac{1}{2\epsilon}\,; \quad (\epsilon \gg 1) \tag{2.18}$$

Can you derive this result using the dominant balance procedure above?

Why is this useful? This whole approximation procedure might seem like a lot of work, given that we could have just directly calculated the answer — for any value of ϵ — from our original quadratic formula in Eq. (2.2). The real payoff comes from the fact that the dominant balance approach still works even when the exact solution is not known. E.g., if we were instead interested in the solution to the quintic equation

$$\epsilon \cdot x^5 + x - 1 = 0 \tag{2.19}$$

then a famous result from group theory tells us that there is *no analytical solution*. However, all of our approximations above still work, and are not really any harder to carry out. This is left as an exercise for the reader. In contrast to some of your other physics and math courses, which tend to emphasize problems that are solvable like Eq. (2.1), we'll see that vast majority of the problems we'll encounter in evolutionary settings will be more like Eq. (2.19). Approximation methods will therefore be very useful for us for understanding what is going on.

While these approximations might seem restrictive, they can be surprisingly useful in practical contexts, e.g. when we are interested in making connections to data. One reason for this is that, given all the possible values of ϵ, the vast majority will have an order-of-magnitude that is either much larger than one or much less than one. Some degree of fine-tuning would be needed for a completely random[3]value of ϵ to be of order 1. This means that in practice, we can

[3]There is some subtlety in this argument, in that it assumes that we are ignorant *order-of-magnitude* of ϵ, in addition to its precise value.

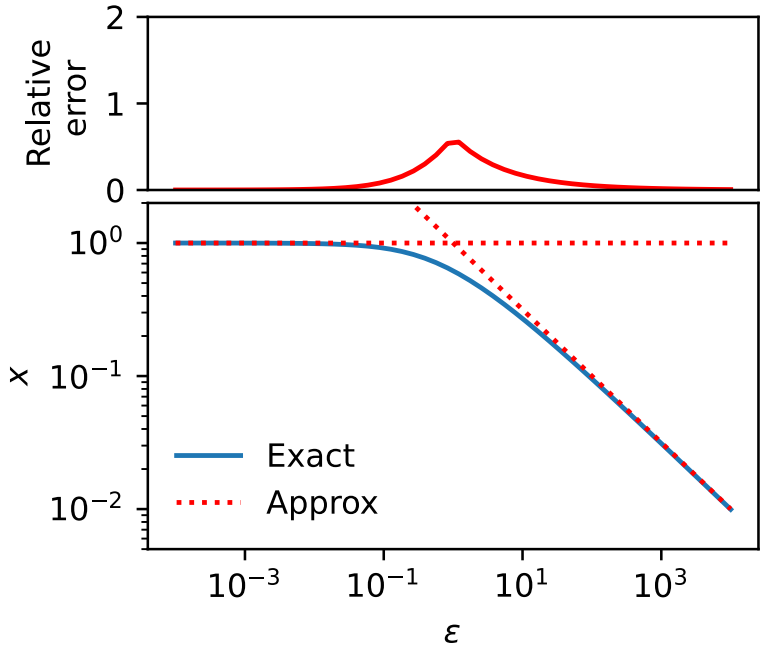


Figure 2.1: Comparing the exact and approximate solutions to the quadratic equation in Eq. (2.1). In this case, the asymptotic approximations are always within a factor of 2 of the exact answer — and often much closer than that.

often get a lot of mileage out of our $\epsilon \ll 1$ and $\epsilon \gg 1$ approximations (Fig. 2.1). Moreover, in the subset of cases where this dichotomy breaks down, it will often signal that there is some additional physical or biological process that is responsible for tuning the value of ϵ to be close to 1. In this way, enumerating the possible asymptotic regimes (and comparing them to data) can be a useful engine of discovery in its own right. We will see several examples of this throughout the course.

Finally, while we have illustrated our dominant balance method using polynomial equations, this same basic approach also works for differential equations, stochastic differential equations, integrals, and many other problems.[4]. So we

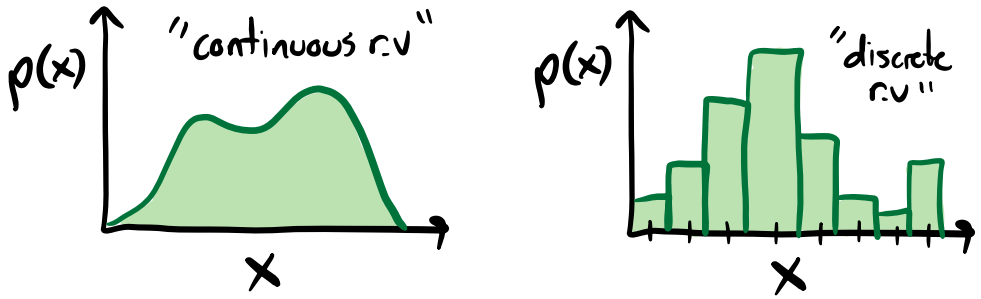
[4] You can take a whole course on these topics. If you are interested in learning more, I highly recommend Michael

will utilize this technique many times throughout this course (and will illustrate with more concrete examples as we go along).

2.2 Randomness and Probability

Since many aspects of evolution are stochastic, the other big tool we'll need is ***probability theory***.

Random variables. We'll assume that you are familiar with the concept of a ***random variable***, $\hat{x}$, which is distributed according to some probability distribution $p(x)$:



We'll often write this as $\hat{x} \sim p(x)$ [pronounced "x is distributed according to the distribution $p(x)$"]. If we're getting sloppy, we might drop the hat.

Means and variances. The ***average*** / ***mean*** / ***expected value*** of $\hat{x}$ will be denoted by

$$\langle x \rangle \equiv \mathbb{E}[x] \equiv \int x \cdot p(x)\, dx\,, \tag{2.20}$$

Brenner's *Physical Mathematics* (`http://esag.harvard.edu/rice/AM201_Brenner,Michael_Course Notes_2010.pdf`) or Hinch's *Perburbation Methods* book.

while the ***variance*** (or ***mean squared deviation***) is defined by

$$\mathrm{Var}(x) \equiv \sigma_x^2 \equiv \langle x^2 \rangle - \langle x \rangle^2 \,. \tag{2.21}$$

and satisfies the scaling property $\mathrm{Var}(c \cdot x) = c^2 \mathrm{Var}(x)$.

Common distributions. We will assume that you are familiar with some common probability distributions. These include discrete distributions like the ***binomial distribution***,

$$n \sim \mathrm{Binomial(N, p)} \implies P(n) = \binom{N}{n} p^n (1-p)^{N-n} \tag{2.22}$$

which models the number of successes in N independent coin flips with success probability p, as well as the ***Poisson distribution***,

$$n \sim \mathrm{Poisson}(\langle n \rangle) \implies P(n) = \frac{\langle n \rangle^n}{n!} e^{-\langle n \rangle} \tag{2.23}$$

which is the limiting form of the Binomial distribution when $N \to \infty$ and $p \to 0$ with $\langle x \rangle = Np$ held fixed. Another common distribution we'll encounter is the ***Gaussian*** or ***Normal distribution***,

$$x \sim \mathrm{Gaussian}(\mu, \sigma^2) \implies p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \,, \tag{2.24}$$

which has mean $\langle x \rangle = \mu$ and variance σ^2. To save space, we will sometimes write this as $x \sim N(\mu, \sigma^2)$.

Note: *Wikipedia is extremely useful for common probability distributions.*[5] *It lists formulas for the means, variances, and other moments (when they are known), as well as useful identities connecting the different distributions.*

[5] e.g. `https://en.wikipedia.org/wiki/Binomial_distribution`.

Joint distributions. We'll also assume you're familiar with the concept of a ***joint distribution*** of 2 (or more) random variables:

$$p(x, y) = \text{“probability that } \hat{x} = x \text{ and } \hat{y} = y \text{ at the same time”} \tag{2.25}$$

If we know the joint distribution, we can calculate the single-variable distribution (also known as the ***marginal distribution***) for one of the variables by integrating over the possible values of the other

$$p(x) \equiv \int p(x, y)\, dy \tag{2.26}$$

which is sometimes known as the ***law of total probability***. We can also define the ***conditional probability***,

$$p(x|y) \equiv \frac{p(x, y)}{p(y)} \equiv \text{“probability of } x \text{ given } y\text{”} \tag{2.27}$$

An important concept is ***statistical independence***, which occurs when the joint distribution factorizes:

$$p(x, y) = p(x)p(y) \tag{2.28}$$

Using the definition of the conditional probability in Eq. (2.27), we can equivalently write this as

$$p(x|y) = p(x) \iff \text{“x is independent of y”} \tag{2.29}$$

Using the definitions above, one can show that for independent random variables,

$$\text{Var}(x + y) = \text{Var}(x) + Var(y) \tag{2.30}$$

(This is left as an exercise for the reader.)

Moment Generating Functions. One topic that might be new is the concept of a ***moment generating function***, defined by

$$H_x(z) \equiv \langle e^{-zx} \rangle = \int e^{-zx} p(x)\, dx\,. \tag{2.31}$$

We can also view this as the Laplace transform of the probability density $p(x)$ (or the Fourier transform if we considered imaginary values of z). One can show that the moment generating function contains all the same information as the probability density $p(x)$, so either one of them will suffice for describing the distribution of x. For example, the moment generating function of a Poisson distribution is given by

$$\text{Poisson}(\langle n \rangle) \iff H_n(z) = e^{-\langle n \rangle (1-e^{-z})} \tag{2.32}$$

while the moment generating function of the Gaussian distribution is

$$\text{Gaussian}(\mu, \sigma^2) \iff H_n(z) = e^{-\mu z + \frac{\sigma^2}{2} z^2} \tag{2.33}$$

Why are moment generating functions useful? As their name implies, they are closely connected to the moments of x. If we expand the exponential in Eq. (2.31) for small values of z, we see that

$$\begin{aligned} H_x(z) &= \int \left[1 - zx + \frac{1}{2} z^2 x^2 + \ldots\right] p(x)\, dx \\ &= 1 - z\langle x \rangle + \frac{z^2}{2} \langle x^2 \rangle + \ldots \end{aligned} \tag{2.34}$$

so expanding $H_x(z)$ around $z = 0$ lets us read off the moments of x. A useful shortcut is that if we can express $H_x(z)$ as an exponential of some function $\phi(x)$, then

$$H_x(z) = e^{\phi(x)} \implies \phi(x) \approx -\langle x \rangle \cdot z + \frac{1}{2} \cdot \text{Var}(x) \cdot z^2 + \ldots \tag{2.35}$$

(This is left as an exercise for the reader.)

The big payoff for moment generating functions is that for *independent random variables*, the generating function of their *sum* satsifies

$$H_{x+y}(z) \equiv \langle e^{-z(x+y)} \rangle = \langle e^{-zx} \cdot e^{-zy} \rangle = \underbrace{\langle e^{-zx} \rangle}_{H_x(z)} \cdot \underbrace{\langle e^{-zy} \rangle}_{H_y(z)} \tag{2.36}$$

This is much easier to calculate than computing the density function of $x + y$ directly.

For related reasons, we will see that in many evolution problems it will often be easier to solve for the generating function $H(z)$ and then invert Eq. (2.31) if we need to find $p(x)$. In practice, this is easiest to do by remembering the generating function for common distributions and then inverting by inspection, as in Eqs. (2.32) and (2.33). (Wikipedia can be extremely useful for this task.)

Central Limit Theorem. Finally, an extremely important result for this course will be the ***central limit theorem***. If $X_1, X_2, \ldots, X_n$ are independent random variables, then for sufficiently large n,

$$\sum_{i=1}^{n} X_i \approx \text{Gaussian} \left[\sum_i \langle x_i \rangle, \sum_i \text{Var}(x_i) \right] \tag{2.37}$$

for certain classes of X_i. If the X_i all the same mean and variance, we will often write this as

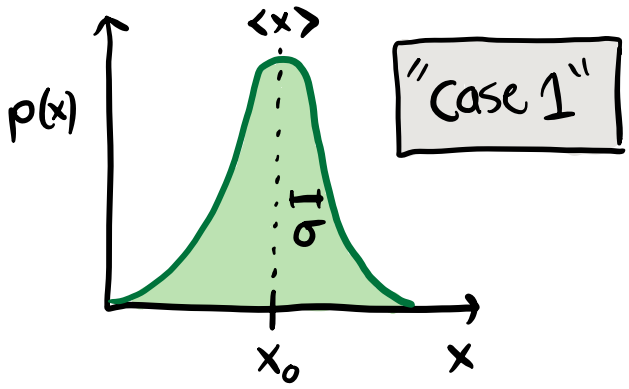
$$\frac{1}{n} \sum_{i=1}^{n} X_i \approx \langle x \rangle \pm \sqrt{\frac{\text{Var}(X)}{n}} \tag{2.38}$$

which shows that the spread of the mean of a bunch of observations scales like $1/\sqrt{n}$ when n is large. We will explore the limits of this approximation when we consider the Luria-Delbrück experiment in Problem Set 1.

2.2.1 Some intuition about random variables

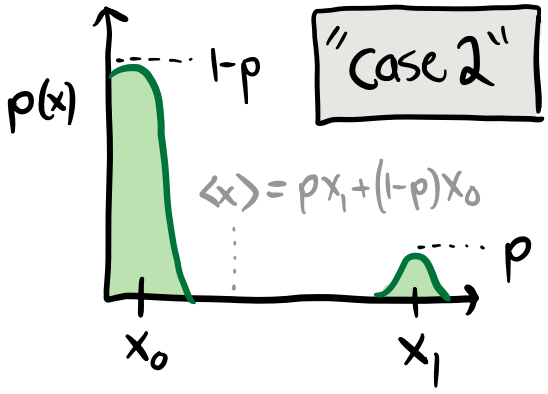
Probability is hard because it forces us to reason a whole range of outcomes all at once. In practice, we'll often want some way of summarizing the ***"typical"*** behavior. There are two main classes of behavior that we will encounter.

- **Case 1 ("fuzzy noise").** In some cases, the distribution of a random variable x will be concentrated around its average value $\langle x \rangle$, with a small amount of spread σ:



  An example would be a $\text{Binomial}(N, p)$ distribution when $Np \gg 1$, or the sum of large number of random variables under the central limit theorem. In these cases, the average of x will often be a good summary of its ***typical*** behavior (similar to our everyday usage of the word "average").

- **Case 2 ("jagged noise").** In other cases, the distribution of x can acquire a bimodal shape:



with most of the probability concentrated at one value (x_0) and a small probability p of having another value (x_1). An example would be the $\text{Binomial}(N, p)$ distribution when $Np \ll 1$. In this case, we can often wind up in a scenario where ***no actual realization of x will be close to the mean value $\langle x \rangle$***. Thus, the average of x is a poor summary of the typical behavior — we would be much better off guessing that $x \approx x_0$, while being prepared for the small possibility that $x \approx x_1$.

This distinction between "fuzzy" and "jagged" noise becomes important if we want to perform an action based on the value of x. For example, we might want to apply a nonlinear function, $F(x)$, which could represent the number of descendants that x mutations leave in a population at some later time.

- In **Case 1 ("fuzzy noise")**, we can often get a lot of mileage out of the approach of treating the noise as a small perturbation, and applying some of the approximation methods described above. Taylor expanding $F(x)$ around $x_0 = \langle x \rangle$, we find that

$$F(x) \approx F(x_0) + F'(x_0)(x - x_0) \approx F(\langle x \rangle) \pm F'(\langle x \rangle)\sigma \,, \qquad (2.39)$$

where we have substituted $x = \langle x \rangle \pm \sigma$. You might recognize this as the ***error propagation formula*** taught in introductory physics or engineering labs. The interpretation here is similar: the "typical behavior" of $F(x)$ is well-approximated by the deterministic portion, $F(\langle x \rangle)$, with the noise introducing a small amount of spread around this value. Using the approaches described in Section 2.1, can now see that this error propagation formula will be a good approximation provided that

$$\sigma \ll \frac{F(\langle x \rangle)}{F'(\langle x \rangle)} \,. \tag{2.40}$$

(Note that depending on the shape of $F(x)$, this could be looser or more stringent than the requirement that $\sigma \ll \langle x \rangle$.)

- In **Case 2 ("jagged noise")**, we'll need to explicitly consider the bifurcating outcomes,

$$F(x) \approx \begin{cases} F(x_0) & \text{w/ prob } 1-p, \\ F(x_1) & \text{w/ prob } p. \end{cases} \tag{2.41}$$

  In this case, the typical behavior is often well approximated by the deterministic value $y \approx F(x_0)$, while the rare exceptions where $y \approx F(x_1)$ must be treated separately.

Much of the randomness we're used to encountering is of the Case 1 variety (e.g., measurement error in an introductory physics lab, or the mass of a given fruit fly). However, when modeling evolutionary dynamics, we will often encounter phenomena that look more like Case 2, and this general strategy of breaking things up into "typical" and "exceptional" outcomes will be useful. (You'll have a chance to work through a concrete example of this in the Luria-Delbrück problem in Homework 1.)

Moreover, just like we saw with the quadratic equation example in Eq. (2.1), if we consider all possible combinations of N and p in a $\text{Binomial}(N, p)$ distribution, the cases where $Np \ll 1$ and $Np \gg 1$ will cover most of the vast

majority of the parameter space. This suggests that this crude dichotomy may often be useful in practice — and that the exceptions where N and p are specifically tuned to have $Np \sim \mathcal{O}(1)$ might signal some that there is some interesting feedback mechanism at play. For these reasons, you may find it useful to keep these two pictures in the back of your head as we deal with random phenomena throughout the course. I'll try to emphasize specific examples as we go along.