APPHYS 237 / BIO 251 – Mathematical Prerequisites

This course is designed to be accessible to students with a broad range of backgrounds in physics, biology, applied mathematics, engineering, etc. The main prerequisites for the course are mathematical, in particular, proficiency in multivariable calculus and basic familiarity with differential equations and probability theory.

The goal of this document is to list some of the mathematical concepts you should be familiar with in order to thrive in the course. Depending on your learning style, it may be possible to pick up one or two of these concepts during the course itself (particularly if you are the kind of person who likes learning through applications). However, we will be building on many of these concepts to introduce new mathematical tools that we will need for the course, so it may be difficult to catch up if you are unfamiliar with too many of the prerequisites.

Mathematical Prerequisites

Calculus

- **Derivatives**, e.g. if $f(x) = \exp\left(-x^2\right)$, then

$$f'(x) \equiv \frac{\partial f}{\partial x} = -2x \exp\left(-x^2\right) \tag{1}$$

$$f''(x) \equiv \frac{\partial^2 f}{\partial x^2} = -2 \exp\left(-x^2\right) + 4x^2 \exp\left(-x^2\right) \tag{2}$$

- **Integrals**, e.g.

$$\int x e^{-x^2} dx = -\frac{1}{2} e^{-x^2} \tag{3}$$

$$\int x e^{-x} dx = -(1+x) e^{-x} \tag{4}$$

- **Series expansions**, e.g. Taylor series around $x = a$:

$$f(x) = \sum_{k=0}^{\infty} \frac{(x-a)^k}{k!} \left.\frac{\partial^k f}{\partial x^k}\right|_{x=a} \tag{5}$$

  e.g., around $x = 0$,

$$e^x = 1 + x + \frac{x^2}{2} + \ldots = \sum_{k=0}^{\infty} \frac{x^k}{k!} \tag{6}$$

$$\frac{1}{1+x} = 1 - x + x^2 + \ldots = \sum_{k=0}^{\infty} (-1)^k x^k \tag{7}$$

$$\log(1+x) = x - \frac{x^2}{2} + \ldots = \sum_{k=1}^{\infty} (-1)^k \frac{x^k}{k} \tag{8}$$

  Eq. 7 also yields a formula for the sum of a geometric series with a *finite* number of terms:

$$\sum_{k=0}^{n} x^k = \frac{1 - x^{n+1}}{1 - x} \tag{9}$$

Using truncated Taylor series as approximations, e.g. when "x is close to 0",

$$e^{-x} \approx 1 - x \tag{10}$$

$$(1+x)^k \approx 1 + kx \tag{11}$$

$$\log(1+x) \approx x \tag{12}$$

Differential equations

- **Ordinary differential equations**, e.g.

$$\textbf{Exponential growth:} \quad \frac{\partial f}{\partial t} = sf \longrightarrow f(t) = f(0)e^{st} \tag{13}$$

$$\textbf{Logistic growth:} \quad \frac{\partial f}{\partial t} = sf(1-f) \rightarrow f(t) = \frac{f(0)e^{st}}{1 + f(0)\left(e^{st} - 1\right)} \tag{14}$$

Verification of solutions by substitution. Finding solutions using **separation of variables** and/or **integrating factors**, e.g.

$$\frac{\partial f}{\partial t} = sf + U(t) \longrightarrow f(0)e^{st} + e^{st}\int_0^t U(t')e^{-st'}dt' \tag{15}$$

- **Partial differential equations**, e.g.

$$\textbf{Advection equation:} \quad \frac{\partial f}{\partial t} = -v\frac{\partial f}{\partial x} \longrightarrow f(x,t) = f(x - vt, 0) \tag{16}$$

$$\textbf{Diffusion equation:} \quad \frac{\partial f}{\partial t} = D\frac{\partial f}{\partial x^2} \rightarrow f(x,t) = \int_{-\infty}^{\infty} \frac{f(x',0)}{\sqrt{4\pi Dt}} e^{-\frac{(x-x')^2}{Dt}} dx' \tag{17}$$

Verification of solutions by substitution. (Advanced) Obtaining solutions via **separation of variables** or **Fourier/Laplace transforms**

Probability and statistics

- **Discrete random variables**, $\Pr[a \le X \le b] = \sum_{k=a}^{b} \Pr[X = k]$, e.g.

$$\textbf{Bernoulli}(p)\textbf{ distribution:} \quad \Pr[X = k] = \begin{cases} p & \text{if } k = 1 \\ 1 - p & \text{if } k = 0 \end{cases}, \qquad k = 0, 1 \tag{18}$$

$$\textbf{Binomial}(n,p)\textbf{ distribution:} \quad \Pr[X = k] = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1 \ldots, n \tag{19}$$

$$\textbf{Poisson}(\lambda)\textbf{ distribution:} \quad \Pr[X = k] = \frac{\lambda^k}{k!} e^{-\lambda}, \qquad k = 0, 1, \ldots, \infty \tag{20}$$

- **Continuous random variables**, $\Pr[a \le X \le b] = \int_a^b p(x)\,dx$, e.g.

  Exponential(λ) distribution: $p(x)dx = \lambda e^{-\lambda x} dx\,, \qquad x \ge 0$ (21)

  Gaussian(μ, σ^2) distribution: $p(x)dx = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx\,, \qquad -\infty < x < \infty$ (22)

  Point distribution / δ-function: $\delta(x-a) \equiv \lim_{\epsilon \to 0} \frac{1}{\sqrt{2\pi\epsilon^2}} e^{-\frac{(x-a)^2}{2\epsilon^2}}$ (23)

  such that

$$\int_{-\infty}^{y} f(x)\delta(x-a)dx = \begin{cases} f(a) & \text{if } y > a \\ 0 & \text{if } y < a \end{cases} \tag{24}$$

  This allows us to write a discrete distribution in continuous terms:

$$p(x)dx = \sum_k \Pr[X = k] \cdot \delta(x-k)dx \tag{25}$$

- **Joint distributions**, $p(x, y)$, including

  Marginalization / law of total probability: $p(x) = \int p(x,y)dy$ (26)

  Conditional probability: $p(x|y) = \frac{p(x,y)}{p(y)}$ (27)

  Independence: $p(x,y) = p(x)p(y)$ (28)

- **Expectation values**, e.g.

  Expected value / ensemble average: $\mathbb{E}\left[g(X)\right] \equiv \langle g(X) \rangle \equiv \int g(x)p(x)\,dx$ (29)

  Conditional expectation: $\mathbb{E}\left[g(X)|\,Y\right] \equiv \langle g(X) \rangle \equiv \int g(x)p(x|y)\,dx$ (30)

  Mean: $\mu_X \equiv \mathbb{E}\left[X\right] \equiv \langle X \rangle \equiv \int xp(x)dx$ (31)

  Variance: $\sigma_X^2 \equiv \mathrm{Var}(X) \equiv \mathbb{E}\left[(X-\mu_X)^2\right] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$ (32)

  Moment generating function / Laplace transform: $H(z) \equiv \mathbb{E}[e^{-zX}]$ (33)

  From the properties of Taylor series, the definition of the moment generating function implies that

$$H(z) = \sum_{k=0}^{\infty} \frac{(-z)^k}{k!} \langle X^k \rangle \tag{34}$$

  or

$$\langle X^k \rangle = (-1)^k \left. \frac{\partial^k H}{\partial z^k} \right|_{z=0} \tag{35}$$

- **Central limit theorem:** For independent and identically distributed random variables $X_1, \ldots, X_n$ with a finite mean and variance, define the **sample average** $\overline{X} \equiv \frac{1}{n}\sum_{i=1}^{n} X_i$. Then

$$\mathbb{E}\left[\overline{X}\right] = \mu_X\,, \quad \mathrm{Var}\left(\overline{X}\right) = \frac{\sigma_X^2}{n} \tag{36}$$

and, as $n \to \infty$, the distribution of $\overline{X}$ converges to a Gaussian distribution with mean $\mu_{\overline{X}} = \mu_X$ and variance $\sigma^2_{\overline{X}} = \sigma^2/n$.

Note: A similar approximation applies for any random variable whose distribution is sufficiently **"strongly peaked"** around a characteristic value x_0. In this case, Taylor expansion of $\ell(x) \equiv \log p(x)$ yields:

$$\ell(x) \approx \ell(x_0) + \ell'(x_0)(x - x_0) + \frac{1}{2}\ell''(x_0)(x - x_0)^2 = \ell(x_0) - \frac{|\ell''(x_0)|}{2}(x - x_0)^2 \tag{37}$$

so that $p(x)$ is approximately Gaussian with mean $\mu_X = x_0$ and variance $\sigma_X^2 = |\ell''(x_0)|^{-1}$.

Programming Prerequisites

The problem sets will also make frequent use of basic programming skills, e.g. the ability to read numerical or textual data from a file, do a few calculations with it, and plot the results in a figure. This is best illustrated with an example from the first problem set. If you feel confident carrying out the programming tasks in this problem in your favorite language (or feel like you could quickly learn these things as you go using the internet), then you will be in good shape for the rest of the course as well.

Example problem: molecular evolution and genetic diversity in the influenza virus

The text file `influenza_HA_dna_sequences.fasta` contains a list of 841 complete DNA sequences of the hemagluttinen (HA) gene in influenza virus samples collected between 1968 and 2005. Hemagluttinen is a surface protein that allows the viruses to enter host cells, making it a primary target for neutralizing antibodies. This creates a strong selection pressure for the HA gene to evolve over time to evade these immune defenses.

(a) Calculate the number of single nucleotide differences between the first sample (`A/Aichi/2/1968`) and the remaining samples, and plot the results as a function of the sampling year. How many differences have accumulated over this ∼40 year period? What fraction of the HA gene does this account for?

(b) Calculate the number of genetic differences between all pairs of strains from the same year, and plot the distribution of this quantity aggregated across all years. Estimate the genetic "turnover time" – i.e., how long would we have to wait for the population to accumulate the same number of genetic differences that typically separate co-circulating strains.