1 Introduction

This course is aimed at introducing computer scientists to uses of computers and computational techniques in other areas of science. The number of ways that computers are used in the sciences are many, varied and often extremely sophisticated. The focus of this course will be on "computational science" which involves constructing mathematical models that can be simulated, analysed and solved using computational methods.

The course is split into two parts: in the first 3-4 weeks, we'll look at techniques for finding the roots of equations, solving systems of linear equations and decomposing matrices. These techniques are basic to areas of research known as computational engineering, numerical analysis and applied linear algebra.

In the remaining 8-9 weeks, we'll turn to computational biology, with a focus on bioinformatics and phylogenetics. There, we see how a wide range of computational and mathematical techniques have revolutionised an area of science and allowed us to analyse and interpret huge amounts of genetic data. This area of study has helped us better understand, among other things, the basic workings of life, our evolutionary history, the causes of inherited diseases and the spread of infectious disease.

From a computational point of view, computational biology is a fascinating and active area of research. The techniques we'll study in this part of the course include stochastic and probabilistic modelling, simulation, dynamic programming, estimation and inference.

CS 369 is more mathematical than many CS courses. This is unavoidable given the subject matter. We assume that students have some background in discrete mathematics (matrices, graphs, linear equations) and continuous mathematics (functions, derivatives, integration), and an understanding of basic probability (discrete and continuous random variables, expectation, conditional probability) . However, we recognise that students come to this course from a variety of backgrounds so will provide explanations from quite a basic level in most cases. We do assume that students have a solid foundation in programming in one of Java, C++, Python, Matlab or R.

2 Mathematical modelling and why we need computers

Mathematical models attempt to precisely describe a system in order to better understand it. A model is usually based on observing the system and is often structured to answer a particular question. It is not an exact replica of the system and is not merely a description of the observations. Recorded observations of the system are known as data. Coupled with data, the model allows us to infer unobserved properties of the model (such as model parameters) and predict future outcomes. Careful comparison of outcomes predicted by the model with data (actual outcomes) tell us how accurate the model is and where it needs to be refined.

This process of modelling and observation has, arguably, been used for thousands of years and certainly for hundreds. The complexity of the models we create and study is somewhat determined by our ability to interpret and “solve” them. Before computers, we were largely limited to using models that were analytically tractable — that is, models for which closed form solutions could be found — or for which good approximations could be made by hand. Our ability to fit models to data was severely limited by our human limitations of collecting, storing and processing information by hand.

With the advent of computers, both of these limitations have eased considerably. It is now possible to collect and store massive amounts of data. For example, Genbank, which stores genetic nucleotide sequences contains over 204 billion nucleotide bases in more than 189 million sequences as at the end of 2015, while CERN’s Large Hadron Collider produces 30 PB ($= 30 \times 10^6$ GB) of data annually. And fast computers allow us process this data and to make almost arbitrarily good approximations to models that are far more complex than could be tackled by hand.

However, even with all the data and computing power in the world we need to be careful to propose useful models and tackle them with efficient techniques if we are to make progress in answering questions that interest us. Bad models, bad data or bad computational techniques could all derail our quest for understanding. In this course, we aim to teach good computational techniques and give some insight into some basic modelling and data analysis techniques that will help to tackle and answer a range of interesting questions.

2.1 Why we need to be clever about our computing

Mathematical problems can be classed into problems that are *well-posed* and *ill-posed.* A problem is *well-posed* if

1. A solution exists
2. the solution is unique
3. A small change in the initial condition induces only a small change in the solution

A problem that is not well-posed is ill-posed.

We are interested in the last criterion which can be termed *sensitivity.* Suppose our problem has inputs x and has a solution (or output) y. An *insensitive* or well-conditioned problem is when a change in x causes a change in y that is of similar relative size. A *sensitive* or *ill-conditioned* problem is one in which the change in solution/output can be large relative the the change in input.

Based on this idea, define the *condition number* by

$$cond = \frac{|\text{relative change in solution}|}{|\text{relative change in input data}|} = \left|\frac{\Delta y/y}{\Delta x/x}\right|.$$

Thus a problem is *ill-conditioned* is $cond \gg 1$.

Example: what is the condition number when we evaluate a function $y = f(x)$ at an approximation of x, $\hat{x} = x + \Delta x$, rather than at the true value x?

Solution:

$$cond = \left|\frac{\Delta y/y}{\Delta x/x}\right| = \left|\frac{f(x+\Delta x) - f(x)/f(x)}{\Delta x/x}\right| = \left|\frac{f(x+\Delta x) - f(x)}{\Delta x}\frac{x}{f(x)}\right| \approx \left|\frac{xf'(x)}{f(x)}\right|.$$

So, depending on the function f and the input x, we could get very large condition numbers. □

Example: What is the condition number for the functions $f(x) = x^n$ and $f(x) = e^x$?

Solution: From above, the condition number is

$$cond \approx \left|\frac{xf'(x)}{f(x)}\right|.$$

When $f(x) = x^n$, $f'(x) = nx^{n-1}$, so

$$cond = \left|\frac{xf'(x)}{f(x)}\right| = \left|\frac{x.nx^{n-1}}{x^n}\right| = \left|\frac{nx^n}{x^n}\right| = |n|.$$

So as the degree of the polynomial increases, the problem becomes increasingly ill-conditioned.

Similarly, when $f(x) = e^x$, $f'(x) = e^x$ so

$$cond = \left|\frac{xf'(x)}{f(x)}\right| = \left|\frac{x.e^x}{e^x}\right| = |x|.$$

In this case, the condition number depends on the input argument, x. If x is very large, the problem can be considered ill-conditioned. □