Doctoral Thesis Research Proposal

Demetri Pananos

February 1, 2019

Contents

1 Introduction

2 Literature Review

2.1 Differential Equations

In this section, I discuss elements of differential equation theory. I begin by delineating between different types of differential equations. I then discuss sufficient conditions for existence and uniqueness of solutions to differential equations, followed by a discussion of computing analytic representations of said solutions. I conclude this section with numerical methods for solving differential equations.

What is a Differential Equation

For the purposes of this proposal, a differential equation is an equation relating an unknown function of a single variable to it's derivative. In general, I will be concerned of differential equations of the form

$$\frac{d\mathbf{y}}{dt} = \mathbf{F}(t, \mathbf{y}(t); \boldsymbol{\theta}) \tag{1}$$

Here, $\mathbf{F} : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ is a vector field, and $\boldsymbol{\theta} \in \mathbb{R}^m$ is a vector of parameters for the differential equation. From here forth, I suppress the dependency on $\boldsymbol{\theta}$, but understand that $\mathbf{F}$ may depend on unknown parameters. Differential equations of these forms are called *ordinary differential equations* (ODEs) since they are deterministic and involve derivatives of a single variable. Often, equation (1) is accompanied by a value of $\mathbf{y}$ evaluated at a point in it's domain. This is called an *initial condition* and is written as $\mathbf{y}(t_0) = \mathbf{y_0}$ for some $t_0 \in \mathbb{R}$. The conjunction of equation (1) and an initial value is referred to as an *initial value problem.*

Various complications to equation (1) yield different types of differential equations. If the equation involves partial or mixed partial derivatives, it is called a *partial differential equation.* If $\mathbf{F}$ is a function of a past state of $\mathbf{y}$, it is called a *delay differential equation.* If one, or more, components of $\mathbf{F}$ is a stochastic process, it is called a *stochastic differential equation.* I will not be concerned with these differential equations in this proposal.

Existence & Uniqueness

Not every differential equation which can be written down has a solution. There are sufficient conditions on $\mathbf{F}$ which guarantee a unique solution exists in a bounded region. I present those conditions here without formal proof.

Consider a differential equation described by equation (1) with the initial condition $\mathbf{y}(t_0) = \mathbf{y_0}$. So long as $\mathbf{F}$ is continuously differentiable in a neighbourhood of $(t_0, \mathbf{y_0}) \in \mathbb{R} \times \mathbb{R}^n$, then there is a neighbourhood of the point t_0 such that a unique solution to equation (1) exists satisfying the initial condition.

For the purposes of this proposal, it need only be checked that $\mathbf{F}$ is continuously differentiable to ensure a solution exists. For a full proof of this theorem, see [1, 2].

Solutions for Ordinary Differential Equations

Not every differential equation which has a solution can have that solution written in terms of algebraic and transcendental functions. In particular, this proposal will be concerned with first order linear differential equations, which do have an analytic representation of their solution. Consider a differential equations of the form

$$\frac{d\mathbf{y}}{dt} = \mathbf{A}(t)\mathbf{y}(t) + \mathbf{g}(t) \,. \tag{2}$$

Here, $\mathbf{A} : \mathbb{R} \to \mathbb{R}^{n\times n}$ which may not necessarily be diagonalizable and $\mathbf{g} : \mathbb{R} \to \mathbb{R}^n$. In general, the solution to equation (2) can be written in terms of fundamental matrices, namely

$$\mathbf{y}(t) = \boldsymbol{\psi}(t)\boldsymbol{\psi}^{-1}(t_0)\mathbf{x}_0 + \boldsymbol{\psi}(t)\int_{t_0}^{t} \boldsymbol{\psi}^{-1}(s)\mathbf{g}(s)\, ds \,, \tag{3}$$

Here, $\boldsymbol{\psi}(t)$ is a fundemantal matrix for equation (2). Our key observation here is that so long as the ODE can be form of equation (2), then the solution can be written in terms of analytic functions. For more on solutions to linear differential equations, see [3].

Numerical Solutions to Ordinary Differential Equations

Not every solution to an ODE which *can* be expressed in terms of analytic functions *should* be expressed in terms of analytic functions. If equation (1) contains many parameters, then equation (3) may be sufficiently complex so that evaluation of $\mathbf{y}(t)$ is not practical. In cases like these, or in cases where the solution can not be found in terms of analytic functions, a rich literature of numerical solutions to differential equations exists.

Consider a scalar form of equation (1)

$$\frac{dy}{dt} = f(t, y) \,, \quad y(t_0) = y_0 \,. \tag{4}$$

By approximating the derivative via a finite difference,

$$\begin{aligned} \frac{y(t_0 + h) - y(t_0)}{h} &\approx f(t_0, y(t_0)) \\ y(t_0 + h) &\approx y(t_0) + h \cdot f(t_0, y(t_0)) \\ y_{n+1} &\equiv y_n + h \cdot f(t_n, y_n) \end{aligned} \tag{5}$$

Equation (5) is known as *Euler's Method* [3]. The method successively approximates the true solution at a finite set of times. Euler's method is equivalent to numerical integration of the ODE. From the fundamental theorem of calculus, equation (4) is equivalent to

$$y(t) - y(t_0) = \int_{t_0}^{t} f(s, y(s))\, ds \,.$$

To evaluate the solution at $t_0 + h$, approximate the integral using a Reimann sum

$$\begin{aligned} y(t_0 + h) &= y(t_0) + \int_{t_0}^{t_0+h} f(s, y(s))\, ds \\ &\approx y(t_0) + (t_0 + h - t_0) \cdot f(t_0, y(t_0)) \end{aligned} \tag{6}$$

In either formulation of Euler's method, h is known as the *step size.* Though in general $y(t_0 + h) \neq y(t_0) + hf(t_0, y(t_0))$, if h is sufficiently small, then the result from Euler's method is an acceptable approximation, assuming that y changes linearly with f as its slope. In principle, in the limit as $h \to 0$, the approximations should can become arbitrarily good[1]. Of course, not all step sizes are practical, which has motivated the study of better numerical methods for approximating the solution to a differential equation.

Quality of a Numerical Solution

Since Euler's method is an approximation, it admits some error between the computed numerical solution and the exact solution. Assessments on the quality of a numerical solution can be examined in terms of the *residual.* For any given ODE, if the analytic solution was known, then it would be the case that $dy/dt - f(t, y(t)) = 0$ identically. Since a numerical methods return approximate solutions on a finite set of points (which can then be interpolated via a desired interpolation scheme) then for an interpolated solution, $z(t)$, it will be the case that $dz/dt - f(t, z(t)) = \Delta(t) \neq 0$ in general $\forall t$. The function $\Delta(t)$ is called the *residual.*

The *order of a method* is defined as $\mathcal{O}(\Delta(t))$ and is usually expressed in terms of the step size h. For instance, Euler's method can be shown to be a $\mathcal{O}(h)$ method [4], which means the residual is proportional to the step size. Smaller steps mean a smaller residual (and thus more accurate solution), but obtaining a desired accuracy may mean taking step sizes too small to be practical.

Most texts on numerical solutions to differential equations use the *local error* as a measure of the quality of a solution. Local error can be interpreted as the error incurred on the n^{th} step when there is no error in the $n - 1^{st}$ step, and is usually expressed as a function of the step size h. Alternatively, the local error is the error between $z(t_1)$ and $y(t_1)$. It can be shown that Euler's method has $\mathcal{O}(h^2)$ local error, assuming f has bounded third derivative [4].

Superior Methods for Numerical Solutions

Euler's method is not usually used in practice for solving differential equations because of it's large local error. Better methods, such as the suite of Runge-Kutta methods, have local error up to $\mathcal{O}(h^5)$ [3]. The fourth order Runge-Kutta method is particularly popular for numerically solving ODEs.

The method involves a weighted average of evaluations of f at various points. The scheme

[1] though some care should be taken to ensure the solution converges [4].

is written as

$$y_{n+1} = y_n + h\left(\frac{k_{n1} + 2k_{n2} + 2k_{n3} + k_{n4}}{6}\right) \tag{7}$$

where

$$\begin{aligned}
k_{n1} &= f(t_n, y_n)\\
k_{n2} &= f(t + 0.5h, y_n + 0.5hk_{n1})\\
k_{n3} &= f(t + 0.5h, y_n + 0.5hk_{n2})\\
k_{n4} &= f(t_n + h, y_n + hk_{n3})
\end{aligned}$$

This method has residual $\mathcal{O}(h^4)$ and local error $\mathcal{O}(h^5)$ [3], which explains why MATLAB's implementation of this method is titled `ode45`.

Other numerical schemes exist which can improve the accuracy further. The methods may utilize adaptive step sizes to control the size of the error [4], or they may use integration schemes that decrease the local error by another order of magnitude, or they may be designed to solve what are known as *stiff problems.* Though these complications are interesting, they are not relevant for the purposes of this proposal.

2.2 Bayesian Statistics

In this section, I introduce some key concepts of Bayesian statistics to be used in the remainder of the proposal. I begin with explaining how Bayesianism differs from Frequentism in philosophy. I then introduce Bayes Nets as a tool for writing down complex Bayesian models in such a way as to preserve economy of thought. Finally, I discuss some finer points of Bayesian modelling, such as model diagnostics and MCMC computation.

Bayesians v. Frequentists

Statistical methods taught in most university classes are Frequentist methods. In Frequentism, probability is understood as the long term relative frequency of an event occurring. Consequently, Statisticians assess estimators by considering the behaviour of the estimator under repeated construction.

This is exemplified by the confidence interval, which is named so not because it has a 95% probability of containing the true estimand[2], but because the long term relative frequency of confidence intervals containing the true estimated is 95%. Thus, Frequentists never make probabilistic statements about any one confidence interval in particular, only about the behaviour of confidence intervals constructed ad infinitum. Frequentistism is strongly contrasted against Bayesianism, where probability represents a strength in a belief [5]. Under the Bayesian paradigm, it is completely acceptable to make probabilistic statements about a particular interval. In fact, all inferences made from a Bayesian data analysis are made in terms of probabilistic statements.

Bayesian Networks

Core to Bayesian statistics is Bayes' Theorem

$$P(\boldsymbol{\theta}|\mathbf{x}) \propto P(\mathbf{x}|\boldsymbol{\theta})P(\boldsymbol{\theta}) \; . \tag{8}$$

Bayesian's refer to $P(\boldsymbol{\theta}|\mathbf{x})$, $P(\mathbf{x}|\boldsymbol{\theta})$, and $P(\boldsymbol{\theta})$ as the *posterior*, *the likelihood*, and *the prior* respectively. Since the product of Bayes' theorem is a probability distribution (i.e. the probability of the parameters conditioned on the data), inferences resulting from a Bayesian analysis are expressed in probabilistic statements. Bayesian modelling begins by specifying a full probability model for the phenomenon. A likelihood for the data generating process is specified, and prior knowledge about the parameters is codified in terms of a probability distribution (i.e. the prior). Conditioning on the observed data is performed, and the posterior is calculated and interpreted. Finally, the resulting model is evaluated and the implications of the resulting posterior are assessed.

Bayesian models can become quite complex, so to ease economy of thought, Bayesian networks (Bayes nets) can be used to exposit the relationship amongst the various parameters and observed data. A Bayes net is a directed acyclic graph which represents a factorization of the

[2] To the dismay of students learning about probability for the first time.

joint probability distribution of the model. The nodes of the graph denote random variables, while the edges denote dependence of the child node on the parent node [6].

Shown in figure 1 is an example of a Bayes net for Gelman's rat tumour example in [5], which I explain here. A total of $N = 71$ experiments on lab rats have been conducted in the past to assess the risk of developing endometrial stromyl polyps. A rat can either develop the endometrial stromal polyp, or not develop the endometrial stromyl polyp, so the number of rats which develop the polyp, y_i, is modelled as binomial. Each of the 71 previous experiments is modelled as having it's own risk of developing the polyp, θ_i, which we postulate are drawn from a beta distribution with parameters α and β.

Traditionally, the model would be written as

$$\begin{aligned}[\alpha, \beta]^T &\sim P(\alpha, \beta)\\ \theta_i &\sim \text{Beta}(\alpha, \beta) \quad i = 1 \ldots N\\ y_i &\sim \text{Binomial}(\theta_i; n_i) \quad i = 1 \ldots N\end{aligned}$$

Here, $P(\alpha, \beta)$ is the prior for the parameters of the beta distribution, and n_i is the number of rats in the i^{th} experiment. This same model is written as a Bayes net in figure 1. The node containing α and β is the parent of θ, indicating that θ relies on α and β, and y is the child of θ, indicating that y relies on θ. The rectangle surrounding θ and y signifies that there are N such copies of these random variables. Instead of explicitly writing out all $N = 71$ of these random variables, we instead place them in what is known as a "plate", and indicate in the bottom corner how many replicates there are. Items in a plate are considered to be independent and identically distributed. Bayes nets make it very easy to write out the posterior distribution of the parameters. We simply follow the net from the bottom up, writing $p(\alpha, \beta, \theta|y) \propto p(y|\theta)p(\theta|\alpha, \beta)p(\alpha, \beta)$, conditioning each node on it's parent nodes.

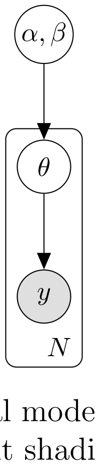

Figure 1: Bayes net for hierarchical binomial model. Note here that nodes with shading correspond to observed data, while nodes without shading are latent.

Bayes nets can be used to do inference, and a literature of algorithms and methods exists for the purposes of doing so [6, 7]. Here, and in subsequent writings, I will use them solely for economy of thought when describing models.

Model Assessment

Once a model is specified, and the posterior for the parameters obtained, the model fit, not only to the data but also to the practitioner's substantive knowledge, must be assessed.

Since the result of a Bayesian analysis is a posterior distribution of the model parameters, it is easy to simulate data from the data generating process. Let y be observed data, and θ be a vector of (hyper)parameters for the model. Denote $\tilde{y}$ as replicated data from the data generating process, or as Gelman writes, "data that we *would* have seen tomorrow if the experiment that generated y today were replicated with the same model and the same value of θ that produced the observed data" [5, page 145]. Then the distribution of the replicated data conditioned on the observed data is

$$p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)\, d\theta \,. \tag{9}$$

The distribution in equation (9) is called the *posterior predictive distribution.* It stands to reason that if the model fits the data well, then observed data should look plausible under the posterior predictive distribution. Simulated data sets are generated from equation (9) and are compared to the observed data. Any systematic differences between observed and simulated may point to areas in which the model can be improved.

Aspects of the observed data can be summarized into a *test quantity* which is then compared against replicated data. This is usually some summary statistic of the observed data. Tail area probability computations can be used to quantify the observed data's departure from the posterior predictive simulations. If $T(y)$ is a test quantity, then the tail area probability is $P(T(\tilde{y}) > T(y))$. This is similar to the Frequentist p-value.

Markov Chain Monte Carlo & Modern Methods

The integrals in Bayesian statistics quickly become intractable when considering unusual models or relaxing assumptions for simple models. Consequently, computational methods have been developed to aid in fitting models. The result is the ability to sample from the posterior distribution without having to know the exact analytical form of the posterior density.

The suit of computational methods for sampling from the posterior are called *Markov Chain Monte Carlo* (MCMC) methods. These methods simulate Markov chains whose limiting distribution is the posterior distribution [8]. The most common MCMC methods for drawing samples from the posterior are The Metropolis-Hastings Algorithm and Gibbs Sampling [5, 9]. Recently however, these algorithms have given way to more efficient algorithms, known as *Hamiltonian Monte Carlo* (HMC) methods. HMC is inspired by Hamiltonian mechanics. The posterior is idealized as a surface on which a particle of mass m rolls after being given a random position and momentum vector. The geometry of the surface influences the movement of the particle, and thus influences the samples obtained. HMC is a marvellously interesting method, spanning topics such as physics, numerical differential equations, and differential geometry. The theory for HMC is quite dense, and more appropriate for a computer science department. I refer interested readers to the following resources [5, 8, 9, 10, 11, 12].

Diagnostics for Hamiltonian Monte Carlo

Provided the simulated Markov chain is geometrically ergodic, a law of large numbers and a central limit theorem exists for these Markov chains [8, 13], thus allowing users to be confident in inferences made from the samples obtained. General conditions under which the chain is and is not geometrically ergodic exist [8], but verification of these properties for applied problems is not usually done analytically. Instead, diagnostics are used to assess the reliability of inferences from the Markov chains.

The first diagnostic most applied Bayesians are introduced to is called the potential scale reduction factor, or Gelman-Rubin diagnostic, or $\widehat{R}$. In MCMC and HMC, several chains are usually initialized and allowed to run for sufficiently long to as to (hopefully) arrive at their stationary distribution. If geometric ergodicity holds, then all chains should arrive at the same stationary distribution, and thus be exploring the same space. The Gelman-Rubin diagnostic measures how well the chains are exploring the space by comparing the within chain variance to the between chain variance [5]. In practice, $1.1 < \widehat{R}$ indicates a problem with the chains, and inference should not be made from the samples drawn [13].

Another diagnostic is the effective sample size, n_{eff}. The Markov chains generate correlated samples, not completely random samples, so using existing theories of convergence of functions of random variables are inappropriate. Effective sample size is heuristic to measure how close the samples are to being independent. Effective sample size is defined as

$$n_{eff} = \frac{n}{1 + 2\displaystyle\sum_k \rho(k)} .$$

Here, $\rho(k)$ is the lag-k within chain correlation [5, 14]. If the chains are autocorrelated, then $n_{eff} < n$, and if the chains are completely independent $n_{eff} = n$. If chains are highly correlated, then $n_{eff} \ll n$, and inferences made from the samples should be avoided because of the bias the correlation would impart.

References

[1] Richard K Miller and Anthony N Michel. *Ordinary differential equations.* Academic Press, 1982.

[2] Morris. Tenenbaum and Harry Pollard. *Ordinary differential equations: an elementary textbook for students of mathematics, engineering, and the sciences.* Dover Publications, 1963.

[3] WE Boyce and RC DiPrima. Differential equations elementary and boundary value problems, 2012.

[4] Robert M Corless and Nicolas Fillion. A graduate introduction to numerical methods. *AMC*, 10:12, 2013.

[5] Andrew Gelman, Hal S Stern, John B Carlin, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis.* Chapman and Hall/CRC, 2013.

[6] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics).* Springer-Verlag, Berlin, Heidelberg, 2006.

[7] Daphne Koller, Nir Friedman, and Francis Bach. *Probabilistic graphical models: principles and techniques.* MIT press, 2009.

[8] Samuel Livingstone, Michael Betancourt, Simon Byrne, and Mark Girolami. On the geometric ergodicity of hamiltonian monte carlo. *arXiv preprint arXiv:1601.08057*, 2016.

[9] Richard McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*, volume 122. CRC Press, 2016.

[10] Radford M Neal et al. Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11):2, 2011.

[11] Matthew D Hoffman and Andrew Gelman. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

[12] Michael Betancourt. A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*, 2017.

[13] Robust statistical workflow with rstan. `https://betanalpha.github.io/assets/case_studies/rstan_workflow.html#25_validating_a_fit_in_stan`. Accessed: 2019-01-31.

[14] Robert E Kass, Bradley P Carlin, Andrew Gelman, and Radford M Neal. Markov chain monte carlo in practice: a roundtable discussion. *The American Statistician*, 52(2):93–100, 1998.