These notes draw heavily on *Numerical Recipes*, a valuable resource for entry-level understanding of numerical methods relevant to physics.

1. Statistical Inference and MCMC

Markov Chain Monte Carlo (MCMC) is a way of drawing inferences from data. If you have some data $\vec{d}$ and a set of model parameters $\vec{\theta}$ within some model M, the output of MCMC are a whole bunches of choices of $\vec{\theta}_i$, which (if everything is done correctly) are distributed according to something called the *posterior distribution.* Roughly speaking, the distribution of points will be peaked near the "best" choice of parameters, and their distribution will be distributed according to our uncertainties about the correct parameters.

Note that as we talk about MCMC and its statistical properties, we are *not* going to be talking about "whether M is correct" or "how 'good a fit' M is under model parameters $\vec{\theta}$." All of this assumes M is the correct model—i.e. there is some set of parameters $\vec{\theta}$ under M which describes the real world.

Mathematically, it is helpful to consider the following distribution:

$$P(\vec{d}, \vec{\theta}), \tag{1}$$

which is the probability distribution of a given set of model parameters $\vec{\theta}$ being correct *and simultaneously* us getting the data $\vec{d}$ from our experiment.

The philosophy of Bayesian statistics is that we want to know the probability distribution of the real θ given the results of our experimental results:

$$P(\vec{\theta}|\vec{d}) \tag{2}$$

and we can clearly write:

$$P(\vec{\theta}, \vec{d}) = P(\vec{\theta}|\vec{d})P(\vec{d}) = P(\vec{d}|\vec{\theta})P(\vec{\theta}), \tag{3}$$

which is Bayes' Theorem. Then we can write:

$$P(\vec{\theta}|\vec{d}) = \frac{P(\vec{d}|\vec{\theta})P(\vec{\theta})}{P(\vec{d})} \tag{4}$$

This is the probability distribution—called the "posterior"—that we want! How can we calculate it? If we have a fully specified model then we might know $P(\vec{d}|\vec{\theta})$, which is called the "likelihood." Note that we might also *not* be able to calculate it easily, which is an interesting situation which there are techniques to handle, like simulation-based inference.

But the other quantities are trickier! $P(\vec{\theta})$ is the "distribution" of model parameters under model M, which is called the "prior." What can this distribution possibly mean?? I'm not 100% sure, but I think it is more-or-less true that it is only truly meaningful if your experiment is done within some well-known external situation. For example, say I am a traffic app trying to estimate the speed of traffic (this is $\vec{\theta}$) on a particular road on a particular day from real-time phone location data (this is $\vec{d}$). I may have made that same measurement lots of times in the past. I *should* have a prior $P(\vec{\theta})$ that is based on all of those previous measurements

On the other hand, what if I am a cosmologist trying to infer or put a limit on the cross-section of annihilation of dark matter particles within some dark matter model (this cross-section is $\vec{\theta}$) based on X-ray data at the center of a cluster of galaxies (this is $\vec{d}$). I'm not even sure whether the model of dark matter is correct, let alone what I should expect for the distribution of cross-sections to annihilation. There's only one value of that in the one observable universe—there may not be a "distribution" of the quantity even in principle. So priors are weird when it comes to constraining physical law. Again, I am not sure, but I think the best we can hope for is that the prior does not matter much.

Meaningless or not, however, we *have* to specify a prior if we want the posterior. We just can't say anything about the posterior otherwise.

The last quantity is $P(\vec{d})$, which is the distribution of data sets we could get over *all* choices of $\vec{\theta}$ (when distributed as the prior). This is most definitely very hard to calculate but we rarely actually need it, as we will see.

A Bayesian analysis seeks the posterior distribution of $\vec{\theta}$ given the obtained data $\vec{d}$, under the model M. If the likelihood can be calculated, then it is pretty straightforward to calculate something proportional to the posterior:

$$P(\vec{\theta}|\vec{d}) \propto P(\vec{d}|\vec{\theta})P(\vec{\theta}) \tag{5}$$

I could make a very big, very fine grid over all of my choices $\vec{\theta}$ covered by a non-zero prior in all the parameter dimensions, and then I'd have a great description of the posterior. I could even estimate the value of $P(\vec{d})$ for the $\vec{d}$ I obtained, since it would just be the integral of the right-hand side of the proportionality (since the left-hand side is a normalized probability distribution). I could find the mode (the most probable value of the parameters), the mean:

$$\left\langle \vec{\theta} \right\rangle = \int \mathrm{d}^N\vec{\theta} P(\vec{\theta}|\vec{d})\vec{\theta} \tag{6}$$

the variances and covariances around the mean:

$$C_{ij} = \langle \theta_i \theta_j \rangle = \int \mathrm{d}^N\vec{\theta} P(\vec{\theta}|\vec{d}) \left(\theta_i - \langle\theta_i\rangle\right) \left(\theta_j - \langle\theta_j\rangle\right) \tag{7}$$

and all the other moments.

That's actually a pretty sensible approach for two and sometimes three-dimensional models. But . . . not for much larger than that! If I have 20 parameters in my model there's no way I'm making a grid to cover it. Also, doing the above integrals is a pain. For high-ish dimensional work MCMC gives a good approach.

Please note that the value of MCMC is when you need this distribution—if you just need the *peak*, just optimize for the peak using BFGS or something like that! Also note that although we use the language of Bayesian inference in this section, you can also do pretty much exactly the same set of calculations and reinterpret it as a frequentist result (that is, characterizing the likelihood).

2. MCMC Concept

MCMC uses a "Markov chain." A Markov chain is a sequence of points $\vec{\theta}_n$ where the distribution of point $n+1$ *only* depends on point n; i.e. we can write this distribution as $p(\vec{\theta}_{n+1}|\vec{\theta}_n)$.

We want this sequence to have the property that it is distributed according to the posterior. Usually this is referred to as "ergodicity," a term which confuses non-mathematicians and as far as I can tell isn't really a well-defined enough term for the mathematicians.

For any desired distribution $\pi(\vec{\theta})$, we can get a Markov chain that will be distributed as π if:

$$\pi(\vec{\theta}_1)p(\vec{\theta}_2|\vec{\theta}_1) = \pi(\vec{\theta}_2)p(\vec{\theta}_1|\vec{\theta}_2) \tag{8}$$

This can be shown by asking, for a set of points $\vec{\theta}_1$ distributed as π, what is the distribution of their subsequent points $\vec{\theta}_2$? This distribution can be calculated as follows if the Markov chain has the property in Equation 8:

$$\begin{aligned} p(\vec{\theta}_2) &= \int \mathrm{d}^N\vec{\theta}_1 p(\vec{\theta}_2|\vec{\theta}_1)\pi(\vec{\theta}_1) \\ &= \int \mathrm{d}^N\vec{\theta}_1 p(\vec{\theta}_1|\vec{\theta}_2)\pi(\vec{\theta}_2) \\ &= \int \pi(\vec{\theta}_2)\mathrm{d}^N\vec{\theta}_1 p(\vec{\theta}_1|\vec{\theta}_2) \\ &= \int \pi(\vec{\theta}_2) \end{aligned} \tag{9}$$

So, if the Equation 8 holds, then the distribution π reproduces itself, i.e. is an equilibrium distribution. This means I can start at any point $\vec{\theta}_0$ and *eventually* the resulting Markov chain will have the posterior distribution. Note that there is a question of "burn-in" time that has to be dealt with, if you happen to start at a very unlikely point in the distribution, which generally you will.

3. Metropolis-Hastings

The Metropolis-Hastings algorithm is designed to produce a Markov chain with the properties above. At each step we have $\vec{\theta}_i$ at which we have evaluated the posterior $\pi(\vec{\theta}_i)$. Then we can:

- Generate a candiate $\vec{\theta}_{i+1}$ by drawing from a "proposal distribution" $q(\vec{\theta}_{i+1}|\vec{\theta}_i)$. A Gaussian distribution is a common choice.

- Calculate an acceptance probability:

$$\alpha(\vec{\theta}_{i+1}, \vec{\theta}_i) = \min\left(1, \frac{\pi(\vec{\theta}_{i+1})q(\vec{\theta}_i|\vec{\theta_{i+1}})}{\pi(\vec{\theta}_i)q(\vec{\theta}_{i+1}|\vec{\theta}_i)}\right) \tag{10}$$

- With a probability α, accept the candidate and move on; otherwise toss it, and try again with a new candidate $\vec{\theta}_{i+1}$.

How does this generate a distribution that satisfies Equation 8? First note that distribution of $\vec{\theta}_{i+1}$ is:

$$p(\vec{\theta}_{i+1}|\vec{\theta}_i) = q(\vec{\theta}_{i+1}|\vec{\theta}_i)\alpha(\vec{\theta}_{i+1}, \vec{\theta}_i) \tag{11}$$

Then we can write:

$$\pi(\vec{\theta}_i)q(\vec{\theta}_{i+1}|\vec{\theta}_i)\alpha(\vec{\theta}_{i+1}, \vec{\theta}_i) = \min\left(\pi(\vec{\theta}_{i+1})q(\vec{\theta}_i|\vec{\theta_{i+1}}), \pi(\vec{\theta}_i)q(\vec{\theta}_{i+1}|\vec{\theta}_i)\right) \tag{12}$$

and we can ask what the same quantity would be if we reverse i and $i+1$:

$$\pi(\vec{\theta}_{i+1})q(\vec{\theta}_i|\vec{\theta_{i+1}})\alpha(\vec{\theta}_i, \vec{\theta}_{i+1}) = \min\left(\pi(\vec{\theta}_i)q(\vec{\theta}_{i+1}|\vec{\theta}_i), \pi(\vec{\theta}_{i+1})q(\vec{\theta}_i|\vec{\theta_{i+1}})\right) \tag{13}$$

So these are the same:

$$\pi(\vec{\theta}_i)q(\vec{\theta}_{i+1}|\vec{\theta}_i)\alpha(\vec{\theta}_{i+1}, \vec{\theta}_i) = \pi(\vec{\theta}_{i+1})q(\vec{\theta}_i|\vec{\theta_{i+1}})\alpha(\vec{\theta}_i, \vec{\theta}_{i+1}) \tag{14}$$

And then using Equation 11:

$$\pi(\vec{\theta}_i)p(\vec{\theta}_{i+1}|\vec{\theta}_i) = \pi(\vec{\theta}_{i+1})p(\vec{\theta}_i|\vec{\theta}_{i+1}) \tag{15}$$