BSTA 670 **Spring 2020**

MCMC Sampling and Integration for Bayesian Computing

Arman Oganisian

1.1 Motivation: Bayesian Logistic Regression

Here I'll introduce a modeling example that motivates the need for sampling methods in Bayesian modeling. Consider observing some binary outcome $Y_i \in \{0, 1\}$ and a covariate vector $X_i \in \mathcal{R} \subset \mathbb{R}^p$. That is, X_i is some $p-$dimensional covariate vector that lives in some subset of $\mathbb{R}$. We observe data for n independent subjects, giving us the complete data $D = \{Y_i, X_i\}_{1:n}$. In Bayesian regression modeling we need to specify both a *conditional distribution* for the outcome $p(Y_i \mid X_i, \theta)$ that is indexed by parameter vector $\theta \in \Theta$. We also need to specify a prior on θ, $p(\theta)$.

$$\begin{aligned} Y_i \mid X_i, \theta &\sim p(Y_i \mid X_i, \theta) \\ \theta &\sim p(\theta) \end{aligned} \tag{1.1}$$

In this case, a *logistic regression* seems appropriate since our outcome is binary:

$$\begin{aligned} Y_i \mid X_i, \theta &\sim Ber\Big(expit(X_i'\theta)\Big) \\ \theta &\sim N_p(0_p, I_p) \end{aligned} \tag{1.2}$$

Here, θ is a $p-$dimensional parameter vector that lives in $\Theta = \mathbb{R}^p$. A $p-$variate standard normal prior is specified with mean vector 0 and $p \times p$ identity matrix as the covariance. We know that $\exp(\theta_j)$ corresponds to an odds ratio (OR). This Gaussian prior is expressing the prior beleif that ORs anywhere from $\exp(-3) \approx .05$ to $\exp(3) \approx 20$ are likely. This is pretty accomodating...rarely would we ever believe (or even encounter) an OR as large as 20. Bayesian inference follows from using Bayes' Rule to find the posterior distribution of θ:

$$\begin{aligned} p(\theta \mid D) &= \frac{p(\theta) \prod_i p(Y_i \mid X_i, \theta)}{p(D)} \\ &\propto p(\theta) \prod_i p(Y_i \mid X_i, \theta) \\ &\propto N_p(\theta; 0_p, I_p) \prod_{i=1}^{n} Ber(Y_i; expit(X_i'\theta)) \end{aligned} \tag{1.3}$$

Above, $N_p(\cdot\,;\cdot)$ and $Ber(\cdot\,;\cdot)$ are the density evaluations. We write the posterior, $p(\theta \mid D)$, is "proportional" to the *unnormalized posterior density* $\tilde{p}(\theta \mid D) = p(\theta) \prod_i p(Y_i \mid X_i, \theta)$, because the posterior is just the $\tilde{p}(\theta \mid D)$ scaled by the constant $\frac{1}{p(D)}$.

In "nice" situations we can re-arrange $\tilde{p}(\theta \mid D)$ and recognize it as a kernel of some *known* distribution - therefore identifying $p(\theta \mid D)$ as some known distribution. Unfortunately, even in this simple logistic regression $p(\theta \mid D)$ is not a known distribution. And even if it were, it might be tedious to find analytically. Suppose for example that we found $p(\theta \mid D)$ to be multivariate Normal with some mean and covariance matrix. Then, making posterior inference on each θ_j is easy. Just take the j^{th} component of the mean vector (the posterior mean) as a point estimate. Take the .025 and .975 percentiles (easy to find for the Normal) as a posterior interval estimate. In the absence of such easy analytical solutions, we need to resort to numerical methods to do posterior inference.

1.2 Markov Chain Monte Carlo (MCMC)

In absence of a posterior that is a known, standard distribution, an alternative is to find ways to *draw* from the posterior distribution. Enter MCMC: a way of obtaining samples from the posterior without knowing the posterior. The idea is to start at some initial guess at the parameter vector $\theta^{(0)}$. Then have an algorithm that moves from $\theta^{(0)}$ to $\theta^{(1)}$, then from $\theta^{(1)}$ to $\theta^{(2)}$ etc. If the moves are generated "correctly", this sequence of draws $\{\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \dots\}$ will eventually walk its way to the posterior distribution. This sequence $\{\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \dots\}$ is called a Monte Carlo Markov Chain. It's "Monte Carlo" in that the draws are determined randomly and it's "Markov" because each draw $\theta^{(t)}$ is only generated dependent on the previous draw $\theta^{(t-1)}$. The "chain" is just a synonym for "sequence" here. Formally, the chain $\{\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \dots\}$ is a *stochastic process*.

The language I just used is pretty vague mathematically. I don't plan on specifying it because it would take too much for one lecture. But I'll outline the ambiguity at least

- What does it mean for the moves to be generated "correctly"? Broadly speaking, we make a move from one point in the chain by first *proposing* a move. Then *accepting* the proposed move with some probability. This *transition probability* needs to satisfy certain criteria so that we can have theoretical guarantees about the chain's convergence to the posterior. Because the chain is a stochastic process, we need to know some stochastic process theory to really understand these criteria. I found reading through Hoel et al. [1] to be helpful. But I think other, more recent books may be better.
- What does it mean for distribution of these draws $\theta^{(t)}$ to "converge" to some distribution? Generally, we talk about convergence as being able to make the "distance" between the distribution of $\theta^{(t)}$ and the posterior $p(\theta \mid D)$ arbitrarily small as $t \to \infty$. Different notions of distance are used in the literature - *total variation distance* is common. You can read more about that here.
- CLTs and LLNs exist that give us asymptotic results in this setting where the $\theta^{(t)}$ are *not* i.i.d. These theorems provide similar results to vanilla CLT and LLN under the Markov dependence we see here. These are all really cool and technical but I'm just focusing on intuition here.

1.2.1 Some History

The history of MCMC goes way back to the Manhattan project - where physicists needed a way to sample from the probability distribution. The earliest MCMC paper people usually cite is Metropolis 1953 [2]. I'll describe the problem outlined in this paper briefly.

Metropolis considered a setting where you had N particles, $X_1, X_2, \dots, X_N$. Each one of these particles has random locations in 2-dimensional space because they're moving around randomly. So if you observe these particles in motion, you can record d_{ij} - the Euclidean distance between X_i and X_j. Let the whole set of pairwise distances be $\Delta = \{d_{ij}, \forall (i,j) \in 1:N \times 1:N\}$. Presumably these "particles" were active molecules in the bomb. Metropolis et al. wanted to compute $\bar{n}(\Delta)$ the average density of the other particles on the surface of any given particle. Presumably, you need a lot of density here to get the thing to blow up...but not too much density so that the detonation is controlled? Idk. I'm not a physicist...but the important thing is that quantities like $\bar{n}(\Delta)$ are functions of relative positions.

The difficulty is that the motion (and therefore the relative position, d_{ij} of these particles) is random. Every time you measure the state, you'll get some variation in d_{ij} - what Metropolis call "statistical fluctuations". Apparently in physics the standard probability distribution use to characterize the uncertainty in relative

positions do to this random motion is

$$p(\Delta) = \frac{1}{C} \exp\Big(- \frac{E(\Delta)}{kT} \Big) \tag{1.4}$$

Above, $C = \int \exp\Big(- \frac{E(d)}{kT} \Big) d\Delta$ is the normalizing constant of the distribution - a very high-dimensional integral. In the paper $E(\Delta)$ is the potential energy of all the molecules (what Metropolis calls "the system"),

$$E(\Delta) = \sum_i \sum_{j \neq i} V(d_{ij})$$

Here, the total energy is a sum of the V - individual pairwise potentials. This potential, V, between two molecules is a function of relative distance. Above, k is the Boltzmann constant and T is the temperature parameter. If temperature T of the system increases,

The issue here is that the normalizing constant C is intractable. But we still want samples of d_{ij} because various properties of the system such as $\bar{n}(\Delta)$ are functions of the position.

So Metropolis proposed an MCMC algorithm (now called the Metropolis Sampler, later refined by Hastings to create the Metropolis-Hastings sampler) to obtain draws $\Delta^{(1)}, \Delta^{(2)}, \ldots, \Delta^{(M)}$ draws from the distribution $p(\Delta)$. The use this to estimate features of the system, such as $\bar{n}(\Delta)$ as Monte Carlo averages of these samples:

$$E[\bar{n}(\Delta)] = \int \bar{n}(\Delta) p(\Delta) d\Delta \approx \sum_{m=1}^{M} \bar{n}(\Delta^{(m)})$$

Where E above denotes expectation, not energy as before.

1.2.2 Metropolis-Hastings (MH)

1.2.3 Adaptive MH

1.2.4 Metropolis Adjusted Langevin Algorithm (MALA)

1.2.5 Parallel Tempering

1.3 Monte Carlo Integration

References

[1] Paul G Hoel, Sidney C Port, and Charles J Stone. *Introduction to stochastic processes.* Waveland Press, 1986.

[2] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.