LECTURE 15
Markov chain Monte Carlo

There are many settings when posterior computation is a challenge in that one does not have a closed form expression for the posterior distribution. Markov chain Monte Carlo methods are a general all purpose method for sampling from a posterior distribution. To explain MCMC we will need to present some general Markov chain theory. However, first we first justify Gibbs sampling, this can be done without the use of any Markov chain theory.

The basic problem is we would like to generate samples from

$$\pi(\theta) \equiv f(\theta \mid x) = \frac{f(x \mid \theta) f(\theta)}{f(x)} \equiv \frac{w(\theta)}{Z},$$

here the normalization constant $Z = \int f(x, \theta) d\theta$ is in general intractable. The objective of our MCMC algorithms will be to sample from $\pi(\theta)$ without ever having to compute Z. The computation of $w(\theta)$ is usually tractable since evaluating the likelihood and prior are typically analytic operations.

15.1. Gibbs sampler

The idea behind a Gibbs sampler is that one wants to sample from a joint posterior distribution. We do not have access to a closed form for the joint, however we do have an analytic form form for the conditionals. Consider a joint posterior $f(\theta \mid x)$ where x is the data and $\theta = \{\theta_1, \theta_2\}$. For ease of notation we will write $\pi(\theta_1, \theta_2) \equiv f(\theta_1, \theta_2 \mid x)$. The idea behind the Gibbs sampling algorithm is that the following procedure will provide samples from the joint distribution $\pi(\theta_1, \theta_2)$:

```
1) Set $\theta_2 \sim$ Unif[support of $\theta_2$]
2) Draw $\theta_1' \sim \pi(\theta_1' \mid \theta_2)$
3) Draw $\theta_2' \sim \pi(\theta_2' \mid \theta_1')$
4) Set $\theta_2 := \theta_2'$
5) Goto step 2.
```

We now show why the draws θ_1', θ_2' from the above algorithm are from $\pi(\theta_1, \theta_2)$. The following chain starts with the joint distribution specified by following the above algorithm and proceeds to show it is the joint distribution

$$\begin{aligned}
p(\theta_1', \theta_2') &= \int \pi(\theta_1, \theta_2) \pi(\theta_1' \mid \theta_2) \pi(\theta_2' \mid \theta_1') d\theta_1 d\theta_2 \\
&= \int \pi(\theta_1, \theta_2) \frac{\pi(\theta_1', \theta_2)}{\pi(\theta_2)} \frac{\pi(\theta_1', \theta_2')}{\pi(\theta_1')} d\theta_1 d\theta_2 \\
&= \int \pi(\theta_1 \mid \theta_2) \pi(\theta_2 \mid \theta_1') \pi(\theta_2', \theta_1') d\theta_1 d\theta_2 \\
&= \pi(\theta_1', \theta_2') \left[\int \pi(\theta_1 \mid \theta_2) \pi(\theta_2 \mid \theta_1') d\theta_1 d\theta_2 \right] \\
&= \pi(\theta_1', \theta_2') \left[\int \pi(\theta_1, \theta_2 \mid \theta_1') d\theta_1 d\theta_2 \right] \\
&= \pi(\theta_1', \theta_2').
\end{aligned}$$

One can also derive the Gibbs sampler from the more general Metropolis-Hastings algorithm. We will leave that as an exercise.

15.2. Markov chains

Before we discuss Markov chain Monte Carlo methods we first have to define some basic properties of Markov chains. The Markov chains we discuss will always be discrete time. The state space or values the chain can take at an time $t = 1, ..., T$ can be either discrete or continuous. We will denote the state space as $\mathcal{S}$ and for almost all examples we will consider a finite discrete state space $\mathcal{S} = \{s_1, ..., s_L\}$, this is so we can use linear algebra rather than operator theory for all our analysis. In the context of Bayesian inference it is natural to think of the state space $\mathcal{S}$ as the space of parameter values Θ and a state s_i corresponding to a parameter value θ_i.

For a discrete state Markov chain we can define a Makov transition matrix $\mathbf{P}$ where each element

$$\mathbf{P}_{s_t \to s_{t+1}} = \Pr(s_t \to s_{t+1}),$$

is the probability of moving from one state to another at any time t. We consider a probability vector ν as a vector of L numbers with $\sum_\ell \nu_\ell = 1$ and $\nu_\ell \geq 0$. We will require our chain to mix and have a unique stationary distribution. This requirement will be captured by two criteria: invariance and irreducibility or ergodicity.

We start with invariance: we would like the chain to have the following property

$$\lim_{T \to \infty} \mathbf{P}^T \nu = \nu^*, \quad \forall \nu$$

the limit ν^* is called the invariant measure or limiting distribution. The existence of a unique invariant distribution piles the following general balance condition

$$\sum_{s'} \mathbf{P}(s' \to s) \, \nu^*(s') = \nu^*(s).$$

There is a simple check for invariance given the transition matrix $\mathbf{P}$ by computing

$$\mathbf{P} = U^T \Lambda U,$$

where if we rank the eigenvalues λ_ℓ from largest to smallest we know the largest eigenvalue $\lambda_1 = 1$. We also know that all the eigenvalues cannot be less than -1.

So we now consider

$$\lim_{N\to\infty}\left[\mathbf{P}^N = \left(\sum_\ell \lambda_\ell^N u_\ell u_\ell^T\right)\right]$$

which will converge as long as no eigenvalues $\lambda_\ell = -1$. In addition, all eigenvalues $\lambda_\ell \in (-1, 1)$ will not have an effect on the limit.

Ergodicity or irreducibility of the chain means the following:
There exists an N such that $\mathbf{P}^N(s' \to s) > 0$ for all s' and $\nu^*(s) > 0$.
Another way of stating the above is that the entire state space is reachable from any point in the state space. Again we can check for irreducibility using linear algebra. We first define the generator of the chain $L = \mathbf{P} - I$. We now look at the eigenvalues of L ordered from smallest to largest. We know the smallest eigenvalue $\lambda_1 = 0$ and has corresponding eigenvector $\mathbf{1}$. If the second eigenvalue $\lambda_2 > 0$ then the chain is irreducible and $\lambda_2 - \lambda_1 = \lambda_2$ is called the spectral gap.

For a Makov chain with a unique invariant measure that is ergodic the following mixing rate holds

$$\sup_\nu \|\nu^* - \mathbf{P}\nu\| = O((1-\lambda)^N).$$

We want our chains to mix.

Algorithmically we will design Markov chains that satisfy what is called detailed balance:

$$\mathbf{P}(s' \to s)\nu^*(s') = \mathbf{P}(s \to s')\nu^*(s), \quad \forall s, s'.$$

Detailed balance is easy to check for in an algorithm and detailed balance plus ergodicity implies that the chain mixes. In the next section we see why detailed balance is easy to verify for the most common MCMC algorithm Metropolis-Hastings.

15.3. Metropolis-Hastings algorithm

We begin with some notation we define a Markov transition probability or Markov transition kernel as

$$Q(s'; s) \equiv f(s' \mid s),$$

as a conditional probability of $s' \mid s$, in the case of a finite state space these values are given by a Markov transition matrix. We also have a state probabilities

$$p(s) \equiv \frac{w(s)}{Z},$$

where we can evaluate $w(s)$ using the prior and likelihood. Note that whenever we write $\frac{p(s)}{p(s')}$ we can use the computation $\frac{w(s)}{w(s')}$ as a replacement.

The following is the Metropolis-Hastings algorithm

```
1) $t = 1$
2) $s^{(t)} \sim$ Unif[support of $s$]
3) Draw $s' \sim Q(s'; s^{(t)})$
4) Compute acceptance probability $\alpha$
       $$\alpha = \min\left(1, \frac{p(s')Q(s; s')}{p(s)Q(s'; s)}\right)$$
5) Accept $s'$ with probability $\alpha$: $u \sim$ Unif$[0,1]$, If $u \le \alpha$ then $\begin{cases} t = t+1 \\ s^{(t)} = s' \end{cases}$
6) If $t < T$ goto step 3 else stop
```

The Metropolis-Hastings algorithm is designed to generate $(s^{(1)}, ..., s^{(T)})$ samples from the posterior distribution $p(s)$. We will show soon that the algorithm satisfies detailed balance. Before that we will state a properties of the above algorithm. A common proposal $Q(s'; s)$ is a random walk proposal $s' \sim \mathrm{N}(s, \sigma^2)$. If σ^2 is very small then typically the acceptance ratio α will be near 1, however in this case two consecutive draws $s^{(t)}, s^{(t+1)}$ will be conditionally dependent. If σ^2 is very large then the acceptance ratio α will be near 0, however in this case two consecutive draws $s^{(t)}, s^{(t+1)}$ will be independent. There is a trick of how local/global the steps should be and what acceptance ratio α is good, some theory suggests $\alpha = .25$ is optimal. It is also the case that the first T_0 samples are not drawn form the stationary distribution, the stationary distribution has not kicked in yet. For this reason one typically does not include the first T_0 samples, this is called the burn-in period.

We now show detailed balance. First observe that $P(s \to s') = \alpha Q(s'; s)$. We start with

$$\begin{aligned}
P(s \to s')\nu^*(s) &= Q(s'; s) \min\left(1, \frac{\nu^*(s')Q(s; s')}{\nu^*(s)Q(s'; s)}\right) \nu^*(s) \\
&= \min\Big(\nu^*(s)Q(s'; s), \nu^*(s')Q(s; s')\Big) \\
&= Q(s; s') \min\left(1, \frac{\nu^*(s)Q(s'; s)}{\nu^*(s')Q(s; s')}\right) \nu^*(s') \\
&= P(s' \to s)\nu^*(s').
\end{aligned}$$

LECTURE 16
Hidden Markov Models

The idea of a hidden Markov model (HMM) is an extension of a Markov chain. The basic formalism is that we have two variables $X_1, ..., X_T$ which are observed and $Z_1, ..., Z_T$ which are hidden states and they have the following conditional dependence structure

$$\begin{aligned} x_{t+1} &= f(x_t; \theta_1) \\ z_{t+1} &= g(x_{t+1}; \theta_2), \end{aligned}$$

where we think of t as time and $f(\cdot)$ and $g(\cdot)$ are conditional distributions. In this case we think of time as discrete. Typically in HMMs we consider the hidden states to be discrete, there are more general state space models where both the hidden variables and the observables are continuous. The parameters of the conditional distribution $g(\cdot)$ is often called the transition probabilities and the parameters for observed distribution $g(x_{t+1}; \theta_2)$ are often called the emission probabilities. We will often use the notation $x_{1:t} \equiv x_1, ..., x_t$.

The questions normally asked using a HMM include:

- Filtering: Given the observations $x_1, ..., x_t$ we want to know the hidden states $z_1, ..., z_t$ so we want to infer – $p(z_{1:t} \mid x_{1:t})$.
- Smoothing: Given the observations $x_1, ..., x_T$ we want to know the hidden states $z_1, ..., z_t$ where $t < T$. Here we are using past and future observation to infer hidden states – $p(z_{1:t})$
- Posterior sampling: $z_{1:T} \sim p(z_{1:T} \mid x_{1:T})$

The hidden variables in an HMM are what make inference challenging. We start by writing down the joint (complete) likelihood

$$\text{Lik}(x_1, ..., x_T, z_1, ..., z_T; \theta_1, \theta_2) = \pi(z_1) \prod_{t=2}^{T} f(z_{t+1} \mid z_t, \theta_1) \prod_{t=1}^{T} g(x_t \mid z_t, \theta_2),$$

here $\pi(\cdot)$ is the probability of the initial state. One can obtain the likelihood of the observed data by marginalization

$$\text{Lik}(x_1, ..., x_T; \theta_1, \theta_2) = \sum_{z_1, \ldots, x_T} \left(\pi(z_1) \prod_{t=2}^{T} f(z_{t+1} \mid z_t, \theta_1) \prod_{t=1}^{T} g(x_t \mid z_t, \theta_2) \right).$$

Naively the above sum is brutal since it consists of all possible hidden trajectories. If we assume N hidden states then we would have N^T possible trajectories. We