Lecture 10

Gibbs Sampling

Contents

1 Introduction

Gibbs sampling was discovered by brothers Stuart and Donald Geman in 1984.[1] It is named in named after American statistical physicist J. W. Gibbs.

Gibbs sampling is an MCMC algorithm for:

- drawing samples from a multivariate joint PDF (jPDF) $\pi(x_1, x_2, ..., x_n)$ by,
- (directly) sampling one random variable at a time from its **conditional distribution**,

$$\pi(x_i|x_1, x_2, ...x_{i-1}, x_{i+1}, ..., x_n).$$

This is a popular method, when the the full set of conditional PDFs are available. Important applications which meet these constraints include Bayesian analysis, and some physics problems.

It can be thought of as a special case of the Metropolis-Hastings MCMC method, where the proposal move is designed to guarantee acceptance.

2 Hammersley-Clifford Theorem

Let's me illustrate a simple form of the theorem that connects joint and conditional PDFs. This provides the philosophical and mathematical basis for Gibbs sampling.

The general statement of the theorem is complicated to state and prove, but proving a surprisingly simple special case is sufficient for illustration.

Theorem

Consider a simple 2D jPDF $\pi(x, y)$. Show that it can be expressed in terms of conditional PDFs:

$$\pi(x, y) = \frac{\pi(y|x)}{\int \frac{\pi(y|x)}{\pi(x|y)} dy} \tag{1}$$

Proof

To prove this, we will need a fundamental relationship between joint, conditional, and marginal distributions:

$$\text{joint} = \text{conditional} \times \text{marginal}.$$

Symbolically,

$$\pi(x, y) = \pi(x|y)\pi(y), \tag{2}$$

where, the marginal distribution of y is,

$$\pi(y) = \int \pi(x, y) dx. \tag{3}$$

[1] Geman and Geman, *IEEE Trans. Pattern Anal. Machine Intel.*, 6(6): 721-741, **1984**

Using eqns. 2 and 3, the denominator in eqn. 1 is:

$$\begin{aligned}\int \frac{\pi(y|x)}{\pi(x|y)} dy &= \int \frac{\frac{\pi(x,y)}{\pi(x)}}{\frac{\pi(x,y)}{\pi(y)}} dy \\ &= \int \frac{\pi(y)}{\pi(x)} dy \\ &= \frac{1}{\pi(x)} \end{aligned} \tag{4}$$

Using equation 4 in 1 implies that we need to prove:

$$\pi(x,y) = \frac{\pi(y|x)}{1/\pi(x)} = \pi(y|x)\pi(x),$$

which is true from the joint = conditional $\times$ marginal relationship.

2.1 Philosophical Importance

Recall that if you know the joint PDF $\pi(x,y)$, you can calculate the marginal and conditional distributions from it.

The Hammersley-Clifford theorem says something in the opposite direction. It says "if you know **all** the conditionals, you can (almost) reconstruct the joint PDF!"[2]

Before we move on to examples and algorithms, it is perhaps helpful to generalize the argument for **higher dimensions**.

Let $\mathbf{x} = (x_1, x_2, ..., x_n)$ be a vector of n random variables, and $\mathbf{x}_{-i} = (x_1, ..., x_{i-1}, x_{i+1}, ..., x_n)$ be $\mathbf{x}$ without the i^{th} element x_i. Thus, $\mathbf{x}_{-i}$ is a $n-1$ dimensional vector.

Let $\pi(\mathbf{x})$ be the jPDF of $\mathbf{x}$. The PDF of $\mathbf{x}_{-i}$ may be obtained by integrating or marginalizing x_i from $\pi(\mathbf{x})$,

$$\pi(\mathbf{x}_{-i}) = \int \pi(\mathbf{x})\, dx_i. \tag{5}$$

The conditional distribution of x_i given all the other variables is related to the joint and marginal distributions,

$$\pi(x_i|\mathbf{x}_{-i}) = \frac{\pi(\mathbf{x})}{\pi(\mathbf{x}_{-i})} \propto \pi(\mathbf{x}), \tag{6}$$

since $\pi(\mathbf{x}_{-i})$ is independent of x_i and may be thought of as a normalization constant.

The individual conditionals are proportional to the joint!

$$\pi(x_i|\mathbf{x}_{-i}) \propto \pi(\mathbf{x}).$$

3 Algorithm

So far, we have talked about how Hammersly-Clifford theorem contends that "it is possible" to sample a jPDF $\pi(\mathbf{x})$ using only the conditional PDFs $\pi(x_i|\mathbf{x}_{-i})$.

[2] subject to compatibility conditions; see Arnold and Press, *J. Am. Stat. Assoc.*, **84**(205), 152-156, **1989**.

We have sufficient ammunition to build a roadmap for how exactly to go about things. For concreteness, let consider the algorithm for $n = 3$ dimensions.[3]

Let $\mathbf{x} = \{x_1, x_2, x_3\}$ represent the state, and $\pi(\mathbf{x})$ be the distribution we seek to sample from.

1. Set $i = 0$; Choose initial state $\mathbf{x}^{(0)} = \{x_1^{(0)}, x_2^{(0)}, x_3^{(0)}\}$
2. Generate the next sample $\mathbf{x}^{(i+1)}$ by sampling each x_j, $(j = 1, 2, 3)$ from its conditional PDF
   - sample $x_1^{(i+1)} \sim \pi(x_1|x_2^{(i)}, x_3^{(i)})$
   - sample $x_2^{(i+1)} \sim \pi(x_2|x_1^{(i+1)}, x_3^{(i)})$
   - sample $x_3^{(i+1)} \sim \pi(x_3|x_1^{(i+1)}, x_2^{(i+1)})$
3. Set $\mathbf{x}^{(i+1)} = \{x_1^{(i+1)}, x_2^{(i+1)}, x_3^{(i+1)}\}$; Set $i = i + 1$.
4. Go to step 2, and repeat.

Notes:

- Step 2 involves sampling from a 1D distribution,[4]
- **caution**: we sample $x_2^{(i+1)} \sim \pi(x_2|x_1^{(i+1)}, x_3^{(i)})$, and not from $x_2^{(i+1)} \sim \pi(x_2|x_1^{(i)}, x_3^{(i)})$ etc.

4 Example 1: Bivariate Normal

Let's consider our old friend the .

$$\pi(\mathbf{x}) = \mathcal{N}(\mu, \Sigma),$$

with the mean and covariance matrix,

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 1.0 & 0.8 \\ 0.8 & 1.0 \end{bmatrix},$$

where the correlation coefficient is $\rho = 0.8$.

The conditional distributions are univariate normal distribution with means and standard deviation given by,[5]

$$\begin{aligned} \pi(x_1|x_2) &= \mathcal{N}(\mu_1 + \frac{\sigma_1}{\sigma_2}\rho(x_2 - \mu_2), \sigma_1^2(1 - \rho^2)) \\ \pi(x_2|x_1) &= \mathcal{N}(\mu_2 + \frac{\sigma_2}{\sigma_1}\rho(x_1 - \mu_1), \sigma_2^2(1 - \rho^2)) \end{aligned} \tag{7}$$

Let's use Gibbs sampling to sample from the bivariate normal, given these conditional distributions

We shall need subroutines to (a) perform a Gibbs update (step 2 of algorithm), and (b) a driver to glue the rest of the steps together. I am exploiting the fact that $\sigma_1 = \sigma_2 = 1$ in this example.

[3] It can be generalized to arbitrary dimensions trivially.

[4] ideally we can sample this using direct MC, without rejection!

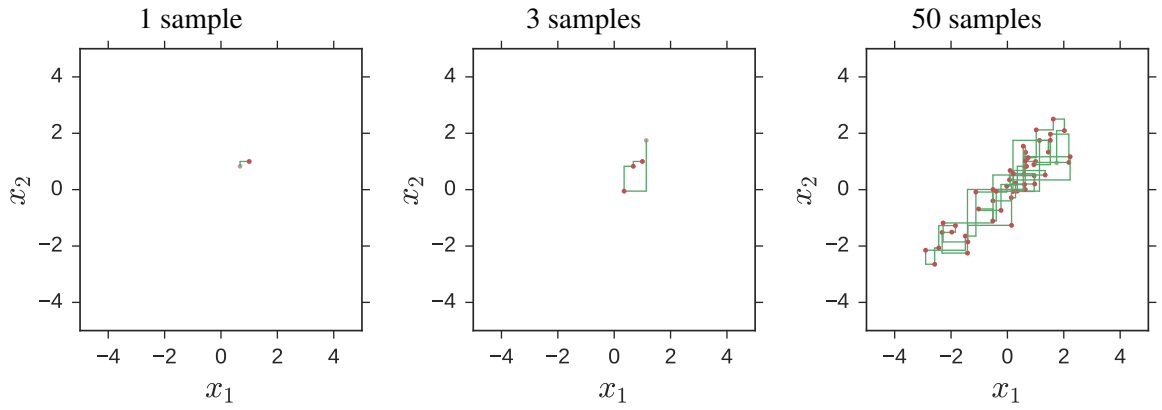
[5] The conditional and marginal PDFs of multivariate normal distributions are also normal. This is a beautiful property. It is also a useful property that is exploited in machine learning using Gaussian processes.

```
def gibbsUpdate(x, mu, rho):
    """use conditional probability formulae"""
    stdv = np.sqrt(1-rho**2)
    newx = np.zeros(x.shape)

    m0      = mu[0] + rho*(x[1]-mu[1])
    newx[0] = np.random.normal(m0, stdv)

    m1      = mu[1] + rho*(newx[0]-mu[0])
    newx[1] = np.random.normal(m1, stdv)

    return newx
```

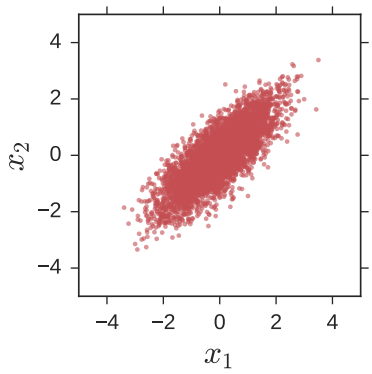
The driver is relatively simple:

```
def driver(x0, nsteps=5000, mu=np.zeros((2,1)), rho = 0.8):

    # initial state
    x[0,:] = x0.reshape(1,2)

    # main loop
    for i in range(1,nsteps):
        newx = gibbsUpdate(x[i-1], mu, rho)
        x[i,:] = newx.reshape(1,2)

    return x
```

We can now run the program with initial state $\mathbf{x}_0 = (1, 1)$, and vary `nsteps`. The green lines are shown to highlight the "sample one dimension at a time" nature of Gibbs sampling.



We can easily generate a lots of samples. A scatterplot of 5000 samples is shown below:



5 Gibbs and MH sampling

Gibbs sampling can be thought of as a particular form of MH sampling.

In MH sampling recall that the accpetance probability is:

$$\text{acc}(o \to n) = \min\left\{1, \frac{\pi_n \pi_{on}}{\pi_o \pi_{no}}\right\}, \tag{8}$$

where the second term inside the braces is the Hastings ratio. The subscripts "o" and "n" stand for "old" and "new" states, and π_{on} is the transition probability from $n \to o$.

Consider a simple 2D example,[6] with a proposal move,

$$\mathbf{x}_o = (x_1, x_2) \to \mathbf{x}_n = (x_1, x_2^*),$$

by drawing $x_2^* \sim \pi(x_2^*|x_1)$ from its conditional PDF. Since Gibbs sampling works on one component at a time, x_1 is the same in both the current and proposed states.

Therefore, the relevant terms in the Hastings ratio, $r = (\pi_n \pi_{on})/(\pi_o \pi_{no})$, are:

$$\begin{aligned}
\pi_n &= \pi(x_1, x_2^*) \\
\pi_o &= \pi(x_1, x_2) \\
\pi_{no} &= \pi(x_2^*|x_1) \\
\pi_{on} &= \pi(x_2|x_1)
\end{aligned}$$

Sticking these expressions in,

$$\begin{aligned}
\frac{\pi_n \pi_{on}}{\pi_o \pi_{no}} &= \frac{\pi(x_1, x_2^*)\pi(x_2|x_1)}{\pi(x_1, x_2)\pi(x_2^*|x_1)} \\
&= \frac{\pi(x_1, x_2^*)}{\pi(x_2^*|x_1)} \frac{\pi(x_2|x_1)}{\pi(x_1, x_2)} \\
&= \pi(x_1) \frac{1}{\pi(x_1)} \\
&= 1,
\end{aligned}$$

where I used the relationship between conditional, joint and marginal distributions for the colored terms. What is the significance of Hastings ratio $r = 1$?

From eqn 8, this means the acceptance probability of proposed moves is 1. Thus, Gibbs sampling is a special case of MH sampling, where the proposal is always accepted.

5.1 Implications

Due to its relationship with MH, Gibbs sampling inherits all the "features" of MH sampling, including:

- burn-in
- correlation in samples

[6] similar to the bivariate Gaussian example

- block-averaging
- convergence diagnostics

The good news is that, we can use the same tools we developed for MH MCMC.

6 Example 2: Sampling from a Uniform Disc

Let us reconsider one more example that we have seen before. Suppose we want to generate samples (x, y) distributed uniformly on a 2D disc of radius 1.

The joint PDF,

$$\pi(x, y) = \frac{1}{\pi}, \quad x^2 + y^2 \leq 1.$$

We have already seen this problem before. Let us consider the following methods for sampling this PDF.

- rejection sampling
- transformation method
- Metropolis MCMC
- Gibbs sampling

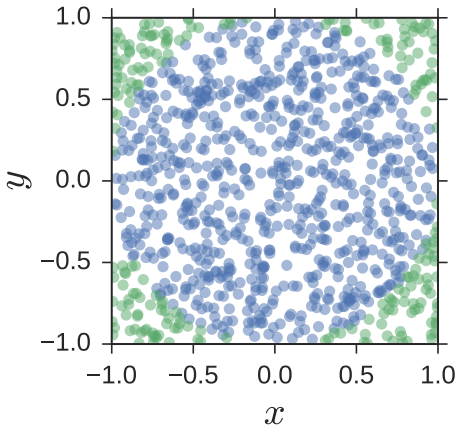
This will allow us to reiterate the relative merits of the different methods.

6.1 Rejection Sampling

Draw samples from independent 1D uniform distributions,

$$x \sim U[-1, 1]$$
$$y \sim U[-1, 1]$$

and retain only those that satisfy $x^2 + y^2 \leq 1$.



The disadvantages of this method are,

- points are rejected
- rejection worse in higher dimensions (n)

$$\text{rejected fraction } = 1 - \frac{\pi^{n/2}}{2^n \Gamma(\frac{n}{2}+1)}$$

6.2 Transformation Method

Transformation between cylindrical (r, θ) and Cartesian (x, y) is well-known. However, sampling from the appropriate distribution is a delicate business, as we saw earlier.

We need to sample r and θ independent from $g(r, \theta) = R(r)T(\theta)$ with,

$$R(r) = 2r, \quad 0 < r < 1$$
$$T(\theta) = \frac{1}{2\pi}, \quad 0 < \theta \leq 2\pi$$

We can sample directly from $R(r)$ by setting $r \sim \sqrt{u}$, where $u \sim U[0, 1]$. This is the most efficient method for this problem.

The disadvantages of this method are: (i) it takes substantial user effort to pose the correct computational problem, and (ii) in higher dimensions, it is almost impossible to use this method.

6.3 Metropolis-Hastings MCMC

It is possible to write a Metropolis sampler to sample from the 2D joint distribution.

$$x_n \sim U[x_c - \delta, x_c + \delta]$$
$$y_n \sim U[y_c - \delta, y_c + \delta]$$

with a suitable δ.

Since probability density is flat, almost all proposals are accepted. Moves to points outside the disc are rejected promptly, and the method does not explore the "corners of the square".[7]

This property helps in higher dimensions, where the sampler does not veer too far away from the "disc" or hypersphere.[8]

The big disadvantage of MCMC is that samples are now correlated!

6.4 Gibbs Sampler

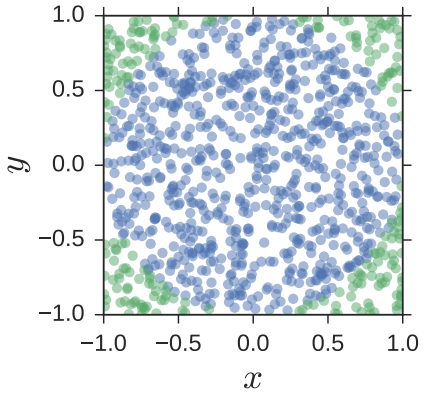
For this problem, it is easy to write down the conditional distributions,

$$\pi(x|y) = U[-\sqrt{1-y^2}, \sqrt{1-y^2}]$$
$$\pi(y|x) = U[-\sqrt{1-x^2}, \sqrt{1-x^2}],$$

since the equation to the circle is $x^2 + y^2 = 1$. Given a value of (say) y, x is forced to lie in the range $-\sqrt{1-y^2} \leq x \leq \sqrt{1-y^2}$.

[7] of course, the MCMC sampler doesn't know of the square

[8] or high-probability probability regions



The advantages are (i) no rejections, (ii) the conditional distribution is easy to sample here, and (iii) the method extends to higher dimensions.

7 Problems

7.1 Thought Questions

(i) Consider the jPDF

$$f(x, y) = kx^2 \exp(-xy^2 - y^2 + 2y - 4x)$$

Show that the conditional distributions are:

$$f(x|y) = \text{Gamma}(3, y^2 + 4)$$
$$f(y|x) = \mathcal{N}\left(\frac{1}{1+x}, \frac{1}{2(1+x)}\right).$$

Hints: $f(x|y)$ is proportional to $f(x, y)$, and terms that do not depend on x can be subsumed into the normalization constant. Thus, $f(x|y) \sim x^2 \exp(-x(y^2 + 4))$, which is similar to the Gamma distribution. Repeat for $f(y|x)$; you will have to complete the square to get it into the right form.

(ii) Consider the 2D Gaussian distribution

$$\pi(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right)\right]$$

Show that the conditional distribution

$$\pi(x|y) = \mathcal{N}(\mu_x + \frac{\sigma_x}{\sigma_y}\rho(y-\mu_y), \sigma_x^2(1-\rho^2)).$$

(iii) To sample points inside a unit circle uniformly using Gibbs sampling, we can use the conditional distributions,

$$\pi(x|y) = U[-\sqrt{1-y^2}, \sqrt{1-y^2}]$$
$$\pi(y|x) = U[-\sqrt{1-x^2}, \sqrt{1-x^2}].$$

Suppose we are at the point $(x_c, y_c) = (0, 0)$ inside the circle. We want to update point to the point (x_n, y_n). Discuss what happens when we incorrectly draw from $\pi(x_n|y_c)$, and $\pi(y_n|x_c)$, instead of $\pi(x_n|y_c)$ and $\pi(y_c|x_n)$. Use a sketch to explain your answer.

7.2 Numerical Problems

(i) Incorrect Gibbs

Consider sampling the bivariate normal distribution using an incorrect Gibbs update.

Suppose we sample from $x_2^{(i+1)} \sim \pi(x_2|x_1^{(i)})$ instead of $x_2^{(i+1)} \sim \pi(x_2|x_1^{(i+1)})$. How does this affect the sampling?

(a) Show contour plots of the correctly and incorrectly sampled distributions

(b) Find the expected value of the distance $E[r = \sqrt{x^2 + y^2}]$ for the correctly and incorrectly sampled distributions.

(ii) From Exam 2018

Consider the 2D joint probability distribution (jPDF) defined over the triangular domain:

$$\pi(x, y) = \begin{cases} e^{-y} & 0 \le x \le y \\ 0 & \text{otherwise.} \end{cases}$$

(a) Find the marginal distributions of x and y, $\pi(x) = \int \pi(x, y)\, dy$ and $\pi(y) = \int \pi(x, y)\, dx$. **Note**: Take care to integrate over the correct domain for x and y. You may verify your setup by ensuring that $\iint \pi(x, y)\, dx\, dy = 1$.

(b) For independent distributions, the joint distribution is the product of the marginals. Test whether $\pi(x, y) = \pi(x)\pi(y)$ and show that x and y are not independent.

(c) Suppose we decide to sample the jPDF using Metropolis MCMC.[9] From a "current" point (x, y) in the domain, we propose a symmetric move,

$$x' \sim U[x - \delta, x + \delta]$$
$$y' \sim U[y - \delta, y + \delta].$$

Write pseudocode/algorithm to decide whether or not to accept the proposed move. What happens if $x' > y'$?

(d) Show that the conditional distributions $\pi(x|y)$ and $\pi(y|x)$ are uniform and exponential, respectively.

(e) Use direct Monte Carlo algorithms to sample 10000 random numbers from $\pi(x|y = 1)$ and $\pi(y|x = 1)$. You may use built-in functions, if convenient. Plot normalized histograms for both cases.[10]

[9] You don't have to actually perform the simulation.

[10] Use 30 bins for all histograms requested for this question.

(f) Starting from $(x, y) = (1, 2)$, write a Gibbs sampler to sample 10000 points from the jPDF $\pi(x, y)$. For burn-in, disregard the first 10% of the samples. Produce a scatter plot of the remaining points. Plot histograms of $\pi(x)$ and $\pi(y)$ and compare with the theoretically derived marginals in part (i).

(iii) Radius of Gyration of a Nonuniform Sphere

The density of a sphere Ω of radius $R = 1$ is given by $\rho(r) = R - r$. The radius of gyration R_g is defined by,

$$R_g^2 = \frac{\int_\Omega \rho(r) r^2 dA}{\int_\Omega \rho(r) dA}$$

We can use MC to estimate R_g^2 by sampling points (x, y, z) from a uniform distribution on the domain Ω. Note that for a uniform sphere $\rho =$ constant, $R_g^2 = 3/5R^2$.

(a) Find the conditional distributions, $\pi(x|y, z)$, $\pi(y|x, z)$, and $\pi(z|x, y)$.

(b) Use Gibbs sampling sample $N = 5000$ points uniformly inside the sphere.

(c) Draw traceplots of the sampled x, y, and z coordinates. Find the autocorrelation of the samples.

(d) Estimate the radius of gyration. Use block-averaging to find the associated uncertainty.