Introduction to Gibbs sampling

Matthew Stephens
University of Chicago
January 23, 2026

See here for a PDF version of this vignette.

Prerequisites

Be familiar with the concept of joint distribution and a conditional distribution. Ideally also with the concept of a Markov chain and its stationary distribution.

Overview

Gibbs sampling is a very useful way of simulating from distributions that are difficult to simulate from directly. However, in this introduction to the key concept, we will use a Gibbs sampler to simulate from a very simple distribution that could be simulated from in other ways.

Gibbs sampling

Suppose X and Y are two binary random variables with joint distribution $\Pr(X = x, Y = y) = p_{X,Y}(x, y)$ given by the following table:

	Y = 0	Y = 1
X = 0	0.60	0.10
X = 1	0.15	0.15

That is, for example, $p_{X,Y}(0, 0) = 0.6$.

The conditional distribution of X given any given value is easy to compute by the usual formula for conditional probability, $\Pr(A \mid B) = \Pr(A \cap B)/\Pr(B)$. For example,

$$\Pr(X = 0 \mid Y = 0) = \Pr(X = 0 \cap Y = 0)/\Pr(Y = 0) = 0.6/0.75 = 0.8,$$

and so

$$\Pr(X = 1 \mid Y = 0) = 1 - 0.8 = 0.2.$$

Similarly,

$$\Pr(X = 0 \mid Y = 1) = 0.1/0.25 = 0.4,$$

and so

$$\Pr(X = 1 \mid Y = 1) = 0.6.$$

We can just as easily compute the conditional distribution of Y for any given value of X:

$$\begin{aligned}
\Pr(Y = 0 \mid X = 0) &= 6/7 \\
\Pr(Y = 1 \mid X = 0) &= 1/7 \\
\Pr(Y = 0 \mid X = 1) &= 1/2 \\
\Pr(Y = 1 \mid X = 1) &= 1/2
\end{aligned}$$

Question: Suppose we start at some value of X, Y and proceed to iterate the following steps:

1. Simulate a new value of X from $\Pr(X \mid Y = y)$, where y is the current value of Y.
2. Simulate a new value of Y from $\Pr(Y \mid X = x)$, where x is the current value of X (that is, the value generated in Step 1.)

What happens? Let's try it.

This function returns 1 with probability p and 0 with probability $1 - p$:

```
rbernoulli <- function (p)
  as.numeric(runif(1) < p)
```

This function samples from the conditional distribution of X given Y:

```
sample_XgivenY <- function(y) {
  if (y == 0)
    x <- rbernoulli(0.2)
  else
    x <- rbernoulli(0.6)
  return(x)
}
```

This function samples from the conditional distribution of Y given X:

```
sample_YgivenX <- function (x) {
  if (x == 0)
    y <- rbernoulli(1/7)
  else
    y <- rbernoulli(0.5)
  return(y)
}
```

Now let's repeat Steps 1 and 2 one thousand times:

```
set.seed(100)
niter <- 1000
X <- rep(0,niter)
Y <- rep(0,niter)
X[1] <- 1
```

```
Y[1] <- 1
for (i in 2:niter) {
  X[i] <- sample_XgivenY(Y[i-1])
  Y[i] <- sample_YgivenX(X[i])
}
res <- data.frame(X = X,Y = Y)
```

Here is what happens for the first 20 iterations:

```
head(res,20)
#    X Y
# 1  1 1
# 2  1 1
# 3  1 1
# 4  1 1
# 5  0 0
# 6  0 0
# 7  0 0
# 8  0 0
# 9  0 0
# 10 0 0
# 11 0 0
# 12 0 0
# 13 0 0
# 14 0 0
# 15 0 0
# 16 0 0
# 17 0 0
# 18 0 0
# 19 0 0
# 20 1 0
```

And here is a summary of what proportion of the rows are of each type:

```
table(res)/niter
#    Y
# X       0     1
#   0 0.617 0.092
#   1 0.154 0.137
```

As you can see, the proportion of iterations in which $X = x$ and $Y = y$ is quite close to $\Pr(X = x, Y = y) = p_{X,Y}(x, y)$. This is not a coincidence!

Explanation

What we have done here is simulate a Markov chain

$$(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), \dots$$

whose *stationary distribution* is $\Pr(X = x, Y = y) = p_{X,Y}(x, y)$.

To see that the pairs (X, Y) form a Markov chain, note that the simulation of X_i is done using only the previous value Y_{i-1}, and the simulation of Y_i is done using only X_i. So simulation of (X_i, Y_i) depends on the previous states only through the immediate previous state (X_{i-1}, Y_{i-1}), which means it is a Markov chain. (And in fact it only depends on Y_{i-1}, but that is not so important.)

To see why it has stationary distribution $p_{X,Y}(x, y)$, imagine simulating X_1, Y_1 from this distribution, so $\Pr(X_1 = x, Y_1 = y) = p_{X,Y}(x, y)$, and in particular $\Pr(Y_1 = y) = \sum_x p_{X,Y}(x, y) = p_Y(y)$.

Now, what is $\Pr(X_2 = x, Y_1 = y)$? Well we know

$$\Pr(X_2 = x, Y_1 = y) = \Pr(Y_1 = y) \Pr(X_2 = x \mid Y_1 = y).$$

And we know from above that

$$\Pr(Y_1 = y) = p_Y(y).$$

And we know that, given $Y_1 = y$, X_2 was simulated from the conditional distribution $p_{X|Y}(x \mid y)$, so

$$\Pr(X_2 = x \mid Y_1 = y) = p_{X|Y}(x \mid y).$$

Putting these together, we have

$$\Pr(X_2 = x, Y_1 = y) = p_Y(y) p_{X|Y}(x \mid y) = p_{X,Y}(x, y).$$

Essentially the same argument shows that $\Pr(X_2 = x, Y_2 = y) = p_{X,Y}(x, y)$. (This is left as an exercise.)

Thus, we have shown that *if* $\Pr(X_1 = x, Y_1 = y) = p_{X,Y}(x, y)$, then also $\Pr(X_2 = x, Y_2 = y) = p_{X,Y}(x, y)$. That is exactly what it means for $p_{X,Y}(x, y)$ to be the "stationary distribution": if we start the chain by simulating from that distribution, then it remains in that distribution after one step, and so it remains in that distribution forever.

Of course, we did *not* start the chain at that distribution. But the above argument shows that this is indeed the stationary distribution. There is a general result that discrete Markov chains "converge" to their stationary distribution provided that they "irreducible and aperiodic" (which this Markov chain is). That is, for large enough n, we should see $\Pr(X_n = x, Y_n = y) \approx p_{X,Y}(x, y)$ no matter where we start. Furthermore, in the long run, the proportion of iterations spent in each state will also converge to this distribution.

This explains the simulation result.