# 7.3 From the Sampled Posterior to the Three Goals

Zongyi Liu

2023-06-15

## 7.3 From the Sampled Posterior to the Three Goals

In Bayesian inference, we need a good description of the posterior distribution. If we cannot achieve that description through formal analysis, nor through dense-grid approximation, then we can generate a lot of representative values from the posterior distribution and use those values to approximate the posterior.

We are trying to estimate our posterior beliefs regarding the underlying probability $\theta$. We start with a prior belief distribution, $p(\theta)$. In the present scenario, $p(\theta)$ is specified by a mathematical function of $\theta$; it is not merely a list of probability masses at discrete values of $\theta$. The value of the mathematical function $p(\theta)$ must be easily computable at any value of $\theta$.

We also have a mathematical likelihood function that specifies $p(D|\theta)$, where the datum $D$ for any one flip is $y = 1$ for heads and $y = 0$ for tails. For a single flip, the likelihood function is the Bernoulli distribution, $p(y|\theta) = \theta^y(1-\theta)^{(1-y)}$.When there are several independent flips, the likelihood is the product of the probabilities of the individual flips.

We start the random walk of the Metropolis algorithm at some candidate value of $\theta$, such as $\theta = 0.5$, and then we propose a jump to a new position. The proposal distribution could be a normal distribution, with some reasonable standard deviation such as $\sigma = 0.2$.

For the choice of $\sigma$; One consideration is that the range of $\theta$ in this application is limited to [0, 1], so certainly we would like the proposal distribution to be narrower than the range of $\theta$. Another consideration is that the proposal distribution should be "tuned" to the width of the posterior, not too wide and not too narrow. When the sample size is small, the posterior is typically not very narrow, and so $\sigma = 0.2$ can work.

An example of the results of the Metropolis algorithm is shown below. The $\theta$ values were generated from using a uniform prior, Bernoulli likelihood, and data in which $z = 11$ and $N = 14$.

### 7.3.1 Estimation

From the heap of representative values of $p(\theta|D)$, we can estimate aspects of the actual $p(\theta|D)$ distribution. For example, to summarize the central tendency of the representative values, it is easy to compute their mean or their median.

#### 7.3.1.1 Highest Density Intervals from Random Samples

Highest density intervals (HDIs) can also be estimated from MCMC samples. One way to do it relies on computing the (relative) posterior probability at each point in the sample.

Essentially, this method finds the water level such that 5% of the points are under water. The remaining points above water represent the 95% HDI region.

If computing the relative posterior probability at each point is onerous or inconvenient (it's not in the present application, but will be in more complex situations), the HDI region can be estimated by another method that uses only the parameter values themselves. The method is limited, however, to treating one parameter at a time, and assumes that the distribution is unimodal.

#### 7.3.1.2 Using a Sample to Estimate an Integral

Suppose we have a large number of representative values from a distribution, then we need to decide a good estimate of the mean of the distribution.

We can express that approximation formally. Let $p(\theta)$ be a distribution over parameter $\theta$. Let $\theta_i$ (notice the subscript $i$) be representative values sampled from the distribution $p(\theta)$. We write $\theta_i \sim p(\theta)$ to indicate that the $\theta_i$ values are sampled according to the probability distribution $p(\theta)$. Then the true mean, which is an integral, is approximated by the mean of the sample:

$$\int d\theta\ \theta\, p(\theta) \approx \frac{1}{N} \sum_{\theta_i \sim p(\theta)}^{N} \theta_i$$

The equation above is just a special case of general principle. For any function $f(\theta)$, the integral of that function, weighted by the probability distribution $p(\theta)$, is approximated by the average of the function values at the sampled points. In math:

$$\int d\theta\, f(\theta)\, p(\theta) \approx \frac{1}{N} \sum_{\theta_i \sim p(\theta)}^{N} f(\theta_i)$$

### 7.3.2 Prediction

The second typical goal for Bayesian inference is predicting subsequent data values. For a given value $y \in \{0, 1\}$, the predicted probability of y is $p(y|D) = \int d\theta p(y|\theta)p(\theta|D)$. Notice that this has the form of the left-hand side of Equation 7.6. Therefore, applying that equation to the predicted probability that the next y equals 1, we have

$$\begin{aligned} p(y{=}1|D) &= \int d\theta\, p(y{=}1|\theta)\, p(\theta|D) \\ &= \int d\theta\, \theta\, p(\theta|D) \\ &\approx \frac{1}{N} \sum_{\theta_i \sim p(\theta|D)}^{N} \theta_i \end{aligned}$$

### 7.3.3 Model Comparison: Estimation of $p(D)$

For the goal of model comparison, we want to compute $p(D) = \int d\theta p(D|\theta)p(\theta)$, where $p(\theta)$ is the prior. In principle, we could just apply Equation 7.6 directly:

$$\begin{aligned} p(D) &= \int d\theta\, p(D|\theta)\, p(\theta) \\ &\approx \frac{1}{N} \sum_{\theta_i \sim p(\theta)}^{N} p(D|\theta_i) \end{aligned}$$

This means that we are getting samples from the prior, $p(\theta)$, perhaps by using a Metropolis algorithm.

Instead of sampling from the prior, we will use our sample from the posterior distribution, in a clever way. First, consider Bayes' rule:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

We can rearrange it to get

$$\frac{1}{p(D)} = \frac{p(\theta|D)}{p(D|\theta)p(\theta)}$$

For any probability density function $h(\theta)$, it is the case that $\int d\theta h(\theta) = 1$. We will multiply the re-arranged Bayes' rule by 1:

$$\begin{aligned}
\frac{1}{p(D)} &= \frac{p(\theta|D)}{p(D|\theta)p(\theta)} \\
&= \frac{p(\theta|D)}{p(D|\theta)p(\theta)} \int d\theta\, h(\theta) \\
&= \int d\theta\, \frac{h(\theta)}{p(D|\theta)p(\theta)}\; p(\theta|D) \\
&\approx \frac{1}{N} \sum_{\theta_i \sim p(\theta|D)}^{N} \frac{h(\theta_i)}{p(D|\theta_i)p(\theta_i)}
\end{aligned}$$

In the derivation above, the transition from first to second lines was just the trick of multiplying by 1. The transition from second to third lines was just algebraic re-arrangement.3 Finally, the transition from third to last lines was application of Equation 7.6.

When the likelihood function is the binomial distribution, it is reasonable that $h(\theta)$ should be a beta distribution with mean and standard deviation corresponding to the mean and standard deviation of the samples from the posterior. The idea is that the posterior will tend to be beta-ish, especially as $N$ gets larger, regardless of the shape of the prior, because the Bernoulli likelihood will overwhelm the prior as $N$ gets large.

In general, there might not be strong theoretical motivations to select a particular $h(\theta)$ density. No matter. All that's needed is any density that reasonably mimics the posterior. In many cases, this can be achieved by first generating a representative sample of the posterior, and then finding an "off-the-shelf" density that describes it reasonably well.