**MCMC**

Suppose we have some distribution \(p(\theta)\) that we can’t compute because the normalizing constant is intractable, but we know it is proportional to \(f(\theta)\), for example \(f(\theta) \equiv \pi(y, \theta)\).

To sample from this distribution, we can set up a Markov chain with transition probabilities \(T(\theta' \mid \theta)\) such that its stationary distribution is \(p\). That is,

\[
\int p(\theta) T(\theta' \mid \theta) \, d\theta = p(\theta')
\]

If we have the transition probabilities, it is usually easier to check the detailed balance equation,

\[
p(\theta) T(\theta' \mid \theta) = p(\theta') T(\theta' \mid \theta)
\]

This is stronger than stationarity. For a proof, assume that \(p(\theta), T(\theta' \mid \theta)\) satisfies the detailed balance equation. Then

\[
\int_{\Theta} p(\theta) T(\theta' \mid \theta) \, d\theta = \int_{\Theta} p(\theta') T(\theta' \mid \theta) \, d\theta
\]

\[
= p(\theta') \int_{\Theta} T(\theta \mid \theta') \, d\theta
\]

\[
= p(\theta')
\]

and so \(p(\theta)\) is a stationary distribution.

**Metropolis-Hastings**

The transition probabilities for the Metropolis-Hastings algorithm are

\[
T(\theta' \mid \theta) = \alpha(\theta', \theta) q(\theta', \theta) + (1 - \alpha(\theta', \theta)) \delta(\theta)
\]

where \(q(\theta', \theta)\) is called the “proposal distribution” and \(\alpha(\theta', \theta)\) is

\[
\alpha(\theta', \theta) \equiv \min \left\{ 1, \frac{f(\theta') q(\theta, \theta')}{f(\theta) q(\theta', \theta)} \right\}
\]

The idea is that the chain tries to go places with high density, but makes sure that it can come back. The chain satisfies the detailed balance equation by construction, as you can check.
Gibbs Sampling

Gibbs sampling is a special case of Metropolis-Hastings. If \( \theta \) is a vector, we update the elements of \( \theta \) one at a time, with

\[
\theta'_i \sim p(\theta_i | \theta_{i:j}) \quad \quad \theta'_{j \neq i} = \theta_j
\]

Then

\[
\alpha(\theta', \theta) = \min \left\{ 1, \frac{p(\theta') p(\theta_i | \theta_{i:j})}{p(\theta) p(\theta'_i | \theta_{i:j})} \right\}
\]

\[
= \min \left\{ 1, \frac{p(\theta') p(\theta)}{p(\theta) p(\theta')} \right\} = 1
\]

And so the proposal is always accepted.

Algorithm 3

Our model is

\[
\pi \sim \text{Dir}(\alpha/k, \ldots, \alpha/k)
\]

\[
z_j \sim \text{cat}(\pi)
\]

\[
\phi_k \sim g(\eta_0)
\]

\[
y_j \sim f(\phi_{z_j})
\]

where \( g(\eta_0) \) is an exponential family that is conjugate to \( f \) as described before. We will integrate out \( \phi \) and \( \pi \) and do Gibbs sampling on \( z \). Since the \( z_i \) are exchangeable, it suffices to consider just the distribution of the last one. We have

\[
P(z_J = k \mid z_{1:J-1}, y) \propto P(z_J = k \mid z_{1:J-1})P(y \mid z)
\]

\[
= P(z_J = k \mid z_{1:J-1}) \prod_{j=1}^{J-1} P(y_j \mid z_{1:j}, y_{1:j-1})P(y_J \mid y_{1:j-1}, z)
\]

\[
\propto P(z_J = k \mid z_{1:J-1})P(y_J \mid y_{1:j-1}, z)
\]

Aside about using auxiliary variables

Suppose we want to find \( p(\theta \mid y) \). We could replace this with \( p(\theta \mid y)p(a \mid \theta, y) \). Then we could use Gibbs sampling with

1. Sample from \( a \mid \theta, y \)
2. Sample from \( \theta_i \mid a, y \)

The idea is that sometimes the extra variables make sampling easier.