10.2 The Metropolis algorithm

Let's consider a very generic situation where we have a sampling model $Y \sim p(y|\theta)$ and a prior distribution $p(\theta)$. Although in most problems $p(y|\theta)$ and $p(\theta)$ can be calculated for any values of y and θ, $p(\theta|y) = p(\theta)p(y|\theta)/\int p(\theta')p(y|\theta')\ d\theta'$ is often hard to calculate due to the integral in the denominator. If we were able to sample from $p(\theta|y)$, then we could generate $\theta^{(1)}, \ldots, \theta^{(S)} \sim$ i.i.d. $p(\theta|y)$ and obtain Monte Carlo approximations to posterior quantities, such as

$$\mathrm{E}[g(\theta)|y] \approx \frac{1}{S}\sum_{s=1}^{S} g(\theta^{(s)}).$$

But what if we cannot sample directly from $p(\theta|y)$? In terms of approximating the posterior distribution, the critical thing is not that we have i.i.d. samples from $p(\theta|y)$ but rather that we are able to construct a large collection of θ-values, $\{\theta^{(1)}, \ldots, \theta^{(S)}\}$, whose empirical distribution approximates $p(\theta|y)$. Roughly speaking, for any two different values θ_a and θ_b we need

$$\frac{\#\{\theta^{(s)}\text{'s in the collection} = \theta_a\}}{\#\{\theta^{(s)}\text{'s in the collection} = \theta_b\}} \approx \frac{p(\theta_a|y)}{p(\theta_b|y)}.$$

Let's think intuitively about how we might construct such a collection. Suppose we have a working collection $\{\theta^{(1)}, \ldots, \theta^{(s)}\}$ to which we would like to add a new value $\theta^{(s+1)}$. Let's consider adding a value θ^* which is nearby $\theta^{(s)}$. Should we include θ^* in the set or not? If $p(\theta^*|y) > p(\theta^{(s)}|y)$ then we want more θ^*'s in the set than $\theta^{(s)}$'s. Since $\theta^{(s)}$ is already in the set, then it seems we should include θ^* as well. On the other hand, if $p(\theta^*|y) < p(\theta^{(s)}|y)$ then it seems we should not necessarily include θ^*. So perhaps our decision to include θ^* or not should be based on a comparison of $p(\theta^*|y)$ to $p(\theta^{(s)}|y)$. Fortunately, this comparison can be made even if we cannot compute $p(\theta|y)$:

$$r = \frac{p(\theta^*|y)}{p(\theta^{(s)}|y)} = \frac{p(y|\theta^*)p(\theta^*)}{p(y)}\frac{p(y)}{p(y|\theta^{(s)})p(\theta^{(s)})} = \frac{p(y|\theta^*)p(\theta^*)}{p(y|\theta^{(s)})p(\theta^{(s)})}. \qquad (10.1)$$

Having computed r, how should we proceed?

If $r > 1$:

Intuition: Since $\theta^{(s)}$ is already in our set, we should include θ^* as it has a higher probability than $\theta^{(s)}$.
Procedure: Accept θ^* into our set, i.e. set $\theta^{(s+1)} = \theta^*$.

If $r < 1$:

Intuition: The relative frequency of θ-values in our set equal to θ^* compared to those equal to $\theta^{(s)}$ should be $p(\theta^*|y)/p(\theta^{(s)}|y) = r$. This means that for every instance of $\theta^{(s)}$, we should have only a "fraction" of an instance of a θ^* value.
Procedure: Set $\theta^{(s+1)}$ equal to either θ^* or $\theta^{(s)}$, with probability r and $1 - r$ respectively.

This is the basic intuition behind the famous *Metropolis algorithm*. The Metropolis algorithm proceeds by sampling a proposal value θ^* nearby the current value $\theta^{(s)}$ using a *symmetric proposal distribution* $J(\theta^*|\theta^{(s)})$. Symmetric here means that $J(\theta_b|\theta_a) = J(\theta_a|\theta_b)$, i.e. the probability of proposing $\theta^* = \theta_b$ given that $\theta^{(s)} = \theta_a$ is equal to the probability of proposing $\theta^* = \theta_a$ given that $\theta^{(s)} = \theta_b$. Usually $J(\theta^*|\theta^{(s)})$ is very simple, with samples from $J(\theta^*|\theta^{(s)})$ being near $\theta^{(s)}$ with high probability. Examples include

- $J(\theta^*|\theta^{(s)}) = \text{uniform}(\theta^{(s)} - \delta, \theta^{(s)} + \delta)$;
- $J(\theta^*|\theta^{(s)}) = \text{normal}(\theta^{(s)}, \delta^2)$.

Source: Hoff PD (2009). *A First Course in Bayesian Statistical Methods.* New York: Springer.