How do we sample from the posterior?

- In general, we want samples from the joint posterior distribution
- *Independent* sampling from $p(\boldsymbol{\theta}|\boldsymbol{y})$ may be difficult
- **BUT** sampling from a *Markov chain* with $p(\boldsymbol{\theta}|\boldsymbol{y})$ as its stationary (equilibrium) distribution turns out to be easier
- A sequence of random variables $\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, ...$ forms a Markov chain if

$$P(\quad |\theta^{(t)}, \theta^{(t-1)}, \ldots) = P(\theta^{(t+1)}| \quad)$$

  - *i.e.* conditional on the value of $\theta^{(t)}$, $\theta^{(t+1)}$ is independent of $\theta^{(t-1)}, ..., \theta^{(0)}$
  - given the , the is independent of the

- There are general theorems that show that
  $\frac{1}{N}\sum_{t=1}^{N} g(\theta^{(t)}) \rightarrow$ as
  when $\theta^{(1)}, ..., \theta^{(N)}$ are sampled from a suitable Markov chain
- Unlike with ordinary Monte Carlo, in the Markov chain case there will be as well as in the empirical mean – both $\rightarrow 0$ as $N \rightarrow \infty$

To summarise, the point of MCMC:

Bayesian posterior inference may be achieved – via Monte Carlo integration – using simulated values of all the unknown quantities in the model, generated from a Markov chain with $p(\boldsymbol{\theta}|\boldsymbol{y})$ as its stationary distribution

5. MCMC recipes: How do we design a Markov chain with $p(\boldsymbol{\theta}|\boldsymbol{y})$ as its unique stationary distribution?

- Note that we are the usual Applied Probability problem (given a Markov chain transition matrix, find the stationary distribution): here we are the stationary distribution and want to a Markov chain that it corresponds to.
- It is surprisingly easy and several standard 'recipes' are available.
- These methods were discovered during World War II as part of the work leading to the atomic bomb: the original published method was by Metropolis *et al.* (1953), a team that worked at Los Alamos.
- The method was generalized by (1970), and an important special case called was introduced in 1984.
- There has since been an explosion of use of these methods – they make Bayesian methods practical on a (not miniature) scale.

Terminology

- While we are discussing the general theory, we will let the vector $\boldsymbol{\theta}$ denote all the and $\boldsymbol{y}$ all the . We will sometimes abbreviate
- $\pi(\boldsymbol{\theta})$ is the of our Markov chain – that is, the chain is designed to have $\pi(\boldsymbol{\theta})$ as its (or equilibrium or invariant) distribution
- We sometimes call a Markov chain used in this way a
- The Markov chain will have typical states $\boldsymbol{\theta}$ – usually not a countable state space (as in AP2), usually a .
- However, in proofs, we will use discrete (countable) notation