Why does the Metropolis algorithm work?

The proof that the sequence of iterations $\theta^1, \theta^2, \ldots$ converges to the target distribution has two steps: first, it is shown that the simulated sequence is a Markov chain with a unique stationary distribution, and second, it is shown that the stationary distribution equals this target distribution. The first step of the proof holds if the Markov chain is irreducible, aperiodic, and not transient. Except for trivial exceptions, the latter two conditions hold for a random walk on any proper distribution, and irreducibility holds as long as the random walk has a positive probability of eventually reaching any state from any other state; that is, the jumping distributions J_t must eventually be able to jump to all states with positive probability.

To see that the target distribution is the stationary distribution of the Markov chain generated by the Metropolis algorithm, consider starting the algorithm at time $t-1$ with a draw θ^{t-1} from the target distribution $p(\theta|y)$. Now consider any two such points θ_a and θ_b, drawn from $p(\theta|y)$ and labeled so that $p(\theta_b|y) \geq p(\theta_a|y)$. The unconditional probability density of a transition from θ_a to θ_b is

$$p(\theta^{t-1}=\theta_a, \theta^t=\theta_b) = p(\theta_a|y)J_t(\theta_b|\theta_a),$$

where the acceptance probability is 1 because of our labeling of a and b, and the unconditional probability density of a transition from θ_b to θ_a is, from (11.1),

$$\begin{aligned} p(\theta^t=\theta_a, \theta^{t-1}=\theta_b) &= p(\theta_b|y)J_t(\theta_a|\theta_b)\left(\frac{p(\theta_a|y)}{p(\theta_b|y)}\right) \\ &= p(\theta_a|y)J_t(\theta_a|\theta_b), \end{aligned}$$

which is the same as the probability of a transition from θ_a to θ_b, since we have required that $J_t(\cdot|\cdot)$ be symmetric. Since their joint distribution is symmetric, θ^t and θ^{t-1} have the same marginal distributions, and so $p(\theta|y)$ is the stationary distribution of the Markov chain of θ. For more detailed theoretical concerns, see the references at the end of this chapter.

Source: Gelman A, Carlin JB, Stern HS, and Rubin DB (2003). Bayesian Data Aanlysis, 2nd ed. London: CRC Press.