Burn-In is Unnecessary

The purpose of this web page is to explain why the practice called ``burn-in'' is not a necessary part of Markov chain Monte Carlo (MCMC).

> *Burn-in is only one method, and not a particularly good method, of finding a good starting point.*

People find this surprising, because many people, including your humble author, have woofed about ``burn-in'' in their MCMC papers. If everyone takes it for granted, what can be wrong with it?

To see clearly why ``burn-in'' is unnecessary (harmless but generally pointless) we need to work through a number of issues.

What is Burn-In?

``Burn-in'' is a colloquial term that describes the practice of throwing away some iterations at the beginning of an MCMC run. The One Long Run web page explains why we can limit the discussion to just one run. The burn-in notion says you start somewhere, say at x, then you run the Markov chain for n steps, from which you throw away all the data (no output). This is the *burn-in* period. After the burn-in you run normally, using each iterate in your MCMC calculations.

The name ``burn-in'' comes from electronics (see the entry in the Jargon File). Many electronics components fail quickly. Those that don't are a more reliable subset. So a burn-in is done at the factory to eliminate the worst ones.

Markov chains don't work the same way. Markov chain ``failure'' (nonconvergence) is different from electronic component failure. Running longer may cure the first, but a dead transistor is dead forever. Thus ``burn-in'' is a bad term in MCMC, but there's more wrong than just the word, there's something fishy about the whole concept.

Asymptotics Contra Burn-In.

There is nothing in MCMC theory, properly understood, that justifies or even motivates burn-in. In MCMC we use the sample average over a run of the Markov chain to approximate an expectation with respect to the equilibrium distribution of the chain. The strong law of large numbers (SLLN) guarantees that the average converges to the expectation with probability one. The central limit theorem (CLT) guarantees that the error will obey the square root law. We need Markov chain versions of the SLLN and CLT, but that's the only difference from ordinary independent-sample Monte Carlo. The theoretical justification is the same for both.

Nowhere in the theory is anything said about burn-in. The SLLN and CLT hold regardless of the distribution of the starting position. More precisely, if they hold for any one initial distribution (the equilibrium distribution, for example) then they hold for every initial distribution. (The technical condition that

guarantees this behavior is called *Harris recurrence*, see Proposition 17.1.6 in Meyn and Tweedie, 1993.)

In fact, from the theoretical point of view, burn-in is just one way of selecting a starting distribution. Suppose the starting point x has the probability distribution m and the burn-in period is n. Then the real starting distribution (the distribution of the first iterate used in calculations) is denoted mP^n, where P is the Markov chain transition probability matrix or kernel.

You don't have to understand this notation to understand the point, which is that burn-in gives a recipe for a probability distribution: start with a sample from m and then run n steps. The official notation for that distribution is mP^n, but that doesn't help us calculate it or say much about it.

So from a theoretical point of view, burn-in is the cultural practice of the MCMC community of only using initial distributions of the form mP^n and no others. This practice has no theoretical motivation. Where did it come from?

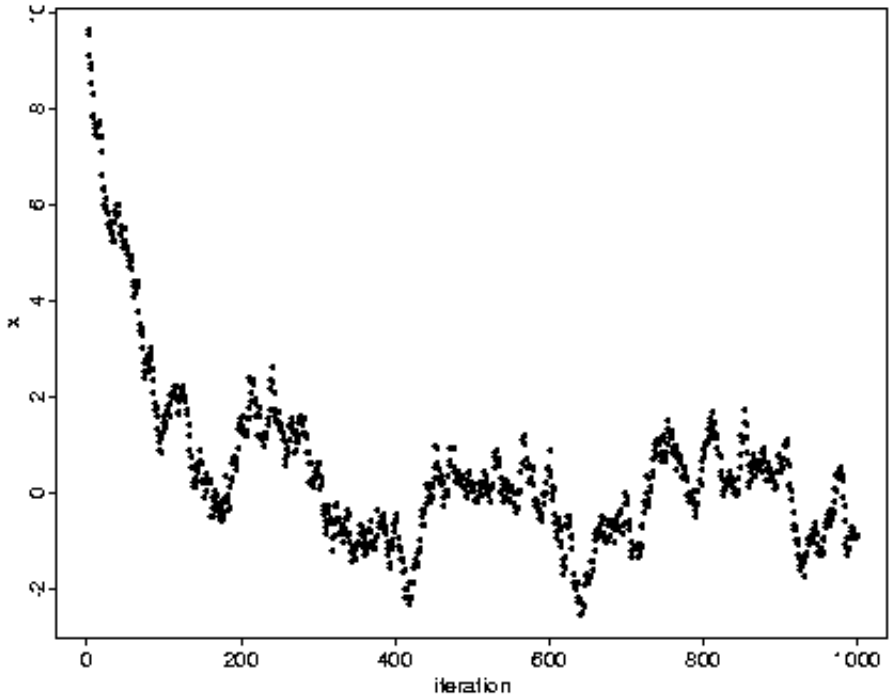
The Problem Burn-In Tries to Solve.

Anyone who has ever done any Markov chain simulation has noticed that some starting points are better than others. Even the simplest and best behaved Markov chains exhibit this phenomenon.

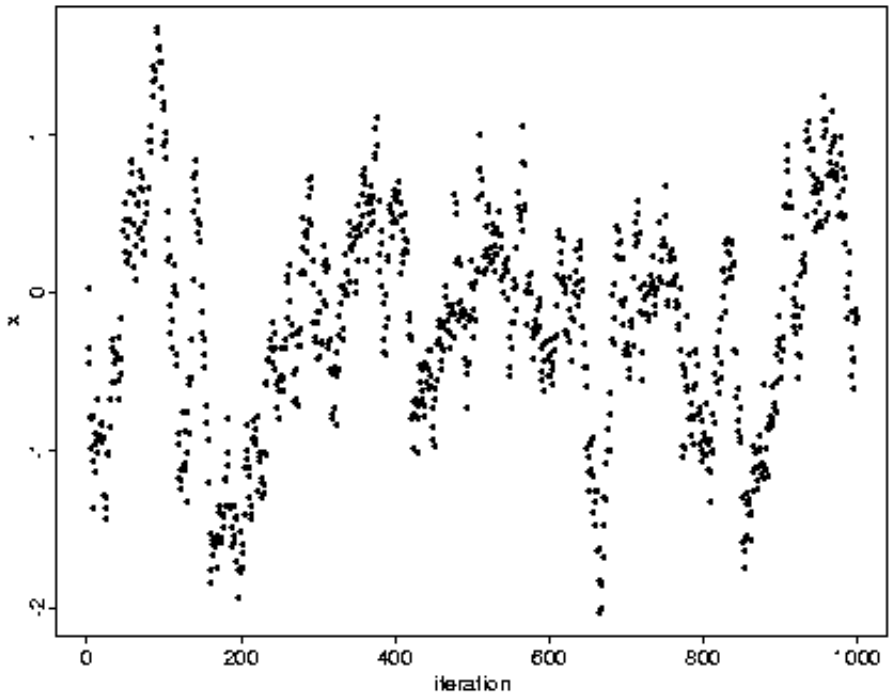
Consider an AR(1) time series, having an update defined by

$$X_{n+1} = r X_n + e_n$$

where the e_n are independent, identically distributed mean-zero normal random variables. The pictures below show an AR(1) sampler with $r = .98$. First a run started at $x = 10$.



Then a run started at $x = 0$.



As any fool can plainly see, the second run is a lot better than the first. One way to say what is wrong with the first is that there is an ``initial transient'' that is unrepresentative of the equilibrium distribution. The obvious cure is to toss the initial 200 iterations, or in other words to use a burn-in period of $n = 200$.

But strictly speaking, the description of the problem using the ``initial transient'' notion is mathematical nonsense. The point $x = 10$ is a possible state which the sampler must eventually visit if run long enough. How long? Under the equilibrium distribution $P(X > 10)$ is less than 10^{-23}. So the problem is not that we shouldn't see $x = 10$ at all, the problem is that we shouldn't see it in a run this short. So what we conclude from examples like this is that we do not want to start way out in the tail of the equilibrium distribution.

But that is not the same as saying that burn-in is a useful or even interesting idea. The second run, started at zero, does not have the same problem. There is no obvious need for burn-in.

Trying to Start in Equilibrium.

Most justifications for burn-in assume a need to start with the chain ``in equilibrium.'' Of course, burn-in doesn't actually do that, but it may come close, and even if it doesn't it is used in hope that it will.

It is no coincidence that most papers that woof about burn-in also appeal to an LLN and CLT from the stationary process literature and were written before the book of Meyn and Tweedie (1993) appeared, or, if written after, were written by authors who had not caught up with the fact that Markov chains do not have to be stationary to have an LLN and CLT.

It is a good thing that stationarity is not necessary for the LLN and CLT because burn-in doesn't actually achieve stationarity. Authors who give this justification for burn-in are really appealing to a non-theorem, the central limit almost-but-not-quite theorem for almost-but-not-quite stationary processes.

A different argument for burn-in uses unbiasedness. If the chain were stationary, then every Monte Carlo estimator would be unbiased because the expectation of a sum is the sum of the expectations regardless of whether the terms of the sum are independent. Burn-in leads to almost-but-not-quite unbiasedness (or so the user hopes). Since unbiasedness is generally held to be a Good Thing, burn-in is necessary.

This unbiasedness argument is rubbish. If you start at x and I start at x then your MCMC run is no better than mine. If you used burn-in and I didn't, then you are entitled to woof about approximate unbiasedness and I am not. But that woof does not make your estimator any better.

It is especially bizarre when a Bayesian makes the unbiasedness argument. Bayesians always insist that inference must condition on the data, that it should not involve possible data that were never observed. In MCMC, this means that we should condition on the actual starting point and not woof about possible starting points that were not used. Of course, one doesn't have to be a Bayesian to see that when one conditions on the starting point actually used the notion of unbiasedness becomes completely irrelevant.

Good Alternatives to Burn-In.

So what should we do instead of burn-in? One rule that is unarguable is

> *Any point you don't mind having in a sample is a good starting point*.

In a typical application, one has no mathematical analysis of the Markov chain that tells where the good starting points are (nor how much burn-in is required to get to a good starting point). All decisions about starting points are based on the output of some preliminary runs that appear to have ``converged'' to stationarity. Any point of the parts of these preliminary runs one believes (hopes?) to be representative of the equilibrium distribution is as good a starting point as any other.

So one rule I often follow is to start the next run where the last run ended. This is the rule most authorities recommend for random number generator seeds. Attempts by the user to start with ``random'' seeds can destroy the randomness properties of a generator. Only when the generator is used as one continuous stream does it have whatever desirable properties are claimed for it. One could do worse than to apply the same principle to MCMC.

Another possible rule is to start at a point, like the mode, known to have reasonably high probability. If no such point is known, this rule is useless, but it could be used more often than it actually is.

I make no claim that either of these rules is the one and only right way to do MCMC. I only claim that burn-in is no better and often worse.

Bibliography

Meyn, S. P. and R. L. Tweedie (1993).
Markov Chains and Stochastic Stability.
London: Springer-Verlag.

Questions or comments to: Charles Geyer charlie@stat.umn.edu

Back to Charlie Geyer's home page.