The need for fast sampling

If you are trying to estimate the best values and uncertainties of a many-parameter model, or if you are trying to compare two models with multiple parameters each, fair and thorough sampling of the parameter space becomes challenging. Imagine, for example, that for a given model you have time to calculate the likelihood of the data for a million different parameter combinations. If you have only one parameter, that means that you can look at a million values of that parameter. If you have two parameters, you could do a grid with a thousand values of each of those parameters. If you have six parameters, you could only afford ten values per parameter; if you have twelve, then you could only sample about three values per parameter. At some point, it's clear that a grid sampling such as this cannot probe the posterior probability thoroughly enough to be useful.

It is, of course, possible to use a maximization routine to find peaks in the posterior. But even if the search for the maximum is successful (and it very well might not be!), it isn't enough: almost always, you would like a sense of the *uncertainties* in your model parameters, not just their best values.

We know that if we somehow obtained the full posterior probability density, we'd be set. It might take a while, but we'd be able to perform any statistical calculation we desired. This isn't practical, though, because if we have many parameters then we suffer from the "curse of dimensionality" outlined above.

What we *can* do is to compute the posterior at any given point in the parameter space. This won't be properly normalized, but the ratio of the posterior density we compute at one point in parameter space, to the posterior density at another point, will be the right ratio. This might not seem terribly helpful until you realize that this means we can calculate posterior densities at various points in the parameter space and "walk upwards", i.e., we can go to higher and higher posterior densities. We always keep in mind that the posterior surface might be especially nasty, with very narrow spikes; if this happens we're doomed, but we can at least imagine various methods to climb to high points in the posterior.

If, however, we wish to end up with a fair sampling of the posterior (if not a *complete* sampling of the posterior), we need to think about another criterion. In particular, we'd like to think of some way to jump back and forth between points in the parameter space in such a fashion that we get to an equilibrium in which over time our points sample the posterior.

As a guide to how we might accomplish this, we can imagine that we've already accomplished that feat. We have some set of points that in some way sample the posterior fairly. If we begin with those initial points, then for our sampling to be representative it must be that if we have a candidate move from a given point, with some probability of accepting the move and some probability of rejecting the move (and thus staying at the original point),

it must be that after all candidate moves are evaluated the new distribution of points must also represent the posterior fairly. This leads us (not accidentally!) to a highly important principle of physics and chemistry and not just statistics. This is the principle of *detailed balance*, which we will now discuss.

Detailed balance

We'll begin with a simple example. Say that you have a million coins lying around, each showing either heads or tails. You flip all of them at once. An individual coin could stay the way it was (at H or at T), or change (from T to H or H to T). Assuming that the coins and flips are fair, on a coin by coin basis there is nothing to distinguish one state from another. But for the ensemble of a million coins, it's another story. If we began with all million coins reading heads, then the state of the system as a whole will change greatly when we go through a round of flips. If instead we began with about half the coins reading heads, then we'd expect to end up after flipping with about half the coins reading heads. The identity of the H coins will have changed, but the ensemble will be more or less as it was.

Now imagine that the coins have been loaded so that in a given flip, there is a 90% chance that the coin will read H. Unsurprisingly, your equilibrium state will have about 900,000 Hs and 100,000 Ts. One way you can justify this conclusion is by noting that when you flip the 900,000 Hs, about $0.1 \times 900,000 = 90,000$ of them will become Ts, and when you flip the 100,000 Ts, about $0.9 \times 100,000 = 90,000$ of them will become Hs. Thus, on average, once your system has reached the 90/10 split, in the next step the system will maintain that split because equal numbers of coins will go from H to T, as from T to H, and thus the relative proportion is steady. In fact, the system is in detailed balance: all processes and their inverses are in balance with each other.

To consider just one application of this in the real world, think about a glass of liquid water in air, with both the water and air in a sealed container. At any given moment, some fraction of the water molecules will evaporate from the liquid to go into the air. If the air initially had no water molecules at all, then this evaporation would just represent a net loss to the water. But as the evaporation proceeds, the air starts to have some concentration of water molecules. Those water molecules have some probability of rejoining the liquid water. Detailed balance would be achieved when the rate at which water molecules evaporate from the liquid is balanced by the rate at which water vapor rejoins the liquid.

That might be interesting (and the applications are endless!) but what does that have to do with our statistics problem?

To understand the relation, it would be useful to think a bit more deeply about what it would mean to have a representative sampling of a full posterior. It would mean that

the probability of having a sample within a fixed small volume of a point in the parameter space is proportional to the posterior probability density there. For example, if the posterior density at some point $\vec{A}$ in parameter space is twice as large as the posterior density at some other point $\vec{B}$ in parameter space, then if we do enough sampling we would expect there to be twice as many samples in some small parameter volume around $\vec{A}$ as in the same small parameter volume around $\vec{B}$. If we have additional points $\vec{C}$, $\vec{D}$, and so on, the same proportionalities would apply.

Thus if we have some procedure in which we walk around parameter space, we want that procedure to maintain this proportionality.

One such procedure was introduced by Metropolis and Hastings. The idea is that we break a possible move into two steps: (1) a proposed move, and (2) a decision about whether to accept that move.

In the Metropolis-Hastings algorithm, proposals are symmetric: it is equally probable to propose a move from $\vec{A}$ to $\vec{B}$ as it is to propose a move from $\vec{B}$ to $\vec{A}$, for any points $\vec{A}$ and $\vec{B}$. But the probability of *acceptance* depends on the posterior probability densities $\text{Post}(\vec{A})$ and $\text{Post}(\vec{B})$. Considering a proposed move from $\vec{A}$ to $\vec{B}$:

1. If $\text{Post}(\vec{B}) \geq \text{Post}(\vec{A})$, we accept the proposed move with 100% probability.
2. Otherwise, accept the proposed move with a probability $\text{Post}(\vec{B})/\text{Post}(\vec{A})$.

To see how this would work in practice, take our previously-considered case in which the posterior density at $\vec{A}$ is twice that at $\vec{B}$. Say that, in fact, there are currently two samples in the near vicinity of $\vec{A}$ and one sample in the near vicinity of $\vec{B}$. A proposed move from $\vec{B}$ to $\vec{A}$ is guaranteed to be accepted, so that would move one sample from $\vec{B}$ to $\vec{A}$. A proposed move from $\vec{A}$ to $\vec{B}$ would be accepted with 50% probability, but because there are two samples there initially, that will also on average move one sample from $\vec{A}$ to $\vec{B}$. Thus the balance is maintained. You can convince yourself that this argument generalizes to any number of locations in parameter space. You can also convince yourself that as long as the *ratio* of accepted moves obeys detailed balance, the distribution will be maintained; for example, in the first case above we could accept the move with 1% probability, and in the second case could accept the move with a probability equal to 1% times the ratio of the posteriors. It would just take longer that way.

Before we close this lecture, we can think about how this would work in practice. We could imagine starting off with some number of points randomly scattered through parameter space. If the data we have analyzed are informative, it is likely that only a small fraction of the parameter space has a large posterior density. Thus most of those initial points will have terribly low posterior densities. As a result, a proposed move has a fine chance of being

accepted. But as the points move toward higher posterior densities, if proposed moves are drawn randomly from the full parameter space, the probability of acceptance will become lower and lower. Thus although the parameter space will indeed be sampled representatively, it will take longer and longer.

Therefore we need a more efficient way to sample. This will be the topic of the next lecture, on Markov chain Monte Carlo methods.