Gibbs Sampling

Carl Edward Rasmussen

October 28th, 2016

Key concepts

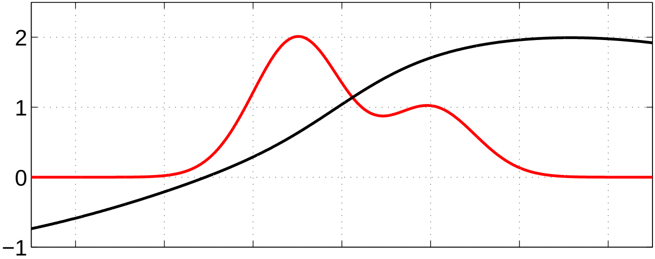
- *inference* requires integrating out variables
- Why may random sampling be useful for integration?
- What happens if the joint distribution is too complicated to sample from?
- Gibbs sampling and conditional distributions

How do we do integrals wrt an intractable posterior?

Approximate expectations of a function $\phi(\mathbf{x})$ wrt probability $p(\mathbf{x})$:

$$\mathbb{E}_{p(x)}[\phi(x)] \; = \; \bar{\phi} \; = \; \int \phi(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}, \quad \text{where} \quad \mathbf{x} \in \mathbb{R}^D,$$

when these are not analytically tractable, and typically $D \gg 1$.



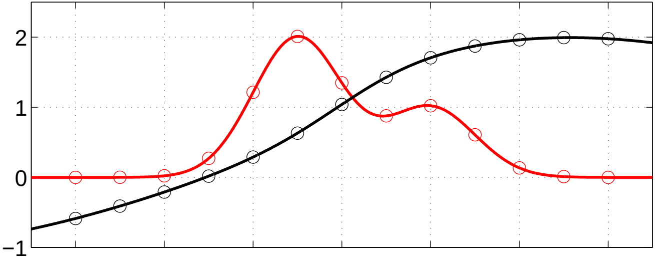
Assume that we can evaluate $\phi(x)$ and $p(x)$.

Numerical integration on a grid

Approximate the integral by a sum of products

$$\int \phi(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \simeq \sum_{\tau=1}^{T} \phi(\mathbf{x}^{(\tau)}) p(\mathbf{x}^{(\tau)}) \Delta\mathbf{x},$$

where the $\mathbf{x}^{(\tau)}$ lie on an equidistant grid (or fancier versions of this).

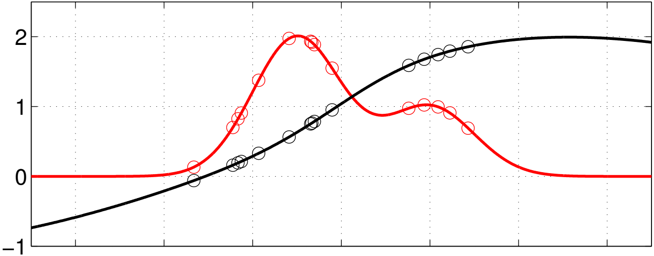


Problem: the number of grid points required, k^D, grows exponentially with the dimension D. Practicable only to $D = 4$ or so.

Monte Carlo

The fundamental basis for Monte Carlo approximations is

$$\mathbb{E}_{p(x)}[\phi(\mathbf{x})] \simeq \hat{\phi} = \frac{1}{T}\sum_{\tau=1}^{T}\phi(\mathbf{x}^{(\tau)}), \quad \text{where} \quad \mathbf{x}^{(\tau)} \sim p(\mathbf{x}).$$



Under mild conditions, $\hat{\phi} \to \mathbb{E}[\phi(\mathbf{x})]$ as $T \to \infty$. For moderate T, $\hat{\phi}$ may still be a good approximation. In fact it is an *unbiased* estimate with

variance of the true phi, not its mean (estimate).

variance of the mean estimate

$$\mathbb{V}[\hat{\phi}] = \frac{\mathbb{V}[\phi]}{T}, \quad \text{where} \quad \mathbb{V}[\phi] = \int \big(\phi(\mathbf{x}) - \bar{\phi}\big)^2 p(\mathbf{x})d\mathbf{x}.$$

Note, that this variance is *independent* of the dimension D of $\mathbf{x}$.

Markov Chain Monte Carlo

This is great, but how do we generate random samples from $p(\mathbf{x})$?

If $p(\mathbf{x})$ has a standard form, we may be able to generate *independent* samples.

Idea: could we design a Markov Chain, $q(\mathbf{x}'|\mathbf{x})$, which generates (dependent) samples from the desired distribution $p(\mathbf{x})$?

$$\mathbf{x} \to \mathbf{x}' \to \mathbf{x}'' \to \mathbf{x}''' \to \dots$$

One such algorithm is called *Gibbs sampling*: for each component i of $\mathbf{x}$ in turn, sample a new value from the conditional distribution of x_i given all other variables:

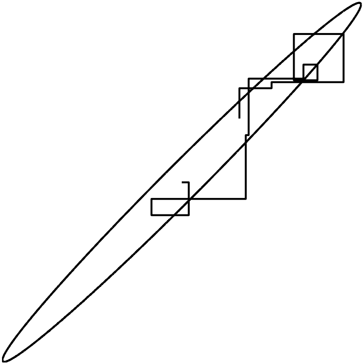
$$x_i' \sim p(x_i|x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_D).$$

It can be shown, that this will eventually generate dependent samples from the joint distribution $p(\mathbf{x})$.

Gibbs sampling reduces the task of sampling from a joint distribution, to sampling from a sequence of univariate conditional distributions.

Gibbs sampling example: Multivariate Gaussian

20 iterations of Gibbs sampling on a bivariate Gaussian; both conditional distributions are Gaussian.



Notice that strong correlations can slow down Gibbs sampling.

Gibbs Sampling

Gibbs sampling is a parameter free algorithm, applicable if we know how to sample from the conditional distributions.

Main disadvantage: depending on the target distribution, there may be very strong correlations between consecutive samples.

To get less dependence, Gibbs sampling is often run for a long time, and the samples are thinned by keeping only every 10th or 100th sample.

Burn-in: often, the initial sequence of samples is discarded, until the chain has converged to the desired distribution. What does *convergence* mean in this context?

It is often challenging to judge the *effective correlation length* of a Gibbs sampler. Sometimes several Gibbs samplers are run from different starting points, to compare results.