The Gibbs Sampler
Lecture 9, October 27, 2004
Notes by Michele Banko

Suppose we have some random variables $x_1, x_2, \ldots, x_k$ over some pdf $P(x_1, x_2, \ldots, x_k)$, and a function $f(x_1, x_2, \ldots x_k)$ for which we wish to compute the expected value, $E(f(x_1, x_2, \ldots, x_k))$.

- Solution 1, Numerical/Analytical approach

  We could try to solve

  $$E(f(x_1, x_2, \ldots, xk)) = \int_{x_1} \int_{x_2} \cdots \int_{x_k} f(x_1, x_2, \ldots, x_k) P(x_1, x_2, \ldots, x_k) dx_1 dx_2 \cdots dx_k$$

  which is difficult/impossible in high dimensions.

- Solution 2, Monte Carlo integration

  Randomly draw n independent samples from the distribution,

  $$\vec{x^{(1)}}, \vec{x^{(2)}}, \cdots \vec{x^{(n)}}$$

  and take the average:

  $$\frac{1}{n} \sum_{i=1}^{n} f(\vec{x^{(i)}})$$

  As n becomes large, we arrive at a good estimate for $E(f(x))$. However, drawing the samples may also be difficult, if the distribution is not straight-forward.

- Solution 3, MCMC

  Suppose we can draw a sample $x^{(t+1)}$ based on the previous sample at time t:

  $$x^{(t+1)} \sim p(x|x^{(t)})$$

  - This is a Markov chain - we don't need to remember the full history, but only the previous step (for a chain of order 1).
  - Now our samples are no longer independent, but this is ok since the computation of the expected value does not depend on samples being independent.
  - While other strategies exist for this type of sampling, **Gibbs Sampling** is widely used:

    Take a random walk within the sample space, simplified such that we reduce the computation of the full conditional distribution

    $$P(x_i|x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k)$$

from one that is multivariate to one that is only univariate, e.g. the computation depends only on variables at time t

- Application to Motif-Finding
  - We have k sequences, $s_1, \ldots, s_k$ and each having one motif of length w, and our motif model is the WMM.
  - Where are the motifs?
  - Let parameter x_i represent where in the sequence is the motif.
  - We build the WMM from all other sequences, and then compute the probability that motif in i^{th} sequence starts at position j given that we know where most of the motifs are located in the other sequences. Do this by usual scanning...
  - See class slides for psuedocode of full algorithm.
- Issues
  - Unlike EM, the Gibbs approach is not guaranteed to improve the model after every iteration.
  - Burnin - for how long should we run the chain to be sure it has "forgotten" it's arbitrary place? Discard these samples.
  - Mixing - how many samples to ensure we cover the space well?