Gibbs Sampler Notes
Lecture 9, October 28, 2005
Steven Balensiefer

How to Average Suppose we have random variables $x_1, x_2, \ldots, x_n$ with a p.d.f $P(x_1, x_2, \ldots, x_k)$ and a particular function $f(x_1, \ldots, x_k)$. Now, we want the expected value of $f(x)$ for the given p.d.f., that is:

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

Analytical solutions are usually difficult or impossibly, and numerical solutions are often unstable, particularly in high-dimensional problems. Therefore, a Monte Carlo integration approach is sometimes preferable.

To perform Monte Carlo integration, randomly draw n independent samples

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

from the distribution and take the average:

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

As n becomes large we get a good estimate of $E(f(x))$. Unfortunately, drawing independent samples isn't always possible, leading to the MCMC approach.

Markov Chain Monte Carlo (MCMC) Independent sampling is hard in many cases, but *not required* for expectation calculations. Instead, we can pick new samples based on the current sample.

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

- This is a 1^{st} order Markov Chain - we compute the next step based only on the previous step.
- Samples are not independent but we don't require independence for our expectation computation.
- Gibbs Sampling approach to MCMC:
  Start at a random point in the sample space. Then take a random walk through the space, in a way that reduces the computation from k-dimensions to a single dimension:

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

  Basically, choose a single dimension at each step, and pick a new value for that dimension by holding all other variables static and taking a random sample from the marginal distribution of the current value given all the values of the other components. On average, more sample will be taken around points of high probability.

Using Gibbs for Motif-Finding

- Given k sequences $s_1, \ldots, s_k$, look for a motif of length w using a WMM model. Assume there is exactly one motif instance in each sequence, starting at an umknown position.
- Variables $x_1, \ldots, x_k$ represent start locations of the motif in each sequence
- For sequence i we construct WMM using $s_1, \ldots, s_{i-1}, s_{i+1}, s_k$ and compute probability that motif in s_i starts at position j using the ordinary scanning method. We update x_i with a position randomly selected according to that probability.

Issues with the Gibbs Sampler

- Convergence Time: Unlike EM, there's no guarantee of improvement with each step. Thus no easy bound on computation can be derived.
- Burn-in: Theoretically, after a number of iterations the initial point will be "forgotten" and samples will truly represent the P.D.F. Samples until that point should be discarded, but it's hard to know where that threshold should be.
- Mixing: How many samples are needed to cover the space well? How fast do we move between (potentially many, widely separated) regions of high probability?

Variants and Extensions

- Phase Shift:
  In some cases sampler gets slightly off of the motif. So check every 100 iterations or so whether shifting over gives better results.
- Pattern Width:
  Standard approach assumes correct motif length. In this case, periodically adjust number of columns by adding or removing columns based on relative entropy measurements.
- Multiple Patterns:
  Attempt to detect 2 or more motifs in sequence of strings

Graph 1 (See Slides) This graph is from the paper by Lawrence et al, and they use it to demonstrate the success of the method, compared to random noise.

1. Authors claim this is a typical run. A fast ramp up, followed by slight variations
2. This is an alternative where an initial plateau is reached, and the final result is less than either of the other two, suggesting an "off-by-one" error
3. This curve has an even longer plateau, but still reaches the same level as Test #1
4. Dotted Line: Represents the control, consisting of generated random sequences that shared the same length and relative frequencies as the actual sequences used.

The basic claim is that the Gibbs sampling technique works well, as shown by the difference between 1-3 and the random sequences used as control.