Overview

Everyone uses Gibbs sampling!
- De facto Markov Chain Monte Carlo method for inference.
- Used for inference and learning in graphical models.
- Works very well in practice.
- Used by many systems such as Factorie, OpenBugs, PGibbs, and DeepDive — including competition-winners.

It’s important for Gibbs sampling to run fast!
- Modern hardware (CPU, GPU, FPGA) is parallel, with many computations running at the same time.
- Gibbs sampling is inherently sequential — all the updates must happen one at a time.

Parallelization without regret: the HOGWILD! strategy
- Big Idea: run multiple threads in parallel without locks.
- We call this asynchronous or HOGWILD! execution.
- The idea comes from stochastic gradient descent (Niu 2011).
- Very successful for SGD — but does it work for Gibbs?

Our contribution: analyzing asynchronous Gibbs
- We identify and study two main challenges.
- First challenge: bias — how far are the samples produced by the chain from the target distribution?
- Second challenge: mixing time — how long do we need to run the chain before we are independent of initial conditions?

What is Gibbs Sampling?

Goal: produce samples from some distribution π
- Typically, it’s too hard to compute π directly.
- But, it’s easy to compute conditional distributions.

Gibbs sampling: Sample from distribution π over variables V

Require: Initial state X, for i ∈ V, number of samples T.
for t = 0 to T − 1 do
    Select i uniformly from V.
    Resample Xi, conditionally from π given Xj (j ≠ i).
    Output sample X for i ∈ V.
end for

The First Challenge: Bias

Known result: sequential Gibbs sampling always approaches the target distribution over time.

Asynchronous Gibbs sampling may have a persistent bias.
- See the example to the right, with two binary variables.
- We plot the results of 2-thread asynchronous Gibbs sampling on this model — 98% of the mass is measured erroneously.
We can get the wrong answer!
- Bad for inference.

Our contribution: bounding bias.
For a model which satisfies Dobrushin’s condition (α < 1), the bias of estimating the marginal distribution of ω variables is

\[ \text{bias}_ω ≤ \frac{1 - \alpha}{1 + \alpha} \]

Even if α > 1, as long as α = O(1) and only O(n) steps are required to get good marginal estimates.

The Second Challenge: Mixing Times

The mixing time t_mix is the number of steps required to be close to independent of initial conditions.
- We need t_mix small for Gibbs sampling to be tractable.

Running HOGWILD! can affect the mixing time!
- See the example to the right.
- Even models with t_mix = O(n) for sequential Gibbs could have t_mix = 2^Ω(n) for asynchronous Gibbs!

Our contribution: bounding t_mix.
For a model which satisfies Dobrushin’s condition (α < 1), we compare the mixing times of HOGWILD! and sequential Gibbs with

\[ \frac{t_{\text{mix}(\text{HOGWILD})}}{t_{\text{mix}(\text{sequential})}} ≤ 1 + \frac{\alpha}{n} \]

Mixing times are about the same!
- HOGWILD! runs much faster.

Experiments

The first two plots show that the experimentally observed mixing times of HOGWILD! Gibbs sampling on two different Ising model graphs match our theoretical predictions. The third plot shows wall-clock performance of asynchronous Gibbs on a real KRP dataset, and compares it to another method, “multi-model” Gibbs, which has similar runtime but produces worse samples.

Dobrushin’s Condition

Total influence α of a model.
- Parameter used to analyze Gibbs sampling.
- Measures the degree to which one variable can depend on the other variables in the model.

Definition:
Let be a probability distribution over some set of variables I. Let Bi be the set of state pairs (x, y) which differ only at variable j. Let denote the conditional distribution in of variable i given all the other variables in state X. Then, the total influence of i, is

\[ \alpha = \max_{i \in \mathcal{I}} \max_{j \in \mathcal{I} \setminus \{i\}} \left| \pi_i(x_i | X_{-i}^j) - \pi_i(x_i | X_{-i}) \right| \]

Model satisfies Dobrushin’s condition if α < 1.
- Condition used in a line of work that ensures the rapid mixing of spin statistics systems.

Definition of mixing time:
Assume we have a process with distribution μ(μ0) at time t, given that we started from distribution μ0. Then its mixing time is

\[ t_{\text{mix}}(\epsilon) = \min \left\{ t \mid \| \mu_t(\mu_0) - \mu(\mu_0) \|_\text{TV} ≤ \epsilon \right\} \]

Roughly this is the time needed for the distribution to become independent of initial conditions.
- Gibbs sampling under this condition is known to mix fast:

Sequential Gibbs sampling on a model that satisfies Dobrushin’s condition has mixing time

\[ t_{\text{mix}}(\epsilon) ≤ \frac{n}{1 - \alpha} \log \left( \frac{2}{\epsilon} \right) \]

Summary

We analyzed HOGWILD!-Gibbs sampling, a heuristic for parallelized Gibbs sampling. We addressed two major concerns, bias and mixing time.

Take away message: in all practical cases, asynchronous Gibbs sampling will work just as well as sequential Gibbs sampling. This means that it can give us parallel speedup without regret.

(This work has been submitted to ICML 2016.)