Many processors, little time: MCMC for partitions via optimal transport couplings

Tin D. Nguyen, Brian L. Trippe, Tamara Broderick

Overview

MCMC is often used for clustering and similar tasks

Wall time is a premium, and MCMC is time consuming
In short wall time, MCMC can be inaccurate

Running parallel short chains and taking the average has small variance, but the bias remains large

We can use Markov chain coupling to debias, but couplings have **not** been developed for partition models

Our contributions: we develop couplings for partition models and demonstrate benefits in time-limited, highly parallel regime

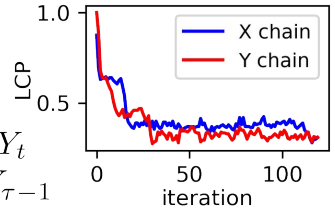
Background

MCMC is often used to characterize the distribution of a random partition Π

E.g. Clustering cells [Prabhakaran et al. 2016]
Report expected proportion of largest component: $H^* = \int h(\Pi) p_\Pi(\Pi) d\Pi$
Get estimate with MCMC: $\widehat{H} \approx H^*$

MCMC with long chains can be expensive & MCMC with short chains can be inaccurate, due to bias

Idea: Use "coupling" [Jacob et al. 2020] to debias: Create two chains $(X_t), (Y_t)$ that are equal in distribution $X_t \stackrel{d}{=} Y_t$ and eventually "meet" $X_\tau = Y_{\tau-1}$



What's missing? Existing work has not shown how to (efficiently) simulate the two chains for partition-valued chains

Our Methods

Greedily make X_{t+1} & Y_t as close to each other as possible

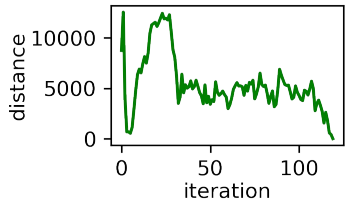
$S(X_{t+1} \mid X_t)$ and $S(Y_t \mid Y_{t-1})$ are marginal transitions

$$S(X_{t+1} = \cdot \mid X_t) = \sum_{k=1}^{K} a^k \delta_{\pi^k}(\cdot)$$
$$S(Y_t = \cdot \mid Y_{t-1}) = \sum_{k'=1}^{K'} b^{k'} \delta_{\nu^{k'}}(\cdot)$$

We quantify the distance between chains with a metric over partitions (rather than over labelings)

We pick Hamming distance, which steadily increase with the dissimilarity of two partitions

We design an optimal transport mechanism to reduce the distance



$$\min_{\gamma} \quad \sum_k \sum_{k'} \gamma(\pi^k, \nu^{k'}) \times d_{\text{Hamming}}(\pi^k, \nu^{k'})$$
$$\text{s.t.} \quad \gamma \geq 0, \sum_k \gamma(\pi^k, \nu^{k'}) = b^{k'}, \sum_{k'} \gamma(\pi^k, \nu^{k'}) = a^k$$

Theoretical Results

A useful Bayesian nonparametrics model is Gaussian Dirichlet process mixture model, and Gibbs sampling is common MCMC
K is the maximum partition size encountered

Our coupling indeed debiases the regular MCMC estimate!

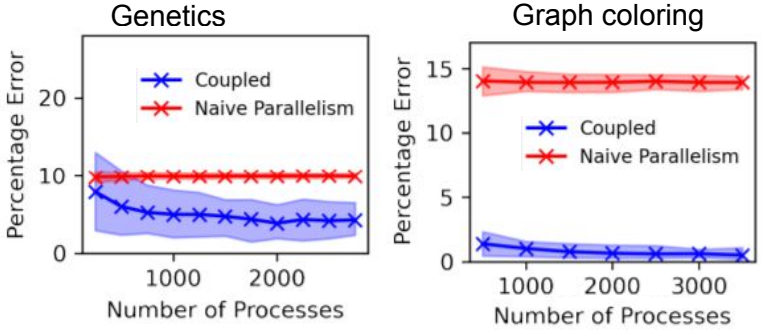
Theorem: Our coupled chains have sub-geometric meeting times and stay faithful after meeting

Computational overhead to implement the coupling is small!

Theorem: The cost of simulating from our coupling is $O(K^3 \log K)$ plus 2 times the cost of a single Gibbs move

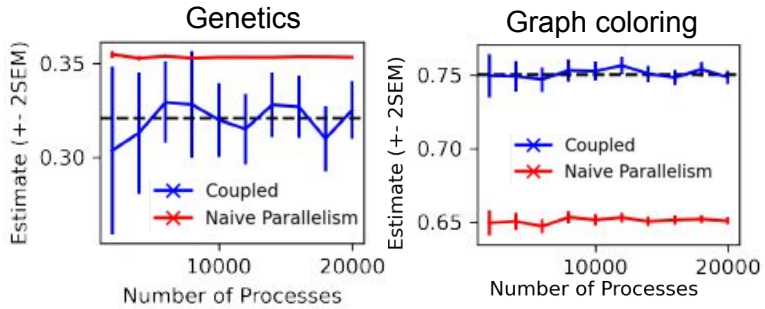
Experiments

Experiment 1: Most runs using couplings are more accurate than most runs using naive parallelism



To capture sampling variability, we run each method 50 times and report 20%, 50% and 80% quantiles
The vast majority of coupled chains runs are better than the vast majority of runs of naive parallel runs

Experiment 2: Confidence intervals from coupled chains provide nominal coverage, unlike naive parallelism



We construct confidence intervals from one run
Intervals from naive parallelism are over-confident, because of biasedness