Massachusetts Institute of Technology

Paper ID: 614

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

Many Processors, Little Time: MCMC for Partitions via Optimal Transport Couplings

AISTATS 2022

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

Markov chain Monte Carlo is often used to characterize the distribution of a random partition Π

- Example: Clustering cells into types [Prabhakaran et al. 2016]
  - Want to report expected proportion of largest component: $H^* = \int h(\Pi) p_\Pi(\Pi) d\Pi$
  - Get estimate with MCMC: $\widehat{H} \approx H^*$
- More examples: co-clustering probability of cells; graph coloring [Chen et al. 2019]

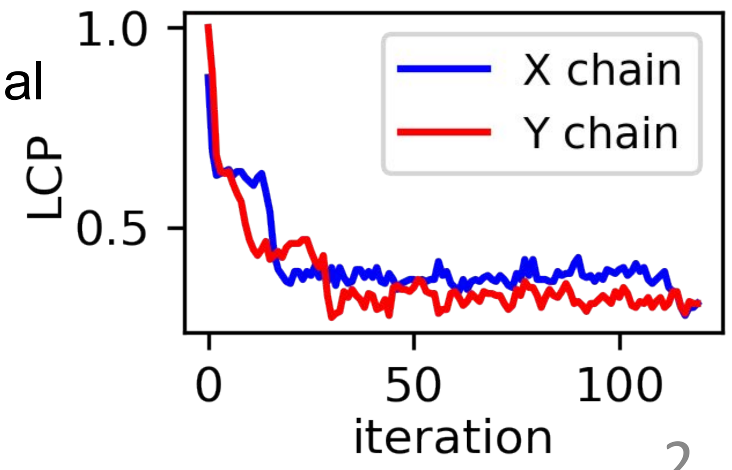
MCMC with long chains can be expensive & MCMC with short chains can be inaccurate

Idea: Run short chains in parallel and average (e.g. thousands of processors)

- **Problem:** bias does not go away from replication

Idea: Use "coupling" to debias: Create two chains $(X_t), (Y_t)$ that are equal in distribution $X_t \overset{d}{=} Y_t$ and eventually "meet" $X_\tau = Y_{\tau-1}$

- Because they have the same marginal distributions, we can roughly subtract out the bias [Jacob et al. 2020]
- **Problem:** [Jacob et al. 2020] does not address coupling for partitions



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

Two options:

1. Use an existing coupling (not previously used for debiasing). **Problem:** (we show) these are inefficient.
2. Develop a *new* coupling. **Challenge:** needs to meet (quickly) in finite time.

Our strategy: 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 need to quantify the distance between chains

Idea: Define a metric over partitions (rather than over labelings)

We pick a metric that steadily increase with dissimilarity of two partitions

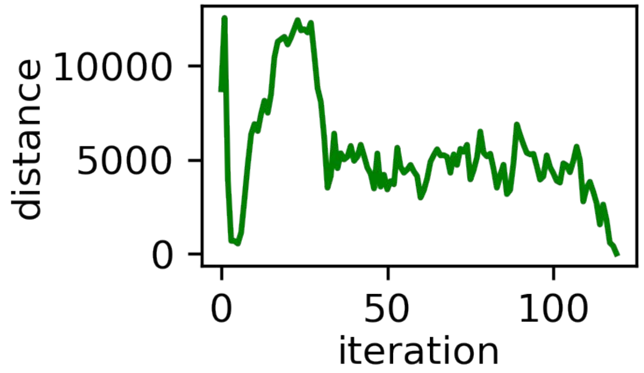
Idea: Pick Hamming distance between adjacency representations

$$d_{\text{Hamming}}(\pi^k, \nu^{k'})$$

We design an optimal transport mechanism to reduce the distance

$$\min_{\gamma} \quad \sum_{k} \sum_{k'} \gamma(\pi^k, \nu^{k'}) \times d_{\text{Hamming}}(\pi^k, \nu^{k'})$$

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

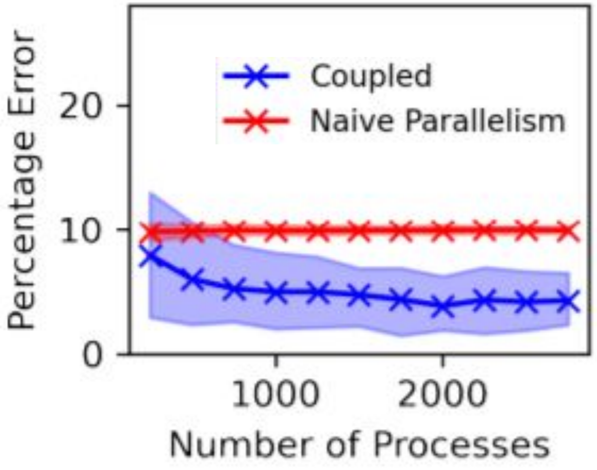


Many processors, little time: MCMC for partitions via optimal transport couplings

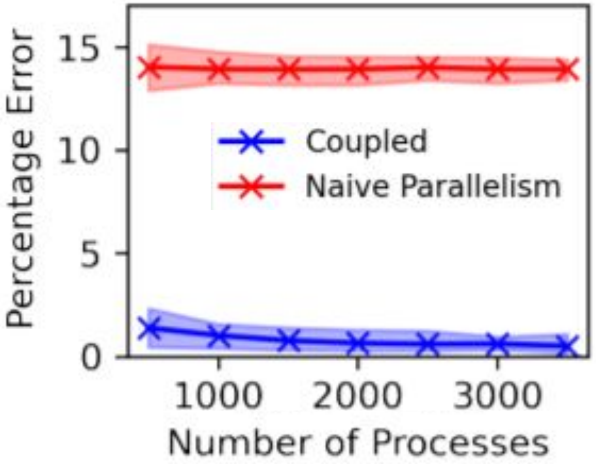
Most runs using couplings are more accurate than most runs using naive parallelism

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

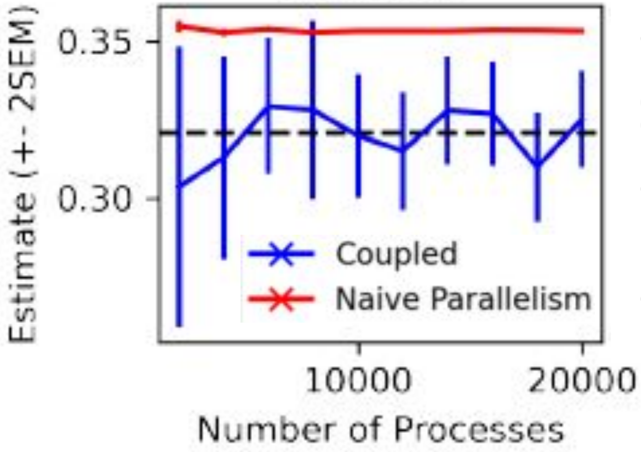
Genetics



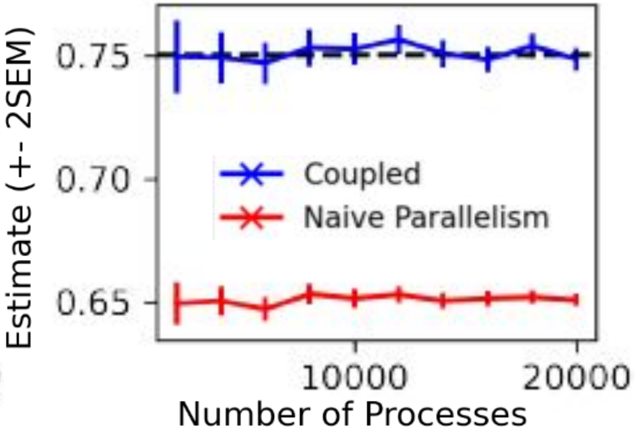
Graph coloring



Genetics



Graph coloring



Conclusion

For partition models, we produce more accurate estimates than standard MCMC in the time-limited, highly parallel regime by using optimal transport coupling.