Non-Reversible Parallel Tempering

Saifuddin Syed1[1], Alexandre Bouchard-Côté[1], George Deligiannidis[2], Arnaud Doucet[2]

THE UNIVERSITY OF BRITISH COLUMBIA

UNIVERSITY OF OXFORD

ArXiv 1905.02939

Summary

Bayesian posterior inference can be very hard when posterior distribution π is high dimensional with well separated modes. Markov Chain Monte Carlo (MCMC) methods get trapped exploring local regions of high probability leading to poor mixing. Parallel tempering (PT) uses parallel computing to improve mixing of MCMC algorithms.

Parallel Tempering

Intuition: Run N additional chains in parallel. Delegate the task of exploration to easy to sample reference distribution π_0 and communicate to chain targeting posterior π through chains targeting annealing distributions π_{β_i}

Annealing distributions: create bridge between reference π_0 and π

$$\pi_{\beta_i}(x) \propto L(x)^{\beta_i}\pi_0(x) \quad \longleftarrow \quad L(x) = \frac{\pi(x)}{\pi_0(x)}$$

$$0 = \beta_0 < \beta_1 < \cdots < \beta_N = 1 \quad \longleftarrow \quad \text{Schedule } \mathcal{P}_N$$

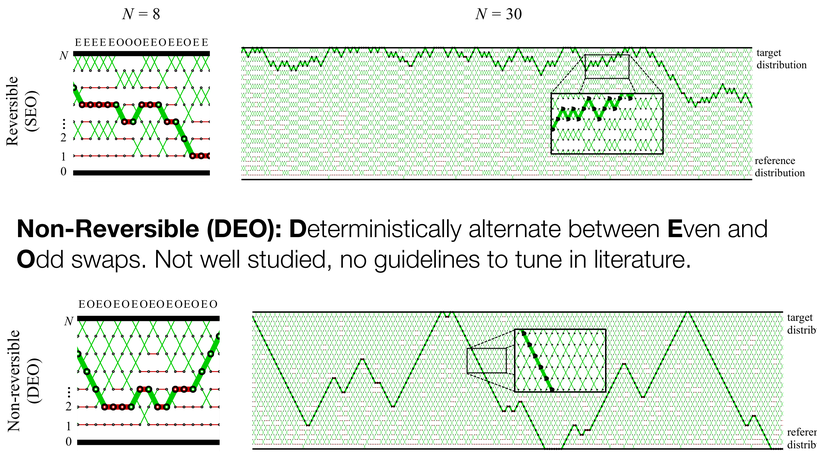
PT involves alternating between local exploration and communication

Local exploration: Update each chain according to MCMC algorithm targeting π_{β_i} (problem specific)

Communication: Propose swap between chains i and $i+1$ and accept with metropolis-hasting acceptance probability. The maximal collection of swaps in parallel when i are **even (E)** or **odd (O)**.

Reversible vs Non-Reversible PT

Reversible (SEO): Stochastically choose **E**ven and **O**dd swaps.



Non-Reversible (DEO): Deterministically alternate between **E**ven and **O**dd swaps. Not well studied, no guidelines to tune in literature.

Round Trip Rate

Objective: Want to maximize the % of samples from reference that reach target, denoted by $\tau(\mathcal{P}_N)$

Theorem 1: Non-reversible **dominates** reversible PT for any schedule and

$$\tau_{SEO}(\mathcal{P}_N) = \frac{1}{2N + E(\mathcal{P}_N)}, \qquad \tau_{DEO}(\mathcal{P}_N) = \frac{1}{2 + 2E(\mathcal{P}_N)}$$

The performance of PT depend on the **communication scheme** and **schedule**. We show DEO is optimal and efficiently tune it.

Communication Barrier

$$\lambda(\beta) = \lim_{\beta' \to \beta} \frac{r(\beta, \beta')}{|\delta|} \quad \longleftarrow \quad \text{Average rejection probability for swap between chains } \beta, \beta'$$

Optimal β_i^* satisfy a constant rejection rate, which by Theorem 2 can be found by integrating λ.

Theorem 2: $$r(\beta_{i-1}, \beta_i) = \int_{\beta_{i-1}}^{\beta_i} \lambda(\beta)d\beta + O\left(\frac{1}{N^3}\right)$$

Can estimate integral by running DEO and tracking rejection rates.

Asymptotic Performance of PT

Theorem 3: As $N \to \infty$,

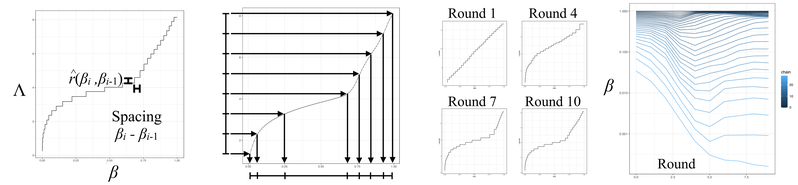
$$\tau_{SEO}(\mathcal{P}_N) \sim \frac{1}{2N}, \qquad \tau_{DEO}(\mathcal{P}_N) \to \bar{\tau} = \frac{1}{2 + 2\Lambda} \quad \Lambda = \int_0^1 \lambda(\beta)d\beta$$

Reversible PT is very **brittle** and decays in performance with parallelization.

Non-reversible PT is very **robust**, and improves in performance with N (limited by Λ). This makes is scalable to **GPUs**.

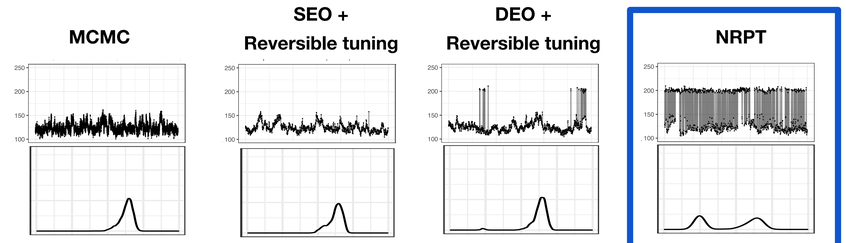
Schedule Optimization (NRPT)

1) Run Non-Reversible PT to compute Monte Carlo estimate $\hat{r}(\beta_{i-1}, \beta_i)$
2) Use Theorem 2 to estimate $\hat{\lambda}$, $\hat{\Lambda}$
3) Use numerical integration to find new schedule
4) Repeat 1-3 until convergence



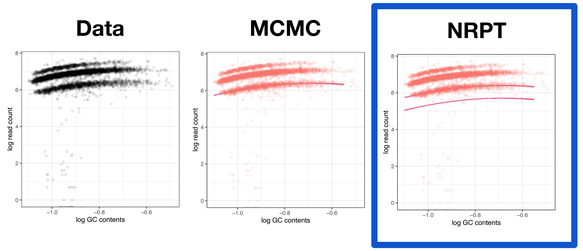
Experiments

Bayesian Mixture Model (305 dimensions)



Copy number inference (whole genome ovarian cancer data)

Used by BC Cancer Research to study evolutionary structure of cancer.



Event Horizon Telescope (EHT)

PT was at the foundation of the computation engine that processed the blackhole photo. The EHT team used NRPT to processing time from approximately **2 days** to **an hour**.

[1] University of British Columbia Deparment of Statistics, [2] University of Oxford Deparment of Statistics

E-Mail: saif.syed@stat.ubc.ca