Parallel tempering with a variational reference

Nikola Surjanovic[1], Saifuddin Syed[2], Alexandre Bouchard-Côté[1], Trevor Campbell[1]

Summary

Parallel tempering (PT) is a class of MCMC algorithms that constructs a path of distributions annealing between a reference, π_0, and intractable target, π_1. States along the path are swapped to improve mixing in the target.

Problem: Past work on PT has only used linear paths with a **fixed reference** that is often different than the target. PT swapping can be computationally expensive.

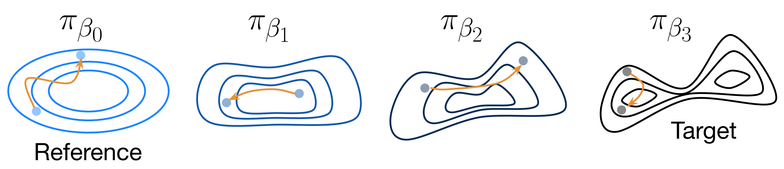
Contribution: We extend to annealing paths with a **variational reference** and optimize the choice of reference. This improves PT swapping significantly.

Parallel tempering

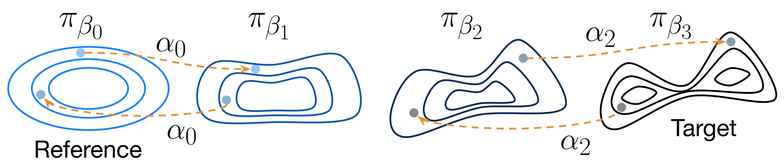
Annealing path: π_β is a path of distributions between π_0 and π_1. A linear path with a **fixed reference** is typically used: $\pi_\beta(x) \propto \pi_0^{1-\beta}(x) \cdot \pi_1^{\beta}(x)$

Run $N+1$ chains targeting π_{β_n}. Alternate between local exploration and communication moves.

Local exploration: Update each chain according to an MCMC algorithm.



Communication: Swap states between chains n and $n+1$ with probability α_n



Objective: Maximize the **restart rate** τ (rate at which samples from the reference chain travel to the target chain)

$$\tau(\pi_0, \pi_1) = \left(2 + 2\sum_{n=0}^{N-1} \frac{r_n}{1 - r_n}\right)^{-1}, \quad r_n = 1 - \mathbb{E}[\alpha_n]$$

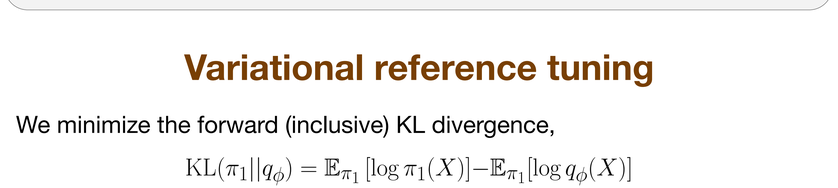
Suboptimality of a fixed reference

When the reference and target do not overlap much, PT will often reject communication swaps between chains.

Proposition (problem with standard PT): For an exchangeable* Bayesian model with m conditionally i.i.d. data points, PT with a fixed reference satisfies

$$\text{restart rate} \xrightarrow{a.s.} 0, \quad m \to \infty$$

*I.e., satisfies usual technical assumptions for a Bernstein-von Mises result, etc.



Annealing paths with a variational reference

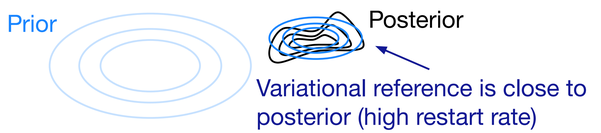
We introduce a **variational reference** family, $\{q_\phi : \phi \in \Phi\}$. The linear annealing path with the modified reference is

$$\pi_{\phi,\beta}(x) \propto q_\phi^{1-\beta}(x) \cdot \pi_1^{\beta}(x)$$

We consider exponential family reference distributions

$$q_\phi(x) = c(\phi)h(x)\exp(\phi^\top \eta(x))$$

and choose a reference distribution close to the target.



Proposition: For an exchangeable Bayesian model with m conditionally i.i.d. data points, PT with a normal variational reference that minimizes the forward KL satisfies

$$\text{restart rate} \xrightarrow{p} 1/2, \quad m \to \infty$$

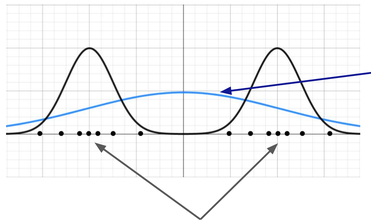
Variational reference tuning

We minimize the forward (inclusive) KL divergence,

$$\mathrm{KL}(\pi_1 || q_\phi) = \mathbb{E}_{\pi_1}[\log \pi_1(X)] - \mathbb{E}_{\pi_1}[\log q_\phi(X)]$$

and use a gradient-free procedure to tune the reference.

Variational reference tuning (...)



Algorithm:

1) Run the non-reversible PT algorithm (**NRPT**) [Syed et al. 2021]
2) Use obtained samples $X_1, X_2, \ldots, X_T$ from the target chain to update the annealing schedule using the procedure in [Syed et al. 2021]
3) Update ϕ so that $\mathbb{E}_{q_\phi}[\eta(X)] = T^{-1}\sum_{t=1}^{T}\eta(X_t)$
4) Repeat 1-3 until the computational budget is depleted

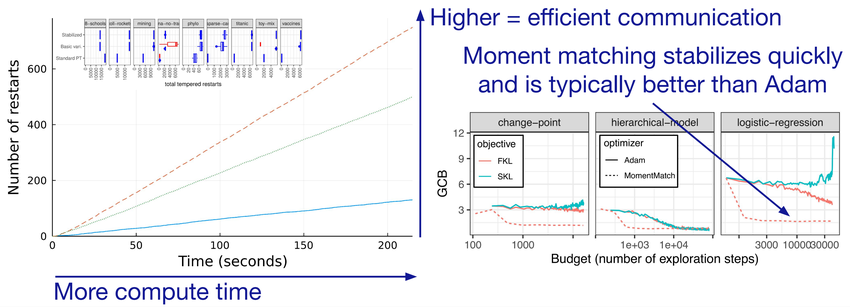
Note: To stabilize the tuning of the variational reference, we introduce PT with two references—one fixed and one variational. (Details provided in paper.)

Theorem (outline): The variational parameter estimates converge to the forward KL minimizer almost surely.

The restart rate at the forward KL minimum is bounded in terms of the flexibility of the variational family. (More details in paper.)

Experiments

PT with a variational reference empirically outperforms PT with a fixed reference. Below: variational PT with a normal reference (mean-field approximation and full covariance) versus NRPT [Syed et al. 2021].



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

Email: nikola.surjanovic@stat.ubc.ca