THE UNIVERSITY OF BRITISH COLUMBIA

Parallel tempering on optimized paths

Saifuddin Syed[1*], Vittorio Romaniello[1*], Trevor Campbell[1], Alexandre Bouchard-Côté[1]

ArXiv
2102.07720v2

Summary

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

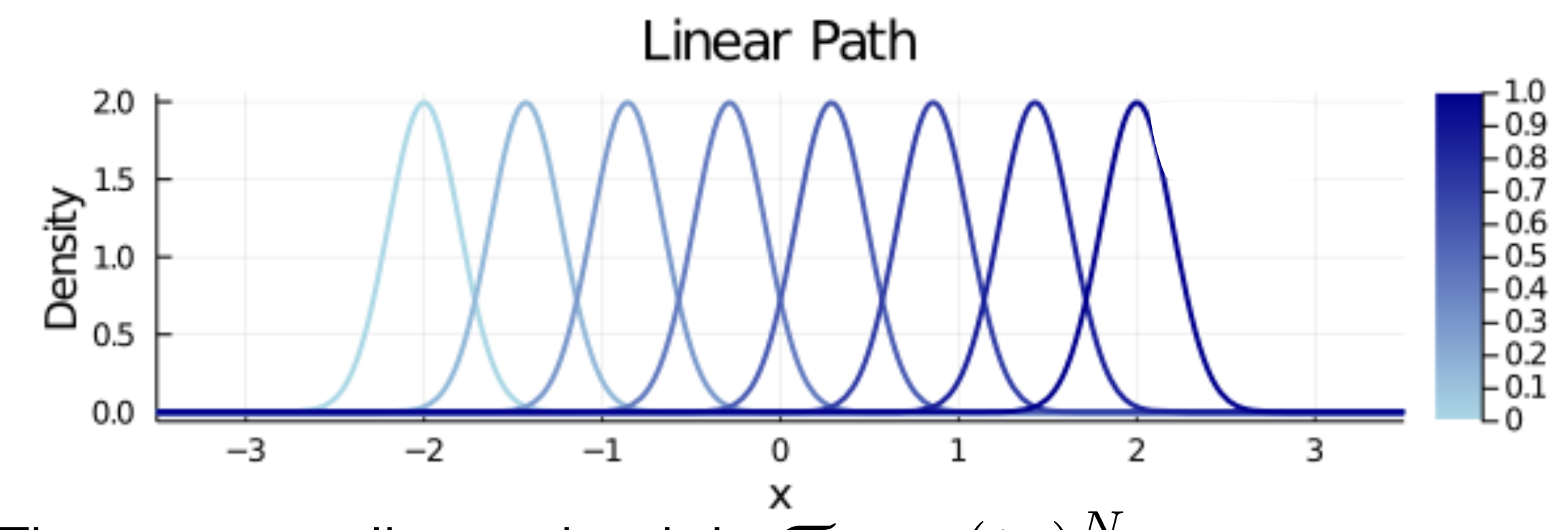
Problem: Past work on PT has only used a suboptimal linear paths constructed from convex combinations of the reference and target log-densities.

Contribution: We extend the theory of PT to non-linear annealing paths and propose a flexible family of paths with a tractable algorithm to optimize over them.

Parallel tempering

Annealing path: π_t is a path of distributions continuously deforming between π_0 and π_1 at $t = 0, 1$

Traditionally a **Linear path** is used: $\pi_t(x) \propto \pi_0(x)^{1-t}\pi_1(x)^t$



Very little overlap between adjacent distributions

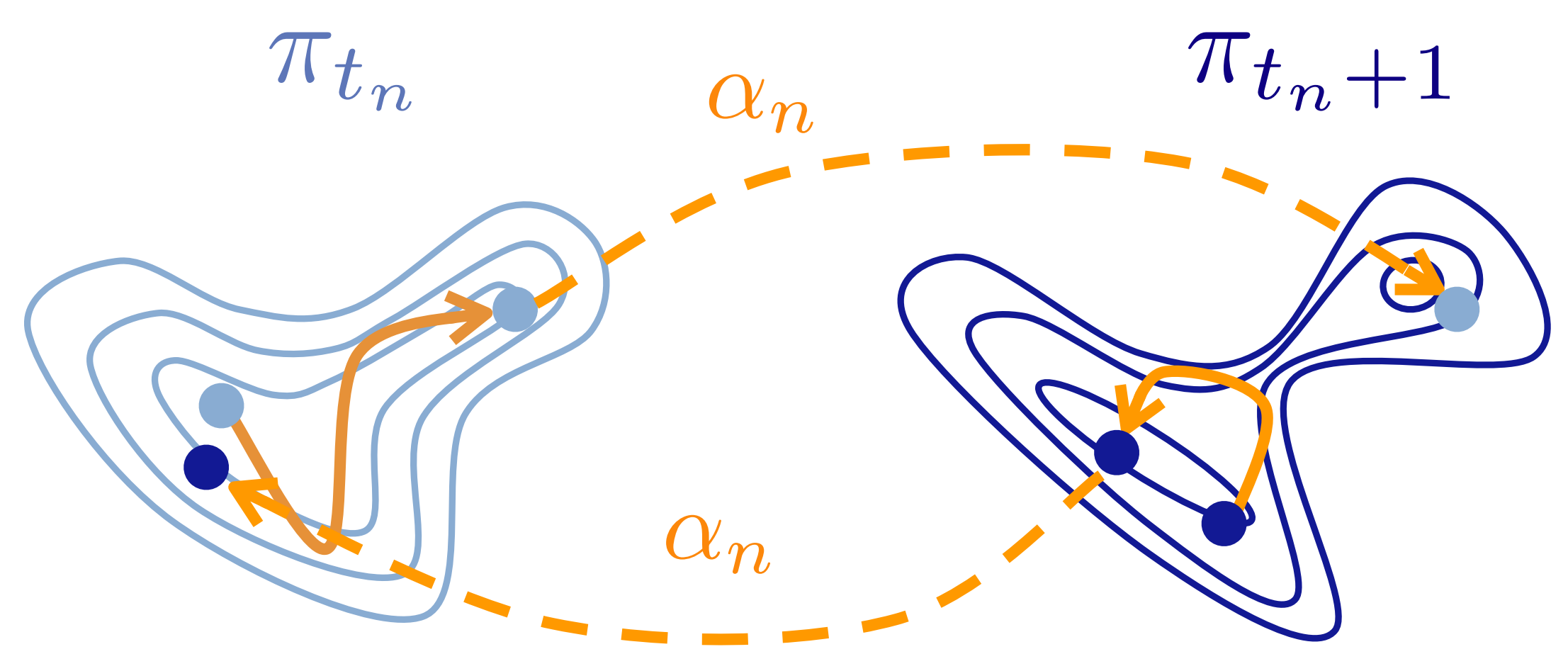
Fix an annealing schedule $\mathcal{T}_N = (t_n)_{n=0}^N$ of interpolating points on the path.

$$0 = t_0 < t_1 < \cdots < t_N = 1, \quad \|\mathcal{T}_N\| = \max_n |t_{n+1} - t_n|$$

Run $N+1$ chains in parallel targeting π_{t_n} along each path and alternate between local exploration and communication moves.

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

Communication: Propose swap between chains n and $n+1$ and accept with Metropolis-Hastings acceptance probability α_n



Objective: Want to maximize the **round trip rate**, $\tau(\mathcal{T}_N)$, the fraction of samples from the reference that reach the target. Round trip rate satisfies [Syed et al. 2019]

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

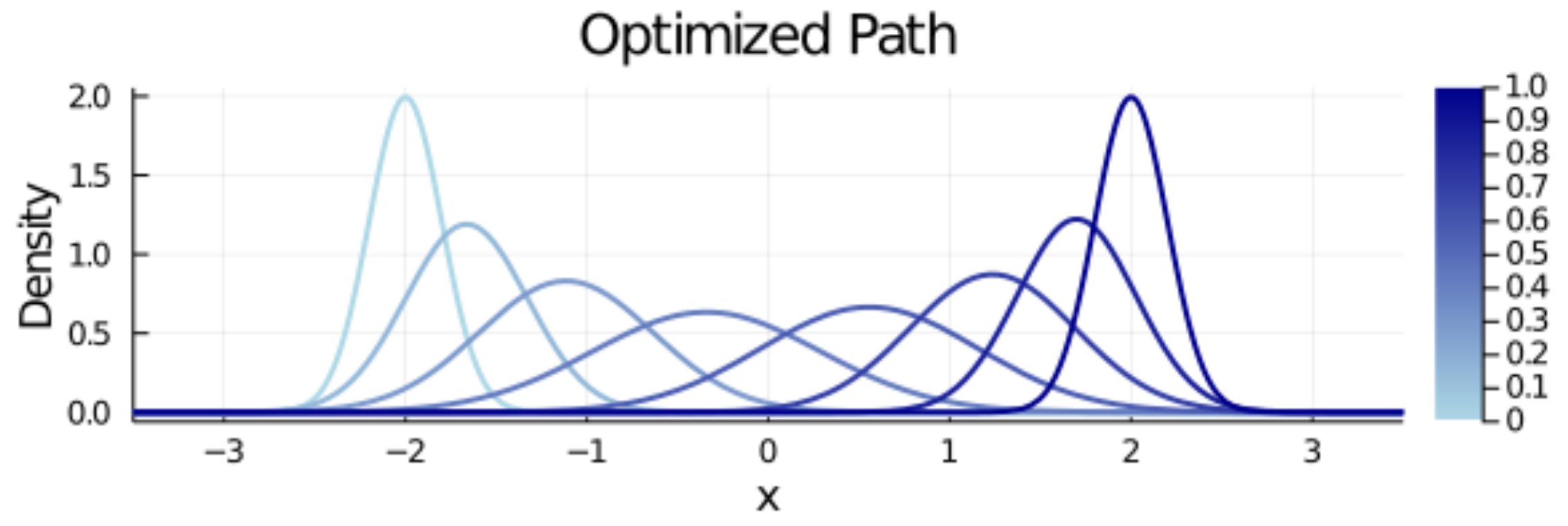
Suboptimality of linear path

Linear path is suboptimal even in the space of Gaussian distributions.

Proposition: Suppose $\pi_0 = N(\mu_0, \sigma^2)$ and $\pi_1 = N(\mu_1, \sigma^2)$ with $z = |\mu_1 - \mu_0|/\sigma$, then as $z \to \infty$:

1. For the **linear** path $\tau = \Theta(1/z)$
2. $\exists$ a **non-linear** annealing path in the space of Gaussians with $\tau = \Omega(1/\log z)$

Optimized path flattens distributions to improve overlap



Communication barrier of annealing path

The round trip rate is sensitive to both the path and schedule. We need an objective for just the path that is robust to the choice of schedule.

Theorem: Given a path π_t, there is a $\Lambda \geq 0$ such that as $N \to \infty$,

$$\lim_{\delta \to 0} \sup_{\mathcal{T}_N : \|\mathcal{T}_N\| \leq \delta} |\tau(\mathcal{T}_N) - (2 + 2\Lambda)^{-1}| = 0$$

If we define $\Lambda(\mathcal{T}_N) = \sum_{n=0}^{N-1} r_n$, then

$$\lim_{\delta \to 0} \sup_{\mathcal{T}_N : \|\mathcal{T}_N\| \leq \delta} |\Lambda - \Lambda(\mathcal{T}_N)| = 0$$

Λ is called the **global communication barrier (GCB)** and controls the asymptotic efficiency of PT. Intuitively can be thought of as the "length" of the path π_t.

Path optimization

Given a family of annealing paths $\{\pi_t^\phi\}_{\phi \in \Phi}$, for large N, the round trip rate is maximized when Λ^ϕ is minimized. The GCB satisfies,

$$2\Lambda^\phi(\mathcal{T}_N)^2 \leq N \sum_{n=0}^{N-1} \mathrm{SKL}(\pi_{t_n}, \pi_{t_{n+1}}) \equiv \mathcal{L}^\phi(\mathcal{T}_N)$$

Admits tractable gradient estimates

Algorithm:

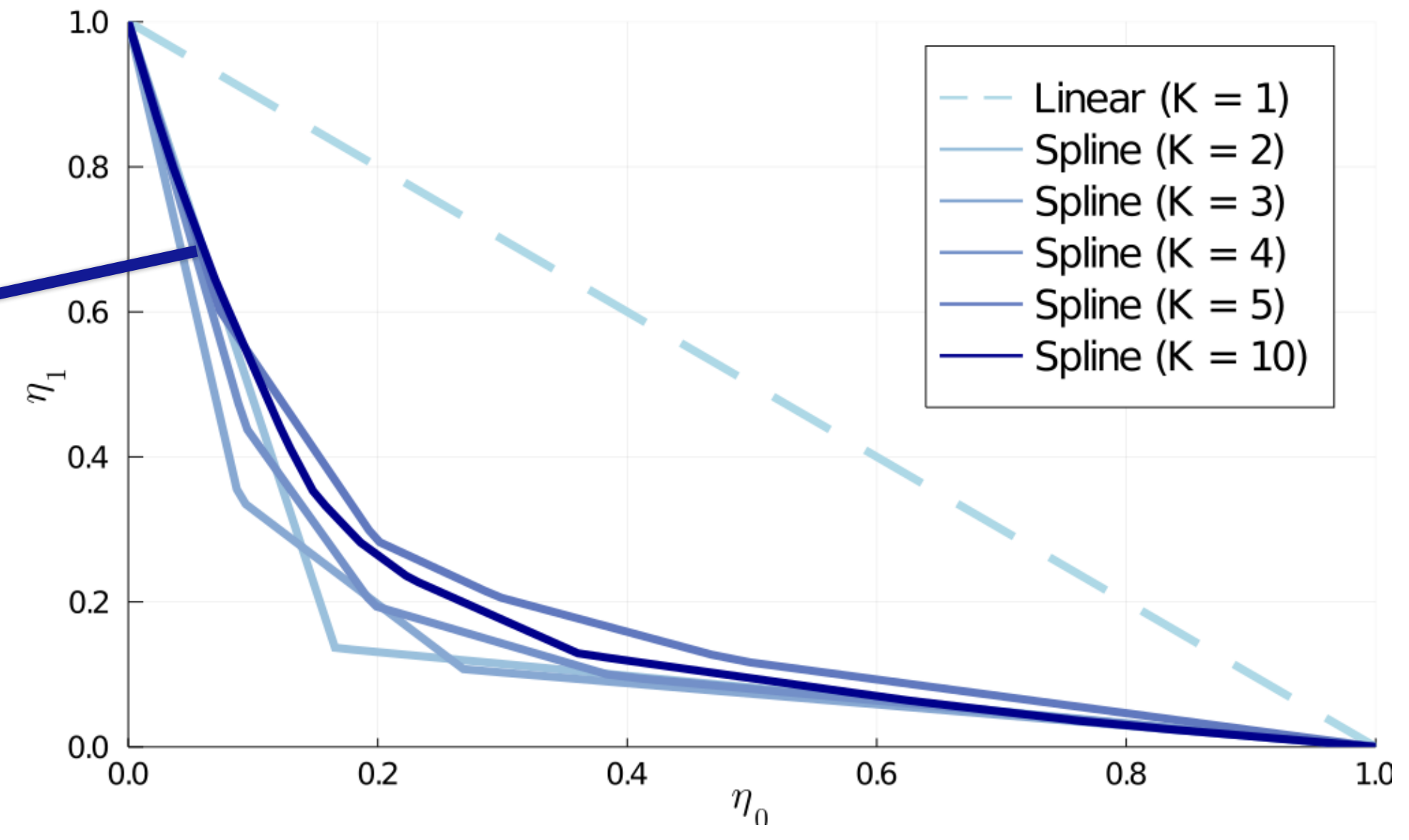
1) Given $\mathcal{T}_N$, ϕ, run non-reversible PT algorithm (**NRPT**) (Syed et al. 2019)
2) Use samples to update schedule $\mathcal{T}_N$ using procedure in (Syed et al. 2019)
3) Update ϕ with gradient decent with loss $\mathcal{L}^\phi(\mathcal{T}_n)$
4) Repeat 1-3 until computational budget is depleted.

Spline path family

For any $\eta(t) = (\eta_0(t), \eta_1(t))$ where $\eta(0) = (1,0)$, $\eta(1) = (0,1)$ we can generate an annealing path:

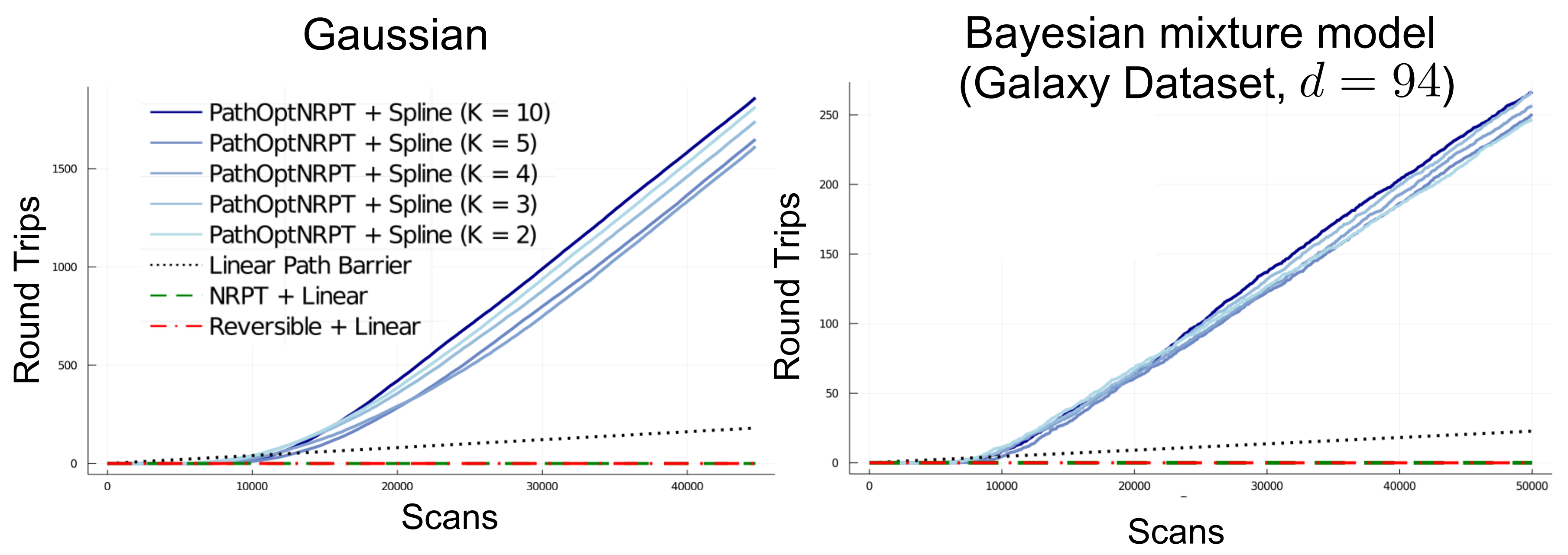
$$\pi_t(x) \propto \pi_0(x)^{\eta_0(t)}\pi_1(x)^{\eta_1(t)}$$

Given $K \geq 2$, we can optimize over the annealing path family generated by $\eta(t)$ in the set of linear splines with K knots.



Experiments

Optimized spline path beats the theoretically optimal performance of the linear path and significantly improves upon the state-of-the-art PT methods.

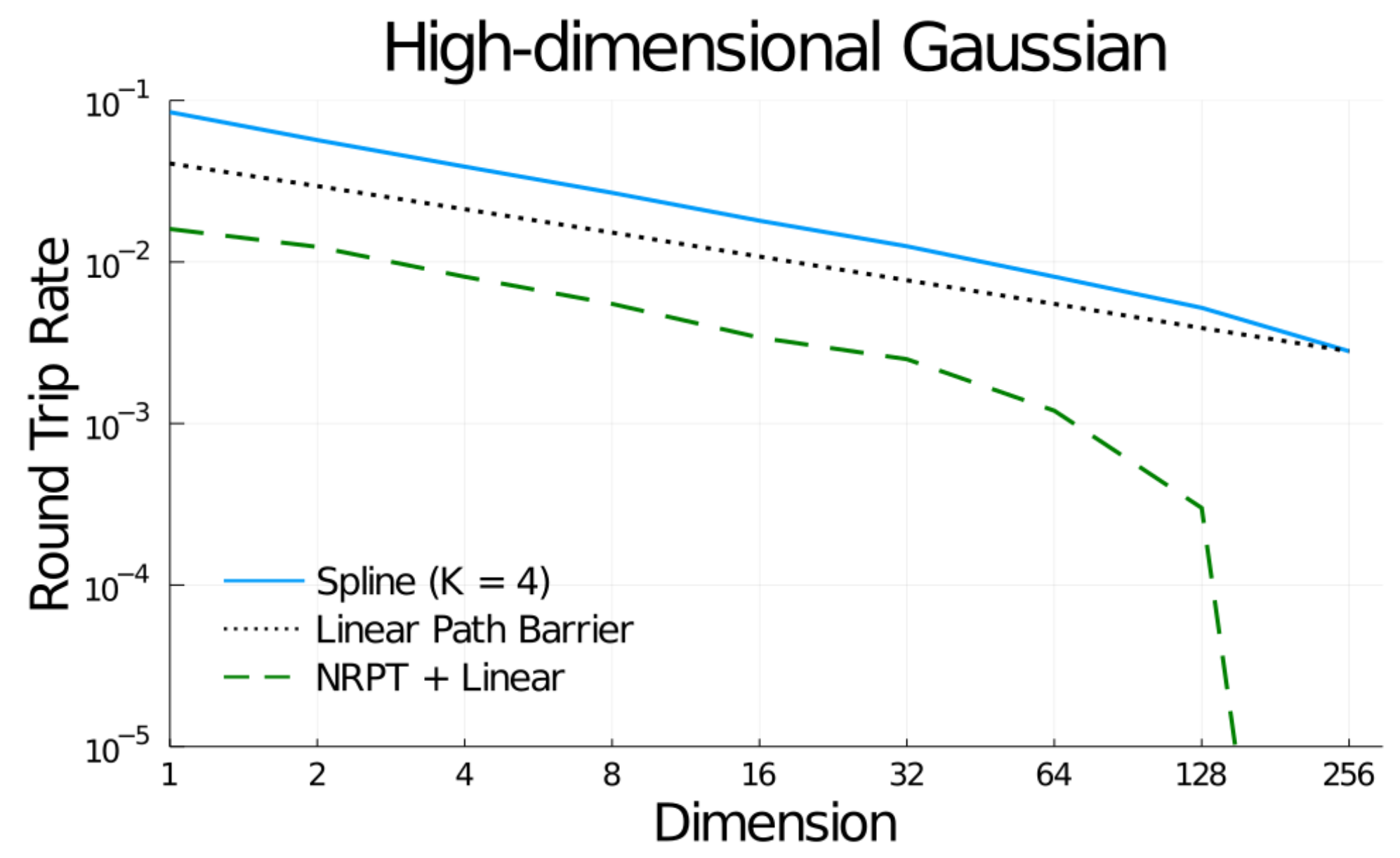


High-dimensional scaling

$$\pi_0 = N(-\mathbf{1}_d, (0.1)^2 I_d)$$

$$\pi_1 = N(\mathbf{1}_d, (0.1)^2 I_d)$$

Optimized spline path scales better than linear path for the same computational budget as dimension increases.



[1] UBC Department of Statistics, * Equal contribution

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