Parallel tempering on optimized paths

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

Abstract

Parallel tempering (PT) is a class of Markov chain Monte Carlo algorithms that constructs a path of distributions annealing between a tractable reference and an intractable target, and then interchanges states along the path to improve mixing in the target. The performance of PT depends on how quickly a sample from the reference distribution makes its way to the target, which in turn depends on the particular path of annealing distributions. However, past work on PT has used only simple paths constructed from convex combinations of the reference and target log-densities. This paper begins by demonstrating that this path performs poorly in the setting where the reference and target are nearly mutually singular. To address this issue, we expand the framework of PT to general families of paths, formulate the choice of path as an optimization problem that admits tractable gradient estimates, and propose a flexible new family of spline interpolation paths for use in practice. Theoretical and empirical results both demonstrate that our proposed methodology breaks previously-established upper performance limits for traditional paths.

1. Introduction

Markov Chain Monte Carlo (MCMC) methods are widely used to approximate intractable expectations with respect to un-normalized probability distributions over general state spaces. For hard problems, MCMC can suffer from poor mixing. For example, faced with well-separated modes, MCMC methods often get trapped exploring local regions of high probability. Parallel tempering (PT) is a widely applicable methodology (Geyer, 1991) to tackle poor mixing of MCMC algorithms.

Suppose we seek to approximate an expectation with respect to an intractable target density $\pi_1$. Denote by $\pi_0$ a reference density defined on the same space, which is assumed to be tractable in the sense of the availability of an efficient sampler. This work is motivated by the case where $\pi_0$ and $\pi_1$ are nearly mutually singular. A typical case is where the target is a Bayesian posterior distribution, the reference is the prior—for which i.i.d. sampling is typically possible—and the prior is misspecified.

PT methods are based on a specific continuum of densities $\pi_t \propto \pi_0^{1-t} \pi_1^t$, $t \in [0, 1]$, bridging $\pi_0$ and $\pi_1$. This path of intermediate distributions is known as the power posterior path in the literature, but in our framework it will be more natural to think of these continua as a linear paths, as they linearly interpolate between log-densities. PT algorithms discretize the path at some $0 = t_0 < \cdots < t_N = 1$ to obtain a sequence of densities $\pi_{t_0}, \pi_{t_1}, \ldots, \pi_{t_N}$. See Figure 1 (top) for an example of a linear path for two nearly mutually singular Gaussian distributions.

Given the path discretization, PT involves running $N + 1$ MCMC chains that together target the product distribution $\pi_{t_0} \pi_{t_1} \cdots \pi_{t_N}$. Based on the assumption that the chain $\pi_0$ can be sampled efficiently, PT uses swap-based interactions between neighbouring chains to propagate the exploration done in $\pi_0$ into improved exploration in the chain of interest $\pi_1$. By designing these swaps as Metropolis–Hastings moves, PT guarantees that the marginal distribution of the $N$th chain converges to $\pi_1$; and in practice, the rate of convergence is often much faster compared to running a single chain (Woodard et al., 2009). PT algorithms are extensively used in hard sampling problems arising in statistics, physics, computational chemistry, phylogenetics, and machine learning (Desjardins et al., 2014; Ballnus et al., 2017; Kamberaj, 2020; Müller & Bouckaert, 2020).

Notwithstanding empirical and theoretical successes, existing PT algorithms also have well-understood theoretical limitations. Earlier work focusing on the theoretical analysis of reversible variants of PT has shown that adding too many intermediate chains can actually deteriorate performance (Lingenheil et al., 2009; Atchadé et al., 2011). Recent work (Syed et al., 2019) has shown that a nonreversible variant of PT (Okabe et al., 2001) is guaranteed to dominate its classical reversible counterpart, and moreover that in the
nonreversible regime adding more chains does not lead to performance collapse. However, even with these more efficient non-reversible PT algorithms, Syed et al. (2019) established that the improvement brought by higher parallelism will asymptote to a fundamental limit known as the global communication barrier.

In this work, we show that by generalizing the class of paths interpolating between \( \pi_0 \) and \( \pi_1 \) from linear to nonlinear, the global communication barrier can be broken, leading to substantial performance improvements. Importantly, the nonlinear path used to demonstrate this breakage is computed using a practical algorithm that can be used in any situation where PT is applicable. An example of a path optimized using our algorithm is shown in Figure 1 (bottom).

We also present a detailed theoretical analysis of parallel tempering algorithms based on nonlinear paths. Using this analysis we prove that the performance gains obtained by going from linear to nonlinear path PT algorithms can be arbitrarily large. Our theoretical analysis also motivates a principled objective function used to optimize over a parametric family of paths.

**Literature review** Beyond parallel tempering, several methods to approximate intractable integrals relay on a path of distributions from a reference to a target distribution, and there is a rich literature on the construction and optimization of nonlinear paths for annealed importance sampling type algorithms (Gelman & Meng, 1998; Rischar et al., 2018; Grosse et al., 2013; Brekelmans et al., 2020). These algorithms are highly parallel; however, for challenging problems, even when combined with adaptive step size procedures (Zhou et al., 2016) they typically suffer from particle degeneracy (Syed et al., 2019, Sec. 7.4). Moreover, these methods use different path optimization criteria which are not well motivated in the context of parallel tempering. Some special cases of non-linear paths have been used in the PT literature (Whitfield et al., 2002; Tawn et al., 2020). Whitfield et al. (2002) construct a non-linear path inspired by the concept of Tsallis entropy, a generalization of Boltzmann-Gibbs entropy, but do not provide algorithms to optimize over this path family. The work of Tawn et al. (2020), also considers a specific example of a nonlinear path distinct from the ones explored in this paper. However, the construction of the nonlinear path in Tawn et al. (2020) requires knowledge of the location of the modes of \( \pi_1 \) and hence makes their algorithm less broadly applicable than standard PT.

**2. Background**

In this section, we provide a brief overview of parallel tempering (PT) (Geyer, 1991), as well as recent results on nonreversible communication (Okabe et al., 2001; Sakai & Hukushima, 2016; Syed et al., 2019). Define a reference unnormalized density function \( \pi_0 \) for which sampling is tractable, and an unnormalized target density function \( \pi_1 \) for which sampling is intractable; the goal is to obtain samples from \( \pi_1 \). Define a path of distributions \( \pi_t \propto \pi_0^{1-t} \pi^t \) for \( t \in [0, 1] \) from the reference to the target. Finally, define the annealing schedule \( T_N \) to be a monotone sequence in \([0,1]\), satisfying

\[
T_N = (t_n)_{n=0}^N, \quad 0 = t_0 \leq t_1 \leq \cdots \leq t_N = 1
\]

The core idea of parallel tempering is to construct a Markov chain \((X_0^0, \ldots, X_N^N)\), \( m = 1, 2, \ldots \), that (1) has invariant distribution \( \pi_{t_0} \cdot \pi_{t_1} \cdots \pi_{t_N} \)—such that we can treat the marginal chain \( X_0^N \) as samples from the target \( \pi_1 \)—and (2) swaps components of the state vector such that independent samples from component \( 0 \) (i.e., the reference \( \pi_0 \)) traverse along the annealing path and aid mixing in component \( N \) (i.e., the target \( \pi_1 \)). This is possible to achieve by iteratively performing a local exploration move followed by a communication move as shown in Algorithm 1.

**Local Exploration** Given \((X_{m-1}^0, \ldots, X_{m-1}^N)\), we obtain an intermediate state \((\tilde{X}_m^0, \ldots, \tilde{X}_m^N)\) by updating the \( n \)th component using any MCMC move targeting \( \pi_{t_n} \) for \( n = 0, \ldots, N \). This move can be performed in parallel across components since each is updated independently.

**Communication** Given the intermediate state \((\tilde{X}_m^0, \ldots, \tilde{X}_m^N)\), we apply pairwise swaps of components \( n \) and \( n + 1 \), \( n \in S_m \subseteq \{0, \ldots, N - 1\} \) for swapped index set \( S_m \). Formally, a swap is a move from \((X_0^n, \ldots, X_N^n)\) to \((x_0^n, \ldots, x_{n+1}^n, x_n^n, \ldots, X_N^n)\), which is accepted with probability

\[
\alpha_n = 1 \wedge \frac{\pi_{t_n}(x_{n+1}^n)\pi_{t_{n+1}}(x_n^n)}{\pi_{t_n}(x^n_n)\pi_{t_{n+1}}(x_{n+1}^n)}.
\]

Since each swap only depends on components \( n, n + 1 \), one can perform all of the swaps in \( S_m \) in parallel, as long as \( n \in S_m \) implies \( (n + 1) \notin S_m \). The largest collection of such non-interfering swaps is \( S_m \in \{S_{even}, S_{odd}\} \), where \( S_{even}, S_{odd} \) are the even and odd subsets of \( \{0, \ldots, N - 1\} \) respectively. In non-reversible PT (NRPT) (Okabe et al., 2001), the swap set \( S_m \) at each step \( m \) is set to

\[
S_m = \begin{cases} S_{\text{even}} & \text{if } m \text{ is even} \\ S_{\text{odd}} & \text{if } m \text{ is odd.} \end{cases}
\]

**Round trips** The performance of PT is sensitive to both the local exploration and communication moves. The quantity commonly used to evaluate the performance of MCMC

---

1We assume all distributions share a common state space \( \mathcal{X} \) throughout, and will often suppress the arguments of (log-)density functions—i.e., \( \pi_1 \) instead of \( \pi_1(x) \)—for notational brevity.
Parallel tempering on optimized paths

We say a round trip rate is a performance measure from the PT literature (Katzgraber et al., 2006; Lingenheil et al., 2009) that is designed to assess communication efficiency alone. We say a round trip has occurred when a new sample from the reference π₀ travels to the target π₁ and then back to π₀; the round trip rate τ(\mathcal{T}_N) is the frequency at which round trips occur. Based on simplifying assumptions on the local exploration moves, the round trip rate may be expressed as (Syed et al., 2019, Section 3.5)

\[
\tau(\mathcal{T}_N) = \left(2 + 2 \sum_{n=0}^{N-1} \frac{r(t_n, t_{n+1})}{1 - r(t_n, t_{n+1})}\right)^{-1},
\]

where \(r(t, t')\) is the expected probability of rejection between chains \(t, t' \in \{0, 1\}\),

\[
r(t, t') = \mathbb{E} \left[ 1 \wedge \frac{\pi_t(X') \pi_{t'}(X)}{\pi_t(X) \pi_{t'}(X')} \right],
\]

and \(X, X'\) have distributions \(X \sim \pi_t, X' \sim \pi_{t'}\). Further, if \(\mathcal{T}_N\) is refined so that \(\|\mathcal{T}_N\| \to 0\) as \(N \to \infty\), we find that the asymptotic (in \(N\)) round trip rate is

\[
\tau_\infty = \lim_{N \to \infty} \tau(\mathcal{T}_N) = (2 + 2\Lambda)^{-1},
\]

where \(\Lambda \geq 0\) is a constant associated with the pair \(\pi_0, \pi_1\) called the global communication barrier (Syed et al., 2019). Note that \(\Lambda\) does not depend on the number of chains \(N\) or discretization schedule \(\mathcal{T}_N\).

3. General annealing paths

In the following, we use the terminology annealing path to describe a continuum of distributions interpolating between \(\pi_0\) and \(\pi_1\); this definition will be formalized shortly. The previous work reviewed in the last section assumes that the annealing path has the form \(\pi_t \propto \pi_0^{1-t} \pi_1^t\), i.e., that the annealing path linearly interpolates between the log densities. A natural question is whether using other paths could lead to an improved round trip rate.

In this work we show that the answer to this question is positive. The following proposition demonstrates that the traditional path \(\pi_t \propto \pi_0^{1-t} \pi_1^t\) suffers from an arbitrarily suboptimal global communication barrier even in simple examples with Gaussian reference and target distributions.

Algorithm 1 NRPT

Require: state \(x_0\), path \(\pi_t\), schedule \(\mathcal{T}_N\), # iterations \(M\)

for \(m = 1\) to \(M\) do

1. \(x_m \leftarrow \text{LocalExploration}(x_{m-1})\)

2. \(S_m \leftarrow S_{\text{even}}\) if \(m\) is even, otherwise \(S_m \leftarrow S_{\text{odd}}\)

for \(n = 0\) to \(N - 1\) do

1. \(\alpha_n \leftarrow 1 \wedge \frac{\pi_{n+1}(x^n) \pi_{n^+1}(x^{n+1})}{\pi_{n^+1}(x^n) \pi_{n}(x^{n+1})}\)

2. \(r_n \leftarrow r_n + (1 - \alpha_n)\)

3. \(U_n \sim \text{Uniform}(0, 1)\)

if \(n \in S_m\) and \(U_n \leq \alpha_n\) then

1. \(\tilde{x}_m, x_{m+1} \leftarrow (\tilde{x}_{m+1}, \tilde{x}_m)\)

end if

end for

end for

\(r_n \leftarrow r_n / M\) for \(n \in \{0, \ldots, N - 1\}\)

Return: \(\{x_m\}_{m=1}^M, \{r_n\}_{n=0}^{N-1}\)

Figure 1. Two annealing paths between a \(\pi_0 = N(-2, 0.2^2)\) (light blue) and \(\pi_1 = N(2, 0.2^2)\) (dark blue): the traditional linear path (top) and an optimized nonlinear path (bottom). While the distributions in the linear path are nearly mutually singular, those in the optimized path overlap substantially, leading to faster round trips.

Proposition 1. Suppose the reference and target distributions are \(\pi_0 = N(\mu_0, \sigma^2)\) and \(\pi_1 = N(\mu_1, \sigma^2)\), and define \(z = |\mu_1 - \mu_0| / \sigma\). Then as \(z \to \infty\),

1. the path \(\pi_t \propto \pi_0^{1-t} \pi_1^t\) has \(\tau_\infty = \Theta(1/z)\), and

2. there exists a path of Gaussians distributions with \(\tau_\infty = \Omega(1 / \log z)\).

Therefore, upon decreasing the variance of the reference and target while holding their means fixed, the traditional linear annealing path obtains an exponentially smaller asymptotic
We introduce the precise general definition of an annealing path. The solution to this issue (bottom) is to allow the distribution of parallel tempering based on that path. For any fixed $N$, these distributions become nearly mutually singular, leading to arbitrarily low round trip rates. The solution to this issue (bottom) is to allow the distributions along the path to have increased variances, thereby maintaining mutual overlap and the ability to swap components with a reasonable probability. This motivates the need to design more general annealing paths. In the following, we introduce the precise general definition of an annealing path, an analysis of path communication efficiency in parallel tempering, and a rigorous formulation of—and solution to—the problem of tuning path parameters to maximize communication efficiency.

3.1. Assumptions

Let $\mathcal{P}(\mathcal{X})$ be the set of probability densities with full support on a state space $\mathcal{X}$. For any collection of densities $\pi_t \in \mathcal{P}(\mathcal{X})$ with index $t \in [0, 1]$, associate to each a log-density $W_t$ such that

$$W_t(x) = \log \pi_t(x), \quad x \in \mathcal{X},$$

where $Z_t = \int_{\mathcal{X}} \exp(W_t(x)) \, dx$ is the normalizing constant. Definition 1 provides the conditions necessary to form a path of distributions from a reference $\pi_0$ to a target $\pi_1$ that are well-behaved for use in parallel tempering.

**Definition 1.** An annealing path is a map $\pi(t) : [0, 1] \to \mathcal{P}(\mathcal{X})$, denoted $t \mapsto \pi_t$, such that for all $x \in \mathcal{X}$, $\pi_t(x)$ is continuous in $t$.

There are many ways to move beyond the standard linear path $\pi_t \propto \pi_0^{1-t} \pi_1^t$. For example, consider a nonlinear path $\pi_t \propto \eta_0(t) \pi_0 + \eta_1(t) \pi_1$ where $\eta_i : [0, 1] \to \mathbb{R}$ are continuous functions such that $\eta_0(0) = \eta_1(1) = 1$ and $\eta_0(1) = \eta_1(0) = 0$. As long as for all $t \in [0, 1]$, $\pi_t$ is a normalizable density this is a valid annealing path between $\pi_0$ and $\pi_1$. Further, note that the path parameter does not necessarily have to appear as an exponent: consider for example the mixture path $\pi_t \propto (1 - t)\pi_0 + t\pi_1$. Section 4 provides a more detailed example based on linear splines.

3.2. Communication efficiency analysis

Given a particular annealing path satisfying Definition 1, we require a method to characterize the round trip rate performance of parallel tempering based on that path. The results presented in this section form the basis of the objective function used to optimize over paths, as well as the foundation for the proof of Proposition 1.

We start with some notation for the rejection rates involved when Algorithm 1 is used with nonlinear paths (Equation 4). For $t, t' \in [0, 1]$, $x \in \mathcal{X}$, define the rejection rate function $r : [0, 1]^2 \to [0, 1]$ to be

$$r(t, t') = 1 - \mathbb{E}\left[\exp\left(\min\{0, A_t(t') (X, X')\}\right)\right]$$

where $X \sim \pi_t$ and $X' \sim \pi_{t'}$ are independent. Assuming that all chains have reached stationarity, and all chains undergo efficient local exploration—i.e., $W_t(X), W_t(X')$ are independent when $X \sim \pi_t$ and $X'$ is generated by local exploration from $X$—then the round trip rate for a particular schedule $T_N$ has the form given earlier in Equation (2). This statement follows from Syed et al. (2019, Cor. 1) without modification, because the proof of that result does not depend on the form of the acceptance ratio.

Our next objective is to characterize the asymptotic communication efficiency of a nonlinear path in the regime where $N \to \infty$ and $\|T_N\| \to 0$—which establishes its fundamental ability to take advantage of parallel computation. In other words, we require a generalization of the asymptotic result in Equation (3). In this case, previous theory relies on the particular form of the acceptance ratio for linear paths (Syed et al., 2019); in the remainder of this section, we provide a generalization of the asymptotic result for nonlinear paths by piecewise approximation by a linear spline.

For $t, t' \in [0, 1]$, define $\Lambda(t, t')$ to be the global communication barrier for the linear secant connecting $\pi_t$ and $\pi_{t'}$,

$$\Lambda(t, t') = \frac{1}{2} \int_0^1 \mathbb{E}[|A_t(t') (X_s, X'_s)|] \, ds,$$

and

$$X_s, X'_s \sim_{\text{i.i.d.}} \frac{1}{Z_s(t, t')} \exp\left( (1 - s)W_t + sW_{t'} \right).$$

Lemma 1 shows that the global communication barrier $\Lambda(t, t')$ along the secant of the path connecting $t$ to $t'$ is a good approximation of the true rejection rate $r(t, t')$ with $O(|t - t'|^3)$ error as $t \to t'$.

**Lemma 1.** Suppose that for all $x \in \mathcal{X}$, $W_t(x)$ is piecewise continuously differentiable in $t$, and that there exists $V_1 : \mathcal{X} \to [0, \infty)$ such that

$$\forall x \in \mathcal{X}, \sup_{t \in [0, 1]} \left| \frac{dW_t}{dt}(x) \right| \leq V_1(x),$$

and

$$\sup_{t \in [0, 1]} \mathbb{E}_{\pi_t}[V_1^3] < \infty.$$  

Then there exists a constant $C < \infty$ independent of $t, t'$ such that for all $t, t' \in [0, 1]$,

$$|r(t, t') - \Lambda(t, t')| \leq C |t - t'|^3.$$
A direct consequence of Lemma 1 is that for any fixed schedule $T_N$,

$$
\sum_{n=0}^{N-1} r(t_n, t_{n+1}) - \Lambda(T_N) \leq C \|T_N\|^2,
$$

where $\Lambda(T_N) = \sum_{n=0}^{N-1} \Lambda(t_n, t_{n+1})$. Intuitively, in the $\|T_N\| \approx 0$ regime where rejection rates are low,

$$
\frac{r(t_n, t_{n+1})}{1 - r(t_n, t_{n+1})} \approx r(t_n, t_{n+1}),
$$

and we have that $\tau(T_N) \approx (2 + 2\Lambda(T_N))^{-1}$. Therefore, $\Lambda(T_N)$ characterizes the communication efficiency of the path in a way that naturally extends the global communication barrier from the linear path case. Theorem 2 provides the precise statement: the convergence is uniform in $T_N$ and depends only on $\|T_N\|$, and $\Lambda(T_N)$ itself converges to a constant $\Lambda$ in the asymptotic regime. We refer to $\Lambda$, defined below in Equation (13) as the global communication barrier for the general annealing path.

**Theorem 2.** Suppose that for all $x \in \mathcal{X}$, $W_i(x)$ is piecewise twice continuously differentiable in $t$, that there exists $V_1 : \mathcal{X} \to [0, \infty)$ satisfying (6) and (7), and that there exists $V_2 : \mathcal{X} \to [0, \infty)$, $\epsilon > 0$ satisfying

$$
\forall x \in \mathcal{X}, \sup_{t \in [0,1]} \left| \frac{d^2 W_i}{d^2 t} (x) \right| \leq V_2(x),
$$

and

$$
\sup_{t \in [0,1]} E_{\pi_1} \left[ (1 + V_1) \exp(\epsilon V_2) \right] < \infty.
$$

Then,

$$
\lim_{\delta \to 0} \sup_{T_N: \|T_N\| \leq \delta} \left| (2 + 2\Lambda(T_N))^{-1} - \tau(T_N) \right| = 0
$$

and

$$
\lim_{\delta \to 0} \sup_{T_N: \|T_N\| \leq \delta} \left| \Lambda(T_N) - \Lambda \right| = 0,
$$

where $\Lambda = \int_0^1 \lambda(t) dt$ for an instantaneous rejection rate function $\lambda : [0, 1] \to [0, \infty)$ given by

$$
\lambda(t) = \lim_{\Delta t \to 0} \frac{r(t + \Delta t, t)}{\Delta t} = \frac{1}{2} \mathbb{E} \left[ \left( \frac{dW_i}{dt} (X_t) - \frac{dW_i}{dt} (X'_t) \right)^2 \right], \quad X_t, X'_t \sim \pi_t.
$$

The integrability condition is required to control the tail behaviour of distributions formed by linearized approximations to the path $\pi_t$. This condition is satisfied by a wide range of annealing paths, e.g., the linear spline paths proposed in this work—since in that case $V_2 = 0$.

### 3.3. Annealing path families and optimization

It is often the case that there are a set of candidate annealing paths in consideration for a particular target $\pi$. For example, if a path has tunable parameters $\phi \in \Phi$ that govern its shape, we can generate a collection of annealing paths that target $\pi$ by varying the parameter $\phi$. We call such collections an annealing path family.

**Definition 3.** An annealing path family for target $\pi_1$ is a collection of annealing paths $\{\pi_1^\phi\}_{\phi \in \Phi}$ such that for all parameters $\phi \in \Phi$, $\pi_1^\phi = \pi_1$.

There are many ways to construct useful annealing path families. For example, if one is provided a parametric family of variational distributions $\{q_\phi : \phi \in \Phi\}$ for some parameter space $\Phi$, one can construct the annealing path family of linear paths $\pi_1^\phi = q_\phi^{-1} \pi_1^1$ from a variational reference $q_\phi$ to the target $\pi_1$. More generally, given $\eta_i(t)$ satisfying the constraints in Section 3.1, $\pi_1^\phi = q_\phi^{-1} \pi_1^1$ defines a nonlinear annealing path family. Another example of an annealing path family used in the context of PT are $q$-paths $\{\pi_1^q\}_{q \in [0, 1]}$ (Whitfield et al., 2002). Given a fixed reference and target $\pi_0, \pi_1$, the path $\pi_1^q$ interpolates between the mixture path ($q = 0$) and the linear path ($q = 1$) (Brekels et al., 2020). In Section 4, we provide a new flexible class of nonlinear paths based on splines that is designed specifically to enhance the performance of parallel tempering.

Since every path in an annealing path family has the desired target distribution $\pi_1$, we are free to optimize the path over the tuning parameter space $\phi \in \Phi$ in addition to optimizing the schedule $T_N$. Motivated by the analysis of Section 3.2, a natural objective function for this optimization to consider is the non-asymptotic round trip rate

$$
\phi^* : T_N^* = \arg \max_{\phi \in \Phi, T_N} \tau^{\phi}(T_N)
$$

$$
= \arg \min_{\phi \in \Phi, T_N} \sum_{n=0}^{N-1} r^{\phi}(t_n, t_{n+1}) \frac{1 - r^{\phi}(t_n, t_{n+1})}{(2 + 2\Lambda(T_N))^2},
$$

where now the round trip rate and rejection rates depend both on the schedule and path parameter, denoted by superscript $\phi$. We solve this optimization using an approximate coordinate-descent procedure, iterating between an update of the schedule $T_N$ for a fixed path parameter $\phi \in \Phi$, followed by a gradient step in $\phi$ based on a surrogate objective function and a fixed schedule. This is summarized in Algorithm 2. We outline the details of schedule and path tuning procedure in the following.

**Tuning the schedule** Fix the value of $\phi$, which fixes the path. We adapt a schedule tuning algorithm from past work...
We propose a surrogate, the symmetric KL divergence, motivated as follows. Consider first the global communication barrier; Theorem 2 guarantees that as long as \( \|T_N\| \approx 0 \), the non-asymptotic round trip rate is maximized when the rejection rates are all equal. The schedule that approximately achieves this satisfies

\[
\forall n \in \{1, \ldots, N-1\}, \quad \frac{1}{N} \int_0^{t_n} \lambda^{\phi}(s) \, ds = \frac{n}{N}. \tag{16}
\]

Following Syed et al. (2019), we use Monte Carlo estimates of the rejection rates \( r^{\phi}(t_n, t_{n+1}) \) to approximate \( t \mapsto \int_0^t \lambda^{\phi}(s) \, ds, \quad s \in [0,1] \) via a monotone cubic spline, and then use bisection search to solve for each \( t_n \) according to Equation (16).

### Optimizing the path

Fix the schedule \( T_N \); we now want to improve the path itself by modifying \( \phi \). However, in challenging problems this is not as simple as taking a gradient step for the objective in Equation (14). In particular, in early iterations—when the path is near its oft-poor initialization—the rejection rates satisfy \( r^{\phi}(t_n, t_{n+1}) \approx 1 \). As demonstrated empirically in Appendix F, gradient estimates in this regime exhibit a low signal-to-noise ratio that precludes their use for optimization.

We propose a surrogate, the symmetric KL divergence, motivated as follows. Consider first the global communication barrier \( \Lambda^{\phi}(T_N) \) for the linear spline approximation to the path; Theorem 2 guarantees that as long as \( \|T_N\| \) is small enough, one can optimize \( \Lambda^{\phi}(T_N) \) in place of the round trip rate \( r^{\phi}(T_N) \). By Jensen’s inequality,

\[
\frac{1}{N^2} \Lambda^{\phi}(T_N)^2 \leq \frac{1}{N} \sum_{n=0}^{N-1} \Lambda^{\phi}(t_n, t_{n+1})^2.
\]

Next, we apply Jensen’s inequality again to the definition of \( \Lambda^{\phi}(t_n, t_{n+1}) \) from (5), which shows that

\[
\Lambda^{\phi}(t_n, t_{n+1})^2 \leq \frac{1}{4} \int_0^{t_n} \mathbb{E}[A_{t_n, t_{n+1}}(X_s, X'_s)^2] \, ds,
\]

where \( X_s \) are defined in (5) are drawn from the linear path between \( \pi_{t_n}^{\phi} \) and \( \pi_{t_{n+1}}^{\phi} \). Finally, we note that the inner expectation is the path integral of the Fisher information metric along the linear path and evaluates to the symmetric KL divergence (Dabak & Johnson, 2002, Result 4),

\[
\int_0^1 \mathbb{E}[A_{t_n, t_{n+1}}(X_s, X'_s)^2] \, ds = 2 \text{SKL}(\pi_{t_n}^{\phi}, \pi_{t_{n+1}}^{\phi}).
\]

Therefore we have

\[
\frac{2 \Lambda^{\phi}(T_N)^2}{N} \leq \sum_{n=0}^{N-1} \text{SKL}(\pi_{t_n}^{\phi}, \pi_{t_{n+1}}^{\phi}). \tag{17}
\]

The slack in the inequality in Equation (17) could potentially depend on \( \phi \) even in the large \( N \) regime. Therefore, during optimization, we recommend monitoring the value of the original objective function (Equation (14)) to ensure that the optimization of the surrogate SKL objective indeed improves it, and hence the round trip rate performance of PT via Equation (2). In the experiments we display the values of both objective functions.

### 4. Spline annealing path family

In this section, we develop a family of annealing paths—the *spline annealing path family*—that offers a practical and flexible improvement upon the traditional linear paths considered in past work. We first define a general family of annealing paths based on the exponential family, and then provide the specific details of the spline family with a discussion of its properties. Empirical results in Section 5 demonstrate that the spline annealing path family resolves the problematic Gaussian annealing example in Figure 1.

#### 4.1. Exponential annealing path family

We begin with the practical desiderata for an annealing path family given a fixed reference \( \pi_0 \) and target \( \pi_1 \) distribution.\(^3\)

First, the traditional linear path \( \pi_t \propto \pi_0^{1-t} \pi_1^t \) should be a member of the family, so that one can achieve at least the round trip rate provided by that path. Second, the family should be broadly applicable and not depend on particular details of either \( \pi_0 \) or \( \pi_1 \). Finally, using the Gaussian example from Figure 1 and Proposition 1 as insight, the family should enable the path to smoothly vary from \( \pi_0 \) to \( \pi_1 \) while inflating / deflating the variance as necessary.

These desiderata motivate the design of the *exponential annealing path family*, in which each annealing path takes the form

\[
\pi_t \propto \pi_0^{\eta_0(t)} \pi_1^{\eta_1(t)} = \exp(\eta(t)^T W(x)),
\]

\(^3\)A natural extension of this discussion would include parametrized variational reference distribution families. For simplicity we restrict to a fixed reference.
for some function \( \eta(t) = (\eta_0(t), \eta_1(t)) \) and reference/target log densities \( W(x) = (W_0(x), W_1(x)) \). Intuitively, \( \eta_0(t) \) and \( \eta_1(t) \) represent the level of annealing for the reference and target respectively along the path. Proposition 2 shows that a broad collection of functions \( \eta \) indeed construct a valid annealing path family including the linear path.

**Proposition 2.** Let \( \Omega \subseteq \mathbb{R}^2 \) be the set
\[
\Omega = \left\{ \xi \in \mathbb{R}^2 : \int \exp(\xi^T W(x))dx < \infty \right\}.
\]
Suppose \((0, 1) \in \Omega \) and \( A \) is a set of piecewise twice continuously differentiable functions \( \eta : [0, 1] \to \Omega \) such that \( \eta(1) = (0, 1) \). Then
\[
\{ \pi_t(x) \propto \exp(\eta(t)^T W(x)) : \eta \in A \}
\]
is an annealing path family for target distribution \( \pi_1 \). If additionally \((1, 0) \in \Omega \), then the linear path \( \eta(t) = (1-t, t) \) may be included in \( A \). Finally, if for every \( \eta \in A \) there exists \( M > 0 \) such that \( \sup_t \max_k \|\eta_k'(t)\|_2, \|\eta_k''(t)\|_2 \leq M \) and Equation (11) holds with \( V_1 = V_2 = M \| W \|_2 \), then each path in the family satisfies the conditions of Theorem 2.

### 4.2. Spline annealing path family

Proposition 2 reduces the problem of designing a general family of paths of probability distributions to the much simpler task of designing paths in \( \mathbb{R}^2 \). We argue that a good choice can be constructed using linear spline paths connecting \( K \) knots \( \phi = (\phi_0, \ldots, \phi_K) \in (\mathbb{R}^2)^{K+1} \), i.e., for all \( k \in \{1, \ldots, K\} \) and \( t \in \left[ \frac{k-1}{K}, \frac{k}{K} \right] \),
\[
\eta^\phi(t) \mapsto (k-1)\phi_{k-1} + (Kt-k+1)\phi_k.
\]
Let \( \Omega \) be defined as in Proposition 2. The \( K \)-knot spline annealing path family is defined as the set of \( K \)-knot linear spline paths such that
\[
\phi_0 = (1, 0), \quad \phi_K = (0, 1), \quad \text{and} \quad \forall k, \phi_k \in \Omega.
\]

**Validity:** Since \( \Omega \) is a convex set per the proof of Proposition 2, we are guaranteed that \( \eta^\phi([0, 1]) \subseteq \Omega \), and so this collection of annealing paths is a subset of the exponential annealing family and hence forms a valid annealing path family targeting \( \pi_1 \).

**Convexity:** Furthermore, convexity of \( \Omega \) implies that tuning the knots \( \phi \in \Omega^{K+1} \) involves optimization within a convex constraint set. In practice, we enforce also that the knots are monotone in each component—i.e., the first component monotonically decreases, \( 1 = \phi_{0,0} \geq \phi_{1,0} \geq \cdots \geq \phi_{K,0} = 0 \) and the second increases, \( 0 = \phi_{0,1} \leq \phi_{1,1} \leq \cdots \leq \phi_{K,1} = 1 \)—such that the path of distributions always moves from the reference to the target. Because monotonicity constraint sets are linear and hence convex, the overall monotonicity-constrained optimization problem has a convex domain.

**Flexibility:** Assuming the family is nonempty, it trivially contains the linear path. Further, given a large enough number of knots \( K \), the spline annealing family well-approximates subsets of the exponential annealing family for fixed \( M > 0 \). In particular,
\[
\sup_{\eta \in A_{\phi}} \inf_{\phi \in K+1} \|\eta - \eta\|_\infty \leq \frac{M}{4K^2}.
\]

Figure 2 provides an illustration of the behaviour of optimized spline paths for a Gaussian reference and target. The path takes a convex curved shape: starting at the bottom right point of the figure (reference), this path corresponds to increasing the variance of the reference, shifting the mean from reference to target, and finally decreasing the variance to match the target. With more knots, this process happens more smoothly.

Appendix D provides an explicit derivation of the stochastic gradient estimates we use to optimize the knots of the spline annealing path family. It is also worth making two practical notes. First, to enforce the monotonicity constraint, we developed an alternative to the usual projection approach, as projecting into the set of monotone splines can cause several knots to become superposed. Instead, we maintain monotonicity as follows: after each stochastic gradient step, we identify a monotone subsequence of knots containing the endpoints, remove the nonmonotone jumps, and linearly interpolate between the knot subsequence with an even spacing. Second, we take gradient steps in a log-transformed space so that knot components are always strictly positive.

### 5. Experiments

In this section, we study the empirical performance of non-reversible PT based on the spline annealing path family \((K \in \{2, 3, 4, 5, 10\})\) from Section 4, with knots and schedule optimized using the tuning method from Section 3. We compare this method to two PT methods.
Parallel tempering on optimized paths

Figure 3. Top: Cumulative round trips averaged over 10 runs for the spline path with \( K = 2, 3, 4, 5, 10 \) (solid blue), NRPT using a linear path (dashed green), and reversible PT with linear path (dash/dot red). The slope of the lines represent the round trip rate. We observe large gains going from linear to non-linear paths \((K > 1)\). For all values of \( K > 1 \), the optimized spline path substantially improves on the theoretical upper bound on round trip rate possible using linear path (dotted black).

Bottom: Non-asymptotic communication barrier from Equation 15 (solid blue) and Symmetric KL (dash orange) as a function of iteration for one run of PathOptNRPT + Spline \((K = 4)\) knots.

These experiments demonstrate two primary conclusions: (1) tuned nonlinear paths provide a substantially higher round trip rate compared to linear paths for the examples considered; and (2) the symmetric KL sum objective (Equation 17) is a good proxy for the round trip rate as a tuning objective.

Based on standard linear paths: non-reversible PT with adaptive schedule (“NRPT+Linear”) (Syed et al., 2019), and reversible PT (“Reversible+Linear”) (Atchadé et al., 2011). Code for the experiments is available at https://github.com/vittrom/PT-pathoptim.

We use the terminology “scan” to denote one iteration of the for loop in Algorithm 2. The computational cost of a scan is comparable for all the methods, since the bottleneck is the local exploration step shared by all methods. These experiments demonstrate two primary conclusions: (1) tuned nonlinear paths provide a substantially higher round trip rate compared to linear paths for the examples considered; and (2) the symmetric KL sum objective (Equation 17) is a good proxy for the round trip rate as a tuning objective.

We run the following benchmark problems; see the supplement for details. Gaussian: a synthetic setup in which the reference distribution is \( \pi_0 = N(-1, 0.01^2) \) and the target is \( \pi_1 = N(1, 0.01^2) \). For this example we used \( N = 50 \) parallel chains and fixed the computational budget to 45000 samples. For Algorithm 2, the computational budget was divided equally over 150 scans, meaning 300 samples were used for every gradient update. “Reversible+Linear” performed 45000 local exploration steps with a communication step after every iteration while for “NRPT+Linear” the computational budget was used to adapt the schedule. The gradient updates were performed using Adagrad (Duchi et al., 2011) with learning rate equal to 0.2. Beta-binomial model: a conjugate Bayesian model with prior \( \pi_0(p) = \text{Beta}(180, 840) \) and likelihood \( L(x|p) = \text{Binomial}(x|n, p) \). We simulated data \( x_1, \ldots, x_{2000} \sim \text{Binomial}(100, 0.7) \) resulting in a posterior \( \pi_1(p) = \text{Beta}(1400180, 60840) \). The prior and posterior are heavily concentrated at 0.2 and 0.7 respectively. We used the same settings as for the Gaussian example.

Galaxy data: A Bayesian Gaussian mixture model applied to the galaxy dataset of (Roeder, 1990). We used six mixture components with mixture proportions \( w_0, \ldots, w_5 \), mixture component densities \( N(\mu_i, 1) \) for mean parameters \( \mu_0, \ldots, \mu_5 \), and a cluster label categorical variable for each data point. We placed a \( \text{Dir}(\mathbf{1}/6) \) prior on the proportions, where \( \mathbf{1}_6 = (1, 1, 1, 1, 1, 1) \) and a \( N(150, 1) \) prior on each of the mean parameters. We did not marginalize the cluster indicators, creating a multi-modal posterior inference problem over 94 latent variables. In this experiment we used \( N = 35 \) chains and fixed the computational budget to 50000 samples, divided into 500 scans using 100 samples each. We optimized the path using Adagrad with a learning rate of 0.3. Exploring the full posterior distribution is challenging in this context due to a combination of misspecification of the prior and label switching. Label switching refers to the invariance of the likelihood under relabelling of the cluster labels. In Bayesian problems label switching can lead to increased difficulty of the sampling problem as it generates symmetric multi-modal posterior distributions. High dimensional Gaussian: a similar setup to the
We fixed the number of spline knots with $K$. The bottom row of Figure 3 shows the value of the surrogate objective from Equation (15). In particular, these figures demonstrate that the SKL provides a surrogate objective that is a reasonable proxy for the non-asymptotic communication barrier, but does not exhibit as large estimation variance in early iterations when there are pairs of chains with rejection rates close to one.

Figure 4 shows the round trip rate as a function of the dimensionality of the problem for Gaussian target distributions. As the dimensionality increases, the sampling problem becomes fundamentally more difficult, explaining the decay in performance of both NRPT with a linear path and an optimized path. Both sampling methods are provided with a fixed computational budget in all runs. Due to this fixed budget, NRPT is unable for $d \geq 64$ to approach the optimal schedule in the allotted time. This leads to an increasing gap between the round trip rates of NRPT with linear path and the spline path for $d \geq 64$.

6. Discussion

In this work, we identified the use of linear paths of distributions as a major bottleneck in the performance of parallel tempering algorithms. To address this limitation, we have provided a theory of parallel tempering on nonlinear paths, a methodology to tune parametrized paths, and finally a practical, flexible family of paths based on linear splines. Future work in this line of research includes extensions to estimate normalization constants, as well as the development of techniques and theory surrounding the use of variational reference distributions.

References


