A. Proof of Proposition 1

Define $\pi_0 = N(\mu_0, \sigma^2)$ and $\pi_1 = N(\mu_1, \sigma^2)$ with $W_i(x) \propto -\frac{1}{2\sigma^2}(x-\mu_i)^2$ (throughout we use the proportionality symbol $\propto$ with log-densities to indicate an unspecified constant, additive with respect to W_i, multiplicative with respect to π_t). Suppose π_t is the linear path $\pi_t(x) \propto \exp(W_t)$ where $W_t = (1-t)W_0 + tW_1$. Note that as a function of x,

$$\begin{aligned} W_t(x) &\propto -\frac{1-t}{2\sigma^2}(x-\mu_0)^2 - \frac{t}{2\sigma^2}(x-\mu_1)^2 \\ &\propto -\frac{1}{2\sigma^2}(x-\mu_t)^2, \quad \mu_t = (1-t)\mu_0 + t\mu_1, \end{aligned}$$

and thus $\pi_t = N(\mu_t, \sigma^2)$. Taking a derivative of W_t, we find that

$$\frac{\mathrm{d}W_t}{\mathrm{d}t} = \frac{(\mu_1-\mu_0)\left(x - \frac{\mu_0+\mu_1}{2}\right)}{\sigma^2}.$$

We will now compute $\lambda(t)$. If $X_t, X_t' \sim \pi_t$, then

$$\begin{aligned} \lambda(t) &= \frac{1}{2}\mathbb{E}\left[\left|\frac{dW}{dt}(X_t) - \frac{dW}{dt}(X_t')\right|\right] \\ &= \frac{|\mu_1-\mu_0|}{2\sigma}\mathbb{E}\left[\left|\frac{X_t - \frac{\mu_0+\mu_1}{2}}{\sigma} - \frac{X_t' - \frac{\mu_0+\mu_1}{2}}{\sigma}\right|\right] \\ &= \frac{|\mu_1-\mu_0|}{2\sigma}\mathbb{E}\left[\left|\frac{X_t-\mu_t}{\sigma} - \frac{X_t'-\mu_t}{\sigma}\right|\right] \\ &= \frac{|\mu_1-\mu_0|}{2\sigma}\mathbb{E}\left[|Z-Z'|\right], \end{aligned}$$

where $Z, Z' \sim N(0,1)$. Thus $Z - Z' \sim N(0,2)$, and $|Z - Z'|$ has a folded normal distribution with expectation $2/\sqrt{\pi}$. This implies $\lambda(t) = z/\sqrt{\pi}$ where $z = |\mu_1 - \mu_0|/\sigma$ and $\Lambda = \int_0^1 \lambda(t)\mathrm{d}t = z/\sqrt{\pi}$. By Theorem 2, the asymptotic round trip rate $\tau_\infty^{\text{linear}}$ for the linear path satisfies,

$$\tau_\infty^{\text{linear}} = \frac{1}{2+2\Lambda} = \frac{1}{2+2z/\sqrt{\pi}} = \Theta\left(\frac{1}{z}\right).$$

We will now establish an upper bound for the communication barrier Λ for a general path π_t. If $X_t, X_t' \sim \pi_t$, then Theorem 2 and Jensen's inequality imply the following:

$$\begin{aligned} \Lambda &= \int_0^1 \frac{1}{2}\mathbb{E}\left[\sqrt{\left(\frac{dW}{dt}(X_t) - \frac{dW}{dt}(X_t')\right)^2}\right]\mathrm{d}t \\ &\le \int_0^1 \frac{1}{2}\sqrt{\mathbb{E}\left[\left(\frac{dW}{dt}(X_t) - \frac{dW}{dt}(X_t')\right)^2\right]}\mathrm{d}t \\ &= \frac{1}{\sqrt{2}}\int_0^1 \sqrt{\mathrm{Var}_{\pi_t}\left[\frac{dW_t}{dt}\right]}\mathrm{d}t \\ &= \frac{1}{\sqrt{2}}\Lambda_F, \end{aligned}$$

where Λ_F is the length of the the path π_t with the Fisher information metric. The geodesic path of Gaussians between π_0 and π_1 that minimizes Λ_F satisfies (Costa et al., 2015, Eq. 11, Sec. 2)

$$\Lambda_F = \sqrt{2}\log\left(1 + \frac{z^2}{4} + \frac{z}{4}\sqrt{8+z^2}\right). \tag{19}$$

Again, by Theorem 2, the asymptotic round trip rate $\tau_\infty^{\text{geodesic}}$ for the geodesic path satisfies

$$\tau_\infty^{\text{geodesic}} = \frac{1}{2+2\Lambda} \ge \frac{1}{2+2\Lambda_F} = \Theta\left(\frac{1}{\log z}\right).$$

B. Proof of Lemma 1

Definition 4. *Given a path π_t and measurable function f, we denote $\|f\|_\pi = \sup_t \mathbb{E}_{\pi_t}[f]$.*

Following the computation in Predescu et al. (2004, Equation (6)), we have

$$r(t,t') = 1 - \frac{\mathbb{E}[\exp(-\frac{1}{2}|A_{t,t'}(\tilde{X}_{1/2}, \tilde{X}'_{1/2})|]}{\mathbb{E}[\exp(-\frac{1}{2}A_{t,t'}(\tilde{X}_{1/2}, \tilde{X}'_{1/2}))]}, \tag{20}$$

where $\tilde{X}_s, \tilde{X}_s' \sim \tilde{\pi}_s = \frac{1}{\tilde{Z}(s)}\exp((1-s)W_t + sW_{t'})$ and

$$A_{t,t'}(x,x') = (W_{t'}(x) - W_t(x)) - (W_{t'}(x') - W_t(x')).$$

In particular, the path of distributions $\tilde{\pi}_s$ for $s \in [0,1]$ is the linear path between π_t and $\pi_{t'}$.

Lemma 2. *Suppose (6) and (7) hold. Then for all $k \le 3$, there is a constant $\tilde{C}_k$ independent of $t, t', \mathcal{T}_N$ such that*

$$\sup_s \mathbb{E}[|A_{t,t'}(\tilde{X}_s, \tilde{X}_s')|^k] \le \tilde{C}_k|t'-t|^k.$$

where $\tilde{X}_s, \tilde{X}_s' \sim \tilde{\pi}_s$.

Proof. The mean-value theorem and (6) imply that $W_t(x)$ is Lipschitz in t,

$$|W_t(x) - W_{t'}(x)| \le V_1(x)|t'-t|.$$

The triangle inequality therefore implies

$$|A_{t,t'}(x,x')| \le (V_1(x) + V_1(x'))|t'-t|.$$

By taking expectations and using the fact $|a+b|^k \le 2^{k-1}(|a|^k + |b|^k)$, we have that

$$\begin{aligned} \mathbb{E}[|A_{t,t'}(\tilde{X}_s, \tilde{X}_s')|^k] &\le 2^k \mathbb{E}_{\tilde{\pi}_s}[V_1^k]|t'-t|^k \\ &\le 2^k \mathbb{E}_{\tilde{\pi}_s}[V_1^3]|t'-t|^k, \end{aligned}$$

where in the last line we use the fact that we can assume $V_1 \ge 1$ without loss of generality. The result follows by taking the supremum on both sides and noting that $\tilde{C}_k = 2^k\|V_1^3\|_{\tilde{\pi}}$ is finite by (7). □

We now begin the proof of Lemma 1. Define $\tilde{\lambda}(s) = \frac{1}{2}\mathbb{E}[|A_{t,t'}(\tilde{X}_s, \tilde{X}'_s)|]$ for $\tilde{X}_s, \tilde{X}'_s \sim \tilde{\pi}_s$. Then a third order Taylor expansion of Equation (20) (Predescu et al., 2004), which contains terms of the form $\mathbb{E}[|A_{t,t'}(\tilde{X}_s, \tilde{X}'_s)|^k]$ that can be controlled via Lemma 2, yields

$$r(t,t') = \tilde{\lambda}(1/2) + R(t,t'), \quad |R(t,t')| \leq C'|t-t'|^3,$$

for some finite constant C' independent of t, t'. By Syed et al. (2019, Prop. 2, Appendix C) we have that in addition, $\tilde{\lambda}(s)$ is in $C^2([0,1])$, and thus there is a constant C'' independent of t, t' such that

$$\sup_s \left|\frac{d^2\tilde{\lambda}}{ds^2}\right| \leq C''|t-t'|^3.$$

The error bound for the midpoint rule implies,

$$\begin{aligned}\left|\tilde{\lambda}(1/2) - \int_0^1 \tilde{\lambda}(s)ds\right| &\leq \frac{1}{24}\sup_s \left|\frac{d^2\tilde{\lambda}}{ds^2}\right| \\ &\leq \frac{C''}{24}|t-t'|^3.\end{aligned}$$

The result follows: there is a finite constant C independent of t, t' such that

$$|r(t,t') - \Lambda(t,t')| = \left|r(t,t') - \int_0^1 \tilde{\lambda}(s)ds\right| \leq C|t'-t|^3.$$

C. Proof of Theorem 2

We first note that without loss of generality we can place an artificial schedule point t_n at each of the finitely many discontinuities in W_t or its first/second derivative. Thus we assume the W_t is C^2 on each interval $[t_{n-1}, t_n]$. Later in the proof it will become clear that the contributions of these artificial schedule points becomes negligible as $\|\mathcal{T}_N\| \to 0$.

Given a schedule $\mathcal{T}_N$, define the path $\tilde{\pi}_t = \frac{1}{\tilde{Z}_t}\exp(\tilde{W}_t)$ with log-likelihood $\tilde{W}_t$ satisfying for each segment $t_{n-1} \leq t \leq t_n$,

$$\tilde{W}_t = W_{t_{n-1}} + \frac{\Delta W_n}{\Delta t_n}(t - t_{n-1}),$$

where $\Delta W_n = W_{t_n} - W_{t_{n-1}}$ and $\Delta t_n = t_n - t_{n-1}$. In particular, $\tilde{W}_t$ agrees with W_t for $t \in \mathcal{T}_N$, linearly interpolates between $W_{t_{n-1}}$ and W_{t_n} for $t \in [t_{n-1}, t_n]$, and for all x, from Taylor's theorem:

$$|\tilde{W}_t(x) - W_t(x)| \leq \frac{1}{2}\sup_{t\in[t_{n-1},t_n]}\left|\frac{d^2W_t}{dt^2}(x)\right|\Delta t_n^2. \quad (21)$$

The following lemma shows that the normalization constant of, and expectations under, $\tilde{\pi}_t$ are comparable to the same for π_t with an error bound that depends on $\|\mathcal{T}_N\|$ and converges to 0 as $\|\mathcal{T}_N\| \to 0$.

Lemma 3. *For measurable functions f and $s > 0$, let*

$$E_t(f,s) = \mathbb{E}_{\pi_t}\left[|f|e^{s^2V_2}\right],$$

and define $E_t(s) = E_t(1,s)$ for brevity.

(a) For any schedule $\mathcal{T}_N$,

$$\left|\frac{\tilde{Z}_t}{Z_t} - 1\right| \leq E_t(\|\mathcal{T}_N\|) - 1,$$

and if $\|\mathcal{T}_N\|$ is small enough that $E_t(\|\mathcal{T}_N\|) < 2$,

$$\left|\frac{Z_t}{\tilde{Z}_t} - 1\right| \leq \frac{E_t(\|\mathcal{T}_N\|) - 1}{2 - E_t(\|\mathcal{T}_N\|)}.$$

(b) For any schedule $\mathcal{T}_N$ and measureable function f, if $\|\mathcal{T}_N\|$ is small enough that $E_t(\|\mathcal{T}_N\|) < 2$,

$$\begin{aligned}|\mathbb{E}_{\tilde{\pi}_t}[f] - \mathbb{E}_{\pi_t}[f]| \leq \frac{E_t(\|\mathcal{T}_N\|) - 1}{2 - E_t(\|\mathcal{T}_N\|)}E_t(f, \|\mathcal{T}_N\|) \\ + E_t(f, \|\mathcal{T}_N\|) - E_t(f, 0).\end{aligned}$$

Proof. (a) We rewrite the expression

$$\begin{aligned}\frac{\tilde{Z}_t}{Z_t} &= \frac{1}{Z_t}\int_{\mathcal{X}} e^{\tilde{W}_t(x)}\mathrm{d}x \\ &= \int_{\mathcal{X}} e^{\tilde{W}_t(x) - W_t(x)}\pi_t(x)\mathrm{d}x \\ &= 1 + \int_{\mathcal{X}}\left(e^{\tilde{W}_t(x) - W_t(x)} - 1\right)\pi_t(x)\mathrm{d}x.\end{aligned}$$

Thus using the inequality $|e^x - 1| \leq e^{|x|} - 1$,

$$\begin{aligned}\left|\frac{\tilde{Z}_t}{Z_t} - 1\right| &\leq \left|\int_{\mathcal{X}}\left(e^{\tilde{W}_t(x) - W_t(x)} - 1\right)\pi_t(x)\mathrm{d}x\right| \\ &\leq \int_{\mathcal{X}}\left(e^{|\tilde{W}_t(x) - W_t(x)|} - 1\right)\pi_t(x)\mathrm{d}x \\ &\leq \int_{\mathcal{X}}\left(e^{V_2(x)\|\mathcal{T}_N\|^2} - 1\right)\pi_t(x)\mathrm{d}x \\ &= \mathbb{E}_{\pi_t}\left[e^{\|\mathcal{T}_N\|^2V_2} - 1\right] \\ &= E_t(\|\mathcal{T}_N\|) - 1.\end{aligned}$$

The bound on $|Z_t/\tilde{Z}_t - 1|$ arises from straightforward algebraic manipulation of the above bound.

(b) We begin by rewriting $\mathbb{E}_{\tilde{\pi}_t}[f]$:

$$\begin{aligned}
&\mathbb{E}_{\tilde{\pi}_t}[f] - \mathbb{E}_{\pi_t}[f] \\
&= \frac{1}{\tilde{Z}_t}\int_{\mathcal{X}} f(x)e^{\tilde{W}_t(x)}\mathrm{d}x - \mathbb{E}_{\pi_t}[f] \\
&= \int_{\mathcal{X}} f(x)\left(\frac{Z_t}{\tilde{Z}_t}e^{\tilde{W}_t(x)-W_t(x)} - 1\right)\pi_t(x)\mathrm{d}x \\
&= \left(\frac{Z_t}{\tilde{Z}_t} - 1\right)\int_{\mathcal{X}} f(x)e^{\tilde{W}_t(x)-W_t(x)}\pi_t(x)\mathrm{d}x \\
&+ \int_{\mathcal{X}} f(x)\left(e^{\tilde{W}_t(x)-W_t(x)} - 1\right)\pi_t(x)\mathrm{d}x.
\end{aligned}$$

Therefore again using $|e^x - 1| \leq e^{|x|} - 1$ and the previous bound,

$$\begin{aligned}
|\mathbb{E}_{\tilde{\pi}_t}[f] - \mathbb{E}_{\pi_t}[f]| \leq \frac{E_t(\|\mathcal{T}_N\|) - 1}{2 - E_t(\|\mathcal{T}_N\|)}E_t(f, \|\mathcal{T}_N\|) \\
+ E_t(f, \|\mathcal{T}_N\|) - E_t(f, 0).
\end{aligned}$$

□

By changing variables via $t = t_{n-1} + s\Delta t_n$ in (5), we can rewrite $\Lambda(t_{n-1}, t_n)$ as

$$\Lambda(t_{n-1}, t_n) = \int_{t_{n-1}}^{t_n} \frac{1}{2}\mathbb{E}\left[\left|\frac{\Delta W_n}{\Delta t_n}(\tilde{X}_t) - \frac{\Delta W_n}{\Delta t_n}(\tilde{X}'_t)\right|\right]\mathrm{d}t,$$

where $\tilde{X}_t, \tilde{X}'_t \sim \tilde{\pi}_t$. Note that by construction for $t \in (t_{n-1}, t_n)$ we have $\frac{d\tilde{W}_t}{dt}$ exists and equals $\frac{\Delta W_n}{\Delta t_n}$. So by summing over n we get,

$$\begin{aligned}
\Lambda(\mathcal{T}_N) &= \sum_{n=1}^{N}\Lambda(t_{n-1}, t_n) \\
&= \int_0^1 \frac{1}{2}\mathbb{E}\left[\left|\frac{d\tilde{W}_t}{dt}(\tilde{X}_t) - \frac{d\tilde{W}_t}{dt}(\tilde{X}'_t)\right|\right]dt \\
&= \int_0^1 \tilde{\lambda}(t)dt
\end{aligned}$$

If we can show that $\sup_t |\tilde{\lambda}(t) - \lambda(t)|$ converges uniformly[4] to 0 as $\|\mathcal{T}_N\| \to 0$ then by dominated convergence theorem $\Lambda(\mathcal{T}_N)$ converges to Λ uniformly as $\|\mathcal{T}_N\| \to 0$. The round trip rate then uniformly converges to $(2+2\Lambda)^{-1}$ by Theorem 3 of (Syed et al., 2019).

Adding and subtracting $\mathbb{E}\left[\left|\frac{d\tilde{W}_t}{dt}(X_t) - \frac{d\tilde{W}_t}{dt}(X'_t)\right|\right]$ within the absolute difference $2|\tilde{\lambda}(t) - \lambda(t)|$ and using the triangle inequality, it can be shown that we require bounds on

$$J_{1,t} = \int \pi_t(x)\pi_t(y)\left|\left|\frac{d\tilde{W}_t}{dt}(x) - \frac{d\tilde{W}_t}{dt}(y)\right| - \left|\frac{dW_t}{dt}(x) - \frac{dW_t}{dt}(y)\right|\right|$$

and

$$J_{2,t} = \int |\pi_t(x)\pi_t(y) - \tilde{\pi}_t(x)\tilde{\pi}_t(y)|\left|\frac{d\tilde{W}_t}{dt}(x) - \frac{d\tilde{W}_t}{dt}(y)\right|.$$

For the first term, the mean value theorem implies that there exist $s, s' \in [t_{n-1}, t_n]$ (potentially functions of x and y, respectively) such that

$$J_{1,t} = \int \pi_t(x)\pi_t(y)\left|\left|\frac{dW_s}{dt}(x) - \frac{dW_{s'}}{dt}(y)\right| - \left|\frac{dW_t}{dt}(x) - \frac{dW_t}{dt}(y)\right|\right|$$

Split the integral into the set A of $x, y \in \mathcal{X}$ where the first term in the absolute value is larger; the same analysis with the same result applies in the other case in A^c. Here, Taylor's theorem and the triangle inequality yield

$$\begin{aligned}
\left|\frac{dW_s}{dt}(x) - \frac{dW_{s'}}{dt}(y)\right| \leq \left|\frac{dW_t}{dt}(x) - \frac{dW_t}{dt}(y)\right| \\
+ (V_2(x) + V_2(y))\|\mathcal{T}_N\|.
\end{aligned}$$

Using this and the same procedure for A^c, we have that

$$\begin{aligned}
J_{1,t} &\leq \int \pi_t(x)\pi_t(y)(V_2(x) + V_2(y))\|\mathcal{T}_N\| \\
&= 2\mathbb{E}_{\pi_t}[V_2]\,\|\mathcal{T}_N\|.
\end{aligned}$$

This converges to 0 as $\|\mathcal{T}_N\| \to 0$.

For the second term $J_{2,t}$, we can again use the mean value theorem to find $s, s' \in [t_{n-1}, t_n]$ where

$$J_{2,t} = \int |\pi_t(x)\pi_t(y) - \tilde{\pi}_t(x)\tilde{\pi}_t(y)|\left|\frac{dW_s}{dt}(x) - \frac{dW_{s'}}{dt}(y)\right|,$$

and therefore via the triangle inequality, symmetry, and the $V_1(x)$ bound on the first path derivative,

$$J_{2,t} \leq 2\int V_1(x)\,|\pi_t(x)\pi_t(y) - \tilde{\pi}_t(x)\tilde{\pi}_t(y)|\,.$$

We then add and subtract $\pi_t(x)\tilde{\pi}_t(y)$ within the absolute value and use the triangle inequality again to find that

$$\begin{aligned}
J_{2,t} &\leq 2\int (V_1(x) + \mathbb{E}_{\pi_t}[V_1])\,|\pi_t(x) - \tilde{\pi}_t(x)| \\
&= 2\int \pi_t(x)\,(V_1(x) + \mathbb{E}_{\pi_t}[V_1])\left|1 - \frac{\tilde{\pi}_t(x)}{\pi_t(x)}\right|.
\end{aligned}$$

Note that by the triangle inequality and the bound $|e^x - 1| \leq e^{|x|} - 1$,

$$\left|1 - \frac{\tilde{\pi}_t(x)}{\pi_t(x)}\right| \leq \left|\frac{Z_t}{\tilde{Z}_t} - 1\right|e^{\|\mathcal{T}_N\|^2 V_2(x)} + e^{\|\mathcal{T}_N\|^2 V_2(x)} - 1.$$

Assume that $\|\mathcal{T}_N\|$ is small enough such that $E_t(\|\mathcal{T}_N\|) < 2$, and let $f = V_1 + \mathbb{E}_{\pi_t}V_1$. Then by Lemma 3,

$$\begin{aligned}
J_{2,t} \leq 2\frac{E_t(\|\mathcal{T}_N\|) - 1}{2 - E_t(\|\mathcal{T}_N\|)}E_t(f, \|\mathcal{T}_N\|) \\
+ E_t(f, \|\mathcal{T}_N\|) - E_t(f, 0).
\end{aligned}$$

[4] We say $a(\mathcal{T}_N)$ converges uniformly to a if for all $\epsilon > 0, \exists\delta > 0$ such that $\|\mathcal{T}_N\| < \delta$ implies $|a(\mathcal{T}_N) - a| < \epsilon$.

By assumption we know that $E_t(f, s)$ is finite for some s small enough. Therefore as $\|\mathcal{T}_N\| \to 0$, by monotone convergence $E_t(f, \|\mathcal{T}_N\|) \to E_t(f, 0)$, and in particular $E_t(\|\mathcal{T}_N\|) \to 1$. Therefore $J_{1,t} + J_{2,t} \to 0$ as $\|\mathcal{T}_N\| \to 0$ and the proof is complete.

D. Objective and Gradient

Here we derive the gradient used to optimize the surrogate SKL objective in Equation 17. First we derive the gradient for the expectation of a general function in Section D.1. Next, in Section D.2, we show the result for the specific case of expectations of linear functions with respect to distributions in the exponential family. Lastly, we show how the result is related to our SKL objective in Sections D.3 and D.4.

D.1. Derivative of parameter-dependent expectation

Here we consider the problem of computing

$$g_\phi(x) = \nabla_\phi \int_{\mathcal{X}} \pi_\phi(x) J_\phi(x) \mathrm{d}x$$

where $\pi_\phi(x) = Z(\phi)^{-1} \exp(W_\phi(x))$, $Z(\phi) = \int_{\mathcal{X}} \exp(W_\phi(x)) \mathrm{d}x$ and $J_\phi(x)$ is a function depending on ϕ. Assuming we can interchange the gradient and the expectation and using the product rule we can rewrite:

$$g_\phi(x) = \int_{\mathcal{X}} \left(J_\phi(x) \nabla_\phi \pi_\phi(x) + \pi_\phi(x) \nabla_\phi J_\phi(x)\right) \mathrm{d}x.$$

Using $\nabla_\phi \pi_\phi(x) = \pi_\phi(x) \nabla_\phi \log \pi_\phi(x)$,

$$g_\phi(x) = \int_{\mathcal{X}} \pi_\phi(x) (J_\phi(x) \nabla_\phi \log \pi_\phi(x) + \nabla_\phi J_\phi(x)) \mathrm{d}x.$$

From the definition of $\pi_\phi(x)$, we can evaluate the score function as

$$\begin{aligned} \nabla_\phi \log \pi_\phi(x) &= -\nabla_\phi \log Z(\phi) + \nabla_\phi W_\phi(x) \\ &= -\mathbb{E}\left[\nabla_\phi W_\phi(x)\right] + \nabla_\phi W_\phi(x). \end{aligned}$$

Substitute this in $g_\phi(x)$ we obtain,

$$\begin{aligned} g_\phi(x) &= \int_{\mathcal{X}} \pi_\phi(x) J_\phi(x) (-\mathbb{E}\left[\nabla_\phi W_\phi(x)\right] + \nabla_\phi W_\phi(x)) \mathrm{d}x \\ &\quad + \int_{\mathcal{X}} \pi_\phi(x) \nabla_\phi J_\phi(x) \mathrm{d}x \\ &= -\mathbb{E}[J_\phi(x)] \mathbb{E}[\nabla_\phi W_\phi(x)] + \mathbb{E}[J_\phi(x) \nabla_\phi W_\phi(x)] \\ &\quad + \mathbb{E}[\nabla_\phi J_\phi(x)] \\ &= \mathrm{Cov}[\nabla_\phi W_\phi(x), J_\phi(x)] + \mathbb{E}[\nabla_\phi J_\phi(x)]. \end{aligned}$$

D.2. Exponential family and linear function

The gradient derived in the previous section can easily be applied to expectations with respect to functions linear in ϕ under distributions in the exponential family. Let $J_\phi(x) = \xi_J(\phi)^T J(x)$ be a linear function in ϕ and suppose $W_\phi(x) = \xi_W(\phi)^T W(x)$ for some functions $\xi_J : \mathbb{R}^d \to \mathbb{R}^n$, $J : \mathcal{X} \to \mathbb{R}^n$ and $\xi_W : \mathbb{R}^d \to \mathbb{R}^m$, $W : \mathcal{X} \to \mathbb{R}^m$. Then

$$\begin{aligned} g_\phi(x) &= \mathrm{Cov}[\nabla_\phi W_\phi(x), J_\phi(x)] + \mathbb{E}[\nabla_\phi J_\phi(x)] \\ &= \nabla_\phi \xi_W(\phi)^T \mathrm{Cov}[W(x), J^T(x)] \xi_J(\phi) \\ &\quad + \nabla_\phi \xi_J(\phi)^T \mathbb{E}[J(x)] \end{aligned}$$

where $\nabla_\phi \xi(\phi)^T$ is the transposed Jacobian of ξ.

D.3. Symmetric KL: general case

Next we show that the symmetric KL divergence of Equation 17 can be rewritten as a sum of expectations over functions parametrized by ϕ, hence falling in the framework presented above.

For path parameter ϕ, the symmetric KL divergence is

$$\begin{aligned} \mathcal{L}_{\mathrm{SKL}}(\phi) &= \sum_{n=0}^{N-1} \mathrm{SKL}(\pi^\phi_{t_n}, \pi^\phi_{t_{n+1}}) \\ &= \sum_{n=0}^{N-1} \mathbb{E}\left[\log \frac{\pi^\phi_{t_{n+1}}(X_{n+1})}{\pi^\phi_{t_n}(X_{n+1})} + \log \frac{\pi^\phi_{t_n}(X_n)}{\pi^\phi_{t_{n+1}}(X_n)}\right] \end{aligned}$$

where $X_n \sim \pi^\phi_{t_n}$. After cancellation of the normalization constants we obtain

$$\begin{aligned} &\mathcal{L}_{\mathrm{SKL}}(\phi) = \\ &\sum_{n=0}^{N-1} \mathbb{E}\big[W^\phi_{t_{n+1}}(X_{n+1}) - W^\phi_{t_n}(X_{n+1}) \\ &+ W^\phi_{t_n}(X_n) - W^\phi_{t_{n+1}}(X_n)\big]. \end{aligned}$$

Collecting expectations under the same distribution and rearranging terms,

$$\begin{aligned} &\mathcal{L}_{\mathrm{SKL}}(\phi) = \\ &\mathbb{E}[W^\phi_{t_0}(X_0) - W^\phi_{t_1}(X_0)] + \\ &\sum_{n=1}^{N-1} \mathbb{E}[2W^\phi_{t_n}(X_n) - W^\phi_{t_{n+1}}(X_n) - W^\phi_{t_{n-1}}(X_n)] + \\ &\mathbb{E}[W^\phi_{t_N}(X_N) - W^\phi_{t_{N-1}}(X_N)]. \end{aligned}$$

Defining for $n = 1, \ldots, N-1$,

$$\begin{aligned} J^\phi_0(x) &= W^\phi_{t_0}(x) - W^\phi_{t_1}(x) \\ J^\phi_n(x) &= 2W^\phi_{t_n}(x) - W^\phi_{t_{n+1}}(x) - W^\phi_{t_{n-1}}(x) \\ J^\phi_N(x) &= W^\phi_{t_N}(x) - W^\phi_{t_{N-1}}(x), \end{aligned}$$

we have that

$$\mathcal{L}_{\mathrm{SKL}}(\phi) = \sum_{n=0}^{N} \mathbb{E}[J^\phi_n(X_n)]$$

and

$$\nabla_\phi \mathcal{L}_{\mathrm{SKL}}(\phi) = \sum_{n=0}^{N} \nabla_\phi \mathbb{E}[J_n^\phi(X_n)]$$

where $\nabla_\phi \mathbb{E}[J_n^\phi(X_n)]$ can be computed using the formula derived in Section D.1.

D.4. Symmetric KL: exponential family case

For the spline family introduce in Section 4, the distributions $\pi_{t_n}^\phi$ are in the exponential family with,

$$W_{t_n}^\phi(x) = \eta^\phi(t_n)^T W(x), \quad n = 0, \dots, N.$$

It follows that the functions J_n^ϕ are linear in ϕ with

$$\begin{aligned} J_0^\phi(x) &= {z_0^\phi}^T W(x) \\ J_n^\phi(x) &= {z_n^\phi}^T W(x), \quad n = 1, \dots, N-1 \\ J_N^\phi(x) &= {z_N^\phi}^T W(x), \end{aligned}$$

where

$$\begin{aligned} z_0^\phi &= \eta^\phi(t_0) - \eta^\phi(t_1) \\ z_n^\phi &= 2\eta^\phi(t_n) - \eta^\phi(t_{n+1}) - \eta^\phi(t_{n-1}), \quad n = 1, \dots, N-1 \\ z_N^\phi &= \eta^\phi(t_N) - \eta^\phi(t_{N-1}). \end{aligned}$$

Given this relation, the stochastic gradient of Equation 17 can be evaluated using s samples from parallel tempering through the formula in Section D.2 defining:

$$\begin{aligned} X &= (X_0, \dots, X_N) \\ W(X) &= [W_0(X_0), W_1(X_0), \dots, W_0(X_N), W_1(X_N)]^T \\ J(X) &= W(X) \\ \xi_W(\phi) &= [\eta_0^\phi(t_0), \eta_1^\phi(t_0), \dots, \eta_0^\phi(t_N), \eta_1^\phi(t_N)]^T \\ \xi_J(\phi) &= [z_{0,0}^\phi, z_{0,1}^\phi, \dots, z_{N,0}^\phi, z_{N,1}^\phi]^T \end{aligned}$$

where X is the $s \times N$ matrix of samples from parallel tempering, $W(X)$ is a $s \times 2N$ matrix evaluating X elementwise at the reference and target distributions W_0 and W_1, $\xi_W(\phi)$ is a $2N \times 1$ vector of annealing coefficients and $\xi_J(\phi)$ is a $2N \times 1$ vector of coefficients defining $J^\phi = [J_0^\phi, \dots, J_N^\phi]$.

E. Proof of proposition 2

For this annealing path family,

$$W_t(x) = \eta(t)^T W(x).$$

Therefore, the piecewise twice continuous differentiability of $\eta(t)$ and endpoint conditions imply that Definition 1 is satisfied. Next, note that if

$$\sup_t \max\{\|\eta'(t)\|_2, \|\eta''(t)\|_2\} \leq M,$$

then

$$\begin{aligned} \left|\frac{\mathrm{d}W_t}{\mathrm{d}t}\right| &= |\eta'(t)^T W(x)| \leq M\|W(x)\|_2 \\ \left|\frac{\mathrm{d}^2 W_t}{\mathrm{d}t^2}\right| &= |\eta''(t)^T W(x)| \leq M\|W(x)\|_2, \end{aligned}$$

and thus by setting $V_1(x) = V_2(x) = M\|W(x)\|_2$ we satisfy Equations (6) and (10). Equation (11) implies Equation (7); so as long as Equation (11) holds, the path η satisfies all of the conditions of Theorem 2.

Finally, note that Ω is a convex subset of $\mathbb{R}^2$: for any non-negative function $G(x)$, vectors $\xi_1, \xi_2 \in \mathbb{R}^2$, and $\lambda \in [0, 1]$,

$$\begin{aligned} &\exp((\lambda\xi_1 + (1-\lambda)\xi_2)^T W(x))G(x) \\ &= \left(\exp(\xi_1^T W)G(x)\right)^\lambda \left(\exp(\xi_1^T W)G(x)\right)^{1-\lambda} \end{aligned}$$

and so Hölder's inequality $\int f^\lambda g^{1-\lambda} \leq (\int f)^\lambda (\int g)^{1-\lambda}$ yields log-convexity (and hence convexity). Therefore as long as the endpoints $(0, 1)$ and $(1, 0)$ are both in Ω, any convex combination of $(0, 1)$ and $(1, 0)$ is also in Ω, and therefore the linear path $\eta(t) = (1-t, t)$ creates a set of normalizable densities and may be included in $\mathcal{A}$.

F. Empirical support for the SKL surrogate objective function

Two objective functions were discussed in Section 3: one based on rejection rate statistics, i.e. Equation (14), and the symmetric KL divergence (SKL). In this section we perform controlled experiments comparing the signal-to-noise ratio of Monte Carlo estimators of the gradient of these two objectives. Let G denote a Monte Carlo estimator of a partial derivative with respect to one of the parameters ϕ_i. Refer to D for details on the stochastic gradient estimators. In this experiment we use i.i.d. samples so that the Monte Carlo estimators are unbiased, justifying the use of the variance as a notion of noise. Hence following Rainforth et al. (2018), we define the signal-to-noise ratio by $\mathrm{SNR} = |\mathbb{E}[G]/\sigma[G]|$, where $\sigma[G]$ denotes the standard deviation of G. We use two chains with one set to a standard Gaussian, the other to a Gaussian with mean ϕ and unit variance. We show the value of the two objective functions in Figure 5 (left). The label "Rejection" refers to the expected rejection of the swap proposal between the two chains, r. We also show the square root of half of the SKL ("SqrtHalfSKL"), to quantify the tightness of the bound in Equation (17), while "Ineff" shows the rejection odds, $r/(1-r)$, called inefficiency in Syed et al. (2019).

Signal-to-noise ratio estimates were computed for each parameter $\phi_i \in \{0, 1/5, 2/5, \dots, 2\}$. Each gradient estimate uses 50 samples, and to approximate the signal-to-noise ratio, the estimation was repeated 1000 times for each ϕ_i

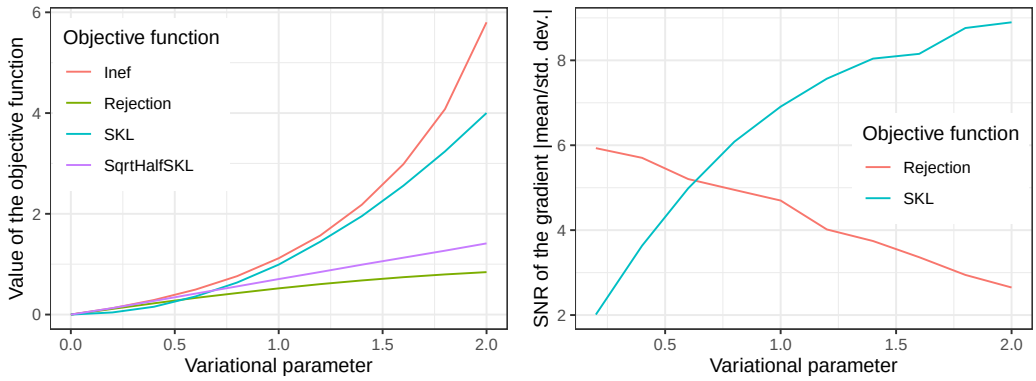

Figure 5. Left: objective functions for path optimization in a controlled experiment as a function of a variational parameter ϕ. Right: signal-to-noise of corresponding gradient estimators on the same range of parameters.

and objective function. The results are shown in Figure 5 (right), and demonstrate that in the regime of small rejection ($\lessapprox 30\%$), the gradient estimator based on the rejection objective has a superior signal-to-noise ratio compared to its SKL counterpart. However as ϕ increases and the two distributions become farther apart, the situation is reversed, providing empirical support for the surrogate objective for challenging path optimization problems.

G. Experimental details

All the experiments were conducted comparing reversible PT, non-reversible PT and non-reversible PT based on the spline family with $K \in \{2, 3, 4, 5, 10\}$.

Every method was initialized at the linear path with equally spaced schedule, i.e. $\pi_t \propto \pi_0^{1-t/N} \pi_1^{t/N}$ with N the number of parallel chains. All methods performed one local exploration step before a communication step.

To ensure a fair comparison of the different algorithms, we fixed the computational budget to a pre-determined number of samples in each experiment. Reversible PT used the budget to perform local exploration steps followed by communication steps. In non-reversible PT the computational budget was used to tune the schedule according to the procedure described in Syed et al. (2019, Section 5.1). For non-reversible PT with path optimization, the computational budget was divided equally over a fixed number of scans of Algorithm 2, where a scan corresponds to one iteration of the for loop.

Optimization of the spline annealing path family was performed using the SKL surrogate objective of Equation 17. Adagrad was used for the optimization. The gradient was scaled elementwise by its absolute value plus the value of the knot component. Such scaling was necessary to limit the gradient in the interval $[-1, 1]$, stabilizing the optimization and avoiding possible exploding gradients due to the transformation to log space.

To mitigate variance in the results due to randomness, we performed 10 runs of each method and averaged the results across the runs.

G.1. Gaussian

This experiment optimized the path between the reference $\pi_0 = N(-1, 0.01^2)$ and the target $\pi_1 = N(1, 0.01^2)$. We used $N = 50$ parallel chains initialized at a state sampled from a standard Gaussian distribution. In this setting, π_t has a closed form that can be shown to be $N\left(\frac{\eta_1(t)-\eta_0(t)}{\eta_0(t)+\eta_1(t)}, \left(\frac{0.01^2}{\eta_0(t)+\eta_1(t)}\right)^2\right)$, therefore, in the local exploration step of parallel tempering we sampled i.i.d. from π_t. The computational budget was fixed at 45000 samples. Non-reversible PT with optimized path divided the budget in 150 scans. Therefore, for every gradient step in Algorithm 2, the gradient was estimated with 300 samples. We used 0.2 as learning rate for Adagrad.

G.2. Beta-binomial model

The second experiment was performed on a conjugate Bayesian model. The model prior was $\pi_0(p) = \text{Beta}(180, 840)$. The likelihood was $L(x|p) = \text{Binomial}(x|n, p)$. We simulated $x_1, \ldots, x_{2000} \sim \text{Binomial}(100, 0.7)$, resulting in a posterior distribution $\pi_1(p) = \text{Beta}(140180, 60840)$. The prior is concentrated at 0.176 with a standard deviation of 0.0119. The posterior distribution is concentrated at 0.697 with a standard deviation of 0.001. We used $N = 50$ parallel chains initialized at 0.5. Also in this experiment it is possible to compute π_t in closed form. Let $S = \sum_{i=1}^{2000} x_i$, $R = 2000 \times 100$ then $\pi_t(p) = \text{Beta}(179\eta_0(t)+(180+S-1)\eta_1(t)+1, 839\eta_0(t)+(840+N-S-1)\eta_1(t)+1)$. Hence, in the local exploration step of parallel tempering we sampled i.i.d. from π_t. The

computational budget was fixed at 45000 samples. Non-reversible PT with optimized path divided the budget in 150 scans. Therefore, for every gradient step in Algorithm 2, the gradient was estimated with 300 samples. We used 0.2 as learning rate for Adagrad.

G.3. Galaxy data

The third experiment was 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 binary cluster label for each data point. We placed a $\mathrm{Dir}(\mathbf{1})$ prior on the proportions, where $\mathbf{1} = (1, 1, 1, 1, 1, 1)$ 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. Mixture proportions were initialized at $1/6$, mean parameters were initialized at 0 and cluster labels were initialized at 0. The local exploration step involved standard Gibbs steps for the means, indicators, and proportions. To improve local mixing, we also included an additional Metropolis-Hastings step for the proportions that approximates a Gibbs step when the indicators are marginalized. We 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.

G.4. Mixture model

The fourth experiment was a Bayesian Gaussian mixture model with mixture proportions w_0, w_1, mixture component densities $N(\mu_i, 10^2)$ for mean parameters μ_0, μ_1, and a binary cluster label for each data point. We placed a $\mathrm{Dir}(1, 1)$ prior on the proportions, and a $N(150, 1)$ prior on each of the two mean parameters. We simulated $n = 1000$ data points from the mixture $0.3N(100, 10^2) + 0.7N(200, 10^2)$. We did not marginalize the cluster indicators, creating a multi-modal posterior over 1004 latent variables. We used $N = 35$ chains. Mixture proportions were initialized at 0.5, mean parameters were initialized at 0 and cluster labels were initialized at 0. The local exploration step involved standard Gibbs steps for the means, indicator variables, and proportions. To improve local mixing, we also included an additional Metropolis-Hastings step for the proportions that approximates a Gibbs step when the indicators are marginalized. The computational budget was fixed at 25000 samples. Non-reversible PT with optimized path divided the budget in 50 scans. Therefore, for every gradient step in Algorithm 2, the gradient was estimated with 500 samples. We used 0.3 as learning rate for Adagrad. Results are shown in Figure 6.

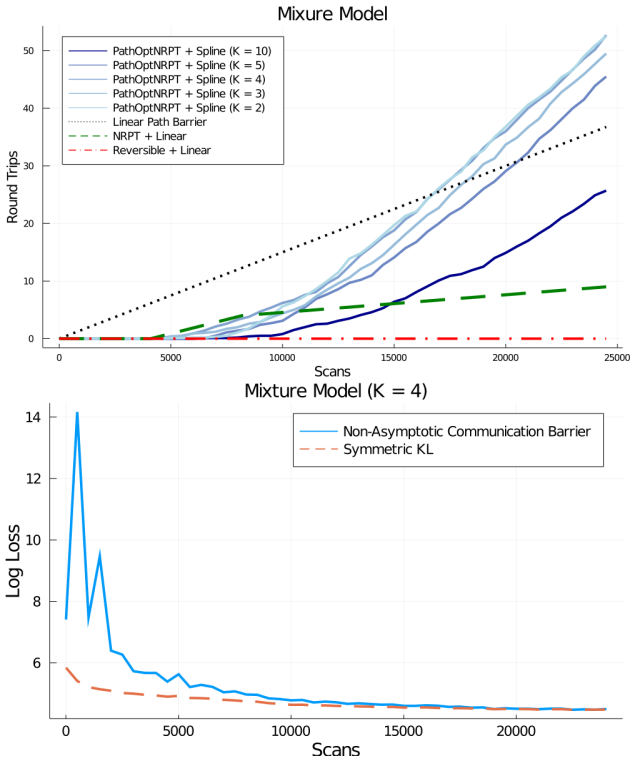

Figure 6. **Top:** Cumulative round trips averaged over 10 runs for the spline path with $K = 2, 3, 4, 5, 10$ (solid blue), NRPT using a linear path (dashed green), and reversible PT with linear path (dash/dot red). The slope of the lines represent the round trip rate. **Bottom:** Non-asymptotic communication 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).