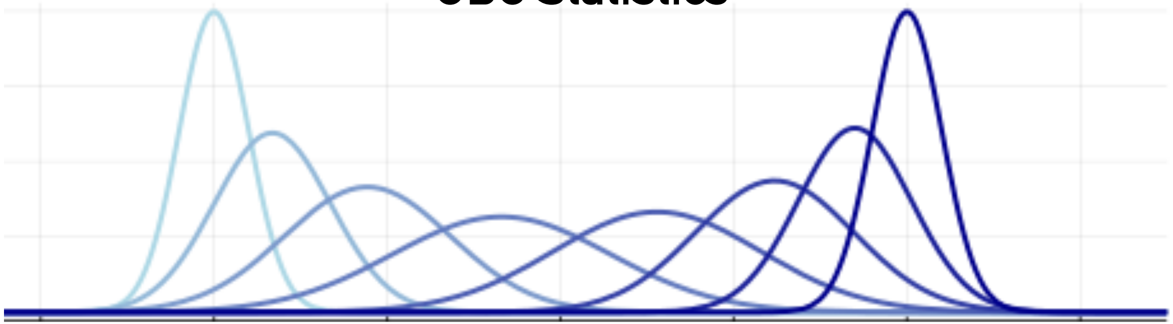
Non-Reversible Parallel Tempering on Optimized Paths

Saifuddin Syed
UBC Statistics



ArXiv:
1905.02939
2102.07720

Motivation

Have some data x, want to infer some unknown parameter θ with posterior

Motivation

Have some data x, want to infer some unknown parameter θ with posterior

$$p(\theta|x) = \frac{1}{p(x)} p(x|\theta)\, p(\theta)$$

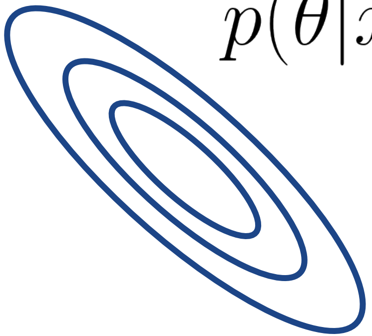
Motivation

Have some data x, want to infer some unknown parameter θ with posterior

$$p(\theta|x) = \frac{1}{p(x)} p(x|\theta)\, p(\theta)$$

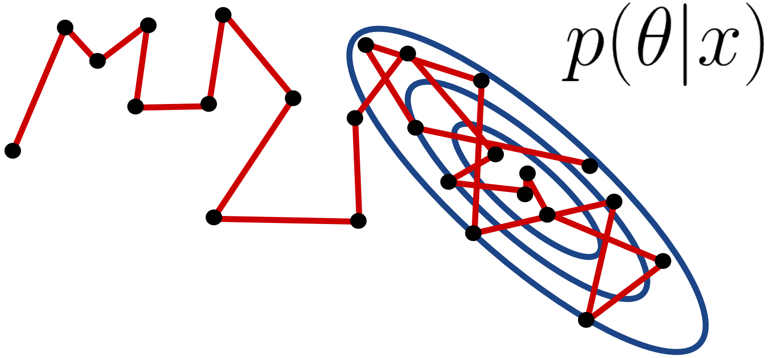
Major Challenge: compute posterior expectations $\mathbb{E}\left[f(\theta)|x\right]$

Markov Chain Monte Carlo



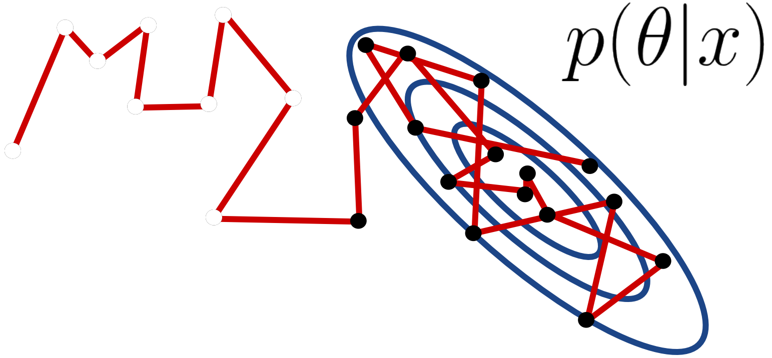
$p(\theta|x)$

Markov Chain Monte Carlo



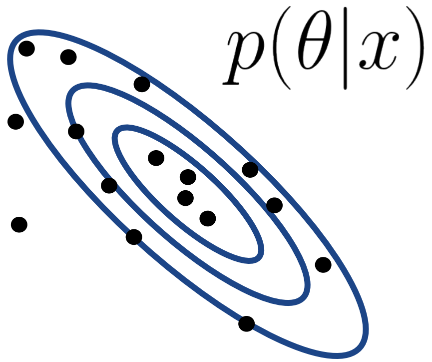
Run a Markov chain whose stationary distribution is the target

Markov Chain Monte Carlo



Run a Markov chain whose stationary distribution is the target

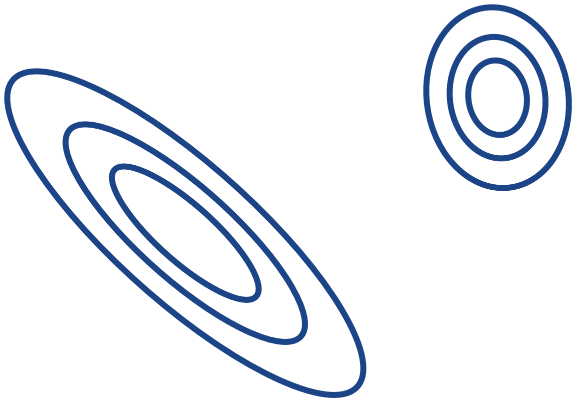
Markov Chain Monte Carlo



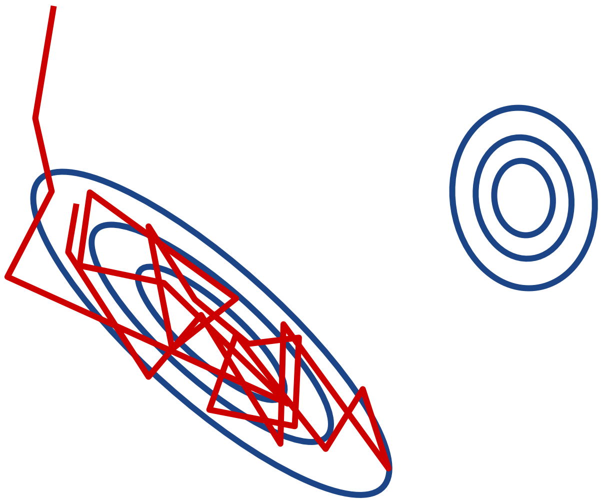
Run a Markov chain whose stationary distribution is the target

Slow convergence to complex, high-dimensional, multi-modal targets

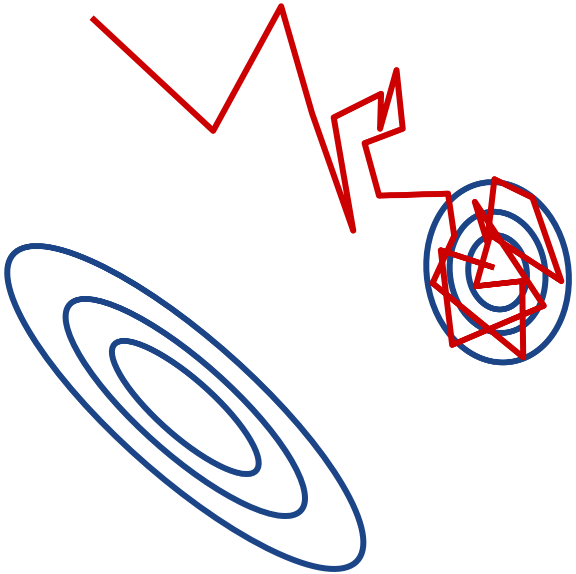
Slow convergence to complex, high-dimensional, multi-modal targets



Slow convergence to complex, high-dimensional, multi-modal targets



Slow convergence to complex, high-dimensional, multi-modal targets

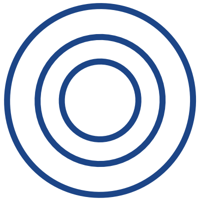


Parallel Tempering

Key Idea: sample from a *path* of distributions, swap states along the path

Parallel Tempering

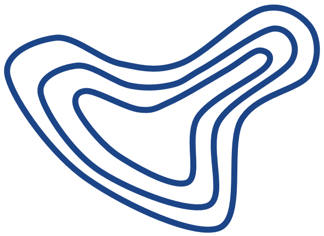
Key Idea: sample from a *path* of distributions, swap states along the path



Parallel Tempering

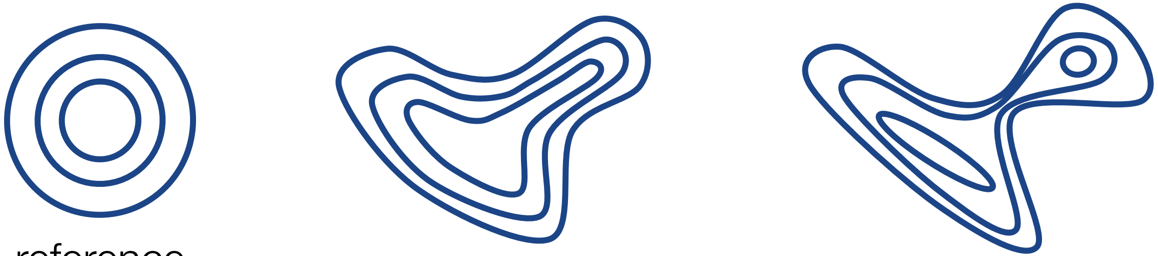
Key Idea: sample from a *path* of distributions, swap states along the path

reference
(eg. Prior)



Parallel Tempering

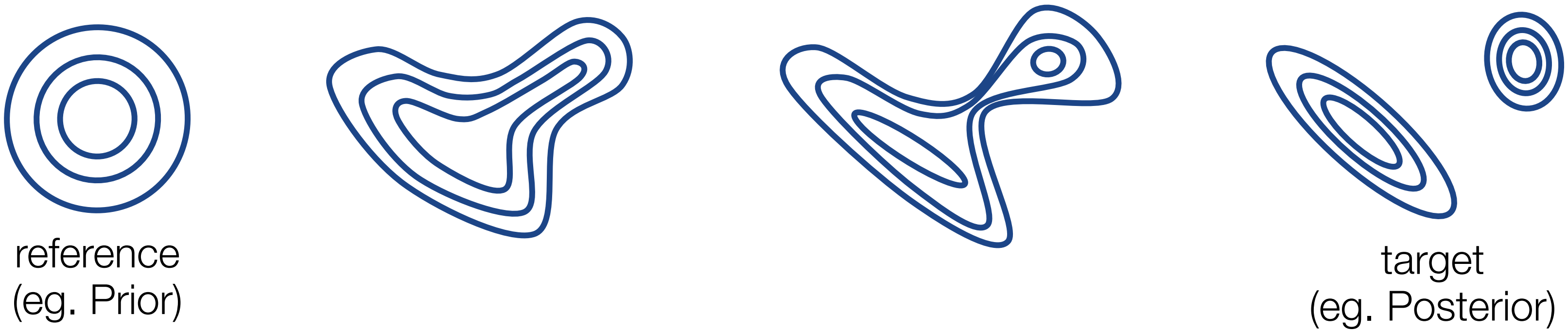
Key Idea: sample from a *path* of distributions, swap states along the path



reference
(eg. Prior)

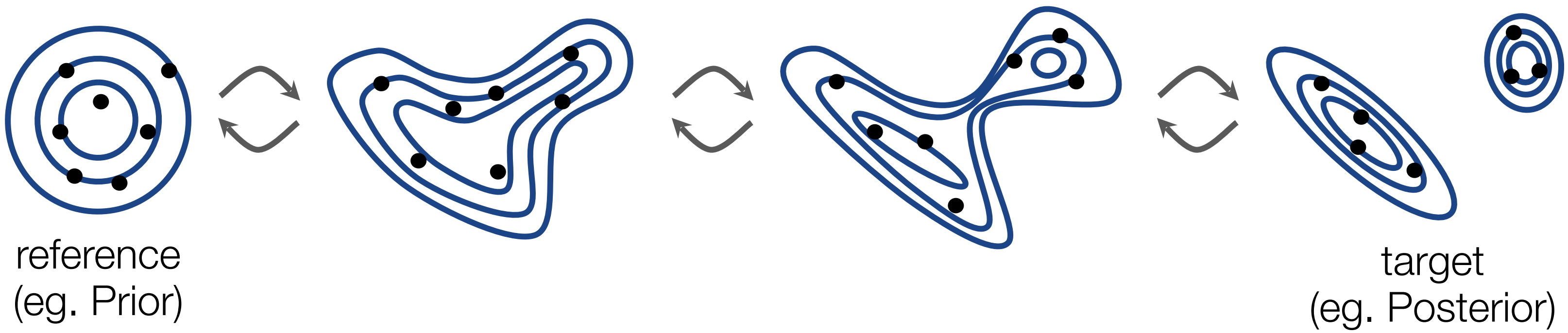
Parallel Tempering

Key Idea: sample from a *path* of distributions, swap states along the path



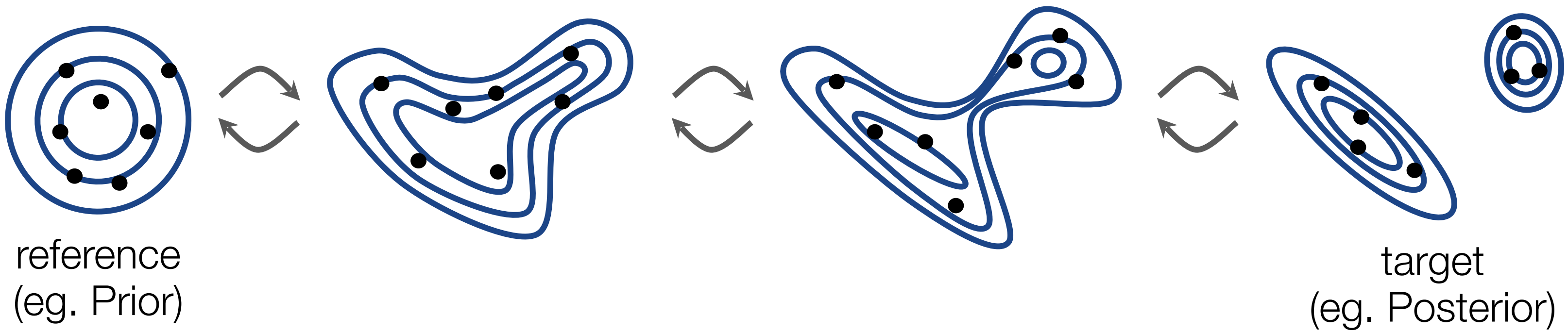
Parallel Tempering

Key Idea: sample from a *path* of distributions, swap states along the path



Parallel Tempering

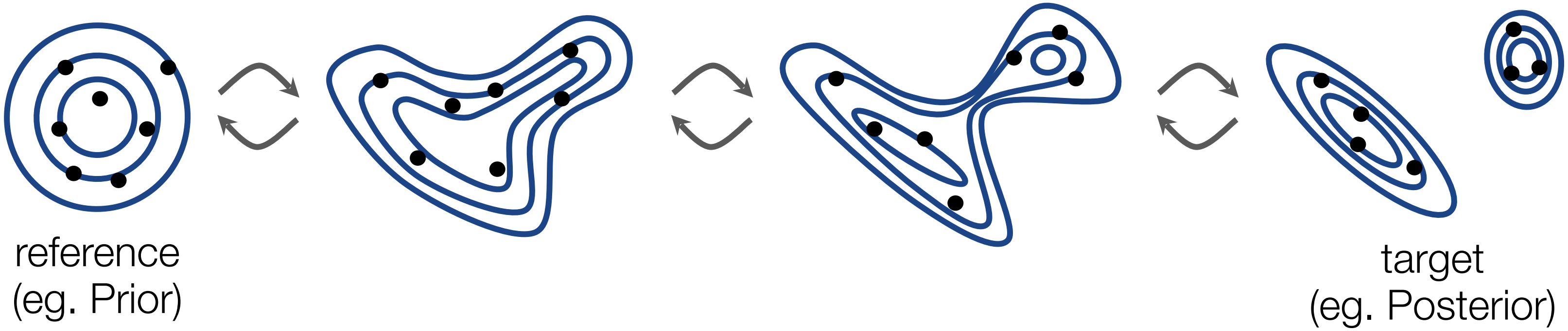
Key Idea: sample from a *path* of distributions, swap states along the path



Annealing path: π_t

Parallel Tempering

Key Idea: sample from a *path* of distributions, swap states along the path

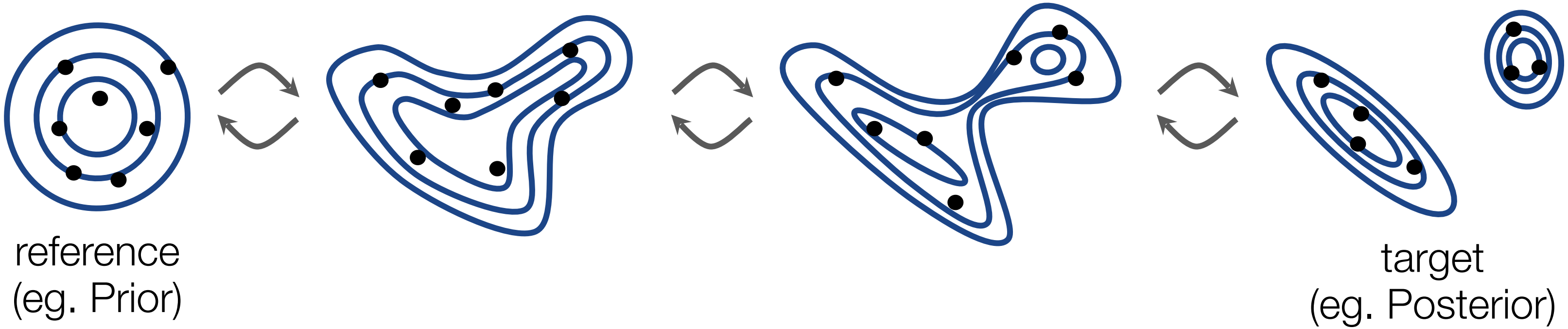


Annealing path: π_t

Schedule: $t_0, \ldots, t_N \in [0, 1]$

Parallel Tempering

Key Idea: sample from a *path* of distributions, swap states along the path



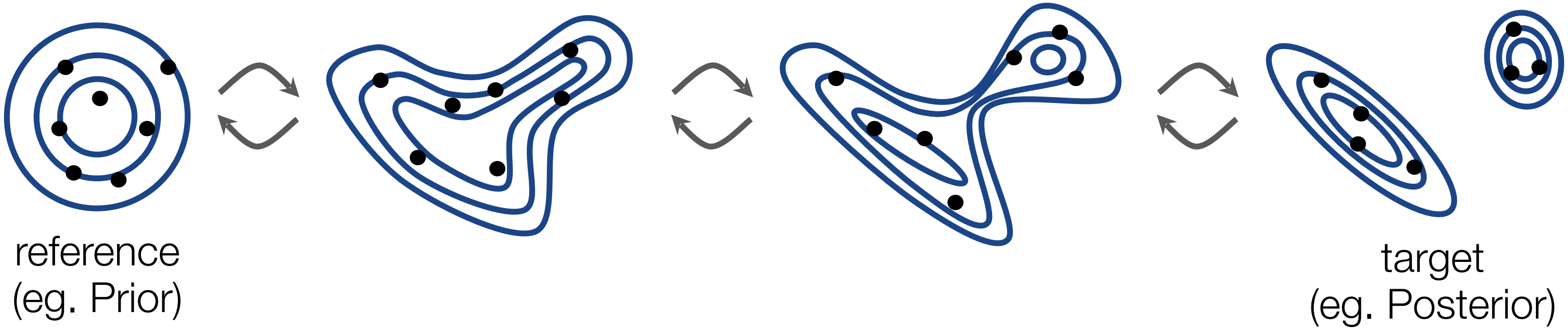
Annealing path: π_t

Schedule: $t_0, \ldots, t_N \in [0, 1]$

Reference: π_0

Parallel Tempering

Key Idea: sample from a *path* of distributions, swap states along the path



Annealing path: π_t

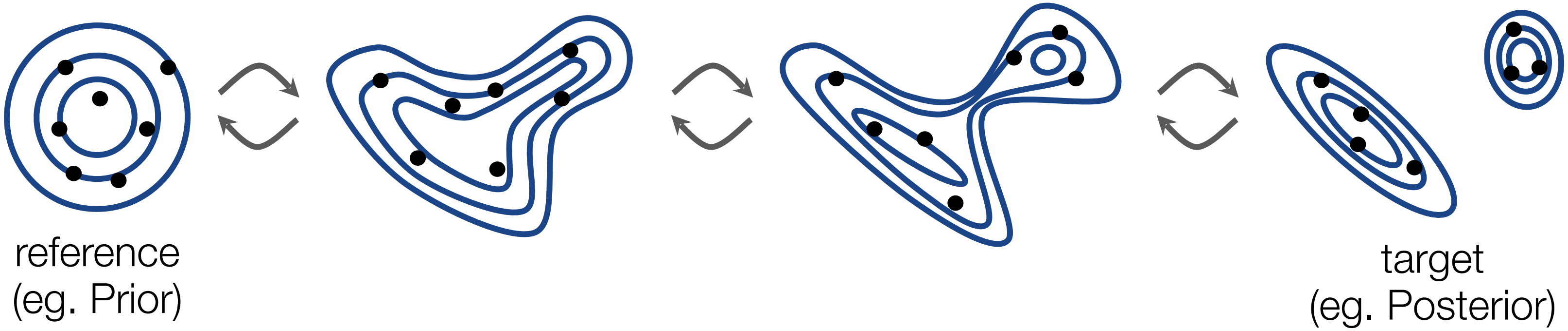
Schedule: $t_0, \ldots, t_N \in [0, 1]$

Reference: π_0

Target: π_1

Parallel Tempering

Key Idea: sample from a *path* of distributions, swap states along the path



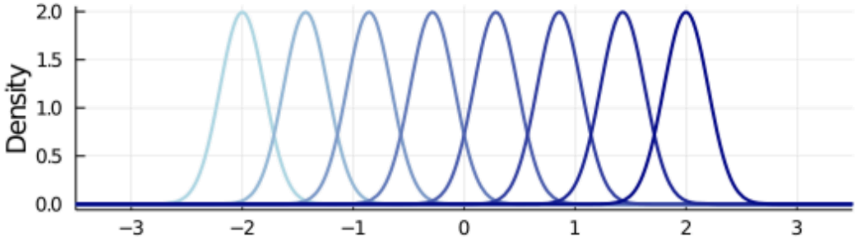
Annealing path: π_t

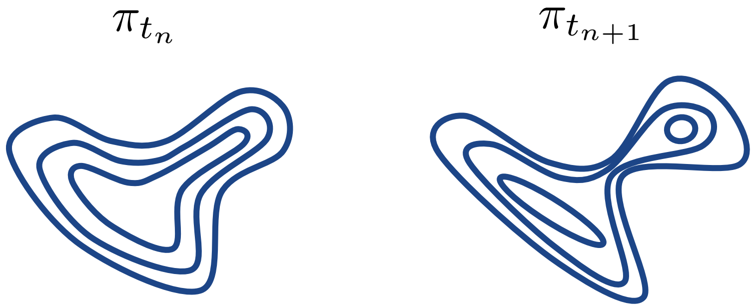
Schedule: $t_0, \ldots, t_N \in [0, 1]$

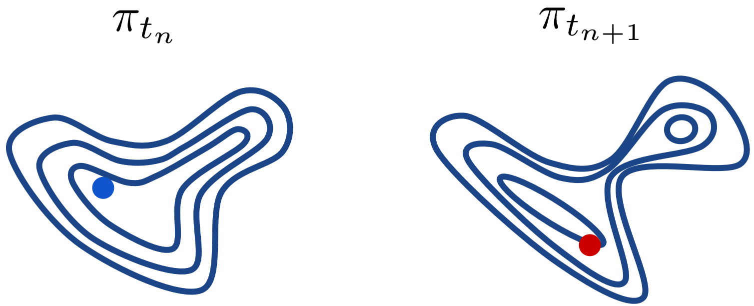
Reference: π_0

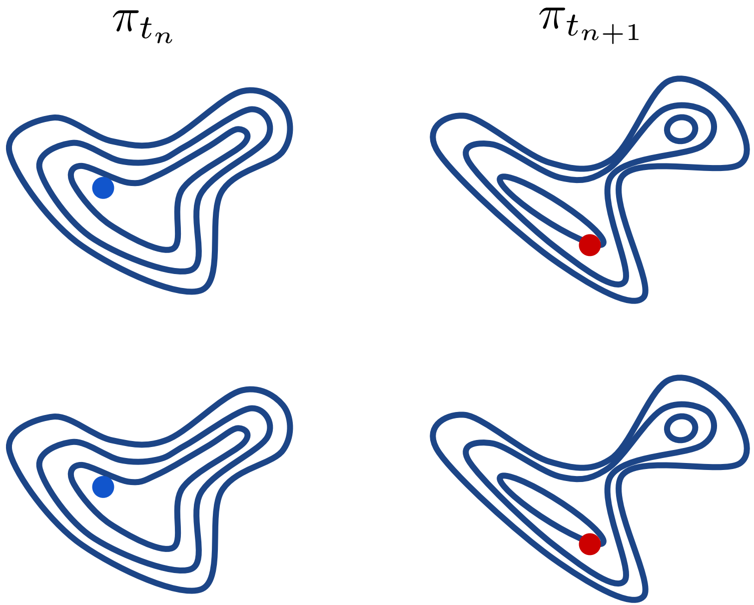
Target: π_1

Linear Path: $\pi_t \propto \pi_0^{1-t} \cdot \pi_1^t$

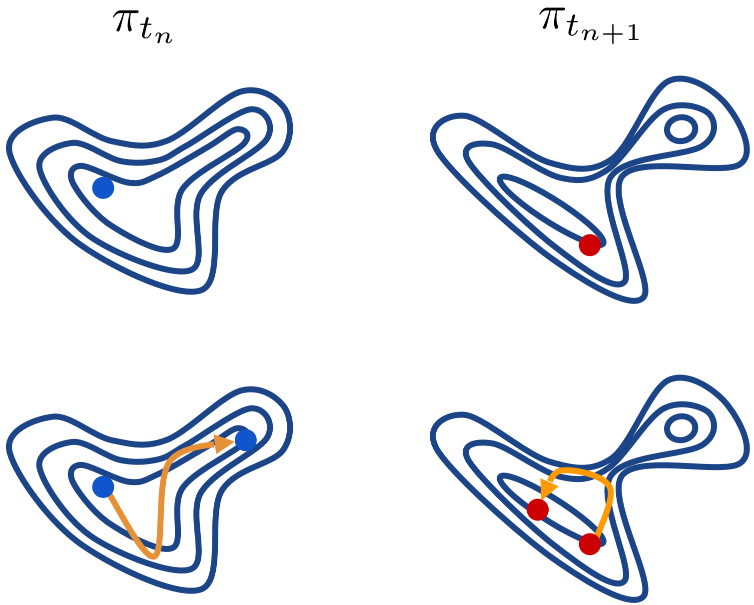



π_{t_n}
$\pi_{t_{n+1}}$

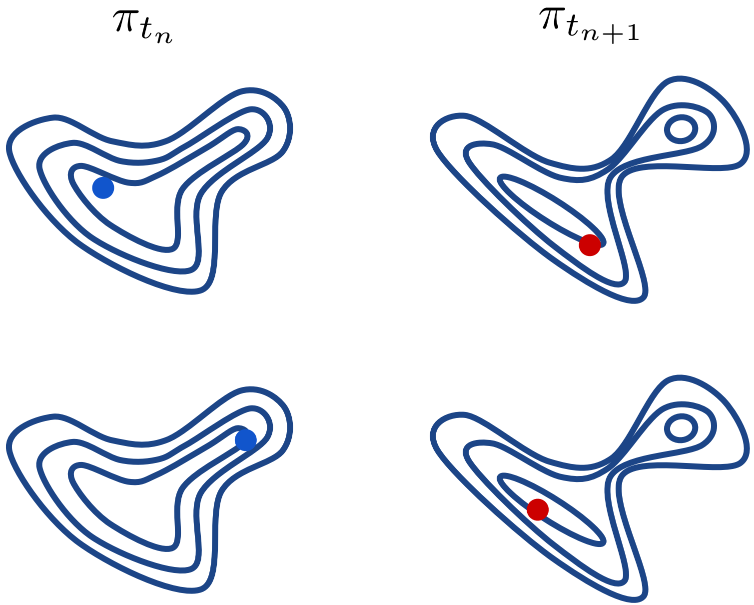

π_{t_n}
$\pi_{t_{n+1}}$



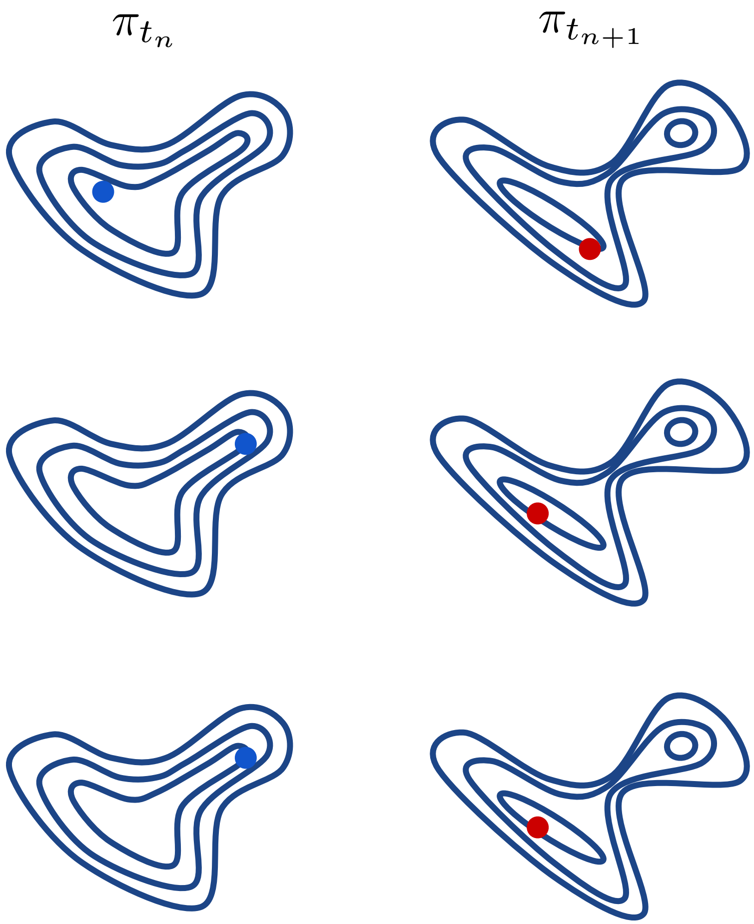
Local Exploration: apply any MCMC update to each chain (eg. HMC, Langevin, MH, etc.)



Local Exploration: apply any MCMC update to each chain (eg. HMC, Langevin, MH, etc.)



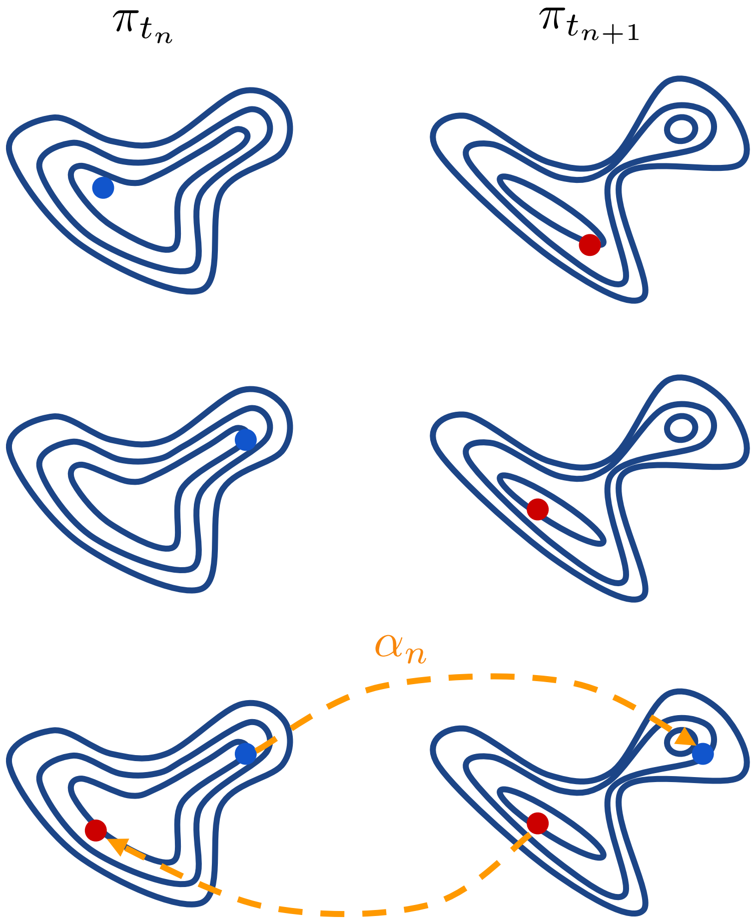
Local Exploration: apply any MCMC update to each chain (eg. HMC, Langevin, MH, etc.)



Local Exploration: apply any MCMC update to each chain (problem specific)

Communication: Metropolis-Hastings move to swap the states of adjacent chains with probability α_n

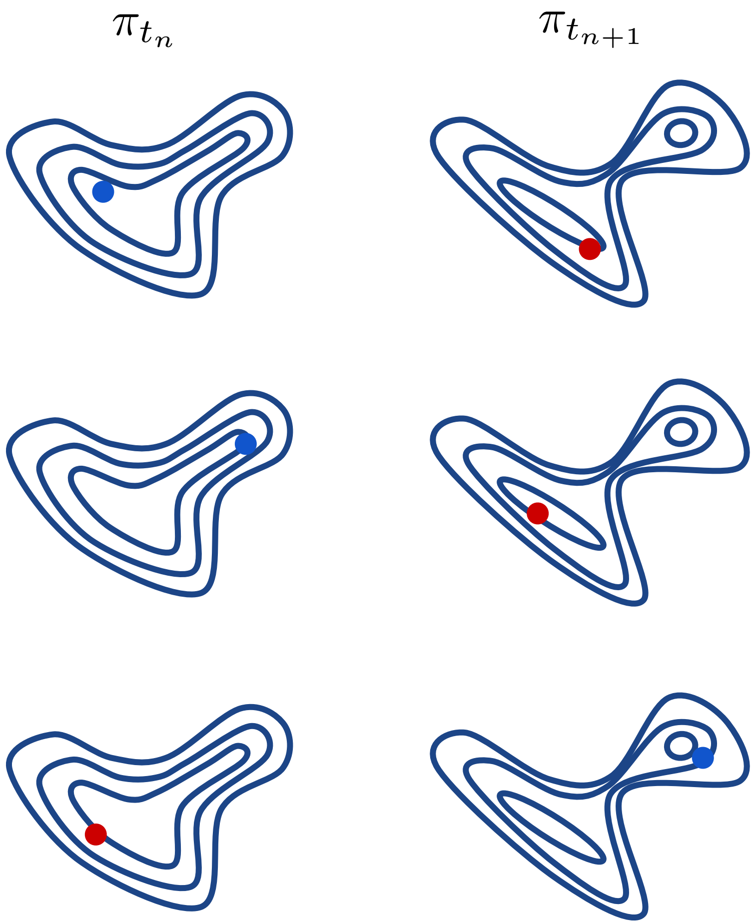
$$\alpha_n = 1 \wedge \frac{\pi_{t_n}(x_{n+1})\pi_{t_{n+1}}(x_n)}{\pi_{t_{n+1}}(x_{n+1})\pi_{t_n}(x_n)}$$



Local Exploration: apply any MCMC update to each chain (problem specific)

Communication: Metropolis-Hastings move to swap the states of adjacent chains with probability α_n

$$\alpha_n = 1 \wedge \frac{\pi_{t_n}(x_{n+1})\pi_{t_{n+1}}(x_n)}{\pi_{t_{n+1}}(x_{n+1})\pi_{t_n}(x_n)}$$



Local Exploration: apply any MCMC update to each chain (problem specific)

Communication: Metropolis-Hastings move to swap the states of adjacent chains with probability α_n

$$\alpha_n = 1 \wedge \frac{\pi_{t_n}(x_{n+1})\pi_{t_{n+1}}(x_n)}{\pi_{t_{n+1}}(x_{n+1})\pi_{t_n}(x_n)}$$

Round trips

How to assess the performance of PT?

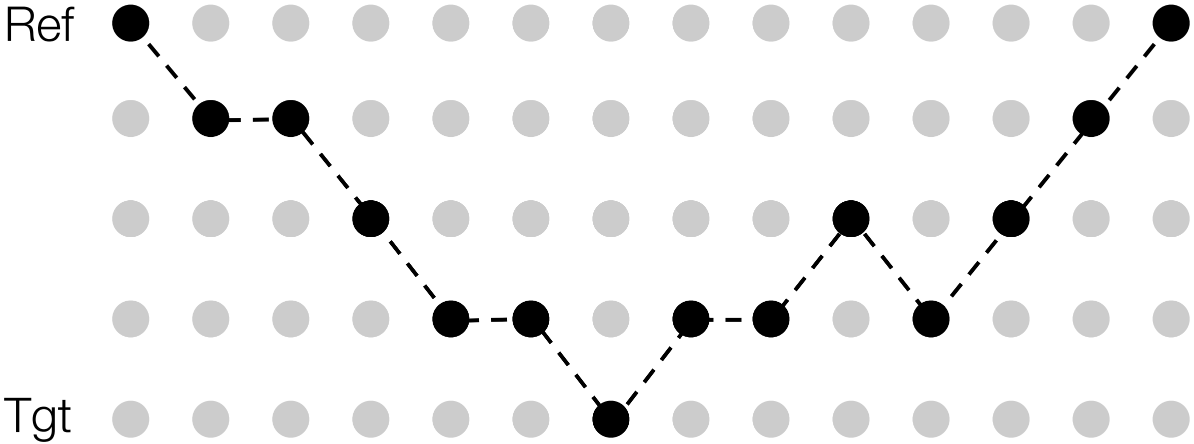
ESS etc influenced by exploration; want to evaluate communication

Round trips

How to assess the performance of PT?

ESS etc influenced by exploration; want to evaluate communication

Round Trip: when a reference state makes it to the target and back

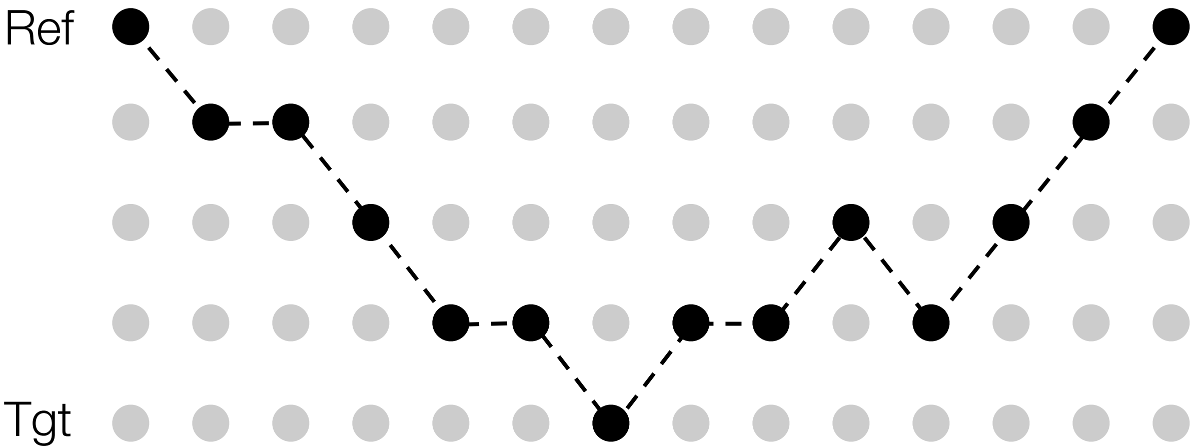


Round trips

How to assess the performance of PT?

ESS etc influenced by exploration; want to evaluate communication

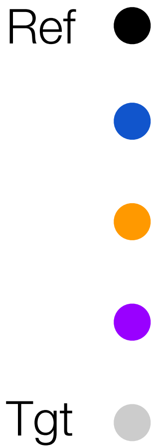
Round Trip: when a reference state makes it to the target and back



Round Trip Rate: the frequency of round trips

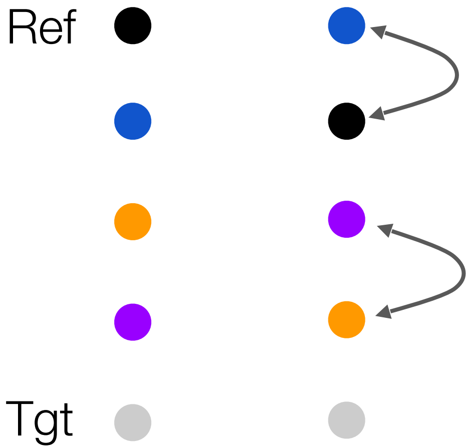
Non-Reversible Parallel Tempering (NRPT)

Deterministically alternate between even and odd, … [OKO+01]



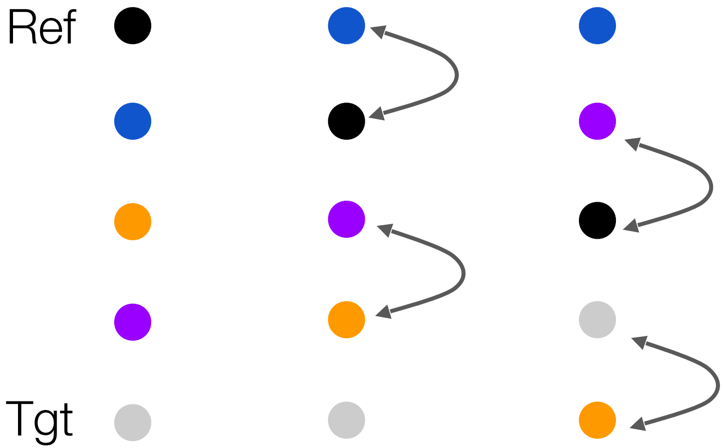
Non-Reversible Parallel Tempering (NRPT)

Deterministically alternate between even and odd, … [OKO+01]



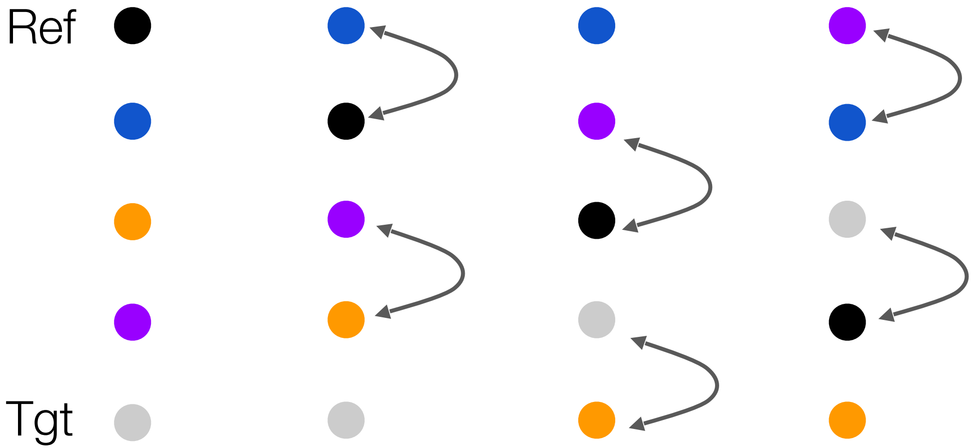
Non-Reversible Parallel Tempering (NRPT)

Deterministically alternate between even and odd, … [OKO+01]



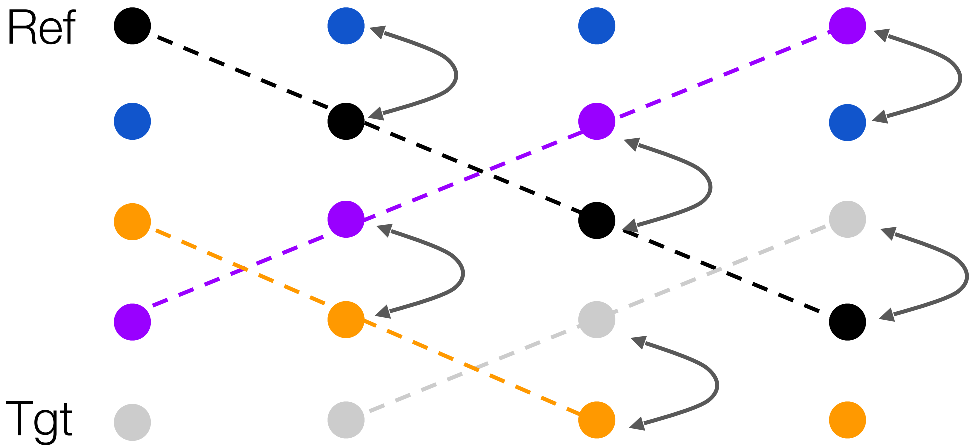
Non-Reversible Parallel Tempering (NRPT)

Deterministically alternate between even and odd, … [OKO+01]

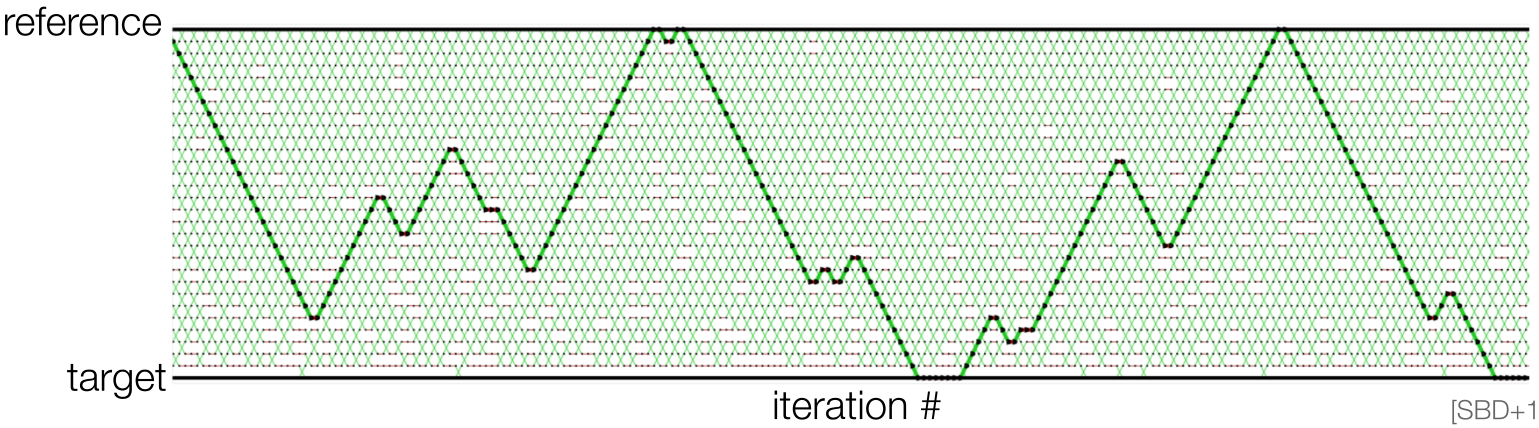


Non-Reversible Parallel Tempering (NRPT)

Deterministically alternate between even and odd, … [OKO+01]



- NRPT eliminates diffusive behaviour provides optimal round trip rate for a given path [SBD+19, JRSS-B]



[SBD+19]
Bayesian changepoint model

- [SBD+19] derives a robust and efficient algorithm to find optimal annealing schedule for a given path

Bayesian Mixture Model (d = 305)

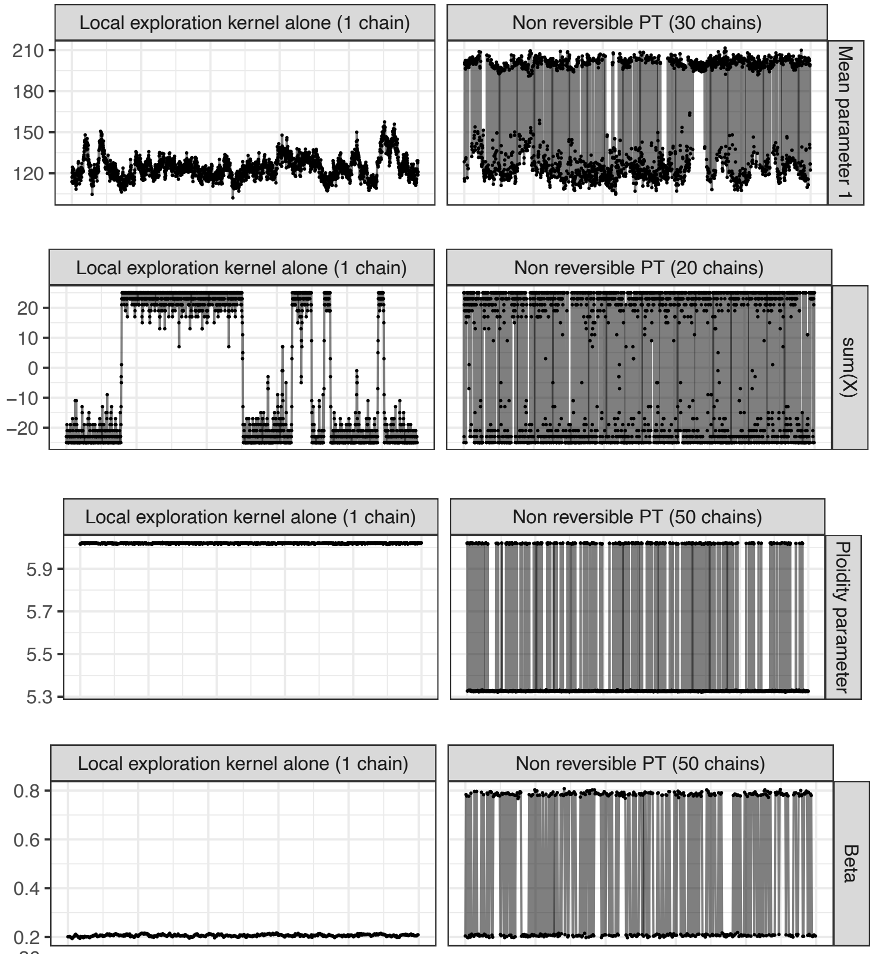
Ising Model (d = 25)

Copy number inference (d = 30)

- Whole genome ovarian cancer data

ODE parameter inference (d = 5)

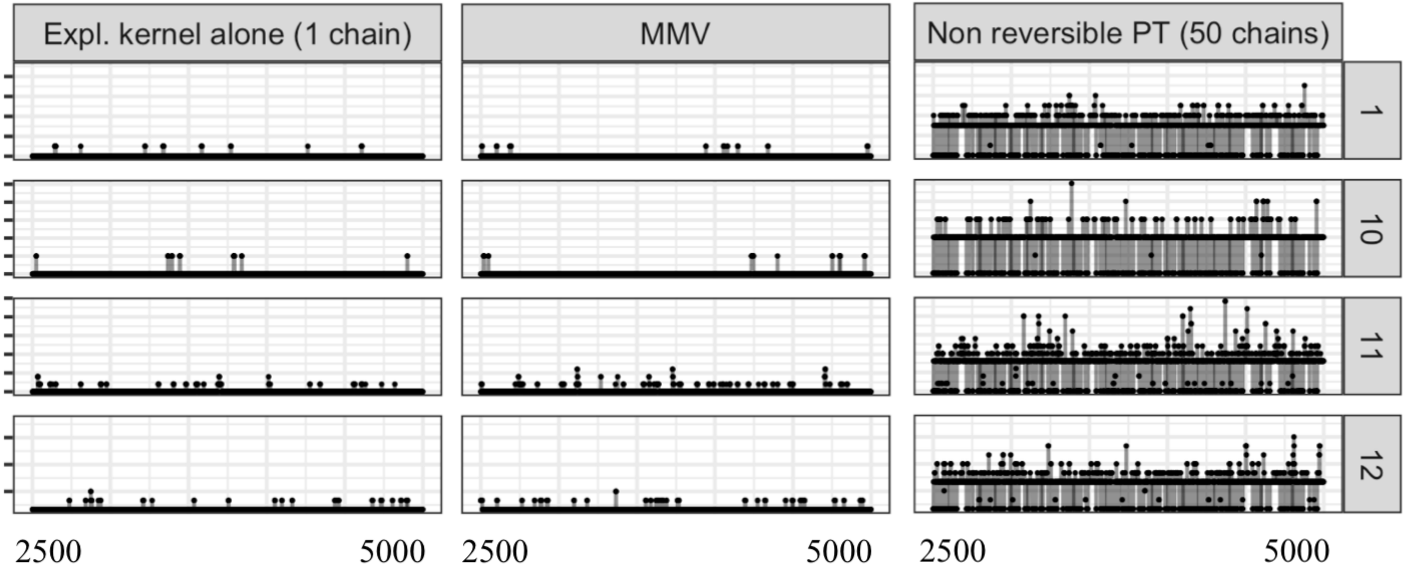
- mRNA data



Copy number inference (d = 30)

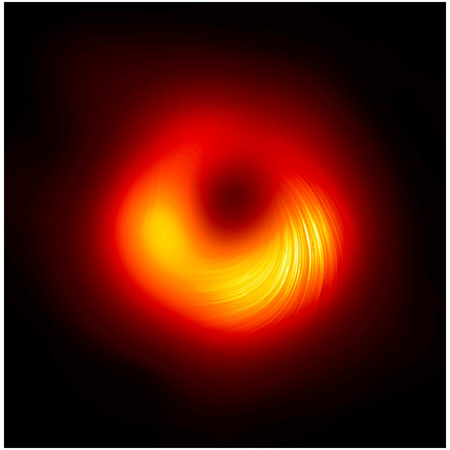
- Whole genome ovarian cancer data

1 Chain | **Reversible PT** | **Non-Reversible PT**



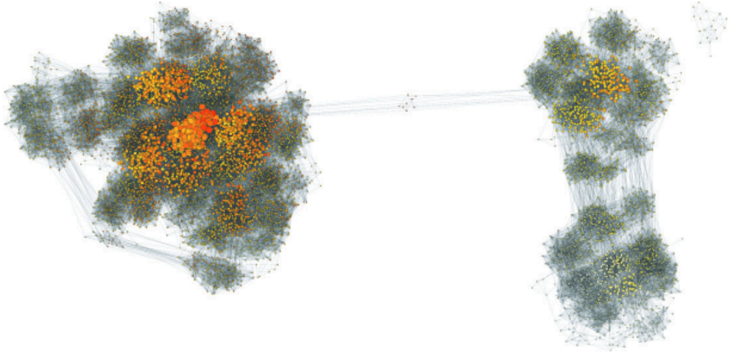
Event horizon telescope collaboration (EHT) used NRPT process original photo from 3 days to 1 hour with higher confidence

Recently EHT used NRPT to discover magnetic polarization in the M87 blackhole!



The EHT Collaboration, 2021. First M87 Event Horizon Telescope Results. VII. Polarization of the Ring.

BC Cancer Research Center used NRPT to improve phylogenetic inference of single cell cancer data by order of 400x [DSC+20]



Communication barrier for a fixed path

Recall: swap probability

$$\alpha_n = 1 \wedge \frac{\pi_{t_n}(x_{n+1})\pi_{t_{n+1}}(x_n)}{\pi_{t_{n+1}}(x_{n+1})\pi_{t_n}(x_n)}$$

Communication barrier for a fixed path

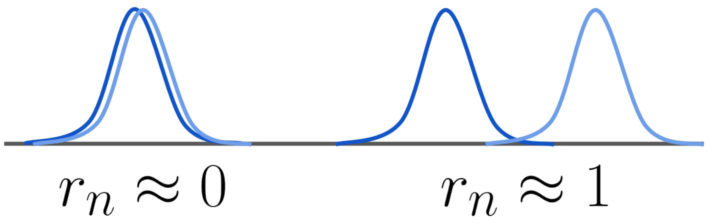
Recall: swap probability

$$\alpha_n = 1 \wedge \frac{\pi_{t_n}(x_{n+1})\pi_{t_{n+1}}(x_n)}{\pi_{t_{n+1}}(x_{n+1})\pi_{t_n}(x_n)}$$

Rejection Rate

$$r_n = 1 - \mathbb{E}\left[\alpha_n\right]$$

$$(X_n, X_{n+1}) \sim \pi_{t_n} \cdot \pi_{t_{n+1}}$$



Communication barrier for a fixed path

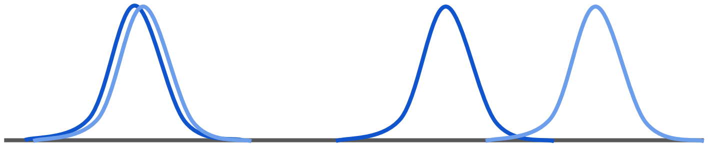
Recall: swap probability

$$\alpha_n = 1 \wedge \frac{\pi_{t_n}(x_{n+1})\pi_{t_{n+1}}(x_n)}{\pi_{t_{n+1}}(x_{n+1})\pi_{t_n}(x_n)}$$

Rejection Rate

$$r_n = 1 - \mathbb{E}\left[\alpha_n\right]$$

$$(X_n, X_{n+1}) \sim \pi_{t_n} \cdot \pi_{t_{n+1}}$$



$r_n \approx 0$ $\qquad$ $r_n \approx 1$

Theorem:
The *round trip rate* is equal to

$$\tau_N = \left(2 + 2\sum_{n=0}^{N-1} \frac{r_n}{1 - r_n}\right)^{-1}$$

Communication barrier for a fixed path

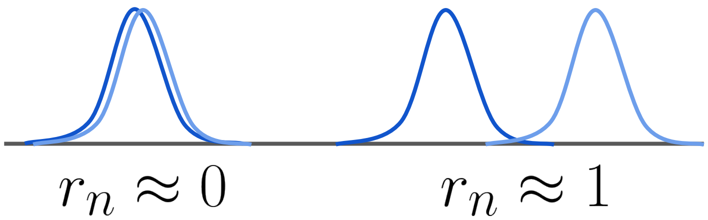
Recall: swap probability

$$\alpha_n = 1 \wedge \frac{\pi_{t_n}(x_{n+1})\pi_{t_{n+1}}(x_n)}{\pi_{t_{n+1}}(x_{n+1})\pi_{t_n}(x_n)}$$

Rejection Rate

$$r_n = 1 - \mathbb{E}[\alpha_n]$$

$$(X_n, X_{n+1}) \sim \pi_{t_n} \cdot \pi_{t_{n+1}}$$



Theorem:
The *round trip rate* is equal to

$$\tau_N = \left(2 + 2\sum_{n=0}^{N-1} \frac{r_n}{1-r_n}\right)^{-1}$$

and in the limit of parallel computation, $N \to \infty$,

$$\tau_N \to (2 + 2\Lambda)^{-1}$$

Λ is the *global communication barrier*

Communication barrier for a fixed path

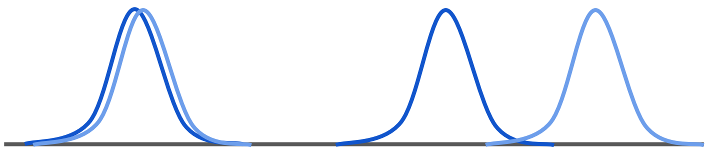
Recall: swap probability

$$\alpha_n = 1 \wedge \frac{\pi_{t_n}(x_{n+1})\pi_{t_{n+1}}(x_n)}{\pi_{t_{n+1}}(x_{n+1})\pi_{t_n}(x_n)}$$

Rejection Rate

$$r_n = 1 - \mathbb{E}\left[\alpha_n\right]$$

$$(X_n, X_{n+1}) \sim \pi_{t_n} \cdot \pi_{t_{n+1}}$$



$r_n \approx 0$ $\quad$ $r_n \approx 1$

Theorem:
The *round trip rate* is equal to

$$\tau_N = \left(2 + 2\sum_{n=0}^{N-1} \frac{r_n}{1 - r_n}\right)^{-1}$$

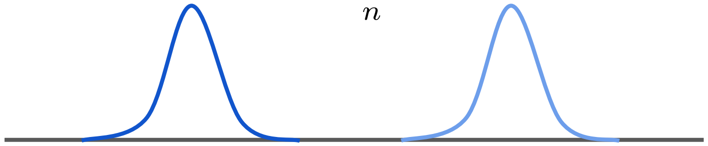
and in the limit of parallel computation, $N \to \infty$,

$$\tau_N \to (2 + 2\Lambda)^{-1} \qquad \Lambda \approx \sum_n r_n$$

Λ is the *global communication barrier*

A Breakdown in Communication

$$\Lambda \approx \sum_n r_n$$

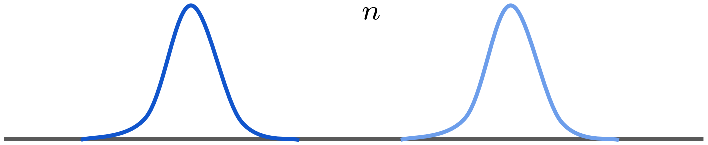


reference and target are *nearly mutually singular*

global communication barrier Λ is large!

A Breakdown in Communication

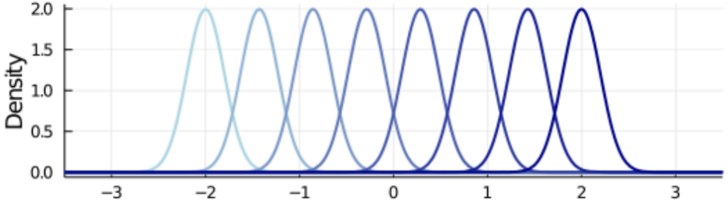
$$\Lambda \approx \sum_n r_n$$



reference and target are *nearly mutually singular*

global communication barrier Λ is large!

Linear Path $\pi_t \propto \pi_0^{1-t} \cdot \pi_1^t$



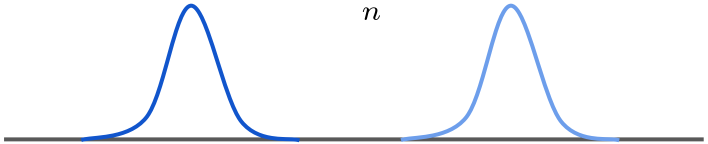
not much overlap between distributions n, n+1

A Breakdown in Communication

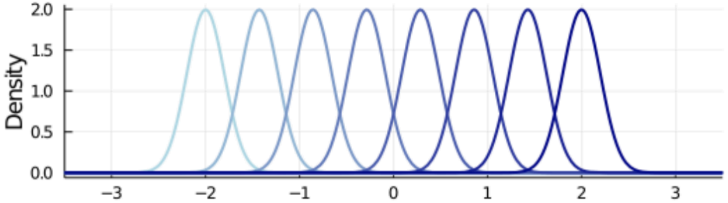
$$\Lambda \approx \sum_n r_n$$

reference and target are *nearly mutually singular*

global communication barrier Λ is large!



Linear Path $\pi_t \propto \pi_0^{1-t} \cdot \pi_1^t$



not much overlap between distributions n, n+1

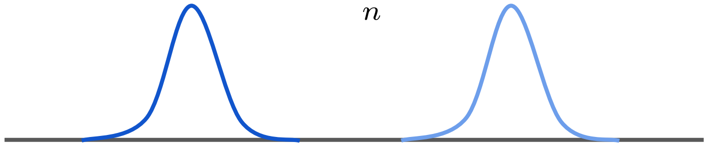
Bad for Bayes: prior (reference) and posterior (target) often nearly mutually singular

A Breakdown in Communication

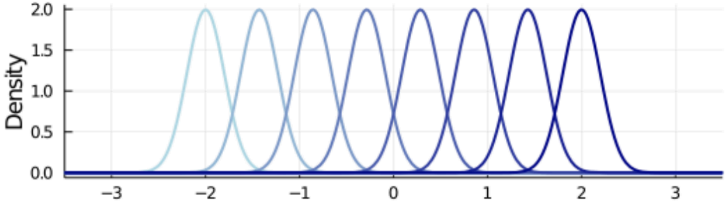
$$\Lambda \approx \sum_n r_n$$

reference and target are *nearly mutually singular*

global communication barrier Λ is large!



Linear Path $\pi_t \propto \pi_0^{1-t} \cdot \pi_1^t$



not much overlap between distributions n, n+1

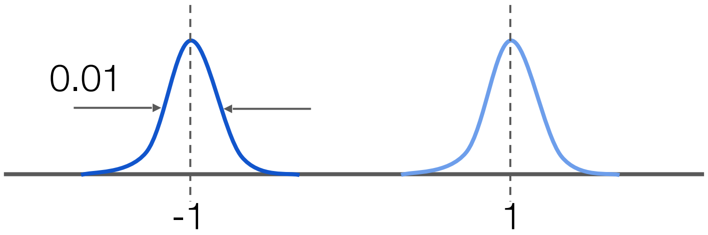
Bad for Bayes: prior (reference) and posterior (target) often nearly mutually singular

Is this problem just fundamentally hard?

(hope not...they're Gaussians…)

Empirical Performance

e.g. Gaussian ref & target



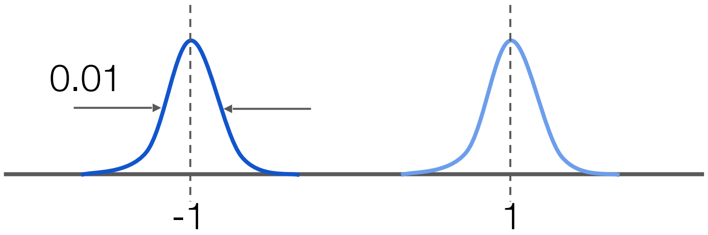
50 chains

i.i.d. exploration

$$\pi_t \propto \pi_0^{1-t} \cdot \pi_1^t$$

Empirical Performance

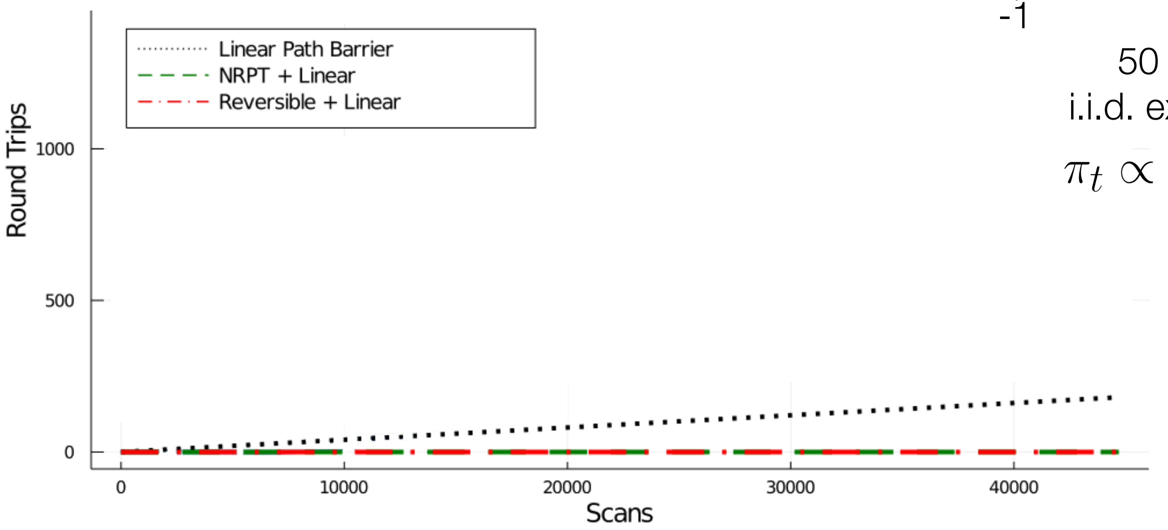
e.g. Gaussian ref & target



50 chains

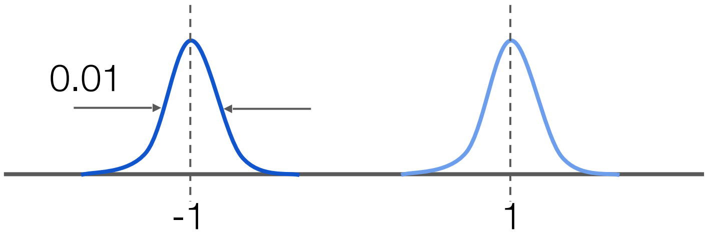
i.i.d. exploration

$$\pi_t \propto \pi_0^{1-t} \cdot \pi_1^t$$



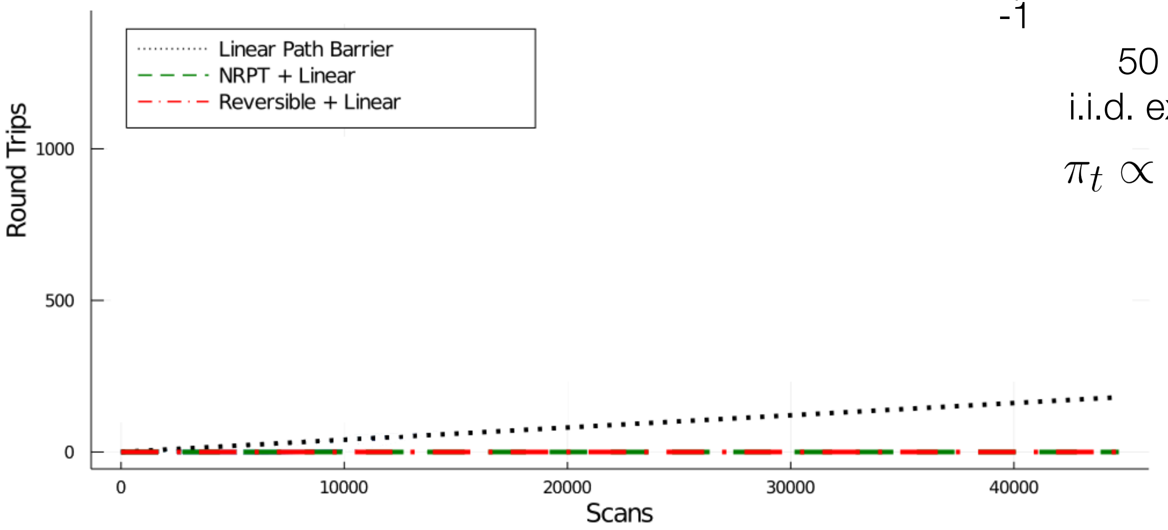
Empirical Performance

e.g. Gaussian ref & target



50 chains
i.i.d. exploration

$$\pi_t \propto \pi_0^{1-t} \cdot \pi_1^t$$



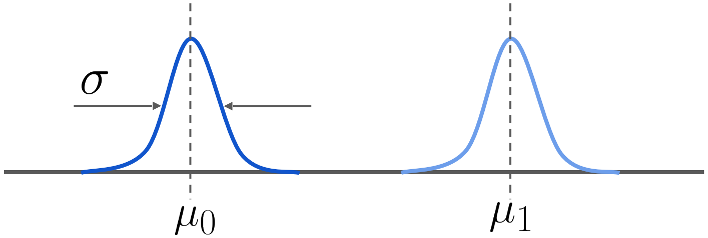
No round trips after 50K steps…

(not to mention upper bound of ~100...)

Can we do better…?

Can we do better…?

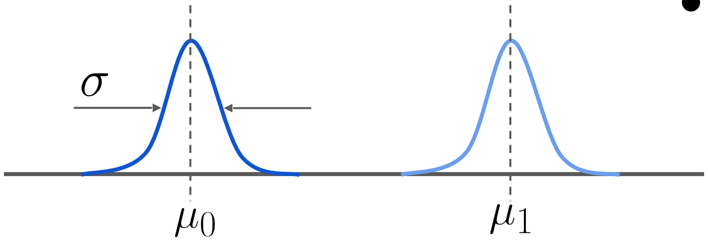
e.g. Gaussian ref & target



$$z = |\mu_0 - \mu_1|/\sigma$$

Can we do better…?

e.g. Gaussian ref & target



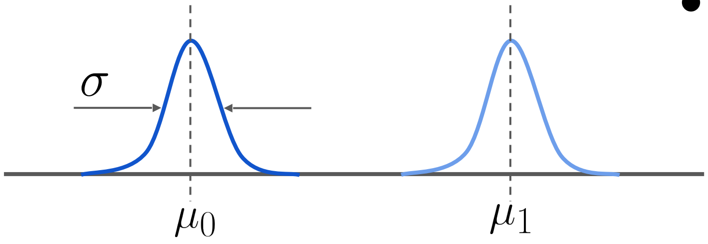
$$z = |\mu_0 - \mu_1|/\sigma$$

Proposition:

- the linear path has $\Lambda = \Theta(z)$
- there exists a path of Gaussians with $\Lambda = O(\log z)$

Can we do better…?

e.g. Gaussian ref & target



$$z = |\mu_0 - \mu_1|/\sigma$$

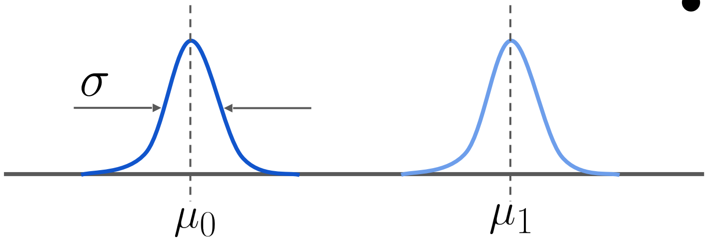
Proposition:

- the linear path has $\Lambda = \Theta(z)$
- there exists a path of Gaussians with $\Lambda = O(\log z)$

we can do at least exponentially better than the standard linear path!

Can we do better…?

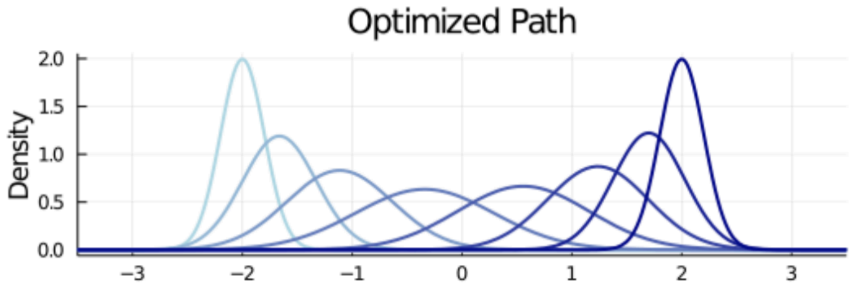
e.g. Gaussian ref & target



$$z = |\mu_0 - \mu_1|/\sigma$$

Proposition:

- the linear path has $\Lambda = \Theta(z)$
- there exists a path of Gaussians with $\Lambda = O(\log z)$



we can do at least exponentially better than the standard linear path!

Exponential Path Family

What kinds of path families should we consider in practice?

Exponential Path Family

What kinds of path families should we consider in practice?

1. Shouldn't be specific to a particular reference, target (e.g. $\pi_t \propto \pi_0^{1-t} \cdot \pi_1^t$)

Exponential Path Family

What kinds of path families should we consider in practice?

1. Shouldn't be specific to a particular reference, target (e.g. $\pi_t \propto \pi_0^{1-t} \cdot \pi_1^t$)
2. Should include the linear path

Exponential Path Family

What kinds of path families should we consider in practice?

1. Shouldn't be specific to a particular reference, target (e.g. $\pi_t \propto \pi_0^{1-t} \cdot \pi_1^t$)
2. Should include the linear path
3. Should deform reference to target while maximizing overlap

Exponential Path Family

What kinds of path families should we consider in practice?

1. Shouldn't be specific to a particular reference, target (e.g. $\pi_t \propto \pi_0^{1-t} \cdot \pi_1^t$)
2. Should include the linear path
3. Should deform reference to target while maximizing overlap

- Family of paths: $\pi_t \propto \pi_0^{\eta_0(t)} \pi_1^{\eta_1(t)}$

Piecewise twice continuously differentiable functions: $\eta : [0, 1] \to \mathbb{R}^2$

Exponential Path Family

What kinds of path families should we consider in practice?

1. Shouldn't be specific to a particular reference, target (e.g. $\pi_t \propto \pi_0^{1-t} \cdot \pi_1^t$)
2. Should include the linear path
3. Should deform reference to target while maximizing overlap

- Family of paths: $\pi_t \propto \pi_0^{\eta_0(t)} \pi_1^{\eta_1(t)}$

Piecewise twice continuously differentiable functions: $\eta : [0, 1] \to \mathbb{R}^2$

Designing path of densities → designing a path in $\mathbb{R}^2$

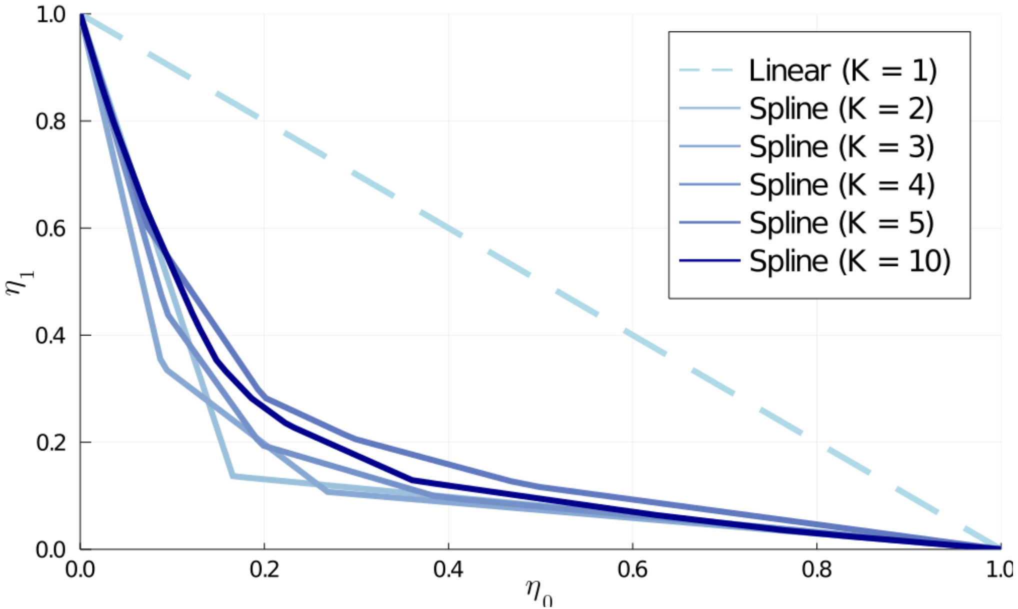
which path?

Spline Path Family

Use a *linear spline* η with K knots

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

Optimize knots with SDG

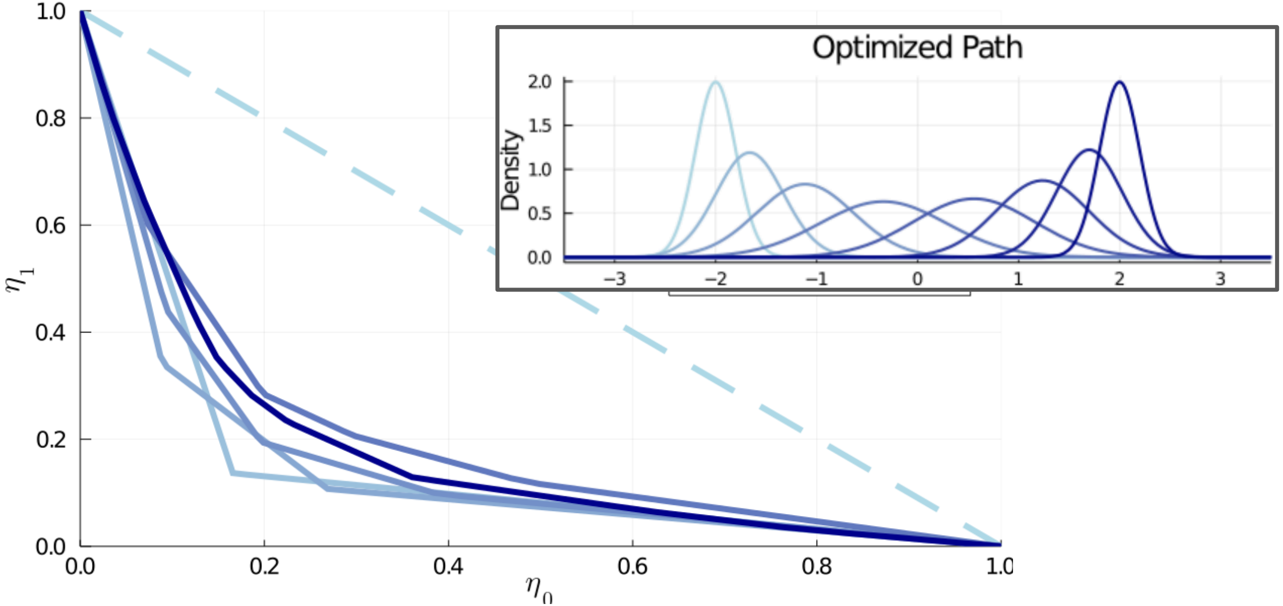


Spline Path Family

Use a *linear spline* η with K knots

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

Optimize knots with SDG

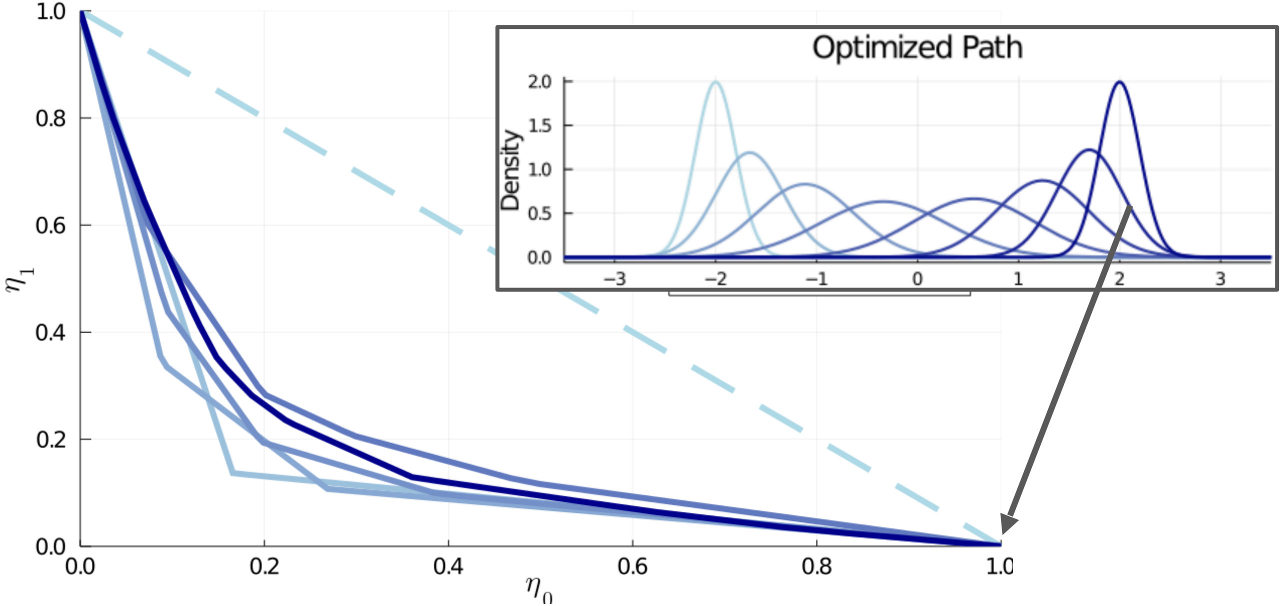


Spline Path Family

Use a *linear spline* η with K knots

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

Optimize knots with SDG

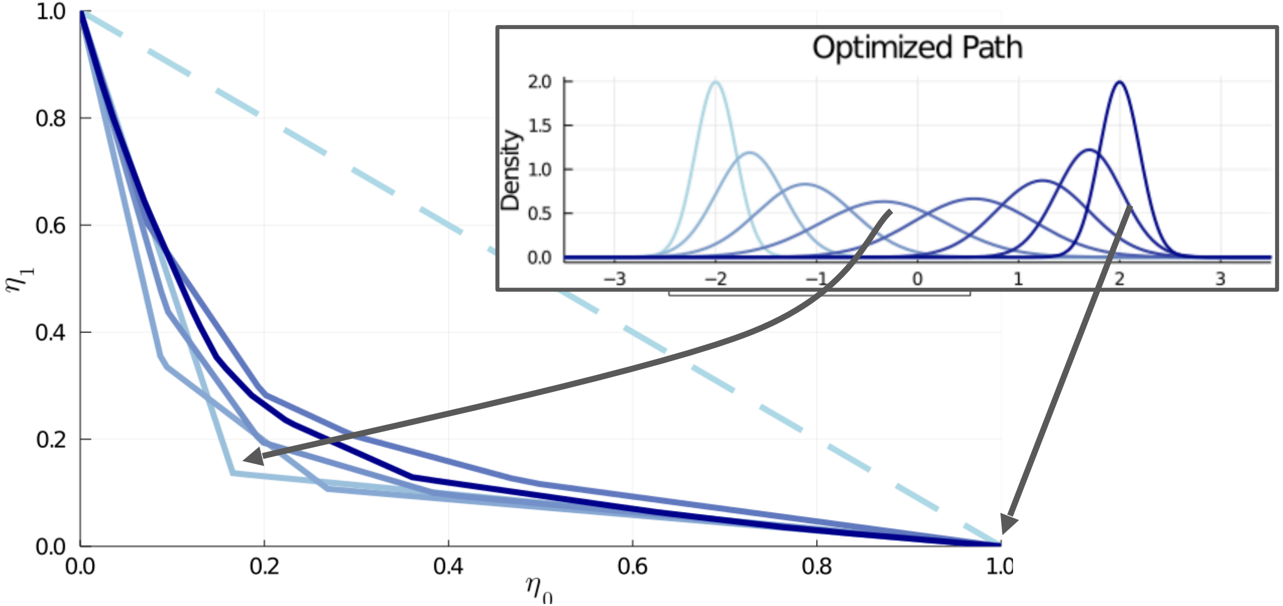


Spline Path Family

Use a *linear spline* η with K knots

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

Optimize knots with SDG

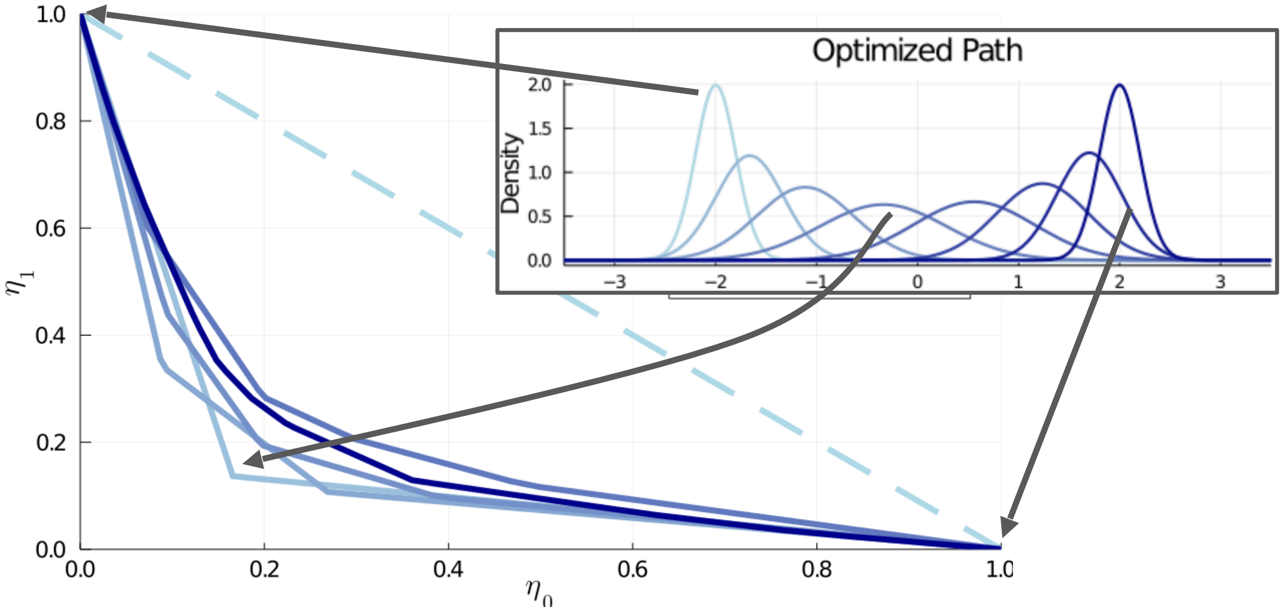


Spline Path Family

Use a *linear spline* η with K knots

Optimize knots with SDG

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



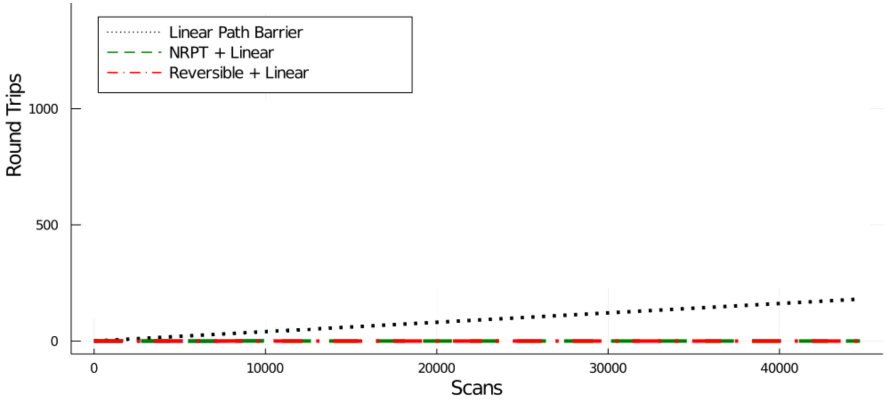
Gaussians

Ref: $N(-1, 10^{-4})$ Tgt: $N(1, 10^{-4})$

red/green: linear path

black: best possible for linear path

blues: optimized spline path



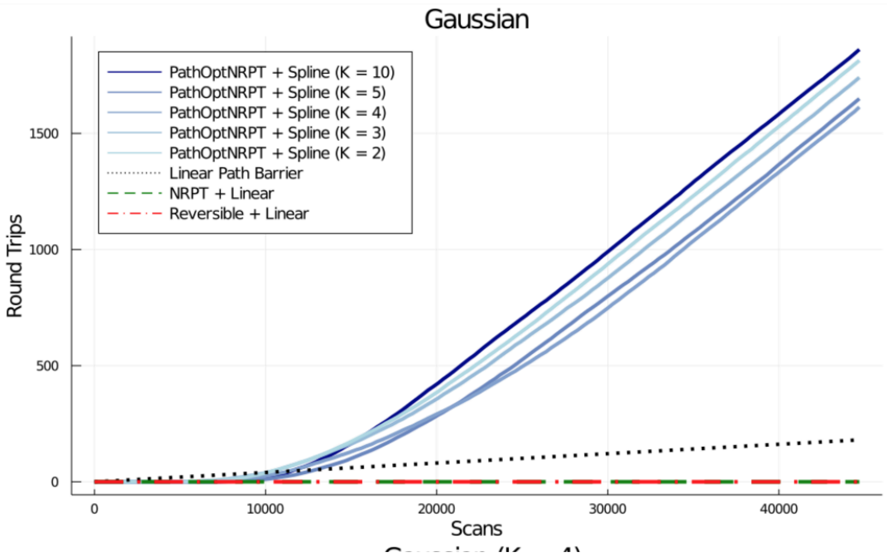
Gaussians

Ref: $N(-1, 10^{-4})$ Tgt: $N(1, 10^{-4})$

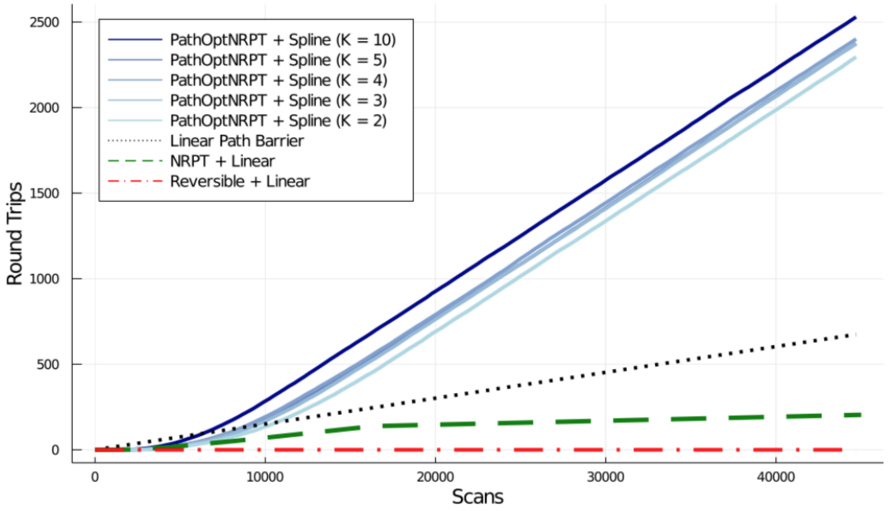
red/green: linear path

black: best possible for linear path

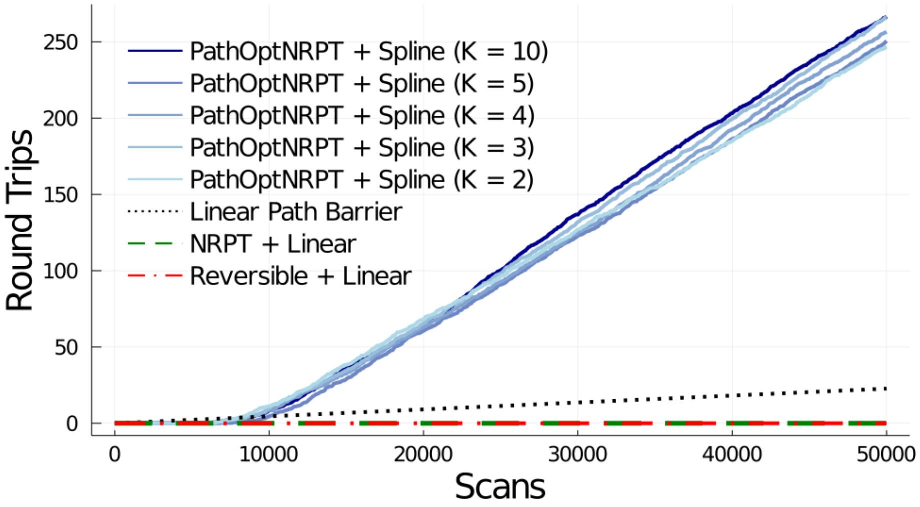
blues: optimized spline path



Beta-Binomial Model



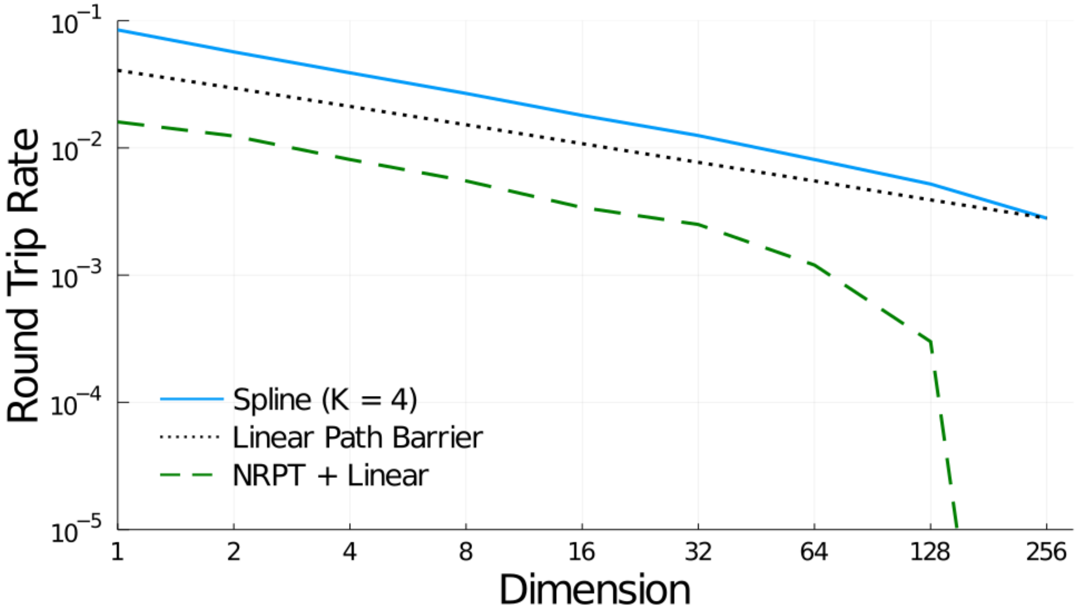
Shapley Galaxy Data (d = 95)



Scaling with Dimension

Reference: $\mathcal{N}((-1,...,-1), 10^{-2}I)$
Target: $\mathcal{N}((1,...,1), 10^{-2}I)$

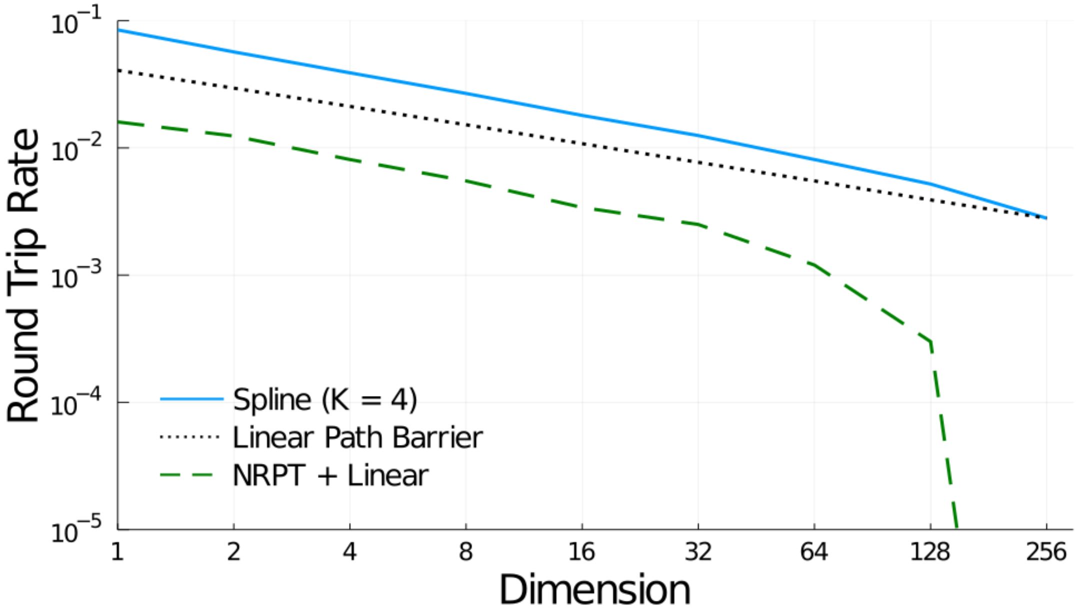
50,000 scans, $15\sqrt{d}$ chains



Scaling with Dimension

Reference: $\mathcal{N}((-1,\ldots,-1),10^{-2}I)$
Target: $\mathcal{N}((1,\ldots,1),10^{-2}I)$

50,000 scans, $15\sqrt{d}$ chains



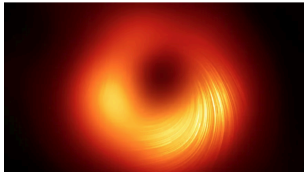
problem gets harder

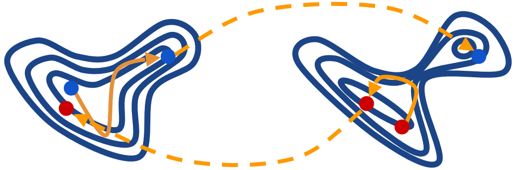
but benefit of using optimized vs linear paths actually *increases*

Conclusion

Conclusion

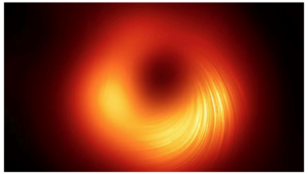
PT enables inference with intractable, multimodal posteriors

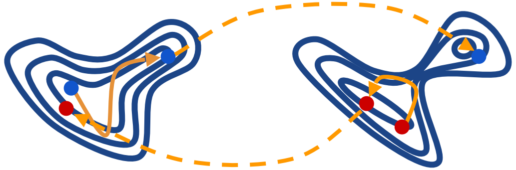




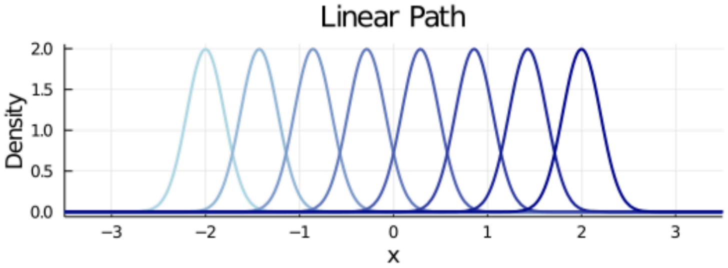
Conclusion

PT enables inference with intractable, multimodal posteriors



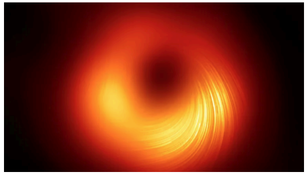


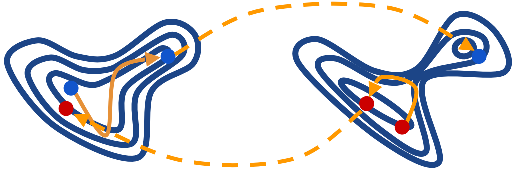
but the standard linear path has suboptimal communication efficiency



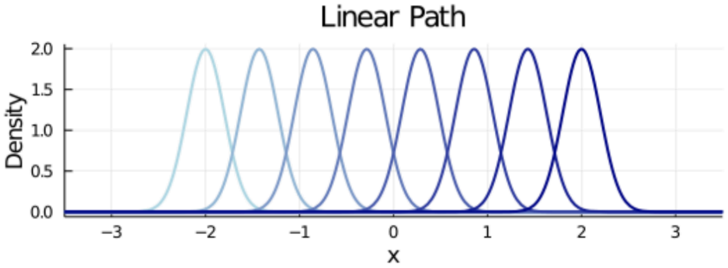
Conclusion

PT enables inference
with intractable,
multimodal posteriors



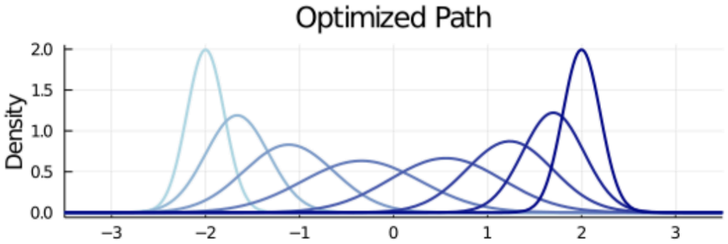


but the standard linear path has
suboptimal communication efficiency



this work: PT on optimized paths

- new theory of general path efficiency
- flexible spline path family
- path tuning algorithm



arXiv preprint:
https://arxiv.org/abs/2102.07720

Questions?

Asymptotic Round Trip Rate

What about asymptotics in N (increasing parallel threads)?

Theorem: $\tau_N \to (2 + 2\Lambda)^{-1}$

Asymptotic Round Trip Rate

What about asymptotics in N (increasing parallel threads)?

Theorem: $\tau_N \to (2 + 2\Lambda)^{-1}$

$$\Lambda = \frac{1}{2} \int_0^1 \mathbb{E} \left| \frac{\mathrm{d}W_t}{\mathrm{d}t}(X_t) - \frac{\mathrm{d}W_t}{\mathrm{d}t}(X_t') \right| \mathrm{d}t$$

$$X_t, X_t' \overset{\text{i.i.d.}}{\sim} \pi_t$$

Asymptotic Round Trip Rate

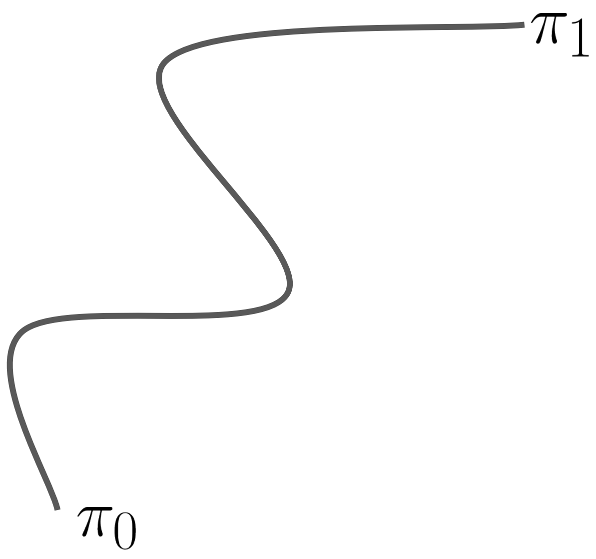
What about asymptotics in N (increasing parallel threads)?

Theorem: $\tau_N \to (2 + 2\Lambda)^{-1}$

$$\Lambda = \frac{1}{2}\int_0^1 \mathbb{E}\left|\frac{\mathrm{d}W_t}{\mathrm{d}t}(X_t) - \frac{\mathrm{d}W_t}{\mathrm{d}t}(X_t')\right| \mathrm{d}t$$

$$X_t, X_t' \overset{\text{i.i.d.}}{\sim} \pi_t$$

generalized communication barrier
looks sort of like "path length" for PT!



Asymptotic Round Trip Rate

What about asymptotics in N (increasing parallel threads)?

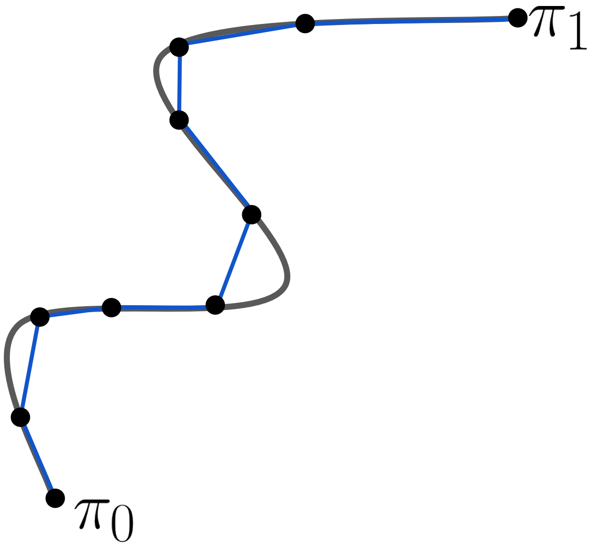
Theorem: $\tau_N \to (2 + 2\Lambda)^{-1}$

$$\Lambda = \frac{1}{2}\int_0^1 \mathbb{E}\left|\frac{\mathrm{d}W_t}{\mathrm{d}t}(X_t) - \frac{\mathrm{d}W_t}{\mathrm{d}t}(X_t')\right| \mathrm{d}t$$

$$X_t, X_t' \overset{\text{i.i.d.}}{\sim} \pi_t$$

generalized communication barrier
looks sort of like “path length” for PT!

In practice we use the path integral Λ_N on the linear spline (we only have discretized path)



Asymptotic Round Trip Rate

What about asymptotics in N (increasing parallel threads)?

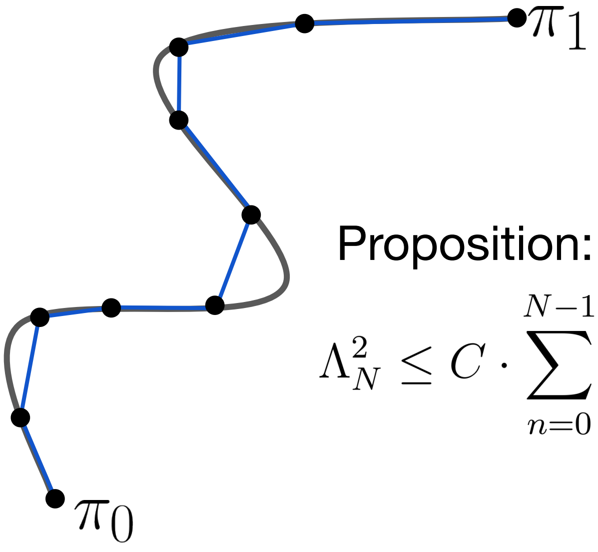
Theorem: $\tau_N \to (2 + 2\Lambda)^{-1}$

$$\Lambda = \frac{1}{2}\int_0^1 \mathbb{E}\left|\frac{\mathrm{d}W_t}{\mathrm{d}t}(X_t) - \frac{\mathrm{d}W_t}{\mathrm{d}t}(X_t')\right| \mathrm{d}t$$

$$X_t, X_t' \overset{\text{i.i.d.}}{\sim} \pi_t$$

generalized communication barrier looks sort of like “path length” for PT!

In practice we use the path integral Λ_N on the linear spline (we only have discretized path)



Proposition:

$$\Lambda_N^2 \leq C \cdot \sum_{n=0}^{N-1} \mathrm{SKL}(\pi_{t_n}, \pi_{t_{n+1}})$$

the round trip rate gradient signal-to-noise ratio is often too low
& gradient noise heavy-tailed
(especially in early iterations)

$$\tau_N = \left(2 + 2\sum_{n=0}^{N-1} \frac{r_n}{1 - r_n}\right)^{-1}$$

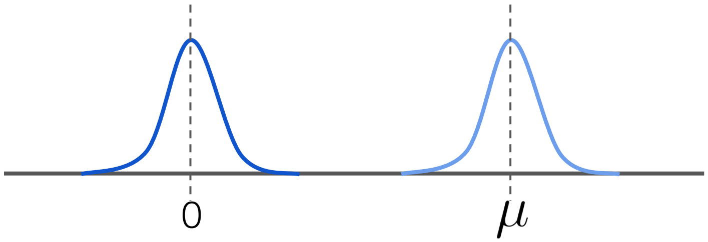
the round trip rate gradient signal-to-noise ratio is often too low
& gradient noise heavy-tailed
(especially in early iterations)

$$\tau_N = \left(2 + 2\sum_{n=0}^{N-1} \frac{r_n}{1 - r_n}\right)^{-1}$$

the round trip rate gradient signal-to-noise ratio is often too low
& gradient noise heavy-tailed
(especially in early iterations)

$$\tau_N = \left(2 + 2\sum_{n=0}^{N-1} \frac{r_n}{1-r_n}\right)^{-1}$$

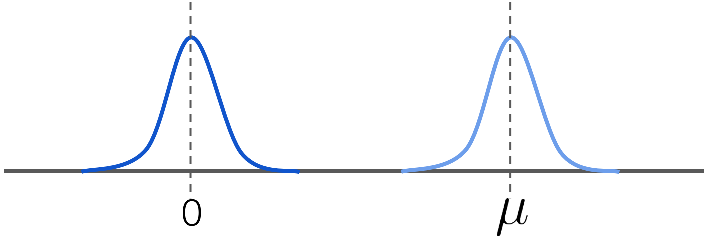
e.g. Gaussian ref & target

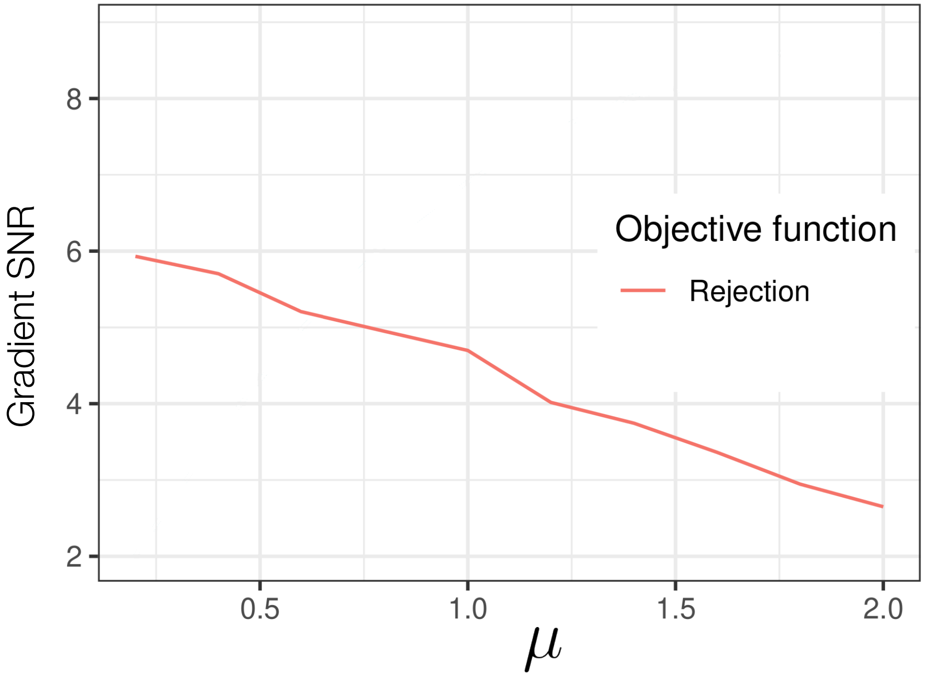


the round trip rate gradient signal-to-noise ratio is often too low
& gradient noise heavy-tailed
(especially in early iterations)

$$\tau_N = \left(2 + 2\sum_{n=0}^{N-1} \frac{r_n}{1 - r_n}\right)^{-1}$$

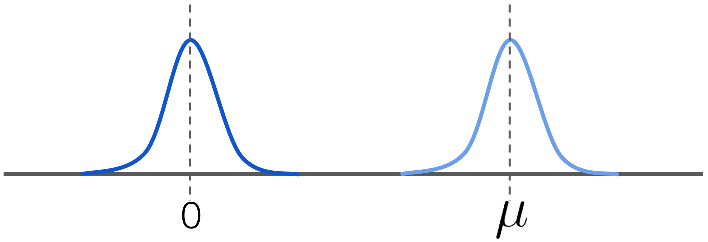
e.g. Gaussian ref & target

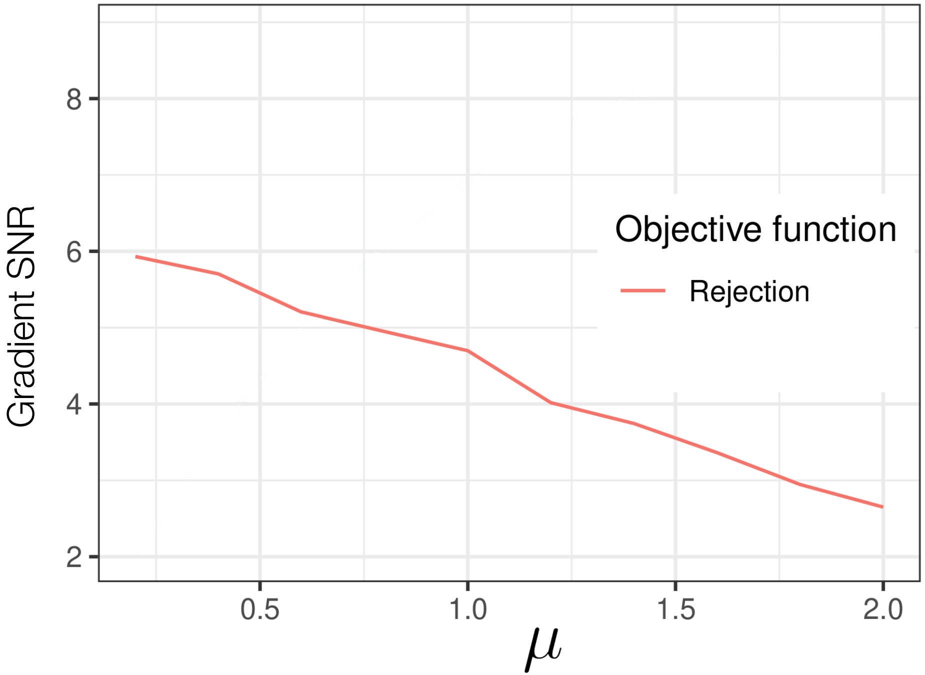




SKL has better gradient signal (early stages)

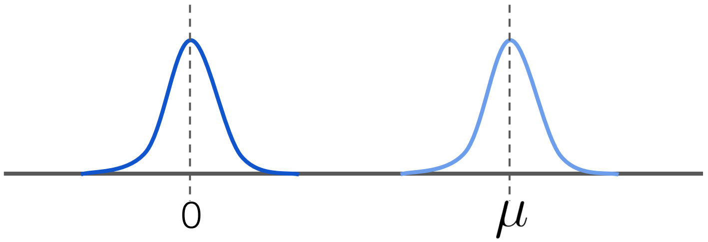
e.g. Gaussian ref & target

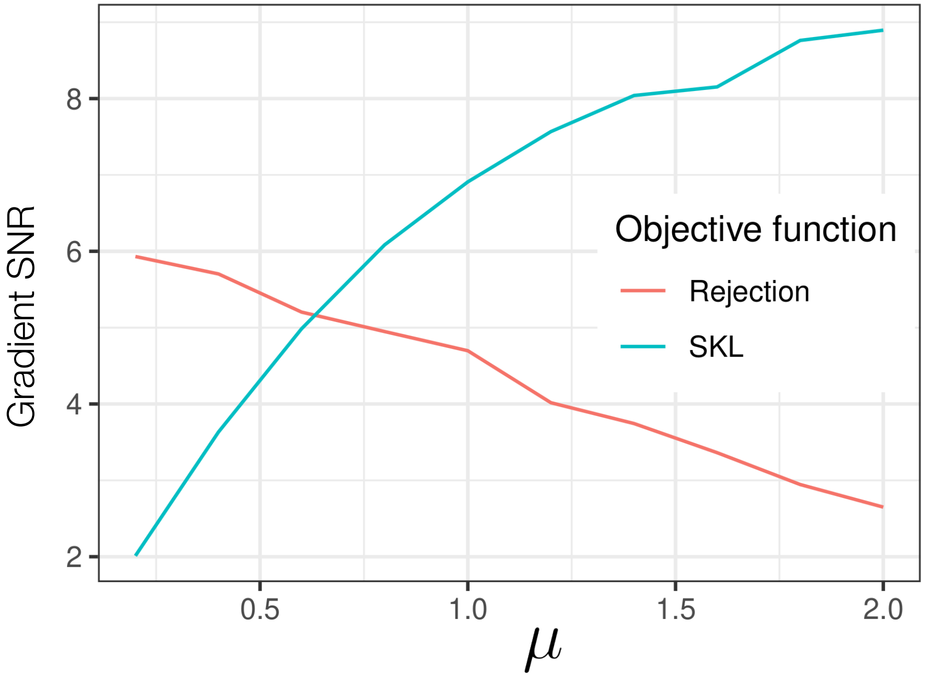




SKL has better gradient signal (early stages)

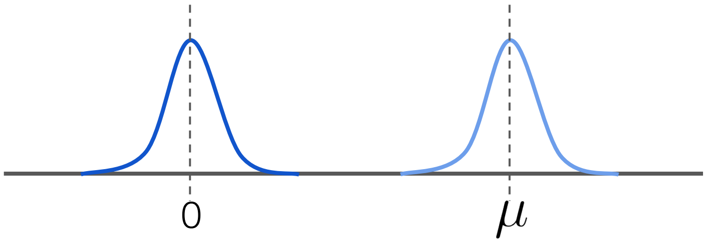
e.g. Gaussian ref & target



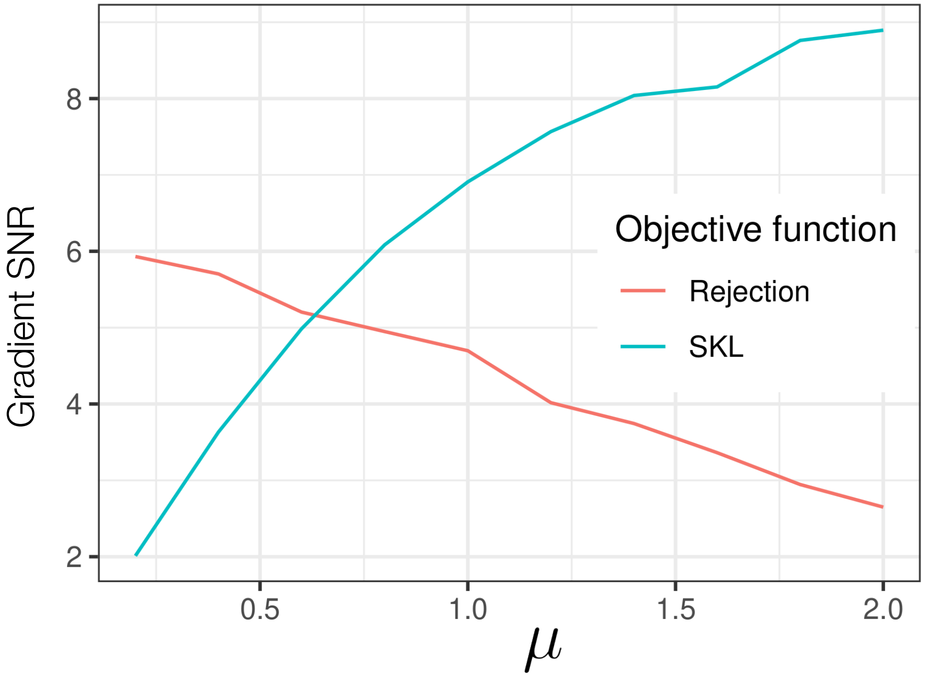


SKL has better gradient signal (early stages)

e.g. Gaussian ref & target

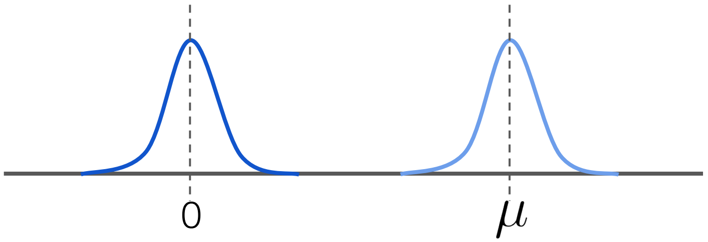


could switch to optimizing the round trip rate in later iterations



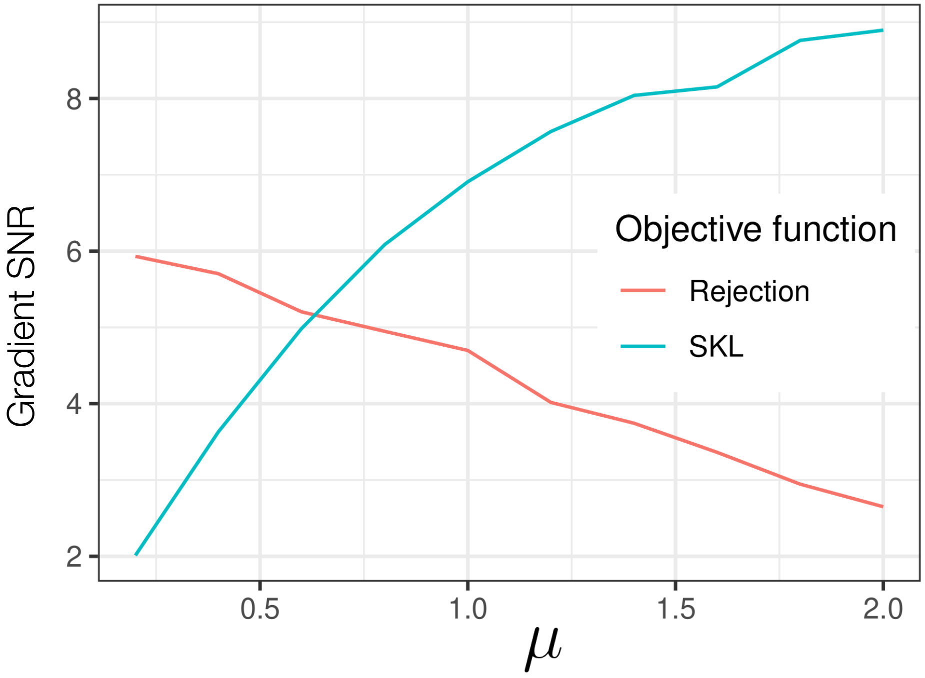
SKL has better gradient signal (early stages)

e.g. Gaussian ref & target



could switch to optimizing the round trip rate in later iterations

we use the schedule tuning procedure from [SBD+19]



Gaussians

Ref: $N(-1, 10^{-4})$ Tgt: $N(1, 10^{-4})$

blue: 1 / (round trip rate)

orange: symmetric KL