Limits of simulated and parallel tempering schemes in high dimensions

Gareth Roberts

University of Warwick

gareth.o.roberts@warwick.ac.uk

July 9, 2021

This is a body of work, jointly with Yves Atchade and mostly with Jeff Rosenthal and Nick Tawn.

It is presented in the series of papers: [Atchadé et al., 2011, Roberts and Rosenthal, 2014, Tawn and Roberts, 2018, Tawn et al., 2018, Roberts et al., 2020, Tawn et al., 2021]

This work is ongoing. Comments on the future directions very welcome!

The Problem

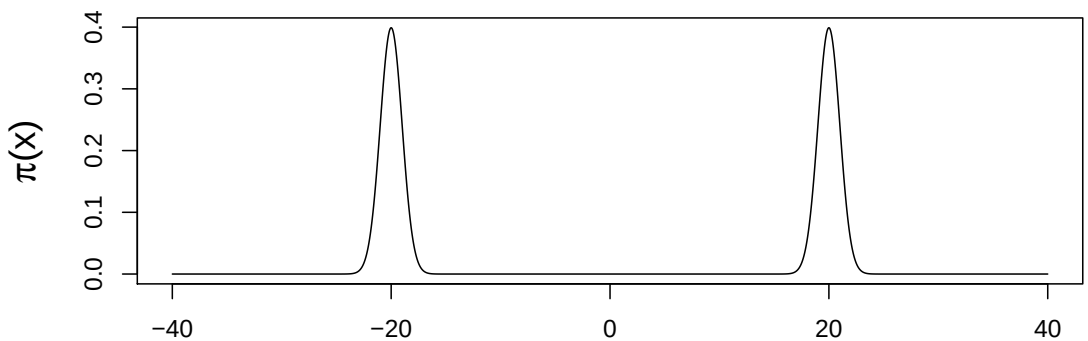
- **Aim:** Evaluate

$$\mathbb{E}_\pi[f(x)] = \int f(x)\pi(x)dx$$

- **Solution:** Simulate $X_1, \ldots, X_K \sim \pi$ then

$$\mathbb{E}_\pi[f(x)] \approx \frac{1}{K}\sum_{k=1}^{K} f(X_k)$$

- **How:** Markov Chain Monte Carlo (MCMC).
- **Problem:** What if π exhibits multimodality?



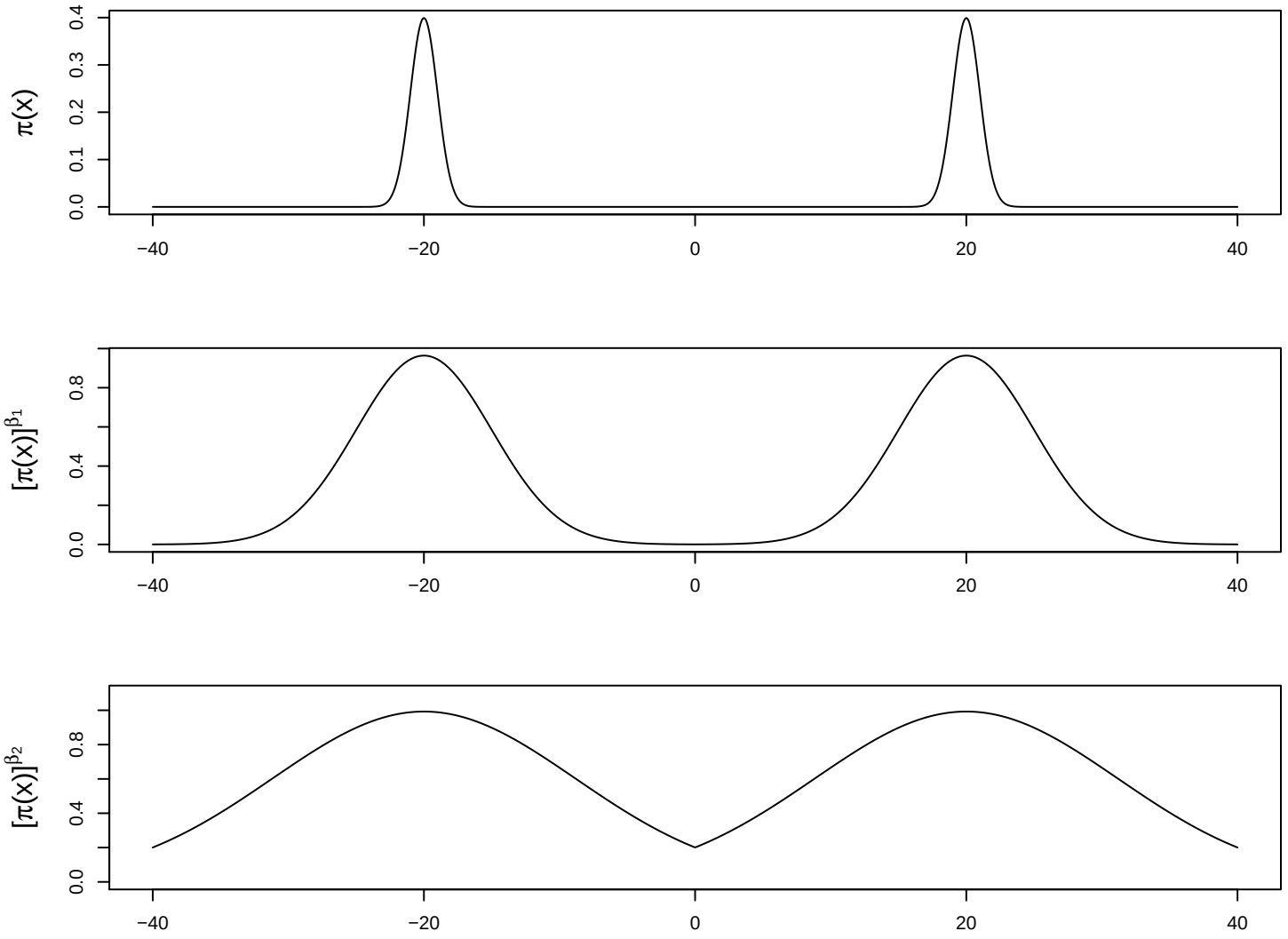
Multimodal Samplers

There is an array of methodology:

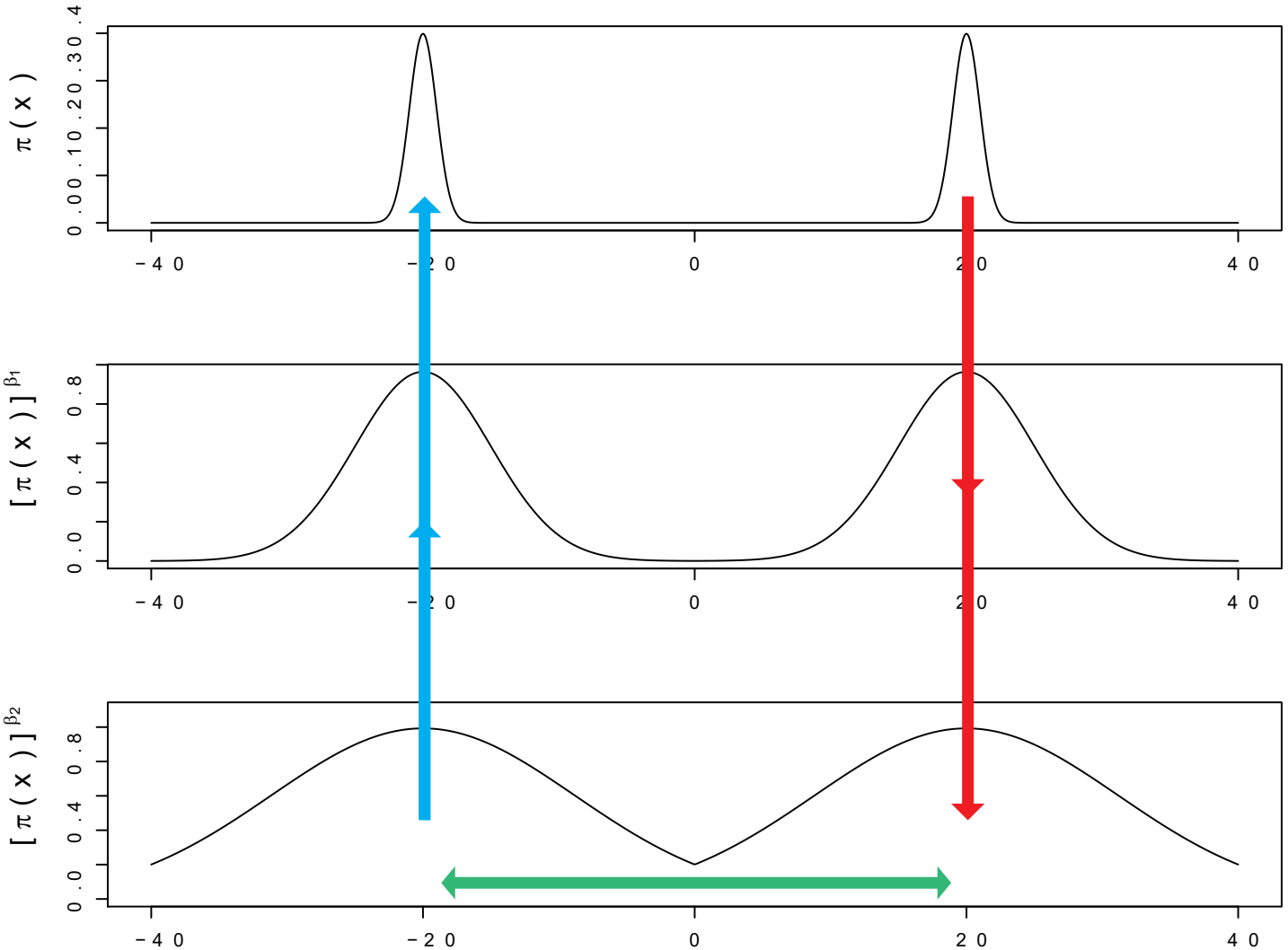
- Simulated (ST) and Parallel Tempering (PT), [Marinari and Parisi, 1992], [Geyer, 1991];
- Tempered Transitions, [Neal, 1996];
- Mode Jumping, [Tjelmeland and Hegstad, 2001];
- Equi-energy Sampler, [Kou et al., 2006];
- Repel-Attract, [Tak et al., 2016];
- Pseudo-extended Tempering, [Nemeth et al., 2017];
- Many More!

Most approaches rely on state space augmentation.

The Tempering Approach



The Tempering Approach



Simulated Tempering

Target density π on $\mathbb{R}^d$.

Take collection of inverse temperatures, $B = \{\beta_0, \ldots \beta_n\}$ with $1 = \beta_0 > \beta_1 > \ldots > \beta_n$.

Simulated tempering constructs a Markov chain, (X, β) on $\mathbb{R}^d \times B$ with invariant distribution $\tilde{\pi}$ with

$$\tilde{\pi}(X, \beta) \propto e^{K(\beta)}(\pi(x))^{\beta} \ .$$

for user-selected constants $K(\beta)$.

Note that a natural choice might select $K(\beta)$ such that $\tilde{\pi}$ assigns equal mass to each temperature, although calculation of such $K(\beta)$ values involve integrals which are typically intractable.

Simulated Tempering

Within temperature move according to a Metropolis-Hastings step with invariant density

$$\tilde{\pi}(x \mid \beta) \propto (\pi(x))^{\beta} \ .$$

Between temperature moves, from (x, β_j) to either (x, β_{j+1}) or (x, β_{j-1}) also according to an appropriate Metropolis-Hastings move. Eg

- With probability 1/2 decide to either propose to move from β_j to either β_{j-1} or β_{j+1}. Call the proposed value β_{new}.
- Accept this move with probability w.p.

$$1 \wedge \frac{\tilde{\pi}(x, \beta_{new})}{\tilde{\pi}(x, \beta)}$$

Simulated tempering and temperature distribution

In ST, the marginal distribution of β depends upon the choice of constants $\{K(\beta);\ \beta \in B\}$.

It seems reasonable to want this distribution to be close to uniform, in order to allow the chain to move easily through each temperature.

This involves setting

$$K(\beta) = -\log \int_{\mathbb{R}^d} (\pi(x))^\beta dx$$

which is difficult to achieve as this integral is typically intractable.

Parallel tempering is an alternative to ST which avoids the need to specify individual normalisation constants ($K(\beta)$) for each inverse temperature β.

Parallel Tempering (or MCMCMC)

PT stores a value at each inverse temperature at each iteration, ie $(X^{(0)}, X^{(1)}, \dots X^{(n)})$.

Within temperature moves are now carried out before (conditionally independently at each inverse temperature).

Between temperature swaps are now also proposed:

- Choose uniformly J from $\{1, 2, \dots n\}$;
- propose to change $(X^{(0)}, X^{(1)}, \dots X^{(J-1)}, X^{(J)}, \dots X^{(n)})$ to $(X^{(0)}, X^{(1)}, \dots X^{(J)}, X^{(J-1)}, \dots X^{(n)})$;
- accept this proposal with acceptance probability

$$1 \wedge \frac{(\pi(X^{(J-1)})^J (\pi(X^{(J)})^{J-1}}{(\pi(X^{(J-1)})^{J-1} (\pi(X^{(J)})^J} \ .$$

It is easy to check that this move is in detailed balance with the invariant density

$$\prod_{j=0}^{n} (\pi(x^{(j)})^{\beta_j} \ .$$

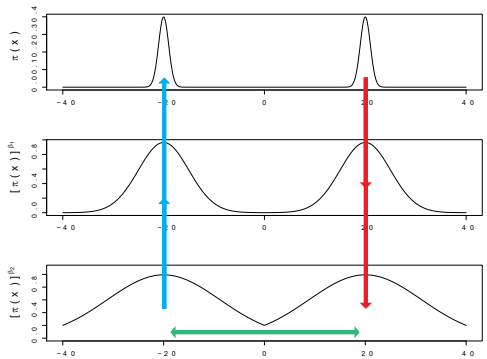
ST and PT

PT does not need user-defined normalisation constants. Therefore it is preferred in practice.

However it can be shown that its convergence properties are closely linked to that of the ST scheme with uniform distribution across temperatures.

Theoretical results are usually simpler to state for ST, so that's what I'll do in this presentation, assuming that equal mass is placed on each inverse temperature.

The temperature choice problem



Choice of B is key for the efficiency of the algorithm:

- β_n must be sufficiently small that within temperature MCMC moves can efficiently traverse the space.
- If between-temperature spacing is too large, then proposed jumps between neighbouring temperatures will be almost always rejected, leading to poor mixing.
- If within-temperature spacing is too large, then moves will be very small and it will take many moves (and substantial computational cost) to traverse between β_0 and β_n.

ST in high-dimensions

Studying ST in finite-dimensional contexts is complicated by the lack of tractability of the target density π.

However more can be said in high-dimensional asymptotic limits.

This is also particularly interesting for practical MCMC contexts.

Of course stylised examples need to be considered to get clean asymptotic results.

The choice of scalings for βs can be couched as an MCMC scaling problem.

The temperature scaling problem was considered for large d in [Atchadé et al., 2011, Roberts and Rosenthal, 2014].

High-dimensional limit: simplified setting

$$\pi_d(x) \;\propto\; e^{dK} \prod_{i=1}^{d} f(x_i)\,,$$

Choose $B(= B^{(d)})$ as follows: $\beta_0^{(d)} = 1$, and

$$\beta_{i+1}^{(d)} = \beta_i^{(d)} - \frac{\ell(\beta_i^{(d)})}{d^{1/2}}$$

for some fixed C^1 function $\ell : [0,1] \to \mathbb{R}_{>0}$, $1 \le i \le k(d)$ where χ is some threshold inverse temperature and

$$k(d) = \sup\{i : \beta_i^{(d)} \ge \chi\}\ .$$

$$\beta_{i+1}^{(d)} = \beta_i^{(d)} - \frac{\ell(\beta_i^{(d)})}{d^{1/2}}$$

So as d increases, $B^{(d)}$ is an increasingly dense discrete subset of $[\chi, 1]$. $|B^{(d)}| = k(d) + 1 = O(d^{1/2})$.

The optimal temperature spacing problem translates to asking what is the optimal choice of the function ℓ.

The factor $d^{-1/2}$ in the temperature spacing represents the intrinsic dimensional cost forcing neighbouring temperatures closer.

Natural to consider convergence of the β process to some stochastic process on $[\chi, 1]$.

Consider a joint process $(\beta_n^{(d)}, X_n)$, with $X_n \in \mathbb{R}^d$, $\beta_n^{(d)} \in B^{(d)}$. defined as follows.

We assume (unrealistically!) that the chain then immediately jumps to stationary at the new temperature, i.e. that mixing within a temperature is infinitely more efficient than mixing between temperatures.

The process $(\beta_n^{(d)}, X_n)$ is thus a Markov chain with stationary density

$$\tilde{\pi}(x, \beta) \;=\; e^{dK(\beta)} \prod_{i=1}^{d} f^{\beta}(x_i)\,,$$

where $K(\beta) = -\log \int f^{\beta}(x)dx$.

Let $I(\beta) = \mathrm{Var}_\beta\left[(\log f)(X)\right]$

Theorem

$\{\beta_n^{(d)}\}$ *speeded up by a factor of* d, *converges weakly as* $d \to \infty$ *to a diffusion limit* $\{Y_t\}_{t\geq 0}$ *satisfying*

$$dY_t = \left[2\ell^2\Phi\left(\frac{-\ell I^{1/2}}{2}\right)\right]^{1/2} dB_t$$

$$+\left[\ell(Y)\ell'(Y)\Phi\left(\frac{-I^{1/2}\ell}{2}\right) - \ell^2\left(\frac{\ell I^{1/2}}{2}\right)'\phi\left(\frac{-I^{1/2}\ell}{2}\right)\right]dt\,,$$

for Y_t *in* $(\chi, 1)$ *with reflecting boundaries at both* χ *and* 1.

For any function ℓ, the limiting diffusion has invariant density on $[\chi, 1]$ proportional to $\ell(y)^{-1}$.

The diffusion speed $\sigma^2(y) = 2\ell(y)^2\Phi\left(\frac{-\ell(y)I(y)^{1/2}}{2}\right)$ can be maximised pointwise (for each $y \in [\chi, 1]$, minimising its Dirichlet form and therefore its convergence rate. (Also minimises Monte Carlo error variances.)

Moreover we can decompose

$$\sigma^2(y) = \ell(y)^2 \times A(y)$$

where

$$A(y) = \mathbb{P}[\beta \text{ move accepted } | \beta = y] \ .$$

Theorem

The speed of this diffusion is maximised, and the asymptotic variance of all L^2 functionals is minimised, when the ℓ is chosen so that the asymptotic temperature acceptance probability at each and every temperature is equal to 0.234.

Comments on the basic convergence result

- Result shows how to tune the spacing between inverse temperatures in B.
- $O(d)$ convergence is a best case scenario under unrealistic conditions. $|B(d)| = O(d^{1/2})$, mixing time is $O(d)$.
- $I(\beta)$ acts as a temperature-dependent *friction*. The lower its value, the more efficient mixing of the temperature can be.
- Under weaker conditions ([Atchadé et al., 2011]) not requiring instantaneous mixing of the within-temperature moves, we can get that the 0.234 strategy optimises local Expected Squared Jumping Distance, $ESJD_\beta$:

$$ESJD_\beta = \mathbb{E}[(\beta_{n+1} - \beta_n)^2 \mid \beta_n = \beta] \ .$$

Critique of ST and PT

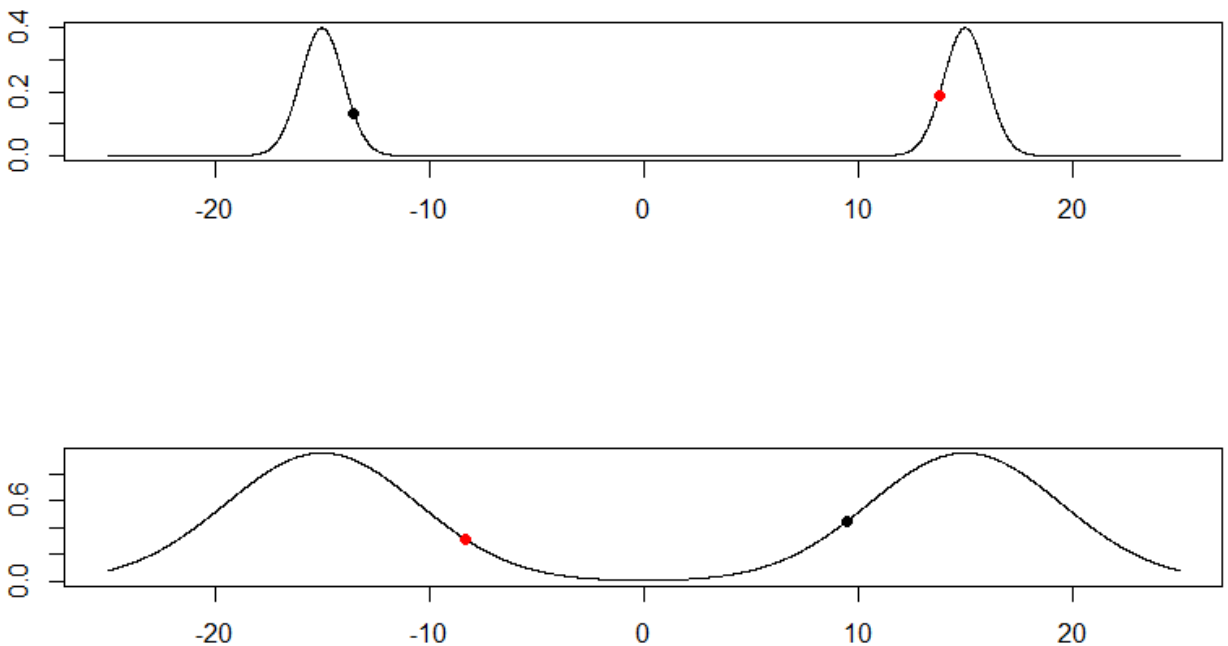
- The Good: Mixing at the hot state.
- The Bad: Costly procedure - number of temperatures increases at least $O(d^{1/2})$ leading to mixing which is at best $O(d)$. (Friction is too large.)
- The Very Bad: Mode mass inconsistency - the proportion of probability mass within each mode is not reserved under tempering.

The Good can help us discover modes. But what about the Bad and Very Bad?

Mitigating the Bad? QuanTA- Accelerated Mixing

[Tawn and Roberts, 2019]
ST and PT inter-temperature moves are inhibited as within mode dispersion varies as temperature changes

QuanTA preserves $(x - \mu_x)\beta^{1/2}$ when updating β, where μ_x denotes the *nearest* mode centre:



Effective? QuanTA- Optimal Scaling

Theorem (Optimal Scaling for the QuanTA Algorithm)

Target, $\pi(x) \propto \prod_{i=1}^{d} f(x_i)$. *Temp spacing* $\beta' - \beta = \epsilon = \ell/d^{1/2}$. *Then the* $ESJD_\beta$ *satisfies*

$$\lim_{d\to\infty} d(ESJD_\beta) = 2\ell^2\Phi\left(-\frac{\ell}{\sqrt{2}}[J(\beta)]^{1/2}\right),$$

where $J(\beta)$ *is an explicit moment of* $X, f(X)$ *and* $f'(X)$.

Optimisation wrt ℓ *induces an associated 0.234 rule.*

Theorem (Cold Temperature Scalings)

Under the setting of Theorem 3 then for large β

$$\textit{friction} = J(\beta) = \mathcal{O}\left(\frac{1}{\beta^k}\right),$$

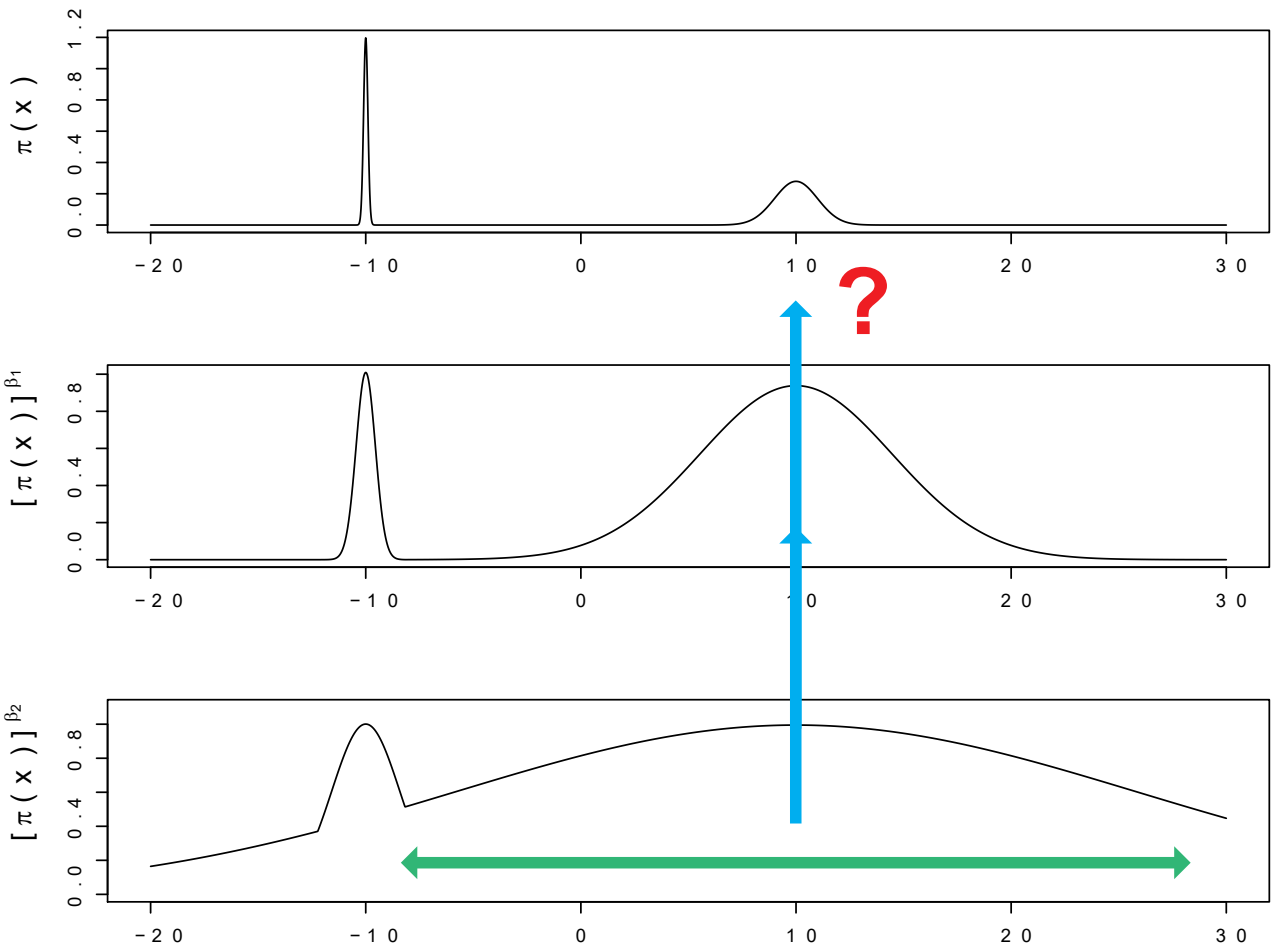
where $k > 2$.

This induces an optimising value $\hat{\ell}$ *such that* $\hat{\ell} = \mathcal{O}\left(\beta^{\frac{k}{2}}\right)$, *showing that at the colder temperatures QuanTA permits higher order behaviour than the standard PT scheme which has* $\hat{\ell} = \mathcal{O}(\beta)$.

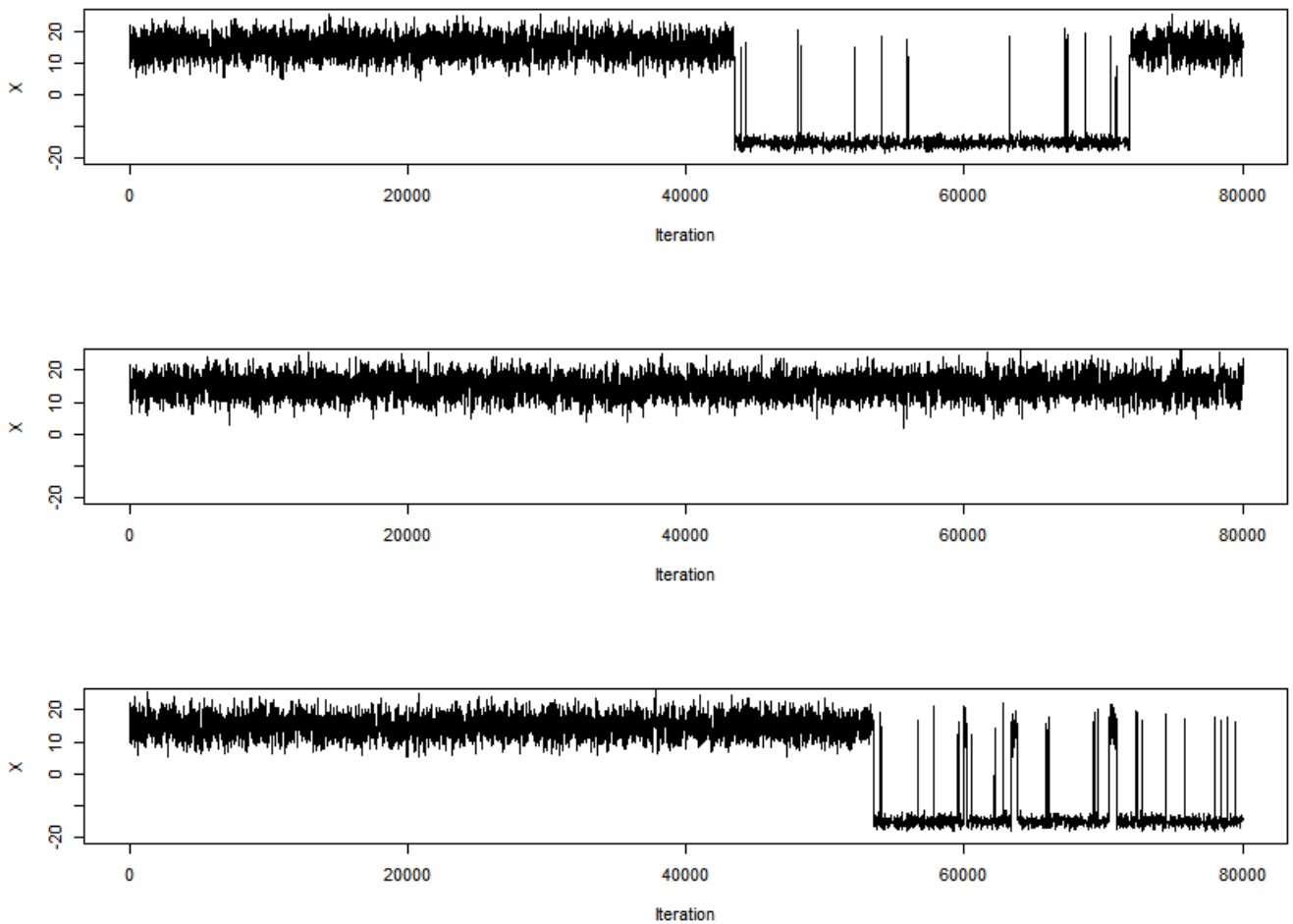
Useful but only guarantees improvements when temperature is cold.

Mass Inconsistency: The very bad

Power-tempering **torpid mixing**, [Woodard et al., 2009]:

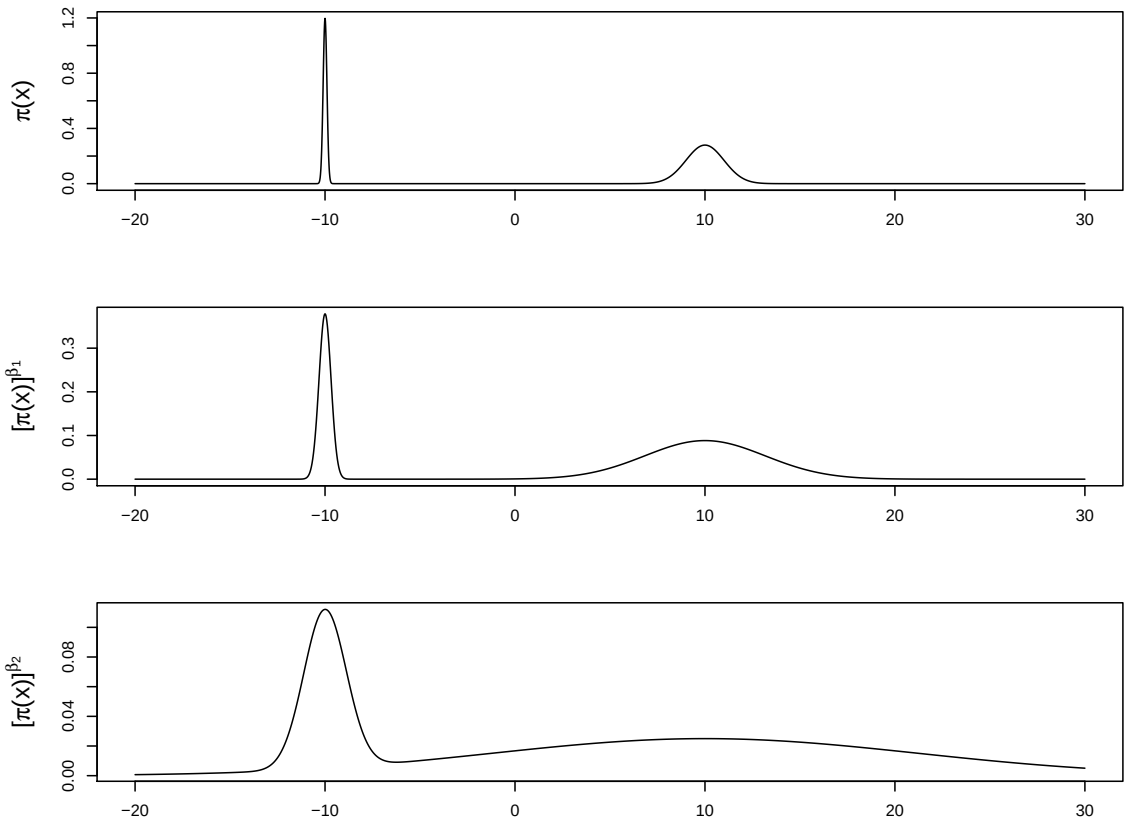


Five-dimensional example



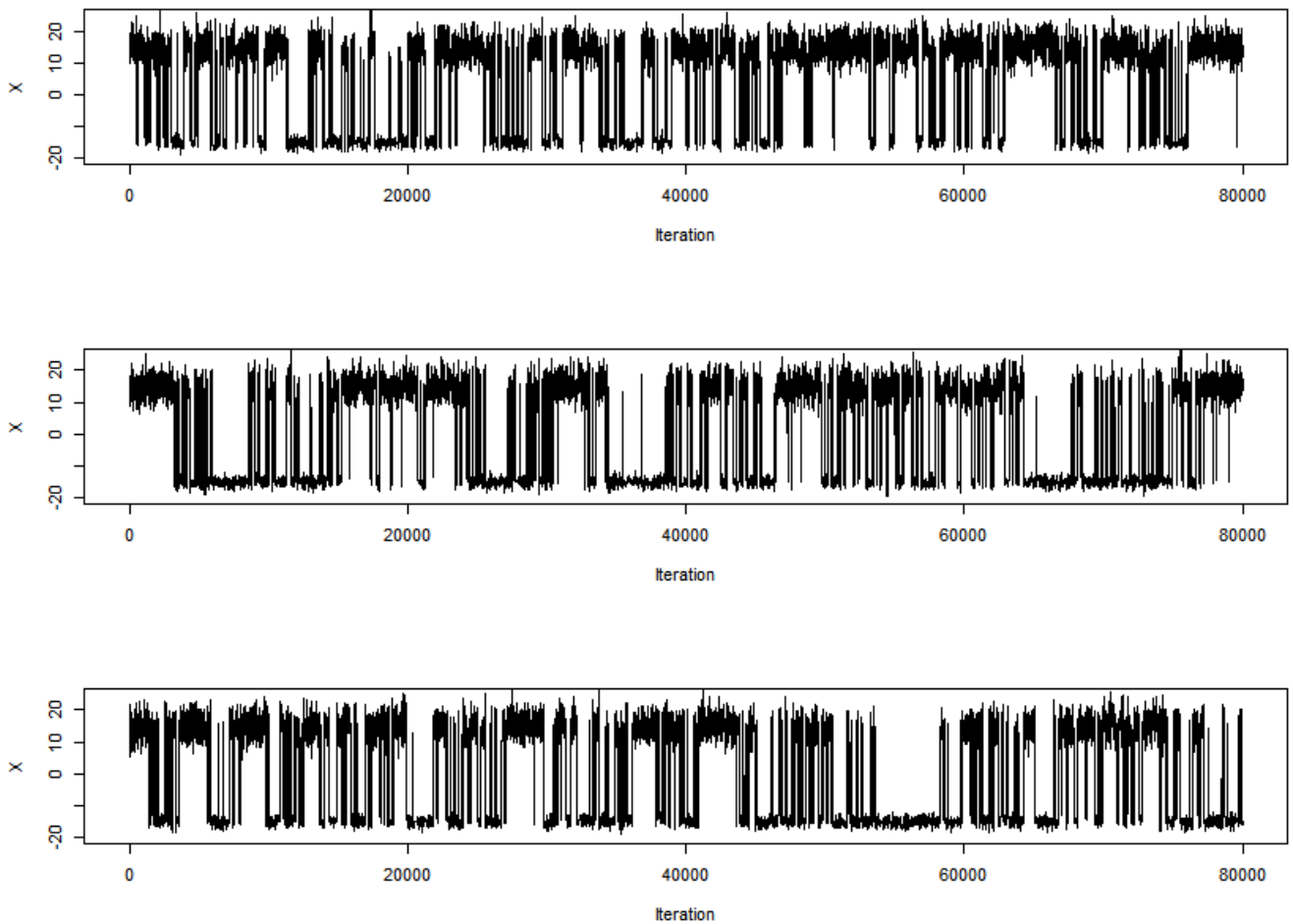
Attempted fix? HAT-Hessian Adjusted Tempering

[Tawn et al., 2018] introduces: $\pi_\beta^{\mathsf{HAT}}(x) = [\pi(x)]^\beta [\pi(\mu_{x,\beta})]^{1-\beta}$

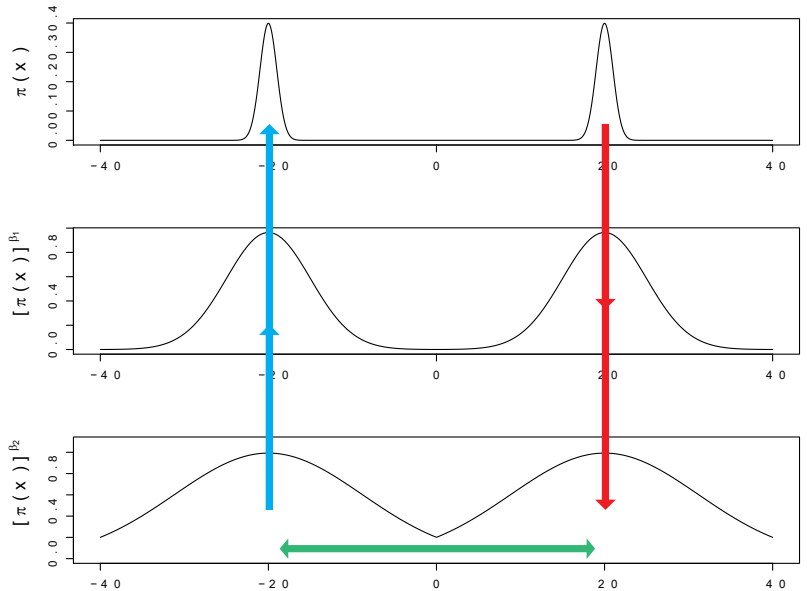


The HAT method approximately preserves mass within each mode as temperature varies.

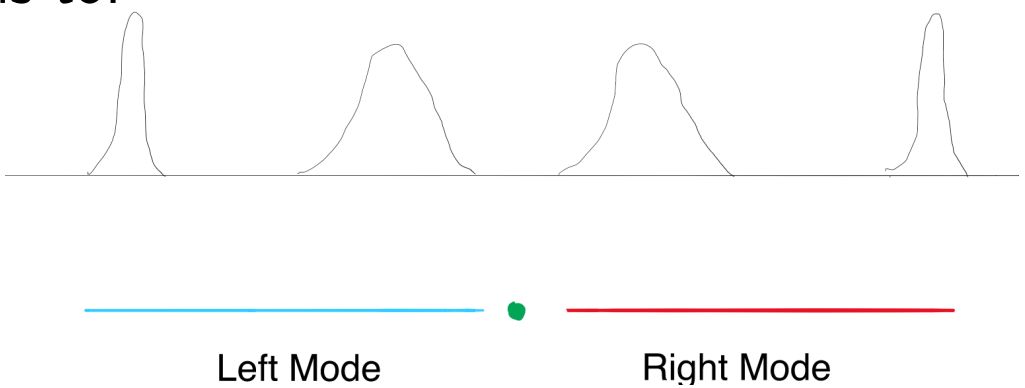
HAT improves mixing between modes



Theory: transforming inverse temperature and mode index



Transform this to:

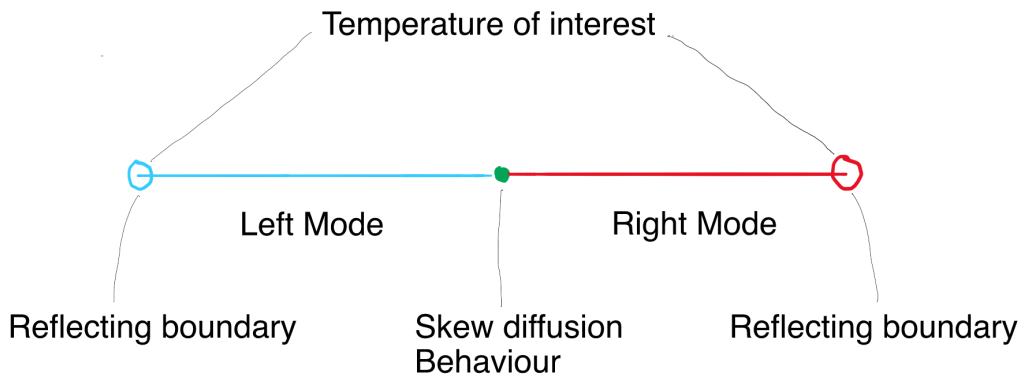


Inverse temperature and mode index, high-d limit

Similar stylised set up as before but with heteroskedasticity of within-mode variation.

Theorem

Suitably scaled and with time speed up of d, HAT converges to a skewed Brownian motion with reflecting boundaries.



Note that friction could be different in each mode and location of reflection points is not generally symmetric.

For more than two modes, the limit is a transformed Walsh Brownian motion.

When modes are not symmetric

- Weight-stabilising work in [Tawn and Roberts, 2018], [Tawn et al., 2018].
- Transformation-aided accelerated mixing in [Tawn and Roberts, 2019].
- Modal-skewness still a major problem: this affects both QUANTA and especially HAT (weights not stable across temperatures.

Tempering and skewness

How does tempering affect skewness?

Tempering and skewness

How does tempering affect skewness?

For small β (high temperature) skewness is increased.

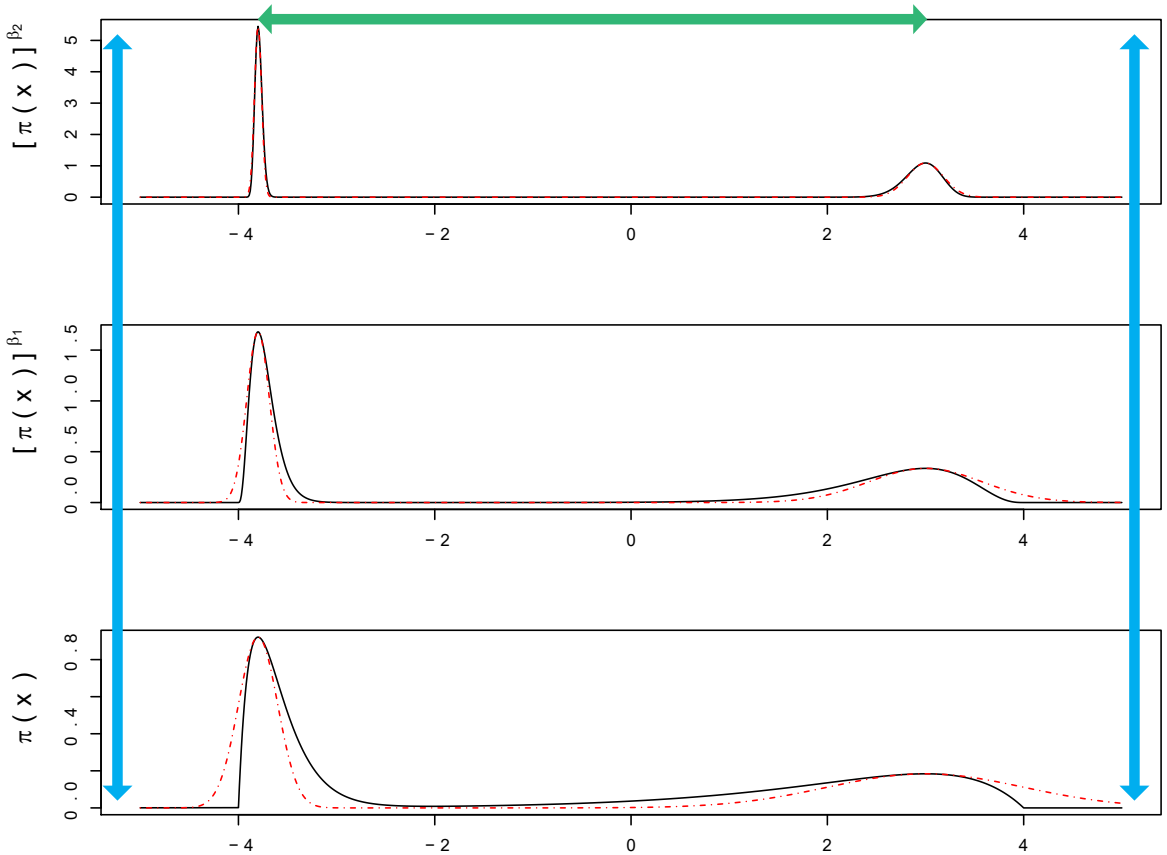
For large β (low temperature) skewness is decreased.

On the other hand if we had completely symmetric, Gaussian-looking modes which we know the location of, we don't need tempering. Instead we can construct highly effective MCMC independence sampler mode hopping moves.

But such an approach would be hopeless in high-d settings with skewed modes.

ALPS- The Heuristic

Consider $B = \{1 = \beta_0 < \beta_1 < \ldots < \beta_n\}$, ie colder temperatures.



At the cold temperature, use independence sampler MCMC move to jump from mode to mode.

ALPS overview

Annealed Leap Point Sampler,
[Roberts et al., 2020, Tawn et al., 2021].

We need

1. a mode-finding algorithm for finding modes. (Typically just use tempering to hot temperatures.
2. a mode-hoping independence sampler algorithm for the very cold temperature. (Details not here, but uses ideas related to the QUANTA algorithm.)
3. Temperature schedule for going cold and how cold?

The ALPS methodology can be carried out in conjunction with the HAT and QUANTA strategies.

ALPS - how cold?

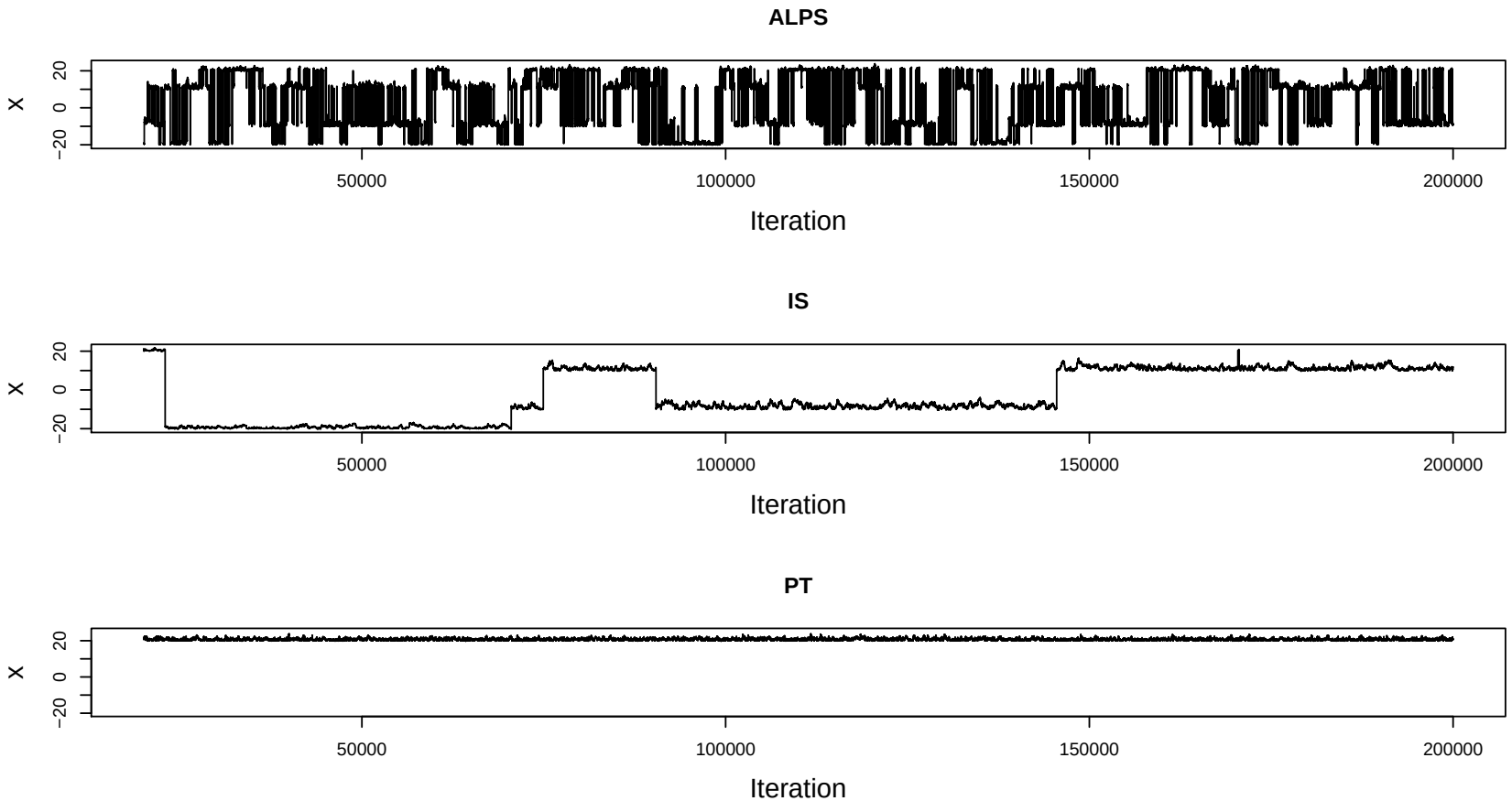
Theorem (Scaling the Coldest Temperature Level)

With β_{max} denoting the coldest temperature level, then as $d \to \infty$ in order to induce a non-degenerate acceptance rate for the β_{max} mode-leaping independence sampler then one must choose $\beta_{max} = \ell d = \mathcal{O}(d)$. Furthermore, if $\beta_{max} = \ell d$, then in the limit as $d \to \infty$ the expected acceptance rate of the leap-mode independence sampler is given by

$$\mathbb{E}_{\pi_{\beta_{max}}}\left(\mathbb{P}(\mathit{Accept})\right) = 2\Phi\left(-\frac{1}{\sqrt{2}}\sqrt{\frac{15h'''(0)^2}{36\ell(-h''(0))^3}}\right)$$

where $h(x) = \log f(x)$ and Φ is the CDF of a standard Gaussian.

ALPS - Empirical Study



Also shows excellent properties on a notoriously hard Bayesian multimodal posterior from a seemingly unrelated regressions model.

ALPS - Theory

In [Tawn et al., 2021] show that (in a stylised setting) $B(d)$ needs to be $O(d^{1/2} \log d)$.

Also $B(d)$ can be improved to $O(d^{1/2})$ if it is used in conjunction with QUANTA.

In [Roberts et al., 2020] establish weak convergence result in the temperature $\times$ mode space.

Thus ALPS with QUANTA can be $O(d)$ despite having to explore much colder temperatures.

Further work

- Can we obtain results which relax the strong assumptions on *within-mode* mixing?
- All this theory relies on smoothness (eg C^3 for ALPS result). What can be said if we relax these assumptions?
- More efficient, between-temperature dynamics, eg see [Faizi et al., 2020]. Can we obtain $O(d^{1/2})$ temperature mixing?
- Extensions to the robustness of ALPS for practitioners.

References I

Atchadé, Y. F., Roberts, G. O., and Rosenthal, J. S. (2011).
Towards Optimal Scaling of Metropolis-Coupled Markov chain Monte Carlo.
Statistics and Computing, 21(4):555–568.

Faizi, F., Buigues, P. J., Deligiannidis, G., and Rosta, E. (2020).
Simulated tempering with irreversible gibbs sampling techniques.
The Journal of Chemical Physics, 153(21):214111.

Geyer, C. J. (1991).
Markov chain Monte Carlo Maximum Likelihood.
Computing Science and Statistics, 23:156–163.

Kou, S., Zhou, Q., and Wong, W. H. (2006).
Equi-energy Sampler with Applications in Statistical Inference and Statistical Mechanics.
The Annals of Statistics, pages 1581–1619.

Marinari, E. and Parisi, G. (1992).
Simulated Tempering: a New Monte Carlo Scheme.
EPL (Europhysics Letters), 19(6):451.

Neal, R. M. (1996).
Sampling from Multimodal Distributions using Tempered Transitions.
Statistics and Computing, 6(4):353–366.

Nemeth, C., Lindsten, F., Filippone, M., and Hensman, J. (2017).
Pseudo-extended Markov Chain Monte Carlo.
ArXiv e-prints.

References II

Roberts, G. O. and Rosenthal, J. S. (2014).

Minimising MCMC Variance via Diffusion limits, with an Application to Simulated Tempering.
The Annals of Applied Probability, 24(1):131–149.

Roberts, G. O., Rosenthal, J. S., and Tawn, N. G. (2020).

Skew brownian motion and complexity of the alps algorithm.
arXiv preprint arXiv:2009.12424.

Tak, H., Meng, X.-L., and van Dyk, D. A. (2016).

A Repulsive-Attractive Metropolis Algorithm for Multimodality.
arXiv preprint arXiv:1601.05633.

Tawn, N. G. and Roberts, G. O. (2018).

Optimal temperature spacing for regionally weight-preserving tempering.

Tawn, N. G. and Roberts, G. O. (2019).

"accelerating parallel tempering: Quantile tempering algorithm (quanta)".
"Advances in Applied Probability—, 51(3):802–834.

Tawn, N. G., Roberts, G. O., and Rosenthal, J. S. (2018).

Weight preserving simulated tempering.
Accepted, to appear in Statistics and Computing.

Tawn, N. G., Roberts, G. O., and Rosenthal, J. S. (2021).

The annealed leap point sampler (alps).
to appear.

References III

Tjelmeland, H. and Hegstad, B. K. (2001).
Mode Jumping Proposals in MCMC.
Scandinavian Journal of Statistics, 28(1):205–223.

Woodard, D. B., Schmidler, S. C., and Huber, M. (2009).
Sufficient Conditions for Torpid Mixing of Parallel and Simulated Tempering.
Electronic Journal of Probability, 14:780–804.