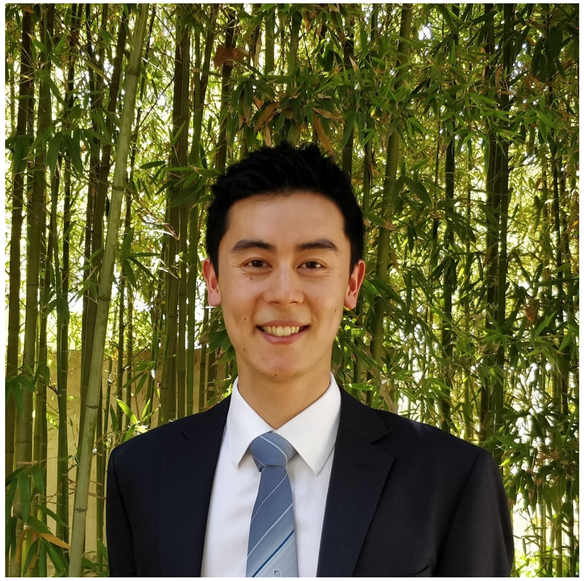
Comparison Theorems for Practical Slice Sampling

Kavli Institute for Cosmology, 5 November 2024

Sam Power, University of Bristol

joint work with: D. Rudolf (Passau), B. Sprungk (Freiberg), A.Q. Wang (Warwick)



feel free to stop me at any point

Markov chain Monte Carlo

- "target" distribution π on $\mathbf{R}^d$
- want samples from π to answer questions
- MCMC: use *iterative* strategy to obtain *approximate* samples

$$X_0 \to X_1 \to X_2 \to \cdots \to X_T \overset{d}{\approx} \pi$$

$$\frac{1}{T} \sum_{0<t\leqslant T} f(X_t) \approx \int \pi(\mathrm{d}x) f(x) =: \pi(f)$$

Some challenges in MCMC

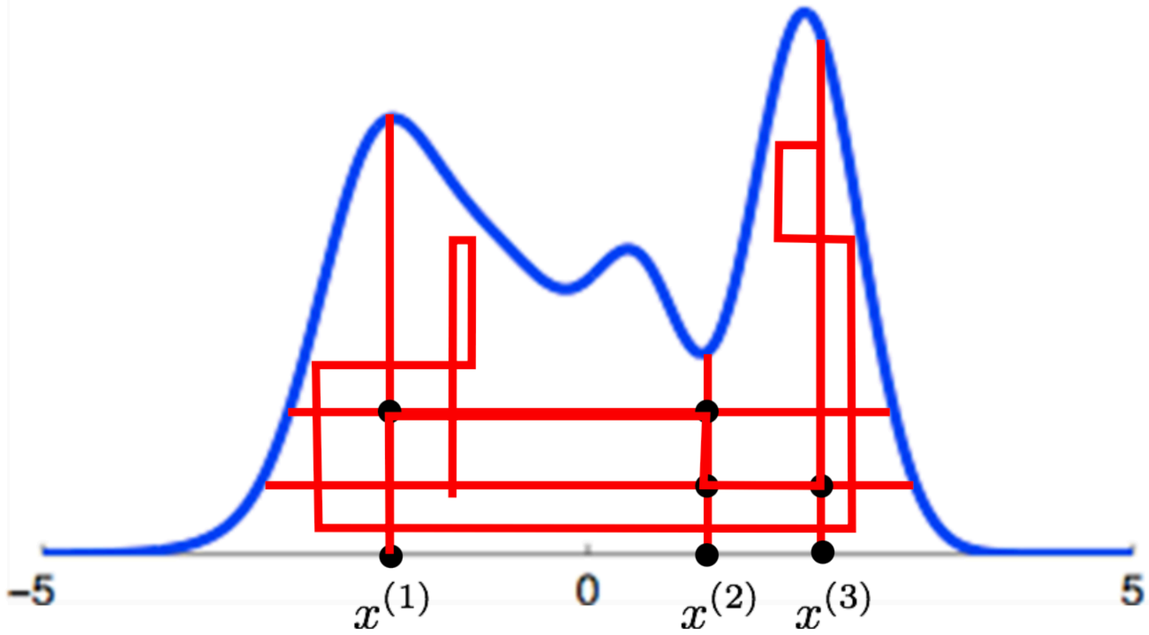
- designing effective Markov kernels
- obtaining and using useful information about π
- tuning of algorithm hyperparameters (step-size, etc.)

Slice Sampling for MCMC

- assume that we can only compute density of π (up to a constant)
- trick: sampling from π is equivalent to sampling *uniformly* under its graph
- mathematically: $\Pi\left(\mathrm{d}x, \mathrm{d}t\right) = \mathbf{1}\left[0 \leqslant t \leqslant \pi(x)\right]\ \mathrm{d}x\,\mathrm{d}t$

Slice Sampling

Define a Markov chain that samples uniformly from the area beneath the curve. This means that we need to introduce a “height” into the MCMC sampler.



(from slides of Ryan Adams)

Some Useful Definitions

- write $\bar{\pi} = \max\left\{\pi(x) : x \in \mathcal{X}\right\}$, $\mathsf{T} = \left[0, \bar{\pi}\right]$ (‘heights’)
- write $\mathsf{G}(t) = \{x : \pi(x) \geqslant t\}$ for the super-level set (‘slice’)
- write $\nu_t = \mathsf{Unif}\left(\mathsf{G}(t)\right)$ (‘on-slice distribution’)
- write $m(t) = \mathrm{Leb}\left(\mathsf{G}(t)\right)$ (‘mass function’)

Slice Sampling: Algorithm

- want to generate sequence $\left\{ \left(X_n, T_n \right) : n \geqslant 1 \right\}$
- so,
  - given $X_{n-1} = x$, sample $T_n \sim \text{Unif} \left(\left[0, \pi(x) \right] \right)$
    - (sample a height)
  - given $T_n = t$, sample $X_n \sim \nu_t$
    - (sample uniformly ‘on the slice’)

Slice Sampling: Qualitatively

- under mild conditions, gives an ergodic, π-invariant Markov chain
  - ⇝ fit for purpose in MCMC
- under still mild conditions, is even *exponentially* convergent
  - ⇝ bonus results, e.g. Markov chain CLT
  - surprisingly hard to break

Slice Sampling: Dualities

- in principle, we could look at the convergence properties of
  1. $\{X_n\}$ alone (of key interest)
  2. $\left\{ \left(X_n, T_n\right) \right\}$ jointly
  3. $\{T_n\}$ alone (univariate)
- in fact, all three processes will equilibrate ‘identically’

Slice Sampling: Invariances

- consider two target distributions π_1, π_2 with mass functions m_1, m_2
- if $m_1 = m_2$, then convergence profiles will be *identical*
  - follows from considering $\{T_n\}$ chain
- convergence is thus (nominally) agnostic to
  - { dimension, linear change of variables, ‘rearrangement’ of mass, … }
- conclusion: can ‘WLOG’ very freely!
- consequence: convergence is ideally quite robust

Slice Sampling: Quantitatively

- for specific π, strong quantitative theory available
  - π spherically-symmetric, log-concave $\rightsquigarrow$ 'relaxation time' $\sim$ dim
  - π multivariate Student-t $\rightsquigarrow$ 'relaxation time' $\sim \text{dim}^2$
  - (other explicit examples can be studied)
- noteworthy: barely slowed down by heavy tails; *rare* property

Implementing Slice Sampling

“given $T_n = t$, sample $X_n \sim \nu_t$”

– Sam Power, Slide 8

Life on the Slice

- recall $\nu_t = \mathrm{Unif}\left(\mathsf{G}(t)\right)$, $\mathsf{G}(t) = \{x : \pi(x) \geqslant t\}$
- if $\mathsf{G}(t)$ is a { ball, box, simplex, … }, then sampling from ν_t is fine
- if not, then we have a new problem

Hybrid Slice Sampling

- instead of
  - "given $T_n = t$, sample $X_n \sim \nu_t$"
- do
  - given $X_{n-1} = x, T_n = t,$
  - sample $X_n \sim \text{MCMC}\left(x \to x'; \text{target} = \nu_t\right)$
    - (call this Markov kernel H_t)

Properties of HSS

- this new algorithm …
  - is still implementable
  - still has the right qualitative behaviour (invariance, ergodicity, …)
  - is not as (statistically) efficient as the ‘ideal’ Slice Sampler
- how do we quantify the cost of approximation?
  - ⇝ Markov chain comparison theory

A word on MCMC theory

- it is (IMO) nice to have an idea of how long it takes for MCMC to converge
- generally, we first seek to understand complexity w.r.t. relevant features
  - { dimension, curvature, dependence, … }
- reasonably useful to compare algorithms; “*which MCMC should I use?*”
- relatively rare (AFAIK) to use such results to e.g. set algorithm runtime

On Details

- technical details about the comparison framework will be suppressed
- general strategy: "L^2 / Hilbert space / Dirichlet form" approach
- "K converges to π at rate γ" $\sim$ "spectral gap" $\geqslant \gamma$, reversible + positive
- this notion of convergence is reasonably strong
- after digesting the definitions, the techniques are relatively easy to use

General Strategy

- let U = Ideal Slice Sampling, H = Hybrid Slice Sampling
- to quantify how quickly H converges to π, we will study
  - how quickly U converges to π, and
  - how well H approximates U
    - (rather, how well H_t approximates ν_t)

Realities of Comparison Theory

- to say that H gives a good Markov chain, we are arguing that
  - U gives a good Markov chain, and
  - H is a good approximation of U
- in principle, H could fail to approximate U well, but still work well
  - our analysis would fail to capture this (∃ examples)

Some Warm-Up Results

- in complete generality, U dominates H
  - $\rightsquigarrow$ all else being equal, always prefer ideal chain
- for experts: any Metropolis(~~-Hastings~~) kernel can be written as such an H
  - $\rightsquigarrow$ all such chains are automatically dominated by (ideal) SS
- conversely: we are interested in when H is **almost as good** as U

A Generic Result

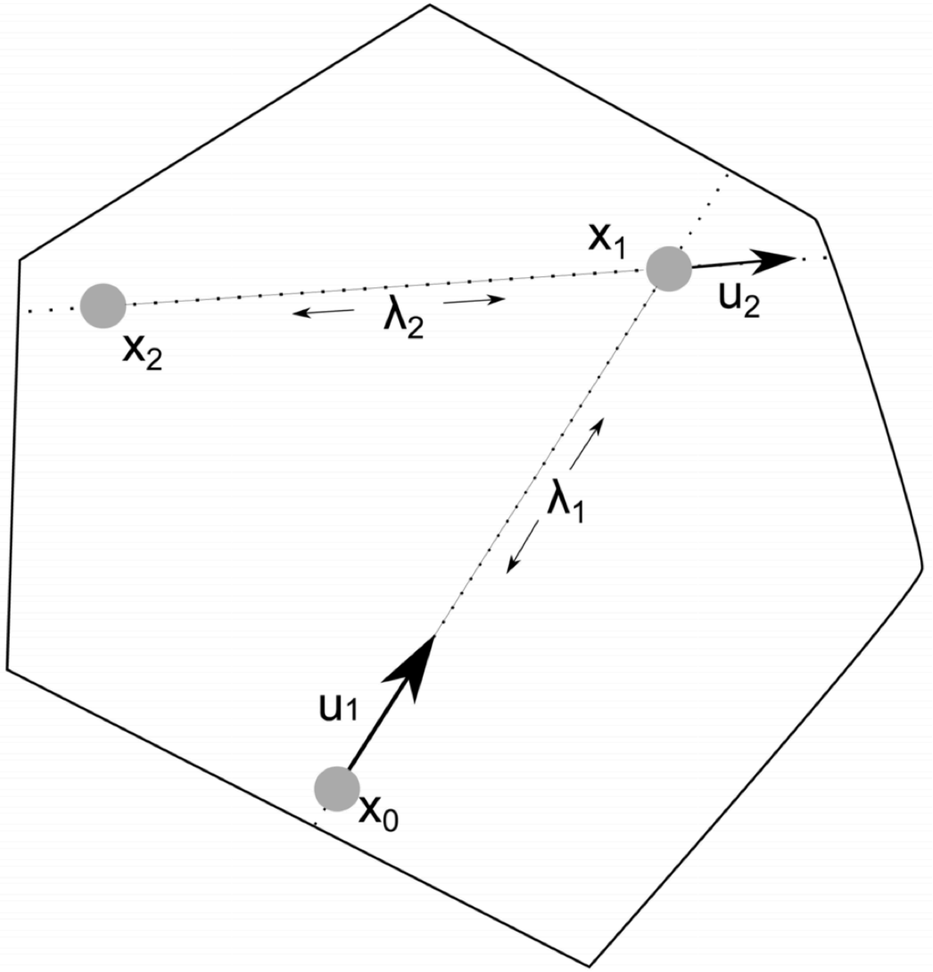
- suppose that
  - for $t \in \mathsf{T}$, the on-slice kernel H_t converges to ν_t at rate $\sigma_t > 0$, and
  - $\sigma_H := \inf_{t \in \mathsf{T}} \sigma_t > 0$
- the convergence rates of H and U then satisfy

$$\gamma_H \gtrsim \sigma_H \cdot \gamma_U,$$

  - interpretation: HSS is at most a factor σ_H ‘worse’ than ideal SS

Case Study: Hit-and-Run on the Slice

- simple method for sampling uniform distributions on convex body G
- at $X_{n-1} = x$,
  - sample $U_n \sim \text{Unif}\left(\mathbb{S}^{d-1}\right)$
  - look at $\left(x + U_n \mathbf{R}\right) \cap \text{G}$
  - move uniformly along this line segment
  - call new location X_n



(diagram from “optGpSampler” paper)

Convergence of Hit-and-Run

- the following is a theorem of Lovász-Vempala from 2004
- let $\mathsf{G} \subset \mathbf{R}^d$ be convex, containing a ball of radius r_{G}, and contained in a ball of radius R_{G}; write $\kappa_{\mathsf{G}} := R_{\mathsf{G}}/r_{\mathsf{G}} \geqslant 1$.
- Then, for some universal $c > 0$, it holds that

$$\gamma_{\mathrm{H\&R}} \geqslant c \cdot d^{-2} \cdot \kappa_{\mathsf{G}}^{-2}.$$

- high dimension is hard, inhomogeneity of scales is hard

Hit-and-Run Hybrid Slice Sampling

- if π has convex super-level sets, then results of LV give us a bound

$$\sigma_t \geqslant c \cdot d^{-2} \cdot \kappa_{\mathsf{G}(t)}^{-2}$$

- interpretation: life is good if super-level sets $\mathsf{G}(t)$ are well-conditioned
  - (if not: worse, though not disastrous)

Well-Conditioned Level Sets

- let $V : \mathbf{R}^d \to \mathbf{R}$ be m-strongly convex and L-smooth
  - i.e. eigs $\left(\text{Hess} V(x)\right) \in [m, L]$
  - write $\kappa_V = L/m \geqslant 1$
- let density π have the form $\pi(x) = \text{decreasing}\left(V(x)\right)$
- then for all t, it holds that $\kappa_{\text{G}(t)} \leqslant \sqrt{\kappa_V}$.

Some Applications

- if π has this form, then $\gamma_H \gtrsim d^{-2} \cdot \kappa_V^{-1} \cdot \gamma_U$
  - H&R-HSS is only worse than ideal SS by factor $d^2 \cdot \kappa_V$
- if e.g. $\pi \propto \exp(-V)$,
  - combine with works on ideal SS, $\rightsquigarrow$ relaxation time of $\lesssim d^3 \cdot \kappa_V$
- if e.g. π is multivariate Student-t, then $\kappa_V = 1, \sigma_H \gtrsim d^{-2}$
  - combine with earlier work, $\rightsquigarrow$ relaxation time of $\lesssim d^4$

Some Recap

- slice sampling performs well in theory, and in practice (when possible)
- hybrid slice sampling performs well in practice, is *typically* possible,
  - … and we provide here some theory to support this
- comparison principles: i) is U good?, ii) is H similar enough to U?
- generally, $H \preceq U$,
  - … but if $H_t \geq \sigma_H \cdot \nu_t$, then $H \geq \sigma_H \cdot U$.

Some Closing Remarks

- today: exponential rates, Hit-and-Run on the slice
- in the paper: slower-than-exponential rates, other examples of on-slice kernels, stepping-out and shrinkage, ‘generalised’ slice sampling with different reference measures, ….
- theoretical framework is very robust to which on-slice kernels are used
- actually, theoretical framework is much more general than slice sampling
  - “Markov chain comparison”, “weak Poincaré inequalities”, …

—Bonus Material—

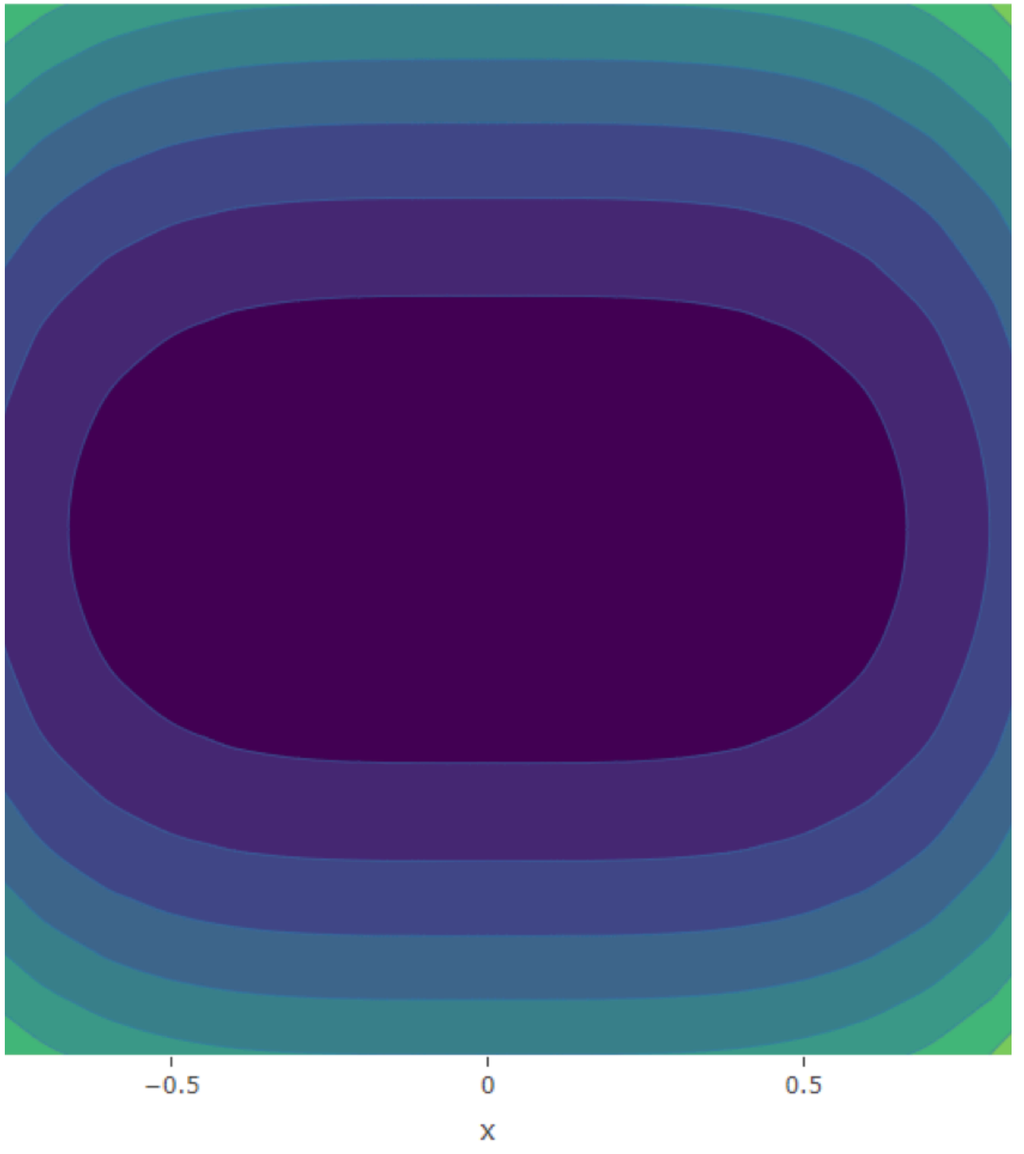
Advanced Applications

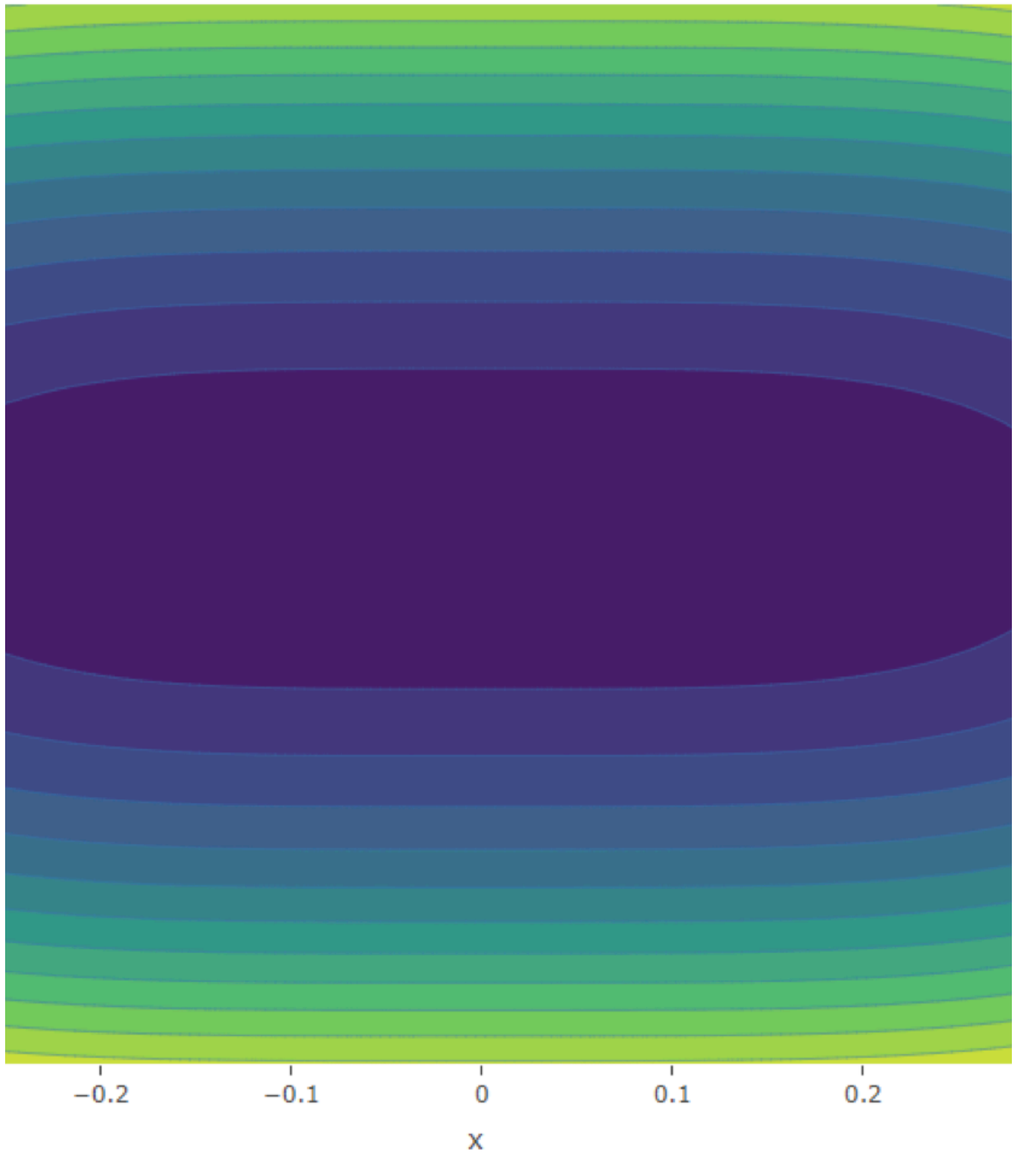
- let $1 \leqslant p_2 \leqslant p_1$, $1 \leqslant q_1 \leqslant q_2$, and suppose that

$$\|x\| \sim 0^+ \implies \|x\|^{p_1} \lesssim V(x) \lesssim \|x\|^{p_2}$$

$$\|x\| \sim \infty \implies \|x\|^{q_1} \lesssim V(x) \lesssim \|x\|^{q_2}$$

- if $p_1 = p_2$, $q_1 \neq q_2$, then convergence rate decays quasi-exponentially
- if $p_1 \neq p_2$, then convergence rate decays only polynomially
- message: in this case, bulk behaviour matters more than tail behaviour


−0.5
0
0.5
x


−0.2
−0.1
0
0.1
0.2
x

$$\kappa_{\mathsf{G}}(t) \leq \begin{cases} c_{\kappa}^{-} \cdot \left(\log\left(\frac{1}{t}\right)\right)^{\theta} & 0 < t \leq \exp(-1); \\ c_{\kappa}^{+} \cdot \left(\log\left(\frac{1}{t}\right)\right)^{-\vartheta} & \exp(-1) \leq t < 1; \end{cases}$$

with $\theta = \frac{1}{q_1} - \frac{1}{q_2}$, $\vartheta = \frac{1}{p_2} - \frac{1}{p_1}$, *and such that the mass function satisfies*

$$m(t) \leq c_m \cdot \left(\log\left(\frac{1}{t}\right)\right)^{d/r}$$

with $r = q_1$. *By application of Proposition 40, we see that for* $p_1 = p_2$, *there holds a WPI with*

$$\beta(s) \leq c^{(1)} \cdot \exp\left(-c^{(2)} \cdot s^{\frac{q_1 \cdot q_1}{q_2 - q_1}}\right),$$

whereas for $p_1 > p_2$, *one instead obtains a WPI with*

$$\beta(s) \leq c^{(3)} \cdot s^{-\left(1+\frac{d}{q_1}\right) \cdot \frac{p_1 \cdot p_2}{p_1 - p_2}}.$$

Quantitative Mode Separation

Definition 33. *Fix a density function* $\varpi : \mathbb{R} \to \mathsf{T}$*, and let* $0 < t_1 \leq t_2$ *be elements of* T*. Say that* ϖ *is* (t_1, t_2)-bimodal *if*

- *for all* $t \in \mathsf{T} \setminus [t_1, t_2)$*, the super-level set* $\mathsf{G}(t)$ *consists of a single interval, and*
- *for all* $t \in [t_1, t_2)$*, the super-level set* $\mathsf{G}(t)$ *consists of a pair of disjoint sub-intervals* $\mathsf{G}(t) = \mathsf{G}_1(t) \sqcup \mathsf{G}_2(t)$ *such that commonly-labelled sub-intervals are nested, i.e. for* $i = 1, 2$ *and* $t_1 \leq s \leq t < t_2$*, there holds the inclusion* $\mathsf{G}_i(t) \subseteq \mathsf{G}_i(s)$*.*

Moreover, given a (t_1, t_2)*-bimodal density* ϖ*, define* $\delta_\varpi : \mathsf{T} \to [0, \infty)$ *by*

$$\delta_\varpi(t) = \begin{cases} \mathsf{dist}\left(\mathsf{G}_1(t), \mathsf{G}_2(t)\right) & t \in [t_1, t_2) \\ 0 & \text{otherwise,} \end{cases}$$

where $\mathsf{dist}\left(\mathsf{G}_1(t), \mathsf{G}_2(t)\right) := \inf\left\{|x - y| : x \in \mathsf{G}_1(t), y \in \mathsf{G}_2(t)\right\}$*, and write* $\Delta_\varpi := \sup\left\{\delta_\varpi(t) : t \in \mathsf{T}\right\}$*.*

Stepping-Out and Shrinkage

Assumption 2. *Let ϖ be a (t_1, t_2)-bimodal density and $h > \Delta_\varpi$ be a stepping-out parameter.*

Under the previous assumption, for $t \in \mathsf{T}$, define the 'stepping-out and shrinkage' on-slice kernel with parameter h by

$$H_t(x, A) = \lambda(t) \cdot \nu_t(A) + \big(1 - \lambda(t)\big) \cdot \Big[\mathbf{1}_{\mathsf{G}_1(t)}(x) \cdot \nu_{t,1}(A) + \mathbf{1}_{\mathsf{G}_2(t)}(x) \cdot \nu_{t,2}(A)\Big],$$

for $x \in \mathbb{R}$, $A \in \mathcal{B}(\mathbb{R})$, where

$$\lambda(t) := \begin{cases} \frac{h - \delta_\varpi(t)}{h} \cdot \frac{m(t)}{m(t) + \delta_\varpi(t)} & t \in [t_1, t_2); \\ 1 & \text{otherwise;} \end{cases}$$

$$\nu_{t,i}(A) := \begin{cases} \frac{\nu\left(\mathsf{G}_i(t) \cap A\right)}{\nu\left(\mathsf{G}_i(t)\right)} & t \in [t_1, t_2); \\ 0 & \text{otherwise;} \end{cases} \qquad \text{for } i = 1, 2.$$

Weak Poincaré Inequalities

Definition 1. *We say that a μ-reversible, positive transition kernel P satisfies a weak Poincaré inequality (WPI) if for all $f \in \mathrm{L}_0^2(\mu)$ we have*

$$\|f\|_\mu^2 \leq s \cdot \mathcal{E}_\mu(P, f) + \beta(s) \cdot \|f\|_{\mathrm{osc}}^2, \tag{3}$$

where $\beta\colon (0,\infty) \to [0,\infty)$ is a decreasing function with $\lim_{s\to\infty} \beta(s) = 0$.

Assumption 1. *We assume that for Lebesgue-almost every $t \in \mathsf{T}$, the kernel H_t is ν_t-reversible, positive and satisfies a WPI, i.e. there is a measurable function $\beta : (0,\infty) \times \mathsf{T} \to [0,\infty)$ with $\beta(\cdot, t)$ satisfying the conditions in Definition 1 for each $t \in \mathsf{T}$, such that for each $s > 0$, $f \in \mathrm{L}^2(\nu_t)$,*

$$\mathrm{Var}_{\nu_t}(f) \leq s \cdot \mathcal{E}_{\nu_t}\left(H_t, f\right) + \beta\left(s, t\right) \cdot \|f\|_{\mathrm{osc}}^2. \tag{8}$$

Theorem 11. *Under Assumption 1, we have the following comparisons for U and H given in (6) and (7):*

For all $f \in \mathrm{L}^2(\pi)$,

$$\mathcal{E}(H, f) \leq \mathcal{E}(U, f), \tag{9}$$

and conversely, for all $s > 0$, $f \in \mathrm{L}^2(\pi)$,

$$\mathcal{E}(U, f) \leq s \cdot \mathcal{E}(H, f) + \beta(s) \cdot \|f\|_{\mathrm{osc}}^2, \tag{10}$$

where $\beta : (0,\infty) \to [0,\infty)$ is given by

$$\beta(s) := c^{-1} \cdot \int_{\mathsf{T}} \beta(s,t) \cdot m\left(t\right) \,\mathrm{d}t.$$

Furthermore, β satisfies the conditions for a WPI in Definition 1.

Metropolis Chains as HSS

4.1 Metropolis chains

Definition 22. *Let π a probability measure admitting a density $\varpi = \frac{\mathrm{d}\pi}{\mathrm{d}\nu}$ with respect to some σ-finite measure ν on G, and let Q be a ν-reversible Markov kernel; we say that such triples (π, ν, Q) are* compatible. *Define* $\mathsf{Metropolis}\,(\pi, \nu, Q)$ *to be the Markov kernel P given by*

$$P\,(x, A) = \int_A Q\,(x, \mathrm{d}y) \cdot \alpha\,(x, y) + \mathbf{1}_A\,(x) \cdot \bar{\alpha}\,(x)\,, \qquad x \in \mathsf{G}, A \in \mathcal{G},$$

where for $x, y \in \mathsf{G}$,

$$\alpha\,(x, y) := 1 \wedge \frac{\varpi\,(y)}{\varpi\,(x)}, \quad \bar{\alpha}\,(x) := 1 - \alpha\,(x)\,, \quad \alpha\,(x) := \int_{\mathsf{G}} Q\,(x, \mathrm{d}y) \cdot \alpha\,(x, y)\,.$$

It is known that all such kernels are π-reversible. In what follows, we will use (π, ν, Q) informally to refer to generic compatible triples in the sense described above.

Lemma 27. *For any compatible triple (π, ν, Q), it holds that* $\mathsf{Metropolis}\,(\pi, \nu, Q) = \mathsf{HybridSlice}\left(\pi, \nu, (H_t)\right)$, *where*

$$H_t\,(x, A) = Q\left(x, A \cap \mathsf{G}\,(t)\right) + \mathbf{1}_A\,(x) \cdot Q\left(x, \mathsf{G}\,(t)^{\complement}\right).$$

Some Metropolis Chains

Example 23. *When $\nu = \mathrm{Leb}$ and Q is a symmetric, ν-reversible kernel, then we can define the* Random Walk Metropolis *(RWM) kernel,*

$$\mathsf{RWM}(\pi, Q) := \mathsf{Metropolis}\left(\pi, \mathrm{Leb}, Q\right).$$

It is conventional to work with $Q_\sigma\left(x, \mathrm{d}y\right) = \mathcal{N}\left(\mathrm{d}y \mid x, \sigma^2 \cdot I_d\right)$ for some step-size $\sigma > 0$; we will work under this assumption going forward. See also Section 6.3.2 of [29].

Example 24. *When ν is a sufficiently-tractable probability measure, we may take $Q\left(x, \cdot\right) = \nu$ directly, independently of x. We can thus define the* Independent Metropolis–Hastings *(IMH) kernel with 'proposal' ν; see [29, Section 6.3.1]:*

$$\mathsf{IMH}(\pi, \nu) := \mathsf{Metropolis}\left(\pi, \nu, \nu\right).$$

Example 25. *When $\nu = \gamma_{\mathsf{m},\mathsf{C}}$ is a Gaussian measure with mean m and covariance operator C, then one may take $\rho, \eta \in (0,1)$ such that $\rho^2 + \eta^2 = 1$ and define the autoregressive proposal $Q_\eta\left(x, \mathrm{d}y\right) = \mathcal{N}\left(\mathrm{d}y \mid \mathsf{m} + \rho \cdot (x - \mathsf{m}), \eta \cdot \mathsf{C}\right)$. The resulting Metropolis chain is known as the* Preconditioned Crank-Nicolson *(pCN) kernel with Gaussian reference $\gamma_{\mathsf{m},\mathsf{C}}$ and step-size η; see e.g. [9]:*

$$\mathsf{pCN}(\pi, \mathsf{m}, \mathsf{C}, \eta) := \mathsf{Metropolis}\left(\pi, \gamma_{\mathsf{m},\mathsf{C}}, Q_\eta\right).$$

Beyond Hit-and-Run

There are various routes left open by this work; we list here a few.

For one, we have largely focused on Simple Slice Sampling as the ideal algorithm, and using Hit-and-Run as on-slice kernels, due to their relative genericity and ease of implementation. In specific settings, other on-slice kernels are of substantial interest. For super-level sets with some coordinate-related structure, Gibbs sampling (also 'Coordinate Hit-and-Run') is a natural option, for which results have recently been obtained in the convex setting; see [17]. For super-level sets which take the form of polytopes, yet more on-Slice Samplers are available, including the Dikin walk [16], the Vaidya and John walks [7], and various gradient-based samplers which introduce additional geometric structure to the problem (e.g. [20]); some of these methods come with theoretical guarantees in the form of estimates on the conductance or spectral gap of the kernel, which can readily be used in our framework.