Comparison Theorems for Practical Slice Sampling

University of Bristol, Statistics Seminar, 8 March 2024

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

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

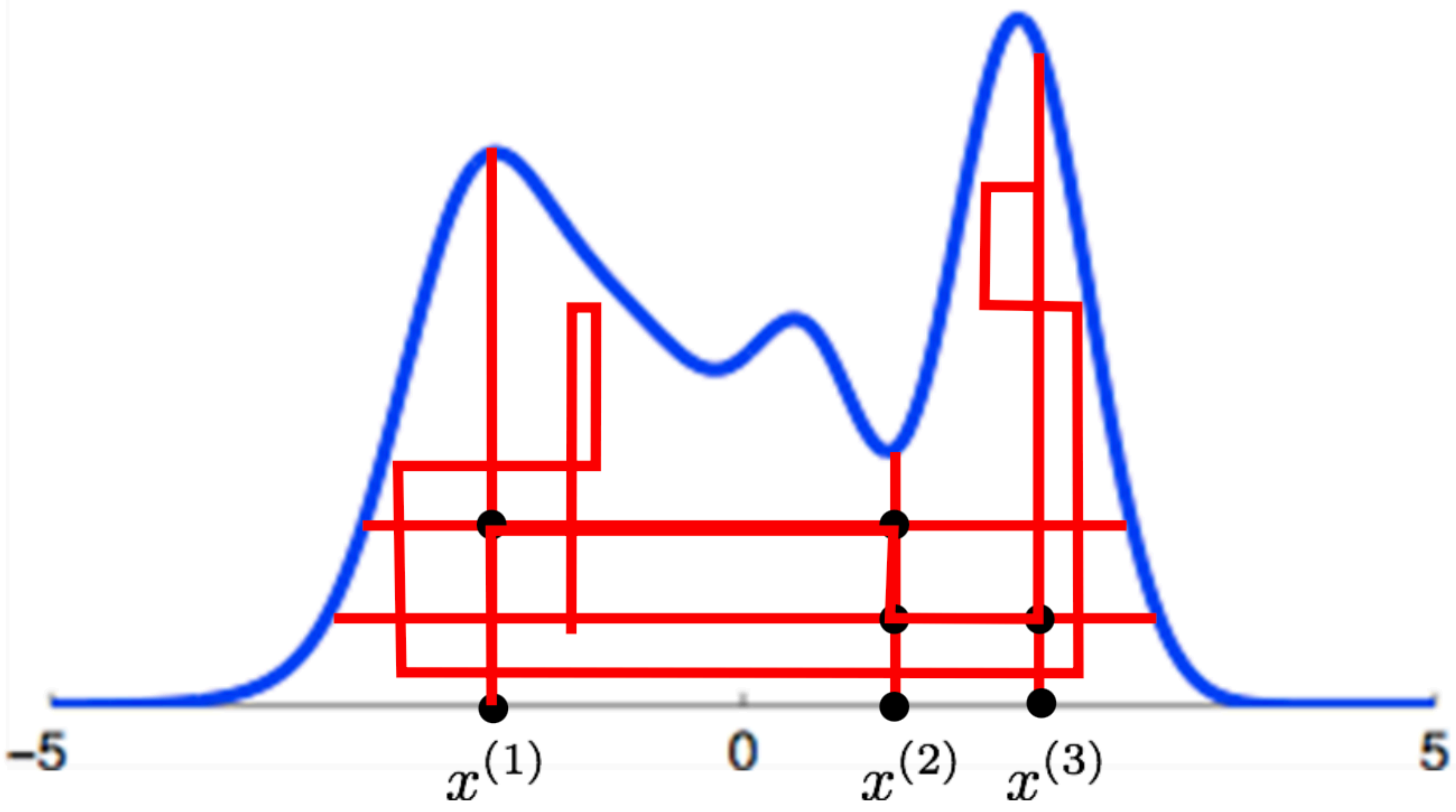
- designing effective Markov kernels
- obtaining and using useful information about π
- tuning of algorithm hyperparameters (step-size, etc.)

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)

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 $X_n \sim \text{Unif}\left(\{x : \pi(x) \geqslant t\}\right)$
    - (sample uniformly ‘across’ this height)

Slice Sampling: Properties 1

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

Slice Sampling: Properties 2

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

Implementing Slice Sampling

“given $T_n = t$, sample $X_n \sim \mathsf{Unif}\left(\{x : \pi(x) \geqslant t\}\right)$”

– Sam Power, Slide 7

Life on the Slice

- write $\mathrm{G}(t) = \{x : \pi(x) \geqslant t\}$ for the super-level set ('slice')
- write $\nu_t = \mathrm{Unif}\left(\mathrm{G}(t)\right)$
- if $\mathrm{G}(t)$ 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 $X_n \sim \text{Unif}\left(\{x : \pi(x) \geqslant t\}\right)$"
- 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 long-time behaviour
  - is <u>not as</u> (statistically) <u>efficient</u> as the ‘ideal’ Slice Sampler
- how do we quantify the cost of approximation?
  - ⇝ Markov chain comparison theory

Convergence of Markov Chains

- many possible approaches
- one approach:
  - let P denote the Markov kernel of interest
  - define $P^n f(x) = \mathbf{E}\left[f(X_n) \mid X_0 = x\right]$
  - by ergodicity, expect that $P^n f(x) \to \pi(f)$ as $n \to \infty$, for all x
  - $\rightsquigarrow$ study how $\mathrm{var}_\pi\left(P^n f\right)$ tends to 0 as $n \to \infty$,

Variance and Energy

- define 'energy' of f as

$$\mathscr{E}_P(f) = \int \pi\left(\mathrm{d}x\right) P\left(x, \mathrm{d}y\right) \cdot \frac{\left[f(x) - f(y)\right]^2}{2}$$

- fact: for (reversible, positive) P,
    - if $\mathscr{E}_P(f) \geqslant \gamma \cdot \mathrm{var}_\pi(f)$ for all f,
    - then $\mathrm{var}_\pi\left(P^n f\right) \leqslant \left(1 - \gamma\right)^{2\cdot n} \cdot \mathrm{var}_\pi\left(f\right)$ for all n, f

Energy and Comparisons

- (rough) interpretation:
  - if $\mathscr{E}_P(f) \geqslant \gamma \cdot \mathrm{var}_\pi(f)$,
  - then it takes γ^{-1} steps of P to get an independent (≈) sample
- extension: let P, Q be two positive, π-reversible kernels
  - if $\mathscr{E}_P(f) \geqslant \delta \cdot \mathscr{E}_Q(f)$,
  - then taking δ^{-1} steps of P is as useful as taking one step of Q

Back to Hybrid Slice Sampling

- let U = Ideal Slice Sampling, H = Hybrid Slice Sampling
- to quantify how $\mathrm{var}_\pi \left(H^n f\right)$ tends to 0, we will study
  - how $\mathrm{var}_\pi \left(U^n f\right)$ tends to 0, 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

Some Warm-Up Results

- in complete generality, $\mathscr{E}_U(f) \geqslant \mathscr{E}_H(f)$
  - ⇝ all else being equal, always prefer ideal chain
- for experts: any Metropolis(~~-Hastings~~) kernel can be written as such an H
  - ⇝ all such chains are automatically dominated by (ideal) SS
- we are interested in when H is **almost as good** as U

A Generic Result

- suppose that for all heights t, there is $\gamma_t > 0$ such that

$$\mathscr{E}_{H_t}(f) \geqslant \gamma_t \cdot \mathrm{var}_{\nu_t}(f),$$

  and that $\gamma_H := \inf_{t \in \mathsf{T}} \gamma_t > 0$

- it then holds that

$$\mathscr{E}_H(f) \geqslant \gamma_H \cdot \mathscr{E}_U(f),$$

  i.e. HSS is only a factor γ_H worse than ideal SS

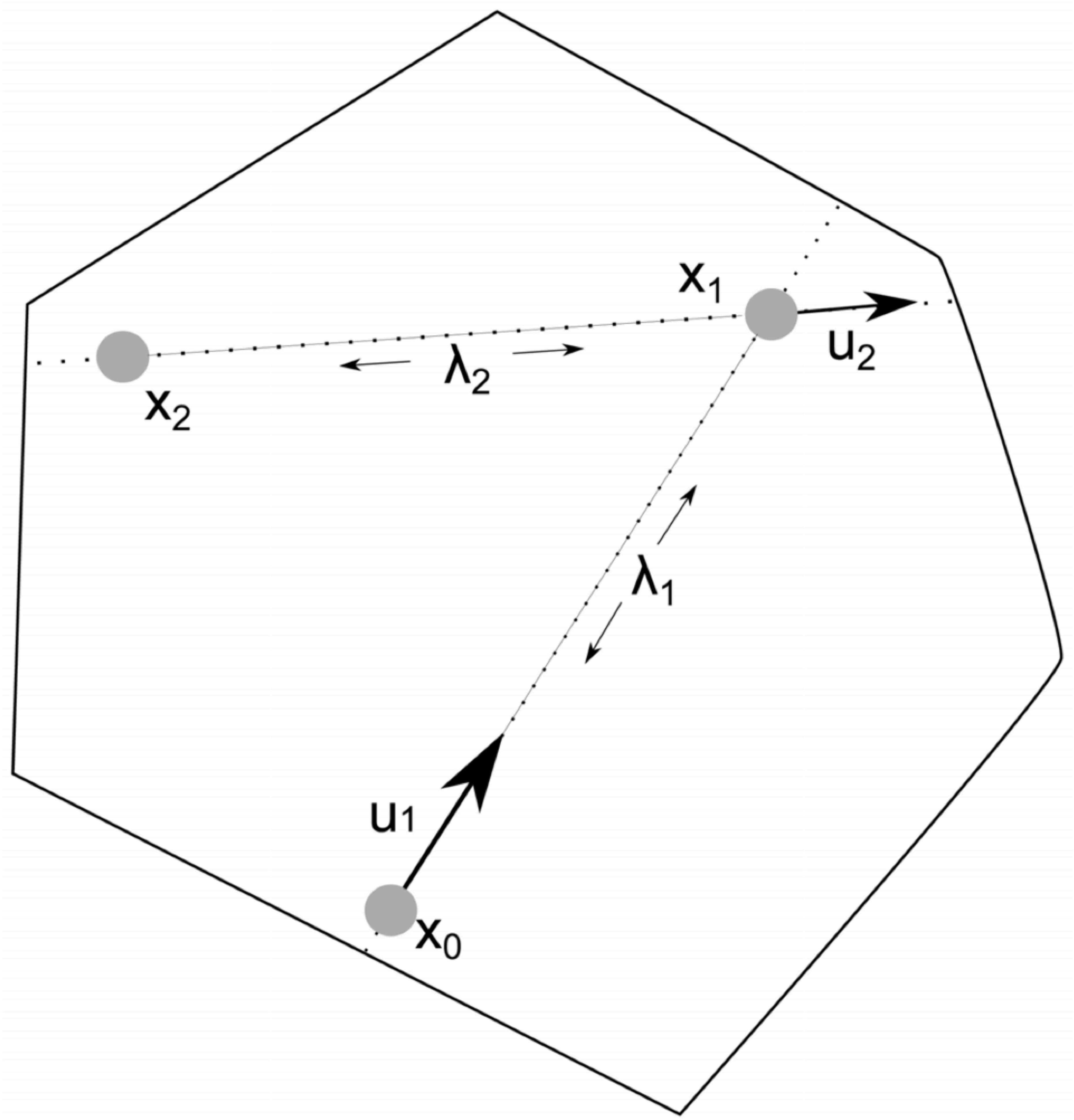
- if U converges exponentially, then so does H; rate only degrades by factor γ_H

On the Slower-than-Exponential Case

- for simplicity, in today's talk, I focus on the setting in which all chains under consideration converge to equilibrium at an exponential rate
- in fact, our theoretical framework also handles very naturally the case of slower-than-exponential convergence
- e.g. if $\gamma_H = 0$, we can still obtain quite explicit convergence rate estimates for H, depending on how badly $\gamma_t \to 0$ in suitable limits

Case Study: Hit-and-Run on the Slice

- simple method for sampling uniform distributions on convex body G
- at $X_{n-1} = x$,
  - sample $U_n \sim$ Unif $\left(\mathbb{S}^{d-1}\right)$
  - look at $\left(x + U_n\mathbf{R}\right) \cap$ 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

$$\mathscr{E}_{\mathrm{H\&R}}(f) \geqslant c \cdot d^{-2} \cdot \kappa_{\mathsf{G}}^{-2} \cdot \mathrm{var}_\pi(f).$$

- 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

$$\mathscr{E}_{H_t}(f) \geqslant \gamma_t \cdot \mathrm{var}_{\nu_t}(f)$$

where $\gamma_t = c \cdot d^{-2} \cdot \kappa_{\mathsf{G}(t)}^{-2}$

- so, convergence will be good if super-level sets $\mathsf{G}(t)$ are well-conditioned

Well-Conditioned Level Sets

- let $V : \mathbf{R}^d \to \mathbf{R}$ be m-strongly convex and L-smooth
  - i.e. eigs $\left(\mathsf{Hess} V(x)\right) \in [m, L]$
  - 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 $\kappa_{\mathsf{G}(t)} \leqslant \sqrt{\kappa_V}$.

Some Applications

- if π has this form, then $\mathscr{E}_H(f) \gtrsim d^{-2} \cdot \kappa_V^{-1} \cdot \mathscr{E}_U(f)$
    - 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, $\rightsquigarrow$ decorrelation time of $\lesssim d^3 \cdot \kappa_V$
- if e.g. π is multivariate Student-t, then $\kappa_V = 1, \gamma_H \lesssim d^2$
    - combine with earlier work, $\rightsquigarrow$ decorrelation 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,
  - 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 $H_t \succeq \gamma_H \cdot \nu_t$, then $H \succeq \gamma_H \cdot U$.

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, ‘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

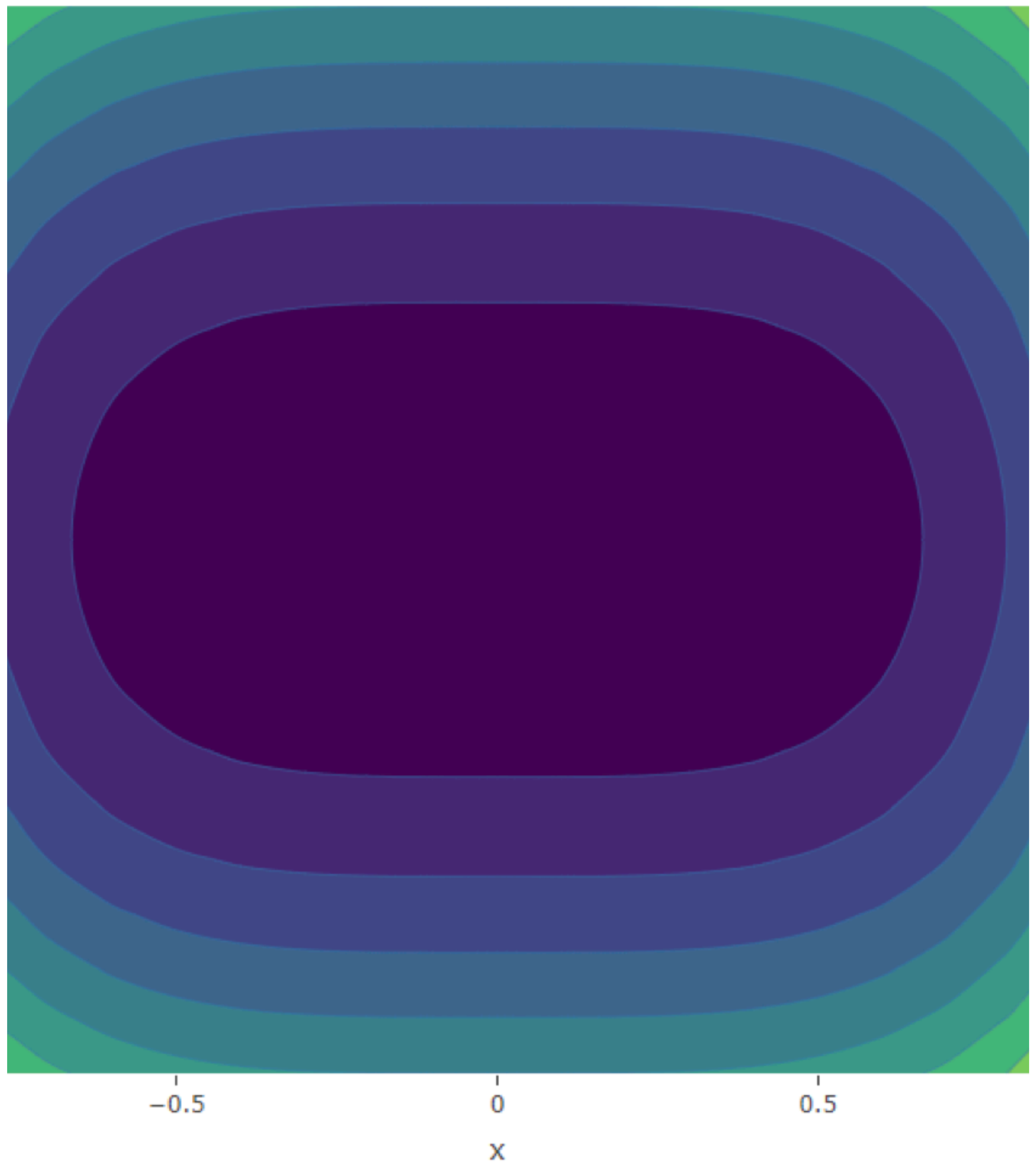
Advanced Applications

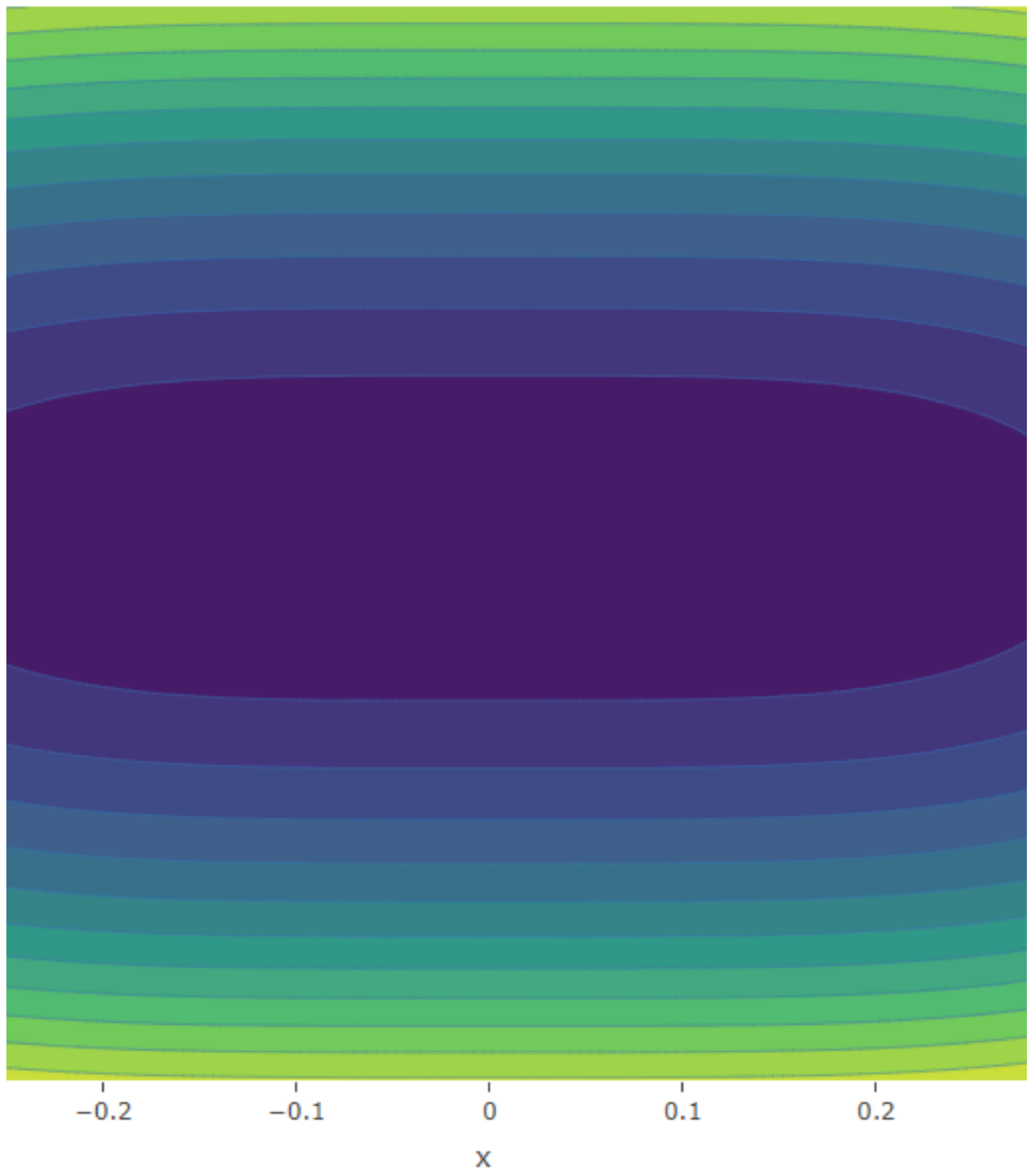
- let $1 \leqslant p_2 \leqslant p_1$, $1 \leqslant q_1 \leqslant q_2$, and suppose that

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

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

- 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*

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

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 $s > 0$, $f \in \mathrm{L}^2(\nu_t)$,*

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

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

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

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

Metropolis Chains as HSS

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 $Q_\eta\left(x, \mathrm{d}y\right) = \mathcal{N}\left(\mathrm{d}y \mid \mathsf{m} + \rho \cdot \left(x - \mathsf{m}\right), \eta \cdot \mathsf{C}\right)$. 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]:*

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