SIAM UQ 2026

Preconditioning for MCMC: Concepts, Practices, Challenges

Max Hird, University of Waterloo

Joint work with Sam Power, University of Bristol

Preconditioners are transformations to π or to the Markov kernel P_π such that mixing time is improved

They attempt to surmount obstacles to sampling efficiency

A Model Process and its Derivatives

Define OLD (π, D) as the solution to

$$\mathrm{d}X_t = \left(D(X_t)\,\nabla \log \pi(X_t) + \operatorname{div}(D(X_t))\right)\mathrm{d}t + \sqrt{2D(X_t)}\mathrm{d}B_t$$

where $D : \mathbb{R}^d \to \mathrm{PD}_{d\times d}$ and $\operatorname{div} : \mathbb{R}^{d\times d} \to \mathbb{R}^d$ is defined with

$$\operatorname{div}(D)_i := \sum_{j\in[d]} \frac{\partial}{\partial x_j} D(x)_{ij}$$

Generator $\mathscr{L}_D = -\nabla^* D\,\nabla$ where ∇^* is the adjoint of ∇ in $\mathrm{L}^2(\pi)$

The Dirichlet form is $\mathscr{E}_{\pi,D}(f) = \mathbb{E}_\pi[\nabla f^\top D\,\nabla f]$

The spectral gap is $\gamma_{\pi,D} = \inf_{\|f\|_\pi=1} \mathscr{E}_{\pi,D}(f)$

Obstacle 1: Scale Mismatch

The scales of π and P_π are mismatched

Case 1: $\text{scale}(\pi) \gg \text{scale}(P_\pi)$

Sampler is Markovian and hence will be ‘too slow’

e.g. [Altschuler and Talwer 2022 Theorem 3.2]: lower bounds on TV mixing of ~ULA

$\gamma_{\pi, h^2 \mathbf{I}_d} \propto h^2$

Case 2: $\text{scale}(\pi) \ll \text{scale}(P_\pi)$

Discretisations will fail

Obstacle 1: Scale Mismatch cont.

Example: ULA with step size h^2 on $\pi = \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$:

$$X_t \mid X_0 \sim \mathcal{N}\left(\left(1 - \frac{h^2}{\sigma^2}\right)^t X_0, 2h^2 \sum_{i=0}^{t-1} \left(1 - \frac{h^2}{\sigma^2}\right)^{2(t-i-1)} \mathbf{I}_d\right)$$

Obstacle 1: Scale Mismatch cont.

Preconditioning

Adjust h^2 according to:

Optimal scaling arguments see, e.g. [Roberts, German and Gilks 1997]

Prior knowledge about π see e.g. non-asymptotic literature

Heuristic arguments: as big as possible and no more

Obstacle 2: Linear Anisotropy

For each direction the target has a single length scale, but these vary across directions

I.e. $f(x, v) := v^\top \nabla^2 \log \pi(x) v$ is const. with fixed v but not with fixed x

If P_π 'is isotropic' then we have scale mismatch along at least one direction

If $f(x) = v^\top x$ then $\mathscr{E}_{\pi, h^2 \mathbf{I}_d}(f) = h^2 \|v\|^2 \left(v^\top \mathrm{Cov}_\pi(X) v\right)^{-1}$

Therefore $\gamma_{\pi, h^2 \mathbf{I}_d} \leq h^2 \lambda_1 \left(\mathrm{Cov}_\pi(X)\right)^{-1}$

Obstacle 2: Linear Anisotropy cont.

Example: ULA with step size h^2 on $\pi = \mathcal{N}(0, \Sigma_\pi)$:

$$W_2(\mathscr{L}(X_t \mid X_0), \pi) \leq \max_{i \in [d]} \{ \, | 1 - h^2 \lambda_i(\Sigma_\pi)^{-1} | \, \}^t W_2(\delta_{X_0}, \pi_{h^2}) + \sqrt{\mathrm{tr}(\Sigma_\pi^{-1})} \left(\sqrt{\frac{2}{2 - h^2 \lambda_1(\Sigma_\pi)^{-1}}} - 1 \right)$$

where π_{h^2} is the equilibrium distribution of ULA

Obstacle 2: Linear Anisotropy cont.

Preconditioning

One may choose a linear change of variables to isotropise π

Use $L = \sigma \operatorname{diag}\left(\hat{\Sigma}_{\pi}\right)$

Used in stan and TensorFlow probability

Computationally cheap to infer and precondition with

There exist cases in which it can increase the anisotropy if the target is already very anisotropic [H. and Livingstone 2025]

Obstacle 2: Linear Anisotropy cont.

Preconditioning

Fully dense preconditioner e.g. $L = \sigma\,\hat{\Sigma}_\pi$ or $\sigma\,\widehat{\mathrm{Cov}}_\pi(\nabla \log \pi)^{-1}$ or $-\sigma\,\nabla^2 \log \pi(x^*)$

Covariance and ‘Fisher’ matrices admit straightforward learning mechanisms

Either can be favourable depending on tailedness of the target

Learning $-\sigma\,\nabla^2 \log \pi(x^*)$ or other problem specific preconditioners is possibly even easier, done e.g. in [Bürkner 2017 ‘brms’]

Obstacle 2: Linear Anisotropy cont.

Recall: $\gamma_{\pi,h^2\mathbf{I}_d} \leq h^2 \|\mathrm{Cov}_\pi\|^{-1}$

KLS conjecture [Kannan, Lovász, Simonovitz 1995]: if π is log-concave $\gamma_{\pi,h^2\mathbf{I}_d} \geq c\|\mathrm{Cov}_\pi\|^{-1}$ for universal $c \in (0,1)$

Proven up to $\sqrt{\log d}$ [Chen 2021, Klartag 2023]

Longest scales of π are the essential barrier to mixing for $\mathrm{OLD}(\pi, h^2\mathbf{I}_d)$

If, however, global shape of π doesn't resemble local shape, linear preconditioners can be problematic:

π is a two component mixture with same covariance Σ but different means μ_1 and μ_2

$$\mathrm{Cov}_\pi(X) = \Sigma + (\mu_1 - \mu_2)(\mu_1 - \mu_2)^\top$$

Obstacle 3: Heavy Tails

There exists a direction v and $R_v > 0$ such that $\mathbb{E}_\pi[\,|v^\top X|^{R_v}] = \infty$ but $\mathbb{E}_\pi[\,|v^\top X|^{r}] < \infty$ for $r < R_v$

We expect $|\nabla_v \log \pi(x)| \to 0$ as $\|x\| \to \infty$

A Markovian sampler will have trouble exploring along direction v

e.g. drift of $\mathrm{OLD}(\pi, h^2\mathbf{I}_d)$ will go to zero along v

Use $f(x) = (w^\top x)^k$ for $k \in \left(\frac{1}{2}R_v, \frac{1}{2}R_v + 1\right)$ to give $\gamma_{\pi,\Sigma} = 0$

Obstacle 3: Heavy Tails cont.

Example:

$$\pi \propto \left(1 + \|x\|^2/R\right)^{-\frac{d+R}{2}} \Rightarrow \nabla \log \pi = -\frac{d+R}{R}\left(1 + \|x\|^2/R\right)^{-1} x$$

$\nabla \log \pi \to 0 \Rightarrow \text{OLD}(\pi, \Sigma)$ *is not geometrically ergodic [Roberts and Stramer 1996 Theorem 2.4]*

[Mousavi-Hosseini et al. 2023 Corollary 8]: mixing times of $\text{OLD}\left(\pi, h^2\mathbf{I}_d\right)$ *and ULA are exponentially dependent on initial error*

Obstacle 3: Heavy Tails cont.

OLD (π, D): $\mathrm{d}X_t = \left(D(X_t)\nabla \log \pi(X_t) + \mathrm{div}(D(X_t))\right)\mathrm{d}t + \sqrt{2D(X_t)}\mathrm{d}B_t$

Preconditioning

Either i) use $D(x)$ to speed up dynamics along v ii) use $T : \mathbb{R}^d \to \mathsf{X}$ such that $T_\sharp \pi$ has lighter tails

In fact $T_\sharp \mathrm{OLD}\left(\pi, \mathbf{I}_d\right) = \mathrm{OLD}\left(T_\sharp \pi, J_T J_T^\top\right)$

[Roberts and Stramer 1996]: use $D(x) = \pi(x)^{-\alpha}\mathbf{I}_d$ for $\alpha \in [0,1]$

Obstacle 3: Heavy Tails cont.

Preconditioning

[Cui, Tong, Zahm 2025]: Use $D(x) = \tau_\pi(x)$ where $\tau_\pi(x)$ is a *Stein Kernel*

$$\mathbb{E}_\pi[\varphi(X)(X - \mu_\pi)] = \mathbb{E}_\pi[\tau_\pi(X)\nabla\varphi(X)]$$

Note: non-unique

[Cui, Tong, Zahm 2025] give a $\tau_\pi(x)$ such that $\gamma_{\pi,\tau_\pi} = 1$

OLD $\left(\pi, \tau_\pi\right)$ has drift $-(X_t - \mu_\pi)$

Obstacle 3: Heavy Tails cont.

Preconditioning

[Johnson and Geyer 2012]: Use a transformation to tame the tails

$$T(x) = f(\|x\|)\frac{x}{\|x\|}$$

Do MCMC on $T_\sharp \pi$

Theoretical results for particular f 's

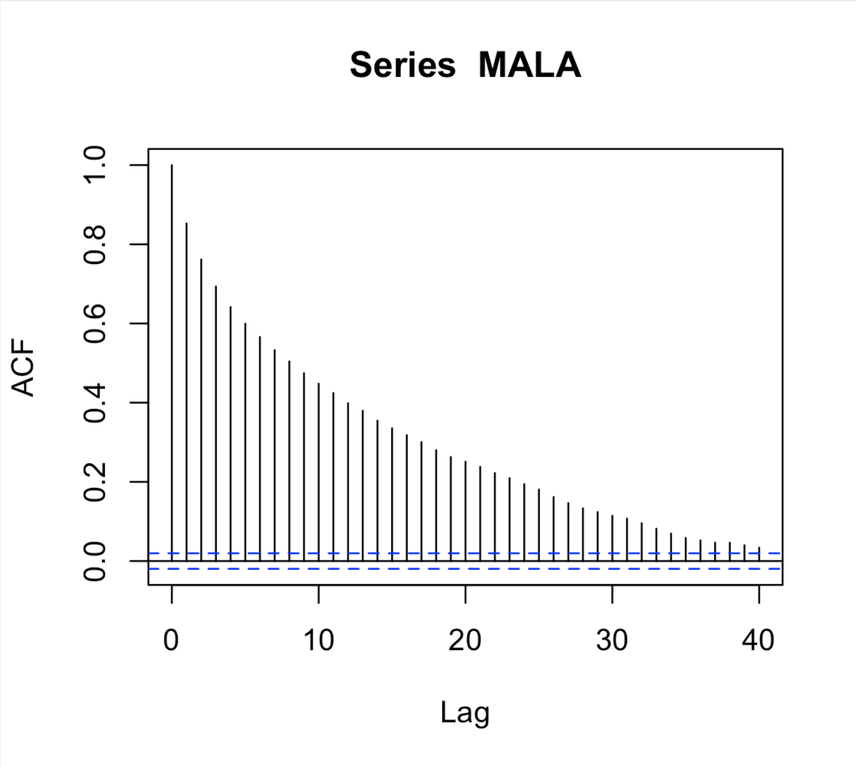
Obstacle 3: Heavy Tails cont.

MALA

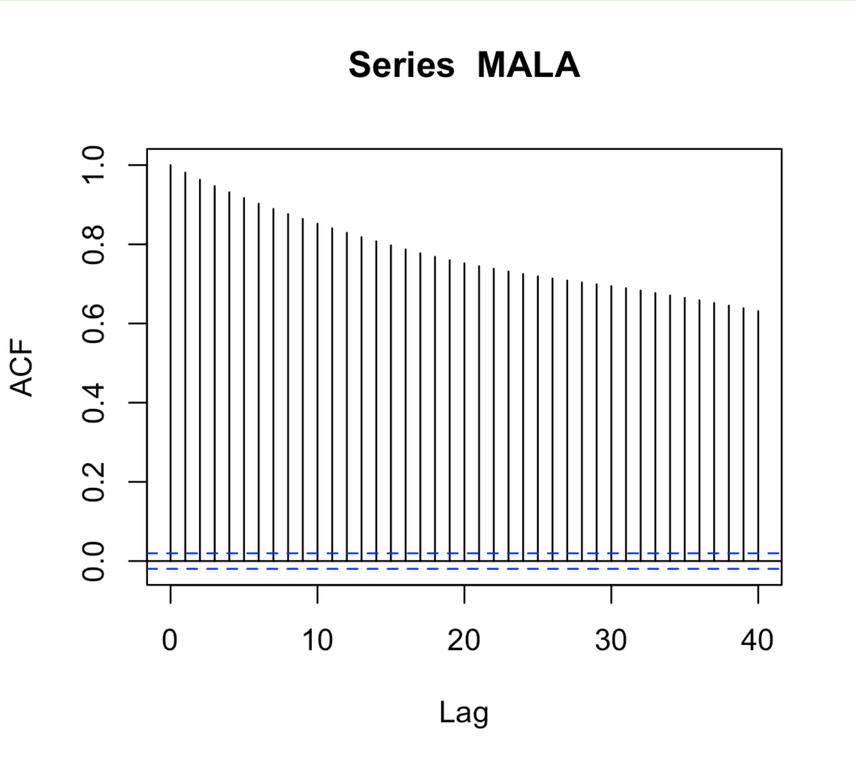
Example:

$$\pi \propto \left(1 + \|x\|^2/R\right)^{-\frac{d+R}{2}}, R = 1.5$$

d=1



d=10



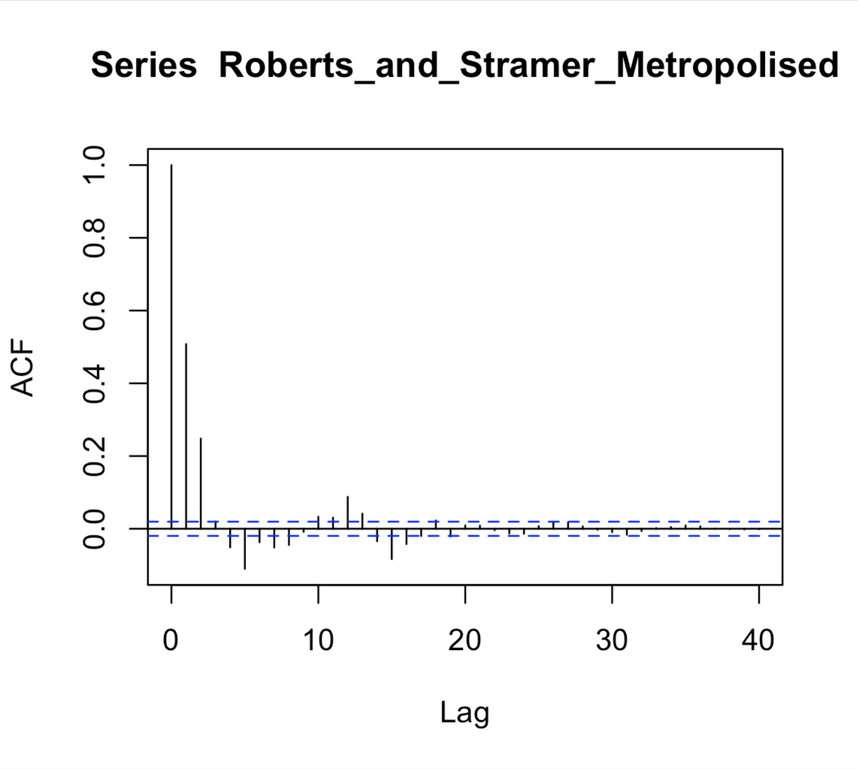
Obstacle 3: Heavy Tails cont.

Roberts + Stramer (Metropolised)

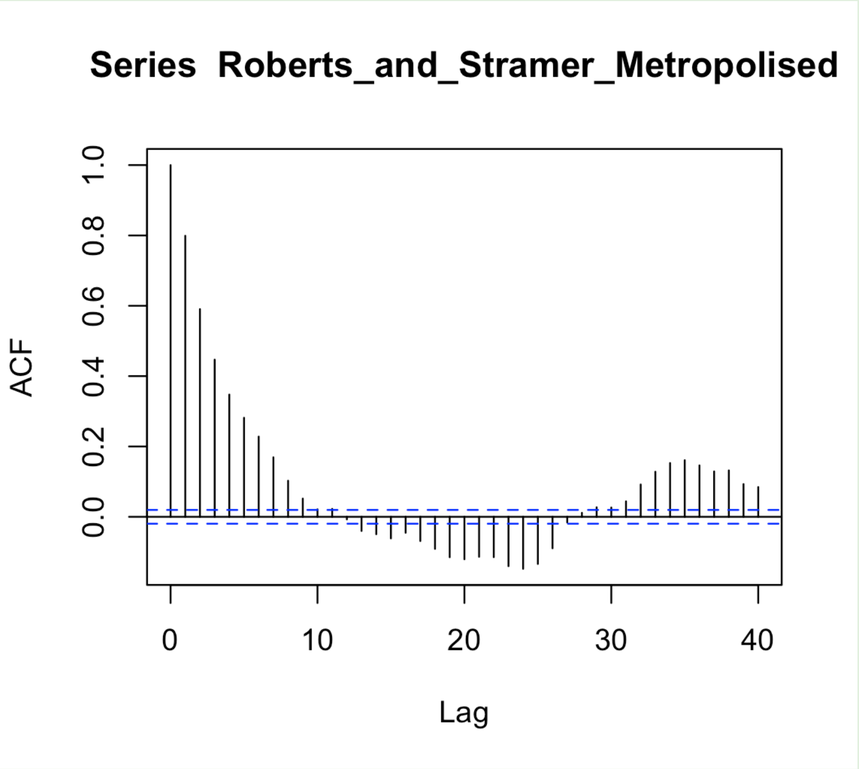
Example:

$$\pi \propto \left(1 + \|x\|^2/R\right)^{-\frac{d+R}{2}}, R = 1.5$$

d=1



d=10



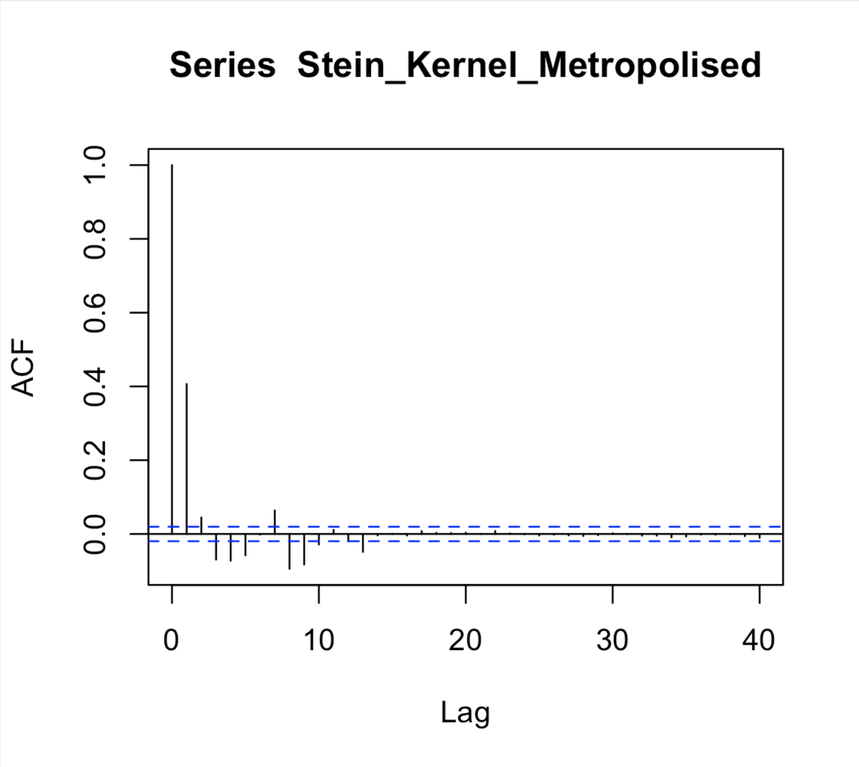
Obstacle 3: Heavy Tails cont.

Stein Kernel (Metropolised)

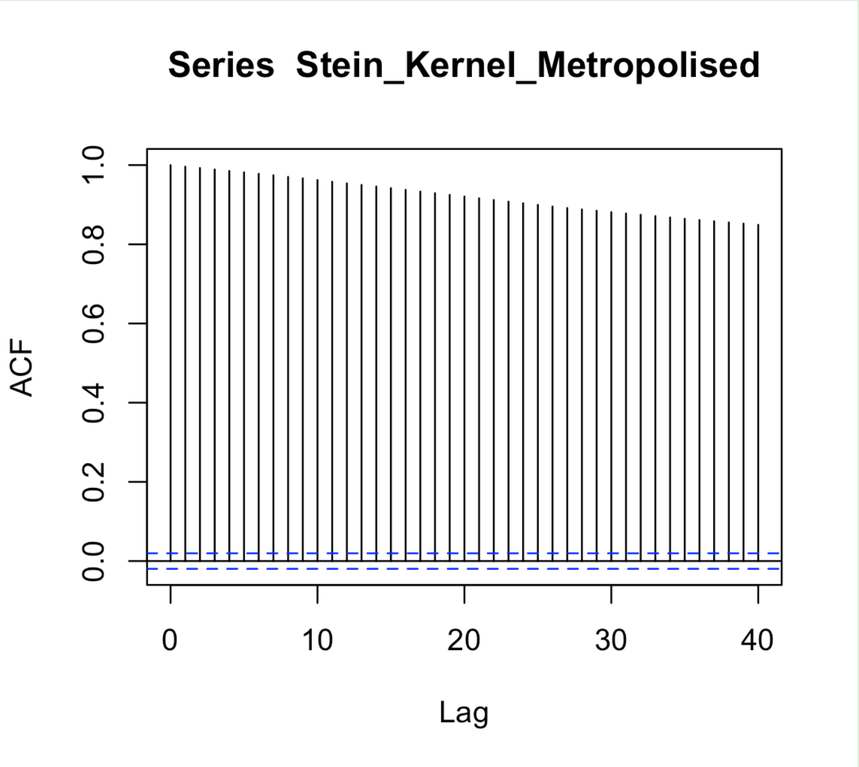
Example:

$$\pi \propto \left(1 + \|x\|^2/R\right)^{-\frac{d+R}{2}}, R = 1.5$$

d=1



d=10



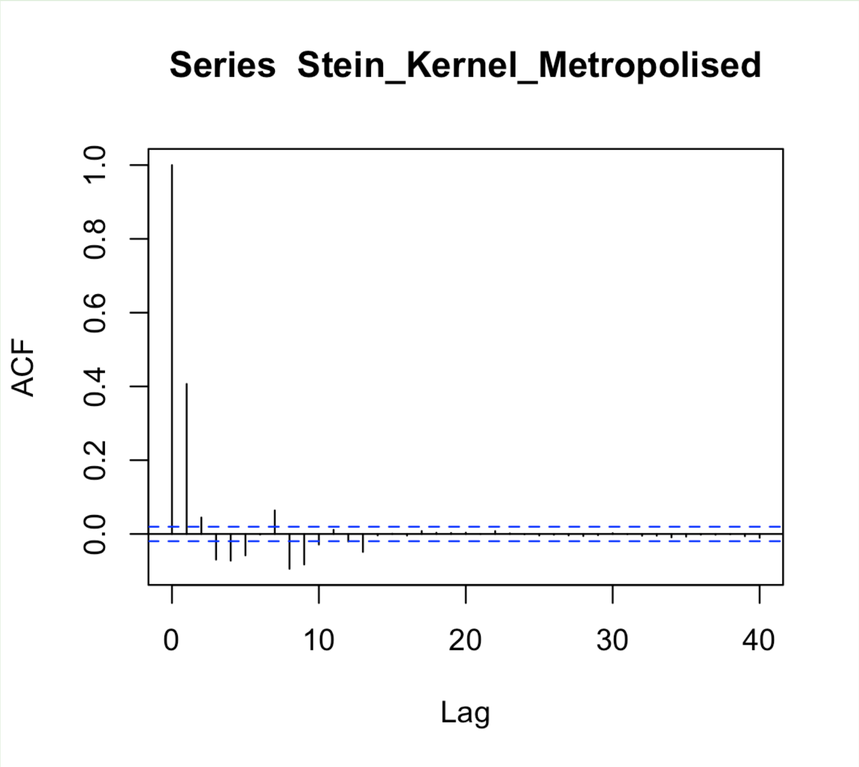
Obstacle 3: Heavy Tails cont.

Stein Kernel (Metropolised)

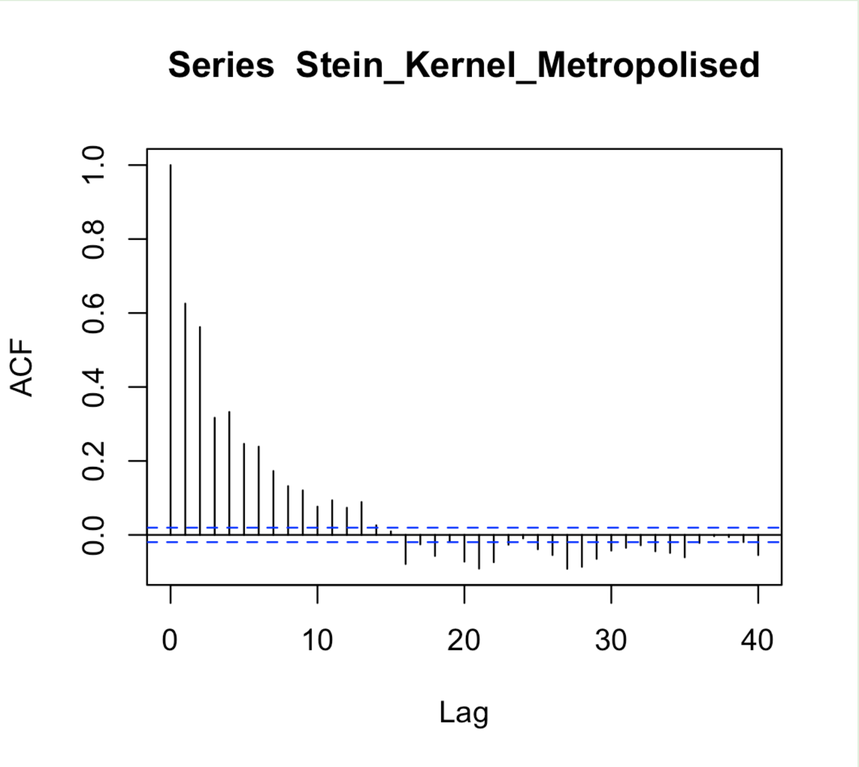
Example:

$$\pi \propto \left(1 + \|x\|^2/R\right)^{-\frac{d+R}{2}}, R = 1.5$$

d=1



d=10, optimised step size



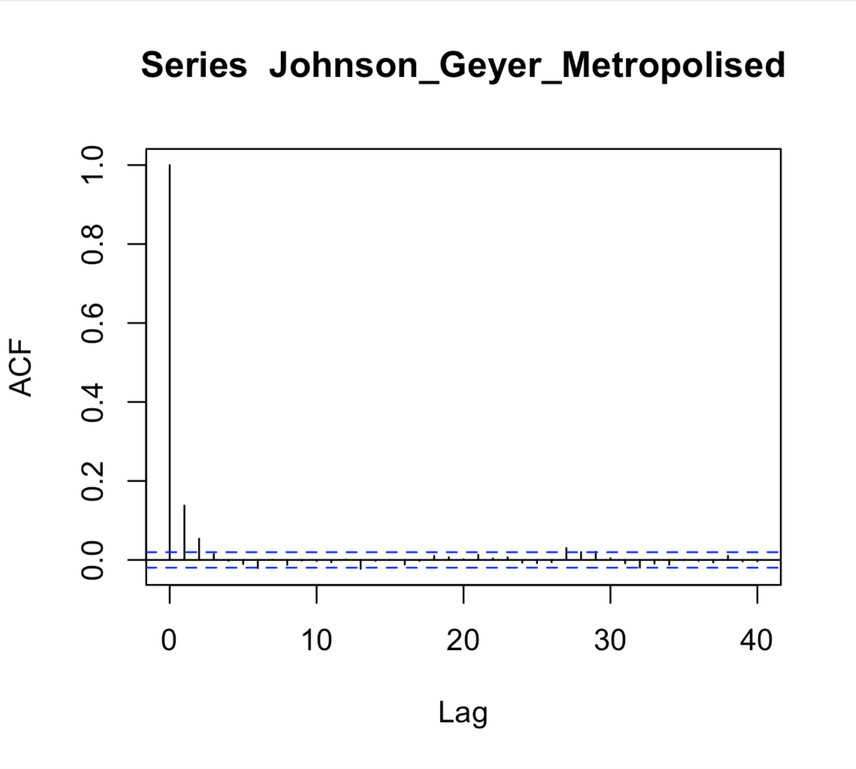
Obstacle 3: Heavy Tails cont.

Johnson and Geyer (Metropolised)

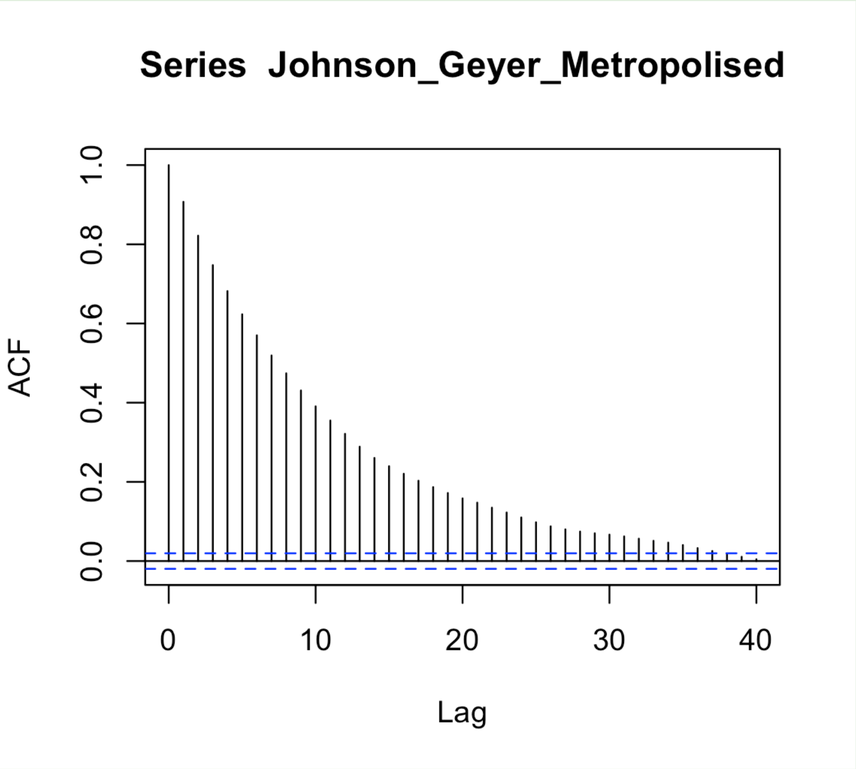
Example:

$$\pi \propto \left(1 + \|x\|^2/R\right)^{-\frac{d+R}{2}}, R = 1.5$$

d=1



d=10



Obstacle 3: Heavy Tails cont.

Non-Stein Kernel Preconditioners can work well given knowledge of centre and tails

Too severe: pathologies à la scale mismatch

Too little: no effect

Difficult to strike a balance in high dimensions

Stein Kernels are good if you can find them

They are a new technology

Obstacle 4: Nonlinear Concentration

π is well concentrated but its contours are no longer convex

Region of high probability under π is ~ non-linear manifold

Without π-specific modification canonical samplers will

i) Fail to explore the full manifold
ii) Exhibit high rejection rates or instability

Obstacle cannot be surmounted with a linear change of variables

Obstacle 4: Nonlinear Concentration cont.

OLD (π, D): $\mathrm{d}X_t = \left(D(X_t)\nabla \log \pi(X_t) + \mathrm{div}(D(X_t))\right)\mathrm{d}t + \sqrt{2D(X_t)}\mathrm{d}B_t$

Preconditioning

Either i) Have the manifold inform D ii) use $T : \mathbb{R}^d \to \mathsf{X}$ such that $T_\sharp \pi$ is $\approx$ standard normal

Transport-based methods can be categorised into
i) methods constructing T using normalising flows
ii) methods constructing T using optimal-transport inspired ‘triangular’ maps AKA measure transport of Youssef Marzouk’s group

Obstacle 4: Nonlinear Concentration cont.

Metric based methods:

Likelihood Informed methods [Cui et al. 2014, Cui et al. 2016]: local to a point in the state space, calculate the directions along which the scale of the posterior differs the most from the prior

Riemannian methods [Girolami and Calderhead 2011] use variants of the Fisher information to construct metrics

Transformation based methods:

Parametrise the space of transformations $\{T_\theta : \theta \in \Theta\}$ and find

$$\theta^* = \arg\min_{\theta \in \Theta} d((T_\theta)_\sharp \pi, \mathcal{N}(0, \mathbf{I}_d))$$

Do MCMC on $(T_{\theta^*})_\sharp \pi$ and transform back

Obstacle 5: Multimodality

π is multimodal with sufficient separation

Going from one ‘basin of attraction’ to another has vanishingly small probability for local methods

Let x and y be modes separated by a unique saddle point at z

Eyring-Kramers [Eyring 1935, Kramers 1940]:

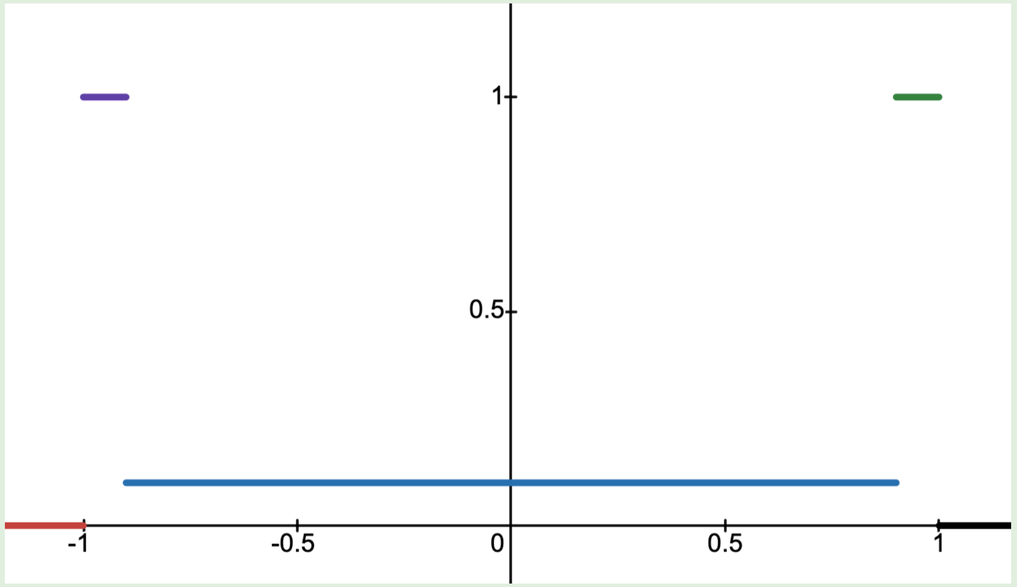
$$\mathbb{E}[\text{time}(x \to y)] \sim \left(\frac{\pi(x)}{\pi(z)}\right)^{1/\varepsilon}$$

Under OLD $\left(\pi, \mathbf{I}_d\right)$ with ε-scaled diffusion, as $\varepsilon \to 0$

Obstacle 5: Multimodality cont.

Example:

$$\pi(x) \propto \varepsilon \times 1\{\,|x| < 1 - r\} + 1\{-1 \le x \le -(1-r)\} + 1\{1 - r \le x \le 1\}$$



Choosing f piecewise linear to vary between modes gives

$$\gamma_{\pi,h^2\mathbf{I}_d} \le h^2 \frac{\varepsilon}{r(1-r)}$$

More in the paper:

Real-world examples

Hard targets require non-regular preconditioners

In solving $Ax = b$ there exist ‘left’ and ‘right’ preconditioners

We define analogous concepts for sampler preconditioners

Gaussian equivalence

More on the mysterious Stein Kernels