Stereographic Multi-Try Metropolis Algorithms for Heavy-tailed Sampling

Zhihao Wang

Department of Mathematical Sciences,
University of Copenhagen

UNIVERSITY OF COPENHAGEN

Stereographic Multi-Try Metropolis

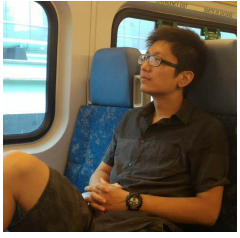
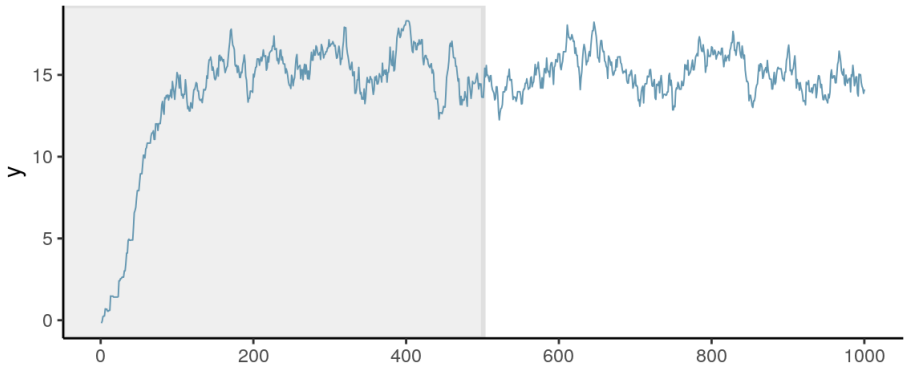


Figure: Jun Yang

- Z.W. and Jun Yang, *Stereographic Multi-Try Metropolis Algorithms for Heavy-tailed Sampling*, arXiv:2505.12487

Markov Chain Monte Carlo (MCMC)



Random Walk Metropolis (RWM)

- Current state $X(t) = x$;
- Draw a proposal $y \sim \mathcal{N}(x, \sigma^2 I)$;
- The proposal is accepted with probability $\min\{1, \frac{\pi(y)}{\pi(x)}\}$;
- $X(t+1) = y$ if accepted; $X(t+1) = x$ otherwise.

✓ Easy to implement;

✓ Gradient-free;

✗ Inefficient exploration in high-dimensional space.

Multi-Try Metropolis (MTM) [LLW'00, PBF'10 et al.]

- Current state $X(t) = x$;
- Draw N new states independently $y_1, \ldots, y_N \sim \mathcal{N}(x, \sigma^2 I)$;
- Select one proposal y_j with probability proportional to weight $\omega(x, y_j)$;
- Draw $N-1$ auxiliary variables independently $z_1, \ldots, z_{N-1} \sim \mathcal{N}(y_j, \sigma^2 I)$;
- The proposal is accepted with probability

$$\min\{1, \frac{\pi(y_j)\omega(y_j, x)/(\sum_{i=1}^{N-1} \omega(y_j, z_i) + \omega(y_j, x))}{\pi(x)\omega(x, y_j)/(\sum_{i=1}^{N} \omega(x, y_i))}\};$$

- $X(t+1) = y_j$ if accepted; $X(t+1) = x$ otherwise.

Multi-Try Metropolis (MTM)

- Choices of weight function:
  - global balance (GB): $\omega(x, y) = \pi(y)/\pi(x)$, [GMZ'23]
  - local balance (LB): $\omega(x, y) = \sqrt{\pi(y)/\pi(x)}$. [GMZ'23]
  - others: $\pi(y)/(\pi(x) + \pi(y))$, $(\pi(y))^2$, $\pi(y)/q(y \mid x)$.

✓ Easy to implement, gradient-free;

✓ Improved exploration;

✗ For a normal target, GB-MTM may be trapped when N is large [Gagnon, et.al.'23].

✗ For a log-concave target, the "performance gain" using N candidates is bounded above by $\log(N)$ [Pozza and Zanella'24];

Heavy-tailed Targets

- Heavy-tailed distribution: heavier than exponential tail
  e.g. multivariate student's t distribution
- Geometric ergodicity

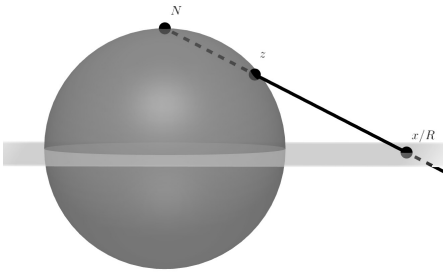
$$\|P^k(x,\cdot) - \pi(\cdot)\|_{\mathsf{TV}} \leq C(x) r^k$$

✗ For any heavy-tailed target, RWM is not geometrically ergodic. [Jarner and Hansen'00]

Theorem [W. and Yang'25]

✗ For any heavy-tailed target, MTM is not geometrically ergodic.

Stereographic MCMC [Yang, Łatuszyński and Roberts'24]



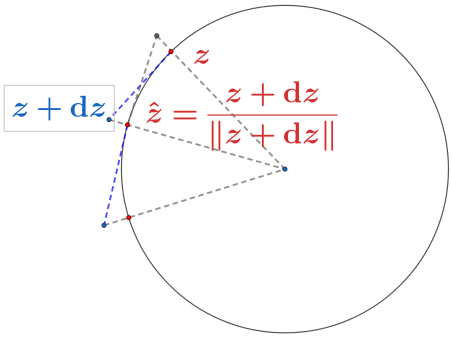
- Bijection: $\mathbb{R}^d \to \mathbb{S}^d \backslash \{\mathscr{N}\}$, where $\mathscr{N} = (0, \ldots, 0, 1)$ is North pole.

Stereographic Random Walk Metropolis (SRWM)

- Let the current state be $X(t) = x$;
- Compute the proposal y:
  - Let $z := \mathsf{SP}^{-1}(x)$;
  - Sample independently $\hat{z} \sim Q_S(z, \cdot)$;
  - The proposal $y := \mathsf{SP}(\hat{z})$.
- $X(t+1) = y$ with probability

$$\min\left\{1, \frac{\pi(y)(R^2 + \|y\|^2)^d}{\pi(x)(R^2 + \|x\|^2)^d}\right\}$$

otherwise $X(t+1) = x$.



Stereographic Multi-Try Metropolis (SMTM)

- Current state $X^d(t) = x$;
- Let $z = \mathsf{SP}^{-1}(x)$;
- Draw $\hat{z}_1, \ldots, \hat{z}_N$ independently from $Q_S(z, \cdot)$;
- Select $\hat{z} = \hat{z}_j$ with probability proportional to $\omega(z, \hat{z}_j)$;
- Draw $z_1^*, \ldots, z_{N-1}^*$ independently from $Q_S(\hat{z}_j, \cdot)$;
- The proposal $\hat{x}_j = \mathsf{SP}(\hat{z}_j)$
- $X^d(t+1) = \hat{x}_j$ with probability

$$\min\{1, \frac{\pi_S(\hat{z}_j)\omega(\hat{z}_j, z)/(\sum_{i=1}^{N-1} \omega(\hat{z}_j, z_i^*) + \omega(\hat{z}_j, z))}{\pi_S(z)\omega(z, \hat{z}_j)/(\sum_{i=1}^{N} \omega(z, \hat{z}_i))}\};$$

otherwise $X^d(t+1) = x$.

Ergodicity for Heavy-Tailed Targets

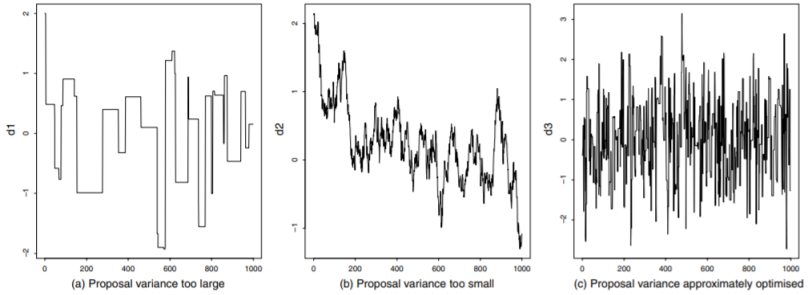
Algorithm	Ergodicity
RWM	not geometrically ergodic [Jarner and Hansen'00]
MTM	not geometrically ergodic [W. and Yang'25]
SRWM	uniformly ergodic* [Yang, Łatuszyński and Roberts'24]
SMTM	uniformly ergodic* [W. and Yang'25]

$*$: For targets with tails heavier than exponential, but lighter than Cauchy. [Grazzi et.al]

Optimal Scaling [Roberts, et.al, 97]

Draw a new state $y \sim \mathcal{N}(x, \sigma^2 I)$.

- If σ is too large, the chain frequently proposes moves into low-probability regions;
- If σ is too small, the Markov chain explores the state space very slowly.



Optimal Scaling [Roberts, et.al, 97]

- Maximizing expected squared jumping distance (ESJD) w.r.t. step size σ:

$$\text{ESJD} = \mathbb{E}_{X^d(t)\sim\pi}\mathbb{E}_{X^d(t+1)\,|\,X^d(t)}\left[\|X^d(t+1) - X^d(t)\|^2\right];$$

- Assume the target distribution $\pi(x)$ has a product i.i.d. form

$$\pi(x) = \prod_{i=1}^{d} f(x_i);$$

- Assume step size σ can be re-parameterized by ℓ through $\sigma = \ell/\sqrt{d}$, and thus σ is scaled as $O(d^{-1/2})$.

Optimal Scaling of RWM and MTM

RWM [RGG'97]	MTM [BDM'12]
ESJD $\approx \ell^2 \mathbb{E}\left[\phi_1(X)\right]$	ESJD $\approx \ell^2 \mathbb{E}\left[\phi_N\left((W_i)_{i=1}^N, (V_i)_{i=1}^{N-1}\right)\right]$
$\phi_1(x) = 1 \wedge e^x$	$\phi_N\left((x_i)_{i=1}^N, (y_i)_{i=1}^{N-1}\right) = \frac{Ne^{x_j}}{\sum_{i=1}^N e^{x_i}} \wedge \frac{Ne^{x_j}}{\sum_{i=1}^{N-1} e^{x_j} e^{y_i} + 1}$
$X \sim \mathcal{N}\left(\frac{-\ell^2}{2} I, \ell^2 I\right)$	$W_i, V_i \sim \mathcal{N}\left(\frac{-\ell^2}{2} I, \ell^2 I\right)$
$\alpha^* \approx 0.23$	$\alpha^* \approx 0.32 (N = 2), \quad 0.37 (N = 3), \quad \ldots$

Optimal Scaling of Stereographc MCMC

- Step size on Euclidean space $\sigma \approx R \cdot h$, where h is step size on unit sphere;
- Assume the parameter of stereographic projection

$$R = \sqrt{\lambda d}$$

- Assume the step size h can be re-parameterized by ℓ through

$$h \approx \frac{\ell}{d} \sqrt{\frac{4\lambda}{(1+\lambda)^2}},$$

  and thus h is scaled as $O(d^{-1})$.
- Step size on Euclidean space is scaled as

$$\sigma \approx R \cdot h = \sqrt{\lambda d} \cdot O(d^{-1}) = O(d^{-1/2}).$$

Optimal Scaling of SRWM and SMTM

SRWM [YŁR'24]	SMTM [W. and Yang'25]
ESJD $\approx \ell^2 \mathbb{E}\left[\phi_1(X)\right]$	ESJD $\approx \ell^2 \mathbb{E}\left[\phi_N((W_i)_{i=1}^N, (V_i)_{i=1}^{N-1})\right]$
$\phi_1(x) = 1 \wedge e^x$	$\phi_N\left((x_i)_{i=1}^N, (y_i)_{i=1}^{N-1}\right) = \frac{Ne^{x_j}}{\sum_{i=1}^N e^{x_i}} \wedge \frac{Ne^{x_j}}{\sum_{i=1}^{N-1} e^{x_j} e^{y_i} + 1}$
$X \sim \mathcal{N}\left(-\frac{\ell^2}{2}(I - \frac{4\lambda}{(1+\lambda)^2}), \ell^2(I - \frac{4\lambda}{(1+\lambda)^2})\right)$	$W_i, V_i \sim \mathcal{N}\left(-\frac{\ell^2}{2}(I - \frac{4\lambda}{(1+\lambda)^2}), \ell^2(I - \frac{4\lambda}{(1+\lambda)^2})\right)$
$\alpha^* \approx 0.23$	$\alpha^* \approx 0.32 (N = 2), \quad 0.37 (N = 3), \quad \ldots$

Maximum ESJD: SMTM versus MTM

- MTM: $\mathcal{N}\left(\frac{-\ell^2}{2}I, \ell^2 I\right)$ SMTM: $\mathcal{N}\left(-\frac{\ell^2}{2}(I - \frac{4\lambda}{(1+\lambda)^2}), \ell^2(I - \frac{4\lambda}{(1+\lambda)^2})\right)$
- Let $\ell'^2 I = \ell^2\left(I - \frac{4\lambda}{(1+\lambda)^2}\right)$, then

$$\mathsf{ESJD}_{\mathsf{SMTM}} \approx \frac{I}{I - \frac{4\lambda}{(1+\lambda)^2}} \ell'^2 \mathbb{E}_{\mathcal{N}\left(\frac{-\ell'^2}{2}I, \ell'^2 I\right)}[\phi_N]$$

- Maximizing ESJD we have

$$\max_{\ell} \mathsf{ESJD}_{\mathsf{MTM}} = \max_{\ell} \ell^2 \mathbb{E}_{\mathcal{N}\left(\frac{-\ell^2}{2}I, \ell^2 I\right)}[\phi_N]$$

$$\max_{\ell} \mathsf{ESJD}_{\mathsf{SMTM}} = \frac{I}{I - \frac{4\lambda}{(1+\lambda)^2}} \max_{\ell} \ell^2 \mathbb{E}_{\mathcal{N}\left(\frac{-\ell^2}{2}I, \ell^2 I\right)}[\phi_N]$$

Robustness to tuning of SMTM

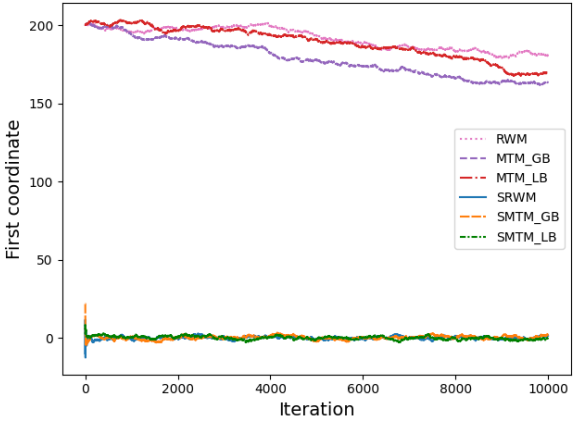
Theorem (W. and Yang'25)

$$\frac{\max_\ell \mathsf{ESJD}_{\mathsf{SMTM}}}{\max_\ell \mathsf{ESJD}_{\mathsf{MTM}}} = \frac{1}{1-\alpha\cdot\beta\cdot\gamma}, \quad \alpha,\beta,\gamma\in[0,1]$$

- $\alpha = \frac{4\lambda}{(1+\lambda)^2}$, which is a penalty for misspecified radius $R=\sqrt{\lambda d}$;
- $\beta = 1-(\mathbb{E}_f[X])^2$, which is a penalty for mis-locating the sphere;
- $\gamma = \frac{1}{(1-(\mathbb{E}_f[X])^2)I}$, which is a distribution-specific penalty.

The optimal acceptance rate for SMTM is the same with MTM.

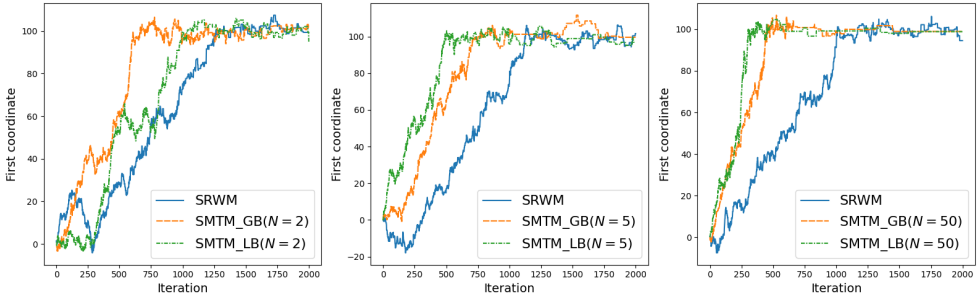
Numerical Experiments (Uniform Ergodicity)



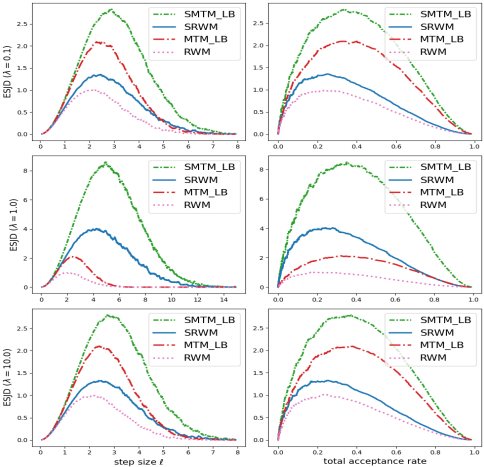
- Target distribution: $\pi(x) = \prod_{i=1}^{100} f(x_i)$,
- f: standard student's t distribution with degree of freedom 101.

Numerical Experiments (Robustness to Locations)

- Center of the sphere: 0
- Mean value of the target: 100



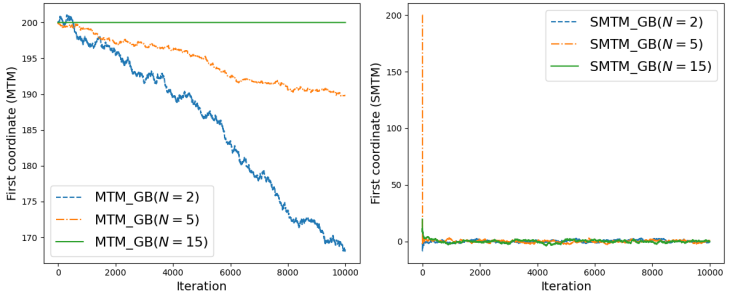
Numerical Experiments (Robustness to Radiuses)



	Proper radius	Improper radius
RWM	✗	✗
MTM	✓	✓
SRWM	✓✓	✗
SMTM	✓✓	✓✓

Other Results and Open Problems

- For a family of heavy-tailed targets, the GB-SMTM chain does not get stuck.



- For a family of heavy-tailed targets, the "performance gain" of SMTM from using N candidates is bounded above by $N^{2/p}$ instead of $\log(N)$. (Matching) lower bound?
- When $N \to \infty$, the acceptance rate of LB-SMTM goes to 1 if $h = O(d^{-1})$. (Almost) gradient-based? Scaling analysis when $h = O(d^{-2/3})$ and $N \to \infty$.

- Stereographic Multi-Try Metropolis;
- Uniform Ergodicity:
  - SMTM is uniformly ergodicity for a family of heavy-tailed targets.
- Optimal Scaling:
  - Larger ESJD than SRWM, MTM and RWM;
  - More robust to parameters tuning.
- Compared to MTM:
  - Not get stuck as $N \to \infty$ in global balance case;
  - (Open) "performance gain" upper bounded by $N^{2/p}$ instead of $\log(N)$;
  - (Open) LB-SMTM performs as (almost) gradient-based when $N \to \infty$.

Thank you!