Optimal scaling of Metropolis-adjusted Langevin algorithm with Laplace target

Alain Durmus, Pablo Jiménez, Eric Moulines, Gareth O. Roberts
alain.durmus@ens-cachan.fr, pablo.jimenez-moreno@polytechnique.edu,
eric.moulines@polytechnique.edu, gareth.o.roberts@warwick.ac.uk

1. The Metropolis-adjusted Langevin algorithm (MALA)

- Goal: to sample from $\pi(x) = \exp[-V(x)]$, a probability distribution over $\mathbb{R}^d$.
- **Langevin SDE**:

$$\mathrm{d}\mathbf{X}_t = -\nabla V(\mathbf{X}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t \,,$$

where $(\mathbf{B}_t)_{t\geqslant 0}$ is a d-dimensional Brownian motion.

– π is invariant,

⚠ difficult to calculate,

⇝ consider a discrete version.

- Euler-Maruyama with step-size $\sigma_d^2 > 0$

⇝**Unadjusted Langevin algorithm**:

$$X_{k+1}^d = X_k^d - \sigma_d^2 \nabla V(X_k^d) + \sqrt{2\sigma_d^2} Z_{k+1}^d \,,$$

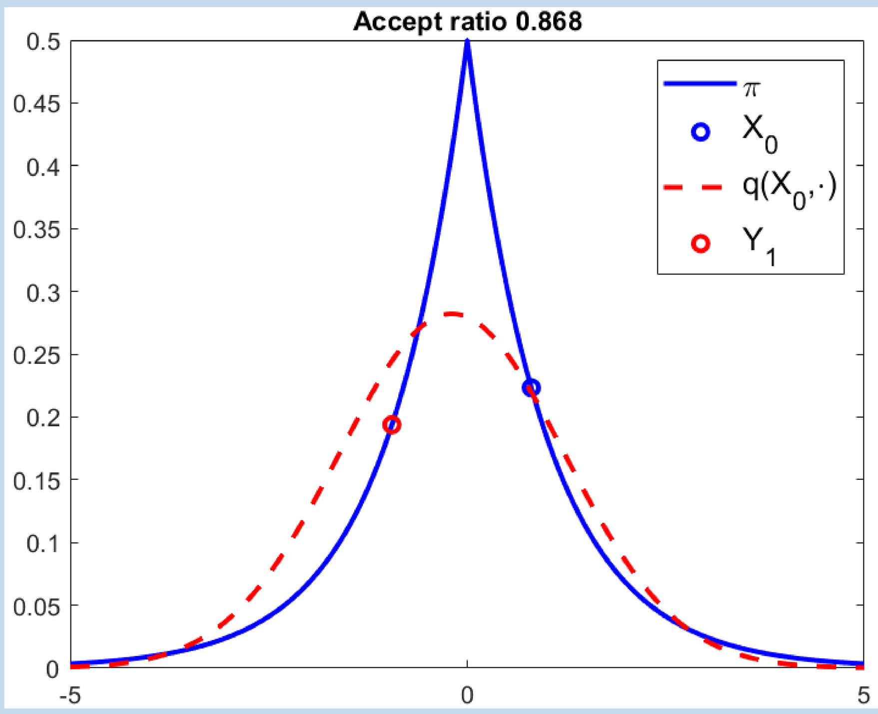
with $(Z_k^d)_{k\in\mathbb{N}}$ i.i.d $\sim \mathrm{N}(0, \mathrm{Id}_d)$.

⚠ invariant measure $\neq \pi$,

⇝ introduce as a Metropolis-Hastings proposal.

- **MALA**:

$$Y_{k+1}^d = X_k^d - \sigma_d^2 \nabla V(X_k^d) + \sqrt{2\sigma_d^2} Z_{k+1}^d \,,$$
$$X_{k+1}^d = Y_{k+1}^d \mathbb{1}_{\mathsf{A}_{k+1}^d} + X_k^d \left(1 - \mathbb{1}_{\mathsf{A}_{k+1}^d}\right) \,,$$



– the *accept/reject* step makes π invariant,

– **biased** proposal ⇝ gradient step,

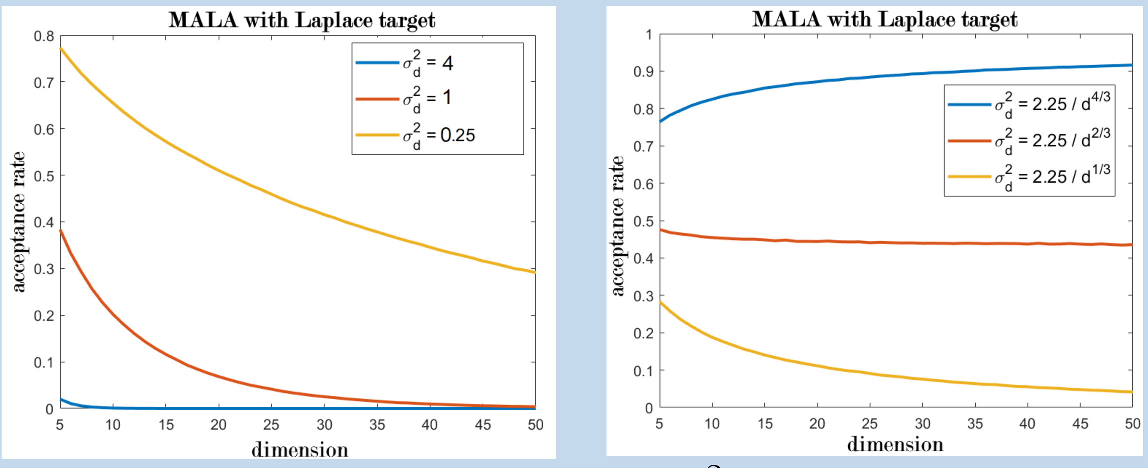
– the method depends on V ⇝ π,

⚠ need to calibrate σ_d^2.

2. How to choose the step size? The optimal scaling problem

- Optimal scaling problem for Laplace target:
  – consider the target sequence $(\pi^d)_{d\geqslant 1}$ given by $\pi^d(x^d) = \exp(-\|x^d\|_1)/2^d$;
  – for each $d \geqslant 1$, $(X_n^d)_{n\geqslant 0}$, the Markov chain from MALA, started at **stationnarity**.
- First criterion for σ_d^2: **scaling**
  – scaling criterion: choose $(\sigma_d^2)_{d\geqslant 1}$ so that $\lim_{d\to+\infty} \mathbb{P}(\mathsf{A}_1^d) \in]0,1[$;

⚠ if $(\sigma_d^2)_{d\geqslant 1}$ is *constant* w.r.t. d, $\lim_{d\to+\infty} \mathbb{P}(\mathsf{A}_1^d) = 0$.

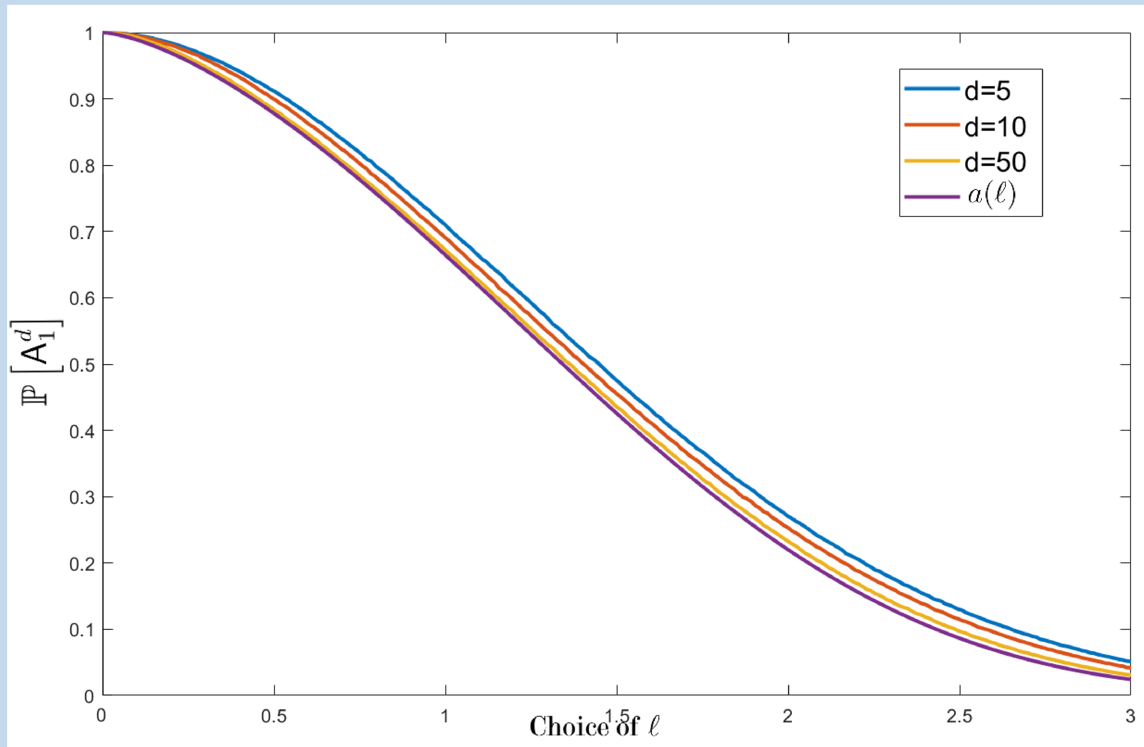


(a) constant σ_d^2 (b) $\sigma_d^2 \to 0$ w.r.t. d

Proposition (*Convergence of the acceptance rate*)
For any $\ell > 0$, $d \in \mathbb{N}^*$, if $\sigma_d^2 = \ell^2/d^{2/3}$, then

$$a(\ell) := \lim_{d\to+\infty} \mathbb{P}(\mathsf{A}_1^d) = 2\Phi\left[-\ell^{3/2}/(3\pi^{1/2})^{1/2}\right] \,,$$

where Φ is the standard Gaussian cdf.



acceptance rate as $d \to +\infty$

3. Fine-tuning through a diffusive limit

- Goal: finding an **optimization** criterion for $\sigma_d^2 = \ell^2/d^{2/3}$,
- **Interpolation** and **scaling** of X^d:

$$\mathbf{Y}_t^d = (\lceil d^{2/3}t \rceil - d^{2/3}t) X_{\lfloor d^{2/3}t \rfloor}^d + (d^{2/3}t - \lfloor d^{2/3}t \rfloor) X_{\lceil d^{2/3}t \rceil}^d \,.$$

Theorem (*Weak convergence to a Langevin diffusion*)
There exist $h : \mathbb{R}_+^* \to \mathbb{R}_+^*$ such that for any $\ell > 0$, $\mathbf{Y}_1^d$ converges weakly towards a one-dimensional process $\mathbf{Y}$, solution to the Langevin SDE:

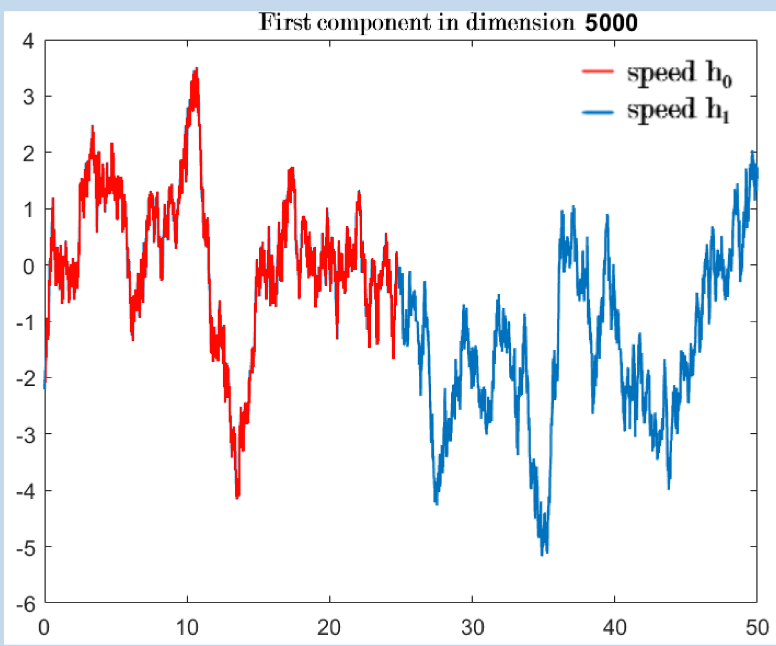
$$\mathrm{d}\mathbf{Y}_t = -h(\ell)\mathrm{sgn}(\mathbf{Y}_t)\mathrm{d}t + \sqrt{2h(\ell)}\mathrm{d}\mathbf{B}_t \,.$$

In addition $h(\ell) = \ell^2 a(\ell)$, where $a(\ell) := \lim_{d\to+\infty} \mathbb{P}(\mathsf{A}_1^d)$.

- maximising h **optimizes the scaling**:
  – h is the limit of the first order efficiency

$$2\,h(\ell) = \lim_{d\to+\infty} d^{2/3} \mathbb{E}\left[(X_{0,1}^d - X_{1,1}^d)^2\right] \,,$$

  – h is also the speed of the diffusion,
  – if $\mathbf{Y}^{(1)}$ satisfies the SDE with $h = 1$, $(\mathbf{Y}_{h(\ell)t}^{(1)})_{t\geqslant 0} \overset{law}{=} (\mathbf{Y}_t)_{t\geqslant 0}$,

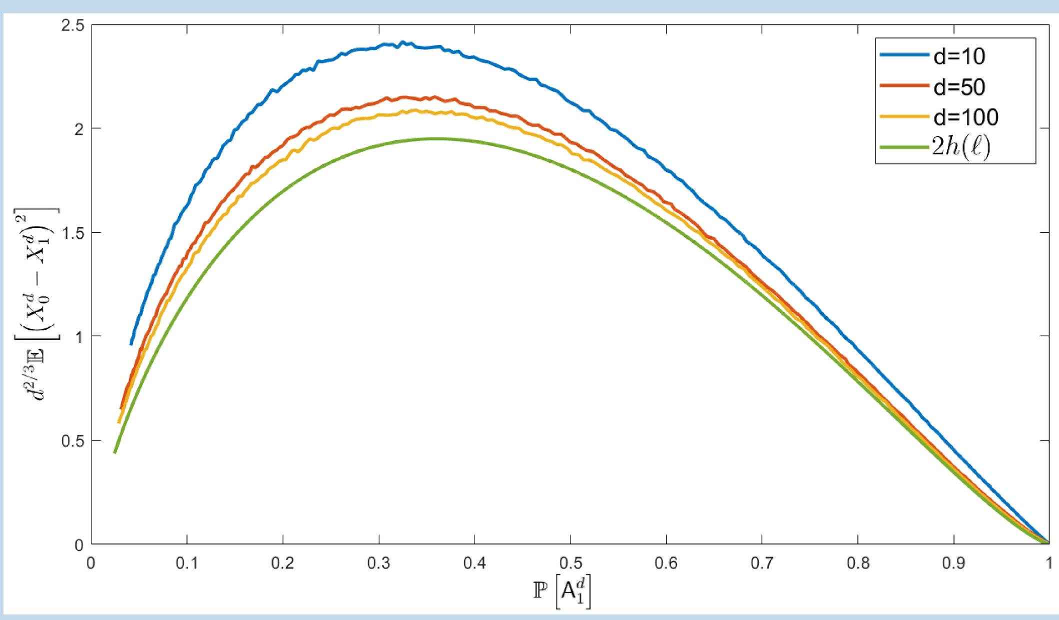


$\neq$ diffusion speeds $h_0 < h_1$ ⇝$\neq$ convergence speeds

⇝ $\mathbf{Y}$ ergodic, π invariant ⇒ **maximising h** gives the best convergence,

⇝ **optimization criterion** for ℓ, therefore σ_d^2,

⇝ $\mathbf{a} = 0.360$ ⇒ **optimal acceptance rate.**



first order efficiency as a function of acceptance rate

4. Comparing to other scalings

- $\alpha = 2/3$ gives an estimate of the **complexity**:
  – scaling $\mathbf{Y}^d$ in $d^{2/3}$ ⇝ $\mathrm{O}(d^{2/3})$ steps to converge (similar idea to speed h),

⇝ **comparing scaling** values allows to compare methods.,

  – see [3] for more details.

	MALA	RWM
Smooth	[4] $\alpha = 1/3$ $\mathbf{a} = 0.574$	[2] $\alpha = 1$ $\mathbf{a} = 0.234$
Non-smooth	⚠ no general result	[1] $\alpha = 1$ $\mathbf{a} = 0.234$

⇝ Laplace target has a new scaling,

⚠ smoothness is important for MALA.

Bibliography

[1] A. Durmus et al. "Optimal scaling of the random walk Metropolis algorithm under Lp mean differentiability". In: *Journal of Applied Probability* 54.4 (2017), pp. 1233–1260. DOI: 10.1017/jpr.2017.61.

[2] G.O. Roberts, A. Gelman, and W.R. Gilks. "Weak convergence and optimal scaling of random walk Metropolis algorithms". In: *The Annals of Applied Probability* 7.1 (1997), pp. 110–120.

[3] G.O. Roberts and J.S. Rosenthal. "Optimal scaling for various Metropolis-Hastings algorithms". In: *Statistical science* 16.4 (2001), pp. 351–367.

[4] G.O. Roberts and J.S. Rosenthal. "Optimal scaling of discrete approximations to Langevin diffusions". In: *J. R. Statist. Soc. B* 60 (1997), pp. 255–268.