Log-concave sampling: Metropolis-Hastings algorithms are fast!

Berkeley | EECS Electrical Engineering and Computer Sciences

Raaz Dwivedi, Yuansi Chen, Martin Wainwright and Bin Yu

{raaz.rsk, yuansi.chen, wainwrig, binyu}@berkeley.edu

University of California, Berkeley
Department of Statistics

Sampling: A fundamental task!

Given a density, with **unknown normalization constant**, draw (approximate) samples from it

$$X_1, \ldots, X_n \overset{i.i.d.}{\sim} \pi^*(x) \propto e^{-f(x)}$$

- **Integration:** $\int g(x)\pi^*(x)dx \approx \frac{1}{n}\sum_{i=1}^{n} g(X_i)$
- **Optimization:** $\min_{x\in\mathcal{K}} g(x) \approx \min\{g(X_1), \ldots, g(X_n)\}$

Set-up and Objectives

Goal: Given access to $f(x)$ and $\nabla f(x)$ for any $x \in \mathbb{R}^d$, derive **non-asymptotic mixing-time bound for k**

$$k_{\text{mix}}(\delta) = \min_{k=1,2,\ldots} \{k \mid \|\mathbb{P}(X_k) - \pi^*\|_{\text{TV}} \le \delta\}$$

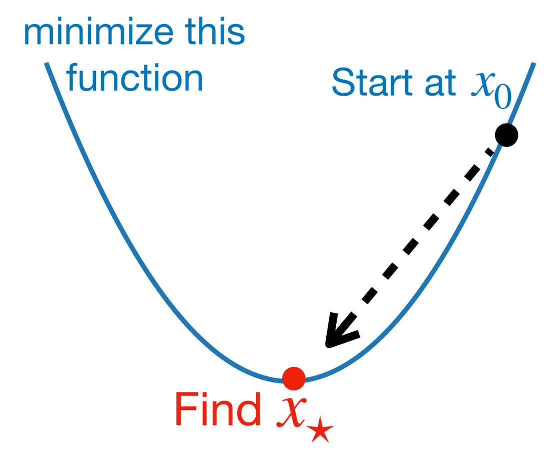
with explicit dependence on d, δ and other parameters.

Algorithms: Optimization vs Sampling

Discretization of gradient flow (ordinary differential equation) vs Langevin diffusion (stochastic differential equation):

Gradient Flow $d\tilde{x}_t = -\nabla f(\tilde{x}_t)dt$

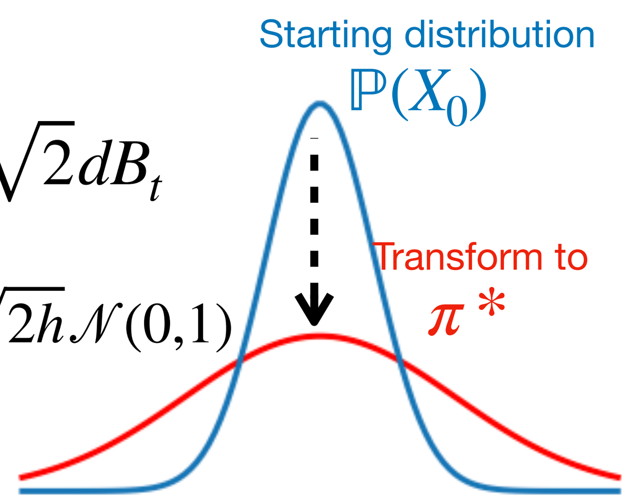
Gradient Descent $x_{k+1} = x_k - h\nabla f(x_k)$



h = step size
k = no. of steps

Langevin Diffusion $d\tilde{X}_t = -\nabla f(\tilde{X}_t)dt + \sqrt{2}dB_t$

Unadjusted Langevin Algorithm $X_{k+1} = X_k - h\nabla f(X_k) + \sqrt{2h}\mathcal{N}(0,1)$



Unadjusted Langevin algorithm (ULA)

- ULA is biased:

$$\|\mathbb{P}(\tilde{X}_t) - \pi^*\|_{\text{TV}} \to 0 \quad \text{but} \quad \|\mathbb{P}(X_k) - \pi^*\|_{\text{TV}} \not\to 0.$$

- Large $h \Rightarrow$ Faster convergence, large asymptotic bias, and, small $h \Rightarrow$ slower convergence, small bias

Metropolis adjusted Langevin algorithm

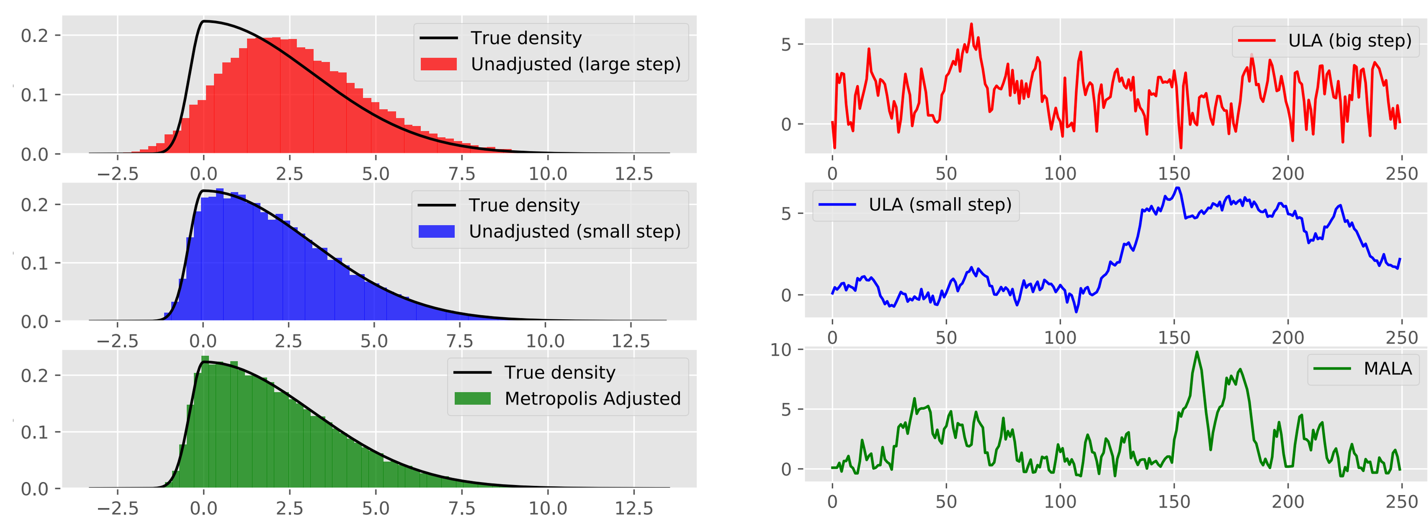
- MALA is made up of two steps

$$Z = X_k - h\nabla f(X_k) + \sqrt{2h}\mathcal{N}(0,1) \quad \text{(proposal step)}$$

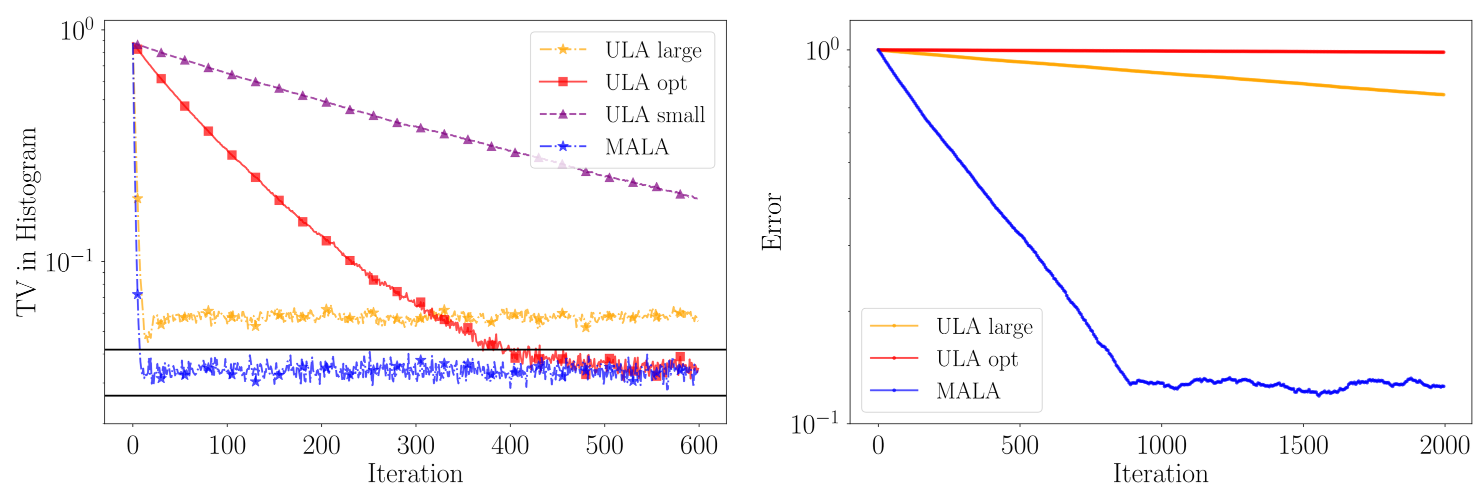
$$\mathbb{P}(X_{k+1} \leftarrow Z) = \min\left\{1, \frac{\pi^*(Z)P(Z\to X_k)}{\pi^*(X_k)P(X_k \to Z)}\right\} \quad \text{(accept-reject step)}$$

If case of rejection, $X_{k+1} \leftarrow X_k$.

- Accept-reject step $\Rightarrow$ Detailed balance $\Rightarrow$ Unbiasedness $\Rightarrow$ larger step size $\Rightarrow$ Faster mixing.



(a) Histogram upon convergence (b) One run of the chain



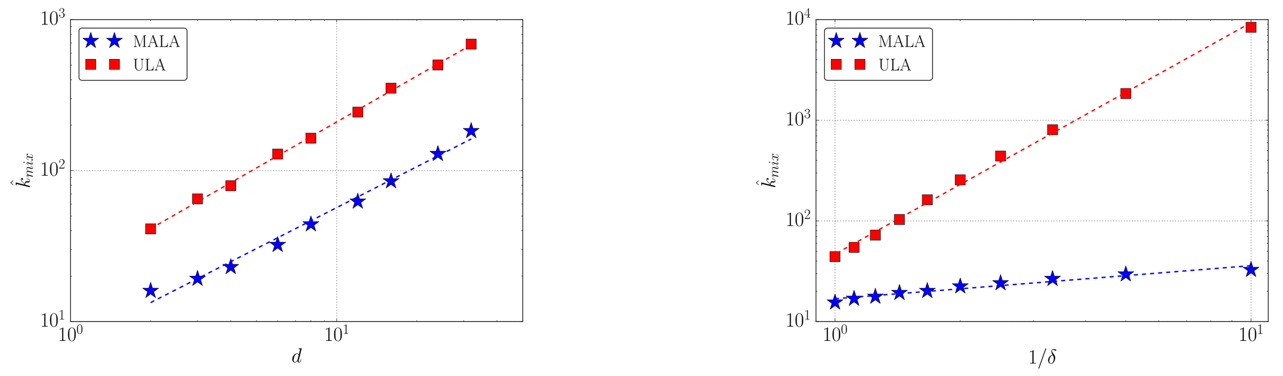
(c) Histogram error with iterations (d) Bayesian posterior mean estimation

Figure 1: Comparison of ULA and MALA

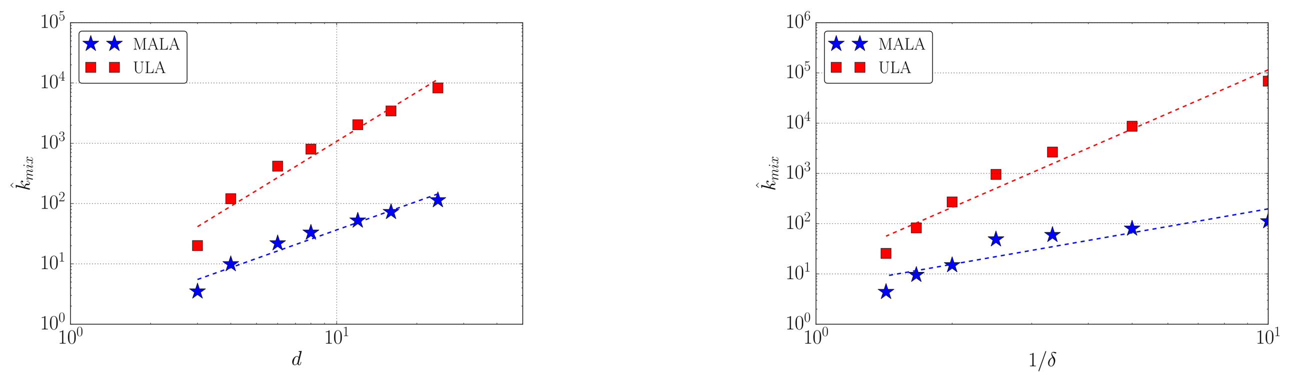
Conditions on f	ULA [1]	MALA [1]
L-smooth, m-strongly convex (strongly log-concave, SLC)	$d\,\frac{L^2}{m^2}\,\frac{1}{\delta^2}$	$d\,\frac{L}{m}\,\log\frac{1}{\delta}$
L-smooth (weakly log-concave, WLC)	$d^3\,L^2\,\frac{1}{\delta^4}$	$d^2\,L^{1.5}\,\frac{1}{\delta^{1.5}}$

Table 1: Our results: Mixing time bounds $k_{\text{mix}}(\delta)$ for MALA.

Numerical Experiments



(a) SLC: d dependency (b) SLC: δ dependency



(c) WLC: d dependency (d) WLC: δ dependency

Figure 2: Scalings in simulations for ULA and MALA

Proof Techniques

High conductance implies fast mixing:

Conductance: $\Phi = \inf_{\pi^*(A_1)<1/2} \int_{u\in A_1} \mathbb{P}(u \to A_1^c)\pi^*(u)du/\pi^*(A_1)$

Mixing time: $k_{\text{mix}}(\delta) \le \mathcal{O}(\log(1/\delta)/\Phi^2)$

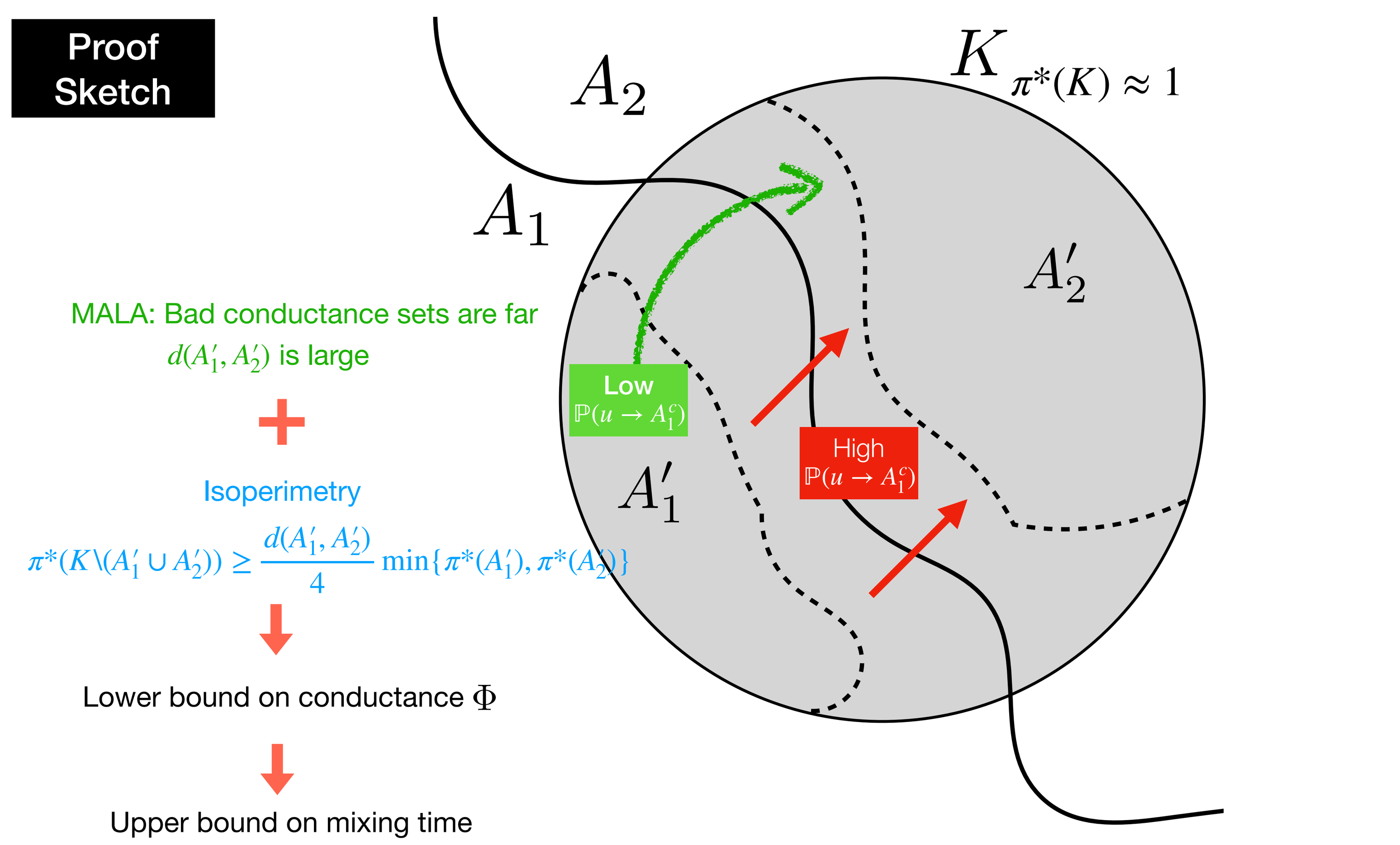


Figure 3: Proof sketch

References

[1] A. S. Dalalyan, *Theoretical guarantees for approximate sampling from smooth and log-concave densities*, JRSSB (2016).

[2] R. Dwivedi, Y. Chen, M. J. Wainwright, B. Yu, *Log-concave sampling: Metropolis-Hastings algorithms are fast!*, extended abstract at COLT (2018), full version at arXiv:1801.02309.