Improved dimension dependence for MALA and lower bounds for sampling

Sinho Chewi

Simons Institute
Geometric Methods in Optimization and Sampling (2021)

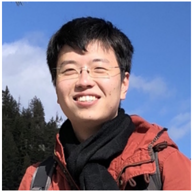
Massachusetts Institute of Technology

Collaborators

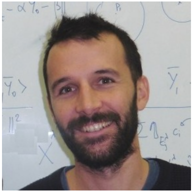
Kwangjun Ahn

Patrik Gerber

Chen Lu



Xiang Cheng



Thibaut Le Gouic

Philippe Rigollet

Massachusetts Institute of Technology

Optimization and Sampling

Optimization	**Sampling**
objective function $f : \mathbb{R}^d \to \mathbb{R}$	target distribution $\pi \propto \exp(-V)$
gradient descent, mirror descent, proximal methods ...	Langevin, mirror-Langevin, proximal Langevin ...
non-asymptotic theory of complexity	in progress!

Complexity of Sampling

Problem: What is the minimum number of queries to V and ∇V needed to output an approximate sample from the target distribution $\pi \propto \exp(-V)$ on $\mathbb{R}^d$?

Throughout, we assume that $\arg\min V = 0$ and

$$\alpha I_d \preceq \nabla^2 V \preceq \beta I_d \,, \qquad \kappa := \frac{\beta}{\alpha} \,.$$

where κ is the condition number.

Outline

an improved complexity bound for the Metropolis-adjusted Langevin algorithm (MALA)

lower bounds for MALA

recent progress towards general sampling lower bounds

Metropolis-Hastings Algorithms

1. initialize at $x_0 \sim \mu_0$
2. for $n = 0, 1, 2, \ldots$:

   propose

$$y_{n+1} \sim \underbrace{Q(x_n, \cdot)}_{\boxed{\text{proposal kernel}}}$$

   accept y_{n+1} with probability

$$a(x_n, y_{n+1}) = 1 \wedge \frac{\pi(y_{n+1})\, Q(y_{n+1}, x_n)}{\pi(x_n)\, Q(x_n, y_{n+1})}$$

Examples

Metropolized random walk (MRW):

$$Q(x, \cdot) = \mathrm{normal}(x, 2hI_d)$$

Metropolis-adjusted Langevin algorithm (MALA):

$$Q(x, \cdot) = \mathrm{normal}\big(x - h\nabla V(x), 2hI_d\big)$$

Metropolized Hamiltonian Monte Carlo (HMC): $Q(x, \cdot) = K$ steps of leapfrog integrator of HMC

Analysis of MH Algorithms

the good:

- Markov chain with correct stationary distribution π
- typically $\mathrm{polylog}(1/\varepsilon)$ dependence on the accuracy ε
- widely used in practice

the bad:

- difficult to control the acceptance probability

What can we say about the non-asymptotic complexity?

Known Results

algorithm	gradient queries
MRW	$\widetilde{O}(d\kappa^2 \log \frac{1}{\varepsilon})$
MALA	$\widetilde{O}(d\kappa \log \frac{1}{\varepsilon})$
MHMC	$\widetilde{O}(d\kappa \operatorname{polylog} \frac{1}{\varepsilon})$

Can we do better?

Non-asymptotic bounds: [Chen, Dwivedi, Wainwright, Yu '19] [Dwivedi, Chen, Wainwright, Yu '19] [Lee, Shen, Tian '20]

Better bounds under higher-order smoothness: [Chen, Dwivedi, Wainwright, Yu '19] [Mangoubi and Vishnoi '19]

Diffusion Scaling Heuristic

Roberts and Rosenthal '98 showed that for product distributions, MALA with step size $\ell/d^{1/3}$ converges ($d \to \infty$) to a Langevin diffusion with speed $s(\ell)$.

Assumption: higher-order regularity of V.

They concluded:

(1) MALA should have dimension dependence $\Theta(d^{1/3})$;

(2) there is an explicit and optimal choice of ℓ.

$\boxed{\text{What can be achieved non-asymptotically?}}$

Our Result

Theorem: Under a $\underbrace{\text{warm start}}$

$$\boxed{\sup \frac{\mu_0}{\pi} \leq O(1)}$$

we obtain an improved mixing time bound for MALA ,

$$\widetilde{O}\big(\sqrt{d}\,\mathrm{poly}\big(\kappa, \log \frac{1}{\varepsilon}\big)\big)\ ,$$

to reach ε-accuracy in any standard metric (TV, W_2, KL, χ^2).

Proof: Conductance

Let T denote the MALA kernel. Define the **conductance**

$$\mathsf{C} := \inf\Big\{ \frac{\int_S T(x, S^{\mathsf{c}})\,\mathrm{d}\pi(x)}{\pi(S)} \;\Big|\; S \subseteq \mathbb{R}^d,\ 0 < \pi(S) < \frac{1}{2} \Big\}\,.$$

Standard result for Markov chain convergence: the mixing time in TV is bounded by

$$n_{\mathrm{mix}} = O\Big(\frac{1}{\mathsf{C}^2} \log \frac{M_0}{\varepsilon}\Big)\,, \qquad M_0 = \text{warm start parameter}\,.$$

[Lovász and Simonovits '93]

Proof: s-Conductance

Let T denote the MALA kernel. Define the s-**conductance**

$$\mathrm{C}_s := \inf\Big\{\frac{\int_S T(x, S^{\mathrm{c}})\,\mathrm{d}\pi(x)}{\pi(S) - s} \;\Big|\; S \subseteq \mathbb{R}^d,\ s < \pi(S) < \frac{1}{2}\Big\}\,.$$

Standard result for Markov chain convergence: the mixing time in TV is bounded by

$$n_{\mathrm{mix}} = O\Big(\frac{1}{\mathrm{C}_s^2}\log\frac{M_0}{\varepsilon}\Big)\,, \qquad M_0 = \text{warm start parameter}\,,$$

where $s = \varepsilon/(2M_0)$. [Lovász and Simonovits '93]

Proof: Conductance Lemma

Lemma [Lee and Vempala, '18]: Suppose that $\|x - y\| \leq r \implies \|T_x - T_y\|_{\mathrm{TV}} \leq \frac{3}{4}$. Then, $\mathsf{C} \gtrsim \sqrt{\alpha}\, r$.

Proof: s-Conductance Lemma

Lemma [Lee and Vempala, '18]: Suppose that $\|x - y\| \le r \implies \|T_x - T_y\|_{\mathrm{TV}} \le \frac{3}{4}$. Then, $\mathsf{C} \gtrsim \sqrt{\alpha}\, r$.

Lemma: Suppose that $\|x - y\| \le r \implies \|T_x - T_y\|_{\mathrm{TV}} \le \frac{3}{4}$ for all x, y in an event of π-probability $\ge 1 - O(rs)$. Then, $\mathsf{C}_s \gtrsim \sqrt{\alpha}\, r$.

$\implies$ $\boxed{\textbf{Goal}\text{: Bound the "overlap" } \|T_x - T_y\|_{\mathrm{TV}} \text{ w.h.p.}}$

Proof: Bounding the Overlap

Prior work used the bound

$$\|T_x - T_y\|_{\mathrm{TV}} \leq \|T_x - Q_x\|_{\mathrm{TV}} + \|Q_x - Q_y\|_{\mathrm{TV}} + \|T_y - Q_y\|_{\mathrm{TV}}$$

where Q is the proposal kernel.

- middle term is easy to bound
- key step: how to bound first and last terms?

Proof: Projection Property

Goal: Bound $\| T_x - Q_x \|_{\mathrm{TV}}$ w.h.p.

Theorem [Billera and Diaconis, '01]: The MH kernel T is the projection of Q to $\{$reversible Markov chains with stationary distribution $\pi\}$.

$$\implies \mathbb{E}_{x\sim\pi} \| T_x - Q_x \|_{\mathrm{TV}} \le 2\, \mathbb{E}_{x\sim\pi} \| \bar{Q}_x - Q_x \|_{\mathrm{TV}}$$

Idea: Take $\bar{Q}$ to be the *continuous-time* Langevin dynamics run for time h.

Proof: Pointwise Projection Property

Goal: Bound $\| T_x - Q_x \|_{\mathrm{TV}}$ w.h.p.

We extend the projection theorem:

Theorem: For any reversible kernel $\bar{Q}$ w.r.t. π and any increasing convex function Φ, for $x \sim \pi$ and $y \sim \bar{Q}_x$,

$$2\,\mathbb{E}\,\Phi(\| T_x - Q_x \|_{\mathrm{TV}}) \leq \mathbb{E}\,\Phi(4\,\| \bar{Q}_x - Q_x \|_{\mathrm{TV}}) + \mathbb{E}\,\Phi\Big(2\,\Big|\frac{Q(x,y)}{\bar{Q}(x,y)} - 1\Big|\Big)\,.$$

Reduces the study of MALA to discretization of Langevin!

Massachusetts Institute of Technology

Recap

improved dimension dependence of MALA to $\widetilde{O}(\sqrt{d})$ under a warm start

new technique for studying Metropolis-Hastings chains which relies on well-studied discretization analysis

Two Questions

1. Can we remove the dependence on the warm start $M_0 := \sup \frac{\mu_0}{\pi}$?

   ▷ [Feasible start: $M_0 = \kappa^{d/2}$.]

2. Are there lower bounds for MALA?

   ▷ [We showed: spectral gap $= O(1/\sqrt{d})$.]

Lower Bounds for MALA

Recently, [Lee, Shen, and Tian '21] show that there exist initializations with $M_0 = \exp(d)$ for which the mixing time of MALA is $\widetilde{\Omega}(d)$.

See also Yuansi's talk on Thursday.

Outline

an improved complexity bound for the Metropolis-adjusted Langevin algorithm (MALA)

lower bounds for MALA

recent progress towards general sampling lower bounds

Sampling Lower Bounds

Key challenge for the theory of sampling:

Can we prove lower complexity bounds for sampling?

Some past work:

- algorithm-specific bounds
  - discretization of underdamped Langevin [Cao, Lu, Wang '20]
  - MALA [Chewi et al. '21] [Lee, Shen, Tian '20, '21]
- stochastic gradient queries [Chatterji, Bartlett, Long '20]
- estimating the normalizing constant [Ge, Lee, Lu '20]

A Result in One Dimension

Theorem: The query complexity of sampling from strongly log-concave distributions in one dimension is $\Theta(\log\log\kappa)$.

Some details:

- performance criterion: sample to within $\frac{1}{64}$ in TV distance
- holds for any oracle evaluating (V, V', V'')
- upper bound achieved via rejection sampling

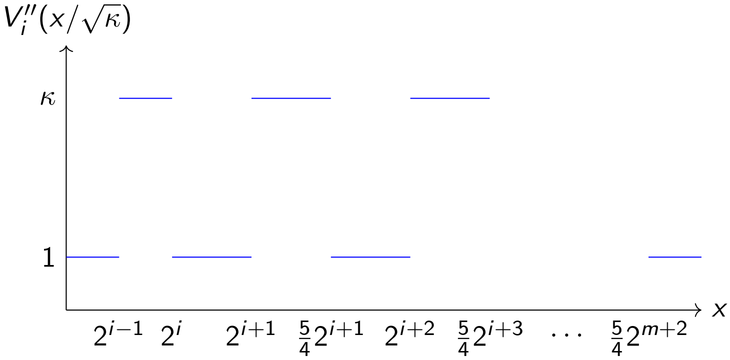
Lower Bound Construction

Strategy of the proof:

Construct family $\mathcal{P}$ of distributions such that

- a single sample from $p \in \mathcal{P}$ *identifies* p, and
- each oracle query reveals only $O(1)$ bits of information.

Lower Bound Construction



Massachusetts Institute of Technology

Open Questions

MALA:

- Can we obtain a warm start?
- What other Metropolis-Hastings algorithms can we analyze?
- How can we Metropolize other algorithms?

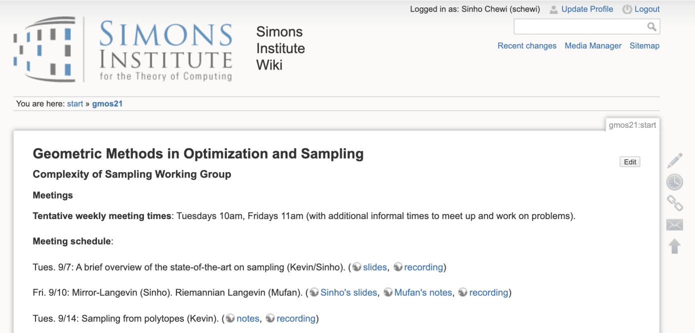
Lower bounds:

- **What is the complexity of sampling?**

Complexity of Sampling Working Group

Meetings: Tuesdays 10am PST, Fridays 11am PST

Email me for a Zoom link!



Simons Wiki (recordings of previous meetings)

Thank You

Sinho Chewi, Chen Lu, Kwangjun Ahn, Xiang Cheng, Thibaut Le Gouic, Philippe Rigollet, *Optimal dimension dependence of the Metropolis-adjusted Langevin algorithm*.

Sinho Chewi, Patrik Gerber, Chen Lu, Thibaut Le Gouic, Philippe Rigollet, *The query complexity of sampling from strongly log-concave distributions in on dimension*.