CPSC 661: Sampling Algorithms in ML

Andre Wibisono

March 8, 2021

Yale University

References

- Dwivedi, Chen, Wainwright, and Yu, *Log-Concave Sampling: Metropolis-Hastings Algorithms are Fast*, Journal of Machine Learning Research, 2019

Recap: To sample from ν on $\mathbb{R}^n$

1. Start from any Markov chain P
2. Apply Metropolis-Hastings filter to get $\tilde{P}$ reversible wrt ν
3. Assume ν is α-SLC, so isoperimetric with $\psi = \Omega(\sqrt{\alpha})$
4. Show $\tilde{P}$ satisfies one-step overlap property:

$$x, y \in \mathcal{R}_s,\ \|x-y\|_2 \leq \Delta_s \quad \Rightarrow \quad \mathrm{TV}(\tilde{P}_x, \tilde{P}_y) \leq \frac{3}{4}$$

Recap: To sample from ν on $\mathbb{R}^n$

1. Start from any Markov chain P
2. Apply Metropolis-Hastings filter to get $\tilde{P}$ reversible wrt ν
3. Assume ν is α-SLC, so isoperimetric with $\psi = \Omega(\sqrt{\alpha})$
4. Show $\tilde{P}$ satisfies one-step overlap property:

$$x, y \in \mathcal{R}_s,\ \|x - y\|_2 \leq \Delta_s \quad \Rightarrow \quad \mathrm{TV}(\tilde{P}_x, \tilde{P}_y) \leq \frac{3}{4}$$

$\Rightarrow$ $\tilde{P}$ has s-conductance $\phi_s = \Omega(\sqrt{\alpha}\,\Delta_s)$

$\Rightarrow$ mixing time in TV distance: $\tau(\epsilon) = O\left(\frac{1}{\alpha\,{\Delta_s}^2} \log \frac{2M}{\epsilon}\right)$

What random walk?

1. $P =$ Brownian motion (Gaussian walk)
   $\Rightarrow \tilde{P} =$ Metropolis Random Walk (MRW)
   (Last time)

2. $P =$ Unadjusted Langevin Algorithm (ULA)
   $\Rightarrow \tilde{P} =$ Metropolis-Adjusted Langevin Algorithm (MALA)
   (Today)

Last time: Metropolis Random Walk (MRW)

To sample from $\nu \propto e^{-f}$ on $\mathbb{R}^n$:

1. From x_k, let
$$y_k = x_k + \sqrt{2\eta}\, z_k$$
where $z_k \sim \mathcal{N}(0, I)$ is independent, $\eta > 0$ is step size.

2. Set
$$x_{k+1} = \begin{cases} y_k & \text{with prob } a_{x_k}(y_k) = \min\left\{1, \dfrac{\nu(y_k)}{\nu(x_k)}\right\} \\ x_k & \text{with prob } 1 - a_{x_k}(y_k). \end{cases}$$

Set up

Assume $\nu \propto e^{-f}$ on $\mathbb{R}^n$ is α-SLC and L-log-smooth:

$$\alpha I \preceq \nabla^2 f(x) \preceq LI$$

Define **condition number**: $\kappa = \dfrac{L}{\alpha} > 1$

- $\nu = \mathcal{N}(\mu, \Sigma)$: $\nabla^2 f(x) = \Sigma^{-1}$, $\alpha = \frac{1}{\lambda_{\max}(\Sigma)}$, $L = \frac{1}{\lambda_{\min}(\Sigma)}$, $\kappa = \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}$

Warm start

$\nu \propto e^{-f}$

Let $x^* = \arg\max_{x \in \mathbb{R}^n} \nu(x) = \arg\min_{x \in \mathbb{R}^n} f(x)$ be the *mode* of ν

Lemma: $\rho_0 = \mathcal{N}\left(x^*, \frac{1}{L} I\right)$ is *warm* with $M = M_\nu^\infty(\rho_0) \leq \kappa^{n/2}$

- With $M = \kappa^{n/2}$, $\log(\frac{2M}{\epsilon}) = O\left(n \log \frac{\kappa}{\epsilon^{1/n}}\right)$
- So mixing time is

$$\tau(\epsilon) = O\left(\frac{n}{\alpha {\Delta_s}^2} \overbrace{\log\left(\frac{\kappa}{\epsilon^{1/n}}\right)}^{\approx \text{ a small constant}}\right) = \tilde{O}\left(\frac{n}{\alpha \Delta_s^2}\right)$$

ignores log factors

Theorem
Under setup above, with step size

$$\eta = c\frac{1}{n\kappa L\log(\frac{\kappa}{\epsilon^{1/n}})} = \tilde{\Theta}\left(\frac{1}{n\kappa L}\right),$$

MRW satisfies one-step overlap with

$$\Delta_s{}^2 \geq \eta = \tilde{\Theta}\left(\frac{1}{n\kappa L}\right)$$

so the mixing time of MRW is

$$\frac{n}{\alpha\Delta_s^2} = \frac{n}{\alpha\eta} = \frac{n}{\alpha}\cdot n\kappa L = n^2\kappa^2$$

$$\tau(\epsilon) = O\left(n^2\kappa^2\log^2\left(\frac{\kappa}{\epsilon^{1/n}}\right)\right) = \tilde{O}\left(n^2\kappa^2\right).$$

Today

1. $P =$ Brownian motion (Gaussian walk)
   $\Rightarrow \tilde{P} =$ Metropolis Random Walk (MRW)
   (Last time)

2. $P =$ Unadjusted Langevin Algorithm (ULA)
   $\Rightarrow \tilde{P} =$ Metropolis-Adjusted Langevin Algorithm (MALA)
   (Today)

Unadjusted Langevin Algorithm (ULA)

To sample from $\nu \propto e^{-f}$ on $\mathbb{R}^n$:

$$x_{k+1} = \underbrace{x_k - \eta \nabla f(x_k)}_{\text{gradient descent}} + \underbrace{\sqrt{2\eta}\, z_k}_{\text{Brownian motion / Gaussian noise}}$$

where $z_k \sim \mathcal{N}(0, I)$ is independent and $\eta > 0$ is step size.

Unadjusted Langevin Algorithm (ULA)

To sample from $\nu \propto e^{-f}$ on $\mathbb{R}^n$:

$$x_{k+1} = x_k - \eta \nabla f(x_k) + \sqrt{2\eta}\, z_k$$

where $z_k \sim \mathcal{N}(0, I)$ is independent and $\eta > 0$ is step size.

Let $P =$ Markov chain for ULA: $P_x = \mathcal{N}(x - \eta \nabla f(x), 2\eta I)$

$$P_x(y) = \frac{1}{(4\pi\eta)^{n/2}} \exp\left(-\frac{\|y - x + \eta \nabla f(x)\|^2}{4\eta}\right)$$

Note: not symmetric: $P_x(y) \neq P_y(x)$

Note: Does *not* converge to ν (even in Gaussian case, see PS1)

Metropolis-Adjusted Langevin Algorithm (MALA)

To sample from $\nu \propto e^{-f}$:

1. From x_k, let

ULA: $$y_k = x_k - \eta \nabla f(x_k) + \sqrt{2\eta}\, z_k$$

   where $z_k \sim \mathcal{N}(0, I)$ is independent, $\eta > 0$ is step size.

2. Set

MH: $$x_{k+1} = \begin{cases} y_k & \text{with prob } a_{x_k}(y_k) = \min\left\{1, \dfrac{\nu(y_k) P_{y_k}(x_k)}{\nu(x_k) P_{x_k}(y_k)}\right\} \\ x_k & \text{with prob } 1 - a_{x_k}(y_k). \end{cases}$$

Mixing time of MALA

Same setup: ν is α-SLC and L-log-smooth, $\kappa = \dfrac{L}{\alpha}$

Warm start $\rho_0 = \mathcal{N}(x^*, \frac{1}{L} I)$ with $M = \kappa^{n/2}$

Assume $\kappa \ll n$ (high-dimensional regime)

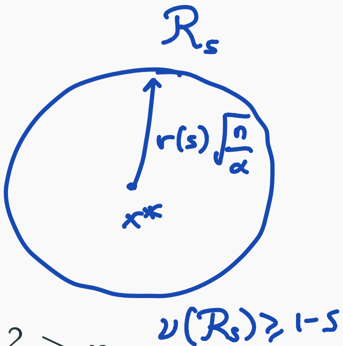
Theorem

With step size $\eta = \Theta\left(\dfrac{1}{nL}\right)$*, MALA has mixing time*

$$\tau(\epsilon) = O\left(n^2 \kappa \log\left(\frac{\kappa}{\epsilon^{1/n}}\right)\right) = \tilde{O}(n^2 \kappa).$$

Show MALA satisfies one-step overlap: (see [DCWY'19, Theorem 1])

Lemma: If $\eta \le c \frac{1}{nL} \min \left\{ 1, \sqrt{\frac{n}{\kappa r(s)^2}} \right\}$, then $\Delta_s^2 \ge \eta$.



- With $s = \frac{\epsilon}{2M}$, $r(s) \sim \sqrt{\frac{1}{n} \log \frac{1}{s}} = \sqrt{\log(\frac{\kappa}{\epsilon^{1/n}})}$
- Assume $\kappa \log(\frac{\kappa}{\epsilon^{1/n}}) \le n$: can take $\eta = \Theta\left(\frac{1}{nL}\right)$ to get ${\Delta_s}^2 \ge \eta$

$\Rightarrow$ mixing time of MALA is

$$\tau(\epsilon) = O\left(\frac{n}{\alpha {\Delta_s}^2} \log\left(\frac{\kappa}{\epsilon^{1/n}} \right) \right) = O\left(n^2 \kappa \log\left(\frac{\kappa}{\epsilon^{1/n}} \right) \right) = \tilde{O}(n^2 \kappa)$$

$\square$

Comparison

To sample from ν on $\mathbb{R}^n$ which is α-SLC and L-log-smooth, $\kappa = \frac{L}{\alpha}$

	Algorithm	Step size	Mixing time
zero-order	**MRW**	$\tilde{\Theta}\left(\frac{1}{n\kappa L}\right)$	$\tilde{O}(n^2\kappa^2)$
first-order (use ∇f)	**MALA**	$\tilde{\Theta}\left(\frac{1}{nL}\right)$	$\tilde{O}(n^2\kappa)$

Comparison: Gaussian

DWIVEDI, CHEN, WAINWRIGHT AND YU

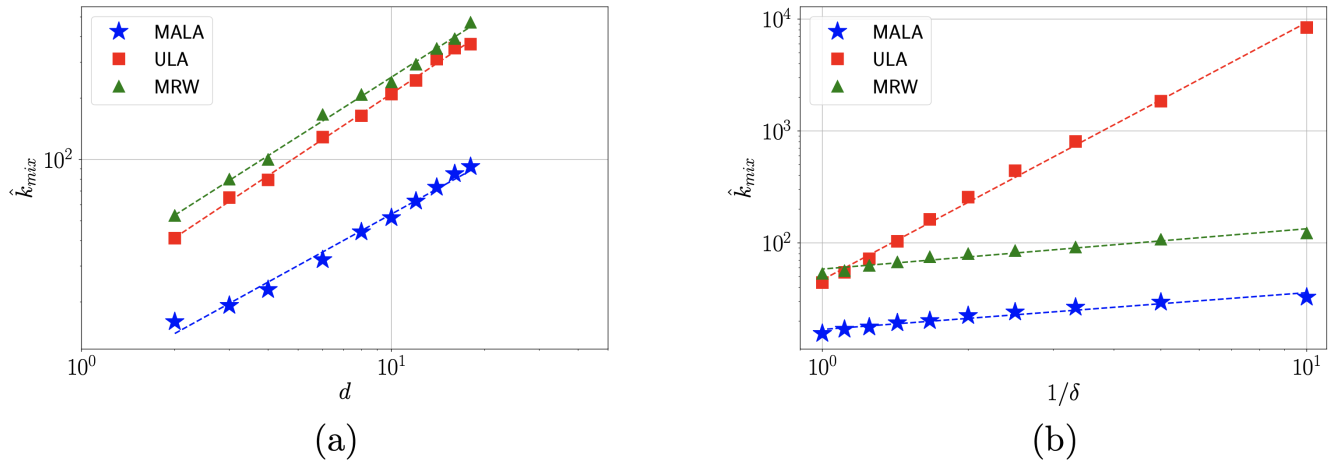


Figure 1. Scaling of the approximate mixing time $\hat{k}_{\mathrm{mix}}$ (refer to the discussion after equation (19) for the definition) on multivariate Gaussian density (19) where the covariance has condition number $\kappa = 4$. (a) Dimension dependency. (b) Error-tolerance dependency.

Comparison: Mixture of Gaussians

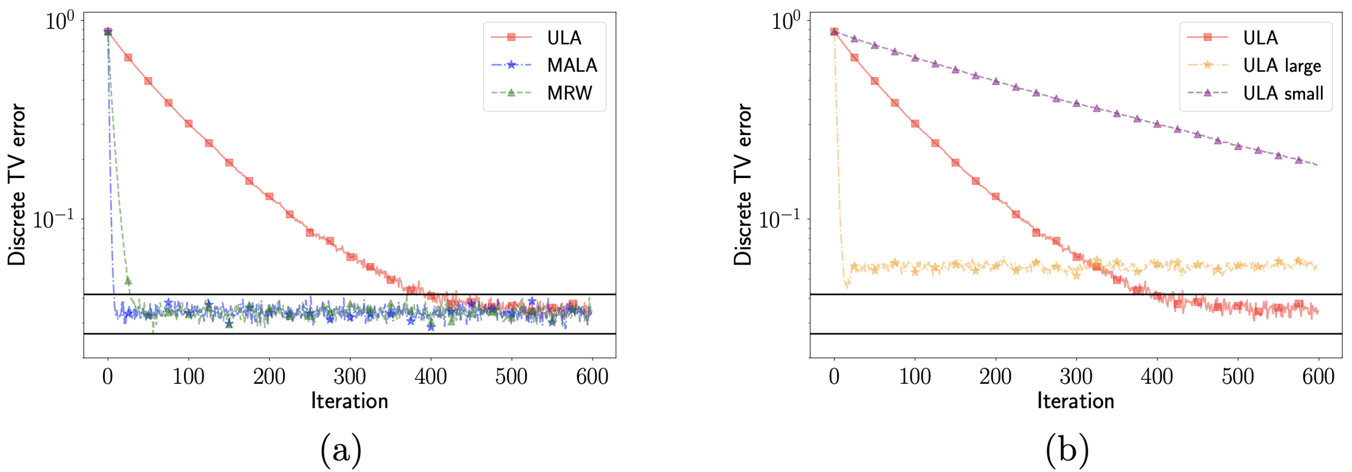


Figure 4. Discrete TV error on a two component Gaussian mixture. (a) Behavior of three different random walks. (b) Behavior of ULA with different choices of step sizes.

Better dimension dependence for MALA

[Chen, Dwivedi, Wainwright, Yu, *Fast mixing of Metropolized Hamiltonian Monte Carlo: Benefits of multi-step gradients*, JMLR, 2020]

- With *conductance profile*, use log-Sobolev instead of Poincaré inequality, reduce dependence $\log M \mapsto \log \log M$
- Mixing time of MALA: $\tilde{O}(n^2\kappa) \mapsto \tilde{O}(n\kappa)$

$M = K^{n/2}$

$\log M \sim n \log K = \tilde{O}(n)$

$\log \log M \sim \log n$

Better dimension dependence for MALA

[Chen, Dwivedi, Wainwright, Yu, *Fast mixing of Metropolized Hamiltonian Monte Carlo: Benefits of multi-step gradients*, JMLR, 2020]

- With *conductance profile*, use log-Sobolev instead of Poincaré inequality, reduce dependence $\log M \mapsto \log\log M$
- Mixing time of MALA: $\tilde{O}(n^2\kappa) \mapsto \tilde{O}(n\kappa)$

[Chewi, Lu, Ahn, Cheng, Gouic, Rigollet, *Optimal dimension dependence of the Metropolis-Adjusted Langevin Algorithm*, arXiv:2012.12810, 2020]

- Use Metropolis-Hastings as TV projection to get better n dependence (but still with $\log M$) $O(n^{1/2})$
- Explicit calculation in Gaussian case: $\eta \sim n^{-1/3} \Rightarrow \tau \sim O(n^{1/3})$
- How to combine them? $\log\log M$

Why MALA?

$$\text{MALA} = \text{ULA} + \text{Metropolis-Hastings}$$

- better than MRW = Brownian Motion + Metropolis-Hastings
- Later also see: MALA as one-step discretization of Hamiltonian Monte Carlo (HMC)

Why ULA?

$$x_{k+1} = \underbrace{x_k - \eta \nabla f(x_k)}_{\text{gradient descent}} + \underbrace{\sqrt{2\eta}\, z_k}_{\text{Gaussian noise}}$$

- Discretization of continuous-time Langevin dynamics

$\eta \to 0$

Why Langevin dynamics?

Brownian motion : $dX_t = \sqrt{2}\, dW_t$

$$\underbrace{dX_t = -\nabla f(X_t)\, dt}_{\text{gradient flow}} + \overbrace{\sqrt{2}\, dW_t}^{\text{Brownian motion}}$$

gradient flow

$$\dot{X}_t = \frac{d}{dt} X_t = -\nabla f(X_t)$$

- The *optimal* dynamics for sampling

Next time: Optimization review

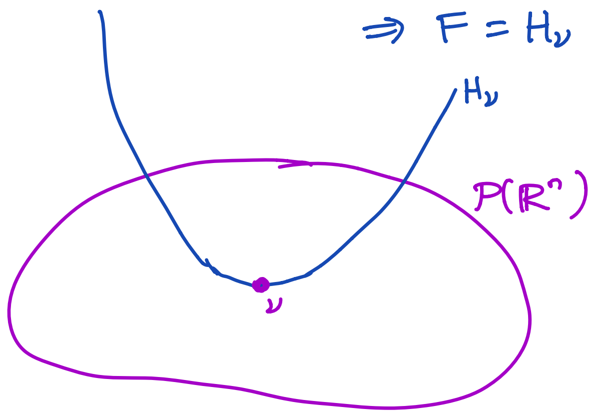
On space of distributions $\mathcal{P}(\mathcal{X}) = \mathcal{P}(\mathbb{R}^n)$

- nice metric: Wasserstein metric $W_2(\rho, \nu)$
- nice functionals: KL divergence $H_\nu(\rho) = F(\rho)$

nice: ν is strongly log-concave

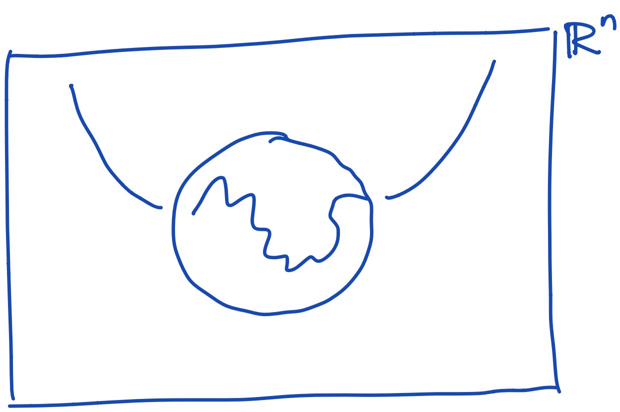
$\Rightarrow F = H_\nu$ is $\underline{\text{geodesically}}$ strongly convex

in the manifold sense



How to get around convexity?

* Some results assume ν is strongly log-concave outside a ball



* Some results use <u>isoperimetry</u>

Nice conditions for $\nu \propto e^{-f}$

1. ν is α- SLC $\quad \left(\nabla^2 f(x) \succeq \alpha I \right)$

   * SLC is enough to get exponential contraction in W_2 distance along Langevin dynamics

$\Downarrow$

2. ν satisfies <u>Log-Sobolev inequality</u>

$$\forall \rho : \quad \boxed{J_\nu(\rho) \geq \frac{\alpha}{2} H_\nu(\rho)}$$

Relative Fisher information

Relative entropy

* LSI is enough to get exp. convergence rate in H_ν along Langevin dynamics

* This is preserved under bounded perturbation

  if $\nu = e^{-f}$ is α-LSI

  then $\tilde{\nu} = e^{-(f+g)}$ is $\tilde{\alpha}$-LSI

  where $\boxed{\tilde{\alpha} = \alpha \cdot e^{-\operatorname{osc}(g)}}$

  where $\operatorname{osc}(g) = \sup_x g(x) - \inf_y g(y)$