CPSC 661: Sampling Algorithms in ML

Andre Wibisono

February 17, 2021

Yale University

Last time

- Reversible Markov chain $\nu(x) \cdot P_x(y) = \nu(y) \cdot P_y(x)$
- Spectral gap $\Leftrightarrow$ Conductance $\frac{\phi^2}{2} \leq \gamma \leq 2\phi$
  $\gamma = \lambda_2(L)$ ϕ
- s-Conductance $\Rightarrow$ Mixing time in TV distance
  ϕ_s, $0 \leq s \leq \frac{1}{2}$ $\tau(\varepsilon) = \tilde{O}\left(\frac{1}{\phi_s^2}\right)$, $s = \frac{\varepsilon}{2M}$
  $M = M_\nu^\infty(Q) = \sup_x \left|\frac{Q(x)}{\nu(x)} - 1\right|$

Questions:

1. How to construct reversible Markov chain P?
   $\Rightarrow$ **Today:** Metropolis-Hastings
2. How to bound s-conductance? (next time)

References

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

- Billera and Diaconis, *A Geometric Interpretation of the Metropolis-Hastings Algorithm*, Statistical Science, 2001

Metropolis Algorithm

THE JOURNAL OF CHEMICAL PHYSICS VOLUME 21, NUMBER 6 JUNE, 1953

Equation of State Calculations by Fast Computing Machines

NICHOLAS METROPOLIS, ARIANNA W. ROSENBLUTH, MARSHALL N. ROSENBLUTH, AND AUGUSTA H. TELLER,
Los Alamos Scientific Laboratory, Los Alamos, New Mexico

AND

EDWARD TELLER,* *Department of Physics, University of Chicago, Chicago, Illinois*

(Received March 6, 1953)

THOSE WE'VE LOST

Arianna Rosenbluth Dies at 93; Pioneering Figure in Data Science

Dr. Rosenbluth, who received her physics Ph.D. at 21, helped create an algorithm that has become a foundation of understanding huge quantities of data. She died of complications of the coronavirus.

By Katie Hafner

Published Feb. 9, 2021 Updated Feb. 15, 2021

The Metropolis algorithm, a technique for generating random samplings, started out as a way to understand a fundamental problem: how atoms rearrange themselves as solids melt.

Over the decades, the Metropolis algorithm and its subsequent variations have been put to a vast number of uses and now serve as an underpinning to understanding critical challenges of our age, including making sense of huge volumes of data, predicting election outcomes and understanding Covid-19's spread.

Dr. Arianna Wright Rosenbluth in 2013. She helped create what has become one of the most important algorithms of all time. via Rosenbluth family

Metropolis Algorithm

To sample from ν on $\mathcal{X}$:

- Start with a symmetric Markov chain $P = (P_x : x \in \mathcal{X})$

  $P_x(y) = P_y(x)$ ← stationary dist for P is uniform

- Do random walk with accept/reject step:

  1. From x, draw $y \sim P_x$ proposal
  2. Accept y with probability $\min\left\{1, \frac{\nu(y)}{\nu(x)}\right\}$

     else reject y

Metropolis filter

new point $x' = \begin{cases} y \sim P_x & \text{w.p. } \alpha_x(y) = \min\left\{1, \frac{\nu(y)}{\nu(x)}\right\} \\ x & \text{w.p. } 1 - \alpha_x(y) \end{cases}$

notes:

1) intritive: y is "betfee" ($\nu(y) \geq \nu(x)$)

$\Rightarrow$ always move

else stay with same probability

2) Zeeo-ordee algorithm: only depends on $\nu(x)$ (not gradient)

& only up to normalizing constant $\left(\frac{\nu(y)}{\nu(x)}\right)$

good for: Bayesian computation

$$p(\theta \mid \text{data}) \propto p(\theta) \cdot p(\text{data} \mid \theta)$$

can't compute $p(\text{data}) = \int_{\Theta} p(\theta) \cdot p(\text{data} \mid \theta)\, d\theta$

don't need

large/exp-size

also in physics, ML, ...

Gibbs distribution $\nu(x) \propto e^{-f(x)}$

from potential function / energy

$f: \mathcal{X} \to \mathbb{R}$

$\begin{pmatrix} q_i \\ p_i \end{pmatrix} \in \mathbb{R}^2 \times \mathbb{R}^2 = \mathbb{R}^4$

eg. • physics: f = Hamiltonian of N particles on $\mathbb{R}^2$

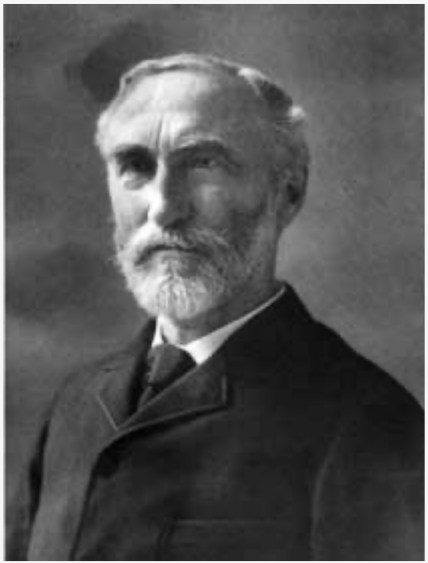
$\mathcal{X} = \mathbb{R}^{4N}$

• convex geometry / analysis

f convex $\longleftrightarrow$ $\nu \propto e^{-f}$ log-concave

Gibbs

- Josiah Willard Gibbs (1839–1903)
- New Haven, CT
- Yale College:
  - ⋆ BS 1858,
  - ⋆ PhD 1863
  - ⋆ Professor of Mathematical Physics 1871
- Thermodynamics, Gibbs free energy, statistical mechanics



Josiah Willard Gibbs

"the greatest mind in American history" (Einstein)


JOSIAH WILLARD GIBBS
BORN FEB. 11, 1839—DIED APR. 28, 1903


JOSIAH WILLARD GIBBS

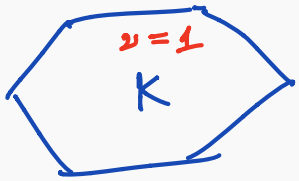
Example: Metropolis Random Walk (MRW)

Start with P = Brownian motion / Gaussian noise with step size η

From x, draw $y = x + \sqrt{2\eta}\, Z$, $Z \sim \mathcal{N}(0, I)$ on $\mathbb{R}^n$

Accept y w.p. $\min\left\{1, \frac{\nu(y)}{\nu(x)}\right\}$, else reject

- eg. ν = uniform on $K \subset \mathbb{R}^n$



$\nu(x) \propto e^{-f(x)}$

$-\log \nu = f = \mathbb{1}_K : \mathbb{R}^n \to \mathbb{R}$

$$\mathbb{1}_K(x) = \begin{cases} 0 & \text{if } x \in K \\ \infty & \text{else} \end{cases}$$

convex if K convex set
function

- eg. ν = Gaussian $\Leftrightarrow$ f quadratic

Generalization: Metropolis-Hastings (MH) Algorithm

To sample from ν:

- Start with any Markov chain P
- Do random walk with accept/reject:
  1. From x, draw $y \sim P_x$
  2. Accept y w.p. $\min\left\{1, \frac{\nu(y)}{\nu(x)}\frac{P_y(x)}{P_x(y)}\right\} = \alpha_x(y)$

     else reject & stay at x

(Metropolis-Hastings alg./filter)

Notes:

* Hastings (1970)
* Zero-order, only needs ν up to constant
* Universal algorithm: can start from any P ← has WRONG stationary dist μ

MH: force $\mu \to \nu$

Why does MH work?

Let P be any Markov chain
Let $\hat{P}$ be MH version wrt. ν

Lemma: $\hat{P}$ is reversible wrt. ν ($\Rightarrow$ ν stationary for $\hat{P}$)

Proof:

$$\hat{P}_x(y) = P_x(y) \cdot \alpha_x(y) + \delta_x(y) \cdot A(x)$$

where $A(x) = P_x(\{x\}) + \int_{\mathcal{X} \setminus \{x\}} (1 - \alpha_x(y)) \cdot P_x(y)\, dy$

$$= 1 - \int_{\mathcal{X} \setminus \{x\}} \alpha_x(y) P_x(y)\, dy$$

WTS: $\nu(x) \cdot \hat{P}_x(y) = \nu(y) \cdot \hat{P}_y(x) \quad \forall\, x, y \in \mathcal{X}$

- if $x = y$, true

- If $x \neq y$: $\nu(x) \cdot \tilde{P}_x(y) = \nu(x) \cdot P_x(y) \cdot \alpha_x(y)$

$$= \nu(x) \cdot P_x(y) \cdot \min\left\{1, \frac{\nu(y)}{\nu(x)} \frac{P_y(x)}{P_x(y)}\right\}$$

$$= \min\{\nu(x) \cdot P_x(y), \nu(y) \cdot P_y(x)\}$$

↗ symmetric in x, y

$$= \nu(y) \cdot \hat{P}_y(x).$$

□

More general family

Given Markov chain $P = (P_x : x \in X)$

Want to make $P \longmapsto \hat{P}$ reversible wrt. ν

Accept / reject step:

1. From x, draw $y \sim P_x$
2. Accept y with some probability $\alpha_x(y)$

} new MC $\hat{P}$

Let $R_x(y) = \dfrac{\nu(y)}{\nu(x)} \cdot \dfrac{P_y(x)}{P_x(y)} = \dfrac{1}{R_y(x)}$

Lemma: Choose any $0 \le \alpha_x(y) \le \min\{1, R_x(y)\}$

[Billera & Diaconis Prop. 3.1] set $\alpha_y(x) = \dfrac{\alpha_x(y)}{R_x(y)}$

Then $\hat{P}$ is reversible wrt. ν

Ex: 1) Metropolis (1950): $\alpha_x(y) = \min\{1, R_x(y)\}$

2) Barker (1965):

$$\alpha_x(y) = \frac{\nu(y) \cdot P_y(x)}{\nu(x) \cdot P_x(y) + \nu(y) \cdot P_y(x)} = \frac{R_x(y)}{1 + R_x(y)}$$

eg. P is rev wrt. $\nu \iff R = 1$

$\Rightarrow$ Metropolis: accept w.p. 1

$\Rightarrow$ Barker: accept w.p. $\frac{1}{2}$

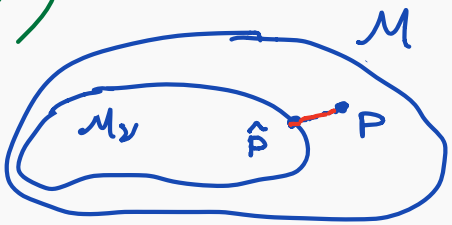
Optimization interpretation of MH

[Billera & Diaconis 2001]

Let $\mathcal{M} = \{$ Markov chains on $\mathcal{X}\}$ ← space of prob. distributions on $\mathcal{X}$

$= \{ P = (P_x : x \in \mathcal{X}), \quad P_x \in \mathcal{P}(\mathcal{X}) \}$

$\left(= \mathcal{P}(\mathcal{X})^{\mathcal{X}} \equiv \{ P : \mathcal{X} \to \mathcal{P}(\mathcal{X}) \} \right)$



Given $\nu \in \mathcal{P}(\mathcal{X})$

Let $\mathcal{M}_\nu = \{ P \in \mathcal{M} : P$ reversible wrt. $\nu \}$

then MH filter: $\mathcal{M} \ni P \longmapsto \hat{P} \in \mathcal{M}_\nu$

Define metric on $\mathcal{M}$:

$$d(P, P') = \int_{\mathcal{X}\times\mathcal{X}\setminus\Delta} |P_x(y) - P'_x(y)|\, \nu(x)\, dx\, dy$$

$\Delta = \{(x,x) : x \in \mathcal{X}\}$

Theorem: [Billera & Diaconis Thm 1]

MH filter is projection in d distance

$$\hat{P} = \underset{\tilde{P}\in\mathcal{M}_\nu}{\arg\min}\ d(\tilde{P}, P)$$

$\hat{P}$: after MH; P: starting MC

(and $\hat{P}$ is unique such MC with $\hat{P}_x(y) \le P_x(y)$)

- Why metric d?

Given P, given ν, define:

$$(\nu P)(x,y) = \nu(x)\cdot P_x(y)$$
$$(\overline{\nu P})(x,y) = \nu(y)\cdot P_y(x)$$
$$\Big\} \in \mathcal{P}(\mathcal{X}\times\mathcal{X})$$

P Rev. wrt. $\nu \iff \nu P = \overline{\nu P}$

then $\boxed{d(P, \mathcal{M}_\nu) = TV(\nu P, \overline{\nu P})}$

Example

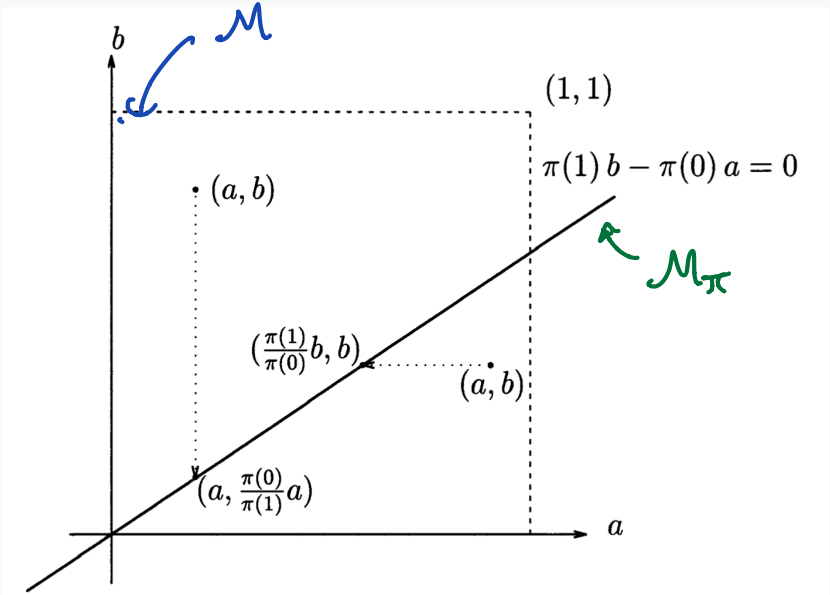
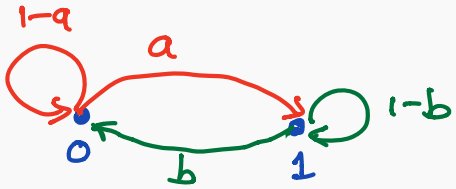


FIG. 1. *Metropolis map in the* 2×2 *case.*

2-point space

$\mathcal{X} = \{0, 1\}$



$$P = \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix}$$

Stationary: $\left(\frac{b}{a+b}, \frac{a}{a+b} \right)$

defines

Given target $\pi = (\pi(0), \pi(1))$

P stationarity wrt. $\pi \iff \pi(0) \cdot a = \pi(1) \cdot b$

Extensions

Projects:

Q: what if projection in other distances? KL divergence, χ^2-div, φ-div

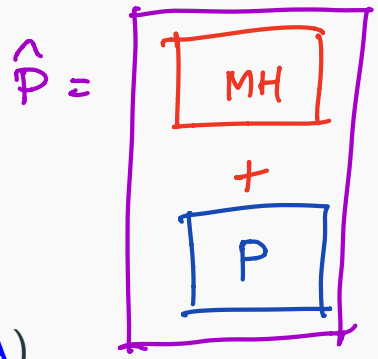
- Diaconis and Miclo, *On characterizations of Metropolis-type algorithms in continuous time*, Latin American Journal of Probability and Mathematical Statistics, 2009

- Choi, *Metropolis-Hastings reversiblizations of non-reversible Markov chains*, Stochastic Processes and their Applications, 2020

  min → max

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

Basic algorithms for sampling



1. Metropolis Random Walk (MRW)
2. Unadjusted Langevin Algorithm (ULA)
3. Metropolis-Adjusted Langevin Algorithm (MALA)

(also underdamped Langevin algorithm
Hamiltonian Monte Carlo
...)

1. Metropolis Random Walk (MRW)

P = Brownian motion

$z_k \sim N(0, I)$

$$X_{k+1} = \begin{cases} y_k = x_k + \sqrt{2\eta}\, z_k & \text{w.p. } \min\left\{1, \dfrac{\nu(y_k)}{\nu(x_a)}\right\} \\ x_k & \text{else} \end{cases}$$

2. Unadjusted Langevin Algorithm (ULA)

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

$$X_{k+1} = X_k - \eta \nabla f(X_k) + \sqrt{2\eta}\, Z_k, \quad Z_k \sim \mathcal{N}(0, I)$$

- why? From continuous time (later)
- has asymptotic bias

  (stationary $\neq \nu$)

3. Metropolis-Adjusted Langevin Algorithm (MALA)

Choose P = ULA

+ MH filter

= MALA +

Comparison: Gaussian

$\nu = \mathcal{N}(0, \Sigma)$

Dwivedi, Chen, Wainwright and Yu

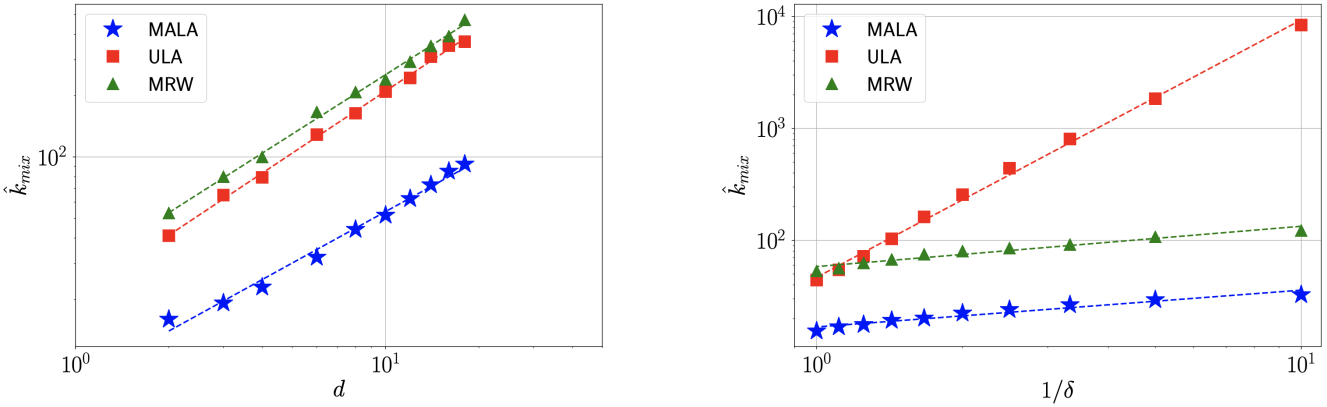
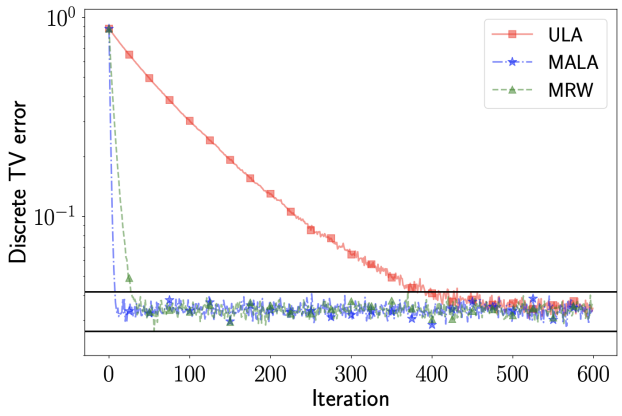


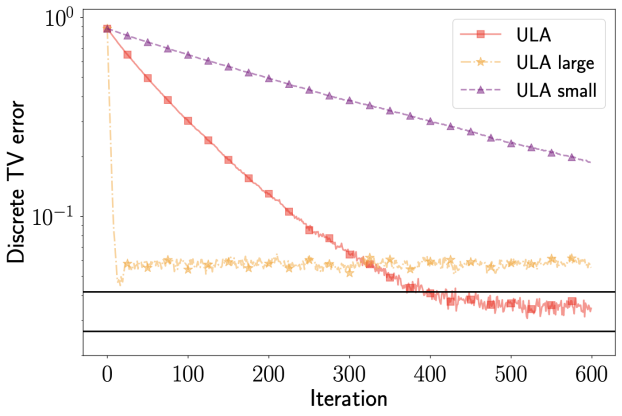
Figure 1. Scaling of the approximate mixing time $\hat{k}_{\text{mix}}$ (refer to the discussion after equation (19) for the definition) on multivariate Gaussian density (19) where the covariance has condition number $\kappa = 4$. (a) Dimension dependency. (b) Error-tolerance dependency.

$\kappa = \dfrac{\lambda_{max}(\Sigma)}{\lambda_{min}(\Sigma)}$

Comparison: Mixture of Gaussians



(a)



(b)

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.

Comparison: Mixing time

Random walk	**Distribution** $\mu_\star$	$t_{\mathbf{mix}}(\delta;\mu_\star)$
ULA (Cheng and Bartlett, 2018)	$\mathcal{N}(x^\star, m^{-1}\mathbb{I}_d)$	$\mathcal{O}\left(\frac{d\kappa^2\log(d\kappa/\delta)}{\delta^2}\right)$
ULA (Dalalyan, 2016)	$\mathcal{N}(x^\star, L^{-1}\mathbb{I}_d)$	$\mathcal{O}\left(\frac{(d^3+d\log^2(1/\delta))\kappa^2}{\delta^2}\right)$
MRW (this work)	$\mathcal{N}(x^\star, L^{-1}\mathbb{I}_d)$	$\mathcal{O}\left(d^2\kappa^2\log^{1.5}\left(\frac{\kappa}{\delta}\right)\right)$
MALA (this work)	$\mathcal{N}(x^\star, L^{-1}\mathbb{I}_d)$	$\mathcal{O}\left(d^2\kappa\log\left(\frac{\kappa}{\delta}\right)\right)$

Table 2. Scalings of upper bounds on δ-mixing time, from the starting distribution $\mu_\star$ given in column two, for different random walks in $\mathbb{R}^d$ with target $\pi \propto e^{-f}$ such that $m\mathbb{I}_d \preceq \nabla^2 f(x) \preceq L\mathbb{I}_d$ for any $x \in \mathbb{R}^d$ and $\kappa := L/m$. Here $x^\star$ denotes the unique mode of the target density π.

Recap

Metropolis-Hastings: Universal filter to construct reversible MC

Next time: How to bound s-conductance?

Please fill out survey: `https://forms.gle/mCTWsWPdsdoTEkvR9`