CPSC 661: Sampling Algorithms in ML

Andre Wibisono

February 24, 2021

Yale University

Last time

- Reversible Markov chain $\nu(x) \cdot P_x(y) = \nu(y) \cdot P_y(x)$
- Spectral gap $\Leftrightarrow$ Conductance
  $\gamma = \lambda_2(L)$ $\quad$ ϕ
  Cheeger: $\frac{\phi^2}{2} \le \gamma \le 2\phi$
- s-Conductance $\Rightarrow$ Mixing time in TV distance
  ϕ_s $\quad$ $\tau(\varepsilon) = \tilde{O}\left(\frac{1}{\phi_s^2}\right)$, $\quad s = \frac{\varepsilon}{2M}$
  $M = M_\nu^\infty(g_0)$ warmness of g_0

Last time

Questions:

1. How to construct reversible Markov chain?
   ⇒ Metropolis-Hastings : accept/reject step

$$\boxed{P} + \boxed{\mathrm{MH}_\nu} = \boxed{\tilde{P}} \quad \text{reversible wrt } \nu$$

from x, draw $y \sim P_x$

set $x' = \begin{cases} y & \text{w.p. } \min\left\{1, \dfrac{\nu(y)\cdot P_y(x)}{\nu(x)\cdot P_x(y)}\right\} = \alpha_x(y) \\ x & \text{w.p. } 1-\alpha_x(y) \end{cases}$

e.g. • P = Brownian motion $\Rightarrow$ $\tilde{P}$ = Metropolis Random Walk (MRW)

• P = Unadjusted Langevin Alg. (ULA) $\Rightarrow$ $\tilde{P}$ = Metropolis-Adjusted Langevin Alg. (MALA)

2. How to bound conductance?
   ⇒ **Today:** Isoperimetry

References

- Vempala, *Geometric Random Walk: A Survey*, Combinatorial and Computational Geometry, 2005

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

From conductance to isoperimetry

Let P be Markov chain reversible wrt ν

Suppose P has large conductance

$$\phi = \inf_{A \subset X} \frac{\Phi(A)}{\min\{\nu(A), 1-\nu(A)\}}$$

where $\Phi(A)$ is ergodic flow

$$= \Pr(X_0 \in A, X_1 \notin A), \quad \begin{matrix} X_0 \sim \nu \\ X_1 | X_0 \sim P_{X_0} \end{matrix}$$

What does this mean for ν?

* In general, nothing: there are always M.C. with large conductance

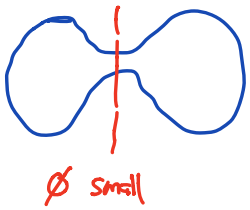
  eg. the ideal MC: $P_x = \nu$

  (random walk: from x, jump to $y \sim \nu$ independent)

  converges in one step (but not implementable)

  Exercise: $\phi = \frac{1}{2}$ (for any ν)

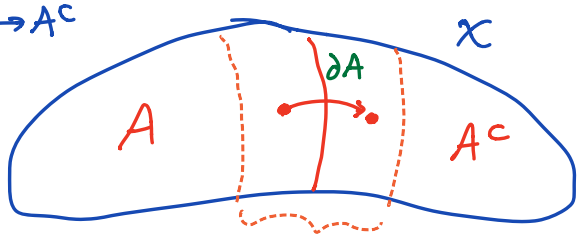
But there are bad targets, eg. ν = uniform on



* Suppose P is local : x and $y \sim P_x$ are close in some distance d

$$d(x, y) \ll 1 \quad \text{(w.h.p.)}$$

then points that cross over $A \to A^c$ are near boundary ∂A



* Large conductance

$\Rightarrow$ region near boundary has large volume

(large sets have large boundaries)

isoperimetry

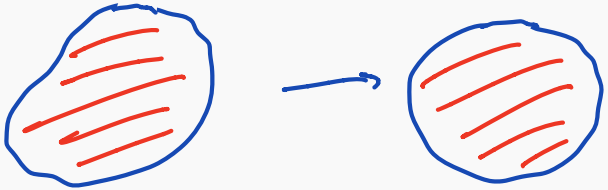
* Conversely, will show isoperimetry + one-step overlap

$\Rightarrow$ large conductance

Isoperimetry

Problem: Among all regions of equal perimeter / area, which has largest volume?

Dido [c. 800 BC]
Queen of Carthage

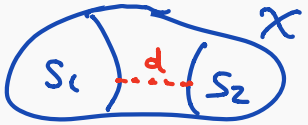


Formally:

Let X have a distance metric $d(x,y)$

(eg. $X = \mathbb{R}^n$, $d(x,y) = \|x-y\|_2$ Euclidean distance)

Given 2 sets $S_1, S_2 \subset X$

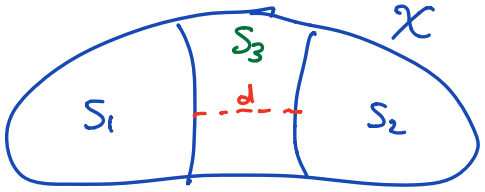


define distance $d(S_1, S_2) = \inf \{ d(x,y) : x \in S_1, y \in S_2 \}$

Definition:

Given a prob dist ν on $\mathcal{X}$, we say ν is isoperimetric (or satisfies isoperimetric inequality) with constant $\psi_\nu > 0$ if $\forall$ partition $\mathcal{X} = S_1 \cup S_2 \cup S_3$:

$$\nu(S_3) \geq \psi_\nu \cdot d(S_1, S_2) \cdot \min\{\nu(S_1), \nu(S_2)\}$$



Notes:

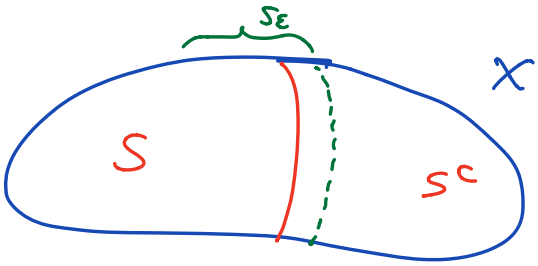
1. Equivalent to (differential formulation) :

$\forall\ S \subset \mathcal{X}$,

$$\nu(\partial S) \geq \psi \cdot \min\{\nu(S), \nu(S^c)\}$$

where $\nu(\partial S) = \lim_{\varepsilon \to 0} \frac{\nu(S_\varepsilon) - \nu(S)}{\varepsilon}$

$$S_\varepsilon = \{x \in \mathcal{X} : d(x, S) \leq \varepsilon\}$$



2. Equivalent to Poincaré inequality (functional form):

$\forall\, g : \mathcal{X} \to \mathbb{R}$

$$\mathbb{E}_\nu\left[\|\nabla g\|^2\right] \geq \frac{\Psi^2}{4}\,\mathrm{Var}_\nu(g)$$

(this is the spectral gap of Dirichlet form of ν)

Equivalently,

$$\Psi \equiv \Psi_\nu = \inf_{S \subset \mathcal{X}} \frac{\nu(\partial S)}{\min\{\nu(S), \nu(S^c)\}}$$

Isoperimetric Theorems

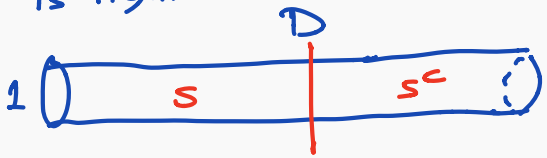
1. Theorem [Dyer & Frieze '91]:

   ν = uniform on convex $K \subset \mathbb{R}^n$ with diameter D.

   $\Rightarrow \quad \psi \geq \frac{2}{D}$

note: dependence on D is tight

e.g., K = cylinder



$\Rightarrow \psi \sim \frac{1}{D}$

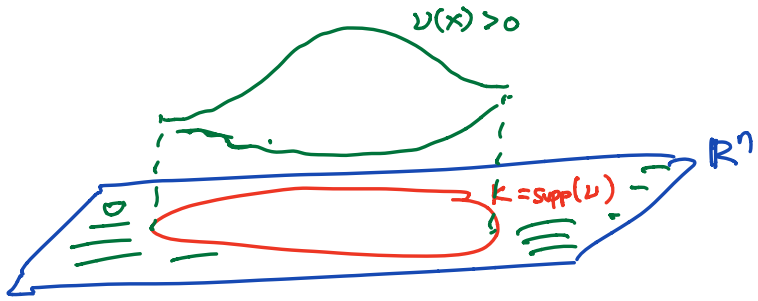
2. Can be generalized to:

   ν = log-concave distribution on $\mathbb{R}^n$ with support of diameter D

$\nu(x) \propto e^{-f(x)}$, $\quad f: \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$

$\text{support}(\nu) = \{x \in \mathbb{R}^n : \nu(x) > 0\}$

$= \{x \in \mathbb{R}^n : f(x) < \infty\} = \text{dom}(f)$



$\Rightarrow \psi \gtrsim \frac{2}{D}$

3. **Theorem** [KLS '95]

ν = log-concave distribution on $\mathbb{R}^n$

$\Rightarrow \psi \gtrsim \frac{\log 2}{M_1(\nu)} \gtrsim \frac{1}{4 M_1(\nu)} \gtrsim \frac{1}{4\sqrt{\text{Var}_\nu(X)}}$ $\quad \log 2 = 0.3 > \frac{1}{4}$

where $\mu = \mathbb{E}_\nu[X]$

$M_1(\nu) = \mathbb{E}_\nu[\|X - \mu\|_2]$

(note: $M_1(\nu) \le \mathbb{E}_\nu[\|X-\mu\|_2^2]^{1/2} = \sqrt{\text{Var}_\nu(X)}$)

eg. for ν isotropic ($\mu = \mathbb{E}_\nu[X] = 0$

$\Sigma = \text{Cov}_\nu(X) = \mathbb{E}_\nu[(X-\mu)(X-\mu)^T] = I$)

$\Rightarrow \text{Var}_\nu(X) = n \Rightarrow$ for isotropic log-concave ν, $\psi \gtrsim \Omega\left(\frac{1}{\sqrt{n}}\right)$.

KLS Conjecture

Let ν be log-concave on $\mathbb{R}^n$

Let $\Sigma = \mathrm{Cov}_\nu(X) \in \mathbb{R}^{n \times n}$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n > 0$.

	$\Sigma = I$
Theorem [KLS '95]: $\psi = \Omega\left(\frac{1}{\sqrt{\sum_{i=1}^n \lambda_i}}\right) = \Omega\left(\frac{1}{\sqrt{\mathrm{Tr}(\Sigma)}}\right)$	$\frac{1}{\sqrt{n}}$
Conjecture [KLS '95]: $\psi \overset{?}{=} \Omega\left(\frac{1}{\sqrt{\lambda_1}}\right) = \Omega\left(\frac{1}{\|\Sigma\|_{op}^{1/2}}\right)$	1 ?
Theorem [LV '17]: $\psi = \Omega\left(\frac{1}{(\sum_{i=1}^n \lambda_i^2)^{1/4}}\right) = \Omega\left(\frac{1}{\|\Sigma\|_{HS}^{1/2}}\right)$	$\frac{1}{n^{1/4}}$
Theorem [C '20]: $\psi = \Omega\left(\frac{1}{n^{o(1)} \cdot \sqrt{\lambda_1}}\right)$	$\frac{1}{n^{o(1)}}$ (little-o)

References for KLS Conjecture

- Kannan, Lovász, & Simonovits, *Isoperimetric problems for convex bodies and a localization lemma*, Discrete & Computational Geometry, 1995

- Lee & Vempala, *The Kannan-Lovász-Simonovits Conjecture*, arXiv:1807.03465, 2018

- Lee & Vempala, *Eldan's Stochastic Localization and the KLS Conjecture: Isoperimetry, Concentration and Mixing*, arXiv:1612.01507v3, 2019

- Chen, *An Almost Constant Lower Bound of the Isoperimetric Coefficient in the KLS Conjecture*, arXiv:2011.13661, 2020

Strongly Log-Concave Distribution

Def: ν is α-strongly log-concave (α-SLC) for some $\alpha > 0$

if $\nu(x) \propto e^{-f(x)}$, $f: \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$

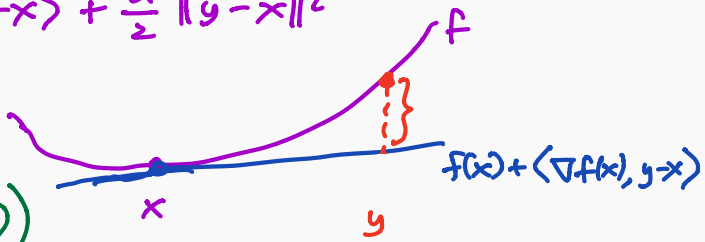
such that f is α-strongly convex

(recall: this means: $\frac{f(x)+f(y)}{2} - f\left(\frac{x+y}{2}\right) \geq \frac{\alpha}{8}\|y-x\|^2$

- if f is differentiable:

$\Leftrightarrow$ $f(y) \geq f(x) + \langle \nabla f(x), y-x \rangle + \frac{\alpha}{2}\|y-x\|^2$

$\nabla f(x)$ = gradient: $\left(\frac{\partial f}{\partial x_1}(x), \dots, \frac{\partial f}{\partial x_n}(x)\right)$



- if f is twice-differentiable:

$\Leftrightarrow$ $\nabla^2 f(x) \geq \alpha I$

$\nabla^2 f(x)$ = Hessian matrix $\left(\frac{\partial^2 f}{\partial x_i \partial x_j}(x)\right)_{i,j=1}^n$

- log-concave is when $\alpha = 0$
- eg. ν = Gaussian $\mathcal{N}(0, \Sigma)$, Σ has eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n > 0$

then ν is α-SLC where $\alpha = \frac{1}{\lambda_1}$

because

$$\nu(x) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \; e^{-\frac{x^T \Sigma^{-1} x}{2}}$$

so $f(x) = -\log \nu(x) = \frac{1}{2} x^T \Sigma^{-1} x + \frac{1}{2} \log \det(2\pi\Sigma)$

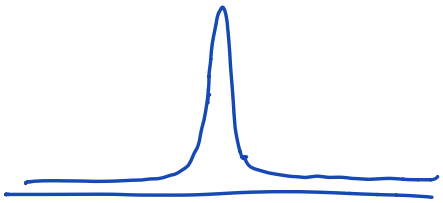
$$\nabla f(x) = \Sigma^{-1} x$$

$$\nabla^2 f(x) = \Sigma^{-1} \succeq \frac{1}{\lambda_1} I$$

↑ has eigenvalues $\frac{1}{\lambda_1} \leq \frac{1}{\lambda_2} \leq \dots \leq \frac{1}{\lambda_n}$

eg. $\Sigma = \lambda I$

- if λ is small:



ν very peaked

$\Rightarrow$ very SLC ($\frac{1}{\lambda}$ large)

- if λ is large:

ν very flat

$\Rightarrow$ not very SLC ($\frac{1}{\lambda}$ small)

Theorem: If ν is α-SLC on $\mathbb{R}^n$

then $\psi \gtrsim (\log 2)\sqrt{\alpha} \gtrsim \frac{\sqrt{\alpha}}{4}$

- [Cousins–Vempala '16]
- Nice: dimension-free (doesn't depend on n)
- (Naively, using earlier bound

$$\psi \gtrsim \frac{\log 2}{M_1(\nu)} \gtrsim \frac{\log 2}{\sqrt{\mathrm{Var}_\nu(x)}} \gtrsim \log 2 \cdot \sqrt{\frac{\alpha}{n}}$$

↑ bad)