CPSC 661: Sampling Algorithms in ML

Andre Wibisono

February 10, 2021

Last time

- Reversible Markov chain P wrt ν

$$\nu(x) \cdot P(x,y) = \nu(y) \cdot P(y,x)$$

- Laplacian $L = I - P \geq 0$

$$L\mathbb{1} = 0 \cdot \mathbb{1} \longleftarrow \lambda_1(L) = 0$$

- Spectral gap $\gamma = \lambda_2(L) = \inf_{f \in L^2(\nu)} \frac{\langle f, Lf \rangle_\nu}{\mathrm{Var}_\nu(f)}$

- Mixing time in χ^2-divergence $\tilde{O}\left(\frac{1}{\gamma}\right)$

Today: Conductance, Cheeger's inequality

References

- Levin, Peres, and Wilmer, *Markov Chains and Mixing Times*, AMS, 2009
- Lovász and Simonovits, *Random Walks in a Convex Body and an Improved Volume Algorithm*, Random Structures and Algorithms, 1993
- Vempala, *Geometric Random Walk: A Survey*, Combinatorial and Computational Geometry, 2005

From last time: Typo

2) What about P general? (P not sym $\Rightarrow$ complex eigenvalues)

Assume reversibility:

$$\tilde{P}(x,y) = \sqrt{\nu(x)} \cdot \frac{P(x,y)}{\sqrt{\nu(y)}} = \tilde{P}(y,x)$$

$\tilde{P}$: symmetric

$\Rightarrow$ spectral theorem

(eg. $\tilde{P} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$

$\tilde{L} = I - \tilde{P} = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$ normalized Laplacian)

should be:

$\tilde{P} = D^{1/2} P D^{-1/2}$ $(P = D^{-1}A)$

$= D^{-1/2} A D^{-1/2}$

this is equivalent to working $L^2(\nu)$

From last time: Mixing time in χ^2-divergence

Theorem
Let P be reversible with respect to ν with spectral gap γ. Along the Markov chain $X_k \sim \rho_k$:

$$\chi^2_\nu(\rho_k) \leq (1-\gamma)^{2k}\, \chi^2_\nu(\rho_0) \leq e^{-2\gamma k}\, \chi^2_\nu(\rho_0)$$

Corollary
The mixing time in χ^2-divergence is:

(assuming n, ε constant)
hides log factor

$$\tau(\epsilon) = \frac{1}{2\gamma} \log \frac{\chi^2_\nu(\rho_0)}{\epsilon} = \tilde{O}\left(\frac{1}{\gamma}\right) \leftarrow \tilde{O}\left(\frac{n}{\gamma}\right)$$

$$\log \tfrac{1}{\varepsilon} + \log \chi^2_\nu(\rho_0)$$

$\log \chi^2_\nu(\rho_0) = O(n)$ in $\mathcal{X} = \mathbb{R}^n$

Proof

Let $h_k = \frac{d\mu_k}{d\nu}$ be density of μ_k wrt ν

P reversible $\Rightarrow$ $h_{k+1} = P h_k$ (exercise)

$\Rightarrow$ $\boxed{h_k = P^k h_0}$

P self-adjoint $\Rightarrow$ has real eigenvalues

$$1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_n > -1$$

(irreducible)

with orthonormal basis of eigenfunctions

$$\mathbb{1} = \phi_1, \phi_2, \ldots, \phi_n$$

Spectral gap $\gamma \Leftrightarrow |\lambda_i| \leq 1-\gamma$ for $i = 2, \ldots, n$

Write $h_0 = \frac{d\beta_0}{d\nu}$ as

$$h_0 = \mathbb{1} + c_2 \phi_2 + \dots + c_n \phi_n$$

where $c_i = \langle h_0, \phi_i \rangle_\nu$ $\left(c_1 = \langle h_0, \phi_1 \rangle_\nu = \langle h_0, \mathbb{1} \rangle_\nu = 1 \right)$

Apply P: $\quad P\phi_i = \lambda_i \phi_i$

$$h_1 = P h_0 = P\mathbb{1} + c_2 P\phi_2 + \dots + c_n P\phi_n$$
$$= \mathbb{1} + c_2 \lambda_2 \phi_2 + \dots + c_n \lambda_n \phi_n$$

Repeat:

$$h_k = P^k h_0 = \mathbb{1} + c_2 \lambda_2^k \phi_2 + \dots + c_n \lambda_n^k \phi_n$$

Then:

$$\chi^2_\nu(\beta_k) = \mathrm{Var}_\nu(h_k) = \| h_k - \mathbb{1} \|^2_\nu$$
$$= \| c_2 \lambda_2^k \phi_2 + \dots + c_n \lambda_n^k \phi_n \|^2_\nu$$
$$= c_2^2 \lambda_2^{2k} + \dots + c_n^2 \lambda_n^{2k} \quad \text{since } \langle \phi_i, \phi_j \rangle_\nu = \begin{cases} 1 & i=j \\ 0 & \text{else} \end{cases}$$

since $|\lambda_i| \le 1-\gamma$:

$$\le (1-\gamma)^{2k} \left(c_2^2 + \dots + c_n^2 \right)$$
$$= (1-\gamma)^{2k} \| c_2 \phi_2 + \dots + c_n \phi_n \|^2_\nu$$
$$= (1-\gamma)^{2k} \| h_0 - \mathbb{1} \|^2_\nu$$
$$= (1-\gamma)^{2k} \chi^2_\nu(\beta_0).$$

□

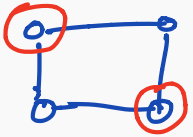
Ex: $\beta_0 = N(0, \alpha I)$, $\nu = N(0, I)$ $\Big\}$ $\chi^2_\nu(\beta_0) = \frac{1}{(\alpha(2-\alpha))^{n/2}}$ $\quad 0 < \alpha < 2$

Examples

Random Walk on Graph:

1) Cycle graph on n vertices

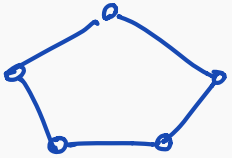
- n even:



$\Rightarrow$ bipartite

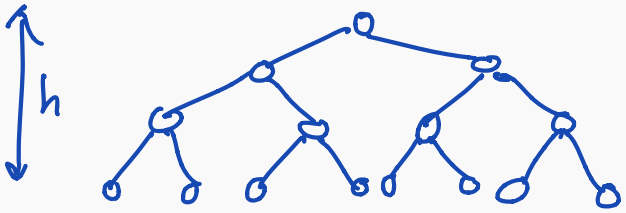
$\Rightarrow \lambda_n = -1 \Rightarrow \gamma = 0$

- n odd:



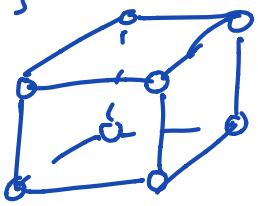
$\gamma = O\left(\frac{1}{n^2}\right) \Rightarrow \tau = O(n^2)$

2) Complete binary tree on $n = 2^h - 1$ vertices



$\gamma = O\left(\frac{1}{n}\right)$

3) Hypercube on $n = 2^h$ vertices $\{0,1\}^h$



$$\gamma = O\left(\frac{1}{h}\right) = O\left(\frac{1}{\log n}\right)$$

For general P:

$$\gamma = \inf_{f \in L^2(\nu)} \frac{\langle f, Lf \rangle_\nu}{\mathrm{Var}_\nu(f)}$$

$\mathrm{Var}_\nu(f) = \mathbb{E}_\nu[(f - \mathbb{E}_\nu[f])^2]$, with $\tilde{f} = f - \mathbb{E}_\nu[f]$

$$= \inf_{f \in L^2(\nu)} \frac{\langle f, Lf \rangle_\nu}{\|f\|_\nu^2}$$

$\mathbb{E}_\nu[f] = \langle f, \mathbb{1} \rangle_\nu = 0$

Exercise: Compute / bound spectral gap of OU:

$$X_{k+1} = e^{-\eta} X_k + \sqrt{(1 - e^{-2\eta})}\, Z \qquad Z \sim \mathcal{N}(0, I)$$

on $\mathcal{X} = \mathbb{R}^n$, $\eta > 0$ is step size

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

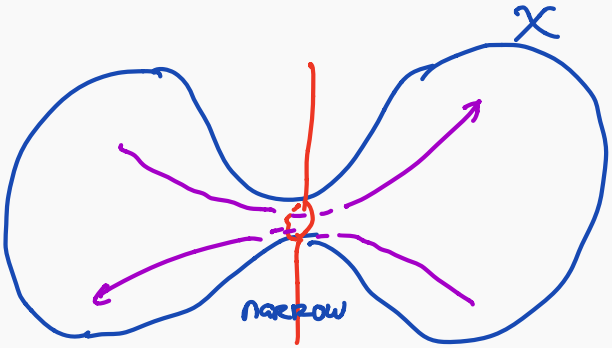
---

now: more intuitive notion

Conductance

Bottleneck

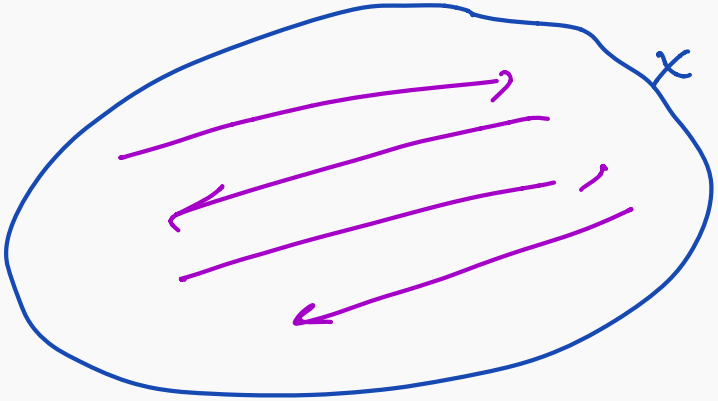
What prevents a random walk from mixing quickly?



bottleneck

⇒ hard to pass through

⇒ RW slow to mix



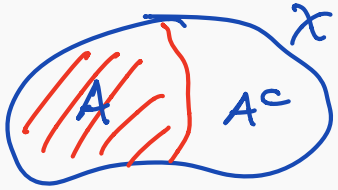
no bottleneck

⇒ RW fast mixing

Ergodic flow

Let P be (reversible) wrt. ν on $\mathcal{X}$

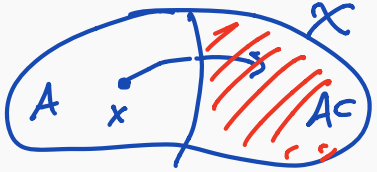
A subset $A \subseteq \mathcal{X}$ defines a partition / cut



$$\mathcal{X} = A \cup A^c$$

"
$\mathcal{X} \setminus A$ complement

Def: The ergodic flow of A is



$$\Phi(A) = \int_A P_x(A^c)\, d\nu(x)$$

$X_0 \sim \nu$

$$= \mathbb{P}_R(X_0 \in A, X_1 \notin A) \qquad X_1 | X_0 \sim P_{X_0}$$

note: Since ν is stationary for P, $\Phi(A) = \Phi(A^c)$

$$\mathbb{P}_R(X_0 \in A, X_1 \notin A) = \mathbb{P}_R(X_0 \notin A, X_1 \in A)$$

(does not need reversibility)

Bad case (bottleneck):

* if have large A with small $\Phi(A)$
* but can make $\Phi(A)$ small by making A small!!

$\Rightarrow$ should normalize: $\dfrac{\Phi(A)}{\nu(A)}$

* also, since $\Phi(A) = \Phi(A^c)$

$$\nu(A) = 1 - \nu(A^c)$$

$\Rightarrow$ should consider:

cut-ratio/ expansion

$$C(A) = \frac{\Phi(A)}{\min\{\nu(A), 1-\nu(A)\}}$$

* Suppose $\nu(A) \leq \frac{1}{2}$:

$$C(A) = \frac{\Phi(A)}{\nu(A)} = \frac{\mathbb{P}_R(X_0 \in A, X_1 \notin A)}{\mathbb{P}_R(X_0 \in A)}$$

$$= \mathbb{P}_R(X_1 \notin A \mid X_0 \in A)$$

Conductance

The **conductance** of Markov chain P is

$$\phi = \inf_{\substack{A \subset \mathcal{X} \\ 0 < \nu(A) < 1}} \frac{\Phi(A)}{\min\{\nu(A), 1 - \nu(A)\}}$$

$$= \inf_{0 < \nu(A) \le \frac{1}{2}} \mathbb{P}_{\nu}\left(X_1 \notin A \mid X_0 \in A\right) \qquad X_0 \sim \nu, \; X_1 \mid X_0 \sim P_{X_0}$$

Large conductance ϕ $\Rightarrow$ every subset has good expansion

$\Rightarrow$ RW mixes fast.

Conductance $\Leftrightarrow$ Spectral gap

Theorem (Cheeger's inequality)
Let P be reversible with spectral gap γ and conductance ϕ. Then:

surprising ↓ trivial ↓

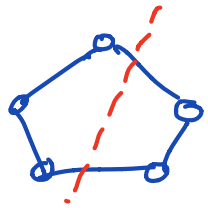
$$\frac{\phi^2}{2} \leq \gamma \leq 2\phi$$

Notes:

* Cheeger [70] for compact Riemannian manifold
* Alon [86] for graph
  $\Rightarrow$ used by Jerrum & Sinclair [88] for approx. permanent
* Lawler & Sokal [88] for general X
* We follow Lovász & Simonovits [93]

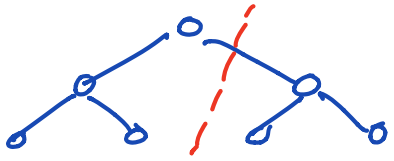
Example: RW on graph

* Cycle on n vertices: $\phi = O(\frac{1}{n})$, $\gamma = O(\frac{1}{n^2})$



$\gamma \sim \phi^2$

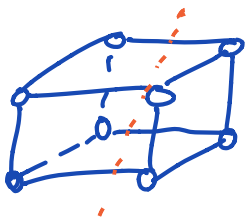
* Complete binary tree on n vertices



$\phi = O(\frac{1}{n})$, $\gamma = O(\frac{1}{n})$

$\gamma \sim \phi$

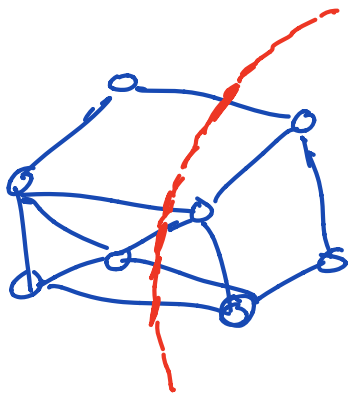
* Hypercube $\{0,1\}^h$ on $n = 2^h$ vertices



$\phi = O(\frac{1}{h}) = O(\frac{1}{\log n})$, $\gamma = O(\frac{1}{\log n})$

$\gamma \sim \phi$

Graph Cut / Partition



Proof of upper bound

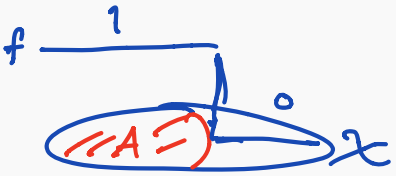
$$\boxed{\gamma \le 2\phi}$$

Proof:

$$\gamma = \inf_{f \in L^2(\nu)} \frac{\langle f, Lf \rangle_\nu}{\mathrm{Var}_\nu(f)} \qquad \phi = \inf_{A \subset X} \frac{\Phi(A)}{\min\{\nu(A), (-\nu(A)\}}$$

Given $A \subset X$, let $f = \mathbb{1}_A \in L^2(\nu)$

$$f(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{else} \end{cases}$$



$$\langle f, Lf \rangle_\nu = \frac{1}{2} \mathbb{E}\left[(f(X_1) - f(X_0))^2 \right]$$

$$f = \mathbb{1}_A : = \frac{1}{2} \mathbb{E}\left[(\mathbb{1}_A(X_1) - \mathbb{1}_A(X_0))^2 \right]$$

$$= \begin{cases} 1 & \text{if } (X_0 \in A, X_1 \notin A) \\ & \text{or } (X_0 \notin A, X_1 \in A) \\ 0 & \text{else} \end{cases}$$

$$= \frac{1}{2}\left(\mathbb{P}_R(X_0 \in A, X_1 \notin A) + \mathbb{P}_R(X_0 \notin A, X_1 \in A)\right)$$

$$= \frac{1}{2}\left(\Phi(A) + \Phi(A^c)\right)$$

$$= \Phi(A)$$

← Bernoulli RV. $\begin{cases} 1 & \text{wp. } \nu(A) \\ 0 & \text{else} \end{cases}$

and $\mathrm{Var}_\nu(f) = \mathrm{Var}_\nu(\mathbb{1}_A)$

$$= \nu(A)\left(1 - \nu(A)\right)$$

$$\geq \frac{1}{2}\min\{\nu(A), 1-\nu(A)\}$$

so $\gamma = \inf_f \dfrac{\langle f, Lf\rangle}{\mathrm{Var}_\nu(f)} \leq \inf_{A \subset X} \dfrac{\Phi(A)}{\frac{1}{2}\min\{\nu(A), 1-\nu(A)\}} = 2\Phi.$

↑ $f = \mathbb{1}_A$

□

Proof of lower bound

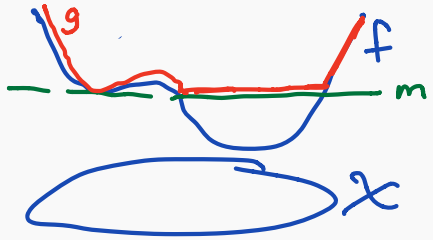
$$\frac{\phi^2}{2} \leq \gamma$$

Proof sketch: from $f \in L^2(\nu)$, need to define a cut.

given f with $\mathbb{E}_\nu[f] = 0$, let m be median of f:

$$\nu(\{x : f(x) > m\}) \leq \frac{1}{2}$$

$$\nu(\{x : f(x) < m\}) \leq \frac{1}{2}$$



Let $g(x) = \max\{f(x) - m, 0\}$

Use g to define cut

See [Lovasz & Simonovits '93 : Lemma 1.6]

□

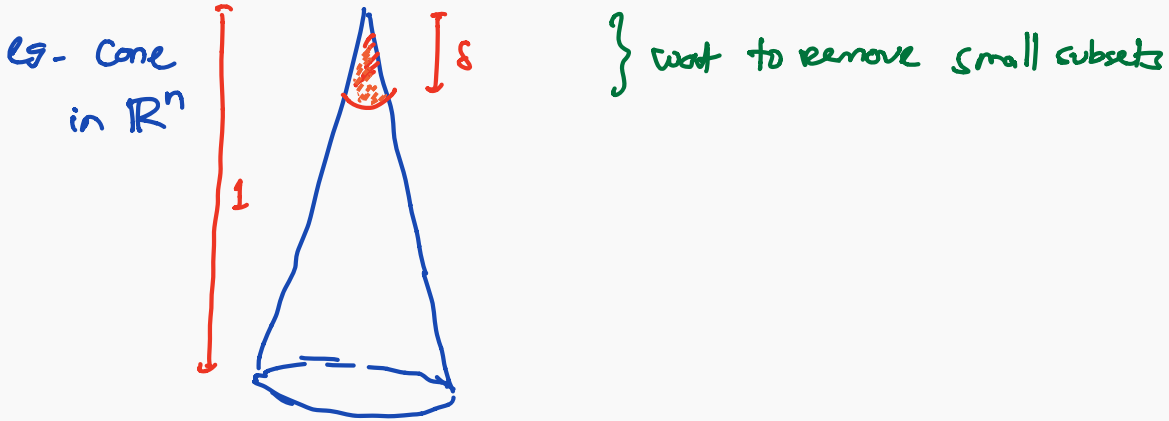
Mixing time bound via conductance

Cheeger: $\gamma \geq \frac{\phi^2}{2}$

$\Rightarrow$ mixing time in χ^2-divergence is

$$\tau = \tilde{O}\left(\frac{1}{\gamma}\right) = \tilde{O}\left(\frac{1}{\phi^2}\right)$$

But sometimes conductance too strict



} want to remove small subsets

s-Conductance

The s-**conductance** of Markov chain P is

$0 \leq s \leq \frac{1}{2}$

$$\phi_s = \inf_{\substack{A \subset \mathcal{X} \\ s < \nu(A) < 1-s}} \frac{\Phi(A)}{\min\{\nu(A) - s, 1 - s - \nu(A)\}}$$

↑ prevent A too small ↑ or too large

* Usual conductance : $s = 0$
* $s \longmapsto \phi_s$ is increasing
* can get mixing time bound in Total Variation distance under warm-start [Lovász & Simonovits '93]