CPSC 661: Sampling Algorithms in ML

Andre Wibisono

February 15, 2021

Yale University

Last time

- Reversible Markov chain $\nu(x) \cdot P(x,y) = \nu(y) \cdot P(y,x)$

  Laplacian $L = I - P \succeq 0$
- Spectral gap

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

  $$\phi = \inf_{A \subseteq X} \frac{\Phi(A) = \Pr(x_0 \in A,\ x_1 \notin A)}{\min\{\nu(A),\ 1-\nu(A)\}}$$
- Cheeger's inequality $\frac{\phi^2}{2} \le \gamma \le 2\phi$
- Mixing time in χ^2-divergence $\tau = \tilde{O}\left(\frac{1}{\gamma}\right) = \tilde{O}\left(\frac{1}{\phi^2}\right)$

Today: s-conductance, mixing time in TV distance

References

- 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

s-Conductance

For $0 \le s \le \frac{1}{2}$, the **s-conductance** of Markov chain P is

$X_0 \sim \nu, \quad X_1 | X_0 \sim P_{X_0}$

$$\phi_s = \inf_{\substack{A \subset \mathcal{X} \\ s < \nu(A) < 1-s}} \frac{\Phi(A) = \Pr(X_0 \in A, X_1 \notin A)}{\min\{\nu(A) - s, 1 - \nu(A) - s\}}$$

- Proposed by Lovász & Simonovits, *Mixing rate of Markov chains, an isoperimetric inequality, and computing the volume*, FOCS 1990

Dyer, Frieze, Kannan : $\tilde{O}(n^{23})$

$\Rightarrow$ Lovász & Simonovits '90: $\tilde{O}(n^{16})$

* usual conductance : $s = 0$ $\quad (\phi = \phi_0)$

* $s \mapsto \phi_s$ is increasing

* Goal: mixing time in TV distance $\checkmark$ $\tau(\varepsilon) = \tilde{O}\left(\frac{1}{\phi_\varepsilon^2}\right)$ under warm start

Total Variation (TV) distance

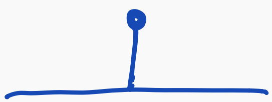
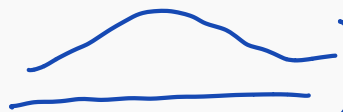
ρ, ν prob. dist. on $\mathcal{X}$ ← state space, σ-algebra $\mathcal{A} \subseteq 2^{\mathcal{X}}$

$$\mathrm{TV}(\rho,\nu) = \sup_{\substack{A\subseteq \mathcal{X} \\ A \in \mathcal{A}}} |\rho(A)-\nu(A)|$$

* A metric (non-negative, symmetric, satisfies triangle ineq.)

* Range: $0 \leq \mathrm{TV}(\rho,\nu) \leq 1$

* works for discrete / continuous distributions

eg. TV( , ) = 1

(whereas $\chi^2_\nu(\rho) = \infty$)

Many ways of writing TV

$$\mathrm{TV}(\rho,\nu) = \sup_{A\subseteq\mathcal{X}} |\rho(A)-\nu(A)|$$

Exercise:

1) $\mathrm{TV}(\rho,\nu) = \sup_{g:\mathcal{X}\to[0,1]} \mathbb{E}_\rho[g] - \mathbb{E}_\nu[g]$

2) $\mathrm{TV}(\rho,\nu) = \inf_{\substack{(X,Y):\\ X\sim\rho,\ Y\sim\nu}} \mathbb{P}_R(X\neq Y) = \mathbb{E}[\underbrace{\mathbb{1}\{X\neq Y\}}_{\text{Hamming distance}}]$

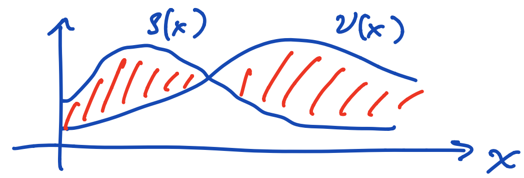
coupling

= joint distribution on $\mathcal{X}\times\mathcal{X}$

with proper marginals

3) $TV(\rho,\nu) = \frac{1}{2}\int_{\mathcal{X}} |\rho(x) - \nu(x)|\, dx$

$= \frac{1}{2} \|\rho - \nu\|_{L^1(dx)}$



4) if ρ has density $h = \frac{d\rho}{d\nu} : \mathcal{X} \to \mathbb{R}$

$TV(\rho,\nu) = \frac{1}{2}\int_{\mathcal{X}} |h(x) - 1|\, d\nu(x)$

$= \frac{1}{2} \| h - \mathbb{1} \|_{L^1(\nu)}$

(Recall: $\chi^2_\nu(\rho) = \| h - \mathbb{1}\|^2_{L^2(\nu)}$)

$\Rightarrow$ Inequality : $\boxed{TV(\rho,\nu) \leq \frac{1}{2}\sqrt{\chi^2_\nu(\rho)}}$

Warm start

The *warmness* of ρ with respect to ν is

$h = \frac{d\rho}{d\nu}$ density

$$M_\nu^\infty(\rho) = \|h - \mathbf{1}\|_{L^\infty(\nu)} = \sup_{x \in \mathcal{X}} |h(x) - 1|$$

* $\rho = \nu$: $M_\nu^\infty(\nu) = 0$

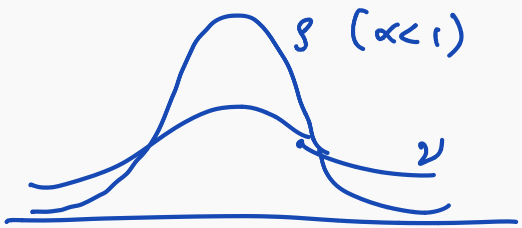
* Relations: 1) $TV(\rho, \nu) \le \frac{1}{2} M_\nu^\infty(\rho)$

2) $\chi_\nu^2(\rho, \nu) \le M_\nu^\infty(\rho)$

3) $\chi_\nu^2(\rho, \nu) \le 2\, TV(\rho, \nu) \cdot M_\nu^\infty(\rho)$

* Ex: $\rho = \mathcal{N}(0, \alpha I)$

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



$M_\nu^\infty(\rho) < \infty \iff 0 < \alpha \le 1$ $\left(\chi_\nu^2(\rho) < \infty \iff 0 < \alpha < 2 \right)$

Mixing time in TV distance

Let P be reversible with respect to ν with s-conductance ϕ_s.

Theorem (Lovász & Simonovits)
Assume ρ_0 is M-warm with respect to ν: $M_\nu^\infty(\rho_0) = M < \infty$.

For all $k \geq 0$ and $0 \leq s \leq \frac{1}{2}$:

$X_k \sim \rho_k$

$$TV(\rho_k, \nu) \leq Ms + M\left(1 - \frac{{\phi_s}^2}{2}\right)^k$$

$\Rightarrow$ to get $TV(\rho_k, \nu) \leq \varepsilon$:

1) set $Ms \leq \frac{\varepsilon}{2}$

$\Leftrightarrow \quad s \leq \frac{\varepsilon}{2M}$

set $s = \frac{\varepsilon}{2M}$

2) $M\left(1 - \frac{\phi_s^2}{2}\right)^k \leq M \cdot e^{-\frac{\phi_s^2 k}{2}} \leq \frac{\varepsilon}{2}$

$\Leftrightarrow \quad k \geq \frac{2}{\phi_s^2} \log \frac{2M}{\varepsilon}$

$\therefore$ mixing time in TV distance

$$\boxed{\tau(\varepsilon) = \frac{2}{\phi_s^2} \log \frac{2M}{\varepsilon} = \tilde{O}\left(\frac{1}{\phi_s^2}\right)}$$

with $s = \frac{\varepsilon}{2M}$, $M = M_2^{\infty}(\rho_0)$

Proof: 1. Definition

Want to bound $TV(\rho, \nu) = \sup_{\substack{(g: \mathcal{X} \to [0,1]) \\ = (g \in \mathcal{G})}} \mathbb{E}_\rho[g] - \mathbb{E}_\nu[g]$

$$= \sup_{A \subset \mathcal{X}} \rho(A) - \nu(A)$$

Let $\mathcal{G} = \{ g : \mathcal{X} \to [0,1] \}$ $\Big\}$ $\mathcal{G} = \bigcup_{0 \le q \le 1} \mathcal{G}_q$

for $0 \le q \le 1$: $\mathcal{G}_q = \{ g \in \mathcal{G} : \mathbb{E}_\nu[g] = q \}$

define "Total Variation at level q":

$$TV_q(\rho, \nu) = \sup_{g \in \mathcal{G}_q} \Big(\mathbb{E}_\rho[g] - \underbrace{q}_{= \mathbb{E}_\nu[g]} \Big)$$

if ν is atom-free

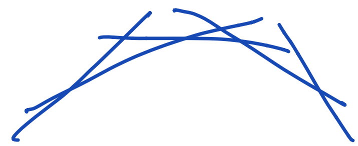
$$= \sup_{A \subset \mathcal{X} : \nu(A) = q} \big(\rho(A) - q \big)$$

so that $\quad TV(\rho,\nu) = \sup_{0\le q\le 1} TV_q(\rho,\nu)$

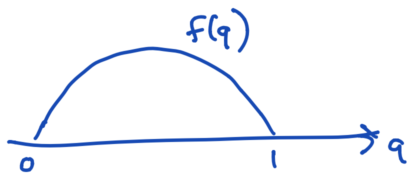
Fix ρ, ν, let $\quad f(q) = TV_q(\rho,\nu)$

Properties: 1) $0 \le f(q) \le 1-q$

2) f is concave in q



(supremum of linear function)

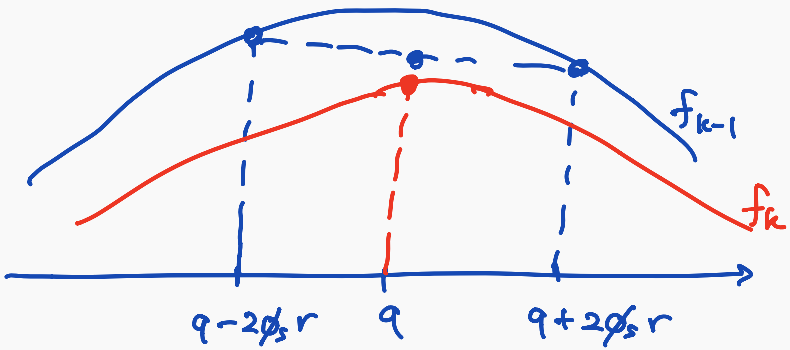


Along Markov chain $X_k \sim \rho_k$, let $f_k(q) = \mathrm{TV}_q(\rho_k, \nu)$

Lemma

Assume ν is atom-free and $0 \le s \le \frac{1}{2}$. For $s \le q \le 1 - s$, let $r = \min\{q - s, 1 - q - s\}$. Then

$$f_k(q) \le \frac{1}{2}\big(f_{k-1}(q - 2\phi_s r) + f_{k-1}(q + 2\phi_s r)\big)$$



Proof: Assume P is lazy $\quad (P_x(\{x\}) \geq \frac{1}{2})$

(can always make P lazy by flip a coin, stay if heads
else follow P)

Say $s \leq q \leq \frac{1}{2}$

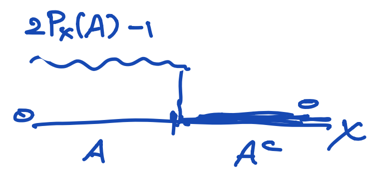
so $r = q - s$

Let A be subset s.t. $\nu(A) = q$

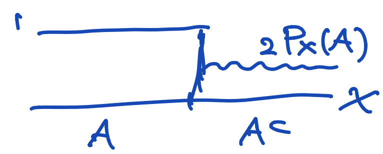
and $f_k(A) = \rho_k(A) - \nu(A)$
$= \rho_k(A) - q$

Define $g_1, g_2 : \mathcal{X} \to [0, 1]$ by

$$g_1(x) = \begin{cases} 2P_x(A) - 1 & \text{if } x \in A \\ 0 & \text{else} \end{cases}$$



$$g_2(x) = \begin{cases} 1 & \text{if } x \in A \\ 2P_x(A) & \text{else} \end{cases}$$



P lazy $\Rightarrow$ $0 \leq g_1(x), g_2(x) \leq 1$

by def: $\boxed{g_1(x) + g_2(x) = 2P_x(A)}$

Let $q_1 = \mathbb{E}_\nu[g_1]$

$q_2 = \mathbb{E}_\nu[g_2]$

ν stat: $\int_A dy \quad \nu(y) = \int_{\mathcal{X}} P_x(y)\, d\nu(x)$

$$\boxed{2q = q_1 + q_2} = \mathbb{E}_\nu[g_1 + g_2] = 2\int_{\mathcal{X}} P_x(A)\, d\nu(x)$$

$$= 2\,\nu(A) \quad \text{by stationarity}$$
$$= 2q$$

$$\rho_k(A) = \int_{\mathcal{X}} \underbrace{P_x(A)}_{\frac{1}{2}\left(g_1(x)+g_2(x)\right)} \, d\rho_{k-1}(x)$$

$$= \frac{1}{2}\left(\mathbb{E}_{\rho_{k-1}}[g_1] + \mathbb{E}_{\rho_{k-1}}[g_2]\right)$$

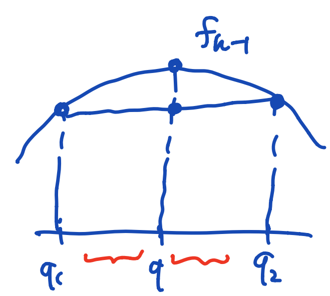
$$q = \frac{1}{2}\left(q_1 + q_2 \right)$$

then

$$f_k(q) = \rho_k(A) - q$$

$$= \underbrace{\frac{1}{2}\left(\mathbb{E}_{\rho_{k-1}}[g_1] - \underbrace{q_1}_{\mathbb{E}_\nu[g_1]}\right)}_{\le \sup\limits_{g \in \mathcal{G}_{q_1}} \left(\mathbb{E}_{\rho_{k-1}}[g_1] - q_1\right) = f_{k-1}(q_1)} + \underbrace{\frac{1}{2}\left(\mathbb{E}_{\rho_{k-1}}[g_2] - \underbrace{q_2}_{\mathbb{E}_\nu[g_2]}\right)}_{\le f_{k-1}(q_2)}$$

$$\therefore \quad \boxed{f_k(q) \le \frac{1}{2}\left(f_{k-1}(q_1) + f_{k-1}(q_2)\right)}$$



note: f_{k-1} concave:

$$\frac{1}{2}\left(f_{k-1}(q_1) + f_{k-1}(q_2)\right) \le f_{k-1}\left(\frac{q_1+q_2}{2}\right) = f_{k-1}(q)$$

$$\Rightarrow \quad f_k(q) \le f_{k-1}(q)$$

Moreover,

$$q_2 - q = \mathbb{E}_\nu[g_2] - q$$

$$= 2\int_{\mathcal{X}\setminus A} P_x(A)\, d\nu(x) + \underbrace{\cancel{\nu(A)}}_{=q} - \cancel{q}$$

$$= 2\Phi(A^c)$$

$$= 2\Phi(A)$$

$$\phi_s \overset{*}{=} \inf_{\substack{A \subset X : \\ s \le \underset{\overset{\shortparallel}{q}}{\nu(A)} \le \frac{1}{2}}} \frac{\Phi(A)}{\nu(A) - s}$$

$$\overset{*}{\geq} 2\phi_s(\nu(A) - s)$$

$$= 2\phi_s(q - s)$$

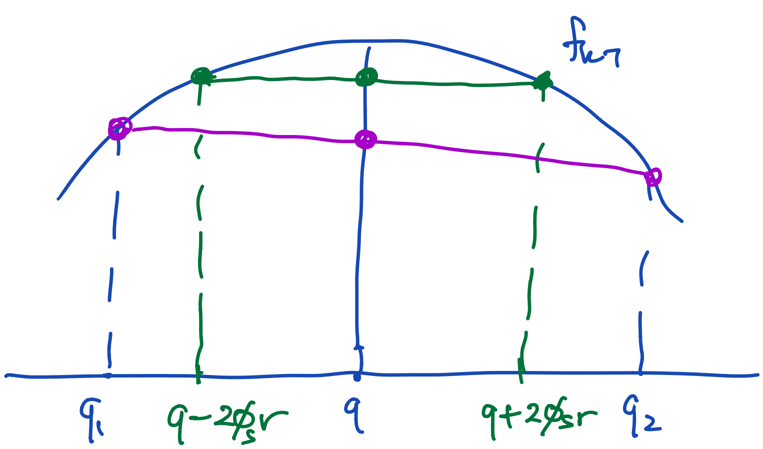
$$\therefore \quad q_2 - q = q - q_1 \geq \underline{2\phi_s(\overbrace{q-s}^{r})} = 2\phi_s \cdot r$$

$\therefore$ by concavity of f_{k-1}:

$$f_k(q) \leq \frac{1}{2}\left(f_{k-1}(q_1) + f_{k-1}(q_2)\right)$$

$$\leq \frac{1}{2}\left(f_{k-1}(q - 2\phi_s \cdot r) + f_{k-1}(q + 2\phi_s r)\right)$$

✓



Similarly for $\frac{1}{2} \leq q \leq 1-s$.

$\square$

3. Induction

Lemma

Let $0 \leq s \leq \frac{1}{2}$. Let C_1, C_2 be such that for all $s \leq q \leq 1 - s$:

$$TV_q(\mathcal{S}_0, \nu) = h_0(q) \leq C_1 + C_2 \min\left\{\sqrt{q - s}, \sqrt{1 - q - s}\right\}$$

Then for all $k \geq 0$ and $s \leq q \leq 1 - s$:

$$TV_q(\mathcal{S}_k, \nu) = h_k(q) \leq C_1 + C_2 \min\left\{\sqrt{q - s}, \sqrt{1 - q - s}\right\}\left(1 - \frac{{\phi_s}^2}{2}\right)^k$$

Proof: Induction.

* Base case $k=0$: Assumed

* Assume for $k-1$

* For k: let $s \le q \le \frac{1}{2}$

by Key Lemma:

$$f_k(q) \overset{*}{\le} \frac{1}{2}\Big(f_{k-1}\big(q - 2\phi_s(q-s)\big) + f_{k-1}\big(q + 2\phi_s(q-s)\big)\Big)$$

hypothesis: $$\le \frac{1}{2}\Bigg(C_1 + C_2 \sqrt{\overbrace{q - 2\phi_s(q-s) - s}^{(q-s)(1-2\phi_s)}}\left(1 - \frac{\phi_s^2}{2}\right)^{k-1} + C_1 + C_2 \sqrt{\underbrace{q + 2\phi_s(q-s) - s}_{(q-s)(1+2\phi_s)}}\left(1 - \frac{\phi_s^2}{2}\right)^{k-1}\Bigg)$$

$$= C_1 + C_2\sqrt{q-s}\left(\frac{\sqrt{1-2\phi_s} + \sqrt{1+2\phi_s}}{2}\right)\left(1 - \frac{\phi_s^2}{2}\right)^{k-1}$$

Lemma: $$\frac{\sqrt{1-a} + \sqrt{1+a}}{2} \le 1 - \frac{a^2}{8} \quad \text{for } 0 \le a \le 1$$

(since
$$\sqrt{1-a} = 1 - \cancel{\frac{a}{2}} - \frac{a^2}{8} - \cancel{\frac{a^3}{16}} - \frac{5}{128}a^4 + \dots$$
$$\sqrt{1+a} = 1 + \cancel{\frac{a}{2}} - \frac{a^2}{8} + \cancel{\frac{a^3}{16}} - \frac{5}{128}a^4 + \dots$$
$+$
$$\frac{\sqrt{1-a} + \sqrt{1+a}}{2} = 1 - \frac{a^2}{8} - \frac{5}{128}a^4 + \dots \le 1 - \frac{a^2}{8}$$

$$f_k(q) \le C_1 + C_2\sqrt{q-s}\left(1 - \frac{4\phi_s^2}{8}\right)\left(1 - \frac{\phi_s^2}{2}\right)^{k-1}$$

$$= C_1 + C_2\sqrt{q-s}\left(1 - \frac{\phi_s^2}{2}\right)^k.$$

$\square$

4. Proof of Theorem

Theorem (Lovász & Simonovits)
Assume ρ_0 is M-warm with respect to ν: $M = M_\nu^\infty(\rho_0) < \infty$.

For all $k \geq 0$ and $0 \leq s \leq \frac{1}{2}$:

$$TV(\rho_k, \nu) \leq Ms + M\left(1 - \frac{{\phi_s}^2}{2}\right)^k$$

Proof:

1) Show $f_0(q) \leq Ms + M \min\{\sqrt{q-s}, \sqrt{1-q-s}\}$

for $s \leq q \leq 1-s$

2) Apply induction:

$$f_k(q) \le Ms + M \min\left\{\sqrt{q-s}, \sqrt{1-q-s}\right\}\left(1-\frac{\phi_s^2}{2}\right)^k$$

for $s \le q \le 1-s$

3) show $f_k(q) \le Ms$ for $0 \le q \le s$

and $1-s \le q \le 1$

4) Conclude:

$$TV(\delta_u, \nu) = \sup_{0 \le q \le 1} f_k(q) = TV_q(\delta_u, \nu)$$

$$\le Ms + M \min\left\{\sqrt{q-s}, \sqrt{1-q-s}\right\}\left(1-\frac{\phi_s^2}{2}\right)^k.$$

$\square$

Recap

Reversible Markov chain P with s-conductance ϕ_s

$\Rightarrow$ Mixing time in TV distance $\tau(\epsilon) = \tilde{O}(1/{\phi_s}^2)$ with $s = \frac{\epsilon}{2M}$

Questions:

1. How to get reversible Markov chain P?
2. How to show large conductance?