CPSC 661: Sampling Algorithms in Machine Learning	Feb. 15, 2021
Lecture 5	
Lecturer: Andre Wibisono	*Scribe: Henry Li*

1 Outline and Definitions

Our goal this lecture is to provide a bound on the mixing time of reversible Markov chains with respect to a certain property of the Markov chain known as the s-conductance. We focus on a key result by Lovasz and Simonovits in *Mixing rate of Markov chains, an isoperimetric inequality, and computing the volume*, FOCS 1990 [1].
We define several mathematical objects useful for our result:

1. s-conductances,
2. the Total Variation (TV) distance,
3. the closely related the Total Variation distance at level q (TV_q), and
4. the *warmness* of a measure with respect to another measure.

1.1 s-conductances

Definition 1. *For $0 < s \leq 1/2$, the s-conductance of a Markov chain P is defined to be*

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

This was proposed by Lovasz and Simonovits in [1].

Note that

1. We recover (regular) conductance when $s = 0$ ($\phi = \phi_0$).
2. $s \mapsto \phi_s$ is monotonically increasing.

Our main goal this lecture is to bound the mixing time of reversible Markov chains by the s-conductance, i.e. in the form

$$\tau(\epsilon) = \widetilde{O}\left(\frac{1}{\phi_\epsilon^2}\right). \tag{2}$$

1.2 Total Variation (TV) distance

Definition 2. *The Total Variation distance between two probability measures* $\rho, \nu : \mathcal{X} \to \mathbb{R}_+$ *is defined to be*

$$TV(\rho, \nu) = \sup_{A \subseteq \mathcal{X}} |\rho(A) - \nu(A)|. \tag{3}$$

TV distance has range $[0, 1]$, works for discrete or continuous (ρ, ν), and satisfies:

1. non-negativity:
$$|\rho(A) - \nu(A)| \geq 0 \qquad \forall A \subset \mathcal{X},$$
2. symmetry:
$$TV(\rho, \nu) = \sup_{A \subseteq \mathcal{X}} |\rho(A) - \nu(A)| = \sup_{A \subseteq \mathcal{X}} |\nu(A) - \rho(A)| = TV(\nu, \rho),$$
3. and the triangle inequality:
$$\begin{aligned} TV(\rho, \nu) &= \sup_{A \subseteq \mathcal{X}} |\rho(A) - \nu(A)| \\ &= \sup_{A \subseteq \mathcal{X}} |\rho(A) - \mu(A) + \mu(A) - \nu(A)| \\ &\leq \sup_{A \subseteq \mathcal{X}} \Big(|\rho(A) - \mu(A)| + |\mu(A) - \nu(A)| \Big) \\ &\leq \sup_{A \subseteq \mathcal{X}} |\rho(A) - \mu(A)| + \sup_{A \subseteq \mathcal{X}} |\mu(A) - \nu(A)| \\ &= TV(\rho, \mu) + TV(\mu, \nu), \end{aligned}$$

therefore $TV(\cdot, \cdot)$ is a metric.

The following are all equivalent forms of TV distance:

1. $TV(\rho, \nu) = \sup_{g:\mathcal{X} \to [0,1]} E_\rho[g] - E_\nu[g]$.
2. $TV(\rho, \nu) = \inf_{\substack{(X,Y): \\ X \sim \rho, Y \sim \nu}} P(X \neq Y)$. [1]
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 $h = \frac{d\rho}{d\nu} : X \to \mathbb{R}$,
$$TV(\rho, \nu) = \frac{1}{2} \int_{\mathcal{X}} |h(x) - \mathbf{1}| d\nu(x) = \frac{1}{2} ||h - \mathbf{1}||_{L^1(\nu)},$$
where $\mathbf{1}$ is the all ones function ($f(x) = 1 \;\; \forall x$). (Recall: $\chi^2_\nu(\rho) = ||h - \mathbf{1}||^2_{L^2(\nu)}$).

[1] Note: this can also be written as the expectation of the hamming distance between X and Y $E[\mathbf{1}_{X \neq Y}]$.

Finally note the relationship between TV distance and χ^2 divergence:

$$TV(\rho, \nu) \leq \frac{1}{2}\sqrt{\chi^2_\nu(\rho)}.$$

1.3 Total Variation at level q

Definition 3. *Given $0 \leq q \leq 1$, the Total Variation at level q between two probability measures $\rho, \nu : \mathcal{X} \to \mathbb{R}_+$ is*

$$TV_q(\rho, \nu) = \sup_{\substack{A \subset \mathcal{X}: \\ \nu(A) = q}} \rho(A) - \nu(A). \tag{4}$$

We note the absence of the absolute value in the *Total Variation at level* q, (c.f. the TV distance). We now provide several properties of TV_q below.

1. **Equivalent formulation**: The Total Variation at level q can also be written as

$$TV_q(\rho, \nu) = \sup_{\substack{g: \mathcal{X} \to [0,1]: \\ E_\nu[g] = q}} \mathbb{E}_\rho[g] - \mathbb{E}_\nu[g].$$

2. **Maximization over q recovers TV**: $TV(\rho, \nu) = \sup_{0 \leq q \leq 1} TV_q(\rho, \nu)$.

   Proof. Since $\{A \subset \mathcal{X} : \nu(A) = q\} \subseteq \{A \subset \mathcal{X}\}$, we have $TV_q(\rho, \nu) \leq TV(\rho, \nu)$ and therefore

$$\sup_q TV_q(\rho, \nu) \leq TV(\rho, \nu).$$

   On the other hand, let $A^* \subset \mathcal{X}$ be such that $|\rho(A^*) - \nu(A^*)| = TV(\rho, \nu)$. Suppose first that $\rho(A^*) \geq \nu(A^*)$. Then $TV_{q^*}(\rho, \nu) = TV(\rho, \nu)$, where $q^* = \nu(A^*)$. Therefore

$$\sup_q TV_q(\rho, \nu) \geq TV(\rho, \nu).$$

   The same follows for $\rho(A^*) < \nu(A^*)$, where we use $q^* = 1 - \nu(A^*)$ instead. Combining both inequalities give our desired result. □

3. **Boundedness (wrt q)**: $0 \leq TV_q(\rho, \nu) \leq 1 - q$.

   Proof.

$$TV_q(\rho, \nu) = \sup_{\substack{A \subset \mathcal{X}: \\ \nu(A) = q}} \rho(A) - \nu(A) \leq 1 - \nu(A) = 1 - q$$

□

4. **Concavity (wrt q)**: Given $\lambda, q, p \in [0,1]$

Proof. Let $A^* \subset \mathcal{X}$ be such that

$$|\rho(A^*) - \nu(A^*)| = \sup_{A \subset \mathcal{X}} |\rho(A) - \nu(A)|,$$

and consider the case where $\rho(A^*) \geq \nu(A^*)$ (the other direction follows nearly the same steps). Furthermore, let $q* = \nu(A^*)$, and analogously define A_q^* such that

$$\rho(A_q^*) - \nu(A_q^*) = \sup_{\substack{A \subset \mathcal{X}: \\ \nu(A) = q}} \rho(A) - \nu(A) = \sup_{A \subset \mathcal{X}} |\rho(A) - \nu(A)|. \tag{5}$$

We make two observations.

First, that for all $q \in [0, q^*)$, $A_q^* \subseteq (A^* \cup A_0)$, where $A_0 := \{x \in \mathcal{X} : \nu(x) = 0\}$. (If not, it contains $x \in \mathcal{X} \setminus A^*$, in which case we can interchange it with some $y \in A^*$, and increase (5), contradicting our definition of A_q^*.) Similarly, we may show that for all $q \in (q^*, 1]$, $A_q^* \subseteq [(\mathcal{X} \setminus A^*) \cup A_0]$.

Second, that there is no $B \in A^*$ such that $\rho(B) < \nu(B)$, and conversely no $B \subset \mathcal{X} \setminus A^*$ such that $\rho(B) > \nu(B)$, as both cases would similarly contradict our definition of A^*.

Therefore TV_q is monotonically increasing for $q \in [0, q^*)$ and monotonically decreasing for $q \in (q^*, 1]$, and thus concave. $\square$

1.4 Warmness

Definition 4. *Let $\rho, \nu : \mathcal{X} \to \mathbb{R}_+$ be two probability measures. The warmness of ρ with respect to ν is*

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

where $h = \frac{d\rho}{d\nu}$.

Example: Suppose $\rho = N(0, \alpha I)$ and $\nu = N(0, I)$. Then $M_\nu^\infty(\rho) < \infty \iff \alpha \in (0, 1]$. [2]

We note the following:

1. If $\rho = \nu$, $M_\nu^\infty(\rho) = 0$.
2. Warmness can be related to TV distance and χ^2 divergence by
   (a) $TV(\rho, \nu) \leq \frac{1}{2} M_\nu^\infty(\rho)$,
   (b) $\chi_\nu^2(\rho, \nu) \leq M_\nu^\infty(\rho)$,
   (c) $\chi_\nu^2(\rho, \nu) \leq 2TV(\rho, \nu) \cdot M_\nu^\infty(\rho)$.

[2] Note that $\chi_\nu^2(\rho) < \infty \iff \alpha \in (0, 2)$.

2 Main result: Mixing time in TV distance

Let P be reversible with respect to ν with s-conductance ϕ_s. We then have the following bound on the TV distance of a Markov chain to its stationary distribution after k steps.

Theorem 1. *(Lovasz & Simonovits) Assume ρ_0 is M-warm with respect to ν: $M_\nu^\infty(\rho_0) = M < \infty$. For all $k \geq 0$ and $0 \leq s \leq \frac{1}{2}$:*

$$TV(\rho_k, \nu) \leq Ms + M\left(1 - \frac{\phi_s^2}{2}\right)^k \tag{6}$$

Now, to obtain our desired mixing time bound (2), we want to pick s and k so that

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

- We set s by: $Ms \leq \frac{\epsilon}{2} \iff s \leq \frac{\epsilon}{2M} \implies s = \frac{\epsilon}{2M}$.
- We set k by: $M\left(1 - \frac{\phi^2_{\frac{\epsilon}{2M}}}{2}\right)^k \leq M \exp -\frac{\phi_s^2 k}{2} \leq \frac{\epsilon}{2} \iff k \geq \frac{2}{\phi^2_{\frac{\epsilon}{2M}}} \log(\frac{2M}{\epsilon}) \implies k = \frac{2}{\phi^2_{\frac{\epsilon}{2M}}} \log\left(\frac{2M}{\epsilon}\right)$.

This gives the mixing time bound

$$\begin{aligned}
\tau(\epsilon) &= \min\{k : TV(\rho_k, \nu) \leq \epsilon\} \\
&\leq \min\left\{k : Ms + M\left(1 - \frac{\phi_s^2}{2}\right)^k \leq \epsilon\right\} \\
&= \frac{2}{\phi^2_{\frac{\epsilon}{2M}}} \log\left(\frac{2M}{\epsilon}\right) \\
&= \widetilde{O}\left(\frac{1}{\phi_\epsilon^2}\right).
\end{aligned}$$

And thus we have accomplished the goal we have set for ourselves this lecture. It remains to prove Theorem 1.

2.1 Proof of Theorem 1

We shall prove Theorem 1 by way of the following induction lemma.

Lemma 1. *Let $0 \le s \le 1/2$. Let C_1, C_2 be such that for all $s \le q \le 1-s$,*

$$f_0(q) \le C_1 + C_2 \min\left\{\sqrt{q-s}, \sqrt{1-q-s}\right\}. \tag{7}$$

Then for all $k \ge 0$ and $s \le q \le 1-s$,

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

Assuming Lemma 1 holds, we let $f_k(q) = TV_q(\rho_k, \nu)$, and note that

$$\begin{aligned} f_0(q) &= TV_q(\rho_0, \nu) \\ &\le TV(\rho_0, \nu) \\ &\le \frac{1}{2}M \\ &\le Ms + \sqrt{2}M \min\left\{\sqrt{q-s}, \sqrt{1-q-s}\right\}, \end{aligned}$$

where the first inequality follows from the definition of the Total Variation at level q, the second inequality from the M-warmness assumption on ρ_0, and the third inequality by inspection from the fact that

$$\frac{1}{2} \le s + \sqrt{2}\min\left\{\sqrt{q-s}, \sqrt{1-q-s}\right\}.$$

Now if $C_1 = Ms$ and $C_2 = \sqrt{2}M$, we see that (7) holds. Therefore, by Lemma 1, we have

$$\begin{aligned} TV(\rho_k, \nu) &= \sup_q TV_q(\rho_k, \nu) \\ &:= \sup_q f_k(q) \\ &\le \sup_q Ms + \sqrt{2}M \min\left\{\sqrt{q-s}, \sqrt{1-q-s}\right\}\left(1 - \frac{\phi_s^2}{2}\right)^k \\ &\le Ms + \sqrt{2}M\sqrt{\frac{1}{2}}\left(1 - \frac{\phi_s^2}{2}\right)^k \\ &= Ms + M\left(1 - \frac{\phi_s^2}{2}\right)^k, \end{aligned}$$

where the second inequality comes from the observation that

$$\min\left\{\sqrt{q-s}, \sqrt{1-q-s}\right\} \le \sqrt{\frac{1}{2}}$$

for all choices of s and $q \in (s, 1-s)$. This completes the proof. It now suffices to show Lemma 1 — however, this requires the following Lemma 2, which holds along any Markov chain $X_k \sim \rho_k$.

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

$$f_k(q) \le \frac{1}{2}(f_{k-1}(q - 2\phi_s r) + f_{k-1}(q + 2\phi_s r)). \tag{9}$$

Note that the concavity of f_{k-1} implies

$$\frac{1}{2}(f_{k-1}(q - 2\phi_s r) + f_{k-1}(q + 2\phi_s r)) \le f_{k-1}(\frac{1}{2}(q - 2\phi_s r + q + 2\phi_s r)), \tag{10}$$

where the right hand side of (10) is greater than $f_k(q)$ (by Lemma 2), and the left hand side is equal to $f_{k-1}(q)$. Therefore an immediate implication of Lemma 2 is that

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

Proof. **(of Lemma 2):** Let us assume

- that P is lazy, and
- that $s \le q \le 1/2$, which means that $r = q - s$.

Due to the atom-free assumption on ν, we have, for all k, the existence of some A that satisfies (4), i.e.

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

All usage of A henceforth will refer to this set.
We now define two functions g_1, g_2 by

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

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

and two constants q_1, q_2 by

$$q_1 = \mathbb{E}_\nu[g_1] \qquad \text{and} \qquad q_2 = \mathbb{E}_\nu[g_2].$$

Note two properties of the above definitions:

1. that $g_1(x) + g_2(x) = 2P_x(A)$, and
2. that $q = \frac{1}{2}(q_1 + q_2)$.

where the first observation is self-evident, and the second is due to

$$\begin{aligned} q_1 + q_2 &= \mathbb{E}_\nu[g_1 + g_2] \\ &= \mathbb{E}_\nu[2P_X(A)] \\ &= 2\int P_x(A) d\nu(x) \\ &= 2\int_A d\nu(x) = 2\nu(A), \end{aligned}$$

and the second to last equality is given by the stationarity of ν under P. Therefore, the third observation follows since $\nu(A) = q$.
Moreover, we also see that

$$\rho_k(A) = \mathbb{E}_{\rho_{k-1}}[P_X(A)] = \frac{1}{2}\mathbb{E}_{\rho_{k-1}}[g_1 + g_2].$$

Together, these properties imply the following inequality:

$$\begin{aligned} f_k(q) &= \rho_k(A) - q \\ &= \frac{1}{2}\left[\mathbb{E}_{\rho_{k-1}}[g_1 + g_2] - (q_1 + q_2)\right] \\ &= \frac{1}{2}\left[(\mathbb{E}_{\rho_{k-1}}[g_1] - \mathbb{E}_\nu[g_1]) + (\mathbb{E}_{\rho_{k-1}}[g_2] - \mathbb{E}_\nu[g_2])\right] \\ &\leq \frac{1}{2}\left[\sup_{\substack{g':\mathcal{X}\to[0,1]: \\ E_\nu[g']=q}} \mathbb{E}_{\rho_{k-1}}[g'] - \mathbb{E}_\nu[g'] + \sup_{\substack{g'':\mathcal{X}\to[0,1]: \\ E_\nu[g'']=q}} \mathbb{E}_{\rho_{k-1}}[g''] - \mathbb{E}_\nu[g'']\right] \\ &= \frac{1}{2}[f_{k-1}(q_1) + f_{k-1}(q_2)]. \end{aligned}$$

Now, since

$$\phi_s = \inf_{\substack{B\subseteq\mathcal{X} \\ s\leq\nu(B)\leq 1/2}} \frac{\Phi(B)}{\nu(B) - s} \implies \phi_s \leq \frac{\Phi(A)}{\nu(A) - s} \implies \phi_s(\nu(A) - s) \leq \Phi(A),$$

we have

$$\phi(A) \geq \phi_s(\nu(A) - s) = \phi_s(q - s) = \phi_s r.$$

And therefore

$$
\begin{aligned}
q_2 - q &= \mathbb{E}_\nu[g_2] - q \\
&= \int_{\mathcal{X}} g_2 d\nu(x) - q \\
&= \left(\int_{A^c} 2P_x(A) d\nu(x) + \int_A 1 d\nu(x)\right) - q \\
&= (2\Phi(A^c) + q) - q \\
&= 2\Phi(A^c) = 2\Phi(A) \\
&\geq 2\phi_s r.
\end{aligned}
$$

Since $\frac{1}{2}(q_1 + q_2) = q \implies q_2 - q = q - q_1$, we have the same inequality for $q - q_1$. This gives us that

$$
q_1 \leq q - 2\phi_s r \qquad \text{and} \qquad q_2 \geq q + 2\phi_s r,
$$

which finally implies that

$$
\frac{1}{2}[f_{k-1}(q_1) + f_{k-1}(q_2)] \geq \frac{1}{2}[f_{k-1}(q - 2\phi_s r) + f_{k-1}(q + 2\phi_s r)],
$$

due to the concavity of f_{k-1}. □

Proof. **(of Lemma 1):** We prove this by induction. To show the base case, note that the inequality

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

holds for $k = 0$ simply due to the lemma assumption.
Now, we prove the inductive step. Assume that the above holds for $k-1$. Then we have

$$
\begin{aligned}
f_k(q) &\leq \frac{1}{2}\left(f_{k-1}(q - 2\phi_s(q-s)) + f_{k-1}(q + 2\phi_s(q-s))\right) \\
&\leq \frac{1}{2}\Bigg(C_1 + C_2 \min\left\{\sqrt{q - 2\phi_s r - s}, \sqrt{1 - (q - 2\phi_s r) - s}\right\} \\
&\quad + C_1 + C_2 \min\left\{\sqrt{q + 2\phi_s(q-s) - s}, \sqrt{1 - (q + 2\phi_s(q-s)) - s}\right\}\Bigg)\left(1 - \frac{\phi_s^2}{2}\right)^{k-1}
\end{aligned}
$$

by Lemma 2 and our inductive assumption on f_{k-1}, respectively. Recall that $r = \min\{q-s, 1-q-s\}$. With some algebra, we can deduce that $\min\left\{\sqrt{q-2\phi_s r - s}, \sqrt{1-(q-2\phi_s r) - s}\right\} = r(1-2\phi_s)$, and $\min\left\{\sqrt{q+2\phi_s r - s}, \sqrt{1-(q+2\phi_s r) - s}\right\} = r(1+2\phi_s)$. Therefore, we have

$$\begin{aligned}
&\leq \frac{1}{2}\Big(C_1 + C_2 r(1-2\phi_s) + C_1 + C_2 r(1+2\phi_s)\Big)\left(1 - \frac{\phi_s^2}{2}\right)^{k-1} \\
&= C_1 + C_2 r \frac{\sqrt{1-2\phi_s}\sqrt{1+2\phi_s}}{2}\left(1 - \frac{\phi_s^2}{2}\right)^{k-1} \\
&= C_1 + C_2 r \left(1 - \frac{(2\phi_s)^2}{8}\right)\left(1 - \frac{\phi_s^2}{2}\right)^{k-1} \\
&= C_1 + C_2 \min\{q-s, 1-q-s\}\left(1 - \frac{\phi_s^2}{2}\right)^{k}.
\end{aligned}$$

□

References

[1] László Lovász and Miklós Simonovits. The mixing rate of markov chains, an isoperimetric inequality, and computing the volume. In *Proceedings [1990] 31st annual symposium on foundations of computer science*, pages 346–354. IEEE, 1990.