Lecture 12 - Continuous time chains and modified log-Sobolev

Monday, September 13

One of the major shortcomings of arguments based on spectral gap is the $\frac{1-\pi_*}{\pi_*}$ term. For example, for the lazy walk on the cycle C_n the mixing time is $\tau(\epsilon) = O(n^2 \log \frac{1}{\epsilon})$, however $1 - \lambda_2 = \Omega(1/n^2)$ and so the eigenvalue bound is only $\tau(\epsilon) = O(n^2 \log \frac{n}{\epsilon})$.

Observe that the asymptotic rate given by λ_2 is of course correct, so the main issue to be addressed here is the "burn-in" to reach distance say $\epsilon = 1/e$. We will look at several methods of studying the burn-in time. These are most easily studied in the context of continuous time Markov chains. We will always work with reversible chains, although these results hold for non-reversible chains as well.

Our discussion is an amalgam of Aldous-Fill [1] Chapter 3 Section 1.2, Jerrum's notes [4] Chapter 5.5, and Section 2 of Bobkov & Tetali's paper on modified log-Sobolev [2].

Definition 12.1. The *continuized chain* associated with a transition matrix P is such that given an infinitesimal dt then $\mathsf{P}(X_{t+dt} = j | X_t = i) = \mathsf{P}(i,j)\, dt$ if $j \neq i$.

One can check that

$$P(X_t = j | X_0 = i) = \left(e^{-t\,(I-\mathsf{P})}\right)_{ij}$$

is a solution to this condition. In particular,

$$X_{dt} = e^{-dt} \sum_{k=0}^{\infty} \mathsf{P}^k \frac{(dt)^k}{k!} = (1-dt)\,(\mathsf{P}^0 + \mathsf{P}\, dt) = (1-dt)\, I + \mathsf{P}\, dt$$

and so $\mathsf{P}(X_{t+dt} = j \neq i | X_t = i) = \mathsf{P}(X_{dt} = j | X_0 = i) = \mathsf{P}(i,j)\, dt$, as desired. An alternative interpretation of this formula is that after each step choose the time t until the next step such that t has mean value 1 (that is $Prob(t \in dx) = e^{-x}\, dx$), wait until time t, make a transition from P, and repeat.

Remark 12.2. *To implement a continuized Markov chain determine the number of steps N that the continuous time chain makes in time T by the equation $Prob(N = n|T) = \frac{T^n\, e^{-T}}{n!}$ (N is Poisson). It is easily checked that $\mathsf{E}(N|T) = T$ and therefore in particular $\mathsf{E}(N|T + dt) = T + dT$, as would be expected. Given the continuous-time mixing-time*

$$\tau_c(\epsilon) = \max_{\sigma} \inf\{t :\ \|\sigma\, e^{-t(I-\mathsf{P})} - \pi\|_{TV} \leq \epsilon\}$$

choose a Poisson random variable T with mean value $\tau_c(\epsilon)$ and run the discrete time chain for T steps. This will be within ϵ is stationary.

Corollary 2.2 of Diaconis and Saloff-Coste's paper on log-Sobolev [3] shows how discrete time and continuous time bounds on L_2 distance are related.

Remark 12.3. *If $(\overrightarrow{u_i})$ is an eigenvector of P then it is easily verified that*

$$(\overrightarrow{u_i})\ e^{-t\,(I-\mathsf{P})} = (\overrightarrow{u_i}) \sum_{k=0}^{\infty} (I-\mathsf{P})^k \frac{(-t)^k}{k!} = (\overrightarrow{u_i}) \sum_{k=0}^{\infty} (1-\lambda_i)^k \frac{(-t)^k}{k!} = (\overrightarrow{u_i})\ e^{-t(1-\lambda_i)}$$

In particular, the continuized chain has the same stationary distribution as P.

Let $H_t = e^{-t(I-P)}$ denote the t-time transition matrix. Then the t-step distribution is found by $\mathsf{p}^{(t)} = \mathsf{p}^{(0)}\, H_t$ (verify this), while the t-step relative probability distribution $f_t(x) := \frac{\mathsf{p}^{(t)}(x)}{\pi(x)}$ can be found by $f_t = H_t\, f_0$, where $\mathsf{p}^{(t)}$ is considered as a row vector and f_t as a column vector.

The *Laplacian* is given by $L = -(I - P)$, and so $H_t = e^{t\,L}$. If a function f on Ω is considered as a column vector then we use the notation $L(f)(x) = (L\,f)(x) = \sum_{y\in\Omega} L(x,y)f(y)$.

The following Lemmas will be key to the remainder of today's lecture.

Lemma 12.4. *If* $f_t = H_t\,f_0$ *then* $\frac{d\,f_t(x)}{dt} = L\,f_t(x)$.

Verify this yourself. Also,

Lemma 12.5. *The Dirichlet form* $\mathcal{E}(f,g)$ *satisfies*

$$\mathcal{E}(f,g) = -\sum_{x\in\Omega} f(x)\,L(g)(x)\,\pi(x)\,.$$

Proof.

$$\begin{aligned}\mathcal{E}(f,g) &= \frac{1}{2}\sum_{x,y\in\Omega} \pi(x)\,\mathsf{P}(x,y)\,(f(x)g(x) + f(y)g(y) - f(x)g(y) - f(y)g(x)) \\ &= <f,g>_\pi - \sum_{x,y\in\Omega} \pi(x)\,\mathsf{P}(x,y)\,f(x)g(y)\end{aligned}$$

But

$$\begin{aligned}\sum_{x\in\Omega} f(x)\,L(g)(x)\,\pi(x) &= \sum_{x,y\in\Omega} f(x)g(y)\,(P-I)(x,y)\,\pi(x) \\ &= -<f,g>_\pi + \sum_{x,y\in\Omega} f(x)g(y)\mathsf{P}(x,y)\pi(x)\end{aligned}$$

□

From these two facts we can show many interesting results.

Lemma 12.6. *If* $f_t = H_t\,f_0$ *then*

$$\frac{d}{dt}Var_\pi(f_t) = -2\,\mathcal{E}(f_t,f_t)$$

Proof. Observe that $Var_\pi(f) = \mathsf{E}f^2 - (\mathsf{E}f)^2 = \sum_{x\in\Omega} f(x)^2\pi(x) - (\mathsf{E}f)^2$, where it is easily checked that $\mathsf{E}f_t = \mathsf{E}f_0$ and is hence constant.

Then

$$\frac{d}{dt}Var_\pi(f_t) = \sum_{x\in\Omega}\frac{d}{dt}f_t^2(x)\,\pi(x) = 2\sum_{x\in\Omega} f_t(x)\,(L\,f_t)(x)\,\pi(x) = -2\,\mathcal{E}(f_t,\,f_t)$$

□

This last statement is especially interesting. Recall that

$$\lambda = \inf_{f\neq constant}\frac{\mathcal{E}(f,f)}{Var_\pi(f)}$$

But by definition of $Var_\pi(f_t)$ and of the L^2 distance it follows that $\|1 - \frac{\mathsf{p}^{(t)}}{\pi}\|_{2,\pi} = Var_\pi(f_t)$ when $f_t = \frac{\mathsf{p}^{(t)}}{\pi}$. Therefore,

$$\frac{d}{dt}\left\|1 - \frac{\mathsf{p}^{(t)}(t)}{\pi}\right\|_{2,\pi}^2 = -2\,\mathcal{E}(f_t,f_t) \leq -2\,\lambda\,Var_\pi(f_t) = -2\,\lambda\left\|1 - \frac{\mathsf{p}^{(t)}(t)}{\pi}\right\|_{2,\pi}^2$$

It follows that the L_2 distance is decreasing at an exponential rate and so

Theorem 12.7.

$$\|\mathsf{p}^{(t)} - \pi\|_{TV} \le \frac{1}{2}\left\|1 - \frac{\mathsf{p}^{(t)}}{\pi}\right\|_{2,\pi} \le \frac{1}{2}\,e^{-\lambda\,t}\left\|1 - \frac{\mathsf{p}^{(0)}}{\pi}\right\|_{2,\pi} \le \frac{1}{2}\,e^{-\lambda\,t}\sqrt{\frac{1-\pi_*}{\pi_*}}\,.$$

This was much easier than the proof in the discrete case that $\lambda = 1 - \lambda_2$ governs mixing time. No need for spectral decomposition or any of that. Moreover, there is no need to worry about λ_n. This is because a continuous time chain is automatically aperiodic since the transition times are random.

Another useful bound on variation distance is given by considering the following.

Definition 12.8. The *informational divergence* is given by

$$\mathsf{D}(\mu\|\pi) = Ent_\pi\left(\frac{\mu}{\pi}\right)$$

where the *entropy*

$$Ent_\pi(f) = \mathsf{E}_\pi f\,\log\frac{f}{\mathsf{E}_\pi f} = \sum_{x\in\Omega}\pi(x)f(x)\,\log\frac{f(x)}{\sum_{y\in\Omega}\pi(y)f(y)}$$

Lemma 12.9.

$$\|\mu - \pi\|_{TV}^2 \le 2\,\mathsf{D}(\mu\|\pi) \le 2\,Var_\pi\left(\frac{\mu}{\pi}\right)$$

Proof. Let $f(v) = \frac{\mu(v)}{\pi(v)}$ below.

For the first inequality,

$$\begin{aligned}
\|\mu - \pi\|_{TV} &= \sum_{\pi(v)\ge\mu(v)}(\pi(v) - \mu(v)) = \mathsf{E}_\pi(1 - f(v))^+ \\
&\le \mathsf{E}_\pi\sqrt{2(1 - f(v) + f(v)\log f(v))} \\
&\le \sqrt{\mathsf{E}_\pi\ 2(1 - f(v) + f(v)\log f(v))} \\
&= \sqrt{2\mathsf{E}_\pi f(v)\log f(v)} \\
&= \sqrt{2\mathsf{D}(\mu\|\pi)}
\end{aligned}$$

where $x^+ = \max\{x, 0\}$. The first inequality follows from $\forall x > 0 :\ (1-x)^+ \le \sqrt{2\,(1 - x + x\log x)}$, the second inequality is Cauchy-Schwartz, and the following equality follows from $\sum_{v\in V}\pi(v) = \sum_{v\in V}\mu(v) = 1$ so $\mathsf{E}_\pi f(v) = 1$.

The second inequality follows easily from $\log x \le x - 1$ and $\mathsf{E}_\pi f\,(f-1) = \mathsf{E}_\pi f^2 - 1 = Var_\pi(f)$. ☐

Lemma 12.10.

$$\frac{d}{dt}\mathsf{D}(\mathsf{p}^{(t)}\|\pi) = -\mathcal{E}(f_t, \log f_t)$$

Proof. This is just like the variance proof.

$$\frac{d}{dt}\mathsf{D}(\mathsf{p}^{(t)}\|\pi) = \sum_{x\in\Omega}\left(\frac{d}{dt}f_t\,\log f_t\right)(x)\,\pi(x) = \sum_{x\in\Omega}\pi(x)\,(\log f_t + 1)\,Lf_t$$

$$= -\mathcal{E}(f_t, \log f_t + 1) = \sum_{x\in\Omega}\pi(x)\,L(\log f_t + 1)\,f_t = -\mathcal{E}(f_t, \log f_t)$$

☐

This suggests considering the following quantity.

Definition 12.11. The *modified log-Sobolev constant* ρ_0 is given by

$$\rho_0 = \inf_{Ent_\pi(f)\neq 0} \frac{\mathcal{E}(f, \log f)}{2\, Ent_\pi(f)}$$

In the discrete Markov setting Bobkov and Tetali [2] initiated the study of modified log-Sobolev. Most of what is know about ρ_0 can be found in their paper.

Then, just as before, we have

Theorem 12.12.

$$\|\mathsf{p}^{(t)} - \pi\|_{TV} \leq \sqrt{2\,\mathsf{D}(\mathsf{p}^{(t)}\|\pi)} \leq e^{-\rho_0\, t}\,\sqrt{2\mathsf{D}(\mathsf{p}^{(0)}\|\pi)} \leq e^{-\rho_0\, t}\,\sqrt{2\log(1/\pi_*)}\,.$$

When bounding the total variation distance this improves substantially over the spectral gap bound because the $\frac{1-\pi_*}{\pi_*}$ term has been replaced by $\log(1/\pi_*)$. For instance, for the walk on the boolean cube $\{0,1\}^d$ then $\pi_* = 2^d$ and $\log(1/\pi_*) = d\log 2$, vs. $\frac{1-\pi_*}{\pi_*} \approx 2^d$ for the spectral bound.

A few examples demonstrating this improvement will be given briefly next class.

References

[1] D. Aldous and J. Fill. Reversible markov chains and random walks on graphs (book to appear). *URL for draft http://www.stat.Berkeley.edu/users/aldous.*

[2] S. Bobkov and P. Tetali. Modified log-sobolev inequalities in discrete settings. *preprint available at http://www.math.gatech.edu/˜tetali/RESEARCH/pubs.html*, 2003.

[3] P. Diaconis and L. Saloff-Coste. Logarithmic sobolev inequalities for finite markov chains. *The Annals of Applied Probability, see http://projecteuclid.org*, 6(3):695–750, 1996.

[4] M. Jerrum. *Counting, sampling and integration: algorithms and complexity.* Birkhauser Boston, also see author's website, 2003.