Lecture 13 - Operator norms and Nash inequalities

Wednesday, September 15

In the previous lecture two ways of upper bounding the mixing time were shown.

$$\tau(\epsilon) \leq \begin{cases} \dfrac{1}{\lambda}\left(\dfrac{1}{2}\log\dfrac{1-\pi_*}{\pi_*} + \log\dfrac{1}{2\epsilon}\right) \\ \dfrac{1}{\rho_0}\left(\dfrac{1}{2}\log\log\dfrac{1}{\pi_*} + \log\dfrac{\sqrt{2}}{\epsilon}\right) \end{cases}$$

The modified log-Sobolev bound can be a major improvement over working with spectral gap λ.

Example 13.1. Consider the lazy random walk on the boolean cube $\{0,1\}^d$. It was previously mentioned that $\lambda = 1-\lambda_2 = 1/2d$ and $\tau(\epsilon) \approx \frac{1}{2}\, d\log d + d\, \log(1/\epsilon)$. It turns out that $\rho_0 = 1/2d$ (to be shown on Friday). The corresponding bounds on mixing time are thus $\tau(\epsilon) = O(d^2 + d\, \log(1/\epsilon))$ via the spectral gap, but the much better $\tau(\epsilon) = \frac{\log(2\log 2)}{2} + d\log d + \log(1/\epsilon) = O(d\log d + d\, \log(1/\epsilon))$ via the modified log-Sobolev.

Example 13.2. Another interesting example is a random walk on the symmetric group S_n with transitions given by choosing a transposition uniformly at random and multiplying the current permutation by this transposition. Then $\lambda = \frac{2}{n-1}$ and $\frac{1}{2(n-1)} \leq \rho_0 \leq \frac{2}{n-1}$ and so the mixing time via a spectral argument is only $\tau(\epsilon) = O(n^2 + n\, \log(1/\epsilon))$ while that due to modified log-Sobolev is correct, at $\tau(\epsilon) = O(n\, \log n + n\log(1/\epsilon))$.

However, ρ_0 seems to be very difficult to compute, largely because of the logarithm in the numerator, $\mathcal{E}(f, \log f)$.

Example 13.3. Consider one of the simplest walks possible, the random walk on the two-point space $\{0,1\}$ with transitions $\mathsf{P} = \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix}$ and $\pi = \left(\frac{b}{a+b}, \frac{a}{a+b}\right)$. Then it is known that $\lambda = a+b$, but the only bounds on ρ_0 are $a+b \geq \rho_0 \geq \frac{a+b}{2} + \sqrt{ab}$.

See Bobkov and Tetali [2] for details on all three of these examples.

In the remainder of this lecture we will work on developing another method for bounding mixing time, the Nash inequality. This will be followed up on Friday with the logarithmic Sobolev inequality. It will take more work to prove the main theorem than it did for modified log-Sobolev, but the results seem easier to do computations with.

Recall from the previous lecture that

$$\frac{d}{dt}\, Var_\pi(f_t) = -2\mathcal{E}(f_t, f_t) \quad and \quad \frac{d}{dt}\, Ent_\pi(f_t) = -\mathcal{E}(f_t, \log f_t)\,.$$

These nice derivatives can be considered as a reason to try studying mixing time via variance (L^2 distance) and entropy (informational divergence). Another related derivative is

Lemma 13.4. *If $f_t = H_t\, f_0$ and $p \geq 1$ then*

$$\frac{d}{dt}\, \|f_t\|_p^p = -p\, \mathcal{E}(f_t, f_t^{p-1})\,.$$

Proof. The proof is almost exactly the same as the variance proof given last class. □

To make use of this lemma it is necessary to relate total variation distance to the p-norm of a function. It was previously shown that the asymptotic behavior of a Markov chain is governed by the spectral gap. Therefore it is reasonable to study the behavior over some initial number of steps, and then leave the asymptotics to the spectral gap.

Observe that if $f_t := \frac{\mathsf{p}^{(t)}}{\pi} = H_t\, f_0$ then

$$\|\mathsf{p}^{(t)} - \pi\|_{TV} \leq \frac{1}{2}\sqrt{Var_\pi(f_t)} = \frac{1}{2}\,\|(H_t - E)\, f_0\|_{2,\pi} = \frac{1}{2}\,\|(H_{t-s} - E)\, H_s f_0\|_{2,\pi}$$

where E is the expectation operator. In our setting the rows of E are all the identical π, and so $E\, f_0 = 1$.

To pull out the initial s step behavior a notion of matrix norms is required. Note that for the next few lectures any time a "norm" is discussed this will always be with respect to the stationary measure π.

Definition 13.5. Given domain Ω with $|\Omega| = n$, an $n \times n$ matrix A and $1 \leq q_1,\, q_2 \leq \infty$ then the *operator norm* is defined by

$$\|\mathsf{A}\|_{q_1 \to q_2} = \sup\{\|\mathsf{A}\, f\|_{q_2} \,:\, \|f\|_{q_1} = 1\} = \sup \frac{\|\mathsf{A}\, f\|_{q_2}}{\|f\|_{q_1}}$$

where the p-norm of a function is as in Lecture 5, $\|f\|_p^p = \sum_{x\in\Omega} |f(x)|^p\, \pi(x)$.

A triangle like inequality follows almost immediately from the definition. Given $1 \leq q_1,\, q_2,\, q_3 \leq \infty$ and two matrices A and B, then

$$\|\mathsf{BA}\|_{q_1 \to q_3} \leq \|\mathsf{A}\|_{q_1 \to q_2} \|\mathsf{B}\|_{q_2 \to q_3}\,.$$

Then

Lemma 13.6. *For an irreducible reversible Markov chain with* $0 \leq s \leq t \in \mathbb{R}^+$ *and* $f_t = \frac{\mathsf{p}^{(t)}}{\pi} = H_t\, \frac{\mathsf{p}^{(0)}}{\pi}$ *then*

$$\sqrt{Var_\pi(f_t)} \leq \|H_s\, f_0\|_2\, \|H_{t-s} - E\|_{2\to 2} \leq \|H_s\|_{1\to 2}\, e^{-(t-s)\lambda}$$

Proof. The first inequality follows from the definition of operator norms. The second inequality contains two parts. First, it is easy to check that $\|f_0\|_1 = 1$. Then $\|H_s\, f_0\|_2 \leq \|H_s\|_{1\to 2}\, \|f_0\|_1 = \|H_s\|_{1\to 2}$. The spectral portion was shown last class in a different form, as the exponential decay of variance. □

Note that if $s = 0$ then $H_0 = I$ and since a point mass is the worst case for convergence then $\|H_s\|_{1\to 2} = \sqrt{\frac{1-\pi_*}{\pi_*}}$. The above lemma is thus a generalization of the standard spectral bound. However, generally it is better to choose the largest value of s such that $\|H_s\|_{1\to 2}$ drops faster in s than does $\|H_s - E\|_{2\to 2}$, and then use the spectral gap to show that distance continues to decay nicely in the final $(t - s)$ steps.

The quantity

$$N(s) := \|H_s\|_{1\to 2} = \max_{i\in\Omega} \left\| \frac{\delta_i}{\pi(i)}\, H_s \right\|_{2,\pi} = \max_{i\in\Omega} \sqrt{\frac{\mathsf{P}^{2s}(i,i)}{\pi(i)}}$$

is generally difficult to study (the second equality is because a point mass is the worst starting point, and the final equality follows by the same argument as in Theorem 5.2).

Definition 13.7. A *Nash inequality* is an inequality of the form

$$\|g\|_2^{2+1/D} \leq C\, \left[\mathcal{E}(g,g) + \frac{1}{T}\|g\|_2^2\right]\, \|g\|_1^{1/D}$$

which holds for all functions $g : \Omega \to \mathbb{R}$ and for some positive constants C, D and T.

The multitude of constants makes sense if one considers a walk on the unbounded space $\mathbb{R}^D$ and wants to study how well a walk has converged in the interval $[-T, T]^D$, with C measuring the level of convergence. Such an argument will be given in a later lecture to study a simple random walk on the d-dimensional grid $[k]^d$ with side length k.

Theorem 13.8. *If a Nash inequality holds then*

$$N(t) \le e\,(DC/t)^D \quad \textit{for } 0 < t \le T$$

For a proof see Chapter 8, Theorem 17 of Aldous-Fill book.

Nash inequalities are not widely used in the rapid mixing community as they seem very difficult to prove. However, if a chain is a product of simple chains then a product theorem applies, or if it can be compared to a simple chain then a comparison theorem works. See Chapter 8, Section 3.2 of Aldous-Fill book [1] for details. The study of Nash inequalities for discrete Markov chains was initiated by Diaconis and Saloff-Coste [3], and most of what is known can be found in their paper.

Roughly speaking, the Nash inequality may be used to show very fast convergence early on, log-Sobolev (next class) will show pretty fast convergence for a while longer, and then spectral gap will kick in for the asymptotics. This will be discussed further during the 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. Nash inequaltites for finite markov chains. *Journal of Theoretical Probability*, 9:459–510, 1996.