CS294 Markov Chain Monte Carlo: Foundations & Applications **Fall 2009**

Lecture 14: October 22

Lecturer: Alistair Sinclair *Scribes: Alistair Sinclair*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

14.1 Relating mixing times using flows

14.1.1 Theory

This section describes a flow-based method for bounding the Poincaré constant (or, in the reversible case, the spectral gap) of a Markov chain based on that of another Markov chain, due to Diaconis and Saloff-Coste [DS93]. Assume we are given two ergodic, lazy Markov chains P and $\tilde{P}$ which share the same stationary distribution π over some state space Ω. Furthermore, assume that we have already proved a lower bound on the Poincaré constant $\tilde{\alpha}$ of $\tilde{P}$ by some other method. By constructing a flow on P that "simulates" the transitions of $\tilde{P}$, we can bound the ratio of the Poincaré constant of P to that of $\tilde{P}$ by a constant determined by the characteristics of the flow.

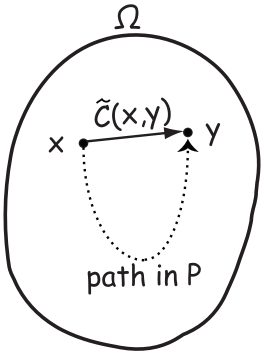

Figure 14.1: A $(P, \tilde{P})$-flow

Definition 14.1 *Let $\mathcal{Q}_{xy}$ denote the set of all simple $x \rightsquigarrow y$ paths in P, and $\mathcal{Q} = \cup_{x,y}\mathcal{Q}_{xy}$. A $(P, \tilde{P})$-flow is a function $f : \mathcal{Q} \to \mathbb{R}^+ \cup \{0\}$ subject to demand constraints*

$$\forall x, y : \quad \sum_{p \in \mathcal{Q}_{xy}} f(p) = \tilde{C}(x, y),$$

where $\tilde{C}(x, y) = \pi(x)\tilde{P}(x, y)$ (the capacity of edge (x, y) in $\tilde{P}$).

Recall that we define the cost of a flow f as

$$\rho(f) = \max_e \frac{f(e)}{C(e)}$$

and the length of a flow as

$$\ell(f) = \max_{p: f(p)>0} |p|.$$

Claim 14.2 *For any $(P, \tilde{P})$ flow f and function $\varphi : \Omega \to \mathbb{R}$, we have*

$$\sum_{x,y} (\varphi(x) - \varphi(y))^2 \, C(x,y) \geq \frac{1}{\rho(f)\ell(f)} \sum_{x,y} (\varphi(x) - \varphi(y))^2 \, \tilde{C}(x,y).$$

The proof of this fact follows that of Theorem 10.6 in Lecture 10; it involves expanding out $\widetilde{C}(x,y)$, applying the Cauchy-Schwartz inequality, and summing over paths.

Exercise 14.3 *Prove Claim 14.2 via pattern matching on the proof of Theorem 10.6.*

Theorem 14.4 *For any $(P, \tilde{P})$-flow f,*

$$\frac{\alpha}{\tilde{\alpha}} \geq \frac{1}{\rho(f)\ell(f)}$$

Proof: Recall from Lecture 10 that

$$\alpha = \inf_{\varphi} \frac{\sum_{x,y} (\varphi(x) - \varphi(y))^2 \, C(x,y)}{\sum_{x,y} (\varphi(x) - \varphi(y))^2 \, \pi(x)\pi(y)},$$

where the inf is taken over non-constant functions $\varphi : \Omega \to \mathbb{R}$. Thus, we can conclude

$$\begin{aligned} \frac{\alpha}{\tilde{\alpha}} &= \frac{\inf_{\varphi} \frac{\sum_{x,y}(\varphi(x)-\varphi(y))^2 C(x,y)}{\sum_{x,y}(\varphi(x)-\varphi(y))^2 \pi(x)\pi(y)}}{\inf_{\varphi} \frac{\sum_{x,y}(\varphi(x)-\varphi(y))^2 \tilde{C}(x,y)}{\sum_{x,y}(\varphi(x)-\varphi(y))^2 \pi(x)\pi(y)}} \\ &\geq \inf_{\varphi} \frac{\sum_{x,y} (\varphi(x) - \varphi(y))^2 \, C(x,y)}{\sum_{x,y} (\varphi(x) - \varphi(y))^2 \, \tilde{C}(x,y)} \\ &\geq \frac{1}{\rho(f)\ell(f)}. \end{aligned}$$

■

Note that the original flow theorem from Lecture 10 can be seen as a special case of Theorem 14.4 in which $\tilde{P}$ is the trivial Markov chain with $\tilde{P}(x,y) = \pi(y)$, which mixes in one step and has a Poincaré constant of 1. Theorem 14.4 can be very useful, because other methods of analysis are often rather sensitive to the details of the Markov chain being analyzed; the theorem allows us to bootstrap the analysis of different, perhaps more complex Markov chains by relating their convergence to that of simpler versions that we can analyze directly.

14.1.2 Example

In this section, the technique just presented will be used to prove the convergence properties of the "vanilla" Markov chain on multi-path routings without tower moves (from Lecture 8).

Recall that there is a one-to-one correspondence between "lozenge tilings" of regular hexagons and "routings" on the rectangular lattice where each point belongs to at most one path (see Figure 14.2). Thus, a Markov

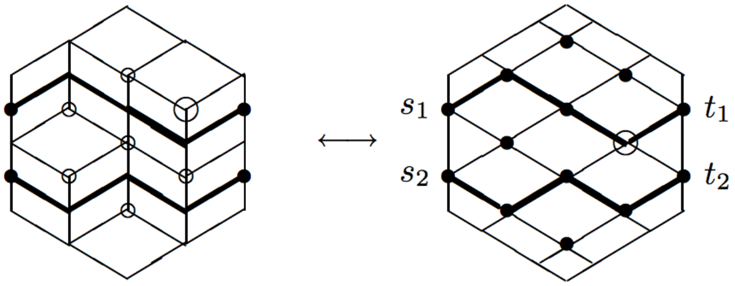

Figure 14.2: Lozenge tiling ≡ routing (reprint of Figure 8.3)

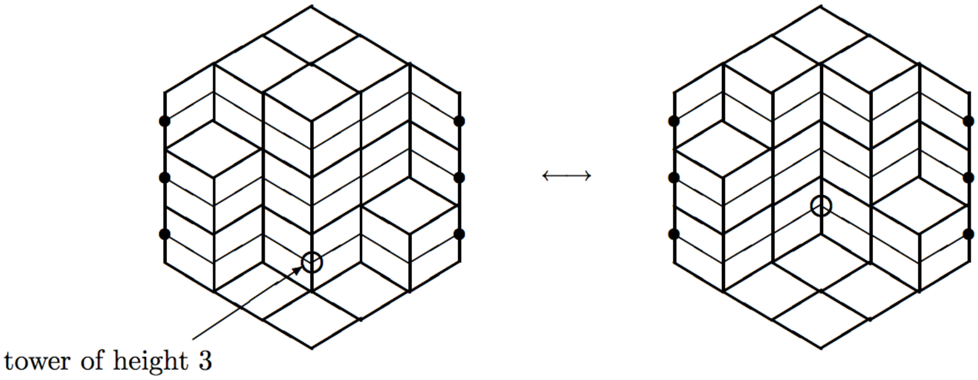

Figure 14.3: Tower rotations (reprint of Figure 8.5)

chain with a stationary distribution that is uniform on such routings can be used to produce uniform samples of lozenge tilings.

In Lecture 8, a "vanilla" Markov chain on multi-path routings was first presented: choose a point p along some path and a direction $d \in \{\uparrow, \downarrow\}$ uniformly at random, and, if possible, move p in direction d. $\uparrow/\downarrow$ moves are possible only for "valleys"/"peaks," and only when the move would not be blocked by another path. This Markov chain was shown to be irreducible, aperiodic and lazy, but its convergence time was difficult to analyze for multi-path routings due to the possibility of blocked moves.

Thus, a related Markov chain with "tower rotations" was introduced: choose a point p and direction d as before, and then rotate the *tower* at p in direction d if possible (i.e., if d is $\downarrow$ and p is a peak or d is $\uparrow$ and p is a valley) with probability $1/2h$, where h is the height of the tower. Because no moves are ever blocked by other paths in this Markov chain, we found it easier to analyze and proved its mixing time to be $O(n^4)$, where n is the total number of points in the lattice. (A more involved analysis by Wilson improves this to $O(n^3)$, which is tight.) However, the analysis of this modified chain told us nothing about the convergence time of the original vanilla Markov chain.

In this section, we will show how to use the technique introduced in Section 14.1.1 to bound the mixing of the vanilla Markov chain in terms of the mixing time of the modified chain with tower moves. The application is fairly straightforward: P is the vanilla chain and $\tilde{P}$ is the chain with tower moves.

To compare P with $\widetilde{P}$, we must specify, for each tower move, a way to "implement" the move by one or more sequences of simpler non-tower moves. In fact we will route all the flow for a tower move along a *single* path of non-tower moves in an obvious way: simply flip one level of the tower at a time, starting from the top in the case of a valley tower, and from the bottom in the case of a peak tower.

To analyze this flow, consider any edge e in P. This edge may lie on the paths corresponding to several tower moves; however, it should be clear that the number of tower moves of height h whose paths can pass through this edge e is at most h. Therefore we have

$$\begin{aligned} f(e) &\leq \sum_h h \times \frac{1}{2Nhn} && (14.1)\\ &\leq \frac{1}{2nN}\sum_h 1 && (14.2)\\ &\leq \frac{m}{2nN}, && (14.3) \end{aligned}$$

where m is the maximum height of any tower.

On the other hand, the capacity of the edge e is

$$C(e) = \frac{1}{N} \times \frac{1}{2n}. \tag{14.4}$$

Hence we have

$$\rho(f) = \max_e \frac{f(e)}{C(e)} \leq m.$$

Since $\ell(f)$ is also bounded by m, and m is bounded above (very crudely) by n, we may apply Theorem 14.4 to conclude that

$$\alpha \geq (\tilde{\alpha}) \times n^{-2} \tag{14.5}$$

Thus, the eigenvalue gap of the vanilla chain is at least an n^2 fraction of that of the enhanced chain. There is a small hitch: we proved the mixing time of the enhanced chain directly using coupling, and never proved a bound on its eigenvalue gap. Previously we showed that the mixing time of a chain is bounded by its eigenvalue gap (or Poincaré constant) via

$$\tau_{mix} \leq \frac{C}{\alpha}\log(\pi_{min}^{-1}).$$

We can also easily prove a bound in the opposite direction:

Claim 14.5 *$\alpha \geq \frac{C'}{\tau_{mix}}$ where C' is some other constant.*

Combining these facts, we can bound the mixing time of the vanilla chain:

$$\begin{aligned} \tau_{mix} &\leq \frac{C}{\alpha}\log(\pi_{min}^{-1})\\ &\leq \frac{Cn}{\alpha}\\ &\leq \frac{Cn^3}{1-\tilde{\alpha}}\\ &\leq C''\tilde{\tau}_{mix}n^3\\ &= \mathrm{O}(n^3)n^3\\ &= \mathrm{O}(n^6). \end{aligned}$$

Note that this analysis was rather loose, and may have introduced several unnecessary factors of n. (Indeed, it is conjectured that the true mixing time of this chain is $O(n^4)$, or even $\tilde{O}(n^3)$.) Notwithstanding, it provided us with a polynomial bound on the mixing time of the vanilla chain, which we were not able to prove directly using coupling.

14.2 Approximate counting

14.2.1 The class $\#P$

Definition 14.6 *$\#P$ is the class of functions $f : \Sigma^* \to \mathbb{N}$ such that there exists a polynomial-time nondeterministic Turing Machine M which, on input x, has exactly $f(x)$ accepting computation paths.*

This is a natural analog of the class NP of search or decision problems, with typical elements of $\#P$ being counting versions of NP search problems: the canonical example is $\#SAT$, which counts the number of satisfying assignments of a boolean formula in CNF. Most natural such problems are $\#P$-complete. Interestingly, this applies not only to counting versions of NP-complete problems, such as $\#SAT$ or $\#COL$ (counting graph colorings), but also to several important problems whose decision version is in P, such as $\#MATCHINGS$ (counting the number of matchings in a graph), or $\#DNFSAT$ (counting the number of satisfying assignments of a DNF formula).

Given this state of affairs, much attention has focused on the design of efficient *approximation algorithms* for problems in $\#P$. We make the following definition:

Definition 14.7 *Let f be a function in $\#P$. A* fully polynomial randomized approximation scheme (fpras) *for f is a randomized algorithm that on input (x, ε), where $x \in \Sigma^*$ and $\varepsilon \in (0, 1]$, outputs a random variable Z such that $\Pr[f(x)(1-\varepsilon) \le Z \le f(x)(1+\varepsilon)] \ge \frac{3}{4}$, and runs in time polynomial in x and ε^{-1}.*

The $\frac{3}{4}$ constant above is chosen for algebraic convenience only:

Claim 14.8 *If there exists an fpras for f, we can boost the confidence from $\frac{3}{4}$ to $1-\delta$ at the cost of a slowdown by a factor of $O(\log \delta^{-1})$.*

Proof: Take $t = O(\log \delta^{-1})$ independent trials of the the fpras and output the *median* of the results. The median falls outside $[f(x)(1-\varepsilon), f(x)(1+\varepsilon)]$ only if at least $t/2$ trials land outside that range, and the probability of this is less than the probability that a coin with $\frac{3}{4}$ probability of landing "heads" comes up heads less than $t/2$ times in t tosses. By a standard Chernoff bound, this probability is bounded by e^{-ct} for a constant c, and thus by δ when $t = O(\log \delta^{-1})$. ■

We have mentioned several times in this course that random sampling (obtained, e.g., by MCMC) can be used to perform approximate counting. In the next section we spell out this connection in more detail for a representative example, namely counting the colorings of a graph.

14.2.2 From sampling to approximate counting: graph coloring

Let us consider the problem of approximately counting q-colorings of a graph $G = (V, E)$. We will restrict our attention to the case $q \ge 2\Delta + 1$, where Δ is the maximum degree of the graph, for which we know that we have an MCMC algorithm to sample colorings uniformly at random with mixing time $O(n \log n)$.

Consider the sequence of graphs $G_0 = (V, E_0), G_1 = (V, E_1), \ldots, G_m = G$, where m=$|E|$, $E_0 = \emptyset$, and each step in the sequence adds one of E's edges, $E_i = E_{i-1} \cup \{e_i\}$.

Let $\Omega(G)$ be the set of all colorings of a graph G. We can represent $|\Omega(G)|$ by a telescoping product:

$$|\Omega(G)| = |\Omega(G_0)| \prod_{i=1}^{m} \frac{|\Omega(G_i)|}{|\Omega(G_{i-1})|}.$$

Since G_0 has no edges, $|\Omega(G_0)| = q^n$ (where $n = |V|$).

We estimate $|\Omega(G_i)| \,/\, |\Omega(G_{i-1})|$ by sampling colorings of G_{i-1} and outputting the proportion of samples which give different colors to the endpoints of e_i, which correspond precisely to the proper colorings of G_i.

To estimate the required sample size, consider the injective, multivalued map μ from the set $\Omega(G_{i-1})\backslash\Omega(G_i)$ of G_{i-1} colorings which aren't proper G_i colorings to the set of proper G_i colorings $\Omega(G_i)$. Define this map to take the lexicographically first endpoint u_i of e_i and recolor it with any color allowed for it in G_i. Since the degree of u_i in G_i is at most Δ, there are at least $q - \Delta \geq \Delta + 1$ such colorings for each coloring in $\Omega(G_{i-1})\backslash\Omega(G_i)$. The map is injective since the original coloring χ in $\Omega(G_{i-1})\backslash\Omega(G_i)$ can be obtained from a coloring $\chi' \in \mu(\chi)$ by changing u_i's color in χ' to the color of the other endpoint of e_i in χ'. We thus have (for $\Delta > 1$; the problem is trivial for graphs with $\Delta \leq 1$):

$$\begin{aligned} |\Omega(G_i)| &\geq (\Delta+1)\,|\Omega(G_{i-1})\backslash\Omega(G_i)| \\ &\geq (\Delta+1)(|\Omega(G_{i-1})| - |\Omega(G_i)|); \\ \frac{|\Omega(G_i)|}{|\Omega(G_{i-1})|} &\geq \frac{\Delta+1}{\Delta+2} \geq \frac{3}{4}. \end{aligned}$$

A standard variance analysis (via Chebyshev's inequality) shows that taking $t \approx \frac{16m^3}{3\varepsilon^2}$ ensures that each ratio in the telescoping product is within $1 \pm \frac{\varepsilon}{2m}$ of its true value with probability at least $1 - \frac{1}{4m}$. (This assumes that the MCMC yields an exactly uniform distribution over colorings, but the deviation from uniformity is of lower order and can easily be absorbed into the above bound.) Since there are m factors in the product, we see that $O(\frac{m^4}{\varepsilon^2})$ samples are sufficient to ensure that the product estimates $|\Omega(G)|$ within ratio $1 \pm \varepsilon$ with probability at least $\frac{3}{4}$ (the confidence follows by a union bound, and the accuracy from the fact that $(1 \pm \frac{\varepsilon}{2m})^m \in [1-\varepsilon, 1+\varepsilon]$). Since each sample takes $O(n \log n)$ time to generate, the overall running time of the algorithm is $O(m^4\varepsilon^{-2}n\log n)$, which is polynomial in n and ε^{-1}, as required for an fpras.

Exercise 14.9 *Verify the variance analysis in the previous paragraph.*

Exercise 14.10 *By analyzing the variance of the product estimator itself (rather than of each term of the product separately as we did above), show that the number of samples required can be reduced by a factor of m^2.*

References

[DS93] P. Diaconis and L. Saloff-Coste, "Comparison theorems for reversible Markov chains," *Annals of Applied Probability* 3 (1993), pp. 696–730.