Ph.D. course on Network Dynamics

Homework 2

To be discussed on Tuesday, October 15, 2013

Exercise 1 (Stochastic matrices are non-expansive in total variation distance)**.** *Let P be a stochastic matrix and μ, ν two probability vectors.*

(a) Prove that

$$||P'\nu - P'\mu||_{TV} \leq ||\nu - \mu||_{TV}$$

Now, assume that, for all i, j there exists k such that $P_{ik}P_{jk} > 0$.

(b) Prove that

$$||P'\nu - P'\mu||_{TV} < ||\nu - \mu||_{TV}$$

Exercise 2 (Lower bound on expected hitting time and Kac's formula)**.** *Let P be a stochastic matrix and $\pi = P'\pi$ a stationary distribution. Let $\mathcal{A} \subseteq \mathcal{V}$ be a subset of states such that $\pi(\mathcal{A}) > 0$, and let $T_\mathcal{A} := \inf\{t \geq 0 : V(t) \in \mathcal{A}\}$ and $T_\mathcal{A}^+ := \inf\{t \geq 1 : V(t) \in \mathcal{A}\}$ be, respectively, the hotting time and the return time on $\mathcal{A}$ for the Markov chain $V(t)$ with transition probability matrix P. Write $\mathbb{P}_\pi(\,\cdot\,) := \sum_{v\in\mathcal{V}} \pi_v \mathbb{P}(\,\cdot\,|V(0) = v)$ for the probability when $V(0)$ is distributed according to π and $\mathbb{E}_{\pi|_\mathcal{A}}(\,\cdot\,) := \pi(\mathcal{A})^{-1} \sum_{a\in\mathcal{A}} \pi_a \mathbb{E}[\,\cdot\,|V(0) = a]$ for the expectation when $V(0)$ is distributed according to $\pi|_\mathcal{A}$, i.e., π conditioned on $\mathcal{A}$.*

(a) Prove that

$$\mathbb{P}_\pi(T_\mathcal{A} < t) \leq t\pi(\mathcal{A})\,, \qquad \forall t \geq 0\,.$$

(hint: use the fact that $\{T_\mathcal{A} < t\} = \cup_{s=0}^{t-1}\{V(s) \in \mathcal{A}\}$ and stationarity)

*(b**) Prove Kac's formula*

$$\mathbb{E}_{\pi|_{\mathcal{A}}}[T^+_{\mathcal{A}}] = \frac{1}{\pi(\mathcal{A})}$$

(hint: observe that

$$\mathbb{P}_\pi(V(1) \notin \mathcal{A}, V(2) \notin \mathcal{A}, \ldots, V(t) \notin \mathcal{A}) = \mathbb{P}_\pi(V(0) \notin \mathcal{A}, V(1) \notin \mathcal{A}, \ldots, V(t-1) \notin \mathcal{A}),$$

for all $t \geq 1$, by stationarity, and use it to prove that

$$\mathbb{P}_\pi(T^+_{\mathcal{A}} = t) = \pi(\mathcal{A})\mathbb{P}_{\pi|_{\mathcal{A}}}(T^+_{\mathcal{A}} \geq t),$$

then sum both sides of the above over all $t \geq 1$.)

Exercise 3 (Hitting times and probabilities on the line)**.** *Consider the simple random walk $V(t)$ on the line graph with nodes $\{0, 1, \ldots, n-1, n\}$. For $0 \leq k \leq n$, let $\tau_k := \mathbb{E}[T_{\{0,n\}}]$ and $p_k = \mathbb{P}(T_n < T_k)$ be respectively the expected hitting time on the set $\{0, n\}$, and the probability of hitting node n before node 0, conditioned on starting from node k.*

(a) Prove that $\tau_k = k(n-k)$ for all $0 \leq k \leq n$;

(b) Prove that $p_k = k/n$ for all $0 \leq k \leq n$.

Exercise 4 (Coupon collector lemma)**.** *Consider the following stochastic process. Let $X_1, X_2, \ldots$ be a sequence of independent random variables, with identical uniform distribution over a finite set $\{1, \ldots, n\}$. For all $t \geq 0$, let $Z_t := |\{X_1, \ldots, X_t\}|$ be number of distinct realizations up to time t, and let $T_n := \inf\{t \geq 0 : Z_t = n\}$ be the first time t that every value in $\{1, \ldots, n\}$ has occurred at least once in the t-tuple $X_1, \ldots, X_t$. This may be thought of as modeling a company issuing n different different types of coupons and of a collector buying every day a random coupon. How long should he/she wait to get all the coupons?*

(a) Prove that

$$n \log n \leq \mathbb{E}[T_n] \leq n(\log n + 1)$$

(hint: conditioned on having collected k coupons, the time to wait before getting a new coupon has geometric distribution with expected value $\frac{n}{n-k}$. Then, use the fact (why?) that $\log n \leq \sum_{k=1}^n 1/k \leq \log n + 1$)

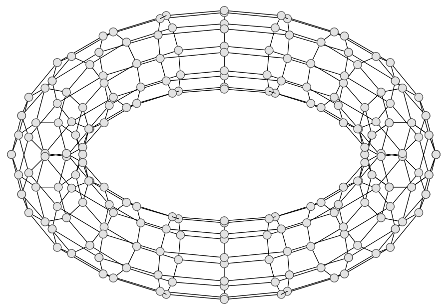

Figure 1: 2-dimensional torus.

(b) Prove that , for all* $c > 0$*,*

$$\mathbb{P}(T_n > \lceil n \log n + cn \rceil) \leq e^{-c} .$$

(hint: let A_i *be the event that coupon* i *has not shown up among the first* $\lceil n \log n + cn \rceil$ *coupons bought. Then, prove that* $\mathbb{P}(T_n > \lceil n \log n + cn \rceil) \leq \sum_{i=1}^{n} \mathbb{P}(A_i)$*, and that* $\mathbb{P}(A_i) \leq e^{-c}/n$ *for all* $1 \leq i \leq n$*)*

Exercise 5 (mixing time on the d-dimensional torus)**.** *The* d*-dimensional torus of size* $n = m^d$*, where* m *and* d *are positive integers, is the graph* $\mathcal{G}$ *with node set* $\mathcal{V} = \{0, 1, \ldots, n-1\}^d$*, and where two nodes* $u, v \in \mathcal{V}$ *are linked to each other if and only if there exists* $1 \leq i \leq k$ *such that* $u_i - v_i \in \{-1, 1\}$ *(where the difference is modulo* n*) and* $u_k = v_k$ *for all* $k \in \{1, \ldots, d\} \setminus \{i\}$*. For* $d = 2$ *the graph is plotted in Fig. 1. Let* P *be the stochastic matrix associated of the lazy random walk on* $\mathcal{G}$*.*

(a) Construct a coupling and use it to prove that the mixing time of P *satisfies*

$$\tau_{\text{mix}} \leq d^2 n^{2/d}$$

(hint: build upon the coupling constructed in class: at each t*, chose* $1 \leq i \leq d$ *uniformly and let* V_1 V_2 *move along component* i *...)*

(b) Prove a lower bound on τ_{mix} *using the conductance bound.*

Exercise 6 (mixing time of Google's PageRank algorithm)**.** *Let* $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ *be the directed graph describing the WWW, whose nodes* $v \in \mathcal{V}$ *correspond to webpages and where there is a directed edge* $(u, v) \in \mathcal{E}$ *whenever page* u *has a hyperlink directed to page* v*. Let* $d_u := |\mathcal{E}_u|$ *and* $\mathcal{E}_u := \{v : (u, v) \in \mathcal{E}\}$ *be the number of hyperlinks and, respectively, the set of linked pages, from page* u*. Define a stochastic matrix* Q *by* $Q_{uv} = 1/n$ *for all* v *if* $d_u = 0$*, and, if* $d_u \geq 1$*, let* $Q_{uv} = 0$ *if* $(u, v) \notin \mathcal{E}$ *and* $Q_{uv} = 1/d_u$ *if* $(u, v) \in \mathcal{E}$*. Also, let* μ *be the uniform distribution over the set of webpages, and* $\beta \in (0, 1)$*.*

(a) Prove that the equation $\pi = (1 - \beta)Q'\pi + \beta\mu$ *has a unique solution* π *which is a probability vector; (hint: it is sufficient to show that the matrix* $W := (I - (1 - \beta)Q')$ *is strictly diagonally dominant, hence nonsingular, so that* $\pi = \beta W^{-1}\mu$ *is the unique solution)*

In fact, π *is the PageRank vector, first introduced by Brin and Page [1] to measure the relative importance of webpages. Typical values of* β *used in practice are about* 0.15*. For general probability distribution* μ*, the vector* π *is referred to as the personalized PageRank [2], and is used in context-sensitive searches. Consider the irreducible stochastic matrix*

$$P := (1 - \beta)Q + \beta\mathbb{1}\mu' \,,$$

and observe that π *is its stationary distribution. In fact, the PageRank vector can be interpreted as the stationary distribution of a random walk on the directed graph* $\mathcal{G}$*, which, at each time* t*, is restarted with probability* β *from a random webpage chosen uniformly from* $\mathcal{V}$*.*

(b) Construct a Markov coupling for P*, and use it to prove that its mixing time satisfies*

$$\tau_{\text{mix}} \leq \left\lceil \frac{-1}{\log(1 - \beta)} \right\rceil \leq \frac{1}{\beta} + 1 \,.$$

References

[1] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the 7th International World Wide Web Conference*, pages 107–117, 1998.

[2] H. Haveliwala. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE Transactions on Knowledge Data Engineering*, 15(4):784–796, 2003.