Fast mixing for Latent Dirichlet Allocation

Johan Jonasson*†

November 30, 2017

Abstract

Markov chain Monte Carlo (MCMC) algorithms are ubiquitous in probability theory in general and in machine learning in particular. A Markov chain is devised so that its stationary distribution is some probability distribution of interest. Then one samples from the given distribution by running the Markov chain for a "long time" until it appears to be stationary and then collects the sample. However these chains are often very complex and there are no theoretical guarantees that stationarity is actually reached. In this paper we study the Gibbs sampler of the posterior distribution of a very simple case of Latent Dirichlet Allocation, the arguably most well known Bayesian unsupervised learning model for text generation and text classification. It is shown that when the corpus consists of two long documents of equal length m and the vocabulary consists of only two different words, the mixing time is at most of order $m^2 \log m$ (which corresponds to $m \log m$ rounds over the corpus). It will be apparent from our analysis that it seems very likely that the mixing time is not much worse in the more relevant case when the number of documents and the size of the vocabulary are also large as long as each word is represented a large number in each document, even though the computations involved may be intractable.

AMS Subject classification : 60J10
Key words and phrases: mixing time, MCMC, Gibbs sampler, conductivity, topic model
Short title: Fast mixing for LDA

*Chalmers University of Technology and University of Gothenburg
†Research supported by the Knut and Alice Wallenberg Foundation, grant no. KAW 2012.0067

1 Introduction

Markov chain Monte Carlo (MCMC) is a powerful tool for sampling from a given probability distribution on a very large state space, where direct sampling is difficult, in part because of the size of the state space and in part because of normalizing constants that are difficult to compute.

In machine learning in particular, MCMC algorithms are extremely common for sampling from posterior distributions of Bayesian probabilistic models. The posterior distribution given observed data is difficult to sample from for the given reasons. One may then design an (irreducible aperiodic) Markov chain whose stationary distribution is precisely the targeted posterior. This is usually fairly easy since the posterior is usually easy to compute up to the normalizing constant (the denominator in Bayes formula). One very often uses Gibbs sampling or the related Metropolis-Hastings algorithm.

Gibbs sampling in general can be described as follows. The state space is a finite set of random variables $X = \{X_a\}_{a \in A}$, where $X_a \in T$ for some measurable space T, so that $X \in T^A$ and the targeted distribution is a given probability measure $\mathbb{P}$ on T^A. In order to sample from $\mathbb{P}$ one starts a Markov chain on T^A whose updates are given by first choosing an index $a \in A$ at random and then choosing a new value of X_a according to the conditional distribution of X_a given all X_b, $b \in A \setminus \{a\}$. Under mild conditions, this Markov chain converges in distribution to $\mathbb{P}$. The chain is then run for a "long time" (the "burn-in") whereupon a sample, hopefully approximately from $\mathbb{P}$, is collected. A key question here is for how long the chain actually has to be run, in order for the distribution after that time to be a good approximation of $\mathbb{P}$. Since A is usually large, the number of steps needed should at least be no more than polynomial in the size of A for Gibbs sampling to be feasible. In almost all practical cases, the structure of the sample space and the probability measure $\mathbb{P}$ is so complex that is virtually impossible to make a rigorous analysis of the mixing rate. However it may be possible to consider some very simplified special cases. In this paper, we will analyse a special case of Latent Dirichlet allocation, henceforth LDA for short, and demonstrate for such a simple special case, the mixing time is polynomial of low degree in the size of the problem.

LDA is a model used to classify documents according to their topics, which introduced by Blei et al [3]. One is faced with a large corpus of documents and wants to determine for each word in each document which topic it belongs to. Knowing this, one can then also classify the documents according to the proportion of words of the different topics it contains. The setup in LDA is the following. The corpus

consists of a fixed number D of documents of lengths N_d, $d = 1, \ldots, D$, a fixed set of topics $t_1, t_2, \ldots, t_s$ and a fixed set of distinct words $w_1, w_2, \ldots, w_v$. These are specified in advance. The number of topics is usually not large, whereas the number of distinct words is. Next, for each document $d = 1, \ldots, D$, independent multinomial distributions $\theta_d = (\theta_d(1), \ldots, \theta_d(s))$ over topics is chosen according to a Dirichlet prior with a known parameter $\alpha = (\alpha_1, \ldots, \alpha_s)$. For each topic t_i, a multinomial distribution $\phi_i = (\phi_i(1), \ldots, \phi_i(v))$ according to a Dirichlet prior with known parameter $\beta = (\beta_1, \ldots, \beta_v)$ independently of each other and of the θ_d:s. Given these, the corpus is then generated by for each position $p = 1, \ldots, N_d$ in each document d, picking a topic z_{dp} according to θ_d and then picking the word w_{dp} at that position according to $\phi_{z_{d,p}}$, doing this independently for all positions. Note that the model is a so called "bag of words" model, i.e. it is invariant under permutations within each document.

In this paper, the very simple special case with $D = s = v = 2$, $N_1 = N_2 = m$ and $\alpha = \beta = (1, 1)$ will be studied with respect to the mixing time asymptotics as $m \to \infty$. We will find that the corresponding Gibbs sampler indeed does mix fast (contrary to what was erroneously claimed in an earlier version of this paper). To simplify the notation, denote the two topics by A and B and the two words by 1 and 2. Define n_{ij} as the number of occurrences of the word j in document i, $i, j = 1, 2$ and write $n_{i\cdot} = n_{i1} + n_{i2}$ (which by assumption equals m in the case of study) and $n_{\cdot j} = n_{1j} + n_{2j}$ and $n_{\cdot\cdot} = \sum_{i,j} n_{ij} = 2m$. We consider the mixing time for Gibbs sampling of the posterior in a seemingly typical case, namely that the number of 1:s in the first document is $3m/10$ and in the second document $6m/10$. (Of course the precise numbers here are of no great significance as long as they are of order m.) Let $R = \{R_t\}_{t=0}^{\infty}$ denote the corresponding Markov chain, whose state space is $\{A, B\}^{2m}$ and let π_R denote the stationary distribution of $\{R_t\}$, i.e. the targeted posterior.

Before stating the main result, Theorem 1.3 below, the concept of mixing time needs to be introduced. Let $X = \{X_t\}_{t=0}^{\infty}$ be a discrete time aperiodic irreducible Markov chain on the finite state space S with transition matrix $[p(x, y)]_{x,y \in S}$. Let $\mathbb{P}_x$ be the underlying probability measure under $X_0 = x$ and let π be the stationary distribution.

Definition 1.1 *Let μ and ν be two probability measures on S. The the total variation distance between μ and ν is given by*

$$\|\mu - \nu\|_{TV} = \max_{A \subset S} (\mu(A) - \nu(A)) = \frac{1}{2} \sum_{x \in S} |\mu(x) - \nu(x)|.$$

Definition 1.2 *For each $\kappa \in (0,1)$ and $x \in S$, the (κ, x)-mixing time of X is given by*

$$\tau_{\mathrm{mix}}(\kappa, x) = \tau^X_{\mathrm{mix}}(\kappa, x) = \min\{t : \|\mathbb{P}_x(X_t \in \cdot) - \pi\|_{TV} \leq \kappa\}.$$

Theorem 1.3 *Consider the case $n_{11} = 3m/10$ and $n_{21} = 6m/10$. Then there is a set $B \subseteq S$ with $\pi_R(B) = 1 - m^{-10}$ such that for each $x \in B$ and each $\kappa = \Omega(1/m)$ there is a $C(\kappa) < \infty$ such that*

$$\tau^R_{\mathrm{mix}}(\kappa, x) \leq C(\kappa) m^2 \log m$$

for every $x \in B$.

The proof of Theorem 1.3 relies heavily on the *conductance* of a Markov chain.

Definition 1.4

- *For $A \subseteq S$, the conductance is given as*

$$Q(A, A^c) = \sum_{x \in A} \sum_{y \in A^c} \pi(x) p(x, y).$$

- *The conductance profile of $\{X_t\}$ is the decreasing function $\Phi : (0, \infty) \to [0, 1]$ given as follows. For $A \subset S$, let $\Phi_A = Q(A, A^c)/\pi(A)$ and for $r \leq 1/2$, set*

$$\Phi(r) = \min\{\Phi_A : \pi(A) \leq r\}$$

and $\Phi(r) = \Phi(1/2)$ for $r > 1/2$.

The following is a consequence of Theorem 1 of [6].

Theorem 1.5 *Let $\pi_* = \min\{\pi(x) : x \in S\}$. Then for any $x \in S$,*

$$\tau_{\mathrm{mix}}(\kappa, x) \leq 1 + \int_{\pi_*}^{4/\kappa} \frac{4}{r \Phi(r)^2} dr.$$

Theorem 1.5 is a refinement of earlier results on conductance bounds on mixing times. Write $\hat{\Phi} = \Phi(1/2)$. The following is from [9, 8].

Theorem 1.6 *Assume that X is reversible. Then*

$$\tau_{\mathrm{mix}}(\kappa, x) \leq \hat{\Phi}^{-2} \left(\log(\pi(x)^{-1}) + \log(\kappa^{-1}) \right).$$

Note that any Gibbs sampler as defined above is reversible.

A powerful method for lower bounding conductance is by *canonical paths*, see [5, 7]. Any reversible Markov chain can be seen as a weighted graph $\mathcal{G}$ by assigning weights $p(x, y)$ to each $e = \{x, y\}$ with $p(x, y) > 0$. For each pair of states x, y, choose a weighted path γ_{xy} in $\mathcal{G}$ from x to y. Even though the direction of the path is not important, it will later on be convenient in the arguments to keep the direction in mind. Write Γ for the collection $\{\gamma_{xy} : x, y \in S\}$. For each $e = \{x, y\}$, let $Q(e) = \pi(x)p(x, y)$ (which equals $\pi(y)p(y, x)$ under reversibility). The idea is to choose these paths in a way that makes the *path congestion* small.

Definition 1.7 *The path congestion of* Γ *is given by*

$$\rho(\Gamma) = \max_e \frac{1}{Q(e)} \sum_{x,y:e\in\gamma_{x,y}} \pi(x)\pi(y).$$

The key result relating path congestion to conductance is the following

Lemma 1.8 *For any collection* Γ *of canonical paths,*

$$\hat{\Phi} \geq \frac{1}{2\rho(\Gamma)}.$$

To finish off the proof, the classical method of *coupling* will be needed. Let Y be a Markov chain with the same transition matrix as X such that Y_0 is distributed according to π. Assume that the transitions of X and Y are made dependent, or *coupled*, in such a way that whenever $X_t = Y_t$ and $r > t$, then $X_r = Y_r$. Then the *coupling inequality* states the following.

Lemma 1.9 *Let* T *be the coupling time of* X *and* Y*, i.e.* $T = \min\{t : X_t = Y_t\}$ *and assume that* $X_0 = x$*. Then*

$$\|\mathbb{P}(X_t \in \cdot) - \pi\|_{TV} \leq \mathbb{P}(T > t).$$

The rest of the paper is devoted to the proof of Theorem 1.3.

Remarks on notation. Let $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$ be two sequences of positive numbers. We say that a_n is at most of order b_n and write $a_n = O(b_n)$ if there is a constant $C < \infty$ such that $a_n < Cb_n$ for all n. We say that a_n is at least of order b_n and write $a_n = \Omega(b_n)$ if $b_n = O(a_n)$. When both $a_n = O(a_n)$ and $a_n = \Omega(b_n)$, we say that a_n is of order b_n and write $a_n = \Theta(b_n)$. When $a_n/b_n \to 0$, we say that a_n is of smaller order than b_n and write $a_n = o(b_n)$.

We will also use the convenient shorthand notation $[n]$ for the set $\{1, 2, \ldots, n\}$.

Other remarks. Today there is a plethora of extensions of the LDA model; e.g. Andrews and Vigliocco [1] consider a hidden Markov model for the topics, Wallach [11] studies a hidden Markov model for the words and Gruber et el. [4] consider a sentence based model. A good summary of the literature can be found in the introduction of Tian et el. [10], who also study a sentence based model. For an introduction to probabilistic topic models, see [2].

2 Proof of Theorem 1.3

The following two lemmas will be needed. The first one is a special case of a well known fact about moments of beta distributions (and more generally of Dirichlet distributions). The second one will only be used with $K = 2$, but the general case comes at no extra cost.

Lemma 2.1 *Let X be a standard uniform random variable and $0 \leq k \leq n$. Then*

$$\mathbb{E}[X^k(1-X)^{n-k}] = \frac{1}{(n+1)\binom{n}{k}}.$$

Lemma 2.2 *For any nonnegative integers $a_{11}, a_{12}, \ldots, a_{1j}, a_{21}, a_{22}, \ldots, a_{2K}$,*

$$\binom{\sum_{i,j} a_{ij}}{\sum_j a_{1j}} \binom{\sum_j a_{1j}}{a_{11}, \ldots, a_{1j}} \binom{\sum_j a_{2j}}{a_{21}, \ldots, a_{2j}}$$

$$= \binom{\sum_{i,j} a_{ij}}{a_{11} + a_{21}, \ldots, a_{1K} + a_{2K}} \prod_j \binom{a_{1j} + a_{2j}}{a_{1j}},$$

where i ranges over $[2]$ and j over $[K]$.

Proof. Both sides of the equality are equal to the multinomial coefficient

$$\binom{\sum_{i,j} a_{ij}}{a_{11}, \ldots, a_{2K}}.$$

□

Let $\mathbf{W} = (w_{11}, w_{12}, \ldots, w_{1m}, w_{21}, \ldots, w_{2m})$ be the words in our corpus, let $\mathbf{Z} = (z_{11}, z_{12}, \ldots, z_{2m})$ be the latent topics and $\mathbf{Z}_d$ be the latent topics in document

d. Let also θ_d be the probability that $z_{d1} = A$, $d = 1, 2$ and let ϕ_t be the conditional probability that $w_{dj} = 1$ given that $z_{dj} = t$, $t = A, B$. In the case under study, these four quantities are all independent standard uniform random variables. We begin by determining the posterior distribution π_R up to a normalizing constant.

Define k_{dj} to be the number words in document d with the topic being A and the word being j and the same dot notation for the k:s as for the n:s.

Proposition 2.3 *The posterior π_R of LDA with two documents, two unique words, two topics and uniform priors is given by*

$$\pi_R(\mathbf{z}) = C \frac{\binom{n_{..}}{k_{..}}}{(k_{..}+1)(n_{..}-k_{..}+1)\binom{n_{1.}}{k_{1.}}\binom{n_{2.}}{k_{2.}}\binom{n_{.1}}{k_{.1}}\binom{n_{.2}}{k_{.2}}},$$

where C is a normalizing constant.

Proof.
By Bayes formula

$$\pi_R(\mathbf{z}) = \mathbb{P}(\mathbf{Z} = \mathbf{z}|\mathbf{W} = \mathbf{w}) \propto \mathbb{P}(\mathbf{W} = \mathbf{w}|\mathbf{Z} = \mathbf{z})\mathbb{P}(\mathbf{Z} = \mathbf{z}).$$

Now

$$\begin{aligned}
\mathbb{P}(\mathbf{Z} = \mathbf{z}) &= \mathbb{E}[\mathbb{P}(\mathbf{Z} = \mathbf{z}|\theta_1, \theta_2)] \\
&= \mathbb{E}[\theta_1^{k_{1\cdot}}(1-\theta_1)^{n_{1\cdot}-k_{1\cdot}}]\mathbb{E}[\theta_2^{k_{2\cdot}}(1-\theta_2)^{n_{2\cdot}-k_{2\cdot}}] \\
&= \frac{1}{(n_{1\cdot}+1)(n_{2\cdot}+1)\binom{n_{1\cdot}}{k_{1\cdot}}\binom{n_{2\cdot}}{k_{2\cdot}}}.
\end{aligned}$$

where the last equality follows from Lemma 2.1. For the second factor we have analogously, again using Lemma 2.1,

$$\begin{aligned}
\mathbb{P}(\mathbf{W} = \mathbf{w}|\mathbf{Z} = \mathbf{z}) &= \mathbb{E}[\phi_A^{k_{\cdot 1}}(1-\phi_A)^{k_{..}-k_{\cdot 1}}]\mathbb{E}[\phi_B^{n_{\cdot 1}-k_{\cdot 1}}(1-\phi_B)^{n_{..}-(n_{\cdot 1}-k_{\cdot 1})}] \\
&= \frac{1}{(k_{..}+1)(n_{..}-k_{..}+1)\binom{k_{..}}{k_{\cdot 1}}\binom{n_{..}-k_{..}}{n_{\cdot 1}-k_{\cdot 1}}}.
\end{aligned}$$

Hence, ignoring factors that do not depend on the k:s, using Lemma 2.2 with $K = 2$, $a_{1j} = k_{\cdot j}$ and $a_{2j} = n_{\cdot j} - k_{\cdot j}$ for the second equality and again ignoring a

factor that does not depend on the k:s,

$$\pi_R(\mathbf{z}) \propto \left((k_{\cdot\cdot}+1)(n_{\cdot\cdot}-k_{\cdot\cdot}+1)\binom{n_{1\cdot}}{k_{1\cdot}}\binom{n_{2\cdot}}{k_{2\cdot}}\binom{k_{\cdot\cdot}}{k_{\cdot 1}}\binom{n_{\cdot\cdot}-k_{\cdot\cdot}}{n_{\cdot 1}-k_{\cdot 1}}\right)^{-1}$$
$$= \frac{\binom{n_{\cdot\cdot}}{k_{\cdot\cdot}}}{(k_{\cdot\cdot}+1)(n_{\cdot\cdot}-k_{\cdot\cdot}+1)\binom{n_{1\cdot}}{k_{1\cdot}}\binom{n_{2\cdot}}{k_{2\cdot}}\binom{n_{\cdot 1}}{k_{\cdot 1}}\binom{n_{\cdot 2}}{k_{\cdot 2}}}.$$

□

The expression for π_R only depends on $\mathbf{z}$ via $\mathbf{k} = \mathbf{k}(\mathbf{z}) := (k_{11}, k_{12}, k_{21}, k_{22})$. Identifying all $\mathbf{z}$ having the same $\mathbf{k}(\mathbf{z})$, we have (regarding,with some abuse of notation, a $\mathbf{k}$ also as the equivalence class consisting of all $\mathbf{z}$:s having that particular k_{dj}:s) that for any $\mathbf{k}_1$ and $\mathbf{k}_2$, all $\mathbf{z} \in \mathbf{k}_1$ have the same probability of transitioning into $\mathbf{k}_2$. Hence the process where we only record the $\mathbf{k}$:s is a lumped Markov chain, whose state space is $[n_{11}]\times[n_{12}]\times[n_{21}]\times[n_{22}]$. Denote this lumped Markov chain by $L = \{L_t\}_{t=0}^{\infty}$. By summing $\pi_R(\mathbf{z})$ over the $\mathbf{z}$:s having the same value of $\mathbf{k}(z)$, gives the following.

Proposition 2.4 *The lumped Gibbs sampler L has the stationary distribution*

$$\pi_L(\mathbf{k}) = \mathbb{P}(\mathbf{K}=\mathbf{k}|\mathbf{W}=\mathbf{w}) = C\frac{\binom{n_{\cdot\cdot}}{k_{\cdot\cdot}}\binom{n_{11}}{k_{11}}\binom{n_{12}}{k_{12}}\binom{n_{21}}{k_{21}}\binom{n_{22}}{k_{22}}}{(k_{\cdot\cdot}+1)(n_{\cdot\cdot}-k_{\cdot\cdot}+1)\binom{n_{1\cdot}}{k_{1\cdot}}\binom{n_{2\cdot}}{k_{2\cdot}}\binom{n_{\cdot 1}}{k_{\cdot 1}}\binom{n_{\cdot 2}}{k_{\cdot 2}}}$$

where C is a normalizing constant and $\mathbf{K} = \mathbf{k}(Z)$.

The most effort will go into analyzing the mixing time of L.

Define

$$h(x) = \left(x^x(1-x)^{1-x}\right)^{-1},\ x \in (0,1).$$

Then, for $x \in [0,1]$, by Stirling's formula,

$$\binom{m}{xm} = C\frac{h(x)^m}{\sqrt{m}\sqrt{(x+1/m)(1-x+1/m))}},$$

where C is of constant order. Hence, by cancelling factors independent of (a,b,c,d) and writing $s(x) = \sqrt{(x+1/m)(1-x+1/m)}$, we get

$$\pi_L(am,bm,cm,dm) = CG(a,b,c,d)^m\frac{s(\frac{10(a+c)}{9})s(\frac{10(b+d)}{11})s(a+b)s(c+d)}{s(\frac{a+b+c+d}{2})^3 s(\frac{10a}{3})s(\frac{10b}{7})s(\frac{5c}{3})s(\frac{5d}{2})},$$

$(a,b,c,d) \in [0,3/10] \times [0,7/10] \times [0,3/5] \times [0,2/5]$, where C is of constant order and

$$G(a,b,c,d) = \left(\frac{h(\frac{(b+d-a-c)}{2})^2 h(\frac{10a}{3})^{3/10} h(\frac{10b}{7})^{7/10} h(\frac{5c}{3})^{3/5} h(\frac{5d}{2})^{2/5}}{h(\frac{10(a+c)}{9})^{9/10} h(\frac{10(b+d)}{11})^{11/10} h(a+b) h(c+d)} \right)^m . \quad (1)$$

We will now analyze the function G. Define $g := \log G$. Intuitively in order to get $3m/10$ ones in the first document and $3m/5$ ones in the second document, one has certain restrictions to the ϕ's and θ's. Working this out, one should expect to have (a,b,c,d) very close to $\mathbf{k}(u,v) = (k_1(u,v), k_2(u,v), k_3(u,v), k_4(u,v))$, $u \in [0,3/10]$, $v \in [3/5,1]$, where

$$k_1(u,v) = \frac{u(v-3/10)}{v-u} \qquad k_2(u,v) = \frac{(1-u)(v-3/10)}{v-u} \quad (2)$$

$$k_3(u,v) = \frac{u(v-3/5)}{v-u} \qquad k_4(u,v) = \frac{(1-u)(v-3/5)}{v-u}, \quad (3)$$

i.e. G should be maximal around the surface

$$\Pi := \{\mathbf{k}(u,v) : (u,v) \in [0,3/10] \times [3/5,1]\}$$

in $[0,3/10] \times [0,7/10] \times [0,3/5] \times [0,2/5]$. To make this more precise, we first claim the following.

Proposition 2.5 *The function G as defined above is constant on Π.*

Proof. The claim is equivalent to saying that all four partial derivatives of g are zero at each $(a,b,c,d) \in \Pi$. It is helpful to spell out g:

$$\begin{aligned} g(a,b,c,d) = {} & 2 \log h \left(\frac{a+b+c+d}{2} \right) \\ & + \frac{3}{10} \log h \left(\frac{10}{3} a \right) + \frac{7}{10} \log h \left(\frac{10}{7} b \right) \\ & + \frac{3}{5} \log h \left(\frac{5}{3} c \right) + \frac{5}{2} \log h \left(\frac{2}{5} d \right) \\ & - \frac{9}{10} \log h \left(\frac{10(a+c)}{9} \right) - \frac{11}{10} \log h \left(\frac{10(b+d)}{11} \right) \\ & - \log h(a+b) - \log h(c+d). \end{aligned} \quad (4)$$

Hence, on observing that $\frac{d}{dx}h(x) = \log((1-x)/x)$,

$$g'_a(a,b,c,d) = \log \frac{(1-\frac{a+b+c+d}{2})(1-\frac{10a}{3})(\frac{10(a+c)}{9})(a+b)}{(\frac{a+b+c+d}{2})(\frac{10a}{3})(1-\frac{10(a+c)}{9})(1-a-b)} \tag{5}$$

For $(a,b,c,d) \in \Pi$, substitute with (2) and (3) and get, after some algebra

$$g'_a(a,b,c,d) = \log \frac{(\frac{9}{20}-u)(v-\frac{10}{3}uv)(\frac{20}{9}uv-u)(v-\frac{3}{10})}{(v-\frac{9}{20})(\frac{10}{3}uv-u)(v-\frac{20}{9}uv)(\frac{3}{10}-u)}$$

which after some further algebra is seen to be 0 for all (u,v). Analogously, one finds that the three other partial derivatives vanish on Π. Since g is clearly differentiable in the interior of its domain, this shows that g is indeed constant on Π. □

Next we study how g behaves as one moves away from Π. In a vicinity of Π, this is up to small order terms captured by the nonzero eigenvalues of the Hessian of g on Π. Write

$$H(u,v) = g''(\mathbf{k}(u,v)),\ 0 < u < \frac{3}{10},\ \frac{3}{5} < v < 1$$

for the Hessian of g at $\mathbf{k}(u,v)$, Since g is constant on Π, it follows that two of the four eigenvalues of $H(u,v)$ are 0. Write $\lambda_1(u,v)$ and $\lambda_2(u,v)$ for the other two eigenvalues. These are the two roots of a second degree polynomial with coefficients depending on (u,v), i.e.

$$\lambda_1(u,v) = \frac{\alpha_0(u,v) - \sqrt{\beta_0(u,v)}}{\delta(u,v)} \tag{6}$$

$$\lambda_2(u,v) = \frac{\alpha_0(u,v) + \sqrt{\beta_0(u,v)}}{\delta(u,v)}, \tag{7}$$

where α_0, β_0 and δ are polynomials in (u,v). The precise expressions of these can be found and analyzed using your favorite mathematical software; we used a combination of Matlab and Maple. It turns out that

$$\delta(u,v) = u(\frac{3}{10}-u)(\frac{9}{20}-u)(v-\frac{9}{20})(v-\frac{3}{5})(1-v).$$

The expressions for α_0 and β_0 are quite extensive and can be found in an appendix. As can be seen there, α_0^2 and β_0 have a factor $(v-u)^4$ in common, and it will be

more convenient to work with $\alpha(u,v) = -10^{-6}\alpha_0(u,v)/(v-u)^2$ and $\beta(u,v) = 10^{-12}\beta_0(u,v)/(v-u)^4$, where the constants are chosen so that the leading terms of α and β have coefficient 1.

Proposition 2.6 *The nonzero eigenvalues $\lambda_1(u,v)$ and $\lambda_2(u,v)$ of $H(u,v)$ are, as functions of (u,v), continuous, negative and bounded away from 0.*

Proof. It is easy to convince oneself of the truth of Proposition 2.6 using some mathematical software. Indeed it seems that $\max(\lambda_1(u,v), \lambda_2(u,v)) < -3$. However a proper proof requires more work. We start with some claims about the limit behavior when approaching the points at which $\delta(u,v)$ vanishes, which are by the expression for δ all points on the boundary of the domain. For the proof of the proposition, it suffices to prove that $\lambda_i(u,v)$ are negative and stay bounded away from 0 as one approaches the boundary, Our claims go a bit further, since this comes with no extra cost in the proof. They are as follows. As $(u,v) \to (0,v_0)$ for some $v_0 \in (3/5,1)$,

$$\lambda_1(u,v) = (1+o(u))\,\frac{C(v_0)}{u},\ \ \lambda_2(u,v) \to C(v_0)$$

where C is a generic notation (i.e. not the same function in each instance) for a continuous negative function. Similarly, as $u \to 3/10$ and $v \to v_0 \in (3/5,1)$,

$$\lambda_1(u,v) = \left(1+o\left(\frac{3}{10}-u\right)\right)\frac{C(v_0)}{\frac{3}{10}-u},\ \ \lambda_2(u,v) \to C(v_0).$$

Analogously

$$\lambda_1(u,v) = \left(1+o\left(v-\frac{3}{5}\right)\right)\frac{C(u_0)}{v-\frac{3}{5}},\ \ \lambda_2(u,v) \to C(v_0)$$

and

$$\lambda_1(u,v) = (1+o(1-v))\frac{C(u_0)}{1-v},\ \ \lambda_2(u,v) \to C(v_0)$$

as $v \to 3/5$ and $v \to 1$ respectively and $u \to u_0 \in (0,3/10)$.

When approaching a corner of the domain, the following holds as $(\mu,\nu) \downarrow$

$(0,0)$,

$$\lambda_1\left(\mu, \frac{3}{5}+\nu\right) = (1+o(\mu\nu))\frac{C}{\mu\nu},$$
$$\lambda_2\left(\mu, \frac{3}{5}+\nu\right) \to C,$$
$$\lambda_1(\mu, 1-\nu) = (1+o(\mu\nu))\frac{C}{\max(\mu,\nu)},$$
$$\lambda_2(\mu, 1-\nu) = (1+o(\mu\nu))\frac{C}{\min(\mu,\nu)},$$
$$\lambda_1\left(\frac{3}{10}-\mu, \frac{3}{5}+\nu\right) = (1+o(\mu\nu))\frac{C}{\max(\mu,\nu)},$$
$$\lambda_2\left(\frac{3}{10}-\mu, \frac{3}{5}+\nu\right) = (1+o(\mu\nu))\frac{C}{\min(\mu,\nu)},$$
$$\lambda_1\left(\frac{3}{10}-\mu, 1-\nu\right) = (1+o(\mu\nu))\frac{C}{\mu\nu},$$
$$\lambda_2\left(\frac{3}{10}-\mu, 1-\nu\right) \to C,$$

where C is here a generic notation for a negative constant. To prove these claims, note first that δ is nonnegative in its domain. Next, let us show how to prove that $\alpha_0 \leq 0$ with strict inequality in the interior of its domain, i.e. that $\alpha \geq 0$ with strict inequality in the interior. Write $\gamma(w,z) = \alpha(3w/10, 1-2z/5)$, $w, z \in [0,1]$. As $\gamma(0,0)$ and $\gamma(1,1)$ are both zero, we use a Taylor expansion around these points. Using e.g. Maple, it is seen that $\gamma'_w(0,0) = 9625/384$ and $\gamma'_z(0,0) = 1625/64$ so that

$$\gamma(w,z) = \frac{9624}{384}w + \frac{1625}{64}z + R$$

where the remainder term is of the form $R = \gamma''_{ww}(w_0,z_0) + 2\gamma''_{wz}(w_0,z_0) + \gamma''_{zz}(w_0,z_0)$ for some point (w_0,z_0) on the line between the origin and (w,z). Summing up the absolute values of the terms in the partial second derivatives, we find that these are in absolute value all bounded by 163 on $[0,1/20]\times[0,1/20]$. Being slightly generous with the linear term coefficients, it follows (since $w \geq w_0 \geq 0$ and $z \geq z_0 \geq 0$) that

$$\gamma(w,z) > 25(w+z) - 163(w+z)^2$$

which is larger than 0 on $[0,1/20]\times[0,1/20]$ except at the origin. Analogously one shows that $\gamma(w,z) > 0$ on $[19/20,1]\times[19/20,1]$ except at $(1,1)$. Next, a

plot in Maple with mesh size 2000×2000 reveals that on the points (w, z) on the mesh and outside these two squares, $\gamma(w, z) > 0.038$. An analogous analysis of the partial first derivatives of γ shows that these are bounded by 48. It follows that the chosen mesh size suffices to draw the desired conclusion.

To prove the negativity of the eigenvalues, it remains to show that $\beta(u, v) \leq \alpha(u, v)^2$ with strict inequality in the interior of the domain. Write $\xi(w, z) = \alpha(3w/10, 1 - 2z/5)^2 - \beta(3w/10, 1 - 2z/5)$, $w, z \in [0, 1]$. Then your favorite software reveals that ξ is quite amenable:

$$\begin{aligned}\xi(w, z) = \frac{729}{125 \cdot 10^{13}} & w(1 - w)(3 - 2w)(2 - w)(10 - 3w) \\ & \cdot z(1 - z)(11 - 8z)(7 - 4z)(5 - 2z)h(w, z),\end{aligned}$$

where

$$\begin{aligned}h(w, z) = {} & 1744w^2z^2 - 4760w^2z - 5280wz^2 + 3475w^2 + 13800wz + 4400z^2 \\ & - 9750w - 11000z + 8125.\end{aligned}$$

Differentiating with respect to w gives

$$h'_w(w, z) = 3488wz^2 - 9520wz - 5280z^2 + 6950w + 13800z - 9750$$

and setting $h'_w(w, z) = 0$ and solving for w gives the solution

$$r(z) = \frac{2640z^2 - 6900z + 4875}{1744z^2 - 4760z + 3475}.$$

Writing $n_r(z)$ and $t_r(z)$ for the numerator and denominator respectively, we find that n_r, t_r and $n_r - t_r$ all have no real roots and that $n_r(0)$, $t_r(0)$ and $n_r(0) - t_r(0)$ are all positive. It follows that $h'_w(\cdot, z)$ has no zero in $[0, 1]$ for an arbitrary fixed z. Since $h'_w(0, 0) < 0$ and $h'_w(0, z)$ has no real roots as a function of z, $h(\cdot, z)$ is decreasing. Since $h(1, z) = 864z^2 - 1960z + 1850$ has no real roots and $h(1, 0) > 0$, it finally follows that $h(\cdot, z)$ is strictly positive for all z, i.e. h is strictly positive for all $w, z \in [0, 1]$. A quick consideration of the product of the other factors in $\phi(w, z)$ now shows that ϕ is 0 on the boundary of $[0, 1] \times [0, 1]$ and strictly positive in the interior.

For the claimed limit behaviors, consider the case $(u, v) = (w, 3/5 + z)$ for small $w \in (0, 3/10)$ and $z \in (0, 2/5)$. It turns out (this can be seen using the expressions for $\alpha(u, v)$ and $\beta(u, v)$ in the appendix) that

$$\beta\left(w, \frac{3}{5} + z\right) + \alpha\left(w, \frac{3}{5} + z\right) = C_1wz + O(w^3 + z^3),$$

$$\beta\left(w, \frac{3}{5}+z\right) - \alpha\left(w, \frac{3}{5}+z\right) = C_2 + O(w+z),$$

and $\alpha(w, \frac{3}{5}+z) = C_3$ for negative constants C_1, C_2, C_3. This proves the claims for $(u,v) \to (0, v_0)$ and $(u,v) \to (u_0, 3/5)$ including when $u_0 = 0$ or $v_0 = 3/5$ (but neither $v_0 = 1$ nor $u_0 = 3/10$).

Consider now $(u,v) = (w, 1-z)$. We have $\alpha(w, 1-z) = C_1 w + C_2 z + O(w^2 + z^2)$ for positive constants C_1 and C_2 and

$$\beta(w, 1-z) - \alpha(w, 1-z)^2 = C_3 w^2 + C_4 z^2 + O(w^3 + z^3)$$

and

$$\beta(w, 1-z) + \alpha(w, 1-z)^2 = C_5 wz + O(w^3 + z^3).$$

This proves the claims for when $(u,v) \to (u_0, 1)$ including $u_0 = 0$. The remaining claims are analogous. □

As a consequence of these results, it follows that for any given unit vector $\mathbf{x}$ orthogonal to Π at some point $\mathbf{k}(u,v) \in \Pi$ there is a $C(u,v,\mathbf{x}) \in [\lambda_1(u,v), \lambda_2(u,v)]$ such that

$$g(\mathbf{k}(u,v) + t\mathbf{x}) = g(\mathbf{k}(u,v))(1 - C(u,v,\mathbf{x})t^2 + O(t^3)),$$

where $C(u,v,\mathbf{x})$ is a continuous function of $(u,v,\mathbf{x})$ bounded away from 0. Hence for any $t = o(1)$,

$$G(\mathbf{k}(u,v) + t\mathbf{x}) = G(\mathbf{k}(u,v))(1 + O(t^3))e^{-C(u,v,\mathbf{x})t^2}.$$

We strongly believe that $G(\mathbf{k}(u,v) + t\mathbf{x})$ is decreasing in t. However the function and its derivative with respect to t seem to be intractable to analyze. Instead we observe that for all t, one has at least some $C(u,v,\mathbf{x})$ bounded away from 0 such that $G(\mathbf{k}(u,v) + t\mathbf{x}) \leq G(\mathbf{k}(u,v))e^{-C(u,v,\mathbf{x})t^2}$. To see this, observe that since the posterior distribution only depends on $\mathbf{W}$ via the number of word 1 tokens in each document. Denote these by O_d and write also A_d for the number that topic A appears in document d. Then the posterior conditioned on (ϕ_A, ϕ_B) can be computed as follows

$$\mathbb{P}(\mathbf{K} = m(a,b,c,d) | O_1 = 0.3m, O_2 = 0.6m, \phi_A, \phi_B)$$

$$= \frac{\mathbb{P}(A_1 = k_{1.}, A_2 = k_{2.})\mathbb{P}(\mathbf{K} = m(a,b,c,d) | A_1 = k_{1.}, A_2 = k_{2.}, \phi_A, \phi_B)}{\mathbb{P}(O_1 = 0.3m, O_2 = 0.6m, \phi_A, \phi_B)}.$$

Since the A_d's are iid uniform and independent of the ϕ's, the first factor of the numerator equals $1/(m+1)^2$. The second factor equals the probability that four independent binomial random variables with parameters $(m(a+b), u)$, $m(1-a-b), v)$, $m(c+d), u)$ and $(m(1-c-d), v)$ respectively equal ma, $m(3/10-a)$, mc and $m(3/5-d)$ respectively. Thus $\mathbb{P}(\mathbf{K} = \mathbf{k}(u,v)|\mathbf{W} = \mathbf{w}, \phi_A, \phi_B) = \Theta(n^{-2})$ for any $(\phi_A, \phi_B) \in [u \pm 1/\sqrt{m}] \times [v \pm 1/\sqrt{m}]$, whereas $\mathbb{P}(\mathbf{K} = \mathbf{k}(u,v) + t\mathbf{x}) \leq e^{-Cmt^2}$ for any (u,v) and any ϕ_A and ϕ_B. This follows from standard Chernoff bounds. The denominator is of order $1/m^2$ for $(\phi_A, \phi_B) \in [0, 3/10] \times [3/5, 1]$.

Consider now the Markov chain $Z = \{Z_t\}$ on $S = [3m/10] \times [7m/10] \times [3m/5] \times [2m/5]$ defined as follows. Its stationary distribution is given by

$$\pi_Z(\mathbf{k}) = CG(\mathbf{k})^m$$

for a normalizing constant C; this differs from π_L only in that π_Z neglects the low order factors in terms of the function s which only make an essential difference close to $\Pi \cap \partial S$. The updates are then made by proposing to change the present state $\mathbf{k}$ to a state $\mathbf{k}'$ which is chosen randomly among the eight states such that $\|\mathbf{k} - \mathbf{k}'\|_1 = 1$ and then making the change with probability $\pi_Z(\mathbf{k}')/(\pi_Z(\mathbf{k}') + \pi_Z(\mathbf{k}))$.

Proposition 2.7 *There is a set B of states such that $\pi_Z(B) \geq 1 - 1/m^{10}$ such that for all $\kappa = \Omega(1/m)$ and all $x \in B$,*

$$\tau^Z_{\text{mix}}(x, \kappa) \leq C(\kappa)m^2.$$

From here on, we will use the notation C for a generic positive constant.

Proof. For $j = 1, 2, \ldots, 2J := 2\lceil 10 \log m \rceil$, let $B_j = \{\mathbf{k} : \pi_Z(\mathbf{k}) \geq e^{-j}\}$. Observe that $\pi_Z(B_J) = 1 - n^{-10}$. We also have that

$$\Phi_Z(r) \geq C\frac{1}{mr^{1/3}} \tag{8}$$

for $r \leq 1/2$ and $\Phi(r) \geq C/m$ for $r > 1/2$ for $\{Z_t\}$ reflected at ∂B_J. This coincides with Z itself as long as Z does not visit a vertex neighboring ∂B_{2L}. However, starting Z from any vertex z_0 within B_J, the expected number of visits to any vertex $x \in B_J^c$ is bounded by m^{-6}. This follows from the well known fact of Markov theory that the expected number of visits to x between two consecutive visits to z_0 is $\pi_Z(x)/\pi_Z(z_0)$. Hence, by Markov's inequality, with probability $1 - m^{-6}$ no such x will be visited. From this and Theorem 1.5, it follows that started from B_J, Z mixes in time Cm^2.

To see that (8) holds, note that $\pi_Z(B_3) > 19/20$. Since any set $A \subseteq B_J$ with $1/10 \leq \pi_Z(A) \leq 1/2$ must have $1/20 \leq \pi_Z(A \cap B_3) \leq 1/2$, it follows that there is a $C > 0$ such that $\Phi_Z(r) \geq C/mr^{1/3}$ for $r \geq 1/10$; this follows on comparing with simple random walk on B_3. On the other hand, this is very easily seen to hold true also for $A \subseteq B_J$ with $\pi_Z(A) < 1/10$. Hence Z has a conductance profile of $C/mr^{1/3}$. Consequently Z reflected at ∂B_{2J} has a mixing time of order m^2 and since Z started within B_J with probability $1 - o(1)$ does not deviate from the reflected version within that time, this goes also for Z itself started from within B_J. □

Next, we would have liked to modify (8) to a lower bound $\Phi(r) \geq C\frac{1}{nr^{1/3}}$ also for L. However, we have not been able to do this formally. Instead we will use a canonical paths argument for the conductance of Z restricted to B_J that can be readily modified in the desired way. First observe that the surface Π can be alternatively expressed as the function surface $\{\mathbf{k}(a,d) = (a, b(a,d), c(a,d), d) : (a,d) \in [0,3/10] \times [0,2/5]\}$. Indeed (using your favorite software)

$$b(a,d) = \frac{\left(-10a + 3d + 3 + \sqrt{r(a,d)}\right)\left(4a + 17d + 3 + \sqrt{r(a,d)}\right)}{-176a + 162d - 6 + 26\sqrt{r(a,d)}}$$

and

$$c(a,d) = \frac{\left(4a - 9d + 3 - \sqrt{r(a,d)}\right)\left(16a + 3d + 3 - \sqrt{r(a,d)}\right)}{-176a + 162d - 6 + 26\sqrt{r(a,d)}}$$

where $r(a,d) = 16a^2 + a(24 - 144d) + 9(d+1)^2$. One can see that these are continuous with bounded partial derivatives. By Proposition 2.6, ∂B_j can be expressed as $h_l(a,d,\omega)$, $(a,d,\omega) \in [0,3/10] \times [0,2/5] \times [0,2\pi)$, where $h_j(a,b,c,d)$ is the distance from $\mathbf{k}(a,d)$ to ∂B_j in the bc-plane in the direction ω and

$$h_j(a,d,\omega) = \left(1 + O\left(\frac{j^{3/2}}{\sqrt{m}}\right)\right)\sqrt{\frac{j}{m}} f(a,d,\omega)$$

for a continuous bounded, and bounded away from 0, function f such that $f(a,d,\cdot)$ describes an ellipse in the bc-plane. In particular the intersection of h_j with the bc-plane is convex for sufficiently large m. Hence we can choose a canonical path in B_j from $x_1 = (a_1,b_1,c_1,d_1)m$ to $x_2 = (a_2,b_2,c_2,d_2)m$, $x_1, x_2 \in B_j$ in the following way. Fix $\alpha \in [0,1]$ and ω_2 such that $x_2 = \alpha h_j(a_2,d_2,\omega_2)$ and write $(a_1,b_0,c_0,d_1) = \alpha h_j(a_1,d_1,\omega_2)$. Start by moving to a nearest neighbor x of x_1 for which the a and d coordinates are odd and then

(i) walk from x to (a_1, b_0, c_1, d_1) via $\{(a_1, b, c_1, d_1) : b \in [b_1, b_0]\}$,

(ii) walk from (a_1, b_0, c_1, d_1) to (a_1, b_0, c_0, d_1) via $\{(a_1, b_0, c, d_1) : c \in [c_1, c_0]\}$,

(iii) walk from (a_1, b_0, c_0, d_0) to x_2 via first walking

$$\{(a, \alpha h_j(a, d_1, \omega_2) \cos \omega_2, \alpha h_j(a, d_1, \omega_2) \sin \omega_2, d_1) : a \in [a_1, a_2]\}$$

and then walking

$$\{(a_2, \alpha h_j(a_2, d, \omega_2) \cos \omega_2, \alpha h_j(a_2, d, \omega_2) \sin \omega_2, d), d : d \in [d_1, d_2]\}.$$

Of course, by the discreteness of the underlying lattice, the paths of step (iii) are chosen as close as possible to the continuous paths indicated, with the extra condition that edges in the b- and c-directions are used only when their a- and d-coordinates are even. This condition together with the starting step makes sure that no edge is ever used in more than one of (i), (ii) or (iii), with the exception that some a- and d-edges will also be used in some paths starting in their neighbors.

Then, as the volume of B_j grows linearly in j, it is easy to see that there is a constant C such that for each $e = \{x, y\} \in B_j \setminus B_{j-1}$,

$$Q(e) \geq \frac{1}{Ce^j m^3}.$$

Since no canonical path between a pair of vertices which are both in a given B_j uses any edge in B_j^c, so that any canonical path using an edge $e \in B_j \setminus B_{j-1}$ has all vertices on at least one side of e in B_{j-1}^c,

$$\sum_{(x,y): e \in \gamma_{x,y}} \pi(x)\pi(y) \leq \frac{C}{e^j m^2}. \tag{9}$$

To spell this out, assume for simplicity that $e \in B_1$. Write π_{before} and π_{after} for the total respective stationary probability masses of vertices that can appear before or after e respectively in the above algorithm for finding canonical paths. Then we have, using a modification of a standard argument for random walk on $\mathbf{Z}_m^4$, that

- if e appears in (i), then $\pi_{before} \leq C/m^{5/2}$ and $\pi_{after} \leq 1$ and so the sum in (9), is bounded by $\pi_{before}\pi_{after} \leq C/m^{5/2}$.
- if e appears in (ii), then $\pi_{before} \leq C/m^2$ and $\pi_{after} \leq C/\sqrt{m}$ and so the sum in (9) is bounded by $C/m^{5/2}$.

- if e appears when adjusting a in (iii), then $\pi_{before} \leq C/m$ and $\pi_{after} \leq C/m$ and so the sum in (9) is bounded by C/m^2.
- if e appears when adjusting b in (iii), then $\pi_{before} \leq 1$ and $\pi_{after} \leq C/m^2$ and so the sum in (9) is bounded by C/m^2.

Adjusting the argument for e in a general B_j is easy. Summing up, we get $\rho(\Gamma) \leq Cm$ and hence

$$\tau^Z_{\text{mix}}(x, \kappa) \leq Cm^2 \log(2m + \kappa^{-1})$$

for all $x \in B_L$ by Theorem 1.6.

Next we observe that Z and L differ in two important ways. One of them is that their stationary distributions are different. The other is that L spends a very long time in some states whereas Z does not. Consider e.g. when $L_t = m(a, b, c, d)$, where $a = 0$. Then by studying π_R, we see that a proposed change from an A to a B of a topic behind a word 1 in document 1 is accepted only with probability $1/(0.3m)$ and hence (since such a change is proposed with probability $0.3/2$) the time taken to change a in the lumped chain will be on average $2m$. For Z however, such a change will always take place with probability at least (say) $1/20$. In general, the probability to make a change in the a-direction for L is $C(a + 1/m)(3/10 - a + 1/m)$ and analogously for the other directions, where C is of constant order.

We will now adjust the arguments for Z to L in two steps, where first step takes care of the latter difference. Consider any Markov chain $X = \{X_t\}$ on B_J with the same stationary distribution as Z for which there are constants C_1, C_2 such that $\mathbb{P}(X_1 = m(a \pm 1/m, b, c, d)|X_0 = m(a, b, c, d)) \in [C_1(a+1/m)(3/10-a+1/m), C_2(a + 1/m)(3/10 - a + 1/m)]$ and analogously for b, c and d. Use the same canonical paths as before. Consider any edge $e = (m(a, b, c, d), m(a, b + 1/m, c, d)) \in B_1$. Unlike for $\{Z_t\}$, $Q(e)$ is now $C(b+1/m)(7/10-b+1/m)/m^3$ rather than C/m^3. However, by the nature of the canonical paths algorithm given, a closer look at (i) reveals that for a path from x_1 to x_2, $x_i = (a_i, b_i, c_i, d_i)$, to use e, either $b_1 \leq mb \leq b_2$ or vice versa. In the former case e will be traversed from left to right (in the b direction) and in the latter it will be traversed from right to left by the canonical path. For the union of all paths traversing e from left to right $\pi_{before} \leq C(b + 1/m)/m^{5/2}$ and $\pi_{after} \leq C(7/10 - b + 1/m)$. This compensates exactly for the extra factor b in $Q(e)$ and hence the conclusion that $\rho(e) \leq C\sqrt{m}$ remains. The union paths traversing b from right to left gets $\pi_{before} \leq C(7/10 - b + 1/m)/m^{5/2}$ and $\pi_{after} \leq C(b + 1/m)$. Steps (ii) and (iii)

are analogous as is the adjustment for general B_j. Hence

$$\tau^X_{\text{mix}}(x,\kappa) \le Cm^2 \log(2m+\kappa^{-1})$$

for all $x \in B_J$.

Next consider L. Obviously L will also with probability $1-o(1)$ not hit the boundary of B_J in the time frame under consideration, so we may consider L restricted to B_J. Again use the same canonical paths as for Z. Let X be defined as the lumped Markov chain of the Gibbs sampler $U = \{U_t\}$ on $\{A,B\}^{2m}$ driven by the stationary distribution

$$\pi_U(\mathbf{z}) = \frac{1}{\binom{\frac{3m}{10}}{am}\binom{\frac{7m}{10}}{bm}\binom{\frac{3m}{5}}{cm}\binom{\frac{2m}{5}}{dm}} G(a,b,c,d)^m$$

for $m(a,b,c,d) = \mathbf{k}(\mathbf{z})$. Then $\{X_t\}$ is a Markov chain of the kind just considered and it is such that the difference between L and X is that the weights corresponding to L as compared to those of X are altered so that the stationary probability $\pi_L(x)$ of vertex $x = (am, bm, cm, dm)$ is $Cs(a,b,c,d)\pi_Z(x)$, where

$$s(a,b,c,d) = \frac{s\left(\frac{10(a+c)}{9}\right) s\left(\frac{10(b+d)}{11}\right) s(a+b)s(c+d)}{s\left(\frac{10a}{3}\right) s\left(\frac{10b}{7}\right) s\left(\frac{5c}{3}\right) s\left(\frac{5d}{2}\right)}$$

and C is of constant order (note that $a+b+c+d$ is of constant order on B_{2J}. Here, for $e = (m(a,b,c,d), m(a,b+1/m,c,d)) \in B_1$,

$$Q(e) = C(b+1/m)(7/10-b+1/m)\pi_L(x).$$

In (i), we get with K being the diameter of the intersection of B_1 with the bc-plane, that for the paths traversing b from left to right

$$\pi_{before} \le C(b+1/m)\pi_L(x)\sum_{i=1}^{K}\frac{\sqrt{K}}{\sqrt{i}} = CK(b+1/m)\pi_L(x) \le Cb\sqrt{m}\pi_L(x)$$

and still $\pi_{after} \le C(7/10-b+1/m)$ and similarly for paths going from right to left. Hence $\rho(e) \le \sqrt{m}$. Again parts (ii) and (iii) are analogous as is the adjustment for general B_j is. Summing up, we have

Proposition 2.8 *There is a set B of states such that $\pi_Z(B) \ge 1-1/m^{10}$ such that for all $\kappa = \Omega(1/m)$ and all $x \in B$,*

$$\tau^L_{\text{mix}}(x,\kappa) \le C(\kappa)m^2 \log m.$$

Finally we take the step from the lumped Markov chain L to the original Gibbs sampler R. Let $\kappa = \Omega(1/m)$, pick C_0 sufficiently large that $\|\mathbb{P}(L_t \in \cdot) - \pi_L\|_{TV} \leq \kappa\}$ whenever $t \geq t_0 := (C_0/2)m^2 \log m$. Start a Gibbs sampler R^0 from π_R and let L^0 be the corresponding lumped Markov chain. It is then possible to design a coupling of L and L^0 such that $\mathbb{P}(L_{t_0} = L^0_{t_0}) \geq 1 - \kappa/2$. Run this coupling up to time t_0. Then if L and L_0 do not agree at that time, let the updates be dependent in an arbitrary way, e.g. simply independent until they meet. Else, $L_{t_0} = L^0_{t_0}$, which means that the underlying Gibbs samplers R and R_0 agree in the number of A's in document d at word j for $d, j = 1, 2$.

Now couple the updates from time t_0 and on in the following way. First pair up the positions in the corpus by pairing each position at which R_t and R^0_t agree with itself and pairing each token (i, j) for which $R_t(i, j) = x$ and $R^0_t(i, j) \neq x$ with a token (i^0, j^0) in the same document with the same word, for which $R^0_t(i^0, j^0) \neq x$ and $R_t(i^0, j^0) = x$, $x = A, B$. Clearly this can be done so that each position is paired with one and only one position. Pick such a pairing.

Next, to pick R_{t+1} and R^0_{t+1}, pick for R_0 the position (i^0, j^0) that is paired with the position (i, j) chosen for R. The key observation now is that the conditional distribution of $R^0(i^0, j^0)$ given $R = (i', j')$, $(i', j') \neq (i^0, j^0)$ is the same as that for $R(i, j)$ given $R(i', j')$, $(i', j') \neq (i, j)$. Hence we may, and do, couple the updates so that R and R^0 still agree at time $t + 1$. This means that no new disagreements between R and R^0 will ever agree and whenever a position (i, j) at which R and R^0 disagree is updated, there is a probability that they agree there, and then also at (i^0, j^0), after the update. This probability might be quite small. However, there are two ways that a new agreement at (i, j) can come about, either by picking (i, j) or by picking (i^0, j^0) to be updated in R, and at least one of these choices gives a probability of at least $1/2$ of agreement after the update (namely the choice that proposes to change a topic that is currently in minority). Hence, if there are D_t positions of disagreement at time $t \geq t_0$, there is a probability of at least $((D_t/2)/(2m)) \cdot (1/2)$ of reducing the number of disagreements by two for one time step. Since there are at most $2m$ disagreements at time t_0, the coupling time T, satisfies that $T - t_0$ is stochastically dominated by the sum $\sum_{j=1}^{m} \xi_j$, where the ξ_j:s are independent and geometric with parameters $(m - 2j)/(4m)$ respectively. Hence $\mathbb{E}[T - t_0] \leq (1 + o(1))2m \log m$ and $\mathbb{V}\mathrm{ar}(T - t_0) < 7m^2$ and so by Chebyshev's inequality $\mathbb{P}(T - t_0 > (C_0/2)m^2 \log m) = O(1/(m^2 \log^2 m)) < \kappa/2$ for sufficiently large m.

Combining the two steps of the coupling argument, it follows that $\mathbb{P}(T > C_0 m^2 \log m) < \kappa$ for sufficiently large m. Now Theorem 1.3 follows from the coupling inequality.

Final remarks. Needless to say the situation considered here, with only two unique words in the corpus, is very unrealistic, as is the fact that we use only two documents. Using only two topics is of course also a bit unrealistic, but since the number of topics is typically very limited compared to the size of the corpus, this is not as serious. However using more words and more documents does not seem to impose any fundamental differences in terms of the analysis here (even though it may become intractable) as long as each word appears a large number of times in each document. Indeed it is easy to generalize Proposition 2.3 to the case with D documents and v distinct words, using the general statement of Lemma 2.2. One gets

$$\pi_R(\mathbf{z}) = C\frac{\binom{n_{..}+2v-2}{k_{..}+v-1}}{\prod_{d=1}^{D}\binom{n_{d.}}{k_{d.}}\prod_{j=1}^{v}\binom{n_{.j}}{k_{.j}}}$$

and for the corresponding lumped Markov chain

$$\pi_L(\mathbf{k}) = C\frac{\binom{n_{..}}{k_{..}}\prod_{d=1}^{D}\prod_{j=1}^{v}\binom{n_{dj}}{k_{dj}}}{(k_{..}+1)(n_{..}-k_{..}+1)\prod_{d=1}^{D}\binom{n_{d.}}{k_{d.}}\prod_{j=1}^{v}\binom{n_{.j}}{k_{.j}}}$$

and when all n_{dj}:s are larger than logarithmic in Dm (having all documents of large equal length m), an analogous treatment seems to be in principle possible. This condition on the n_{dj}'s is of course also not true in most practical situations. What happens in such situations is an important question worthy of further study.

Another assumption we made, was that the topic distribution per document prior α and the word distribution per topic prior β were both $(1, 1)$. It is common to use symmetric priors but usually with smaller parameters. In case of β this does not seem to be likely to make much of a difference as long as we keep the number of words very low compared to the length of the documents. For α, the value used will typically depend on the context and may be optimized through cross validation. In any case, I conjecture (or rather speculate) that with $\alpha = (\epsilon, \epsilon)$ for a small constant ϵ independent of m, the mixing time for $\epsilon < 1/2$ is of order $m^{3-2\epsilon}\log m$.

References

[1] Mark Andrews and Gabriella Vigliocco (1989), The hidden markov topic Model: a probabilistic model of semantic representation, *Topics in Cognitive Science* **2**, 101-113.

[2] Davi M. Blei (2011), Introduction to Probabilistic Topic Models, Princeton University.

[3] David M. Blei, Andrew Y. Ng and Michael I. Jordan (2003), Latent Dirichlet Allocation, *Journal of Machine Learning Research* **93**, 993-1022.

[4] Amit Gruber, Yair Weiss and Michal Rosen-Zvi (2007), Hidden topic markov models, *International Conference on Artificial Intelligence and Statistics* 163170.

[5] M. Jerrum and A. Sinclair (1989), Approximating the permanent, *SIAM Journal on Computing* **18**, 1149-1178.

[6] Ben Morris and Yuval Peres (2005), Evolving sets, mixing and heat kernel bounds, *Probability Theory and Related Fields* **133**, 245-266.

[7] A. Sinclair (1988), Algorithms for random generation and counting: a Markov chain approach. Ph.D thesis, University of Edinburgh, June 1988.

[8] A. Sinclair (1992), Improved bounds for mixing rates of Markov chains and multicommodity flow, *Combinatorics, Probability and Computing* **1**, 351-370.

[9] A. Sinclair and M. Jerrum (1989), Approximate counting, uniform generation, and rapidly mixing Markov chains, *Information and Computation*, 93-133.

[10] F. Tian, B. Gao, D. He, T. Liu (2016), Sentence level recurrent topic model: letting topics speak for themselves, https://arxiv.org/abs/1604.02038

[11] H. Wallach (2006), Topic modeling: beyond bag-of-words, *ICML2006*.

3 Appendix

The functions α and β are given by

$$\alpha_0(u,v) = \frac{25}{2}(v-u)^2\Big(80000\,u^4v^4 - 152000\,u^4v^3 - 152000\,u^3v^4 + 90000\,u^4v^2 + 252400\,u^3v^3 + 90000\,u^2v^4 - 18000\,u^4v - 122220\,u^3v^2 - 122220\,u^2v^3 - 18000\,uv^4 + 12420\,u^3v + 27738\,u^2v^2 + 12420\,uv^3 + 3240\,u^3 + 16182\,u^2v + 16182\,uv^2 + 3240\,v^3 - 6156\,u^2 - 15633\,vu - 6156\,v^2 + 3645\,u + 3645\,v - 729\Big).$$

$$\begin{aligned}\beta_0(u,v) = \Big(\frac{25}{2}\Big)^2(v-u)^4\Big(&6400000000\,u^8v^8 - 24320000000\,u^8v^7 - 24320000000\,u^7v^8 + 37504000000\,u^8v^6 + 93568000000\,u^7v^7 \\ &+37504000000\,u^6v^8 - 30240000000\,u^8v^5 - 146374400000\,u^7v^6 - 146374400000\,u^6v^7 \\ &-30240000000\,u^5v^8 + 13572000000\,u^8v^4 + 120018240000\,u^7v^5 + 233005760000\,u^6v^6 \\ &+120018240000\,u^5v^7 + 13572000000\,u^4v^8 - 3240000000\,u^8v^3 - 54971856000\,u^7v^4 \\ &-195046704000\,u^6v^5 - 195046704000\,u^5v^6 - 54971856000\,u^4v^7 - 3240000000\,u^3v^8 \\ &+324000000\,u^8v^2 + 13500432000\,u^7v^3 + 91833066000\,u^6v^4 + 167842288800\,u^5v^5 \\ &+91833066000\,u^4v^6 + 13500432000\,u^3v^7 + 324000000\,u^2v^8 - 1432080000\,u^7v^2 \\ &-23668200000\,u^6v^3 - 82693612080\,u^5v^4 - 82693612080\,u^4v^5 - 23668200000\,u^3v^6 \\ &-1432080000\,u^2v^7 + 11664000\,u^7v + 2901646800\,u^6v^2 + 23482334160\,u^5v^3 \\ &+44511382260\,u^4v^4 + 23482334160\,u^3v^5 + 2901646800\,u^2v^6 + 11664000\,uv^7 \\ &-127720800\,u^6v - 3799373040\,u^5v^2 - 15209217720\,u^4v^3 - 15209217720\,u^3v^4 \\ &-3799373040\,u^2v^5 - 127720800\,uv^6 + 10497600\,u^6 + 408414960\,u^5v \\ &+3551084388\,u^4v^2 + 6965123256\,u^3v^3 + 3551084388\,u^2v^4 + 408414960\,uv^5 \\ &+10497600\,v^6 - 39890880\,u^5 - 606551328\,u^4v - 2281385088\,u^3v^2 \\ &-2281385088\,u^2v^3 - 606551328\,uv^4 - 39890880\,v^5 + 61515936\,u^4 \\ &+485146584\,u^3v + 933840981\,u^2v^2 + 485146584\,uv^3 + 61515936\,v^4 \\ &-49601160\,u^3 - 218074518\,u^2v - 218074518\,uv^2 - 49601160\,v^3 \\ &+22261473\,u^2 + 52435512\,vu + 22261473\,v^2 \\ &-5314410\,u - 5314410\,v + 531441\Big)\end{aligned}$$