Slow mixing for Latent Dirichlet Allocation - Corrigendum

Johan Jonasson[*][†]

November 30, 2017

Abstract

This note is a correction of the main theorem of ”Slow mixing for latent Dirichlet allocation”, *Statist. Probab. Letters*, Vol. 129 (2017), pp 96-100, along with a corresponding modification of the proof.

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

In the original paper, we considered the case $D = s = v = 2$, $N_1 = N_2 = m$ and $\alpha = \beta = (1, 1)$. As it turns out, the claim there, that the function g has strict local maxima at the two points $(3/10, 7/10, 0, 0)$ and $3/10, 0, 6/10, 0)$, is false (and the mistake in the proof was a failure to observe that g is not differentiable on the boundary of the domain). We need to modify by changing to $D = v = 3$ and for convenience we also change to $N_1 = N_2 = N_3 = 10m$. The notation used here is in obvious analogy with the original paper.

Theorem 0.1 *Consider the case* $n_{11} = 9m$, $n_{12} = m$, $n_{22} = 10m$, $n_{33} = 10$ *and the remaining* $n_{ij} = 0$. *Then there exists a* $\lambda > 0$ *such that for each* $0 < \kappa < 1$,

$$\tau_{\text{mix}}(\kappa) > e^{\lambda n}.$$

[*]Chalmers University of Technology and University of Gothenburg
[†]Research supported by the Knut and Alice Wallenberg Foundation

One may think of this situation in the following way. The corpus actually consists of three topics, but we insist to classify them as two topics. Then the analysis will have to lump two of them together. If there are different proportions of the three topics in the documents and different overlap in terms of words between the topics, this will result in that there are classifications that are different in essential ways and that the Gibbs sampler will have a very hard time to go from one to the other.

It is likely that this problem persists whenever one tries to classify into fewer topics than there really are and that it is important to use ”many enough” topics. It is not clear to me if this resolves the problem and it would indeed be interesting to answer that question. (Of course it is in practice not at all clear what number of topics that there are, as real text is usually very hierarchical in nature, with topics within topics within topics etc. There are generalizations of LDA that try to take this into account.)

The formula for the stationary distribution of the lumped Markov chain generalizes to

$$f(\mathbf{k}) \propto \frac{\binom{n_{..}+2v-2}{k_{..}+v-1} \prod_{d=1}^{D} \prod_{j=1}^{v} \binom{n_{dj}}{k_{dj}}}{\prod_{d=1}^{D} \binom{n_{d.}}{k_{d.}} \prod_{j=1}^{v} \binom{n_{.j}}{k_{.j}}}.$$

For the present setting, it follows in analogy with the original paper that

$$f(\mathbf{k}) \propto C(m) \left(\frac{H(10, a+b)H(11, b+c)H(10, d)}{H(1, b)H(30, a+b+c+d)} \right)^m,$$

where $C(m)$ is polynomial in m, $k_{11} = am$, $k_{12} = bm$, $k_{22} = cm$, $k_{33} = dm$ and

$$H(L, x) = x^x (L-x)^{L-x},\ x \in [0, L].$$

Write

$$G(a, b, c, d) = \frac{H(10, a+b)H(11, b+c)H(10, d)}{H(1, b)H(30, a+b+c+d)},\ (a, b, c, d) \in \Omega.$$

where $\Omega = [0, 9] \times [0, 1] \times [0, 10] \times [0, 10]$.

We claim that this function has strict local maxima at $(9, 1, 10, 0)$ and $(9, 0, 0, 10)$. Once this is done, the theorem follows exactly as in the original paper.

Consider first the point $\mathbf{k}_0 = (9, 0, 0, 10)$. We compute the partial derivatives and find that G'_a and G'_d are continuous at $\mathbf{k}_0$, $G'_a(\mathbf{k}_0) > 0$ and $G'_d(\mathbf{k}_0) > 0$.

Hence there is a neighborhood Ω_0 of $\mathbf{k}_0$ such that $G'_a(\mathbf{k}) > 0$ and $G'_d(\mathbf{k}) > 0$ for all $\mathbf{k} \in \Omega_0$. Also

$$G(9, b, c, 10) = \frac{(b+c)^{b+c}(b+9)^{b+9}}{b^b(b+c+19)^{b+c+19}}.$$

This function is strictly decreasing in both its variables. To see this, it suffices to observe that the map $[0, \infty) \ni y \to y/(y+C)$ for some arbitrary positive constant C is strictly decreasing. To see this in turn, find that the derivative of the logarithm is $\log(y/(y+C))$, which is negative and tends to ∞ as $y \downarrow 0$. It follows that for $(a, b, c, d) \in \Omega_0$,

$$G(a, b, c, d) \leq G(9, b, c, d) \leq G(9, b, c, 10) \leq G(9, 0, c, 10) \leq G(9, 0, 0, 10)$$

with strict inequality whenever there is strict inequality in the corresponding variable.

Consider now $\mathbf{k}_1 = (9, 1, 10, 0)$. This more or less analogous, but slightly more inconvenient since all the partial derivatives are undefined on the boundary near $\mathbf{k}_1$. However

$$G'_a(a, 1, 10, 0) = \frac{(a+1)^{a+1}(9-a)^{9-a}(19-a)^{19-a}}{(a+11)^{a+11}}.$$

The derivative of the logarithm is $-2 + \log((a+1)/((9-a)(19-a)(a+11)))$, which is positive for all a sufficiently close to 9 and tends to ∞ as $a \uparrow 9$. Hence $G(a, 1, 10, 0) < G(\mathbf{k}_1)$ whenever a is sufficiently close to 9.

Next consider $G(a, 1, c, 0)$ and find that the expression contains a factor $(10 - c)^{10-c}$ in the denominator. From this it follows analogously that $G(a, 1, c, 0)$ is increasing in c for c sufficiently close to 10. Consider then $G(a, b, c, 0)$ as a function of b and find analogously that it is increasing in b for b sufficiently close to 1. Finally consider $G(a, b, c, d)$ as a function of d. In this case one finds that the expression contains a factor d^d in the numerator, from which it follows analogously that it is decreasing as a function of d for d sufficiently close to 0. Now it follows as for $\mathbf{k}_0$ that also $\mathbf{k}_1$ is a strict local maximum.

Finally observe that $G(9, 1, 10, 0)$ exceeds $G(9, 0, 0, 1)$ by orders of magnitude. Hence a true posterior sample from the LDA model would almost certainly classify the entire documents 1 and 2 as one topic and the third document as the other, or something very close to that. However, starting from a classification that nearly entirely classifies all instances of word 1 and 3 as one topic and word 2 as the other, it will take an astronomical time to leave a small neighborhood of it.