Supplementary Materials for “Rethinking LDA: Why Priors Matter”

Hanna M. Wallach **David Mimno** **Andrew McCallum**
Department of Computer Science
University of Massachusetts Amherst
Amherst, MA 01003
{wallach,mimno,mccallum}@cs.umass.edu

1 Conditional Posterior Probabilities for LDA

With symmetric Dirichlet priors over $\Theta = \{\boldsymbol{\theta}_1, \dots \boldsymbol{\theta}_D\}$ and $\Phi = \{\boldsymbol{\phi}_1, \dots \boldsymbol{\phi}_T\}$, the conditional posterior probability, or predictive probability, of topic t occurring in document d given the corresponding topic assignments $\mathcal{Z} = \{\boldsymbol{z}^{(d)}\}_{d=1}^D$ for a corpus of documents $\mathcal{W} = \{\boldsymbol{w}^{(d)}\}_{d=1}^D$ is as follows:

$$P(z_{N_d+1}^{(d)} = t \,|\, \mathcal{Z}, \alpha\boldsymbol{u}) = \int \mathrm{d}\boldsymbol{\theta}_d \, P(t \,|\, \boldsymbol{\theta}_d) \, P(\boldsymbol{\theta}_d \,|\, \mathcal{Z}, \alpha\boldsymbol{u}) = \frac{N_{t|d} + \frac{\alpha}{T}}{N_d + \alpha}, \tag{1}$$

where topic t occurs $N_{t|d}$ times in $\boldsymbol{z}^{(d)}$ of length $N_d = \sum_t N_{t|d}$. In other words, the conditional posterior distribution over topics for document d is a Pólya conditional distribution.

The conditional posterior distribution over words for topic t is also a Pólya conditional distribution.

2 Joint Distributions for LDA

With symmetric priors, the joint distribution over topic assignments $\mathcal{Z}$ for documents $\mathcal{W}$ is

$$\begin{aligned} P(\mathcal{Z} \,|\, \alpha\boldsymbol{u}) &= \textstyle\prod_d \prod_n P(z_n^{(d)} \,|\, \mathcal{Z}_{<d,n}, \alpha\boldsymbol{u}) \\ &= \textstyle\prod_d \prod_n \dfrac{N_{z_n^{(d)}|d}^{<d,n} + \frac{\alpha}{T}}{N_d^{<d,n} + \alpha} = \prod_d \dfrac{\Gamma(\alpha)}{\Gamma(N_d + \alpha)} \prod_t \dfrac{\Gamma(N_{t|d} + \frac{\alpha}{T})}{\Gamma(\frac{\alpha}{T})}, \end{aligned} \tag{2}$$

where “$< d, n$” denotes a quantity involving data from documents $1, \dots, d$ and, for document d, positions $1, \dots, n-1$ only. In other words, the joint distribution over $\mathcal{Z}$ is a Pólya distribution.

The joint distribution over $\mathcal{W}$ given $\mathcal{Z}$ is also a Pólya distribution.

3 Variation of Information for Topic Models

The similarity between two sets of topic assignments $\mathcal{Z}$ and $\mathcal{Z}'$ for documents $\mathcal{W}$ can be measured using *variation of information*, introduced by Meilă [2] and recently used by Goldwater and Griffiths in the context of text processing [1]. Given two sets of topic assignments $\mathcal{Z}$ and $\mathcal{Z}'$ for some $\mathcal{W}$ (with T and T' topics, respectively), computing the variation of information between $\mathcal{Z}$ and $\mathcal{Z}'$, denoted VI $(\mathcal{Z}, \mathcal{Z}')$, requires three distributions: $P(z)$ over the T topics in $\mathcal{Z}$, proportional to $\{N_t\}_{t=1}^T$ for $\mathcal{Z}$; $P(z')$ over the T' topics in $\mathcal{Z}'$, proportional to $\{N_{t'}\}_{t'=1}^{T'}$ for $\mathcal{Z}'$; and $P(z, z')$, proportional to the number of tokens assigned to topic t in $\mathcal{Z}$ and topic t' in $\mathcal{Z}'$. VI $(\mathcal{Z}, \mathcal{Z}')$ is then

$$\begin{aligned} \mathrm{VI}\,(\mathcal{Z}, \mathcal{Z}') &= H(z) + H(z') - 2I(z, z') \\ &= H(z \,|\, z') + H(z' \,|\, z), \end{aligned} \tag{3}$$

where $H(\cdot)$ denotes the entropy of a random variable and $I(\cdot\,,\cdot)$ denotes the mutual information between two random variables. If two sets of topic assignments $\mathcal{Z}$ and $\mathcal{Z}'$ are identical, then VI $(\mathcal{Z}, \mathcal{Z}')$ will be zero. The higher the value of VI $(\mathcal{Z}, \mathcal{Z}')$, the greater the dissimilarity between $\mathcal{Z}$ and $\mathcal{Z}'$.

4 Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval, in part by CIA, NSA and NSF under NSF grant number IIS-0326249, and in part by subcontract number B582467 from Lawrence Livermore National Security, LLC under prime contract number DE-AC52-07NA27344 from DOE/NNSA. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

References

[1] S. Goldwater and T. L. Griffiths. A fully Bayesian approach to unsupervised part-of-speech tagging. In *Association for Computational Linguistics*, 2007.

[2] M. Meilă. Comparing clusterings by the variation of information. In *Conference on Learning Theory*, 2003.