Pólya Urn Latent Dirichlet Allocation

using sparsity to reduce parallel MCMC complexity in natural language processing

Alexander Terenin (Imperial College London) · Måns Magnusson (Linköping University, SE),
Leif Jonsson (Linköping University, SE) · David Draper (University of California, Santa Cruz)

Imperial College London

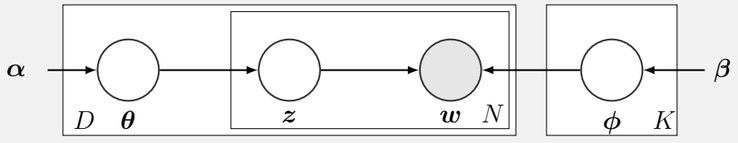
LINKÖPING UNIVERSITY

Abstract

Dirichlet distributions are an essential building block in many Bayesian models, particularly those in natural language processing. We propose the Poisson Pólya urn distribution, a novel sparse approximation to the Dirichlet distribution. We prove its asymptotic exactness under large data sets. We show that in the popular Latent Dirichlet Allocation topic model, we can use the Poisson Pólya urn to define a Gibbs sampler that is massively parallel in all steps, and uses only sparse data structures. We show that this sampler faster than the current state-of-the-art, both theoretically and empirically.

Latent Dirichlet Allocation

The canonical topic model
- Used to infer topics in a collection of documents
- Very popular: 20,000+ citations on Google Scholar



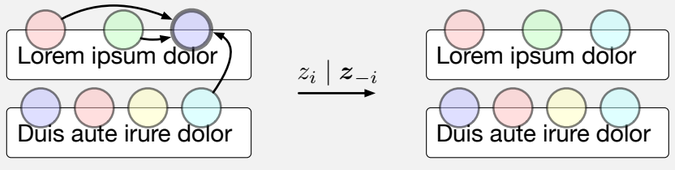
α, β, V	hyperparameters	K	number of topics
$\boldsymbol{z}$	topic indicators	$\boldsymbol{\Phi}$	document-topic probabilities
$\mathbf{n}$	topic-word statistic	$\mathbf{m}$	document-topic statistic

$K_d^{(\mathbf{m})}$ number of nonzero topics in document d

$K_v^{(\cdot)}$ number of nonzero topics assigned to word type v in $\mathbf{n}, \boldsymbol{\Phi}$

Sparse Collapsed Gibbs Sampling

(previous state-of-the-art)



$$z_i \mid \boldsymbol{z}_{-i} \propto \frac{n_{k,v(i)}^{-i} + \beta}{n_{k,\cdot}^{-i} + V\beta}(m_{d,k}^{-i} + \alpha) \qquad O\left[\max\left\{K_d^{(\mathbf{m})}, K_v^{(\mathbf{n})}\right\}\right]$$

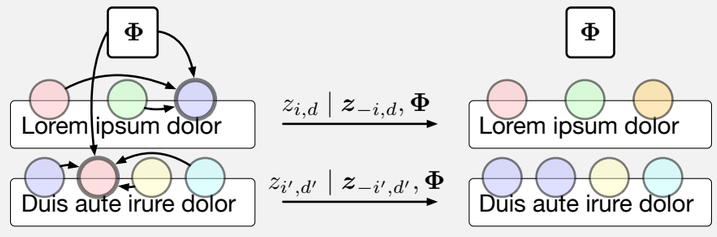
Performance limited by sparsity in $\mathbf{m}$ or $\mathbf{n}$
Not parallelizable without losing convergence theory
Alternatives require Metropolis-Hastings steps – slower mixing

Poisson Pólya Urn approximation to the Dirichlet distribution

Dirichlet		Poisson Pólya Urn
$\boldsymbol{x} = \left[\frac{\gamma_1}{\sum_{i=1}^k \gamma_i}, .., \frac{\gamma_k}{\sum_{i=1}^k \gamma_i}\right]$	$\rightarrow$	$\boldsymbol{y} = \left[\frac{\tilde{\gamma}_1}{\sum_{i=1}^k \tilde{\gamma}_i}, .., \frac{\tilde{\gamma}_k}{\sum_{i=1}^k \tilde{\gamma}_i}\right]$
$\gamma_i \sim \text{Gamma}(\varpi F_i, 1)$		$\tilde{\gamma}_i \sim \text{Poisson}(\varpi F_i)$
Dense		Sparse

Theorem. Let $\boldsymbol{x} \sim \text{Dir}(\varpi, \boldsymbol{F})$ and $\boldsymbol{y} \sim \text{PPU}(\varpi, \boldsymbol{F})$. Then for all $\boldsymbol{F}$ we have $||\boldsymbol{x} - \boldsymbol{y}|| \to 0$ as $\varpi \to \infty$ in the Levy-Prokhorov metric. [1]

Parallel Doubly Sparse Partially Collapsed Gibbs Sampling



$$z_{i,d} \mid \boldsymbol{z}_{-i,d}, \boldsymbol{\Phi} \propto \phi_{k,v(i)}(m_{d,k}^{-i} + \alpha)$$

Partially Collapsed LDA [2]		Pólya Urn LDA
$\boldsymbol{\phi}_k \mid \boldsymbol{z} \sim \text{Dir}(\boldsymbol{n}_k + \boldsymbol{\beta})$	$\rightarrow$	$\boldsymbol{\phi}_k \mid \boldsymbol{z} \sim \text{PPU}(\boldsymbol{n}_k + \boldsymbol{\beta})$
$O\left[K_d^{(\mathbf{m})}\right]$		$O\left[\min\left\{K_d^{(\mathbf{m})}, K_v^{(\boldsymbol{\Phi})}\right\}\right]$

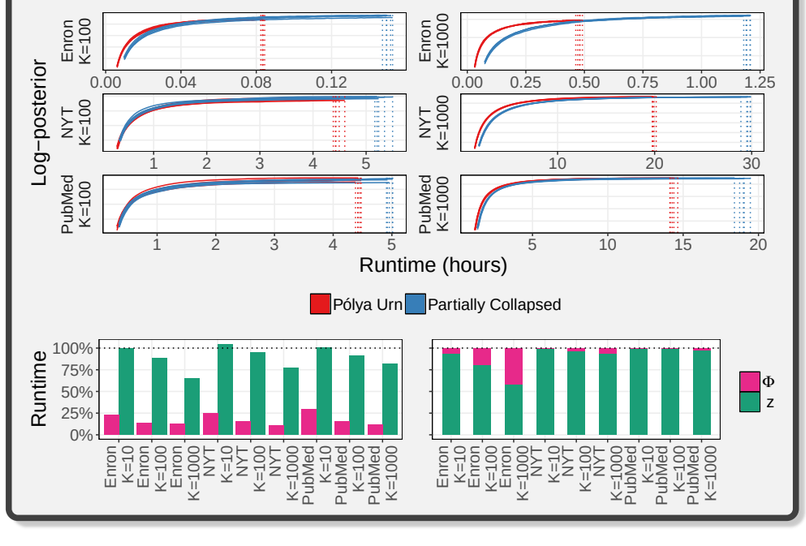
Uses sparsity in $\mathbf{m}$ and $\mathbf{n}$ – the maximum possible
Conditional independence allows parallelism in all documents

References

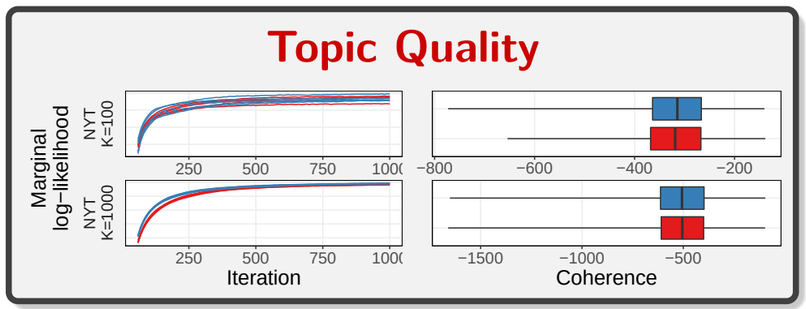
[1] A. Terenin, M. Magnusson, L. Jonsson, and D. Draper. Pólya Urn Latent Dirichlet Allocation: a doubly sparse massively parallel sampler. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[2] M. Magnusson, L. Jonsson, M. Villani, and D. Broman. Sparse Partially Collapsed MCMC for Parallel Inference in Topic Models. *Journal of Computational and Graphical Statistics*, 26(4), 2017.

Runtime Performance



Topic Quality



Parallelism