Pólya Urn Latent Dirichlet Allocation

Alexander Terenin
Imperial College London

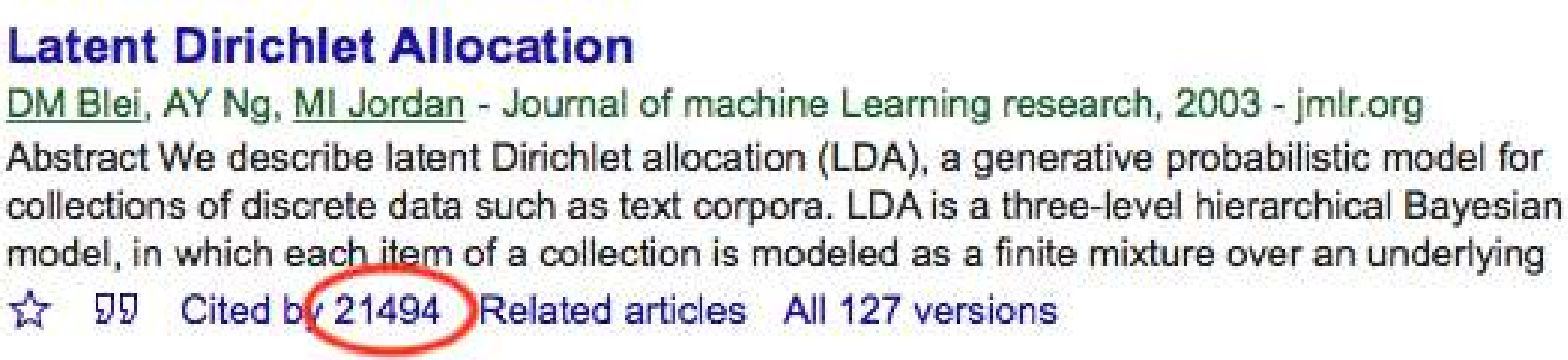
Joint work with Måns Magnusson,
Leif Jonsson, and David Draper

January 22, 2018

Talk for Carnegie Mellon University

arXiv:1704.03581

Latent Dirichlet Allocation



The canonical topic model - everybody uses it!

Topic Modeling

Topic modeling is an area of natural language processing.

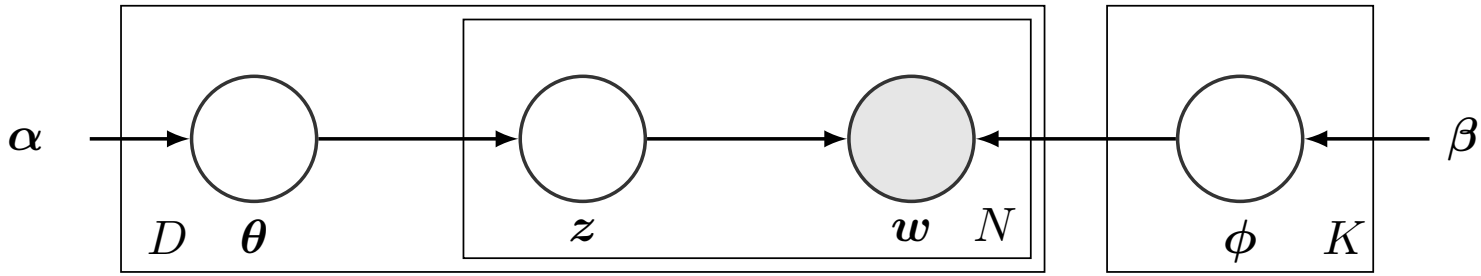
Setup

- We have a dataset consisting of *documents*.
- Each document is a collection of *word tokens*.

Goal: group documents into topics.

Approach: Bayesian learning.

Latent Dirichlet Allocation



Define a *topic* as a probability distribution over word tokens.

Assume each *document* has a probability distribution over topics.

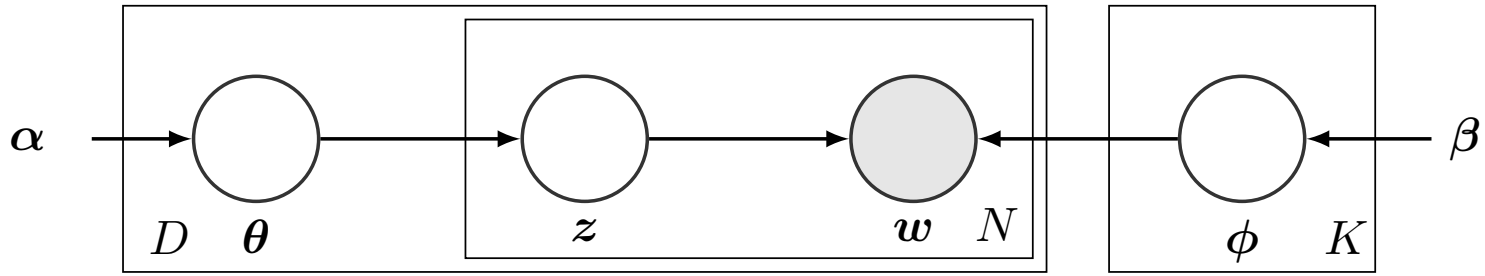
Assume documents are exchangeable, assume bag of words:

$\overset{\text{de Finetti's theorem}}{\Longrightarrow}$ conditionally multinomial likelihood.

$\underset{K \times V}{\boldsymbol{\Phi}}$: topic-word proportions. $\boldsymbol{z}$: unobserved topic indicators.

$\underset{D \times K}{\boldsymbol{\Theta}}$: document-topic proportions. $\boldsymbol{w}$: observed word tokens.

Latent Dirichlet Allocation



$\underset{K \times V}{\Phi}$: topic-word proportions. z : unobserved topic indicators.

$\underset{D \times K}{\Theta}$: document-topic proportions. w : observed word tokens.

Need to learn Θ, Φ, z.

Bayesian Learning: (prior, likelihood) $\rightarrow$ posterior.

$$\theta_k \sim \mathrm{Dir}(\alpha) \qquad \phi_k \sim \mathrm{Dir}(\beta)$$

Dirichlet distribution: conjugate with multinomial.

Training algorithm for $\boldsymbol{\Theta}, \boldsymbol{\Phi}, \boldsymbol{z}$

We only care about $\boldsymbol{z}$: integrate $\boldsymbol{\Theta}, \boldsymbol{\Phi}$ out $\implies$ $z \sim \text{discrete}$.

Algorithm – sparse fully collapsed Gibbs sampling:

$$z_{i,d} \mid \boldsymbol{z}_{-i,d} \propto \frac{n_{k,v(i)}^{-i} + \beta}{n_{k,\cdot}^{-i} + V\beta}(m_{d,k}^{-i} + \alpha).$$

Many alternative methods: variational, geometric, spectral, etc.

MCMC approaches tend to give best-quality topics.

Problem: only uses sparsity in $\mathbf{m}$ $\implies$ $O\left[\sum_{i=1}^{N} K_{d_i}^{(\mathbf{m})}\right]$ per iteration.

Problem: completely sequential $\implies$ slow.

Asynchronous Distributed LDA

Solution: pretend that

$$z_{i,d} \mid \boldsymbol{z}_{-i,d} \propto \frac{n^{-i}_{k,v(i)} + \beta}{n^{-i}_{k,\cdot} + V\beta}(m^{-i}_{d,k} + \alpha).$$

isn't sequential and just compute asynchronously.

Convergence analysis: much more complicated![1]

[1] A. Terenin, D. Simpson, and D. Draper. Asynchronous Gibbs Sampling. In revision at Journal of the American Statistical Association: theory and methods. arXiv:1509.08999, 2017.

Partially Collapsed LDA

Can we do better? What if we don't integrate out $\mathbf{\Phi}$?

Algorithm: sparse partially collapsed Gibbs sampling[2].

(1) Sample $\boldsymbol{\phi}_k \sim \mathrm{Dir}(\boldsymbol{n}_k + \boldsymbol{\beta})$.

(2) Sample $z_{i,d} \propto \phi_{k,v(i)}(m^{-i}_{d,k} + \alpha)$.

Step (1) is parallel by topics, step (2) is parallel by documents.

Empirical convergence rate is good: similar to fully collapsed LDA.

Problem: still only uses sparsity in $\mathbf{m} \implies O\left[\sum_{i=1}^{N} K_{d_i}^{(\mathbf{m})}\right]$.

Problem: requires $O(K \times V)$ memory.

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

Key idea

This work: completely eliminate both problems.

Observation: the problem is that $\boldsymbol{x} \sim \mathrm{Dir}(\boldsymbol{\varpi})$ is a dense vector.

What if we made $\boldsymbol{x}$ sparse?

We'd get a complexity of $O\left[\sum_{i=1}^{N} \min\left\{K_{d(i)}^{(\mathbf{m})}, K_{v(i)}^{(\boldsymbol{\Phi})}\right\}\right]$ – better!

Can we do that in a principled manner?

Dirichlet Processes and Pólya Urns

Pólya Urn: we have an urn with colored balls.

We take out one ball, and replace it with two balls of the same color.

Result: the de Finetti measure of a Pólya Urn is a Dirichlet Process.[3]

So, to sample from a Dirichlet process, simulate a large Pólya Urn.

Large in the sense of number of balls in the urn:

$$\implies \text{concentration parameter of } DP(\varpi, F).$$

Henceforth: reparametrize $\mathrm{Dir}(\boldsymbol{\varpi})$ as $\mathrm{Dir}(\varpi, \boldsymbol{F})$ with $\boldsymbol{\varpi} = \varpi \boldsymbol{F}$.

[3] D. Blackwell and J. MacQueen. Ferguson distributions via Pólya Urn schemes. The Annals of Statistics, 1973.

Pólya Urns and Bootstrap distributions

If we have 5 red balls and 2 green balls, what's the Pólya Urn's distribution? Exactly that: 5 red balls and 2 green balls:

$$\implies \mathrm{MN}(\varpi, \boldsymbol{F}).$$

Problem: large multinomials are slow, cumbersome to work with.

Observation: same problem arises in distributed Bootstrap methods.

Idea: replace multinomial with Poisson process on an extended space.

The Poisson Pólya Urn distribution

Consider an n-dimensional probability vector $\boldsymbol{x}$ given by

$$x_i = \frac{\gamma_i}{\sum_{j=1}^{n} \gamma_j}$$

where $\gamma_i \sim \mathrm{Pois}(\varpi_i)$. Call this the Poisson Pólya Urn distribution.

This is *almost* the Gamma representation of a Dirichlet distribution. We just replaced the Gammas with Poissons.

This is sparse and scalable. Does it actually work?

The Poisson Pólya Urn distribution

Theorem:[4] if $\boldsymbol{x} \sim \mathrm{PPU}(\varpi, \boldsymbol{F})$, and $\boldsymbol{x}^* \sim \mathrm{Dir}(\varpi, \boldsymbol{F})$, we have

$$||\boldsymbol{x} - \boldsymbol{x}^*|| \to 0$$

as $\varpi \to \infty$ in the Lévy-Prokhorov metric, and therefore both random vectors convergence in distribution.

In LDA, ϖ is driven by the number of word tokens, which is enormous.

Moreover, approximation is good even without asymptotics:

$$\mathbb{E}(x) = \mathbb{E}(x^*) \qquad \mathrm{Var}(x) \cong \mathrm{Var}(x^*).$$

[4] A. Terenin, M. Magnusson, L. Jonsson, and D. Draper. Pólya Urn Latent Dirichlet Allocation: a doubly sparse massively parallel sampler. In revision at IEEE Transactions on Pattern Analysis and Machine Intelligence. arXiv:1704.03581, 2017.

Pólya Urn LDA

Algorithm: Poisson Pólya Urn doubly sparse partially collapsed Gibbs sampling[5].

(1) Sample $\boldsymbol{\phi}_k \sim \mathrm{PPU}(\boldsymbol{n}_k + \boldsymbol{\beta})$.

(2) Sample $z_{i,d} \propto \phi_{k,v(i)}(m_{d,k}^{-i} + \alpha)$.

Complexity: $O\left[\sum_{i=1}^{N} \min\left\{K_{d(i)}^{(\mathbf{m})}, K_{v(i)}^{(\mathbf{\Phi})}\right\}\right]$.

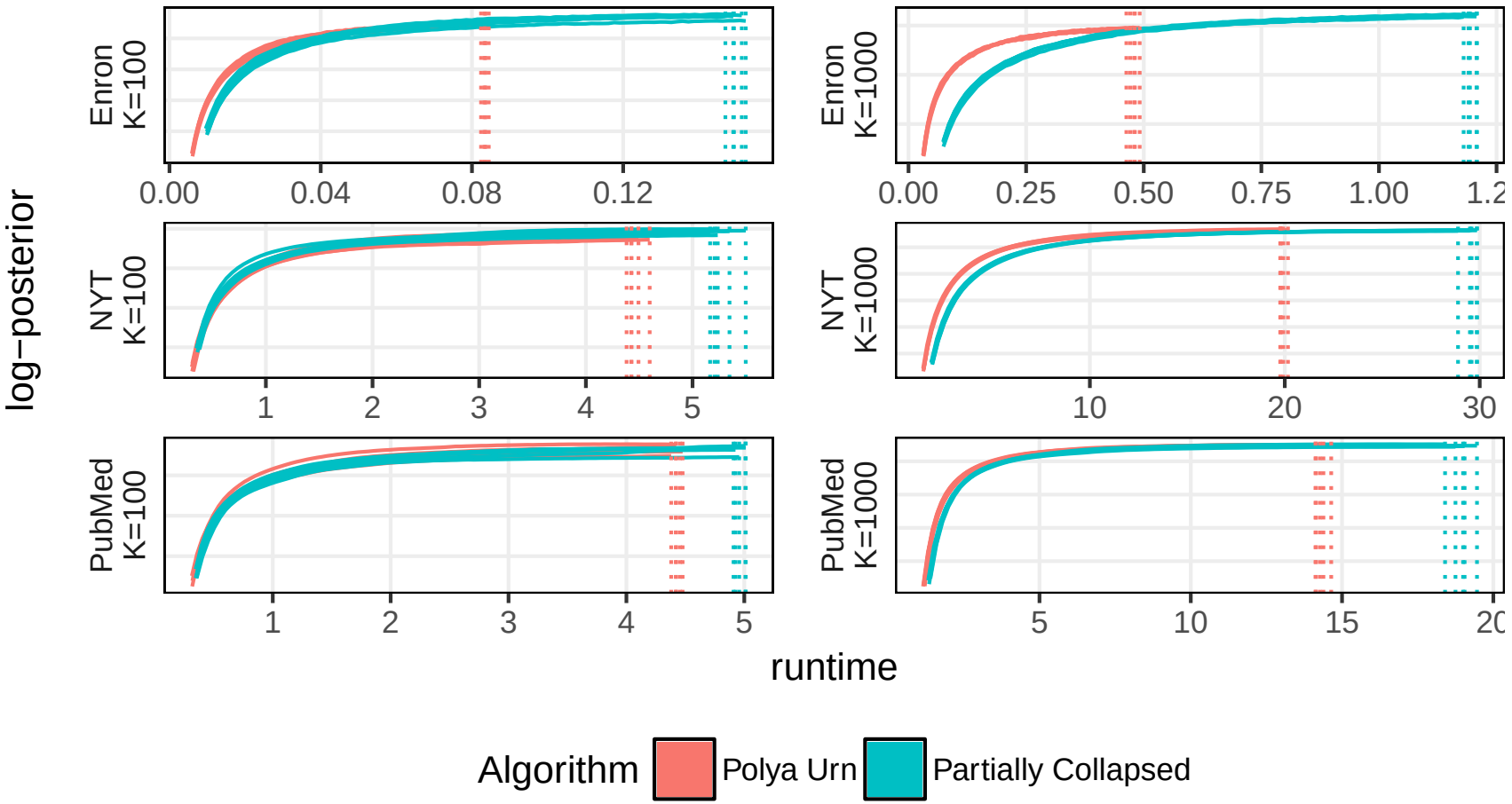
Everything is massively parallel.

Empirical convergence rate is good.

Same memory requirements as fully collapsed LDA.

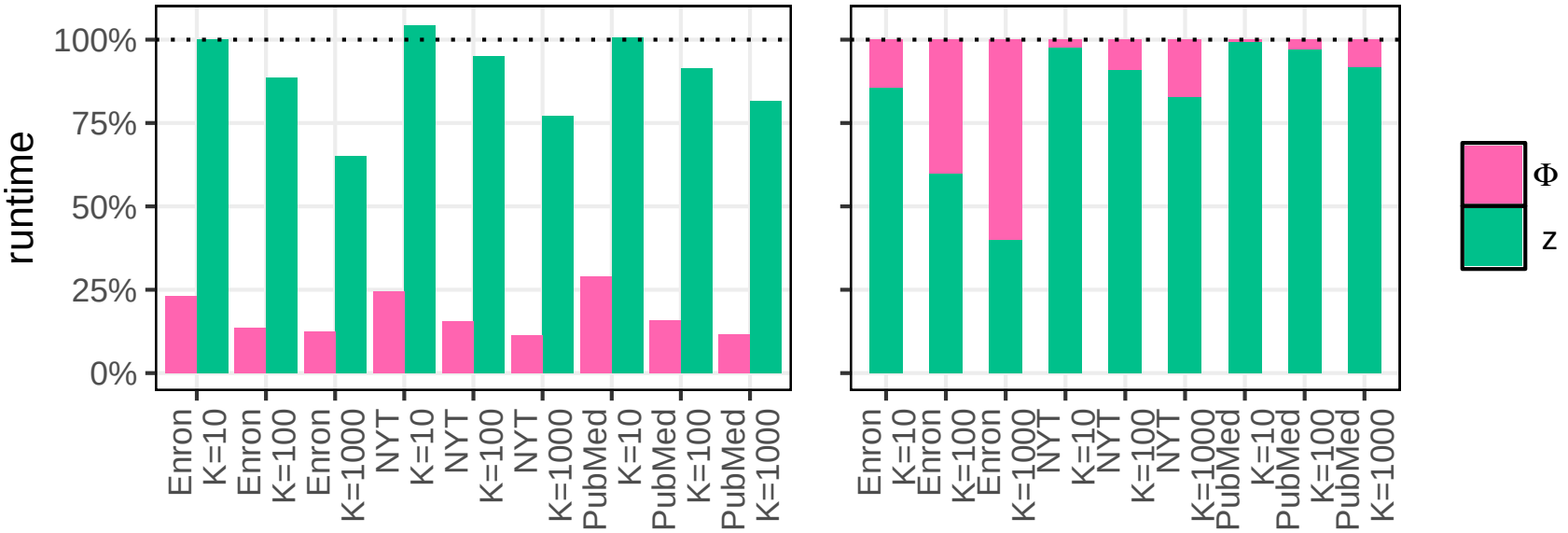
[5] A. Terenin, M. Magnusson, L. Jonsson, and D. Draper. Pólya Urn Latent Dirichlet Allocation: a doubly sparse massively parallel sampler. In revision at IEEE Transactions on Pattern Analysis and Machine Intelligence. arXiv:1704.03581, 2017.

Results



In a parallel setting, performance is state-of-the-art

Results

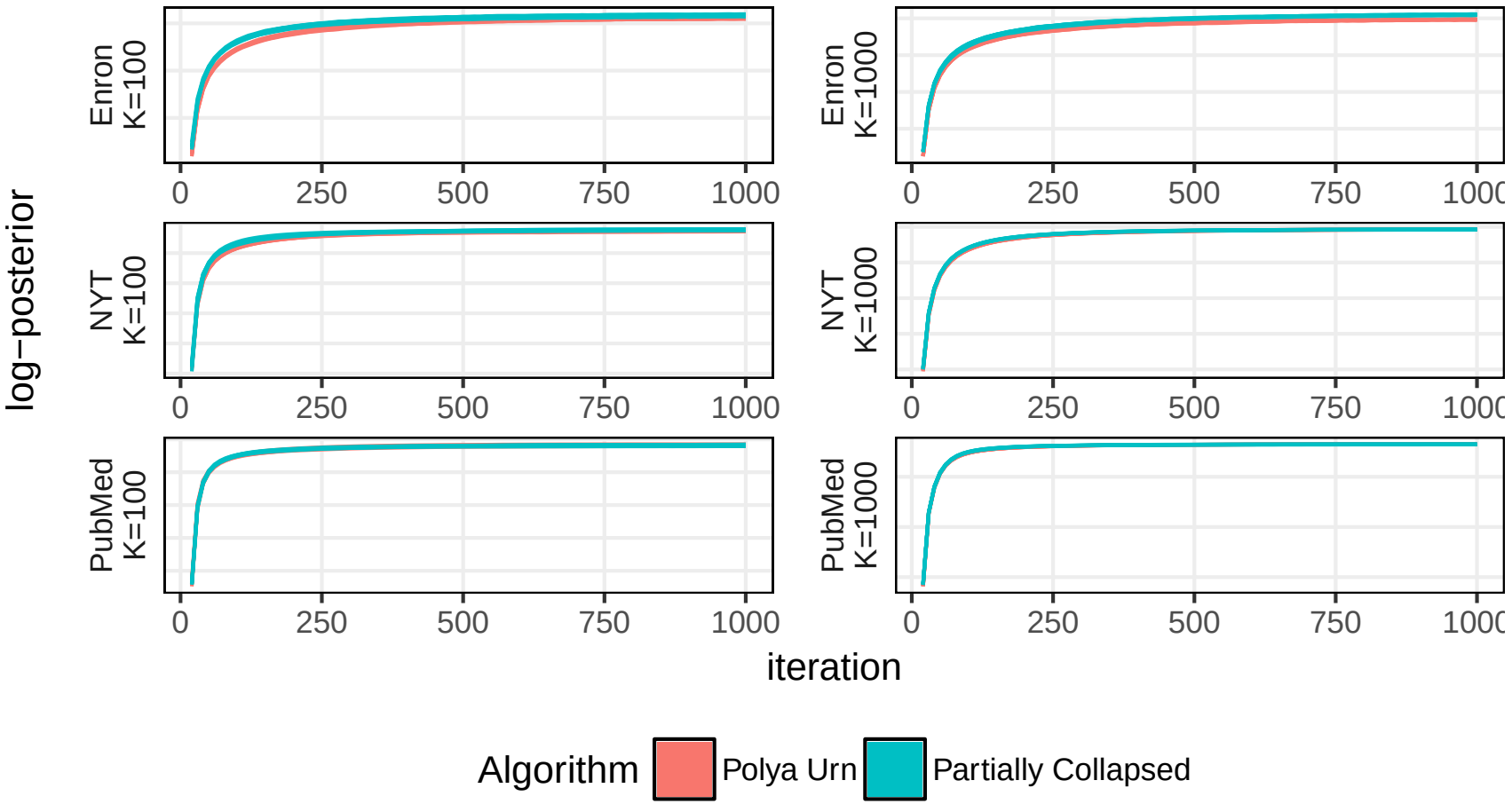


Where does the speedup come from?

Faster z due to additional sparsity.

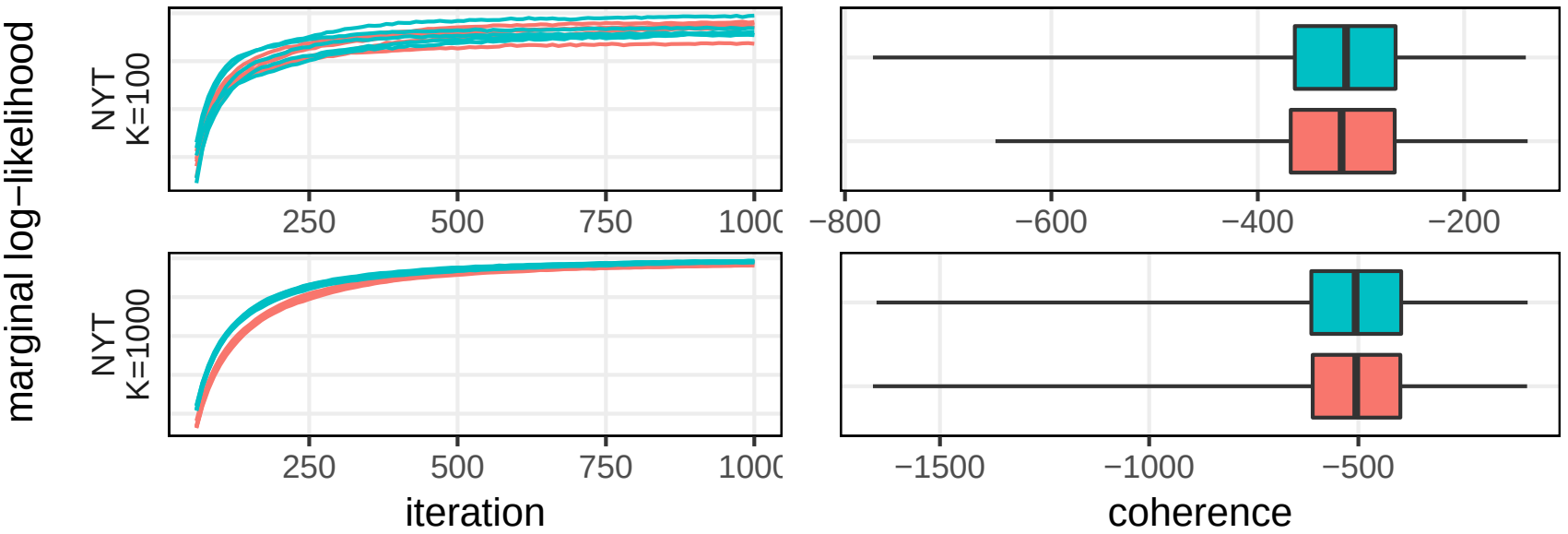
Faster Φ due to fast Poisson trick.

Results



Does it converge slower?

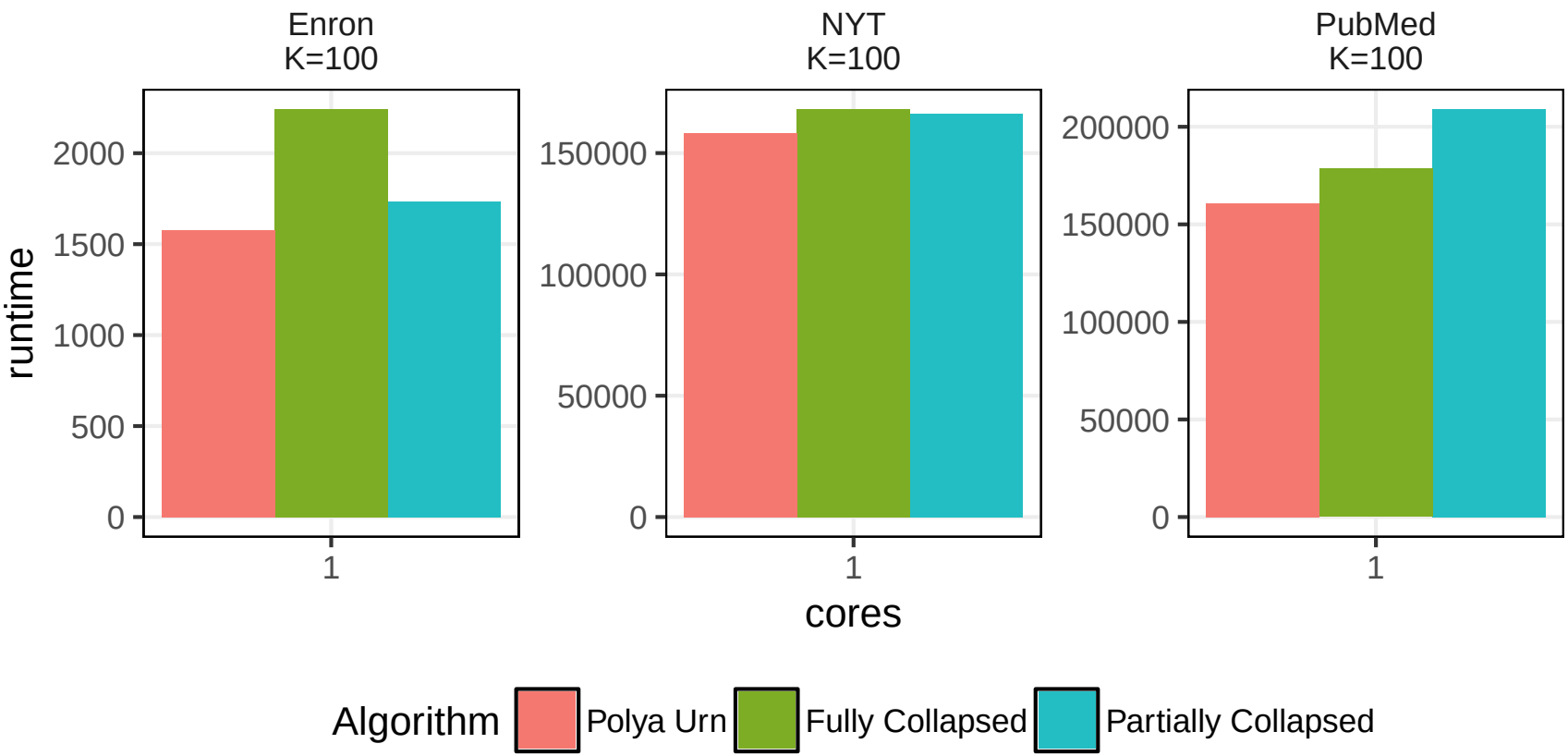
Results



Does it give good quality topics?

Yes: similar topic coherence and test set log marginal likelihood for z.

Results



Is single-core performance good?

Yes: also state-of-the-art among MCMC-based methods.

Results

Any downsides?

Approximation may not be good for small N and large K.

Sparse-on-sparse arithmetic is hard to implement well.

Conclusions

Poisson Pólya Urn: generic asymptotic approximation to Dirichlet.

Used in this work for doubly sparse massively parallel LDA sampler.

Theorem:[6] if $\boldsymbol{x} \sim \mathrm{PPU}(\varpi, \boldsymbol{F})$, and $\boldsymbol{x}^* \sim \mathrm{Dir}(\varpi, \boldsymbol{F})$, we have

$$||\boldsymbol{x} - \boldsymbol{x}^*|| \to 0$$

as $\varpi \to \infty$ in the Lévy-Prokhorov metric, and therefore both random vectors convergence in distribution.

Looks promising for other areas – try it yourself!

[6] A. Terenin, M. Magnusson, L. Jonsson, and D. Draper. Pólya Urn Latent Dirichlet Allocation: a doubly sparse massively parallel sampler. In revision at IEEE Transactions on Pattern Analysis and Machine Intelligence. arXiv:1704.03581, 2017.