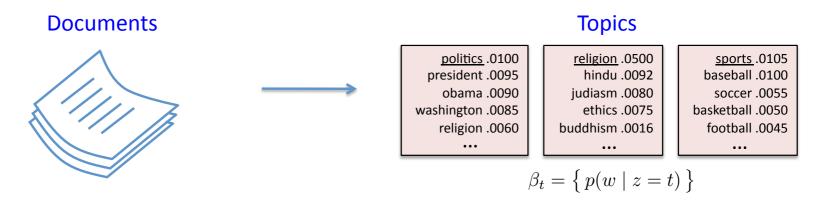
Complexity of Inference in Latent Dirichlet Allocation

David Sontag, Daniel Roy (NYU, Cambridge)

W66

Topic models are powerful tools for exploring large data sets and for making inferences about the content of documents



Almost all uses of topic models (e.g., for unsupervised learning, information retrieval, classification) require **probabilistic inference**:



Complexity of Inference in Latent Dirichlet Allocation

David Sontag, Daniel Roy (NYU, Cambridge)

W66

We study the complexity of probabilistic inference in Latent Dirichlet Allocation

Input: new document with words $w_{1:N}$ topic-word distributions $\beta_t, t = 1, 2, ..., T$ and Dirichlet hyper-parameters $\alpha_{1:T}$

Generative model

① $\theta \sim \mathrm{Dirichlet}(\alpha_{1:T})$ Choose a distribution over the T topics

2 For each word i,

 $z_i \mid \theta \sim \theta$ Choose a topic for i'th word

 $w_i \mid z_i \sim \beta_{z_i}$ Sample a word

Popular inference problems

- 1. Maximize $p(z_{1:N} \mid w_{1:N})$. \leftarrow Discrete. Classification
- 2. Maximize $p(\theta \mid w_{1:N})$. \leftarrow Dimensionality reduction, IR
- 3. Sample from $p(\theta \mid w_{1:N})$. \leftarrow Useful for learning

Complexity of Inference in Latent Dirichlet Allocation

David Sontag, Daniel Roy (NYU, Cambridge)

W66

Main Results

Maximize $p(z_{1:N} \mid w_{1:N})$

For any α

Most common→
setting

# topics in MAP assignment	Complexity	Intuition
Small	Easy	First choose topic sizes, then match words to topics
Large	NP-hard	Reduction from set packing

Maximize $p(\theta \mid w_{1:N})$

Most common setting

Dirichlet hyper-parameters	Complexity	Intuition
$\alpha_t \ge 1$	Easy	Maximizing concave function
$\alpha_t < 1$	NP-hard	Reduction from set cover

Sample from $p(\theta \mid w_{1:N})$

Dirichlet hyper-parameters	Complexity	Intuition
$\alpha_t \ge 1$	Easy	Log-concave distribution
$\alpha_t \approx 0$	NP-hard	Reduction from set cover