Complexity of Inference in Latent Dirichlet Allocation

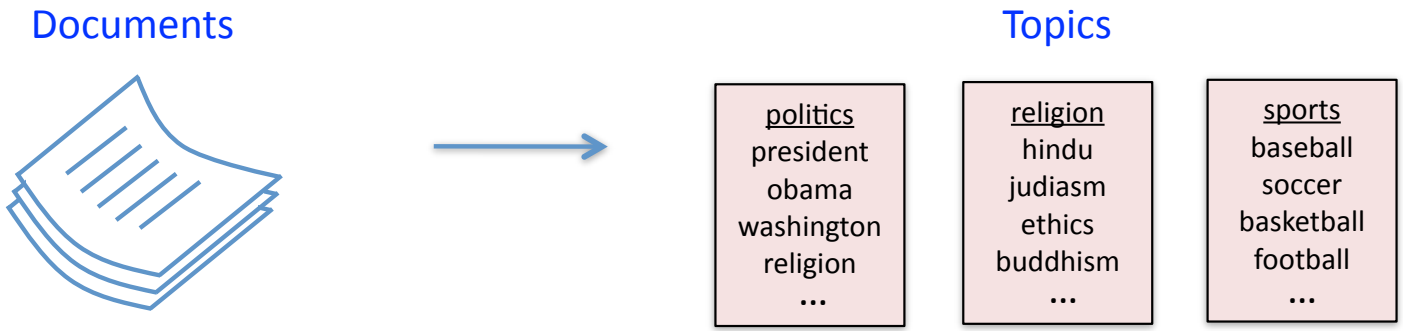
David Sontag (NYU)

Joint work with Daniel Roy (Cambridge)

March 7, 2012

Latent Dirichlet allocation (LDA)

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



- Many applications in information retrieval, document summarization, and classification

- LDA is one of the simplest and most widely used topic models

Generative model for a document in LDA

1. Sample the document's **topic distribution** θ (aka topic vector)

$$\theta \sim \text{Dirichlet}(\alpha_{1:T})$$

where the $\{\alpha_t\}_{t=1}^T$ are fixed hyperparameters. Thus θ is a distribution over T topics with mean $\theta_t = \alpha_t / \sum_{t'} \alpha_{t'}$

2. For $i = 1$ to N, sample the **topic** z_i of the i'th word

$$z_i | \theta \sim \theta$$

3. ... and then sample the actual **word** w_i from the z_i'th topic

$$w_i | z_i, \ldots \sim \beta_{z_i}$$

where $\{\beta_t\}_{t=1}^T$ are the *topics* (a fixed collection of distributions on words)

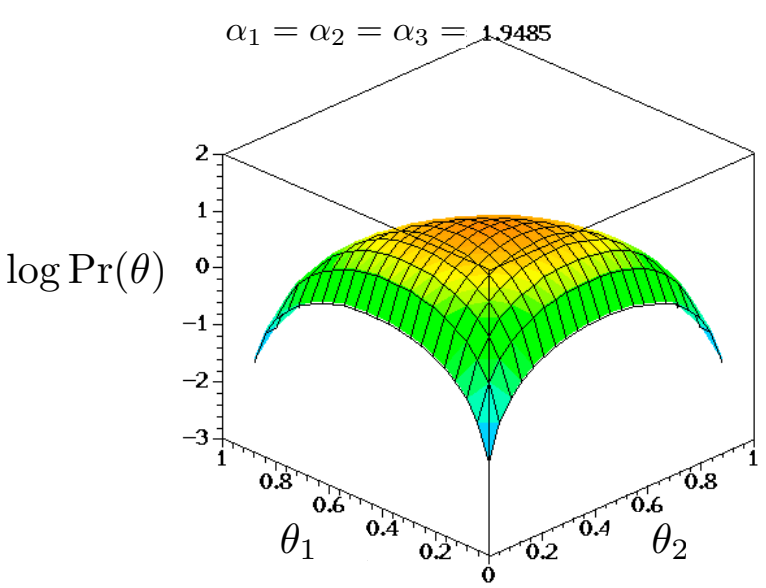
Generative model for a document in LDA

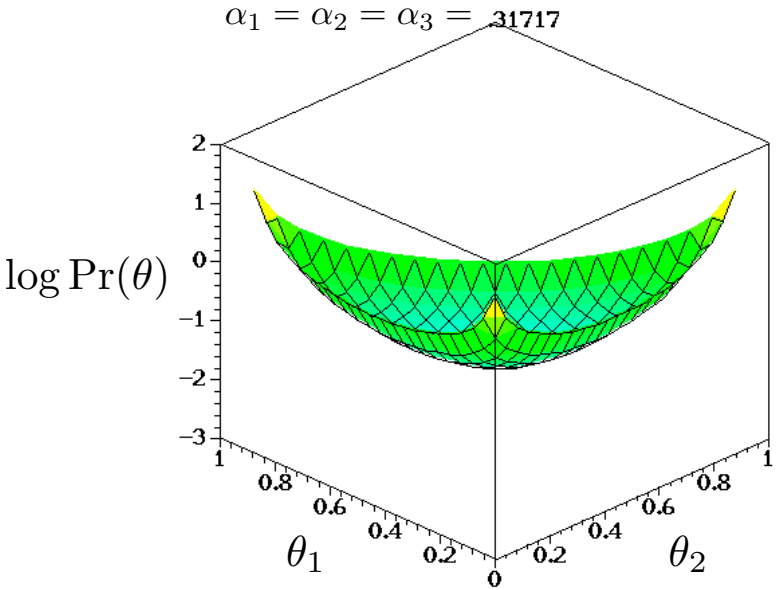
1. Sample the document's **topic distribution** θ (aka topic vector)

$$\theta \sim \text{Dirichlet}(\alpha_{1:T})$$

where the $\{\alpha_t\}_{t=1}^T$ are hyperparameters. The Dirichlet density is:

$$\Pr(\theta_1, \ldots, \theta_T) \propto \prod_{t=1}^{T} \theta_t^{\alpha_t - 1}$$



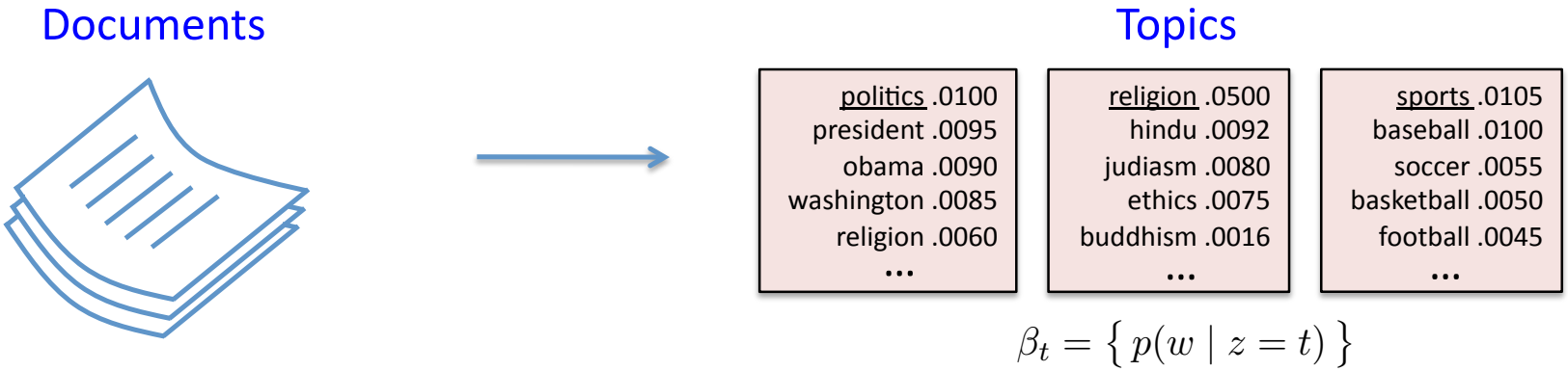


Generative model for a document in LDA

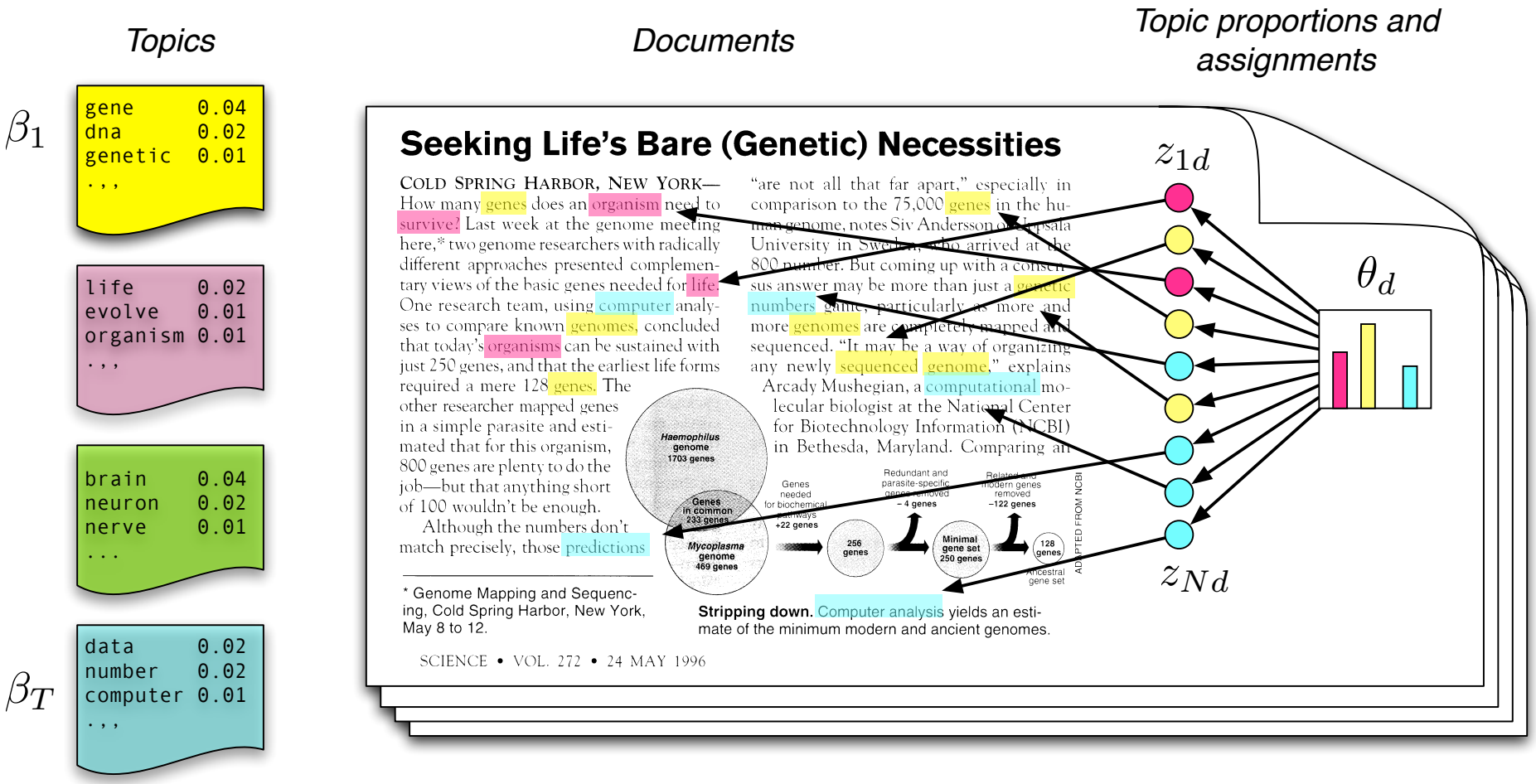
3. ... and then sample the actual **word** w_i from the z_i'th topic

$$w_i | z_i, \ldots \sim \beta_{z_i}$$

where $\{\beta_t\}_{t=1}^T$ are the *topics* (a fixed collection of distributions on words)



Example of using LDA



(Blei, *Introduction to Probabilistic Topic Models*, 2011)

Probabilistic inference in LDA (this talk)

- MAP word-topic assignment (discrete optimization, classification)
  $\max_{z_{1:N}} p(z_{1:N}|w_{1:N})$ (For any $\alpha > 0$)

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

- MAP topic distribution (dimensionality reduction, information retrieval)
  $\max_{\theta} p(\theta|w_{1:N})$

	Dirichlet hyper-parameters	Complexity	Intuition
Most common setting →	$\alpha_t \geq 1$	Easy	Maximizing concave function
	$\alpha_t < 1$	NP-hard	Reduction from set cover

- Sample topic distribution (useful for learning, capturing uncertainty)
  $\sim p(\theta|w_{1:N})$

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

What this talk is **not** about

- This talk is not about **learning**, i.e. the task of finding the topic-word distributions:

politics .0100	religion .0500	sports .0105
president .0095	hindu .0092	baseball .0100
obama .0090	judiasm .0080	soccer .0055
washington .0085	ethics .0075	basketball .0050
religion .0060	buddhism .0016	football .0045
...	...	...

$$\beta_t = \{ p(w \mid z = t) \}$$

- Learning in LDA is also a very interesting question (and open), but is of a different nature:
  - Possible to succeed in learning but still have difficulty with inference
  - For example, often reasonable to assume that there are *some* documents in corpora that are generated from a **single** topic

MAP word-topic assignment – $\max_{z_{1:N}} p(z_{1:N}|w_{1:N})$

- Let n_t be the total # of words assigned to topic t, i.e. $n_t = \sum_{i=1}^{N} 1[z_i = t]$
- The conditional probability of topic assignment $z_{1:N}$ given words $w_{1:N}$ is:

$$\begin{aligned} \Pr(z_1, \ldots, z_N|\mathbf{w}) &\propto \Pr(\mathbf{z}) \Pr(\mathbf{w} \mid \mathbf{z}) \\ &= \frac{\Gamma(\sum_t \alpha_t)}{\prod_t \Gamma(\alpha_t)} \frac{\prod_t \Gamma(n_t + \alpha_t)}{\Gamma(\sum_t \alpha_t + N)} \prod_i \Pr(w_i|z_i) \end{aligned}$$

- Let $l_{it} = \log \Pr(w_i|z_i = t)$ and define $x_{it} = 1[z_i = t]$
- The MAP word-topic assignment problem WORD-LDA(α) is:

$$\begin{aligned} \Phi = \max_{x_{it} \in \{0,1\}, n_t} & \quad \sum_t \log \Gamma(n_t + \alpha_t) + \sum_{i,t} x_{it} l_{it} \\ \text{subject to} & \quad \sum_t x_{it} = 1, \quad \sum_i x_{it} = n_t, \end{aligned}$$

Exact maximization for small # of *effective* topics

$$\Phi = \max_{x_{it} \in \{0,1\}, n_t} \sum_t \log \Gamma(n_t + \alpha_t) + \sum_{i,t} x_{it} l_{it}$$
$$\text{subject to} \quad \sum_t x_{it} = 1, \quad \sum_i x_{it} = n_t,$$

- If topic counts n_t are known, then this is a weighted **b**-matching problem (solvable in polynomial time)
- Suppose τ is the # of topics in the MAP assignment **and is small**
- Try all $\binom{T}{\tau}$ choices for the support of **n**!

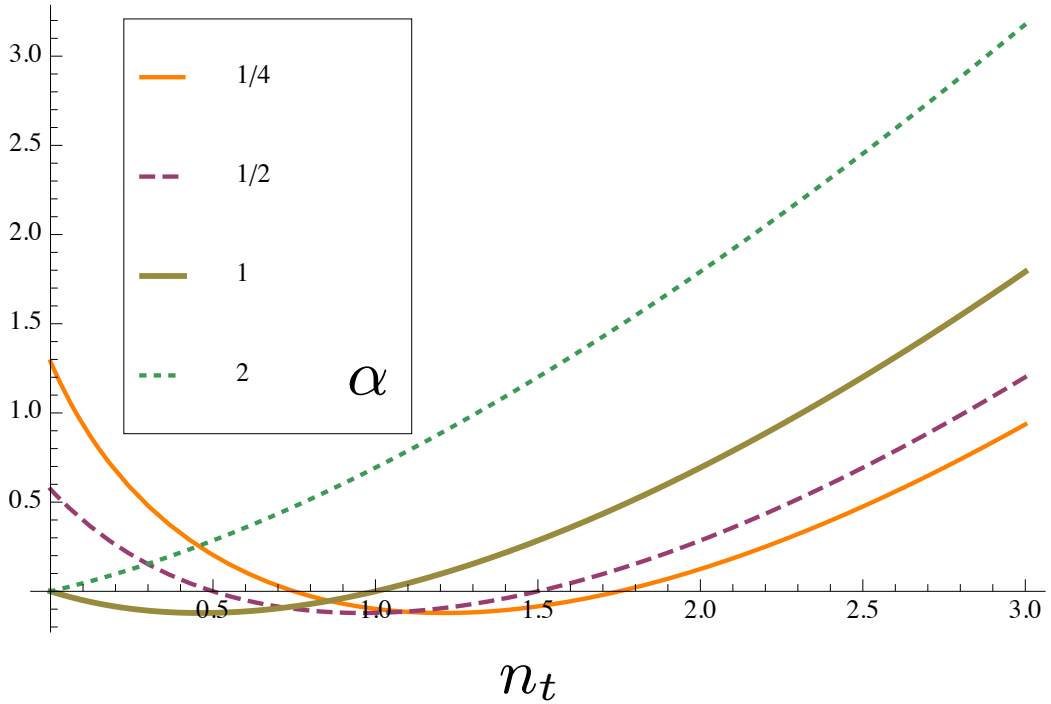
for all subsets $A \subseteq [T]$ such that $|A| = \tau$ **do**
  for all valid partitions $\mathbf{n} = (n_1, n_2, \ldots, n_T)$, i.e., $n_t = 0$ for $t \notin A$ **do**
    $\Phi_{A,\mathbf{n}} \leftarrow \textsc{Weighted-B-Matching}(A, \mathbf{n}, l) + \sum_t \log \Gamma(n_t + \alpha_t)$
  end for
end for
return $\arg\max_{A,\mathbf{n}} \Phi_{A,\mathbf{n}}$

- Total running time is $O((NT)^\tau (N+\tau)^3)$, polynomial in N and T for fixed τ

NP-hard for large # of effective topics

$$\Phi = \max_{x_{it} \in \{0,1\}, n_t} \sum_t \log \Gamma(n_t + \alpha_t) + \sum_{i,t} x_{it} l_{it}$$
$$\text{subject to} \quad \sum_t x_{it} = 1, \quad \sum_i x_{it} = n_t,$$

$$F(n_t) = \log \Gamma(n_t + \alpha)$$

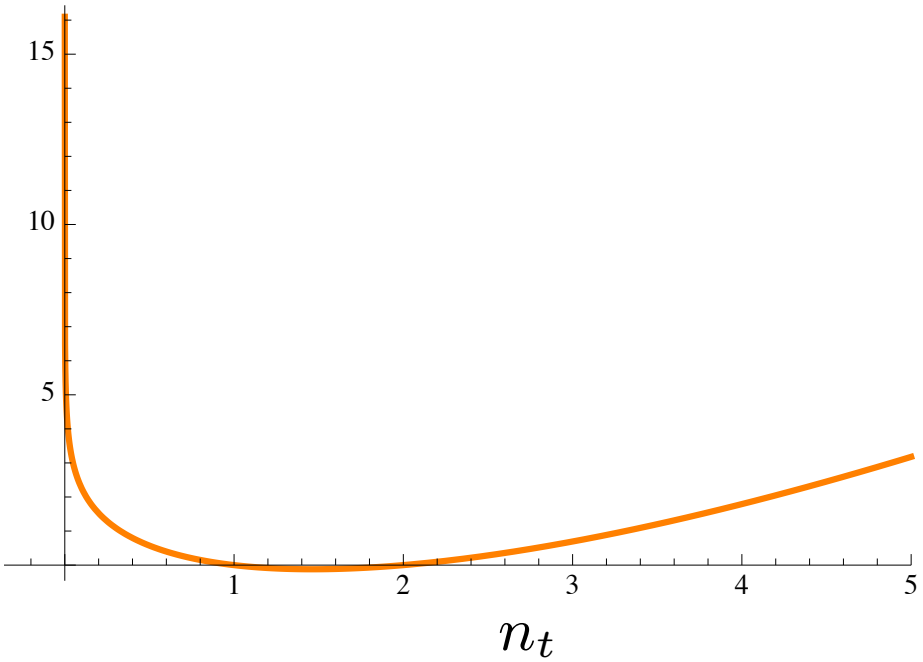


- $F(n_t)$ is strictly convex for $n_t \geq 1$. Preference for larger topics
- When $\alpha < 1$, $F(0)$ is large, giving a strong sparsity reward

NP-hard for large # of effective topics

$$\Phi = \max_{x_{it} \in \{0,1\}, n_t} \sum_t \log \Gamma(n_t + \alpha_t) + \sum_{i,t} x_{it} l_{it}$$
$$\text{subject to} \quad \sum_t x_{it} = 1, \quad \sum_i x_{it} = n_t,$$

Behavior of $\log \Gamma(n_t + \alpha)$ as $\alpha \to 0$



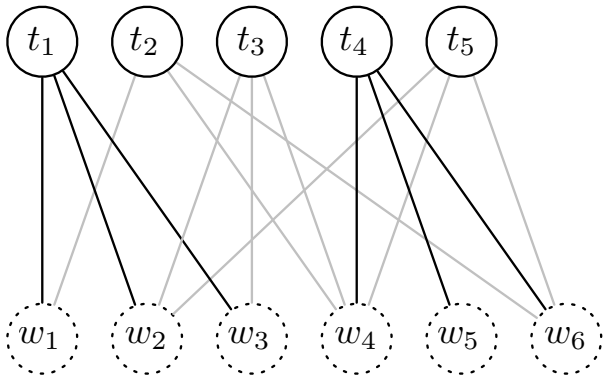
- $F(n_t)$ is strictly convex for $n_t \geq 1$. Preference for larger topics
- When $\alpha < 1$, $F(0)$ is large, giving a strong sparsity reward

NP-hard for large # of effective topics

- Reduction from k-set packing: given a collection of k-element sets, find largest collection of *disjoint* sets:

$$\{\mathbf{1},\mathbf{2},\mathbf{3}\}\ \ \{1,2,4\}\ \ \{\mathbf{4},\mathbf{5},\mathbf{6}\}\ \ \{1,3,5\}$$

- For some constant $c > 1$, NP-hard to decide whether there is a solution with n/k disjoint sets (covering *all* elements), or at most cn/k disjoint sets
- Reduction is as follows (document consists of all words):



One topic for each set

$\Pr(w \mid t) = 0$ if word not in set

$= \frac{1}{k}$ otherwise

One word for each element

- If a perfect matching exists, MAP assignment will find it (because it uses as few topics as possible)

MAP topic distribution – $\max_\theta p(\theta|w_{1:N})$

- Let $\psi_{it} = \Pr(w_i|z_i = t)$. By Bayes' rule, we have

$$\begin{aligned} p(\theta|\mathbf{w}) &\propto p(\theta)\prod_i p(w_i \mid \theta) \\ &\propto (\prod_t \theta_t^{\alpha_t - 1})(\prod_i \sum_t \theta_t \psi_{it}) \end{aligned}$$

- Taking the log and ignoring constants, we obtain the MAP topic distribution problem:

$$\begin{aligned} \max_\theta \ & \textstyle\sum_t (\alpha_t - 1)\log(\theta_t) + \sum_i \log(\sum_t \theta_t \psi_{it}) \\ \text{subject to } & \textstyle\sum_t \theta_t = 1, \quad 0 \le \theta_t \le 1 \end{aligned}$$

- When $\alpha_t \geq 1$ for $t = 1 \ldots T$, objective is **concave** in θ
- Can solve in polynomial time, e.g. using exponentiated gradient (Kivinen and Warmuth, 1995)
- When $\alpha_t < 1$, objective becomes degenerate – left-hand side becomes ∞ for $\theta_t = 0$, overwhelming the likelihood term

MAP topic distribution – $\max_\theta p(\theta|w_{1:N})$

- To prevent this degeneracy, we restrict θ_t to be bounded below by ϵ. The TOPIC-LDA(ϵ, α) problem is:

$$\begin{aligned} \max_\theta \ & \sum_t (\alpha_t - 1) \log(\theta_t) + \sum_i \log(\sum_t \theta_t \psi_{it}) \\ \text{subject to } & \sum_t \theta_t = 1, \quad \epsilon \leq \theta_t \leq 1. \end{aligned}$$

- Most common scenario is $\alpha < 1$. For example, learning LDA model on corpus of NIPS abstracts with $T = 200$, median value is $\alpha_t = 0.01$
- Even though non-convex for $\alpha < 1$, useful approximate inference algorithms may still be obtained by performing local search
- One applicable algorithm, for example, is the Concave-Convex Procedure (Yuille and Rangarajan, 2003)

MAP topic distribution – $\max_\theta p(\theta|w_{1:N})$

$$\max_\theta \ \sum_t (\alpha_t - 1)\log(\theta_t) + \sum_i \log(\sum_t \theta_t \psi_{it})$$
$$\text{subject to } \sum_t \theta_t = 1, \quad \epsilon \le \theta_t \le 1.$$

- Define the *dynamic range* of word w_i to be $\kappa_i = \max_{t,t':\psi_{it},\psi_{it'}>0} \frac{\psi_{it}}{\psi_{it'}}$
- Let $\kappa = \max_i \kappa_i$
- **Small hyperparameters encourage sparsity**:

Theorem

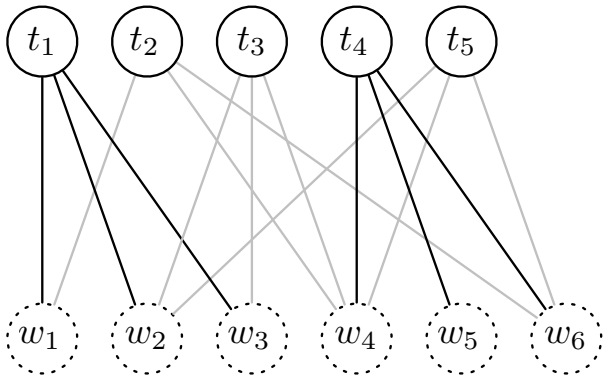
For $\alpha < 1$, all optimal solutions to TOPIC-LDA(ϵ, α) have $\theta_t \le \left(e^{\frac{1}{1-\alpha}+2}\right)\epsilon$ or $\theta_t \ge \kappa^{-1} e^{-3/\alpha} N^{-2} T^{-1/\alpha}$.

- Thus, solving TOPIC-LDA(ϵ, α) corresponds to finding the non-trivial **support** of θ
- Motivates greedy algorithms for approximately maximizing TOPIC-LDA(ϵ, α), analogous to set cover

TOPIC-LDA(ϵ, α) is NP-hard for $\alpha < 1$ and $\epsilon = o\left((NT)^{-T}\right)$

$$\max_{\theta} \ \sum_t (\alpha_t - 1) \log(\theta_t) + \sum_i \log(\sum_t \theta_t \psi_{it})$$
$$\text{subject to } \sum_t \theta_t = 1, \quad \epsilon \leq \theta_t \leq 1.$$

- Reduction from **set cover** (again, document consists of all elements):



One topic for each set

$\Pr(w \mid t) = 0$ if word not in set

$= c$ otherwise

One word for each element

- Introduce dummy words (not in document) to force $\Pr(w \mid t)$ to be a constant
- Support of the MAP topic distribution θ (topics having non-negligible probability) corresponds to the minimal set cover
- Proof requires ϵ to be exponentially small in # words N and # topics T

Sampling from the posterior

- $\alpha_t \geq 1$: Can approximately sample from $p(\theta|w_{1:N})$ in polynomial time
  - Density is log-concave when $\alpha_t \geq 1$
  - Use algorithm from Lovasz and Vempala (2006) based on random walks
- $\alpha_t < (NT)^{-N}$: NP-hard to approximately sample from $p(\theta|w_{1:N})$
  - Reduction from set cover
  - Non-trivial posterior probability given to sparsest possible θ vectors, so set cover can be read off from marginals
  - Would need a *very large and unusual corpus* to learn such a small α
- **Open**: computational complexity for α constant (less than 1)

Discussion

- Possible to give a poly-time approx. algorithm for MAP $p(\theta \mid \mathbf{w})$ when effective number of topics per document is constant
- $\alpha \geq 1$: MAP for $p(\mathbf{z} \mid \mathbf{w})$ NP-hard, whereas $p(\theta \mid \mathbf{w})$ easy. Why?
- Can approximately *sample* from $p(\mathbf{z} \mid \mathbf{w})$ in polynomial time for $\alpha \geq 1$
- Connections between inference in topic models and sparse signal recovery (see also recent work by Zhu & Xing, UAI '11)
- Motivates study of greedy algorithms for MAP inference of topic distribution, analogous to those used for set cover