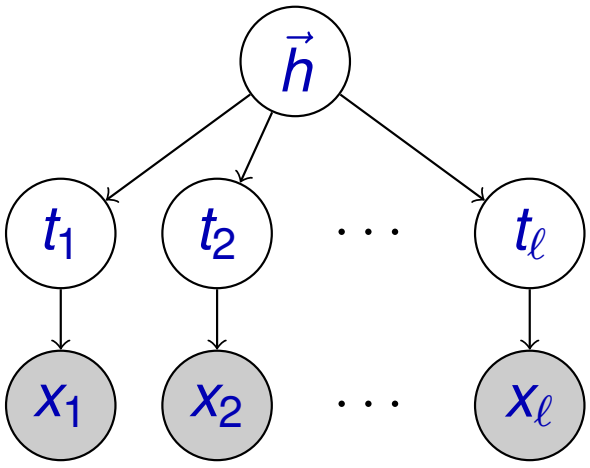
A Spectral Algorithm for Latent Dirichlet Allocation

Anima Anandkumar Dean P. Foster **Daniel Hsu**
Sham M. Kakade Yi-Kai Liu

Simplified LDA model



Parameters ($k = $ # topics, $d = $ # words)

$$\vec{\alpha} \in \mathbb{R}^k_{>0}, \quad \underbrace{\vec{\beta}_1, \vec{\beta}_2, \ldots, \vec{\beta}_k}_{\text{topic-word distributions}} \in \Delta^{d-1}$$

$$\vec{h} \sim \text{Dirichlet}(\vec{\alpha});$$

$$t_i | \vec{h} \sim \text{Discrete}(\vec{h}), \quad i \in [\ell];$$

$$x_i | t_i \sim \text{Discrete}(\vec{\beta}_{t_i}), \quad i \in [\ell].$$

(Assume each document has ℓ words.)

Learning LDA parameters

Questions:
1. Can we learn LDA parameters from documents? **Yes.**
2. How long do the documents have to be?

- **One** word / document: impossible.
- **Two** words / document: possible, under separation conditions (Arora-Ge-Moitra, FOCS 2012).
- **Three** words / document: possible, when each document is about a single topic (Anandkumar-Hsu-Kakade, COLT 2012).

We show that **three** words / document suffices even for LDA, for any $\vec{\alpha}$, without any separation condition.

Main result

Write $\vec{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_k)$ and define $\alpha_0 := \sum_{j=1}^{k} \alpha_j$.

Theorem

There is an efficient algorithm that — given α_0 and third-order (cross) moments of $x_1, x_2, \ldots, x_\ell$ — recovers the model parameters $\vec{\alpha}$ and $\{\vec{\beta}_1, \vec{\beta}_2, \ldots, \vec{\beta}_k\}$ exactly.

- Third-order cross moments $\equiv$ frequencies of word-triples.
- Also have sample complexity bound that is **polynomial** in relevant parameters.
- Computation is **linear** in number of non-zero entries of term-document matrix.

Poster W66