**Provably and Practically Learning Topic Models**

Yoni Halpern¹, Sanjeev Arora², Rong Ge², Ankur Moitra², David Mimno², Yichen Wu², Michael Zhu² and David Sontag¹


**Abstract**

We improve on the provable method of learning topic models with anchor words described in Arora et al. (2012).

The new algorithm combines strong theoretical results and practicality:

- **Provable** - Robust to noise
- **Fast** - High quality topics

**Representation**

- Describe each word by a distribution over co-occurring words.
- Simply computed from counts.

**Finding Anchor Words**

In this representation, anchors lie at the corners of a low dimensional simplex.

A search for extreme points finds anchors in $O(K\log V)$.

**Recovering Topics**

Probability of topic conditioned on observing word $i$ is related to the **position** of a word in the simplex.

$$P(z|w = i) = \arg\min_\alpha \|Q_i - \sum_k \alpha_k Q_k\|$$

s.t. $\alpha_k \geq 0$, $\sum_k \alpha_k = 1$

- Convex optimization problem solved using the **exponentiated gradient** algorithm.
- Inversion with Bayes rule.

**Guarantees**

- Method-of-moments algorithm to recover topics from a corpus of documents.
- If data is drawn from topic model distribution: the original distribution can be recovered to **arbitrary accuracy** with a polynomial number of samples.

**Simulations**

- Real Data Evaluation

Example topics in New York Times
- cell, stem, research, Bush, patent
- cup, minutes, add water, tablespoon oil
- oil, prices, percent, million, United States
- anthrax, official, mail, letter, worker, attack

**References**

- Arora, S., Ge, R., and Moitra, A. Learning Topic Models – Going Beyond SVD. In FOCS, 2012

---

**Topic Models**

Explore
- Generative model to describe collections of documents.
- Each “topic” describes a distribution of words.
- Bag of words for each document from a distribution of topics.

Summarize/Annotate
- Additional anchor assumptions: For every topic there is at least one word that can only come from that topic.