

Relative Performance Guarantees for Approximate Inference in Latent Dirichlet Allocation

Indraneel Mukherjee, David M. Blei

Princeton University, USA

Introduction

Hierarchical probabilistic modeling of discrete data has emerged as a powerful tool for text analysis. Posterior inference in such models is intractable, and practitioners rely on approximate posterior inference methods such as variational inference or Gibbs sampling. We analyze the improvement that the recently proposed collapsed variational inference (CVB) provides over mean field variational inference (VB) in latent Dirichlet allocation. We prove that the advantage is lost for long documents but increases with the number of topics.

Latent Dirichlet Allocation

- ▶ Models text corpora as topic mixtures.
- ▶ Each topic is a distribution over vocabulary.
- ▶ Dirichlet prior over topic mixtures.
- ▶ Documents are a bag of m words $x_{1:m}$, and generated as follows:
 - ▶ Sample topic mixture θ from the Dirichlet $\mathcal{D}(\vec{\alpha})$
 - ▶ For each word, sample topic assignment $z_i \sim \theta$
 - ▶ Sample word x_i from topic z_i . This describes a joint probability distribution of the observed and latent variables $p(\vec{x}, \vec{z}, \theta | \vec{\alpha}, \beta)$.

Inference in LDA

- ▶ Given document \vec{x} , compute posterior distribution over latent variables $p(\theta, \vec{z} | \vec{x})$.
- ▶ Difficulty: computationally intractable normalizing constant $p(\vec{x})$.
- ▶ Variational approximation: Output *best* option q from tractable family of distributions over latent variables. The best choice minimizes:
 - ▶ Relative entropy distance to true posterior, $\text{RE}(q(\theta, \vec{z}) || p(\theta, \vec{z} | \vec{x}))$, or equivalently,
 - ▶ minimizes the upper-bound approximation, also known as **variational free energy**, to the negative log-normalization constant:

$$\mathbb{E}_q \left[\log \frac{q(\theta, \vec{z})}{p(\theta, \vec{z}, \vec{x})} \right]$$

Variational Inference

In the variational inference algorithm for LDA (VB), the posterior $p(\theta, \vec{z} | \vec{x})$ is approximated by a fully-factorized variational distribution

$$q(\theta, \vec{z} | \vec{\gamma}, \phi_{1:m}) = q(\theta | \vec{\gamma}) \prod_i q(z_i | \phi_i).$$

Collapsed Variational Inference

The collapsed variational inference algorithm (CVB) reformulates the LDA model by marginalizing out the topic proportions θ . This yields a formulation where the topic assignments z are fully dependent, but where the dimensionality of the latent space has been reduced. The variational family in CVB is a fully-factorized product of multinomial distributions,

$$q(z | \phi_{1:m}) = \prod_i q(z_i | \phi_i).$$

Theorem

Consider any LDA model with k topics, and a document consisting of m words x_1, \dots, x_m , where m is sufficiently large. Let $VB(\vec{x})$ and $CVB(\vec{x})$ be the free energies measured by VB and CVB respectively. Then,

$$0 \leq [VB(\vec{x}) - CVB(\vec{x})] \leq O(k-1) + o(1) \quad (1)$$

for this model. Here $o(1)$ goes to 0 at least as fast as $\sqrt{\frac{\log m}{m}}$.

Discussion

- ▶ Our main result implies that the **per word free energy change**, as well as the percentage free energy change, **between VB and CVB goes to zero with the length of the document**.
- ▶ Our results are stated in log-space. Since the probability of a document falls exponentially fast with the number of words, the additive difference in the probability estimates of VB and CVB is again negligible for large documents.
- ▶ A strength of the above **Theorem** is that it **holds for any document, and not necessarily one generated by an LDA model**.
- ▶ The upper-bound in (1) is nearly tight. When all topics are uniform distributions, the difference in the free energy estimates is $\Omega(k)$ for long documents.

Proof ingredients

- ▶ In principle, both CVB and VB try to approximate $p(\theta, \vec{z} | \vec{x}) = p(\vec{z} | \vec{x})p(\theta | \vec{z})$. By collapsing out θ , CVB can exactly approximate $p(\theta | \vec{z})$, whereas VB has to commit to a fixed distribution $q(\theta)$. The additional error suffered by VB is then bounded by

$$\mathbb{E}_{q(\vec{z}, \theta)} \log \frac{q(\theta)}{p(\theta | \vec{z})}.$$

i.e. **the average deviation of $q(\theta)$ from $p(\theta | \vec{z})$** for the best choice of $q(\theta)$.

- ▶ The sufficient statistic $T(\vec{z})$ of \vec{z} wrt θ is the k -tuple denoting the occurrence frequency of each topic in \vec{z} . This tuple exhibits a **concentration type phenomenon**

$$\Pr[|T(\vec{z}) - \mu| > c\sigma] < \exp(-c^2) \\ \sigma^2 = O(\mu)$$

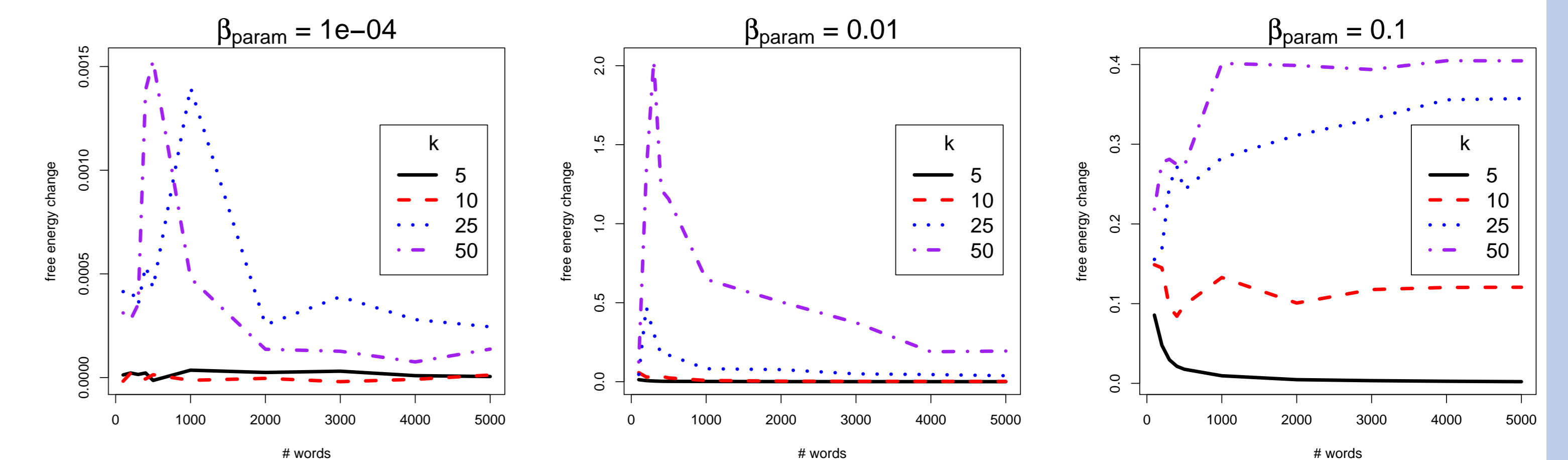
where μ, σ are the mean and s.d. of $T(\vec{z})$ under $q(\vec{z})$.

- ▶ **The function $\log p(\vec{z})$ is weakly convex**. This translates to upper-bounds on the eigenvalues of the Hessian of this function.

Summary and Future Work

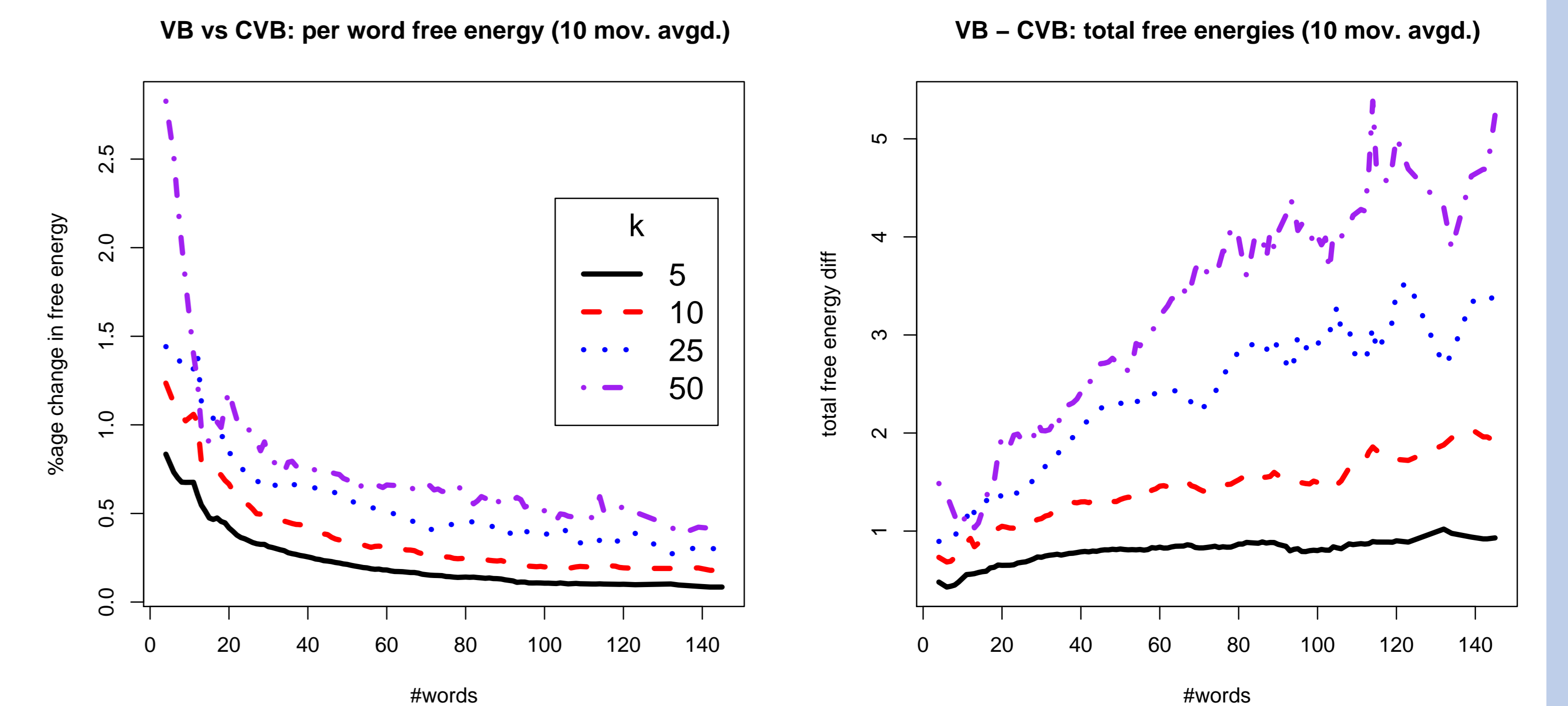
- ▶ The following practical guideline emerges from our work: CVB possesses a distinct advantage over VB for short documents, but for longer documents one may switch to the significantly faster VB algorithm with very little added penalty.
- ▶ Our results hold more generally for exponential families satisfying some conditions, in particular Hierarchical Dirichlet Processes.
- ▶ In the future, we plan to examine more closely the effect of sparsity of topics on the relative performance of VB and CVB.

Experiments on synthetic data



Results on synthetic text data. We sample k topics from a symmetric Dirichlet distribution with parameter β_{param} , and generate documents from LDA models with these topics. The VB and CVB per-word free energies for different prefix lengths are averaged over these documents. The curves obtained show how the advantage of CVB over VB changes with the length of a document, number of topics and sparsity of topics.

Experiments on real data



(Above) Arxiv data-set [2000 short documents]

(Below) Yale-Law data-set [200 documents of length 1000 – 10000].

Left and right columns contain per-word and total free energy difference plots, resp.

