Xi-ming Li, Ji-hong Ouyang, You Lu, 2015. Topic modeling for large-scale text data. *Frontiers of Information Technology & Electronic Engineering*, **16**(6):457-465. [doi:10.1631/FITEE.1400352]

Front Inform Technol & Electron Eng

Topic modeling for large-scale text data

Key words: Latent Dirichlet allocation (LDA), Topic modeling, Online learning, Moving average

Contact: Xi-ming Li
E-mail: liximing86@gmail.com
ORCID: http://orcid.org/0000-0001-8190-5087

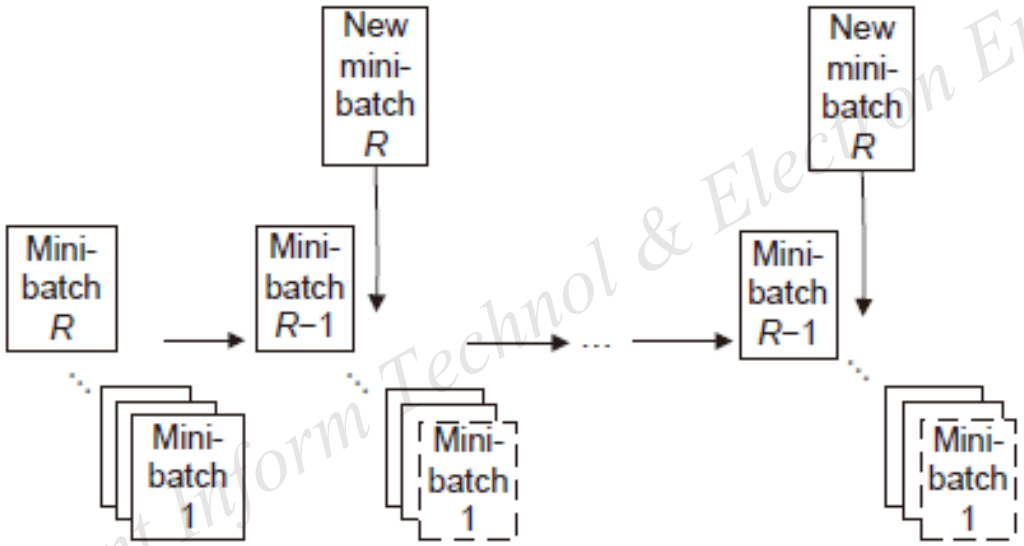
Introduction

- Topic models such as latent Dirichlet allocation (LDA) are mainstays for modern data analysis.

- Stochastic variational inference (SVI) can efficiently infer topic models for large-scale and online data. However, the noise of stochastic natural gradients is commonly large.

Front Inform Technol & Electron Eng

Moving average stochastic variational inference



Front Inform Technol & Electron Eng

Fig. 2 A graphic illustration for the moving average scheme (at each iteration, we sample a new mini-batch R and discard old mini-batch 1)

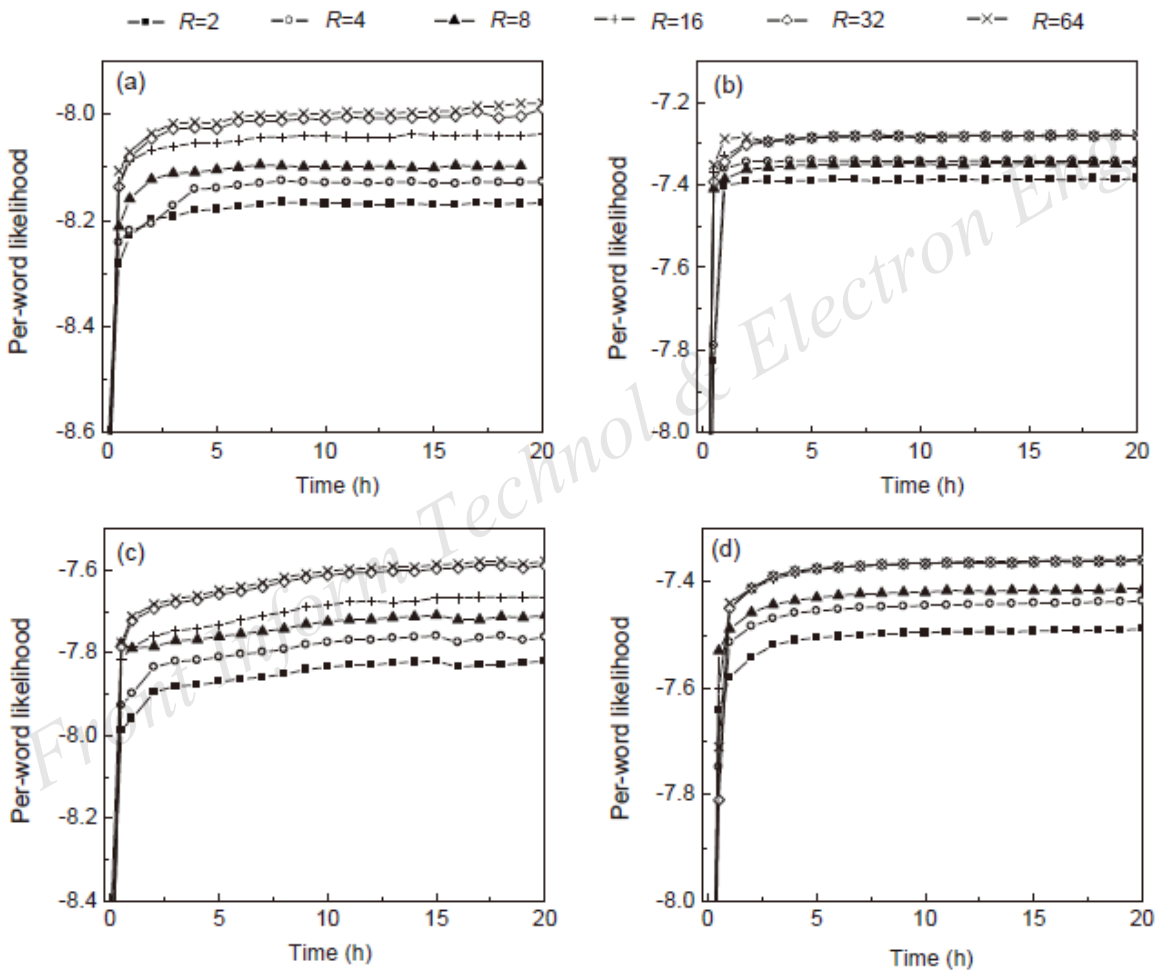
Moving average stochastic variational inference

Algorithm 1 MASVI for LDA

1: Initialize parameters, including α, β, ρ, M, and R
2: Generate $\tilde{\beta}^{(0)}$, and then initialize $f^{(0)}$
3: **For** $t = 1, 2, \cdots, \infty$ **do**
4: Sample M documents
5: **For** d=1 to M **do**
6: Compute $\tilde{\alpha}_d$ and $\tilde{\theta}_d$ using Eqs. (2) and (3)
7: **End for**
8: Update the moving average $f^{(t)}$ using Eq. (7)
9: Update $\tilde{\beta}^{(t)}$ using Eq. (9)
10: **End for**

Front Inform Technol & Electron Eng

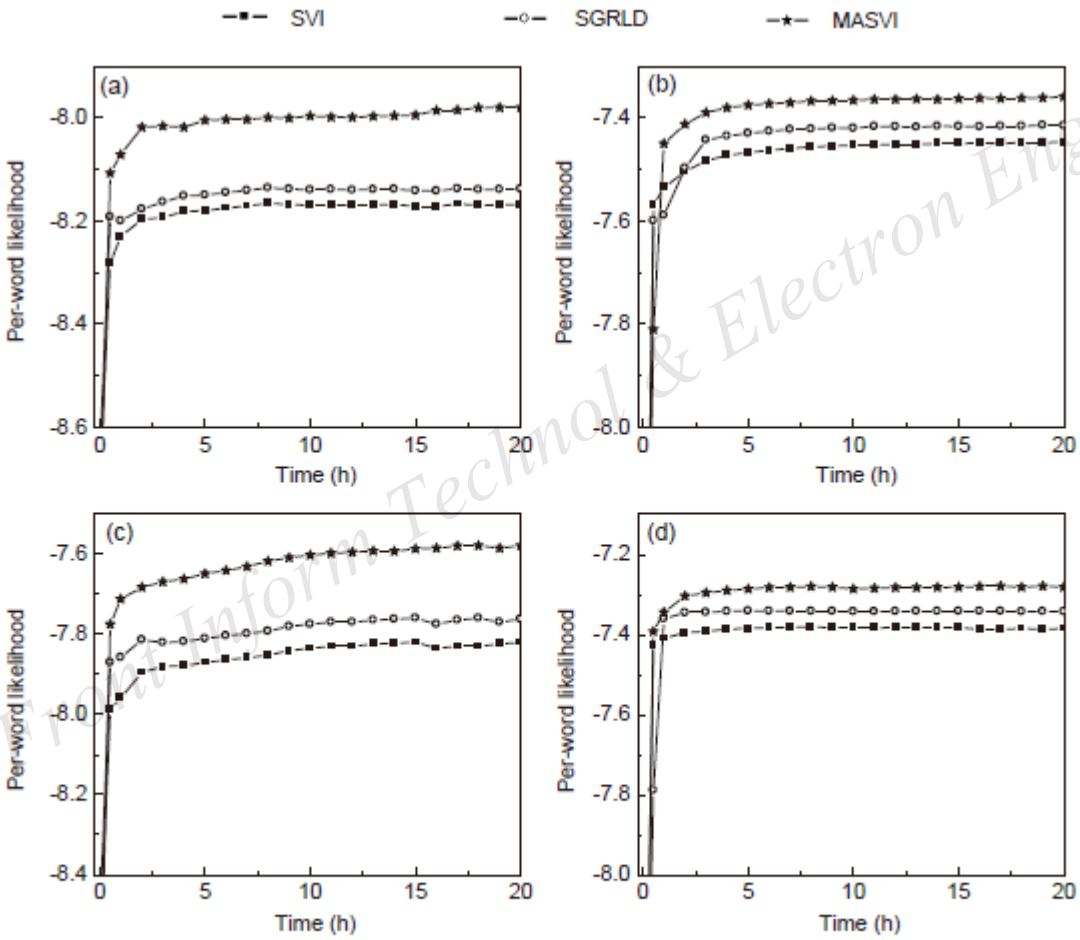
Experimental results (1)



Front. Optoelectron. Technol. & Electron Eng.

Fig. 3 Experiments on moving frequency R: (a) PubMed ($M = 100$); (b) Wikipedia ($M = 100$); (c) PubMed ($M = 500$); (d) Wikipedia ($M = 500$)

Experimental results (2)



Front Inform Technol & Electron Eng

Fig. 4 A 100-topic LDA inference: (a) PubMed ($M = 100$); (b) Wikipedia ($M = 100$); (c) PubMed ($M = 500$); (d) Wikipedia ($M = 500$)

Conclusions

- We developed a novel online inference algorithm for LDA, namely moving average stochastic variational inference.
- The experimental results showed that the proposed algorithm outperforms the state-of-the-art SVI and SGRLD.

Front Inform Technol & Electron Eng