A trust-region method for stochastic variational inference with applications to streaming data

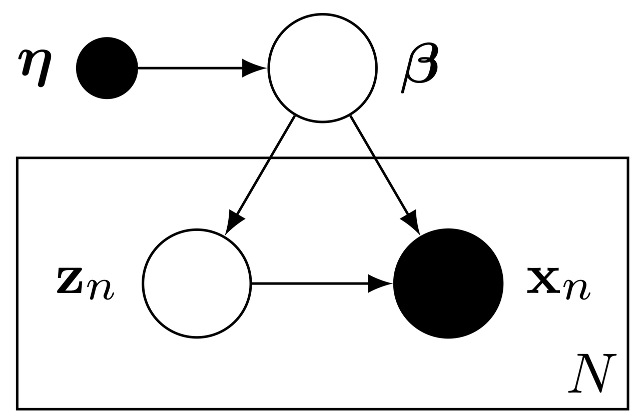
Lucas Theis[1], Matthew D. Hoffman[2]

[1] Graduate School of Neural Information Processing, University of Tübingen
[2] Adobe Research, San Francisco

Introduction

Stochastic variational inference (SVI) has enabled fast posterior inference on massive datasets for a large class of complex Bayesian models. However, SVI can be sensitive to the choice of hyper-parameters and is prone to local optima [1, 2, 3]. We introduce a new trust region based variant of SVI which is more robust. We also show how to apply SVI to streaming data and find that here our trust-region method is crucial.

Models



We here assume an exponential family over **local parameters** $\mathbf{z}$ and observations $\mathbf{x}$ and a conjugate exponential family prior over **global parameters** $\boldsymbol{\beta}$.

$$p(\boldsymbol{\beta}) = h(\boldsymbol{\beta}) \exp\left(\boldsymbol{\eta}^\top t(\boldsymbol{\beta}) - a(\boldsymbol{\eta})\right)$$
$$p(\mathbf{x}, \mathbf{z} \mid \boldsymbol{\beta}) = \prod_n h(\mathbf{x}_n, \mathbf{z}_n) \exp\left(t(\boldsymbol{\beta})^\top f(\mathbf{x}_n, \mathbf{z}_n)\right)$$

Instances of this model include, for example, latent Dirichlet allocation, mixture models, HMMs, probabilistic matrix factorizations, and hierarchical linear and probit regression models.

Mean-field approximation

$$q(\boldsymbol{\beta}, \mathbf{z}) = q(\boldsymbol{\beta}) \prod_{n,m} q(z_{nm}),$$
$$q(\boldsymbol{\beta}) = h(\boldsymbol{\beta}) \exp\left(\boldsymbol{\lambda}^\top t(\boldsymbol{\beta}) - a(\boldsymbol{\lambda})\right),$$

where $q(z_{nm})$ is controlled by ϕ_{nm}. Our goal is to optimize the lower bound on the log-likelihood:

$$\mathcal{L}(\boldsymbol{\lambda}) = \mathbb{E}_q\left[\log \frac{p(\boldsymbol{\beta})}{q(\boldsymbol{\beta})}\right] + \max_{\phi} \sum_{n=1}^{N} \mathbb{E}_q\left[\log \frac{p(\mathbf{x}_n, \mathbf{z}_n \mid \boldsymbol{\beta})}{q(\mathbf{z}_n)}\right]$$

Stochastic variational inference

Stochastic variational inference optimizes the lower bound by following a stochastic approximation,

$$\mathcal{L}_n(\boldsymbol{\lambda}) = \mathbb{E}_q\left[\log \frac{p(\boldsymbol{\beta})}{q(\boldsymbol{\beta})}\right] + N \max_{\phi_n} \mathbb{E}_q\left[\log \frac{p(\mathbf{x}_n, \mathbf{z}_n \mid \boldsymbol{\beta})}{q(\mathbf{z}_n)}\right],$$

via natural gradient ascent,

$$\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t + \rho_t \mathcal{I}(\boldsymbol{\lambda}_t)^{-1} \nabla \mathcal{L}_n(\boldsymbol{\lambda}_t).$$

repeat
    select n
    $\phi_n^* \leftarrow \operatorname{argmax}_{\phi_n} \mathcal{L}_n(\boldsymbol{\lambda}_t, \phi_n)$
    $\boldsymbol{\lambda}_{t+1} \leftarrow (1-\rho_t)\boldsymbol{\lambda}_t + \rho_t\left(\boldsymbol{\eta} + N\mathbb{E}_{\phi^*}[f(\mathbf{x}_n, \mathbf{z}_n)]\right)$

Trust-region method

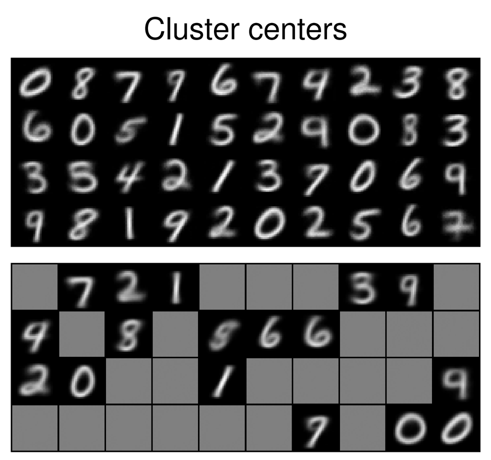
We propose to perform the following update instead:

$$\boldsymbol{\lambda}_{t+1} = \operatorname{argmax}_{\boldsymbol{\lambda}} \left\{\mathcal{L}_n(\boldsymbol{\lambda}) - \xi_t D_{\mathrm{KL}}(\boldsymbol{\lambda}, \boldsymbol{\lambda}_t)\right\}$$

repeat
    select n
    initialize ϕ_n^*
    repeat
        $\boldsymbol{\lambda} \leftarrow (1-\rho_t)\boldsymbol{\lambda}_t + \rho_t\left(\boldsymbol{\eta} + N\mathbb{E}_{\phi^*}[f(\mathbf{x}_n, \mathbf{z}_n)]\right)$
        $\phi_n^* \leftarrow \operatorname{argmax}_{\phi_n} \mathcal{L}_n(\boldsymbol{\lambda}, \phi_n)$
    $\boldsymbol{\lambda}_{t+1} \leftarrow \boldsymbol{\lambda}$

MNIST

We trained a mixture of multivariate Bernoulli distributions on binarized MNIST digits. Left are the expected cluster means.



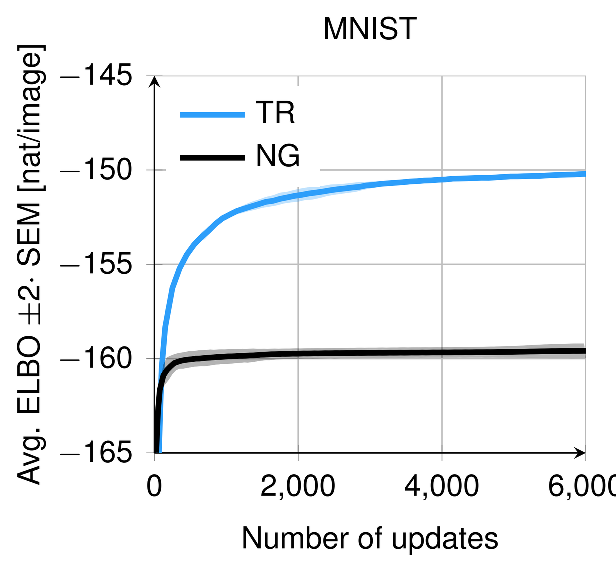
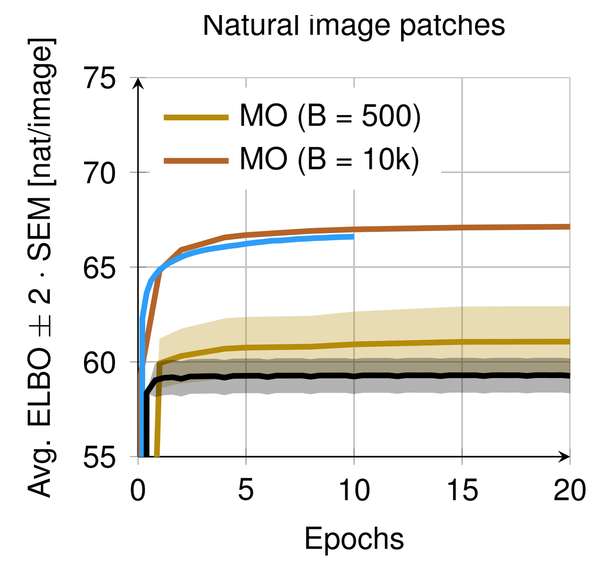


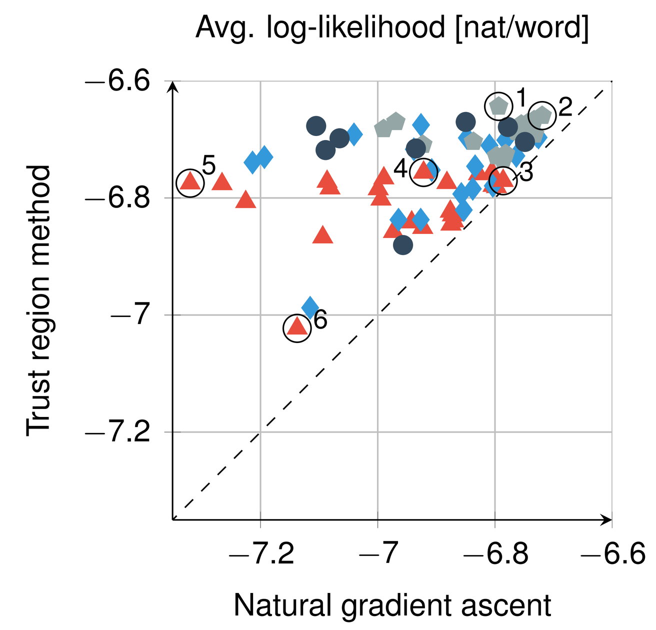
Image patches

We trained Gaussian mixtures on 8x8 image patches and compared our method to **memoized variational inference** (MO) [2].



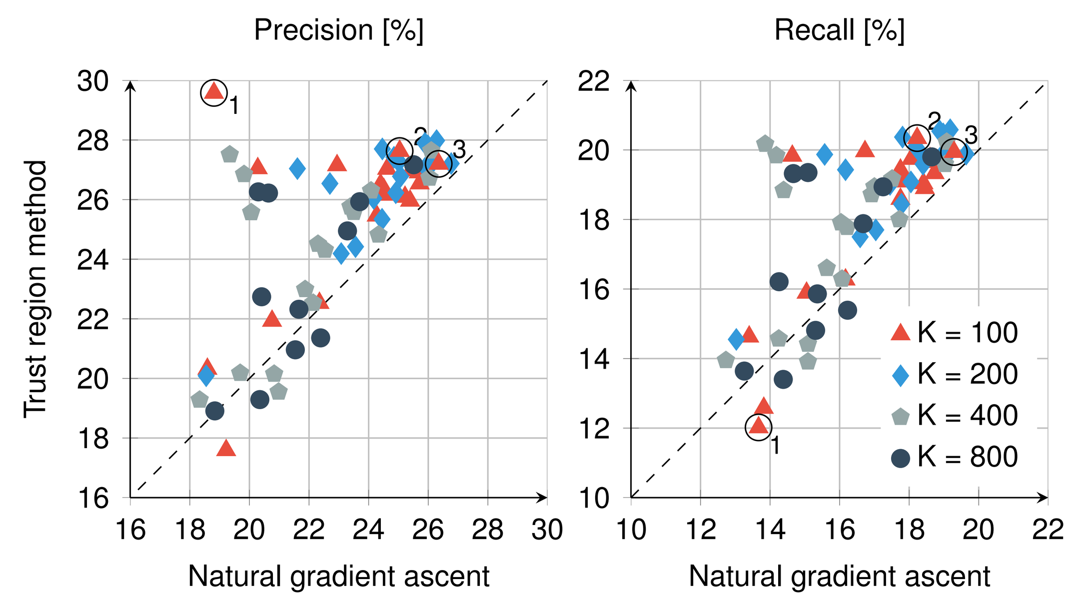
Topic modeling

We trained **latent Dirichlet allocation** (LDA) on Wikipedia articles using the same random hyperparameter settings for natural gradient and trust-region updates.



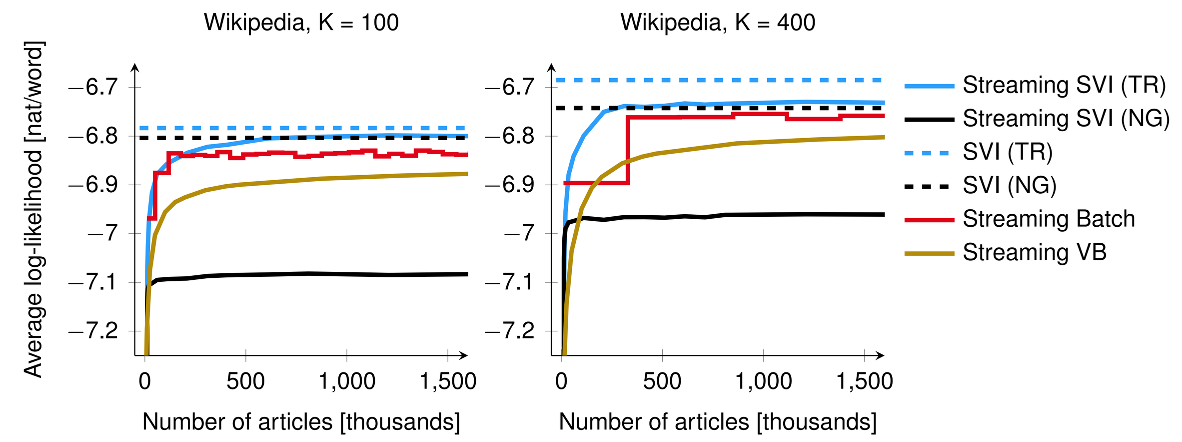
Collaborative filtering

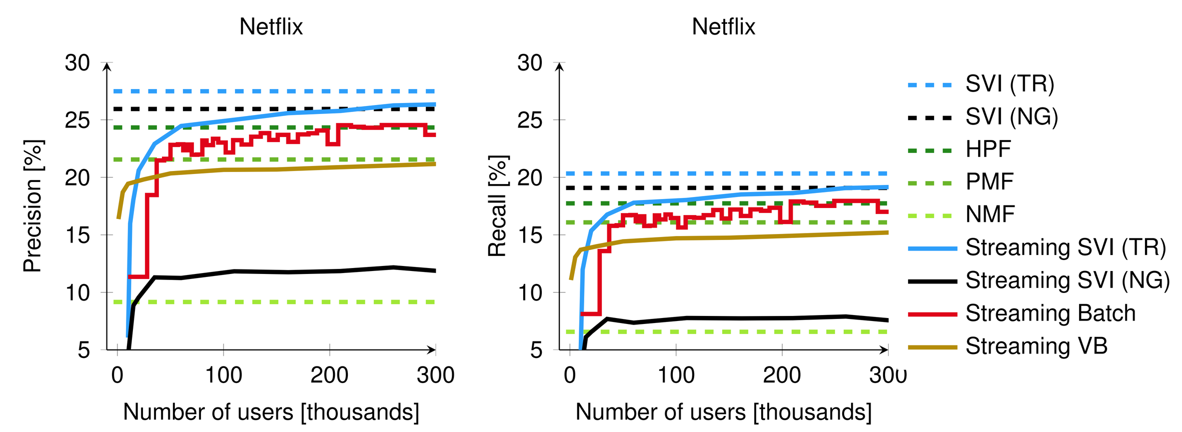
As above, but this time applying LDA to Netflix.



Streaming SVI

Instead of applying SVI to a fixed dataset, we apply SVI to a continuously growing dataset (word by word, or movie by movie). We compare with **streaming variational Bayes** [4]. **Empirical Bayes** was used to automatically tune some of the hyperparameters.





Conclusions and remarks

- SVI with trust-region updates is more **robust** to the choice of hyperparameters and generally performs a little better.
- SVI works well for **streaming** data with trust-region updates but not with natural gradient steps.
- What's missing: **automatic tuning of learning rates** for streaming data.

Resources

Code for training latent Dirichlet allocation:

http://github.com/lucastheis/trlda/

References

[1] R. Ranganath et al., ICML, 2013
[2] M.C. Hughes and E. B. Sudderth, NIPS, 2013
[3] M. D. Hoffman and D. M. Blei, JMLR, 2014
[4] T. Broderick et al., NIPS, 2013