LSVI: On-line supplement

Anonymous

March 29, 2025

Abstract

This document reports the results from extra experiments that were performed to reply some of the comments of the referees.

1 Subsampling

Following Ref. 8cK4's suggestion and Ref. W5YG point about scalability, we report here extra results on using subsampling within LSVI.

Assuming that the log target density may be written as a sum over n terms as follows:

$$\log \pi(x) = \sum_{i=0}^{n} f_i(x)$$

one may derive an unbiased mini-batch estimate of $\log \pi(x)$ by

$$\widehat{\log \pi}(x) = \frac{n}{n_{\text{batch}}} \sum_{i \in \mathcal{S}} f_i(x)$$

where the random set $\mathcal{S} \subset \{1, \dots, n\}$ is obtained by sampling with replacement.

We implement LSVI with $\log \pi(x)$ replaced by its mini-batch (stochastic) approximation, to showcase that LSVI may be applied to big data scenario, i.e. n is large. We consider the same settings in our first experiment in the paper (logistic regression, Gaussian prior, predictors are pre-processed in the same way), and apply LSVI to the Census dataset https://archive.ics.uci.edu/dataset/20/census+income), for which $n \approx 49,000$. TODO number of regresors. We set $n_{\text{batch}} = 1000$. A new batch is drawn at each iteration.

Figure 1 reports how the KL loss (up to an unknown constant) evolves across iterations (x-axis) and over repeated runs (red lines) for two different schedules. As expected, using a mini-batch approximation makes the results more noisy, however it seems to converge quickly relative to the number of epochs (where an epoch is a block of k successive iterations, with $k = n/n_{\text{batch}}$).

2 Comparison with NGD (natural gradient descent)

As requested by Refs. W5YG and 8cK4, we report here the performance of NGD in the same settings as in our logistic regression example; see Figure 2b. The loss is evaluated via Monte-Carlo sampling, and the gradient is obtained via jax.grad.

TODO: comments

3 Variability over repeated runs

As recommended by Ref. hBY5, we provide plots for the mean losses obtained over 100 trials with a one standard deviation interval for the mean-field case with Sonar dataset in the exact same setting as Figure 2 in the manuscript. We plan on replacing Figures 1., 2. and 5. with similar plots.

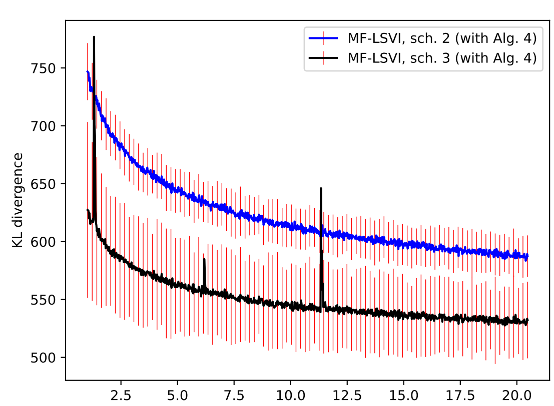
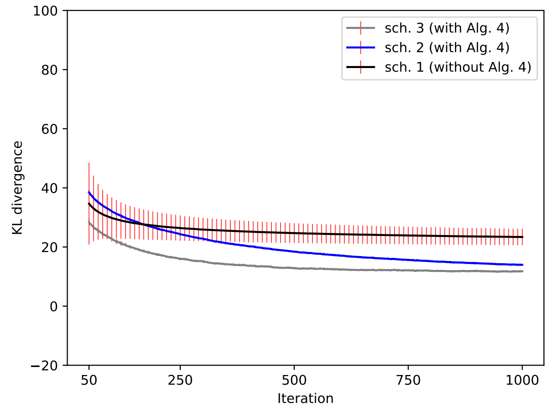
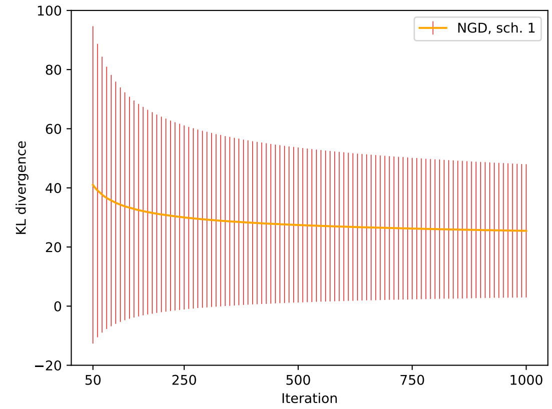


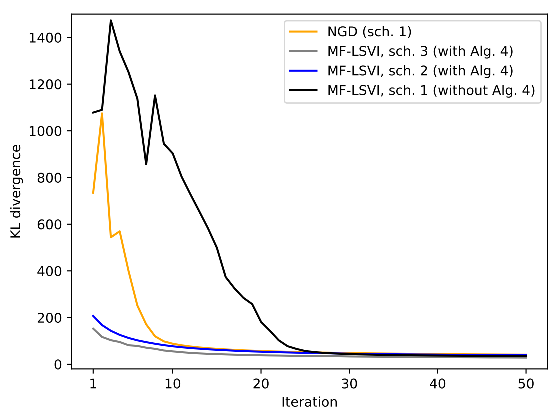
Figure 1: Logistic regression posterior with sub-sampling, Census-income dataset, diagonal covariance approximation, for MF-LSVI with schedules 2 and 3 (see supplement for details on schedules). Mean and one-standard deviation over 100 trials, $N = 10^4$ samples. Truncated from iteration 50 for better readability.



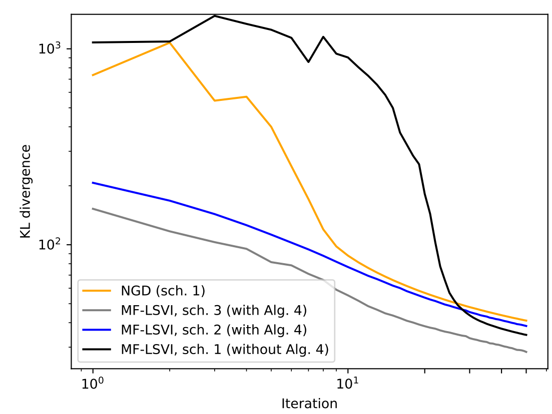
(a) Logistic regression posterior, Sonar data, diagonal covariance approximation, for MF-LSVI (Alg. 2) with different schedules. Mean and one-standard deviation over 100 trials, $N = 10^4$ samples. Truncated from iteration 50 for better readability.



(b) Logistic regression posterior, Sonar data, diagonal covariance approximation, for NGD with schedule 1. Mean and one-standard deviation over 100 trials, $N = 10^4$ samples. Truncated from iteration 50 for better readability.



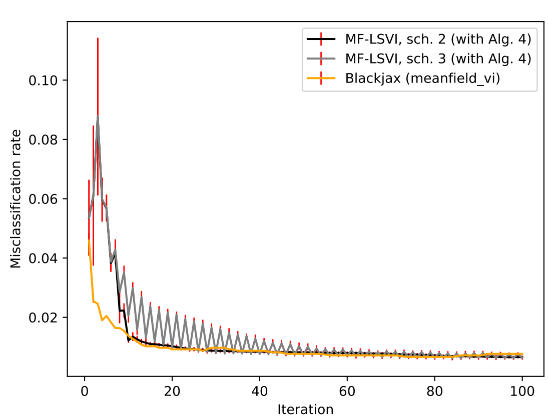
(a) Logistic regression posterior, Sonar data, diagonal covariance approximation, for MF-LSVI (Alg. 2), and NGD with different schedules. Mean over 100 trials, $N = 10^4$ samples. Truncated to iteration 50 for better readability.



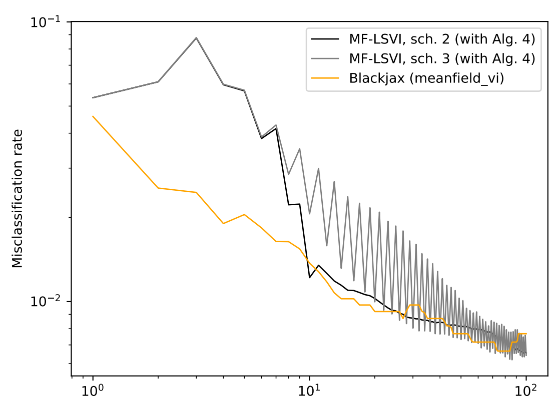
(b) Logistic regression posterior, Sonar data, diagonal covariance approximation, MF-LSVI (Alg. 2), NGD, log-log scale. Mean over 100 trials, $N = 10^4$ samples. Truncated to iteration 50 for better readability.

4 Reporting missclassification rates

Following hBY5's comment on mis-classification rate, Fig. 4b reports this rate across iterations in the MNIST logistic regression problem. The experiment is ran over 20 independent runs to obtain confidence intervals. The predictors are taken to be the means of the variational approximations.



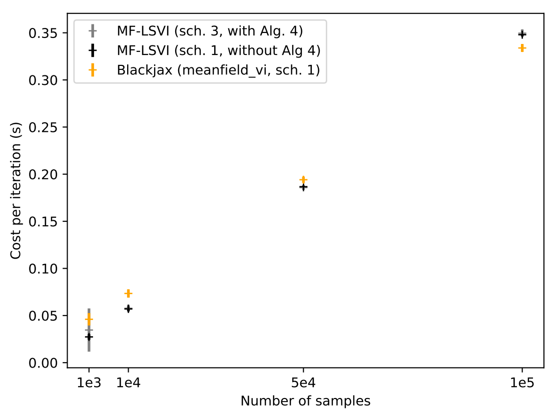
(a) Missclassification rate for the MNIST dataset. Diagonal covariance approximation, MF-LSVI (Alg. 2) vs Blackjax mean-field VI. Mean and one std interval over 20 trials, $N = 10^4$ samples. Truncated to iteration 100 for better readability.



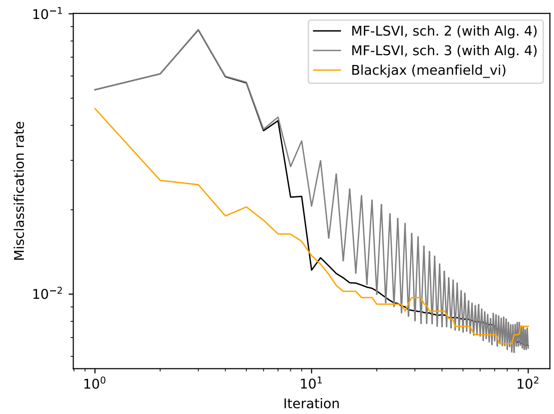
(b) Missclassification rates for the MNIST dataset. Diagonal covariance approximation, MF-LSVI (Alg. 2) vs Blackjax mean-field VI. Mean over 20 trials, log-log scale, $N = 10^4$ samples. Truncated to iteration 100 for better readability.

5 Wall-clock time

Following hBY5's comments, we ran additional experiments to monitor the actual cost per iteration (in seconds) of our proposed methods in comparison to standard implementations (i.e., the default pyMC3 implementation of ADVI, and Blackjax implementation, blackjax.meanfield_vi). We found our method to have similar costs to the existing implementations.



(a) Time per iteration on the MNIST dataset. Mean-field approximation (MF-LSVI vs blackjax.meanfield_vi). Minima, maxima, means and one-standard deviation intervals, over 5 trials, and 100 iterations.



(b) Time per iteration on the Sonar dataset. Full-covariance approximation (FC-LSVI vs pyMC3 default implementation of ADVI). Minima, maxima, means and one-standard deviation intervals, over 5 trials, and 100 iterations.

TODO: plot on the right