Distributed Bayesian Learning with Stochastic Natural-gradient Expectation Propagation

Leonard Hasenclever University of Oxford
Stefan Webb University of Oxford
Thibaut Lienart University of Oxford
Sebastian Vollmer University of Oxford

Balaji Lakshminarayanan Google DeepMind
Charles Blundell Google DeepMind
Yee Whye Teh University of Oxford

Abstract

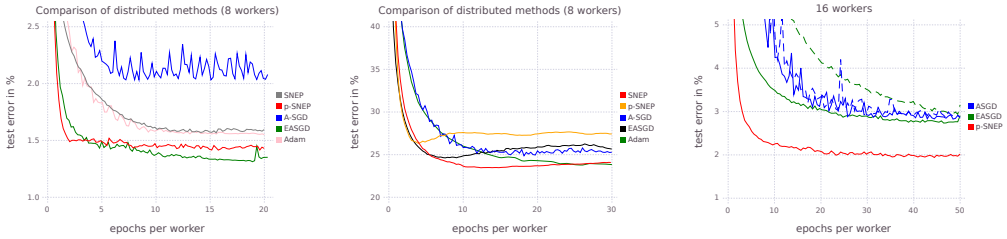
This paper makes two contributions to Bayesian machine learning algorithms. Firstly, we propose stochastic natural gradient expectation propagation (SNEP), a novel black box variational algorithm that is an alternative to expectation propagation (EP). In contrast to EP which has no guarantee of convergence, SNEP can be shown to be convergent, even when using Monte Carlo moment estimates. Secondly, we propose a novel architecture for distributed Bayesian learning which we call the posterior server, implementing a distributed asynchronous version of SNEP, which allows scalable and robust Bayesian learning in cases where a dataset is stored in a distributed manner across a cluster. An independent Monte Carlo sampler is run on each compute node which targets an approximation to the global posterior distribution given all data across the whole cluster. We demonstrate SNEP and the posterior server on distributed Bayesian learning of neural networks.

1 Introduction

Deep neural networks have recently led to advances in fields such as computer vision [KSH12] and reinforcement learning [MKS+15, SHM+16]. While deep neural networks perform well in many tasks, they tend to be overconfident in their predictions and do not provide well calibrated uncertainty estimates. Bayesian deep learning tackles this problem in a principled way. There are two different approaches to Bayesian deep learning, variational inference methods, which seek to approximate the posterior with a tractable distribution [HLA15, BCKW15, GG16] and sampling methods which obtain approximate samples from the posterior distribution [LCCC16, KRMW15]. Another challenging problem in large-scale machine learning is how to parallelize learning. This is a well-studied problem in deep learning [DCM+12, ZCL15] but there has been little work on this in the Bayesian context. In this paper we present stochastic natural gradient expectation propagation (SNEP), a novel black-box variational inference algorithm related to expectation propagation (EP, [Min01]), which lends itself well to parallelization. Due to space constraints, we will describe the high level idea and present experimental results. For more details, we refer to [HWL+16].

In SNEP the data set is partitioned into N disjoint parts belonging to N worker processes. On each worker we run an MCMC sampler targeting an approximation to the posterior distribution where the likelihood from data on other workers has been replaced with an exponential family distribution (co-incidentally the EP tilted distributions). We will also maintain a likelihood approximation on each worker which is sent infrequently across the network to update the target distribution of other workers. Instead of using parallel EP updates, an approach proposed by [XLT+14], which does not perform well for complex models, we augment the Power EP [Min04] objective with auxiliary variables (as in [HZ02]) to derive a (convergent) double-loop algorithm. The inner loop update is a stochastic natural gradient ([Ama98, RM15]) update for the parameters of local likelihood approximation using

Workshop on Bayesian Deep Learning, NIPS 2016, Barcelona, Spain.



(a) (b) (c)

Figure 1: Comparing (p-)SNEP to non-Bayesian distributed learning algorithms with: (a) a shallow dense architecture on MNIST; (b) a deep dense architecture with twenty hidden layers on MNIST; (c) LeNeT-5 on CIFAR-10. In (b), the dashed lines for A-SGD and EASGD indicate that a prelearning phase was performed, whilst for p-SNEP no prelearning was found necessary.

an MCMC sampler targeting the EP tilted distribution. By updating the likelihood approximations between workers we obtain an asynchronous distributed Bayesian algorithm in which each worker runs an MCMC sampler targeting a local approximation to the global posterior. While double loop algorithms can be slow we found in experiments that a small number of inner loop iterations per outer loop iteration was sufficient for convergence.

2 Experiments

In this section we report experimental results applying SNEP to distributed Bayesian learning of neural networks. The underlying models are neural networks with Gaussian priors on the weights and and we are trying to learn the posterior distribution over the weights. We compare our algorithm to Adam [KB15], a state-of-the-art stochastic gradient descent (SGD) algorithm with access to the whole dataset on a single computer, as well as several state-of-the-art distributed SGD algorithms: asynchronous SGD (A-SGD) [DCM$^+$12] and elastic averaging SGD (EASGD) [ZCL15]. In our experiments we used stochastic gradient Langevin dynamics [WT11] with an preconditioning scheme reminiscent of Adam as the MCMC sampler [LCCC16].

In a first set of experiments we applied our algorithm to the MNIST data set of handwritten digits using a deep neural network with two hidden layers of 500 and 300 hidden units (see learning curves in figure 1(a)). There are two versions of our algorithm, SNEP and pSNEP, corresponding to slightly different objectives. We tuned all algorithms for optimal performance. As can be seen in the figure, pSNEP is competitive with EASGD in this experiment. Both algorithms outperform A-SGD. Figure 1(b) shows the same comparison on CIFAR10 with a small CNN. Here SNEP performs better than pSNEP and converges faster than all other algorithms.

In another set of experiments, we compared p-SNEP to a deep feedforward network with twenty hidden layers of dimension fifty (see Figure 1(c)). [NVL$^+$15] recently used this architecture to demonstrate the advantages of adding noise to standard SGD. We found that while adding noise to SGD, A-SGD, and EASGD did help some runs escape suboptimal solutions, it did not allow any of these methods to obtain a solution like that found by p-SNEP with extra percent of accuracy. Thus, this suggests that the benefits to learning with SNEP cannot entirely be put down to the addition of noise. Further experiments in [HWL$^+$16] show that SNEP is robust to the length of the communication intervals.

3 Conclusion

We introduced SNEP, an asynchronous distributed algorithm for Bayesian learning in complex models and presented experiments showing its performance in Bayesian neural networks. It is competitive with other distributed algorithms but further research is needed to fully understand its properties and explore its performance on larger models. SNEP could also be extended in various ways.

References

[Ama98] S. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10:251–276, 1998.

[BCKW15] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight Uncertainty in Neural Network. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 1613–1622, 2015.

[DCM+12] Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Andrew Senior, Paul Tucker, Ke Yang, Quoc V Le, and Andrew Y. Ng. Large scale distributed deep networks. In *Advances in Neural Information Processing Systems*, pages 1223–1231, 2012.

[GG16] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 1050–1059, 2016.

[HLA15] Jose Miguel Hernandez-Lobato and Ryan Adams. Probabilistic Backpropagation for Scalable Learning of Bayesian Neural Networks. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 1861–1869, 2015.

[HWL+16] Leonard Hasenclever, Stefan Webb, Thibaut Lienart, Sebastian Vollmer, Balaji Lakshminarayanan, Charles Blundell, and Yee Whye Teh. Distributed Bayesian Learning with Stochastic Natural-gradient Expectation Propagation and the Posterior Server. *arxiv preprint arxiv:1512.09327*, 2016.

[HZ02] T. Heskes and O. Zoeter. Expectation propagation for approximate inference in dynamic Bayesian networks. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, volume 18, 2002.

[KB15] D. P. Kingma and J. L. Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*, 2015.

[KRMW15] Anoop Korattikara, Vivek Rathod, Kevin P Murphy, and Max Welling. Bayesian dark knowledge. In C Cortes, N D Lawrence, D D Lee, M Sugiyama, and R Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3438–3446. Curran Associates, Inc., 2015.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.

[LCCC16] Chunyuan Li, Changyou Chen, David Carlson, and Lawrence Carin. Preconditioned Stochastic Gradient Langevin Dynamics for Deep Neural Networks. *AAAI*, dec 2016.

[Min01] T. P. Minka. *A Family of Algorithms for Approximate Bayesian Inference*. PhD thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, 2001.

[Min04] T. Minka. Power EP. Technical report, Microsoft Research, 2004.

[MKS+15] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, feb 2015.

[NVL+15] Arvind Neelakantan, Luke Vilnis, Quoc V Le, Ilya Sutskever, Lukasz Kaiser, Karol Kurach, and James Martens. Adding gradient noise improves learning for very deep networks. *arXiv preprint arXiv:1511.06807*, 2015.

[RM15] G. Raskutti and S. Mukherjee. The information geometry of mirror descent. *IEEE Transactions on Information Theory*, 61:1451–1457, 2015.

[SHM+16] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, jan 2016.

[WT11] M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the International Conference on Machine Learning*, 2011.

[XLT+14] M. Xu, B. Lakshminarayanan, Y. W. Teh, J. Zhu, and B. Zhang. Distributed Bayesian posterior sampling via moment sharing. In *Advances in Neural Information Processing Systems*, 2014.

[ZCL15] S. Zhang, A. Choromanska, and Y. LeCun. Deep learning with elastic averaging SGD. *Advances in Neural Information Processing Systems*, 2015.