UNIVERSITY OF OXFORD

AIMS
Autonomous Intelligent Machines & Systems

EPSRC
Engineering and Physical Sciences Research Council

Distributed Bayesian learning of deep neural networks

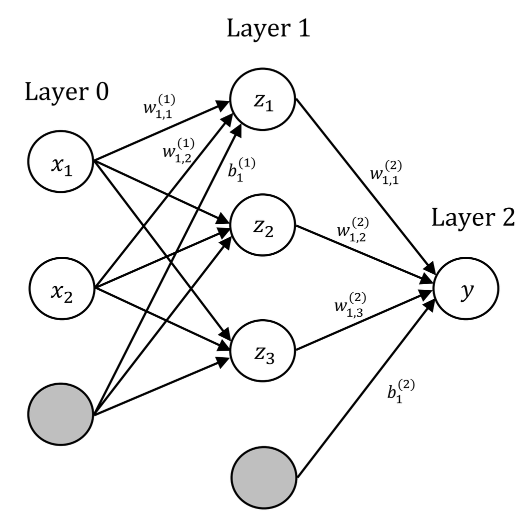
Stefan Webb working with Yee Whye Teh, Thibaut Lienart, Sebastian Vollmer, Minjie Xu, Balaji Lakshminarayanan, Charles Blundell, and Leonard Hasenclever

Introduction

- Increasing the scale of neural networks (NNs) with respect to the number of parameters and samples can drastically improve classification accuracy $\Rightarrow$ explosion of interest in deep learning
- Learning large models requires distributed computing and storage
- Distributed Bayesian methods are underdeveloped compared to those based on variants of stochastic gradient descent (SGD)
- (Wang and Dunson, 2013), (Scott et al., 2013), and (Neiswanger et al. 2013) run independent Markov chains without communication and have an expensive final combination step to merge the samples
- Our approach builds on the Bayesian posterior server of (Xu et al. 2013) which is based on EP, fixing its limitations
- EP converges poorly with moderate stochasticity in the moment estimates $\Rightarrow$ long runs of MCMC are required, need relaxation of optimization problem
- No guarantees of convergence are provided for standard EP $\Rightarrow$ need to convexify the variational optimization problem
- Key insight is that by being Bayesian about the communication between workers, the uncertainty each worker node has about the full model can be properly aggregated and taken care of

Problem

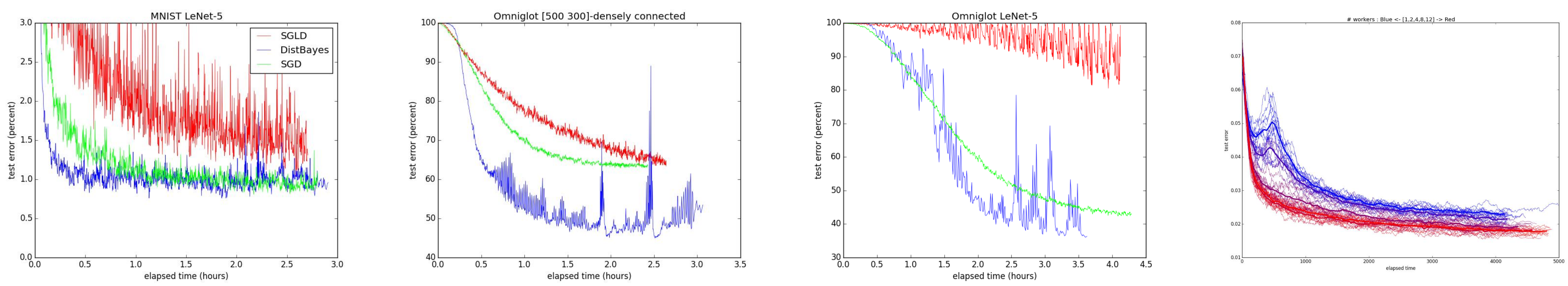
- Supervised learning for classification of small handwritten images
- The data is split between m computing nodes, each of which keeps a separate set of the model parameters
- A feedforward NN is used to map the inputs to the parameters for the assumed distribution of the targets



$$a_1 = W^{(1)}x + b^{(1)}$$
$$z = \sigma(a_1)$$
$$a_2 = W^{(2)}z + b^{(2)}$$
$$y = \sigma(a_2)$$
$$t \sim f(\cdot|y)$$

- For classification, there is an output node for each class that represents the probability of that class.
- A prior is placed over the parameters and our goal is to calculate the mean of the posterior

Results



Method

- In the most general setting, a prior from the exponential family is placed over the weights and biases of the NN
- The expectation propagation algorithm can be understood as a Lagrangian approach for solving the following relaxed variational principle,

$$\max_{\{\tau,(\eta_i,\tilde{\tau}_i)\}} \left(\langle \tau, \theta \rangle + \sum_{i=1}^{M} \langle \tilde{\tau}_i, \tilde{\theta}_i \rangle + H(\tau) + \sum_{i=1}^{M} \beta_i \left(H(\eta_i, \tilde{\tau}_i) - H(\eta_i) \right) \right)$$

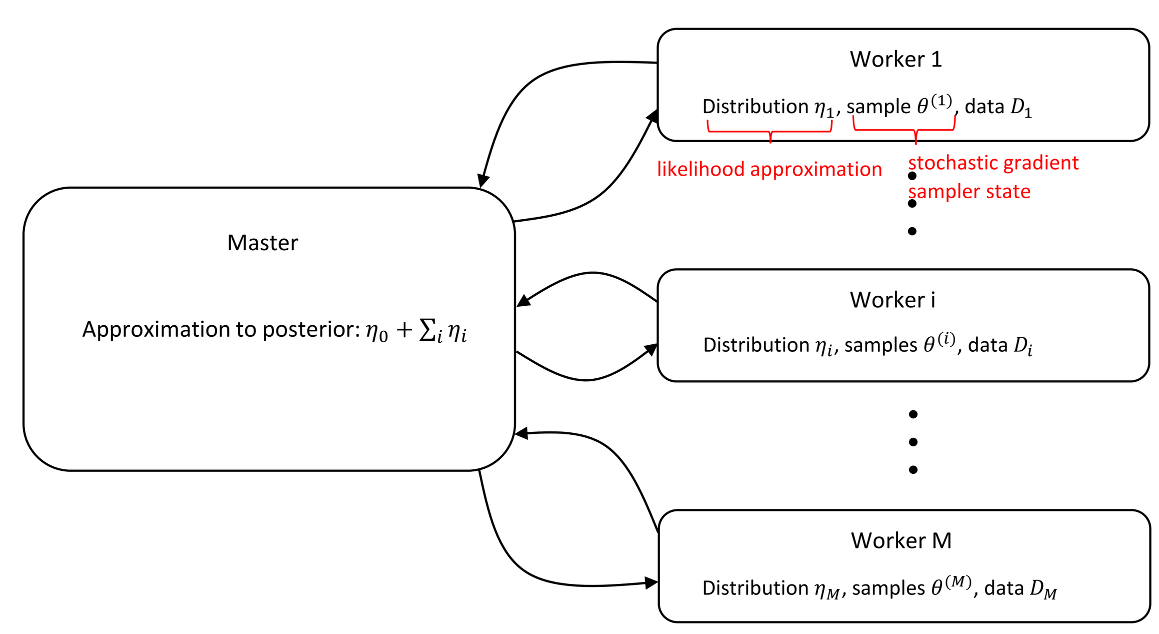
subject to $\tau = \eta_i$ and domain constraints

- We wish to convexify this problem in a way that improves convergence when there is stochasticity in the estimate of the moments
- To these ends we introduce dummy variables θ_i', subtract $\sum_{i=1}^{M} \beta_i KL(\eta_i \,||\, \theta_i')$ and take the supremum over θ_i',

$$\max_{\{\tau,(\eta_i,\tilde{\tau}_i),\theta_i'\}} \left(\langle \tau, \theta \rangle + \sum_{i=1}^{M} \langle \tilde{\tau}_i, \tilde{\theta}_i \rangle + H(\tau) + \sum_{i=1}^{M} \beta_i \left(H(\eta_i, \tilde{\tau}_i) - A(\theta_i') + \langle \eta_i, \theta_i' \rangle \right) \right)$$

subject to $\tau = \eta_i$ and domain constraints

- Changing the order of maximization, an algorithm can be derived, similarly to EP, to iteratively calculate the fixed point of the variational problem (details to be published soon)



Conclusions

- Appears competitive with advanced non-distributed SGD for models on the MNIST and Omniglot data sets
- More workers up to 8 improves the solution to which it is converging
- Combines variational and MCMC algorithms to produce *a major advance in distributed Bayesian learning*
- More work required for adaptive step sizes and increased synchronization intervals
- Method is applicable to all Bayesian models, not just feedforward networks
- Future applications include Bayesian matrix factorization and RNNs