Bayesian Deep Learning via Expectation Maximization and Turbo Deep Approximate Message Passing

Seminar Report

Submitted in partial fulfillment for the award of the degree of

MASTER OF TECHNOLOGY

in

COMPUTER SCIENCE & ENGINEERING

of

APJ Abdul Kalam Technological University

Submitted by

PAUL JOSE
Reg.No : MAC24CSCE07

Department of Computer Science and Engineering
Mar Athanasius College of Engineering (Govt. Aided & Autonomous)
Kothamangalam

April 2025

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
MAR ATHANASIUS COLLEGE OF ENGINEERING
(Govt. Aided & Autonomous)
KOTHAMANGALAM

CERTIFICATE

This is to certify that the report entitled **Bayesian Deep Learning via Expectation Maximization and Turbo Deep Approximate Message Passing** submitted by **Mr. PAUL JOSE, Reg. No: MAC24CSCE07** to the APJ Abdul Kalam Technological University in partial fulfillment of the requirements for the award of the Degree of Master of Technology in Computer Science and Engineering is a bonafide record of the seminar presented by her under our supervision and guidance. This report in any form has not been submitted to any other University or Institute for any purpose.

....	**.................................**
Prof. Pristy Paul T	**Prof. Joby George**
Seminar Coordinator	**HEAD OF THE DEPT.**

ACKNOWLEDGEMENT

First and foremost, I sincerely thank **God Almighty** for his grace for the successful and timely completion of the micro project. I express my sincere gratitude and thanks to the Principal **Dr. Bos Mathew Jos** and Head of the Department **Prof. Joby George** for providing the necessary facilities and their encouragement and support. I owe special thanks to the faculty in charge **Prof. Pristy Paul T** for their corrections, suggestions and efforts to coordinate the micro project under a tight schedule. I also express my gratitude to the staff members in the Department of Computer Science and Engineering who have taken sincere efforts in helping me to completing this microproject. Finally, I would like to acknowledge the tremendous support given to me by our dear friends without whose support this work would have been all the more difficult to accomplish.

ABSTRACT

This report presents a detailed examination of the EM-TDAMP framework, a novel Bayesian approach for simultaneous learning and structured compression of deep neural networks (DNNs). By formulating DNN training as sparse Bayesian inference with group-sparse priors, EM-TDAMP replaces standard stochastic gradient descent (SGD) with an Expectation–Maximization (EM) algorithm in which the E-step leverages Turbo Deep Approximate Message Passing (TDAMP) for efficient posterior inference, and the M-step adaptively updates hyperparameters—group activations and output noise variance—via closed-form and Gumbel-based approximations. The report further extends EM-TDAMP to federated learning, enabling communication-efficient aggregation of client-side posterior parameters through weighted geometric averaging. Comprehensive theoretical exposition covers factor graph modeling, turbo message-passing modules, PasP mini-batch fusion, and algorithmic complexity. Empirical evaluations on Boston housing regression and MNIST classification demonstrate accelerated convergence (2× faster than Adam), robust performance at high sparsity (1% accuracy drop at 80% pruning), and reduced federated rounds (¡10 vs. 50 for FedAvg). This work highlights EM-TDAMP's potential for uncertainty-aware, resource-constrained DNN deployment and outlines future extensions to convolutional architectures and asynchronous federated protocols.

Contents

List of Figures

CHAPTER 1

INTRODUCTION

Background and Motivation

Deep neural networks (DNNs) have catalyzed advances in computer vision, natural language processing, and signal processing due to their capacity to approximate complex functions. Traditional training leverages stochastic gradient descent (SGD) with backpropagation to fit large parameter sets to vast datasets. However, as DNN architectures deepen and widen, three critical issues arise:

1. Optimization Challenges: Vanishing and exploding gradients hinder stable training, especially in deep architectures.
2. Convergence Speed: SGD-based methods often require thousands of epochs, limiting real-time and resource-constrained applications.
3. Model Compression: To deploy on edge devices, pruning and quantization are applied post-training, but they struggle to achieve exact sparsity targets and impair uncertainty estimation.

Bayesian approaches offer principled uncertainty quantification, framing learning as posterior inference. However, exact Bayesian inference in DNNs is intractable, leading to approximations such as variational Bayesian inference (VBI). VBI relies on SGD to fine-tune variational parameters, inheriting many SGD drawbacks.

Limitations of Existing Methods

- SGD-Based Training: Susceptible to poor local minima and slow convergence, with heuristic learning rate schedules.
- Post-Hoc Pruning: Magnitude-based methods (e.g., L1 regularization) lack precise sparsity control and degrade calibration.

- Message Passing Methods: Early works apply approximate message passing (AMP) to DNN training but do not support structured compression and face numerical instability.
- Federated Learning: FedAvg and variants aggregate SGD updates, suffering from heterogeneous data distributions and high communication overhead.

Overview of EM-TDAMP Contributions

The EM-TDAMP framework addresses these gaps by:

- Jointly learning and compressing DNNs via a Bayesian inference problem with group-sparse priors.
- Employing an EM algorithm where the E-step uses a novel Turbo Deep Approximate Message Passing (TDAMP) to infer posterior distributions efficiently.
- Introducing PasP to fuse minibatch information, accelerating convergence without manual scheduling.
- Extending to a federated setting, reducing communication rounds through compact posterior parameter exchange and weighted geometric aggregation.

CHAPTER 2

LITERATURE REVIEW

Bayesian Deep Learning

Bayesian neural networks (BNNs) treat weights as random variables with priors, yielding posterior distributions over parameters that naturally capture uncertainty. VBI methods approximate the posterior via an explicit variational distribution and optimize via SGD, e.g., Bayes by Backprop. Monte Carlo dropout offers a practical but heuristic alternative.

Message Passing in DNNs

Belief propagation and AMP have been explored for training graphical models but extend poorly to multilayer DNNs due to loops and high connectivity. BPI, mean-field approximations, and AMP variants achieve some gains but lack compression mechanisms.

Structured Model Compression

Group-sparse regularization prunes entire neurons or filters by enforcing structured L1 penalties. Filter pruning and lottery ticket hypothesis studies reveal that pre-training mask selection can preserve accuracy but require multi-phase training.

Federated Learning Paradigms

Federated averaging (FedAvg) iterates local SGD and weighted averaging of model updates. Recent compression techniques in FL include quantized updates and sparse aggregation to reduce communication. Bayesian FL explores posterior aggregation but often neglects structured sparsity.

CHAPTER 3

THEORETICAL FRAMEWORK

Bayesian Formulation of DNN Learning

Given data $D = \{(x_i, y_i)\}_{i=1}^{N}$, we seek the posterior:

$$p(\theta, z^L \mid D) \propto p(\theta)\, p(z^L \mid x, \theta)\, p(y \mid z^L) \tag{3.1}$$

Here, $\theta = \{W^l, b^l\}_{l=1}^{L}$ and z^L is the network's top-layer pre-activation.

Group-Sparse Priors and Structured Sparsity

To prune neuron outputs, we define groups G_i as all outgoing weights of neuron i. A Bernoulli–Gaussian prior for group i:

$$p(W_{G_i}) = \rho_i\, \mathcal{N}(W_{G_i}; 0, \Sigma_i) + (1 - \rho_i)\, \delta(W_{G_i}) \tag{3.2}$$

ρ_i controls the expected active neurons; updating these via EM achieves targeted sparsity.

Likelihood Models for Regression and Classification

- Regression: Gaussian noise added to output:

$$p(y \mid z^L) = \mathcal{N}(y; z^L, v\mathbf{I}) \tag{3.3}$$

- Classification: Probit-product likelihood approximates argmax with additive noise and Q-function evaluation.

CHAPTER 4

EXPECTATION-MAXIMIZATION (EM)

EM Algorithm Fundamentals

EM alternates between E-step: computing

$$Q(\psi, v) = \mathbb{E}_{p(\theta, z^L|D)}\left[\log p(\theta, z^L, y)\right] \tag{4.1}$$

and M-step: maximizing Q w.r.t. hyperparameters ψ, v. For large D, a minibatch version with PasP fuses posterior into priors.

E-Step: Posterior Inference via TDAMP

Approximate inference on factor graph via turbo: Module A infers with independent priors (DAMP); Module B reimposes structured sparsity (SPMP). Messages circulate until convergence.

M-Step: Hyperparameter Updates

- Prior params $\{ho_i, \sum_i\}$ set to posterior moments (group-level means and variances).
- Noise variance v closed-form update for regression (mean squared deviation) and Gumbel-based update for classification (moment-matched approximation).

CHAPTER 5

TURBO DEEP APPROXIMATE MESSAGE PASSING (TDAMP)

Factor Graph Construction

Nodes represent variables u^l, z^l, θ. Edges per layer form loops, making exact sum-product infeasible.

Module A: Approximate Message Passing (DAMP)

BiG-AMP inspired linearization and Gaussian projections reduce high-dimensional integrals to sequential matrix–vector operations with Onsager corrections.

Module B: Structured Prior (SPMP)

Exact tree-based sum-product on group-sparse factors updates group inclusion probabilities ρ_i and variances.

PasP Rule for Minibatch Fusion

After each minibatch, set prior

$$p(\theta) \leftarrow p(\theta \mid D_{\text{mini}}) \tag{5.1}$$

(PasP) to incorporate observed data quickly; akin to learning rate scheduling.

Algorithmic Complexity

Per-iteration cost $O(B\sum_l N_{l-1}N_l)$, where B is minibatch size, matching SGD's matrix multiplications.

CHAPTER 6

FEDERATED EM-TDAMP EXTENSION

Federated Learning Overview

Clients hold D_k; server aggregates posteriors. Privacy is preserved by only exchanging group-level parameters (ϕ_k,σ_k).

Client-Side Inference and Compression

Each client runs τ_{max} TDAMP iterations, then sends:

- Active group probabilities $\rho_i^{(k)}$
- Posterior means/variances of active weights
- Local noise statistics σ_k

Server-Side Aggregation and Updates

Global posterior approximated via weighted geometric average of local posteriors; hyperparameters updated accordingly, minimizing global negative log-posterior.

Communication Efficiency Analysis

Exchange per round: $O(B\sum_l N_{l-1}N_l)$ scalars, compressed by sparsity and group-averaging. Rounds < 20 suffice for convergence in experiments.

CHAPTER 7

COMPARATIVE ANALYSIS

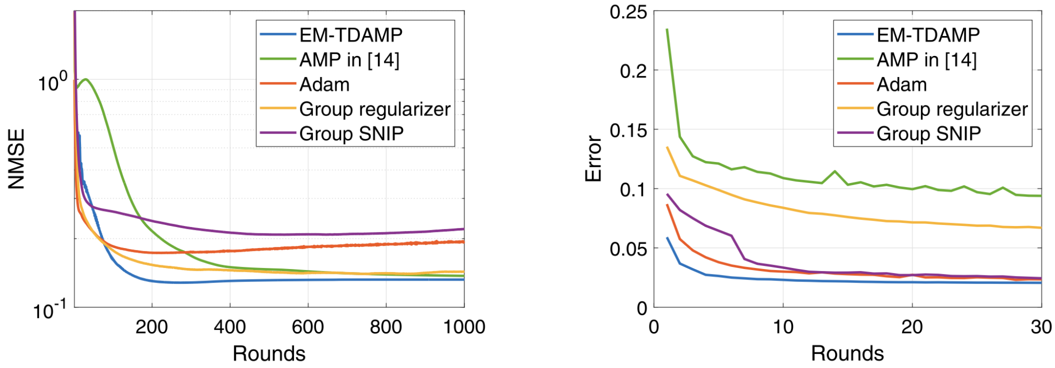


Figure 7.1: In centralized learning case, training curves of the proposed EM-TDAMP compared to baselines.

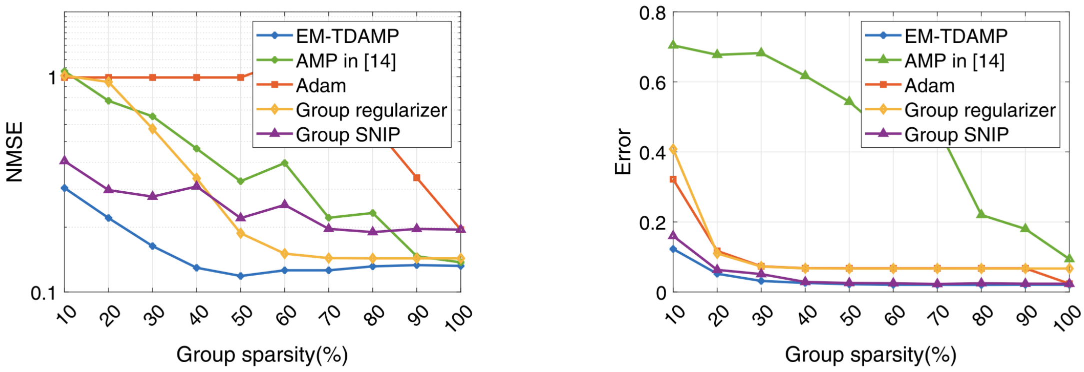
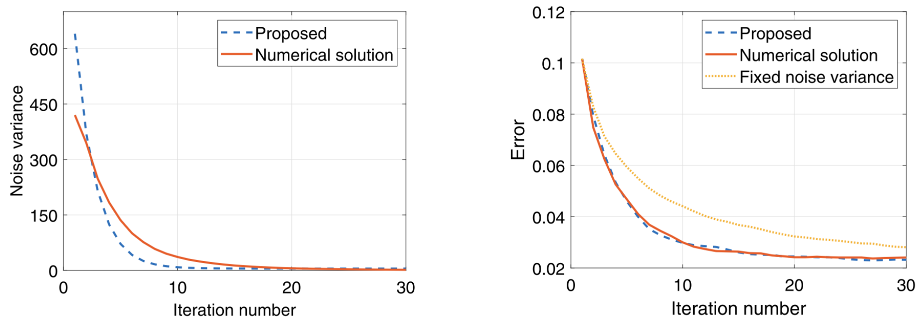


Figure 7.2: In centralized learning case, converged performance of proposed EM-TDAMP compared to baselines at different sparsity.



(a) Noise variance curve. (b) Test error curve.

Figure 7.3: Comparison of different noise variance updating methods during training.

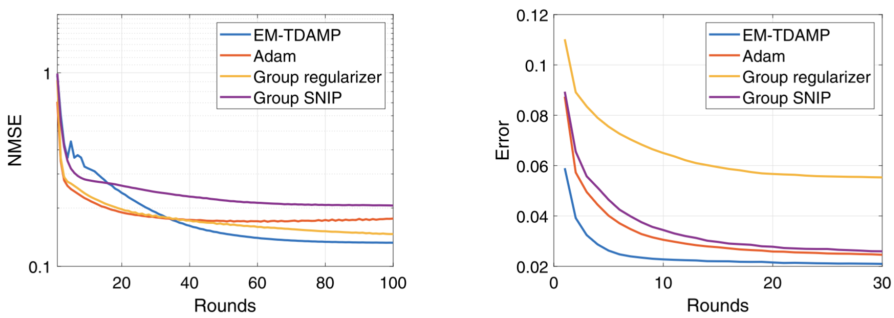
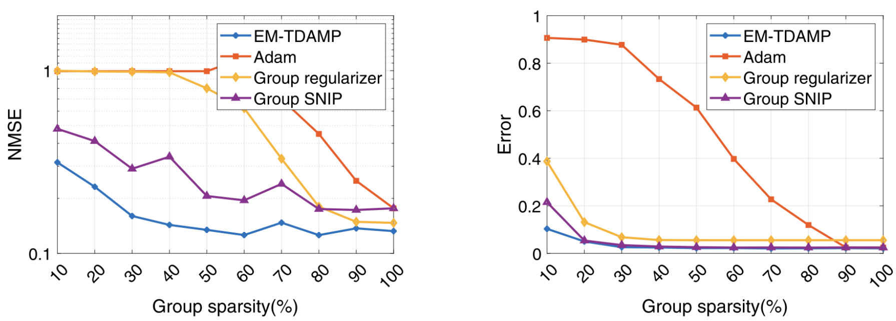


Figure 7.4: In federated learning case, training curves of the proposed EM-TDAMP compared to baselines.



(a) Test NMSE in Boston housing price prediction. (b) Test error in handwriting recognition.

Figure 7.5: In federated learning case, converged performance of proposed EM-TDAMP compared to baselines at different sparsity.

CHAPTER 8

DISCUSSION AND INSIGHTS

Advantages of EM-TDAMP

- Joint training & pruning with uncertainty quantification.
- Adaptive noise control removes manual LR tuning.
- Structured sparsity suitable for hardware acceleration.

Limitations and Potential Improvements

- Current focus on fully connected layers; extension to CNNs needed.
- Computational overhead of message updates vs. SGD in extremely large models.

Broader Impact in Edge and Federated AI

EM-TDAMP can enable robust, efficient DNN deployment on IoT devices and in privacy-sensitive federated environments.

CHAPTER 9

CONCLUSION

EM-TDAMP presents a unified Bayesian framework for DNN learning and compression, outperforming SGD-based and AMP baselines in convergence speed and sparsity-accuracy tradeoff. Its federated extension dramatically reduces communication rounds while preserving performance. Future work will adapt TDAMP to convolutional architectures and investigate asynchronous federated protocols.

REFERENCES

[1] S. Kim and E. P. Xing, "Tree-guided group lasso for multi-response regression with structured sparsity, with an application to eQTL mapping," *Ann. Appl. Statist.*, vol. 6, no. 3, pp. 1095–1117.

[2] K. Mitsuno, J. Miyao, and T. Kurita, "Hierarchical group sparse regularization for deep convolutional neural networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2020, pp. 1–8.

[3] S. Kuutti, R. Bowden, Y. Jin, P. Barber, and S. Fallah, "A survey of deep learning applications to autonomous vehicle control," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 2, pp. 712–733, Feb. 2021.

[4] D. Wang, F. Weiping, Q. Song, and J. Zhou, "Potential risk assessment for safe driving of autonomous vehicles under occluded vision," *Sci. Rep.*, vol. 12, p. 4981, Mar. 2022.

[5] A. A. Abdullah, M. M. Hassan, and Y. T. Mustafa, "A review on Bayesian deep learning in healthcare: Applications and challenges," *IEEE Access*, vol. 10, pp. 36538–36562, 2022.