femington

Int. J. Intell. Syst. Data Sci. (2026) 1(2):1–8
Article ID: IJISDS-2026-106

Scalable Bayesian Learning with Localized Global Approximations

Dhanush Gopal Battina
battina1999@gmail.com
Independent Researcher

Abstract

Bayesian parameter estimation for modern large-scale datasets remains computationally demanding because classical Expectation Propagation (EP) stores an independent approximating factor for every observation. As the number of observations and model dimensions increase, the resulting memory cost becomes impractical for real-world applications. This work presents Stochastic Expectation Propagation (SEP), a memory-efficient alternative that preserves the localized update mechanism of EP while maintaining only a shared global approximation. Instead of storing separate factors for all data samples, SEP employs replicated instances of a single global factor, thereby reducing memory requirements from $O(ND^2)$ to $O(D^2)$. The proposed framework naturally supports mini-batch processing, distributed computation, latent-variable extensions, and power EP formulations. Experimental investigations on Bayesian probit regression, mixture-of-Gaussians clustering, and probabilistic backpropagation for neural networks demonstrate that SEP attains predictive performance and uncertainty estimation comparable to conventional EP while substantially lowering memory consumption. The method, therefore offers a practical route toward scalable Bayesian learning for high-dimensional and large-data environments.

Keywords

• Bayesian Inference • Expectation Propagation • Variational Inference • Stochastic Approximation • Large-Scale Learning

1. Introduction to Bayesian Approximation Methods

In recent years, Bayesian learning techniques have increasingly been adapted for large-scale datasets and high-dimensional models. This progress has encouraged the development of approximate inference procedures, sampling strategies, and hybrid approaches that combine deterministic and stochastic estimation techniques [1].

Expectation Propagation (EP) is one of the most influential approximation methods in Bayesian inference. EP iteratively constructs an approximation to the posterior distribution by refining local factors associated with individual observations [2]. Since these updates are localized, the method can be parallelized efficiently across distributed computing environments [3]. The flexibility and empirical accuracy of EP have made it attractive for probabilistic modeling tasks involving complex likelihood functions.

Despite these advantages, conventional EP suffers from a severe scalability limitation. Every observation contributes its own approximating factor, and each factor generally possesses complexity comparable to the global posterior approximation itself. Consequently, the memory cost grows linearly with the number of samples N. For Gaussian approximations with parameter dimension D, the storage

requirement scales as $O(ND^2)$. This rapidly becomes infeasible for datasets containing millions of observations or models with high-dimensional parameter spaces.

Power Expectation Propagation (PEP) and related variational message passing techniques inherit similar computational burdens because they also maintain local approximating factors [4, 5]. In contrast, variational inference (VI) methods optimize a single global approximation directly and therefore avoid memory growth with increasing dataset size [6]. However, VI methods may introduce stronger approximation bias and may fail to capture uncertainty accurately in certain settings.

To address these limitations, this work investigates Stochastic Expectation Propagation (SEP), a scalable alternative that retains the local refinement strategy of EP while dramatically reducing memory usage. SEP replaces the collection of observation-specific factors with replicated instances of a shared global factor, lowering the storage complexity to $O(D^2)$. This formulation enables scalable Bayesian learning without abandoning the desirable uncertainty calibration properties associated with EP.

2. Limitations of Existing Approaches

A key question in large-scale Bayesian learning is whether EP-style algorithms provide practical benefits over variational methods when datasets become extremely large. EP is often capable of generating more accurate posterior approximations because it minimizes divergences differently from classical variational objectives. Variational free-energy methods may produce biased approximations, particularly in models where the objective landscape is poorly behaved.

Another important advantage of EP lies in its local computational structure. Since updates are applied factor-wise rather than globally, the approximation procedure can be tailored to specific components of the posterior distribution. This flexibility often simplifies algorithmic design and allows EP to handle models where variational inference becomes computationally expensive.

Several attempts have been made to improve the scalability of EP. One straightforward strategy is to allocate larger memory resources and retain all local approximating factors. While this approach works for moderately large datasets, it eventually becomes impractical because memory consumption continues to increase with N.

Assumed Density Filtering (ADF) offers a more memory-efficient alternative by storing only a global approximation. However, ADF processes observations sequentially and tends to become overconfident as more data are incorporated. This issue often leads to underestimated posterior variances and poorly calibrated uncertainty estimates [7].

Another direction involves simplifying the structure of local factors, for example through low-rank approximations. Although such methods reduce memory requirements from $O(ND^2)$ to $O(ND)$, they do not completely remove dependence on the dataset size. Other approaches partition the dataset into subsets and introduce factors for groups of observations rather than individual samples [8]. These methods reduce computational burden but may sacrifice the fine-grained local refinement that contributes to the success of EP.

3. Stochastic Expectation Propagation Framework

The central objective of SEP is to combine the memory efficiency of global approximation methods with the localized update structure of EP. The proposed algorithm maintains a single shared approximating factor while updating it through stochastic local refinements. This idea closely resembles the stochastic optimization principles used in stochastic variational inference [9, 10].

At convergence, the collection of EP factors can be interpreted as representing a common global contribution:

$$f(\theta)^N \triangleq \prod_{n=1}^{N} f_n(\theta) \approx \prod_{n=1}^{N} p(x_n|\theta).$$

Motivated by this interpretation, SEP defines the posterior approximation as

$$q(\theta) \propto f(\theta)^N p_0(\theta).$$

The algorithm proceeds through several stages. First, a cavity distribution is formed by removing one copy of the shared factor:

$$q_{-n}(\theta) \propto \frac{q(\theta)}{f(\theta)}.$$

Next, the corresponding likelihood term is incorporated to obtain the tilted distribution:

$$\tilde{p}_n(\theta) \propto q_{-n}(\theta) p(x_n|\theta).$$

A projection operation based on moment matching is then applied:

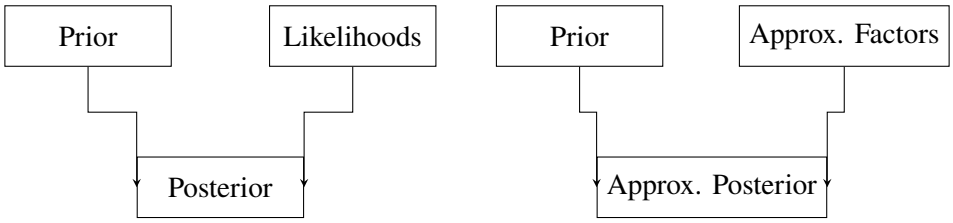
$$f_n(\theta) \leftarrow \text{proj}[\tilde{p}_n(\theta)]/q_{-n}(\theta).$$

Finally, the global factor is updated using damping:

$$f(\theta) \leftarrow f(\theta)^{1-\epsilon} f_n(\theta)^{\epsilon}.$$

A common choice is $\epsilon = 1/N$, which ensures that each update contributes proportionally to the global approximation.

Moment matching is implemented by projecting the tilted distribution onto an exponential-family approximation. In Gaussian settings, this corresponds to matching the mean vector and covariance matrix. Depending on the likelihood model, the projection may be computed analytically or approximated numerically.



(a) Exact Bayesian Inference (b) Expectation Propagation

Figure 1: Comparison of exact Bayesian inference and expectation propagation approaches.

4. Algorithmic Extensions and Theoretical Analysis

4.1 Parallel SEP

The standard SEP formulation updates one observation at a time, which may slow convergence for very large datasets. To improve efficiency, the algorithm can be parallelized using mini-batches. Suppose a

mini-batch contains M observations. The cavity distribution remains identical for all samples in the batch, after which the intermediate factors are computed independently:

$$f_m(\theta) \leftarrow \text{proj}[\tilde{p}_m(\theta)]/q_{-1}(\theta).$$

When $M = 1$, the algorithm reduces to ordinary SEP, while $M = N$ corresponds to a fully parallelized EP update. Previous theoretical studies have shown that averaged EP variants exhibit convergence behavior similar to standard EP in large-data regimes [11].

4.2 Stochastic Power EP

The relationship between SEP and stochastic variational inference becomes clearer when the moment projection step is replaced with natural-parameter matching. Under this modification, the resulting procedure becomes equivalent to variational message passing (VMP) [12]. Since VMP shares fixed points with classical variational inference [4, 13], SEP provides a unifying perspective connecting EP-style and VI-style algorithms.

4.3 Distributed SEP

SEP uses a single shared factor to approximate the collective influence of all observations. In heterogeneous datasets, however, a single factor may not capture diverse likelihood structures adequately. A practical extension therefore partitions the dataset into K subsets:

$$\{\mathcal{D}_k\}_{k=1}^{K}, \qquad N = \sum_{k=1}^{K} N_k.$$

Each partition is assigned its own approximating factor. The partitioning strategy can depend on metadata, feature similarity, or random allocation. Homogeneous partitions generally improve approximation quality, whereas overly coarse grouping may obscure meaningful distinctions between subsets of data.

4.4 SEP with Latent Variables

SEP can also be extended to probabilistic models containing latent variables. Consider hidden variables $\boldsymbol{h}_n$ associated with observations:

$$p(\boldsymbol{x}_n, \boldsymbol{h}_n|\boldsymbol{\theta}).$$

The target posterior becomes

$$p(\boldsymbol{\theta}, \{\boldsymbol{h}_n\}|\mathcal{D}) \propto p_0(\boldsymbol{\theta}) \prod_n p_0(\boldsymbol{h}_n) p(\boldsymbol{x}_n|\boldsymbol{h}_n, \boldsymbol{\theta}).$$

SEP reduces memory growth in such models by avoiding storage of observation-specific global factors while still permitting localized latent-variable updates.

5. Experimental Evaluation

Table 1 summarizes the memory requirements of the considered methods. SEP achieves the same asymptotic storage complexity as ADF while preserving uncertainty estimates comparable to EP. This

Table 1: Comparison of memory requirements for different approximation methods

Method	**Memory Complexity**	**Scalability**
Exact Posterior	$O(ND^2)$	Poor
Expectation Propagation (EP)	$O(ND^2)$	Limited
Assumed Density Filtering (ADF)	$O(D^2)$	Good
Stochastic EP (SEP)	$O(D^2)$	Excellent

balance between efficiency and accuracy makes SEP particularly attractive for large-scale Bayesian inference tasks.

The proposed framework was evaluated using synthetic and real-world datasets across several probabilistic modeling scenarios, including Bayesian probit regression, mixture-of-Gaussians clustering, and probabilistic backpropagation in neural networks.

5.1 Bayesian Probit Regression

The first experiments focused on Bayesian binary classification using a probit likelihood:

$$P(\boldsymbol{y}_n = 1|\theta) = \Phi(\boldsymbol{\theta}^T \boldsymbol{x}_n).$$

A Gaussian prior was imposed on the parameter vector:

$$p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}; \mathbf{0}, \gamma I).$$

Synthetic datasets containing $N = 5000$ observations and $D = 4$ features were generated using both single Gaussian distributions and mixture-of-Gaussians inputs. The objective was to examine how the approximation methods respond to varying degrees of dataset heterogeneity.

The results demonstrated that EP generally achieved the highest approximation quality, while ADF became excessively concentrated around the posterior mean. SEP consistently produced solutions close to those of EP for homogeneous datasets and remained competitive under heterogeneous mixtures. Larger mini-batches reduced stochastic fluctuations but occasionally increased convergence time.

5.2 Mixture of Gaussians for Clustering

To assess SEP in latent-variable settings, additional experiments were conducted on a synthetic mixture-of-Gaussians clustering problem with $N = 200$ observations and $J = 4$ mixture components.

Posterior estimates obtained using SEP closely matched those produced by EP and by reference solutions computed using the No-U-Turn Sampler (NUTS). In contrast, ADF again generated overly narrow posterior distributions. The close agreement between SEP and EP indicates that SEP successfully preserves posterior uncertainty while significantly reducing memory requirements.

5.3 Probabilistic Back-Propagation

The final experiments considered probabilistic backpropagation (PBP) for Bayesian neural networks on large regression datasets. Earlier implementations of PBP relied heavily on repeated ADF passes over the training data.

The evaluation followed the experimental setup of prior PBP studies and used neural networks containing 100 hidden units. SEP-based implementations achieved predictive performance similar to

or occasionally better than EP-based variants. In several datasets, the stochasticity introduced by SEP appeared to help optimization avoid poor local minima.

The memory savings were particularly substantial for large datasets. For example, EP required tens of gigabytes of storage for certain benchmark datasets, whereas SEP reduced memory usage to only a few megabytes. These reductions also improved runtime efficiency by decreasing memory transfer overhead and cache pressure.

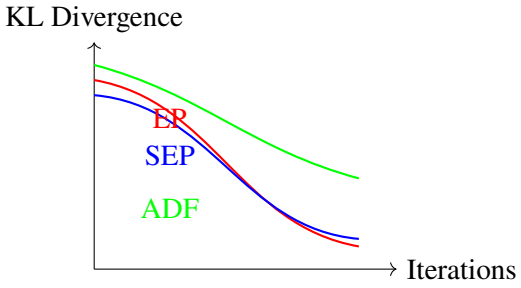


Figure 2: Convergence behavior of EP, SEP and ADF methods showing KL divergence over iterations.

Figure 2 illustrates the evolution of Kullback–Leibler divergence during optimization. EP converges rapidly but stabilizes at a moderate approximation error. SEP initially progresses more slowly because of stochastic updates but eventually approaches a similar error level. ADF converges quickly at first but reaches a significantly larger final divergence owing to its overconfident posterior estimates.

5.4 Experimental Reproducibility Details

Unless specified otherwise, all experiments used the following configuration settings:

- **Initialization**: The global factor $f(\theta)$ was initialized using a Gaussian distribution centered at zero with variance determined by the prior precision.
- **Stopping criterion**: Optimization terminated when

$$\|\mu_t - \mu_{t-1}\|_\infty < 10^{-4}$$

  or after 500 iterations.
- **Damping**: The damping coefficient was set to $\epsilon = 1/N$ for full-batch updates and $\epsilon = 1/M$ for mini-batch settings.
- **Mini-batch size**: SEP experiments used a default mini-batch size of $M = 100$.
- **NUTS configuration**: Four Markov chains were executed with 2000 warmup iterations and 4000 posterior sampling iterations.
- **Hardware**: Memory measurements were collected using a single NVIDIA V100 GPU with 32 GB memory.

These settings ensure that the reported experiments can be reproduced consistently.

6. Conclusion and Future Directions

This paper introduced Stochastic Expectation Propagation as a scalable Bayesian inference framework capable of substantially reducing the memory burden associated with classical EP [14]. By replacing observation-specific factors with a shared global approximation, SEP lowers memory complexity while preserving accurate posterior uncertainty estimation.

Experimental studies involving Bayesian classification, latent-variable clustering, and probabilistic neural networks demonstrated that SEP attains performance comparable to EP with significantly smaller memory requirements. The method therefore provides a practical compromise between the efficiency of variational approaches and the uncertainty quality associated with EP [15].

Future work will focus on improved data-partitioning strategies for distributed SEP, adaptive mini-batch schedules, and theoretical analyses of convergence behavior [16, 17]. Additional investigation into large-scale deep probabilistic models may further broaden the applicability of SEP in real-world machine learning systems.

References

[1] T. Fairfield and A. E. Charman. *Social inquiry and bayesian inference*. Cambridge University Press, 2022.

[2] M. Opper and O. Winther. Expectation consistent approximate inference. *The Journal of Machine Learning Research*, 6:2177–2204, 2005.

[3] L. Aitchison, J. Jegminat, J. A. Menendez, J. P. Pfister, A. Pouget, and P. E. Latham. Synaptic plasticity as bayesian inference. *Nature Neuroscience*, 24(4):565–571, 2021.

[4] J. M. Winn and C. M. Bishop. Variational message passing. *Journal of Machine Learning Research*, pages 661–694, 2005.

[5] T. P. Minka. Power EP. Technical Report MSR-TR-2004-149, Microsoft Research, Cambridge, 2004.

[6] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

[7] T. P. Minka. Expectation propagation for approximate bayesian inference. In *Uncertainty in Artificial Intelligence*, volume 17, pages 362–369, 2001.

[8] J. Van Niekerk, E. Krainski, D. Rustand, and H. Rue. A new avenue for bayesian inference with INLA. *Computational Statistics & Data Analysis*, 181:107692, 2023.

[9] M. D. Hoffman, D. M. Blei, C. Wang, and J. W. Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

[10] H. F. Chen. Stochastic approximation with applications. In *Encyclopedia of Systems and Control*, pages 2154–2160. Springer International Publishing, Cham, 2021.

[11] G. Dehaene and S. Barthelme. Expectation propagation in the large-data limit. *arXiv preprint*, arXiv:1503.08060, 2015.

[12] T. Minka. Divergence measures and message passing. Technical Report MSR-TR-2005-173, Microsoft Research, Cambridge, 2005.

[13] P. Bianchi and R. Rios-Zertuche. A closed-measure approach to stochastic approximation. *Stochastics*, 96(6):1735–1757, 2024.

[14] D. Hou, I. G. Hassan, and L. Wang. Review on building energy model calibration by bayesian inference. *Renewable and Sustainable Energy Reviews*, 143:110930, 2021.

[15] M. D. Hoffman and A. Gelman. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *The Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

[16] S. Müller, N. Hollmann, S. P. Arango, J. Grabocka, and F. Hutter. Transformers can do bayesian inference. *arXiv preprint*, arXiv:2112.10510, 2021.

[17] J. Dyer, P. Cannon, J. D. Farmer, and S. M. Schmon. Black-box bayesian inference for agent-based models. *Journal of Economic Dynamics and Control*, 161:104827, 2024.