DP-SEP: Differentially private stochastic expectation propagation

Margarita Vinaroz
Max Planck Institute for Intelligent Systems
University of Tübingen
mvinaroz@tuebingen.mpg.de

Mijung Park
University of British Columbia
mijungp@cs.ubc.ca

Abstract

We are interested in privatizing an approximate posterior inference algorithm called Expectation Propagation (EP). EP approximates the posterior by iteratively refining approximations to the local likelihoods, and is known to provide better posterior uncertainties than those by variational inference. However, using EP for large-scale datasets imposes a challenge in terms of memory requirements as it needs to maintain each of the local approximates in memory. To overcome this problem, stochastic expectation propagation (SEP) was proposed, which only considers a unique local factor that captures the average effect of each likelihood term to the posterior and refines it in a way analogous to EP. Therefore in this work, we focus on developing a differentially private stochastic expectation propagation (DP-SEP) algorithm, which outputs differentially private natural parameters of the exponential-family posteriors in each step of SEP.

1 Introduction

Bayesian learning provides a level of certainty about the parameters of a model, which then provides reasoning about how certain the model is about its output through the posterior predictive distribution. Variational inference (VI) [2, 5] is a popular Bayesian inference method that refines a global approximation of the posterior and scales well to large applications. However, VI often underestimates the variance of the posterior and poor performance for models with non-smooth likelihoods [3, 12].

In contrast, expectation Propagation (EP) is known to provide better posterior uncertainties than VI [8, 9]. EP constructs the posterior approximation by iterating local computations that refine approximating factors which capture each likelihood contribution to the true posterior. With large datasets, however, using EP imposes challenges as maintaining each of the local approximates in memory is costly. Stochastic Expectation Propagation (SEP) [7] overcomes this challenge by iteratively refining a single approximated factor that is repeated as many times as number of datapoints that are in the dataset. This makes the algorithm suitable for large-scale datasets as it only has to maintain in memory the global approximation in contrast to EP, that needs to keep in memory each of the approximating factors.

More importantly, in terms of applying DP, the SEP algorithm is more suitable. To apply DP to EP, a difficulty arises in sensitivity analysis: at each step of the algorithm, the approximating factor that is being refined depends on the rest, and so the sensitivity of the approximated posterior depends not only the particular factor that is being refined but also the rest of the factors that contribute to the posterior. On the other hand, in every SEP step it considers a single approximating factor at a time while all the other factors are fixed to the initial value. Hence, the sensitivity analysis of the approximate posterior becomes straightforward as we demonstrate in the following sections.

Preprint. Under review.

This work provides a framework to perform diferentilly private Bayesian learning for conjugate-exponential family models. To the best of our knowledge, there exists some previous works that desing differentially private VI methods [6, 10, 11] but this is the first differentially private algorithm that focuses on the approximating family of EP algorithms.

2 Background

We begin describing relevant background information.

Algorithm 1 EP

1: Choose a factor f_n to refine
2: Compute the cavity distribution
$q_{-n}(\boldsymbol{\theta}) \propto q(\boldsymbol{\theta})/f_n(\boldsymbol{\theta})$
3: compute tilted distribution
$\tilde{p}_n(\boldsymbol{\theta}) \propto p(\boldsymbol{x}_n|\boldsymbol{\theta})q_{-n}(\boldsymbol{\theta})$
4: moment matching
$f_n(\boldsymbol{\theta}) \leftarrow \text{proj}[\tilde{p}_n(\boldsymbol{\theta})]/q_{-n}(\boldsymbol{\theta})$
5: inclusion
$q(\boldsymbol{\theta}) \leftarrow q_{-n}(\boldsymbol{\theta})f_n(\boldsymbol{\theta})$

Algorithm 2 SEP

1: Choose a datapoint $\boldsymbol{x}_n \sim \mathcal{D}$
2: Compute the cavity distribution
$q_{-1}(\boldsymbol{\theta}) \propto q(\boldsymbol{\theta})/f(\boldsymbol{\theta})$
3: compute the tilted distribution
$\tilde{p}_n(\boldsymbol{\theta}) \propto p(\boldsymbol{x}_n|\boldsymbol{\theta})q_{-1}(\boldsymbol{\theta})$
4: moment matching
$f_n(\boldsymbol{\theta}) \leftarrow \text{proj}[\tilde{p}_n(\boldsymbol{\theta})]/q_{-1}(\boldsymbol{\theta})$
5: implicit update
$f(\boldsymbol{\theta}) \leftarrow f(\boldsymbol{\theta})^{1-\frac{1}{N}} f_n(\boldsymbol{\theta})^{\frac{1}{N}}$
6: inclusion
$q(\boldsymbol{\theta}) \leftarrow q_{-1}(\boldsymbol{\theta})f(\boldsymbol{\theta})$

Expectation propagation (EP) and Stochastic EP (SEP) Consider a dataset $\mathcal{D} = \{\boldsymbol{x}_n\}_{n=1}^N$ containing N i.i.d samples and the parametric probabilistic model given by the prior $p_0(\boldsymbol{\theta})$ of the unknown parameters $\boldsymbol{\theta}$ and the likelihood $p(\boldsymbol{x}|\boldsymbol{\theta})$. The true (intractable) posterior in Bayesian inference can be computed by:

$$p(\boldsymbol{\theta}|\mathcal{D}) \propto p_0(\boldsymbol{\theta}) \prod_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{\theta}) \approx q(\boldsymbol{\theta}) \propto p_0(\boldsymbol{\theta}) \prod_{n=1}^{N} f_n(\boldsymbol{\theta}). \tag{1}$$

EP is an iterative algorithm that produces a simpler and tractable approximating posterior distribution, $q(\boldsymbol{\theta})$, by refining the approximating factors $f_n(\boldsymbol{\theta})$. The process that EP follows to refine itelativelely these factors can be depicted in four steps. As shown in Algorithm 1, we initialize the approximating factors and form the cavity distribution $q_{-n}(\boldsymbol{\theta})$ by taking the n-th approximating factor out from the approximated posterior (i.e $q_{-n}(\boldsymbol{\theta}) \propto q(\boldsymbol{\theta})/f_n(\boldsymbol{\theta})$). On second step, the tilted distribution, $\tilde{p}_n(\boldsymbol{\theta})$, is computed by including the corresponding likelihood term to the cavity distribution: $\tilde{p}_n(\boldsymbol{\theta}) \propto q_{-n}(\boldsymbol{\theta})p(\boldsymbol{x}_n|\boldsymbol{\theta})$. The third step updates the approximating factor by minimizing the Kullback-Leibler (KL) divergence between the tilted distribution and $q_n(\boldsymbol{\theta})f_n(\boldsymbol{\theta})$ in order to capture the likelihood term contribution to the posterior. When the approximating distribution belongs to the exponential family, the KL minimization is reduced to moment matching [1] step, denoted by:$f_n(\boldsymbol{\theta}) \leftarrow \text{proj}[\tilde{p}(\boldsymbol{\theta})]/q_{-1}(\boldsymbol{\theta})$.

A major difference between EP and SEP is that SEP constructs an approximate posterior, $q(\boldsymbol{\theta})$, by iteratively refining N copies of an unique factor, $f(\boldsymbol{\theta})$, such that $\prod_{n=1}^N p(\boldsymbol{x}_n|\boldsymbol{\theta}) \approx f(\boldsymbol{\theta})^N$. The intuition behind SEP is that the approximating factor captures the average effect of a likelihood term on the posterior distribution since updates are performed analogously to EP. Similar to EP, as shown in Algorithm 2, SEP algorithm starts by initializing the approximating factor and computing the cavity distribution by removing one copy of the approximating factor from the approximate posterior: $q_{-1}(\boldsymbol{\theta}) \propto q(\boldsymbol{\theta})/f(\boldsymbol{\theta})$. Then, it calculates the tilted distribution in the same way as EP by $\tilde{p}(\boldsymbol{\theta}) \propto q_{-1}(\boldsymbol{\theta})p(\boldsymbol{x}_n|\boldsymbol{\theta})$. In the third step, SEP minimizes the KL-divergence between the tilted distribution and $q_{-1}(\boldsymbol{\theta})f_n(\boldsymbol{\theta})$ to find an intermediate factor approximate, $f_n(\boldsymbol{\theta})$.

Differential privacy Given privacy parameters $\epsilon \geq 0, \delta \geq 0$ randomized algorithm, $\mathcal{M}$, is said to be (ϵ, δ)-DP [4] if for all possible sets of mechanism's outputs S and for all neighboring datasets $\mathcal{D}, \mathcal{D}'$ differing in an only single entry ($d(\mathcal{D}, \mathcal{D}') \leq 1$), the following inequality holds: $\Pr[\mathcal{M}(\mathcal{D}) \in S] \leq e^\epsilon \cdot \Pr[\mathcal{M}(\mathcal{D}') \in S] + \delta$. The definition states that the amount of information revealed by a randomized algorithm about any individual's participation is limited. In this work we will use the *Gaussian mechanism* to privatize the natural parameters of the posterior distribution and use [13] for computing the cumulative privacy loss.

Algorithm 3 DP-SEP

Require: Dataset $\mathcal{D}$. Initial natural parameters (bounded by C)damping value γ, and the privacy parameter σ.
Ensure: (ϵ, δ)-DP natural parameters of the approximated posterior

for $t = 1, \ldots, T$ **do**
  for $n \in \{1, \ldots, N\}$, uniformly random without replacement **do**
    Choose a datapoint $\boldsymbol{x}_n \sim \mathcal{D}$
    Compute cavity distribution $q_{-1}(\boldsymbol{\theta}) \propto q(\boldsymbol{\theta})/f(\boldsymbol{\theta})$
    Compute tilted distribution $\tilde{p}_n(\boldsymbol{\theta}) \propto q_{-1}(\boldsymbol{\theta})p(\boldsymbol{x}_n|\boldsymbol{\theta})$
    Moment matching $f_n(\boldsymbol{\theta}) \leftarrow \text{Proj}[\tilde{p}_n(\boldsymbol{\theta})]/q_{-1}(\boldsymbol{\theta})$ and clip its natural parameters: $\|\boldsymbol{\theta}_{f_n}\|_2 \leq C$
    Update approximate posterior $q^{\text{new}}(\boldsymbol{\theta}) \leftarrow f_n(\boldsymbol{\theta})^{\frac{\gamma}{N}} f(\boldsymbol{\theta})^{1-\frac{\gamma}{N}} q_{-1}(\boldsymbol{\theta})$
    Add noise to natural parameters: $\tilde{\boldsymbol{\theta}}_{\text{new}} = \boldsymbol{\theta}_{\text{new}} + \mathbf{n}$ where $\mathbf{n} \sim \mathcal{N}(0, \sigma^2 \Delta^2_{\boldsymbol{\theta}_{\text{new}}} I)$
    Update $f(\boldsymbol{\theta}) \propto \left(q^{\text{new}}(\tilde{\boldsymbol{\theta}}_{\text{new}})/p_0(\boldsymbol{\theta})\right)^{\frac{1}{N}}$ and clip its natural parameters: $\|\tilde{\boldsymbol{\theta}}_{f^{\text{new}}}\|_2 \leq C$
  end for
end for

3 Our algorithm: DP-SEP

Here we introduce and describe our proposed method: DP-SEP. As shown in Algorithm 3, we output the differentially private natural parameters of the approximate posterior computed in SEP. First, our algorithm initialize the approximating factor, $f(\boldsymbol{\theta})$, such that the norm of it's natural parameters, $\boldsymbol{\theta}_f$, and prior natural parameters, $\boldsymbol{\theta}_0$, are bounded by a constant C (i.e. $\|\boldsymbol{\theta}_f\|_2 \leq C$, $\|\boldsymbol{\theta}_0\|_2 \leq C$). We need to take in account this consideration in order to further compute the sensitivity of the natural parameters for the global approximate $q(\boldsymbol{\theta})$. As the approximating distribution is in the exponential family, we can express the approximate posterior natural parameters, $\boldsymbol{\theta}$, as a linear combination of the natural parameters of the approximating factor and the prior (i.e. $\boldsymbol{\theta} = N\boldsymbol{\theta}_f + \boldsymbol{\theta}_0$).

At each run of the algorithm, we first subsample uniformly without replacent one datapoint from the dataset, $\boldsymbol{x}_n \in \mathcal{D}$, then compute the cavity distribution $q_{-1}(\boldsymbol{\theta})$, the tilted distribution $\tilde{p}_n(\boldsymbol{\theta})$ and the intermediate factor approximation $f_n(\boldsymbol{\theta})$ for $\boldsymbol{x}_n$ as in SEP algorithm. The computation of $f_n(\boldsymbol{\theta})$ is reduced to a *moment matching* step as we considered in the begining that the approximating factor is in the exponential family. Once $f_n(\boldsymbol{\theta})$ is computed, we need to ensure that it's natural parameters, $\boldsymbol{\theta}_{f_n}$, are also norm bounded by C (i.e $\|\boldsymbol{\theta}_{f_n}\|_2 \leq C$). This is due to the fact that the approximate posterior update also takes into account this intermediate factor natural parameters. After that, the algorithm updates the approximate posterior natural parameters by making a partial update of the approximating factor and the cavity distribution: $q^{\text{new}}(\boldsymbol{\theta}) \leftarrow f_n(\boldsymbol{\theta})^{\frac{\gamma}{N}} f(\boldsymbol{\theta})^{1-\frac{\gamma}{N}} q_{-1}(\boldsymbol{\theta})$. Note that all those distributions belong to the exponential family and thus $\boldsymbol{\theta}_{\text{new}} = \frac{\gamma}{N}\boldsymbol{\theta}_{f_n} + \left(N - \frac{\gamma}{N}\right)\boldsymbol{\theta}_f + \boldsymbol{\theta}_0$. In the next step, DP-SEP pertubs $\boldsymbol{\theta}_{\text{new}}$ by adding Gaussian noise. The last step ensures that its norm is also bounded by C. The DPSEP algorithm is summarized in Algorithm 3.

The following propositions state that (1) the sensitivity of the natural parameters is $\frac{2\gamma C}{N}$ and (2) the resulting algorithm is differentially private.

Proposition 1. *The natural parameters, $\boldsymbol{\theta}_{new}$, on DP-SEP algorithm have sensitivity $\Delta_{\boldsymbol{\theta}_{new}} = \frac{2\gamma C}{N}$*

Proposition 2. *The DP-SEP algorithm produces (ϵ, δ)-DP approximate posterior distributions.*

See Appendix for proof. Note that given a chosen privacy level ϵ, δ, and the number of repetitions N, T, we use the auto-dp package by [13] to compute the privacy parameter σ.

4 Experiments

We reproduced Mixture of Gaussian for clustering problem presented in the SEP paper and tested it on DP-SEP. The problem considers a synthetic dataset containing $N = 1000$ datapoints drawn from $J = 4$ Gaussians with the following assumptions: each mean is sampled from a Gaussian distribution $p(\boldsymbol{\mu}_j) = \mathcal{N}(\boldsymbol{\mu}; \mathbf{m}, I)$, each mixture component is isotropic $p(\mathbf{x}|\mathbf{h}_n) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{\mathbf{h}_n}, 0.5^2 I)$ and the cluster identity variables are sampled from a categorial uniform distribution $p(\mathbf{h}_n = j) = \frac{1}{4}$. Figure 1 visualizes the posterior means after 100 iterations for the true labels, EP, SEP and DP-SEP at different values of ϵ with clipping norm set to $C = 1$. For SEP and DP-SEP we fixed the damping value, $\gamma = 1$, i.e., $\gamma/N = 1/1000$. The figure shows that for a restrictive privacy regime $\epsilon = 1$, the clusters

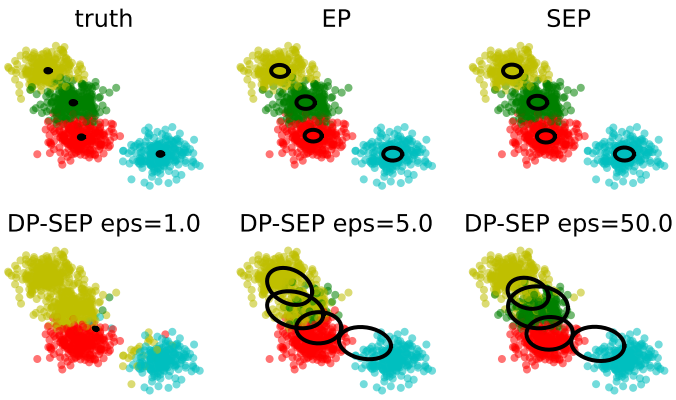

Figure 1: Mean posterior approximation for the Gaussian components (black rings indicate 98 % confidence level). The top row shows the the true labels (left), EP (middle) and SEP (right). The bottom row shows the labels for DPSEP with $\delta = 10^{-5}$ and $\epsilon = 1, 5.5, 50$.

obtained by DP-SEP are overlapping. As we increase the privacy loss, the performance of DP-SEP gets closer to the non-private ones (SEP and EP) and the ground truth. The posterior from DP-SEP exhibits a higher uncertainty than the other non-private methods due to the added noise to the mean and covariance during training.

In Table 4, we also provide a quantitative analysis of the results above in terms of F-norm of the difference between the ground truth parameters and the estimated parameters by each method. In addition, we use KL divergence between the ground truth posterior and the posterior obtained by each method. Under the mixture of Gaussians model, there is no closed form KL divergence. We instead use a proxy to the KL divergence in the following way: We first pair two Gaussians in terms of their mean locations (i.e., from a given Gaussian in ground truth, which estimated Gaussian is closest in terms of the mean estimate), and the compute the KL divergence between the paired Gaussians and averaged over those KL divergences across four paired Gaussians.

Method	F-norm	KL-divergence (proxy)
SEP ($\epsilon = \infty$)	0.0007	4.3524
DP-SEP ($\epsilon = 50$)	0.5650	516.8689
DP-SEP ($\epsilon = 5.5$)	1.6237	955.0533
DP-SEP ($\epsilon = 1$)	4.2722	4162.9041

5 Conclusions

In this work we have introduced DP-SEP, the first method for performing differentially private stochastic expectation propagation. The current experimental results follow the sensitivity we derived, i.e., the noise scales with the number of datapoints in the dataset. We will continue testing this algorithm to other models such as linear and logistic regression. A meaningful step to pursue is a theoretical understanding of the effect of noise added for privacy to the accuracy of the posterior estimates under these models.

References

[1] Shun-ichi Amari and Hiroshi Nagaoka. Methods of information geometry. volume 191, 2000.

[2] Matthew James Beal. *Variational algorithms for approximate Bayesian inference*. PhD thesis, University of London, 2003.

[3] John P. Cunningham, Philipp Hennig, and Simon Lacoste-Julien. Gaussian probabilities and expectation propagation, 2013.

[4] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9:211–407, August 2014.

[5] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, November 1999.

[6] Joonas Jälkö, Onur Dikmen, and Antti Honkela. Differentially private variational inference for non-conjugate models, 2017.

[7] Yingzhen Li, José Miguel Hernández-Lobato, and Richard E. Turner. Stochastic expectation propagation. In *Advances in Neural Information Processing Systems 29*, 2015.

[8] Thomas P. Minka. Expectation propagation for approximate bayesian inference. *In Uncertainty in Artificial Intelligence*, 17:362–369, 2001.

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

[10] Mijung Park, James Foulds, Kamalika Chaudhuri, and Max Welling. Variational bayes in private settings (VIPS), 2018.

[11] Mrinank Sharma, Michael Hutchinson, Siddharth Swaroop, Antti Honkela, and Richard E Turner. Differentially private federated variational inference. *arXiv preprint arXiv:1911.10563*, 2019.

[12] Richard Turner and Maneesh Sahani. Probabilistic amplitude and frequency demodulation. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

[13] Yu-Xiang Wang, Borja Balle, and Shiva Prasad Kasiviswanathan. Subsampled rényi differential privacy and analytical moments accountant. PMLR, 2019.

A Appendix

A.1 Proof of Prop. 1

Proof. Consider two neighboring databases, $\mathcal{D}, \mathcal{D}'$ difering only in the $\boldsymbol{x}_n, n \in \{1, \cdots, N\}$ datapoint with respective updated approximated posterior parameters $\boldsymbol{\theta}_{\text{new}}$ and $\boldsymbol{\theta}'_{\text{new}}$.

$$\Delta_2 \boldsymbol{\theta}_{\text{new}} = \max_{D,D'} \|\boldsymbol{\theta}_{\text{new}} - \boldsymbol{\theta}'_{\text{new}}\|_2$$

$$= \max_{D,D'} \left\| \left(\frac{\gamma}{N} \boldsymbol{\theta}_{f_n} + \left(N - \frac{\gamma}{N} \right) \boldsymbol{\theta}_f + \boldsymbol{\theta}_0 \right) - \left(\frac{\gamma}{N} \boldsymbol{\theta}'_{f_n} + \left(N - \frac{\gamma}{N} \right) \boldsymbol{\theta}_f + \boldsymbol{\theta}_0 \right) \right\|_2$$

$$= \frac{\gamma}{N} \max_{D,D'} \|\boldsymbol{\theta}_{f_N} - \boldsymbol{\theta}'_{f_N}\|_2 \text{ by the triangle inequality:}$$

$$\leq \frac{\gamma}{N} \max_{D,D'} \left(\|\boldsymbol{\theta}_{f_n}\|_2 + \| - \boldsymbol{\theta}'_{f_n}\|_2 \right) = \frac{\gamma}{N} \max_{D,D'} \left(\|\boldsymbol{\theta}_{f_n}\|_2 + \|\boldsymbol{\theta}'_{f_n}\|_2 \right) \leq \frac{\gamma}{N}(C + C) = \frac{2C\gamma}{N}$$

□

A.2 Proof of Prop. 2

Proof. Due to the Gaussian mechanism, the natural parameters after each perturbation are DP. By composing these with the subsampled RDP composition [13], the final natural parameters are (ϵ, δ)-DP, where the exact relationship between (ϵ, δ), T (how many repetitions SEP runs), N (how many datapoints a dataset has), and σ (the privacy parameter) follows the analysis of [13]. □