Weekly Report KARIMI 2021-08-13

My work this week has mainly been towards finishing NeurIPS rebuttals and making progress on the EM paper.

1. Reviews for NeurIPS papers on Monda and Tuesday.
2. Federated and Distributed EM paper
3. Polishing the Stanley paper for AAAI

1 NeurIPS 2021 Reviews

All rebuttals have been posted on OpenReview.

2 STANLEy paper

Several typos and modifications resulting from ICCV reviews have been made on overleaf.

3 EM paper

Main progress include a clear understanding of what the paper will include. Both motivations for distributing the E-step (MovieLens data for instance is huge and the E step can be overwhelming so there is a need to parallelize it) and for making it private and efficient (considering sensible data and low bandwidth devices, making the E step private and not heavy computationally is also important).

I have been studying two important references which are [2] and [1].

In the first paper, they develop a simple parallelization of the E-step for a particular class of network models and massive data. We could argue that massive data can be solved using simple incremental methods since then the complexity of the algorithm would become independent of the number of observations n. Yet, when several workers are at our disposal, being able to compute complete gradients/expectations is always better (in terms of variance and accuracy). They develop an asymptotic convergence result. Working in our case with exponential family will allow us to develop a better theory. Numerical applications are mainly nonlinear models applied to MovieLens dataset.

In the second reference, the problem is slightly different. Workers only can share information with their neighbours. This setting will not be the one of our paper but looking at how it works in such a setup is interesting.

In our case, we first develop the basic Decentralized EM for the exponential family. The algorithm is straight forward. Later, we present its extension to the federated settings where the statistics and quantized and the latent samples, drawn from MCMC for generality, are compressed.

Another important difference in our work is that the expectations are not tractable (making it more flexible to complex models). Hence, the sampling step being costly and in general not private, the need to tackle those issues is important.

Please see the overleaf project for details on these points.

TODO:

- Develop a theory for the distributed settings first. Then move to the Federated settings (more challenging since quantization and compression in the mix)
- Plots for at least one nonlinear model and on other type of model (bi-factor or pLSA).

References

[1] Gemma Morral, Pascal Bianchi, and Jérémie Jakubowicz. On-line gossip-based distributed expectation maximization algorithm. In *2012 IEEE Statistical Signal Processing Workshop (SSP)*, pages 305–308. IEEE, 2012.

[2] Sanvesh Srivastava, Glen DePalma, and Chuanhai Liu. An asynchronous distributed expectation maximization algorithm for massive data: the dem algorithm. *Journal of Computational and Graphical Statistics*, 28(2):233–243, 2019.