CS208 Final Project Proposal

Charlie Harrington
charlesharrington@g.harvard.edu

Jeffrey Mayolo
jeffmayolo@g.harvard.edu

April 8, 2022

1 Introduction

Throughout the semester, we have examined various forms of differential privacy. The focus of these methods has been how to release information about the data in a way that provides a reasonable guarantee of privacy for the individuals in the dataset. This released information has been static. Whether it be summary statistics, histograms, or gradients, these releases have been deterministic. There are many purposes, however, for which stochastic representations of sensitive data could be of great value. In many underdetermined problems, having functional representations of data provides the opportunity for analysis and inference that could not be accomplished otherwise. There are numerous ways to construct such functional representations: parametric and non-parametric density estimation, kernel density estimation (KDE), and Gaussian Mixture Models (GMMs), to name a few. In this project, we will focus on GMMs and how to construct them from sensitive data in a differentially private manner.

GMMs behave as a hybrid between clustering and density estimators that have a large variety of practical uses, such as the one outlined above. A GMM is comprised of $k \in \{1, \ldots, K\}$ Gaussians, K being the number of clusters in the data [1]. Whereas KDE assigns each datapoint its own Gaussian component, resulting in a non-parametric estimation, GMM assigns a cluster of points to a single component. Each component has a mean μ, and covariance Σ, and a mixing probability π that defines how big or small the Gaussian will be (all K of these mixing coefficients must add to one). Building GMMs from raw data is an iterative process that utilizes the expectation-maximization (EM) algorithm [2]. The result is a non-parametric representation of the data distribution that can then be used for a number of further applications.

2 Research Question

Our research question is: How can we build differentially private Gaussian Mixture Models in a manner that maximizes privacy while minimizing error and computational cost?

3 Relation to Previous Work

There are a number of existing works that focus on differentially private releases of marginal statistics from multi-dimensional queries. While this problem definition does not align with ours, the literature does discuss some relevant topics. Wang et al. propose using local differential privacy (LDP) encodings to answer analytical queries [3], while Yang et al. consider answering range queries through the use of a frequency oracle protocol that allows for frequency counts under LDP. In their proposed CALM framework, Zhang et al. provide a way to construct marginal tables, which capture correlations among a set of attributes [4]. They demonstrate how to use the EM algorithm within a differentially private context and explicitly outline why our research question is relevant, stating "in the LDP setting, we cannot compute a noisy marginal by adding Laplace noise to the true marginal." All of these papers demonstrate relevant approaches to marginal estimation under LDP settings.

Other related work involves ensuring that the EM algorithm is conducted in a differentially private manner [5]. Park et al. address the inherent challenge in iterative methods such as this, this being that the noise needed increases with each iteration. To overcome this, the authors propose a new EM algorithm that uses a novel moment perturbation formulation as well as the moments accountant (MA) and zero-mean concentrated differential provacy (zCDP) to bound the privacy cost of multiple iterations.

The most directly-related work is a paper titled "Differentially Private Algorithms for Learning Mixtures of Separated Gaussians" by Kamath et al. [6], in which they propose a new algorithm that matches non-private algorithms in terms of sample complexity. From a top-level perspective, their algorithm projects the data into a low-dimensional space and then recursively clustering the datapoints using clustering techniques that are amenable to privacy.

4 Our Approach

We will begin with a narrow problem scope by first focusing on building a single differentially private marginal distribution, i.e., our dataset will be of form $\mathbf{x} \in \mathbb{R}^N_{\geq 0}$, where N is the size of the dataset. Time permitting, we'll expand this to other data domains.

There are two possible approaches to our research question. One is to use a central DP model and focus on how to implement the EM algorithm in a differentially private manner. The other is to employ local differential privacy and execute standard GMM construction on the randomized response data. The first order of work in this project will be to formalize

the theoretical methodologies of these methods and to analyze their performance in terms of accuracy and computational cost. As has been the case with most concepts in this course, we expect that there will be tradeoffs between the two. We'll discuss these tradeoffs and select one methodology for implementation, recognizing that the preferred location of the trust barrier may ultimately be the deciding factor in reality.

We will implement our chosen approach in a Jupyter notebook and test it on a dummy dataset that we create to meet our data domain criteria. From this, we will analyze the empirical performance of the method. Finally, we'll demonstrate the implementation on the Public Use Microdata Sample dataset, likely using a one-hot encoding on the feature of interest.

5 Timeline

- April 8th — **Project Description Due**
- April 11th — **Literature Review Complete:** By this date we will have completely and thoroughly analyse all the papers we can find related to this topic (most literature review has already been completed).
- April 15th — **Theoretical Methodology complete:** Formalize the analysis and evaluation of a central DP model compared to a local DP model for constructing a GMM and selecting our desired methodology.
- April 17th — **Make Any Adjustments to Project Vision:** By this time we hope to have received feedback on our project description and make any necessary adjustments to our project description.
- April 23rd — **Baseline Model Running:** We want a model that successfully runs to create a DP Gaussian Mixture Model on a small data set.
- April 27th — **Coding and analysis complete:** At this time we plan to have a complete jupyter notebook that runs DP-GMMs on different datasets and is evaluated on our proposed metrics. At this time we will begin writing our paper.
- April 30th — **Draft of Paper Complete:** All sections of our initial paper will be complete. From this point we will only need to review the sections.
- May 2nd — **Initial Paper Due**
- May 9th — **Presentation**
- May 13th — **Revised Paper Due**

6 Fall Back Plan

In the case that we are unable to effectively build a DP Gaussian Mixture model that satisfies our goals, we will continue to explore deferentially private distributions in the form of Markov Chain Monte Carlo simulations. Research has previously been done to develop $(\epsilon, \delta) - DP$ MCMC algorithms that promise realistic Markov Chain convergence assumptions [7]. The authors do this using subsamples and the Barker acceptance criteria as well as Rényi divergence. For our backup plan we will use the same methodology outlined in this paper to recreate their models and explore the re-producability of their results. Additionally, we will be evaluating the accuracy of their MCMC techniques on metrics not considered in the paper such as posterior standard deviation accuracy and convergence time as a function of sample size.

References

[1] O. C. Carrasco, *Gaussian Mixture Models Explained*, en, Feb. 2020. [Online]. Available: https://towardsdatascience.com/gaussian-mixture-models-explained-6986aaf5a95.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data Via the EM Algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977. DOI: https://doi.org/10.1111/j.2517-6161.1977.tb01600.x. [Online]. Available: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1977.tb01600.x.

[3] T. Wang, B. Ding, J. Zhou, *et al.*, "Answering Multi-Dimensional Analytical Queries under Local Differential Privacy," in *Proceedings of the 2019 International Conference on Management of Data*, ser. SIGMOD '19, New York, NY, USA: Association for Computing Machinery, 2019, pp. 159–176, ISBN: 978-1-4503-5643-5. DOI: 10.1145/3299869.3319891. [Online]. Available: https://doi.org/10.1145/3299869.3319891.

[4] Z. Zhang, T. Wang, N. Li, S. He, and J. Chen, "CALM: Consistent Adaptive Local Marginal for Marginal Release under Local Differential Privacy," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '18, New York, NY, USA: Association for Computing Machinery, 2018, pp. 212–229, ISBN: 978-1-4503-5693-0. DOI: 10.1145/3243734.3243742. [Online]. Available: https://doi.org/10.1145/3243734.3243742.

[5] M. Park, J. Foulds, K. Choudhary, and M. Welling, "DP-EM: Differentially Private Expectation Maximization," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, A. Singh and J. Zhu, Eds., ser. Proceedings of Machine Learning Research, vol. 54, PMLR, Apr. 2017, pp. 896–904. [Online]. Available: https://proceedings.mlr.press/v54/park17c.html.

[6] G. Kamath, O. Sheffet, V. Singhal, and J. Ullman, "Differentially Private Algorithms for Learning Mixtures of Separated Gaussians," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d. Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32, Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/68d30a9594728bc39aa24be94b319d21-Paper.pdf.

[7] M. Heikkilä, J. Jälkö, O. Dikmen, and A. Honkela, "Differentially Private Markov Chain Monte Carlo," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d. Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32, Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/074177d3eb6371e32c16c55a3b8f706b-Paper.pdf.