Privacy in the Interpolation Regime

Context

Interpolation Regime: Modern machine learning models are often over-parametrized, i.e., the number of model parameters typically far exceed the number of data points by orders of magnitude. In the over-parametrized setting, learning algorithms typically achieve close to 100% training accuracy and are said to be in the "interpolation regime" where the algorithm almost perfectly fits all the training data [1].

Conventional statistical wisdom suggests that such overfitting may inhibit the generalization capacity of learning algorithm i.e. the ability of an algorithm to perform well on unseen data outside of the training set. However, recent progress on over-parametrized models challenge such wisdom and indicate that generalization and overfitting may not be in conflict [2]. For long-tail data distributions (which are common in image and text data), interpolation may be necessary for generalization [3]. In addition, there are practical advantages to interpolation. Recent work shows that optimization algorithms like the Stochastic Gradient Descent (SGD) converge faster in the interpolation regime [4].

Privacy: Recent research shows that deep learning models are vulnerable to a variety of privacy attacks. In a *membership inference attack*, an adversary is able to identify if a specific data point is used in training the DL model [5]. In a *reconstruction attack*, an adversary is able to approximately reconstruct the data used for training the model [6]. Such attacks pose a serious privacy risk to the use of DL models, especially when it comes to sensitive applications like healthcare.

Proposal

Despite the practical advantages and pervasiveness, learning in the interpolation regime poses strong challenges to data privacy. By nature, interpolating models essentially memorize almost all the training samples. This makes them particularly vulnerable to membership inference attacks [5], which guess whether a particular data point was used for training the model, and reconstruction attacks, which approximately reconstruct the data points that were used for training [7, 6, 8]. For example, simply by observing that an interpolating model incurs a near-zero loss on a data point, one can guess the membership of this point to the training set.

Existing defenses against privacy attacks aim to reduce overfitting through a variety of techniques such as data augmentation, regularization, randomization, and modifications to the optimization algorithm. For instance, DP-SGD, a state-of-the-art approach to guarantee differential privacy in deep learning models reduces overfitting to training data by adding noise to (clipped) gradient updates [9]. However, defenses often come with a significant drop in accuracy, indicating the necessity of sacrificing utility to guarantee privacy in the interpolation regime. The loss in accuracy is particularly steep in healthcare applications due to the high-dimensional and long-tail distributions [10]. This insight is in sharp contrast to some recent theoretical results on the generalization benefits of differential privacy [11, 12].

The ubiquity of interpolating algorithms in modern machine learning combined with their unique drawbacks in data privacy, calls for a focused study of privacy in the interpolation regime.

Goals

- **Reevaluate the privacy utility trade-off in the interpolation regime:** To achieve this objective, new bounds need to be derived that more precisely capture the trade-off between privacy and generalization in the interpolation regime. Prior works use the connection between algorithmic stability and differential privacy to show low generalization error [11]. However, empirical findings in healthcare applications indicate a steep trade-off between generalization and privacy in overparametrized settings [10]. Hence, there is a strong need to reexamine the trade-off in the interpolation regime. Generalization bounds for interpolating models cannot use arguments based on algorithmic stability, and typically rely on margin-based arguments in simpler models [13]. Therefore, new theoretical tools are needed to analyze privacy in the interpolation regime. A good starting point for this is the recent work on differentially private learning with margin guarantees [14].

- **Propose privacy defenses suited to interpolating models:** To achieve this objective, new private algorithms need to be proposed that work in conjunction with interpolating models. One approach for such algorithms is based on aggregation. Private aggregation of interpolating models trained on disjoint datasets has had recent success [15]. Another approach is to train interpolating models on noisy data. Such an approach has been used in theoretical works to explain generalization of interpolating models in simple settings [16], but there has not been an analysis of the privacy guarantees offered by such an approach.

Supervisors

- Muni Sreenivas Pydi, PSL AI Fellow, LAMSADE Laboratory

  - Contact: muni.pydi@lamsade.dauphine.fr

- Jamal Atif, Professor, Head of MILES Team, LAMSADE Laboratory
  - Contact: jamal.atif@lamsade.dauphine.fr

Profile

The ideal candidate will meet the following criteria.

- Masters degree or equivalent in Computer Science, Mathematics, Data Science or Electrical Engineering
- Mathematical maturity, and a strong theoretical background in probability theory, statistics and machine learning
- Experience in programming (Python)
- Exposure to differential privacy is a plus

References

[1] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, 2021.

[2] M. Belkin, D. Hsu, S. Ma, and S. Mandal, "Reconciling modern machine-learning practice and the classical bias–variance trade-off," *Proceedings of the National Academy of Sciences*, vol. 116, no. 32, pp. 15849–15854, 2019.

[3] V. Feldman, "Does learning require memorization? a short tale about a long tail," in *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 954–959, 2020.

[4] S. Ma, R. Bassily, and M. Belkin, "The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning," in *International Conference on Machine Learning*, pp. 3325–3334, PMLR, 2018.

[5] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramer, "Membership inference attacks from first principles," in *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1897–1914, IEEE, 2022.

[6] N. Haim, G. Vardi, G. Yehudai, O. Shamir, and M. Irani, "Reconstructing training data from trained neural networks," *Conference on Neural Information Processing Systems*, 2022.

[7] C. Dwork, A. Smith, T. Steinke, and J. Ullman, "Exposed! a survey of attacks on private data," *Annu. Rev. Stat. Appl*, vol. 4, no. 1, pp. 61–84, 2017.

[8] B. Balle, G. Cherubin, and J. Hayes, "Reconstructing training data with informed adversaries," *IEEE Symposium on Security and Privacy*, 2022.

[9] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

[10] V. M. Suriyakumar, N. Papernot, A. Goldenberg, and M. Ghassemi, "Chasing your long tails: Differentially private prediction in health care settings," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 723–734, 2021.

[11] R. Bassily, K. Nissim, A. Smith, T. Steinke, U. Stemmer, and J. Ullman, "Algorithmic stability for adaptive data analysis," in *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pp. 1046–1059, 2016.

[12] C. Jung, K. Ligett, S. Neel, A. Roth, S. Sharifi-Malvajerdi, and M. Shenfeld, "A new analysis of differential privacy's generalization guarantees (invited paper)," in *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2021, (New York, NY, USA), p. 9, Association for Computing Machinery, 2021.

[13] V. Muthukumar, A. Narang, V. Subramanian, M. Belkin, D. Hsu, and A. Sahai, "Classification vs regression in overparameterized regimes: Does the loss function matter?," *The Journal of Machine Learning Research*, vol. 22, no. 1, pp. 10104–10172, 2021.

[14] R. Bassily, M. Mohri, and A. T. Suresh, "Differentially private learning with margin guarantees," *Conference on Neural Information Processing Systems*, 2022.

[15] N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow, and K. Talwar, "Semi-supervised knowledge transfer for deep learning from private training data," *International Conference on Learning Representations*, 2017.

[16] V. Muthukumar, K. Vodrahalli, V. Subramanian, and A. Sahai, "Harmless interpolation of noisy data in regression," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 67–83, 2020.