Label Leakage from Regression Models Gradients in Federated Learning
[Talk submission]

Jean Leprovost, Pierre Jobic, and Aurélien Mayoue

Université Paris-Saclay, CEA, LIST, F-91120, Palaiseau, France

Abstract. Federated learning (FL) is one of the most popular way to collaboratively train models while preserving privacy. Participants train their model locally and share only their gradients instead of their personal data. However, recent gradient attacks have shaken this guarantee of "privacy by design" by reconstructing the participants data from the shared gradients. Serious improvements have been achieved by first inferring the labels of the data, making it easier to then reconstruct the input data. Until now these attacks have been studied only in the context of classification models, leaving the regression case unaddressed. In this paper we develop a gradient-based attack on labels in the context of a regression model being trained under a FL framework. This attack relies on solving an approximated linear system of equations of gradients and labels, calibrated using auxiliary data. Our experiments show promising results about inferring labels and will be extended until October.

Keywords: Label inference · Gradient attack · Federated learning · Regression models.

1 Motivation

Federated learning [2] is a ML approach allowing to train a model on various user data without accessing their personal data. Concretely, a central coordinator called the *server* first send an untrained model to the participants who hold the data, called the *clients* (typically hospitals or personal devices like smartphones, holding personal data that can be sensitive). Then each client train the model locally on their personal data and send back the updated model (or eventually the gradients) to the server. These updated models are then aggregated together by the server, *e.g.* by computing the mean or median of the parameters. This pattern called a round of communication is repeated successively as many times as needed to train the model. Therefore, privacy seems guaranteed "by design" since clients never share their personal data directly. However, recent work has shaken this belief showing how client data can be reconstructed from the shared update or gradient by an honest-but-curious server.

Gradient-based Attacks like Zhu et al. [6] search for the optimal pairs of input and label best matching the shared gradients. Since the communication of gradients can be encrypted, these attacks most likely occur under the "honest

but curious" server context. Therefore the attacker may reasonably also have access to the initial model and to auxiliary data similar to the clients data.

Label inference. One way to improve gradient-based attacks proposed by [6] is to first reconstruct the labels to constrain the searching space of the input data. It has been shown ([5], [4], [3], [1]) that labels can be inferred more or less analytically from the gradients of the last layer of the model.

While the current literature label reconstruction attacks focus on classification models, we here investigate the case of a regression model where labels are continuous values that can be as sensitive as the input data *e.g.* if the clients are companies training a model to predict the salary of its employees.

2 Methods and Results

We consider the case where the local training of the client consists in a single gradient descent step over its whole dataset of K samples, and we try to reconstruct the client labels from this batch averaged gradient shared to the server.

Methods. When $K = 1$, it is possible to derive an exact formula for the label $y = w^T \frac{\nabla w}{\nabla b} + b - \frac{\nabla b}{2}$ where (w, b) are the weights and bias of the last fully connected layer. In the general case when $K > 1$, the problem is more delicate since the gradients $\nabla p = (\nabla w, \nabla b)^T$ verify a linear system $\nabla p = Ax$ where $x \in \mathbb{R}^K$ is the unknown vector containing the labels information, and $A \in \mathbb{R}^{m \times K}$ is an unknown matrix (involving the outputs of the second to last layer). However we can approximate A by transforming the problem in the following way: instead of reconstructing the labels sample-wise, we assume that the data can be clustered in n "meta-classes" of samples that share similarities (e.g. the gender for face images, the diploma for employees data) and we try to reconstruct the mean label for these n classes in the client batch. Therefore we now need to solve the transformed system $\nabla p = \bar{A}\bar{x}$ where $\bar{A} \in \mathbb{R}^{m \times n}$ can now be estimated by passing auxiliary data through the model.

Two methods are investigated to solve this linear least squares problem. We first use the Moore-Penrose pseudoinverse directly (MP), but since this problem can often be ill-conditioned, we also solve it adding Tikhonov regularization (T).

Protocol. The dataset contains the salary (target label) and other predictors. We use the "diploma" feature to create 4 "meta-classes" (high school, bachelor, master, PhD). We split it in $N_{train} = 5000$ and $N_{test} = 500$ and use the test set as the auxiliary dataset. We attack a $K = 40$ client batch to reconstruct its 4 mean salaries at different epochs of training (simulating communication rounds). We repeat this experiment 100 times with different batches to get averaged metrics. The metric is the mean absolute error (MAE) between the ground-truth vs. reconstructed mean labels. Attacks are compared to the naive attack that give the mean predicted mean labels of the current model on the auxiliary dataset.

Results. Figure 1 shows that the attack with regularization (T) give the best results (lowest error). As salaries are expressed in 10k$ per year (and range from 25k$ to 250k$), a MAE of 1 correspond to an error of this magnitude on average on the 4 reconstructed mean labels.

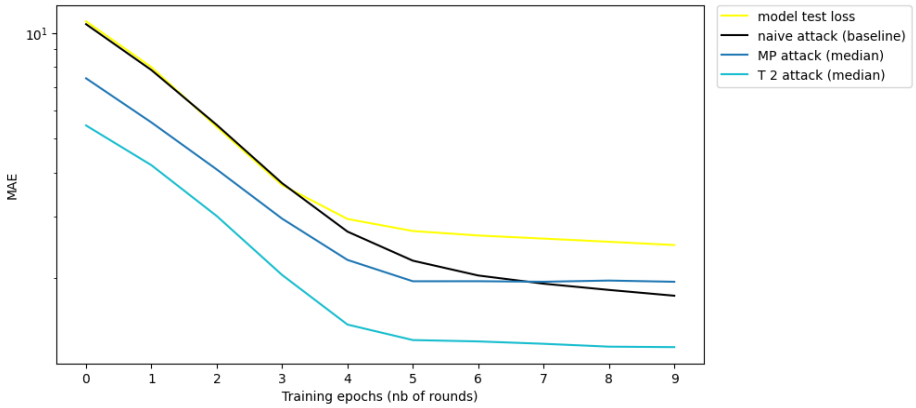

Fig. 1. Median MAE of standard (MP) and regularized (T) attacks over 100 experiments. In yellow, the test loss of the model.

3 Conclusion & Perspective

Our method to reconstruct the labels of the client data class-wise reveal how shared gradients of a regression model trained under FL could leak information. Further research may improve this attack by imposing less restrictive conditions and lowering the error.

References

1. Ma, K., Sun, Y., Cui, J., Li, D., Guan, Z., Liu, J.: Instance-wise batch label restoration via gradients in federated learning. In: The Eleventh International Conference on Learning Representations (2023)
2. McMahan, H.B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: International Conference on Artificial Intelligence and Statistics (2016)
3. Wainakh, A., Ventola, F., Müßig, T., Keim, J., Garcia Cordero, C., Zimmer, E., Grube, T., Kersting, K., Mühlhäuser, M.: User-level label leakage from gradients in federated learning. Proceedings on Privacy Enhancing Technologies pp. 227–244 (04 2022)
4. Yin, H., Mallya, A., Vahdat, A., Alvarez, J.M., Kautz, J., Molchanov, P.: See through gradients: Image batch recovery via gradinversion. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16332–16341. IEEE Computer Society, Los Alamitos, CA, USA (jun 2021)
5. Zhao, B., Mopuri, K.R., Bilen, H.: idlg: Improved deep leakage from gradients. arXiv preprint arXiv:2001.02610 (2020)
6. Zhu, L., Liu, Z., Han, S.: Deep leakage from gradients. Curran Associates Inc., Red Hook, NY, USA (2019)