

On the Behavior of PATE Under Class Imbalance and Non-IID Data

Alexander Bogdanov[1], Valery Khvatov[2], Yueh Chiang[1(✉)], and Nadezhda Shchegoleva[1]

[1] St. Petersburg University, St. Petersburg, Russia
st080802@spbu.ru
[2] DGT Technologies AG, Toronto, Canada

Abstract. Through systematic experimental analysis, this study investigates the behavior of PATE (Private Aggregation of Teacher Ensembles) in a federated learning (FL) context, focusing on highly class-imbalanced data and Dirichlet non-IID partitions, with the aim of exploring the applicability range of PATE. The results show that aggregation with differential privacy (DP) noise requires a sufficient number of teacher models to accumulate an adequate voting margin. However, under non-IID partitions, increasing the number of teachers dilutes minority-class samples, leading to teacher blinding and causing the margin to decrease rather than increase. When the margin falls below the noise magnitude, the aggregation outcome becomes noise-dominated. There are multi-level interactions among these factors; for tabular data with weak classification signals and severe class imbalance, the aggregated signal may fall below the noise threshold, causing a noise-dominated regime. Importantly, this behavior is attributable to the characteristics of the evaluation domain rather than to an inherent limitation of the PATE mechanism. To assist practitioners in identifying suitable configurations in advance, we propose an (S, σ/n) binning strategy as a pre-training screening indicator.

Keywords: Federated Learning · PATE · Rényi Differential Privacy · DP-SGD · Class Imbalance · Non-IID Partition

1 Introduction

Privacy regulations and cross-institutional data governance requirements limit the feasibility of centralized data collection and processing [1–3]. In this context, machine learning is shifting towards distributed, collaborative training paradigms where data remains local, such as FL [4]. FL enables participating parties to jointly train a model while retaining their raw data locally, making it a significant approach for balancing data governance and model performance. However, data localization does not eliminate privacy risks. Existing research indicates that model updates and prediction outputs can still leak information about training data, leading to attack risks such as membership inference and attribute inference [5]. Current FL methods typically collaborate through

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16766, pp. 327–337, 2027.
https://doi.org/10.1007/978-3-032-30539-8_25

model parameters, gradients, or updates, such as FedAvg and FedProx [4, 6]; other studies further integrate DP mechanisms, like DP-SGD, using gradient clipping and random perturbation to provide formal privacy guarantees [7]. However, under conditions of data heterogeneity and noise interference, model utility can still degrade significantly.

To reduce reliance on parameter or gradient exchange, PATE uses an ensemble of teacher models to label query samples and releases only the DP-protected aggregated labels [8]. Its utility therefore depends not only on the teacher models but also on teacher partitioning, voting structure, and the noise mechanism. Existing studies show that when teacher data shifts from approximately IID to heterogeneous partitions, teacher prediction consistency decreases, reducing the quality of aggregated labels and making ensemble decisions unstable [9].

This paper studies the applicability of PATE under federated data conditions, where class imbalance and non-IID distributions often coexist. We simulate non-IID settings using Dirichlet partitioning and control heterogeneity through the concentration parameter. Through systematic experiments, we analyze how the number of teachers, privacy budget, and data heterogeneity affect PATE performance, and identify when teacher blinding occurs. Based on this analysis, we propose an (S, σ/n) binning strategy as a practical pre-training indicator for screening PATE applicability and unstable configurations.

2 Foundational Analysis

2.1 Basic Definition of DP

Two datasets D and $D^{'}$ differing in at most one record are called neighboring datasets, denoted as $D \sim D^{'}$. A randomized mechanism M satisfies ϵ-DP if for any set of measurable events S, the following condition holds [10].

$$\Pr[M(D) \in S] \leq e^{\epsilon} \Pr\left[M\left(D'\right) \in S\right] \tag{1}$$

To calibrate the noise, the maximum impact of a single record on the query output must be quantified. For a vector-valued query function f, its L_1 sensitivity is defined as follows.

$$\Delta_1(f) = \max_{D \sim D'} \|f(D) - f\left(D'\right)\|_1 \tag{2}$$

2.2 Laplace and Gaussian Mechanisms

Under the L_1 sensitivity setting, the query result can be released using the Laplace mechanism:

$$M(D) = f(D) + Y, Y_i \sim Lap\left(\frac{\Delta_1(f)}{\epsilon}\right) \tag{3}$$

When a small failure probability δ is permissible, (ϵ, δ)-DP can be employed:

$$\Pr[M(D) \in S] \leq e^{\epsilon} \Pr\left[M\left(D'\right) \in S\right] + \delta \tag{4}$$

Correspondingly, the L_2 sensitivity of the query is defined as follows.

$$\Delta_2(f) = \max_{D \sim D'} \|f(D) - f\left(D'\right)\|_2 \tag{5}$$

If Gaussian noise is added to the output, the mechanism can be written as:

$$M(D) = f(D) + Z, Z \sim N\left(0, \sigma^2 I\right) \tag{6}$$

Compared to the Laplace mechanism, the Gaussian mechanism is more suitable for analyzing vector-valued vote statistics and multiple sequential queries.

2.3 Composition of Repeated Queries and RDP Accounting

In scenarios involving multiple sequential queries, the privacy cost accumulates with the number of queries. For (ϵ, δ)-DP, if the Gaussian mechanism is applied k times to the same dataset, their sequential composition results in $(k\epsilon, k\delta)$, where ϵ is the privacy parameter for a single query. Under the RDP framework [11], the privacy loss of the Gaussian mechanism can be expressed as:

$$\epsilon_{RDP}(\alpha) = \frac{\alpha \Delta^2}{2\sigma^2} \tag{7}$$

If the total number of queries is k, the cumulative cost is

$$\epsilon_{RDP}^{(k)}(\alpha) = k \cdot \frac{\alpha \Delta^2}{2\sigma^2} \tag{8}$$

This can then be converted back to (ϵ, δ)-DP using:

$$\epsilon = \epsilon_{RDP}^{(k)}(\alpha) + \frac{\log(1/\delta)}{\alpha - 1} \tag{9}$$

Since ϵ depends on α, it can be minimized for $\alpha > 1$ to obtain the corresponding privacy loss estimate. This method supports a larger number of queries under the same privacy budget.

2.4 PATE Voting Query and Sensitivity

In PATE, n teachers classify an unlabeled sample x. The vote count for class c can be represented as [8]:

$$V_c(x) = \sum\nolimits_{k=1}^{n} 1\left[f_k(x) = c\right] \tag{10}$$

This paper uses the vote histogram consisting of the vote counts for each class as the query output and adds Gaussian noise before releasing the final label:

$$M(x) = \operatorname*{argmax}_c (V_c(x) + Z_c), Z_c \sim N\left(0, \sigma^2\right) \tag{11}$$

If a single teacher's prediction for sample x changes from class i to class j, the count for one class decreases by one and the count for the other increases by one, resulting in a sensitivity $\Delta_2 = \sqrt{2}$.

2.5 Voting Margin Analysis

Consider a binary classification problem. Let the total number of training samples be N and the proportion of the minority class be π. Then the total number of minority class samples is $N_1 = \lfloor \pi N \rfloor$. We are concerned with whether the minority class samples become excessively diluted due to partitioning, thus failing to form a reliable aggregation capable of resisting noise.

Under class-wise Dirichlet partitioning, the allocation of minority class samples is a two-step process. First, the data proportion for each teacher is sampled from a Dirichlet distribution. Second, the N_1 minority class samples are assigned to each teacher according to this proportion. Therefore, the number of minority class samples C_k assigned to a single teacher follows a Beta-Binomial distribution:

$$P(C_k = j) = \binom{N_1}{j} \frac{B(j+\beta, N_1 - j + (n-1)\beta)}{B(\beta, (n-1)\beta)} \tag{12}$$

To simplify the analysis, taking logistic regression as an example, we use the Events Per Variable (EPV) criterion as a condition to determine whether a teacher can effectively learn the minority class. Each teacher needs at least $10d$ minority class samples for effective learning, where d is the number of predictor variables. In this paper, we set the threshold $k_{EPV} = 30$, which corresponds to the conservative engineering setting $d = 3$ effective predictors. For datasets with higher feature dimensionality, the threshold scales proportionally: at $d = 10$, $k_{EPV} \approx 100$; at $d = 20$, $k_{EPV} \approx 200$. In all experiments reported here, d is set to match the feature count of the controlled dataset generated by sklearn. make classification, so the threshold is re-evaluated per dataset rather than applied globally. Define $L_k = 1_{C_k \geq 30}$ as an indicator of whether teacher k is qualified, and $L = \sum_{k=1}^{n} L_k$ as the total number of qualified teachers.

Under class-wise Dirichlet partitioning, the number of minority class samples C_k assigned to a single teacher follows a Beta-Binomial distribution, so the probability that any given teacher is qualified is (Fig. 1):

$$p_{30}(n) = P(C_k \geq 30) = 1 - \sum_{j=0}^{29} \binom{N_1}{j} \frac{B(j+\beta, N_1 - j + (n-1)\beta)}{B(\beta, (n-1)\beta)} \tag{13}$$

The expected number of qualified teachers is $E[L] = np_{30}(n)$.

Adopting a most optimistic voting assumption: qualified teachers correctly vote for the minority class, while unqualified teachers vote for the majority class. Then, given L, the vote count for the minority class is $V_1 = L$, and the vote count for the majority class is $V_0 = n - L$. The voting margin is defined as $m = V_1 - V_0 = 2L - n$. Since L is a random variable, we define its expected voting margin as $E[m] = 2E[L] - n$. Let $p_{30}(n) = P(C_k \geq 30)$, then $E[L] = np_{30}(n)$, thus $E[m] = n(2p_{30}(n) - 1)$.

During PATE aggregation, independent Gaussian noise $Z_1, Z_0 \sim N(0, \sigma^2)$ is added to the vote counts for both classes. The condition for correct decision-making is $V_1 + Z_1 > V_0 + Z_0$, i.e., $Z_0 - Z_1 < m$. Since $Z_0 - Z_1 \sim N(0, 2\sigma^2)$, the accuracy given m is

$$P(correct \vee m) = \Phi\left(\frac{m}{\sqrt{2}\sigma}\right) \tag{14}$$

Requiring accuracy at least $1-\delta$ necessitates $m \geq \sqrt{2}\sigma\Phi^{-1}(1-\delta)$. Define $\tau_{\sigma,\delta} = \sqrt{2}\sigma\Phi^{-1}(1-\delta)$. Even without considering data partitioning, the maximum margin $m \leq n$ requires $n \geq \tau_{\sigma,\delta}$. Considering data partitioning, we need

$$n(2p_{30}(n)-1) \geq \tau_{\sigma,\delta} \tag{15}$$

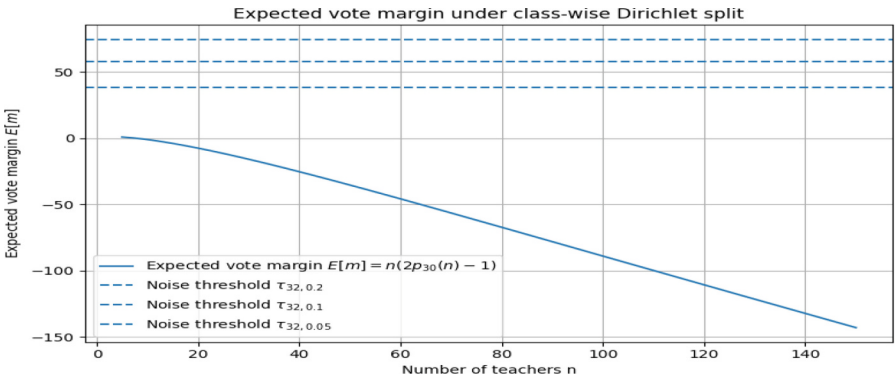

Fig. 1. Voting Margin vs. Number of Teachers

This figure shows the change in the expected voting margin $E[m] = n(2p_{30}(n) - 1)$ as the number of teachers n increases under class-wise Dirichlet partitioning ($\beta = 0.3$). It can be observed that as n increases, $E[m]$ monotonically decreases and quickly becomes negative, indicating that most teachers tend to vote for the majority class due to insufficient minority class samples ($C_k < k_{epv}$), causing the overall vote to lean towards the incorrect class.

Additionally, the figure uses dashed lines to mark the noise thresholds $\tau_{32,\delta}$ of the Gaussian mechanism for different δ values. Across the entire range of n examined, $E[m]$ is far below $\tau_{32,\delta}$, and is even negative in most regions, demonstrating that under these conditions of data imbalance ($\pi = 0.067$) and partitioning, the aggregated signal cannot form an effective margin sufficient to resist noise.

This result indicates that when the number of teachers increases, the dilution effect on the minority class caused by Dirichlet partitioning dominates the system's behavior, driving the PATE aggregation into a noise-dominated regime in practice and making reliable decision-making impossible. It is worth noting that, for tabular data, when the number of predictor variables is larger or the number of classes in the classification problem is higher, the model requires more minority-class samples to learn a stable decision boundary; however, the required sample size is difficult to estimate in an engineering manner. Therefore, this paper subsequently proposes the (S, σ/n) binning strategy as a practical screening indicator for evaluating the applicability of PATE.

3 Experimental Evaluation

3.1 Experimental Setup

The first part of the experiment uses one public financial benchmark dataset: GiveMeSomeCredit (sample size N = 10,000, feature dimension d = 10, minority-class proportion π =0.067) [12]. The dataset achieved a centralized no-privacy baseline AUC of 0.784 using an 80/20 stratified train-test split. This part illustrates that, when the sample size is limited, classes are imbalanced, and the privacy budget is low, uncalibrated parameter settings may drive PATE into a performance degradation regime. The second part further includes German Credit (N = 1,000, d = 20, π =0.30) [13], Bank Marketing (N = 41,188, d = 20, π =approximately 0.113) [14], and the CDC Diabetes Dataset (N = 253,680, d = 21, π =0.139) [15], together with the previous dataset as auxiliary experiments for illustrating mitigation strategies under engineering settings. Their centralized no-privacy baseline AUCs under the same 80/20 stratified train-test split were 0.779, 0.951, and 0.823, respectively.

To simulate a non-IID environment, this paper uses the Dirichlet distribution for data partitioning, with the concentration parameter fixed at $\alpha = 0.3$. All methods are compared under the same data partitioning, with privacy settings following (ε, δ)-DP, where $\delta = 10^{-5}$.

3.2 Baseline Comparison with DP-SGD

The first part of the experiment uses the GiveMeSomeCredit dataset and logistic regression teacher models, with three teacher-count settings (10, 30, and 50), query counts of 20, 30, and 30, respectively, and privacy budgets $\varepsilon \in \{1, 2, 5, 10\}$.

With fixed data volume, the left panel of Fig. 2 compares the logistic regression PATE results with different numbers of teachers on GiveMeSomeCredit. DP-SGD maintains an AUC between approximately 0.64 and 0.69 across privacy budgets. In contrast, the AUC for the three PATE-LR configurations is generally lower, ranging approximately from 0.31 to 0.55. With 10 teachers, the AUC is approximately 0.55 at $\varepsilon = 2$; with 30 teachers, it ranges approximately from 0.40 to 0.44; with 50 teachers, it ranges approximately from 0.35 to 0.45. Overall, under this setting, increasing the number of teachers does not show a consistent improvement in performance.

On the other hand, the right panel of Fig. 2 presents the proportion of blind teacher models. Under Dirichlet partitioning, as the number of teachers increases, the blind-teacher proportion rises from 0.07 with 10 teachers to 0.28 and 0.27 with 30 and 50 teachers, respectively, consistent with the expected dilution of minority-class samples.

3.3 Mitigation Strategy under Engineering Conditions

The preceding results indicate that, under weak class signals and non-IID partitioning, PATE may enter a performance degradation regime due to teacher blinding and insufficient voting margins. We therefore evaluate whether the average cosine similarity S between teacher data distributions and σ/n can serve as practical binning indicators for screening PATE applicability. Here, S describes the degree of similarity among teacher

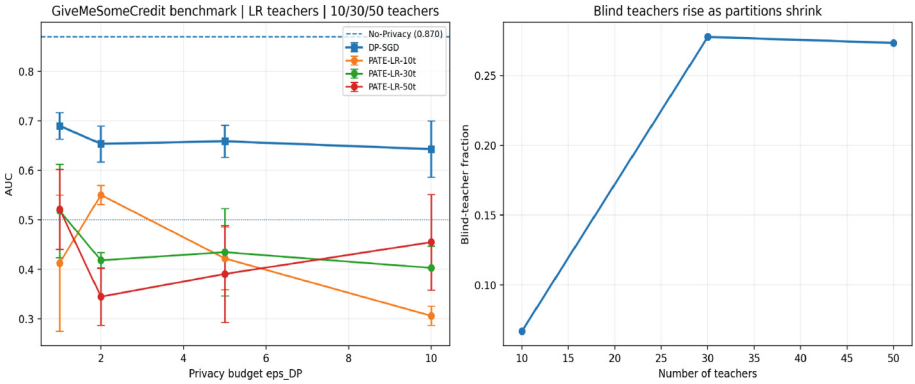

Fig. 2. Teacher Count vs. Performance and Blind Ratio

data distributions, while σ/n reflects the ratio of noise intensity to the number of teachers. For each real dataset, we generate a controlled dataset using sklearn.make_classification, matched in minority-class proportion π, sample size N, and feature dimension d, with class_sep adjusted according to the original-data baseline AUC. German Credit uses a query pool of 200 samples, while the other datasets use 2,000 samples. Query samples are clustered before selection to avoid all-negative query batches. Experiments are conducted under a fixed privacy budget with $\alpha \in \{0.1, 0.3, \infty\}$, $n \in \{20, 40, 60, 80, 100, 125, 150, 175, 200\}$, $Q \in \{20, 40, 60, 80, 100\}$, and 16 random seeds. A configuration is treated as a failure if any student model has an AUC below 0.7, and failure ratios are computed over (S, σ/n) bins to construct the heatmaps.

As shown in the figures, the heatmaps for all datasets exhibit continuous regional structures rather than completely random distributions, indicating that the (S, σ/n) binning strategy has a certain ability to distinguish regions where PATE performance degrades. The learnable regions usually appear as contiguous areas rather than scattered points, suggesting that stable PATE performance is concentrated mainly within specific parameter ranges. Although the S-bin distributions of the real and controlled data differ and the grid cells do not align perfectly, the heatmap structures remain clearly similar.

The analysis of the GiveMeSomeCredit dataset in Fig. 3 echoes the preceding results: for a low-data-volume dataset, there is almost no learnable region. German Credit in Fig. 4 is an even smaller dataset, and the conclusion is the same. In contrast, Bank Marketing in Fig. 5 and CDC Diabetes in Figure 6 retain a wider range of low-failure-rate regions as the data volume increases and the degree of imbalance decreases. Although the grid cells do not align perfectly, this binning approach still presents regional trends close to those of the real data, indicating that even under simplified distributional conditions, controlled data can provide an approximate evaluation of the learnable and degraded regions of PATE.

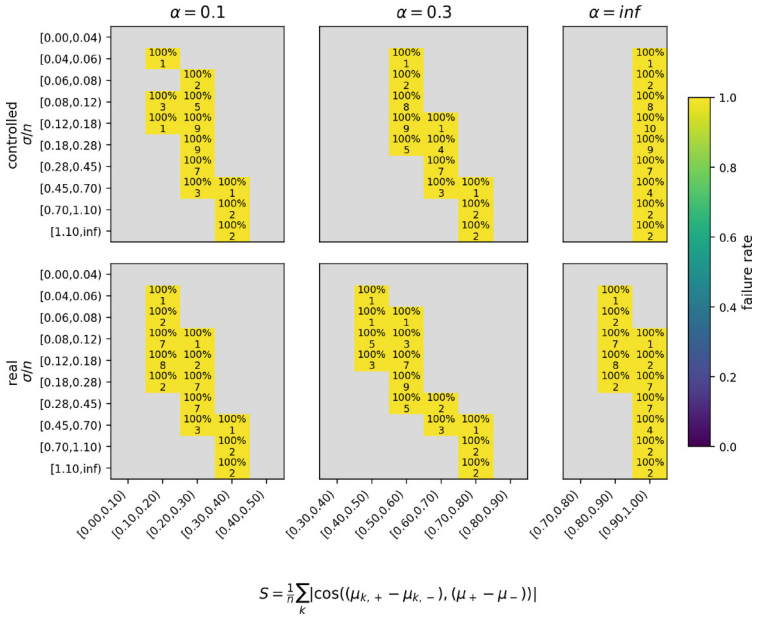

Fig. 3. Heatmap for GiveMeSomeCredit

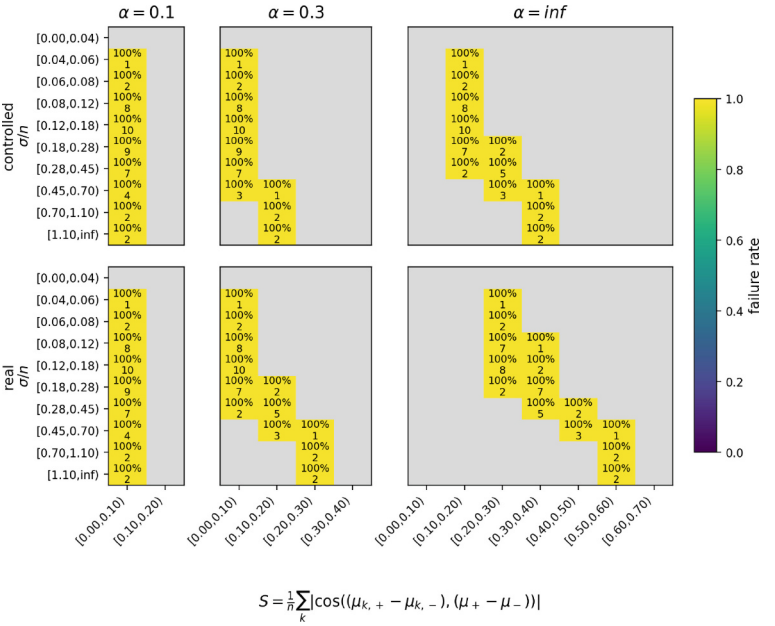

Fig. 4. Heatmap for German Credit

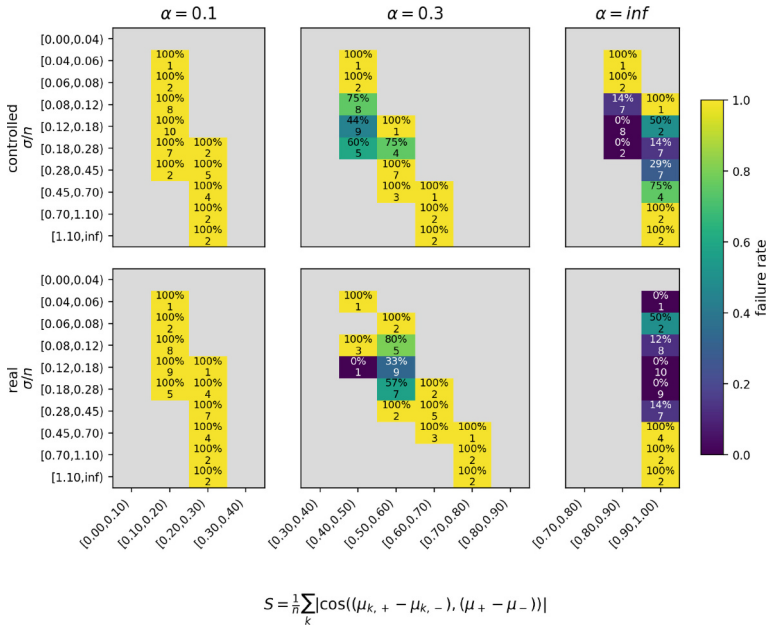

Fig. 5. Heatmap for Bank Marketing

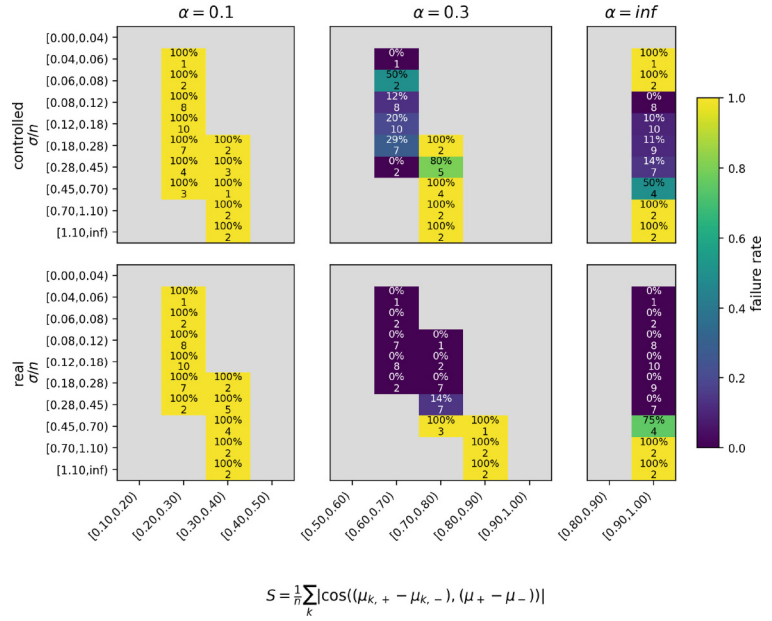

Fig. 6. Heatmap for CDC Diabetes

4 Conclusion

In summary, this study demonstrates that the performance of PATE in FL scenarios is closely related to data distribution, class structure, and signal redundancy. When minority-class samples are diluted due to Dirichlet partitioning, the teacher voting margin decreases as the number of teachers increases, causing the aggregation process to enter a noise-dominated region. On tabular data with weak class signals, this phenomenon may lead to noticeable performance degradation.

To assist practitioners in determining whether a specific dataset is suitable for PATE, this paper proposes an (S, σ/n)-based binning strategy as a pre-training applicability screening indicator. We find that the failure-rate heatmaps generated by this binning approach exhibit similar regional structures, indicating that the strategy can effectively distinguish the learnable regions of PATE from regions of performance degradation.

However, the results of this paper do not imply that PATE is unsuitable for high-dimensional data. On the contrary, the original PATE work was mainly built on datasets such as MNIST, which contain high-dimensional redundant signals, and subsequent studies such as SeqPATE [16] are likewise based on high-dimensional semantic structures in language models. In such data, even in the presence of DP noise and distributed partitioning, teacher models may still maintain relatively high consistency and stable voting margins.

Therefore, this paper argues that the performance degradation observed in this study mainly results from voting-margin decay jointly caused by weak classification signals, class imbalance, and non-IID partitioning, rather than from a failure of the PATE method itself. Based on this observation, future work will further explore federated text-learning scenarios with stronger semantic redundancy and classification signals, in order to analyze the applicability and stability of PATE under different data conditions.

References

1. European Parliament and Council of the European Union: Regulation (EU) 2016/679 (General Data Protection Regulation). Off. J. Eur. Union (2016). https://eur-lex.europa.eu/eli/reg/2016/679/oj
2. Standing Committee of the National People's Congress: Personal Information Protection Law of the People's Republic of China (2021). https://www.cac.gov.cn/2021-08/20/c_1631050028355286.htm
3. European Parliament and Council of the European Union: Regulation (EU) 2024/1689 (Artificial Intelligence Act). Off. J. Eur. Union (2024). https://eur-lex.europa.eu/eli/reg/2024/1689/oj
4. McMahan, H.B., Moore, E., Ramage, D., Hampson, S., Aguera y Arcas, B.: Communication-efficient learning of deep networks from decentralized data. In: Proceedings of 20th International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 1273–1282 (2017)
5. Shirvani, G., Ghasemshirazi, S., Beigzadeh, B.: Federated learning: attacks, defenses, opportunities, and challenges. In: Proceedings of the 11th International Symposium on Telecommunications (IST), pp. 470–479 (2024)
6. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. In: Proceedings of Machine Learning and Systems (MLSys), pp. 429–450 (2020)

7. Abadi, M., et al.: Deep learning with differential privacy. In: Proceedings of the ACM Conference on Computer and Communications Security (CCS), pp. 308–318 (2016)
8. Papernot, N., Abadi, M., Erlingsson, Ú., Goodfellow, I.J., Talwar, K.: Semi-supervised knowledge transfer for deep learning from private training data. In: International Conference on Learning Representations (ICLR) (2017)
9. Dodwadmath, A., Stich, S.U.: Preserving privacy with PATE for heterogeneous data. In: NeurIPS 2022 Workshop on Distribution Shifts (DistShift) (2022) https://openreview.net/forum?id=NvlUqen8Cya
10. Dwork, C., Roth, A.: The algorithmic foundations of differential privacy. Found. Trends Theor. Comput. Sci. **9**(3–4), 211–407 (2014)
11. Mironov, I.: Rényi differential privacy. In: Proceedings of the IEEE Computer Security Foundations Symposium (CSF), pp. 263–275 (2017)
12. Credit Fusion, Cukierski, W.: Give me some credit [Dataset]. Kaggle Competition (2011). https://kaggle.com/competitions/GiveMeSomeCredit
13. Hofmann, H.: Statlog (German Credit Data) [Dataset]. UCI Machine Learning Repository (1994). https://doi.org/10.24432/C5NC77
14. Moro, S., Rita, P., Cortez, P.: Bank Marketing [Dataset]. UCI Machine Learning Repository (2014). https://doi.org/10.24432/C5K306
15. Centers for Disease Control and Prevention (CDC): Behavioral Risk Factor Surveillance System 2015 [Dataset] (2015). https://www.cdc.gov/brfss/annual_data/annual_2015.html
16. Tian, Z., Zhao, Y., Huang, Z., Wang, Y.X., Zhang, N.L., He, H.: SeqPATE: differentially private text generation via knowledge distillation. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 35, pp. 11117–11130 (2022)