A Closer Look at the Calibration of Differentially Private Learners

Hanlin Zhang[1✉] Xuechen Li[2] Prithviraj Sen[3]
Salim Roukos[3] Tatsunori Hashimoto[2✉]

[1]Carnegie Mellon University [2]Stanford University [3]IBM Research ✉Correspondence

TL;DR

◇ Differentially private stochastic gradient descent (DP-SGD) gives rise to miscalibration due to per-example gradient clipping operation.
◇ Recalibration methods can be easily adapted to improve the privacy-calibration tradeoff with negligible utility cost.

Background

Differential Privacy (DP)

Definition

(Approximate-DP). A randomized algorithm $\mathcal{M}: \mathcal{X} \to \mathcal{Y}$ is (ϵ, δ)-DP if for all neighboring datasets $X, X' \in \mathcal{X}$ that differ on a single element and all measurable $Y \subset \mathcal{Y}, \mathbb{P}(\mathcal{M}(X) \in Y) \leq \exp(\epsilon)\mathbb{P}(\mathcal{M}(X') \in Y) + \delta$.

DP-SGD

Per-example gradient clipping + Gradient noise injection.

Calibration

Intuition

A calibrated model should give predictions that can truthfully reflect the predictive uncertainty, e.g., among the samples to which a calibrated classifier gives a confidence 0.1 for class k, 10% of the samples actually belong to class k.

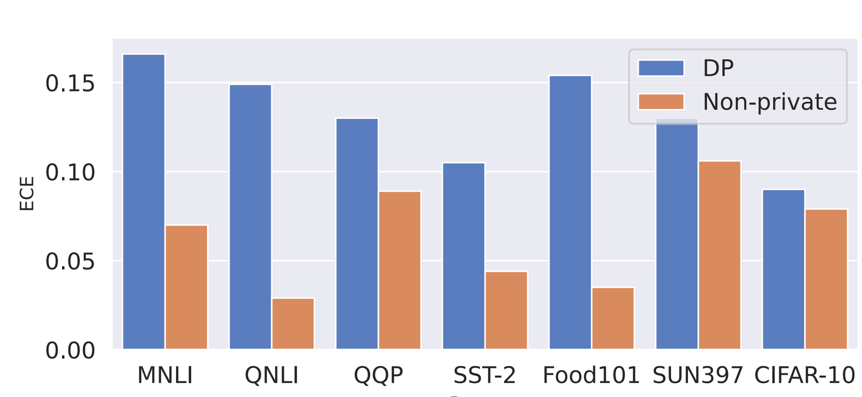
Definition

$\mathrm{ECE} - \sum_{m=1}^{M} \sum_{n} |B_m| \,\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\ |$, where $\mathrm{conf}(B_m) = \sum_{i \in B_m} \hat{p}_i \,/\, |B_m| and \mathrm{acc}(B_m) = \sum_{i \in B_m} \mathbf{1}(\hat{y}_i = y_i) / |B_m|$.

Motivations

Significance: Accessing model uncertainty is important for deploying models in safety-critical scenarios like healthcare and law where explainability (Cosmides & Tooby, 1996) and risk control (Van Calster et al., 2019) are needed in addition to privacy (Knolle et al., 2021).

Universality: DP is expected to generalize through stability but DP-SGD gives rise to miscalibration over a wide range of settings even if we use state-of-the-art pre-trained backbones.



(a) Non-separable Gaussian Data (b) Calibration comparison of logistic regression w and w/o DP

Algorithm 1: Differentially Private Recalibration

Input: $X = \{(\mathbf{x_1}, y_1), ..., (\mathbf{x_n}, y_n)\}$, validation ratio α.
Initial: Parameters of models h_θ, recalibrator g_ϕ.

1. $X_{\text{train}}, X_{\text{recal}} = \mathrm{RandomSplit}(X, \alpha)$
2. Train $h_\theta(\mathbf{x})$ using DP-SGD to optimize $\min_\theta \mathbb{E}\left[\ell\left(\mathrm{softmax}\left(h_\theta(\mathbf{x})\right), y\right)\right]$ with X_{train}
3. Train g_ϕ using DP-SGD to optimize $\min_\phi \mathbb{E}\left[\ell\left(\mathrm{softmax}\left(g_\phi \circ h_\theta(\mathbf{x})\right), y\right)\right]$ with X_{recal}

Output: $g_\phi \circ h_\theta(\cdot)$

Mitigation of Miscalibration

Post-hoc Recalibration (Algorithm 1)

Adjust the calibration of classifier h_θ by learning a g_ϕ that adjusts the log probabilities and produces a better calibrated forecast softmax $(g_\phi \circ h_\theta(\mathbf{x}))$ by solving $\min_\phi \mathbb{E}\left[\ell\left(\mathrm{softmax}\left(g_\phi \circ h_\theta(\mathbf{x})\right), y\right)\right]$.
We consider the differentially private variants of temperature scaling (DP-TS) $(g_\phi(\mathbf{x}) = \mathbf{x}/T)$ and the Platt scaling (DP-PS) $(g_\phi(\mathbf{x}) = \mathbf{W}\mathbf{x} + \mathbf{b})$ with the choice of log loss for ℓ.

Expriments

In-domain Evaluation DP trained models display consistently higher ECE than their non-private counterparts. The overall trend of miscalibration is clear across datasets and modalities. DP-TS and DP-PS perform consistently well, with a minor percentage drop of accuracy.

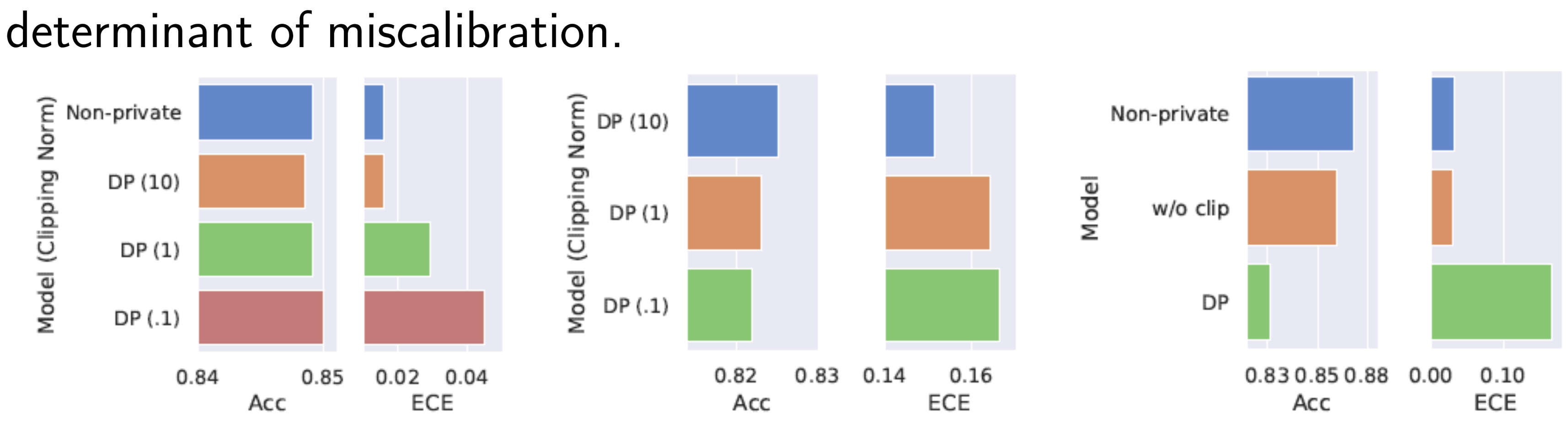
Category	Model	CIFAR-10		SUN397		Food101	
		Accuracy	**ECE**	**Accuracy**	**ECE**	**Accuracy**	**ECE**
Baseline	DP	0.7951	0.0903	0.6844	0.1302	0.7582	0.154
	DP-SGLD	0.7122	0.1331	0.6062	0.1952	0.6476	0.2416
	Global Clipping	0.7712	0.0804	0.6215	0.1125	0.7451	0.1017
Recalibration	DP-PS	0.789	**0.012**	0.674	0.104	0.7543	0.0554
	DP-TS	0.789	0.0221	0.674	**0.0763**	0.7543	**0.0540**
Non-private	DP+Non-private-TS	0.789	0.0222	0.674	0.0764	0.7543	0.0539
	Non-private	0.83	0.0794	0.7044	0.1062	0.8245	0.0349

Category	Model	MNLI		QNLI		QQP		SST-2	
		Accuracy	**ECE**	**Accuracy**	**ECE**	**Accuracy**	**ECE**	**Accuracy**	**ECE**
Baseline	DP	0.8281	0.166	0.8503	0.149	0.8685	0.13	0.8922	0.105
	DP-SGLD	0.7188	0.2625	0.7787	0.2138	0.7917	0.2009	0.82	0.1742
	Global Clipping	0.8236	0.1667	0.8502	0.1491	0.8685	0.1296	0.8922	0.1047
Recalibration	DP-PS	0.826	**0.0487**	0.8464	**0.0305**	0.8659	0.0672	0.8842	**0.0201**
	DP-TS	0.826	0.0849	0.8464	0.0915	0.8659	**0.0635**	0.8842	0.0665
Non-private	DP+Non-private-TS	0.826	0.0849	0.8464	0.0915	0.8659	0.0635	0.8842	0.0665
	Non-private	0.8642	0.0699	0.914	0.028	0.9042	0.0891	0.9323	0.0425

Out-of-domain Evaluation OOD results are consistent with the in-domain evaluations.

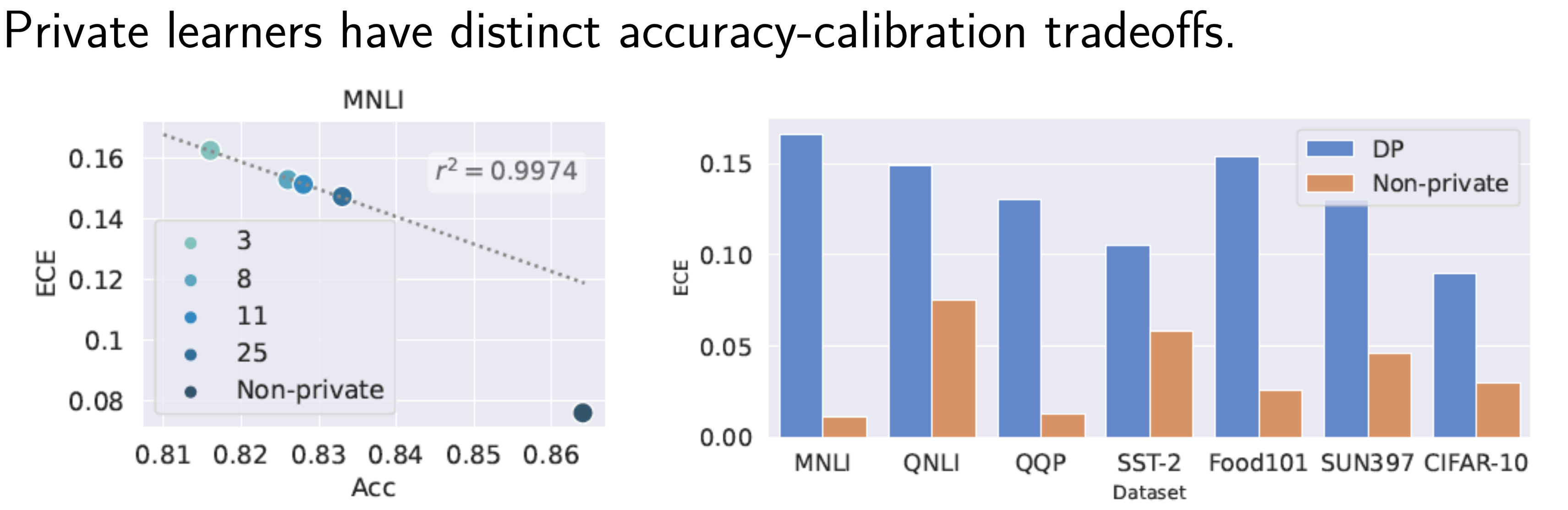
Dataset	Category	Model	Hans		Scitail		RTE		WNLI	
			Accuracy	**ECE**	**Accuracy**	**ECE**	**Accuracy**	**ECE**	**Accuracy**	**ECE**
MNLI	Baseline	DP	0.5195	0.4786	0.7761	0.2172	0.7437	0.2541	0.4507	0.5492
		DP-SGLD	0.4996	0.4995	0.7515	0.233	0.6498	0.3169	0.4507	0.5491
		Global Clipping	0.5221	0.4747	0.7845	0.2051	0.7076	0.2737	0.4366	0.5632
	Recalibration	DP-PS	0.5237	**0.348**	0.7707	**0.1089**	0.7220	**0.1516**	0.4366	**0.4416**
		DP-TS	0.5237	0.3544	0.7707	0.1168	0.7220	0.1593	0.4366	0.4495
	Non-private	DP+Non-private-TS	0.5237	0.3544	0.7707	0.1168	0.7220	0.1593	0.4366	0.4495
		Non-private	0.668	0.2687	0.7853	0.1348	0.7906	0.1518	0.507	0.4677
QNLI	Baseline	DP	0.5046	0.4932	0.729	0.2666	0.5657	0.4407	0.4724	0.5215
		DP-SGLD	0.5	0.4986	0.7209	0.2723	0.5668	0.4266	0.4225	0.5738
		Global Clipping	0.5025	0.4971	0.7293	0.2684	0.5199	0.4761	0.4789	0.52
	Recalibration	DP-PS	0.5002	**0.3244**	0.7377	**0.0832**	0.5632	**0.2578**	0.4648	**0.3464**
		DP-TS	0.5002	0.385	0.7377	0.1353	0.5632	0.3121	0.4648	0.404
	Non-private	DP+Non-private-TS	0.5002	0.385	0.7377	0.1353	0.5632	0.3121	0.4648	0.404
		Non-private	0.538	0.1969	0.7454	0.0690	0.5199	0.3036	0.5493	0.2438

Controlled Studies Per-example gradient clipping is identified as a major determinant of miscalibration.



(a) 2D synthetic data (b) MNLI (c) Gradient clipping ablation

Private learners have distinct accuracy-calibration tradeoffs.



(a) Privacy budget ϵ ablation (b) Comparison with non-private models with matched accuracies

NeurIPS 2022 Workshop on Algorithmic Fairness through the Lens of Causality and Privacy

✉ hanlinzh@cs.cmu.edu