Crowding Out The Noise: Algorithmic Collective Action Under Differential Privacy

Rushabh Solanki[1] **Meghana Bhange**[2] **Ulrich Aïvodji**[2] **Elliot Creager**[1]

[1]University of Waterloo, Vector Institute
[2]ÉTS Montréal, Mila

r7solank@uwaterloo.ca, creager@uwaterloo.ca
meghana-shashikant.bhange.1@etsmtl.net, ulrich.aivodji@etsmtl.ca

Abstract

The integration of AI into daily life has generated considerable attention and excitement, while also raising concerns about automating algorithmic harms and re-entrenching existing social inequities. While top-down solutions such as regulatory policies and improved algorithm design are common, the fact that AI trains on social data creates an opportunity for a grassroots approach, *Algorithmic Collective Action*, where users deliberately modify the data they share to steer a platform's learning process in their favor. This paper considers how these efforts interact with a firm's use of a differentially private model to protect user data, motivated by the growing regulatory focus on privacy and data protection. In particular, we investigate how the use of Differentially Private Stochastic Gradient Descent (DP-SGD) affects the collective's ability to influence the learning process. Our findings show that while differential privacy contributes to the protection of individual data, it introduces challenges for effective algorithmic collective action. We characterize lower bounds on the success of these actions as a function of the collective's size and the firm's privacy parameters, verifying these trends experimentally by training deep neural network classifiers across several datasets.

1 Introduction

The rapid proliferation of AI systems across multiple domains has been propelled by the ability of AI firms to collect vast amounts of data for training purposes, sourced from public websites, users of the firm's products, and crowd workers. This dependence on personal data has introduced pressing concerns about algorithmic harms, such as threats to privacy, exposure of sensitive information, and biased decision-making that perpetuates social disparities. In response to these concerns, firms have increasingly adopted fairness checks, privacy audits, and adversarial testing throughout the model development pipeline to build "trustworthy AI" [Barocas et al., 2023]. At the same time, these measures can conflict with goals like maximizing performance and user engagement. Meanwhile, regulations such as the EU's GDPR [European Parliament and Council of the European Union, 2016], Canada's PIPEDA [Government of Canada, 2000], and California's CPRA [State of California, 2020] set baseline privacy rules, but legal compliance alone does not ensure socially responsible outcomes [Selbst et al., 2019, Utz et al., 2019], underscoring the need for complementary strategies beyond organizational and regulatory measures.

In addition to these top-down measures, a grassroots effort in *Algorithmic Collective Action* (ACA) is taking shape [Hardt et al., 2023], where users actively organize and contribute their data in a coordinated manner to strategically influence model behavior "from below" [DeVrio et al., 2024]. ACA [Hardt et al., 2023, Olson, 1971] provides a principled framework for understanding how a

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Algorithmic Collective Action.

group of individuals, through coordinated changes in their data, can impact the behavior of deployed models. Prior work has explored this under assumptions like Bayes optimality, empirical risk minimization [Hardt et al., 2023], and robust optimization [Ben-Dov et al., 2024]. However, the interaction between the actions of coordinated users and privacy-preserving techniques employed by the model owners remains largely unexplored.

In this paper, we investigate the interplay between ACA and platform-side privacy measures, focusing on Differential Privacy (DP), a widely used method for protecting individual-level data through the injection of calibrated noise into the learning process. In particular, we study the application of differential privacy in deep learning settings through Differentially Private Stochastic Gradient Descent (DP-SGD), a common approach for preserving privacy during model training. Motivated by the strengthening of regulatory frameworks and growing consumer demand for privacy guarantees, we seek to understand how differential privacy affects an algorithm's responsiveness to collective action and the success rate of such interventions. We put forward a theory that examines the impact of differential privacy constraints on the effectiveness of a collective taking action on the firm's learning algorithm.

Our contributions are summarized as follows:

- We identify and characterize a trade-off between Differential Privacy and Algorithmic Collective Action. Our theoretical model characterizes lower bounds on the collective's success under DP constraints, in terms of the collective's size and the privacy parameters.
- These theoretical findings are validated through extensive experiments on multiple datasets, showing that DP reduces the collective's ability to influence the behavior of the model.
- We measure empirical privacy risk through the lens of membership inference attacks and observe minor variations associated with the presence of the collective in the data distribution.

2 Background

This section provides a formal introduction to algorithmic collective action and privacy-preserving training, and defines the notation used throughout the paper.

2.1 Collective Action

We adopt the framework of Hardt et al. [2023], where a collective, representing a fraction α of the data, aims to influence a firm's learning algorithm. The collective's goal is to steer the firm's model parameters θ_t toward a desired target θ^*. They implement a *feature-label strategy* by mapping each of their data points $(x,\, y)$ to a new point $(g(x),\, y^*)$, where g is an input transformation and y^* is a fixed target label. The firm's algorithm then trains on a mixture distribution $\mathcal{P} = \alpha\mathcal{P}^* + (1-\alpha)\mathcal{P}_0$, where $\mathcal{P}_0$ is the base distribution and $\mathcal{P}^*$ is the collective's modified distribution.

While there are several learning-theoretic settings explored in Hardt et al. [2023], this work focuses on characterizing the success criteria of the collective in the context of gradient-based optimization, where the learner essentially selects a model from a parameterized family $\{f_\theta\}_{\theta\in\Theta}$. The collective's success after t steps is measured by the proximity to the target model: $S_t(\alpha) = -\left\|\theta_t - \theta^*\right\|$. We are interested in finding the smallest size of the collective that can achieve a desired level of success, which is referred to as the *critical mass*, denoted by α^*.

At each time step t, the learner updates its parameters via gradient descent using the expected gradient under the current distribution $\mathcal{P}_t$ with $\theta_{t+1} = \theta_t - \eta\, g_{\mathcal{P}_t}(\theta_t)$, where $g_{\mathcal{P}_t}(\theta_t) = \mathbb{E}_{z\sim\mathcal{P}_t}\left[\nabla\ell(\theta_t; z)\right]$. To influence these updates, the collective constructs a gradient-redirecting distribution $\mathcal{P}'$ (defined in Appendix A.1) that both counteracts the original gradient $g_{\mathcal{P}_0}(\theta_t)$ and nudges the model toward a target θ^*. Using this strategy, Hardt et al. [2023] provide a lower bound on the collective's success and show that this success increases with the collective's size α, formalized as follows.

Theorem 1 (Theorem 10 from Hardt et al. [2023])**.** *Assume the collective can implement the gradient-redirecting strategy at all $\lambda\theta_0 + (1-\lambda)\theta^*, \lambda \in [0,1]$. Then, there exists $C(\alpha) > 0$ such that the success of the gradient-redirecting strategy after T steps is lower bounded by,*

$$S_T(\alpha) \geq -\left(1 - \eta C(\alpha)\right)^T \left\|\theta_0 - \theta^*\right\|.$$

where $C(\alpha)$ is directly proportional to collective's size α.

2.2 Privacy-preserving Training

To incorporate privacy constraints, we use Differential Privacy, a formal framework that ensures an algorithm's output does not reveal significant information about any individual in the training data [Dwork et al., 2006]. An algorithm (or "mechanism") is said to be differentially private if, for any two neighboring datasets that differ in a single record, the probability of any output changes only slightly. The formal definition of DP from Dwork et al. [2006] is presented as follows.

Definition 1 ((ε, δ)-Differential Privacy)**.** A randomized mechanism $M : D \to R$ with domain D and range R satisfies (ε, δ)-differential privacy if for any two neighboring inputs $d, d' \in D$ and for any subset of outputs $S \subseteq R$, it holds that

$$\Pr[M(d) \in S] \leq e^{\varepsilon} \Pr[M(d') \in S] + \delta.$$

At a high level, ε quantifies the extent to which a single data point can influence the algorithm's output, while δ accounts for a small probability of exceeding the bound. We use Differentially Private Stochastic Gradient Descent (DP-SGD) [Abadi et al., 2016], a practical algorithm that achieves DP by clipping per-sample gradients and adding calibrated Gaussian noise to the batch gradient at each training step. The amount of noise is controlled by a multiplier σ, which is inversely related to the privacy budget ε. Consequently, stronger privacy (smaller ε) requires more noise, which can degrade model utility and potentially interfere with the collective's efforts.

3 Collective Action under Differential Privacy

In this section, we provide a theoretical framework that characterizes bounds on the success of the collective action under DP constraints. Our approach builds on and extends the foundational work of Hardt et al. [2023], who initiated a principled study of the collective interacting with the firm's learning algorithm.

Problem setup Given a data distribution $\mathcal{P}$, we assume the firm employs a private learning algorithm $\mathcal{A}$ to produce a model $f = \mathcal{A}(\mathcal{P})$. We consider a realistic learning scenario without convexity assumptions on the objective function, using Differentially Private Stochastic Gradient Descent (DP-SGD), a widely used algorithm for training models under (ε, δ)-differential privacy constraints. At each time step t, the collective interacts with the learner by choosing a distribution $\mathcal{P}_t^*$ [Hardt et al., 2023]. This models the best-case scenario for the collective, enabling us to analyze the potential effectiveness of its strategy under ideal conditions. Unlike prior work that analyzed this interaction using expected gradients, we model the explicit stochasticity of the process: the learner observes a batch $\mathcal{B}_t$ sampled i.i.d. from $\mathcal{P}_t^*$. Given a clipping threshold C and a noise scale σ, the model parameters are updated by taking the gradient step computed according to the DP-SGD:

$$\begin{aligned}\theta_{t+1} &= \theta_t - \eta \frac{1}{|\mathcal{B}_t|} \left(\left(\sum_{z \in \mathcal{B}_t} \text{clip}(\nabla \ell(\theta_t; z),\, C) \right) + \mathcal{N}(0,\, \sigma^2 C^2 I_d) \right) \\ &= \theta_t - \eta \left(\bar{g}_{\mathcal{B}_t}^{\text{clip}}(\theta_t) + \mathcal{N}\left(0,\, \frac{\sigma^2 C^2}{|\mathcal{B}_t|^2} I_d\right) \right) = \theta_t - \eta\, \bar{g}_{\mathcal{B}_t}^{\text{DP}}(\theta_t),\end{aligned}$$

where $\bar{g}_{\mathcal{B}_t}$ denotes the average gradients over batch $\mathcal{B}_t$ and $\text{clip}(g, C) = g \cdot \min(1, C/\|g\|)$ denotes the gradient clipping operation, which scales the gradient g so that it has norm of at most C.

Theoretical results The most intuitive factor that limits the success of the collective when the firm uses DP-SGD is the algorithm's inherent ability to limit the influence of any individual data point on the model's output. Gradient clipping reduces the collective's ability to align the gradients with their desired direction, while the injected noise further deflects this directional push. As a result, the signal that the collective is trying to correlate with the target label also gets attenuated. This is equivalent to the collective introducing a noisy signal, which in turn increases the efforts required for the collective to influence the outcome. We now formalize this idea.

Theorem 2. *Assume that the collective can implement the gradient-redirecting strategy from Definition 2 at all $\lambda\theta_0 + (1-\lambda)\theta^*$, where $\lambda \in [0,1]$ and $\theta_0, \theta^* \in \mathbb{R}^d$. Then, for a given clipping threshold C and noise multiplier σ, and letting $|\mathcal{B}|_{\min}$ denote the minimum batch size observed over*

T steps, there exists $B_{\alpha,C} > 0$, such that the success of the gradient-control strategy after T steps is lower-bounded with probability greater than $1 - \delta$ by,

$$S_T(\alpha, \sigma, C) \geq -\left(1 - \eta B_{\alpha,C}\right)^T \|\theta_0 - \theta^*\| - \frac{\sigma\, C}{|\mathcal{B}|^2_{\min}} \Gamma_T\, \Delta_{d,\delta}. \tag{1}$$

Here, $B_{\alpha,C}$ is directly proportional to the collective's size α and clipping threshold C, the function Γ_T denotes convergence-dependent noise-accumulation factor, and $\Delta_{d,\delta}$ is a high-probability upper bound on the Euclidean norm of a d-dimensional standard Gaussian vector. See subsection A.3 for the proof and Equation (12) for the closed-form expressions. Setting $C = \infty$, which corresponds to no clipping being applied to the gradient and considering the expected gradients, recovers $C(\alpha)$–stated in Theorem 1–from $B(\alpha, C)$. In addition, setting the noise scale $\sigma = 0$, reducing the learner to the standard SGD algorithm, eliminates the second term and reconstructs the bound from Theorem 1.

Relation between privacy parameters and success Theorem 2 shows that the success of the collective is inversely proportional to noise scale σ. From Section 2.2, we know that increasing σ leads to lower privacy loss ε, meaning stronger privacy guarantees. Therefore, tightening privacy constraints by increasing σ adversely affects the collective's success.

Next, we examine the clipping threshold C, set by the firm, which influences the collective's success through a more complex relationship than σ. Since the upper bound of the clipped gradient, $B(\alpha, C)$, increases with C, the term $-\left(1 - \eta B(\alpha, C)\right)^T$ also increases, positively contributing to the collective's success. However, the second term introduces two opposing effects: while the linear dependence on C tends to reduce success as C increases, the factor Γ_T (which itself decreases with C) counteracts this negative trend. As a result, the overall impact of C on the success of the collective is determined by the interplay between these competing influences.

4 Experiments

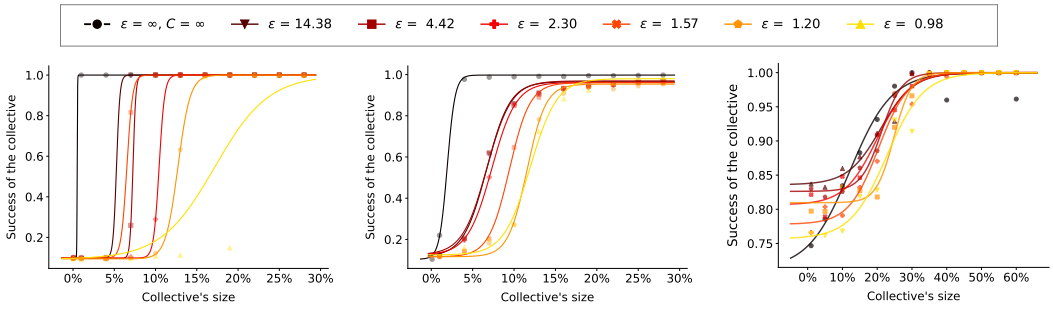
This section presents our experimental evaluation of how the critical mass of the collective changes when using a differentially private learning algorithm. We perform experiments on image benchmarks, MNIST [LeCun and Cortes, 2010] and CIFAR-10 [Krizhevsky and Hinton, 2009], as well as a tabular Bank Marketing dataset [Moro et al., 2014]. We further evaluate how the collective's presence and size affect vulnerability to membership inference attacks under private and non-private settings.

4.1 Strategy and Success of the Collective

As discussed in Section 2.1, we consider a setting where a collective controls a fraction $\alpha \in [0, 1]$ of the training data and seeks to influence the model by embedding a signal g into the data they control, steering predictions on transformed inputs toward a target label y^*. The collective's success is measured by $S(\alpha) = \Pr_{x \sim \mathcal{P}_0}\{f(g(x)) = y^*\}$, representing the probability that the trained model predicts the target label on evaluation data with the planted signal. In practice, this corresponds to the accuracy on a modified test set where the signal is planted in all test examples. Our goal is to determine the *critical mass* α^*, the smallest fraction of controlled data needed to achieve a desired success rate, which we estimate by training models on datasets with varying levels of collective control and evaluating their performance on signal-planted test data.

4.2 Experimental Setup

We conduct experiments on three datasets: MNIST, CIFAR-10, and the UCI Bank Marketing dataset. For the image datasets, we use a ResNet-18 model with group normalization replacing batch normalization [Kurakin et al., 2022, Luo et al., 2021]. The model used to train CIFAR-10 is pre-trained on CIFAR-100 to improve performance under differential privacy. For the tabular Bank Marketing data, we use a simple feedforward neural network. The transformation g for MNIST sets the top-left 2×2 pixel patch to 50, with the target label reassigned to class "8" (digit 8). For CIFAR-10, g alters every second pixel in every second row by a magnitude of 2 (handling overflow to stay within the $[0, 255]$ range), and relabels images as class "8" (ship). Finally, for the Bank Marketing task, g adds an offset of 50 to a single feature, and the transformed data is relabeled to class "0". Further experimental details are provided in Appendix B.



(a) MNIST (b) CIFAR-10 (c) Bank Marketing

Figure 1: The success of the collective across ε on MNIST, CIFAR-10, and Bank Marketing datasets with fixed clipped threshold ($C = 1$). For each plot, we evaluate the collective's success under different values of privacy budget ε and compare it with the baseline case ($\varepsilon = \infty, C = \infty$), which corresponds to SGD without any privacy constraints. Collective size $\alpha \in [0, 1]$ is reported as a percentage of the overall training dataset. For results with other clipping thresholds, see Appendix C.1.

4.3 Results

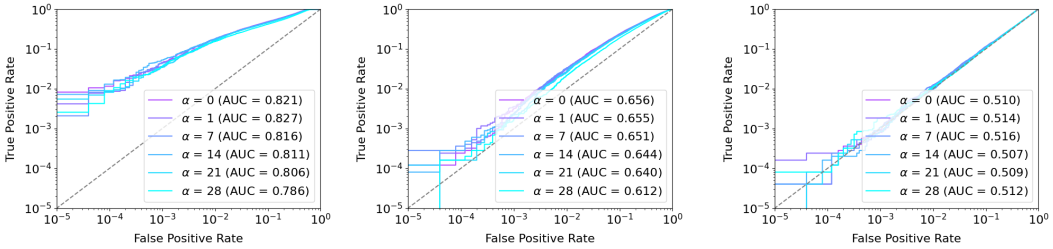
We evaluate the success of the collective by training multiple models, each using a dataset where the collective controls a different amount of the data. Figure 1 shows a clear trend for all three datasets: as the privacy loss decreases (corresponding to higher privacy), the critical mass required for success increases. This observation aligns with the theoretical results in Section 3, where the collective's success in Theorem 2 is inversely proportional to the noise scale σ, which appears in the second term of the bound. Consequently, when a firm deploys a model that prioritizes privacy at the expense of accuracy, it negatively raises the threshold for effective collective action. In such scenarios, greater coordination and organizational strength are required for the collective to accomplish its objective. These findings reveal a trade-off between DP and ACA. While stricter privacy protections are beneficial from regulatory or accountability perspectives, they increase the burden on groups of individuals adversely affected by model outcomes who aim to influence the model's behavior.

4.4 Membership Inference Attack Evaluations

Although DP offers formal guarantees against information leakage, models trained without it usually quantify privacy risks using empirical methods, such as membership inference. We use Likelihood Ratio Attack (LiRA) Carlini et al. [2022] to evaluate the membership status of a data point in the training set. Figure 2 presents the ROC curves showing LiRA's success rates under various configurations on the CIFAR-10 dataset, while Table 1 reports the corresponding AUC scores and the TPR @ 0.1% FPR values. We find that the collective action during training slightly decreases the attack success rate. This phenomenon can be attributed to the impact of the collective's impact on the model's ability to overfit the original training distribution. As the collective's size increases, the model is increasingly exposed to data containing conflicting samples, which effectively act as a form of implicit regularization. Moreover, the introduction of the collective does not appear to compromise the robustness to MIA already provided by DP training across different privacy levels. Consistent results on the SVHN dataset are presented in Appendix C.2.

5 Related Works

Collective Action Algorithmic Collective Action (ACA) adapts classic theories of collective action [Olson, 1971] to modern socio-technical systems, defining a setting where users coordinate to steer a machine learning model's output toward a group objective [Hardt et al., 2023]. This builds on the idea of Data Leverage, where individuals do not treat their data as passive inputs, but rather as levers that can be used to influence the outcome of the algorithmic system [Vincent et al., 2020]. ACA may also face challenges similar to those faced with other modes of collective action (i.e. in non-algorithmic settings), such as free-riding, where non-participants benefit from the collective



(a) No privacy ($varepsilon = 0$) (b) Fixed privacy loss ($\varepsilon = 15$) (c) Fixed privacy loss ($\varepsilon = 1$)

Figure 2: Success rate of Likelihood Ratio Attack (LiRA) [Carlini et al., 2022] evaluation on CIFAR-10 dataset. Each figure corresponds to a different setting of privacy constraints with privacy increasing from left to right.

effort [Sigg et al., 2024]. Moreover, its effectiveness is strongly influenced by the characteristics of the learning algorithm under consideration [Ben-Dov et al., 2024].

Data Poisoning At a technical level, the strategy followed for ACA is closely related to *data poisoning* attacks, where the adversary manipulates the training dataset to degrade the performance of a predictive model. Comprehensive surveys on data poisoning and backdoor attacks are provided by Tian et al. [2022] and Guo et al. [2022]. While data poisoning involves malicious manipulation, ACA is not inherently adversarial and often pursues constructive objectives. In addition, ACA emphasizes coordination among the collective, often to align with societal or personal objectives. Nonetheless, foundational work in poisoning shows that such data manipulation is feasible against various models, including those using semi-supervised learning or differential privacy [Gu et al., 2019, Shejwalkar et al., 2023, Ma et al., 2019].

Private Machine Learning DP is a gold standard for privacy in machine learning [Cummings et al., 2024], achieved through methods like output, objective, or gradient perturbation [Chaudhuri et al., 2011, Kifer et al., 2012, Bassily et al., 2014], with DP-SGD being prevalent for deep learning [Abadi et al., 2016]. A differentially private mechanism with privacy budget ε implicitly offers group privacy with a privacy $k\varepsilon$ for any group of size k [Dwork et al., 2014, Thm 2.2]. However, for real-world scenarios with possibly large group sizes, differential privacy offers limited protection. To account for these settings, variants such as attribute differential privacy [Zhang et al., 2022] have been proposed, but their integration in modern machine learning training algorithms remains challenging.

Trade-offs in Trustworthy ML Trustworthy machine learning integrates principles such as security, privacy, and fairness, but several recent studies have shown that they can be in tension with each other. For instance, fairness can decrease utility [Menon and Williamson, 2018, Yaghini et al., 2023], robustness can weaken privacy [Song et al., 2019], and explainability can be exploited for "fairwashing" or model stealing [Aïvodji et al., 2019, Milli et al., 2019]. This work contributes to raising awareness of the trade-off between privacy and ACA's effectiveness.

6 Conclusion

In this paper, we focus on the intersection of Algorithmic Collective Action and Differential Privacy. Specifically, we examined how privacy-preserving training using DP-SGD affects the ability of a collective to influence model behavior through coordinated data contributions. Our key contributions are a theoretical characterization and empirical validation of the limitations that differential privacy imposes on collective action, highlighting how the collective's success depends on the model's privacy parameters. We further evaluated empirical privacy through membership inference attacks and investigated the trends arising from the collective's presence in the training data. More broadly, this work offers a novel perspective on the societal implications of using privacy-preserving techniques in machine learning, highlighting important trade-offs between individual data protection and the capacity for collective influence over decision-making systems.

Acknowledgements

The resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute www.vectorinstitute.ai/partnerships/. The authors thank the Digital Research Alliance of Canada for computing resources. Ulrich Aïvodji is supported by NSERC Discovery grant (RGPIN-2022-04006) and IVADO's Canada First Research Excellence Fund to develop Robust, Reasoning and Responsible Artificial Intelligence (R^3AI) grant (RG-2024-290714).

References

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, October 2016. doi: 10.1145/2976749.2978318. URL http://dx.doi.org/10.1145/2976749.2978318.

Ulrich Aïvodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp. Fairwashing: the risk of rationalization. In *International Conference on Machine Learning*, pages 161–170. PMLR, 2019.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT press, 2023.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Differentially private empirical risk minimization: Efficient algorithms and tight error bounds, 2014. URL https://arxiv.org/abs/1405.7085.

Omri Ben-Dov, Jake Fawkes, Samira Samadi, and Amartya Sanyal. The role of learning algorithms in collective action. In *International Conference on Machine Learning*, pages 3443–3461. PMLR, 2024.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *2022 IEEE symposium on security and privacy (SP)*, pages 1897–1914. IEEE, 2022.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

Rachel Cummings, Damien Desfontaines, David Evans, Roxana Geambasu, Yangsibo Huang, Matthew Jagielski, Peter Kairouz, Gautam Kamath, Sewoong Oh, Olga Ohrimenko, et al. Advancing differential privacy: Where we are now and future directions for real-world deployment. *Harvard Data Science Review*, 6(1), 2024.

Alicia DeVrio, Motahhare Eslami, and Kenneth Holstein. Building, shifting, & employing power: A taxonomy of responses from below to algorithmic harm. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1093–1106, 2024.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In Serge Vaudenay, editor, *Advances in Cryptology - EUROCRYPT 2006*, pages 486–503, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-34547-3.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union, L 119, pp. 1–88, 4 May 2016, 2016. URL https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679.

Government of Canada. Personal information protection and electronic documents act. https://laws-lois.justice.gc.ca/eng/acts/P-8.6/, 2000. URL https://laws-lois.justice.gc.ca/eng/acts/P-8.6/. S.C. 2000, c. 5.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain, 2019. URL https://arxiv.org/abs/1708.06733.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

Wei Guo, Benedetta Tondi, and Mauro Barni. An overview of backdoor attacks against deep neural networks and possible defences. *IEEE Open Journal of Signal Processing*, 3:261–287, 2022.

Moritz Hardt, Eric Mazumdar, Celestine Mendler-Dünner, and Tijana Zrnic. Algorithmic collective action in machine learning. In *International Conference on Machine Learning*, pages 12570–12586. PMLR, 2023.

Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private convex empirical risk minimization and high-dimensional regression. In Shie Mannor, Nathan Srebro, and Robert C. Williamson, editors, *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 25.1–25.40, Edinburgh, Scotland, 25–27 Jun 2012. PMLR. URL https://proceedings.mlr.press/v23/kifer12.html.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Alexey Kurakin, Shuang Song, Steve Chien, Roxana Geambasu, Andreas Terzis, and Abhradeep Thakurta. Toward training at imagenet scale with differential privacy, 2022. URL https://arxiv.org/abs/2201.12328.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302 – 1338, 2000. doi: 10.1214/aos/1015957395. URL https://doi.org/10.1214/aos/1015957395.

Yann LeCun and Corinna Cortes. MNIST handwritten digit database. http://yann.lecun.com/exdb/mnist/, 2010. URL http://yann.lecun.com/exdb/mnist/.

Zelun Luo, Daniel J. Wu, Ehsan Adeli, and Li Fei-Fei. Scalable differential privacy with sparse network finetuning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5057–5066, 2021. doi: 10.1109/CVPR46437.2021.00502.

Yuzhe Ma, Xiaojin Zhu, and Justin Hsu. Data poisoning against differentially-private learners: Attacks and defenses. In *International Joint Conference on Artificial Intelligence*, 2019.

Aditya Krishna Menon and Robert C Williamson. The cost of fairness in binary classification. In *Conference on Fairness, accountability and transparency*, pages 107–118. PMLR, 2018.

Smitha Milli, Ludwig Schmidt, Anca D Dragan, and Moritz Hardt. Model reconstruction from model explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 1–9, 2019.

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, page 263–275. IEEE, August 2017. doi: 10.1109/csf.2017.11. URL http://dx.doi.org/10.1109/CSF.2017.11.

Sérgio Moro, Paulo Cortez, and Paulo Rita. Bank Marketing. UCI Machine Learning Repository, 2014. DOI: https://doi.org/10.24432/C5K306.

Mancur Olson. *The Logic of Collective Action: Public Goods and the Theory of Groups, Second Printing with a New Preface and Appendix*. Harvard University Press, 1971. ISBN 9780674537507. URL http://www.jstor.org/stable/j.ctvjsf3ts.

Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020.

Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 59–68, 2019.

Virat Shejwalkar, Lingjuan Lyu, and Amir Houmansadr. The perils of learning from unlabeled data: Backdoor attacks on semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4730–4740, 2023.

Dorothee Sigg, Moritz Hardt, and Celestine Mendler-Dünner. Decline now: A combinatorial model for algorithmic collective action. *ArXiv*, abs/2410.12633, 2024.

Liwei Song, Reza Shokri, and Prateek Mittal. Membership inference attacks against adversarially robust deep learning models. In *2019 IEEE Security and Privacy Workshops (SPW)*, pages 50–56. IEEE, 2019.

State of California. California privacy rights act of 2020 (cpra). `https://oag.ca.gov/privacy/ccpa`, 2020. URL `https://oag.ca.gov/privacy/ccpa`. Proposition 24, approved November 3, 2020.

Zhiyi Tian, Lei Cui, Jie Liang, and Shui Yu. A comprehensive survey on poisoning attacks and countermeasures in machine learning. *ACM Computing Surveys*, 55(8):1–35, 2022.

Christine Utz, Martin Degeling, Sascha Fahl, Florian Schaub, and Thorsten Holz. (un) informed consent: Studying gdpr consent notices in the field. In *Proceedings of the 2019 acm sigsac conference on computer and communications security*, pages 973–990, 2019.

Nicholas Vincent, Hanlin Li, Nicole Tilly, Stevie Chancellor, and Brent J. Hecht. Data leverage: A framework for empowering the public in its relationship with technology companies. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2020.

Mohammad Yaghini, Patty Liu, Franziska Boenisch, and Nicolas Papernot. Learning with impartiality to walk on the pareto frontier of fairness, privacy, and utility. *arXiv preprint arXiv:2302.09183*, 2023.

Wanrong Zhang, Olga Ohrimenko, and Rachel Cummings. Attribute privacy: Framework and mechanisms. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 757–766, 2022.

A Theoretical Results

A.1 Gradient-redirecting distribution

Definition 2 (Gradient-redirecting distribution from Hardt et al. [2023])**.** Given an observed model θ and a target model θ^*, the collective finds a *gradient-redirecting distribution* $\mathcal{P}'$ for θ where:

$$g_{\mathcal{P}'}(\theta) = -\frac{1-\alpha}{\alpha} g_{\mathcal{P}_0}(\theta) + \xi \cdot (\theta - \theta^*),$$

for some $\xi \in \left(0, \frac{1}{\alpha\eta}\right)$. Once such a distribution is identified, we can sample modified data $z' \sim \mathcal{P}'$ to guide the optimization process by setting $h(z) = z'$.

Intuitively, the gradient under distribution $\mathcal{P}'$ is composed of two terms—one that reverses and rescales the original gradients $g_{\mathcal{P}_0}(\theta)$, and another that pushes the parameters toward the collective's desired model θ^*.

A.2 Tail bounds on norm of scaled standard Gaussian distribution

Lemma 1. *Let $Y_1, \ldots, Y_D \sim \mathcal{N}(0, \sigma^2)$ be independent Gaussian random variable, and define the scaled chi-squared distribution as,*

$$S = \|Y\|_2 = \sqrt{\sum_{i=1}^{D} Y_i^2}$$

Then, for any $\delta \in (0, 1)$, with probability of $1 - \delta$,

$$S \leq \sigma \left(\sqrt{D} + \sqrt{2 \log 1/\delta}\right) \tag{2}$$

Proof. Since each $Y_i \sim \mathcal{N}(0, \sigma^2)$, we can write $Y_i = \sigma Z_i$, where $Z_i \sim \mathcal{N}(0, 1)$. Then,

$$S = \sqrt{\sum_{i=1}^{D} Y_i^2} = \sigma \sqrt{\sum_{i=1}^{D} Z_i^2} = \sigma\sqrt{U}$$

Let $U = \sum_{i=1}^{D} Z_i^2$. A standard tail bound for the chi-squared distribution (refer to the Corollary 1 in Laurent and Massart [2000]) gives, for any $t > 0$,

$$\mathbb{P}\left(U \geq D + 2\sqrt{Dt} + 2t\right) \leq e^{-t}$$

Substituting $t = \log(1/\delta)$ we can obtain, with probability at least $1 - \delta$,

$$\begin{aligned} U &\leq D + 2\sqrt{D \log(1/\delta)} + 2\log(1/\delta) \\ &\leq D + 2\sqrt{2D \log(1/\delta)} + 2\log(1/\delta) \\ &= \left(\sqrt{D} + \sqrt{2\log(1/\delta)}\right)^2 \end{aligned}$$

Taking square root and multiplying by σ on both sides, we get the required bounds.

A.3 Proof for Theorem 2

Assuming the collective implements the gradient-redirecting strategy (Definition 2), we can express the average clipped gradient $\bar{g}$ over the samples $\mathcal{B}_t$ drawn from the collective's distribution $\mathcal{P}_t$ as[1]:

$$\bar{g}^{\text{clip}}_{\mathcal{B}_t}(\theta_t) = \frac{1}{|\mathcal{B}_t|}\left(\sum_{z\in\mathcal{B}'_t}\text{clip}(\nabla\,\ell(\theta_t;z),\,C) + \sum_{z\in\mathcal{B}_0}\text{clip}(\nabla\,\ell(\theta_t;z),\,C)\right) \tag{3}$$

$$= \alpha\,\bar{g}^{\text{clip}}_{\mathcal{B}'_t}(\theta_t) + (1-\alpha)\,\bar{g}^{\text{clip}}_{\mathcal{B}_0}(\theta_t) \tag{4}$$

$$\bar{g}^{\text{DP}}_{\mathcal{B}_t}(\theta_t) = \alpha\,\bar{g}^{\text{clip}}_{\mathcal{B}'_t}(\theta_t) + (1-\alpha)\,\bar{g}^{\text{clip}}_{\mathcal{B}_0}(\theta_t) + \mathcal{N}\left(0,\,\frac{\sigma^2C^2}{|\mathcal{B}_t|^2}I_d\right) \tag{5}$$

$$= \alpha\,\xi^c(\theta_t)\,(\theta_t-\theta^*) + \mathcal{N}\left(0,\,\frac{\sigma^2C^2}{|\mathcal{B}_t|^2}I_d\right) \tag{6}$$

$$\text{where}\quad \xi^c(\theta_t) = \frac{\left\|\bar{g}^{\text{clip}}_{\mathcal{B}'_t}(\theta_t) + \frac{1-\alpha}{\alpha}\,\bar{g}^{\text{clip}}_{\mathcal{B}_0}(\theta_t)\right\|}{\|\theta_t-\theta^*\|},$$

where to get Equation 6, we start by following a similar strategy to Hardt et al. [2023], but in our case we express the average clipped gradients as a scalar multiple of the model update direction $(\theta_t - \theta^*)$. Refer to Definition 2. With $\xi^c_{min} = \min_{\lambda\in[0,1]}\xi(\lambda\theta_0 + (1-\lambda)\theta^*)$, and using parameters update equation, we can derive an upper bound on the difference between the learned and optimal parameter as follows:

$$\begin{aligned}
\|\theta_T-\theta^*\| &\le \left\|\theta_{T-1} - \eta\left(\bar{g}^{\text{clip}}_{\mathcal{B}_T}(\theta_{T-1}) + \mathcal{N}\left(0,\,\frac{\sigma^2C^2}{|\mathcal{B}_T|^2}I_d\right)\right) - \theta^*\right\| \\
&\le \left\|\theta_{T-1} - \eta\left(\alpha\,\xi^c\,(\theta_{T-1})\,(\theta_{T-1}-\theta^*) + \mathcal{N}\left(0,\,\frac{\sigma^2C^2}{|\mathcal{B}_T|^2}I_d\right)\right) - \theta^*\right\| \\
&= \left\|(1-\eta\,\alpha\,\xi^c(\theta_{T-1}))\,(\theta_{T-1}-\theta^*) - \eta\,\mathcal{N}\left(0,\,\frac{\sigma^2C^2}{|\mathcal{B}_T|^2}I_d\right)\right\| \\
&\le \left\|(1-\eta\,\alpha\,\xi^c_{\text{min}})\,(\theta_{T-1}-\theta^*) - \eta\,\mathcal{N}\left(0,\,\frac{\sigma^2C^2}{|\mathcal{B}_T|^2}I_d\right)\right\| \qquad (7) \\
&= \left\|(1-\eta\,\alpha\,\xi^c_{\text{min}})^2\,(\theta_{T-2}-\theta^*) - \eta\,\mathcal{N}(0,\,\sigma^2C^2I_d)\left(\frac{1}{|\mathcal{B}_T|} + \frac{(1-\eta\,\alpha\,\xi^c_{\text{min}})}{|\mathcal{B}_{T-1}|}\right)\right\| \\
&= \left\|(1-\eta\,\alpha\,\xi^c_{\text{min}})^T\,(\theta_0-\theta^*) - \eta\,\mathcal{N}(0,\,\sigma^2C^2I_d)\sum_{k=0}^{T-1}\frac{(1-\eta\,\alpha\,\xi^c_{\text{min}})^k}{|\mathcal{B}_{T-k}|}\right\| \qquad (8) \\
&\overset{d}{=} \left\|(1-\eta\,\alpha\,\xi^c_{\text{min}})^T\,(\theta_0-\theta^*) + \eta\,\mathcal{N}(0,\,\sigma^2C^2I_d)\sum_{k=0}^{T-1}\frac{(1-\eta\,\alpha\,\xi^c_{\text{min}})^k}{|\mathcal{B}_{T-k}|}\right\| \qquad (9) \\
&\le (1-\eta\,\alpha\,\xi^c_{\text{min}})^T\,\|\theta_0-\theta^*\| + \eta\sum_{k=0}^{T-1}\frac{(1-\eta\,\alpha\,\xi^c_{\text{min}})^k}{|\mathcal{B}_{T-k}|}\,\|\mathcal{N}(0,\,\sigma^2C^2I_d)\| \qquad (10)
\end{aligned}$$

[1]This implicitly assumes that the collective has some knowledge of the firm's clipping operation when selecting the gradient-redirecting distribution.

Applying Lemma 1, the right-hand side can be further upper-bounded with probability at least $1-\delta$, for θ with d degrees of freedom,

$$\begin{aligned}
&\leq (1-\eta\,\alpha\,\xi_{\min}^{c})^{T}\,\|\theta_0-\theta^*\| + \eta\sum_{k=0}^{T-1}\frac{(1-\eta\,\alpha\,\xi_{\min}^{c})^{k}}{|\mathcal{B}_{T-k}|}\,\sigma\,C\left(\sqrt{d}+\sqrt{2\log 1/\delta}\right)\\
&\leq (1-\eta\,\alpha\,\xi_{\min}^{c})^{T}\,\|\theta_0-\theta^*\| + \eta\left(\sum_{k=0}^{T-1}(1-\eta\,\alpha\,\xi_{\min}^{c})^{k}\right)\frac{\sigma\,C}{|\mathcal{B}|_{\min}^{2}}\left(\sqrt{d}+\sqrt{2\log 1/\delta}\right) \qquad (11)\\
&= (1-\eta\,\alpha\,\xi_{\min}^{c})^{T}\,\|\theta_0-\theta^*\| + \eta\,\frac{1-(1-\eta\,\alpha\,\xi_{\min}^{c})^{T}}{\alpha\,\xi_{\min}^{c}}\,\frac{\sigma\,C}{|\mathcal{B}|_{\min}^{2}}\left(\sqrt{d}+\sqrt{2\log 1/\delta}\right)
\end{aligned}$$

Let $B_{\alpha,C}=\alpha\xi_{\min}^{c}$, $\Gamma_T=\eta\left(1-(1-\eta B_{\alpha,C})^{T}\right)/B_{\alpha,C}$ and $\Delta_{d,\delta}=\left(\sqrt{d}+\sqrt{2\log 1/\delta}\right)$; then we have

$$\|\theta_T-\theta^*\| \leq (1-\eta\,B_{\alpha,C})^{T}\,\|\theta_0-\theta^*\| + \frac{\sigma\,C}{|\mathcal{B}|_{\min}^{2}}\,\Gamma_T\,\Delta_{d,\delta} \qquad (12)$$

Multiplying both sides by -1 transforms the left-hand side into the collective's success. This converts the upper bound on the parameter norm difference into a lower bound on collective's success, resulting in the final bound:

$$S_T(\alpha,\sigma,C) \geq -\,(1-\eta\,B_{\alpha,C})^{T}\,\|\theta_0-\theta^*\| - \frac{\sigma\,C}{|\mathcal{B}|_{\min}^{2}}\,\Gamma_T\,\Delta_{d,\delta}.$$

Additional details for some steps of the proof are provided for clarity. In step (7), we substitute θ_{T-1} with a smaller value, yielding a relaxed upper bound. Step (8) involves unrolling the gradient-update recursion, which leads to a geometric series whose first term is 1 and common ratio $1-\eta\,\alpha\,\xi_{\min}^{c}$. In step (9), we use the fact that adding or subtracting a zero-centered Gaussian random variable results in random variables that are equal in distribution. In step (10), we apply the triangle inequality, where the sum of norms is greater than or equal to the norm of the sum. Finally, in step (11), we bound the sum by assuming a minimum batch size $|\mathcal{B}|_{\min}$ for all T steps.

B Experimental Details

For all experiments, we use the Stochastic Gradient Descent (SGD) optimizer for non-private training and the Differentially Private SGD (DP-SGD) optimizer in privacy-preserving settings. In the ResNet-18 models, batch normalization layers are replaced with group normalization to ensure compatibility with DP-SGD, as this modification enables accurate per-sample gradient computation required for enforcing differential privacy [Kurakin et al., 2022, Luo et al., 2021].

The experimental datasets are constructed as follows. For MNIST, we generate a balanced training set by uniformly sampling 5,000 images per class, yielding 50,000 training samples, while the test set remains unchanged at 10,000 images. The CIFAR-10 dataset is used in its standard split of 50,000 training and 10,000 test images. In addition, CIFAR-100, comprising 60,000 images (32×32 resolution) across 100 classes, is used for pre-training the CIFAR-10 model. For tabular data, we utilize the UCI Bank Marketing dataset, which contains 45,211 samples with 17 features for a binary classification task. The corresponding feedforward neural network consists of a single hidden layer with 128 ReLU units, followed by a fully connected output layer.

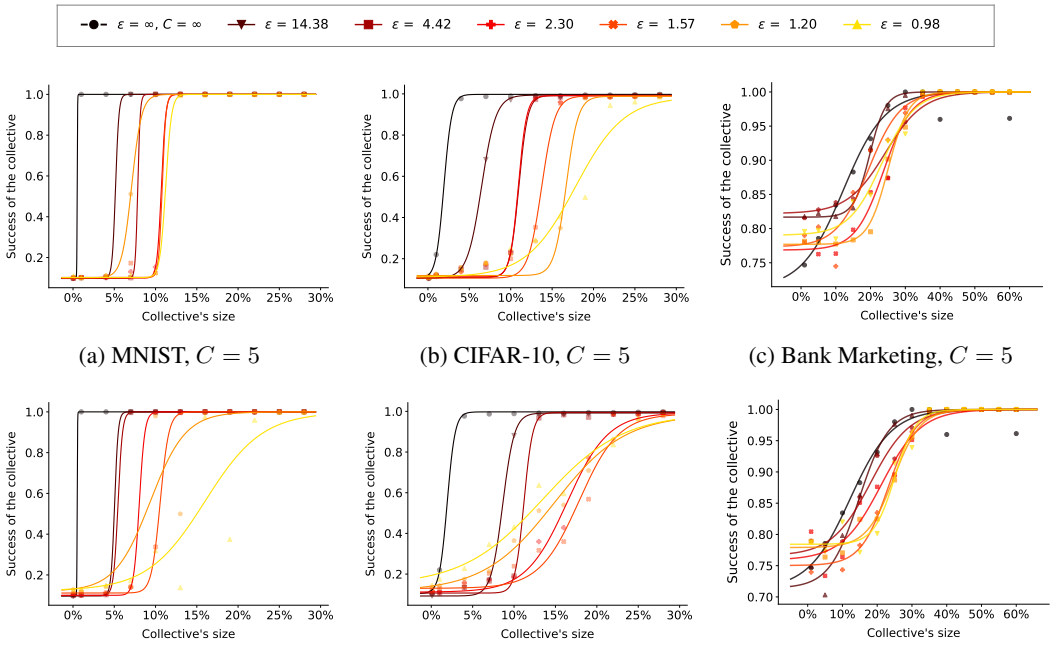
We utilize the PyTorch library for model implementation and training. For differentially private training, we use the Opacus framework,[2] built on top of PyTorch, with its default configurations, which uses Rényi Differential Privacy (RDP) [Mironov, 2017] accounting for DP-SGD. All the models are trained for 30 epochs, and we report baseline accuracies in Table 3. See Appendix D for examples of the signals (which are designed to be difficult to detect by humans) inserted onto samples from the image classification datasets.

[2] `https://github.com/pytorch/opacus`

C Additional Results

C.1 Collective's success across clipping thresholds

Figure 3, which evaluates collective success with additional clipping thresholds $C = 5$ and $C = 10$, also shows a clear trend of higher critical mass with increased privacy, similar to Figure 1 for $C = 1$.



(a) MNIST, $C = 5$ (b) CIFAR-10, $C = 5$ (c) Bank Marketing, $C = 5$

(d) MNIST, $C = 10$ (e) CIFAR-10, $C = 10$ (f) Bank Marketing, $C = 10$

Figure 3: The success of the collective across ε on MNIST, CIFAR-10, and Bank Marketing with $C = 5$ and $C = 10$. Each column shows a different dataset, and each row represents a different clipping threshold. Each plot on a row corresponds to a different clipping threshold. For each plot, we evaluate the collective's success under different values of privacy budget ε and compare it with the baseline case ($\varepsilon = \infty, C = \infty$), which corresponds to SGD without any privacy constraints. Collective size $\alpha \in [0, 1]$ is reported as a percentage of the overall training dataset.

Collective's size (α)	No privacy		$\varepsilon = 15$		$\varepsilon = 1$	
	TPR @ 0.1% FPR	AUC	TPR @ 0.1% FPR	AUC	TPR @ 0.1% FPR	AUC
0%	4.5736%	82.1%	0.3212%	65.6%	0.0803%	51.0%
1%	3.6420%	82.7%	0.3574%	65.5%	0.0924%	51.4%
7%	4.4210%	81.6%	0.2289%	65.1%	0.0883%	51.6%
14%	5.2642%	81.1%	0.2409%	64.4%	0.1084%	50.7%
21%	3.4774%	80.6%	0.2208%	64.0%	0.0843%	50.9%
28%	3.7183%	78.6%	0.1767%	61.2%	0.1847%	51.2%

Table 1: Evaluation of LiRA under varying privacy constraints using AUC and TPR at 0.1% FPR on CIFAR-10 dataset. Lower TPR 0.1% FPR indicates better robustness to MIA, while with AuC the desired metric is as close to 50% (random chance) as possible (i.e. 50.02% is better than 49.89%, which are both better than 81.78%).

C.2 MIA Evaluations

Experimental details For LiRA, we train 16 shadow models on CIFAR-10 using the same architecture described in Section 4.2. Each shadow model is trained on half of the dataset (25,000 samples), with the remaining half used as non-members. We consider six collective sizes, $\alpha \in \{0\%, 1\%, 7\%, 14\%, 21\%, 28\%\}$, and three learning algorithm configurations: $(\varepsilon = 1, C = 1)$,

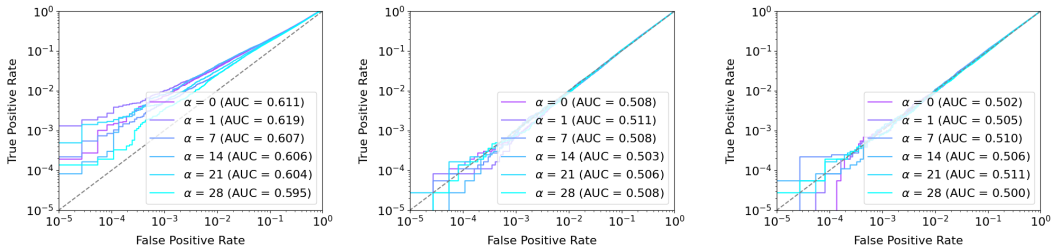


Figure 4: Success rate of LiRA [Carlini et al., 2022] evaluation on SVHN dataset. Each figure corresponds to a different setting of privacy constraints with privacy increasing from left to right.

Collective's size α	No privacy		$\varepsilon = 15$		$\varepsilon = 1$	
	TPR @ 0.1% FPR	AUC	TPR @ 0.1% FPR	AUC	TPR @ 0.1% FPR	AUC
0%	0.73%	61.1%	0.068%	50.8%	0.114%	50.2%
1%	1.00%	61.9%	0.076%	51.1%	0.114%	50.5%
7%	0.50%	60.7%	0.087%	50.8%	0.084%	51.0%
14%	0.79%	60.6%	0.092%	50.3%	0.111%	50.6%
21%	0.52%	60.4%	0.100%	50.6%	0.122%	51.1%
28%	0.33%	59.5%	0.076%	50.8%	0.079%	50.0%

Table 2: Evaluation of LiRA under varying privacy constraints using AUC and TPR at 0.1% FPR on SVHN dataset. Lower TPR 0.1% FPR indicates better robustness to MIA, while with AuC the desired metric is as close to 50% (random chance) as possible.

$(\varepsilon = 15, C = 1)$, and $(\varepsilon = \infty, C = \infty$ for no privacy$)$. This results in 288 training runs in total. Table 1 reports the Area under the ROC Curve (AUC) scores and the True Positive Rate (TPR) at a 0.1% False Positive Rate (FPR) for our experiments. Table 1 reports the corresponding AUC scores and the TPR @ 0.1% FPR values. Figure 4 and Table 2 extend our results from Section 4.4 to include MIA results on the SVHN dataset.

Why does ACA lead to marginally improved empirical privacy? We speculate that this due to how ACA indirectly affects model confidences. LiRA relies on likelihood ratios computed using the model's predictive distribution. The collective inserts a signal over a data subspace, meaning that the label function is no longer smooth. Whereas model trained with supervised learning typically saturate their predictive confidences [Guo et al., 2017, Papyan et al., 2020], a model trained on this new labeling function may hedge away from high-confidence predictions, making it more difficult to determine training data membership based solely on model confidences.

C.3 Predictive Performance

Table 3 shows baseline predictive accuracies for DP-trained classifiers on CIFAR-10.

Privacy loss ε	Noise multiplier σ	$C = 1$	$C = 5$	$C = 10$
14.38	0.5	72%	61%	55%
4.42	0.7	66%	52%	49%
2.30	0.9	59%	49%	46%
1.57	1.1	56%	48%	42%
1.20	1.3	52%	45%	37%
0.98	1.5	46%	43%	35%

Table 3: Test set accuracies under different privacy configurations for models trained on CIFAR-10. The non-private baseline ($\varepsilon = \infty,\ \sigma = 0,\ C = \infty$) achieves 85% accuracy.

D Data Visualization

Figure 5 shows examples of MNIST data points with and without the signal inserted by the collective. Figure 6 visualizes the same thing, but for data from the CIFAR-10 dataset.

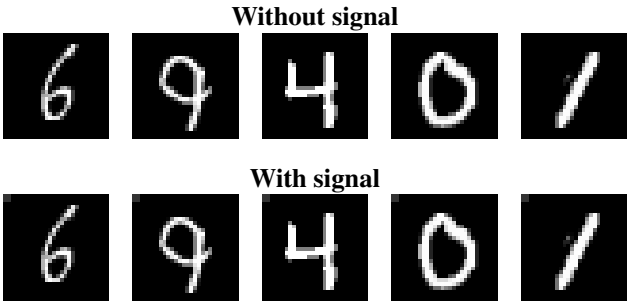


Figure 5: MNIST samples with and without adding application of transformation g

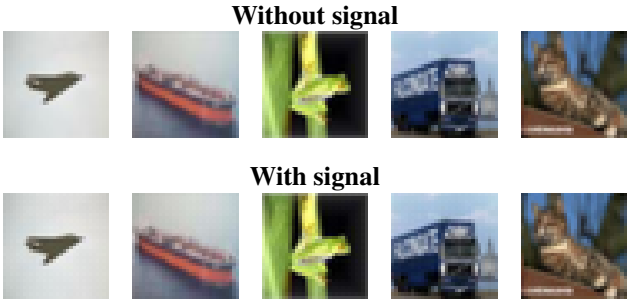


Figure 6: CIFAR-10 samples with and without adding application of transformation g