

Crowding Out The Noise: Algorithmic Collective Action Under Differential Privacy

Rushabh Solanki[‡], Meghana Bhangé[§], Ulrich Aivodji[§], and Elliot Creager[‡]

[‡]University of Waterloo, Vector Institute

[§]ÉTS Montréal, Mila

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 the responsible deployment of trustworthy AI systems is a worthy goal, there are many possible ways to realize it, from policy and regulation to improved algorithm design and evaluation. In fact, since AI trains on social data, there is even a possibility for everyday users, citizens, or workers to directly steer its behavior through *Algorithmic Collective Action*, by deliberately modifying the data they share with a platform to drive its learning process in their favor. This paper considers how these grassroots efforts to influence AI interact with methods already used by AI firms and governments to improve model trustworthiness. In particular, we focus on the setting where the AI firm deploys a differentially private model, motivated by the growing regulatory focus on privacy and data protection. We investigate how the use of Differentially Private Stochastic Gradient Descent (DPSGD) 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 algorithmic collective action under differential privacy as a function of the collective’s size and the firm’s privacy parameters, and verify these trends experimentally by simulating collective action during the training of 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, which is sourced from public websites, users of the firm’s products, and crowd workers. By leveraging these large-scale data, the firms are able to train increasingly sophisticated model that not only improves their predictive capabilities but also expand the range of problems that they can address. Despite its advantages, the extensive use of personal data in training machine learning models 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, various solutions have been proposed and implemented at different stages of the model development pipeline. At the firm level, efforts towards building "trustworthy AI" often involve fairness assessments, bias mitigation techniques, privacy auditing, and adversarial evaluations like red teaming across multiple stages from data collection to model training and post-processing [Barocas et al., 2023]. However, implementing these techniques may introduce trade-offs with the firm’s broader objective of maximizing predictive performance and enhancing user engagement for more data. On the other hand, several regional regulations such as the European Union’s General Data Protection Regulation (GDPR) [European Parliament and Council of the European Union, 2016], Canada’s Personal Information Protection and Electronic Document Act (PIPEDA) [Government of Canada, 2000] and The California Privacy Rights Act (CPRA) [State of California, 2020] establish baseline privacy protections, yet compliance with these laws alone does not guarantee socially responsible outcomes [Selbst et al., 2019, Utz et al., 2019]. In parallel with organization and regulatory measures, grassroots efforts of *Algorithmic Collective Action* is

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

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].

Algorithmic collective action (ACA) [Hardt et al., 2023, OLSON, 1971] provides a principled framework for understanding how a group of individuals, through coordinated changes in their data, can impact the behavior of deployed models. Prior work has provided theoretical insights under assumptions such as Bayes optimality, empirical risk minimization [Hardt et al., 2023], or robust optimization [Ben-Dov et al., 2024], offering an informed view of how these assumptions can affect the effectiveness of collective action on model behavior. 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 this intersection, 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 strengthening of regulatory frameworks and growing consumer demand for privacy guarantees, we seek to understand how differential privacy affects algorithm’s responsiveness to collective action and success rate of such interventions.

We put forward a theory that examines the impact of differential privacy constraints on the effectiveness of collective taking action on the firm’s learning algorithm. We operationalize this framework in practical deep learning scenarios, and perform extensive experiments on multiple benchmark datasets showing that while differential privacy provides strong guarantees to protect individual data, it inadvertently reduces the collective’s ability to coordinate and alter the behavior of the firm’s model. This work offers a new lens on the societal implications of using privacy-preserving techniques in machine learning, through the combination of theoretical insight and empirical validation.

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 differential privacy 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 differential privacy reduces the collective’s ability to influence the behavior of the model.
- We also measure empirical privacy through the lens of membership inference attacks, and find that the collective’s presence in the data distribution offers some degree of empirical privacy.

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

Hardt et al. [2023] proposed a theoretical framework to model the dynamics between a firm’s learning algorithm and a collective. Within this framework, the size of the collective is represented by a parameter $\alpha > 0$, which denotes the proportion of individuals within the data drawn from the base distribution \mathcal{P}_0 . The collective selects a strategy $h : \mathcal{Z} \rightarrow \mathcal{Z}$, representing allowable modifications to data, where $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ denotes the feature-label domain. Applying h to data drawn from \mathcal{P}_0 induces the collective’s distribution \mathcal{P}^* . The firm’s learning algorithm \mathcal{A} then would encounter the following mixture of data distribution:

$$\mathcal{P} = \alpha\mathcal{P}^* + (1 - \alpha)\mathcal{P}_0,$$

As a result, the firm will deploy the classifier $f = \mathcal{A}(\mathcal{P}) : \mathcal{X} \rightarrow \mathcal{Y}$, mapping from features to labels.

Planting a signal In this work, we focus on the collective’s goal of influencing the firm’s learning behavior by modifying both the features and labels for all data under their control—the *feature-label* strategy from [Hardt et al. \[2023\]](#). The data is modified in such a way that the classifier f learns to associate the transformed version of features with the chosen target label y^* , where the transformation is defined by the function $g : \mathcal{X} \rightarrow \mathcal{X}$, resulting in the strategy $h(x, y) = (g(x), y^*)$.

Definition of success 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 $\{f_\theta\}_{\theta \in \Theta}$. By defining the target model θ^* that the collective desires to achieve by influencing the firm’s model θ , we measure the success of the collective after t steps as:

$$S_t(\alpha) = -\|\theta_t - \theta^*\|.$$

Critical mass 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*. Formally, for a given target success level S^* , the critical mass is the smallest value α such that the achieved success $S(\alpha) \geq S^*$.

Consistent with the prior setup of [\[Hardt et al., 2023\]](#), we do not impose any convexity assumptions on the objective function and consider a learner that observes the distribution \mathcal{P}_t at each time step. Let $g_{\mathcal{P}_t}(\theta_t) = \mathbb{E}_{z \sim \mathcal{P}_t} \nabla \ell(\theta_t; z)$ be the expected gradient of the loss over the distribution \mathcal{P}_t , measured at the parameter $\theta_t \in \Theta$. The learner then performs the following gradient descent update:

$$\theta_{t+1} = \theta_t - \eta g_{\mathcal{P}_t}(\theta_t).$$

With this, the collective aims to steer the firm’s model θ toward a target θ^* by influencing the overall gradient to align it with their desired direction.

Definition 1 (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^* . The following theorem formalizes the lower bound on the success of the collective when applying the gradient-redirecting strategy.

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,*

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

where $C(\alpha)$ is directly proportional to collective’s size α . As α increases (and consequently $C(\alpha)$), the lower bound on the collective’s success also increases. This result also implies that the collective can attain any desired model θ^* , provided a continuous path exists from θ_0 to θ^* that does not encounter large gradients with respect to the initial distribution \mathcal{P}_0 .

2.2 Privacy-preserving Training

In machine learning applications that involve sensitive data, it is essential to ensure the privacy of individual records, especially if that model is to be deployed publicly. A common approach to formalize privacy guarantees is through differential privacy (DP), which provides a mathematical framework for limiting the information that a learned model can reveal about any single data point. DP is based on the concept of *neighboring* datasets, which are defined as two datasets that differ in the data of a single record. An algorithm (or “mechanism”) is said to be differentially private if it admits nearly the same statistical inference for two neighboring datasets. The formal definition of DP from [Dwork et al. \[2006\]](#) is presented as follows.

Definition 2 ((ϵ, δ) -Differential Privacy). *A randomized mechanism $M : D \rightarrow R$ 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*

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

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. DP is often applied to learning algorithms by considering how the parameter identified is affected by the addition of carefully calibrated noise throughout learning: the mechanism $M(d)$ is a learning algorithm run on some dataset d , while the outcome S is a particular parameter value θ found by the learning algorithm. To characterize the success of the collective in a more practical setting, we use Differentially Private Stochastic Gradient Descent (DPSGD), proposed by [Abadi et al. \[2016\]](#), which guarantees (ϵ, δ) -differential privacy, as detailed in Algorithm 1.

Algorithm 1 DPSGD Algorithm from [Abadi et al. \[2016\]](#)

Input: Dataset \mathcal{D} , loss function ℓ , learning rate η , batch size \mathcal{B} , noise scale σ , clipping threshold C , initial model θ_0

for $t \in [T]$ **do**

Uniformly draw mini-batch \mathcal{B}_t from \mathcal{D}

For each $z_i \in \mathcal{B}_t$, $g_i^{\text{clip}}(\theta_t) = \text{clip}(\nabla \ell(\theta; z_i), C)$

$g^{\text{DP}}(\theta_t) = \frac{1}{|\mathcal{B}_t|} \left(\left(\sum_i g_i^{\text{clip}}(\theta_t) \right) + \mathcal{N}(0, \sigma^2 C^2 I) \right)$

$\theta_{t+1} = \theta_t - \eta g^{\text{DP}}(\theta_t)$

Return: θ_T and the overall privacy cost (ϵ, δ)

In a nutshell, DPSGD modifies standard stochastic gradient descent by adding noise to the gradient updates, using the Gaussian mechanism [[Dwork et al., 2014, Appendix A](#)]. The noise multiplier σ , which controls the scale of Gaussian noise added to the clipped gradients, is inversely proportional¹ to the privacy loss ϵ . A higher σ introduces more noise, offering stronger privacy guarantees (meaning smaller ϵ), but this incurs the cost of the reduced utility of the model.

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 We assume that the firm deploys a private learning algorithm \mathcal{A} with the objective of preserving user data privacy. Given a data distribution \mathcal{P} and a parameter space Θ , $f = \mathcal{A}(\mathcal{P}) \in \Theta$ represents the model chosen by the firm. We consider a realistic learning scenario without convexity assumptions on the

¹For a simple application of the Gaussian mechanism, this inverse relationship has a closed-form expression [[Dwork et al., 2014](#)]. However, in DPSGD which involves repeated application of the mechanism across training iterations, the cumulative privacy loss is tracked using *privacy accountant* [[Abadi et al., 2016](#), [Mironov, 2017](#)]

objective function, where gradient-based learning algorithms are typically used. In particular, we focus on Differentially Private Stochastic Gradient Descent (DPSGD), a widely used algorithm for training models under (ϵ, δ) -differential privacy constraints, and examine how this choice affects the success of collective action.

At each time step t , we assume the learner observes the current data distribution \mathcal{P}_t , allowing the collective to adaptively interact with the learner by choosing \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. Given a clipping threshold C and a noise scale σ , the model parameters are updated by taking the gradient step computed according to the DPSGD:

$$\begin{aligned}\theta_{t+1} &= \theta_t - \eta \left(\mathbb{E}_{z \sim \mathcal{P}_t} [\text{clip}(\nabla \ell(\theta_t; z), C)] + \mathcal{N}(0, \sigma^2 C^2 I) \right) \\ &= \theta_t - \eta \left(g_{\mathcal{P}_t}^{\text{clip}}(\theta_t) + \mathcal{N}(0, \sigma^2 C^2 I) \right) \\ &= \theta_t - \eta g_{\mathcal{P}_t}^{\text{DP}}(\theta_t)\end{aligned}$$

where $\text{clip}(g, C) = g \cdot \min(1, C/\|g\|)$ denotes the gradient clipping operation, which scales the gradient g to have norm of at most C . To maintain consistency with the analysis from Hardt et al. [2023], we use the expectation of (clipped) gradients when performing each gradient descent update.

Theoretical results The most intuitive factor that limits the success of the collective when the firm uses DPSGD 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 1 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 , 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,*

$$S_T(\alpha, \sigma, C) \geq -(1 - \eta B(\alpha, C))^T \|\theta_0 - \theta^*\| - \sigma C \cdot f_1(B(\alpha, C), T, \eta) \cdot f_2(d, \delta),$$

where $B(\alpha, C)$ here is directly proportional to the collective’s size α and clipping threshold C , as it depends on the norm of the clipped gradients. Therefore, as the clipping threshold increases, so does the norm of the clipped gradient. The function f_1 is the convergence-dependent scaling factor, while f_2 quantifies how much noise we might expect in high dimensions with high confidence (see Appendix A.2 for a full expression of these two functions). Setting $C = \infty$, which corresponds to no clipping being applied to the gradient, recovers $C(\alpha)$ –stated in Theorem 1–from $B(\alpha, C)$. In addition, setting the noise scale $\sigma = 0$, reducing the learner to standard SGD algorithm, eliminates the second term entirely and reconstructs the bound from Theorem 1. The formal proof for this theorem can be found in Appendix A.2.

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 role of clipping threshold C , set by the firm, on the success of the collective. In contrast to σ , this relationship involves more nuanced dynamics. Since $B(\alpha, C)$ depends directly on the clipped gradient, its upper bound is an increasing function of C . This, in turn, causes the expression $-(1 - \eta B(\alpha, C))^T$ in the first term of the sound bound to increase with clipping threshold C , leading to a positive contribution to the collective’s success. However, second term introduces two opposing effects–while

the linear dependence on C tends to reduce success as C increases, the function f_1 , which itself decreases with C , counteracts the negative trend. As a result, the overall impact of the clipping threshold 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 standard image-based multi-class classification benchmarks, including MNIST [LeCun and Cortes, 2010] and CIFAR-10 [Krizhevsky and Hinton, 2009], as well as a tabular binary classification task using the Bank Marketing dataset [Moro et al., 2014]. We further evaluate how both the presence and the size of the collective influence vulnerability to membership inference attacks under both private and non-private settings.

4.1 Strategy and Success of the Collective

As discussed in Section 2.1, we assume the collective comprises a proportion $\alpha \in [0, 1]$ of the training data and aims to influence the algorithm’s behavior by planting a signal $g(x)$ within the data they control. They desire to steer the model’s prediction on transformed data points $g(x)$ towards a desired target label y^* . Specifically, we assess the effectiveness of the *feature-label* strategy, where the collective modifies the input data—such as pixel values in images or entries in tabular data—and assigns a chosen target label y^* to these modified examples. Given a transformation $g : \mathcal{X} \rightarrow \mathcal{X}$, the collective aims to maximize the following measure of success:

$$S(\alpha) = \Pr_{x \sim \mathcal{P}_0} \{f(g(x)) = y^*\}.$$

That is, collective’s success is defined in terms of how the model’s predictions agree with the collective’s chosen target label for evaluation data where the signal has been planted.

A straightforward way to measure $S(\alpha)$ in the experiments would be to plant a signal in all test points and count how often the model successfully predicts the desired output. This corresponds to the accuracy on the modified test data. Our objective is to determine the *critical mass*, denoted as α^* , the smallest size of the collective required to achieve a fixed target success rate, by evaluating model accuracy on test data with the planted signal. We do this by training multiple models on datasets where the collective controls different amounts of data.

4.2 Experimental Setup

The datasets used, corresponding model architectures, and the specific data transformations applied by collective to each dataset are detailed as follows.

MNIST We begin by performing experiments using the MNIST dataset, which contains grayscale images of handwritten digits, each sized 28×28 pixels. We balance the training dataset by randomly sampling 5,000 data points per class, resulting in a total of 50,000 samples across 10 classes. The test set is left unchanged with 10,000 samples.

We use the standard ResNet18 [He et al., 2016] architecture but replace batch normalization with group normalization [Wu and He, 2018] to ensure consistency when applying DPSGD, following prior work [Kurakin et al., 2022, Luo et al., 2021]. This modification allows for accurate per-sample gradient computation, which is required for enforcing differential privacy during training. For all experiments conducted in this work, we use the SGD optimizer in the non-privacy setting and DPSGD for the privacy-preserving setting.

The transformation g applied on the input space modifies each image by setting the pixel values within the 2×2 patch on the top-left corner to a fixed value of 50. To carry out the full *feature-label* strategy, the label of each transformed image is reassigned to a class “8”, corresponding to the digit 8. We evaluate the collective’s success by applying this transformation to all test samples and measuring the classification accuracy, treating the target label 8 as the ground truth.

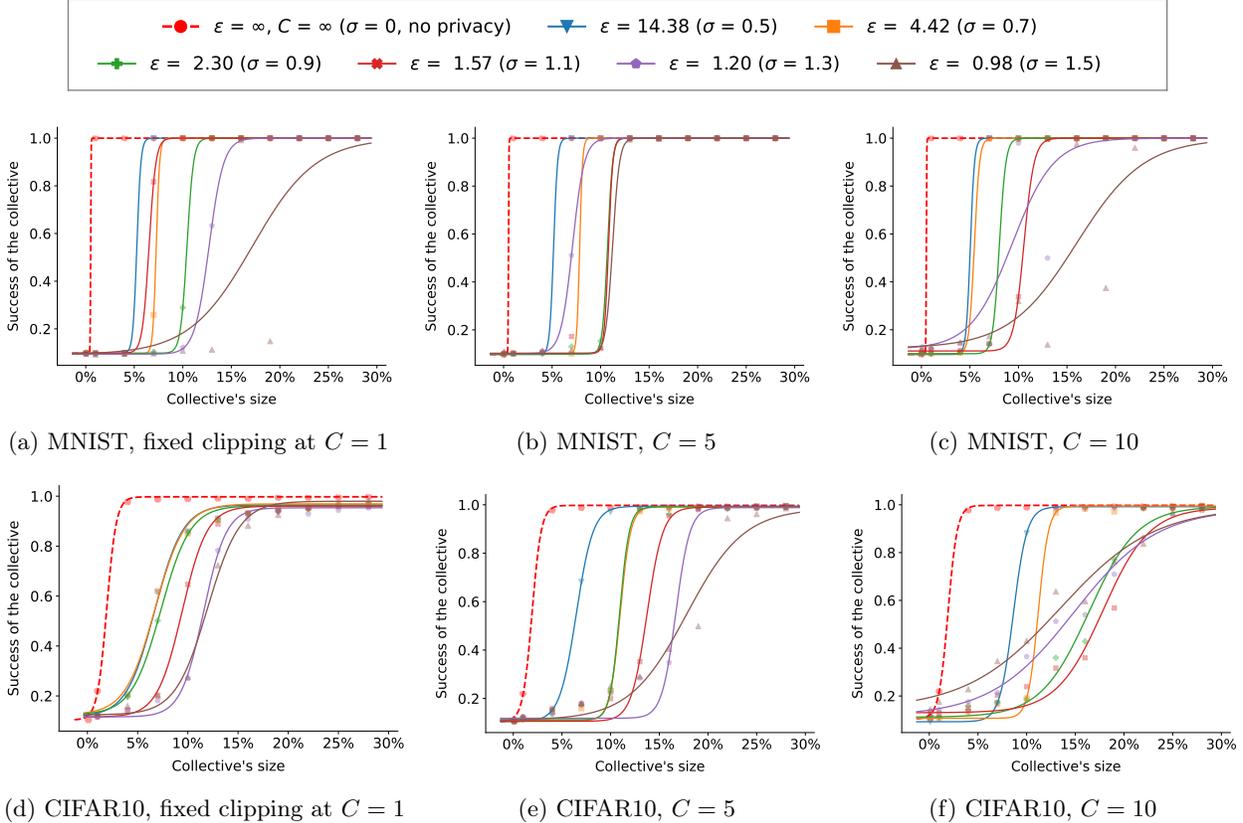


Figure 1: The success of the collective across ϵ . The *top* row shows results on the MNIST dataset, while the *bottom* one on CIFAR-10. Each column corresponds to a different clipping threshold. For each plot, we evaluate the collective’s success under different values of privacy budget ϵ and compare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

CIFAR-10 We also conduct experiments on the CIFAR-10 dataset, which consists of 60,000 color images divided into 10 mutually exclusive classes, with 6,000 images per class. The dataset is partitioned into 50,000 training images and 10,000 test images. Each image is sized at 32×32 pixels.

We apply the same architecture as used for MNIST experiments, that is, ResNet18 with batch normalization layers replaced by group normalization layers. In addition, we pre-train the ResNet-18 on the CIFAR-100 dataset, an extension of CIFAR-10 that consists of 60,000 32×32 color images categorized into 100 fine-grained classes, with 600 images per class. We use this pre-trained model as the initialization for each CIFAR-10 training experiment in order to achieve improved accuracy under differential privacy constraints.

The transformation g here is a structured alteration to the image by modifying pixel intensities on a regular grid, where every second pixel along every second row is adjusted by the magnitude of 2 [Ben-Dov et al., 2024]. To ensure that the pixel values remain within the valid range $[0, 255]$, any pixel that would overflow when increased by 2 is instead decreased by 2. Each altered image is relabeled with a target class “8”, corresponding to the label “ship”.

Bank Marketing Lastly, we perform experiments on the UCI Bank Marketing dataset [Moro et al., 2014], which contains 45,211 samples with 17 features describing client demographics and details of past marketing campaigns. While the other datasets used in our experiments involve multiclass classification, the objective of this dataset is to perform binary classification to predict whether a client will subscribe to term deposit.

For this dataset, we use a simple feedforward neural network with a single hidden layer of 128 units, followed by a ReLU activation function [Agarap, 2019], and a final fully connected layer that maps the representation to the number of classes. For the transformation g , we restrict the collective’s ability to update only a specific feature of data they control, and add a fixed offset of 50 to its value. We then reassign the label of the collective’s data to the target class “0”.

Implementation details 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 DPSGD. All the models are trained for 30 epochs, and we report baseline accuracies in Table 2. See Appendix C for examples of the signals (which are designed to be difficult to detect by humans) inserted onto samples from the image classification datasets.

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. Our aim is to find the smallest collective’s size that reaches close to 100% accuracy on the altered test set. This evaluation is performed for multiple values of clipping threshold C ; for each value of C , we vary the privacy loss ϵ that the firm intends to tolerate.

Figure 1 (for MNIST and CIFAR-10 datasets) and Figure 2 (for Bank Marketing dataset) show a clear trend: as the privacy loss decreases (corresponding to higher privacy), the critical mass required for the 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. This trend is consistent across different values of the clipping threshold C , as shown for $C = 1, 5$, and 10 in each column of Figure 1. 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 differential privacy and algorithmic collective action. 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

DP is appealing because it provides theoretical guarantees over how much privacy leakage is admitted by a given learning algorithm. These guarantees hold for arbitrarily strong adversaries attempting to predict an individual’s membership in the training data, even for adversaries who have access to side information [Dwork et al., 2014]. Of course, not all machine learning models are trained with DP (especially considering that the addition of noise can adversely affect model utility). To assess the privacy risks of a broader class of learning algorithms, especially those trained without DP where theoretical guarantees cannot be established, estimates of the *empirical* privacy leakage can be used [Hu et al., 2022].

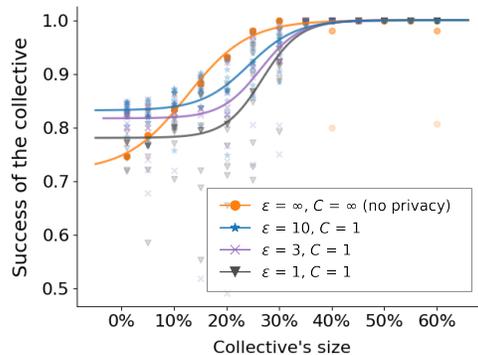


Figure 2: Success of the collective across ϵ on Bank Marketing dataset [Moro et al., 2014]. We evaluate collective’s success under different values of privacy loss ϵ and compare it with baseline case ($\epsilon = \infty, C = \infty$), which corresponds to SGD without any privacy constraints.

²<https://github.com/pytorch/opacus>

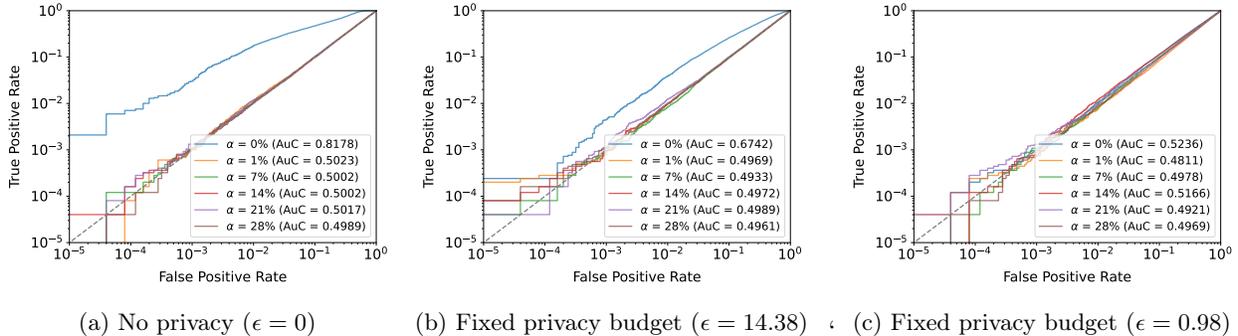


Figure 3: Success rate of Likelihood Ratio Attack (LiRA) [Carlini et al., 2022] evaluation on CIFAR-10 dataset. Each figure corresponds to a different setting of privacy constraints with privacy increasing from left to right.

Collective's size (α)	$\epsilon = \infty$		$\epsilon = 14.38$		$\epsilon = 0.98$	
	TPR @ 0.1% FPR	AuC	TPR @ 0.1% FPR	AuC	TPR @ 0.1% FPR	AuC
0%	3.02%	81.78%	0.45%	67.42%	0.10%	52.36%
1%	0.11%	50.23%	0.08%	49.69%	0.09%	48.11%
7%	0.11%	50.02%	0.09%	49.33%	0.11%	49.78%
14%	0.10%	50.02%	0.10%	49.72%	0.08%	51.66%
21%	0.13%	50.17%	0.11%	49.89%	0.13%	49.21%
28%	0.10%	49.89%	0.11%	49.61%	0.07%	49.69%

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%).

In keeping with this investigation of empirical privacy risks, we now investigate how the presence of a collective can impact the vulnerability of the model to Membership Inference Attack (MIA) [Shokri et al., 2017]. The goal of the MIAs is to determine whether a specific data point was included in the training set. The collective controlling some part of the training data can alter the model's decision boundaries and learning pattern, which changes the likelihood of successful attacks. Our analysis aims to explore the correlation between the size of the collective in the training data and the model's susceptibility to MIA given a learning algorithm used by the firm.

Experimental Setup We use Likelihood Ratio Attack (LiRA) [Carlini et al., 2022] to evaluate the membership of the data point in the training set. LiRA determines the membership by comparing the model output probabilities of the target model to that of several *shadow models*, trained on different subsets of the data. We train 16 shadow models on the CIFAR-10 dataset with the same architectural configuration mentioned in Section 4.2. Each shadow model is trained on half of the dataset (25,000 data points), leaving another half for evaluation as non-members. We evaluate the success rates of the attack under different settings of the collective's size and learning algorithm on the CIFAR-10 dataset. We select six different sizes of the collective $\alpha \in \{0\%, 1\%, 7\%, 14\%, 21\%, 28\%\}$ and three configurations of the learning algorithm—($\epsilon = 0.98, C = 1$), ($\epsilon = 14.38, C = 1$), and ($\epsilon = \infty, C = \infty$, implying no privacy), totaling to 288 training runs. To effectively determine the success of the attack, we report true-positive rates (TPR) at a very low false-positive rate of 0.1% in addition to the area under the curve (AuC) of the receiver operating characteristics (ROC).

Results Figure 3 presents the ROC curves showing the LiRA's success rates under various configurations, while Table 1 reports the corresponding AuC scores and the TPR @ 0.1% FPR values. We find that the

collective action during training *improves* empirical privacy by increasing the robustness to MIA (pushing MIA success close to 50%, or random chance), *even for models trained without DP*. Specifically, we observe that the presence of a collective comprising of as little as 1% of the dataset, there is noticeable improvement in robustness against MIA for the models trained without any privacy constraints (Figure 3a) or with low privacy when $\epsilon = 14.38$ (Figure 3b) and that with high privacy when $\epsilon = 0.98$ (Figure 3c). Moreover, the introduction of collective does not appear to compromise the robustness to MIA already provided by DP training across different privacy levels.

Why does ACA lead to 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, Pappan 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.

5 Related Works

5.1 Collective Action

Collective action problems were central to 20th century social scientific inquiry [OLSON, 1971, Hardin, 1982, Marwell and Oliver, 1993], where various disciplinary perspectives were adopted to characterize the circumstances under which a small yet organized group (e.g. a political action committee, labor union, or voting block) could have an out-sized effect on social outcomes. More recently collective action problems have been revisited in the context of socio-technical systems involving algorithms, especially those that rely on data-driven prediction and decision-making.

As algorithmic systems become part of high-stakes decision-making, they can also cause socio-technical harms that reinforce existing social inequalities, marginalize vulnerable groups, or create an inequitable environment [Shelby et al., 2022]. In response, users of these systems may aim to collectively influence algorithmic outcomes to be more equitable when traditional channels of accountability are not available. Hardt et al. formalized the concept of Algorithmic Collective Action (ACA) as a setting in which users of a system can steer the output of the machine learning algorithm to achieve 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). One such challenge is free-riding, where individuals benefit from a group’s effort without participating, as mentioned in Olson’s theory of collective action [OLSON, 1971]. Sigg et al. refines this view in their model of the #DeclineNow campaign, where DoorDash workers coordinated to reject low-paying jobs [Sigg et al., 2024]. Their results show that collective strategies remain rational for individuals under labor undersupply, but in oversupplied markets, free-riding becomes attractive, undermining participation.

Even when participation and incentives align, another challenge to the success of collective action could depend on the type of learning algorithm. Ben-Dov et al. show that the success of ACA is highly dependent on the properties of the learning algorithms, and recommend studying the learning algorithm when taking into account the success of collective action [Ben-Dov et al., 2024].

5.2 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. While data poisoning involves malicious manipulation, ACA is not inherently adversarial and often pursues constructive objectives. Moreover, ACA emphasizes coordination among the collective, often to align with societal or personal objectives. We refer to the comprehensive surveys by Tian et al. [2022] and Guo et al. [2022] which provide a detailed overview of data poisoning and backdoor attack techniques, respectively, along with corresponding defense mechanisms. Foundational work by Gu et al. [2019] demonstrated that models can

effectively mislead the classifier in realistic scenarios using special signals. Shejwalkar et al. [2023] showed that semi-supervised learning models, which aim to preserve privacy by relying on unlabeled data, are still vulnerable to such attacks. Ma et al. [2019] studies the robustness of differentially private learners against data poisoning and shows that attackers can poison models effectively if they have access to sufficient portion of the training data.

5.3 Private Machine Learning

Differential Privacy (DP) has emerged as a gold standard technique providing a formal privacy guarantee in machine learning and data analysis [Cummings et al., 2024]. A range of techniques have been developed to achieve DP, particularly for convex learning problems, including output perturbation [Chaudhuri et al., 2011], objective perturbation [Chaudhuri et al., 2011, Kifer et al., 2012], and gradient perturbation [Bassily et al., 2014]. In non-convex learning problems, especially in deep learning, DPSGD has become the prevailing method [Abadi et al., 2016], due to its conceptual simplicity. By design, a differentially private mechanism with privacy budget ϵ implicitly offers group privacy with a privacy $k\epsilon$ for any group of size k [Dwork et al., 2014, Thm 2.2]. However, for real-world scenarios with possibly large group size, 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. Privacy accounting techniques for tracking privacy parameters (ϵ_i, δ_i) for individual data points $\{x_i\}$ has also been proposed [Yu et al., 2022], suggesting that the global bounds can be improved upon, although these improvements may be differentially distributed across constituent groups within the training data.

5.4 Trade-offs in Trustworthy ML

Trustworthy machine learning focuses on the integration of trustworthiness principles, such as security, privacy, fairness, robustness, and explainability, into the development of machine learning models. While all these values are widely recognized as important, several recent studies have shown that they can be in tension with each other. For instance, increasing fairness can decrease utility [Menon and Williamson, 2018, Yaghini et al., 2023], adversarial robustness can make ML systems more vulnerable to privacy inference attacks [Song et al., 2019], explanation algorithms can be arbitrarily manipulated to give fake evidence of fairness [Aïvodji et al., 2019], or leveraged to perform powerful model stealing attacks [Milli et al., 2019]. This work contributes to raising awareness of the trade-off between privacy and ACA’s effectiveness.

6 Discussion

6.1 Tensions Between Privacy and Other Trustworthy ML Goals

To accommodate the regulatory needs for the protection of individual privacy, such as being compliant with GDPR [European Parliament and Council of the European Union, 2016], companies can implement differentially private algorithms. DPSGD builds on the SGD algorithm by introducing a clipping threshold and sensitivity of the gradients computed from individual data points to a certain norm and adding noise to the aggregated gradient before updating the global model. Although algorithms trained using DP protect user privacy, this could also affect the success of ACA. This work exists alongside the larger literature on works that study tensions between privacy and other trustworthy AI interventions [Ferry et al., 2023]. Examples include tensions between differential privacy and fairness [Bagdasaryan et al., 2019, Fioretto et al., 2022] as well as tension between differential privacy and explainability. Bagdasaryan et al. [2019] that models trained using DPSGD can disproportionately reduce the accuracy for underrepresented groups such as having lower accuracy for black faces than for white faces when compared to the non-DP model. This disparate impact may also lead to vulnerability towards MIAs, as empirically shown in Kulynych et al. [2019]. Fioretto et al. [2022] explores situations in which privacy and fairness may have goals that are similar or different, and studies the reasons behind how DP may exacerbate bias and unfairness.

This work adds to the broader research discussion on trade-offs in trustworthy machine learning, where we explore the tensions between privacy and algorithmic influence.

6.2 Privacy as Anti Cooperation Strategy

Data privacy is used to protect users from possible harm involving misuse of personal information. This could also motivate users to participate in collective action, as they are less likely to be identified. However, some institutions may also use privacy interventions as a pretext to limit accountability and transparency [Van Loo, 2022]. Privacy and data protection laws can also be exploited to strengthen the surveillance infrastructure [Yew et al., 2024]. Additionally, privacy laws can be used to withhold workplace data from worker representatives during collective bargaining citing legal data privacy responsibilities [Gould, 2024]. This work characterizes the impact of platform privacy interventions on the success of ACA. Although such interventions can be justified, platforms can also adopt privacy as a pretext [Van Loo, 2022] to defend against ACA. Although we have found no specific evidence that privacy is being used as a shield against ACA, there are real-world cases in which Big Tech firms use privacy to justify exclusionary conduct [Chen, 2022, Tůmová, 2024].

6.3 Cooperation as an Implicit Privacy Strategy

As observed in Section 4.4, algorithmic collective action can inadvertently provide empirical privacy against membership inference attacks. The *feature-label* strategy employed by collective can be said to mirror the randomized-response technique in collecting survey samples, where each respondent flips their true answer with some probability so that the collector only sees a noisy signal, yet aggregate statistics remain accurate [Blair et al., 2015]. Similarly, when collective “flips” its label in the coordinated manner, it injects randomness into training labels. Although, in the current over-parameterized regime of deep learning, the model is able to correlate the transformation g applied by the collective with the target label y^* . While the collective action with this strategy may offer empirical resistance to MIA, it does not provide any formal privacy guarantees in the sense of differential privacy.

6.4 Role of Pre-training and Fine-tuning in Trustworthy ML

The final privacy parameters (ϵ, δ) for DPSGD are derived through a composition analysis, by considering the cumulative effect of applying the Gaussian mechanism to parameter gradients at every step of training. Because the overall privacy leakage scales with the number of training steps, and because training deep neural networks typically requires many epochs of model updates, practitioners have tended towards applying DPSGD to the *fine-tuning* stage of model training, assuming that a suitable model initialization θ_0 is available through pre-training on publicly available data [Papernot et al., 2020].³ Our experiments have covered both popular settings for DPSGD—private training from scratch (as in our MNIST experiments) and private fine-tuning (as in our CIFAR-10 experiments)—and applied ACA in each case. However, there are reasons to prefer the pre-training/fine-tuning paradigm beyond just privacy considerations, as the model utility has been shown to scale with dataset size and parameter count [Kaplan et al., 2020]. Indeed, AI practitioners have moved strongly towards the use and on-the-fly adaptation of pre-trained models in recent years [Bommasani et al., 2021]. This raises interesting questions about the role of ACA in shaping the behavior of modern models: are collectives most effective when inserting signals into pre-training data, fine-tuning data, preference data used for post-training, or some combination of these options?

6.5 Broader Impacts

ACA allows collectives to influence the outcomes of algorithmic systems and mitigate harms without directly relying on service providers and can be seen as a “*response from below*” [DeVrio et al., 2024] strategy. At the same time, depending on the motivation of the collective, ACA also has the potential of being misused either as data poisoning attacks or to exacerbate the preexisting harms. A system may also have multiple competing collectives with conflicting goals. This makes the motivation of the collectives an important factor in understanding the border social impact of ACA.

³This approach has also been critiqued for eschewing privacy considerations for individuals whose data comprises the pre-training dataset [Tramèr et al., 2024].

One possible motivation for collectives to organize could be the introduction of privacy-preserving techniques. Although these techniques offer data privacy and prevent the misuse of personal data, they could also have unintended consequences [Calvi et al., 2024]. These consequences could be in the forms of disparate impact [Bagdasaryan et al., 2019, Kulynych et al., 2019] or exacerbating bias [Fioretto et al., 2022]. Consequently, firms may also strategically adopt such privacy-preserving techniques not only to protect individual data but also to weaken the influence of groups acting on their learning algorithm. There is also a risk of fairwashing [Aïvodji et al., 2019] or using privacy as a pretext to limit accountability [Van Loo, 2022]. Paradoxically, knowing that DP is used could empower collective action. If individuals believe that their actions are masked by DP, they may be more willing to participate in collective action.

Our work takes a step towards understanding the competing tensions between privacy-preserving training and how differential privacy may affect the success of collective action.

7 Conclusion

In this paper, we focus on the intersection of Algorithmic Collective Action and Differential Privacy. Specifically, we investigated how privacy-preserving training using DPSGD 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 observed that the collective’s presence in the training data can provide some privacy benefits. 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.

Having established and characterized the tradeoff between DP and ACA, there are several promising directions for future work that might shed further light on the relationship between these two concepts. This includes an examination of alternative DPSGD design choices, such as the choice of clipping threshold or privacy accountant. Another direction of research is to investigate the potential of using activation functions specifically designed for privacy-preserving training, which are shown to improve the privacy-utility trade-offs [Papernot et al., 2021] and could also enhance the collective’s success under differential privacy.

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 (R³AI) 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>.
- Abien Fred Agarap. Deep learning using rectified linear units (relu), 2019. URL <https://arxiv.org/abs/1803.08375>.
- 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.

- Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. *Advances in neural information processing systems*, 32, 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.
- Graeme Blair, Kosuke Imai, and Yang-Yang Zhou and. Design and analysis of the randomized response technique. *Journal of the American Statistical Association*, 110(511):1304–1319, 2015. doi: 10.1080/01621459.2015.1050028. URL <https://doi.org/10.1080/01621459.2015.1050028>.
- Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.
- Alessandra Calvi, Gianclaudio Malgieri, and Dimitris Kotzinos. The unfair side of privacy enhancing technologies: Addressing the trade-offs between pets and fairness. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2047–2059, 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.
- Sammi Chen. The latest interface: Using data privacy as a sword and shield in antitrust litigation. *Hastings LJ*, 74:551, 2022.
- 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>.
- Julien Ferry, Ulrich Aïvodji, Sébastien Gambs, Marie-José Huguet, and Mohamed Siala. Sok: Taming the triangle—on the interplays between fairness, interpretability and privacy in machine learning. *arXiv preprint arXiv:2312.16191*, 2023.

- Ferdinando Fioretto, Cuong Tran, Pascal Van Hentenryck, and Keyu Zhu. Differential privacy and fairness in decisions and learning tasks: A survey. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2022.
- Sandy J.J. Gould. Differential Privacy and Collective Bargaining over Workplace Data. *Italian Labour Law e-Journal*, 17(2):133–144, December 2024. doi: 10.6092/ISSN.1561-8048/20838. URL <https://illej.unibo.it/article/view/20838>. Publisher: Italian Labour Law e-Journal.
- 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.
- Russell Hardin. *Collective action*. Rff Press, 1982.
- 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.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54(11s):1–37, 2022.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
- 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>.
- Bogdan Kulynych, Mohammad Yaghini, Giovanni Cherubin, Michael Veale, and Carmela Troncoso. Disparate vulnerability to membership inference attacks. *Proceedings on Privacy Enhancing Technologies*, 2022:460 – 480, 2019.
- 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. 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.
- Gerald Marwell and Pamela Oliver. *The critical mass in collective action*. Cambridge University Press, 1993.
- 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.ctvjjsf3ts>.
- Nicolas Papernot, Steve Chien, Shuang Song, Abhradeep Thakurta, and Ulfar Erlingsson. Making the shoe fit: Architectures, initializations, and tuning for learning with privacy. 2020.
- Nicolas Papernot, Abhradeep Thakurta, Shuang Song, Steve Chien, and Úlfar Erlingsson. Tempered sigmoid activations for deep learning with differential privacy. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(10):9312–9321, May 2021. doi: 10.1609/aaai.v35i10.17123. URL <https://ojs.aaai.org/index.php/AAAI/article/view/17123>.
- Vardan Pappayan, 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.
- Renee Marie Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N’Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 2022.
- Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.
- Dorothee Sigg, Moritz Hardt, and Celestine Mender-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.
- Florian Tramèr, Gautam Kamath, and Nicholas Carlini. Position: Considerations for differentially private learning with large-scale public pretraining. In *International Conference on Machine Learning*, pages 48453–48467. PMLR, 2024.
- Natálie Tůmová. Data privacy: A handy shield against anti-competitive behaviour in eu digital markets? *Charles University in Prague Faculty of Law Research Paper No*, 2024.
- 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.
- Rory Van Loo. Privacy pretexts. *Cornell L. Rev.*, 108:1, 2022.
- 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.
- Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.
- 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.
- Rui-Jie Yew, Lucy Qin, and Suresh Venkatasubramanian. You still see me: How data protection supports the architecture of ml surveillance. In *AAAI/ACM Conference on AI, Ethics, and Society*, 2024.
- Da Yu, Gautam Kamath, Janardhan Kulkarni, Tie-Yan Liu, Jian Yin, and Huishuai Zhang. Individual privacy accounting for differentially private stochastic gradient descent. *arXiv preprint arXiv:2206.02617*, 2022.
- 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 Tail bounds on norm of scaled standard Gaussian distribution

Lemma 1. Let $Y_1, \dots, Y_D \sim \mathcal{N}(0, \sigma^2)$ 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 $1 - \delta$,

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

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.2 Proof for Theorem 2

The gradient-redirecting strategy induces the following gradient evaluated on \mathcal{P}_t ,

$$\begin{aligned} g_{\mathcal{P}_t}^{\text{DP}}(\theta_t) &= \alpha g_{\mathcal{P}_t'}^{\text{DP}}(\theta_t) + (1 - \alpha) g_{\mathcal{P}_0}^{\text{DP}}(\theta_t) \\ &= \alpha g_{\mathcal{P}_t'}^{\text{clip}} + (1 - \alpha) g_{\mathcal{P}_0}^{\text{clip}} + \mathcal{N}(0, \sigma^2 C^2 I) \\ &= \alpha \xi^c(\theta_t) (\theta_t - \theta^*) + \mathcal{N}(0, \sigma^2 C^2 I) \end{aligned} \quad (2)$$

$$\text{where } \xi^c(\theta_t) = \frac{\|g_{\mathcal{P}_t'}^{\text{clip}}(\theta_t) + \frac{1-\alpha}{\alpha} g_{\mathcal{P}_0}^{\text{clip}}(\theta_t)\|}{\|\theta_t - \theta^*\|},$$

where to get Equation 2, we start by following a similar strategy to [Hardt et al. \[2023\]](#), which expresses the sum of expected gradients as a scalar multiple of the model update direction $(\theta_t - \theta^*)$. Refer to Definition 1. With $\xi_{\min}^c = \min_{\lambda \in [0, 1]} \xi(\lambda \theta_0 + (1 - \lambda) \theta^*)$, and using parameters update equation, we can derive an upper

bound on the difference between the learned and optimal parameter as follows:

$$\begin{aligned} \|\theta_T - \theta^*\| &\leq \|\theta_{T-1} - \eta(\alpha\xi^c(\theta_{T-1})(\theta_{T-1} - \theta^*) + \mathcal{N}(0, \sigma^2 C^2 I)) - \theta^*\| \\ &= \|(1 - \eta\alpha\xi^c(\theta_{T-1}))(\theta_{T-1} - \theta^*) - \eta\mathcal{N}(0, \sigma^2 C^2 I)\| \\ &\leq \|(1 - \eta\alpha\xi_{\min}^c)(\theta_{T-1} - \theta^*) - \eta\mathcal{N}(0, \sigma^2 C^2 I)\| \end{aligned} \quad (3)$$

$$\leq \left\| (1 - \eta\alpha\xi_{\min}^c)^T (\theta_0 - \theta^*) - \eta\mathcal{N}(0, \sigma^2 C^2 I) \left(1 + (1 - \eta\alpha\xi_{\min}^c) + \dots + (1 - \eta\alpha\xi_{\min}^c)^{T-1} \right) \right\| \quad (4)$$

$$= \left\| (1 - \eta\alpha\xi_{\min}^c)^T (\theta_0 - \theta^*) - \mathcal{N}(0, \sigma^2 C^2 I) \frac{(1 - (1 - \eta\alpha\xi_{\min}^c)^T)}{\alpha\xi_{\min}^c} \right\| \quad (5)$$

$$\stackrel{d}{=} \left\| (1 - \eta\alpha\xi_{\min}^c)^T (\theta_0 - \theta^*) + \mathcal{N}(0, \sigma^2 C^2 I) \frac{(1 - (1 - \eta\alpha\xi_{\min}^c)^T)}{\alpha\xi_{\min}^c} \right\| \quad (6)$$

$$\leq (1 - \eta\alpha\xi_{\min}^c)^T \|\theta_0 - \theta^*\| + \frac{1 - (1 - \eta\alpha\xi_{\min}^c)^T}{\alpha\xi_{\min}^c} \|\mathcal{N}(0, \sigma^2 C^2 I)\|. \quad (7)$$

By applying Lemma 1, we can further upper bound the right-hand side with probability greater than $1 - \delta$, assuming that θ has d degrees of freedom,

$$\|\theta_T - \theta^*\| \leq (1 - \eta\alpha\xi_{\min}^c)^T \|\theta_0 - \theta^*\| + \frac{\sigma C (1 - (1 - \eta\alpha\xi_{\min}^c)^T)}{\alpha\xi_{\min}^c} (\sqrt{d} + \sqrt{2 \log 1/\delta})$$

Collective success is defined simply as the negative difference norm between the learned and optimal parameters. Therefore our upper bound on the parameter norm difference becomes a lower bound on collective success:

$$\begin{aligned} S_T(\alpha, C, \sigma) &\geq -\|\theta_T - \theta^*\| \\ &= -(1 - \eta B(\alpha, C))^T \|\theta_0 - \theta^*\| - \sigma C \cdot f_1(B(\alpha, C)T, \eta) \cdot f_2(d, \delta), \end{aligned} \quad (8)$$

where $B(\alpha, C) = \alpha\xi_{\min}^c$, $f_1(B(\alpha, C), T, \eta) = \frac{(1 - (1 - \eta\alpha\xi_{\min}^c)^T)}{\alpha\xi_{\min}^c}$ and $f_2(d, \delta) = (\sqrt{d} + \sqrt{2 \log 1/\delta})$, which gives us the final bound. Additional details for some steps of the proof are provided for clarity. In step 3, we substitute θ_{T-1} with a smaller value, which results in a relaxed upper bound. Step 4 involves unrolling the gradient-update recursion leading to a geometric series whose first term is 1 and common ratio $1 - \eta\xi_{\min}^c$. In step 6, we use the fact that adding or subtracting a zero-centered Gaussian random variable results in random variables that are equal in distribution. Finally, in step 7, we apply the triangle inequality, where the sum of norms is greater than or equal to the norm of the sum.

B Additional Experimental Results

Table 2 shows baseline predictive accuracies for DP-trained classifiers on CIFAR-10. Figure 4 and Table 3 extend our results from Section 4.4 to include MIA results on the SVHN dataset.

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

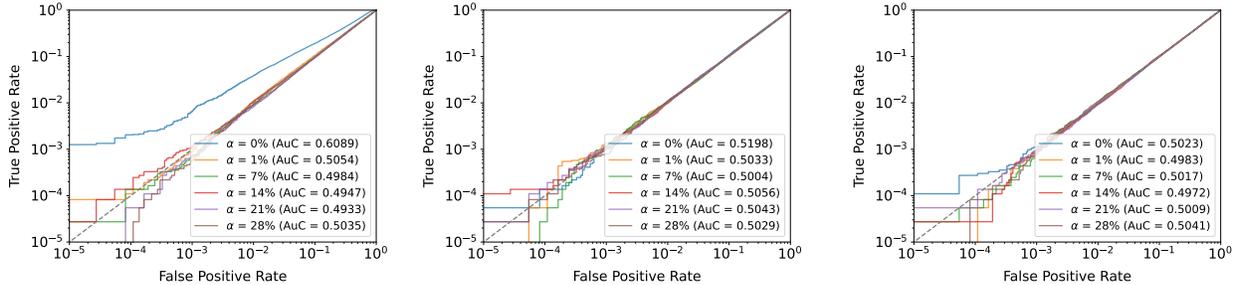


Figure 4: Success rate of Likelihood Ratio Attack (LiRA) [Carlini et al., 2022] evaluation on SVHN dataset. Each figure corresponds to a different setting of privacy constraints with privacy increasing from left to right.

Collective's size α	$\epsilon = \infty$		$\epsilon = 14.38$		$\epsilon = 0.98$	
	TPR @ 0.1% FPR	AuC	TPR @ 0.1% FPR	AuC	TPR @ 0.1% FPR	AuC
0%	0.65%	60.89%	0.11%	51.98%	0.11%	50.23%
1%	0.09%	50.54%	0.13%	50.33%	0.08%	49.83%
7%	0.07%	50.13%	0.11%	50.04%	0.08%	50.17%
14%	0.11%	49.47%	0.10%	50.56%	0.10%	49.72%
21%	0.07%	50.01%	0.10%	50.43%	0.08%	50.09%
28%	0.06%	50.35%	0.08%	50.29%	0.10%	50.41%

Table 3: 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.

C 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.

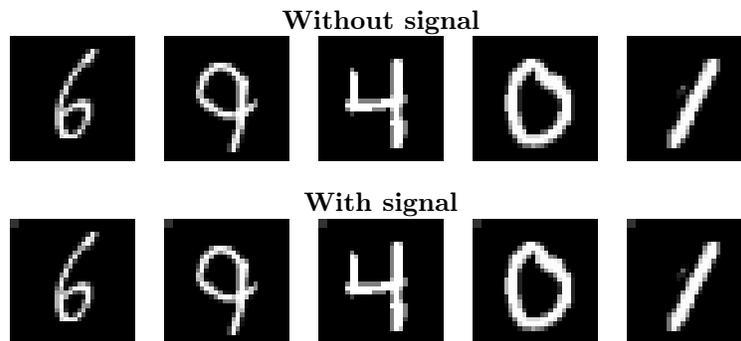


Figure 5: MNIST samples with and without adding application of transformation g

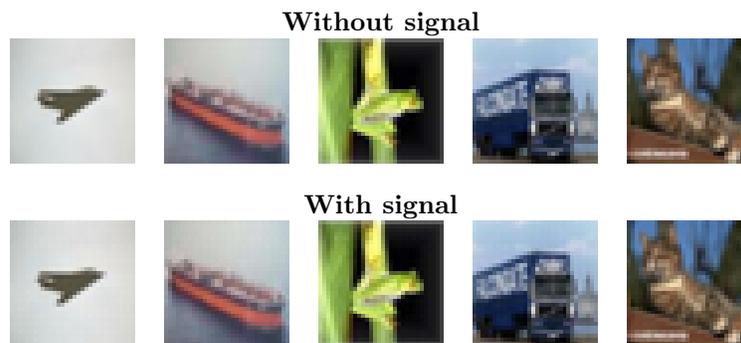


Figure 6: CIFAR-10 samples with and without adding application of transformation g