Enhancing One-run Privacy Auditing with Quantile Regression-Based Membership Inference

Anonymous authors
Paper under double-blind review

Abstract

Differential privacy (DP) auditing aims to provide empirical lower bounds on the privacy guarantees of DP mechanisms like DP-SGD. While some existing techniques require many training runs that are prohibitively costly, recent work introduces one-run auditing approaches that effectively audit DP-SGD in white-box settings while still being computationally efficient. However, in the more practical black-box setting where gradients cannot be manipulated during training and only the last model iterate is observed, prior work shows that there is still a large gap between the empirical lower bounds and theoretical upper bounds. Consequently, in this work, we study how incorporating approaches for stronger membership inference attacks (MIA) can improve one-run auditing in the black-box setting. Evaluating on image classification models trained on CIFAR-10 with DP-SGD, we demonstrate that our proposed approach, which utilizes quantile regression for MIA, achieves tighter bounds while *crucially* maintaining the computational efficiency of one-run methods.

1 Introduction

Differential privacy (DP) (Dwork et al., 2006) has become an effective, practical framework for specifying and ensuring privacy guarantees of statistical algorithms, including stochastic gradient descent (DP-SGD) for training large models privately (Chaudhuri et al., 2011; Abadi et al., 2016). While DP provides an upper bound on the privacy guarantee ε of the algorithm, it is useful to additionally have a *lower bound* on ε to validate it in practice and potentially detect errors in implementations (Ding et al., 2018; Jagielski et al., 2020; Tramer et al., 2022). This lower bound is derived empirically through *privacy auditing*.

DP Auditing often requires training a model hundreds—if not thousands—of times, inducing heavy computational requirements that simply don't scale when auditing larger models (Tramer et al., 2022). These costs are further exacerbated by the computational costs of calculating per-example gradients in DP-SGD. Despite recent advancements in computational efficiency (Nasr et al., 2023), multiple-run auditing still incurs overheads that can lead to prohibitively costly experiments (Muthu Selva Annamalai & De Cristofaro, 2024). In light of these problems, Steinke et al. (2023) introduce a new framework requiring only a single run. Framed as a guessing game, the goal is to identify among a set of "canary" examples the ones that were seen during training. If one is able to make more guesses correctly, then one can establish higher empirical lower bounds on ε.

We view these types of guessing games for DP auditing as a form of membership inference (Shokri et al., 2017), where the goal is determine if a given sample was used in training a machine learning model. However, Steinke et al. (2023) and Mahloujifar et al. (2024) introduce and evaluate their auditing schemes using only the simplest strategy for MIA, which can be summarized as looking at some score function (i.e., loss of the canary) and sorting (i.e., predicting that it was used in training if the loss is small and vice versa). We posit, however, that in applying this naive strategy, these auditing procedures may underestimate the empirical lower bounds for DP-SGD.

Contributions. In this work, we evaluate to what extent using strong MIA methods for privacy auditing in the one-run setting can tighten empirical privacy estimates. Given that the purpose of such one-run auditing

Algorithm 1: Differentially Private Stochastic Gradient Descent (DP-SGD)

Input: $x \in \mathcal{X}^n$

Requires: Loss function $f : \mathbb{R}^d \times \mathcal{X} \to \mathbb{R}$

Parameters: Number of iterations ℓ, learning rate η, clipping threshold $c > 0$, noise multiplier $\sigma > 0$, sampling probability $q \in (0, 1]$

Initialize $w_0 \in \mathbb{R}^d$;

for $t = 1, \ldots, \ell$ **do**

  Sample $S^t \subseteq [n]$ where each $i \in [n]$ is included independently with probability q;

  Compute $g_i^t = \nabla_{w^{t-1}} f(w^{t-1}, x_i) \in \mathbb{R}^d$ for all $i \in S^t$;

  Clip $\tilde{g}_i^t = \min\left\{1, \frac{c}{\|g_i^t\|_2}\right\} \cdot g_i^t \in \mathbb{R}^d$ for all $i \in S^t$;

  Sample $\xi^t \in \mathbb{R}^d$ from $\mathcal{N}(0, \sigma^2 c^2 I)$;

  Sum $\tilde{g}^t = \xi^t + \sum_{i \in S^t} \tilde{g}_i^t \in \mathbb{R}^d$;

  Update $w^t = w^{t-1} - \eta \cdot \tilde{g}^t \in \mathbb{R}^d$;

Output: $w^0, w^1, \ldots, w^\ell$

procedures is to assess privacy mechanisms while maintaining efficiency, we specifically adopt approaches for MIA introduced in Bertran et al. (2023), who introduce a class of attacks that compete with state-of-the-art shadow model approaches for MIA (Shokri et al., 2017; Carlini et al., 2022) while being computationally efficient (i.e., also require one training run).

We consider the black-box setting for auditing, where the auditor can only access the model at the final training step. Evaluating on image classification models trained on CIFAR-10 using DP-SGD, we demonstrate that MIA significantly improves empirical lower bounds estimated from one-run procedures introduced by Steinke et al. (2023) and Mahloujifar et al. (2024). Furthermore, we find that the advantage holds across a wide range of data settings (i.e., the number of training examples and proportion of canaries inserted into training), improving the lower bound on ε by up to **3x** in some cases.

1.1 Additional Related Works

In addition to those mentioned above, there have many other works that have recently studied private auditing under various scenarios. For example, rather than auditing models that are made private during training, Chadha et al. (2024) audit methods that are made private during inference. Pillutla et al. (2023), on the other hand, introduce the definition of Lifted Differential Privacy (LiDP) and propose a multi-run auditing procedure that can utilize multiple, randomized canaries (similar to our one-run auditing setting, in which the auditor also inserts many canaries). Furthermore, a variety of works have recently studied private auditing specifically under the constraints of black-box model access. Steinke et al. (2024) study black-box auditing of DP-SGD for models with linear structure, proposing a heuristic that predicts the outcome of an audit performed on only the last training iterate. Muthu Selva Annamalai & De Cristofaro (2024) show that empirical lower bounds for black-box auditing are much tighter when models are initialized to worst-case parameters, and lastly, Cebere et al. (2024) study black-box auditing in the case where an adversary can inject sequences of gradients that are crafted ahead of training.

2 Preliminaries

Update: We have substantially revised this entire section to provide more details about the one-run auditing methods of Steinke et al. (2023) and Mahloujifar et al. (2024) such that the information can be more digestible to readers unfamiliar with the prior work. Therefore for this section, we only highlight in red the changed text that is referenced directly in our review responses.

At a high level, differential privacy provides a mathematical guarantee that the output distribution of an algorithm is not heavily influenced by any single data point. Formally, it is defined as the following:

Definition 2.1 (Differential Privacy (DP) (Dwork et al., 2006))**.** A randomized algorithm $\mathcal{M} : \mathcal{X}^N \to \mathbb{R}$ satisfies (ε, δ)-differential privacy if for all neighboring datasets D, D' and for all outcomes $S \subseteq \mathbb{R}$ we have

$$P(\mathcal{M}(D) \in S) \leq e^{\varepsilon} P(\mathcal{M}(D') \in S) + \delta$$

To train deep learning models with privacy guarantees, differentially private form of stochastic gradient descent (DP-SGD) (Song et al., 2013; Bassily et al., 2014; Abadi et al., 2016), which the algorithm noises and clips gradients before every update step, is typically used. In our work, we aim to audit such models trained with DP-SGD. We present DP-SGD in more detail in Algorithm 1.

2.1 Auditing Differentially Privacy

Differentially private mechanisms, including DP-SGD, are accompanied by some proof that upper bounds the privacy parameters ε and δ. While the theoretical upper bound on ε guarantees that privacy loss (or leakage) cannot exceed some threshold, this bound is not tight. Privacy auditing instead provides an empirical *lower bound* on ε, using some form of membership inference and statistical testing to empirically show that (with some probably p), the privacy loss of a DP mechanism must be at least some amount (i.e., showing that ε must exceed some lower bound). At very high level, the more successful a membership inference attack on a model trained with DP-SGD is (i.e., the more privacy leakage occurs), the higher ε must be. Our work thus explores whether incorporating stronger membership inference attacks in privacy auditing can produce higher lower bound estimates for ε.

Black-box auditing. Nasr et al. (2023) presents two main threat models for privacy audits:

- White-box access: The auditor has full access throughout the training process to both model's weights and gradients, being able to inject arbitrarily-designed gradients at each update step
- Black-box access (with input space canaries): This approach is more restrictive. The auditor is only able to insert training samples in the dataset and observe the model at the end of training.

In our work, we study the *black-box* setting that does not allow modifications to the training procedure (i.e., modifying gradients like in white-box setting with Dirac gradients (Nasr et al., 2023; Steinke et al., 2023; Mahloujifar et al., 2024) or in an alternative black-box setting studied in Cebere et al. (2024) that allows gradient sequences to be inserted). This threat model is often more practically relevant and includes settings such as publishing the final weights of an open-sourced model. As shown in Nasr et al. (2023) and Steinke et al. (2023), the gap between the empirical lower bound and theoretical upper bound is generally still large in the black-box setting, suggesting that this area of research may still be underexplored.[1]

2.2 One-run auditing

To audit models trained using DP-SGD, we consider the "one-run" auditing procedures proposed by Steinke et al. (2023) and Mahloujifar et al. (2024), which we present in Algorithms 2 and 3, respectively, for black-box auditing. At a high level for both procedures, the auditor uses the private model w to compute some $\texttt{score}(w, c_{i,1})$ (e.g., negative cross entropy loss) for each canary c_i. Guesses are then made based on these scores (described below), and the final empirical lower bound is estimated based on the accuracy of the guesses.

We note that in Algorithms 2 and 3, we make minor changes to the notation compared to how they were original introduced in their respective works (Steinke et al., 2023; Mahloujifar et al., 2024). In this way, we make the notation of the two algorithms consistent with each other. For example, we now let n denote the total number of examples used in training (rather than the total number of auditing and non-auditing examples in Steinke et al. (2023)) and m be the total number of canaries (rather than canary sets in Mahloujifar et al. (2024)). In Algorithm 2, exactly half of the canaries are randomly sampled such that the data partitioning is exactly equivalent to Mahloujifar et al. (2024) when the canary set size is $K = 2$.

[1] Mahloujifar et al. (2024), for example, do not evaluate their proposed method in the black-box setting at all.

Algorithm 2: Black-box Auditing - One Run (Steinke et al., 2023)

Input: probability threshold τ, privacy parameter δ, training algorithm $\mathcal{A}$, dataset D, set of m canaries $C = \{c_1, \ldots, c_m\}$

Requires: scoring function `score`

Parameters: number of positive and negative guesses k_+ and k_-

1 Randomly split canaries C into two equally-sized sets C_{IN} and C_{OUT}

2 Let $S = \{s_i\}_{i=1}^m$, where $s_i = \begin{cases} 1 & \text{if } c_i \in C_{\text{IN}} \\ -1 & \text{if } c_i \in C_{\text{OUT}} \end{cases}$

3 Train model $w \leftarrow \mathcal{A}(D \cup C_{\text{IN}})$

4 Compute vector of scores $Y = \{\texttt{score}(w, c_i)\}_{i=1}^m$

5 Sort scores in ascending order $Y' \leftarrow \texttt{sort}(Y)$

6 Construct vector of guesses $T = \{t_i\}_{i=1}^m$, where

$$t_i = \begin{cases} 1 & \text{if } Y_i \text{ is among the top } k_+ \text{ scores in } Y \text{ (i.e., } Y_i \geq Y'_{m-k_+}\text{) // guess } c_i \in C_{\text{IN}} \\ -1 & \text{if } Y_i \text{ is among the bottom } k_- \text{ scores in } Y \text{ (i.e., } Y_i \leq Y'_{k_-}\text{) // guess } c_i \in C_{\text{OUT}} \\ 0 & \text{otherwise // abstain} \end{cases}$$

7 Compute empirical epsilon $\tilde{\varepsilon}$ (i.e., find the largest $\tilde{\varepsilon}$ such that S, T, τ, and δ satisfy Theorem 1)

Output: $\tilde{\varepsilon}$

2.2.1 Auditing with One Run (Steinke et. al, 2023)

Steinke et al. (2023). Steinke et al. (2023) first developed the notion of auditing in one training run. Rather than training many models on neighboring datasets that differ on *single* examples, their auditing scheme requires training only a single model on a dataset with *many* "canary" examples. We summarize this procedure in Algorithm 2. At a high level, half of the canaries are randomly sampled from a larger set of m canaries and included in the training set. The auditor then predicts which of the m canaries were in and not in the training set by sorting the canaries by their score and guessing that the top k_+ are in the training set and bottom k_- are not (while abstaining for the remaining canaries). The final empirical lower bound on ε is determined by how many total guesses were made and how many were correct. Specifically, Steinke et al. (2023) use binary search to empirically estimate the largest value for ε such that Theorem 1 is still satisfied.

Theorem 1 (Analytic result for approximate DP (Steinke et al., 2023))**.** Suppose $\mathcal{A} : \{-1, 1\}^m \to \{-1, 0, 1\}^m$ satisfy (ε, δ)-DP. Let $S \in \{-1, 1\}^m$ be uniformly random and $T = \mathcal{A}(S)$. Suppose $\mathbb{P}[\|T\|_1 \leq r] = 1$. Then, for all $v \in \mathbb{R}$,

$$\mathbb{P}_{\substack{S \leftarrow \{-1,1\}^m \\ T \leftarrow M(S)}} \left[\sum_{i=1}^m \max\{0, T_i \cdot S_i\} \geq v \right] \leq f(v) + 2m\delta \cdot \max_{i \in \{1,\ldots,m\}} \left\{ \frac{f(v-i) - f(v)}{i} \right\},$$

where

$$f(v) := \mathbb{P}_{\tilde{W} \leftarrow \text{Binomial}\left(r, \frac{e^\varepsilon}{e^\varepsilon + 1}\right)} \left[\tilde{W} \geq v \right].$$

At a high level, Theorem 1 bounds the success rate (number of correct guesses v) of the auditor for the privacy parameter ε. Given the success rate of some membership inference attack on m canaries, the auditor can then check whether some ε would violate this bound. In more detail, $\mathcal{A}$ is some randomized mechanism (i.e., DP-SGD in our work) that takes in as input some set of m canaries that are labeled as being included ($S = 1$) or excluded ($S = -1$) from the training set. Observing the outputs of $\mathcal{A}$, the auditor then makes guesses $T \in \{-1, 0, 1\}^m$ for the m canaries where $T_i = 0$ means that the auditor abstains from making a guess for a canary i. $T_i \cdot S_i = 1$ if a guess i is correct and $\sum_{i=1}^m \max\{0, T_i \cdot S_i\}$ counts the total number of correct guesses. Thus, Theorem 1 bounds the probability of making at least v correct guesses.

Algorithm 3: Black-box Auditing - One Run (Mahloujifar et al., 2024)

Input: privacy parameter δ, training algorithm $\mathcal{A}$, dataset D, set of m canaries $C = \{c_1, \ldots, c_m\}$
Requires: scoring function `score`
Parameters: number of guesses k

Randomly split canaries C into two equally-sized sets C_{IN} and C_{OUT}

Create disjoint canary sets $E = \{e_i\}_{i=1}^{m/2}$ by randomly pairing canaries from C_{IN} and C_{OUT} such that $e_i = (c_{i,1}, c_{i,2})$ for $c_{i,1} \in C_{\text{IN}}$ and $c_{i,2} \in C_{\text{OUT}}$ (each canary $c \in C$ appears in **exactly** one set e_i)

Train model $w \leftarrow \mathcal{A}(D \cup C_{\text{IN}})$

Compute vector of scores $Y = \{|\texttt{score}(w, c_{i,1}) - \texttt{score}(w, c_{i,2})|\}_{i=1}^{m/2}$

Sort scores in ascending order $Y' \leftarrow \texttt{sort}(Y)$

Construct vector of guesses $T = \{t_i\}_{i=1}^{m/2}$, where

$$t_i = \begin{cases} 1 & \text{if } Y_i \text{ is among the top } k \text{ values in } Y \text{ (i.e., } Y_i \geq Y'_{m-k}) \\ & \quad \text{and } \texttt{score}(w, c_{i,1}) > \texttt{score}(w, c_{i,2}) \texttt{// guess } c_{i,1} \in C_{\text{IN}} \\ -1 & \text{if } Y_i \text{ is among the top } k \text{ values in } Y \text{ (i.e., } Y_i \geq Y'_{m-k}) \\ & \quad \text{and } \texttt{score}(w, c_{i,1}) \leq \texttt{score}(w, c_{i,2}) \texttt{// guess } c_{i,2} \in C_{\text{IN}} \\ 0 & \text{otherwise } \texttt{// abstain} \end{cases}$$

Let number of correct guesses $k' = \sum_{i=1}^{m/2} \mathbb{1}\{t_i = 1\}$

Compute empirical epsilon $\tilde{\varepsilon}$ (i.e., find the largest $\tilde{\varepsilon}$ whose corresponding f-DP function f passes Algorithm 4 for m, k, k' τ, and δ.)

Output: $\tilde{\varepsilon}$

Algorithm 4: Upper bound probability of making correct guesses (Mahloujifar et al., 2024)

Input: probability threshold τ, functions f and f^{-1}, number of guesses k, number of correct guesses k', number of samples m, alphabet size s

$\forall 0 < i < k'$ set $h[i] = 0$, and $r[i] = 0$

Set $r[k'] = \tau \cdot \frac{c}{m}$

Set $h[k'] = \tau \cdot \frac{c'-c}{m}$

for $i \in [k'-1, \ldots, 0]$ **do**

  $h[i] = (s-1) f^{-1}(r[i+1])$

  $r[i] = r[i+1] + \frac{i}{k-i} \cdot (h[i] - h[i+1])$

if $r[0] + h[0] \geq \frac{k}{m}$ **then**

  Return True (probability of k' correct guesses (out of k) is less than τ)

else

  Return False (probability of having k' correct guesses (out of k) could be more than τ)

2.2.2 Auditing f-DP with One Run (Mahloujifar et. al, 2023)

More recently, Mahloujifar et al. (2024) present an alternative approach based on f-DP (Dong et al., 2022), which they show provides better privacy estimates in the one-run setting. Rather than having a single set of canaries, Mahloujifar et al. (2024)'s method first constructs a set of canary sets of size K, where a random example in each canary set is using in training. Here, the goal is to guess which of the K canaries in each set was used in training. Both in Mahloujifar et al. (2024)'s and our experiments, the canary set size is $K = 2$. Thus to make guesses, we calculate the absolute difference in scores between the canaries in each pair and sort these sets by this difference, making guesses for the k pairs with the largest absolute difference. For each of these k pairs, we guess that the canary with the higher score was included in the training set. Finally, like in Steinke et al. (2023), the empirical lower bound is determined by the number of guesses made and the number that are correct. We present their procedure in Algorithms 3 and 4.

To explain the auditing process in more detail, we first define f-differential privacy, which Mahloujifar et al. (2024) use to estimate the empirical lower bound.

Definition 2.2 (f-Differential Privacy (Dong et al., 2022))**.** A mechanism $\mathcal{M}$ is f-DP if for all neighboring datasets $\mathcal{S}, \mathcal{S}'$ and all measurable sets T with $|\mathcal{S} \triangle \mathcal{S}'| = 1$, we have

$$\Pr[\mathcal{M}(\mathcal{S}) \in T] \leq \bar{f}\left(\Pr[\mathcal{M}(\mathcal{S}') \in T]\right). \tag{1}$$

We note that f-DP relates to (ε, δ)-DP in the following way:

Proposition 1. A mechanism is (ε, δ)-DP if it is f-DP with respect to $\bar{f}(x) = e^{\varepsilon}x + \delta$, where $\bar{f}(x) = 1 - f(x)$.

In Algorithm 3, the auditor makes and evaluates some set of k guesses and then computes and empirical lower bound using Algorithm 4. To estimate the lower bound, the auditor first construct a set of candidate values for ε and a corresponding f-DP function for each. For each value of ε (and function f), Algorithm 4 runs a hypothesis test for the number of correct guesses k'. The final empirical lower bound is the maximum ε that passes this hypothesis test.

3 Applying (Efficient) MIA to Privacy Auditing

As discussed in the previous section, single-one run auditing requires running a membership inference attack on the canaries, where the auditor makes guesses under the assumption that the higher the `score` is, the more likely the canary was seen during training. In the formulation of their algorithm, Steinke et al. (2023)[2] use the simplest approach for membership inference—taking the loss directly as the `score` function (e.g., calculating negative cross entropy loss). However, this simple approach can be naive. For example, if a canary has a low cross-entropy loss, is it due to it being seen in training or it being an easy image to classify?

"Stronger" membership inference attacks more effectively account for the characteristics of the target sample. For instance, the most effective approaches (Carlini et al., 2022) train multiple shadow models on random subsets of data that either include or exclude the sample. These methods generate a distribution of losses for models that have seen the sample during training and for those that have not. By performing a likelihood-ratio test on this distribution, these approaches estimate the likelihood that the observed loss came from a model trained on the sample. Since this likelihood-ratio test is conducted separately for each sample, the resulting scores are inherently better calibrated to each individual sample, leading to more accurate membership inference.

This shadow-model approach, while effective, requires high computational demands that do not align with the goals of one-run auditing.[3] In contrast, Bertran et al. (2023) introduce a new class of MIA methods that relies on training a single quantile regressor on holdout data only. In doing so, they predict a threshold for determining membership that like shadow-model approaches, is calibrated to the difficulty of each sample. Moreover, Bertran et al. (2023) show that this lightweight approach both outperforms marginal thresholds, which are equivalent to the sort and rank style procedure Steinke et al. (2023) and Mahloujifar et al. (2024), and are competitive with state-of-the-art shadow-model approaches.

3.1 Membership Inference via Quantile Regression

Formally, the quantile regressor can be written as the following:

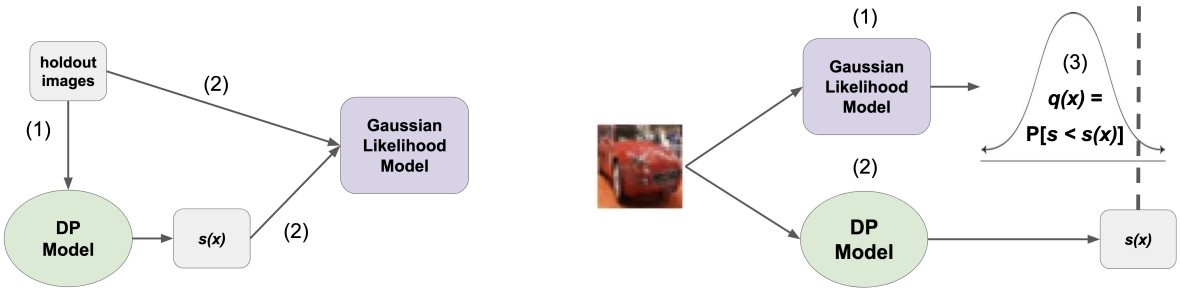
Definition 3.1 (Quantile Regressor)**.** Given a target false positive rate α, a quantile regressor is a model $q : \mathcal{X} \to \mathbb{R}$ trained on an holdout dataset $\mathcal{P}$ to predict the $(1-\alpha)$-quantile for the score distribution associated to each given sample:

$$\forall (x, s) \in \mathcal{P},\ Pr[y \leq q(x)] = 1 - \alpha$$

Given the relatively small sample size in image datasets like CIFAR-10, Bertran et al. (2023) propose an alternative method for outputting quantile thresholds in which they train a model that instead predicts the

[2]Note that Mahloujifar et al. (2024) do not conduct black-box auditing experiments.

[3]For example, one run of DP-SGD for our experiments at $\varepsilon = 8.0$ takes approximately 12 hours on a single GPU, making training even tens of shadow models (with DP) for each of the m canaries infeasible for us.



(a) Train "MIA" regressor

(b) "Rescore" canaries

Figure 1: We provide a high-level diagram describing our quantile-regression based MIA approach to auditing. In the first part **(a)**, we **(1)** calculate the score $s(x)$ (e.g., loss) using the privately trained model and **(2)**, train the Gaussian likelihood model on the images x themselves and $s(x)$. Once the Gaussian likelihood model has been trained on the holdout set, in part **(b)**, we take each canary and **(1)** use the Gaussian likelihood model to output the parameters of a Gaussian distribution (i.e., μ, σ). **(2)** Next, we again feed the canary into the private model to obtain $s(x)$. **(3)** Finally, we calculate our new score, $q(x) = P(s < s(x) \mid \mu, \sigma)$.

mean $\mu(x)$ and the standard deviation $\sigma(x)$ of the score $s(x)$ (e.g., loss of the model to be attacked) associated with each example x. The per-example threshold is then calculated based on this normal distribution (i.e., $P(s < s(x) \mid \mu, \sigma)$).

The loss can then be written as the following:

Definition 3.2 (Negative Log-Likelihood for Gaussian Distributions)**.** The negative log-likelihood loss for a Gaussian distribution with mean μ and standard deviation σ is given by:

$$\mathcal{L}_{\text{NLL}} = \mathbb{E}_{x \sim p(x)} \left[\frac{(x-\mu)^2}{2\sigma^2} + \log \sigma \right]$$

where $x \sim p(x)$ represents samples from some underlying data distribution (e.g., losses from an image classification model).

3.2 Proposed Methodology

At a high level, our approach can be summarized as proposing a new scoring function `score` that is used in Algorithms 2 and 3 to produce guesses for auditing. In this way, we reduce the problem integrating different membership inference attacks into existing one-run auditing procedures to simply changing `score`. Importantly, by simply modifying how black-box auditing traditionally scores the canaries (i.e., directly calculating some metric like cross entropy loss of the canary), our proposed methodology inherits the statistical guarantees of the original auditing algorithms. Moreover, this approach—unlike shadow model based ones—requires no additional runs of DP-SGD.

We present our proposed method in Figure 1. In more detail, let $s(x)$ be some base scoring function. Adapting the MIA approach of Bertran et al. (2023), we train a neural network that takes in as input the canary c_i and predicts a Gaussian distribution for $s(c_i)$. However, unlike Bertran et al. (2023), who use the Gaussian distribution to predict the threshold of the q-quantile for some predetermined value of q, we calculate q directly as the CDF: $q(c_i) = P(s < s(x) \mid \mu, \sigma)$. We then use q as the input `score` function for Algorithms.

In the following sections discussing our empirical evaluation, we will refer to the base scoring function $s(x)$ (i.e., the baseline) as $\texttt{score}_{\texttt{base}}$ and the quantile scoring function $q(x)$ (i.e., our method) as $\texttt{score}_{\texttt{quantile}}$.

Table 1: We list the hyperparameters for training Wide ResNet 16-4 models using DP-SGD with $\varepsilon = 8.0$ and $\delta = 10^{-5}$. We also report the average test accuracy of the trained models in our experiments.

Hyperparameter	$n = 47500$	$n = 20000$	$n = 10000$	$n = 5000$
Augmentation multiplicity	16	16	16	16
Batch size	4096	2048	1024	512
Clipping norm	1.0	1.0	1.0	1.0
Learning rate	4.0	2.0	1.0	1.0
Noise multiplier	3.0	2.5	2.0	2.0
Test Accuracy	79.73	68.11	58.30	50.75

4 Empirical Evaluation

Auditing setup. For our empirical evaluation, we follow the experimental set up in prior work (Nasr et al., 2023; Steinke et al., 2023; Mahloujifar et al., 2024) and train Wide ResNet (WRN) 16-4 models (Zagoruyko & Komodakis, 2016) from scratch using DP-SGD (at $\varepsilon = 8.0, \delta = 10^{-5}$) on the CIFAR-10 dataset (Krizhevsky et al., 2009). In these experiments, differential privacy is defined at the sample-level privacy with add/remove adjacency. All models are trained using code provided by Balle et al. (2022), which implements training of state-of-the-art DP CIFAR-10 models presented in De et al. (2022) using a Rényi DP (Mironov et al., 2019) privacy accountant.

As conducted in Steinke et al. (2023) and Mahloujifar et al. (2024), we randomly sample 5000 examples from the training set to be canaries and use the remaining 45000 as non-canaries, giving us in total $n = 47500$ (i.e., $45000 + \frac{5000}{2}$) examples in the training set. We also run additional experiments varying the size of the training set, using $n = 5000, 10000$ and 20000 (discussed further in Section 5. We report the hyperparameters of DP-SGD and test accuracy of the trained models for each value of n in Table 1.

Note that while Steinke et al. (2023) experimented with black-box canaries with both flipped and unperturbed class labels, we found early on that flipping labels did not improve the lower bound. Thus, given that perturbing the labels can only hurt the final DP model's accuracy, we do not flip the canary labels in our experiments.

Choice of number of guesses k. In general, empirical lower bounds on ε can be quite sensitive to the number guesses made (Mahloujifar et al., 2024). However, it is unclear from both Steinke et al. (2023) and Mahloujifar et al. (2024) how the number of guesses was chosen to produce their main results. For example, Steinke et al. (2023) state that they "evaluate different values of k_+ and k_- and only report the highest auditing results," but do not specify what exact values were tested. We reached out to the authors, who told us that some values between 10 and 1000 were chosen (but not exactly how many values of k were tested). Consequently, we evaluate all methods in our experiments from 10 to the maximum number of guesses possible in multiples of 10, and like prior work (Nasr et al., 2023; Steinke et al., 2023; Mahloujifar et al., 2024), report the highest auditing results for each run. We note that, as stated in Steinke et al. (2023) (Section 6), while evaluating different guesses is equivalent to running multiple hypothesis tests (thereby reducing the confidence value of experiments), the practice is commonly used in prior work on privacy auditing.

Quantile regressor. Following Bertran et al. (2023), we use a pretrained ConvNeXt (Liu et al., 2022) as our model architecture for the quantile regressor. We train for 5 epochs using Adam with a learning rate 10^{-4} and batch size of 128. Similarly, we use as our base score function (i.e., $\texttt{score}_{\texttt{base}}$) the difference in logit of the true class label and the sum of the remaining logits. As shown in Carlini et al. (2022), this score—in contrast to cross-entropy loss—follows a normal distribution empirically, making it a natural choice for our approach.

Table 2: We present the empirical lower bounds estimated the baseline method ($\texttt{score}_{\texttt{base}}$) and quantile regression ($\texttt{score}_{\texttt{quantile}}$). ε_{or} corresponds to Steinke et al. (2023), $\varepsilon_{\text{or-fdp}}$ corresponds to Mahloujifar et al. (2024), and $\varepsilon_{\max}$ corresponds to max of the two. We calculate ε for 5 different runs and report the average.

n	method	$r = 45000$, $m = 5000$		
		ε_{or}	$\varepsilon_{\text{or-fdp}}$	$\varepsilon_{\max}$
47500	$\texttt{score}_{\texttt{base}}$	0.159	**0.147**	0.208
	$\texttt{score}_{\texttt{quantile}}$ (*ours*)	**0.210**	0.134	**0.253**

5 Results

All results reported in Tables 2 and 3 are averages over the maximum lower bound (with 95% confidence) over 5 different runs, each of which is conducted by randomly partitioning the dataset into canary and non-canary sets. In these tables, ε_{or} corresponds to Steinke et al. (2023) and $\varepsilon_{\text{or-fdp}}$ corresponds to Mahloujifar et al. (2024). In addition, we consider the setting in which one considers the choice of auditing procedure (i.e., Steinke et al. (2023) vs Mahloujifar et al. (2024)) as an additional parameter that can be chosen by the auditor.[4] In this case, we take the max of ε_{or} and $\varepsilon_{\text{or-fdp}}$ for each run, which we denote as $\varepsilon_{\max}$, and again report the average over 5 runs in Tables 2 and 3.

In Table 2, we present our results when auditing a model trained with $n = 47500$ examples where $m = 5000$ and $r = 45000$[5]. For our method, we use the remaining 10000 holdout set examples to train the quantile regression model. In Table 3, we run experiments similar to those found in Steinke et al. (2023) for the black-box setting, where the number of training examples n is smaller. For each choice of n, we run experiments for both when $r = 0$ (all training examples are canaries) and $r = \frac{n}{2}$ (half of the training examples are canaries). In these experiments, we randomly sample 20000 examples out of the remaining holdout set examples to train our quantile regression model.

In most cases, we find that our method, $\texttt{score}_{\texttt{quantile}}$, achieves tighter privacy estimates, estimating higher empirical lower bounds on ε across various data settings (i.e., choices of n, m, and r) and auditing procedures (ε_{or}, $\varepsilon_{\text{or-fdp}}$, and $\varepsilon_{\max}$). In cases where the baseline performs better, the difference between it and our method is small (e.g., difference of 0.20 for $n = 5000$, $r = \frac{n}{2}$). Our results strongly indicate that better member inference attacks can improve DP-SGD auditing and suggest that in general, MIA methods should be incorporated into auditing experiments when applicable.

Additional insights. Lastly, we present additional observations about how one-run auditing procedures operate in the black-box setting. First, we note that generally speaking, we observe no clear winner between Steinke et al. (2023) and Mahloujifar et al. (2024) in the black-box setting, in contrast to the white-box setting in which Mahloujifar et al. (2024) achieves much tighter auditing results compared to Steinke et al. (2023). In all cases, the average $\varepsilon_{\max}$ strictly dominates both ε_{or} and $\varepsilon_{\text{or-fdp}}$, further suggesting that one auditing procedure does not consistently outperform the other. In addition, while Steinke et al. (2023) posit that when all training examples are canaries ($r = 0$), one can achieve higher auditing results, Table 3 does not clearly corroborate this hypothesis (if anything, the auditing procedures estimate slightly higher lower bounds when $r = \frac{n}{2}$). We leave further investigation of such observations to future work.

6 Conclusion

We study auditing of differential privacy in the black-box setting, empirical auditing image classification models when trained with DP-SGD. Focusing on one-run auditing methods, we make the observation that quantile-regression based MIA approaches complement the computationally efficient nature of one-run

[4]Similarly to how Steinke et al. (2023) report the maximum over lower bounds produced by flipping and not flipping labels.

[5]We note that this data setup corresponds to the experiments described in Steinke et al. (2023) under their notation of $n = 50000$ and $m = 5000$. While both Steinke et al. (2023) and Mahloujifar et al. (2024) audit this model in the white-box setting, neither report results for it in the black-box setting.

Table 3: We present the empirical lower bounds estimated using the baseline method ($\texttt{score}_{\texttt{base}}$) and quantile regression ($\texttt{score}_{\texttt{quantile}}$) for various data settings, including when the canaries make up all ($r = 0$) and half ($r = \frac{n}{2}$ of the training examples. ε_{or} corresponds to Steinke et al. (2023), $\varepsilon_{\text{or-fdp}}$ corresponds to Mahloujifar et al. (2024), and $\varepsilon_{\max}$ corresponds to max of the two. We calculate ε for 5 different runs and report the average.

n	method	$r = 0$, $m = 2n$			$r = \frac{n}{2}$, $m = n$		
		ε_{or}	$\varepsilon_{\text{or-fdp}}$	$\varepsilon_{\max}$	ε_{or}	$\varepsilon_{\text{or-fdp}}$	$\varepsilon_{\max}$
5000	$\texttt{score}_{\texttt{base}}$	0.181	0.175	0.237	**0.299**	0.234	0.393
	$\texttt{score}_{\texttt{quantile}}$ (*ours*)	**0.280**	**0.240**	**0.364**	0.279	**0.486**	**0.503**
10000	$\texttt{score}_{\texttt{base}}$	**0.202**	0.172	0.216	0.227	0.115	0.241
	$\texttt{score}_{\texttt{quantile}}$ (*ours*)	0.201	**0.339**	**0.364**	**0.341**	**0.217**	**0.356**
20000	$\texttt{score}_{\texttt{base}}$	0.055	0.086	0.098	0.128	0.191	0.204
	$\texttt{score}_{\texttt{quantile}}$ (*ours*)	**0.146**	**0.246**	**0.268**	**0.165**	**0.313**	**0.324**

procedures introduced by Steinke et al. (2023) and Mahloujifar et al. (2024). Empirically, we demonstrate that our quantile-regression based approach improves the baseline procedures across a wide range of data settings. We recognize, however, that our experiments show that large gap between empirical and theoretical privacy still exists. We hope that our work will help inspire future studies that may attempt to further close this gap in the black-box auditing setting.

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*, pp. 308–318, 2016.

Borja Balle, Leonard Berrada, Soham De, Sahra Ghalebikesabi, Jamie Hayes, Aneesh Pappu, Samuel L Smith, and Robert Stanforth. JAX-Privacy: Algorithms for privacy-preserving machine learning in jax, 2022. URL `http://github.com/google-deepmind/jax_privacy`.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Differentially private empirical risk minimization: Efficient algorithms and tight error bounds. *arXiv preprint arXiv:1405.7085*, 2014.

Martin Bertran, Shuai Tang, Aaron Roth, Michael Kearns, Jamie H Morgenstern, and Steven Z Wu. Scalable membership inference attacks via quantile regression. *Advances in Neural Information Processing Systems*, 36:314–330, 2023.

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)*, pp. 1897–1914. IEEE, 2022.

Tudor Cebere, Aurélien Bellet, and Nicolas Papernot. Tighter privacy auditing of dp-sgd in the hidden state threat model. *arXiv preprint arXiv:2405.14457*, 2024.

Karan Chadha, Matthew Jagielski, Nicolas Papernot, Christopher Choquette-Choo, and Milad Nasr. Auditing private prediction. *arXiv preprint arXiv:2402.09403*, 2024.

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

Soham De, Leonard Berrada, Jamie Hayes, Samuel L Smith, and Borja Balle. Unlocking high-accuracy differentially private image classification through scale. *arXiv preprint arXiv:2204.13650*, 2022.

Zeyu Ding, Yuxin Wang, Guanhong Wang, Danfeng Zhang, and Daniel Kifer. Detecting violations of differential privacy. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pp. 475–489, 2018.

Jinshuo Dong, Aaron Roth, and Weijie J. Su. Gaussian differential privacy. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(1):3–37, 2022.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006.

Matthew Jagielski, Jonathan Ullman, and Alina Oprea. Auditing differentially private machine learning: How private is private sgd? *Advances in Neural Information Processing Systems*, 33:22205–22216, 2020.

Alex Krizhevsky et al. Learning multiple layers of features from tiny images, 2009.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11976–11986, 2022.

Saeed Mahloujifar, Luca Melis, and Kamalika Chaudhuri. Auditing f-differential privacy in one run. *arXiv preprint arXiv:2410.22235*, 2024.

Ilya Mironov, Kunal Talwar, and Li Zhang. R\'enyi differential privacy of the sampled gaussian mechanism. *arXiv preprint arXiv:1908.10530*, 2019.

Meenatchi Sundaram Muthu Selva Annamalai and Emiliano De Cristofaro. Nearly tight black-box auditing of differentially private machine learning. *Advances in Neural Information Processing Systems*, 37:131482–131502, 2024.

Milad Nasr, Jamie Hayes, Thomas Steinke, Borja Balle, Florian Tramèr, Matthew Jagielski, Nicholas Carlini, and Andreas Terzis. Tight auditing of differentially private machine learning. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 1631–1648, 2023.

Krishna Pillutla, Galen Andrew, Peter Kairouz, H Brendan McMahan, Alina Oprea, and Sewoong Oh. Unleashing the power of randomization in auditing differentially private ml. *Advances in Neural Information Processing Systems*, 36:66201–66238, 2023.

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)*, pp. 3–18. IEEE, 2017.

Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. Stochastic gradient descent with differentially private updates. In *2013 IEEE global conference on signal and information processing*, pp. 245–248. IEEE, 2013.

Thomas Steinke, Milad Nasr, and Matthew Jagielski. Privacy auditing with one (1) training run. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pp. 49268–49280, 2023.

Thomas Steinke, Milad Nasr, Arun Ganesh, Borja Balle, Christopher A Choquette-Choo, Matthew Jagielski, Jamie Hayes, Abhradeep Guha Thakurta, Adam Smith, and Andreas Terzis. The last iterate advantage: Empirical auditing and principled heuristic analysis of differentially private sgd. *arXiv preprint arXiv:2410.06186*, 2024.

Florian Tramer, Andreas Terzis, Thomas Steinke, Shuang Song, Matthew Jagielski, and Nicholas Carlini. Debugging differential privacy: A case study for privacy auditing, 2022. URL `https://arxiv.org/abs/2202.12219`.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.