

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

Terrance Liu
Carnegie Mellon University

terrancel@cmu.edu

Matteo Boglioni*
ETH Zurich

mboglioni@ethz.ch

Yiwei Fu*
Carnegie Mellon University

yiweif@cmu.edu

Shengyuan Hu*
Carnegie Mellon University

shengyuanhu@cmu.edu

Pratiksha Thaker*
Carnegie Mellon University

pthaker@cmu.edu

Zhiwei Steven Wu*
Carnegie Mellon University

zstevenwu@cmu.edu

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

*Order alphabetically by last name.

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 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 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

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 $\mathcal{M} : \mathcal{X}^N \rightarrow \mathbb{R}$ 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, Abadi et al. (2016) propose a differentially private form of stochastic gradient descent (DP-SGD), in which the algorithm noises and clips gradients before every update step.

Algorithm 1 Auditor with One Training Run

Require: Algorithm to audit \mathcal{A} , target number of examples to train on n , scoring function SCORE

Require: Number of positive and negative guesses k_+ and k_- respectively,

Require: $x \in \mathcal{X}^{m+r}$ consisting of m auditing examples (a.k.a. canaries) and r non-auditing examples, where $n = r + \frac{m}{2}$

- 1: Randomly assign $S_i = +1$ to half of the m canaries and $S_i = -1$ to the other half. Set $S_i = 1$ for all remaining examples $i \in [m+r] \setminus [m]$.
 - 2: Partition x into $x_{\text{IN}} \in \mathcal{X}^{n_{\text{IN}}}$ and $x_{\text{OUT}} \in \mathcal{X}^{n_{\text{OUT}}}$ according to S , where $n_{\text{IN}} + n_{\text{OUT}} = n$. Namely, if $S_i = 1$, then x_i is in x_{IN} ; and, if $S_i = -1$, then x_i is in x_{OUT} .
 - 3: Run \mathcal{A} on input x_{IN} with appropriate parameters, outputting w .
 - 4: Compute the vector of scores $Y = (\text{SCORE}(x_i, w) : i \in [m]) \in \mathbb{R}^m$
 - 5: (i.e., $T \in \{-1, 0, +1\}^m$ maximizes $\sum_i T_i \cdot Y_i$ subject to $\sum_i |T_i| = k_+ + k_-$ and $\sum_i T_i = k_+ - k_-$).
 - 6: **return** The vector $S \in \{-1, +1\}^m$ indicating the true selection and the guesses $T \in \{-1, 0, +1\}^m$.
-

2.1 One-run auditing

To audit models trained using DP-SGD, we consider the following “one-run” auditing procedures:

1. **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. Specifically, these canaries are randomly sampled from a larger set of canaries. The auditor then attempts to predict which canaries were in and not in the training set (with abstentions are allowed). The final empirical lower bound is determined by how many guesses were made and how many were correct.
2. **Mahloujifar et al. (2024).** More recently, Mahloujifar et al. (2024) present an alternative approach, 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. As in Steinke et al. (2023), abstentions are also allowed, and again, the empirical lower bound is determined by the number of guesses made and the number that are correct.

We present in Algorithms 1 and 2 the auditing procedures for Steinke et al. (2023) and Mahloujifar et al. (2024), respectively.

We note that in Algorithms 1 and 2, 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 1, 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 $K = 2$.

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

- **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;

Algorithm 2 Reconstruction in one run game

Require: Algorithm to audit \mathcal{A} , target number of examples to train on n , scoring function SCORE

Require: Number of guesses k respectively

Require: $M = \frac{m}{K}$ sets (of size K) of canary examples and r non-auditing examples, where $n = r + \frac{m}{K}$ (assume that m is a multiple of K).

- 1: Let $\mathcal{C} = \{x_j^i\}_{i \in [M], j \in [K]}$ be the matrix of canaries
 - 2: Let $u = (u_1, \dots, u_M)$ be a vector uniformly sampled from $[K]^M$.
 - 3: Let $S = \{x_{u_i}^i : i \in [M]\}$.
 - 4: Run mechanism \mathcal{A} on $S \cup \mathcal{T}$ to get output w .
 - 5: Compute the matrix of scores $Y = (\text{SCORE}(x_j^i, w) : i \in [M], j \in [K]) \in \mathbb{R}^{M \times K}$
 - 6: Use scores Y to run a reconstruction attack on w to obtain a vector $v = (v_1, \dots, v_M) \in ([K] \cup \{\perp\})^M$ in which the number of guesses is k (i.e., $k = \sum_i^M \mathbf{1}\{v_i \neq \perp\}$)
 - 7: (i.e., v is a vector guessing the index of the canary in each set that was used in training. \perp indicates an abstention.)
 - 8: **return** The vector v
-

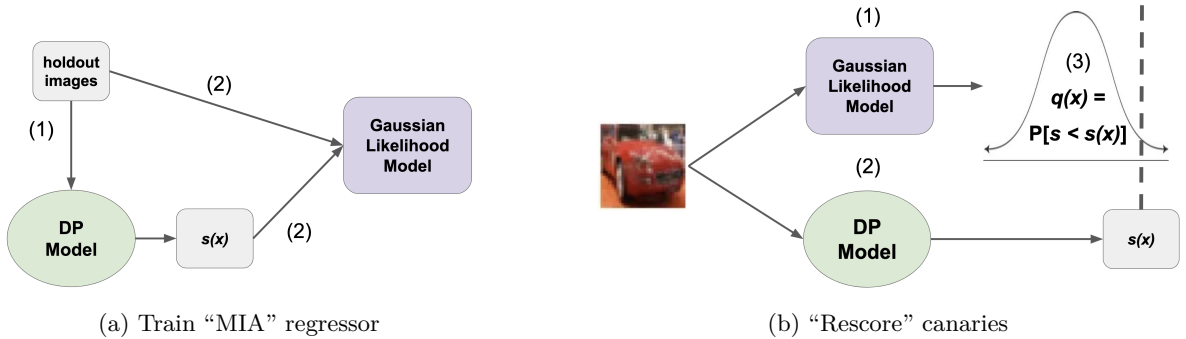


Figure 1: We provide a high-level diagram describing our quantile-regression based MIA approach to auditing. In the first part **(a)**, we **(1)** calculate the score $s(x)$ (e.g., loss) using the privately trained model and **(2)**, train the Gaussian likelihood model on the images x themselves and $s(x)$. 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, $q(x) = P(s < s(x) \mid \mu, \sigma)$.

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

3 Applying (Efficient) MIA to Privacy Auditing

Membership inference often requires the attacker to train several shadow models on a random subsets of data (Carlini et al., 2022). This approach, while effective, requires high computational demands that do not align with the goals of one-run auditing. 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 sample-specific threshold for determining membership that outperforms marginal thresholds, which are equivalent to the sort and rank (by loss) procedure employed in Steinke et al. (2023) and Mahloujifar et al. (2024).

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

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

Definition 3.1 (Quantile Regressor). Given a target false positive rate α , a quantile regressor is a model $q : \mathcal{X} \rightarrow \mathbb{R}$ 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 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).

In our proposed method (See Figure 1), we also adapt this approach and train a neural network to output a Gaussian distribution for each canary image. However, rather than using as a threshold the q -quantile for some predetermined value of q (Bertran et al., 2023; Tang et al., 2024), we calculate q directly (i.e., the CDF $P(s < s(x) \mid \mu, \sigma)$). We then use q as the input SCORE function for Algorithms 1 and 2.

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; Mahlouljifar et al., 2024) and train Wide ResNet models (Zagoruyko & Komodakis, 2016) using DP-SGD on the CIFAR-10 dataset (Krizhevsky et al., 2009). 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). 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.

Quantile regressor. Following Bertran et al. (2023), we use ConvNeXt (Liu et al., 2022) as our model architecture for the quantile regressor. Similarly, we use as our score function 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 function—in contrast to cross-entropy loss—follows a normal distribution, making it a natural choice for our approach.

Choice of number of guesses k . In general, empirical lower bounds on ϵ can be quite sensitive to the number guesses made (Mahlouljifar et al., 2024). However, it is unclear from both Steinke et al. (2023) and Mahlouljifar 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; Mahlouljifar et al., 2024), report the highest auditing results for each run.

Table 1: We present the empirical lower bounds estimated using baseline method and quantile regression (*ours*). ε_{or} corresponds to Steinke et al. (2023), $\varepsilon_{\text{or-fdp}}$ corresponds to Mahloujifar et al. (2024), and $\varepsilon_{\text{or-max}}$ 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}	$\varepsilon_{\text{or-fdp}}$	ε_{max}
47500	baseline	0.159	0.147	0.208
	<i>ours</i>	0.210	0.134	0.253

Table 2: We present the empirical lower bounds estimated using baseline method and quantile regression (*ours*) 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 $\varepsilon_{\text{or-max}}$ 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}	$\varepsilon_{\text{or-fdp}}$	ε_{max}	ε_{or}	$\varepsilon_{\text{or-fdp}}$	ε_{max}
5000	baseline	0.181	0.175	0.237	0.299	0.234	0.393
	<i>ours</i>	0.280	0.240	0.364	0.279	0.486	0.503
10000	baseline	0.202	0.172	0.216	0.227	0.115	0.241
	<i>ours</i>	0.201	0.339	0.364	0.341	0.217	0.356
20000	baseline	0.055	0.086	0.098	0.128	0.191	0.204
	<i>ours</i>	0.146	0.246	0.268	0.165	0.313	0.324

5 Results

All results reported in Tables 1 and 2 are averages over the maximum lower bound (with 95% confidence) over 5 different runs, each of which is conducted on a different random sample of the dataset. 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.² In this case, we take the max of ε_{or} and $\varepsilon_{\text{or-fdp}}$ for each run, which we denote as ε_{max} , and again report the average over 5 runs in Tables 1 and 2.

In Table 1, we present our results when auditing a model trained with $n = 47500$ examples where $m = 5000$ and $r = 45000$ ³. For our method, we use the remaining 10000 holdout set examples to train the quantile regression model. In Table 2, 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 achieves higher auditing results, regardless of data setting (i.e., choices of n , m , and r) and all auditing procedures (ε_{or} , $\varepsilon_{\text{or-fdp}}$, and ε_{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.

²Similarly to how Steinke et al. (2023) report the maximum over lower bounds produced by flipping and not flipping labels.

³We 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.

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 ε_{\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 2 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 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.
- 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.
- 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.
- 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.
- 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.
- Shuai Tang, Steven Wu, Sergul Aydore, Michael Kearns, and Aaron Roth. Membership inference attacks on diffusion models via quantile regression. In *International Conference on Machine Learning*, pp. 47819–47829. PMLR, 2024.
- Florian Tramèr, 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.