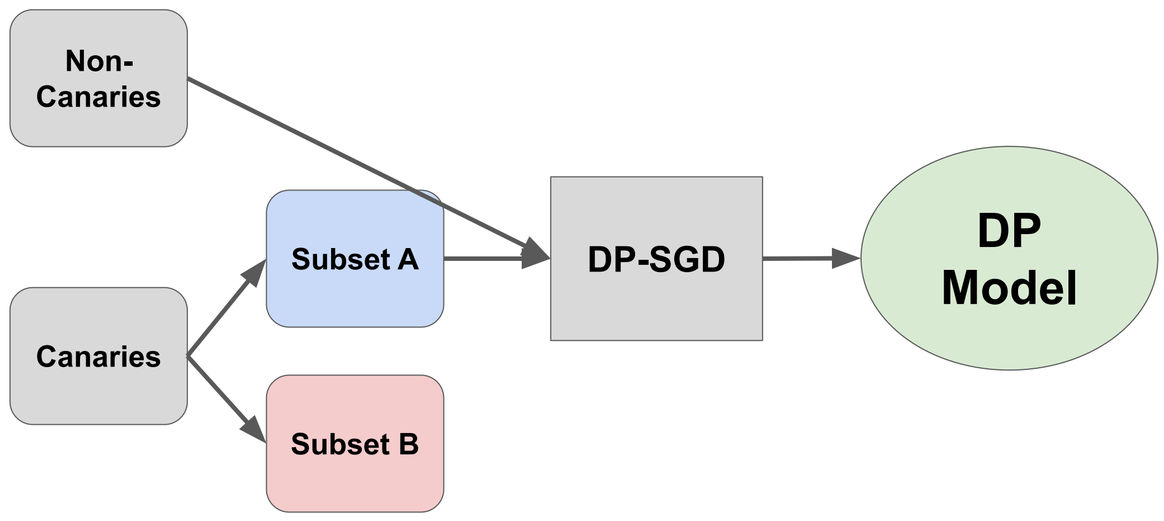
Enhancing One-run Privacy Auditing with Quantile Regression-Based Membership Inference

Terrance Liu, Matteo Boglioni, Yiwei Fu, Shengyuan Hu, Pratiksha Thaker, Zhiwei Steven Wu

Auditing Differential Privacy in One Run

Goal (DP auditing): Differentially private mechanisms (e.g., DP-SGD) provide a **theoretical upper bound** on the privacy budget ε. **Auditing** derives an **empirical lower bound**.

Steinke et al. [2023] first developed the notion of auditing in **one training run** (prior work proposed auditing procedures that required running a DP mechanism *many* times to estimate a lower bound).



Using the DP model only, **guess** whether a canary was used in training (i.e., part of subset A or B). More accurate guesses => higher empirical ε.

Membership Inference Attacks via Quantile Regression

Key Insight: the guessing game for one-run auditing is a *membership inference* problem.

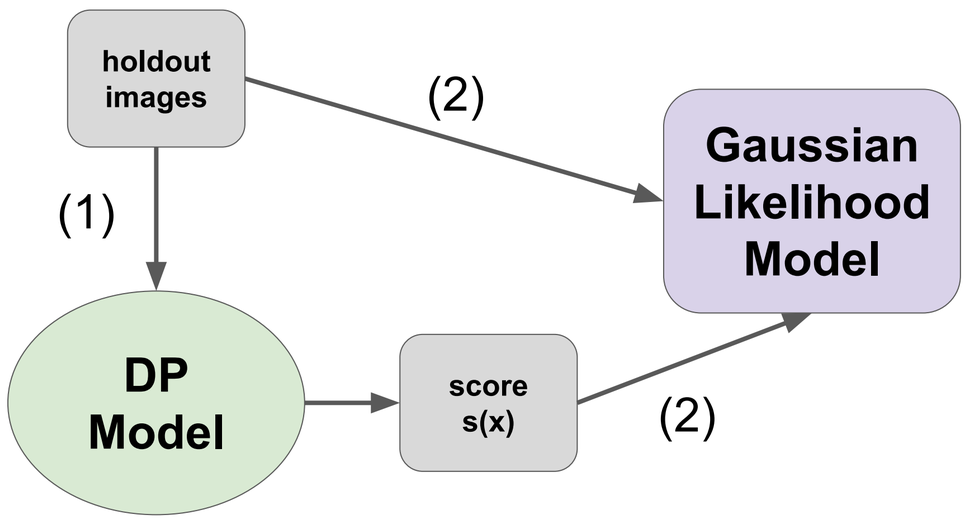
Therefore, can we leverage **efficient** membership inference attacks to improve auditing (i.e., make better guesses)?

Initial idea: shadow models are a state-of-the-art MIA approach
Problem: requires training **many** shadow models : **in conflict** with the spirit of *one-run* auditing approaches

Ours: Quantile regression based MIA methods (Bertran et. al [2023]) achieve similar performance to shadow model approaches but only require training a separate "quantile" model **once** on held-out data.

High-level overview (based on Bertran et. al [2023])

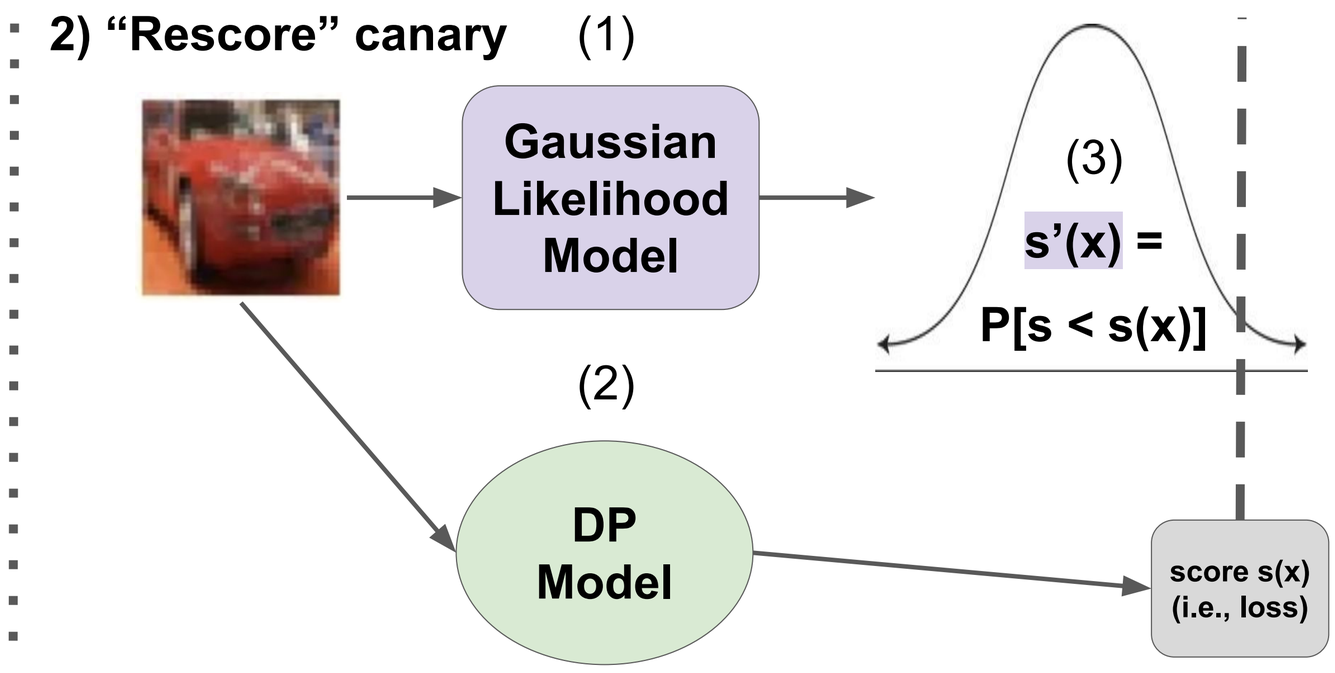
1) Train "MIA" regressor



1. For each image in the hold out set, calculate the score s(x) (e.g., loss) using the DP model
2. Train the Gaussian likelihood model, feeding in the raw image (input) and s(x) (target)

2) "Rescore" canary

use **s'(x)** as new score



1. Pass a canary image into the Gaussian likelihood model, which returns the parameters for some Gaussian distribution
2. Pass the image to the DP model to get s(x)
3. Calculate the CDF of s(x) under the predicted Gaussian distribution

Empirical Evaluation

- We conduct **black-box auditing** on a WRN-28-2 model, trained on **CIFAR-10** using DP-SGD with **ε=8**
- We evaluate two one-run auditing procedures
  - **Steinke et al. [2023]**
  - **Mahloujifar et al. [2024]**
- **Baseline** (for both approaches)
  - use the score s(x) directly to make guesses on the canaries
  - i.e., a high negative loss means the canary was seen in training
- **Ours**
  - replaces s(x) with s'(x), which calculated via the Gaussian likelihood model described above

We present our results in **Tables 1 and 2**

- ***r*** : the number of non-canary examples
- **m** : number of canaries (half of which are used in training)

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
	ours	**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
	ours	**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
	ours	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
	ours	**0.146**	**0.246**	**0.268**	**0.165**	**0.313**	**0.324**