LET'S ASK GAUSS: IMPROVED ONE-RUN PRIVACY AUDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

Privacy auditing provides an important safeguard by estimating the actual information leaked by a model, thus ensuring that theoretical privacy guarantees hold in practice. We study empirical privacy auditing for differentially private (DP) machine learning, focusing on efficient one-run methods for mechanisms such as DP-SGD. Prior one-run approaches threshold training examples or "canaries" into binary membership guesses, which discards useful information. We show that, in the white-box DP-SGD setting, canary-aligned signals naturally form a sequence of random variables whose normalized sum is asymptotically Gaussian. We show that this gaussian perspective can also be naturally extended to other composition mechanisms that satisfy a classical, triangular-array, or martingale central limit theorem. Leveraging this distributional perspective, we develop a DP-auditing framework that leads to tighter privacy lower bounds from a single training run.

1 INTRODUCTION

As data becomes increasingly valuable, a central challenge is to extract statistical utility from potentially sensitive data without compromising the privacy of the individuals who contribute to it. Differential privacy (DP) (Dwork et al., 2006; Dwork & Roth, 2014) has emerged as a broadly accepted notion for addressing this challenge and has enabled a wide range of privacy-preserving applications, including private deep learning (Abadi et al., 2016), synthetic data generation (Kurakin et al., 2024; Zhang et al., 2021; Torkzadehmahani et al., 2019; Yoon et al., 2019; McKenna et al., 2019), the release of aggregate statistics (Dwork et al., 2006; Li et al., 2015; Chan et al., 2011) etc.

The growing use of DP in machine learning and data analysis has motivated a parallel need for *privacy auditing*: methods that test whether a concrete implementation actually satisfies its claimed privacy specification. Such discrepancies often arise due to reasons ranging from simple coding errors to incorrect parameter settings. This need has motivated a growing body of work that ranges from using classical bug detection approaches to statistically estimating whether some privacy-relevant event can distinguish neighboring distributions more strongly than the claimed DP upper bound (see for example Lu et al. (2024); Askin et al. (2025); Nasr et al. (2023)).

Classical auditing approaches to private deep learning mechanisms, including DP-SGD, verify the privacy of the mechanism by running it multiple times and applying a statistical test on the (aggregate) outputs (Jagielski et al., 2020; Nasr et al., 2021). A challenge here is that DP offers worst-case guarantees, which requires finding worst case (neighboring) datasets to test the mechanism. A valuable approach suggested by Carlini et al. (2019) is to inject structured examples referred to as *canaries* to the training data and test the effect that these canaries have on the output. However, even this approach does not solve the problem of scalability: it requires repeating the learning process (with or without canaries) many times, making it computationally intensive, and often prohibitive.

Recent work has addressed whether meaningful privacy lower bounds can be obtained while avoiding the overhead of repeatedly running the training algorithms. Steinke et al. (2023) accomplished this using only one training run by embedding many independent canaries in parallel to define many neighboring dataset pairs simultaneously. The advantage of such an auditing method is evident: the cost of auditing the privacy of a model reduces to the cost of simply training the model. They demonstrate that the gradient projections of the canary examples can be thresholded into (binary) membership guesses which give a statistically valid privacy lower bound. Subsequent work has sharpened the analysis of the canary-aligned observations by using f-DP tradeoff curves, modeling canary

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

inclusion bits as messages in a noisy channel, and using order statistics to select high-confidence guesses (Mahloujifar et al., 2025; Xiang et al., 2025a;b). This one-shot paradigm has proven especially critical for federated learning, where the distributed nature of client data makes traditional, multi-run auditing impractical (Andrew et al., 2024). We defer a more detailed discussion of privacy auditing, canary-based auditors, and one-run auditing to Appendix A.

Existing one-run methods for DP-SGD share a common bottleneck: each per-canary score is ultimately collapsed to a single binary membership guess before aggregation, discarding the rich distributional information of the canaries and the underlying mechanism when designing statistical tests. While these approaches are proven to be very promising, it naturally brings forth the following questions:

> What are the distributional properties that underlie the general success of one-run DP-SGD auditing? And, can we characterize this distribution precisely, and exploit it to develop improved auditing?

To understand the distributional properties underlying these prior works, it is worth abstracting their common core: the canary-aligned observation is generated sequentially, with each training step contributing a bounded clipped signal (when the canary is sampled) plus a fresh Gaussian noise term. The final canary score accumulates these noisy observations across training, so by the Central Limit Theorem its distribution converges to a Gaussian. This intuition is already implicit in prior auditors, which use mean-like statistics on canary-aligned observations. We conjecture that this asymptotic Gaussian behavior is central to the accuracy of canary-based one-run auditing, and exploiting it directly rather than via a binary thresholding intermediate as in prior works, gives us tighter audits.

Our contributions.

A sequence-level distributional view and a Gaussian-pair auditor. We show that the normalized canary score in one-run DP-SGD auditing is a sum of sequential random variables with a Gaussian limiting law, and we give its mean and variance in closed form as functions of the DP-SGD hyperparameters. Building on this, we design a one-run white-box auditor that models the absent- and present-canary score distributions as a pair of one-dimensional Gaussians. We use the closed-form hockey-stick divergence between Gaussians (Andrew et al., 2024; Lu et al., 2023) to obtain a much tighter lower bound on the DP distance.

Quantitative convergence guarantees. We address whether the Gaussian asymptotics manifest within a practical number of training steps (T). The bound decays as $O(T^{-1/2})$ in general, and decays even faster as $O(T^{-1})$ in the small-sampling-rate regime $q = O(T^{-1/2})$ that is expected for DP-SGD. Applying this bound with our experimental parameters, we obtain a deviation three orders of magnitude smaller than the chosen value of δ.

Benchmark and Comparison. We evaluate our auditor on DP-SGD, where the sequential canary-aligned scores are asymptotically Gaussian. On CIFAR-10 DP-SGD at theoretical $\varepsilon = 8$, our auditor recovers an empirical lower bound of ≈ 6.7 (about 84% of the analytic upper bound), compared with ≈ 4.7 for the f-DP audit of (Mahloujifar et al., 2025) and ≈ 3.3 for (Steinke et al., 2023). We obtain similar 1–$2\times$ improvements across all ε regimes. We demonstrate that our auditor detects privacy violations that previous auditors miss due to significantly looser bounds. We also extend our auditor to other composition algorithms like matrix factorization (including DP-FTRL).

2 Preliminaries

2.1 Differential privacy and hockey-stick divergence

Differential privacy (DP) (Dwork et al., 2006) is the most widely accepted privacy definition for the release of (queries on) sensitive data. Let $D \sim D'$ denote neighboring databases, meaning that D and D' differ in the addition or removal of one record. A randomized mechanism $\mathcal{M}$ with output space $\mathcal{Y}$ is (ε, δ)-DP if, for every neighboring pair $D \sim D'$ and every measurable event $\mathcal{S} \subseteq \mathcal{Y}$,

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

For a fixed pair of neighboring output distributions $\mathcal{M}(D), \mathcal{M}(D')$, the smallest admissible value of δ for a given ε is well known as the hockey-stick divergence (Balle et al., 2018) between $\mathcal{M}(D)$

and $\mathcal{M}(D')$ at level e^{ε}. Formally,

$$\delta_{\mathcal{M}(D),\mathcal{M}(D')}(\varepsilon) = \sup_{\mathcal{S}\in\mathbb{R}^T} \Big(\Pr\left[\mathcal{M}(D)\in\mathcal{S}\right] - e^{\varepsilon}\Pr\left[\mathcal{M}(D')\in\mathcal{S}\right]\Big)_+,$$

where we write $(x)_+ \stackrel{\text{def}}{=} \max\{x, 0\}$, meaning $\delta(\varepsilon) \geq 0$.

In our auditor, the induced neighboring distributions are represented by a pair of one-dimensional Gaussians. For $\theta = (\mu_1, \sigma_1, \mu_2, \sigma_2)$ with $\sigma_1, \sigma_2 > 0$, it is therefore enough to compute the hockey-stick divergence

$$\delta_\theta(\varepsilon) \stackrel{\text{def}}{=} \delta_{\mathcal{N}\left(\mu_1,\sigma_1^2\right),\mathcal{N}\left(\mu_2,\sigma_2^2\right)}(\varepsilon) \tag{1}$$

The following closed form is the one-dimensional specialization of the hockey-stick divergence calculation for arbitrary pairs of Gaussians:

Lemma 1 (Hockey-stick divergence between one-dimensional Gaussians (Andrew et al., 2024; Lu et al., 2023))**.** *Let $\theta = (\mu_1, \sigma_1, \mu_2, \sigma_2)$ with $\mu_1, \mu_2 \in \mathbb{R}$ and $0 < \sigma_1^2 < \sigma_2^2$, and fix $\varepsilon \geq 0$. Let*

$$\tau \stackrel{\text{def}}{=} \frac{\sigma_1^2}{\sigma_2^2}, \qquad a \stackrel{\text{def}}{=} 1-\tau, \qquad b \stackrel{\text{def}}{=} -\frac{\sigma_1(\mu_1-\mu_2)}{\sigma_2^2},$$
$$c \stackrel{\text{def}}{=} \varepsilon + \frac{1}{2}\log\tau - \frac{(\mu_1-\mu_2)^2}{2\sigma_2^2}, \qquad \Delta \stackrel{\text{def}}{=} b^2 - 2ac, \qquad m \stackrel{\text{def}}{=} -\frac{\mu_1-\mu_2}{\sigma_1}.$$

If $\Delta \leq 0$, then $\delta_\theta(\varepsilon) = 0$. If $\Delta > 0$, define $z_\pm \stackrel{\text{def}}{=} (-b \pm \sqrt{\Delta})/a$. Then, writing Φ for the standard Gaussian CDF,

$$\delta_\theta(\varepsilon) = \Big(\Phi(z_+) - \Phi(z_-)\Big) - e^{\varepsilon}\Big(\Phi\big((z_+ - m)\sqrt{\tau}\big) - \Phi\big((z_- - m)\sqrt{\tau}\big)\Big).$$

Equivalently, given a fixed δ, we define the function $\varepsilon_\theta(\delta)$ as the smallest privacy parameter ε for which the Gaussian pair has hockey-stick divergence at most δ. Formally,

$$\varepsilon_\theta(\delta) \stackrel{\text{def}}{=} \inf\left\{\varepsilon \geq 0 : \delta_\theta(\varepsilon) \leq \delta\right\}. \tag{2}$$

2.2 DP-SGD TRANSCRIPTS AND CANARY-ALIGNED OBSERVATIONS

The seminal DP-SGD algorithm (Abadi et al., 2016) is a differentially private variant of the standard stochastic-gradient method. In each iteration, the algorithm computes per-example gradients on a sampled minibatch, clips each gradient to have ℓ_2 norm at most C, aggregates the clipped gradients, and adds Gaussian noise $Z_t \sim \mathcal{N}\left(0, \sigma^2 C^2 I_d\right)$ before updating the model parameters θ_t.

Broadly speaking, there are two general types of DP auditing for ML training mechanisms: In *black-box* auditing, the auditor may observe only the final model, or possibly query access to its predictions. In *white-box* auditing, the auditor sees the internal state of training; we focus on the standard instantiation in which the auditor observes the noisy averaged gradient $\widetilde{g}_t^{(b)}$ at each step, consistent with the privacy analysis of Abadi et al. (2016). We use the bit $b \in \{0,1\}$ to indicate whether a target canary $x^\star$ is absent ($b = 0$) or present ($b = 1$) in the training dataset, following Nasr et al. (2023), and write $\mathsf{Tr}_T^{(b)} = (\widetilde{g}_t^{(b)}, \theta_t^{(b)})_{t\in[T]}$ for the resulting white-box transcript. The full procedure is formalized as Algorithm 2 in Appendix B, identical to transcript in Steinke et al. (2023).

Definition 1 formally defines the sequential observation used by the white-box audit. Intuitively, the auditor projects the privatized gradient at each iteration onto the direction of the canary gradient. Thus, the observation in our analysis is not the model iterate itself, but a one-dimensional canary-aligned projection of the noisy gradient update.

Definition 1 (Canary-aligned sequential observation)**.** (Following prior works (Steinke et al., 2023; Mahloujifar et al., 2025)) Let $b \in \{0,1\}$, and let $\mathsf{Tr}_T^{(b)} = \left(\widetilde{g}_t^{(b)}, \theta_t^{(b)}\right)_{t\in[T]}$ be the white-box DP-SGD transcript from Algorithm 2. Let $\bar{u}$ be a deterministic unit vector, and for each iteration $t \in [T]$, define

$$c_t^{(b)} \stackrel{\text{def}}{=} \mathsf{clip}_C\left(\nabla_\theta \ell(\theta_{t-1}^{(b)}; x^\star)\right), \qquad u_t^{(b)} \stackrel{\text{def}}{=} \begin{cases} c_t^{(b)} / \left\|c_t^{(b)}\right\|_2 & \text{if } c_t^{(b)} \neq 0, \\ \bar{u} & \text{otherwise.} \end{cases}$$

The *canary-aligned sequential observation* induced by $\mathsf{Tr}_T^{(b)}$ and $x^\star$ is

$$X_{1:T}^{(b)} \stackrel{\text{def}}{=} \left(X_1^{(b)}, \ldots, X_T^{(b)}\right) \qquad \text{where,} \quad X_t^{(b)} \stackrel{\text{def}}{=} \langle u_t^{(b)}, \widetilde{g}_t^{(b)} \rangle$$

3 METHODOLOGY

In this section, we present our one-run white-box auditing procedure. Given a fixed privacy parameter δ, the auditor outputs an empirical lower bound on the corresponding privacy parameter ε for a DP-SGD implementation and can be extended to other mechanisms. At a high level, the procedure has three steps. First, the auditor runs white-box DP-SGD once with many independently inserted canaries, following the transcript model in Algorithm 2. Each canary induces a sequential observation $X_{1:T}^{(b)}$ from the resulting transcript, as defined in Definition 1. Second, for each canary, the auditor aggregates and normalizes its sequential observation into a single scalar audit score $S_T^{(b)}$. Across many canaries, these scores form two empirical samples: one from the absent-canary world and one from the present-canary world. Third, the auditor models these two samples as independent draws from two Gaussian distributions, and the hockey-stick divergence between these two Gaussians yields the empirical lower bound on ε.

Definition 2 introduces the scalar score produced by each canary. Recall that Definition 1 associates each canary with a sequential observation $X_{1:T}^{(b)}$, where $X_t^{(b)}$ is the projection of the noisy gradient at iteration t onto the canary's direction. The auditor compresses this sequence into a single normalized score by summing the projected observations over the training trajectory and scaling by $1/\sqrt{T}$.

Definition 2 (Canary score)**.** Let $X_{1:T}^{(b)}$ be as in Definition 1. The *canary score* induced by this sequential observation is

$$S_T^{(b)} \stackrel{\text{def}}{=} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} X_t^{(b)}.$$

The per-iteration observation $X_t^{(b)}$ admits a natural signal-plus-noise interpretation. Intuitively, when the canary is absent, the projected observations X_t^0 contain only background contributions and projected DP noise. When the canary is present, X_t^1 additionally contains a clipped signal if the canary is sampled. Thus, the canary score X_t^b is expected to shift between the neighboring worlds $b = 0$ and $b = 1$ with an amount of Bernoulli-sampled canary contributions.

This motivates a Gaussian model for the normalized canary score $S_T^{(b)} = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} X_t^{(b)}$. In the world where the canary is absent, the score is dominated by sum of independent DP noise. In the world where the canary is present, the score additionally accumulates Bernoulli-sampled canary contributions. By the Central Limit Theorem, the normalized sum of many i.i.d. finite-variance observations is well approximated by a one-dimensional Gaussian. We therefore approximate the canary-score distributions under the two neighboring worlds by a Gaussian pair

$$G_0 \stackrel{\text{def}}{=} \mathcal{N}\left(\mu_0, \sigma_0^2\right), \qquad G_1 \stackrel{\text{def}}{=} \mathcal{N}\left(\mu_1, \sigma_1^2\right),$$

where G_0 models the distribution of $S_T^{(0)}$ and G_1 models the distribution of $S_T^{(1)}$.

We next present our modeling choice for the canary-score distribution $S_T^{(b)}$ explicitly. This model isolates the two dominant factors: the projected Gaussian DP noise (present in both worlds) and the clipped canary contribution (present only when the canary is sampled).

Model 1 (Canary-observation model)**.** For each $t \in [T]$, the absent- and present-canary observations take the form

$$X_t^{(0)} = Z_t^{(0)}, \qquad X_t^{(1)} = B_t C + Z_t^{(1)}, \qquad t \in [T],$$

where $B_t \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(q)$ and $Z_t^{(0)}, Z_t^{(1)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ are mutually independent.

Under Model 1, the absent-canary score is exactly Gaussian, while the present-canary score is a normalized sum of independent Bernoulli-Gaussian random variables and is therefore asymptotically Gaussian. Proposition 1 gives the corresponding mean and variance of the Gaussian surrogate in terms of the DP-SGD parameters C, T, q, and τ^2.

Proposition 1. *Let $S_T^{(0)}$ and $S_T^{(1)}$ be the canary scores (see Definition 2). Then under Model 1,*

$$S_T^{(0)} \sim \mathcal{N}(0, \sigma^2), \qquad \mathbb{E}[S_T^{(1)}] = \sqrt{T} qC, \qquad \mathrm{Var}(S_T^{(1)}) = \sigma^2 + q(1-q)C^2$$

Moreover, $S_T^{(1)}$ asymptotically follows the distribution $\mathcal{N}\left(\sqrt{T}\, qC, \sigma^2 + q(1-q)C^2\right)$ as $T \to \infty$.

From adaptive DP-SGD trajectory to Model 1. In a real DP-SGD trajectory, canary magnitude, and canary direction can depend on the training history. Appendix C gives the corresponding martingale decomposition and establishes the following reduction.

Proposition 2 (Reduction of adaptive scores to Model 1)**.** *Consider the adaptive DP-SGD score process of Appendix C. Suppose that (i) background gradients have zero projection onto the canary direction at every step, and (ii) the present-canary clipped gradient has norm C at every step. Then the adaptive score process has exactly the formulation of Model 1.*

A fixed-coordinate Dirac gradient canary enforces exact norm saturation C by construction, and we show that background suppression holds empirically in Section 4.4.1. The formal statements and proofs, including the adaptive martingale decomposition, are deferred to Appendices C and D.

Algorithm 1 One-run white-box DP-SGD auditor (under Model 1)

Input: Privacy parameter δ; absent-canary scores $\mathcal{S}_0 = \{s_i^{(0)}\}_{i=1}^{m_0}$; present-canary scores $\mathcal{S}_1 = \{s_j^{(1)}\}_{j=1}^{m_1}$; parameter $m = m_1 + m_0$; confidence level α.
Output: Empirical lower bound $\widehat{\varepsilon}_{\mathrm{lb}}$ for privacy parameter ε.
1: Using $\mathcal{S}_0$ and $\mathcal{S}_1$, construct a $(1-\alpha)$-confidence region

$$C_\alpha \subseteq \Theta \qquad \text{for} \qquad \theta^\star = (\mu_0, \sigma_0, \mu_1, \sigma_1),$$

where $\mathcal{N}\left(\mu_0, \sigma_0^2\right)$ and $\mathcal{N}\left(\mu_1, \sigma_1^2\right)$ are the absent- and present-canary score distributions.
2: Compute $\widehat{\varepsilon}_{\mathrm{lb}}(\alpha) \leftarrow \inf_{\theta \in C_\alpha} \varepsilon_\theta(\delta)$, where $\varepsilon_\theta(\delta)$ is defined in Equation (2).
3: **return** $\widehat{\varepsilon}_{\mathrm{lb}}(\alpha)$.

The above discussion reduces the one-run white-box auditing problem to estimating the parameters of a one-dimensional Gaussian pair. Algorithm 1 presents our one-run white-box DP-SGD auditor based on this reduction. Given the absent- and present-canary score samples, the auditor first constructs a $(1-\alpha)$-confidence region C_α for the unknown Gaussian parameters $\theta^\star = (\mu_0, \sigma_0, \mu_1, \sigma_1)$. It then evaluates the most conservative privacy lower bound over all Gaussian pairs in this confidence region by computing $\widehat{\varepsilon}_{\mathrm{lb}}(\alpha) \overset{\mathrm{def}}{=} \inf_{\theta \in C_\alpha} \varepsilon_\theta(\delta)$. Consequently, the auditor returns $\widehat{\varepsilon}_{\mathrm{lb}}(\alpha)$ as a confidence-adjusted lower bound of the true privacy parameter ε under Model 1.

3.1 Convergence Rate of the Gaussian Approximation

The methodology in Section 3 models the canary score $S_T^{(b)}$ as a one-dimensional Gaussian under both the absent-canary and present-canary worlds. Under the idealized scalar observation, the absent-canary score $S_T^{(0)} \sim \mathcal{N}\left(0, \sigma^2\right)$ is exactly Gaussian for every T. The present-canary score $S_T^{(1)}$ however, is only asymptotically Gaussian: each summand $X_t^{(1)} = B_t C + Z_t$ mixes a Bernoulli signal with Gaussian noise. In Appendix E we present a general theorem that captures the exact rate of convergence for normalized i.i.d. sum of Gaussian-mixture random variables. In this subsection we specifically apply it to the DP auditing setting to compute the rate of convergence of $S_T^{(1)}$.

In the DP auditing setting, the per-step audit observation is i.i.d.: $X_t^{(1)} = B_t C + Z_t$ with $B_t \overset{\text{i.i.d.}}{\sim} \mathrm{Bernoulli}(q)$ and $Z_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma^2\right)$. Centering by the mean, $\widetilde{X}_t \overset{\mathrm{def}}{=} X_t^{(1)} - qC$, we can write

$$\widetilde{X}_t \sim \mathrm{GM}(1-q,\, q,\, -qC,\, (1-q)C,\, \sigma)$$

where $\mathrm{GM}(u,\, v,\, \mu_1,\, \mu_2,\, \sigma)$ represents a Gaussian mixture distribution with weights u and v (where $u + v = 1$) and the two Gaussians have same standard deviation σ and different means μ_1 and μ_2

respectively. We can further define:

$$\widetilde{S}_T \stackrel{\text{def}}{=} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \widetilde{X}_t = S_T^{(1)} - \sqrt{T}\, qC$$

For the above normalized sum of Gaussian mixture distribution we derive the exact cumulants and hence its rate of convergence in Appendix E, giving us the following result:

Theorem 1 (Closed-form Kolmogorov-distance bound for the DP audit statistic)**.** *Consider the canary-observation model 1 of Algorithm 1 with T training steps, sampling rate $q \in (0, 1)$, clipping norm $C > 0$, and per-step noise standard deviation $\sigma > 0$. Let $\widetilde{Q}_T$ denote the true law of the centered audit statistic $\widetilde{S}_T$ and $Q_T \stackrel{\text{def}}{=} \mathcal{N}\left(0, \sigma^2 + q(1-q)C^2\right)$ denote its Gaussian model. Then*

$$d_{\mathrm{K}}\Big(\widetilde{Q}_T,\, Q_T\Big) = \frac{q(1-q)\,|1-2q|\,C^3}{6\sqrt{2\pi\,T}\left[\sigma^2 + q(1-q)C^2\right]^{3/2}} + O(1/T).$$

While in general the audit statistic distribution converges to a Gaussian at rate $O(1/\sqrt{T})$, for asymptotically small sampling rate (e.g. $q = O(1/\sqrt{T})$), as is the case for DP-SGD, the distributions converge even faster (rate at least $O(1/T)$).

Concrete estimates. To assess the accuracy of the Gaussian approximation in our experiments, Appendix E derives a tail-localized refinement of Theorem 1 (see Corollary 2). For our experimental parameters $T = 2500$, $q = 0.0819$, $C = 1$, $\sigma = 2.6245$, the leading-order CDF-deviation estimate is 2.29×10^{-8}, which is three orders of magnitude smaller than our audit parameter $\delta = 10^{-5}$. These estimates give additional confidence in the Gaussian approximation for the empirical privacy lower bound in our experimental setup.

3.2 Gaussian parameter estimation and confidence regions

The convergence rate study in Section 3.1 shows the Gaussian approximation closely fits the finite-T present-canary score distribution in the DP-SGD settings. In this section, we address a separate source of error: the finite number of canary scores used to estimate the Gaussian parameters.

Proposition 3 states that our auditor's output (Algorithm 1) serves as a conservative lower bound of privacy parameter ε with high probability under the Gaussian score model.

Proposition 3 (Confidence guarantee for Algorithm 1)**.** *Let $\widehat{\varepsilon}_{\mathrm{lb}}(\alpha)$ be the output of Algorithm 1. Then*

$$\Pr\left[\widehat{\varepsilon}_{\mathrm{lb}}(\alpha) \le \varepsilon^\star\right] \ge 1 - \alpha.$$

This directly follows from the fact that on event $\{\theta^\star \in C_\alpha\}$, the output of Algorithm 1 satisfies $\widehat{\varepsilon}_{\mathrm{lb}}(\alpha) = \inf_{\theta \in C_\alpha} \varepsilon_\theta(\delta) \le \varepsilon_{\theta^\star}(\delta) = \varepsilon^\star$.

Hypothesis test. Equivalently, Algorithm 1 tests $H_0 : \varepsilon \le \varepsilon_{\mathrm{claim}}$ vs. $H_1 : \varepsilon > \varepsilon_{\mathrm{claim}}$, rejecting H_0 when $\widehat{\varepsilon}_{\mathrm{lb}}(\alpha) > \varepsilon_{\mathrm{claim}}$. Since $\varepsilon^\star \le \varepsilon_{\mathrm{claim}}$ under H_0, Proposition 3 gives $\Pr\left[\widehat{\varepsilon}_{\mathrm{lb}}(\alpha) > \varepsilon_{\mathrm{claim}}\right] \le \alpha$, and so the test has level α.

Constructing C_α. In our implementation, C_α can be instantiated using standard confidence-region constructions for Gaussian parameters. A simple finite-sample choice is the Bonferroni rectangle obtained from the usual Student-t confidence intervals for the means and chi-squared confidence intervals for the variances, applied separately to the absent- and present-canary score samples (Casella & Berger, 2002). This construction is valid under the Gaussian score model, but can be conservative because it treats the four coordinates of $\theta^\star$ separately.

As a less conservative alternative, one can use a bootstrap ellipsoid based on the joint covariance of the estimator $\widehat{\theta} = (\widehat{\mu}_0, \widehat{\sigma}_0, \widehat{\mu}_1, \widehat{\sigma}_1)$, where the covariance is estimated from bootstrap resamples of the canary scores (Efron & Tibshirani, 1994). This ellipsoidal region captures correlations among the estimated Gaussian parameters and has asymptotic coverage under standard bootstrap regularity conditions.

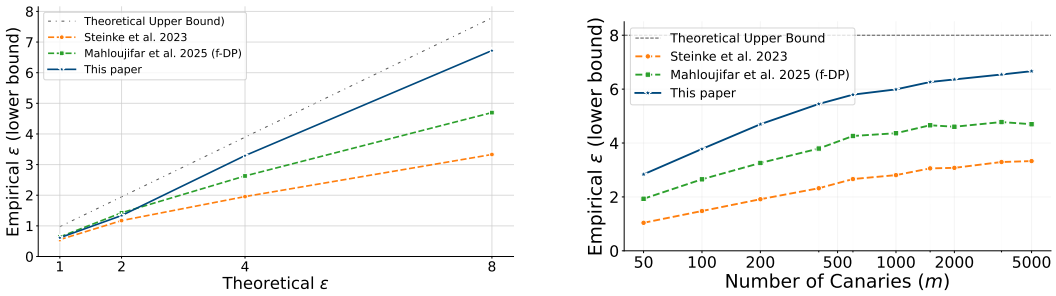


Figure 1: DP-SGD auditing results: (a) Left: shows empirical lower bounds across different theoretical ϵ, and (b) Right: shows the impact of canary count m on auditing tightness compared to Steinke et al. (2023) and Mahloujifar et al. (2025). All lower bounds are computed at a 95% confidence.

4 Experiments and Results

We empirically validate our auditing method on composition DP mechanisms where the audit statistic converges to a Gaussian distribution. We first audit DP-SGD with Poisson subsampling and analytic-Gaussian accounting, showing that our distributional perspective substantially improves auditing power and enables detection of privacy violations that previous auditors miss. We then apply our method to DP-FTRL and DP-MF-FTRL, which use correlated noise. We ran our experiments on a shared HPC cluster utilizing NVIDIA H200-100GB GPUs. Our code is available here.

4.1 Auditing for DP-SGD

To verify the tightness of our estimated lower bounds, we follow the white-box auditing protocol established in Steinke et al. (2023) and Mahloujifar et al. (2025). We train a WideResNet-16-4 on CIFAR-10 (Batch size $B = 4096$, Augmentation $K = 16$, $T = 2500$ steps). At a theoretical $\varepsilon = 8$, the model achieves $\approx 70\%$ accuracy with a negligible utility drop ($\approx 1\%$) upon the inclusion of 5,000 canaries. We follow the auditing procedure described in Algorithm 1. We evaluate all three methods at a 95% confidence level. While baselines suffer Bonferroni penalties from threshold searching, our method avoids thresholding entirely by extracting worst-case parameters from a continuous 95% bootstrap ellipsoid (See Sec 3.2). As illustrated in Figure 1, our method yields significantly tighter lower bounds than existing one-run baselines (by $1-2\times$) across all ε regimes. Our method narrows the gap to the analytic upper bound, capturing 63%–84% of the theoretical limit. Crucially, these gains are achieved without modifying the underlying training or auditing pipeline, simply by leveraging the distributional information inherent in the sequential gradient observations.

We also vary the number of canaries, starting at $m = 50$. As shown in Figure 1, all three methods are relatively stable beyond $m = 1000$, but the baselines converge to significantly looser bounds. Our method recovers a larger fraction of the theoretical privacy budget, obtaining tighter bounds with only 100 canaries than competing methods achieve with nearly 2500. Appendix G.1 also extends our auditor to DP-SGD with adaptive clipping.

Table 1: Detecting privacy violations incurred due to DP-SGD implementation errors

Implementation errors	True ε	Steinke et al. (2023)	Mahloujifar et al. (2025)	Ours
Correct implementation	8.00	3.330	4.697	6.651
Noise scaled for clipping bound $C/2$, trained with C	21.98	5.542	**9.404**	**15.835**
Training steps doubled, noise not recalibrated	12.12	4.315	6.654	**9.778**
Sampling rate $2\times$ the accounted rate	18.88	5.528	**9.365**	**16.162**

4.2 Detecting privacy violations for implementation errors

We evaluate our auditor and the baselines on three synthetic implementation errors in DP-SGD: a clipping-sensitivity mismatch, incorrect noise calibration, and a sub-sampling rate mismatch. These errors are modeled after documented failures in practice (Tramer et al., 2022; Cebere et al., 2026; Chua et al., 2024).

The errors substantially increase the true privacy loss, from the claimed $\epsilon = 8$ to $12.12 - 21.98$ as shown in Table 1. Our tighter lower bound exceeds the claimed bound for all three tested errors and therefore detects each violation in this experiment. Steinke et al. (2023) detects none, while the auditor of Mahloujifar et al. (2025) detects two of the three. These results demonstrate that common calibration and accounting errors can change the parameters of the score distributions without destroying their Gaussian shape, allowing our auditor to audit them with greater power than existing auditors.

4.3 Auditing for applications beyond DP-SGD

Our framework is not tied to DP-SGD as a specific algorithm. It applies whenever the auditor can construct an audit statistic whose distributions under the neighboring worlds are well approximated by a pair of Gaussians. Matrix-factorization (MF) mechanisms, which include DP-FTRL, can be audited with Gaussian statistics.

Let $x \in \mathbb{R}^{T\times d}$ stack the per-round sums of clipped gradients and let A be the prefix-sum workload. An MF mechanism factors $A = \mathbf{BC}$, releases $y = \mathbf{C}x + Z$ with i.i.d. $\mathcal{N}(0, \sigma^2)$ entries in Z, and decodes $\mathbf{B}y = Ax + \mathbf{B}Z$.

Definition 3 (Canary score for MF)**.** Let canary i be a Dirac gradient canary on coordinate c_i, inserted at round t_i. We use the likelihood-ratio score $S_i = \langle \mathbf{C}_{:,t_i},\, y_{:,c_i}\rangle / \|\mathbf{C}_{:,t_i}\|_2$.

Under Assumption 1 (background suppression), coordinate c_i of y is $\mathcal{N}(0, \sigma^2 I)$ when the canary is absent and $\mathcal{N}(C\,\mathbf{C}_{:,t_i}, \sigma^2 I)$ when it is present. Hence, the Gaussian convergence is exact with no approximation errors.

The tree-based DP-FTRL of Kairouz et al. (2021) is the special case in which $\mathbf{C}$ is the binary-tree encoder, whereas for DP-MF-FTRL the encoder comes from a factorization $\mathbf{A} = \mathbf{BC}$ of the prefix-sum workload. In Figures 2 and 3, we compare our auditor against the one-shot auditing method of Andrew et al. (2024). While their approach also utilizes a closed form (ϵ, δ) estimator using Gaussian parameters, it relies on a maximum-score statistic over training steps. In contrast, our framework leverages the entire joint distribution of sequential observations. Details of the audit statistics, gaussian pairs and experimental setup are in Appendix F.

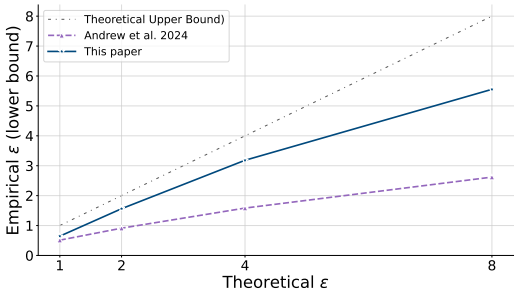


Figure 2: Auditing results for DP-FTRL across multiple ε values compared to Andrew et al. (2024)

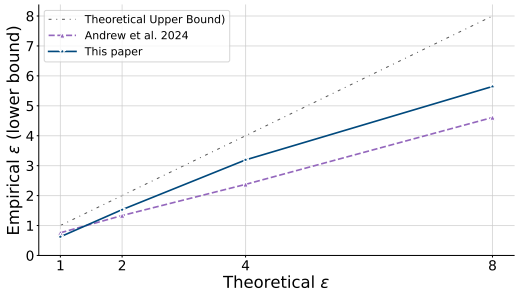


Figure 3: Auditing results for DP-MF-FTRL across multiple ε values compared to Andrew et al. (2024)

4.4 Ablations

We run additional ablations to test our assumptions and robustness across the tested hyperparameter ranges, and the observed behavior agrees with the dependence predicted by Proposition 1. We add detailed experiments varying the clipping bound, batch size and learning rate in Appendix G.

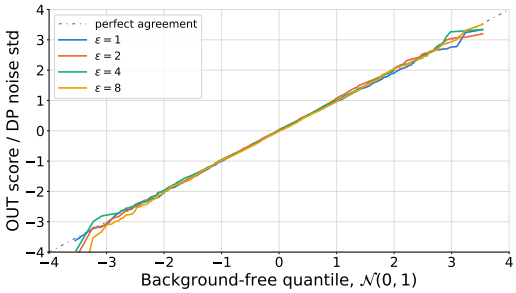


Figure 4: Q-Q plot of real canary scores against the background-free prediction.

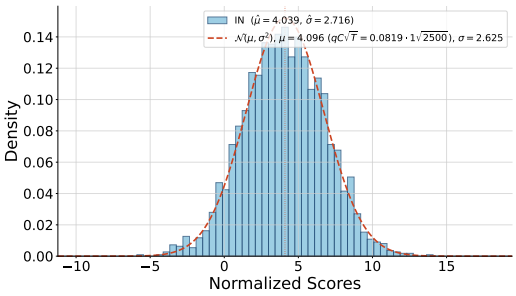


Figure 5: Histogram to visualize Gaussianity of canary-present scores against analytical parameters.

4.4.1 Background suppression

Our Model 1 assumes each canary score is only DP noise plus the canary's own signal, with no contribution from the real training records at the canary coordinate. To test this, we compare the real audit statistic from a DP-SGD run against 500 background-free simulations. The simulations use the same (σ, C, T, q) and the same presence split. Figure 4 compares the scores from the absent canary world, divided by the DP noise standard deviation σC, with the background-free prediction $\mathcal{N}(0, 1)$. At every ε the quantiles lie on the line of perfect agreement; the only deviations are in the extreme tails, where the number of samples is low in both directions. This indicates that any background contribution is therefore below sampling noise and does not affect the audit. We also add a hexbin density against canary index plot in Figure 7 of Appendix G.

4.4.2 Convergence and Geometry in DP-SGD

Gaussian Approximation In Figure 5, we see that the histogram of distribution of estimated canary-in scores compared against the analytical Gaussian parameters is very close. We also evaluate the stability of the bound across training steps $T \in \{50, 100, \ldots, 2500\}$ in Figure 6a of Appendix G. The empirical bound remains robust even at low step counts, confirming that the Gaussian convergence of the gradient sums happens rapidly in the DP-SGD regime.

Confidence Region Geometry We compare the Parametric Bonferroni and Bootstrap Ellipsoid confidence regions. As shown in the Appendix G, the Ellipsoid consistently recovers higher ε values by capturing the cross-correlation between sample means and variances, which the axis-aligned Bonferroni rectangle ignores.

5 Discussion

We presented a one-run, white-box auditor that departs from the binary-thresholding paradigm of prior work and instead it uses the full distributional information in canary-aligned observations. Our key insight is that for DP-SGD, after normalization the canary score distribution is asymptotically Gaussian, which lets us convert these scores directly into a privacy lower bound via the hockey-stick divergence between a Gaussian pair. We also extend to other composition algorithms and our experiments show $1 - 2\times$ tighter lower bounds than existing one-run baselines.

Our auditing framework uses a distributional perspective on the canary scores, whereas threshold-based one-run audits (Steinke et al., 2023; Mahloujifar et al., 2025) are distribution-free. As noted by Steinke et al. (2023), exploiting the structure of a specific mechanism can yield stronger auditing results. We show that our gaussian perspective improves auditing power significantly. We hope this distributional perspective serves as a useful building block for future auditing schemes in settings beyond those discussed here, leading to meaningful ways to detect privacy violations in practice.

AI USE STATEMENT

In this work, we used generative AI tools for implementing our method and polishing the writing of our research paper.

We have not used generative AI tools for developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments or interpreting results.

The tasks of generating synthetic datasets, assisting with translation, cleaning and reformatting datasets, supporting qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for editing software code. We have reviewed all AI-assisted work. LLM-generated code and edited paper write-up were verified and tested for correctness by 2 authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The work focuses on designing privacy auditing tools that strengthen verification of differential privacy guarantees in ML systems. It uses only the publicly released CIFAR-10 dataset, does not involve human subjects or crowdsourcing, does not release any model that poses misuse risk, and preserves authorship anonymity in this submission.

REPRODUCIBILITY STATEMENT

We provide extensive details about our implementation in the Experiments Section and Appendix. We also provide an anonymized link to our code.

REFERENCES

Martín Abadi, Andy Chu, Ian J. 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. ACM, 2016. doi: 10.1145/2976749.2978318. URL https://doi.org/10.1145/2976749.2978318.

Mathieu Andreux, Tomás Angles, Georgios Exarchakis, Roberto Leonarduzzi, Gaspar Rochette, Louis Thiry, John Zarka, Stéphane Mallat, Joakim Andén, Eugene Belilovsky, et al. Kymatio: Scattering transforms in python. *Journal of Machine Learning Research*, 21(60):1–6, 2020.

Galen Andrew, Om Thakkar, Hugh Brendan McMahan, and Swaroop Ramaswamy. Differentially private learning with adaptive clipping. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=RUQ1zwZR8_.

Galen Andrew, Peter Kairouz, Sewoong Oh, Alina Oprea, H. Brendan McMahan, and Vinith M. Suriyakumar. One-shot empirical privacy estimation for federated learning. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Meenatchi Sundaram Muthu Selva Annamalai and Emiliano De Cristofaro. Nearly tight black-box auditing of differentially private machine learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Önder Askin, Holger Dette, Martin Dunsche, Tim Kutta, Yun Lu, Yu Wei, and Vassilis Zikas. General-purpose f-DP estimation and auditing in a black-box setting. *CoRR*, abs/2502.07066, 2025. doi: 10.48550/arXiv.2502.07066. URL https://arxiv.org/abs/2502.07066.

Borja Balle and Yu-Xiang Wang. Improving the Gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 394–403. PMLR, 2018. URL https://proceedings.mlr.press/v80/balle18a.html.

Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. In *Advances in Neural Information Processing Systems*, volume 31, pp. 6280–6290, 2018.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)*, pp. 267–284, 2019.

George Casella and Roger L. Berger. *Statistical Inference*. Duxbury, 2002.

Tudor Cebere, David Erb, Damien Desfontaines, Aurélien Bellet, and Jack Fitzsimons. Privacy in theory, bugs in practice: Grey-box auditing of differential privacy libraries. *arXiv preprint arXiv:2602.17454*, 2026.

T.-H. Hubert Chan, Elaine Shi, and Dawn Song. Private and continual release of statistics. *ACM Transactions on Information and System Security*, 14(3):1–24, 2011. doi: 10.1145/2043621.2043626. URL https://doi.org/10.1145/2043621.2043626.

Lynn Chua, Badih Ghazi, Pritish Kamath, Ravi Kumar, Pasin Manurangsi, Amer Sinha, and Chiyuan Zhang. How private are DP-SGD implementations? In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8904–8918. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/chua24a.html.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 2014.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, volume 3876 of *Lecture Notes in Computer Science*, pp. 265–284. Springer, 2006. doi: 10.1007/11681878_14. URL https://doi.org/10.1007/11681878_14.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1994.

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.

Peter Kairouz, Brendan McMahan, Shuang Song, Om Thakkar, Abhradeep Thakurta, and Zheng Xu. Practical and private (deep) learning without sampling or shuffling. In *International Conference on Machine Learning*, pp. 5213–5225. PMLR, 2021.

Amit Keinan, Moshe Shenfeld, and Katrina Ligett. How well can differential privacy be audited in one run? *arXiv preprint arXiv:2503.07199*, 2025.

Alexey Kurakin, Natalia Ponomareva, Umar Syed, Liam MacDermed, and Andreas Terzis. Harnessing large-language models to generate private synthetic text, 2024. URL https://arxiv.org/abs/2306.01684.

Chao Li, Gerome Miklau, Michael Hay, Andrew McGregor, and Vibhor Rastogi. The matrix mechanism: Optimizing linear counting queries under differential privacy. *The VLDB Journal*, 24 (6):757–781, 2015. doi: 10.1007/s00778-015-0391-1. URL https://doi.org/10.1007/s00778-015-0391-1.

Terrance Liu, Matteo Boglioni, Yiwei Fu, Shengyuan Hu, Pratiksha Thaker, and Zhiwei Steven Wu. Enhancing one-run privacy auditing with quantile regression-based membership inference. *arXiv preprint arXiv:2506.15349*, 2025.

Yun Lu, Malik Magdon-Ismail, Yu Wei, and Vassilis Zikas. The normal distributions indistinguishability spectrum and its application to privacy-preserving machine learning. *arXiv preprint arXiv:2309.01243*, 2023. Preliminary version to appear at the Theory and Practice of Differential Privacy (TPDP) 2026 workshop.

Yun Lu, Malik Magdon-Ismail, Yu Wei, and Vassilis Zikas. Eureka: A general framework for black-box differential privacy estimators. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 913–931. IEEE, 2024.

Samuel Maddock, Alexandre Sablayrolles, and Pierre Stock. CANIFE: Crafting canaries for empirical privacy measurement in federated learning. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Saeed Mahloujifar, Luca Melis, and Kamalika Chaudhuri. Auditing f-differential privacy in one run. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 42615–42641. PMLR, 2025. URL https://proceedings.mlr.press/v267/mahloujifar25a.html.

Ryan McKenna, Daniel Sheldon, and Gerome Miklau. Graphical-model based estimation and inference for differential privacy. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4435–4444. PMLR, 2019. URL https://proceedings.mlr.press/v97/mckenna19a.html.

Milad Nasr, Shuang Songi, Abhradeep Thakurta, Nicolas Papernot, and Nicholas Carlin. Adversary instantiation: Lower bounds for differentially private machine learning. In *2021 IEEE Symposium on security and privacy (SP)*, pp. 866–882. IEEE, 2021.

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. USENIX Association, 2023. URL https://www.usenix.org/conference/usenixsecurity23/presentation/nasr.

Valentin V Petrov. Classical-type limit theorems for sums of independent random variables. In *Limit theorems of probability theory*, pp. 1–24. Springer, 2000.

Valentin V Petrov. *Sums of independent random variables*. Springer Science & Business Media, 2012.

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 *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/9a6f6e0d6781d1cb8689192408946d73-Abstract-Conference.html.

Reihaneh Torkzadehmahani, Peter Kairouz, and Benedict Paten. DP-CGAN: Differentially private synthetic data and label generation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 98–104. IEEE Computer Society, 2019. doi: 10.1109/CVPRW.2019.00018. URL https://doi.org/10.1109/CVPRW.2019.00018.

Florian Tramer, Andreas Terzis, Thomas Steinke, Shuang Song, Matthew Jagielski, and Nicholas Carlini. Debugging differential privacy: A case study for privacy auditing. *arXiv preprint arXiv:2202.12219*, 2022.

Zihang Xiang, Tianhao Wang, and Di Wang. Privacy audit as bits transmission: (Im)possibilities for audit by one run. In *34th USENIX Security Symposium (USENIX Security 25)*. USENIX Association, 2025a. URL https://www.usenix.org/conference/usenixsecurity25/presentation/xiang-zihang.

Zihang Xiang, Tianhao Wang, Hanshen Xiao, Yuan Tian, and Di Wang. Tight privacy audit in one run. *CoRR*, abs/2509.08704, 2025b. doi: 10.48550/arXiv.2509.08704. URL https://arxiv.org/abs/2509.08704.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pp. 268–282. IEEE, 2018.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=S1zk9iRqF7.

Santiago Zanella-Béguelin, Lukas Wutschitz, Shruti Tople, Ahmed Salem, Victor Rühle, Andrew Paverd, Mohammad Naseri, Boris Köpf, and Daniel Jones. Bayesian estimation of differential privacy. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Zhikun Zhang, Tianhao Wang, Ninghui Li, Jean Honorio, Michael Backes, Shibo He, Jiming Chen, and Yang Zhang. PrivSyn: Differentially private data synthesis. In *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, 2021. URL https://www.usenix.org/conference/usenixsecurity21/presentation/zhang-zhikun.

A Related Work

The goal of privacy auditing is to investigate the extent to which the implementations of private mechanisms adhere to theoretical guarantees of privacy. Membership inference attacks, formalized by Shokri et al. (2017) and further analyzed by Yeom et al. (2018), serve as the fundamental primitive for auditing data privacy in machine learning, quantifying the extent to which a model leaks information about its training set. Carlini et al. (2019) introduced the concept of injecting unique, identifiable "canaries" into the training data to measure unintended memorization.

Building on these foundations, a substantial line of work audits DP-SGD by repeatedly training the mechanism under neighboring datasets and converting attack success rates into privacy lower bounds, beginning with the data-poisoning audits of Jagielski et al. (2020) and the first tight white-box DP-SGD audit of Nasr et al. (2021). Subsequent work has refined the statistical conversion via Bayesian credible intervals Zanella-Béguelin et al. (2023), generic black-box estimators Lu et al. (2024); Askin et al. (2025), and worst-case initializations for tight black-box bounds Annamalai & Cristofaro (2024). Most relevant to our work, Nasr et al. (2023) fit a Gaussian DP model to canary-aligned dot products, but their exact analysis applies to a single, non-subsampled Gaussian step and requires thousands of training runs. We instead prove that the composed, subsampled canary score is asymptotically Gaussian and exploit this in a single training run.

Steinke et al.Steinke et al. (2023) introduced the idea that, by parallelizing across many independent canaries, a single training run can produce enough statistical evidence for a meaningful privacy lower bound. Their construction thresholds each canary's score into a binary membership guess and aggregates these guesses into a confidence interval on ε. Mahloujifar et al. (2025) sharpened this analysis by working directly with f-DP trade-off curves, which enables auditing of the full privacy trade-off curve, rather than only testing a localized (ε, δ) point. Xiang et al. (2025a;b) reinterpret one-run auditing as a noisy-channel coding problem and use order statistics to select high-confidence guesses, while Liu et al. (2025) strengthen the underlying membership-inference attack via quantile regression. Keinan et al. (2025) characterize the limits of distribution-free, guessing-based one-run auditing. Our work differs from these works by taking a different analysis route: instead of thresholding the canary score to a binary random variable, we exploit a parametric Gaussian model for real-valued canary scores. We also note that Steinke et al. (2023) observe that their analysis is generic and may be loose for realistic Gaussian-noise mechanisms. They point to algorithm-specific analyses exploiting the iterative structure of DP-SGD as an important future direction, and our work pursues that direction and gives an affirmative answer.

In federated learning, Maddock et al. (2023) and Andrew et al. (2024) studied the problem of empirical privacy estimation in both per-round and one-shot settings. We note that, while Andrew et al. (2024) also uses a Gaussian approximation, it is applied to cosine-angle statistics between canary directions and noised gradient, rather than to the aggregate canary score. Our Gaussian model instead characterizes the normalized sum of canary observations over the full training trajectory, which is crucial in composed mechanisms such as DP-FTRL and DP-SGD and, in experiments, we show a clear empirical advantage.

Algorithm 2 White-box DP-SGD transcript, following the clipped-and-noised DP-SGD update of Abadi et al. (2016).

Input: Background database D, target canary $x^\star$, bit $b \in \{0,1\}$, loss function ℓ, clipping norm C, subsampling rate q, noise multiplier σ, step sizes $(\eta_t)_{t=1}^T$, initial parameter θ_0, and number of iterations T.

Output: White-box transcript $\mathsf{Tr}_T^{(b)}$.

Set $D^{(0)} \stackrel{\text{def}}{=} D$, $D^{(1)} \stackrel{\text{def}}{=} D \cup \{x^\star\}$, $n_b \stackrel{\text{def}}{=} |D^{(b)}|$.
Initialize $\theta_0^{(b)} \leftarrow \theta_0$ and $\mathsf{Tr}_0^{(b)} \leftarrow (\theta_0^{(b)})$.
for $t = 1, 2, \ldots, T$ **do**
  for each record $i \in [n_b]$ **do**
    Sample $B_{t,i}^{(b)} \sim \text{Bernoulli}(q)$ independently.
  Set the minibatch $\mathcal{L}_t^{(b)} \leftarrow \{D^{(b)}[i] : B_{t,i}^{(b)} = 1,\ i \in [n_b]\}$.
  Compute the noisy averaged gradient

$$\widetilde{g}_t^{(b)} = \frac{1}{qn_b}\left(\sum_{z \in \mathcal{L}_t^{(b)}} \mathsf{clip}_C\left(\nabla_\theta \ell(\theta_{t-1}^{(b)}; z)\right) + Z_t^{(b)}\right), \qquad Z_t^{(b)} \sim \mathcal{N}\left(0, \sigma^2 C^2 I_d\right).$$

  Update $\theta_t^{(b)} \leftarrow \theta_{t-1}^{(b)} - \eta_t \widetilde{g}_t^{(b)}$.
  Append $(\widetilde{g}_t^{(b)}, \theta_t^{(b)})$ to $\mathsf{Tr}_{t-1}^{(b)}$.
return $\mathsf{Tr}_T^{(b)}$.

B White-box DP-SGD Transcript

C Adaptive DP-SGD Scores: A Martingale Extension

Recall from Algorithm 2 and definition 1 that $b \in \{0,1\}$ indicates whether the canary is absent or present, $q \in (0,1)$ is the Poisson-subsampling rate, $\theta_{t-1}^{(b)}$ is the model iterate after step $t-1$, and $u_t^{(b)}$ is the unit direction induced by its clipped canary gradient. Fix a neighboring world b and a rate q. Let the pre-sampling sigma-field $\mathcal{F}_{t-1}^{(b)}$ contain the fixed data and canary, all training randomness through step $t-1$, and any auxiliary randomness used to form the candidate per-record gradients at step t (such as data augmentations or adaptive clipping randomness). Thus $\theta_{t-1}^{(b)}$, $u_t^{(b)}$, and all quantities defined below are $\mathcal{F}_{t-1}^{(b)}$-measurable. Let $\mathcal{F}_t^{(b)}$ additionally reveal the step-t sampling indicators $I_{t,z}^{(b)}$ for $z \in D$ and $B_t^{(b)}$ for the canary x^*, together with the fresh Gaussian noise. For a background record $z \in D$, write

$$g_{t,z}^{(b)} \stackrel{\text{def}}{=} \mathsf{clip}_C\left(\nabla_\theta \ell(\theta_{t-1}^{(b)}; z)\right), \qquad h_{t,z}^{(b)} \stackrel{\text{def}}{=} \langle u_t^{(b)}, g_{t,z}^{(b)} \rangle, \qquad a_t^{(b)} \stackrel{\text{def}}{=} \left\| c_t^{(b)} \right\|_2.$$

Thus $|h_{t,z}^{(b)}| \le C$ and $0 \le a_t^{(b)} \le C$. Let $I_{t,z}^{(b)} \sim \text{Bernoulli}(q)$ be the Poisson-sampling indicator of background record z, and let $B_t^{(b)} \sim \text{Bernoulli}(q)$ be the indicator for the canary. Conditional on $\mathcal{F}_{t-1}^{(b)}$, these indicators and the fresh Gaussian mechanism noise are mutually independent.

Proposition 4 (Proof is in Appendix D)**.** *For every $t \in [T]$ and $b \in \{0,1\}$, the canary observation takes the form*

$$X_t^{(b)} = m_t^{(b)} + \xi_t^{(b)},$$

where

$$m_t^{(b)} = q\left[\sum_{z \in D} h_{t,z}^{(b)} + b\, a_t^{(b)}\right],$$

$$\xi_t^{(b)} = \sum_{z \in D} (I_{t,z}^{(b)} - q) h_{t,z}^{(b)} + b(B_t^{(b)} - q) a_t^{(b)} + \sigma C G_t^{(b)}, \quad G_t^{(b)} \mid \mathcal{F}_{t-1}^{(b)} \sim \mathcal{N}(0,1)$$

Consequently,

$$\mathbf{Ex}\left[\xi_t^{(b)} \mid \mathcal{F}_{t-1}^{(b)}\right] = 0, \quad \mathbf{Ex}\left[(\xi_t^{(b)})^2 \mid \mathcal{F}_{t-1}^{(b)}\right] = q(1-q)\left[\sum_{z\in D}(h_{t,z}^{(b)})^2 + b(a_t^{(b)})^2\right] + \sigma^2 C^2.$$

and $(\xi_t^{(b)}, \mathcal{F}_t^{(b)})_{t\le T}$ *is a martingale-difference sequence.*

The canary score aggregates the one-step canary observations in Proposition 4. The following definition introduces the relevant notation for its deterministic mean $\mu_{b,T}$, centered fluctuation $D_{b,T}$, martingale sum $M_{b,T}$, unconditional martingale variance $v_{b,T}$, and predictable quadratic variation $Q_{b,T}$.

Definition 4 (Score decomposition)**.** Fix a world $b \in \{0,1\}$ and a number of training steps T. Write $X_{T,t}^{(b)} = m_{T,t}^{(b)} + \xi_{T,t}^{(b)}$ by Proposition 4, and define

$$S_T^{(b)} \stackrel{\text{def}}{=} \frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_{T,t}^{(b)}, \qquad \mu_{b,T} \stackrel{\text{def}}{=} \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\mathbf{Ex}\left[m_{T,t}^{(b)}\right]$$

$$D_{b,T} \stackrel{\text{def}}{=} \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left(m_{T,t}^{(b)} - \mathbf{Ex}\left[m_{T,t}^{(b)}\right]\right), \qquad M_{b,T} \stackrel{\text{def}}{=} \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\xi_{T,t}^{(b)}$$

$$v_{b,T} \stackrel{\text{def}}{=} \frac{1}{T}\sum_{t=1}^{T}\mathbf{Ex}\left[(\xi_{T,t}^{(b)})^2\right], \qquad Q_{b,T} \stackrel{\text{def}}{=} \frac{1}{T}\sum_{t=1}^{T}\mathbf{Ex}\left[(\xi_{T,t}^{(b)})^2 \mid \mathcal{F}_{T,t-1}^{(b)}\right].$$

The score admits the decomposition

$$S_T^{(b)} = \mu_{b,T} + D_{b,T} + M_{b,T},$$

Definition 4 gives the score representation with the full adaptive background trajectory. To distinguish the DP-SGD noise multiplier from the standard deviation of its one-dimensional projection, write $\tau \stackrel{\text{def}}{=} \sigma C > 0$.

From the adaptive decomposition to Model 1. The preceding decomposition allows the background projections, canary magnitude, and canary direction to depend on the training history. The next two assumptions separate the roles of background suppression and canary saturation. A fixed-coordinate Dirac gradient canary enforces exact saturation by construction, but it does not by itself suppress the projections of ordinary clipped gradients; background suppression is a separate condition. Once the background is suppressed, a predictable change in direction does not alter the scalar noise law: isotropy implies that the projection of each fresh Gaussian noise vector onto any past-determined unit direction has law $\mathcal{N}\left(0,\tau^2\right)$ and is independent of the past. Exact saturation therefore gives a finite-T reduction to Model 1.

Assumption 1 (Background suppression)**.** *For every* $t \in [T]$*,* $z \in D$*, and* $b \in \{0,1\}$*,*

$$h_{t,z}^{(b)} = 0 \qquad \textit{almost surely}.$$

Lemma 2 (Exact canary saturation)**.** *For every* $t \in [T]$*, the present-world canary magnitude satisfies*

$$a_t^{(1)} = C$$

Proposition 5 (Fixed-magnitude reduction; the proof is in Section D)**.** *Under Assumption 1 and lemma 2, for each fixed* $b \in \{0,1\}$*, the sequence* $\left(B_t^{(b)}, G_t^{(b)}\right)_{t=1}^{T}$ *is i.i.d. with common law* $\mathrm{Bernoulli}(q) \otimes \mathcal{N}(0,1)$*, and*

$$X_t^{(b)} = bB_t^{(b)}C + \tau G_t^{(b)}.$$

Consequently, for the scores defined in Definition 4,

$$S_T^{(0)} \sim \mathcal{N}\left(0,\tau^2\right),$$

$$S_T^{(1)} = \frac{C}{\sqrt{T}}\sum_{t=1}^{T}B_t^{(1)} + \frac{\tau}{\sqrt{T}}\sum_{t=1}^{T}G_t^{(1)}, \quad \mathbf{Ex}\left[S_T^{(1)}\right] = \sqrt{T}\,qC, \quad \mathrm{Var}S_T^{(1)} = \tau^2 + q(1-q)C^2.$$

D MISSING PROOFS FOR APPENDIX C

Proof of Proposition 4. Fix $t \in [T]$ and $b \in \{0,1\}$. By Algorithm 2 and definition 1,

$$X_t^{(b)} = qn\langle u_t^{(b)}, \widetilde{g}_t^{(b)} \rangle = \sum_{z \in D} I_{t,z}^{(b)} h_{t,z}^{(b)} + bB_t^{(b)} a_t^{(b)} + \langle u_t^{(b)}, Z_t^{(b)} \rangle.$$

The direction $u_t^{(b)}$ is $\mathcal{F}_{t-1}^{(b)}$-measurable and has unit norm, with $Z_t^{(b)} \mid \mathcal{F}_{t-1}^{(b)} \sim \mathcal{N}\left(0, \sigma^2 C^2 I_d\right)$. Rotational invariance therefore gives a random variable $G_t^{(b)}$ such that

$$\langle u_t^{(b)}, Z_t^{(b)} \rangle = \sigma C G_t^{(b)}, \qquad G_t^{(b)} \mid \mathcal{F}_{t-1}^{(b)} \sim \mathcal{N}(0,1).$$

Adding and subtracting the conditional means of the sampling indicators, we have

$$\begin{aligned} X_t^{(b)} &= q\left[\sum_{z \in D} h_{t,z}^{(b)} + b\, a_t^{(b)}\right] + \sum_{z \in D} (I_{t,z}^{(b)} - q) h_{t,z}^{(b)} + b(B_t^{(b)} - q) a_t^{(b)} + \sigma C G_t^{(b)} \\ &= m_t^{(b)} + \xi_t^{(b)}. \end{aligned}$$

Hence $m_t^{(b)}$ is the predictable conditional mean of $X_t^{(b)}$, whereas $\xi_t^{(b)} = X_t^{(b)} - m_t^{(b)}$ is its centered perturbation.

Conditionally on $\mathcal{F}_{t-1}^{(b)}$, the variables $I_{t,z}^{(b)} \sim \text{Bernoulli}(q)$, $B_t^{(b)} \sim \text{Bernoulli}(q)$, and $G_t^{(b)} \sim \mathcal{N}(0,1)$ are mutually independent, while the coefficients $h_{t,z}^{(b)}$ and $a_t^{(b)}$ are measurable. Thus the conditional expectations below average only over the fresh step-t sampling indicators and Gaussian noise, with the past fixed. It follows that $\mathbf{Ex}\left[\xi_t^{(b)} \middle| \mathcal{F}_{t-1}^{(b)}\right] = 0$. Also, we have

$$\begin{aligned} \mathbf{Ex}\left[(\xi_t^{(b)})^2 \middle| \mathcal{F}_{t-1}^{(b)}\right] &= q(1-q) \sum_{z \in D} (h_{t,z}^{(b)})^2 + b^2 q(1-q)(a_t^{(b)})^2 + \sigma^2 C^2 \\ &= q(1-q)\left[\sum_{z \in D} (h_{t,z}^{(b)})^2 + b(a_t^{(b)})^2\right] + \sigma^2 C^2, \end{aligned}$$

where the final equality uses $b^2 = b$ for $b \in \{0,1\}$.

Finally, since $\xi_t^{(b)}$ is $\mathcal{F}_t^{(b)}$-measurable and square-integrable: $\left|h_{t,z}^{(b)}\right| \le C$, $0 \le a_t^{(b)} \le C$, and the Gaussian term has finite second moment. Together with its zero conditional mean, this proves that $(\xi_t^{(b)}, \mathcal{F}_t^{(b)})_{t \le T}$ is a martingale-difference sequence. □

Proof of Proposition 5. Under Assumption 1 and Lemma 2, the general formulas in Proposition 4 reduce to

$$m_t^{(b)} = bqC, \qquad \xi_t^{(b)} = b(B_t^{(b)} - q)C + \tau G_t^{(b)}.$$

Here the factor b makes the value of $a_t^{(0)}$ irrelevant in the absent world. Adding these two terms gives

$$X_t^{(b)} = bB_t^{(b)} C + \tau G_t^{(b)}.$$

It remains to justify that the history-dependent direction does not create scalar temporal dependence. Conditional on $\mathcal{F}_{t-1}^{(b)}$, the vector $u_t^{(b)}$ is fixed and has unit norm, whereas the fresh Gaussian noise is isotropic with law $\mathcal{N}\left(0, \tau^2 I_d\right)$. Therefore $G_t^{(b)} \mid \mathcal{F}_{t-1}^{(b)} \sim \mathcal{N}(0,1)$. To see that $\left(B_t^{(b)}, G_t^{(b)}\right)_{t=1}^T$ is i.i.d. with common law $\text{Bernoulli}(q) \otimes \mathcal{N}(0,1)$, examine the conditional characteristic function of the pair at time t. For every $r, s \in \mathbb{R}$,

$$\mathbf{Ex}\left[\exp\left(irB_t^{(b)} + isG_t^{(b)}\right) \mid \mathcal{F}_{t-1}^{(b)}\right] = \left[(1-q) + qe^{ir}\right] e^{-s^2/2}.$$

The right-hand side is deterministic, does not depend on $\mathcal{F}_{t-1}^{(b)}$, and is the characteristic function of $\mathrm{Bernoulli}(q)\otimes\mathcal{N}(0,1)$. Iterating this identity with the tower property proves the joint characteristic-function factorization across time, and hence the asserted temporal independence.

Summing the one-step representation gives

$$S_T^{(b)} = \frac{bC}{\sqrt{T}}\sum_{t=1}^{T} B_t^{(b)} + \frac{\tau}{\sqrt{T}}\sum_{t=1}^{T} G_t^{(b)}.$$

The absent-world formula is exactly $\mathcal{N}\left(0,\tau^2\right)$. In the present world, independence and the Bernoulli and Gaussian moments give

$$\mathbf{Ex}\left[S_T^{(1)}\right] = \sqrt{T}\,qC, \qquad \mathrm{Var}S_T^{(1)} = q(1-q)C^2 + \tau^2.$$

Finally, $m_t^{(b)} = bqC$ is deterministic, so $D_{b,T} = 0$, and the conditional second-moment formula in Proposition 4 becomes the deterministic value $\tau^2 + bq(1-q)C^2$. Its time average and expectation are therefore both the same value, which proves $Q_{b,T} = v_{b,T} = \tau^2 + bq(1-q)C^2$ and completes the proof. □

E Convergence theory for sum of Gaussian mixture

Definition 5 (Two-component centered Gaussian mixture)**.** A real random variable X is a *centered two-component Gaussian mixture* with parameters $(w, v, \mu, \rho, \sigma)$, written $X \sim \mathrm{GM}(w,\, v,\, \mu,\, \rho,\, \sigma)$, if its density function is

$$p_X(x) \;=\; \frac{w}{\sqrt{2\pi}\,\sigma}\exp\left(-(x-\mu)^2/(2\sigma^2)\right) \;+\; \frac{v}{\sqrt{2\pi}\,\sigma}\exp\left(-(x-\rho)^2/(2\sigma^2)\right),$$

where $w, v \geq 0$, $w + v = 1$, $\sigma > 0$, and the centering condition $w\mu + v\rho = 0$ holds, so that $\mathbf{Ex}[X] = 0$.

For a real random variable Y, the moment generating function $M_Y(t) \stackrel{\mathrm{def}}{=} \mathbf{Ex}\left[\exp\left(tY\right)\right]$ and the cumulant generating function $K_Y(t) \stackrel{\mathrm{def}}{=} \log M_Y(t)$. Further define ℓ-th cumulant of Y as

$$\kappa_\ell(Y) \stackrel{\mathrm{def}}{=} \lim_{t\to 0}\frac{d^\ell}{dt^\ell}K_Y(t).$$

For a Gaussian distribution $G \sim \mathcal{N}\left(0,\tau^2\right)$, $K_G(t) = \tau^2 t^2/2$, and so $\kappa_2(G) = \tau^2$ and $\kappa_\ell(G) = 0$ for all $\ell \geq 3$.

Because the MGF of a two-component Gaussian mixture is finite in a neighborhood of $t = 0$, the moment problem for Y is determinate: the cumulant sequence $\{\kappa_l(Y)\}_{l\geq 1}$ uniquely determines the law of Y. Hence, if a sequence of mixture-derived random variables has cumulants of order ≥ 3 shrinking to zero, the limit law is necessarily Gaussian, and the rate of cumulant decay quantifies the rate of convergence.

Theorem 2 (Cumulant expansion of normalized mixture sums)**.** *Let $X_1, \ldots, X_n$ be independent with $X_i \sim \mathrm{GM}(w_i,\, v_i,\, \mu_i,\, \rho_i,\, \sigma_i)$ for each $i \in [n]$, and define*

$$S_n \stackrel{\mathrm{def}}{=} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i.$$

Then $\kappa_0(S_n) = \kappa_1(S_n) = 0$, and

$$\kappa_2(S_n) = \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2 + \frac{1}{n}\sum_{i=1}^{n}\left(w_i\mu_i^2 + v_i\rho_i^2\right),$$

$$\kappa_3(S_n) = \frac{1}{n^{3/2}}\sum_{i=1}^{n}\left(w_i\mu_i^3 + v_i\rho_i^3\right),$$

$$\kappa_4(S_n) = \frac{1}{n^2}\sum_{i=1}^{n}\Big[w_i\mu_i^4 + v_i\rho_i^4 - 3\left(w_i\mu_i^2 + v_i\rho_i^2\right)^2\Big],$$

and more generally, for every $\ell \geq 3$,

$$\kappa_\ell(S_n) = \frac{\alpha_\ell}{n^{\ell/2-1}}, \qquad \alpha_\ell \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \beta_{\ell,i},$$

where $\beta_{\ell,i}$ is a polynomial of degree ℓ in (μ_i, ρ_i) with coefficients depending only on (w_i, v_i).

Proof. For $X_i \sim \mathrm{GM}(w_i, v_i, \mu_i, \rho_i, \sigma_i)$, conditioning on the mixture component and using $\mathbf{Ex}\left[\exp\left(tN\right)\right] = \exp\left(mt + s^2t^2/2\right)$ for $N \sim \mathcal{N}\left(m, s^2\right)$ gives

$$M_{X_i}(t) = \exp\left(\sigma_i^2 t^2/2\right)\Big(w_i \exp\left(\mu_i t\right) + v_i \exp\left(\rho_i t\right)\Big)$$

$$\implies M_{S_n}(t) = \prod_{i=1}^{n} M_{X_i}\left(\frac{t}{\sqrt{n}}\right) = \exp\left(\frac{t^2}{2n}\sum_{i=1}^{n}\sigma_i^2\right)\prod_{i=1}^{n}\Big(w_i \exp\left(\mu_i t/\sqrt{n}\right) + v_i \exp\left(\rho_i t/\sqrt{n}\right)\Big).$$

(by independence)

Taking logarithms,

$$K_{S_n}(t) = \frac{t^2}{2n}\sum_{i=1}^{n}\sigma_i^2 + \sum_{i=1}^{n}\log\Big(w_i \exp\left(\mu_i t/\sqrt{n}\right) + v_i \exp\left(\rho_i t/\sqrt{n}\right)\Big).$$

Expanding the exponentials as power series:

$$w_i \exp\left(\mu_i t/\sqrt{n}\right) + v_i \exp\left(\rho_i t/\sqrt{n}\right) = \sum_{k=0}^{\infty}\frac{t^k}{k!\,n^{k/2}}\left(w_i\mu_i^k + v_i\rho_i^k\right).$$

The $k = 0$ term is $w_i + v_i = 1$, and the $k = 1$ term is $\frac{t}{\sqrt{n}}(w_i\mu_i + v_i\rho_i) = 0$ by the centering condition. Hence we can rewrite:

$$w_i \exp\left(\mu_i t/\sqrt{n}\right) + v_i \exp\left(\rho_i t/\sqrt{n}\right) = 1 + f_i(t), \qquad \text{where } f_i(t) \stackrel{\text{def}}{=} \sum_{k=2}^{\infty}\frac{t^k}{k!\,n^{k/2}}\left(w_i\mu_i^k + v_i\rho_i^k\right).$$

Applying Taylor series for log we get:

$$K_{S_n}(t) = \frac{t^2}{2n}\sum_{i=1}^{n}\sigma_i^2 + \sum_{i=1}^{n}\sum_{\ell=1}^{\infty}\frac{(-1)^{\ell+1}}{\ell}\big[f_i(t)\big]^\ell.$$

Given the above equation, we can compute $m!$ times the coefficient of t^m to extract the cumulants of S_n.

- Case $m = 2$: Two contributions: the explicit Gaussian-noise term $\frac{t^2}{2n}\sum_i \sigma_i^2$, and the $\ell = 1$, $k = 2$ term of $f_i(t)$, which is $\frac{t^2}{2n}\sum_i(w_i\mu_i^2 + v_i\rho_i^2)$. The $\ell \geq 2$ contributions vanish at order t^2, since $f_i(t)^\ell$ already starts at $t^{2\ell}$.
- Case $m = 3$: For $\ell \geq 2$, $f_i(t)^\ell$ starts at $t^{2\ell} \geq t^4$, contributing nothing at order t^3. Hence only $\ell = 1$, $k = 3$ contributes:

  $$\sum_{i=1}^{n}\frac{1}{3!\,n^{3/2}}(w_i\mu_i^3 + v_i\rho_i^3).$$

  Multiplying by 3! gives the exact form.
- Case (general $m \geq 3$): The lowest-order term of $f_i(t)^\ell$ is $t^{2\ell}/n^\ell$, so only finitely many values of ℓ contribute to t^m, namely $\ell \in \{1, 2, \ldots, \lfloor m/2 \rfloor\}$. Each such contribution to t^m carries a factor $1/n^{m/2}$ (from the $f_i(t)^\ell$ expansion at order t^m, where the powers of t and $1/\sqrt{n}$ track together). The per-i coefficient is therefore of the form $\beta_{\ell,i}/n^{m/2}$ for a polynomial $\beta_{\ell,i}$ in (μ_i, ρ_i), and summing over $i \in [n]$ produces a factor of n. After multiplication by $m!$, the cumulant takes the form (some bounded scalar)$/n^{m/2-1}$. This gives the general formula for the cumulants of a Gaussian mixture.

□

Applying the above theorem to the case where X_i are i.i.d. samples gives us the following asymptotic rate of convergence for all cumulants:

Corollary 1 (Rate of convergence)**.** *In Theorem 2 if the summands are i.i.d., i.e.,* $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{GM}(w,\, v,\, \mu,\, \rho,\, \sigma)$ *for some fixed parameters* $(w, v, \mu, \rho, \sigma)$*. Then* $\kappa_\ell(S_n) = O(n^{-(\ell/2-1)})$ *for every* $\ell \geq 3$*, with the leading non-vanishing cumulant of order* ≥ 3 *determined as follows.*

- *If* $w\mu^3 + v\rho^3 \neq 0$*, then*
$$\kappa_3(S_n) = \Theta(n^{-1/2}),$$
and the third cumulant is the leading non-vanishing one.
- *If* $w\mu^3 + v\rho^3 = 0$*, then* $\kappa_3(S_n) = 0$ *and*
$$\kappa_4(S_n) = \Theta(n^{-1}),$$
and the fourth (or higher) cumulant is the leading non-vanishing one.

Quantitative Kolmogorov-distance estimate The cumulant decay above is not just an asymptotic statement: it directly controls the Kolmogorov distance between the true law $\widetilde{Q}_T$ of the standardized audit statistic and its Gaussian model Q_T via the classical Edgeworth expansion:

Lemma 3 (Edgeworth expansion, Petrov (2000; 2012))**.** *Let* $Y_1, \ldots, Y_n$ *be i.i.d. random variables with mean* 0*, variance* $\tau^2 > 0$*, and distribution satisfying Cramér's condition* $\limsup_{|t|\to\infty} |\mathbf{E}\mathbf{x}\left[e^{itY_1}\right]| < 1$*. Let* F_n *be the CDF of* $(\tau\sqrt{n})^{-1}\sum_{i=1}^{n} Y_i$*. Then*

$$F_n(x) - \Phi(x) = -\varphi(x)\left[\frac{\gamma_1}{6\sqrt{n}}\, h_2(x) + \frac{1}{n}\left(\frac{\gamma_2}{24}\, h_3(x) + \frac{\gamma_1^2}{72}\, h_5(x)\right)\right] + O\big(n^{-3/2}\big), \tag{3}$$

where $\gamma_1 \overset{\text{def}}{=} \kappa_3(Y_1)/\tau^3$*,* $\gamma_2 \overset{\text{def}}{=} \kappa_4(Y_1)/\tau^4$*,* φ *is the standard Gaussian density, and* $h_2(x) = x^2 - 1$*,* $h_3(x) = x^3 - 3x$*,* $h_5(x) = x^5 - 10x^3 + 15x$ *are probabilists' Hermite polynomials.*

The expansion does not require any computation beyond the cumulants of $\widetilde{S}_T$ already derived: setting $Y_t = \widetilde{X}_t$ and $n = T$, the Edgeworth coefficients equal the standardized cumulants of $\widetilde{S}_T$ itself,

$$\frac{\gamma_1}{\sqrt{T}} = \frac{\kappa_3\Big(\widetilde{S}_T\Big)}{\kappa_2\Big(\widetilde{S}_T\Big)^{3/2}}, \qquad \frac{\gamma_2}{T} = \frac{\kappa_4\Big(\widetilde{S}_T\Big)}{\kappa_2\Big(\widetilde{S}_T\Big)^{2}}.$$

Using the above general statement we can derive the closed-form Kolmogorov-distance bound for the DP audit statistic:

Proof of Theorem 1. The summands in normalized distribution $\widetilde{S}_T = T^{-1/2}\sum_{t=1}^{T} \widetilde{X}_t$ satisfy $\widetilde{X}_t \overset{\text{i.i.d.}}{\sim} \mathrm{GM}(1-q,\, q,\, -qC,\, (1-q)C,\, \sigma)$, an i.i.d. centered Gaussian mixture. Its distribution has a smooth density, so Cramér's condition holds and Lemma 3 applies with $Y_t = \widetilde{X}_t$, $\tau^2 = \kappa_2\Big(\widetilde{X}_t\Big) = \sigma^2 + q(1-q)C^2$, and $n = T$. Using the cumulant formulas already derived,

$$\gamma_1 = \frac{q(1-q)(1-2q)\, C^3}{\big[\sigma^2 + q(1-q)C^2\big]^{3/2}}.$$

The Kolmogorov distance is invariant under common affine rescaling of $\widetilde{Q}_T$ and Q_T, so it equals the standardized $\sup_{x\in\mathbb{R}} |F_T(x) - \Phi(x)|$ from equation 3. The function $\varphi(x)(x^2-1)$ attains its maximal absolute value at $x = 0$ with $|\varphi(0)\cdot(-1)| = 1/\sqrt{2\pi}$, and the next-order Edgeworth contributions are uniformly $O(1/T)$. Hence

$$d_{\mathrm{K}}\Big(\widetilde{Q}_T,\, Q_T\Big) = \frac{|\gamma_1|}{6\sqrt{T}} \cdot \frac{1}{\sqrt{2\pi}} + O(1/T)$$

□

Improved concrete estimate for convergence rate For the DP-SGD auditing parameters, we can in fact get tighter bound than the above global Kolmogorov-distance estimate. The intuition is that the audit only ever queries the CDF in one specific region, and hence we need the Gaussian approximation to be accurate only where the audit looks.

Let F_T be the CDF of the standardized audit statistic $\widetilde{S}_T$; and let the functions φ and Φ respectively represent the standard Gaussian density function and its corresponding CDF. Then we have the following tighter bound:

Corollary 2 (Tail-localized CDF deviation)**.** *Under the setting of Theorem 1, for every threshold* $x_0 \geq \sqrt{3}$*,*

$$\sup_{|x|\geq x_0} |F_T(x) - \Phi(x)| = \frac{q(1-q)\,|1-2q|\,C^3}{6\sqrt{T}\left[\sigma^2 + q(1-q)C^2\right]^{3/2}} \cdot \varphi(x_0)\,(x_0^2 - 1) + O(1/T).$$

Proof of corollary 2. The Edgeworth expansion Lemma 3 gives the pointwise expansion $F_T(x) - \Phi(x) = -\varphi(x)(x^2-1)\,\gamma_1/(6\sqrt{T}) + O(1/T)$ uniformly in x. By the monotonicity of $\varphi(x)(x^2-1)$ on $|x| \geq \sqrt{3}$, the supremum on $\{|x| \geq x_0\}$ is attained at $|x| = x_0$, giving the stated identity after substituting $\gamma_1 = q(1-q)(1-2q)C^3/[\sigma^2 + q(1-q)C^2]^{3/2}$. □

For concrete parameters used in our experiments: $T = 2500$, $\epsilon = 8$, $\sigma = 2.6245$, $q = 0.0819$, and $C = 1$, the prefactor $\frac{q(1-q)\,|1-2q|\,C^3}{6\sqrt{T}\,(\sigma^2+q(1-q)C^2)^{3/2}} \approx 1.14 \times 10^{-5}$, and the tail factor $\varphi(x_0)(x_0^2-1)$ controls how small the deviation becomes as follows:

x_0	$\varphi(x_0)(x_0^2-1)$	$\sup_{\lvert x\rvert\geq x_0} \lvert F_T(x) - \Phi(x)\rvert$
3	3.55×10^{-2}	4.04×10^{-7}
4	2.01×10^{-3}	2.29×10^{-8}
5	3.58×10^{-5}	4.08×10^{-10}

The threshold $x_0 = 4$ corresponds to a Gaussian tail mass $\Phi(-4) \approx 3 \times 10^{-5}$, which is the regime relevant for our $\delta = 10^{-5}$ audit. Within this audit-relevant tail the Gaussian model matches the true distribution to within $\sim 10^{-8}$, which is three orders of magnitude smaller. This gives more confidence that the empirical privacy lower bound produced by the auditor closely tracks the theoretical guarantee in our experimental setup.

F Experimental details for DP-FTRL and DP-MF-FTRL

For DP-FTRL, $\mathbf{C}_{\nu,t} = 1$ iff leaf t lies below node ν, and each node is released with independent noise. Every leaf has exactly $L = \lceil \log_2 T \rceil + 1$ ancestors, so $\|\mathbf{C}_{:,t}\|_2 = \sqrt{L}$ for all t, and the score is the sum of the released node values on the ancestor path of t_i, divided by $\sqrt{L}$. The Gaussian pair is

$$G_0 = \mathcal{N}\big(0,\ \sigma^2_{\text{node}}\big), \qquad G_1 = \mathcal{N}\big(\sqrt{L}\,C,\ \sigma^2_{\text{node}}\big),$$

and every leaf is a worst-case placement.

Whereas in DP-FTRL, although the released updates exhibit correlated noise, the decoder $\mathbf{B}$ is invertible, so the auditor can recover y from the decoded releases.

We calibrate the noise with the analytic Gaussian mechanism (Balle & Wang, 2018), at sensitivity $\sqrt{L}\,C$ for the tree and at sensitivity $C \max_t \|\mathbf{C}_{:,t}\|_2$ for the factorization. As the encoder we use the square root of the prefix-sum matrix, which is lower triangular and Toeplitz, has a closed form, and needs no numerical optimization. Both mechanisms train the same ScatterLinear classifier on CIFAR-10, a $J = 2$ scattering transform (Andreux et al., 2020) followed by GroupNorm and a linear layer, in a single pass with $T = 128$ rounds, logical batch size $B = 380$, and $m = 5000$ Dirac canaries. At $T = 128$ the factorization needs 1.75 times less noise for the same privacy guarantee. The per-canary statistic of Andrew et al. (2024) is the round-wise max cosine $\tilde{s}_i = \max_{t\in[0,T)} \langle e_{c_i},\, G_t\rangle / \|G_t\|_2$, where G_t is the released cumulative iterate at round t.

G Additional Experiments

This section provides supplementary plots referenced in Section 4, including adaptive clipping for DP-SGD, robustness to hyperparameters, canary-out score distribution, a hexbin density visualization of canary direction independence, and a comparison of confidence region constructions across varying canary counts.

G.1 Adaptive clipping in DP-SGD

In the adaptive clipping algorithm, the clipping bound at each step (C_t) is noised and released publicly (Andrew et al., 2021). Since the adaptive clipping threshold is released publicly, the auditor can normalize each canary score by the corresponding clipping bound. This reduces the score process to the same normalized model analyzed in Model 1. The nearly identical empirical results in the table below suggest that the background gradient channel remains unchanged.

	Steinke et al.	Mahloujifar et al.	Our auditor	True epsilon
Fixed clipping	3.330	4.697	6.651	8
Adaptive clipping	3.095	4.464	6.769	8

G.2 Robustness to hyperparameters

Clipping norm: Proposition 1 shows that when the DP noise is calibrated proportional to the clipping norm, changing the clipping norm scales both neighboring score distributions by the same factor. Because the hockey-stick divergence is invariant under a common invertible rescaling, the inferred privacy bound should remain essentially unchanged. Our experiment below confirms this prediction: increasing the clipping from 1 to 10 doesn't change our estimate too much, and it remains substantially tighter than both baselines.

Table 2: Empirical lower bounds on ε for different clipping bounds C (claimed $\varepsilon = 8$)

Clipping bound C	Steinke et al. (2023)	Mahloujifar et al. (2025)	Ours
1	3.330	4.697	6.679
10	3.486	5.115	6.678

Batch size: The batch size affects the sampling probability. A smaller batch therefore produces fewer canary contributions over training and makes the audit statistically harder. This expected trend appears for all methods. Nevertheless, our auditor remains the strongest across all tested batch sizes, as shown below.

Table 3: Empirical lower bounds on ε for different batch sizes (claimed $\varepsilon = 8$)

Batch size	Steinke et al. (2023)	Mahloujifar et al. (2025)	Ours
1024	2.410	3.236	4.518
2048	3.427	4.877	6.108
4096	3.330	4.697	6.679

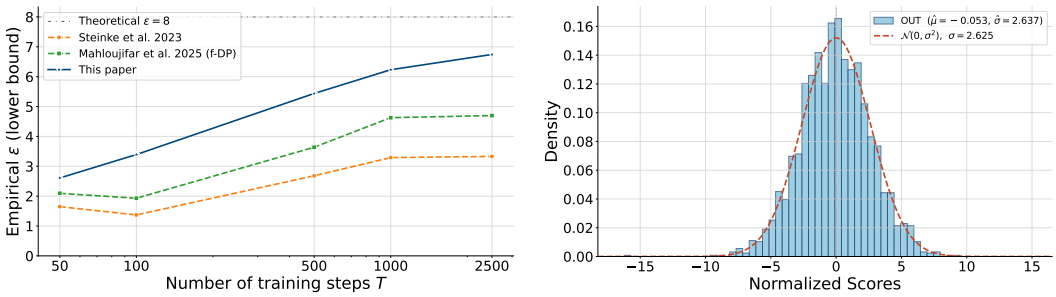
Learning rate: The learning rate has no substantial effect on any of the three auditors' estimates. This invariance is exactly what Model 1 predicts. The close agreement between this prediction and the empirical results across a range of learning rates provides additional evidence that Model 1's idealization captures the relevant dependencies well in practice.

Table 4: Empirical lower bounds on ε for different learning rates (claimed $\varepsilon = 8$)

Learning rate	Steinke et al. (2023)	Mahloujifar et al. (2025)	Ours
2	3.330	4.697	6.645
4	3.330	4.697	6.679
8	3.346	4.727	6.699
16	3.324	4.687	6.674

G.3 Gaussianity of Canary Scores

Complementing Figure 5 in the main text, which visualizes the distribution of canary-present (in) scores, Figure 6b shows the empirical distribution of canary-absent (out) scores against the analytical Gaussian parameters.



(a) Auditing performance versus training steps.

(b) Canary-absent scores at $T = 2500$.

Figure 6: DP-SGD ablations at theoretical $\epsilon = 8$ and $m = 5000$ canaries. In panel (b), the empirical score distribution closely matches the analytical Gaussian distribution $\mathcal{N}(0, \sigma^2)$.

G.4 Identical Distribution Across Canary Index

Figure 7 plots the standardized score $z_i = (s_i - \mu)/\sigma$ for each canary $i \in [1, m]$ as a hexbin density. If the marginal distribution of s_i drifted with i due to canary ordering correlating with training dynamics or embedding-space artifacts — we would expect to see vertical banding, trends, or shifts in the central mass. Instead, the density is uniform across the canary index axis, peaks symmetrically near zero, and remains contained within $[-3, 3]$, supporting the assumption that $s_i \sim \mathcal{N}(\mu, \sigma^2)$ uniformly across the canary population.

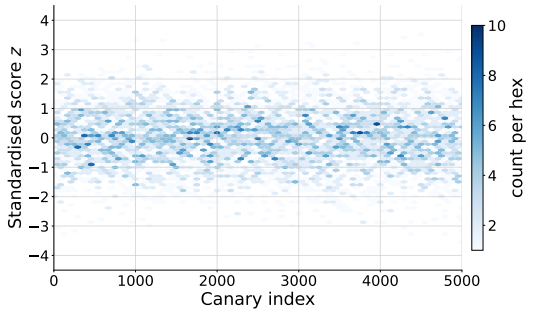


Figure 7: Hexbin density of standardized canary scores $z_i = (s_i - \mu)/\sigma$ plotted against canary index for DP-SGD at $\epsilon = 8$, $m = 5000$, $T = 2500$.

G.5 Confidence Region Geometry vs. Sample Complexity

As discussed in the *Confidence Region Geometry* paragraph of Section 4, we compare the Parametric Bonferroni and Bootstrap Ellipsoid confidence regions. Figure 8 shows how the recovered empirical

ε varies with the number of canaries m under each construction. The Ellipsoid consistently recovers higher ε values across all sample sizes by capturing the cross-correlation between sample means and variances, which the axis-aligned Bonferroni rectangle ignores. The gap is especially pronounced at smaller m, where efficient use of statistical information matters most.

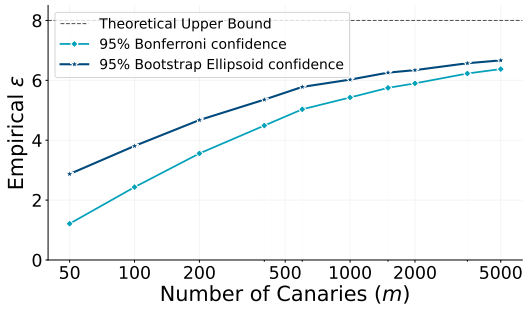


Figure 8: Comparison of empirical ε recovered by the Parametric Bonferroni rectangle versus the Bootstrap Ellipsoid confidence region across varying canary counts m. The Ellipsoid yields tighter (higher) lower bounds at all sample sizes by exploiting the joint geometry of the mean-variance estimates.