# Bayesian Estimation of Differential Privacy

**Santiago Zanella-Béguelin** [*1] **Lukas Wutschitz** [*1] **Shruti Tople** [1] **Ahmed Salem** [1] **Victor Rühle** [1]
**Andrew Paverd** [1] **Mohammad Naseri** [2] **Boris Köpf** [1] **Daniel Jones** [1]

## Abstract

Algorithms such as Differentially Private SGD enable training machine learning models with formal privacy guarantees. However, because these guarantees hold with respect to unrealistic adversaries, the protection afforded against practical attacks is typically much better. An emerging strand of work empirically estimates the protection afforded by differentially private training as a confidence interval for the privacy budget $\hat{\varepsilon}$ spent with respect to specific threat models. Existing approaches derive confidence intervals for $\hat{\varepsilon}$ from confidence intervals for false positive and false negative rates of membership inference attacks, which requires training an impractically large number of models to get intervals that can be acted upon. We propose a novel, more efficient Bayesian approach that brings privacy estimates within the reach of practitioners. Our approach reduces sample size by computing a posterior for $\hat{\varepsilon}$ (not just a confidence interval) from the joint posterior of the false positive and false negative rates of membership inference attacks. We implement an end-to-end system for privacy estimation that integrates our approach and state-of-the-art membership inference attacks, and evaluate it on text and vision classification tasks. For the same number of samples, we see a reduction in interval width of up to $40\,\%$ compared to prior work.

## 1. Introduction

The use of machine learning in industries such as healthcare and finance requires strong and auditable safeguards against leakage of sensitive training data. Differentially Private (DP) training using algorithms such as DP-SGD (Abadi et al., 2016; Song et al., 2013) and PATE (Papernot et al., 2017) partially addresses this concern by bounding the amount of information that models can leak. However, there is a gap between the degree of protection that DP training offers *in theory*, and the protection it offers *in practice*. For example, DP training with a privacy budget of $\varepsilon = 4$, a common choice in practice (Desfontaines, 2022), cannot rule out membership inference attacks (Humphries et al., 2020). Nonetheless, DP training with such large budgets effectively defeats attacks in many practical scenarios (Carlini et al., 2019; Jayaraman & Evans, 2019; Song & Shmatikov, 2019; Zanella-Béguelin et al., 2020). The reason for this discrepancy is that provable DP bounds (Gopi et al., 2021) hold up to extremely powerful threat models (e.g., in the case of DP-SGD, adversaries that can observe and tamper with intermediate model updates) and so overestimate the privacy risks of weaker adversaries that matter in practice.

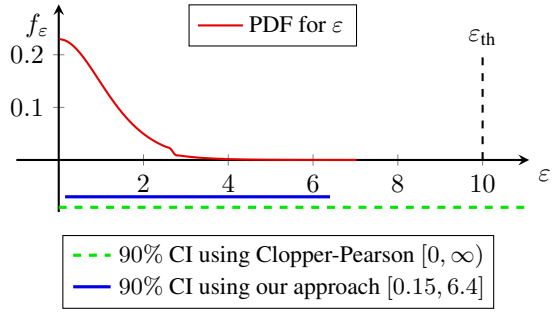



*Figure 1.* Comparison of the posterior PDF $f_\varepsilon$ using our Bayesian approach and the upper bound $\varepsilon_{\text{th}}$ obtained from a state-of-the-art DP accountant (Gopi et al., 2021) for a CNN trained on CIFAR-10 with $\delta$ of $10^{-5}$. Empirical evaluation suggests stronger privacy than what can be proven using accountants. Below the plot, we illustrate the reduction in uncertainty of the $90\,\%$ credible interval of our Bayesian approach over Clopper-Pearson intervals.

Without any information beyond provable DP bounds, practitioners must either err on the side of caution and use unnecessarily small privacy budgets, which hurt utility, or risk using larger budgets based on a guess of the privacy they provide. To resolve this conflict, an emerging strand of work measures the empirical protection afforded by DP training against specific adversaries by computing statistical estimates for the privacy budget spent (Hyland & Tople, 2019;

[*]Equal contribution [1]Microsoft, Cambridge, UK [2]University College London, London, UK. Correspondence to: Santiago Zanella-Béguelin <santiago@microsoft.com>, Lukas Wutschitz <luwutsch@microsoft.com>.

*Proceedings of the $40^{th}$ International Conference on Machine Learning*, Honolulu, Hawaii, USA. PMLR 202, 2023. 

Jagielski et al., 2020; Malek et al., 2021; Nasr et al., 2021). Existing approaches compute a confidence interval for the privacy budget $\hat{\varepsilon}$ spent by a training pipeline from estimates of the false positive and false negative rates of membership inference attacks against models trained using it. Practitioners can then make informed decisions based on $\hat{\varepsilon}$ rather than $\varepsilon$ and adjust training hyperparameters accordingly.

However, existing approaches have statistical and computational limitations that prevent their broader applicability.

1. On the statistical side, current approaches bound the false positive and false negative rates separately using Clopper-Pearson (CP) confidence intervals. We show that this largely underestimates coverage (see Figure 3) and so requires a large sample size to draw high confidence conclusions. In fact, for sample sizes considered in prior work, confidence intervals for $\hat{\varepsilon}$ derived from CP intervals are often so wide that they include both 0 and provable upper bounds for DP models (Nasr et al., 2021, Fig. 1), i.e., they do not support drawing more informed conclusions.

2. On the computational side, the (typically thousands of) samples required for statistical estimation need to be drawn from independently trained models. This makes it challenging to scale the approach to large models, or to a large number of models, as required for architecture search or hyperparameter tuning.

**Bayesian Approach** To overcome these limitations, we propose a novel Bayesian approach that is more precise and thus requires fewer samples to obtain meaningful estimates. In line with prior art (Jagielski et al., 2020; Nasr et al., 2021), we derive an estimate $\hat{\varepsilon}$ from estimates of the false positive and false negative rates of membership inference attacks. Unlike previous approaches that derive estimates from *separate* confidence intervals for each rate, we model their *joint distribution*. Exploiting the hypothesis testing interpretation of differential privacy, we use this joint distribution to compute a posterior distribution for $\hat{\varepsilon}$, from which we derive significantly tighter credible intervals.

**End-to-End System for Privacy Estimation** To address computational challenges we implement a modular end-to-end system for privacy estimation that incorporates many conveniences and optimizations, including 1. parallelization of model training and attacks, 2. caching of models and intermediate results including attack scores.

Given a training pipeline, membership inference attack, and desired confidence level, the system selects challenge points for membership inference, trains the required models in parallel, runs the attack, and produces a confidence interval for $\hat{\varepsilon}$. We implement the system as an Azure ML pipeline, allowing for an efficient utilization of large GPU clusters, but the system can make use of more modest resources and its design is generic enough to be ported to any other ML framework. The design enables swapping modules, e.g., using a white-box rather than a black-box attack to reflect a threat model where models are deployed on clients devices rather than on the cloud. Plugging in our Bayesian approach and state-of-the-art black-box membership inference attacks (Carlini et al., 2022) into this system brings privacy estimation within the reach of practitioners.

**Evaluation** We first demonstrate the gains of our Bayesian approach via a numerical simulation:

• We compare the sample size required for a desired confidence. For this, we align equal-tailed credible intervals for $\hat{\varepsilon}$ obtained using the Bayesian approach with confidence intervals derived from Clopper-Pearson and Jeffreys intervals for false positive and false negative rates. The comparison shows that a Bayesian approach enables us to draw conclusions that are as significant as prior work with only a fraction of the samples. For example, for an estimate within $\pm 0.15$ with $90\,\%$ confidence, our approach reduces the number of samples required from approximately $1500$ to just $500$.

• We compare confidence interval width varying attack accuracy for a fixed sample size, showing that a Bayesian approach provides the narrowest intervals for all FPR and FNR combinations. For $1000$ samples, our approach reduces the interval size by up to $32\,\%$ compared to Jeffreys and up to $52\,\%$ compared to Clopper-Pearson intervals.

• We evaluate our system for privacy estimation on text (SST-2) and vision (CIFAR-10) classification, where we observe a reduction in interval width of up to $\mathbf{40}\,\%$ w.r.t. prior work for a fixed $(1000)$ number of samples. These results confirm the gains observed using numeric simulation. Figure 1 illustrates the gains for CIFAR-10 and $\varepsilon_{\text{th}} = 10$.

## 2. Preliminaries

In this section we introduce notation, recall the definition of $(\varepsilon, \delta)$-differential privacy and its hypothesis testing interpretation, and overview membership inference attacks and their relation to differential privacy.

### 2.1. Notation

We use calligraphic font for randomized algorithms (e.g., $\mathcal{T}$) and distributions (e.g., $\mathcal{D}$), and uppercase serif font for lists and sets (e.g., $S$). We use $z \sim \mathcal{D}$ to denote an example $z$ drawn from $\mathcal{D}$ and $S \sim \mathcal{D}^n$ to denote a list $S$ of $n$ examples independently drawn from $\mathcal{D}$. $b \sim \{0,1\}$ denotes a fair coin flip, i.e., a bit $b$ sampled uniformly from $\{0,1\}$. Adversary algorithms (e.g., $\mathcal{A}_1, \mathcal{A}_2$) are randomized procedures that share mutable state, although for clarity we often include re-

dundant arguments. We formalize probabilistic experiments as sequential pseudocode and write $\Pr[\mathrm{Exp}(\cdots) : A]$ for the probability of event $A$ in experiment Exp. Table 1 summarizes the notation used throughout the paper.

*Table 1.* Summary of notation

| Notation | Description |
|---|---|
| $\mathcal{T}$ | A stochastic training algorithm |
| $\mathcal{D}$ | Distribution over samples |
| $\mathcal{D}^n$ | Distribution of $n$ independent samples from $\mathcal{D}$ |
| $\mathcal{A}, \mathcal{A}_1, \mathcal{A}_2$ | Adversary procedures sharing mutable state |
| $z \sim \mathcal{D}$ | Draw an example $z$ from $\mathcal{D}$ |
| $S \sim \mathcal{D}^n$ | Draw $n$ examples $S$ independently from $\mathcal{D}$ |
| $b \sim \{0, 1\}$ | Sample a bit $b$ uniformly |
| $y \leftarrow \mathcal{P}(\vec{x})$ | Call $\mathcal{P}$ with arguments $\vec{x}$ and assign result to $y$ |

### 2.2. Approximate Differential Privacy

**Definition 2.1** (Approximate Differential Privacy)**.** Let $\varepsilon > 0$ and $\delta \in [0, 1]$. A mechanism $\mathcal{T} : X \to Y$ is $(\varepsilon, \delta)$-differentially private with respect to an adjacency relation $\sim$ on $X$ if for any $D_0 \sim D_1$ and any $O \subseteq Y$,

$$\Pr[\mathcal{T}(D_0) \in O] \leq e^{\varepsilon} \Pr[\mathcal{T}(D_1) \in O] + \delta.$$

The mechanisms that we study are machine learning training algorithms $\mathcal{T}$ that stochastically map a dataset $S$ of examples from $X$ to model weights $\theta$. We refer to $S$ as the training dataset of $\theta$, which is typically composed of i.i.d. examples drawn from some underlying distribution $\mathcal{D}$ with support $X$. We use the *add/remove one* adjacency relation: two training datasets are adjacent if one can be obtained from the other by adding or removing a single example. This corresponds to *unbounded differential privacy* (Kifer & Machanavajjhala, 2011).

### 2.3. Hypothesis Testing Characterization of DP

Consider a run of a mechanism $\mathcal{T} : X \to Y$ that outputs some $y \in Y$ when given one of two adjacent inputs $D_0, D_1$. We can recast the differential privacy of $\mathcal{T}$ as a hypothesis test where the null hypothesis is that the input was $D_0$ and the alternative hypothesis is that it was $D_1$. A deterministic test rejects the null hypothesis when $y$ is in a rejection region $R$. A Type-I error (false positive) occurs when the null hypothesis is true but is rejected, with probability $\Pr[\mathcal{T}(D_0) \in R]$. A Type-II error (false negative) occurs when the null hypothesis is false but is not rejected, with probability $\Pr\left[\mathcal{T}(D_1) \in \overline{R}\right]$.

The following theorem from (Kairouz et al., 2017) characterizes $(\varepsilon, \delta)$-differential privacy in terms of conditions on the false positive and false negative rates of hypothesis tests.

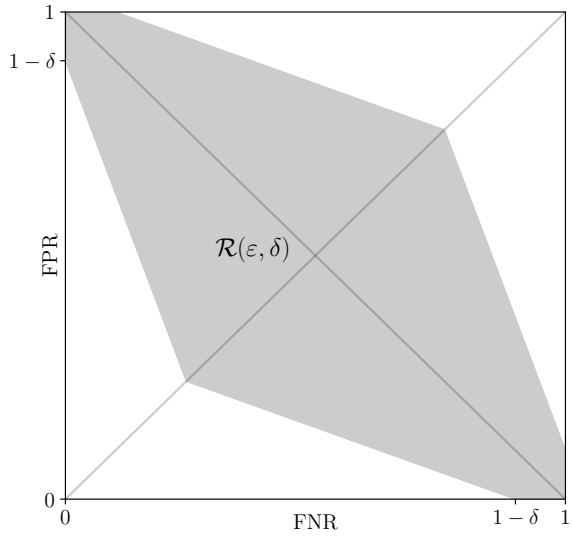



*Figure 2.* Privacy region $\mathcal{R}(\varepsilon, \delta)$. The region grows with $\varepsilon$ and covers the unit square as $\varepsilon$ tends towards $\infty$.

This extends an earlier result from (Hall et al., 2013) that only shows that the conditions are necessary.

**Theorem 2.2.** *A mechanism $\mathcal{T} : X \to Y$ is $(\varepsilon, \delta)$-differentially private if and only if for all adjacent inputs $D_0 \sim D_1$ and all $R \subseteq Y$, the following conditions are met*

$$\Pr[\mathcal{T}(D_0) \in R] + \mathrm{e}^{\varepsilon} \Pr\left[\mathcal{T}(D_1) \in \overline{R}\right] \geq 1 - \delta\ ,$$
$$\Pr\left[\mathcal{T}(D_1) \in \overline{R}\right] + \mathrm{e}^{\varepsilon} \Pr[\mathcal{T}(D_0) \in R] \geq 1 - \delta\ .$$

A distinguisher that observes the output of an $(\varepsilon, \delta)$-differentially private mechanism $\mathcal{T}$ and makes a guess as to which hypothesis is true implicitly defines a rejection region. The set of false positive and false negative rates achievable by distinguishers, or equivalently, the set of Type-I and Type-II errors for any rejection region must be included in the *privacy region* $\mathcal{R}(\varepsilon, \delta)$, defined as follows:

$$\mathcal{R}(\varepsilon, \delta) = \{(x, y) \,|\, x + \mathrm{e}^{\varepsilon} y \geq 1 - \delta \wedge y + \mathrm{e}^{\varepsilon} x \geq 1 - \delta \wedge y + \mathrm{e}^{\varepsilon} x \leq \mathrm{e}^{\varepsilon} + \delta \wedge x + \mathrm{e}^{\varepsilon} y \leq \mathrm{e}^{\varepsilon} + \delta\}\ .$$

Figure 2 illustrates the privacy region $\mathcal{R}(\varepsilon, \delta)$. It is symmetric w.r.t. the FNR $= 1 -$ FPR line because if a rejection region $Y$ achieves (FNR, FPR), its complement $\overline{Y}$ achieves $(1 - \text{FNR}, 1 - \text{FPR})$. It is symmetric w.r.t. the FNR $=$ FPR line because the adjacency relation is symmetric and so positive and negative instances are interchangeable.

### 2.4. Privacy Estimates from Membership Inference

Membership inference attacks (MIA) try to determine whether samples belong to the training dataset of a model. Yeom et al. (2018) formalize membership inference with balanced priors as a game equivalent to Experiment 1 below.

**Experiment 1:** MIA

**Input:** $\mathcal{T}, \mathcal{D}, n, \mathcal{A}$
$S \sim \mathcal{D}^{n-1}$; $z_0, z_1 \sim \mathcal{D}^2$
$S_0, S_1 \leftarrow S \cup \{z_0\}, S \cup \{z_1\}$
$b \sim \{0, 1\}$
$\theta \leftarrow \mathcal{T}(S_b)$
$\tilde{b} \leftarrow \mathcal{A}(\mathcal{T}, \mathcal{D}, n, \theta, z_0)$

Experiment 2 adapts this to the unbounded DP setting using the *add/remove one* adjacency relation, and considers general DP distinguishers which choose the base training dataset $S$ and the challenge $z$.

**Experiment 2:** IND-MIA

**Input:** $\mathcal{T}, \mathcal{D}, n, \mathcal{A}$
$S, z \leftarrow \mathcal{A}_1(\mathcal{T}, \mathcal{D}, n)$ // $|S| = n - 1$
$S_0, S_1 \leftarrow S, S \cup \{z\}$
$b \sim \{0, 1\}$
$\theta \leftarrow \mathcal{T}(S_b)$
$\tilde{b} \leftarrow \mathcal{A}_2(\mathcal{T}, \mathcal{D}, n, \theta, S, z)$

A MIA such as Experiment 2 defines a hypothesis test with false negative and false positive rates

$$\text{FNR} \coloneqq \Pr\left[\text{IND-MIA} : \tilde{b} = 0 \mid b = 1\right],$$
$$\text{FPR} \coloneqq \Pr\left[\text{IND-MIA} : \tilde{b} = 1 \mid b = 0\right].$$

We use this interpretation to bound the empirical privacy parameter $\hat{\varepsilon}$ of a training algorithm for a fixed $\delta$. The key idea is that any pair (FNR, FPR) serves as a counterexample for the training pipeline being $(\varepsilon, \delta)$-differentially private for any $\varepsilon$ such that $(\text{FNR}, \text{FPR}) \notin \mathcal{R}(\varepsilon, \delta)$. So, a lower bound for $\hat{\varepsilon}$ is given by

$$\hat{\varepsilon}_- = \inf\{\varepsilon \in \mathbb{R}^+ \mid (\text{FNR}, \text{FPR}) \in \mathcal{R}(\varepsilon, \delta)\}$$

Assuming $\text{FNR}, \text{FPR} \neq 0$ and $\text{FNR}, \text{FPR} \leq 1 - \delta$, this is

$$\hat{\varepsilon}_- = \max\left\{\log\frac{1-\delta-\text{FPR}}{\text{FNR}}, \log\frac{1-\delta-\text{FNR}}{\text{FPR}}\right\} \quad (1)$$

### 2.5. Privacy Estimates from Confidence Intervals

Previous work (Carlini et al., 2021; Jagielski et al., 2020) uses a Monte Carlo approach to estimate FPR and FNR with Clopper-Pearson confidence intervals and then uses these to estimate $\hat{\varepsilon}$. Given samples $\{b_i, \tilde{b}_i\}$ from runs of Experiment 2, the first step is to obtain estimates and intervals for FPR and FNR:

$$\overline{\text{FPR}} = \frac{\sum_{i=1}^m \left[\tilde{b}_i \neq b_i \wedge b_i = 0\right]}{\sum_{i=1}^m [b_i = 0]} \in [\text{FPR}_-, \text{FPR}_+]$$
$$\overline{\text{FNR}} = \frac{\sum_{i=1}^m \left[\tilde{b}_i \neq b_i \wedge b_i = 1\right]}{\sum_{i=1}^m [b_i = 1]} \in [\text{FNR}_-, \text{FNR}_+]$$

A lower bound for $\hat{\varepsilon}$ can be computed minimizing Equation (1) over these confidence intervals (where the terms are well-defined).[1] An upper bound $\hat{\varepsilon}_+$ can be computed analogously, but is less interesting since it does not bound the privacy afforded by the training pipeline w.r.t. more powerful adversaries.

From the union bound, the significance of the confidence interval for $\hat{\varepsilon}$ is double the significance of the confidence intervals for $\overline{\text{FPR}}$ and $\overline{\text{FNR}}$ used to derive it. For instance, when using $95\,\%$ confidence intervals for $\overline{\text{FPR}}$ and $\overline{\text{FNR}}$, the derived confidence interval $[\hat{\varepsilon}_-, \hat{\varepsilon}_+]$ has $90\,\%$ confidence.

### 2.6. Clopper-Pearson Confidence Intervals

Sample false negative (FN) and false positive counts (FP) can be modeled as the number of successes of two binomial distributions with respective unknown success probabilities FNR and FPR. Given $k$ observed successes in $N$ trials, the lower and upper limits of the two-sided $100(1-\alpha)\,\%$ Clopper-Pearson interval are respectively the solutions $p$ to the equations $\Pr[\text{Bin}(N, p) \geq k] = \alpha/2$ and $\Pr[\text{Bin}(N, p) \leq k] = \alpha/2$. The interval can be succinctly written in terms of quantiles of Beta distributions as $[\text{B}(\alpha/2, k, N-k+1), \text{B}(1-\alpha/2, k+1, N-k)]$, where $\text{B}(q, a, b)$ is the $q$ quantile of $\text{Beta}(a, b)$.

Clopper-Pearson intervals are guaranteed to reach nominal coverage. However, they typically exceed it, which results in privacy estimates that are overly conservative. We next present a Bayesian approach that addresses this problem.[2]

## 3. A Bayesian Approach to Privacy Estimates

In this section we present a novel Bayesian approach to privacy estimates that models false positive and false negative rates as independent binomial proportions with non-

[1] Carlini et al. (2021, Eq. 5) simply take the value at $(\text{FNR}_+, \text{FPR}_+)$, but special care should be taken when either $\text{FNR}_-$ or $\text{FPR}_-$ is 0 as the minimum can occur at e.g., $(\text{FNR}_+, 0)$. For example, when TP, FP, TN, FN = (90, 0, 100, 10) and $\delta = 10^{-5}$ using $90\,\%$ Clopper-Pearson intervals, the value of Eq. (1) at $(\text{FNR}_+, \text{FPR}_+)$ is 3.124, while the minimum 1.736 occurs at $(\text{FNR}_+, 0)$.

[2] An obvious improvement over the state-of-the-art approach to lower bound $\hat{\varepsilon}$ (Carlini et al., 2021) is to use one-sided Clopper-Pearson intervals since $\hat{\varepsilon}_-$ only depends on their upper-limit ($\text{FPR}_-$ and $\text{FNR}_-$ are only 0 when FP or FN are exactly 0). This effectively halves the significance of estimates.

informative Jeffreys priors. We first present Jeffreys intervals, derived from the same model, as an alternative to Clopper-Pearson intervals. We then present a much more precise method that directly computes credible intervals from the posterior distribution of (FNR, FPR).

### 3.1. Jeffreys Intervals

Jeffreys intervals have roots in Bayesian analysis, achieve good probability matching properties, and are particularly recommended as one-sided intervals (Tony Cai, 2005, p.68). Their Bayesian derivation uses a non-informative conjugate prior for the binomial proportion $p$, resulting in the model

$$\begin{aligned} p &\sim \text{Beta}(1/2, 1/2) \\ k|p &\sim \text{Bin}(N, p) \\ p|k &\sim \text{Beta}(1/2 + k, 1/2 + N - k) \end{aligned} \tag{2}$$

The upper-limit of the one-sided $100(1-\alpha)\,\%$ Jeffreys interval is the $1-\alpha$ quantile of the posterior $p|k$, that is $\text{B}(1-\alpha, 1/2+k, 1/2+N-k)$. When $k=0$ the lower limit is set to 0 and when $k=N$ the upper limit is set to 1 to avoid the coverage tending to 0 as $p$ tends to 0 or 1.

Using the techniques of Nasr et al. (2021), one-sided Jeffreys intervals for $\overline{\text{FPR}}$ and $\overline{\text{FNR}}$ yield narrower confidence intervals for $\hat{\varepsilon}$ than previous approaches using two-sided Clopper-Pearson intervals. For instance, an attack with perfect accuracy over 2000 trials with $\delta = 10^{-5}$ results in a $90\,\%$ confidence $\hat{\varepsilon}_-$ of 5.6 using two-sided CP intervals, 5.81 using one-sided CP intervals, and 6.25 using one-sided Jeffreys intervals.[3]

### 3.2. Estimates from the Posterior Joint Distribution

We show how to improve estimates using the joint posterior of (FNR, FPR) to derive a credible interval for $\hat{\varepsilon}$.

**Intuition** Figure 3 provides an intuitive graphical explanation in the (FNR,FPR) space of the advantage of using the joint posterior for estimating $\hat{\varepsilon}$. The blue rectangle is given by Jeffreys confidence intervals for the false positive and false negative rate of a membership inference attack. This rectangle covers $1-\alpha$ of the density of the joint distribution of (FNR,FPR), but it fits in-between two privacy regions whose difference covers strictly more density. A $100(1-\alpha)\,\%$ confidence interval for $\hat{\varepsilon}$ derived using this method will have larger than nominal coverage because the additional density in $\mathcal{R}(\hat{\varepsilon}'_+, \delta) \setminus \mathcal{R}(\hat{\varepsilon}'_-, \delta)$ outside the rectangle is unaccounted for. Instead, we integrate the probability density $f_{(\text{FNR,FPR})}$ over the exact area between regions and derive a credible interval for $\hat{\varepsilon}$ with nominal coverage.

[3] Carlini et al. (2021) report the first figure of 5.6 for 1000 trials, but it clearly is only achievable with 2000 trials.

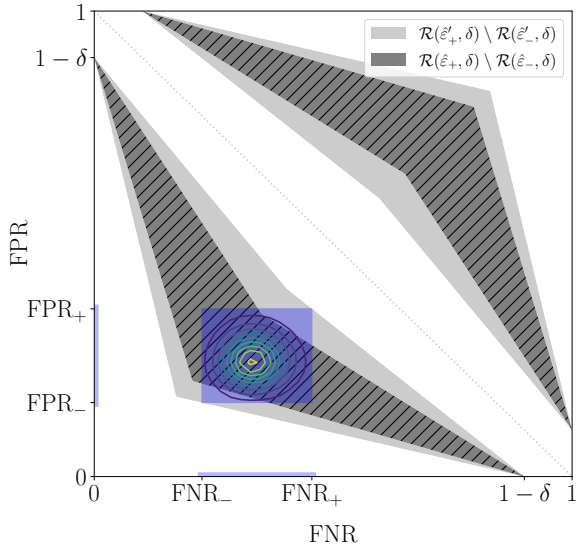



*Figure 3.* Graphical interpretation of intervals for $\hat{\varepsilon}$ obtained using a joint binomial model ($[\hat{\varepsilon}_-, \hat{\varepsilon}_+]$) and Jeffreys confidence intervals ($[\hat{\varepsilon}'_-, \hat{\varepsilon}'_+]$). The contour plot of the density $f_{(\text{FNR,FPR})}$ and the rectangle determined by Jeffreys intervals match closely.

**Computation** Given the probability density function $f_{(\text{FNR,FPR})}$ of the joint posterior of (FNR, FPR), we obtain the cumulative distribution of $\hat{\varepsilon}$.

**Definition 3.1** (Cumulative Distribution Function of $\hat{\varepsilon}$)**.** Let $\delta \in [0,1]$ and $f_{(\text{FNR,FPR})}$ be the density function of the posterior joint distribution of (FNR, FPR) given observed counts of FN, TP, FP, TN from Experiment 2. The value of the cumulative distribution function of $\hat{\varepsilon}$ at $\varepsilon$ is the integral of $f_{(\text{FNR,FPR})}$ over the privacy region $\mathcal{R}(\varepsilon, \delta)$:

$$\begin{aligned} F_{\hat{\varepsilon}}(\varepsilon) &= \Pr\left[(\text{FNR}, \text{FPR}) \in \mathcal{R}(\varepsilon, \delta)\right] \\ &= \iint_{\mathcal{R}(\varepsilon,\delta)} f_{(\text{FNR,FPR})}(x, y)\, \mathrm{d}x\, \mathrm{d}y\ . \end{aligned} \tag{3}$$

Equipped with $F_{\hat{\varepsilon}}$ we can compute the $100(1-\alpha)\%$ equal-tailed *credible interval* $[\hat{\varepsilon}_-, \hat{\varepsilon}_+]$ given by:

$$\hat{\varepsilon}_- = \sup\{\varepsilon \mid F_{\hat{\varepsilon}}(\varepsilon) \leq \alpha/2\}\ , \tag{4}$$

$$\hat{\varepsilon}_+ = \inf\{\varepsilon \mid F_{\hat{\varepsilon}}(\varepsilon) \geq 1-\alpha/2\}\ . \tag{5}$$

The Bayesian model we presented above in Equation (2) gives us the densities of the posteriors FNR|FN and FPR|FP. Since the populations of positive and negative instances are independent, it is natural to model these posteriors as independent, yielding a joint distribution we can plug into Equation (3):

$$f_{(\text{FNR,FPR})}(x, y) \coloneqq f_{\text{FNR}|\text{FN,TP}}(x)\ f_{\text{FPR}|\text{FP,TN}}(y) \tag{6}$$

The resulting integral in Equation (3) cannot be expressed

in analytical form so we approximate it numerically using SciPy's `dblquad`, based on QUADPACK's `qagse`.

**Example** Suppose we run 200 times Experiment 2, collecting samples $\{b_i, \tilde{b}_i\}$ with a tally $\text{FN} = 35, \text{TP} = 65, \text{FP} = 25, \text{TN} = 75$. To derive a 95 % credible interval for $\hat{\varepsilon}$ at $\delta = 0.05$, we construct the cumulative distribution function of $\hat{\varepsilon}$ by integrating $f_{(\text{FNR},\text{FPR})}$ and solve Equations (4) and (5), which yields the interval $[0.522, 1.268]$. In contrast, deriving a 95 % confidence interval by taking the minimum and maximum of Equation (1) over the rectangle defined by the two-sided Jeffreys intervals for FNR and FPR yields a much larger interval $[0.321, 1.456]$. (Clopper-Pearson intervals will give an even larger interval $[0.295, 1.489]$.) In terms of Figure 3, the rectangle covers 96.3 % of the density of $f_{(\text{FNR},\text{FPR})}$, but is enclosed in an area in-between two privacy regions covering 99.8%. So, the actual coverage of the interval computed from independent Jeffreys intervals is much larger than desired. In comparison, the smaller hatched area corresponding to the Bayesian credible interval has 95 % coverage by definition.

### 3.3. Summary

We discussed three ways to estimate $\hat{\varepsilon}_-$ at significance $\alpha$ from the confusion matrix $(\text{FN}, \text{TP}, \text{FP}, \text{TN})$ of an attack, obtained from multiple runs of Experiment 2:

1. From Clopper-Pearson confidence intervals for FNR and FPR (as proposed in prior work);
2. From Jeffreys confidence intervals for FNR and FPR (as a drop-in replacement for CP intervals);
3. From the joint distribution of FNR and FPR.

Below we combine the various steps in self-contained expressions for $\hat{\varepsilon}_-$ given $\delta$ using the first and third methods.

**Estimation from Clopper-Pearson Intervals** Ignoring corner cases, the first method yields the following expression for $\hat{\varepsilon}_-$ in terms of $(\text{FN}, \text{TP}, \text{FP}, \text{TN})$, $\alpha$, and $\delta$:

$$\hat{\varepsilon}_- = \max \left\{ \begin{array}{l} \log \dfrac{1-\delta-B(1-\alpha/2, \text{FN}+1, \text{TP})}{B(1-\alpha/2, \text{FP}+1, \text{TN})}, \\ \log \dfrac{1-\delta-B(1-\alpha/2, \text{FP}+1, \text{TN})}{B(1-\alpha/2, \text{FN}+1, \text{TP})} \end{array} \right\}$$

where $\text{B}(q, a, b)$ is the percent point function of $\text{Beta}(a, b)$ (i.e., the inverse of its cdf). This expression is obtained by evaluating Equation (1) at the upper limit of one-sided Clopper-Pearson intervals for FPR and FNR with significance $\alpha/2$. The significance must be halved because one needs to apply the union bound to combine the intervals.

**Estimation from Joint Distribution** The third method yields $\hat{\varepsilon}_- = F_{\hat{\varepsilon}}^{-1}(\alpha)$, where

$$F_{\hat{\varepsilon}}(\varepsilon) = \iint_{\mathcal{R}(\varepsilon,\delta)} \frac{g(1/2+\text{FN}, 1/2+\text{TP}, x)}{g(1/2+\text{FP}, 1/2+\text{TN}, x)} \, \mathrm{d}x \, \mathrm{d}y$$

and $g(a, b, x)$ is the pdf of $\text{Beta}(a, b)$. This expression is obtained from Equation (3) by modelling the joint pdf of FNR,FPR as in Equation (6) and the posteriors $\text{FNR}|\text{FN}, \text{TP}$ and $\text{FPR}|\text{FP}, \text{TN}$ as in Equation (2).

**About Closed Forms** While the third method requires computing an integral with no closed form, the first method yields a closed form. However, this *closed form* expression involves the inverse of the cdf of Beta distributions. This requires evaluating the inverse of the regularized incomplete Beta function, which must also be done numerically. An analytical comparison of both methods is a significant endeavor which would require computing expansions and bounding error terms. We provide below a numerical comparison.

### 3.4. Numeric Evaluation of the Bayesian Approach

We evaluate the performance of our Bayesian approach in numeric simulations. For this, we compare equal-tailed credible intervals for $\hat{\varepsilon}$ obtained using our new Bayesian approach with confidence intervals for $\hat{\varepsilon}$ derived from two-sided Clopper-Pearson and Jeffreys intervals.

**Varying Number of Samples** We assume a hypothetical attack with a balanced accuracy of 60%, from which we derive FPR and FNR. We evaluate the reduction in uncertainty by comparing confidence interval sizes for $\hat{\varepsilon}$ (assuming a fixed $\delta$) for different numbers of samples using Clopper-Pearson intervals, Jeffreys intervals, and our Bayesian approach. We also quantify the improvement in computational cost by comparing the number of samples required to achieve a given confidence interval ($\pm 0.15$) using the different methods.

Figure 4 shows the results of this comparison. Here we are interested in an estimate for $\hat{\varepsilon}$ within $\pm 0.15$ with a significance level of $\alpha = 10\%$. The Clopper-Pearson approach requires approximately 1500 samples. Jeffreys intervals marginally reduce the number of samples. Using our Bayesian approach, we can significantly reduce the number of samples to just over 500 thereby reducing the computational cost by ⅔.

**Varying Attack Accuracy** We assume a fixed number of samples (1000) and a suite of hypothetical attacks with varying attack accuracy, from which we derive FPR and FNR. We evaluate the reduction in uncertainty by comparing the confidence interval sizes for $\hat{\varepsilon}$ (assuming a fixed $\delta$).

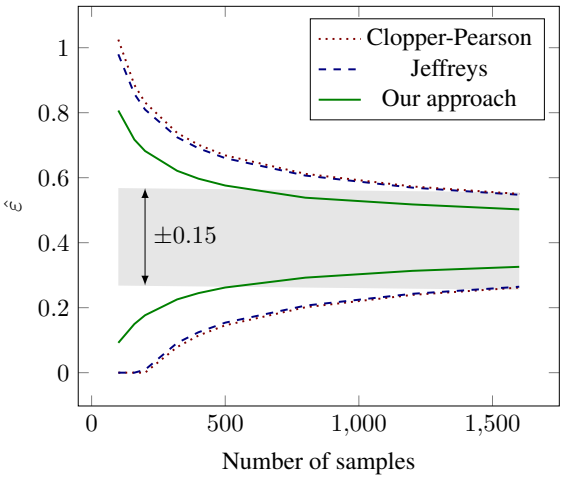



*Figure 4.* Interval endpoints as a function of number of samples. We compare three estimation techniques at 90 % confidence ($\alpha = 0.1$) for an attack with FPR = FNR = 40 %. For a given number of samples, the distance between lower and upper endpoints illustrates the reduction in interval width. For a given interval width (e.g., as illustrated by the shaded area), the difference in the $x$-coordinates at which the curves intersect with the shaded area illustrates the reduction in the number of samples. The shaded area illustrates an estimate $\hat{\varepsilon}$ within $\pm 0.15$ with 90 % confidence.

Compared to either Clopper-Pearson or Jeffreys intervals, our Bayesian approach provides the narrowest interval for *all* combinations of FPR and FNR. For each of the three methods, Figure 5 shows the interval width as a function of FPR and FNR. We observed the most prevalent differences at the corners of the region (e.g. at low FPR and high FNR), which is also the target area for recent membership inference attacks (Carlini et al., 2022). For the FPR and FNR ranges in Figure 5, our approach decreases interval width by 20 % to 32 % compared to Jeffreys and 36 % to 52 % compared to Clopper-Pearson intervals.

## 4. End-to-End Privacy Estimation

In this section, we present a modular end-to-end system for estimating $\hat{\varepsilon}$. Figure 6 illustrates how data flows through the different components. We discuss next the key components of the system and how they could be implemented. The experimental evaluation in Section 4.1 uses a particular implementation that plugs in our Bayesian estimation method and a state-of-the-art black-box membership inference attack to measure *worst-case* privacy of text and vision classifiers. The implementation of the core Bayesian estimation method used to produce the results reported in the paper can be found at `https://aka.ms/privacy-estimates`.

**Selecting Challenge Points** Model training pipelines have access to a limited amount of data that must be partitioned to train, validate, and test models. Accordingly, we estimate privacy w.r.t. to a specific dataset $D$, limiting the universe of possible adjacent datasets in the definition of differential privacy.[4] Since we are mostly interested in measuring the empirical DP budget spent by a pipeline w.r.t. to a given attack, we select challenge points that maximize the attack performance. To do this, we train $M$ shadow models with a random split of $D$. For each point $z \in D$ and shadow model $\theta$, we compute the confidence score that the attack assigns for $z$ being a member of the training dataset of $\theta$, and aggregate these scores and the ground truth membership information to obtain a weight for $z$ (e.g. the accuracy of the attack). Algorithm 3 in Appendix A.1 shows pseudocode for this challenge point selection procedure.

**Choosing a Membership Inference Attack** The system is parametric on the membership inference attack and can be used together with both black-box and white-box attacks. In practice, the attack module should be chosen based on the relevant threat model. For example, if models are deployed on the cloud behind an API, a black-box attack is appropriate, while a white-box attack would be more appropriate when models are deployed on untrusted devices.

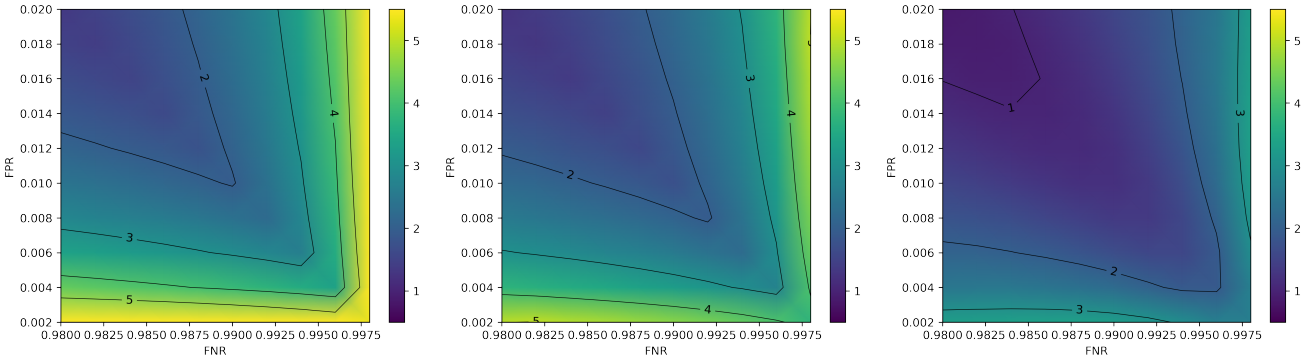
**Privacy Estimation** To estimate $\hat{\varepsilon}$ at a desired significance $\alpha$, we take as input the the ground truth membership information (the challenge bits $b$ in Experiment 2) and confidence scores from the membership inference attack. With these, we construct a ROC curve. For $N$ samples, this curve determines $N + 1$ decision thresholds for membership. We compute a lower bound $\hat{\varepsilon}_{-}$ using our Bayesian approach according to Equation (4) for each of these thresholds and select the one that yields the largest lower bound. This last step can be easily adapted to use e.g. Jeffreys intervals rather our Bayesian approach. We do this when comparing to prior methods in our experimental evaluation.

### 4.1. Implementation and Evaluation

**Choice of Attack** We use the state of the art likelihood ratio membership inference attack (LiRA) of Carlini et al. (2022). This attack requires to collect loss statistics for each training data point. In our experiments, we train 512 shadow models on random splits of the training and validation sets.

We re-use the shadow models required for the LiRA attack to find the most vulnerable training samples. We split the 512 shadow models using a train to test split of 80:20. We then use the training samples where LiRA achieves the highest accuracy on the test set as challenge points.

[4] Alternatively, one can introduce poisoned points as done by Jagielski et al. if this reflects better the threat model.




(a) Clopper-Pearson (b) Jeffreys (c) Our approach

*Figure 5.* Interval width (i.e., difference between $\hat{\varepsilon}$ interval endpoints) as a function of FPR and FNR for 1000 samples.

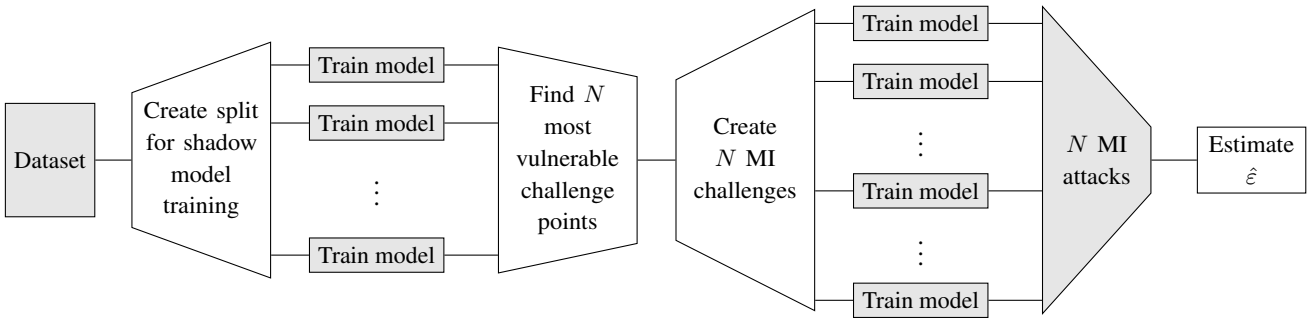



*Figure 6.* An end-to-end system for estimating privacy.

**Parallelizing Model Training and Attack Execution** We implement the end-to-end pipeline depicted in Figure 6 in Azure ML. Our pipeline requires training multiple models for i) selection of challenge points and ii) gathering membership inference guesses and, for LiRA, iii) training shadow models. We do this efficiently using Azure ML pipelines, but our original implementation used open-source libraries (Ray Tune). Although not as costly, we similarly parallelize the execution of attacks and the linear traversal over ROC curve thresholds to estimate optimal lower bounds for $\hat{\varepsilon}$.

### 4.2. Experiments on Text and Vision Classifiers

We evaluate the performance of the Bayesian approach on vision (CIFAR-10) and text (SST-2) classifiers:

- CIFAR-10 (Krizhevsky, 2009) consists of 60 000 labeled (50 000 training, 10 000 test) labeled images in 10 object classes, with 6000 images per class. We use a 4-layer CNN with 974 K parameters and $\tanh$ activations with average pooling and max pooling units, which we train for 50 epochs. We reach 65 % accuracy with $\varepsilon = 8$, $\delta = 10^{-5}$ compared to 71 % in the non-private baseline.
- SST 2 (Socher et al., 2013) is a binary sentiment text classification dataset consisting of 67 349 training samples and 1821 test samples. We fine-tune a RoBERTa base model with a classification head for 3 epochs (Liu et al., 2021). We reach 90 % accuracy with $\varepsilon = 8$, $\delta = 10^{-5}$ compared to 93 % in the non-private baseline.

Table 2 summarizes the results of this comparison on text and vision tasks using 1024 samples. We compute the width of confidence intervals using each method and the reduction in interval width relative to the Clopper-Pearson method. We observe reductions in width of between **40 % to 48 %** for the same number of samples. Importantly, our approach is successful in computing meaningful confidence intervals when other methods result in trivial $(0, \infty)$ intervals.

## 5. Related Work

**Empirical Privacy Estimates** Hyland & Tople (2019) estimate DP bounds based on an empirical estimate of the sensitivity of SGD. Jagielski et al. (2020) derive estimates from black-box membership inference attacks, using clipping-aware poisoning attacks against DP-SGD. Nasr et al. (2021) use similar techniques but consider a hierarchy of adversaries, ranging from black-box membership inference to

*Table 2.* Comparison of intervals obtained from different estimation methods for text and vision models trained with and without DP ($\delta = 10^{-5}$) using worst-case challenge points. For each method, the bounds and widths are for equal-tailed intervals at $\alpha = 0.1$.

| | | **Clopper-Pearson** | | **Jeffreys** | | | **Bayesian Approach** | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Interval | Width | Interval | Width | vs. CP | Interval | Width | vs. CP |
| **SST-2** | No DP | $(0, \infty)$ | $\infty$ | $(3.57, \infty)$ | $\infty$ | – | (4.0, 9.2) | 5.25 | **–** |
| | $\varepsilon = 8$ | $(0, \infty)$ | $\infty$ | $(0, \infty)$ | $\infty$ | – | (0.12, 8.5) | 8.4 | **–** |
| **CIFAR-10** | No DP | (1.8, 5.3) | 3.5 | (1.9, 5.0) | 3.1 | -11% | (2.2, 4.5) | 2.3 | **-40%** |
| | $\varepsilon = 8$ | (0.23, 7.7) | 7.5 | (0.43, 6.0) | 5.6 | -25% | (1.1, 5.0) | 3.9 | **-48%** |

distinguishers that craft worst-case datasets. Both works derive estimates from Clopper-Pearson confidence intervals of the false positive and false negative rates of attacks. Our Bayesian approach is applicable in the same settings and yields tighter estimates for the same number of samples.

**DP Violations** Several approaches (Bichsel et al., 2021; Ding et al., 2018) find violations of DP by constructing counterexamples (i.e., adjacent inputs together with a distinguishing test). These approaches aim to falsify a conjectured guarantee, whereas we aim to estimate an unknown guarantee for a given threat model. More fundamentally, these approaches are applicable to DP mechanisms beyond ML training but require the search space to be sufficiently constrained for the counterexample search to succeed. In contrast, we compute estimates with respect to a given class of parametrized distinguishers which allows us to run a much more efficient search over relatively small parameter space. Lu et al. (2022) build a general framework for auditing DP training extending techniques from Bichsel et al. (2021) and Jagielski et al. (2020); when applied to DP-SGD, results are comparable with the ClipBKD method of Jagielski et al. (2020).

**Membership Inference Attacks** Our approach is parametric on the choice of membership inference attack. Early membership inference attacks relied on training shadow models (Shokri et al., 2017). Threshold-based attacks were introduced by Yeom et al. (2018). Ye et al. (2021) compare different strategies to choose loss thresholds. In our evaluation, we choose model-dependent thresholds as they offer an attractive trade-off between accuracy and computational cost. Carlini et al. (2022) challenge the use of attack accuracy as a meaningful way to evaluate empirical privacy and instead propose to measure false positive rates at low false negative rates. Our evaluation shows that our Bayesian approach performs particularly well in this regime. It also obtains meaningful estimates where prior approaches would result in intervals including 0 and the known theoretical bound (see e.g., Table 2). Yaghini et al. (2022) show that different cohorts of samples can exhibit disparate vulnerability to membership inference and prove that differential privacy bounds the magnitude of the disparity. It would be interesting to study how this disparity correlates with empirical estimates of differential privacy.

**Provable DP Bounds** Since the introduction of the Moments Accountant (Abadi et al., 2016), there have been steady improvements in privacy accounting techniques, resulting in tighter and tighter privacy budget accounting for DP-SGD. However, this trend cannot continue as state-of-the-art accountants are tight (Doroshenko et al.; Gopi et al., 2021; Koskela et al., 2021). Further improvements require different algorithms such as PATE (Papernot et al., 2017), or the introduction of additional assumptions such as weaker adversary models or convexity (Chourasia et al., 2021).

## 6. Conclusion

We propose a novel Bayesian approach that yields high-confidence estimates of the differential privacy budget spent by training pipelines w.r.t. a given class of attacks. We implement an efficient end-to-end system for privacy estimation incorporating this approach in Azure ML. We demonstrate on text and image classifiers that the system gives tighter estimates than using prior approaches at a fraction of the computational cost.

One interesting direction for future work is to further reduce the computational cost of privacy estimation by utilizing heuristics that reduce the number of samples required for high-confidence estimates. For instance, Malek et al. (2021) proposed an heuristic to evaluate label-only DP that draws multiple samples from a single model. We found mixed results when attempting to adapt this heuristic to record-level DP, and we could not yet identify conditions under which it behaves consistently across datasets and attacks. Another avenue for future research is extending statistical estimates to other privacy metrics beyond $\hat{\varepsilon}$, such as membership inference advantage (Yeom et al., 2020).

## References


Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *23rd ACM SIGSAC Conference on Computer and Communications Security, CCS 2016*, pp. 308–318. ACM, 2016. doi:10.1145/2976749.2978318.

Bichsel, B., Steffen, S., Bogunovic, I., and Vechev, M. DP-Sniper: Black-box discovery of differential privacy violations using classifiers. In *42nd IEEE Symposium on Security and Privacy, S&P 2021*, pp. 391–409. IEEE Computer Society, 2021. doi:10.1109/SP40001.2021.00081.

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., and Song, D. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium, USENIX Security 2019*, pp. 267–284. USENIX Association, 2019. URL https://www.usenix.org/conference/usenixsecurity19/presentation/carlini.

Carlini, N., Deng, S., Garg, S., Jha, S., Mahloujifar, S., Mahmoody, M., Song, S., Thakurta, A., and Tramèr, F. Is private learning possible with instance encoding? In *42nd IEEE Symposium on Security and Privacy, S&P 2021*, pp. 410–427. IEEE Computer Society, 2021. doi:10.1109/SP40001.2021.00099.

Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., and Tramer, F. Membership inference attacks from first principles. In *43nd IEEE Symposium on Security and Privacy, S&P 2022*, pp. 1546–1564. IEEE Computer Society, 2022. doi:10.1109/SP46214.2022.00090.

Chourasia, R., Ye, J., and Shokri, R. Differential privacy dynamics of Langevin diffusion and noisy gradient descent. In *Advances in Neural Information Processing Systems, NeurIPS 2021*, volume 34, pp. 14771–14781. Curran Associates, Inc., 2021.

Desfontaines, D. A list of real-world uses of differential privacy. Ted is writing things, blog, Jan 2022. URL https://desfontain.es/privacy/real-world-differential-privacy.html. Accessed May 19, 2022 [Online].

Ding, Z., Wang, Y., Wang, G., Zhang, D., and Kifer, D. Detecting violations of differential privacy. In *25th ACM SIGSAC Conference on Computer and Communications Security, CCS 2018*, pp. 475–489. ACM, 2018. doi:10.1145/3243734.3243818.

Doroshenko, V., Ghazi, B., Kamath, P., Kumar, R., and Manurangsi, P. Connect the dots: Tighter discrete approximations of privacy loss distributions. *Proceedings on Privacy Enhancing Technologies*, 2022:552–570. doi:10.56553/popets-2022-0122.

Gopi, S., Lee, Y. T., and Wutschitz, L. Numerical composition of differential privacy. In *Advances in Neural Information Processing Systems, NeurIPS 2021*, volume 34, pp. 11631–11642. Curran Associates, Inc., 2021.

Hall, R., Rinaldo, A., and Wasserman, L. A. Differential privacy for functions and functional data. *J. Mach. Learn. Res.*, 14(1):703–727, 2013. URL https://jmlr.csail.mit.edu/papers/v14/hall13a.html.

Humphries, T., Oya, S., Tulloch, L., Rafuse, M., Goldberg, I., Hengartner, U., and Kerschbaum, F. Investigating membership inference attacks under data dependencies. *arXiv preprint arXiv:2010.12112 [cs.CR]*, 2020. doi:10.48550/arXiv.2010.12112.

Hyland, S. L. and Tople, S. On the intrinsic privacy of stochastic gradient descent. *arXiv preprint arXiv:1912.02919 [cs.LG]*, 2019. doi:10.48550/arXiv.1912.02919.

Jagielski, M., Ullman, J., and Oprea, A. Auditing differentially private machine learning: How private is private SGD? In *Advances in Neural Information Processing Systems, NeurIPS 2020*, volume 33, pp. 22205–22216. Curran Associates, Inc., 2020.

Jayaraman, B. and Evans, D. Evaluating differentially private machine learning in practice. In *28th USENIX Security Symposium, USENIX Security 2019*, pp. 1895–1912. USENIX Association, 2019. URL https://www.usenix.org/conference/usenixsecurity19/presentation/jayaraman.

Kairouz, P., Oh, S., and Viswanath, P. The composition theorem for differential privacy. *IEEE Transactions on Information Theory*, 63(6):4037–4049, 2017. doi:10.1109/TIT.2017.2685505.

Kifer, D. and Machanavajjhala, A. No free lunch in data privacy. In *2011 ACM SIGMOD International Conference on Management of Data, SIGMOD 2011*, pp. 193–204. ACM, 2011. doi:10.1145/1989323.1989345.

Koskela, A., Jälkö, J., Prediger, L., and Honkela, A. Tight differential privacy for discrete-valued mechanisms and for the subsampled Gaussian mechanism using FFT. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2021*, volume 130 of *Proceedings of Machine Learning Research*, pp. 3358–3366. PMLR, 2021. URL http://proceedings.mlr.press/v130/koskela21a.html.

Krizhevsky, A. Learning multiple layers of features from tiny images. Master's thesis, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Liu, Z., Lin, W., Shi, Y., and Zhao, J. A robustly optimized BERT pre-training approach with post-training. In *20th Chinese National Conference on Computational Linguistics, CCL 2021*, pp. 1218–1227. Chinese Information Processing Society of China, 2021. URL https://aclanthology.org/2021.ccl-1.108.

Lu, F., Munoz, J., Fuchs, M., LeBlond, T., Zaresky-Williams, E. V., Raff, E., Ferraro, F., and Testa, B. A general framework for auditing differentially private machine learning. In *Advances in Neural Information Processing Systems, NeurIPS 2022*, volume 35. Curran Associates, Inc., 2022. To appear.

Malek, M., Mironov, I., Prasad, K., Shilov, I., and Tramèr, F. Antipodes of label differential privacy: PATE and ALIBI. *arXiv preprint arXiv:2106.03408 [cs.LG]*, 2021. doi:10.48550/ARXIV.2106.03408.

Nasr, M., Songi, S., Thakurta, A., Papemoti, N., and Carlini, N. Adversary instantiation: Lower bounds for differentially private machine learning. In *42nd IEEE Symposium on Security and Privacy, S&P 2021*, pp. 866–882. IEEE Computer Society, 2021. doi:10.1109/SP40001.2021.00069.

Papernot, N., Abadi, M., Erlingsson, Ú., Goodfellow, I., and Talwar, K. Semi-supervised knowledge transfer for deep learning from private training data. In *5th International Conference on Learning Representations, ICLR 2017*. OpenReview.net, 2017. URL https://openreview.net/forum?id=HkwoSDPgg.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *38th IEEE Symposium on Security and Privacy, S&P 2017*, pp. 3–18. IEEE Computer Society, 2017. doi:10.1109/SP.2017.41.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. Recursive deep models for semantic compositionality over a sentiment treebank. In *2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013*, pp. 1631–1642. ACL, 2013. URL https://www.aclweb.org/anthology/D13-1170.

Song, C. and Shmatikov, V. Auditing data provenance in text-generation models. In *25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019*, pp. 196–206. ACM, 2019. doi:10.1145/3292500.3330885.

Song, S., Chaudhuri, K., and Sarwate, A. D. Stochastic gradient descent with differentially private updates. In *IEEE Global Conference on Signal and Information Processing, GlobalSIP 2013*, pp. 245–248, 2013. doi:10.1109/GlobalSIP.2013.6736861.

Tony Cai, T. One-sided confidence intervals in discrete distributions. *J. Stat. Plan. Inference*, 131(1):63–88, 2005. doi:10.1016/j.jspi.2004.01.005.

Yaghini, M., Kulynych, B., Cherubin, G., Veale, M., and Troncoso, C. Disparate vulnerability to membership inference attacks. *Proceedings on Privacy Enhancing Technologies*, 2022(1):460–480, 2022. doi:10.2478/popets-2022-0023.

Ye, J., Maddi, A., Murakonda, S. K., and Shokri, R. Enhanced membership inference attacks against machine learning models. *arXiv preprint arXiv:2111.09679 [cs.LG]*, 2021. doi:10.48550/arXiv.2111.09679.

Yeom, S., Giacomelli, I., Fredrikson, M., and Jha, S. Privacy risk in machine learning: Analyzing the connection to overfitting. In *31st IEEE Computer Security Foundations Symposium, CSF 2018*, pp. 268–282. IEEE Computer Society, 2018. doi:10.1109/CSF.2018.00027.

Yeom, S., Giacomelli, I., Menaged, A., Fredrikson, M., and Jha, S. Overfitting, robustness, and malicious algorithms: A study of potential causes of privacy risk in machine learning. *Journal of Computer Security*, 28(1):35–70, 2020.

Zanella-Béguelin, S., Wutschitz, L., Tople, S., Rühle, V., Paverd, A., Ohrimenko, O., Köpf, B., and Brockschmidt, M. Analyzing information leakage of updates to natural language models. In *27th ACM SIGSAC Conference on Computer and Communications Security, CCS 2020*, pp. 363—375. ACM, 2020. doi:10.1145/3372297.3417880.

# A. Appendix

## A.1. Omitted algorithms

**Algorithm 3:** `SelectWorst`

**Input:** $\mathcal{T}, D, n, k, M, \mathcal{A}$
**for** $i \leftarrow 1$ **to** $M$ **do**
  $S_i^+, S_i^- \leftarrow \texttt{Split}(D, n)$
  $\theta_i \leftarrow \mathcal{T}(S_i^+)$
  **foreach** $z \in D$ **do**
    $s[z][i] \leftarrow \texttt{Score}(\mathcal{A}, \theta_i, z)$
    $b[z][i] \leftarrow z \in S_i^+$
  **end**
**end**
**foreach** $z \in D$ **do**
  $w[z] \leftarrow \texttt{AggregateWeight}(s[z], b[z])$
**end**
**return** $\{z_i\}_k$ with largest $k$ aggregate weights $w$

## A.2. Probability Density Function of $\hat{\varepsilon}$

We describe here how to derive a probability density function for $\hat{\varepsilon}$. The derivative of the cumulative distribution function $F_{\hat{\varepsilon}}$ is given by

$$\hat{f}_\varepsilon(\varepsilon) = \frac{\mathrm{d}}{\mathrm{d}\varepsilon} F_\varepsilon(\varepsilon) = \frac{\mathrm{d}}{\mathrm{d}\varepsilon} \iint_{\mathcal{R}(\varepsilon,\delta)} f_{(\text{FNR,FPR})}(x,y)\mathrm{d}x\mathrm{d}y = \oint_{\partial\mathcal{R}(\varepsilon,\delta)} f_{(\text{FNR,FPR})} \boldsymbol{v}_\varepsilon \cdot \boldsymbol{n}\mathrm{d}L \tag{7}$$

where we have used Reynolds transport theorem in the last equation.

The symbol $\boldsymbol{v}_\varepsilon$ denotes the derivative of the boundary with respect to $\varepsilon$ and $\boldsymbol{n}$ is the outward pointing normal vector of a boundary element.

In order to make this more concrete, let us parameterize the boundary of the privacy region using the following curves

$$\boldsymbol{R}_{\text{LO}}(\varepsilon,\delta,x) := \begin{bmatrix} x \\ \max\{0, 1-\delta-\mathrm{e}^\varepsilon x, (1-\delta-x)\mathrm{e}^{-\varepsilon}\} \end{bmatrix}$$
$$\boldsymbol{R}_{\text{HI}}(\varepsilon,\delta,x) := \begin{bmatrix} x \\ \min\{1, (\delta-x)\mathrm{e}^{-\varepsilon}, \delta+(1-x)\mathrm{e}^\varepsilon\} \end{bmatrix} .$$

Note that $\partial\mathcal{R}(\varepsilon,\delta) = \boldsymbol{R}_{\text{LO}}(\varepsilon,\delta,[0,1]) \cup \boldsymbol{R}_{\text{HI}}(\varepsilon,\delta,[0,1])$.

Applying this to compute $\boldsymbol{v}_\varepsilon$ and $\boldsymbol{n}$ from Eq. (7) gives

$$\boldsymbol{v}_{\text{LO}} = \partial_\varepsilon \boldsymbol{R}_{\text{LO}}(\varepsilon,\delta,x) , \tag{8}$$
$$\boldsymbol{n}_{\text{LO}} = \frac{Q\partial_x \boldsymbol{R}_{\text{LO}}(\varepsilon,\delta,x)}{\|\partial_x \boldsymbol{R}_{\text{LO}}(\varepsilon,\delta,x)\|} , \tag{9}$$

where $Q$ denotes a rotation matrix performing a clockwise rotation by $\pi/2$. Similarly, we have

$$\boldsymbol{v}_{\text{HI}} = \partial_\varepsilon \boldsymbol{R}_{\text{HI}}(\varepsilon,\delta,x) , \tag{10}$$
$$\boldsymbol{n}_{\text{HI}} = \frac{Q\partial_{-x} \boldsymbol{R}_{\text{HI}}(\varepsilon,\delta,x)}{\|\partial_x \boldsymbol{R}_{\text{LO}}(\varepsilon,\delta,x)\|} . \tag{11}$$

We can then plug this expression into Eq. (7)). Splitting the closed line integral into an integral over the upper and lower path gives

$$\hat{f}_\varepsilon(\varepsilon) = \int_0^1 f_{(\text{FNR,FPR})}(\boldsymbol{R}_{\text{LO}}(\varepsilon,\delta,x)) \boldsymbol{v}_{\text{LO}} \cdot \boldsymbol{n}_{\text{LO}}\mathrm{d}x + \int_1^0 f_{(\text{FNR,FPR})}(\boldsymbol{R}_{\text{HI}}(\varepsilon,\delta,x)) \boldsymbol{v}_{\text{HI}} \cdot \boldsymbol{n}_{\text{HI}}\mathrm{d}x . \tag{12}$$

Note, however that $\hat{f}_\varepsilon$ is not a probability density function since it is not normalized. The mass of the privacy region for $\varepsilon = 0$ is missing: $\int_0^\infty f_\varepsilon(\varepsilon)\mathrm{d}\varepsilon = 1 - F_\varepsilon(0) \neq 1$. We can correct for that by adding a point mass at $\varepsilon = 0$ which gives a final expression for the probability density of $\varepsilon$

$$f_\varepsilon(\varepsilon) = F_\varepsilon(0)\delta(\varepsilon) + \hat{f}_\varepsilon(\varepsilon) \ , \tag{13}$$

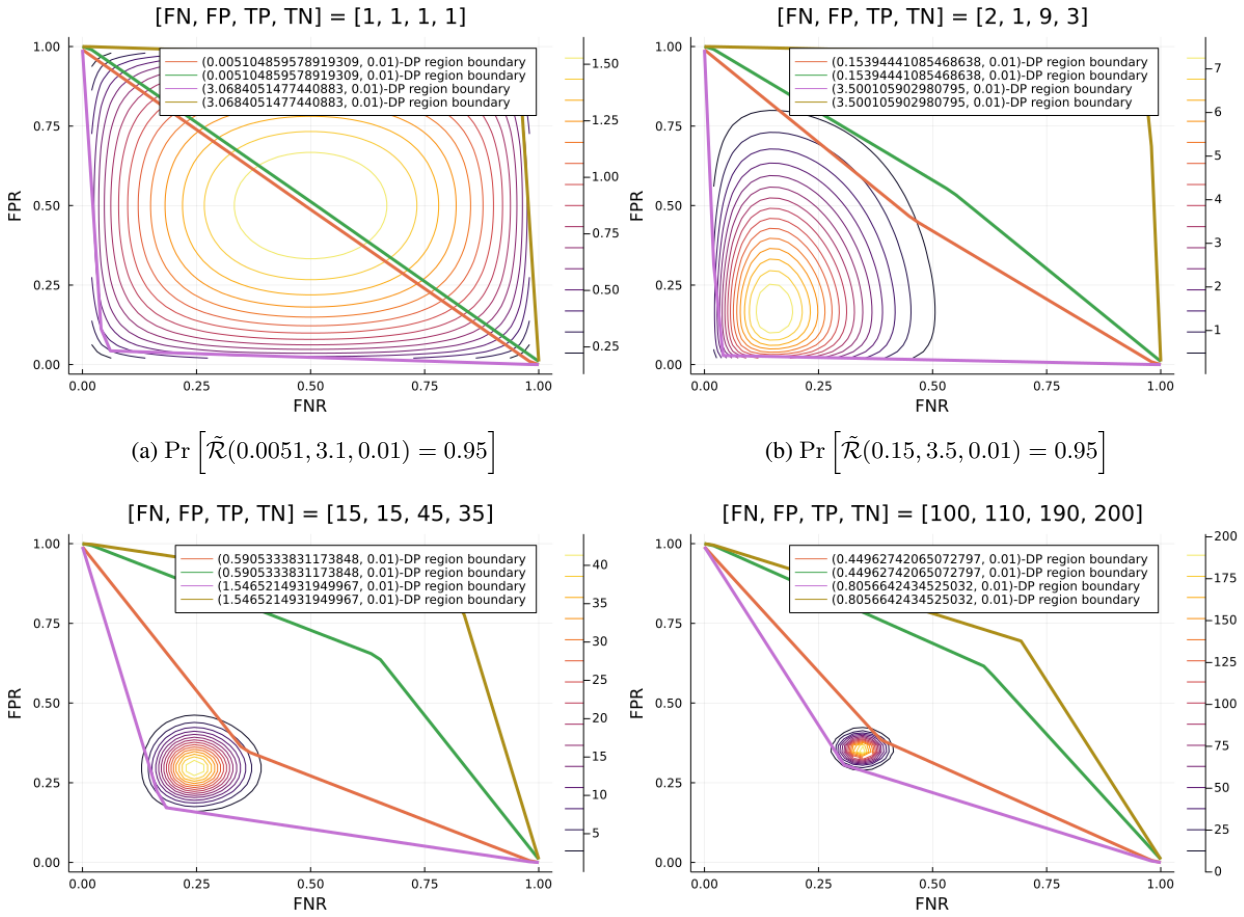
where $\delta$ is the Dirac $\delta$ distribution.

### A.3. Illustration of the Joint Posterior as $N$ grows

For this illustration of the joint posterior as the number of samples $N$ grows. For this illustration we find an interval of possible values of $\varepsilon$ in which the true $\varepsilon$ lies with a given probability. For convenience, we define the two-sided privacy region $\tilde{\mathcal{R}}$ as follows

$$\tilde{\mathcal{R}}(\varepsilon_-, \varepsilon_+, \delta) := \mathcal{R}(\varepsilon_+, \delta) \setminus \mathcal{R}(\varepsilon_-, \delta) \ . \tag{14}$$

The results are illustrated in Figure 7. Initially, we look at privacy regions after only 4 trials. As expected, the two-sided privacy region is fairly large and covers almost the entire unit square. As we see more and more samples our confidence increases and the two-sided privacy region shrinks.




(a) $\Pr\left[\tilde{\mathcal{R}}(0.0051, 3.1, 0.01) = 0.95\right]$

(b) $\Pr\left[\tilde{\mathcal{R}}(0.15, 3.5, 0.01) = 0.95\right]$

(c) $\Pr\left[\tilde{\mathcal{R}}(0.59, 1.5, 0.01) = 0.95\right]$

(d) $\Pr\left[\tilde{\mathcal{R}}(0.45, 0.80, 0.01) = 0.95\right]$

*Figure 7.* Convergence of the joint posterior $f_{(\mathrm{FNR},\mathrm{FPR})}$ as the number of samples grows.