

Privacy Auditing with One (1) Training Run

Thomas Steinke* Milad Nasr* Matthew Jagielski*

Abstract

We propose a scheme for auditing differentially private machine learning systems with a single training run. This exploits the parallelism of being able to add or remove multiple training examples independently. We analyze this using the connection between differential privacy and statistical generalization, which avoids the cost of group privacy. Our auditing scheme requires minimal assumptions about the algorithm and can be applied in the black-box or white-box setting.

1 Introduction

Differential privacy (DP) [DMNS06] provides a quantifiable privacy guarantee by ensuring that no person’s data significantly affects the probability of any outcome. Formally, a randomized algorithm M satisfies (ϵ, δ) -DP if, for any pair of inputs x, x' differing only by the addition or removal of one person’s data and any measurable S , we have

$$\mathbb{P}[M(x) \in S] \leq e^\epsilon \cdot \mathbb{P}[M(x') \in S] + \delta. \quad (1)$$

A DP algorithm is accompanied by a mathematical proof giving an *upper bound* on the privacy parameters ϵ and δ . In contrast, a *privacy audit* provides an empirical *lower bound* on the privacy parameters. Privacy audits allow us to assess the tightness of the mathematical analysis [JUO20; NHSBTJCT23] or, if the lower and upper bounds are contradictory, to detect errors in the analysis or in the algorithm’s implementation [TTSSJC22].

Typically, privacy audits obtain a lower bound on the privacy parameters directly from the DP definition (1). That is, we construct a pair of inputs x, x' and a set of outcomes S and we estimate the probabilities $\mathbb{P}[M(x) \in S]$ and $\mathbb{P}[M(x') \in S]$. However, estimating these probabilities requires running the algorithm M hundreds of times. This approach to privacy auditing is computationally expensive, which raises the question

Can we perform privacy auditing using a single run of the algorithm M ?

This is the question we address in our work.

*Google. {steinke,srxzr,jagielski}@google.com. Reverse alphabetical author order.

1.1 Our Contributions

Our approach (§2): The DP definition (1) considers adding or removing a single person’s data to or from the dataset. We consider multiple people’s data and the dataset independently includes or excludes each person’s data point. Our analysis exploits the parallelism of multiple independent data points in a single run of the algorithm in lieu of multiple independent runs.

Our auditing procedure operates as follows. We identify m data points (i.e., training examples or “canaries”) to either include or exclude and we flip m independent unbiased coins to decide which of them to include or exclude. We then run the algorithm on the randomly selected dataset. Based on the output of the algorithm, the auditor “guesses” whether or not each data point was included or excluded (or it can abstain from guessing for some data points). We obtain a lower bound on the privacy parameters from the fraction of guesses that were correct.

Intuitively, if the algorithm is $(\epsilon, 0)$ -DP, then the auditor can correctly guess each inclusion/exclusion coin flip with probability at most $\frac{e^\epsilon}{e^\epsilon+1}$. Thus DP implies a high-probability upper bound on the fraction of correct guesses and, conversely, a large fraction of correct guesses implies a high-probability lower bound on the privacy parameters.

Our analysis (§5): Naïvely, analyzing the addition or removal of multiple data elements would rely on group privacy; but this does not exploit the fact that the data items were included or excluded independently. Instead, we leverage the connection between DP and generalization [DFHPRR15b; DFHPRR15a; BNSSSU16; RRST16; JLNRSMS19; SZ20]. Our main theoretical contribution is an improved analysis of this connection that is tailored to yield nearly tight bounds in our setting.

Informally, if we run a DP algorithm on i.i.d. samples from some distribution, then, conditioned on the output of the algorithm, the samples are still “close” to being i.i.d. samples from that distribution. There is some technicality in making this precise, but, roughly speaking, we show that including or excluding m data points independently for one run is essentially as good as having m independent runs (as long as δ is small).

Our results (§6): We implement our new auditing framework to audit DP-SGD training on a WideResNet model, trained on the CIFAR10 dataset across multiple configurations. Our approach successfully achieves an empirical lower bound of $\epsilon \geq 1.8$, compared to a theoretical upper bound of $\epsilon \leq 4$ in the white-box setting. The m examples we insert for auditing (known in the literature as “canaries”) do not significantly impact the accuracy of the final model (less than a 5% decrease in accuracy) and our procedure only requires a single end-to-end training run. Such results were previously unattainable in the setting where only one model could be trained.

Algorithm 1 Auditor with One Training Run

- 1: **Data:** $x \in \mathcal{X}^n$ consisting of m auditing examples (a.k.a. canaries) and $n-m$ non-auditing examples.
 - 2: **Parameters:** Algorithm to audit \mathcal{A} , number of examples to randomize m , number of positive k_+ and negative k_- guesses.
 - 3: For $i \in [m]$ sample $S_i \in \{-1, +1\}$ independently with $\mathbb{E}[S_i] = 0$. Set $S_i = 1$ for all $i \in [n] \setminus [m]$.
 - 4: Partition x into $x_{\text{IN}} \in \mathcal{X}^{n_{\text{IN}}}$ and $x_{\text{OUT}} \in \mathcal{X}^{n_{\text{OUT}}}$ according to S , where $n_{\text{IN}} + n_{\text{OUT}} = n$. Namely, if $S_i = 1$, then x_i is in x_{IN} ; and, if $S_i = -1$, then x_i is in x_{OUT} .
 - 5: Run \mathcal{A} on input x_{IN} with appropriate parameters, outputting w .
 - 6: Compute the vector of scores $Y = (\text{SCORE}(x_i, w) : i \in [m]) \in \mathbb{R}^m$.
 - 7: Sort the scores Y . Let $T \in \{-1, 0, +1\}^m$ be $+1$ for the largest k_+ scores and -1 for the smallest k_- scores.
 - 8: (I.e., $T \in \{-1, 0, +1\}^m$ maximizes $\sum_i^m T_i \cdot Y_i$ subject to $\sum_i^m |T_i| = k_+ + k_-$ and $\sum_i^m T_i = k_+ - k_-$.)
 - 9: **Return:** The vector $S \in \{-1, +1\}^m$ indicating the true selection and the guesses $T \in \{-1, 0, +1\}^m$.
-

2 Our Auditing Procedure

We now present our auditing procedure in Algorithm 1. We independently include each of the first m examples with 50% probability and exclude it otherwise.¹ Our approach is applicable to both white-box auditing in the sense that the adversary has access to all intermediate values of the model weights and black-box auditing in the sense that the adversary only sees the final model weights (or can only query the final model). In both cases we compute a “score” for each example and “guess” whether the example is included or excluded based on these scores. Specifically, we guess that the examples with the k_+ highest scores are included and the examples with the k_- lowest scores are excluded, and we abstain from guessing for the remaining $m - k_+ - k_-$ auditing examples; the setting of these parameters will depend on the application.

Note that we only randomize the first m examples x_1, \dots, x_m (which we refer to as “auditing examples” or “canaries”); the last $n-m$ examples x_{m+1}, \dots, x_n are always included and, thus, we do not make any guesses about them. To get the strongest auditing results we would set $m = n$, but we usually want to set $m < n$. For example, computing the score of all n examples may be computationally prohibitive, so we only compute the scores of m examples. Also we may wish to artificially construct m examples to be easy to identify (i.e., canaries), but still include $n - m$ “real” examples to ensure that \mathcal{A} still produces a useful model. (I.e., having more training examples improves the performance of the model.)

¹Alternatively, we could also consider a different probability of inclusion; our theoretical results can handle this (see Proposition 5.7). However, this seems unlikely to be useful, as it intuitively lowers the signal-to-noise ratio. Another alternative is to non-independently choose which points to include to ensure x_{IN} has a fixed size; see Appendix A.

Intuitively, the vector of scores Y should be correlated with the true selection S , but too strong a correlation would violate DP. This is the basis of our audit. Specifically, the auditor computes T from Y which is a “guess” at S . By the postprocessing property of DP, the guesses T are a differentially private function of the true S , which means that they cannot be too accurate.

To obtain a lower bound on the DP parameters, in Section 5, we show that DP implies a high-probability upper bound on the number of correct guesses $W := \sum_i^m \max\{0, T_i \cdot S_i\}$. The observed value of W then yields a high-probability lower bound on the DP parameters. To be more precise, we have the following guarantee.

Theorem 2.1 (Informal version of Theorem 5.2). *Let $(S, T) \in \{-1, +1\}^m \times \{-1, 0, +1\}^m$ be the output of Algorithm 1. Assume the algorithm to audit \mathcal{A} satisfies (ε, δ) -DP. Let $r := k_+ + k_- = \|T\|_1$ be the number of guesses. Then, for all $v \in \mathbb{R}$,*

$$\mathbb{P}_{\substack{S \leftarrow \{-1, +1\}^m \\ T \leftarrow M(S)}} \left[\sum_i^m \max\{0, T_i \cdot S_i\} \geq v \right] \leq \mathbb{P}_{\check{W} \leftarrow \text{Binomial}(r, \frac{e^\varepsilon}{e^\varepsilon + 1})} [\check{W} \geq v] + O(\delta). \quad (2)$$

If we ignore δ for the moment, Theorem 2.1 says that the number of correct guesses is stochastically dominated by $\text{Binomial}(r, \frac{e^\varepsilon}{e^\varepsilon + 1})$, where $r = k_+ + k_-$ is the total number of guesses. This binomial distribution is precisely the distribution of correct guesses we would get if T was obtained by independently performing $(\varepsilon, 0)$ -DP randomized response on r bits of S . In other words, the theorem says that $(\varepsilon, 0)$ -DP randomized response is the worst-case algorithm in terms of the number of correct guesses. In particular, this means the theorem is tight (when $\delta = 0$)

The binomial distribution is well-concentrated. In particular, for all $\beta \in (0, 1)$, we have

$$\mathbb{P}_{\check{W} \leftarrow \text{Binomial}(r, \frac{e^\varepsilon}{e^\varepsilon + 1})} \left[\check{W} \geq \underbrace{\frac{r \cdot e^\varepsilon}{e^\varepsilon + 1} + \sqrt{\frac{1}{2} \cdot r \cdot \log(1/\beta)}}_{=v} \right] \leq \beta. \quad (3)$$

There is an additional $O(\delta)$ term in the guarantee (2). The exact expression for this term is somewhat complex. It is always $\leq 2m\delta$, but it is much smaller than this for reasonable parameter values. In particular, for v as in Equation 3 with $\beta \leq 1/r^4$, this term is $\leq O(\frac{m}{r}\delta)$.

Theorem 2.1 gives us a hypothesis test: If \mathcal{A} is (ε, δ) -DP, then the number of correct guesses W is $\leq \frac{r \cdot e^\varepsilon}{e^\varepsilon + 1} + O(\sqrt{r})$ with high probability. Thus, if the observed number of correct guesses v is larger than this, we can reject the hypothesis that \mathcal{A} satisfies (ε, δ) -DP. We can convert this hypothesis test into a confidence interval (i.e., a lower bound on ε) by finding the largest ε that we can reject at a desired level of confidence; see Section 4.3.

3 Related Work

The goal of privacy auditing is to empirically estimate the privacy provided by an algorithm, typically to accompany a formal privacy guarantee. Early work on auditing has often been

motivated by trying to identify bugs in the implementations of differentially private data analysis algorithms [DWWZK18; BGDCTV18].

Techniques for auditing differentially private machine learning typically rely on conducting some form of membership inference attack [SSSS17];² these attacks are designed to detect the presence or absence of an individual example in the training set. Essentially, a membership inference attack which achieves some true positive rate (TPR) and false positive rate (FPR) gives a lower bound on the privacy parameter $\epsilon \geq \log_e(\text{TPR}/\text{FPR})$ (after ensuring statistical validity of the TPR and FPR estimates).

Jayaraman and Evans [JE19] use standard membership inference attacks to evaluate different privacy analysis algorithms. Jagielski, Ullman, and Oprea [JUO20] consider inferring membership of worst-case “poisoning” examples to conduct stronger membership inference attacks and understand the tightness of privacy analysis. Nasr, Song, Thakurta, Papernot, and Carlini [NSTPC21] measure the tightness of privacy analysis under a variety of threat models, including showing that the DP-SGD analysis is tight in the threat model assumed by the standard DP-SGD analysis.

Improvements to auditing have been made in a variety of directions. For example, Nasr, Hayes, Steinke, Balle, Tramèr, Jagielski, Carlini, and Terzis [NHSBTJCT23] and Maddock, Sablayrolles, and Stock [MSS22] take advantage of the iterative nature of DP-SGD, auditing individual steps to understand privacy of the end-to-end algorithm. Improvements have also been made to the basic statistical techniques for estimating the ϵ parameter, for example by using Log-Katz confidence intervals [LMFLZWRFT22], Bayesian techniques [ZBWT-SRPNK22], or auditing algorithms in different privacy definitions [NHSBTJCT23]. Andrew, Kairouz, Oh, Oprea, McMahan, and Suriyakumar [AKOOMS23] build on the observation that, when performing membership inference, analyzing the case where the data is not included does not require re-running the algorithm; instead we can re-sample the excluded data point; if the data points are i.i.d. from a nice distribution, this permits closed-form analysis of the excluded case.

A recent heuristic proposed to improve the efficiency of auditing is performing membership inference on multiple examples simultaneously. This heuristic was proposed by Malek Esmaili, Mironov, Prasad, Shilov, and Tramer [MEMPST21], and evaluated more rigorously by Zanella-Béguelin, Wutschitz, Tople, Salem, Rühle, Paverd, Naseri, and Köpf [ZBWTSRPNK22]. However, this heuristic is not theoretically justified, as the TPR and FPR estimates are not based on independent samples. In our work, we provide a proof of the validity of this heuristic. In fact, with this proof, we show for the first time that standard membership inference attacks, which attack multiple examples per training run, can be used for auditing analysis; prior work using these attacks must make an independence assumption. As a result, auditing can take advantage of progress in the membership inference field [CCNSTT22; WBKBGGG22].

²Shokri, Stronati, Song, and Shmatikov [SSSS17] coined the term “membership inference attack” and were the first to apply such attacks to machine learning systems. However, similar attacks were developed for applications to genetic data [HSRDTMPSNC08; SOJH09; DSSUV15] and in cryptography [BS98; Tar08].

4 Background

We briefly review some standard background material. Readers may wish to skip to the next section and revisit this only if necessary.

4.1 Differential Privacy

They We recite the definitions of differential privacy and some relevant relaxations. For detailed background, see the tutorial by Vadhan [Vad17] or the textbook by Dwork and Roth [DR14].

Definition 4.1 (Differential Privacy [DMNS06; DKMMN06]). *Let $M : \mathcal{X}^* \rightarrow \mathcal{Y}$ be a randomized algorithm, where $\mathcal{X}^* = \bigcup_{n \geq 0} \mathcal{X}^n$. We say M is (ε, δ) -differentially private $((\varepsilon, \delta)$ -DP) if, for all $x, x' \in \mathcal{X}^*$ differing only by the addition or removal of one element, we have*

$$\forall S \subset \mathcal{Y} \quad \mathbb{P}[M(x) \in S] \leq e^\varepsilon \cdot \mathbb{P}[M(x') \in S] + \delta.$$

Definition 4.2 (Rényi Differential Privacy [Mir17]). *We say $M : \mathcal{X}^* \rightarrow \mathcal{Y}$ is $(\alpha, \tilde{\varepsilon})$ -Rényi differentially private $((\alpha, \tilde{\varepsilon})$ -RDP) if, for all $x, x' \in \mathcal{X}^*$ differing only by the addition or removal of one element, we have*

$$D_\alpha(M(x) \| M(x')) \leq \tilde{\varepsilon},$$

where $D_\alpha(P \| Q) := \frac{1}{\alpha-1} \log \mathbb{E}_{Y \leftarrow P} \left[\left(\frac{P(Y)}{Q(Y)} \right)^{\alpha-1} \right]$ denotes the Rényi divergence of order α .

Definition 4.3 (Concentrated Differential Privacy [DR16; BS16]). *We say $M : \mathcal{X}^* \rightarrow \mathcal{Y}$ is ρ -zero concentrated differentially private (ρ -zCDP) if, for all $x, x' \in \mathcal{X}^*$ differing only by the addition or removal of one element, we have*

$$\forall \alpha > 1 \quad D_\alpha(M(x) \| M(x')) \leq \alpha \cdot \rho.$$

Remark 4.4. *In this paper, we focus on to the addition or removal notion of DP, rather than replacement. (In Appendix A, we consider replacement.) Note that, in our theoretical analysis, we consider DP algorithms of the form $M : \{0, 1\}^m \rightarrow \mathcal{Y}$. In this case, DP is with respect to flipping one of the input bits, as each bit indicates whether some example is included or excluded.*

The main property of DP that we use is invariance under *postprocessing*. That is, if $M : \mathcal{X}^* \rightarrow \mathcal{Y}$ satisfies DP and $F : \mathcal{Y} \rightarrow \mathcal{Z}$ is an arbitrary function, then $F \circ M : \mathcal{X}^* \rightarrow \mathcal{Z}$ also satisfies DP with the same parameters.

Gaussian Mechanism A common method for achieving DP is Gaussian noise addition. The following gives the optimal DP guarantee for the Gaussian mechanism.

Lemma 4.5 ([BW18, Theorem 8]). *Let $q : \mathcal{X}^* \rightarrow \mathbb{R}$ be a function with sensitivity $\Delta := \sup_{x,x'} |q(x) - q(x')|$. (In the supremum $x, x' \in \mathcal{X}^*$ are restricted to differ only by the addition or removal of one element.) Fix $\sigma^2 > 0$ and let $\rho := \Delta^2/2\sigma^2$. Define $M : \mathcal{X}^* \rightarrow \mathbb{R}$ by $M(x) = \mathcal{N}(q(x), \sigma^2)$. Then, for any $\varepsilon \geq 0$, the algorithm M satisfies (ε, δ) -DP with*

$$\delta = \bar{\Phi}\left(\frac{\varepsilon - \rho}{\sqrt{2\rho}}\right) - e^\varepsilon \cdot \bar{\Phi}\left(\frac{\varepsilon + \rho}{\sqrt{2\rho}}\right),$$

where $\bar{\Phi}(z) := \mathbb{P}_{Z \leftarrow \mathcal{N}(0,1)}[Z > z] = \frac{1}{\sqrt{2\pi}} \int_z^\infty \exp(-x^2/2) dx$. Furthermore, M satisfies ρ -zCDP.

4.2 DP-SGD – Differentially Private Stochastic Gradient Descent

The algorithm whose privacy we are most interested in auditing is Differentially Private Stochastic Gradient Descent (DP-SGD, Algorithm 2). This is the workhorse of private machine learning both in theory [BST14] and in practice [ACGMMTZ16].

Algorithm 2 DP-SGD – Differentially Private Stochastic Gradient Descent

- 1: **Input:** $x \in \mathcal{X}^n$
 - 2: **Model:** Loss function $f : \mathbb{R}^d \times \mathcal{X} \rightarrow \mathbb{R}$.
 - 3: **Parameters:** Number of iterations $\ell \geq 1$, clipping threshold $c > 0$, noise multiplier $\sigma > 0$, sampling probability $q \in (0, 1]$, learning rate $\eta > 0$.
 - 4: Initialize $w_0 \in \mathbb{R}^d$.
 - 5: **for** $t = 1, \dots, \ell$ **do**
 - 6: Sample $S^t \subseteq [n]$ where each $i \in [n]$ is included independently with probability q .
 - 7: Compute $g_i^t = \nabla_{w^{t-1}} f(w^{t-1}, x_i) \in \mathbb{R}^d$ for all $i \in S^t$.
 - 8: Clip $\hat{g}_i^t = \min\left\{1, \frac{c}{\|g_i^t\|_2}\right\} \cdot g_i^t \in \mathbb{R}^d$ for all $i \in S^t$.
 - 9: Sample $\xi^t \in \mathbb{R}^d$ from $\mathcal{N}(0, \sigma^2 c^2 I)$.
 - 10: Sum $\tilde{g}^t = \xi^t + \sum_{i \in S^t} \hat{g}_i^t \in \mathbb{R}^d$.
 - 11: Update $w^t = w^{t-1} - \eta \cdot \tilde{g}^t \in \mathbb{R}^d$.
 - 12: **end for**
 - 13: **Output:** w^0, w^1, \dots, w^ℓ .
-

DP-SGD satisfies differential privacy. Much ink has been spilled precisely quantifying its privacy properties [MTZ19; WBK19; KJH20; GLW21; ZDW22, etc.]. A simple guarantee is the following.

Proposition 4.6 ([MTZ19; Ste22]). *DP-SGD (Algorithm 2) satisfies $(2, \varepsilon)$ -RDP for*

$$\varepsilon = \ell \cdot \log\left(1 + q^2 \cdot (\exp(1/\sigma^2) - 1)\right) \approx \ell \cdot q^2 \cdot \frac{1}{\sigma^2}.$$

If $\tilde{\varepsilon} \leq 1$, then DP-SGD should provide meaningful privacy protection. In particular, $(2, \tilde{\varepsilon})$ -RDP implies that membership inference has a maximum accuracy (in the balanced case) of

$$\frac{1}{2} + \frac{1}{2} \sqrt{\frac{e^{\tilde{\varepsilon}} - 1}{e^{\tilde{\varepsilon}} + 3}} \approx \frac{1}{2} + \frac{1}{4} \sqrt{\tilde{\varepsilon}}. \quad (4)$$

Our goal is to audit this guarantee.

4.3 Hypothesis Testing & Statistical Estimation

Our goal is to estimate the privacy parameters of the algorithm that we are auditing. As prior work has noted [DWWZK18; JUO20], this task can be framed as statistical estimation, with a goal of outputting a statistical lower bound on the privacy parameters. These lower bounds will have a corresponding confidence level, roughly representing the probability that the lower bound could have been produced even when analyzing an algorithm with perfect privacy. As empirical methods, it is impossible to have 100% confidence in our methods, so we will generally use 95% confidence in our experiments, comparable to the use of $p < 0.05$ in science literature.

To be precise, our auditor runs the algorithm M and outputs $\varepsilon_{\text{LB}} \geq 0$ with the following guarantee. If M satisfies $(\varepsilon_{\text{true}}, \delta)$ -DP, then, with probability at least $1 - \beta$, we have $\varepsilon_{\text{LB}} \leq \varepsilon_{\text{true}}$. Here $1 - \beta$ is the confidence level and $\delta \geq 0$ is fixed. Note that this is a frequentist guarantee, rather than a Bayesian guarantee. That is, the probability is with respect to our auditing procedure, rather than a statement about our beliefs about M .

We can also view this in terms of hypothesis testing. Here we start with a “null hypothesis” that M satisfies $(\varepsilon_{\text{null}}, \delta)$ -DP and the auditor’s goal is to test this hypothesis by running M . If the auditor rejects this null hypothesis, then this gives us a lower bound $\varepsilon_{\text{LB}} = \varepsilon_{\text{null}}$.

The difference between hypothesis testing and statistical estimation is that a hypothesis test starts with a given $\varepsilon_{\text{null}}$ and outputs a binary decision to reject or not, while an estimator outputs a number ε_{LB} . However, we can convert between these:

Lemma 4.7. *For each M , let $A_M \in \Omega$ be a random variable and let $P_M \in \mathbb{R}$ be a fixed number. For each $\varepsilon, \beta > 0$, let $T_{\varepsilon, \beta} \subset \Omega$ satisfy*

$$\forall M \quad (P_M = \varepsilon \implies \mathbb{P}[A_M \in T_{\varepsilon, \beta}] \leq \beta). \quad (5)$$

Further suppose that, if $\varepsilon_1 \leq \varepsilon_2$, then $T_{\varepsilon_1, \beta} \supset T_{\varepsilon_2, \beta}$. Then, for all M and all $\beta > 0$,

$$\mathbb{P}[P_M \geq \sup \{\varepsilon > 0 : A_M \in T_{\varepsilon, \beta}\}] \geq 1 - \beta. \quad (6)$$

Proof. Fix a realization of A_M and suppose $P_M < \sup \{\varepsilon > 0 : A_M \in T_{\varepsilon, \beta}\}$. Then there exists some $\varepsilon \geq P_M$ with $A_M \in T_{\varepsilon, \beta}$ and, hence,

$$A_M \in \bigcup_{\varepsilon \geq P_M} T_{\varepsilon, \beta} = T_{P_M, \beta}.$$

The equality above follows from our monotonicity assumption on T . Thus

$$\mathbb{P}[P_M < \sup\{\varepsilon > 0 : A_M \in T_{\varepsilon, \beta}\}] \leq \mathbb{P}[A_M \in T_{P_M, \beta}] \leq \beta,$$

as required. \square

To interpret Lemma 4.7, M is an algorithm and P_M is the “true” privacy parameter ε that it satisfies. (We’re considering δ to be fixed.) The random variable A_M is the output of our auditing procedure applied to M . (This is our test statistic in the language of hypothesis testing.) The hypothesis test’s rejection set is $T_{\varepsilon, \beta}$ and Equation 5 guarantees that, if M is indeed (ε, δ) -DP (i.e., the null hypothesis is true), then the probability that we reject the null hypothesis is at most β . Equation 6 then shows how to estimate the true privacy parameter P_M from A_M ; we simply take the largest ε for which we can reject the corresponding null hypothesis.

Note that Lemma 4.7 needs to make a technical monotonicity assumption. In our setting this simply means that, if a given realization of the test statistic A_M allows us to reject the null hypothesis that M is (ε_2, δ) -DP and $\varepsilon_1 \leq \varepsilon_2$, then we can also reject the null hypothesis that M is (ε_1, δ) -DP.

4.4 Stochastic Dominance

In our theoretical analysis we use the concept of stochastic dominance. Specifically, we use this to formalize the “worst-case” DP algorithm for auditing.

Definition 4.8 (Stochastic Dominance). *Let $X, Y \in \mathbb{R}$ be random variables. We say X is stochastically dominated by Y (or Y stochastically dominates X) if $\mathbb{P}[X > t] \leq \mathbb{P}[Y > t]$ for all $t \in \mathbb{R}$. Equivalently, X is stochastically dominated by Y if there exists a coupling (i.e., a joint distribution that matches the marginal distributions of X and Y) such that $\mathbb{P}[X \leq Y] = 1$.*

Stochastic dominance is preserved under sums/convolutions:

Lemma 4.9. *Suppose X_1 is stochastically dominated by Y_1 . Suppose that, for all $x \in \mathbb{R}$, the conditional distribution $X_2|X_1 = x$ is stochastically dominated by Y_2 . Assume that Y_1 and Y_2 are independent. Then $X_1 + X_2$ is stochastically dominated by $Y_1 + Y_2$.*

Proof. For all $t \in \mathbb{R}$, we have

$$\begin{aligned} \mathbb{P}[X_1 + X_2 > t] &= \mathbb{E}_{X_1} \left[\mathbb{P}_{X_2} [X_2 > t - X_1 | X_1] \right] \\ &\leq \mathbb{E}_{X_1} \left[\mathbb{P}_{Y_2} [Y_2 > t - X_1] \right] && (Y_2 \text{ dominates } X_2 | X_1) \\ &= \mathbb{E}_{Y_2} \left[\mathbb{P}_{X_1} [X_1 > t - Y_2] \right] \\ &\leq \mathbb{E}_{Y_2} \left[\mathbb{P}_{Y_1} [Y_1 > t - Y_2] \right] && (Y_1 \text{ dominates } X_1 \text{ \& independence}) \\ &= \mathbb{P}[Y_1 + Y_2 > t]. \end{aligned}$$

□

5 Theoretical Analysis

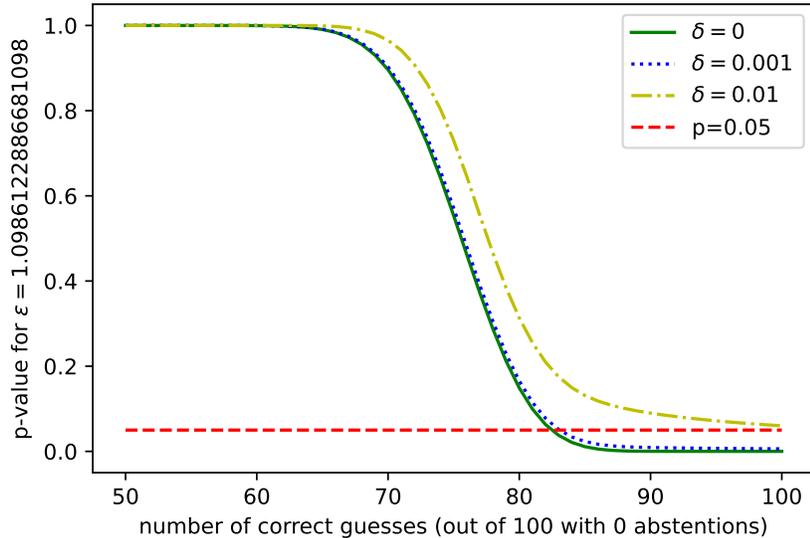


Figure 1: Theorem 5.2’s p-value as the number of correct guesses changes for fixed $\epsilon = \log 3$ (i.e., ideally 75% of guesses correct). The total number of examples and guesses is 100.

To analyze the results of our audit, we leverage the connection between DP and generalization [DFHPRR15b; DFHPRR15a; BNSSSU16; RRS16; JLNRSMS19; SZ20]. Unfortunately, directly applying the existing results from the literature is unlikely to yield meaningful results, as the constants are not optimal. Thus we provide an analysis of DP’s generalization guarantees that is suitable for our application and which has sharp constants.

We consider the following formalism. The algorithm $M : \{-1, +1\}^m \rightarrow \mathbb{R}^m$ takes in a vector of bits and outputs a vector of “guesses”. Each input bit indicates whether or not a particular example is included in or excluded from the dataset. In particular, the DP guarantee ensures that the outputs are indistinguishable if we flip one bit, which corresponds to adding or removing the corresponding data point. Each coordinate of the output is a guess for the corresponding input bit; the sign of the score should match the corresponding input bit, while the magnitude is a reflection of the confidence.

The algorithm M represents both the “real” algorithm (e.g., DP-SGD) and the auditor which postprocesses the output of the real algorithm into guesses. In this formalism, the examples themselves are considered fixed and not part of the input – i.e., the examples are “hardcoded” into M . The algorithm M is an abstraction for our analysis, rather than a realistic system.

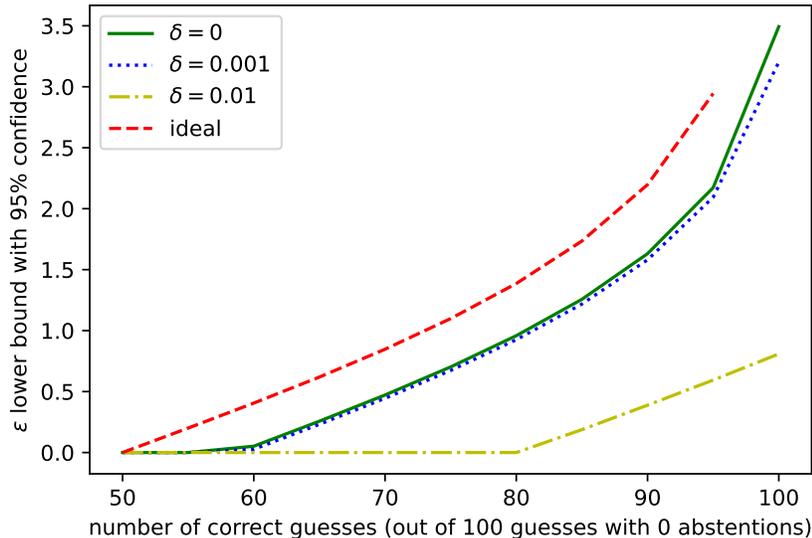


Figure 2: Lower bound on the privacy parameter ϵ given by Theorem 5.2 with 95% confidence as the number of correct guesses changes. The total number of examples and guesses is 100. For comparison, we plot the **ideal** ϵ that gives $100 \cdot \frac{e^\epsilon}{e^\epsilon + 1}$ correct guesses.

We evaluate the quantity

$$W := \sum_i^m \max\{0, T_i \cdot S_i\},$$

where S is uniform on $\{-1, +1\}^m$ and $T = M(S)$. If T_i and S_i disagree in sign (i.e., the guess is wrong), then $\max\{0, T_i \cdot S_i\} = 0$; if they agree (i.e., the guess is right), then $\max\{0, T_i \cdot S_i\} = |T_i|$. That is, W increases when we guess correctly and the increase is proportional to how much “weight” we placed on that guess. The auditor seeks to maximize W and then we compare it to a baseline that is consistent with DP. (The analysis in this section focuses on computing this baseline.) Incorrect guesses do not increase W , but they do increase the baseline. Note that we can guess $T_i = 0$, which amounts to abstaining from making a guess; this doesn’t increase W , but also doesn’t increase the baseline.

Our formalism is inspired by that of Steinke and Zakyntinou [SZ20], who also restrict to binary inputs. In contrast, most of the work connecting DP and generalization does not do this. The benefit of restricting to binary inputs which represent inclusion or exclusion of a data point is that it simplifies our analysis.

5.1 Pure DP Analysis

We first consider the pure DP ($\delta = 0$) case, as it is considerably simpler than the general case. We follow the analysis of Jung, Ligett, Neel, Roth, Sharifi-Malvajerdi, and Shenfeld

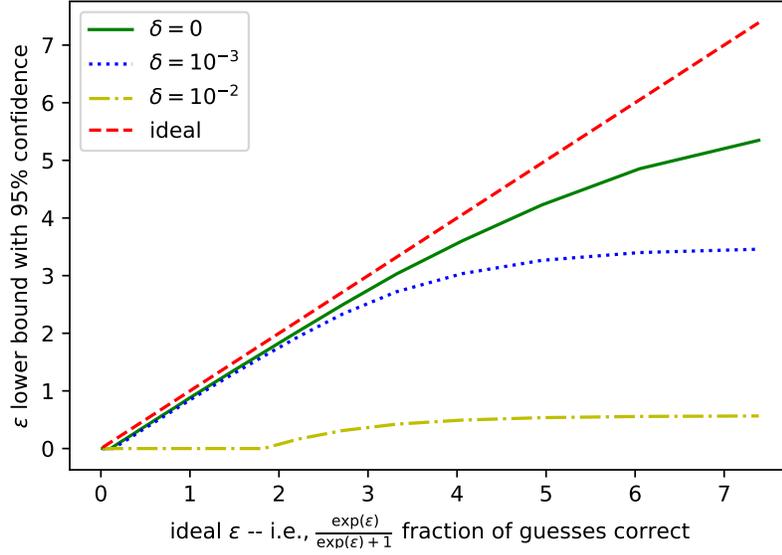


Figure 3: Lower bound on the privacy parameter ε given by Theorem 5.2 with 95% confidence as the number of correct guesses changes. The total number of examples and guesses is 1000 (with no abstentions). Here we plot the ideal ε on the horizontal axis, so that the number of correct guesses is $1000 \cdot \frac{e^\varepsilon}{e^\varepsilon+1}$.

[JLNRSMS19] with some refinement. Specifically, rather than relying on a Hoeffding bound, we show that it is stochastically dominated by a Binomial distribution. This result is tight – i.e., if M independently performs a randomized response for each input bit, then the inequality becomes an equality.

Proposition 5.1 (Pure DP Version of Main Result). *Let $M : \{-1, +1\}^m \rightarrow \mathbb{R}^m$ satisfy $(\varepsilon, 0)$ -DP. Let $S \in \{-1, +1\}^m$ be uniformly random. Let $T = M(S) \in \mathbb{R}^m$. Then, for all $v \in \mathbb{R}$ and all $t \in \mathbb{R}^m$ in the support of T ,*³

$$\mathbb{P}_{\substack{S \leftarrow \{-1, 1\}^m \\ T \leftarrow M(S)}} \left[\sum_i^m \max\{0, T_i \cdot S_i\} \geq v \mid T = t \right] \leq \mathbb{P}_{\check{S} \leftarrow \text{Bernoulli}\left(\frac{e^\varepsilon}{e^\varepsilon+1}\right)^m} \left[\sum_i^m \check{S}_i \cdot |t_i| \geq v \right] =: \beta(m, \varepsilon, v, t).$$

Proposition 5.1 is Bayesian: We condition on the output and then consider the probability that each guess was right. The vector \check{S} should be seen as indicating whether each guess was right. The proposition says that, in the worst case, each guess is correct independently with probability $\frac{e^\varepsilon}{e^\varepsilon+1}$.

³To be precise, this holds with probability 1, but may fail for t in a set of measure zero under T . Note that $S \leftarrow \{-1, 1\}^m$ denotes that S is uniform on the set $\{-1, 1\}^m$ and $\check{S} \leftarrow \text{Bernoulli}\left(\frac{e^\varepsilon}{e^\varepsilon+1}\right)$ denotes that $\check{S} \in \{0, 1\}^m$ has a product distribution with each coordinate having expectation $\frac{e^\varepsilon}{e^\varepsilon+1}$.

How do we use this result? Suppose we have conducted an audit and observed s and t as the output of Algorithm 3. Let $v = \sum_i^m \max\{0, s_i \cdot t_i\}$. Following Lemma 4.7, we choose a desired confidence $1 - \beta < 1$ (e.g., $\beta = 0.05$) and then we choose $\varepsilon \geq 0$ so that $\beta(m, \varepsilon, v, t) = \beta$. Then this value of ε is our lower bound.

In the language of hypothesis testing, $W = \sum_i^m \max\{0, T_i \cdot S_i\}$ is the test statistic and our null hypothesis is that M is ε -DP. Under the null hypothesis we have $\mathbb{P}[W \geq v] \leq \beta(m, \varepsilon, v, t)$. Thus, if v is the observed value of the test statistic, then $\beta(m, \varepsilon, v, t)$ is our p-value. And we can reject the null hypothesis if, say, $\beta(m, \varepsilon, v) \leq 0.05$.

Proof. Fix some $t \in \mathbb{R}^m$. We now analyze the distribution of S conditioned on $M(S) = t$. Note that the unconditional distribution of S is uniform on $\{-1, +1\}^m$ and M is $(\varepsilon, 0)$ -DP. We perform the analysis one bit at a time. Fix some $i \in [m]$ and $s_{<i} \in \{-1, +1\}^{i-1}$. By Bayes' law and $(\varepsilon, 0)$ -DP,

$$\begin{aligned}
& \mathbb{P}[S_i = 1 | M(S) = t, S_{<i} = s_{<i}] \\
&= \frac{\mathbb{P}[M(S) = t | S_i = 1, S_{<i} = s_{<i}] \cdot \mathbb{P}[S_i = 1 | S_{<i} = s_{<i}]}{\mathbb{P}[M(S) = t | S_{<i} = s_{<i}]} \\
&= \frac{\mathbb{P}[M(S) = t | S_i = 1, S_{<i} = s_{<i}] \cdot \mathbb{P}[S_i = 1]}{\mathbb{P}[M(S) = t | S_i = 1, S_{<i} = s_{<i}] \cdot \mathbb{P}[S_i = 1] + \mathbb{P}[M(S) = t | S_i = -1, S_{<i} = s_{<i}] \cdot \mathbb{P}[S_i = -1]} \\
&= \frac{\mathbb{P}[M(S) = t | S_i = 1, S_{<i} = s_{<i}] \cdot \frac{1}{2}}{\mathbb{P}[M(S) = t | S_i = 1, S_{<i} = s_{<i}] \cdot \frac{1}{2} + \mathbb{P}[M(S) = t | S_i = -1, S_{<i} = s_{<i}] \cdot \frac{1}{2}} \\
&= \frac{1}{1 + \mathbb{P}[M(S) = t | S_i = -1, S_{<i} = s_{<i}] / \mathbb{P}[M(S) = t | S_i = 1, S_{<i} = s_{<i}]} \\
&\in \left[\frac{1}{1 + e^\varepsilon}, \frac{1}{1 + e^{-\varepsilon}} \right].
\end{aligned}$$

Thus $\mathbb{P}[S_i = \text{sign}(T_i) | T = t, S_{<i} = s_{<i}] \leq \frac{1}{1 + e^{-\varepsilon}} = \frac{e^\varepsilon}{e^\varepsilon + 1}$.

With this in hand, we can prove the result by induction. We assume inductively that $W_{m-1} := \sum_i^{m-1} \max\{0, T_i \cdot S_i\}$ is stochastically dominated by $\check{W}_{m-1} := \sum_i^{m-1} \check{S}_i \cdot |t_i|$ where $\check{S} \leftarrow \text{Bernoulli}\left(\frac{e^\varepsilon}{e^\varepsilon + 1}\right)^{m-1}$. As above, conditioned on the value of W_{m-1} , the variable $\max\{0, T_m \cdot S_m\} = |T_m| \cdot \mathbb{I}[S_m = \text{sign}(T_m)]$ is stochastically dominated by $|T_m| \cdot \text{Bernoulli}\left(\frac{e^\varepsilon}{e^\varepsilon + 1}\right)$. By Lemma 4.9, $W_m = W_{m-1} + \max\{0, T_m \cdot S_m\}$ is stochastically dominated by $\check{W}_m := \sum_i^m \check{S}_i \cdot |t_i|$ where $\check{S} \leftarrow \text{Bernoulli}\left(\frac{e^\varepsilon}{e^\varepsilon + 1}\right)^m$. \square

5.2 Approximate DP Analysis

We extend the pure DP analysis (§5.1) to approximate DP ($\delta > 0$). This becomes quite messy. In the pure DP case, we can condition on an arbitrary output t . In the approximate DP case, some outputs are “bad” in the sense that the privacy loss is unbounded. To handle this we do two things: First, we require the guesses to be bounded (i.e., $T \in [-1, +1]^m$ instead of $T \in \mathbb{R}^m$), which ensures that bad outputs cannot skew things too much. Second, the guarantees we prove have an additional failure probability that depends on δ .

Our analysis most closely resembles that of Rogers, Roth, Smith, and Thakkar [RRST16]. Essentially, we repeat the analysis for the pure DP case, but add some failure events, and carefully account for how much they can distort the results.

Theorem 5.2 (Main Result). *Let $M : \{-1, +1\}^m \rightarrow [-1, +1]^m$ satisfy (ε, δ) -DP. Let $S \in \{-1, +1\}^m$ be uniformly random. Let $T = M(S) \in [-1, +1]^m$. Then, for all $v \in \mathbb{R}$,*

$$\mathbb{P}_{\substack{S \leftarrow \{-1, +1\}^m \\ T \leftarrow M(S)}} \left[\sum_i^m \max\{0, T_i \cdot S_i\} \geq v \right] \leq \beta + \alpha \cdot 2m \cdot \delta, \quad (7)$$

where

$$\beta = \mathbb{P}_{\check{W}^*} [\check{W}^* \geq v], \quad (8)$$

$$\alpha = \max \left\{ \frac{1}{i} \left(\mathbb{P}_{\check{W}^*} [\check{W}^* \geq v - i] - \beta \right) : i \in \{1, 2, \dots, m\} \right\}. \quad (9)$$

Here \check{W}^* is any distribution on \mathbb{R} that stochastically dominates $\check{W}(t) := \sum_i^m \check{S}_i |t_i|$ for $\check{S} \leftarrow \text{Bernoulli} \left(\frac{e^\varepsilon}{e^\varepsilon + 1} \right)^m$ for all t in the support of T .

To evaluate the bound of Theorem 5.2, we need to identify \check{W}^* and compute its distribution. We can set $\mathbb{P}_{\check{W}^*} [\check{W}^* \geq v] = \sup_{t \in \text{support}(T)} \mathbb{P}_{\check{W}} [\check{W}(t) \geq v]$. This can be difficult to compute, depending on what we know about the support of T . If the support of T is nice, we can compute this explicitly; e.g., see Corollary 5.4. There are other things we can do. For example, if we have bounds on $\sup_{t \in \text{support}(T)} \|t\|_2$ and $\sup_{t \in \text{support}(T)} \|t\|_1$, then we can use a concentration inequality to bound $\sup_{t \in \text{support}(T)} \mathbb{P}_{\check{W}} [\check{W}(t) \geq v]$ and then use this bound as the distribution of \check{W}^* . This yields the following corollary.

Corollary 5.3 (Analytic Version of Main Result). *Let $M : \{-1, +1\}^m \rightarrow [-1, +1]^m$ satisfy (ε, δ) -DP. Let $S \in \{-1, +1\}^m$ be uniformly random. Let $T = M(S) \in [-1, +1]^m$. Suppose $\mathbb{P} [\|T\|_2 \leq r_2] = 1$ and $\mathbb{P} [\|T\|_1 \leq r_1] = 1$. Then, for all $v \geq \frac{e^\varepsilon}{e^\varepsilon + 1} \cdot r_1 + 2$, we have*

$$\mathbb{P}_{\substack{S \leftarrow \{-1, +1\}^m \\ T \leftarrow M(S)}} \left[\sum_i^m \max\{0, T_i \cdot S_i\} \geq v \right] \leq f(v) + 2m \cdot \delta \cdot \max \left\{ \frac{f(v - i) - f(v)}{i} : i \in [m] \right\} \quad (10)$$

$$\leq f(v) + 2m\delta \cdot \max \left\{ \frac{2}{v - \frac{e^\varepsilon}{e^\varepsilon + 1} r_1}, f \left(\frac{1}{2} \left(v + \frac{e^\varepsilon}{e^\varepsilon + 1} r_1 \right) \right) \right\}, \quad (11)$$

where

$$f(v) := \begin{cases} \exp \left(\frac{-2}{r_2^2} \left(v - \frac{e^\varepsilon}{e^\varepsilon + 1} r_1 \right)^2 \right) & \text{if } v \geq \frac{e^\varepsilon}{e^\varepsilon + 1} r_1 \\ 1 & \text{if } v < \frac{e^\varepsilon}{e^\varepsilon + 1} r_1 \end{cases}.$$

In particular, if we substitute $v = \frac{e^\varepsilon}{e^\varepsilon+1}r_1 + r_2 \cdot \sqrt{\frac{1}{2} \log(1/\beta)}$ into Equation 11, we get

$$\mathbb{P}_{\substack{S \leftarrow \{-1,+1\}^m \\ T \leftarrow M(S)}} \left[\sum_i^m \max\{0, T_i \cdot S_i\} \geq v \right] \leq \beta + 2m \cdot \delta \cdot \max \left\{ \frac{1}{r_2 \sqrt{\frac{1}{2} \log(1/\beta)}}, \beta^{1/4} \right\}. \quad (12)$$

Proof. Fix an arbitrary t the support of T . Define $\check{W}(t) := \sum_i^m \check{S}_i |t_i|$ for $\check{S} \leftarrow \text{Bernoulli} \left(\frac{e^\varepsilon}{e^\varepsilon+1} \right)^m$. Then $\mathbb{E} [\check{W}(t)] = \frac{e^\varepsilon}{e^\varepsilon+1} \|t\|_1$. By Hoeffding's inequality, for all $\lambda \geq 0$,

$$\mathbb{P} \left[\check{W}(t) \geq \frac{e^\varepsilon}{e^\varepsilon+1} \|t\|_1 + \lambda \right] \leq \exp \left(\frac{-2\lambda^2}{\|t\|_2^2} \right).$$

Now define \check{W}^* by

$$\mathbb{P} [\check{W}^* \geq v] := f(v) := \begin{cases} \exp \left(\frac{-2}{r_2^2} \left(v - \frac{e^\varepsilon}{e^\varepsilon+1} r_1 \right)^2 \right) & \text{if } v \geq \frac{e^\varepsilon}{e^\varepsilon+1} r_1 \\ 1 & \text{if } v < \frac{e^\varepsilon}{e^\varepsilon+1} r_1 \end{cases}.$$

Since \check{W}^* stochastically dominates $\check{W}(t)$ for all t in the support of T , we can apply Theorem 5.2 to obtain the first part of the result (10).

Next, for any $c \geq 1$, we have

$$\begin{aligned} & \max \left\{ \frac{f(v-i) - f(v)}{i} : i \in [m] \right\} \\ & \leq \max \left\{ \frac{f(v-x)}{x} : x \in [1, \infty) \right\} \quad (f(v) \geq 0 \text{ and } [m] \subset [1, \infty)) \\ & = \max \left\{ \max \left\{ \frac{f(v-x)}{x} : x \in [1, c] \right\}, \max \left\{ \frac{f(v-x)}{x} : x \in [c, \infty) \right\} \right\} \\ & \leq \max \left\{ \max \left\{ \frac{f(v-x)}{1} : x \in [1, c] \right\}, \max \left\{ \frac{1}{x} : x \in [c, \infty) \right\} \right\} \\ & = \max \left\{ f(v-c), \frac{1}{c} \right\}. \end{aligned}$$

Setting $c = \frac{1}{2} \left(v - \frac{e^\varepsilon}{e^\varepsilon+1} r_1 \right)$ yields the second part of the result (11) \square

In the next corollary we restrict M to ternary outputs, so it must either guess ($T_i = \pm 1$) or abstain ($T_i = 0$). We bound the number of guesses by r . In this case the dominating distribution \check{W}^* is a binomial distribution, which is relatively easy to compute. This is the form of Theorem 5.2 that we use in all of our experimental results. We provide pseudocode in Appendix D.

Corollary 5.4 (Ternary Guesses). *Let $M : \{-1,+1\}^m \rightarrow \{-1,0,+1\}^m$ satisfy (ε, δ) -DP. Let $S \in \{-1,+1\}^m$ be uniformly random. Let $T = M(S) \in \{-1,0,+1\}^m$. Suppose $\mathbb{P} [\|T\|_1 \leq r] = 1$. Then, for all $v \in \mathbb{R}$,*

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

where

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

Now we delve into the proof of Theorem 5.2. We use a decomposition result of Kairouz, Oh, and Viswanath [KOV15] (see also [MV15] & [Ste22, Corollary 24]).

Lemma 5.5. *Let P and Q be probability distributions over \mathcal{Y} . Fix $\varepsilon, \delta \geq 0$. Suppose that, for all measurable $S \subset \mathcal{Y}$, we have $P(S) \leq e^\varepsilon \cdot Q(S) + \delta$ and $Q(S) \leq e^\varepsilon P(S) + \delta$.*

Then there exist $\delta' \in [0, \delta]$ and distributions $P', Q', P'',$ and Q'' over \mathcal{Y} such that the following three properties are all satisfied. First, we can express P and Q as convex combinations:

$$\begin{aligned} P &= (1 - \delta')P' + \delta'P'', \\ Q &= (1 - \delta')Q' + \delta'Q''. \end{aligned}$$

Second, for all measurable $S \subset \mathcal{Y}$, we have $e^{-\varepsilon}P'(S) \leq Q'(S) \leq e^\varepsilon P'(S)$. Third, there exist measurable $S, T \subset \mathcal{Y}$ such that $P''(S) = 1$, $Q''(T) = 1$, $\forall S' \subset S P(S') \geq Q(S')$, and $\forall T' \subset T Q(T') \geq P(T')$.

Proof. This proof follows that of Steinke [Ste22]. We begin with some formalities: Fix some base measure such that P and Q are absolutely continuous with respect to the base measure. (If P and Q are discrete distributions, this can be the counting measure. If they are continuous distributions, this can be the Lebesgue measure. In general, $P + Q$ serves as such a measure.) For $y \in \mathcal{Y}$, let $P(y)$ and $Q(y)$ denote the Radon-Nikodym derivative of P and, respectively, Q with respect to this base measure.

If $e^{-\varepsilon} \cdot Q(S) \leq P(S) \leq e^\varepsilon \cdot Q(S)$ for all measurable S , then the result follows trivially by setting $\delta' = 0$, $P' = P$ and $Q' = Q$, and choosing P'' and Q'' to be arbitrary distributions supported on $S = \{y \in \mathcal{Y} : P(y) \geq Q(y)\}$ and $T = \{y \in \mathcal{Y} : P(y) \leq Q(y)\}$ respectively. Thus we assume that this is not the case and, hence, that $\delta > 0$ and $d_{\text{TV}}(P, Q) > 0$.

Similarly, if $\delta \geq 1$ and $d_{\text{TV}}(P, Q) = 1$, then the result follows trivially by setting $\delta' = 1$, $P'' = P$, $Q'' = Q$, and $P' = Q'$ arbitrary. Thus we assume that $\min\{\delta, d_{\text{TV}}(P, Q)\} < 1$.

Fix $\varepsilon_1, \varepsilon_2 \in [0, \varepsilon]$ to be determined later. Define distributions $P', P'', Q',$ and Q'' (in terms of their Radon-Nikodym derivatives) as follows. For all points $y \in \mathcal{Y}$,

$$\begin{aligned} P'(y) &= \frac{\min\{P(y), e^{\varepsilon_1} \cdot Q(y)\}}{1 - \delta_1}, \\ P''(y) &= \frac{P(y) - (1 - \delta_1)P'(y)}{\delta_1} = \frac{\max\{0, P(y) - e^{\varepsilon_1} \cdot Q(y)\}}{\delta_1}, \\ Q'(y) &= \frac{\min\{Q(y), e^{\varepsilon_2} \cdot P(y)\}}{1 - \delta_2}, \\ Q''(y) &= \frac{Q(y) - (1 - \delta_2)Q'(y)}{\delta_2} = \frac{\max\{0, Q(y) - e^{\varepsilon_2} \cdot P(y)\}}{\delta_2}, \end{aligned}$$

where $\delta_1, \delta_2 \in (0, 1)$ are appropriate normalizing constants. (We will choose ε_1 to avoid $\delta_1 \in \{0, 1\}$ and, likewise, we will choose ε_2 to avoid $\delta_2 \in \{0, 1\}$.)

By construction, $(1 - \delta_1)P' + \delta_1P'' = P$ and $(1 - \delta_2)Q' + \delta_2Q'' = Q$, so the first property is satisfied. Note that P'' is supported on $S = \{y \in \mathcal{Y} : P(y) > e^{\varepsilon_1} \cdot Q(y)\}$ and Q'' is supported on $T = \{y \in \mathcal{Y} : Q(y) > e^{\varepsilon_2} \cdot P(y)\}$, which implies the third property.

If $0 < \delta_1 = \delta_2 \leq \delta$, then we have the appropriate decomposition (with $\delta' = \delta_1 = \delta_2$) and, for all $y \in \mathcal{Y}$, we have

$$e^{-\varepsilon} \leq e^{-\varepsilon_2} \leq \frac{P'(y)}{Q'(y)} = \frac{\min\{P(y), e^{\varepsilon_1} \cdot Q(y)\}}{\min\{Q(y), e^{\varepsilon_2} \cdot P(y)\}} \leq e^{\varepsilon_1} \leq e^{\varepsilon},$$

as required for the second property.

It only remains to show that we can ensure that $0 < \delta_1 = \delta_2 \leq \delta$ by appropriately setting $\varepsilon_1, \varepsilon_2 \in [0, \varepsilon]$. We have

$$\delta_1 = \int_{\mathcal{Y}} \max\{0, P(y) - e^{\varepsilon_1} \cdot Q(y)\} dy = \int_S P(y) - e^{\varepsilon_1} \cdot Q(y) dy = P(S) - e^{\varepsilon_1} Q(S),$$

where $S = \{y \in \mathcal{Y} : P(y) \geq e^{\varepsilon_1} \cdot Q(y)\}$. If $\varepsilon_1 = \varepsilon$, then $\delta_1 \leq \delta$ by assumption. If $\varepsilon_1 = 0$, then $\delta_1 = d_{\text{TV}}(P, Q) > 0$. By decreasing ε_1 , we continuously increase δ_1 . Thus, by starting at $\varepsilon_1 = \varepsilon$ and decreasing ε_1 until either $\varepsilon_1 = 0$ or $\delta_1 = \delta$, we can pick $\varepsilon_1 \in [0, \varepsilon]$ such that $\delta_1 = \min\{\delta, d_{\text{TV}}(P, Q)\} \in (0, 1)$. Similarly, we can pick $\varepsilon_2 \in [0, \varepsilon]$, such that $\delta_2 = \min\{\delta, d_{\text{TV}}(P, Q)\}$. \square

We need a Bayesian version of this decomposition. I.e., suppose we observe a sample from either P or Q and we have a prior on these two possibilities, what is the posterior distribution on possibilities? The following gives such a result. However, it introduces an event $E_{P,Q}$. Intuitively, when $E_{P,Q}(Y) = 1$, then we get the result we would get under pure DP. But $E_{P,Q}(Y) = 0$ with probability δ , in which case things can fail arbitrarily.

Kasiswathan and Smith [KS14, Lemma 3.4] provide a similar result. Ours improves the constant factors and is also stated slightly differently.

Lemma 5.6. *Let P and Q be probability distributions over \mathcal{Y} . Fix $\varepsilon, \delta \geq 0$. Suppose that, for all measurable $S \subset \mathcal{Y}$, we have $P(S) \leq e^\varepsilon \cdot Q(S) + \delta$ and $Q(S) \leq e^\varepsilon P(S) + \delta$.*

Then there exists a randomized function $E_{P,Q} : \mathcal{Y} \rightarrow \{0, 1\}$ with the following properties.

Fix $p \in [0, 1]$ and suppose $X \leftarrow \text{Bernoulli}(p)$. If $X = 1$, sample $Y \leftarrow P$; and, if $X = 0$, sample $Y \leftarrow Q$. Then, for all $y \in \mathcal{Y}$, we have

$$\mathbb{P}_{\substack{X \leftarrow \text{Bernoulli}(p) \\ Y \leftarrow XP + (1-X)Q}} [X = 1 \wedge E_{P,Q}(Y) = 1 | Y = y] \leq \frac{p}{p + (1-p)e^{-\varepsilon}}.$$

Furthermore,

$$\mathbb{E}_{Y \leftarrow P} [E_{P,Q}(Y)] \geq 1 - \delta \quad \text{and} \quad \mathbb{E}_{Y \leftarrow Q} [E_{P,Q}(Y)] \geq 1 - \delta.$$

Proof. We apply the decomposition from Lemma 5.5: There exist distributions P' , Q' , P'' , and Q'' over \mathcal{Y} and $\delta' \in [0, \delta]$ such that

$$\begin{aligned} P &= (1 - \delta')P' + \delta'P'', \\ Q &= (1 - \delta')Q' + \delta'Q'', \end{aligned}$$

and, for all $y \in \mathcal{Y}$, $e^{-\varepsilon}P'(y) \leq Q'(y) \leq e^\varepsilon P'(y)$ and $P''(y) > 0 \implies P(y) \geq Q(y)$ and $Q''(y) > 0 \implies P(y) \leq Q(y)$. (Here $P(\cdot)$ denotes the Radon-Nikodym derivative of the distribution P with respect to some appropriate base measure and similarly for the other distributions.)

We define $E_{P,Q} : \mathcal{Y} \rightarrow \{0, 1\}$ by

$$\mathbb{P}[E_{P,Q}(y) = 1] = (1 - \delta') \cdot \frac{P'(y)}{P(y)} = 1 - \delta' \cdot \frac{P''(y)}{P(y)}.$$

Clearly, $\mathbb{E}_{Y \leftarrow P, E_{P,Q}} [E_{P,Q}(Y)] = \int_{\mathcal{Y}} P(y) \mathbb{P}[E_{P,Q}(Y) = 1] dy = \int_{\mathcal{Y}} (1 - \delta') P'(y) dy = 1 - \delta' \geq 1 - \delta$.

Also

$$\begin{aligned} \mathbb{E}_{Y \leftarrow Q} [E_{P,Q}(Y)] &= 1 - \delta' \mathbb{E}_{Y \leftarrow Q} \left[\frac{P''(y)}{P(y)} \right] \\ &= 1 - \delta' \int_{\mathcal{Y}} \frac{Q(y)}{P(y)} \cdot P''(y) dy \\ &\geq 1 - \delta' \int_{\mathcal{Y}} P''(y) dy \\ &= 1 - \delta' \geq 1 - \delta, \end{aligned}$$

since $P''(y) > 0 \implies P(y) \geq Q(y)$. For any $y \in \mathcal{Y}$, we have

$$\begin{aligned}
& \mathbb{P}[X = 1 \wedge E_{P,Q}(Y) = 1 | Y = y] \\
&= \mathbb{P}[X = 1 | Y = y] \cdot \mathbb{P}[E_{P,Q}(y) = 1] \\
&= \frac{\mathbb{P}[Y = y | X = 1] \cdot \mathbb{P}[X = 1]}{\mathbb{P}[Y = y]} \cdot \mathbb{P}[E_{P,Q}(y) = 1] \\
&= \frac{P(y) \cdot p}{pP(y) + (1-p)Q(y)} \cdot \mathbb{P}[E_{P,Q}(y) = 1] \\
&= \frac{p(1-\delta')P'(y) + p\delta'P''(y)}{p(1-\delta')P'(y) + p\delta'P''(y) + (1-p)(1-\delta')Q'(y) + (1-p)\delta'Q''(y)} \cdot \mathbb{P}[E_{P,Q}(y) = 1] \\
&= \frac{p + p\frac{\delta'P''(y)}{(1-\delta')P'(y)}}{p + p\frac{\delta'P''(y)}{(1-\delta')P'(y)} + (1-p)\frac{Q'(y)}{P'(y)} + (1-p)\frac{\delta'Q''(y)}{(1-\delta')P'(y)}} \cdot \mathbb{P}[E_{P,Q}(y) = 1] \\
&= \frac{p + p\frac{\delta'}{1-\delta'}\frac{P''(y)}{P'(y)}}{p + (1-p)\frac{Q'(y)}{P'(y)} + \frac{\delta'}{1-\delta'} \cdot \frac{pP''(y) + (1-p)Q''(y)}{P'(y)}} \cdot \mathbb{P}[E_{P,Q}(y) = 1] \\
&\leq \frac{p + p\frac{\delta'}{1-\delta'}\frac{P''(y)}{P'(y)}}{p + (1-p)e^{-\varepsilon} + 0} \cdot \mathbb{P}[E_{P,Q}(y) = 1] \\
&= \frac{p}{p + (1-p)e^{-\varepsilon}} \cdot \left(1 + \frac{\delta'}{1-\delta'}\frac{P''(y)}{P'(y)}\right) \cdot \mathbb{P}[E_{P,Q}(y) = 1] \\
&= \frac{p}{p + (1-p)e^{-\varepsilon}} \cdot \left(\frac{P(y)}{(1-\delta')P'(y)}\right) \cdot \mathbb{P}[E_{P,Q}(y) = 1] \\
&= \frac{p}{p + (1-p)e^{-\varepsilon}}.
\end{aligned}$$

□

Now we can prove an analog of Proposition 5.1 for the (ε, δ) -DP setting.

Proposition 5.7 (General Form of Main Result). *Let $M : \{-1, +1\}^m \rightarrow [-1, +1]^m$ satisfy (ε, δ) -DP. Let $S \in \{-1, +1\}^m$ be m independent samples from $2\text{Bernoulli}(p) - 1$ - i.e., $\mathbb{P}[S_i = 1] = p$ independently for each $i \in [m]$. Let $T = M(S) \in [-1, +1]^m$. Then, for all $v \in \mathbb{R}$ and all $t \in [-1, +1]^m$,*

$$\begin{aligned}
& \mathbb{P}_{\substack{S \leftarrow (2\text{Bernoulli}(p)-1)^m, \\ T \leftarrow M(S)}} \left[\sum_i^m \max\{0, T_i \cdot S_i\} \geq v \mid T = t \right] \leq \\
& \mathbb{P}_{\substack{\check{S}^+ \leftarrow \text{Bernoulli}\left(\frac{p \cdot e^{-\varepsilon}}{p \cdot e^{-\varepsilon} + 1 - p}\right)^m, \\ \check{S}^- \leftarrow \text{Bernoulli}\left(\frac{(1-p) \cdot e^{\varepsilon}}{(1-p) \cdot e^{\varepsilon} + p}\right)^m, F}} \left[F(t) + \sum_{i \in [m]: t_i > 0} t_i \cdot \check{S}_i^+ + \sum_{i \in [m]: t_i < 0} -t_i \cdot \check{S}_i^- \geq v \right],
\end{aligned}$$

where $F : [-1, 1]^m \rightarrow \{0, 1, \dots, m\}$ is independent from \check{S}^+ and \check{S}^- and satisfies $\mathbb{E}_{T,F}[F(T)] \leq 2m \cdot \delta$.

Proof. For $i \in [m] \cup \{0\}$ and $s_{\leq i} \in \{-1, 1\}^i$, let $M(s_{\leq i})$ denote the distribution on $[-1, 1]^m$ obtained by conditioning $M(S)$ on $S_{\leq i} = s_{\leq i}$. We can express this as a convex combination:

$$M(s_{\leq i}) = \sum_{s_{>i} \in \{-1, 1\}^{m-i}} M(s_{\leq i}, s_{>i}) \cdot \mathbb{P}_{S_{>i} \leftarrow (2\text{Bernoulli}(p)-1)^{m-i}} [S_{>i} = s_{>i}].$$

For distributions P and Q on $[-1, 1]^m$, let $E_{P,Q} : [-1, 1]^m \rightarrow \{0, 1\}$ be the randomized function promised by Lemma 5.6. In our analysis, the internal randomness of $E_{P,Q}$ is independent from everything else – i.e., the only dependence is induced by its input. Specifically, for all $i \in [m]$, all $s_{<i} \in \{-1, 1\}^{i-1}$, and all $t \in [-1, 1]^m$, we have

$$\begin{aligned} \mathbb{P}_{\substack{S \leftarrow (2\text{Bernoulli}(p)-1)^n, \\ T \leftarrow M(S), E}} [S_i = 1 \wedge E_{M(s_{<i}, 1), M(s_{<i}, -1)}(T) = 1 \mid S_{<i} = s_{<i}, T = t] &\leq \frac{p \cdot e^\varepsilon}{p \cdot e^\varepsilon + 1 - p}, \\ \mathbb{E}_{\substack{S \leftarrow (2\text{Bernoulli}(p)-1)^n, \\ T \leftarrow M(S), E}} [E_{M(s_{<i}, 1), M(s_{<i}, -1)}(T) \mid S_{\leq i} = (s_{<i}, 1)] &\geq 1 - \delta, \\ \mathbb{E}_{\substack{S \leftarrow (2\text{Bernoulli}(p)-1)^n, \\ T \leftarrow M(S), E}} [E_{M(s_{<i}, 1), M(s_{<i}, -1)}(T) \mid S_{\leq i} = (s_{<i}, -1)] &\geq 1 - \delta. \end{aligned}$$

Symmetrically, for all $i \in [m]$, all $s_{<i} \in \{-1, 1\}^{i-1}$, and all $t \in [-1, 1]^m$, we have

$$\begin{aligned} \mathbb{P}_{\substack{S \leftarrow (2\text{Bernoulli}(p)-1)^n, \\ T \leftarrow M(S), E}} [S_i = -1 \wedge E_{M(s_{<i}, -1), M(s_{<i}, 1)}(T) = 1 \mid S_{<i} = s_{<i}, T = t] &\leq \frac{(1-p) \cdot e^\varepsilon}{(1-p) \cdot e^\varepsilon + p}, \\ \mathbb{E}_{\substack{S \leftarrow (2\text{Bernoulli}(p)-1)^n, \\ T \leftarrow M(S), E}} [E_{M(s_{<i}, -1), M(s_{<i}, 1)}(T) \mid S_{\leq i} = (s_{<i}, -1)] &\geq 1 - \delta, \\ \mathbb{E}_{\substack{S \leftarrow (2\text{Bernoulli}(p)-1)^n, \\ T \leftarrow M(S), E}} [E_{M(s_{<i}, -1), M(s_{<i}, 1)}(T) \mid S_{\leq i} = (s_{<i}, 1)] &\geq 1 - \delta. \end{aligned}$$

For simplicity, we define a symmetric event: $E_P^Q(y) = E_Q^P(y) := E_{P,Q}(y) \cdot E_{Q,P}(y)$, where the internal randomnesses are again independent. Combining these, we have, for all $i \in [m]$, all $s_{<i} \in \{-1, 1\}^{i-1}$, and all $t \in [-1, 1]^m$,

$$\mathbb{P}_{\substack{S \leftarrow (2\text{Bernoulli}(p)-1)^n, \\ T \leftarrow M(S), E}} [S_i = \text{sign}(T_i) \wedge E_{M(s_{<i}, -1), M(s_{<i}, 1)}^M(T) = 1 \mid T = t, S_{<i} = s_{<i}] \leq \begin{cases} \frac{p \cdot e^\varepsilon}{p \cdot e^\varepsilon + 1 - p} & \text{if } t_i > 0 \\ \frac{(1-p) \cdot e^\varepsilon}{(1-p) \cdot e^\varepsilon + p} & \text{if } t_i < 0 \end{cases}$$

and, for $b \in \{-1, 1\}$, we have

$$\mathbb{E}_{\substack{S \leftarrow (2\text{Bernoulli}(p)-1)^n, \\ T \leftarrow M(S), E}} [E_{M(s_{<i}, -1), M(s_{<i}, 1)}^M(T) \mid S_{\leq i} = (s_{<i}, b)] \geq 1 - 2\delta.$$

For $k \in [m]$, $s \in \{-1, 1\}^m$, and $t \in [-1, 1]^m$, define

$$\widetilde{W}_k(s, t) := \sum_{i \in [k]} \max\{0, t_i \cdot s_i\} \cdot E_{M(s_{<i}, -1), M(s_{<i}, 1)}^M(t) = 1 = \sum_{i \in [k]} |t_i| \cdot \mathbb{I}[s_i = \text{sign}(t_i) \wedge E_{M(s_{<i}, -1), M(s_{<i}, 1)}^M(t) = 1]$$

and

$$\check{W}_k(t) = \sum_{i \in [k]} \check{S}_i(t) \cdot |t_i|,$$

where, for each $i \in [k]$ independently, $\check{S}(t)_i \leftarrow \text{Bernoulli}(\frac{p \cdot e^\varepsilon}{p \cdot e^\varepsilon + 1 - p})$ if $t_i > 0$ and $\check{S}(t)_i \leftarrow \text{Bernoulli}(\frac{(1-p) \cdot e^\varepsilon}{(1-p) \cdot e^\varepsilon + p})$ if $t_i < 0$.

By induction and Lemma 4.9, for any $k \in [m]$ and $t \in [-1, 1]^m$, the conditional distribution $(\check{W}_k(S, t) | M(S) = t)$ where $S \leftarrow (2\text{Bernoulli}(p) - 1)^m$ is stochastically dominated by $\check{W}_k(t)$.

For $s \in \{-1, 1\}^m$ and $t \in [-1, 1]^m$, define

$$F(s, t) := \sum_i^m \mathbb{I} \left[E_{M(s_{<i,1})}^{M(s_{<i,-1})}(t) = 0 \right],$$

so that

$$W_m(s, t) := \sum_{i \in [m]} \max\{0, t_i \cdot s_i\} \leq \check{W}_m(s, t) + F(s, t).$$

Since the conditional distribution $(\check{W}_k(S, t) | M(S) = t)$ where $S \leftarrow (2\text{Bernoulli}(p) - 1)^m$ is stochastically dominated by $\check{W}_k(t)$, W_m is stochastically dominated by the convolution $\check{W}_m(T) + F(S, T)$.

Finally $F(s, t)$ is supported on $\{0, 1, \dots, m\}$ and

$$\mathbb{E}[F(s, t)] = \sum_i^m \mathbb{P} \left[E_{M(s_{<i,1})}^{M(s_{<i,-1})}(T) = 0 \right] \leq 2m \cdot \delta.$$

Since $\check{W}_m(T)$ does not depend on S , the input S does not contribute to the dependence between $F(S, T)$ and $\check{W}_m(T)$, so we can elide this input in the statement – i.e., $F(T) = F(S, T)$ for S drawn from an appropriate distribution. \square

Proposition 5.7 is rather unwieldy. It can be simplified by setting $p = \frac{1}{2}$ and identifying the optimal distribution $F(T)$, which yields Theorem 5.2.

Proof of Theorem 5.2. Let $M : \{-1, 1\}^m \rightarrow [-1, 1]^m$ satisfy (ε, δ) -DP. Let $S \in \{-1, 1\}^m$ be uniformly random. Let $T = M(S) \in [-1, 1]^m$. Setting $p = \frac{1}{2}$ in Proposition 5.7 and averaging over T , we have, for all $v \in \mathbb{Z}$,

$$\mathbb{P}_{\substack{S \leftarrow \{-1, 1\}^m \\ T \leftarrow M(S)}} \left[\sum_i^m \max\{0, T_i \cdot S_i\} \geq v \right] \leq \mathbb{P}_{\substack{S \leftarrow \{-1, 1\}^m, T \leftarrow M(S), \\ \check{S} \leftarrow \text{Bernoulli}(\frac{e^\varepsilon}{e^\varepsilon + 1})^m, F}} \left[F(T) + \sum_i^m \check{S}_i \cdot |T_i| \geq v \right],$$

where F is arbitrary – but independent from \check{S} – except for the constraints that $F(T)$ is supported on $\{0, 1, \dots, m\}$ and $\mathbb{E}[F(T)] \leq 2m \cdot \delta$.

Given these constraints, we can formulate finding the optimal distribution $F(t)$ for a given $t \in [-1, 1]^m$ and $v \in \mathbb{R}$ as a linear program:

$$\begin{aligned}
& \text{maximize} && \mathbb{P}_{\check{W}, F} [\check{W}(t) + F(t) \geq v] = \sum_{i=0}^m \mathbb{P}_F [F(t) = i] \cdot \mathbb{P}_{\check{W}} [\check{W}(t) \geq v - i] \\
& \text{subject to} && \mathbb{E}_F [F(t)] = \sum_{i=0}^m \mathbb{P}_F [F(t) = i] \cdot i \leq 2m \cdot \delta, \\
& && \sum_{i=0}^m \mathbb{P}_F [F(t) = i] = 1, \text{ and} \\
& && \mathbb{P}_F [F(t) = i] \geq 0 \quad \forall i \in \{0, 1, \dots, m\},
\end{aligned}$$

where $\check{W}(t) := \sum_i^m \check{S}_i |t_i|$ for $\check{S} \leftarrow \text{Bernoulli} \left(\frac{e^\varepsilon}{e^\varepsilon + 1} \right)^m$.

By strong duality, the linear program above has the same value as its dual:

$$\begin{aligned}
& \text{minimize} && 2m\delta\alpha + \beta \\
& \text{subject to} && \alpha \cdot i + \beta \geq \mathbb{P}_{\check{W}} [\check{W}(t) \geq v - i] \quad \forall i \in \{0, 1, \dots, m\}, \\
& && \alpha \geq 0.
\end{aligned}$$

Any feasible solution to the dual gives an upper bound on the primal. So, in particular, we can use the solution given by

$$\begin{aligned}
\beta &= \mathbb{P}_{\check{W}^*} [\check{W}^* \geq v], \\
\alpha &= \max \left(\{0\} \cup \left\{ \frac{1}{i} \left(\mathbb{P}_{\check{W}^*} [\check{W}^* \geq v - i] - \beta \right) : i \in \{1, 2, \dots, m\} \right\} \right),
\end{aligned}$$

where \check{W}^* is a distribution on \mathbb{R} that satisfies $\mathbb{P}_{\check{W}^*} [\check{W}^* \geq v - i] \geq \mathbb{P}_{\check{W}} [\check{W}(t) \geq v - i]$ for all $i \in \{0, 1, \dots, m\}$ and all t in the support of T . \square

Theorem 5.2 gives a worst-case bound in terms of T . Specifically, \check{W}^* must uniformly bound $\check{W}(t)$ for all t in the support of T . Proposition 5.7 is more general than this. Thus we give another corollary that allows us to have the bound adjust to T . In particular, this result allows the auditing procedure (Algorithm 1 or 3) to dynamically choose the number of guesses $r = k_+ + k_-$.

Corollary 5.8 (Variant of Main Result). *Let $M : \{-1, +1\}^m \rightarrow [-1, +1]^m$ satisfy (ε, δ) -DP. Let $S \in \{-1, +1\}^m$ be uniformly random. Let $T = M(S) \in [-1, +1]^m$. Then, for all $\gamma \in [0, 1]$ and $\tau > 0$,*

$$\mathbb{P}_{\substack{S \leftarrow \{-1, +1\}^m \\ T \leftarrow M(S)}} \left[\sum_i^m \max\{0, T_i \cdot S_i\} \geq g_{m, \varepsilon}(T, \gamma) + \tau \right] \leq \gamma + \frac{2m\delta}{\tau},$$

where $g_{m,\varepsilon} : [-1, +1]^m \times [0, 1] \rightarrow \mathbb{R}$ is an arbitrary function satisfying

$$\forall t \in [-1, 1]^m \quad \forall \gamma \in [0, 1] \quad \mathbb{P}_{\check{S} \leftarrow \text{Bernoulli}\left(\frac{e^\varepsilon}{e^\varepsilon + 1}\right)^m} \left[\sum_i^m |t_i| \cdot \check{S}_i \geq g_{m,\varepsilon}(t, \gamma) \right] \leq \gamma.$$

Proof. Setting $p = \frac{1}{2}$ in Proposition 5.7 yields

$$\begin{aligned} \forall v \in \mathbb{R} \quad \forall t \in [-1, +1]^m \quad & \mathbb{P}_{\substack{S \leftarrow \{-1, +1\}^m \\ T \leftarrow M(S)}} \left[\sum_i^m \max\{0, T_i \cdot S_i\} \geq v \mid T = t \right] \\ & \leq \mathbb{P}_{\check{S} \leftarrow \text{Bernoulli}\left(\frac{e^\varepsilon}{e^\varepsilon + 1}\right), F} \left[F(t) + \sum_i^m |t_i| \cdot \check{S}_i \geq v \right], \end{aligned}$$

where $F : [-1, 1]^m \rightarrow \{0, 1, \dots, m\}$ satisfies $\mathbb{E}_{\substack{S \leftarrow \{-1, +1\}^m \\ T \leftarrow M(S), F}} [F(T)] \leq 2m\delta$.

By a union bound and Markov's inequality, we have, for all $t \in [-1, 1]^m$, all $\gamma \in [0, 1]$, and all $\tau > 0$,

$$\begin{aligned} & \mathbb{P}_{\check{S} \leftarrow \text{Bernoulli}\left(\frac{e^\varepsilon}{e^\varepsilon + 1}\right), F} \left[F(t) + \sum_i^m |t_i| \cdot \check{S}_i \geq g_{m,\varepsilon}(t, \gamma) + \tau \right] \\ & \leq \mathbb{P}_{\check{S} \leftarrow \text{Bernoulli}\left(\frac{e^\varepsilon}{e^\varepsilon + 1}\right), F} \left[\tau + \sum_i^m |t_i| \cdot \check{S}_i \geq g_{m,\varepsilon}(t, \gamma) + \tau \right] + \mathbb{P}_F [F(t) > \tau] \\ & \leq \mathbb{P}_{\check{S} \leftarrow \text{Bernoulli}\left(\frac{e^\varepsilon}{e^\varepsilon + 1}\right), F} \left[\sum_i^m |t_i| \cdot \check{S}_i \geq g_{m,\varepsilon}(t, \gamma) \right] + \frac{\mathbb{E}_F [F(t)]}{\tau} \\ & \leq \gamma + \frac{\mathbb{E}_F [F(t)]}{\tau}. \end{aligned}$$

We combine inequalities, set $v = g_{m,\varepsilon}(t, \gamma) + \tau$, and average over T to obtain

$$\begin{aligned} \forall \gamma \in [0, 1] \quad \forall \tau > 0 \quad & \mathbb{P}_{\substack{S \leftarrow \{-1, +1\}^m \\ T \leftarrow M(S)}} \left[\sum_i^m \max\{0, T_i \cdot S_i\} \geq g_{m,\varepsilon}(T, \gamma) + \tau \right] \\ & \leq \gamma + \frac{\mathbb{E}_{\substack{S \leftarrow \{-1, +1\}^m \\ T \leftarrow M(S), F}} [F(T)]}{\tau} \\ & \leq \gamma + \frac{2m\delta}{\tau}. \end{aligned}$$

□

Algorithm 3 DP-SGD Auditor (Instantiation of Algorithm 1)

- 1: **Data:** $x \in \mathcal{X}^n$ consisting of m auditing examples (a.k.a. canaries) and $n - m$ non-auditing examples.
 - 2: **Parameters:** Number of examples to randomize m for audit, number of positive k_+ and negative k_- guesses **audit-type** (either **black-box** or **white-box**).
 - 3: For $i \in [m]$ sample $S_i \in \{-1, +1\}$ independently with $\mathbb{E}[S_i] = 0$. Set $S_i = 1$ for all $i \in [n] \setminus [m]$.
 - 4: Split x into $x_{\text{IN}} \in \mathcal{X}^{n_{\text{IN}}}$ and $x_{\text{OUT}} \in \mathcal{X}^{n_{\text{OUT}}}$ according to S , where $n_{\text{IN}} + n_{\text{OUT}} = n$. Namely, if $S_i = 1$, then x_i is in x_{IN} ; and, if $S_i = -1$, then x_i is in x_{OUT} .
 - 5: Run DP-SGD (Algorithm 2) on input x_{IN} with appropriate parameters.
 - 6: Let ℓ be the number of iterations and let $f : \mathbb{R}^d \times \mathcal{X} \rightarrow \mathbb{R}$ be the loss.
 - 7: Let $w^0, \dots, w^\ell \in \mathbb{R}^d$ be the output of DP-SGD.
 - 8: **if audit-type = black-box then**
 - 9: Define $\text{SCORE}(x_i, w^\ell) = \ell(w^0, x_i) - \ell(w^\ell, x_i)$ for all $i \in [m]$.
 - 10: Compute the vector of scores $Y = (\text{SCORE}(x_i, w^\ell) : i \in [m]) \in \mathbb{R}^m$.
 - 11: **else if audit-type = white-box then**
 - 12: **procedure** $\text{SCORE}(x_*, w^1, \dots, w^\ell)$
 - 13: **for** $t = 1, \dots, \ell$ **do**
 - 14: Compute $g^t = \nabla_{w_{t-1}} \ell(w_{t-1}, x_*) \in \mathbb{R}^d$.
 - 15: Clip $\hat{g}^t = \min \left\{ 1, \frac{c}{\|g^t\|_2} \right\} \cdot g^t \in \mathbb{R}^d$.
 - 16: Let $v^t = \langle w^{t-1} - w^t, \hat{g}^t \rangle \in \mathbb{R}$.
 - 17: **end for**
 - 18: Return $\sum_{t=1}^\ell v^t \in \mathbb{R}$.
 - 19: **end procedure**
 - 20: Compute the vector of scores $Y = (\text{SCORE}(x_i, w^0, w^1, \dots, w^\ell) : i \in [m]) \in \mathbb{R}^m$.
 - 21: **end if**
 - 22: Sort the scores Y . Let $T \in \{-1, 0, +1\}^m$ be $+1$ for the largest k_+ scores and -1 for the smallest k_- scores.
 - 23: (I.e., $T \in \{-1, 0, +1\}^m$ maximizes $\sum_i^m T_i \cdot Y_i$ subject to $\sum_i^m |T_i| = k_+ + k_-$ and $\sum_i^m T_i = k_+ - k_-$.)
 - 24: **Return:** The vector $S \in \{-1, +1\}^m$ indicating the true selection and the guesses $T \in \{-1, 0, +1\}^m$.
-

6 Experiments

Experiment Setup Our contributions are focused on improved analysis of an existing privacy attack, and are therefore orthogonal to the design of an attack. As a result, we rely on the experimental setup of the recent auditing procedure of Nasr, Hayes, Steinke, Balle, Tramèr, Jagielski, Carlini, and Terzis [NHSBTJCT23].

We run DP-SGD on the CIFAR-10 dataset with Wide ResNet (WRN-16) [ZK16], we followed the experimental setup from Nasr et al. [NHSBTJCT23]. Our experiments reach 76% test accuracy at $(\varepsilon = 8, \delta = 10^{-5})$ -DP, which is comparable with the state-of-the-art [DBHSB22]. Unless specified otherwise, all lower bounds are presented with 95% confidence. Following Nasr et al. [NHSBTJCT23], we refer to the setting where the adversary has access to all intermediate steps as “white-box” and when the adversary can only see the last iteration as “black-box.” We experiment with both settings.

Algorithm 3 summarizes our approach for auditing DP-SGD. The results are converted into lower bounds on the privacy parameters using Theorem 5.2 / Corollary 5.4.

We also experiment with both the gradient and input attacks proposed by Nasr et al. [NHSBTJCT23]. In particular, for the gradient attack we use the strongest attack they

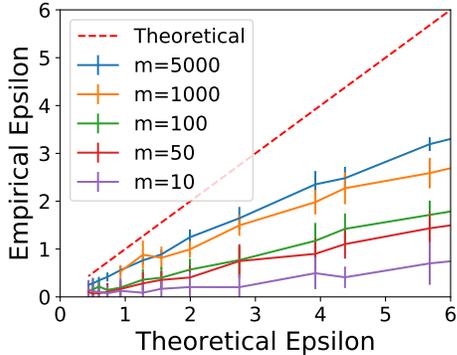


Figure 4: Effect of the number of auditing examples (m) in the white-box setting. By increasing the number of the auditing examples we are able to achieve tighter empirical lower bounds.

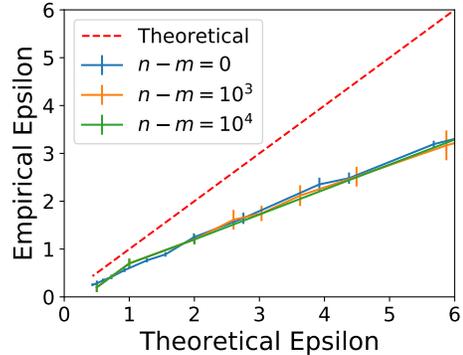


Figure 5: Effect of the number of additional examples ($n - m$) in the white-box setting. Importantly, adding additional examples does not impact the auditing results in the white-box setting.

proposed – the “Dirac canary” approach – which sets all gradients to zero except at a single random index. In our setting where we need to create multiple auditing examples (canaries) we make sure the indices selected in our experiments do not have any repetitions. To compute the score for gradient space attacks, we use the dot product between the gradient update and auditing gradient. When auditing in input space, we leverage two different types of injected examples as:

1. **Mislabeled example:** We select a random subset of the test set and randomly relabel them (ensuring the new label is not the same as the original label).
2. **In-distribution example:** We select a random subset of the test set.

For input space audits, we use the loss of the input example as the score. In our experiments we report the attack with the highest lower bound.

In our experiments, we evaluate different values of k_+ and k_- and only report the highest auditing results. Since this is doing multiple hypothesis testing on the same data, we are reducing the confidence value of our results. However, this is commonly used in the previous works [ZBWTSRPNK22; MSS22] and can be easily improved by using a different set of observations to select the parameters for the auditing and another set of the data for the auditing itself (see also Corollary 5.8).

6.1 Gradient Space attacks

We start with the strongest attack: We assume white-box access – i.e., the auditor sees all intermediate iterates of DP-SGD – and that the auditor can insert examples with arbitrary gradients into the training procedure. First, we evaluate the effect of the number of the auditing example on the tightness. Figure 4 demonstrates that as the number of examples

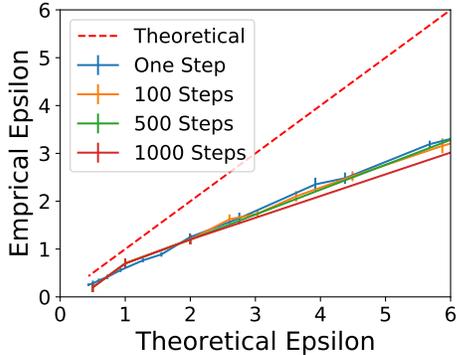


Figure 6: Effect of number of iterations in the white-box setting. Increasing the number of the steps (while keeping the same overall privacy by increasing the added noise) will not effect the auditing results.

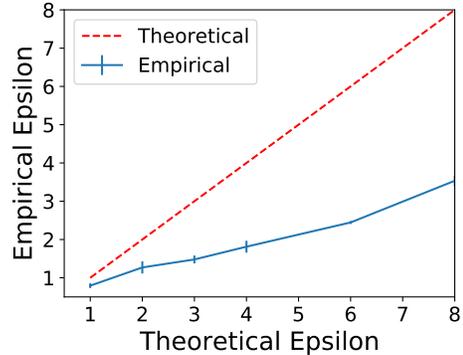


Figure 7: Auditing CIFAR10 SoTA in white-box setting using gradient attacks. Our auditing framework can achieve meaningful empirical privacy lower bounds for SoTA models.

increases, the auditing becomes tighter. However, the impact of the additional examples eventually diminishes. Intriguingly, adding more non-auditing training examples (resulting in a larger n compared to m) does not seem to influence the tightness of the auditing, as depicted in Figure 5. This can be primarily due to the fact that gradient attacks proposed in prior studies can generate near-worst-case datasets, irrespective of the presence of other data points.

Another parameter that might affect the auditing results is the number of iterations ℓ in the DP-SGD algorithm. As shown in Figure 6 we compare the extreme setting of having one iteration to multiple iterations and we do not observe any significant difference in the auditing when auditing for the equivalent privacy guarantees (by increasing the noise). The results confirm the tightness of composition and that the number of iterations does not have significant effect on auditing in white-box setting.

Now we directly use the parameters used in the training CIFAR10 models. Figure 7 summarizes results for the CIFAR10 models. We used $m = 5000$ and all of the training dataset from CIFAR10 ($n = 50,000$) for the attack. We were able to achieve 76% accuracy for $\varepsilon = 8$ ($\delta = 10^{-5}$, compared to 78% when not auditing). We are able to achieve an empirical lower bound of 0.7, 1.2, 1.8, 3.5 for theoretical epsilon of 1, 2, 4, 8 respectively. While our results are not as tight as the prior works, we only require a single run of training which is not possible using the existing techniques. In the era of exponentially expanding machine learning models, the computational and financial costs of training these colossal architectures even once are significant. Expecting any individual or entity to shoulder the burden of training such models thousands of times for the sake of auditing or experimental purposes is both unrealistic and economically infeasible. Our method offers a unique advantage by facilitating the auditing of these models, allowing for an estimation of privacy leakage in a

white-box setting without significantly affecting performance.

6.2 Input Space Attacks

Now we evaluate the effect of input space attacks in the black-box setting. In this attack, the auditor can only insert actual images into the training procedure and cannot control any of the aspects of the training. Then, the adversary can observe the final model as mentioned in Algorithm 3. This is the weakest attack setting.

For simplicity we start with the setting where $m = n$; in other words, all of the examples used to train the model are randomly included or excluded and can be used for auditing. Figure 8 illustrates the result of this setting. As we see from the figure, unlike the white-box attack we do not observe a monotonic relationship between the number of auditing examples and the tightness of the auditing. Intuitively, when the number of auditing examples are low then we do not have enough observations to have high confidence lower bounds for epsilon. On the other hand, when the number of auditing examples are high, the model does not have enough capacity to “memorize” all of the auditing examples which reduces the tightness of the auditing. However, this can be improved by designing better black-box attacks which we reiterate in the next section.

We also evaluate the effect of adding additional training data to the auditing in Figure 9. We see that adding superfluous training data significantly reduces the effectiveness of auditing. The observed reduction in auditing effectiveness with the addition of more training data could be attributed to several factors. One interpretation could be that the theoretical privacy analysis in a black-box setting tends to be considerably more loose when the adversary is constrained to this setting. This could potentially result in an overestimation of the privacy bounds. Conversely, it is also plausible that the results are due to the weak black-box attacks and can be improved in the future.

7 Discussion

Our main contribution is showing that we can audit the differential privacy guarantees of an algorithm with a single run. In contrast, prior methods require hundreds – if not thousands – of runs, which is computationally prohibitive for all but the simplest algorithms. Our experimental results demonstrate that in practical settings our methods are able to give meaningful lower bounds on the privacy parameter ϵ .

However, while we win on computational efficiency, we lose on tightness of our lower bounds. We now illustrate the limitations of our approach and discuss the extent to which this is inherent, and what lessons we can learn.

But, first, we illustrate that our method can give tight lower bounds. In Figure 10, we consider an idealized setting where the number of guesses changes and the fraction that are correct is fixed at $\frac{e^\epsilon}{e^\epsilon+1}$ for $\epsilon = 4$ – i.e., 98.2% of guesses are correct.⁴ This is the maximum expected fraction of correct guesses compatible with $(4, 0)$ -DP. In this setting the lower

⁴The number of correct guesses is rounded down to an integer (which results in the lines being jagged).

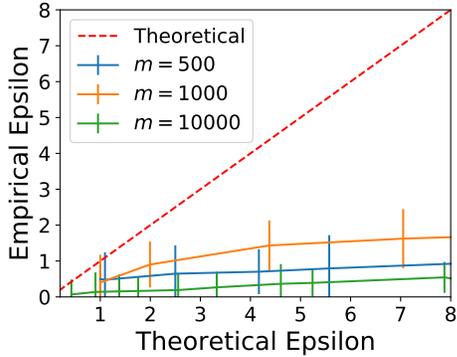


Figure 8: Effect of the number of auditing examples (m) in the black-box setting. Black-box auditing is very sensitive to the number of auditing examples.

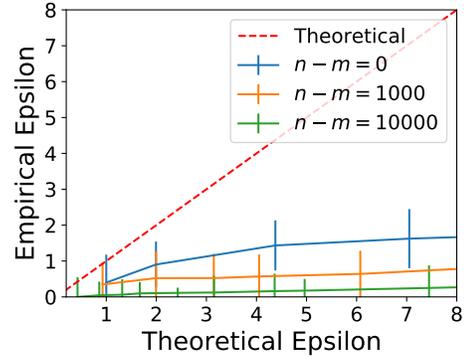


Figure 9: Effect of the number of additional example on auditing ($n - m$) in the black-box setting. By increasing the number of additional examples, the auditing results get significantly looser.

bound on ε does indeed come close to 4. With 10,000 guesses we get $\varepsilon \geq 3.87$ with 95% confidence.

Note that the lower bound in Figure 10 improves as we increase the number of guesses. This is simply accounting for sampling error – to get a lower bound with 95% confidence, we must underestimate to account for the fact that the number of correct guesses may have been inflated by chance. As we get more guesses, the relative size of chance deviations reduces.

Limitations: Next we consider a different idealized setting – one that is arguably more realistic – where our method does not give tight lower bounds. Suppose $S_i \in \{-1, +1\}$ indicates whether example $i \in [n]$ is included or excluded. In Figure 11, we consider Gaussian noise addition. That is, we release a sample from $\mathcal{N}(S_i, 4)$. (In contrast, Figure 10 considers randomized response on S_i .) Lemma 4.5 gives an upper bound of $(4.38, 10^{-5})$ -DP. Unlike for randomized response, abstentions matter here. We consider 100,000 examples, each of which has a score sampled from $\mathcal{N}(S_i, 4)$, where $S_i \in \{-1, +1\}$ is uniformly random. We pick the largest $r/2$ scores and guess $S_i = +1$. Similarly we guess $S_i = -1$ for the smallest $r/2$ scores. We abstain for the remaining $100,000 - r$ examples. If we make more guesses (i.e., increase r), then the accuracy goes down and so does our lower bound. We must trade off between more guesses being less accurate on average and more guesses having smaller relative sampling error.

In Figure 11, the highest value of the lower bound is $\varepsilon \geq 2.675$ for $\delta = 10^{-5}$, which is attained by 1439 correct guesses out of 1510. In contrast, the upper bound is $\varepsilon = 4.38$ for $\delta = 10^{-5}$. To get a matching upper bound of $\varepsilon \leq 2.675$ we would need to set $\delta = 0.0039334$. In other words, the gap between the upper and lower bounds is a factor of $393 \times$ in δ .

Figure 12 considers the same idealized setting as Figure 11, but we fix the number of guesses to 1,500 out of 100,000 (of which 1,429 are correct); instead we vary δ and consider

There are no abstentions.

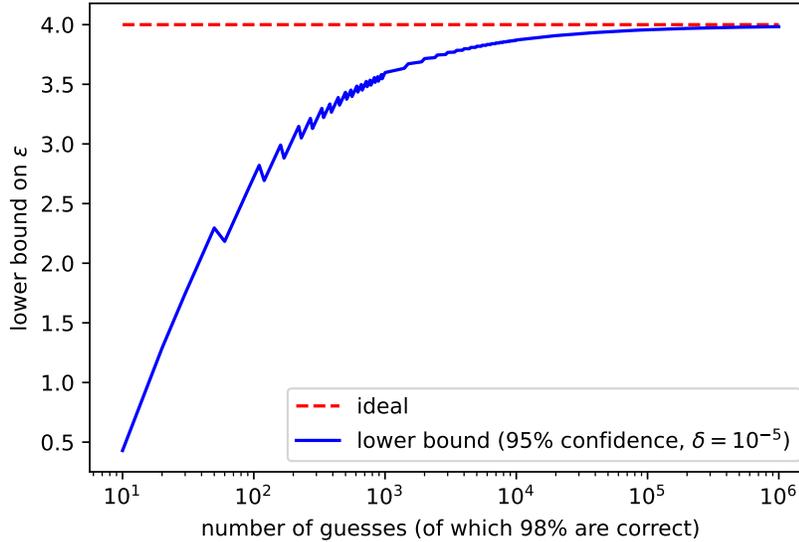


Figure 10: Comparison of upper and lower bounds for idealized setting with varying number of guesses. The fraction of correct guesses is always $\frac{e^\epsilon}{e^\epsilon+1}$ for $\epsilon = 4$ (i.e., 98.2%).

different confidence levels.

Are these limitations inherent? Figures 11 & 12 illustrate the limitations of our approach. They also hint at the causes: The number of guesses versus abstentions, the δ parameter, and the confidence all have a large effect on the tightness of our lower bound.

Our theoretical analysis is fairly tight; there is little room to improve Theorem 5.2. We argue that the inherent problem is a mismatch between “realistic” DP algorithms and the “pathological” DP algorithms for which our analysis is nearly tight. This mismatch makes our lower bound much more sensitive to δ than it “should” be.

To be concrete about what we consider pathological, consider $M : \{-1, +1\}^m \rightarrow \{-1, 0, +1\}^m$ defined by Algorithm 4. This algorithm satisfies (ϵ, δ) -DP and makes r guesses with $m-r$ abstentions. In the $X = 1$ case, the expected fraction of correct guesses is $\frac{m\delta}{r\beta} + \left(1 - \frac{m\delta}{r\beta}\right) \cdot \frac{e^\epsilon}{e^\epsilon+1}$. This is higher than the average fraction of correct guesses, but if we want confidence $1 - \beta$ in our lower bound, we must consider this case, as $X = 1$ happens with probability β .

Intuitively, the contribution from δ to the fraction of correct guesses should be negligible. However, we see that δ is multiplied by $m/r\beta$. That is to say, in the settings we consider, δ is multiplied by a factor on the order of $100\times$ or $1000\times$, which means $\delta = 10^{-5}$ makes a non-negligible contribution to the fraction of correct guesses.

It is tempting to try to circumvent this problem by simply setting δ to be very small. However, as shown in Figure 12, the upper bound on ϵ also increases as $\delta \rightarrow 0$.

Unfortunately, there is no obvious general way to rule out algorithms that behave like Algorithm 4. The fundamental issue is that the privacy losses of the m examples are not

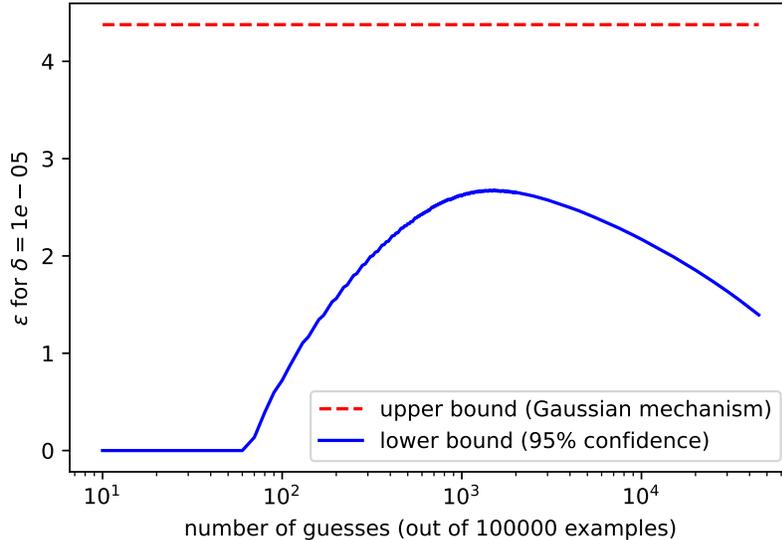


Figure 11: Comparison of upper and lower bounds for idealized setting with varying number of guesses. For each example $i \in [m]$, we release $S_i + \xi_i$, where $\xi_i \leftarrow \mathcal{N}(0, 4)$ and $S_i \in \{-1, +1\}$ is independently uniformly random and indicates whether the sample is included/excluded. For the upper bound, we compute the exact $(4.38, 10^{-5})$ -DP guarantee for the Gaussian mechanism (Lemma 4.5). For the lower bound, we plot the bound of Theorem 5.2 with 95% confidence for varying numbers of guesses r . We consider a total of $m = 100,000$ randomized examples; we guess $T_i = +1$ for the largest $r/2$ scores and we guess $T_i = -1$ for the smallest $r/2$ scores; we guess $T_i = 0$ for the remaining $m - r$ examples. The number of correct guesses is set to $\lceil r \cdot \mathbb{P}[S_i = +1 | S_i + \xi_i > c] \rceil$, where c is a threshold such that $\mathbb{P}[S_i + \xi_i > c] = \frac{r}{2m}$.

independent; we shouldn't expect them to be independent, but we also shouldn't expect them to be pathologically dependent in reality.

Directions for further work: Our work highlights several questions for further exploration:

- **Improved attacks:** Our experimental evaluation uses existing attack methods. Any improvements to membership inference attacks could be combined with our results to yield improved privacy auditing.

One limitation of our attacks is that some examples may be “harder” than others and the scores we compute do not account for this. When we have many runs, we can account for the hardness of individual examples [CCNSTT22], but in our setting it is not obvious how to do this.

- **Algorithm-specific analyses:** Our methods are generic – they can be applied to essentially any DP algorithm. This is a strength, but there is also the possibility that we could obtain stronger results by exploiting the structure of specific algorithms. A

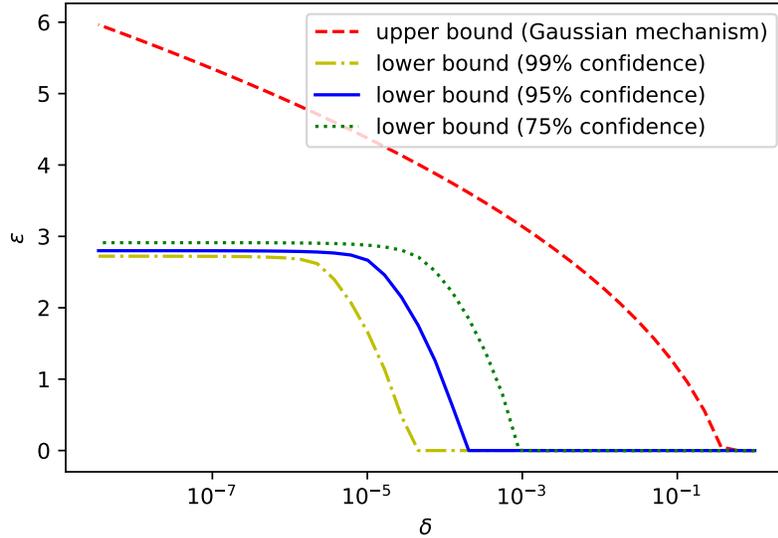


Figure 12: Comparison of upper and lower bounds for idealized setting with varying δ . For each example $i \in [m]$, we release $S_i + \xi_i$, where $\xi_i \leftarrow \mathcal{N}(0, 4)$ and $S_i \in \{-1, +1\}$ is independently uniformly random and indicates whether the sample is included/excluded. For the upper bound, we compute the exact $(4.38, 10^{-5})$ -DP guarantee for the Gaussian mechanism (Lemma 4.5). For the lower bound, we plot the bound of Theorem 5.2 with 75%, 95%, and 99% confidence. We consider $m = 100,000$ randomized examples and 1,500 guesses of which 1,429 are correct. This corresponds to guessing $T_i = +1$ for the largest 750 scores, $T_i = -1$ for the smallest 750 scores, and $T_i = 0$ for the remaining 98,500 examples.

natural example of such structure is the iterative nature of DP-SGD. That is, we can view one run of DP-SGD as the composition of multiple independent DP algorithms which are run sequentially.

- **Multiple runs & multiple examples:** Our method performs auditing by including or excluding multiple examples in a single training run, while most prior work performs multiple training runs with a single example included or excluded. Can we get the best of both worlds? If we use multiple examples and multiple runs, we should be able to get tighter results with fewer runs.
- **Other measures of privacy:** Our theoretical analysis is tailored to the standard definition of differential privacy. But there are other definitions of differential privacy such as Rényi DP. And, in particular, many of the upper bounds (e.g., Proposition 4.6) are stated in this language. Hence it would make sense for the lower bounds also to be stated in this language.
- **Beyond lower bounds:** Privacy auditing produces empirical lower bounds on the

Algorithm 4 Pathological Algorithm

```
1: Input:  $s \in \{-1, +1\}^m$ 
2: Parameters:  $r \in [m]$ ,  $\varepsilon, \delta \geq 0$ ,  $\beta \in [0, 1]$ . Assume  $0 < m\delta \leq r\beta$ .
3: Select  $U \subset [m]$  of size  $|U| = r$  uniformly at random.
4: Set  $T_i = 0$  for all  $i \notin U$ .
5: Sample  $X \leftarrow \text{Bernoulli}(\beta)$ .
6: if  $X = 1$  then
7:   for  $i \in U$  do
8:     Independently sample  $T_i \in \{-1, +1\}$  with  $\mathbb{P}[T_i = s_i] = \frac{m\delta}{r\beta} + \left(1 - \frac{m\delta}{r\beta}\right) \cdot \frac{e^\varepsilon}{e^\varepsilon + 1}$ .
9:   end for
10: else if  $X = 0$  then
11:   for  $i \in U$  do
12:     Independently sample  $T_i \in \{-1, +1\}$  with  $\mathbb{P}[T_i = s_i] = \frac{e^\varepsilon}{e^\varepsilon + 1}$ .
13:   end for
14: end if
15: Output:  $T \in \{-1, 0, +1\}^m$ .
```

privacy parameters. In contrast, mathematical analysis produces upper bounds. Both are necessarily conservative, which leaves a large gap between the upper and lower bounds. A natural question is to find some middle ground – an estimate which is neither a lower nor upper bound, but provides some meaningful estimate of the “true” privacy loss. However, it is unclear what kind of guarantee such an estimate should satisfy, or what interpretation the estimate should permit.

References

- [ACGMMTZ16] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. “Deep learning with differential privacy”. In: *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 2016, pp. 308–318. URL: <https://arxiv.org/abs/1607.00133> (cit. on p. 7).
- [AKOOMS23] G. Andrew, P. Kairouz, S. Oh, A. Oprea, H. B. McMahan, and V. Suriyakumar. “One-shot Empirical Privacy Estimation for Federated Learning”. In: *arXiv preprint arXiv:2302.03098* (2023). URL: <https://arxiv.org/abs/2302.03098> (cit. on p. 5).
- [BGDCTV18] B. Bichsel, T. Gehr, D. Drachler-Cohen, P. Tsankov, and M. Vechev. “Dp-finder: Finding differential privacy violations by sampling and optimization”. In: *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. 2018, pp. 508–524 (cit. on p. 5).

- [BNSSSU16] R. Bassily, K. Nissim, A. Smith, T. Steinke, U. Stemmer, and J. Ullman. “Algorithmic stability for adaptive data analysis”. In: *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*. 2016, pp. 1046–1059. URL: <https://arxiv.org/abs/1511.02513> (cit. on pp. 2, 10, 38, 41).
- [BS16] M. Bun and T. Steinke. “Concentrated differential privacy: Simplifications, extensions, and lower bounds”. In: *Theory of Cryptography: 14th International Conference, TCC 2016-B, Beijing, China, October 31–November 3, 2016, Proceedings, Part I*. Springer. 2016, pp. 635–658. URL: <https://arxiv.org/abs/1605.02065> (cit. on pp. 6, 43).
- [BS98] D. Boneh and J. Shaw. “Collusion-secure fingerprinting for digital data”. In: *IEEE Transactions on Information Theory* 44.5 (1998), pp. 1897–1905 (cit. on p. 5).
- [BST14] R. Bassily, A. Smith, and A. Thakurta. “Private empirical risk minimization: Efficient algorithms and tight error bounds”. In: *2014 IEEE 55th annual symposium on foundations of computer science*. IEEE. 2014, pp. 464–473. URL: <https://arxiv.org/abs/1405.7085> (cit. on p. 7).
- [BW18] B. Balle and Y.-X. Wang. “Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising”. In: *International Conference on Machine Learning*. PMLR. 2018, pp. 394–403. URL: <https://arxiv.org/abs/1805.06530> (cit. on p. 7).
- [CCNSTT22] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramer. “Membership inference attacks from first principles”. In: *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2022, pp. 1897–1914. URL: <https://arxiv.org/abs/2112.03570> (cit. on pp. 5, 30).
- [DBHSB22] S. De, L. Berrada, J. Hayes, S. L. Smith, and B. Balle. “Unlocking high-accuracy differentially private image classification through scale”. In: *arXiv preprint arXiv:2204.13650* (2022) (cit. on p. 24).
- [DFHPRR15a] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth. “Generalization in adaptive data analysis and holdout reuse”. In: *Advances in Neural Information Processing Systems* 28 (2015). URL: <https://arxiv.org/abs/1506.02629> (cit. on pp. 2, 10, 38, 41, 43).
- [DFHPRR15b] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. L. Roth. “Preserving statistical validity in adaptive data analysis”. In: *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*. 2015, pp. 117–126. URL: <https://arxiv.org/abs/1411.2664> (cit. on pp. 2, 10, 38, 41).

- [DKMMN06] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. “Our data, ourselves: Privacy via distributed noise generation”. In: *Advances in Cryptology-EUROCRYPT 2006: 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques, St. Petersburg, Russia, May 28-June 1, 2006. Proceedings 25*. Springer. 2006, pp. 486–503 (cit. on p. 6).
- [DMNS06] C. Dwork, F. McSherry, K. Nissim, and A. Smith. “Calibrating noise to sensitivity in private data analysis”. In: *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*. Springer. 2006, pp. 265–284. URL: <https://www.iacr.org/archive/tcc2006/38760266/38760266.pdf> (cit. on pp. 1, 6).
- [DR14] C. Dwork and A. Roth. “The algorithmic foundations of differential privacy”. In: *Foundations and Trends® in Theoretical Computer Science* 9.3–4 (2014), pp. 211–407. URL: <https://www.cis.upenn.edu/~aaroht/privacybook.html> (cit. on p. 6).
- [DR16] C. Dwork and G. N. Rothblum. “Concentrated differential privacy”. In: *arXiv preprint arXiv:1603.01887* (2016). URL: <https://arxiv.org/abs/1603.01887> (cit. on p. 6).
- [DSSUV15] C. Dwork, A. Smith, T. Steinke, J. Ullman, and S. Vadhan. “Robust traceability from trace amounts”. In: *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*. IEEE. 2015, pp. 650–669 (cit. on p. 5).
- [DWWZK18] Z. Ding, Y. Wang, G. Wang, D. Zhang, and D. Kifer. “Detecting violations of differential privacy”. In: *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. 2018, pp. 475–489 (cit. on pp. 5, 8).
- [FS17] V. Feldman and T. Steinke. “Generalization for adaptively-chosen estimators via stable median”. In: *Conference on learning theory*. PMLR. 2017, pp. 728–757. URL: <https://arxiv.org/abs/1706.05069> (cit. on pp. 38, 41).
- [FS18] V. Feldman and T. Steinke. “Calibrating noise to variance in adaptive data analysis”. In: *Conference On Learning Theory*. PMLR. 2018, pp. 535–544. URL: <https://arxiv.org/abs/1712.07196> (cit. on p. 43).
- [GLW21] S. Gopi, Y. T. Lee, and L. Wutschitz. “Numerical composition of differential privacy”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 11631–11642. URL: <https://arxiv.org/abs/2106.02848> (cit. on p. 7).

- [HSRDTMPSNC08] N. Homer, S. Szelingner, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig. “Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays”. In: *PLoS genetics* 4.8 (2008), e1000167 (cit. on p. 5).
- [JE19] B. Jayaraman and D. Evans. “Evaluating differentially private machine learning in practice”. In: *USENIX Security Symposium*. 2019 (cit. on p. 5).
- [JLNRSMS19] C. Jung, K. Ligett, S. Neel, A. Roth, S. Sharifi-Malvajerdi, and M. Shenfeld. “A new analysis of differential privacy’s generalization guarantees”. In: *arXiv preprint arXiv:1909.03577* (2019). URL: <https://arxiv.org/abs/1909.03577> (cit. on pp. 2, 10, 11, 38, 41, 42).
- [JUO20] M. Jagielski, J. Ullman, and A. Oprea. “Auditing differentially private machine learning: How private is private sgd?” In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 22205–22216 (cit. on pp. 1, 5, 8).
- [KJH20] A. Koskela, J. Jälkö, and A. Honkela. “Computing tight differential privacy guarantees using fft”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 2560–2569. URL: <https://arxiv.org/abs/1906.03049> (cit. on p. 7).
- [KOV15] P. Kairouz, S. Oh, and P. Viswanath. “The composition theorem for differential privacy”. In: *International conference on machine learning*. PMLR. 2015, pp. 1376–1385. URL: <https://arxiv.org/abs/1311.0776> (cit. on pp. 16, 44).
- [KS14] S. P. Kasiviswanathan and A. Smith. “On the ‘semantics’ of differential privacy: A bayesian formulation”. In: *Journal of Privacy and Confidentiality* 6.1 (2014). URL: <https://arxiv.org/abs/0803.3946> (cit. on p. 17).
- [LMFLZWRFT22] F. Lu, J. Munoz, M. Fuchs, T. LeBlond, E. Zaresky-Williams, E. Raff, F. Ferraro, and B. Testa. “A General Framework for Auditing Differentially Private Machine Learning”. In: *arXiv preprint arXiv:2210.08643* (2022) (cit. on p. 5).
- [MEMPST21] M. Malek Esmaeili, I. Mironov, K. Prasad, I. Shilov, and F. Tramer. “Antipodes of label differential privacy: Pate and alibi”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 6934–6945 (cit. on p. 5).
- [Mir17] I. Mironov. “Rényi differential privacy”. In: *2017 IEEE 30th computer security foundations symposium (CSF)*. IEEE. 2017, pp. 263–275. URL: <https://arxiv.org/abs/1702.07476> (cit. on p. 6).

- [MSS22] S. Maddock, A. Sablayrolles, and P. Stock. “CANIFE: Crafting Canaries for Empirical Privacy Measurement in Federated Learning”. In: *arXiv preprint arXiv:2210.02912* (2022) (cit. on pp. 5, 25).
- [MTZ19] I. Mironov, K. Talwar, and L. Zhang. “Renyi differential privacy of the sampled gaussian mechanism”. In: *arXiv preprint arXiv:1908.10530* (2019). URL: <https://arxiv.org/abs/1908.10530> (cit. on p. 7).
- [MV15] J. Murtagh and S. Vadhan. “The complexity of computing the optimal composition of differential privacy”. In: *Theory of Cryptography: 13th International Conference, TCC 2016-A, Tel Aviv, Israel, January 10-13, 2016, Proceedings, Part I*. Springer. 2015, pp. 157–175. URL: <https://arxiv.org/abs/1507.03113> (cit. on p. 16).
- [NHSBTJCT23] M. Nasr, J. Hayes, T. Steinke, B. Balle, F. Tramèr, M. Jagielski, N. Carlini, and A. Terzis. “Tight Auditing of Differentially Private Machine Learning”. In: *arXiv preprint arXiv:2302.07956* (2023). URL: <https://arxiv.org/abs/2302.07956> (cit. on pp. 1, 5, 24).
- [NSTPC21] M. Nasr, S. Song, A. Thakurta, N. Papernot, and N. Carlini. “Adversary instantiation: Lower bounds for differentially private machine learning”. In: *2021 IEEE Symposium on security and privacy (SP)*. IEEE. 2021, pp. 866–882. URL: <https://arxiv.org/abs/2101.04535> (cit. on p. 5).
- [RRST16] R. Rogers, A. Roth, A. Smith, and O. Thakkar. “Max-information, differential privacy, and post-selection hypothesis testing”. In: *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2016, pp. 487–494. URL: <https://arxiv.org/abs/1604.03924> (cit. on pp. 2, 10, 14).
- [SOJH09] S. Sankararaman, G. Obozinski, M. I. Jordan, and E. Halperin. “Genomic privacy and limits of individual detection in a pool”. In: *Nature genetics* 41.9 (2009), pp. 965–967 (cit. on p. 5).
- [SSSS17] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. “Membership inference attacks against machine learning models”. In: *2017 IEEE symposium on security and privacy (SP)*. IEEE. 2017, pp. 3–18 (cit. on p. 5).
- [Ste22] T. Steinke. “Composition of Differential Privacy & Privacy Amplification by Subsampling”. In: *arXiv preprint arXiv:2210.00597* (2022). URL: <https://arxiv.org/abs/2210.00597> (cit. on pp. 7, 16).
- [SZ20] T. Steinke and L. Zakyntinou. “Reasoning about generalization via conditional mutual information”. In: *Conference on Learning Theory*. PMLR. 2020, pp. 3437–3452. URL: <https://arxiv.org/abs/2001.09122> (cit. on pp. 2, 10, 11, 38, 43).

- [Tar08] G. Tardos. “Optimal probabilistic fingerprint codes”. In: *Journal of the ACM (JACM)* 55.2 (2008), pp. 1–24 (cit. on p. 5).
- [TTSSJC22] F. Tramer, A. Terzis, T. Steinke, S. Song, M. Jagielski, and N. Carlini. “Debugging differential privacy: A case study for privacy auditing”. In: *arXiv preprint arXiv:2202.12219* (2022). URL: <https://arxiv.org/abs/2202.12219> (cit. on p. 1).
- [Vad17] S. Vadhan. “The complexity of differential privacy”. In: *Tutorials on the Foundations of Cryptography: Dedicated to Oded Goldreich* (2017), pp. 347–450. URL: https://privacytools.seas.harvard.edu/files/privacytools/files/manuscript_2016.pdf (cit. on p. 6).
- [WBK19] Y.-X. Wang, B. Balle, and S. P. Kasiviswanathan. “Subsampled rényi differential privacy and analytical moments accountant”. In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR. 2019, pp. 1226–1235. URL: <https://arxiv.org/abs/1808.00087> (cit. on p. 7).
- [WBKBGGG22] Y. Wen, A. Bansal, H. Kazemi, E. Borgnia, M. Goldblum, J. Geiping, and T. Goldstein. “Canary in a Coalmine: Better Membership Inference with Ensembled Adversarial Queries”. In: *arXiv preprint arXiv:2210.10750* (2022) (cit. on p. 5).
- [ZBWTSRPNK22] S. Zanella-Béguelin, L. Wutschitz, S. Tople, A. Salem, V. Rühle, A. Paverd, M. Naseri, and B. Köpf. “Bayesian estimation of differential privacy”. In: *arXiv preprint arXiv:2206.05199* (2022) (cit. on pp. 5, 25).
- [ZDW22] Y. Zhu, J. Dong, and Y.-X. Wang. “Optimal accounting of differential privacy via characteristic function”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2022, pp. 4782–4817. URL: <https://arxiv.org/abs/2106.08567> (cit. on p. 7).
- [ZK16] S. Zagoruyko and N. Komodakis. “Wide residual networks”. In: *arXiv preprint arXiv:1605.07146* (2016) (cit. on p. 24).

A Sampling a Fixed-Size Dataset

Our auditing framework considers randomly including or excluding m examples independently. This means that the size of the dataset is random. This may be undesirable.

Fortunately, we can fix this, without changing our theoretical analysis. But it does require more examples and it requires changing the definition of DP to consider pairs of datasets differing by the replacement of one person’s data, rather than the addition or removal of one person’s data (cf. Remark 4.4).

Recall that our auditing framework starts with m examples x_1, \dots, x_m and then samples $S \in \{-1, +1\}^m$ uniformly at random. Then each example x_i is included in the dataset if $S_i = +1$ and excluded if $S_i = -1$. Thus flipping S_i corresponds to adding or removing x_i .

Instead we can start with $2m$ examples x_1, \dots, x_{2m} and then sample $S \in \{-1, +1\}^m$ uniformly. Now, if $S_i = +1$, we include x_{2i} in the dataset and, if $S_i = -1$, we include x_{2i-1} instead. Thus flipping S_i corresponds to replacing x_{2i} with x_{2i-1} or vice versa.

This alternative approach ensures that we always include m out of the $2m$ examples – i.e., the dataset size is not random. This still fits the formalism of our theoretical analysis (§5). However, the DP guarantee of the algorithm being audited (e.g., DP-SGD) must now be with respect to replacement of one example, rather than addition or removal.⁵ The auditor also needs to change slightly; rather than being given x_i and needing to guess whether or not it is included in the datasets, the auditor is given both x_{2i} and x_{2i-1} and must guess which of the two is included.

B Generalization from Differential Privacy

Our analysis builds on the connection between DP and generalization [DFHPRR15b; DFH-PRR15a; BNSSSU16; FS17; JLNRSMS19]. We now extend our theoretical results (§5) to this setting. The main difference between our analysis in Section 5 and the prior work on DP and generalization is that we restrict to i.i.d. binary inputs with a uniform distribution, while prior work considers i.i.d. inputs from an arbitrary set with an arbitrary distribution. Thus the prior work is more general, but, as we now show, we can reduce the general case to the binary case.

Theorem B.1 (DP implies Generalization). *Let $A : \mathcal{X}^n \rightarrow \mathcal{Y} \times [0, 1]$ be (ε, δ) -DP (with respect to replacement). Let P be a distribution on \mathcal{X} . Let $q : \mathcal{Y} \times \mathcal{X} \rightarrow [0, 1]$. For $x \in \mathcal{X}^n$ and $y \in \mathcal{Y}$, denote $q(y, x) = \frac{1}{n} \sum_i q(y, x_i) \in [0, 1]$ and $q(y, P) = \mathbb{E}_{X \leftarrow P} [q(y, X)] \in [0, 1]$. Then, for all $\gamma \geq \frac{3}{2}\eta \geq 0$, we have*

$$\mathbb{P}_{\substack{X \leftarrow P^n \\ Y \leftarrow A(X)}} [q(Y, X) - q(Y, P) \geq \gamma] \leq \mathbb{P} \left[\check{W} \geq \frac{(1+\gamma-\frac{3}{2}\eta)n}{2} \right] + 2 \cdot e^{-n\eta^2/2} + \max_{i \in [n]} \frac{2n\delta}{i} \mathbb{P} \left[\frac{(1+\gamma-\frac{3}{2}\eta)n}{2} > \check{W} \geq \frac{(1+\gamma-\frac{3}{2}\eta)n}{2} - i \right],$$

where $\check{W} \leftarrow \text{Binomial} \left(n, \frac{e^\varepsilon}{e^\varepsilon + 1} \right)$.

The proof of Theorem B.1 relies on the following technical lemma. This is using what is known as the “ghost samples” symmetrization technique [SZ20, Footnote 2].

Lemma B.2. *Let $x^+, x^- \in \mathcal{X}^n$. For $s \in \{-1, +1\}^n$, define $x^s \in \mathcal{X}^n$ by $x_i^s = x_i^+$ if $s_i = +1$ and $x_i^s = x_i^-$ if $s_i = -1$. Let $A : \mathcal{X}^n \rightarrow \mathcal{Y}$ be (ε, δ) -DP (for replacement). Let $q : \mathcal{Y} \times \mathcal{X} \rightarrow [0, 1]$, and denote $q(y, x) = \frac{1}{n} \sum_i q(y, x_i) \in [0, 1]$ for $y \in \mathcal{Y}$ and $x \in \mathcal{X}^n$.*

⁵By group privacy, (ε, δ) -DP for addition or removal implies $(2\varepsilon, (e^\varepsilon + 1) \cdot \delta)$ -DP for replacement.

Let $S \in \{-1, +1\}$ be uniform. Then, for all $v, r \geq 0$,

$$\begin{aligned} & \mathbb{P}_{\substack{S \leftarrow \{-1, +1\}^n \\ Y \leftarrow A(x^S)}} \left[q(Y, x^S) - q(Y, x^{-S}) \geq \frac{v}{n} \right] \\ & \leq \mathbb{P} \left[\check{W} \geq \frac{v - r + n}{2} \right] + \max_{i \in [n]} \frac{2n\delta}{i} \mathbb{P} \left[\frac{v - r + n}{2} > \check{W} \geq \frac{v - r + n}{2} - i \right] + e^{-r^2/2n}, \end{aligned}$$

where $\check{W} \leftarrow \text{Binomial} \left(n, \frac{e^\varepsilon}{e^\varepsilon + 1} \right)$.

Proof. Let $R : [-1, +1] \rightarrow \{-1, +1\}$ denote the randomized rounding function. I.e., $\mathbb{E}[R(x)] = x$ for all $x \in [-1, 1]$. We define $M : \{-1, +1\}^n \rightarrow \{-1, +1\}^n$ as follows. The inputs $x^+, x^- \in \mathcal{X}^n$ are ‘‘hardcoded’’ into M and, for this analysis, we do not consider them private. Instead the input is $s \in \{-1, +1\}^n$. The algorithm $M(s)$ first runs $A(x^s)$ and then postprocesses the output using the hardcoded information. Specifically, given $A(s) = y$, the output $M(s) \in \{-1, +1\}^n$ has a product distribution with $M(s)_i = R(q(y, x_i^+) - q(y, x_i^-)) \in \{-1, +1\}$ for all $y \in \mathcal{Y}$ and all $i \in [n]$. That is, for each coordinate $i \in [n]$, we independently randomly round $q(y, x_i^+) - q(y, x_i^-) \in [-1, +1]$ to $\{-1, +1\}$, where y is the output of $A(x^s)$. By postprocessing, M is (ε, δ) -DP. Thus we can apply Theorem 5.2 to M . We have, for all $r, v \geq 0$,

$$\begin{aligned} & \mathbb{P}_{\substack{S \leftarrow \{-1, +1\}^n \\ Y \leftarrow A(x^S)}} \left[q(Y, x^S) - q(Y, x^{-S}) \geq \frac{v}{n} \right] \\ & = \mathbb{P}_{\substack{S \leftarrow \{-1, +1\}^n \\ Y \leftarrow A(x^S)}} \left[\sum_i^n q(Y, x_i^S) - q(Y, x_i^{-S}) \geq v \right] \\ & = \mathbb{P}_{\substack{S \leftarrow \{-1, +1\}^n \\ Y \leftarrow A(x^S)}} \left[\sum_i^n (q(Y, x_i^+) - q(Y, x_i^-)) \cdot S_i \geq v \right] \\ & = \mathbb{P}_{\substack{S \leftarrow \{-1, +1\}^n \\ Y \leftarrow A(x^S)}} \left[\mathbb{E}_R \left[\sum_i^n R(q(Y, x_i^+) - q(Y, x_i^-)) \cdot S_i \right] \geq v \right] \\ & \leq \mathbb{P}_{\substack{S \leftarrow \{-1, +1\}^n \\ Y \leftarrow A(x^S), R}} \left[\sum_i^n R(q(Y, x_i^+) - q(Y, x_i^-)) \cdot S_i \geq v - r \right] + e^{-r^2/2n} \quad (\text{Hoeffding \& union}) \\ & = \mathbb{P}_{\substack{S \leftarrow \{-1, +1\}^n \\ T \leftarrow M(S)}} \left[\sum_i^n T_i \cdot S_i \geq v - r \right] + e^{-r^2/2n} \\ & = \mathbb{P}_{\substack{S \leftarrow \{-1, +1\}^n \\ T \leftarrow M(S)}} \left[\sum_i^n 2 \max\{0, T_i \cdot S_i\} - |T_i| \geq v - r \right] + e^{-r^2/2n} \quad (S_i \in \{-1, +1\}) \\ & = \mathbb{P}_{\substack{S \leftarrow \{-1, +1\}^n \\ T \leftarrow M(S)}} \left[\sum_i^n \max\{0, T_i \cdot S_i\} \geq \frac{v - r + n}{2} \right] + e^{-r^2/2n} \quad (|T_i| = 1) \\ & \leq \mathbb{P} \left[\check{W} \geq \frac{v - r + n}{2} \right] + \max_{i \in [n]} \frac{2n\delta}{i} \mathbb{P} \left[\frac{v - r + n}{2} > \check{W} \geq \frac{v - r + n}{2} - i \right] + e^{-r^2/2n}, \end{aligned}$$

where $\check{W} \leftarrow \text{Binomial}(n, \frac{e^\varepsilon}{e^\varepsilon+1})$. Note that Theorem 5.2 applies with any distribution \check{W}^* satisfying

$$\forall v \in \mathbb{R} \quad \mathbb{P}[\check{W}^* > v] \geq \sup_{t \in \text{support}(M(s))} \mathbb{P}^{\check{S} \leftarrow \text{Bernoulli}(\frac{e^\varepsilon}{e^\varepsilon+1})^n} \left[\sum_i^n \check{S}_i \cdot |t_i| > v \right].$$

If $t \in \text{support}(M(s))$, then $|t_i| = 1$ for all $i \in [n]$, which implies \check{W} satisfies this requirement. In the analysis above, we used Hoeffding's inequality to show that the sum of randomized roundings is close to (within r of) the unrounded sum with high probability and we carry this failure probability $e^{-r^2/2n}$ into the final result. \square

Proposition B.3. *Let $A : \mathcal{X}^n \rightarrow \mathcal{Y}$ be (ε, δ) -DP (with respect to replacement). Let $q : \mathcal{Y} \times \mathcal{X} \rightarrow [0, 1]$, and denote $q(y, x) = \frac{1}{n} \sum_i^n q(y, x_i) \in [0, 1]$ for $y \in \mathcal{Y}$ and $x \in \mathcal{X}^n$. Let P be a distribution on \mathcal{X} . Then, for all $\gamma, \eta \geq 0$, we have*

$$\begin{aligned} & \mathbb{P}_{\substack{X, \tilde{X} \leftarrow P^n \\ Y \leftarrow A(X)}} \left[q(Y, X) - q(Y, \tilde{X}) \geq \gamma \right] \\ & \leq \mathbb{P} \left[\check{W} \geq \frac{(1 + \gamma - \eta)n}{2} \right] + \max_{i \in [n]} \frac{2n\delta}{i} \mathbb{P} \left[\frac{(1 + \gamma - \eta)n}{2} > \check{W} \geq \frac{(1 + \gamma - \eta)n}{2} - i \right] + e^{-n\eta^2/2}, \end{aligned}$$

where $\check{W} \leftarrow \text{Binomial}(n, \frac{e^\varepsilon}{e^\varepsilon+1})$.

Proof. The proof relies on Lemma B.2, which considers $x^+, x^- \in \mathcal{X}^n$ to be fixed. We now average the lemma over these being i.i.d. samples from P , which gives

$$\begin{aligned} & \mathbb{E}_{X^+, X^- \leftarrow P^n} \left[\mathbb{P}_{\substack{S \leftarrow \{-1, +1\}^n \\ Y \leftarrow A(X^S)}} \left[q(Y, X^S) - q(Y, X^{-S}) \geq \frac{v}{n} \right] \right] \\ & \leq \mathbb{P} \left[\check{W} \geq \frac{v - r + n}{2} \right] + \max_{i \in [n]} \frac{2n\delta}{i} \mathbb{P} \left[\frac{v - r + n}{2} > \check{W} \geq \frac{v - r + n}{2} - i \right] + e^{-r^2/2n}, \end{aligned}$$

where $\check{W} \leftarrow \text{Binomial}(n, \frac{e^\varepsilon}{e^\varepsilon+1})$. Since the samples from P are independent, the coordinates of X^+ and X^- are interchangeable, so

$$\mathbb{P}_{\substack{X, \tilde{X} \leftarrow P^n \\ Y \leftarrow A(X)}} \left[q(Y, X) - q(Y, \tilde{X}) \geq \gamma \right] = \mathbb{E}_{X^+, X^- \leftarrow P^n} \left[\mathbb{P}_{\substack{S \leftarrow \{-1, +1\}^n \\ Y \leftarrow A(X^S)}} \left[q(Y, X^S) - q(Y, X^{-S}) \geq \frac{v}{n} \right] \right]$$

for $v = \gamma n \geq 0$. Setting $r = n\eta$ yields the result. \square

Proof of Theorem B.1. Let $X, \tilde{X} \leftarrow P^n$ be two independent samples. Let $Y \leftarrow A(X)$. Let $\check{W} \leftarrow \text{Binomial}(n, \frac{e^\varepsilon}{e^\varepsilon+1})$. By Proposition B.3, for all $\gamma, \eta \geq 0$, we have

$$\begin{aligned} & \mathbb{P}_{\substack{X, \tilde{X} \leftarrow P^n \\ Y \leftarrow A(X)}} \left[q(Y, X) - q(Y, \tilde{X}) \geq \gamma \right] \\ & \leq \mathbb{P} \left[\check{W} \geq \frac{(1 + \gamma - \eta)n}{2} \right] + \max_{i \in [n]} \frac{2n\delta}{i} \mathbb{P} \left[\frac{(1 + \gamma - \eta)n}{2} > \check{W} \geq \frac{(1 + \gamma - \eta)n}{2} - i \right] + e^{-n\eta^2/2}, \end{aligned}$$

By Hoeffding's inequality, for all $\eta \geq 0$, we have

$$\forall y \in \mathcal{Y} \quad \mathbb{P}_{\tilde{X} \leftarrow P^n} \left[q(y, \tilde{X}) - q(y, P) \geq \frac{\eta}{2} \right] \leq \exp(-n\eta^2/2).$$

By a union bound, for all $\gamma \geq \eta \geq 0$, we have

$$\mathbb{P}_{\substack{X \leftarrow P^n \\ Y \leftarrow A(X)}} [q(Y, X) - q(Y, P) \geq \gamma] \leq \mathbb{P}_{\substack{X, \tilde{X} \leftarrow P^n \\ Y \leftarrow A(X)}} [q(Y, \tilde{X}) - q(Y, P) \geq \gamma - \eta/2] + \mathbb{P}_{\substack{X, \tilde{X} \leftarrow P^n \\ Y \leftarrow A(X)}} [q(Y, X) - q(Y, \tilde{X}) \geq \eta/2].$$

Combining inequalities yields the result:

$$\begin{aligned} & \mathbb{P}_{\substack{X \leftarrow P^n \\ Y \leftarrow A(X)}} [q(Y, X) - q(Y, P) \geq \gamma] \\ & \leq \mathbb{P} \left[\check{W} \geq \frac{(1 + \gamma - \eta/2 - \eta)n}{2} \right] + 2 \cdot e^{-n\eta^2/2} \\ & \quad + \max_{i \in [n]} \frac{2n\delta}{i} \mathbb{P} \left[\frac{(1 + \gamma - \eta/2 - \eta)n}{2} > \check{W} \geq \frac{(1 + \gamma - \eta/2 - \eta)n}{2} - i \right]. \end{aligned}$$

□

B.1 Comparison to Prior Work on DP & Generalization

We now briefly compare our results to the prior work on the connection between DP and generalization [DFHPRR15b; DFHPRR15a; BNSSSU16; FS17; JLNRSMS19]. We focus on the work of Jung, Ligett, Neel, Roth, Sharifi-Malvajerdi, and Shenfeld [JLNRSMS19] as it has the sharpest results in the literature.

Note that the prior work is focused on the setting of adaptive data analysis, while we are focused on the setting of auditing. This difference is mostly cosmetic, but there is a material difference when the prior results are applied to our setting: In addition to outputting guesses, the prior works assume that the algorithm outputs a differentially private estimate of the number of correct guesses. The guarantee then is that this differentially private estimate is close to the distributional average (i.e., only half of the guesses being correct). In contrast, for auditing, we want the true number of correct guesses to be close to the distributional average and don't produce a DP estimate. We can convert between these two settings using the triangle inequality.

Below we state the accuracy guarantee that we compare against, followed by a corollary of Theorem B.1 that applies the triangle inequality and a union bound to ensure that it is directly comparable.

Theorem B.4 ([JLNRSMS19, Theorem 3.5]). *Let $A : \mathcal{X}^n \rightarrow \mathcal{Y} \times [0, 1]$ be (ε, δ) -DP (with respect to replacement). Let P be a distribution on \mathcal{X} . Let $q : \mathcal{Y} \times \mathcal{X} \rightarrow [0, 1]$. For $x \in \mathcal{X}^n$*

and $y \in \mathcal{Y}$, denote $q(y, x) = \frac{1}{n} \sum_i^n q(y, x_i) \in [0, 1]$ and $q(y, P) = \mathbb{E}_{X \leftarrow P} [q(y, X)] \in [0, 1]$. Suppose

$$\mathbb{P}_{\substack{X \leftarrow P^n \\ (Y, Z) \leftarrow A(X)}} [|Z - q(Y, X)| \geq \alpha] \leq \beta.$$

Then, for any $c, d > 0$, we have

$$\mathbb{P}_{\substack{X \leftarrow P^n \\ (Y, Z) \leftarrow A(X)}} [|Z - q(Y, P)| > \alpha + e^\varepsilon - 1 + c + 2d] \leq \frac{\beta}{c} + \frac{\delta}{d}. \quad (13)$$

Corollary B.5 (Theorem B.1, triangle inequality, & union bound). Let $A : \mathcal{X}^n \rightarrow \mathcal{Y} \times [0, 1]$ be (ε, δ) -DP (with respect to replacement). Let P be a distribution on \mathcal{X} . Let $q : \mathcal{Y} \times \mathcal{X} \rightarrow [0, 1]$. For $x \in \mathcal{X}^n$ and $y \in \mathcal{Y}$, denote $q(y, x) = \frac{1}{n} \sum_i^n q(y, x_i) \in [0, 1]$ and $q(y, P) = \mathbb{E}_{X \leftarrow P} [q(y, X)] \in [0, 1]$. Suppose

$$\mathbb{P}_{\substack{X \leftarrow P^n \\ (Y, Z) \leftarrow M(X)}} [|Z - q(Y, X)| \geq \alpha] \leq \beta.$$

Then, for all $\gamma \geq \frac{3}{2}\eta \geq 0$, we have

$$\mathbb{P}_{\substack{X \leftarrow P^n \\ (Y, Z) \leftarrow A(X)}} [|Z - q(Y, P)| \geq \alpha + \gamma] \leq \beta + \mathbb{P} \left[\check{W} \geq \frac{(1+\gamma-\frac{3}{2}\eta)n}{2} \right] + 2 \cdot e^{-n\eta^2/2} + \max_{i \in [n]} \frac{2n\delta}{i} \mathbb{P} \left[\frac{(1+\gamma-\frac{3}{2}\eta)n}{2} > \check{W} \geq \frac{(1+\gamma-\frac{3}{2}\eta)n}{2} - i \right], \quad (14)$$

where $\check{W} \leftarrow \text{Binomial} \left(n, \frac{e^\varepsilon}{e^\varepsilon + 1} \right)$.

Equations 13 and 14 are directly comparable, but it is not immediately obvious how they compare. By setting $\delta = 0$, $\gamma = \frac{e^\varepsilon - 1}{e^\varepsilon + 1} + c$, $\eta = \frac{2}{5}c$, and applying Hoeffding's inequality to \check{W} , we can simplify Equation 14 to

$$\mathbb{P}_{\substack{X \leftarrow P^n \\ (Y, Z) \leftarrow A(X)}} \left[|Z - q(Y, P)| \geq \alpha + \frac{e^\varepsilon - 1}{e^\varepsilon + 1} + c \right] \leq \beta + 3 \cdot e^{-n \frac{2}{25} c^2} \quad (15)$$

For comparison, setting $\delta = 0$ in Equation 13 gives

$$\mathbb{P}_{\substack{X \leftarrow P^n \\ (Y, Z) \leftarrow A(X)}} [|Z - q(Y, P)| > \alpha + e^\varepsilon - 1 + c] \leq \frac{\beta}{c}. \quad (16)$$

Now we can compare the results more easily. The $e^\varepsilon - 1$ term in the accuracy bound of Jung, Ligett, Neel, Roth, Sharifi-Malvajerdi, and Shenfeld [JLNRSMS19] is improved to $\frac{e^\varepsilon - 1}{e^\varepsilon + 1}$ in our result, which is an improvement by a factor of at least two. This is (arguably) the dominant term, so our result is a significant improvement. In particular, if $\varepsilon \geq \log 2$, then Equation 13 gives a vacuous bound (since the value of q is always in $[0, 1]$ anyway), while our bound can be non-vacuous for any value of ε (as $\frac{e^\varepsilon - 1}{e^\varepsilon + 1} < 1$).

However, there is another term in the accuracy bound – i.e., c . The failure probability either has a $1/c$ *multiplicative* factor or a $3 \cdot e^{-n \frac{2}{25} c^2}$ *additive* factor. How these compare depends on the value of β . To give a concrete comparison, suppose $\varepsilon = 1/3$, $n = 2000$, $\beta = \delta = 10^{-5}$, and we want a final failure probability of 0.05; then Theorem B.4 gives an error guarantee of $\alpha + 0.397$, while Corollary B.5 gives $\alpha + 0.308$.

C Mutual Information Bounds from DP

Our framework for the theoretical analysis (§5) is inspired by that of Steinke and Zakyntinou [SZ20]. In this appendix, we use our analysis to also improve one of their results. Specifically, they show that if $M : \{-1, 1\}^m \rightarrow \mathcal{Y}$ satisfies (ε, δ) -DP and $S \in \{-1, 1\}^m$ is uniformly random, then

$$I(S; M(S)) \leq (e^\varepsilon - 1 + \delta) \cdot m \cdot \log e, \quad (17)$$

where $I(\cdot; \cdot)$ denotes the mutual information.⁶ Prior work [DFHPRR15a; BS16] showed that, if $M : \mathcal{X}^m \rightarrow \mathcal{Y}$ satisfies $(\varepsilon, 0)$ -DP and $S \in \mathcal{X}^m$ has as product distribution, then

$$I(S; M(S)) \leq \frac{1}{2}\varepsilon^2 \cdot m \cdot \log e. \quad (18)$$

The latter result is numerically better than the former result, but only holds for pure DP. (The latter result is also not restricted to binary inputs. However, if we do not restrict the input at all, then it is not possible to prove bounds under approximate DP.)

We improve the bound to the following. If $M : \{-1, 1\}^m \rightarrow \mathcal{Y}$ satisfies (ε, δ) -DP and $S \in \{-1, 1\}^m$ is uniformly random, then

$$I(S; M(S)) \leq \frac{1}{8}\varepsilon^2 \cdot m \cdot \log e + \delta \cdot m \cdot \log 2. \quad (19)$$

Proposition C.1. *Let $M : \{0, 1\}^n \rightarrow \mathcal{Y}$ satisfy (ε, δ) -DP. Let $S \in \{0, 1\}^n$ be sampled from Bernoulli(p)ⁿ. Then*

$$I(S; M(S)) \leq n\delta h(p) + n(1 - \delta)h\left(\frac{p \cdot e^\varepsilon + 1 - p}{e^\varepsilon + 1}\right) - n(1 - \delta) \cdot \left(\log(1 + e^{-\varepsilon}) + \frac{\log(e^\varepsilon)}{e^\varepsilon + 1}\right),$$

where $h(p) := p \log(1/p) + (1 - p) \log(1/(1 - p))$ is the binary Entropy function.

In particular, if $p = \frac{1}{2}$, then

$$\begin{aligned} I(S; M(S)) &\leq n\delta \log 2 + n(1 - \delta) \cdot \left(\log 2 - \log(1 + e^{-\varepsilon}) - \frac{\log(e^\varepsilon)}{e^\varepsilon + 1}\right) \\ &\leq n\delta \log 2 + n(1 - \delta) \frac{\varepsilon^2}{8} \log e. \end{aligned}$$

Proof. We apply the chain rule and convexity of KL divergence [FS18, Lemma 3.7]:

$$I(S; M(S)) = \sum_i^n I(S_i; M(S)|S_{<i}) \leq \sum_i^n I(S_i; M(S)|S_{-i}).$$

⁶Throughout this paper we use natural logarithms (so $\log e = 1$), including when defining information-theoretic quantities like mutual information. However, it is common to use base-2 logarithms in information theory (i.e., $\log_2 e \approx 1.44$). To avoid confusion, the statements (outside proofs) in this section are stated in a redundant way so that they would be correct regardless of the base of the logarithm, as long as we are consistent.

Fix $i \in [n]$ and fix $s_{-i} \in \{0, 1\}^{n-1}$. Now we must analyze

$$\begin{aligned} I(S_i; M(S)|S_{-i} = s_{-i}) &= I(S_i; M(S_i, s_{-i})) \\ &= pD_1(M(1, s_{-i}) \| pM(1, s_{-i}) + (1-p)M(0, s_{-i})) \\ &\quad + (1-p)D_1(M(1, s_{-i}) \| pM(1, s_{-i}) + (1-p)M(0, s_{-i})) \\ &= pD_1(Q_1 \| Q_p) + (1-p)D_1(Q_0 \| Q_p), \end{aligned}$$

where $Q_t := tM(1, s_{-i}) + (1-t)M(0, s_{-i})$ for $t \in [0, 1]$.

Since M is (ε, δ) -DP, we have $Q_t(S) \leq e^\varepsilon \cdot Q_{1-t}(S) + \delta$ for all measurable $S \subset \mathcal{Y}$ and $t \in \{0, 1\}$. Thus we can apply Lemma 5.5: There exist distributions Q'_0, Q''_0, Q'_1, Q''_1 on \mathcal{Y} such that $Q_0 = (1-\delta) \cdot Q'_0 + \delta \cdot Q''_0$ and $Q_1 = (1-\delta) \cdot Q'_1 + \delta \cdot Q''_1$ and $e^{-\varepsilon} \cdot Q'_0(S) \leq Q'_1(S) \leq e^\varepsilon \cdot Q'_0(S)$ for all measurable $S \subset \mathcal{Y}$.

Define distributions

$$R_0 := \frac{e^\varepsilon \cdot Q'_0 - Q'_1}{e^\varepsilon - 1} \quad \text{and} \quad R_1 := \frac{e^\varepsilon \cdot Q'_1 - Q'_0}{e^\varepsilon - 1},$$

so that $Q'_0 = \frac{e^\varepsilon R_0 + R_1}{e^\varepsilon + 1}$ and $Q'_1 = \frac{e^\varepsilon R_1 + R_0}{e^\varepsilon + 1}$. Hence

$$Q_0 = \frac{e^\varepsilon(1-\delta)}{e^\varepsilon + 1} \cdot R_0 + \frac{1-\delta}{e^\varepsilon + 1} \cdot R_1 + \delta \cdot Q''_0$$

and

$$Q_1 = \frac{e^\varepsilon(1-\delta)}{e^\varepsilon + 1} \cdot R_1 + \frac{1-\delta}{e^\varepsilon + 1} \cdot R_0 + \delta \cdot Q''_1.$$

This decomposition (which was first used by Kairouz, Oh, and Viswanath [KOV15]) states that we can view $Q_{s_i} = M(s_i, s_{-i})$ as a postprocessing of an (ε, δ) -DP randomized response on the bit s_i . That is, with probability δ , we output the bit s_i with a flag indicating certainty; with probability $\frac{e^\varepsilon(1-\delta)}{e^\varepsilon + 1}$, we output s_i with an uncertain flag; and, with probability $\frac{1-\delta}{e^\varepsilon + 1}$, we output $1 - s_i$ with the uncertain flag. We can postprocess this to generate a sample from $Q_{s_i} = M(s_i, s_{-i})$ as follows. If we receive $b \in \{0, 1\}$ with the uncertain flag, then output a sample from R_b . If we receive $b \in \{0, 1\}$ with the certain flag, then output a sample from Q''_b .

To be formal, define two distributions on the set $[4] = \{1, 2, 3, 4\}$ by

$$\begin{aligned} \tilde{Q}_0 &= \left(\frac{e^\varepsilon(1-\delta)}{e^\varepsilon + 1}, \frac{1-\delta}{e^\varepsilon + 1}, \delta, 0 \right), \\ \tilde{Q}_1 &= \left(\frac{1-\delta}{e^\varepsilon + 1}, \frac{e^\varepsilon(1-\delta)}{e^\varepsilon + 1}, 0, \delta \right). \end{aligned}$$

Define the a randomized postprocessing function $F : [4] \rightarrow \mathcal{Y}$ by $F(1) = R_0$, $F(2) = R_1$, $F(3) = Q''_0$, and $F(4) = Q''_1$. Then we have $F(\tilde{Q}_0) = Q_0$ and $F(\tilde{Q}_1) = Q_1$.

Now we use the postprocessing property (a.k.a. the data processing inequality):

$$\begin{aligned} I(S_i; M(S)|S_{-i} = s_{-i}) &= pD_1(Q_1 \| Q_p) + (1-p)D_1(Q_0 \| Q_p) \\ &\leq pD_1(\tilde{Q}_1 \| \tilde{Q}_p) + (1-p)D_1(\tilde{Q}_0 \| \tilde{Q}_p). \end{aligned}$$

A tedious calculation now yields the bound:

$$\begin{aligned}
& pD_1\left(\tilde{Q}_1\|\tilde{Q}_p\right) + (1-p)D_1\left(\tilde{Q}_0\|\tilde{Q}_p\right) \\
&= p\left(\frac{1-\delta}{e^\varepsilon+1}\log\left(\frac{\frac{1-\delta}{e^\varepsilon+1}}{p\frac{1-\delta}{e^\varepsilon+1}+(1-p)\frac{e^\varepsilon(1-\delta)}{e^\varepsilon+1}}\right) + \frac{e^\varepsilon(1-\delta)}{e^\varepsilon+1}\log\left(\frac{\frac{e^\varepsilon(1-\delta)}{e^\varepsilon+1}}{p\frac{e^\varepsilon(1-\delta)}{e^\varepsilon+1}+(1-p)\frac{1-\delta}{e^\varepsilon+1}}\right) + \delta\log\left(\frac{\delta}{p\delta}\right)\right) \\
&+ (1-p)\left(\frac{e^\varepsilon(1-\delta)}{e^\varepsilon+1}\log\left(\frac{\frac{e^\varepsilon(1-\delta)}{e^\varepsilon+1}}{p\frac{1-\delta}{e^\varepsilon+1}+(1-p)\frac{e^\varepsilon(1-\delta)}{e^\varepsilon+1}}\right) + \frac{1-\delta}{e^\varepsilon+1}\log\left(\frac{\frac{1-\delta}{e^\varepsilon+1}}{p\frac{e^\varepsilon(1-\delta)}{e^\varepsilon+1}+(1-p)\frac{1-\delta}{e^\varepsilon+1}}\right) + \delta\log\left(\frac{\delta}{(1-p)\delta}\right)\right) \\
&= p\frac{1-\delta}{e^\varepsilon+1}\left(\log\left(\frac{1}{p+(1-p)e^\varepsilon}\right) + e^\varepsilon\log\left(\frac{e^\varepsilon}{pe^\varepsilon+(1-p)}\right)\right) \\
&+ (1-p)\frac{1-\delta}{e^\varepsilon+1}\left(e^\varepsilon\log\left(\frac{e^\varepsilon}{p+(1-p)e^\varepsilon}\right) + \log\left(\frac{1}{pe^\varepsilon+(1-p)}\right)\right) \\
&+ \delta(p\log(1/p) + (1-p)\log(1/(1-p))) \\
&= \frac{1-\delta}{e^\varepsilon+1}\left((p+(1-p)e^\varepsilon)\log\left(\frac{1}{p+(1-p)e^\varepsilon}\right) + (1-p)e^\varepsilon\varepsilon + (pe^\varepsilon+1-p)\log\left(\frac{1}{pe^\varepsilon+1-p}\right) + pe^\varepsilon\varepsilon\right) \\
&+ \delta h(p) \\
&= (1-\delta)\left(\frac{p+(1-p)e^\varepsilon}{e^\varepsilon+1}\log\left(\frac{e^\varepsilon+1}{p+(1-p)e^\varepsilon}\right) + \frac{pe^\varepsilon+1-p}{e^\varepsilon+1}\log\left(\frac{e^\varepsilon+1}{pe^\varepsilon+1-p}\right) - \log(e^\varepsilon+1) + \frac{e^\varepsilon\varepsilon}{e^\varepsilon+1}\right) \\
&+ \delta h(p) \\
&= (1-\delta)\left(h\left(\frac{p+(1-p)e^\varepsilon}{e^\varepsilon+1}\right) - \log(e^\varepsilon+1) + \frac{e^\varepsilon\varepsilon}{e^\varepsilon+1}\right) + \delta h(p) \\
&= (1-\delta)\left(h\left(\frac{pe^\varepsilon+(1-p)}{e^\varepsilon+1}\right) - \log(1+e^{-\varepsilon}) - \frac{\varepsilon}{e^\varepsilon+1}\right) + \delta h(p).
\end{aligned}$$

Combining inequalities and summing over $i \in [n]$ yields the first part of the result. The final part of the result is the bound

$$\forall \varepsilon \geq 0 \quad g(\varepsilon) := \log 2 - \log(1 + e^{-\varepsilon}) - \frac{\varepsilon}{e^\varepsilon + 1} \leq \frac{\varepsilon^2}{8},$$

which can be verified by showing that $g(0) = g'(0) = 0$ and $\forall \varepsilon \geq 0 \quad g''(\varepsilon) \leq \frac{1}{4}$ (or by plotting it). \square

D Implementation of Theorem 5.2

On the next page is Python pseudocode implementing Corollary 5.4. Some example usage:

- Suppose the auditor correctly guesses $v = 75$ out of $m = r = 100$ examples, with no abstentions. We have $\frac{75}{100} = \frac{3}{4} = \frac{e^{\log 3}}{e^{\log 3} + 1}$. So we would expect this to correspond roughly to $\varepsilon = \log 3 \approx 1.09$. Theorem 5.2 gives p-value of 0.553 for the null hypothesis $\varepsilon \leq \log 3$ and $\delta = 0$; to obtain this result call `p_value_DP_audit(100, 100, 75, math.log(3), 0)` in the code below. If we want 95% confidence, we obtain the lower bound $\varepsilon \geq 0.702$

by calling `get_eps_audit(100,100,75,0,0.05)`. If we set $\delta = 10^{-4}$, we obtain the weaker lower bound $\varepsilon \geq 0.699$ by calling `get_eps_audit(100,100,75,1e-4,0.05)`.

- Suppose the auditor correctly guesses $v = 75$ out of $r = 100$ guesses, but with a total of $m = 1000$ examples. I.e., the auditor abstains on $m - r = 900$ examples. We obtain a lower bound of $\varepsilon \geq 0.673$ for $\delta = 10^{-4}$ and 95% confidence. (This is slightly weaker than the $\varepsilon \geq 0.699$ lower bound we get when there are no abstentions.) This is obtained by calling `get_eps_audit(1000,100,75,1e-4,0.05)`.

```
# m = number of examples, each included independently with probability 0.5
# r = number of guesses (i.e. excluding abstentions)
# v = number of correct guesses by auditor
# eps,delta = DP guarantee of null hypothesis
# output: p-value = probability of >=v correct guesses under null hypothesis
def p_value_DP_audit(m, r, v, eps, delta):
    assert 0 <= v <= r <= m
    assert eps >= 0
    assert 0 <= delta <= 1
    q = 1/(1+math.exp(-eps)) # accuracy of eps-DP randomized response
    beta = scipy.stats.binom.sf(v-1, r, q) # = P[Binomial(r, q) >= v]
    alpha = 0
    sum = 0 # = P[v > Binomial(r, q) >= v - i]
    for i in range(1, v + 1):
        sum = sum + scipy.stats.binom.pmf(v - i, r, q)
        if sum > i * alpha:
            alpha = sum / i
    p = beta + alpha * delta * 2 * m
    return min(p, 1)

# m = number of examples, each included independently with probability 0.5
# r = number of guesses (i.e. excluding abstentions)
# v = number of correct guesses by auditor
# p = 1-confidence e.g. p=0.05 corresponds to 95%
# output: lower bound on eps i.e. algorithm is not (eps,delta)-DP
def get_eps_audit(m, r, v, delta, p):
    assert 0 <= v <= r <= m
    assert 0 <= delta <= 1
    assert 0 < p < 1
    eps_min = 0 # maintain p_value_DP(eps_min) < p
    eps_max = 1 # maintain p_value_DP(eps_max) >= p
    while p_value_DP_audit(m, r, v, eps_max, delta) < p: eps_max = eps_max + 1
    for _ in range(30): # binary search
        eps = (eps_min + eps_max) / 2
        if p_value_DP_audit(m, r, v, eps, delta) < p:
            eps_min = eps
        else:
            eps_max = eps
    return eps_min
```