Privacy Auditing with One (1) Training Run

Thomas Steinke [1] **Milad Nasr** [1] **Matthew Jagielski** [1]

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 (i.e., central DP) or white-box (i.e., federated learning) setting. We demonstrate the effectiveness of our framework by applying it to DP-SGD, where we can achieve meaningful empirical privacy lower bounds by training only *one model*, where standard methods would require training hundreds of models.

1. Introduction

Differential privacy (DP) (Dwork et al., 2006) 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 event S, we have

$$\mathbb{P}\left[M(x) \in S\right] \leq e^{\varepsilon} \cdot \mathbb{P}\left[M(x') \in S\right] + \delta. \qquad (1)$$

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 (Jagielski et al., 2020; Nasr et al., 2023) or, if the lower and upper bounds are contradictory, to detect errors in the analysis or in the algorithm's implementation (Ding et al., 2018; Bichsel et al., 2018; Tramer et al., 2022).

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 S and we estimate the probabilities $\mathbb{P}\left[M(x) \in S\right]$ and $\mathbb{P}\left[M(x') \in S\right]$. However, estimating these probabilities requires running M hundreds of times. This approach to privacy auditing is computationally expensive. *Can we perform privacy auditing using a single run of the algorithm M?*

1.1. Our Contributions

Our approach: 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. This approach is commonly used as an *unproven* heuristic in prior work (Malek Esmaeili et al., 2021; Zanella-Béguelin et al., 2022).

Our auditing procedure operates as follows. We identify m data points (i.e., training examples or "canaries") to 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 $(\varepsilon, 0)$-DP, then the auditor can correctly guess each inclusion/exclusion coin flip with probability $\leq \frac{e^{\varepsilon}}{e^{\varepsilon}+1}$. Thus DP implies a high-probability upper bound on the fraction of correct guesses and, conversely, the fraction of correct guesses implies a high-probability lower bound on the privacy parameters.

Our analysis: 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 (Dwork et al., 2015b;a; Bassily et al., 2016; Rogers et al., 2016; Jung et al., 2019; Steinke & Zakynthinou, 2020). 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

[*] Equal contribution [1] Google. Correspondence to: Thomas Steinke <steinke@google.com>, Milad Nasr <srxzr@google.com>, Matthew Jagielski <jagielski@google.com>.

Federated Learning and Analytics in Practice: Algorithms, Systems, Applications, and Opportunities Workshop at ICML 2023. Honolulu, Hawaii, USA.

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, we show that including or excluding m data points independently for one run is almost as good as having m independent runs (for small δ).

Our results: As an application of our auditing framework, we 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 $\varepsilon \geq 1.8$, compared to a theoretical upper bound of $\varepsilon \leq 4$ in the white-box setting – i.e., we assume the adversary has access to intermediate updates, as is the case in federated learning. 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.

2. Our Auditing Procedure

Algorithm 1 Auditor with One Training Run

```
Data: $x \in \mathcal{X}^n$ consisting of $m$ auditing examples (a.k.a. canaries) $x_1, \cdots, x_m$ and $n - m$ non-auditing examples $x_{m+1}, \cdots, x_n$.
Parameters: Algorithm to audit $\mathcal{A}$, number of examples to randomize $m$, number of positive $k_+$ and negative $k_-$ guesses.
For $i \in [m]$ sample $S_i \in \{-1, +1\}$ uniformly and independently. Set $S_i = 1$ for all $i \in [n] \setminus [m]$.
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 $x_{\text{IN}}$; and, if $S_i = -1$, then $x_i$ is in $x_{\text{OUT}}$.
Run $\mathcal{A}$ on input $x_{\text{IN}}$ with appropriate parameters, outputting $w$.
Compute the vector of scores $Y = (\text{SCORE}(x_i, w) : i \in [m]) \in \mathbb{R}^m$.
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. (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 $\sum_i^m T_i = k_+ - k_-$.)
Return: $S \in \{-1, +1\}^m$ indicating the true selection and the guesses $T \in \{-1, 0, +1\}^m$.
```

We present our auditing procedure in Algorithm 1. We independently include each of the first m examples with 50% probability and exclude it otherwise. When applied to DP-SGD, our approach is applicable to both “white-box” auditing, where the adversary can access to all intermediate values of the model weights, and “black-box” auditing, where the adversary only sees the final model weights (or can only query the final model). In both cases we simply compute a “score” for each example and “guess” whether the example is included or excluded based on these scores.

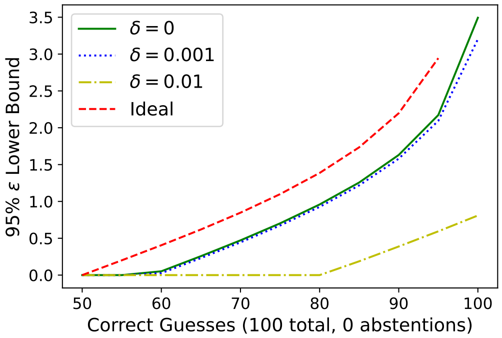

Figure 1. Lower bound on the privacy parameter ε of Theorem 3.1 with 95% confidence as the number of correct guesses changes. The total number of examples and guesses is 100. For comparison, we plot the `ideal` ε that gives $100 \cdot \frac{e^\varepsilon}{e^\varepsilon+1}$ correct guesses.

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.

Note that we only randomize the first m examples $x_1, \cdots, x_m$ (which we refer to as “auditing examples” or “canaries”); the last $n - m$ examples $x_{m+1}, \cdots, x_n$ 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. We may wish to artificially construct m examples to be easy to identify (i.e., canaries), but also include $n - m$ “real” examples to ensure that $\mathcal{A}$ still produces a useful model.

The score function is arbitrary and will depend on the application. For black-box auditing, we use the loss of the final model w on the example x_i – i.e., $\text{SCORE}(x_i, w) = -\text{loss}(w, x_i)$. For white-box auditing, $\text{SCORE}(x_i, w) = \sum_t^\ell \langle w^{t-1} - w^t, \nabla_{w^{t-1}} \text{loss}(w^{t-1}, x_i) \rangle$ is the sum of the inner products of updates with the (clipped) gradients of the loss on the example. In federated learning settings, a realistic adversary should be assumed to have access to many (if not all) of the intermediate model weights w^t. Intuitively, the vector of scores Y should be correlated with the true selection S, but too strong a correlation would violate DP. The auditor computes T from Y which is a “guess” at S. By postprocessing, T is a DP function of S.

3. Theoretical Analysis

To obtain a lower bound on the DP parameters we show that DP implies a high-probability upper bound on the number of correct guesses $W := \sum_i^m \max\{0, T_i \cdot S_i\}$ of our auditing procedure (Algorithm 1). 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 3.1 (Main Result)**.** *Let* $(S, T) \in \{-1, +1\}^m \times \{-1, 0, +1\}^m$ *be the output of Algorithm 1. Assume the*

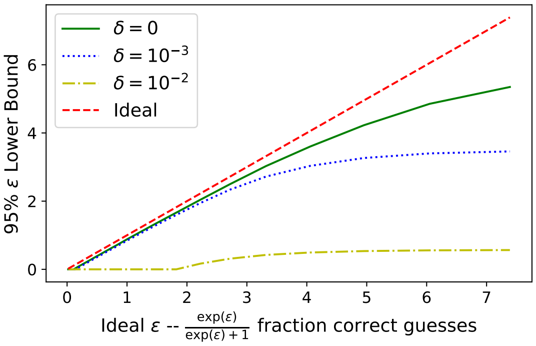

Figure 2. Lower bound on the privacy parameter ε given by Theorem 3.1 with 95% confidence as the number of correct guesses changes. The total number of examples and guesses is 1000 (with no abstentions). Here we plot the ideal ε on the horizontal axis, so that the number of correct guesses is $1000 \cdot \frac{e^\varepsilon}{e^\varepsilon+1}$.

algorithm to audit $\mathcal{A}$ satisfies (ε, δ)-DP. Let $r := k_+ + k_- = \|T\|_1$ be the number of guesses. Then, for all $v \in \mathbb{R}$,

$$\mathbb{P}\left[\sum_i^m \max\{0, T_i \cdot S_i\} \geq v\right] \leq \mathbb{P}\left[\check{W} \geq v\right] + \delta \cdot m \cdot \alpha, \quad (2)$$

where $\check{W} \leftarrow \mathsf{Binomial}\left(r, \frac{e^\varepsilon}{e^\varepsilon+1}\right)$ and

$$\alpha = \max_{i \in [m]} \frac{2}{i} \cdot \mathbb{P}\left[v > \check{W} \geq v - i\right]. \quad (3)$$

If we ignore δ for the moment, Theorem 3.1 says that the number of correct guesses is stochastically dominated by $\mathsf{Binomial}\left(r, \frac{e^\varepsilon}{e^\varepsilon+1}\right)$, 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

$$\mathbb{P}_{\check{W} \leftarrow \mathsf{Binomial}\left(r, \frac{e^\varepsilon}{e^\varepsilon+1}\right)}\left[\check{W} \geq \underbrace{\frac{r \cdot e^\varepsilon}{e^\varepsilon+1} + \sqrt{\frac{1}{2} \cdot r \cdot \log(\tfrac{1}{\beta})}}_{=v}\right] \leq \beta. \quad (4)$$

There is an additional $O(\delta)$ term in the guarantee (2). The exact expression (3) 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 4 with $\beta \leq 1/r^4$, we have $\alpha \leq O(1/r)$.

Theorem 3.1 gives 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 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.

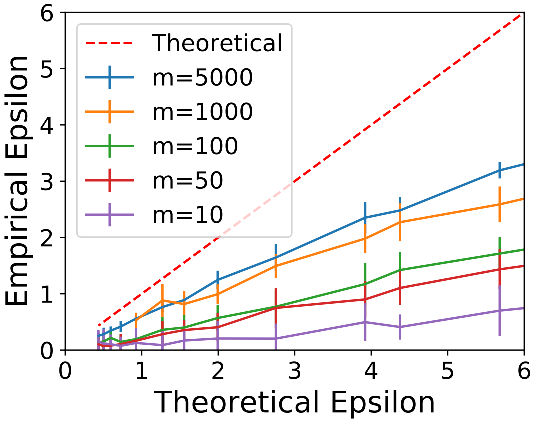

Figure 3. Effect of the number of auditing examples (m) in the white-box setting. By increasing the number of the auditing examples we are able to achieve tighter empirical lower bounds.

Proof of Theorem 3.1: Due to space limitations, the proof is deferred to the full version. But we outline the main ideas: First note that $S \in \{-1, +1\}^m$ is uniform and T is a DP function of S. Consider the distribution of S conditioned on $T = t$. If $\mathcal{A}$ is pure $(\varepsilon, 0)$-DP, then we can easily analyze the conditional distribution using Bayes' law to conclude that each guess has probability $\leq \frac{e^\varepsilon}{e^\varepsilon+1}$ of being correct. Furthermore, this holds even if we condition on the other guesses, which allows us to inductively prove that the number of correct guesses is stochastically dominated by $\mathsf{Binomial}\left(r, \frac{e^\varepsilon}{e^\varepsilon+1}\right)$. Handling approximate DP ($\delta > 0$) introduces additional complexity – some outputs T are "bad" in the sense that the conditional distribution of S_i could be arbitrary. Fortunately, such bad outputs are rare (Kasiviswanathan & Smith, 2014). What we can show is that the number of correct guesses is stochastically dominated by $\check{W} + F(T)$, where $\check{W} \leftarrow \mathsf{Binomial}\left(r, \frac{e^\varepsilon}{e^\varepsilon+1}\right)$ is as before and $F(T) \in \{0, 1, \cdots, m\}$ indicates how many bad events happened. We do not know the exact distribution of $F(T)$, but we do know $\mathbb{E}\left[F(T)\right] \leq 2m\delta$, which suffices to prove our result. Equation 3 comes from looking for the worst-case $F(T)$; essentially the worst case is $\mathbb{P}\left[F(T) = i\right] = 2m\delta/i$ and $\mathbb{P}\left[F(T) = 0\right] = 1 - 2m\delta/i$ for some $i \in [m]$.

4. Experiments

We rely on the experimental setup of Nasr et al. (2023). We run DP-SGD on the CIFAR-10 dataset with Wide ResNet (WRN-16) (Zagoruyko & Komodakis, 2016). Our experiments reach 76% test accuracy at $(\varepsilon = 8, \delta = 10^{-5})$-DP, which is comparable with the state-of-the-art (De et al., 2022). Unless specified otherwise, all lower bounds are with 95% confidence. We use Algorithm 1 to audit DP-SGD and we convert the results into lower bounds on the privacy parameters using Theorem 3.1.

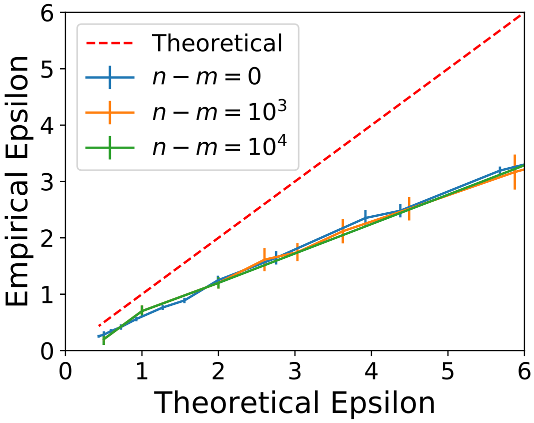

Figure 4. Effect of the number of additional examples $(n-m)$ in the white-box setting. Importantly, adding additional examples does not impact the auditing results in the white-box setting.

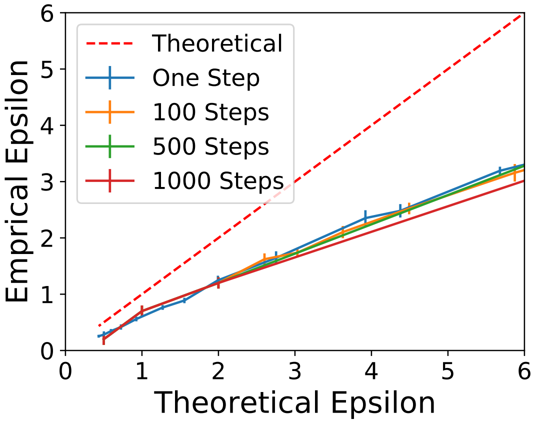

Figure 5. Effect of number of iterations in the white-box setting. Increasing the number of the steps (while incresing noise to keep overall privacy fixed) will not affect the auditing results.

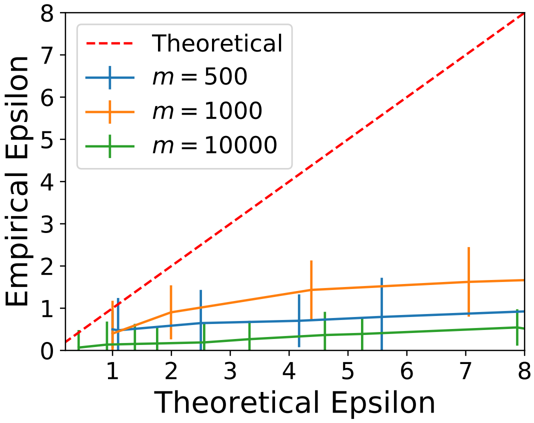

Figure 6. Effect of the number of auditing examples (m) in the black-box setting. Black-box auditing is very sensitive to the number of auditing examples.

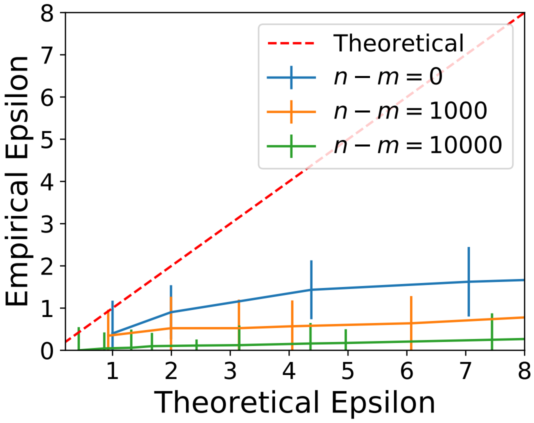

Figure 7. Effect of the number of additional example on auditing $(n-m)$ in the black-box setting. By increasing the number of additional examples, the auditing results get significantly looser.

Following Nasr et al. (2023), we consider both gradient and input attacks. In the white-box setting, we consider gradient attacks, which injects artificial gradients; to compute the score, we use the inner product between the gradient update and auditing gradient. In the black-box setting, we inject different types of injected examples – either randomly mislabelled examples or in-distribution examples; we use the loss of the input example as the score. We evaluate different values of k_+ and k_- and report the best auditing results.

Figures 3, 4, 5, 6, and 7 summarize our experimental results.

5. Discussion

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

In the idealized setting where the fraction of guesses that are correct is fixed at $\frac{e^\varepsilon}{e^\varepsilon+1}$, then the auditing lower bound approaches the true ε as the number of guesses increases. E.g., if $\varepsilon = 4$ in this setting, then, with 10,000 guesses, we get $\varepsilon \geq 3.87$ with 95% confidence.

Limitations: However, for realistic mechanisms – i.e., those based on the Gaussian mechanism (e.g., DP-SGD) – the fraction of correct guesses is not simply $\frac{e^\varepsilon}{e^\varepsilon+1}$; it depends on the number of guesses versus abstentions. And both the upper and lower bounds on ε depend on δ. Thus we cannot get tight bounds using our method. If we audit the Gaussian mechanism (i.e., adding $\mathcal{N}(0,1)$ to a sensitivity-1 value) for $\delta = 10^{-5}$ with $m = 10^5$ auditing examples, we get a lower bound of $\varepsilon \geq 2.675$ versus an upper bound of $\varepsilon = 4.38$. This limitation is inherent because the worst-case mechanisms for which Theorem 3.1 is tight do not correspond to realistic mechanisms.

References

Bassily, R., Nissim, K., Smith, A., Steinke, T., Stemmer, U., and Ullman, J. Algorithmic stability for adaptive data analysis. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pp. 1046–1059, 2016. URL https://arxiv.org/abs/1511.02513.

Bichsel, B., Gehr, T., Drachsler-Cohen, D., Tsankov, P., and Vechev, M. Dp-finder: Finding differential privacy violations by sampling and optimization. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pp. 508–524, 2018.

De, S., Berrada, L., Hayes, J., Smith, S. L., and Balle, B. Unlocking high-accuracy differentially private image classification through scale. *arXiv preprint arXiv:2204.13650*, 2022.

Ding, Z., Wang, Y., Wang, G., Zhang, D., and Kifer, D. Detecting violations of differential privacy. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pp. 475–489, 2018.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. 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*, pp. 265–284. Springer, 2006. URL https://www.iacr.org/archive/tcc2006/38760266/38760266.pdf.

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., and Roth, A. Generalization in adaptive data analysis and holdout reuse. *Advances in Neural Information Processing Systems*, 28, 2015a. URL https://arxiv.org/abs/1506.02629.

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., and Roth, A. L. Preserving statistical validity in adaptive data analysis. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pp. 117–126, 2015b. URL https://arxiv.org/abs/1411.2664.

Jagielski, M., Ullman, J., and Oprea, A. Auditing differentially private machine learning: How private is private sgd? *Advances in Neural Information Processing Systems*, 33:22205–22216, 2020.

Jung, C., Ligett, K., Neel, S., Roth, A., Sharifi-Malvajerdi, S., and Shenfeld, M. A new analysis of differential privacy's generalization guarantees. *arXiv preprint arXiv:1909.03577*, 2019. URL https://arxiv.org/abs/1909.03577.

Kasiviswanathan, S. P. and Smith, A. On the'semantics' of differential privacy: A bayesian formulation. *Journal of Privacy and Confidentiality*, 6(1), 2014. URL https://arxiv.org/abs/0803.3946.

Malek Esmaeili, M., Mironov, I., Prasad, K., Shilov, I., and Tramer, F. Antipodes of label differential privacy: Pate and alibi. *Advances in Neural Information Processing Systems*, 34:6934–6945, 2021.

Nasr, M., Hayes, J., Steinke, T., Balle, B., Tramèr, F., Jagielski, M., Carlini, N., and Terzis, A. Tight auditing of differentially private machine learning. *arXiv preprint arXiv:2302.07956*, 2023. URL https://arxiv.org/abs/2302.07956.

Rogers, R., Roth, A., Smith, A., and Thakkar, O. Max-information, differential privacy, and post-selection hypothesis testing. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 487–494. IEEE, 2016. URL https://arxiv.org/abs/1604.03924.

Steinke, T. and Zakynthinou, L. Reasoning about generalization via conditional mutual information. In *Conference on Learning Theory*, pp. 3437–3452. PMLR, 2020. URL https://arxiv.org/abs/2001.09122.

Tramer, F., Terzis, A., Steinke, T., Song, S., Jagielski, M., and Carlini, N. Debugging differential privacy: A case study for privacy auditing. *arXiv preprint arXiv:2202.12219*, 2022. URL https://arxiv.org/abs/2202.12219.

Zagoruyko, S. and Komodakis, N. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Zanella-Béguelin, S., Wutschitz, L., Tople, S., Salem, A., Rühle, V., Paverd, A., Naseri, M., and Köpf, B. Bayesian estimation of differential privacy. *arXiv preprint arXiv:2206.05199*, 2022.

A. Full Version

https://arxiv.org/abs/2305.08846