Bounding User Contributions: A Bias-Variance Trade-off in Differential Privacy

Kareem Amin [1] Alex Kulesza [1] Andres Muñoz Medina [1] Sergei Vassilvitskii [1]

Abstract

Differentially private learning algorithms protect individual participants in the training dataset by guaranteeing that their presence does not significantly change the resulting model. In order to make this promise, such algorithms need to know the maximum contribution that can be made by a single user: the more data an individual can contribute, the more noise will need to be added to protect them. While most existing analyses assume that the maximum contribution is known and fixed in advance—indeed, it is often assumed that each user contributes only a single example—we argue that in practice there is a meaningful choice to be made. On the one hand, if we allow users to contribute large amounts of data, we may end up adding excessive noise to protect a few outliers, even when the majority contribute only modestly. On the other hand, limiting users to small contributions keeps noise levels low at the cost of potentially discarding significant amounts of excess data, thus introducing bias. Here, we characterize this trade-off for an empirical risk minimization setting, showing that in general there is a "sweet spot" that depends on measurable properties of the dataset, but that there is also a concrete cost to privacy that cannot be avoided simply by collecting more data.

1. Introduction

Differential privacy (Dwork & Roth, 2014) has emerged as the standard framework for quantifying information revealed by an algorithm about the users that supply its underlying data. A differentially private algorithm guarantees that the presence of any single user in the dataset cannot be accurately predicted from the algorithm's output; this is achieved by perturbing the result using random noise. Differential privacy is built around a rigorous theory and has strong formal properties; for example, the protection it affords cannot be broken by any kind of post-processing.

While a variety of mechanisms for generating differentially private algorithms are now known—and, increasingly, used in practice—significant challenges remain. In applying differential privacy to location data, Pyrgelis et al. (2018) lamented that "differentially private mechanisms . . . yield a significant loss in utility." Describing their use of differential privacy at the U.S. Census Bureau, Garfinkel et al. (2018) wrote that their chosen value for the privacy parameter ϵ (for which smaller values indicate stronger privacy protection) was "far higher than those envisioned by the creators of differential privacy." In order to achieve their desired utility, they were forced to accept a less than ideal level of protection.

While a variety of factors contribute to these kinds of problems, we focus here on a particular difficulty arising from the need to add noise sufficient to mask the largest effect of any individual user. In typical applications, this maximum effect can be quite large or potentially unbounded: even when typical users contribute only a modest amount of data, there can be extreme outliers, and they must be protected too. (Arguably, the protection of outliers is even more important.) And, making things worse, we must protect not only the users already in the dataset, but also the hypothetical users who might have elected not to contribute—otherwise an attacker could infer their absence.

Formally, the magnitude of the noise usually must be calibrated to match the *sensitivity* of the analysis with respect to a single user. Most existing theoretical work assumes that the sensitivity is fixed and known in advance; for instance, in differentially private learning it is often assumed that each user can contribute only a single example (Chaudhuri et al., 2011; Bassily et al., 2014). In reality, of course, users often contribute many examples, with different users contributing at vastly different rates; a single user might thus be responsible for a disproportionately large fraction of the dataset.

[1] Google Research New York, NY, USA. Correspondence to: Kareem Amin <kamin@google.com>, Alex Kulesza <kulesza@google.com>, Andres Muñoz Medina <ammedina@google.com>, Sergei Vassilvitskii <sergeiv@google.com>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

This phenomenon, where the amount of data contributed by a user follows a heavy-tailed distribution, is often referred to as a "power law" (or by other names in other contexts, e.g., Zipf's law in linguistics). Power laws are extremely common in real datasets across diverse domains, whether counting the number of movies rated by a user (Harper & Konstan, 2015) or the number of connections a user has in a social network (Leskovec & Krevl, 2014).

For differentially private algorithms, the dependence on sensitivity can lead to huge amounts of noise in such situations. Practitioners sometimes compensate by raising ϵ, but this results in reduced privacy protection. Here, we investigate a common alternative approach: limiting the contributions of individual users in order to reduce the sensitivity. (Indeed, this can be required when the sensitivity would otherwise be unbounded.)

A fundamental question is how to choose a value for the maximum allowed contribution. If set too high, the noise level may be so great that any utility in the result is lost. If set too low, we will be forced to discard large amounts of data. This not only reduces our sample size, but also adds bias: users who contributed more than the limit are now under-represented. As highly active users often behave quite differently from occasional users, this is a non-trivial concern.

In this paper we investigate this bias-variance trade-off in detail, showing that in general there is an intermediate contribution limit for which the expected error of differentially private empirical risk minimization is optimal. That is, a biased training set can actually be *preferable* when the learning algorithm is differentially private. We identify the relevant characteristics of the domain that control this trade-off, showing that in some scenarios they can be measured or approximated using prior information.[1]

Along the way, our analysis reveals that there is no contribution limit that suffices to eliminate both bias and variance, even in the limit of infinite data. This suggests an explanation for the difficulties encountered by practitioners, among whom contribution limiting techniques are common. Moreover, it leaves open the question of whether there exist better generic preprocessing methods for differentially private learning algorithms than simple contribution bounding.

1.1. Related Work

Bias-variance trade-offs are well documented across the machine learning and statistics literature: from structural risk minimization and regularization (Vapnik, 1998) to bandit algorithms (Bubeck & Cesa-Bianchi, 2012), biasing the training distribution in order to reduce variance is a commonly used technique. There are now general theoretical tools like VC dimension (Vapnik, 1998) and Rademacher complexity (Mohri et al., 2012) that can be used to understand this trade-off in learning problems.

In contrast to these general bias-variance trade-offs, the one studied in this paper is quite specific to differential privacy: it arises from capping user contributions to a data set in order to reduce the sensitivity of a learning algorithm. In this specialized scenario we know the exact form of the noise (generally Laplacian or Gaussian) and can therefore provide explicit bounds.

The difficulties associated with sensitivity bounding are well-known in the differential privacy literature. Previous work has sought to address cases where the *typical* sensitivity is expected to be much lower than the theoretical sensitivity; for instance, when computing medians, one can construct datasets where the removal of a single value can change the result arbitrarily, but in practice, most data sets are dense near the median, and so deleting a single datapoint has almost no effect.

Using *smooth sensitivity* techniques that depend on the actual dataset (as opposed to the worst-case dataset), Nissim et al. (2007) showed how it is possible to significantly reduce noise in these cases. Similarly, Dwork & Lei (2009) developed techniques for measuring whether the actual dataset has acceptable sensitivity, and rejecting the analysis if it does not. Both of these approaches fail, however, when the typical sensitivity is large, as is the case when computing sums or averages. More generally, quantifying their utility loss remains a challenge.

Other efforts to reduce the noise required to make a dataset differentially private include biasing the loss function by adding a regularization penalty or minimizing a noisy version of the loss function (Chaudhuri et al., 2011). However, this work still relies on the common assumption that every user contributes a single record to a database. Another line of work similar to the one introduced here is that of Smith (2011). The authors analyzed the bias and variance of private estimators, but for a subset of statistics that they call *generically normal*.

Our work complements these results by analyzing the effects of removing the single instance per user assumption. The high level idea is to formally analyze the effects of limiting the number of times a user is allowed to contribute to the dataset, quantifying the bias and variance that this process generates.

Similar ideas have been proposed in the context of graph analysis, with privacy guarantees with respect to a single node. Kasiviswanathan et al. (2013) gave a method for

[1] Note that we do not attempt to optimize the trade-off for general distributions in a differentially private way; this remains an open problem. Here we are most interested in understanding the forces at play.

computing private statistics when the degree of the node can be arbitrary, and their work proceeds by bounding the node degree and analyzing the bias this introduces. However, their results require specific assumptions on degree distributions, whereas our results are much more general and hold for any distribution of user contributions. On the other hand, while the work of Blocki et al. (2013) does not require additional assumptions, their method for finding a bounded degree graph is not computationally efficient. Moreover, neither of these easily extends to the empirical risk minimization domain.

Other examples of bounding contributions to reduce noise can be found in private training methods for neural networks. For instance, Abadi et al. (2016) and Geyer et al. (2017) truncate the gradient of a neural network to control the sensitivity of the sum of gradients. Nevertheless, this work fails to provide a detailed analysis of how to choose the truncation level for the gradient norm, instead suggesting using the median of observed gradients. We show that, in fact, using the median (or any fixed quantile independent of the privacy parameter ϵ) as a cap can yield suboptimal estimates of a sum.

We emphasize that our goal is to provide a formal analysis of the trade-off introduced by bounding user contributions, and express these bounds in terms of observable and computable quantities. Previous work either ignored the question of privately computing the trade-off completely (Abadi et al. (2016) and Geyer et al. (2017) simply use the empirical median), required strong assumptions on the data (Kasiviswanathan et al. (2013) need the graphs to have specific degree distributions), or relied on quantities not easily computable (Chaudhuri et al. (2011) provide an optimal selection of the parameter that depends on the norm of the optimal hypothesis).

2. Preliminaries

For our purposes, a dataset $\mathcal{S} \in \mathbb{S}$ is a collection of contributions made by individual users. For instance, a dataset might comprise a set of training examples, each contributed by a particular user. Each user might be able to contribute any number of examples.

Definition 1. *We say two datasets $\mathcal{S}, \mathcal{S}' \in \mathbb{S}$ are neighbors and write $\mathcal{S} \sim \mathcal{S}'$ if one can be recovered from the other by removing only the data corresponding to a single user.*

Definition 2. *Let H be a hypothesis space. An algorithm $\mathcal{A} : \mathbb{S} \to H$ is said to be ϵ-differentially private if, for every pair of neighboring datasets $\mathcal{S} \sim \mathcal{S}'$ and every $U \subseteq H$,*

$$\Pr(\mathcal{A}(\mathcal{S}) \in U) \leq e^{\epsilon} \Pr(\mathcal{A}(\mathcal{S}') \in U) \,.$$

A differentially private algorithm produces approximately the same output distribution whether or not any single user chooses to participate in the dataset. This provides a form of plausible deniability: an adversary viewing the result cannot determine with high confidence whether a user was even present in the dataset, let alone any details of the user's data. Of course, a large number of users can, in aggregate, have a large effect on the algorithm's output; thus, differential privacy allows us to learn population-level information while protecting the privacy of individual users.

As implied by Definition 2, any nontrivial differentially private algorithm must be stochastic; that is, it must involve some kind of noise. Determining the scale of this noise is critical. If the noise level is too low, the algorithm will not be differentially private; if it is too high, the utility of the result will be unnecessarily degraded. Ideally, we should use just the amount of noise needed to obscure the effect of any single user. This idea is formalized using the concept of *sensitivity*.

Definition 3. *The (ℓ_1) sensitivity of a function $f : \mathbb{S} \to \mathbb{R}$ is given by*

$$\Delta_f = \sup_{\mathcal{S} \sim \mathcal{S}'} |f(\mathcal{S}) - f(\mathcal{S}')| \,. \tag{1}$$

Δ_f is the maximum amount that adding or removing a user from the dataset can change the value of f.

To see how sensitivity drives noise, consider the Laplace mechanism, a simple technique for approximating a function f in a differentially private way (Dwork et al., 2006).

Definition 4 (Laplace mechanism)**.** *Given a target function $f : \mathbb{S} \to \mathbb{R}$ and a fixed $\epsilon > 0$, the Laplace mechanism $\mathrm{Lap}_{f,\epsilon}(\mathcal{S})$ returns $f(\mathcal{S}) + \eta$, where η is a random noise variable with density proportional to $\exp(-\epsilon|\eta|/\Delta_f)$. (That is, η is a Laplace variable with scale parameter Δ_f/ϵ.)*

The Laplace mechanism is ϵ-differentially private. Note that Δ_f (along with ϵ) controls the noise level: when the target function f is highly sensitive, more noise is required to ensure the same level of privacy.

3. A Simple Example

Before proceeding to our main result, we illustrate the underlying concepts in a simpler setting. Suppose that $\mathcal{S}$ is a collection of n non-negative real numbers $x_1, x_2, \ldots, x_n$, each contributed by a unique user. We would like to estimate the sum of the numbers in our dataset in a differentially private way while minimizing the absolute error. (This setup has connections to federated learning, for instance, where each x_i is a vector representing a gradient and the learner is interested in the sum of these gradients.)

Naïvely, we might try to do this by applying the Laplace mechanism to the function $g(\mathcal{S}) = \sum_{i=1}^n x_i$. But there is a problem: since a single user can contribute an arbitrarily large value, the sensitivity of g, and therefore the scale of the noise, is infinite. To fix this, we will introduce a cap T on the maximum size of a user's contribution, instead applying the Laplace mechanism to the function $g_T(\mathcal{S}) = \sum_{i=1}^{n} \min(x_i, T)$. This will bias our estimated sum, of course, but it also reduces the amount of added noise, as the sensitivity is now

$$\max_{\mathcal{S}\sim\mathcal{S}'} |g_T(\mathcal{S}) - g_T(\mathcal{S}')| = T\ . \tag{2}$$

So how should we choose T? In practice (Abadi et al., 2016), a rule of thumb is to set T equal to the median of the observed points. Is this choice optimal? This is an instance of the basic question we aim to understand in this paper.

First, recall from Definition 4 that the noise added to g_T follows a Laplace distribution with scale parameter Δ_{g_T}/ϵ. We can decompose the expected error of the estimate $\hat{g}$ produced by $\mathrm{Lap}_{g_T,\epsilon}(\mathcal{S})$ into a variance term (due to the noise) and a bias term (due to the contribution limit):

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\hat{g}}|\hat{g} - g(\mathcal{S})| && (3)\\
&\quad\le \mathop{\mathbb{E}}_{\hat{g}} |\hat{g} - g_T(\mathcal{S})| + |g_T(\mathcal{S}) - g(\mathcal{S})| \\
&\quad= \Delta_{g_T}/\epsilon + |g_T(\mathcal{S}) - g(\mathcal{S})| && (4)\\
&\quad= T/\epsilon + \sum_{i=1}^{n} \max(0, x_i - T)\ , && (5)
\end{aligned}$$

where we use the fact that the mean absolute deviation of a Laplace variable is equal to its scale parameter. Note that, if T is very small, the variance is almost zero but the bias is approximately $g(\mathcal{S})$, rendering the estimate useless. On the other hand, as T gets large, the bias drops to zero but the noise increases without bound. We can find the optimal T by noting that the bound is convex with sub-derivative

$$\frac{1}{\epsilon} - |\{i : x_i > T\}|\ , \tag{6}$$

thus the minimum is achieved when T is equal to the $\lceil 1/\epsilon \rceil$th largest value in $\mathcal{S}$. Note that the analysis is nearly tight, and it is easy to show that

$$\mathop{\mathbb{E}}_{\hat{g}} |\hat{g} - g(\mathcal{S})| \ge \frac{1}{2}\left(T/\epsilon + \sum_{i=1}^{n} \max(0, x_i - T)\right).$$

This is in some ways a counter-intuitive result. It says that the limit we should impose on user contributions is just the $(1 - 1/n\epsilon)$-quantile of the contributions themselves. It does not matter how large or small the contributions are above or below the cutoff, only that a fixed number of values are clipped. We do not need detailed knowledge of the distribution of user values; a simple statistic suffices.

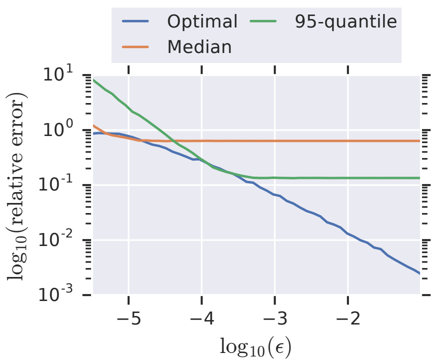


Figure 1. Error when privately estimating the number of ratings in the MovieLens dataset using different truncation strategies, averaged over 1000 runs. Fixed quantiles achieve the optimal trade-off for a single ϵ, but remain strongly suboptimal for other values of ϵ.

This last point is especially important given that any information about the dataset used to determine T must itself be computed privately. (That is, choosing T based on exact statistics of the dataset might indirectly reveal user information.) Luckily, there are a variety of differentially private algorithms that can be applied to approximating quantiles (Nissim et al., 2007; Dwork & Lei, 2009; Smith, 2011). Moreover, a quantile is an intuitive property of the dataset about which a practitioner might have strong prior knowledge, enabling the selection of a good T without reference to the dataset at all. We will see in Section 6 that a similarly intuitive statistic also appears in the more general setting.

Example. The MovieLens 20M dataset[2] consists of 20 million movie ratings from 138 thousand users. Consider the problem of releasing the total number of ratings in a differentially private way. That is, if x_i is the number of movies rated by user i, we are interested in releasing $g = \sum_i x_i$.

We compare three methods of contribution bounding, followed by the Laplace mechanism: (1) truncation of $\{x_i\}$ to the $(1 - 1/n\epsilon)$-quantile, as suggested by our analysis, (2) the commonly used truncation at the median, and (3) truncation at the 95% quantile. Figure 1 shows the relative error in the sum as a function of the privacy parameter ϵ. It is easy to observe from this simple experiment that no fixed quantile is universally good for capping contributions when estimating a sum; for a given ϵ it will either over-truncate contributions, leading to bias (the flat part of the error curve), or the sensitivity will be too high, leading to high levels of noise (the linear part of the error curve). By optimizing the threshold we balance this trade-off and essentially always outperform capping at a fixed quantile.

[2] https://grouplens.org/datasets/movielens/

4. Contribution Bounding for Learning

In the remainder of this work, we analyze a similar bias-variance tension that arises during differentially private learning. Suppose now that $\mathcal{S}$ is a collection of examples $z_1, z_2, \ldots, z_n$ contributed by a population of users. A single user may have provided all n examples, or each z_i could have been provided by a unique user. Assuming $z_i = (x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, we wish to deploy some learning algorithm that outputs a good hypothesis $h \in H$ with respect to a loss function $L : H \times \mathcal{Z} \to [0, 1]$.

In particular, we would like to do empirical risk minimization (ERM). Given an arbitrary set of example vectors $\mathcal{S}$, we can define the empirical risk of an arbitrary hypothesis h as $\mathcal{L}_{\mathcal{S}}(h) = \frac{1}{|\mathcal{S}|}\sum_{z \in \mathcal{S}} L(h, z)$. A differentially private ERM algorithm selects $h_{\text{priv}} \in H$ in a manner that satisfies Definition 2 while keeping $\mathcal{L}_{\mathcal{S}}(h_{\text{priv}})$ close to the actual minimal empirical risk $\mathcal{L}^*_{\mathcal{S}} = \inf_{h \in H} \mathcal{L}_{\mathcal{S}}(h)$.

A variety of algorithms have been proposed for differentially private ERM, including input perturbation, output perturbation, and the exponential mechanism with utility measured by $\mathcal{L}_{\mathcal{S}}(h)$ (Chaudhuri et al., 2011; Bassily et al., 2014). The specific utility guarantees depend on the method, as well as assumptions on L. For simplicity, most previous work assumes that users are guaranteed to contribute at most one example to the dataset. However, these algorithms are easily adapted to settings where users contribute multiple examples. Generically, the utility will depend on the fraction of the dataset contributed by a single user; here, we use the following definition.

Definition 5 (F-utility)**.** *Fix a dataset $\mathcal{S}$ of size n and a loss function $\mathcal{L}_{\mathcal{S}}$, and let τ be the maximum number of examples in $\mathcal{S}$ contributed by any single user. We say that an ϵ-differentially private ERM algorithm gives an F-utility guarantee if, with probability at least $1 - \delta$, for every hypothesis h,*

$$\mathcal{L}_{\mathcal{S}}(h_{\text{priv}}) \le \mathcal{L}_{\mathcal{S}}(h) + F(\tau/n, 1/\epsilon, 1/\delta) ,$$

where the probability is taken over the randomness in the differentially private ERM algorithm, and $F : \mathbb{R} \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ captures the growth of the error with respect to the different parameters.

Typically, F will be polynomial in $\tau/n, 1/\epsilon$ and polylogarithmic in $1/\delta$. For example, when $\tau = 1$, Lemma 3 in (Chaudhuri et al., 2011), concludes that output perturbation on a sufficiently smooth and regularized loss function satisfies the F-utility guarantee with $F(1/n, 1/\epsilon, 1/\delta) = C_F \frac{\ln^2(1/\delta)}{n^2\epsilon^2}$, for a constant C_F specific to their setting (depending on, e.g., dimensionality constants, and the smoothness and regularization parameters of L). It is an easy modification of their Corollary 1 to show that if users can contribute up to τ examples, then $F(\tau/n, 1/\epsilon, 1/\delta) = \tau^2 F(1/n, 1/\epsilon, 1/\delta)$.

In datasets where the participation of a single user is unbounded, the only *a priori* bound on τ is n, and the F-utility guarantee given above is vacuous in the sense that it is greater than 1 and does not decay with n. A standard approach used in practice is therefore to first bound the contribution of any single user to the dataset by setting a hard contribution threshold of τ. If a user contributes more than τ examples to the original dataset $\mathcal{S}$, we only take the first τ to produce a modified dataset $\mathcal{S}_\tau$. This bounds the sensitivity of the learning algorithm to the presence or absence of any particular user. As in our sum example from Section 3, a small τ introduces bias; the contribution-bounded dataset $\mathcal{S}_\tau$ no longer resembles $\mathcal{S}$. On the other hand, a large τ increases the sensitivity of the algorithm to a single user, requiring more noise and eventually making the F-utility bound vacuous. What exactly is the trade-off between bias and variance as a function of τ? We provide an answer to this question in the following section.

5. Setting and Main Results

Before stating the main results of our work, we introduce some additional structure to the problem. Let $\mathbb{N}$ index a (possibly infinite) set of users. Each example z_i in $\mathcal{S}$ is generated by first selecting a user from $\mathbb{N}$, then generating an example from a distribution specific to that user. We assume that both stages of this process are i.i.d., so that there is a well-defined distribution D generating the examples in $\mathcal{S}$. That is, z_i is generated by first drawing a user J_i from the *participation distribution* P, and then drawing z_i according to the *user-data distribution* D^{J_i}. We can then define the risk $\mathcal{L}(h) = \mathbb{E}_{z \sim D}[L(h, z)]$ of a hypothesis in H. We also parametrize P by $p_j = \mathbb{P}_{J \sim P}[J = j]$.

As introduced in Section 4, we are interested in selecting τ so that running private ERM on $\mathcal{S}_\tau$ produces a good hypothesis. It is reasonable to have to this threshold grow as the size of the dataset grows (more on this in Section 7), and so we will let $\tau = \tau_0 n$ for some $\tau_0 \in [0, 1]$. (Although τ is a function of n, we will drop this for clarity of notation in what follows). τ_0 represents the maximum fraction of the original dataset that a single user can contribute to $\mathcal{S}_\tau$. We are therefore interested in the following procedure.

Algorithm 1 Contribution-bounded ERM.

```
**Input:** Dataset $\mathcal{S} = (z_1, \ldots, z_n)$ drawn from $D$, contribution fraction $\tau_0 > 0$, privacy parameter $\epsilon > 0$.
Construct $\mathcal{S}_\tau$ where $\tau = \tau_0 n$.
Run $\epsilon$-differentially private ERM on $\mathcal{S}_\tau$ to obtain $h_{\text{priv}}$.
**return** $h_{\text{priv}}$.
```

Notice that the privacy of h_{priv} holds regardless of the process that generated $\mathcal{S}$. However, in this setting we will be able to characterize the bias-variance trade-off induced by the choice of τ_0. In particular, the trade-off will depend on properties of the participation distribution, P.

Let n_j denote the number of times user j was observed in the data. That is, $n_j = \sum_{i=1}^n \mathbb{1}_{J_i = j}$. Define also $n_{j\tau} = \min(n_j, \tau)$ and $n_\tau = \sum_j n_{j\tau}$. The quantity n_τ corresponds to the *effective sample size* of the data. Finally, let $\mathcal{L}_j(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}^j}[L(h,y)]$ indicate the risk with respect to the j-th user's data distribution. We can now state our main theorem.

Theorem 1. *Suppose Algorithm 1 is run with an ϵ-differentially private ERM algorithm admitting an F-utility bound. Suppose L is 1-Lipschitz, $\|L\|_\infty \leq 1$, $d = \mathrm{VCDim}(H)$. For every n, $\tau > 0$ and $\delta > 0$ with probability at least $1-\delta$:*

$$\mathcal{L}(h_{\text{priv}}) \leq \inf_{h\in H} \mathcal{L}(h) + \underbrace{2\sup_{h\in H}\left|\sum_j \left(\frac{n_{\tau j}}{n_\tau} - \frac{n_j}{n}\right)\mathcal{L}_j(h)\right|}_{\textit{bias}}$$
$$+ \underbrace{O\left(\sqrt{\frac{d\log\frac{n}{\delta}}{n_\tau}}\right)}_{\textit{finite sample variance}} + \underbrace{F\left(\frac{\tau}{n_\tau}, 1/\epsilon, 3/\delta\right)}_{\textit{privacy noise variance}} .$$

We can already make some high-level qualitative observations. As $\tau \to n$ we expect a number of things to happen. First, the size of the dataset $\mathcal{S}_\tau$ should approach n, and so the finite sample variance should approach $O(\sqrt{1/n})$ as we retain all the data. Second, the process that generated $\mathcal{S}_\tau$ should start to resemble the i.i.d. process that generated $\mathcal{S}$, and the bias introduced by using $\mathcal{S}_\tau$ in lieu of $\mathcal{S}$ should disappear. Indeed, as $\tau \to n$, we have $n_{j\tau} \to n_j$ and $n_\tau \to n$. On the other hand, increasing the threshold comes at a cost. Indeed, notice that as $\tau \to n$ then the term $\frac{\tau}{n_\tau}$ tends to 1 making the bound on the privacy noise vacuous. This reflects the fact that the privacy mechanism has no *a priori* bound on how much a single user could have contributed to $\mathcal{S}$.

On the other hand, as $\tau \to 1$ we truncate the contribution of every user to a handful of examples. If on top of that the number of users we observe in the sample is in $\Theta(n)$ then $\frac{\tau}{n_\tau}$ will be in $O(1/n)$ and we recover the original bound on the privacy noise needed when every users is known to contribute a single example. On the other hand, by truncating more data we are likely to increase the bias term.

This suggests that there is an optimal choice of τ that will perfectly trade-off the bias introduced by truncation and the error introduced by making the output private. In order to find this value, we must first better understand the bias term. The bias term corresponds to the difference in expectation of a loss under two different empirical probabilities. As such, we could in principle leverage the literature in domain adaptation to bound this term (Blitzer et al., 2007; Cortes et al., 2015). However, these general purpose techniques are not illuminating with respect to the dependence on τ. Instead, we perform additional analysis to quantify the bias, both in terms of an easily computed data-dependent quantity and as a simple function of τ.

6. Understanding the Bias Term

We now show how to bound the bias term in Theorem 1. We provide both a data dependent bound as well as a high probability bound with a very simple dependence on τ_0. These bounds illustrate the types of distributions that are more amenable to truncation, and provide us with a way to analytically determine the optimal truncation parameter τ_0 for certain families of distributions.

To motivate the definitions throughout this section, we first identify some desirable properties of the underlying distribution that would make the term

$$\left|\sum_{j=1}^n \left(\frac{n_{j\tau}}{n_\tau} - \frac{n_j}{n}\right)\mathcal{L}_j(h)\right|$$

small. Notice that if $n_{j\tau} = n_j$ for all j, i.e., if we keep all of the examples, then the term vanishes for all h. Thus any bound on the bias should decrease to 0 as $\tau_0 \to 1$. On the other hand, suppose all of the users are statistically the same, that is, $\mathcal{L}_j(h)$ is constant with respect to j; then reducing the contribution from any user does not incur bias, and the term should also vanish for any τ_0. This intuition suggests the following definition.

Definition 6. *The empirical variance of a hypothesis class H is*

$$\mathrm{Var}(H) = \sup_{h\in H} \mathrm{Var}(h) ,$$

where for any $h \in H$, $\mathrm{Var}(h) = \sum_j \frac{n_j}{n}(\mathcal{L}_j(h) - \mathcal{L}(h))^2$.

The empirical variance quantifies how close $\mathcal{L}_j(h)$ is to being constant across users. We can now state the main result of this section. The proof of this statement can be found in the appendix.

Proposition 1. *If $\|L\|_\infty \leq 1$ and $\tau = \tau_0 n$ for $\tau_0 \in (0,1]$, then the bias term in Theorem 1 can be bounded as follows:*

$$\sup_{h\in H}\left|\sum_{j=1}^n \left(\frac{n_{j\tau}}{n_\tau} - \frac{n_j}{n}\right)\mathcal{L}_j(h)\right|$$
$$\leq \min\left(\sqrt{\frac{1}{2}\log\left(\frac{n}{n_\tau}\right)}, \sqrt{\frac{2n}{n_\tau}\mathrm{Var}(H)}\right)$$
$$\leq \min\left(\sqrt{\frac{1}{2}\log\left(\frac{1}{\tau_0}\right)}, \sqrt{\frac{2\mathrm{Var}(H)}{\tau_0}}\right) .$$

Proposition 1 reduces the complexity of the bias term to understanding a single ratio: $\frac{n_\tau}{n}$, the proportion of the original sample that is kept after capping individual user contributions. This simple statistic can be calculated from data, and can in principle guide the choice of τ_0.

The second bound in the proposition is a crude estimate of this ratio (see the appendix), but is sufficient to capture the intuition we described at the beginning of the section: as $\tau_0 \to 0$ the bias term goes to zero as $\sqrt{\log(1/\tau_0)}$. On the other hand, if D_j does not vary across users, then $\mathrm{Var}(H)$ is zero and the bias term vanishes regardless of τ_0. We can now provide explicit guarantees on the generalization ability of private ERM.

To get a better understanding of the privacy noise term, recall from the discussion following Definition 5 that (Chaudhuri et al., 2011) showed how to instantiate $F(\tau/n, 1/\epsilon, 1/\delta)$ to $O(\frac{\tau^2}{n^2\epsilon^2}\log^2(1/\delta))$.

Substituting these into Theorem 1 we get:

Corollary 1. *Under the assumptions of Theorem 1, the setting of Chaudhuri et al. (2011), and that $n_\tau/n \geq 0.25$, for any δ, with probability at least $1-\delta$:*

$$\mathcal{L}(h_{\mathrm{priv}}) \leq \inf_{h\in H} \mathcal{L}(h) + \underbrace{2\sqrt{(1-n_\tau/n)}}_{\textit{bias}}$$
$$+ \underbrace{O\left(\sqrt{\frac{d\log\frac{n}{\delta}}{n_\tau}}\right)}_{\textit{finite sample variance}} + \underbrace{O\left(\frac{\tau^2}{n_\tau^2\epsilon^2}\log^2(1/\delta)\right)}_{\textit{privacy noise variance}}.$$

The proof follows by substitution of Proposition 1, observing that for $x \leq 3/4$, $\ln\frac{1}{1-x} \leq 2x$, and taking $x = 1 - \frac{n_\tau}{n}$.

The key to understanding the bound is untangling the relationship between τ, n_τ, and n. It can be shown (see Proposition 5 in the appendix) that if $\tau = \tau_0 n$ then $\frac{n_\tau}{n} = \Theta\big(\sum_{j:p_j\leq\tau_0} p_j\big)$. Therefore if there is a small choice of τ such that the majority of the participation probability mass is on users who are not capped ($p_j \leq \tau_0$), then $n_\tau/n \approx 1$ making the bias term small, since τ is also small we achieve a good bias-variance trade-off. In other words, Corollary 1 tells us that private learning is easier when the contribution is spread amongst multiple users and becomes significantly harder if the contribution is concentrated on a small number of users. While this might be intuitively clear, the previous corollary quantifies this effect. In Figure 2 we plot $\tau_0 \mapsto \sum_{j:p_j\leq\tau_0} p_j$ for probabilities of the form $p_j \propto \frac{1}{j^{1+\alpha}}$ for different values of α.

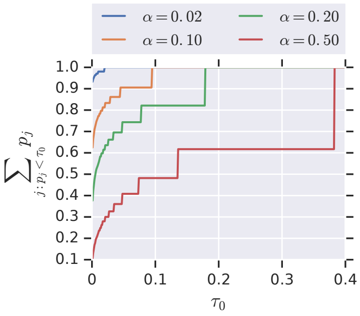


Figure 2. Sum of probability mass of uncapped users.

7. Cost of Privacy

In this section, we draw attention to an interesting consequence of Theorem 1 and Corollary 1. For a finite n, τ_0 and thus τ can be selected to optimize the bound, trading off bias and variance. However, unlike traditional learning bounds, there is no choice of threshold that causes both the bias and variance terms to vanish as $n \to \infty$.

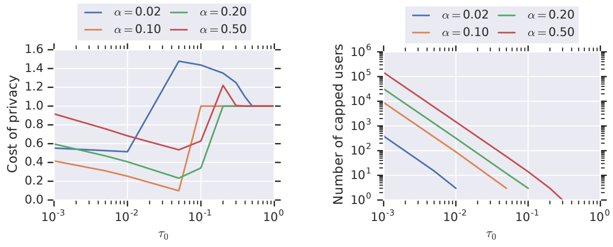
We first discuss this phenomenon and then give an explanation as to why this is not just an artifact of our bound but is indeed fundamental to contribution bounding for differentially private learning, a phenomenon we call the *cost of privacy*.

Any non-trivial F-utility guarantee for a differentially private learning algorithm will be increasing in its first argument. Thus to reduce the privacy variance term to 0 it is critical to make sure the threshold, τ grows sublinearly with n. Otherwise, in case of a constant τ_0 the bound contains a term that does not vanish with the introduction of more data. This, however, makes the bias term increase, as n/n_τ grows as τ_0 shrinks.

The cost of privacy appears not to be an artifact of our bound. In particular, differentially private learning algorithms require both that the dataset is growing and that an individual's contribution to the dataset is bounded by some constant τ in order for the privacy variance to be vanishing. If we keep a constant τ, then as n grows, the number of users whose contribution is capped increases, until $\mathcal{S}_\tau$ appears to have been drawn from a distribution that is uniform over these users i. In all but the luckiest of circumstances this biases the data in $\mathcal{S}_\tau$ away from the data-generation process.

We leave it as an open question whether any mechanism can make the cost of privacy vanish as $n \to \infty$, but we suspect it is unavoidable in general, simply because individual users contribute constant fractions of the dataset. We can either add constant noise to protect these users, or suppress their contributions to a level that leaves the dataset strongly biased.

To give another sense of this cost, we show in Figure 3 the sum of the bias and privacy noise for $\epsilon = 1.0$, different values of τ_0 and different participation distributions P. We consider Pareto distributions where $p_j \propto \frac{1}{j^{1+\alpha}}$. To generate



(a)

(b)

Figure 3. (a) Cost of privacy as a function of τ_0. The minimum error that we can add to a private ERM estimator is given by the minimum of each curve. (b) Number of users whose contribution is capped as a function of τ_0

these curves we use the empirical bound in Proposition 1, simulating samples from the Pareto distributions and averaging the ratio $\frac{n}{n_\tau}$ over 100 runs. Figure 3 (a) shows the cost of privacy for various choices of τ_0; note that the minimum of each curve is significantly above zero. Figure 3 (b) shows the number of users whose contribution is expected to be limited as a function of τ_0. Notice that for the optimal parameter τ_0, the number of users whose contribution are capped is very small (2 to 8 users). This result mirrors that from Section 3, where we capped the contribution of only the top $1/\epsilon$ users.

8. Proofs

We now prove Theorem 1. Throughout this section we assume that L is 1-Lipchitz. Recall that J_i is the random identity of the user selected for example i in $\mathcal{S}$. The first lemma states that the empirical risk converges to a reweighted version of the true risk, conditioned on the outcomes of $\{J_i\}$. The proof follows from standard arguments in learning theory and can be found in the Appendix.

Lemma 1. *Conditioned on the outcomes of $\{J_i\}$, with probability at least $1-\delta$ the following holds uniformly over $h \in H$:*

$$\left| \mathcal{L}_{\mathcal{S}_\tau}(h) - \sum_j \frac{n_{\tau j}}{n_j} \mathcal{L}_j(h) \right| \leq \sqrt{\frac{2d \log \frac{en}{d}}{\tau_0 n}} + \sqrt{\frac{\log(1/\delta)}{2\tau_0 n}}$$

Next, we use Lemma 1 to bound the difference between the empirical risk on our thresholded data set $\mathcal{S}_\tau$ and the true risk.

Lemma 2. *Fix $\delta > 0$ and let $d = \mathrm{VCdim}(H)$. Then with probability at least $1-\delta$, the following inequality holds uniformly for h in H.*

$$|\mathcal{L}_{\mathcal{S}_\tau}(h) - \mathcal{L}(h)| \leq \sqrt{\frac{2d \log \frac{en}{d}}{\tau_0 n}} + \sqrt{\frac{\log(2/\delta)}{2\tau_0 n}}$$
$$+ \left| \sum_j \left(\frac{n_{\tau j}}{n_\tau} - \frac{n_j}{n} \right) \mathcal{L}_j(h) \right| + \sqrt{\frac{\log \frac{4}{\delta}}{2n}} .$$

The proof of this lemma can also be found in the appendix.

Theorem 1. *Suppose Algorithm 1 is run with an ϵ-differentially private ERM algorithm admitting an F-utility bound. Let $d = \mathrm{VCDim}(H)$, and fix n, $\tau_0 > 0$ and $\delta > 0$. Then with probability at least $1-\delta$:*

$$\mathcal{L}(h_{\mathrm{priv}}) \leq \inf_{h \in H} \mathcal{L}(h) + \underbrace{\sup_{h \in H} 2 \left| \sum_j \left(\frac{n_{\tau j}}{n_\tau} - \frac{n_j}{n} \right) \mathcal{L}_j(h) \right|}_{\textit{bias}}$$
$$+ \underbrace{O\left(\sqrt{\frac{d \log \frac{n}{\delta}}{n_\tau}} \right)}_{\textit{finite sample variance}} + \underbrace{F\left(\frac{\tau}{n_\tau}, 1/\epsilon, 3/\delta \right)}_{\textit{privacy noise variance}} .$$

Proof. Let h_{priv} be the hypothesis returned by our algorithm and h^* the hypothesis optimizing $\mathcal{L}$. Notice that h_{priv} is obtained by running Algorithm 1 on a data set of size n_τ. Thus by definition of F-utility we have

$$\mathcal{L}_{\mathcal{S}_\tau}(h_{\mathrm{priv}}) \leq \mathcal{L}_{\mathcal{S}_\tau}(h^*) + F(\frac{\tau}{n_\tau}, \frac{1}{\epsilon}, \frac{3}{\delta}).$$

Next we can use Lemma 2 to bound $\mathcal{L}_{\mathcal{S}_\tau}(h^*)$ in terms of $\mathcal{L}(h^*)$. Similarly we can lower bound $\mathcal{L}_{\mathcal{S}_\tau}(h_{\mathrm{priv}})$ in terms of $\mathcal{L}(h_{\mathrm{priv}})$. Combining both bounds and using the union bound yields the proof. □

9. Conclusion

We have provided a detailed analysis of the bias-variance trade-off which arises as a result of contribution bounding datasets during differentially private learning, specifically in settings where users are allowed to contribute more than one example to the dataset. Our bounds give the practitioner a way to carefully tune contribution-bounding as a function of statistical properties of the data. Moreover, we show that there is a fundamental cost to privacy which does not vanish even with large datasets. We leave as interesting open questions how to estimate the aforementioned statistical properties *privately.* We also leave open the search for other pre-processing techniques that avoid the cost of privacy imposed by naive contribution-bounding.

References

Abadi, M., Chu, A., Goodfellow, I. J., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–318, 2016.

Bassily, R., Smith, A. D., and Thakurta, A. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *55th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2014, Philadelphia, PA, USA, October 18-21, 2014*, pp. 464–473, 2014.

Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Wortman, J. Learning bounds for domain adaptation. In *Proccedings of NIPS*, pp. 129–136, 2007.

Blocki, J., Blum, A., Datta, A., and Sheffet, O. Differentially private data analysis of social networks via restricted sensitivity. In *Proceedings of the 4th Conference on Innovations in Theoretical Computer Science*, ITCS '13, pp. 87–96, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1859-4.

Bubeck, S. and Cesa-Bianchi, N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(Mar):1069–1109, 2011.

Cortes, C., Mohri, M., and Medina, A. M. Adaptation algorithm and theory based on generalized discrepancy. In *Proceedings SIGKDD*, pp. 169–178, 2015.

Dwork, C. and Lei, J. Differential privacy and robust statistics. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pp. 371–380. ACM, 2009.

Dwork, C. and Roth, A. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9:211–407, August 2014.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006.

Garfinkel, S. L., Abowd, J. M., and Powazek, S. Issues encountered deploying differential privacy. *CoRR*, abs/1809.02201, 2018. URL http://arxiv.org/abs/1809.02201.

Geyer, R. C., Klein, T., and Nabi, M. Differentially private federated learning: A client level perspective. *CoRR*, abs/1712.07557, 2017.

Greenberg, S. and Mohri, M. Tight lower bound on the probability of a binomial exceeding its expectation. *CoRR*, abs/1306.1433, 2013. URL http://arxiv.org/abs/1306.1433.

Harper, F. M. and Konstan, J. A. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5 (4):19:1–19:19, December 2015. ISSN 2160-6455.

Kasiviswanathan, S. P., Nissim, K., Raskhodnikova, S., and Smith, A. D. Analyzing graphs with node differential privacy. In *Proceedings of the 10th Theory of Cryptography Conference, TCC*, pp. 457–476, 2013.

Leskovec, J. and Krevl, A. SNAP Datasets: Stanford large network dataset collection. http://snap.stanford.edu/data, June 2014.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of Machine Learning*. Adaptive computation and machine learning. MIT Press, 2012.

Nissim, K., Raskhodnikova, S., and Smith, A. D. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the 39th Annual ACM Symposium on Theory of Computing, San Diego, California, USA, June 11-13, 2007*, pp. 75–84, 2007.

Pyrgelis, A., Troncoso, C., and Cristofaro, E. D. Knock knock, who's there? membership inference on aggregate location data. In *25th Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, February 18-21, 2018*, 2018.

Smith, A. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pp. 813–822. ACM, 2011.

Vapnik, V. *Statistical learning theory*. Wiley, 1998. ISBN 978-0-471-03003-4.

A. Concentration bounds

In this section we include a series of well known concentration bounds used in the statistical learning literature. In order to prove this bounds we will use the notion of Rademacher complexity.

Definition 7. *Given a sample $z_1, \ldots, z_m \in \mathcal{Z}$ and a class of functions G mapping $\mathcal{Z}$ to $[0,1]$, we define the empirical Rademacher complexity of G as*

$$\Re_m(G) = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\Big[\sup_{g \in G} \sum_{i=1}^{m} g(z_i)\sigma_i\Big],$$

where σ_i are i.i.d. uniform random varialbes over the set $\{-1, 1\}$.

The Rademacher complexity of a class is closely related to its VC dimension. The following Lemma can be found in (Mohri et al., 2012).

Lemma 3. *Let G be a function class with VC dimension $\mathrm{VCdim}(h) = d$ then*

$$\Re(G) \leq \sqrt{2md \log \frac{em}{d}}$$

Lemma 4. *Let L be K-Lipchitz and let $\delta > 0$. Conditioned on the choice of users belonging to the sample the following bound holds with probability at least $1 - \delta$ for for all $h \in H$*

$$\Big|\sum_j \sum_{i=1}^{n_{\tau j}} L(h(x_{ij}), y_{ij}) - \sum_j n_{\tau j} \mathcal{L}_j(h)\Big| \leq 2K\Re_{n_\tau}(H) + \sqrt{\frac{n_\tau \log \frac{1}{\delta}}{2}}$$

Proof. Relabeling the samples we notice that the left hand side of the above inequality is given by

$$\Big|\sum_{i=1}^{n_\tau} L(h(x_i), y_i) - \mathbb{E}[\sum_{i=1}^{n_\tau} L(h(x_i), y_i)]\Big|.$$

Let $H_L = \{(x, y) \mapsto L(h(x), y) | h \in H\}$, using the fact that (x_i, y_i) are independent conditioned on the choice of users and a standard learning theory bound (Mohri et al., 2012) we have with probability at least $1 - \delta$

$$\Big|\sum_{i=1}^{n_\tau} L(h(x_i), y_i) - \mathbb{E}[\sum_{i=1}^{n_\tau} L(h(x_i), y_i)]\Big| \leq \Re_{n_\tau}(H_L) + \sqrt{\frac{n_\tau \log \frac{1}{\delta}}{2}}.$$

Finally by Talagrand's contraction lemma (Mohri et al., 2012) we know that $\Re_{n_\tau}(H_L) \leq K\Re_{n_\tau}(H)$ which concludes the proof. □

Lemma 1. *Conditioned on the outcomes of $\{J_i\}$, with probability at least $1 - \delta$ the following holds uniformly over $h \in H$:*

$$\Big|\mathcal{L}_{\mathcal{S}_\tau}(h) - \sum_j \frac{n_{\tau j}}{n_\tau} \mathcal{L}_j(h)\Big| \leq \sqrt{\frac{2d \log \frac{en}{d}}{\tau_0 n}} + \sqrt{\frac{\log(1/\delta)}{2\tau_0 n}}$$

Proof. The proof follows directly from the previous proposition and a standard bound on the Rademacher complexity by the VC dimension (Mohri et al., 2012). □

Lemma 2. *Fix $\delta > 0$ and let $d = \mathrm{VCdim}(H)$. Then with probability at least $1 - \delta$, the following inequality holds uniformly for h in H.*

$$|\mathcal{L}_{\mathcal{S}_\tau}(h) - \mathcal{L}(h)| \leq \sqrt{\frac{2d \log \frac{en}{d}}{\tau_0 n}} + \sqrt{\frac{\log(2/\delta)}{2\tau_0 n}} + \Big|\sum_j \Big(\frac{n_{\tau j}}{n_\tau} - \frac{n_j}{n}\Big) \mathcal{L}_j(h)\Big| + \sqrt{\frac{\log \frac{4}{\delta}}{2n}}.$$

Proof. We begin by decomposing the loss into three parts.

$$|\mathcal{L}_{\mathcal{S}_\tau}(h) - \mathcal{L}(h)| \leq \Big|\mathcal{L}_{\mathcal{S}_\tau}(h) - \sum_j \frac{n_{\tau j}}{n_\tau} \mathcal{L}_j(h)\Big| \tag{7}$$

$$+ \Big|\sum_j \Big(\frac{n_{\tau j}}{n_\tau} - \frac{n_j}{n}\Big) \mathcal{L}_j(h)\Big| \tag{8}$$

$$+ \Big|\sum_j \Big(\frac{n_j}{n} - p_j\Big) \mathcal{L}_j(h)\Big|. \tag{9}$$

Eq. (7) is the generalization error of our empirical loss, conditioned on the outcomes of $\{J_i\}$. We bound it by applying Lemma 1 with $\frac{\delta}{2}$.

Eq. (8) is the error attributable to differences between the original dataset $\mathcal{S}$ and the thresholded data set $\mathcal{S}_\tau$; it appears directly in the bound.

Finally, Eq. (9) is the finite sample error due to the randomness in $\{J_i\}$. Observe that

$$\Big|\sum_j \Big(\frac{n_j}{n} - p_j\Big) \mathcal{L}_j(h)\Big| = \Big|\frac{1}{n} \sum_{i=1}^{n} L_{J_i}(h) - \sum_j p_j \mathcal{L}_j(h)\Big|,$$

which is just the difference between the sample mean of n i.i.d. random variables bounded in $[0,1]$ and their true mean. Hoeffding's inequality thus bounds (9) by $\sqrt{\frac{\log \frac{4}{\delta}}{2n}}$ with probability $1 - \frac{\delta}{2}$.

Combining these results under a union bound completes the proof. □

B. Bias bounds

Proposition 2. *Let r_j for $j \in \mathbb{N}$ be such that $r_j \geq 0$ and $\sum_{j=1}^n r_j = 1$. Let $0 \leq q_j \leq r_j$, $Q = \sum_j q_j$. Finally let $q'_j = \frac{q_j}{Q}$. If $|L(h,z)| \leq 1$, then the following bound holds for all hypotheses h.*

$$\Big| \sum_j \Big(q'_j - r_j\Big)\mathcal{L}_j(h)\Big| \leq \sqrt{\frac{1}{2}\log\Big(\frac{1}{Q}\Big)}$$

Proof. Using the fact that $\mathcal{L}_j(h) \leq 1$ we have

$$\Big|\sum_j (q'_j - r_j)\mathcal{L}_j(h)\Big| \leq \sum_j \Big|q'_j - r_j\Big| \tag{10}$$

Let $\mathbf{r}$ and $\mathbf{q}'$ denote the distributions induced by r_j and q'_j respectively. By Pinsker's inequality we know

$$\sum_{j=1} \Big|q'_j - r_j\Big| \leq \sqrt{\frac{1}{2}\mathrm{KL}(\mathbf{r}||\mathbf{q}')}\,,$$

where $\mathrm{KL}(\mathbf{r}||\mathbf{q}')$ denotes the Kullback-Leibler divergence between the two distributions. We can bound this divergence as follows:

$$\mathrm{KL}(\mathbf{r}||\mathbf{q}') = \frac{1}{Q}\sum_j q_j \log\Big(\frac{q_j}{Qr_j}\Big) \leq \frac{1}{Q}\sum_j q_j \log\Big(\frac{1}{Q}\Big) = \log\Big(\frac{1}{Q}\Big),$$

where we have used the fact that $q_j < r_j$ for the first inequality. Substituting this bound back in (10) yields the statement of the proposition. □

We now define a more general version of the variance term introduced in Section 6.

Definition 8. *Given a distribution $\mathbf{r}$ over $\mathbb{N}$ and a hypothesis $h \in H$ we define the variance of h with respect to $\mathbf{r}$ as*

$$\mathrm{Var}(h, \mathbf{r}) = \sum_j r_j(\mathcal{L}_j(h) - \mathcal{L}_h)^2.$$

Proposition 3. *Under the notation and assumptions of Proposition 2, the following bound holds for every h:*

$$\Big|\sum_j (q'_j - r_j)\mathcal{L}_j(h)\Big| \leq \sqrt{\frac{2\mathrm{Var}(h,\mathbf{r})}{Q}}$$

Proof. The proof relies on the simple fact that:

$$\sum_i \sum_j (\mathcal{L}_j(h) - \mathcal{L}_i(h))r_i q'_j = \sum_j \mathcal{L}_j(h)q'_j - \sum_i \mathcal{L}_i(h) r_i.$$

This is easy to verify using the fact that $\sum r_i = 1$ and $\sum q'_j = 1$. We can now apply the Cauchy-Schwarz inequality as follows:

$$\begin{aligned}
&\Big|\sum_j (q'_j - r_j)\mathcal{L}_j(h)\Big| \\
&= \Big|\sum_i \sum_j (\mathcal{L}_j(h) - \mathcal{L}_i(h))q'_j r_i\Big| \\
&= \Big|\sum_i \sum_j (\mathcal{L}_j(h) - \mathcal{L}_i(h))\sqrt{r_i r_j}\frac{q'_j}{\sqrt{r_j}}\sqrt{r_i}\Big| \\
&\leq \sqrt{\sum_i \sum_j (\mathcal{L}_j(h) - \mathcal{L}_i(h))^2 r_i r_j}\sqrt{\sum_i \sum_j \frac{(q'_j)^2}{r_j} r_i} \\
&= \sqrt{\sum_i \sum_j (\mathcal{L}_j(h) - \mathcal{L}_i(h))^2 r_i r_j}\sqrt{\sum_j \frac{(q'_j)^2}{r_j}}
\end{aligned}$$

A simple calculation shows that the first term in the above expression is in fact equal to $2\mathrm{Var}(h,\mathbf{r})$. Therefore we need only to prove that the second term is bounded by $\frac{1}{Q}$. We have

$$\begin{aligned}
\sum_j \frac{(q'_j)^2}{r_j} &= \frac{1}{Q^2}\sum_j \frac{q_j^2}{r_j} \\
&\leq \frac{1}{Q^2}\sum_j q_j = \frac{1}{Q},
\end{aligned}$$

where we used the fact that $q_j \leq r_j$. □

The proof of Proposition 1 is easily derived from Propositions 2 and 3. Indeed, letting $r_j = \frac{n_j}{n}$ and $q_j = \frac{n_{j\tau}}{n}$ we have $q_j \leq r_j$, and thus the result follows.

C. Additional bounds

Proposition 4. *Let $\tau \leq n$ be the cap on user contributions. Then $n_\tau > \tau$.*

Proof. There are only two possibilities: either $n_j < \tau$ for all j or $n_j \geq \tau$ for some j. In the latter case $n_\tau \geq n_j = \tau$ by definition. On the other hand, if $n_j < \tau$ for all j then

$$n_\tau = \sum_j n_{j\tau} = \sum_j n_j = n \geq \tau.$$

□

Proposition 5. *Let $1 > \tau_0 > 0$ and $\tau = \tau_0 n$. Let $K(\tau_0) = |\{j \mid p_j > \tau_0\}|$ and let $\delta > 0$. With probability at least $1 - \delta$,*

$$\frac{n_\tau}{n} \geq \frac{\tau_0 K(\tau_0)}{4} - \sqrt{\frac{\log(1/\delta)}{2n}}\,.$$

Proof. Recall that J_i is the random variable that denotes the user corresponding to example i. We know that $n_j = \sum_{i=1}^n \mathbb{1}_{J_i=j}$ and $n_\tau = \sum_{i=1}^n \min(n_i, \tau)$. Let $\phi(J_1, \dots, J_n) = \frac{n_\tau}{n}$. We want to bound the change in ϕ as we perturb a single coordinate:

$$|\phi(J_1, \dots, J_n) - \phi(J_1', \dots, J_n)|.$$

If we change only one point in the sample then, clearly, we change the contribution of at most two users i_1 and i_2. Let n'_{i_1} and n'_{i_2} denote the user contributions under the perturbation. Then the above expression is equal to

$$\frac{1}{n}|\min(n_{i_1},\tau)-\min(n'_{i_1},\tau)+\min(n_{i_2},\tau)-\min(n'_{i_2},\tau)|. \tag{11}$$

Let us assume w.l.o.g. that $n_{i_1} \geq n'_{i_1}$; this implies that $n_{i_2} \leq n'_{i_2}$. Therefore $0 \leq \min(n_{i_1}, \tau) - \min(n'_{i_1}, \tau) \leq 1$ and $0 \geq \min(n_{i_2}, \tau) - \min(n'_{i_2}, \tau) \geq -1$. This readily implies that (11) is bounded by $\frac{1}{n}$. We can now apply McDiarmid's inequality and see that for any $\eta > 0$

$$P\Big(\frac{n_\tau}{n} \leq \frac{1}{n}\mathbb{E}[n_\tau] - \eta\Big) \leq e^{-2n\eta^2}. \tag{12}$$

Now let $Q(\tau_0) = \sum_{j=1}^n \min(p_j, \tau_0)$. It is easy to see that

$$Q(\tau_0) = \sum_{j:p_j>\tau_0} \tau_0 + \sum_{j:p_j\leq\tau_0} p_j \geq K(\tau_0).$$

Therefore from Corollary 2 we know that

$$\begin{aligned} P\Big(\frac{n_\tau}{n} \leq \frac{\tau_0 K(\tau_0)}{4} - \eta\Big) &\leq P\Big(\frac{n_\tau}{n} \leq \frac{Q(\tau_0)}{4} - \eta\Big) \\ &\leq P\Big(\frac{n_\tau}{n} \leq \frac{1}{n}\mathbb{E}[n_\tau] - \eta\Big) \end{aligned}$$

The result follows from (12) by setting $\delta = e^{-2n\eta^2}$ and solving for η. □

Lemma 2. *Let $S_n = \sum_{i=1}^N X_i$ be a sum of i.i.d. Bernoulli random variables with $P(X_i = 1) = p$. Then*

$$\mathbb{E}[\min(S_n, \tau)] \geq \frac{1}{4}\min(pn, \tau) \tag{13}$$

Proof. First let us assume that $\tau < np$ in that case we have:

$$\begin{aligned} \mathbb{E}[\min(S_n, \tau)] &= \mathbb{E}[S_n \mathbb{1}_{S_n<\tau}] + \tau P(S_n > \tau) \\ &\geq \tau P(S_n > \tau) \\ &\geq \tau P(S_n > np) \geq \frac{\tau}{4}, \end{aligned}$$

where we used the fact that $P(S_n > np) > \frac{1}{4}$ (Greenberg & Mohri, 2013; Vapnik, 1998).

On the other hand if $\tau > np$ then

$$\begin{aligned} \mathbb{E}[\min(S_n, \tau)] \geq \mathbb{E}[S_n \mathbb{1}_{S_n<\tau}] &\geq \mathbb{E}[S_n \mathbb{1}_{S_n>np}] \\ &= \int_0^\infty P(S_n \mathbb{1}_{S_n>np} > t)dt \\ &= \int_0^{np} P(S_n > t)dt \\ &\geq \int_0^{np} P(S_n > np)dt \\ &\geq \frac{1}{4}np \end{aligned}$$

Combining the two cases yields the statement of the proposition. □

Corollary 2. *Let J_k, $k = 1, \dots, n$ be a random variable in $\mathbb{N}$ such that $P(J_k = j) = p_j$. Let $n_j = \sum_{i=1}^n \mathbb{1}_{J_k=j}$, $\tau_0 > 0$ and $\tau = \tau_0 n$. Finally, let $n_\tau = \sum_j \min(n_j, \tau)$; then we have*

$$\frac{1}{n}\mathbb{E}[n_\tau] \geq \frac{1}{4}\sum_j \min(p_j, \tau_0)$$

Proof. By Fubini's theorem,

$$\mathbb{E}[n_\tau] = \mathbb{E}[\sum_j \min(n_j, \tau)] = \sum_j \mathbb{E}[\min(n_j, \tau)] \,.$$

On the other hand, n_j is a sum of independent Bernoulli random variables with probability p_j. So from the previous proposition we have

$$\begin{aligned} \frac{1}{n}\sum_j \mathbb{E}[\min(n_j, \tau)] &\geq \frac{1}{4n}\sum_j \min(p_j n, \tau) \\ &= \frac{1}{4}\sum_j \min(p_j, \tau_0) \end{aligned}$$

□