Bias Variance Trade-off in Differential Privacy

Kareem Amin
Google Research NY
kamin@google.com

Alex Kulesza
Google Research NY
kulesza@google.com

Andrés Muñoz Medina
Google Research NY
ammedina@google.com

Sergei Vassilvitskii
Google Research NY
sergeiv@google.com

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 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. 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 tradeoff for an empirical risk minimization setting, showing that in general there is a "sweet spot" that depends on measurable properties of the dataset.

1 Introduction

Differential privacy [Dwork and 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. While a variety of mechanisms for generating differentially private algorithms are now known—and, increasingly, used in practice—significant challenges remain.

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. 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 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.

When the sensitivity is high, practitioners sometimes compensate by raising ϵ, but this results in reduced privacy protection. Here, we advocate a common alternative approach: limiting the contributions of individual users in order to reduce the sensitivity. A fundamental question is how to

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

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 empricial 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 they are often intuitive and can plausibly be measured or approximated using prior information.

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 collection 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 $U \subseteq H$,*

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

The noise added by a differentially private algorithm is typically calibrated using *sensitivity*.

Definition 3. *The (ℓ_1) sensitivity of a function $f : \mathbb{S} \to \mathbb{R}$ is given by $\Delta_f = \sup_{\mathcal{S} \sim \mathcal{S}'} |f(\mathcal{S}) - f(\mathcal{S}')|$.*

Definition 4 (Laplace mechanism [Dwork et al., 2006])**.** *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 $\exp(-\epsilon|\eta|/\Delta_f)$. The Laplace mechanism is ϵ-differentially private.*

The Laplace mechanism applies for any real-valued function f. A somewhat more general technique for constructing differentially private algorithms is the exponential mechanism [McSherry and Talwar, 2007]. The exponential mechanism is parameterized by a utility function u, where $u_{\mathcal{S}}(h) \in \mathbb{R}$ denotes the utility of hypothesis h (which need not be numeric) on dataset $\mathcal{S}$. Sensitivity of u is measured with respect to the dataset, maximized over all hypotheses:

$$\Delta_u = \sup_{h \in H} \sup_{\mathcal{S} \sim \mathcal{S}'} |u_{\mathcal{S}}(h) - u_{\mathcal{S}'}(h)| \ . \tag{1}$$

Definition 5 (Exponential mechanism)**.** *Let u be a utility function, and fix $\epsilon > 0$. The exponential mechanism $\mathrm{Exp}_{u,\epsilon}(\mathcal{S})$ returns a hypothesis $h \in H$ with probability proportional to $\exp\left(\frac{\epsilon u_{\mathcal{S}}(h)}{2\Delta_u}\right)$. The exponential mechanism is also ϵ-differentially private.*

In both mechanisms, sensitivity controls the noise level. In practice, it often is possible to explicitly limit sensitivity by introducing some bias into the utility function; here we are interested in understanding that trade-off when the utility function is an empirical loss measure.

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 nonnegative 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.

Naïvely, we might try to do this by applying the Laplace mechanism to the function $\mu(\mathcal{S}) = \sum_{i=1}^{n} x_i$. But there is a problem: since a single user can contribute an arbitrarily large value, the sensitivity of μ, and therefore the scale of the noise, is infinite. To fix this, we will introduce a cap τ on the maximum size of a user's contribution, instead applying the Laplace mechanism to the function

$\mu_\tau(\mathcal{S}) = \sum_{i=1}^{n} \min(x_i, \tau)$. This will bias our estimated sum, of course, but it also reduces the amount of added noise, as the sensitivity is now τ.

So how should we choose τ? We can decompose the expected error of the estimate $\hat{\mu}$ produced by $\text{Lap}_{\mu_\tau,\epsilon}(\mathcal{S})$ into a variance term (due to the noise) and a bias term (due to the contribution limit):

$$\mathop{\mathbb{E}}_{\hat{\mu}} |\hat{\mu} - \mu(\mathcal{S})| = \Delta_{\mu_\tau}/\epsilon + |\mu_\tau(\mathcal{S}) - \mu(\mathcal{S})| = \tau/\epsilon + \sum_{i=1}^{n} \max(0, x_i - \tau) \; . \tag{2}$$

The minimum is achieved when τ is equal to the $\lceil 1/\epsilon \rceil$th largest value in $\mathcal{S}$. Note that it does not matter how large or small the contributions are above or below the cutoff. This is important given that any information about the dataset used to determine τ must itself be computed privately. Luckily, there are a variety of differentially private algorithms that can be applied to approximating quantiles [Nissim et al., 2007, Dwork and Lei, 2009, Smith, 2011]. We will see in Section 4.1 that a similarly intuitive statistic also appears in the more general setting.

4 A Generalization Bound

Consider an infinite set of users identified by the natural numbers. We consider a data generation process that proceeds in rounds; on each round $i \leq n$, a user $J_i \in \mathbb{N}$ is drawn from the fixed *participation distribution* μ. Each user $j \in \mathbb{N}$ is equipped with its own *data distribution* D^j over some example space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. The example $z_i = (x_i, y_i)$ generated on the ith round is an iid sample from D^{J_i}. Let D denote the resulting mixture distribution, and $S_n = \{z_1, \cdots, z_n\}$ denote a sample of size n. Observe that since draws from the participation and data distribution are each independent, the z_i are iid.

Let $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be a loss function measuring the quality of the predication made by a hypothesis $h \in H$, and denote the true risk by $\mathcal{L}(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[L(h(x), y)]$. We assume that L is bounded, and for simplicity that bound is $||L||_\infty < 1$.We are also interested in the the true risk on just the ith user's distribution, which we denote by $\mathcal{L}_i(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}^i}[L(h(x), y)]$. Given an arbitrary dataset S containing examples from $\mathcal{Z}$, the empirical loss of h with respect to S is defined as $\mathcal{L}_S(h) = \frac{1}{|S|} \sum_{(x,y)\in S} L(h(x), y)$.

Algorithm. A differentially private learning algorithm maps S_n to a choice of hypothesis h_{priv} in a manner that satisfies Definition 2. Our goal is to approximately optimize the expected loss $\mathcal{L}(h)$. We will follow the work of Bassily et al. [2014] and use the exponential mechanism on the utility function $u_S(h) = -\mathcal{L}_S(h)$. The authors show that this algorithm can be efficiently implemented and that it has near optimal accuracy. Nevertheless, notice that since our model could in principle allow for a user to contribute all n samples, the sensitivity of the utility function is $\Delta_u = 1$, rendering any hypothesis output by this mechanism useless.

To reduce the impact of noise we cap the contribution of every user to τ. More precisely, given n examples from D, let $n_j = \sum_{i=1}^{n} \mathbf{1}[J_i = j]$ be the number of examples contributed by user j. Let $n_{\tau j} = \min\{\tau, n_j\}$ denote the number of examples contributed by user j after truncation, and let $n_\tau = \sum_{j\in\mathbb{N}} n_{\tau j}$. Given a sample S_n, let S_τ contain, the first $n_{\tau j}$ examples in S_n generated by user j. Our algorithm will sample h according to the exponential mechanism with utility function $u_{S_\tau} = -\mathcal{L}_{S_\tau}$.

Notice that since we are truncating each user contribution to τ the sensitivity of our utility function Δ_{u_τ} is given by $\frac{\tau}{n_\tau}$. The following lemma from [Bassily et al., 2014] provides a bound on the empirical error of the hypothesis returned by our algorithm when $H = B^d(0, 1)$, the ball in $\mathbb{R}^d$ of radius 1. Similar guarantees can be given for hypothesis sets with a finite number of hypotheses, and we conjecture that the extension to general VC classes has the same dependency on τ.

Lemma 1. *Let $h_{\text{priv}} \in B^d(0, 1)$ be a hypothesis sampled proportional to $\exp\Big(-\frac{\epsilon \mathcal{L}_{S_\tau}(h)}{2\tau}\Big)$, and h_τ the hypothesis minimizing $\mathcal{L}_{S_\tau}$. Then*

$$\mathbb{E}[\mathcal{L}_{S_\tau}(h_{\text{priv}})] \leq \mathcal{L}_{S_\tau}(h_\tau) + \frac{8\tau}{n_\tau \epsilon}\Big((d+1)\log(3) + \log\Big(\frac{1}{\delta}\Big)\Big), \tag{3}$$

where the expectation is taken only over the randomness of the exponential mechanism.

The lemma shows that in order for our hypothesis to be close to the optimal empirical risk minimizer τ needs to be small. However, truncating the user contribution too much will bias the empirical error.

For interpretability, assume $\tau = \gamma n$ for some $\gamma \in [0, 1]$. Let $p_j = \mathbb{P}[J = j]$, where J follows the participation distribution μ, and let $q_j = \min\{\gamma, p_j\}$. Our first theorem bounds $|\mathcal{L}_{S_\tau}(h) - \mathcal{L}(h)|$ uniformly for all $h \in H$. This bound includes a bias term introduced by the fact that we are thresholding a user's contribution by τ and consequently S_τ is not an iid sample from D. It also contains a variance term introduced by both standard finite sample effects and also the fact that thresholding results in further data loss as the sample size is decreased from n to n_τ.

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

$$|\mathcal{L}_{\mathcal{S}_\tau}(h) - \mathcal{L}(h)| \leq O\left(\frac{1}{\gamma^2}\sqrt{\frac{d \log \frac{n}{d}}{n}}\right) + \Big|\sum_j \Big(\frac{q_j}{\sum q_j} - p_j\Big)\mathcal{L}_j(h)\Big| \tag{4}$$

The proof is available in the supplement. Recall that instead of minimizing $\mathcal{L}_{S_\tau}$ directly we are sampling h_{priv} according to the exponential mechanism. Two applications of the previous theorem as well as Lemma 1 allow us to state the following corollary for linear hypothesis in $B^d(0, 1)$.

Corollary 1. *Let $\delta > 0$ and $H = B^d(0, 1)$. Let $\eta = \sqrt{\frac{4 \log(n/\delta)}{n}}$. Let $K(\gamma) = |\{i : p_i > \gamma + \eta\}|$; then with probability at least $1 - \delta$ the following inequality holds uniformly for $h \in H$.*

$$\mathbb{E}[\mathcal{L}(h_{\mathrm{priv}})] \leq \inf_{h \in H} \mathcal{L}(h) + \underbrace{\sup_{h \in H} \Big|\sum_j \Big(\frac{q_j}{\sum q_j} - p_j\Big)\mathcal{L}_j(h)\Big|}_{\textit{bias term}} + \underbrace{O\left(\frac{1}{\gamma^2}\sqrt{\frac{d \log \frac{n}{d}}{n}}\right)}_{\textit{finite sample variance}} + \underbrace{O\left(\frac{d}{K(\gamma)\epsilon}\right)}_{\textit{privacy variance}}.$$

This bound shows that if the bias is small, depending on how fast $K(\gamma)$ decreases, we can obtain non-trivial error bounds. For instance, suppose the distribution of users is uniformly supported on $N = O(n^{1/6})$ users; then we can set $\gamma \approx 1/N$, making $K(\gamma) = N$, and the previous bound becomes

$$\mathbb{E}[\mathcal{L}(h_{\mathrm{priv}})] \leq \inf_{h \in H} \mathcal{L}(h) + 2 \sup_{h \in H} \Big|\sum_j \Big(\frac{q_j}{\sum q_j} - p_j\Big)\mathcal{L}_j(h)\Big| + O\left(\sqrt{\frac{d \log \frac{n}{d}}{n^{1/3}}}\right) + O\left(\frac{d}{n^{1/6}\epsilon}\right).$$

4.1 Understanding the bias

The bias term $\Big|\sum_j \Big(\frac{q_j}{\sum_j q_j} - p_j\Big)\mathcal{L}_j(h)\Big|$ does not depend on the sample size n and therefore does not vanish as $n \to \infty$. The bias can be seen as the difference of expected losses under two different distributions, allowing us to apply bounds from the domain adaptation literature, e.g., using the $d_{H\Delta H}$-distance [Blitzer et al., 2007] or the $\mathcal{Y}$-discrepancy [Cortes et al., 2015]. One of the advantages of using the $d_{H\Delta H}$-distance is that it is estimatable from unlabeled samples.

However, ultimately we must trade off the bias error with the privacy variance, which is on the order of $O(\frac{\gamma}{\epsilon})$. To do so, we need to better understand how the bias varies as a function of γ; to that end we provide upper bounds on the bias that depend on simple statistical properties of the data distribution.

Proposition 1. *For all hypotheses h:* $\Big|\sum_j \Big(\frac{q_j}{\sum_j q_j} - p_j\Big)\mathcal{L}_j(h)\Big| \leq \sqrt{\frac{1}{2}\log\Big(\frac{1}{\gamma}\Big)}$.

Using the fact that $\log(1/\gamma) \leq \frac{1-\gamma}{\gamma}$ we can conclude that truncating user contribution to $\tau = \gamma n$ induces a bias that is $O(\sqrt{\frac{1-\gamma}{\gamma}})$. However, this can be overly pessimistic. In particular, if $\mathcal{L}_(h)$ is constant (that is, the expected error is the same for all users), then the bias is zero for any γ. This suggests that the bias should be bounded in terms of a measure of the spread in the losses $\mathcal{L}_j(h)$.

Definition 6. *For any hypothesis h, we define the variance of its loss across users by* $\mathrm{Var}(h) = \sum_j (\mathcal{L}_j(h) - \mathcal{L}(h))^2 p_j$.

Proposition 2. *The following bound holds for every h:* $\Big|\sum_j \Big(\frac{q_j}{\sum_j q_j} - p_j\Big)\mathcal{L}_j(h)\Big| \leq \sqrt{\frac{2\mathrm{Var}(h)}{\gamma}}$

Notice that this bound correctly captures the fact that if $\mathcal{L}_j(h)$ is constant then the bias is zero. It also shows that if the distribution between users does not differ wildly, then by Corollary 1 we can achieve non-trivial learning guarantees. Like the quantiles in Section 3, the variance of the losses across users is an intuitive quantity that can plausibly be measured or approximated by prior knowledge.

References

R. Bassily, A. D. Smith, and A. Thakurta. 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*, pages 464–473, 2014.

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

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

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

K. Crammer, M. Kearns, and J. Wortman. Learning from multiple sources. *Journal of Machine Learning Research*, 9(Aug):1757–1774, 2008.

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

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

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

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

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

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

F. McSherry and K. Talwar. Mechanism design via differential privacy. In *Foundations of Computer Science, 2007. FOCS'07. 48th Annual IEEE Symposium on*, pages 94–103. IEEE, 2007.

K. Nissim, S. Raskhodnikova, and A. D. Smith. 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*, pages 75–84, 2007.

A. Pyrgelis, C. Troncoso, and E. D. Cristofaro. 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.

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