Proceedings of Machine Learning Research vol 75:1–10, 2018 31st Annual Conference on Learning Theory

Calibrating Noise to Variance in Adaptive Data Analysis

Vitaly Feldman VITALY@POST.HARVARD.EDU
Google Brain

Thomas Steinke ALKLS@THOMAS-STEINKE.NET
IBM Research – Almaden

Editors: Sebastien Bubeck, Vianney Perchet and Philippe Rigollet

Abstract

Datasets are often used multiple times and each successive analysis may depend on the outcome of previous analyses. Standard techniques for ensuring generalization and statistical validity do not account for this adaptive dependence. A recent line of work studies the challenges that arise from such adaptive data reuse by considering the problem of answering a sequence of "queries" about the data distribution where each query may depend arbitrarily on answers to previous queries.

The strongest results obtained for this problem rely on differential privacy – a strong notion of algorithmic stability with the important property that it "composes" well when data is reused. However the notion is rather strict, as it requires stability under replacement of an arbitrary data element. The simplest algorithm is to add Gaussian (or Laplace) noise to distort the empirical answers. However, analysing this technique using differential privacy yields suboptimal accuracy guarantees when the queries have low variance.

Here we propose a relaxed notion of stability based on KL divergence that also composes adaptively. We show that our notion of stability implies a bound on the mutual information between the dataset and the output of the algorithm and then derive new generalization guarantees implied by bounded mutual information. We demonstrate that a simple and natural algorithm based on adding noise scaled to the standard deviation of the query provides our notion of stability. This implies an algorithm that can answer statistical queries about the dataset with substantially improved accuracy guarantees for low-variance queries. The only previous approach that provides such accuracy guarantees is based on a more involved differentially private median-of-means algorithm and its analysis exploits stronger "group" stability of the algorithm.

1. Introduction

The central challenge in statistical data analysis is to infer the properties of some unknown population given only a small number of samples from that population. While a plethora of techniques for guaranteeing statistical validity are available, few techniques can account for the effects of *adaptivity*. Namely, if a single dataset is used multiple times, then the choice of which subsequent analyses to perform may depend on the outcomes of previous analyses. This adaptive dependence increases the risk of overfitting — that is, inferring a conclusion that does not generalize to the underlying population.

0. Extended abstract. Full version appears as (Feldman and Steinke, 2017b, v2).

To formalize this problem, Dwork et al. (2014) and subsequent works (Hardt and Ullman, 2014; Steinke and Ullman, 2015; Bassily et al., 2016; Feldman and Steinke, 2017a, etc.) study the following question: How many data samples are necessary to accurately answer a sequence of queries about the data distribution when the queries are chosen adaptively – that is, each query can depend on answers to previous queries? Each query corresponds to a procedure that the analyst wishes to execute on the data. The goal is to design an algorithm that provides answers to these adaptive queries that are close to answers that would have been obtained had each corresponding analysis been run on independent samples freshly drawn from the data distribution.

A common and relatively simple class of queries are statistical queries (Kearns, 1998). A statistical query is specified by a function $\psi : \mathcal{X} \to [0,1]$ and corresponds to analyst wishing to compute the true mean $\mathbf{E}_{X\sim\mathcal{P}}[\psi(X)]$ of ψ on the data distribution $\mathcal{P}$. (This is usually done by using the empirical mean $\frac{1}{n}\sum_{i=1}^{n}\psi(S_i)$ on a dataset S consisting of n i.i.d. draws from the distribution $\mathcal{P}$.) For example, such queries can be used to to estimate the true loss (or error) of a predictor, the gradient of the loss function, or the moments of the data distribution. Standard concentration results imply that, given n independent samples from $\mathcal{P}$, k fixed (i.e. not adaptively-chosen) statistical queries can be answered with an additive error of at most $O\left(\sqrt{\log(k)/n}\right)$ with high probability by simply using the empirical mean of each query. At the same time it is not hard to show that, for a variety of simple adaptive sequences of queries, using the empirical mean to estimate the expectation leads to an error of $\Omega(\sqrt{k/n})$ (Dwork et al., 2014). Equivalently, in the adaptive setting, the number of samples required to ensure fixed error scales linearly (rather than logarithmically in the non-adaptive setting) with the number of queries and, in particular, in the worst case, using empirical estimates gives the same guarantees as using fresh samples for every query (by splitting the dataset into k parts).

Dwork et al. (2014) showed that, remarkably, it is possible to quadratically improve the dependence on k in the adaptive setting by simply perturbing the empirical answers. Specifically, let $S \in \mathcal{X}^n$ denote a dataset consisting of n i.i.d. samples from some (unknown) probability distribution $\mathcal{P}$. Given S, the algorithm receives k adaptively-chosen statistical queries $\psi_1, \ldots, \psi_k : \mathcal{X} \to [0,1]$ one-by-one and provides k approximate answers $v_1, \ldots, v_k \in \mathbb{R}$. Namely, $v_j = \frac{1}{n}\sum_{i=1}^{n}\psi_j(S_i) + \xi_j$, where each "noise" variable ξ_j is drawn independently from $\mathcal{N}(0,\sigma^2)$. The results of Dwork et al. (2014) and subsequent sharper analyses (Bassily et al., 2016; Steinke, 2016) show that, with high probability (over the drawing of the sample $S \sim \mathcal{P}^n$, the noise ξ, and the choice of queries), we have the following guarantee

$$\forall j \in \{1, \cdots, k\} \qquad \left| v_j - \mathop{\mathbf{E}}_{X\sim\mathcal{P}}[\psi_j(X)] \right| \leq O\left(\sqrt{\frac{\sqrt{k}\log k}{n}}\right). \tag{1}$$

This quadratic relationship between n and k was also shown to be optimal in the worst case (Hardt and Ullman, 2014; Steinke and Ullman, 2015).

The approach of Dwork et al. (2014) relies on properties of differential privacy (Dwork et al., 2006a,b) and known differentially private algorithms. Differential privacy is a stability property of an algorithm, namely it requires that replacing any element in the input dataset results in a small change in the output distribution of the algorithm. Specifically, a randomized algorithm $M : \mathcal{X}^n \to \mathcal{Y}$ is (ε, δ)-differentially private if, for all datasets $s, s' \in \mathcal{X}^n$

that differ on a single element and all events $E \subseteq \mathcal{Y}$,

$$\mathbf{Pr}\left[M(s) \in E\right] \leq e^{\varepsilon} \mathbf{Pr}\left[M(s') \in E\right] + \delta.$$

This stability notion implies that a function output by a differentially private algorithm on a given dataset generalizes to the underlying distribution (Dwork et al., 2014; Bassily et al., 2016). Specifically, if a differentially private algorithm is run on a dataset drawn i.i.d from any distribution and the algorithm outputs a function, then the empirical mean of that function on the input dataset is close to the expectation of that function on sample from the same distribution.

The second crucial property of differential privacy is that it composes adaptively: running several differentially private algorithms on the same dataset still is differentially private (with somewhat worse parameters) even if each algorithm depends on the output of all the previous algorithms. This property makes it possible to answer adaptively-chosen queries with differential privacy and a number of algorithms have been developed for answering different types of queries. The generalization property of differential privacy then implies that such algorithms can be used to provide answers to adaptively-chosen queries while ensuring generalization (Dwork et al., 2014). Specifically, the algorithm for answering statistical queries mentioned above is based on the most basic differentially private algorithm: perturbation by adding Laplace or Gaussian noise (Dwork et al., 2006a).

Differential privacy requires that the output distribution of an algorithm does not change much when any element of a dataset is replaced with an arbitrary other element in the domain $\mathcal{X}$. As a result, the amount of noise that needs to be added to ensure differential privacy scales linearly with the range of the function ψ whose expectation needs to be estimated. If the range of ψ is comparable to the standard deviation of $\psi(x)$ on x drawn from $\mathcal{P}$ (such as when ψ has range $\{0,1\}$ and mean $1/2$) then the error resulting from addition of noise is comparable to the standard deviation of ψ. However, for queries whose standard deviation is much lower than the range, the error introduced by noise is much worse than the sampling error. Variance is much smaller than the range for a variety of common settings, for example, difference between candidate predictors for the same problem or individual input features when the input is usually sparse.

Achieving error guarantees in the adaptive setting that scale with the standard deviation instead of range is a natural problem. Recently, Feldman and Steinke (2017a) gave a different algorithm that achieves such a guarantee. Specifically, their algorithm ensures that with probability at least $1-\beta$,

$$\forall j \in \{1, \cdots, k\} \qquad \left| v_j - \mathop{\mathbf{E}}_{X \sim \mathcal{P}}\left[\psi_j(X)\right] \right| \leq \mathrm{sd}(\psi_j(\mathcal{P})) \cdot O\left(\sqrt{\frac{\sqrt{k \log^3(k/\beta)}}{n}}\right) + \beta, \qquad (2)$$

where $\mathrm{sd}(\psi_j(\mathcal{P})) = \sqrt{\mathbf{E}_{Y \sim \mathcal{P}}\left[(\psi_j(Y) - \mathbf{E}_{X \sim \mathcal{P}}\left[\psi_j(X)\right])^2\right]}$ is the standard deviation of ψ_j on the distribution $\mathcal{P}$ and $\beta > 0$ can be chosen arbitrarily. Their algorithm is based on an approximate version of the median of means algorithm and its analysis still relies on differential privacy. (Their results extend beyond statistical queries, but we restrict our attention to statistical queries in this paper.)

In this work, we ask: does the natural algorithm that perturbs the empirical answers with noise scaled to the standard deviation suffice to answer adaptive queries with accuracy scaling to sampling error? To answer this seemingly simple question, we address a more fundamental problem: does there exist a notion of stability that has the advantages of differential privacy (namely, allows adaptive composition and implies generalization) but avoids the poor dependence on the worst-case sensitivity of the query. This algorithm was analyzed by Bassily and Freund (2016) via a notion of typical stability they introduced. Their analysis shows that the algorithm will ensure the correct scaling of the error with standard deviation but it does not improve on the naive mechanisms in terms of scaling with k. Several works have considered relaxations of differential privacy in this context. For example, Bassily et al. (2016) considered a notion of stability based on using KL divergence or total variation distance in place of differential privacy (which can be defined in terms of approximate max divergence). Wang et al. (2016) considered the expected KL divergence between the output of the algorithm when run on a random i.i.d dataset versus the same dataset with one element replaced by a fresh sample; unfortunately, their stability definition does *not* compose adaptively. Notions based on the mutual information between the dataset and the output of the algorithm and their relationship to differential privacy have also been studied (Dwork et al., 2015; Russo and Zou, 2016; Rogers et al., 2016; Raginsky et al., 2016; Xu and Raginsky, 2017). However, to the best of our knowledge, these approaches do not give a way to analyze the calibrated noise addition that ensures correct dependence on k.

1.1. Our Contributions

We introduce new stability-based and information-theoretic tools for analysis of the generalization of algorithms in the adaptive setting. The stability notion we introduce is easier to satisfy than differential privacy, yet has the properties crucial for application in adaptive data analysis. These tools allow us to demonstrate that calibrating the variance of the perturbation to the empirical variance of the query suffices to ensure generalization, as long as the noise rate does not become too small. To ensure this lower bound on the noise rate we simply add a second order term to the variance of the perturbation. Specifically, our algorithm is described in Figure 1. The only difference between our algorithm and previous work (Dwork et al., 2014; Bassily et al., 2016) is that in prior work the variance of the Gaussian perturbation is fixed.

We prove that this algorithm has the following accuracy guarantee.

Theorem 1.1 (Main Theorem) *Let $\mathcal{P}$ be a distribution on $\mathcal{X}$ and let M be our algorithm from Figure 1 instantiated with $T = n^2/k$ and $t = n\sqrt{2\ln(2k)/k}$. Suppose M is given a sample $S \sim \mathcal{P}^n$ and is asked adaptive statistical queries $\psi_1, \cdots, \psi_k : \mathcal{X} \to [0,1]$. Then M produces answers $v_1, \cdots, v_k \in \mathbb{R}$ satisfying the following.*

$$\mathbf{E}\left[\max_{j=1}^{k} \frac{|v_j - \mathbf{E}_{X\sim\mathcal{P}}[\psi_j(X)]|}{\max\{\tau \cdot \mathrm{sd}(\psi_j(\mathcal{P})), \tau^2\}}\right] \leq 4, \qquad \textit{where} \qquad \tau = \sqrt{\frac{\sqrt{2k\ln(2k)}}{n}}.$$

Intuitively (that is, ignoring the second term in the maximum), the conclusion of Theorem 1.1 states that, with good probability, the error in each answer scales as the standard deviation of the query multiplied by $\tilde{O}\left(\sqrt{\sqrt{k}/n}\right)$ — which is what would be expected if we used

Parameters: $t, T > 0$.
Input: $s \in \mathcal{X}^n$.
For $j = 1, 2, \cdots, k$ do:
  Receive a statistical query $\psi_j : \mathcal{X} \to [0,1]$.
  Compute $\mu_j = \frac{1}{n}\sum_{i=1}^n \psi_j(s_i)$ and $\sigma_j^2 = \frac{1}{n}\sum_{i=1}^n (\psi_j(s_i) - \mu_j)^2$.
  Sample $\xi_j \sim \mathcal{N}(0,1)$.
  Let $v_j = \mu_j + \xi_j \cdot \sqrt{\max\left\{\sigma_j^2/t, 1/T\right\}}$.
  Output answer v_j.

Figure 1: Calibrating noise to variance for answering adaptive queries.

$n/\sqrt{k}$ fresh samples for each query. The $\ln k$ factor arises from the fact that we take a union bound over the k queries.

More precisely, applying Markov's inequality to the conclusion of Theorem 1.1, shows that, with probability at least 90%,

$$\forall j \quad \left|v_j - \mathop{\mathbf{E}}_{X\sim\mathcal{P}}[\psi_j(X)]\right| \le 40 \cdot \max\left\{\tau \cdot \mathrm{sd}(\psi_j(\mathcal{P})), \tau^2\right\} \le \mathrm{sd}(\psi_j(\mathcal{P})) \cdot 40\sqrt{\frac{\sqrt{2k\ln(2k)}}{n}} + 40\frac{\sqrt{2k\ln(2k)}}{n}. \tag{3}$$

This guarantee is directly comparable to the earlier bound (2) of Feldman and Steinke (2017a) – though it is weaker in two ways: First, Theorem 1.1 is a bound on the expectation and does not readily yield high probability bounds (other than via Markov's inequality). Second, the second term in the maximum (which we think of as a low-order term) still depends linearly on the sensitivity and is potentially larger. The advantage of this algorithm is that it is substantially simpler than the earlier work.

Now we turn to the analysis tools that we introduce. Clearly the *empirical error* of our algorithm — that is $|v_j - \mu_j|$ — scales with the empirical standard deviation σ_j. However, we must bound the *true error*, namely $|v_j - \mathbf{E}_{X\sim\mathcal{P}}[\psi_j(X)]|$. By the triangle inequality, it suffices to bound the *generalization error* $|\mu_j - \mathbf{E}_{X\sim\mathcal{P}}[\psi_j(X)]|$ in terms of standard deviation and to relate the empirical standard deviation σ_j to the true standard deviation $\mathrm{sd}(\psi_j(\mathcal{P}))$.

1.1.1. Average leave-one-out KL stability and generalization

The key to our analysis is the following stability notion.

Definition 1.2 (Average Leave-one-out KL stability) *An algorithm $M : (\mathcal{X}^n \cup \mathcal{X}^{n-1}) \to \mathcal{Y}$ is ε-ALKL stable if, for all $s \in \mathcal{X}^n$,*

$$\frac{1}{n}\sum_{i\in[n]} \mathrm{D}\left(M(s) \| M(s_{-i})\right) \le \varepsilon,$$

where $s_{-i} \in \mathcal{X}^{n-1}$ denotes s with the i^{th} element removed. Here $\mathrm{D}(\cdot\|\cdot)$ denotes the Kullback-Leibler divergence.

Our notion differs from differential privacy in three significant ways.[1] First, we use stability to leaving one out (LOO) rather than replacing one element. Second, we average the stability parameter across the n dataset elements. Third, we use KL divergence instead of (approximate) max divergence. This is necessary to obtain stronger bounds for our calibrated noise addition as our algorithm does not satisfy differential privacy with parameters that would be suitable to ensure generalization. We note that average LOO stability is a well-studied way to define algorithmic stability for the loss function (*e.g.* , (Bousquet and Elisseeff, 2002; Poggio et al., 2004)). The use of KL divergence appears to be necessary to ensure adaptive composition of our averaged notion. Specifically, the following composition result is easy to prove.

Lemma 1.3 (Composition) *Suppose $M : (\mathcal{X}^n \cup \mathcal{X}^{n-1}) \to \mathcal{Y}$ is ε-ALKL stable and $M' : \mathcal{Y} \times (\mathcal{X}^n \cup \mathcal{X}^{n-1}) \to \mathcal{Z}$ is such that $M'(y, \cdot) : (\mathcal{X}^n \cup \mathcal{X}^{n-1}) \to \mathcal{Z}$ is ε'-ALKL stable for all $y \in \mathcal{Y}$. Then the composition $s \mapsto M'(M(s), s)$ is $(\varepsilon + \varepsilon')$-ALKL stable.*

Using composition, we can show that our algorithm (Figure 1, with the parameters set as in Theorem 1.1) is $\frac{kt}{n^2}$-ALKL stable. In particular, we show that each one of the k answers is computed in a way that is $\frac{t}{n^2}$-ALKL stable. This follows from the properties of the KL divergence between Gaussian distributions and the way we calibrate the noise. (Alternatively, we could use Laplace noise to obtain similar results.)

We note that $\sqrt{2\varepsilon}$-differential privacy (Dwork et al., 2006a), notions based on Renyi differential privacy (Bun and Steinke, 2016; Mironov, 2017), and ε-KL-stability (Bassily et al., 2016) all imply ε-ALKL stability[2] Thus we can also compose any ALKL stable algorithm with any of the many algorithms satisfying one of the aforementioned definitions.

Crucially, average KL-divergence is strong enough to provide a generalization guarantee that scales with the standard deviation of the queries, as we require. Our proof is based on the high-level approach introduced by Dwork et al. (2015) who first convert a stability guarantee to an upper bound on information between the input dataset and the output of the algorithm and then derive generalization bounds from the bound on information. Here, we demonstrate that ALKL stability implies a bound on the mutual information between the input and output of the algorithm when run on independent samples and then derive generalization guarantees from the bound on mutual information.[3]

Proposition 1.4 *Let $M : (\mathcal{X}^n \cup \mathcal{X}^{n-1}) \to \mathcal{Q}$ be ε-ALKL stable. Let $S \in \mathcal{X}^n$ consist of n independent samples from some distribution $\mathcal{P}$. Then*

$$I(S; M(S)) \leq \varepsilon n, \tag{4}$$

where I denotes mutual information.

1. These relaxations mean that ALKL stability is *not* a good privacy definition, in contrast to differential privacy. In particular, because of the averaging, ALKL stability cannot distinguish between an algorithm that offers good privacy to all individuals and one that offers great privacy for $n-1$ individuals but terrible privacy for the last individual. Compromising a single data point is, however, not an issue for generalization.
2. It may be necessary to extend an algorithm satisfying one of these definitions to inputs of size $n-1$ to satisfy ALKL stability. This can be done by simply padding such an input with one arbitrary item.
3. We thank Adam Smith for suggesting that we try this approach to proving generalization for ALKL stable algorithms.

To prove Proposition 1.4, we introduce an intermediate notion of stability:

Definition 1.5 (Mutual Information Stability) *A randomized algorithm $M : \mathcal{X}^n \to \mathcal{Y}$ is ε-MI stable if, for any random variable S distributed over $\mathcal{X}^n$ (including non-product distributions),*

$$\frac{1}{n}\sum_{i=1}^{n} I(M(S); S_i | S_{-i}) \leq \varepsilon.$$

This notion is based on the notion of stability studied in (Raginsky et al., 2016) that considers only product distributions over the datasets and, as a result, does not compose adaptively.

We prove Proposition 1.4 by combining the following two facts.

(i) ε-ALKL stability implies ε-MI stability. To show this, we express $I(M(S); S_i|S_{-i})$ as the expectation over S of the KL divergence of the distribution (over the randomness of M) of $M(S)$ from an appropriately weighted convex combination of distributions $M(S')$. (Specifically, S' is S with S_i "resampled.") The "mean-as-minimizer" property of KL divergence means we can simply replace this convex combination with $M(S_{-i})$ to complete the proof.

(ii) ε-MI stability implies the mutual information bound (4). To prove this, we invoke the chain rule for mutual information along with the fact that S_i is independent from S_{-i} (which helps resolve the conditioning).

Further, we point out that mutual information stability composes adaptively in the same way as ALKL stability and hence could be useful for understanding adaptive data analysis for more general queries (*e.g.* unlike ALKL stability it does not require $M(S_{-i})$ to be defined).

As first shown in the context of PAC-Bayes bounds (McAllester, 2013) and more recently in (Russo and Zou, 2016), a bound on mutual information implies generalization results. Using a similar technique, we show that, if the mutual information $I(S; \psi_j)$ is small (with S consisting of n i.i.d. draws from $\mathcal{P}$), we have $\mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n}\psi_j(S_i)\right] \approx \mathbf{E}_{X\sim\mathcal{P}}\left[\psi_j(X)\right]$. Moreover, the quality of the approximation scales with the standard deviation. (Specifically, the approximation bound depends on the moment generating function $\mathbf{E}\left[e^{\lambda\mu_j}\right]$ of $\mu_j = \frac{1}{n}\sum_{i=1}^{n}\psi_j(S_i)$, which we bound using both the variance and the range of μ_j.) We can similarly relate the empirical variance σ_j^2 to the true variance. Thus a bound on mutual information suffices to bound generalization error and, thus, prove Theorem 1.1.

Another known implication of bounded mutual information is that any event that would happen with sufficiently low probability on fresh data will still happen with low probability (Russo and Zou, 2016; Rogers et al., 2016). In particular, if E is some "bad" event – such as overfitting the data or making a false discovery – and we know that we are exponentially unlikely to overfit fresh data S', then the probability of M overfitting its input data S is also small, provided the mutual information is small.

One downside of using mutual information is that does not allow us to prove high probability bounds, as can be done with differential privacy and the notion of approximate max-information (Dwork et al., 2015). We note, however, that our analysis still upper bounds the expectation of the largest error among all the queries that were asked. In other words, a union bound over queries is built into the guarantees of the algorithm. Using known

techniques, the confidence can be amplified at the expense of a somewhat more complicated algorithm. In addition, our algorithm yields stronger stability guarantees than just ALKL stability. For example, the minimum noise level of $1/T$ ensures differential privacy (albeit with relatively large parameters[4]). The parameters can be improved using the averaging over the indices that we use in ALKL stability but that leads to a notion that does not appear to compose adaptively. Using a different analysis technique it might be possible to exploit the stronger stability properties of our algorithm to prove high probability generalization bounds. We leave this as an open problem. On the other hand, stability with KL divergence is easier to analyze and allows a potentially wider range of algorithms to be used.

1.2. Related work

Our use of mutual information to derive generalization bounds is closely related to PAC-Bayes bounds first introduced by McAllester (1999) and extended in a number of subsequent works (see (McAllester, 2013) for an overview). In this line of work, the expected generalization error of a predictive model (such as classifier) randomly chosen from some data-dependent distribution $\mathcal{Q}(S)$ is upper-bounded by the KL divergence between $\mathcal{Q}$ and an arbitrary data-independent prior distribution $\mathcal{P}_0$. One natural choice of $\mathcal{Q}(S)$ is the output distribution of a randomized learning algorithm $\mathcal{A}$ on S. By choosing the prior $\mathcal{P}_0$ to be the distribution of the output of $\mathcal{A}$ on a dataset drawn from $\mathcal{P}^n$ one obtains that the expected generalization error is upper-bounded by the expected KL divergence between $\mathcal{Q}(S)$ an $\mathcal{P}_0$ (McAllester, 2013). While this has not been pointed out in (McAllester, 2013), this is exactly the mutual information between S and $\mathcal{A}(S)$.

Recently, interest in using information-based generalization bounds was revived by applications in adaptive data analysis (Dwork et al., 2015). Specifically, Dwork et al. (2015) demonstrate that approximate max-information between the input dataset and the output of the algorithm (a notion based on the infinity divergence between the joint distribution and the product of marginals) implies generalization bounds with high probability. They also showed that $(\epsilon, 0)$-differential privacy implies an upper bound on approximate max-information (and this later extended to (ϵ, δ)-differential privacy by Rogers et al. (2016)). Russo and Zou (2016) show that mutual information can also be used to derive bounds on expected generalization error and discuss several applications of these bounds. Xu and Raginsky (2017) show how to derive "low-probability" bounds on the generalization error in this context. (We note that (Russo and Zou, 2016; Xu and Raginsky, 2017) use the same technique as that used in PAC-Bayes bounds and appear to have overlooked the direct connection between their results and the PAC-Bayes line of work.)

Recent work (Bassily et al., 2018) studies learning algorithms in the PAC model whose output has low mutual information with the input dataset. They also discuss generalization bounds based on mutual information and (independently) derive results similar to those we give for low-probability events.

4. Specifically, with the parameter setting from Theorem 1.1, our algorithm satisfies $\left(O\left(\sqrt{\log(1/\delta)}\right), \delta\right)$-differential privacy for all $\delta > k^{-\Omega(k)}$.

Acknowledgements

We thank Adam Smith for his suggestion to analyze the generalization of ALKL stable algorithms via mutual information. This insight greatly simplified our initial analysis and allowed us to derive additional corollaries. We also thank Nati Srebro for pointing out the connection between our results and the PAC-Bayes generalization bounds. Part of this work was done while Vitaly Feldman was at IBM Research – Almaden and while visiting the Simons Institute, UC Berkeley.

References

Raef Bassily and Yoav Freund. Typicality-based stability and privacy. *CoRR*, abs/1604.03336, 2016. URL http://arxiv.org/abs/1604.03336.

Raef Bassily, Kobbi Nissim, Adam D. Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *STOC*, pages 1046–1059, 2016. URL http://arxiv.org/abs/1511.02513.

Raef Bassily, Shay Moran, Ido Nachum, Jonathan Shafer, and Amir Yehudayoff. Learners that use little information. In *ALT*, pages 25–55, 2018. URL http://proceedings.mlr.press/v83/bassily18a.html.

Olivier Bousquet and André Elisseeff. Stability and generalization. *JMLR*, 2:499–526, 2002. URL http://www.jmlr.org/papers/v2/bousquet02a.html.

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer Berlin Heidelberg, 2016. URL https://arxiv.org/abs/1605.02065.

C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, pages 265–284, 2006a. URL http://repository.cmu.edu/jpc/vol7/iss3/2.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 486–503. Springer Berlin Heidelberg, 2006b. URL https://www.iacr.org/archive/eurocrypt2006/40040493/40040493.pdf.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Preserving statistical validity in adaptive data analysis. *CoRR*, abs/1411.2664, 2014. Extended abstract in STOC 2015.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. *CoRR*, abs/1506, 2015. Extended abstract in NIPS 2015.

Vitaly Feldman and Thomas Steinke. Generalization for adaptively-chosen estimators via stable median. In *Conference on Learning Theory (COLT)*, 2017a. URL https://arxiv.org/abs/1706.05069.

Vitaly Feldman and Thomas Steinke. Calibrating noise to variance in adaptive data analysis. *CoRR*, abs/1712.07196, 2017b. URL http://arxiv.org/abs/1712.07196.

M. Hardt and J. Ullman. Preventing false discovery in interactive data analysis is hard. In *FOCS*, pages 454–463, 2014. URL https://arxiv.org/abs/1408.1655.

M. Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 45 (6):983–1006, 1998.

David McAllester. A pac-bayesian tutorial with a dropout bound. *arXiv preprint arXiv:1307.2118*, 2013. URL https://arxiv.org/abs/1307.2118.

David A. McAllester. Pac-bayesian model averaging. In *COLT*, pages 164–170, 1999. doi: 10.1145/307400.307435.

Ilya Mironov. Rényi differential privacy. In *Computer Security Foundations Symposium, CSF*, pages 263–275, 2017.

Tomaso Poggio, Ryan Rifkin, Sayan Mukherjee, and Partha Niyogi. General conditions for predictivity in learning theory. *Nature*, 428(6981):419–422, 2004.

Maxim Raginsky, Alexander Rakhlin, Matthew Tsao, Yihong Wu, and Aolin Xu. Information-theoretic analysis of stability and bias of learning algorithms. In *2016 IEEE Information Theory Workshop, ITW 2016, Cambridge, United Kingdom, September 11-14, 2016*, pages 26–30, 2016.

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

Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS*, 2016. URL https://arxiv.org/abs/1511.05219.

Thomas Steinke. Adaptive data analysis. 2016. URL http://people.seas.harvard.edu/~madhusudan/courses/Spring2016/notes/thomas-notes-ada.pdf. Lecture Notes.

Thomas Steinke and Jonathan Ullman. Interactive fingerprinting codes and the hardness of preventing false discovery. In *COLT*, pages 1588–1628, 2015. URL http://jmlr.org/proceedings/papers/v40/Steinke15.html.

Yu-Xiang Wang, Jing Lei, and Stephen E Fienberg. On-average kl-privacy and its equivalence to generalization for max-entropy mechanisms. In *International Conference on Privacy in Statistical Databases*, pages 121–134. Springer, 2016. URL https://arxiv.org/abs/1605.02277.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *CoRR*, abs/1705.07809, 2017. URL http://arxiv.org/abs/1705.07809.