SIAM J. COMPUT.
Vol. 50, No. 3, pp. STOC16-377–STOC16-405

ALGORITHMIC STABILITY FOR ADAPTIVE DATA ANALYSIS*

RAEF BASSILY†, KOBBI NISSIM‡, ADAM SMITH§, THOMAS STEINKE¶, URI STEMMER‖, AND JONATHAN ULLMAN#

Abstract. Adaptivity is an important feature of data analysis—the choice of questions to ask about a dataset often depends on previous interactions with the same dataset. However, statistical validity is typically studied in a nonadaptive model, where all questions are specified before the dataset is drawn. Recent work by Dwork et al. [*Proceedings of STOC*, ACM, 2015, pp.117–126] and Hardt and Ullman [*Proceedings of FOCS*, IEEE, 2014, pp. 454–463] initiated the formal study of this problem and gave the first upper and lower bounds on the achievable generalization error for adaptive data analysis. Specifically, suppose there is an unknown distribution $\mathbf{P}$ and a set of n independent samples $\mathbf{x}$ is drawn from $\mathbf{P}$. We seek an algorithm that, given $\mathbf{x}$ as input, accurately answers a sequence of adaptively chosen "queries" about the unknown distribution $\mathbf{P}$. How many samples n must we draw from the distribution, as a function of the type of queries, the number of queries, and the desired level of accuracy? In this work we make two new contributions toward resolving this question: 1. We give upper bounds on the number of samples n that are needed to answer *statistical queries*. The bounds improve and simplify the work of Dwork et al. and have been applied in subsequent work by those authors [*Science*, 349 (2015), pp. 636–638; *Proceedings of NIPS*, 2015, pp. 2350–2358]. 2. We prove the first upper bounds on the number of samples required to answer more general families of queries. These include arbitrary *low-sensitivity queries* and an important class of *optimization queries* (alternatively, *risk minimization queries*). As in Dwork et al., our algorithms are based on a connection with *algorithmic stability* in the form of *differential privacy*. We extend their work by giving a quantitatively optimal, more general, and simpler proof of their main theorem that stable algorithms of the kind guaranteed by differential privacy imply low generalization error. We also show that weaker stability guarantees such as bounded Kullback–Leibler divergence and total variation distance lead to correspondingly weaker generalization guarantees.

Key words. adaptive data analysis, algorithmic stability, differential privacy, statistical queries

AMS subject classification. 68Q01

DOI. 10.1137/16M1103646

*Received by the editors November 15, 2016; accepted for publication (in revised form) January 4, 2021; published electronically April 20, 2021. This work unifies and subsumes two arXiv manuscripts arXiv:1503.04843, 2015 and arXiv:1504.05800, 2015. A preliminary version of the joint work appeared in *Proceedings of STOC*, 2016.

https://doi.org/10.1137/16M1103646

Funding: The first author was supported by NSF award CDI-0941553. The second author was supported by a Simons Investigator grant to Salil Vadhan and by NSF grant CNS-1237235. The third author was supported by NSF award IIS-1447700 and a Google Faculty award. The second and third authors were supported by the Sloan Foundation. The fourth author was supported by NSF grants CCF-1116616, CCF-1420938, and CNS-1237235. The fifth author was supported by a gift from Google. The sixth author was supported by a Junior Fellowship from the Simons Society of Fellows.

†Center for Information Theory and Applications and Department of Computer Science and Engineering, University of California San Diego, San Diego, CA 92093 USA (rbassily@ucsd.edu).

‡Department of Computer Science, Ben-Gurion University of the Negev, Beer Sheva, 8410501, Israel, and Center for Research on Computation and Society (CRCS), Harvard University, Cambridge, MA 02138 USA (kobbi@cs.bgu.ac.il).

§Department of Computer Science and Engineering, Penn State University, State College, PA 16802 USA (asmith@psu.edu).

¶John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138 USA (tsteinke@seas.harvard.edu).

‖Department of Computer Science, Ben-Gurion University of the Negev, Beer Sheva, 8410501, Israel (stemmer@cs.bgu.ac.il).

#College of Computer and Information Science, Northeastern University, Boston, MA 02115 USA (jullman@gmail.com).

1. Introduction. Multiple hypothesis testing is a ubiquitous task in empirical research. A finite sample of data is drawn from some unknown population, and several analyses are performed on that sample. The outcome of an analysis is deemed significant if it is unlikely to have occurred by chance alone, and a "false discovery" occurs if the analyst incorrectly declares an outcome to be significant. False discovery has been identified as a substantial problem in the scientific community (see, e.g., [Ioa05, GL14]). This problem persists despite decades of research by statisticians on methods for preventing false discovery, such as the widely used Bonferroni correction [Bon36, Dun61] and the Benjamini–Hochberg procedure [BH95].

False discovery is often attributed to misuse of statistics. An alternative explanation is that the prevalence of false discovery arises from the inherent *adaptivity* in the data analysis process—the fact that the choice of analyses to perform depends on previous interactions with the data (see, e.g., [GL14]). Adaptivity is essentially unavoidable when a sequence of research groups publish research papers based on overlapping datasets. Adaptivity also arises naturally in other settings, for example, in multistage inference algorithms where data are preprocessed (say, to select features or restrict to a principal subspace) before the main analysis is performed, in scoring data-based competitions [BH15], and in the reuse of holdout or test data [DFH+15c, DFH+15a].

The general problem of adaptive data analysis was formally modeled and studied in recent papers by Dwork et al. [DFH+15b] and by Hardt and Ullman [HU14]. The striking results of Dwork et al. [DFH+15b] gave the first nontrivial algorithms for provably ensuring statistical validity in adaptive data analysis, allowing for even an *exponential* number of tests against the same sample. In contrast, [HU14, SU15b] showed inherent statistical and computational barriers to preventing false discovery in adaptive settings.

The key ingredient in Dwork et al. is a notion of "algorithmic stability" that is suitable for adaptive analysis. Informally, changing one input to a stable algorithm does not change its output too much. Traditionally, stability was measured via the change in the generalization error of an algorithm's output, and algorithms stable according to such a criterion have long been known to ensure statistical validity in nonadaptive analysis [DW79a, DW79b, KR99, BE02, SSSS10]. Following a connection first suggested by McSherry,[1] Dwork et al. showed that a stronger stability condition designed to ensure data privacy—called *differential privacy* [DMNS06, Dwo06]—guarantees statistical validity in adaptive data analysis. This allowed them to repurpose known differential privacy algorithms to prevent false discovery. A crucial difference from traditional notions of stability is that differential privacy requires a change in one input lead to a small change in the *probability distribution* on the outputs (in particular, differentially private algorithms must be randomized). In this paper, we refer to differential privacy as *max-KL stability* (Definition 2.3) to emphasize the relationship to the literature on algorithmic stability, and to the other notions of stability that we study (Kullback–Leibler (KL) and total variation (TV) stability, in particular).

In this work, we extend the results of Dwork et al. along two axes. First, we give an *optimal* analysis of the statistical validity of max-KL stable algorithms. As a consequence, we immediately obtain the best known bounds on the *sample complexity* (equivalently, the *convergence rate*) of adaptive data analysis. Second, we generalize the connection between max-KL stability and statistical validity to a much larger family of statistics. Our proofs are also significantly simpler than those

[1] See, e.g., [McS14], although the observation itself dates back at least to 2008 (personal communication).

of Dwork et al., and clarify the role of different stability notions in the adaptive setting.

1.1. Overview of results.

Adaptivity and statistical queries. Following the previous work on this subject [DFH+15b], we formalize the problem of adaptive data analysis as follows. There is a *distribution* $\mathbf{P}$ over some finite universe $\mathcal{X}$, and a *mechanism* $\mathcal{M}$ that does not know $\mathbf{P}$, but is given a set $\mathbf{x}$ consisting of n samples from $\mathbf{P}$. Using its sample, the mechanism must answer *queries* on $\mathbf{P}$. Here, a query q, coming from some family Q, maps a distribution $\mathbf{P}$ to a real-valued answer. The mechanism's answer a to a query q is α-*accurate* if $|a - q(\mathbf{P})| \leq \alpha$ with high probability. Importantly, the mechanism's goal is to provide answers that "generalize" to the underlying distribution, rather than answers that are specific to its sample.

We model adaptivity by allowing a *data analyst* to ask a sequence of queries $q_1, q_2, \ldots, q_k \in Q$ to the mechanism, which responds with answers $a_1, a_2, \ldots, a_k$. In the adaptive setting, the query q_j may depend on the previous queries and answers $q_1, a_1, \ldots, q_{j-1}, a_{j-1}$ arbitrarily. We say the mechanism is α-*accurate* given n samples for k adaptively chosen queries if, with high probability,[2] when given a vector $\mathbf{x}$ of n samples from an arbitrary distribution $\mathbf{P}$, the mechanism accurately responds to any adaptive analyst that makes at most k queries.

Dwork et al. [DFH+15b] considered the family of *statistical queries* [Kea93]. A statistical query q asks for the expected value of some function on random draws from the distribution. That is, the query is specified by a function $p : \mathcal{X} \to [0, 1]$ and its answer is $q(\mathbf{P}) = \mathbb{E}_{z \leftarrow_{\mathrm{R}} \mathbf{P}}[p(z)]$.

The most natural way to answer a statistical query is to compute the *empirical answer* $\mathbb{E}_{z \leftarrow_{\mathrm{R}} \mathbf{x}}[p(z)]$, which is just the average value of the function on the given sample $\mathbf{x}$.[3] It is simple to show that when k queries are specified *nonadaptively* (i.e., independent of previous answers), then the empirical answer is within $q(\mathbf{P}) \pm \alpha$ (henceforth, "α-accurate") with high probability so long as the sample has size $n \gtrsim \log(k)/\alpha^2$.[4,5] However, when the queries can be chosen adaptively, the empirical average performs much worse. In particular, there is an algorithm (based on [DN03]) that, after seeing the empirical answer to $k = O(\alpha^2 n)$ random queries, can find a query such that the empirical answer and the correct answer differ by α. Thus, the empirical average is guaranteed to be accurate only when $n \gtrsim k/\alpha^2$, and so exponentially more samples are required to guarantee accuracy when the queries may be adaptive.

Answering adaptive statistical queries. Surprisingly, Dwork et al. [DFH+15b], showed there are mechanisms that are much more effective than naïvely outputting the empirical answer. They show that "stable" mechanisms are accurate and, by applying a stable mechanism from the literature on differential privacy, they obtain mechanisms that are accurate given only $n \gtrsim \sqrt{k}/\alpha^{2.5}$ samples, which is a significant

[2] By "with high probability," we mean that we are primarily interested in accuracy statements that hold with probability $1 - \beta$ for arbitrarily small β. Typically β will decay exponentially in the number of samples.

[3] For convenience, we will often use $\mathbf{x}$ as shorthand for the empirical distribution over $\mathbf{x}$. We use $z \leftarrow_{\mathrm{R}} \mathbf{x}$ to mean a random element chosen from the uniform distribution over the elements of $\mathbf{x}$.

[4] This guarantee follows from bounding the error of each query using a Chernoff bound and then taking a union bound over all queries. The $\log k$ term corresponds to the Bonferroni correction in classical statistics.

[5] To simplify notation in this introduction, we write $n \gtrsim f(k, \alpha)$ to denote that n must be at least as large as some quantity that is approximately $f(k, \alpha)$ to within polylogarithmic factors.

improvement over the naïve mechanism when α is not too small. (See Table 1 for more detailed statements of their results, including results that achieve an *exponential* improvement in the sample complexity when $|\mathcal{X}|$ is bounded.)

Our first contribution is to give a simpler and quantitatively optimal analysis of the generalization properties of stable algorithms, which immediately yields new accuracy bounds for adaptive statistical queries. In particular, we show that $n \gtrsim \sqrt{k}/\alpha^2$ samples suffice. Since $1/\alpha^2$ samples are required to answer a single nonadaptive query, our dependence on α is optimal.

Beyond statistical queries. Although statistical queries are surprisingly general [Kea93], we would like to be able to ask more general queries on the distribution $\mathbf{P}$ that capture a wider variety of machine learning and data mining tasks. To this end, we give the first bounds on the sample complexity required to answer large numbers of adaptively chosen *low-sensitivity queries* and *optimization queries,* which we now describe.

Low-sensitivity queries are a generalization of statistical queries. A query is specified by an arbitrary function $p : \mathcal{X}^n \to \mathbb{R}$ satisfying $|p(\mathbf{x}) - p(\mathbf{x}')| \le 1/n$ for every $\mathbf{x}, \mathbf{x}' \in \mathcal{X}^n$ differing on exactly one element. The query applied to the population is defined to be $q(\mathbf{P}) = \mathbb{E}_{\mathbf{x} \leftarrow_{\mathrm{R}} \mathbf{P}^n}[p(\mathbf{x})]$. Examples include *distance queries* (e.g., "How far is the sample from being well-clustered?") and maxima of statistical queries (e.g., "What is the classification error of the best k-node decision tree?")

Optimization queries are a broad generalization of low-sensitivity queries to arbitrary output domains. The query is specified by a loss function $L : \mathcal{X}^n \times \Theta \to \mathbb{R}$ that is low-sensitivity in its first parameter, and the goal is to output $\theta \in \Theta$ that is "best" in the sense that it minimizes the average loss. Specifically, $q(\mathbf{P}) = \arg\min_{\theta \in \Theta} \mathbb{E}_{\mathbf{z} \leftarrow_{\mathrm{R}} \mathbf{P}^n}[L(\mathbf{z}; \theta)]$. An important special case is when $\Theta \subseteq \mathbb{R}^d$ is convex and L is convex in θ, which captures many fundamental regression and classification problems.

Our sample complexity bounds are summarized in Table 1.

Subsequent work. Our bounds were applied in subsequent work of Dwork et al. [DFH+15c, DFH+15a] in the analysis of their "reusable holdout" construction.

TABLE 1

Summary of results. Here k = number of queries, n = number of samples, α = desired accuracy, $\mathcal{X}$ = universe of possible samples, d = dimension of parameter space Θ.

Query type	Sample complexity		Time per query
	[DFH+15b]	This Work	
Statistical ($k \ll n^2$)	$\tilde{O}\left(\frac{\sqrt{k}}{\alpha^{2.5}}\right)$	$\tilde{O}\left(\frac{\sqrt{k}}{\alpha^2}\right)$	$\mathrm{poly}(n, \log \lvert\mathcal{X}\rvert)$
Statistical ($k \gg n^2$)	$\tilde{O}\left(\frac{\sqrt{\log \lvert\mathcal{X}\rvert} \cdot \log^{3/2} k}{\alpha^{3.5}}\right)$	$\tilde{O}\left(\frac{\sqrt{\log \lvert\mathcal{X}\rvert} \cdot \log k}{\alpha^3}\right)$	$\mathrm{poly}(n, \lvert\mathcal{X}\rvert)$
Low sensitivity ($k \ll n^2$)	—	$\tilde{O}\left(\frac{\sqrt{k}}{\alpha^2}\right)$	$\mathrm{poly}(n, \log \lvert\mathcal{X}\rvert)$
Low sensitivity ($k \gg n^2$)	—	$\tilde{O}\left(\frac{\log \lvert\mathcal{X}\rvert \cdot \log k}{\alpha^3}\right)$	$\mathrm{poly}(\lvert\mathcal{X}\rvert^n)$
Convex min. ($k \ll n^2$)	—	$\tilde{O}\left(\frac{\sqrt{dk}}{\alpha^2}\right)$	$\mathrm{poly}(n, d, \log \lvert\mathcal{X}\rvert)$
Convex min. ($k \gg n^2$)	—	$\tilde{O}\left(\frac{(\sqrt{d} + \log k) \cdot \sqrt{\log \lvert\mathcal{X}\rvert}}{\alpha^3}\right)$	$\mathrm{poly}(n, d, \lvert\mathcal{X}\rvert)$

1.2. Overview of techniques. Our main result is a new proof, with optimal parameters, that a stable algorithm that provides answers to adaptive queries that are close to the empirical value on the sample gives answers that generalize to the underlying distribution. In particular, we prove the following.

THEOREM 1.1 (main "transfer theorem"). *Let $\mathcal{M}$ be a mechanism that takes a sample $\mathbf{x} \in \mathcal{X}^n$ and answers k adaptively chosen low-sensitivity queries. Suppose that $\mathcal{M}$ satisfies the following for some $\alpha, \beta > 0$:*

1. *For every sample $\mathbf{x}$, $\mathcal{M}$'s answers are $(\alpha, \alpha\beta)$-accurate with respect to the sample $\mathbf{x}$. That is, $\mathbb{P}\left[\max_{j \in k} |q_j(\mathbf{x}) - a_j| \leq \alpha\right] \geq 1 - \alpha\beta$, where $q_1, \ldots, q_k : \mathcal{X}^n \to \mathbb{R}$ are the low-sensitivity queries that are asked and $a_1, \ldots, a_k \in \mathbb{R}$ are the answers given. The probability is taken only over $\mathcal{M}$'s random coins.*
2. *$\mathcal{M}$ satisfies $(\alpha, \alpha\beta)$-max-KL stability (Definition 2.3, identical to $(\alpha, \alpha\beta)$-differential privacy).*

Then, if $\mathbf{x}$ consists of n samples from an arbitrary distribution $\mathbf{P}$ over $\mathcal{X}$, $\mathcal{M}$'s answers are $(O(\alpha), O(\beta))$-accurate with respect to $\mathbf{P}$. That is, $\mathbb{P}\left[\max_{j \in k} |q_j(\mathbf{P}) - a_j| \leq O(\alpha)\right] \geq 1 - O(\beta)$, where the probability is taken only over the choice of $\mathbf{x} \leftarrow_R \mathbf{P}^n$ and $\mathcal{M}$'s random coins.

Our actual result is somewhat more general than Theorem 1.1. We show that the population-level error of a stable algorithm is close to its error on the sample, whether or not that error is low. Put glibly, stable algorithms cannot be wrong without realizing it.

Compared to the results of [DFH+15b], Theorem 1.1 requires a quantitatively weaker stability guarantee—$(\alpha, \alpha\beta)$-stability, instead of $(\alpha, (\beta/k)^{1/\alpha})$-stability. It also applies to arbitrary low-sensitivity queries as opposed to the special case of statistical queries.

Our analysis differs from that of Dwork et al. in two key ways. First, we give a better bound on the probability with which a *single* low-sensitivity query output by a max-KL stable algorithm has good generalization error. Second, we show a reduction from the case of *many* queries to the case of a single query that has no loss in parameters (in contrast, previous work took a union bound over queries, leading to a dependence on k, the number of queries).

Both steps rely on a thought experiment in which several "real" executions of a stable algorithm are simulated inside another algorithm, called a *monitor*, which outputs a function of the "real" transcripts. Because stability is closed under post-processing, the monitor is itself stable. Because it exists only as a thought experiment, the monitor can be given knowledge of the true distribution from which the data are drawn and can use this knowledge to process the outputs of the simulated "real" runs. The monitor technique allows us to start from a basic guarantee, which states that a single query has good generalization error with constant probability, and amplify the guarantee so that (a) the generalization error holds with very high probability, and (b) the guarantee holds simultaneously over all queries in a sequence. The proof of the basic guarantee follows the lines of existing proofs using algorithmic stability (e.g., [DW79a]), while the monitor technique and the resulting amplification statements are new.

The amplification of success probability is the more technically sophisticated of the two key steps. The idea is to run many (about $1/\beta$, using the notation of Theorem 1.1) copies of a stable mechanism on independently selected datasets. Each of these interactions results in a sequence of queries and answers. The monitor then

selects the query and answer pair from among all of the sequences that has the largest error. It then outputs this query as well as the index of the interaction that produced it. Our main technical lemma shows that the monitor will find a "bad" query/dataset pair (one where the true and empirical values of the query differ) with at most constant probability. This implies that each of the real executions outputs a bad query with probability $O(\beta)$. Relative to previous work, the resulting argument yields better bounds, applies to more general classes of queries, and even generalizes to other notions of stability.

Optimality. In general, we cannot prove that our bounds are optimal. Recently, [HU14, SU15b] showed that $n \gtrsim \min\{\sqrt{k}, \sqrt{\log|\mathcal{X}|}\}/\alpha$ samples are necessary to answer adaptively chosen statistical queries. In addition, clearly $n \gtrsim \log(k)/\alpha^2$ are necessary, even for nonadaptive queries. However, the gap between the upper and lower bounds is still significant.

However, we can show that our connection between max-KL stability and generalization is optimal (see section 7 for details). Moreover, for every family of queries we consider, no max-KL stable algorithm can achieve better sample complexity [BUV14, BST14]. Thus, any significant improvement to our bounds must come from using a weaker notion of stability or some entirely different approach.

Computational complexity. Throughout, we will assume that the analyst only issues queries q such that the empirical answer $q(\mathbf{x})$ can be evaluated in time $\text{poly}(n, \log|\mathcal{X}|)$. When $k \ll n^2$ our algorithms have similar running time. However, when answering $k \gg n^2$ queries, our algorithms suffer running time at least $\text{poly}(n, |\mathcal{X}|)$. Since the mechanism's input is of size $n \cdot \log|\mathcal{X}|$, these algorithms cannot be considered computationally efficient. For example, if $\mathcal{X} = \{0,1\}^d$ for some dimension d, then in the nonadaptive setting $\text{poly}(n, d)$ running time would suffice, whereas our algorithms require $\text{poly}(n, 2^d)$ running time. Unfortunately, this running time is known to be optimal, as [HU14, SU15b] (building on hardness results in privacy [Ull13]) showed that, assuming exponentially hard one-way functions exist, any $\text{poly}(n, 2^{o(d)})$ time mechanism that answers $k = \omega(n^2)$ statistical queries is not even 1/3-accurate.

Stable/differentially private mechanisms. Each of our results requires instantiating the mechanism with a suitable stable/differentially private algorithm. For statistical queries, the optimal mechanisms are the well-known Gaussian and Laplace mechanisms (slightly refined by [SU15a]) when k is small and the private multiplicative weights mechanism [HR10] when k is large. For arbitrary low-sensitivity queries, the Gaussian or Laplace mechanism is again optimal when k is small, and for large k we can use the median mechanism [RR10].

When considering arbitrary search queries over an arbitrary finite range, the optimal algorithm is the exponential mechanism [MT07]. For the special case of convex minimization queries over an infinite domain, we use the optimal algorithm of [BST14] when k is small, and when k is large we use an algorithm of [Ull15] that accurately answers exponentially many such queries.

Other notions of stability. Our techniques apply to notions of distributional stability other than max-KL/differential privacy. In particular, defining stability in terms of TV or of KL divergence leads to bounds on the generalization error that have polynomially, rather than exponentially, decreasing tails. See section 4 for details.

2. Preliminaries.

2.1. Queries. Given a distribution $\mathbf{P}$ over $\mathcal{X}$ or a sample $\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$, we would like to answer *queries* about $\mathbf{P}$ or $\mathbf{x}$ from some family Q. We will often want to bound the "sensitivity" of the queries with respect to changing one element

of the sample. To this end, we use $\mathbf{x} \sim \mathbf{x}'$ to denote that $\mathbf{x}$ and $\mathbf{x}' \in \mathcal{X}^n$ differ on at most one entry. We will consider several different families of queries:

- **Statistical queries:** These queries are specified by a function $q : \mathcal{X} \to [0,1]$ and (abusing notation) are defined as

$$q(\mathbf{P}) = \mathop{\mathbb{E}}_{z \leftarrow_{\mathrm{R}} \mathbf{P}}[q(z)] \qquad \text{and} \qquad q(\mathbf{x}) = \frac{1}{n}\sum_{i \in [n]} q(x_i).$$

  The error of an answer a to a statistical query q with respect to $\mathbf{P}$ or $\mathbf{x}$ is defined to be

$$\mathrm{err}_{\mathbf{x}}(q,a) = a - q(\mathbf{x}) \qquad \text{and} \qquad \mathrm{err}^{\mathbf{P}}(q,a) = a - q(\mathbf{P}).$$

- **Δ-sensitive queries:** For $\Delta \in [0,1]$, $n \in \mathbb{N}$, these queries are specified by a function $q : \mathcal{X}^n \to \mathbb{R}$ satisfying $|q(\mathbf{x}) - q(\mathbf{x}')| \le \Delta$ for every pair $\mathbf{x}, \mathbf{x}' \in \mathcal{X}^n$ differing in only one entry. Abusing notation, let

$$q(\mathbf{P}) = \mathop{\mathbb{E}}_{\mathbf{z} \leftarrow_{\mathrm{R}} \mathbf{P}^n}[q(\mathbf{z})].$$

  The error of an answer a to a Δ-sensitive query q with respect to $\mathbf{P}$ or $\mathbf{x}$ is defined to be

$$\mathrm{err}_{\mathbf{x}}(q,a) = a - q(\mathbf{x}) \qquad \text{and} \qquad \mathrm{err}^{\mathbf{P}}(q,a) = \mathop{\mathbb{E}}_{\mathbf{z} \leftarrow_{\mathrm{R}} \mathbf{P}^n}[\mathrm{err}_{\mathbf{z}}(q,a)] = a - q(\mathbf{P}).$$

  We denote the set of all Δ-sensitive queries by Q_Δ. If $\Delta = O(1/n)$ we say the query is *low sensitivity*. Note that $1/n$-sensitive queries are a strict generalization of statistical queries.

- **Minimization queries:** These queries are specified by a loss function $L : \mathcal{X}^n \times \Theta \to \mathbb{R}$. We require that L has sensitivity Δ with respect to its first parameter, that is,

$$\sup_{\theta \in \Theta, \mathbf{x}, \mathbf{x}' \in \mathcal{X}^n, \mathbf{x} \sim \mathbf{x}'} |L(\mathbf{x};\theta) - L(\mathbf{x}';\theta)| \le \Delta.$$

  Here Θ is an arbitrary set of items (sometimes called "parameter values") among which we aim to choose the item ("parameter") with minimal loss, either with respect to a particular input dataset $\mathbf{x}$ or with respect to expectation over a distribution $\mathbf{P}$.

  The error of an answer $\theta \in \Theta$ to a minimization query $L : \mathcal{X}^n \times \Theta \to \mathbb{R}$ with respect to $\mathbf{x}$ is defined to be

$$\mathrm{err}_{\mathbf{x}}(L,\theta) = L(\mathbf{x},\theta) - \min_{\theta^* \in \Theta} L(\mathbf{x},\theta^*)$$

  and, with respect to $\mathbf{P}$, is

$$\mathrm{err}^{\mathbf{P}}(L,\theta) = \mathop{\mathbb{E}}_{\mathbf{z} \leftarrow_{\mathrm{R}} \mathbf{P}^n}[\mathrm{err}_{\mathbf{z}}(L,\theta)] = \mathop{\mathbb{E}}_{\mathbf{z} \leftarrow_{\mathrm{R}} \mathbf{P}^n}[L(\mathbf{z},\theta)] - \mathop{\mathbb{E}}_{\mathbf{z} \leftarrow_{\mathrm{R}} \mathbf{P}^n}\left[\min_{\theta^* \in \Theta} L(\mathbf{z},\theta^*)\right].$$

  Note that $\min_{\theta^* \in \Theta} \mathbb{E}_{\mathbf{z} \leftarrow_{\mathrm{R}} \mathbf{P}^n}[L(\mathbf{z},\theta^*)] \ge \mathbb{E}_{\mathbf{z} \leftarrow_{\mathrm{R}} \mathbf{P}^n}[\min_{\theta^* \in \Theta} L(\mathbf{z},\theta^*)]$, whence

$$\mathop{\mathbb{E}}_{\mathbf{z} \leftarrow_{\mathrm{R}} \mathbf{P}^n}[L(\mathbf{z},\theta)] - \min_{\theta^* \in \Theta} \mathop{\mathbb{E}}_{\mathbf{z} \leftarrow_{\mathrm{R}} \mathbf{P}^n}[L(\mathbf{z},\theta^*)] \le \mathrm{err}^{\mathbf{P}}(L,\theta),$$

  so the quantity $\mathrm{err}^{\mathbf{P}}(L,\theta)$ upper bounds the standard notion of population error. Note that minimization queries (with $\Theta = \mathbb{R}$) generalize low-sensitivity queries: given a Δ-sensitive $q : \mathcal{X}^n \to \mathbb{R}$, we can define $L(\mathbf{x};\theta) = |\theta - q(\mathbf{x})|$ to obtain a minimization query with the same answer.

$\mathcal{A}$ chooses a distribution $\mathbf{P}$ over $\mathcal{X}$.
Sample $x_1, \ldots, x_n \leftarrow_{\mathrm{R}} \mathbf{P}$, let $\mathbf{x} = (x_1, \ldots, x_n)$. (Note that $\mathcal{A}$ does not know $\mathbf{x}$.)
For $j = 1, \ldots, k$
  $\mathcal{A}$ outputs a query $q_j \in Q$.
  $\mathcal{M}(\mathbf{x}, q_j)$ outputs a_j.
  (As $\mathcal{A}$ and $\mathcal{M}$ are stateful, q_j and a_j may depend on the history $q_1, a_1, \ldots, q_{j-1}, a_{j-1}$.)

FIG. 1. *The accuracy game* $\mathsf{Acc}_{n,k,Q}[\mathcal{M}, \mathcal{A}]$.

We denote the set of minimization queries by Q_{min}. We highlight two special cases:
- *Minimization for finite sets:* We denote by $Q_{min,D}$ the set of minimization queries where Θ is finite with size at most D.
- *Convex minimization queries:* If $\Theta \subset \mathbb{R}^d$ is closed and convex and $L(\mathbf{x}; \cdot)$ is convex on Θ for every dataset $\mathbf{x}$, then the query can be answered nonprivately up to any desired error α, in time polynomial in d and α. We denote the set of all convex minimization queries by Q_{CM}.

2.2. Mechanisms for adaptive queries. Our goal is to design a *mechanism* $\mathcal{M}$ that answers queries on $\mathbf{P}$ using only independent samples $x_1, \ldots, x_n \leftarrow_{\mathrm{R}} \mathbf{P}$. Our focus is the case where the queries are chosen adaptively and adversarially.

Specifically, $\mathcal{M}$ is a stateful algorithm that holds a sample $x_1, \ldots, x_n \in \mathcal{X}$, takes a query q from some family Q as input, and returns an answer a. We require that when $x_1, \ldots, x_n$ are independent draws from $\mathbf{P}$, the answer a is "close" to $q(\mathbf{P})$ in a sense that is appropriate for the family of queries. Moreover we require that this condition holds for every query in an adaptively chosen sequence $q_1, \ldots, q_k$. Formally, we define an accuracy game between a mechanism $\mathcal{M}$ and a stateful *data analyst* $\mathcal{A}$ in Figure 1.

DEFINITION 2.1 (accuracy). *A mechanism* $\mathcal{M}$ *is* (α, β)-accurate with respect to the population for k adaptively chosen queries from Q given n samples in $\mathcal{X}$ *if for every adversarial data analyist*[6] $\mathcal{A}$,

$$\mathop{\mathbb{P}}_{\mathsf{Acc}_{n,k,Q}[\mathcal{M},\mathcal{A}]} \left[\max_{j \in [k]} \left| \mathrm{err}^{\mathbf{P}}(q_j, a_j) \right| \le \alpha \right] \ge 1 - \beta.$$

We will also use a definition of accuracy relative to the sample given to the mechanism, described in Figure 2.

DEFINITION 2.2 (sample accuracy). *A mechanism* $\mathcal{M}$ *is* (α, β)-accurate with respect to samples of size n from $\mathcal{X}$ for k adaptively chosen queries from Q *if for every adversary* $\mathcal{A}$,

$$\mathop{\mathbb{P}}_{\mathsf{SampAcc}_{n,k,Q}[\mathcal{M},\mathcal{A}]} \left[\max_{j \in [k]} \left| \mathrm{err}_{\mathbf{x}}(q_j, a_j) \right| \le \alpha \right] \ge 1 - \beta.$$

[6] As the data analyst is assumed to be arbitrary, it is often convenient to think of it as an *adversary* and thus we will sometimes interchange the terms adversary and data analyst.

$\mathcal{A}$ chooses $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathcal{X}^n$.
For $j = 1, \ldots, k$
  $\mathcal{A}$ outputs a query $q_j \in Q$.
  $\mathcal{M}(\mathbf{x}, q_j)$ outputs a_j.
  (q_j and a_j may depend on the history $q_1, a_1, \ldots, q_{j-1}, a_{j-1}$ and on $\mathbf{x}$.)

FIG. 2. *The sample accuracy game* $\mathsf{SampAcc}_{n,k,Q}[\mathcal{M}, \mathcal{A}]$.

2.3. Max-KL stability (a.k.a. differential privacy). Informally, an algorithm is "stable" if changing one of its inputs does not change its output too much. For our results, we will consider randomized algorithms and require that changing one input does not change the *distribution* of the algorithm's outputs too much. With this in mind, we will define here one notion of algorithmic stability that is related to the well-known notion of KL-divergence between distributions. In section 4.1, we will give other related notions of algorithmic stability based on different notions of closeness between distributions.

DEFINITION 2.3 (max-KL stability). *Let $\mathcal{W} : \mathcal{X}^n \to \mathcal{R}$ be a randomized algorithm. We say that $\mathcal{W}$ is* (ε, δ)-max-KL stable *if for every pair of samples $\mathbf{x}, \mathbf{x}'$ that differ on exactly one element, and every $R \subseteq \mathcal{R}$,*

$$\mathbb{P}\left[\mathcal{W}(\mathbf{x}) \in R\right] \leq e^{\varepsilon} \cdot \mathbb{P}\left[\mathcal{W}(\mathbf{x}') \in R\right] + \delta.$$

This notion of (ε, δ)-max-KL stability is also commonly known as (ε, δ)-*differential privacy* [DMNS06]; however, in this context we choose the term max-KL stability to emphasize the conceptual relationship between this notion and other notions of algorithmic stability that have been studied in machine learning. We also emphasize that our work has a very different motivation to the motivation of differential privacy—stable algorithms are desirable even when privacy is not a concern, such as when the data does not concern humans.

In our analysis, we will make crucial use of the fact that max-KL stability (as well as the other notions of stability discussed in section 4.1) is *closed under postprocessing*.

LEMMA 2.4 (postprocessing). *Let $\mathcal{W} : \mathcal{X}^n \to \mathcal{R}$ and $f : \mathcal{R} \to \mathcal{R}'$ be a pair of randomized algorithms. If $\mathcal{W}(\mathbf{x})$ is (ε, δ)-max-KL stable, then the algorithm $f(\mathcal{W}(\mathbf{x}))$ is (ε, δ)-max-KL stable.*

2.3.1. Stability for interactive mechanisms. The definition we gave above does not immediately apply to algorithms that interact with a data analyst to answer adaptively chosen queries. Such a mechanism does not simply take a sample $\mathbf{x}$ as input and produce an output. Instead, in the interactive setting, there is a mechanism $\mathcal{M}$ that holds a sample $\mathbf{x}$ and interacts with some algorithm $\mathcal{A}$. We can view this entire interaction between $\mathcal{M}$ and $\mathcal{A}$ as a single noninteractive meta algorithm that outputs the transcript of the interaction and define stability with respect to that meta algorithm. Specifically, we define the algorithm $\mathcal{W}[\mathcal{M}, \mathcal{A}](\mathbf{x})$ that simulates the interaction between $\mathcal{M}(\mathbf{x})$ and $\mathcal{A}$ and outputs the messages sent between them. The simulation is also parameterized by n, k, Q, although we will frequently omit these parameters when they are clear from context.

Input: A sample $\mathbf{x} \in \mathcal{X}^n$
For $j = 1, \ldots, k$
  Feed a_{j-1} to $\mathcal{A}$ and get a query $q_j \in Q$.
  Feed q_j to $\mathcal{M}(\mathbf{x})$ and get an answer $a_j \in \mathcal{R}$.
Output $((q_1, a_1), \ldots, (q_k, a_k))$.

Note that $\mathcal{W}[\mathcal{M}, \mathcal{A}]$ is a noninteractive mechanism, and its output is just the query-answer pairs of $\mathcal{M}$ and $\mathcal{A}$ in the sample accuracy game, subject to the mechanism being given the sample $\mathbf{x}$. Now we can define the stability of an interactive mechanism $\mathcal{M}$ using $\mathcal{W}$.

DEFINITION 2.5 (stability of for interactive mechanism). *We say an interactive mechanism $\mathcal{M}$ is* (ε, δ)-max-KL stable for k queries from Q *if for every adversary $\mathcal{A}$, the algorithm $\mathcal{W}_{n,k,Q}[\mathcal{M}, \mathcal{A}](\cdot) : \mathcal{X}^n \to (Q \times \mathcal{R})^k$ is (ε, δ)-max-KL stable.*

2.3.2. Composition of max-KL stability. The definition above allows for *adaptive composition*. This follows directly from composition results of (ε, δ)-*differentially private* algorithms. A mechanism that is (ε, δ)-max-KL stable for 1 query is $(\approx \varepsilon\sqrt{k}, \approx \delta k)$-stable for k adaptively chosen queries [DMNS06, DRV10]. More precisely, for every $0 \leq \varepsilon \leq 1$ and $\delta, \delta' > 0$, if a mechanism that is (ε, δ)-max-KL stable for 1 query is used to answer k adaptively chosen queries, it remains $(\varepsilon\sqrt{k \log(1/\delta')} + 2\varepsilon^2 k,\ \delta' + k\delta)$-max-KL stable [DRV10].

3. From max-KL stability to accuracy for low-sensitivity queries. In this section we prove our main result that any mechanism that both is accurate with respect to the sample and satisfies max-KL stability (with suitable parameters) is also accurate with respect to the population. The proof proceeds in two main steps. First, we prove a lemma that says that there is no max-KL stable mechanism that takes several independent sets of samples from the distribution and finds a query and a set of samples such that the answer to that query on that set of samples is very different from the answer to that query on the population. In section 3.2 we prove this lemma for the simpler case of statistical queries and then in section 3.3 we extend the proof to the more general case of low-sensitivity queries.

The second step is to introduce a *monitoring algorithm*. This monitoring algorithm will simulate the interaction between the mechanism and the adversary on multiple independent sets of samples. It will then output the least accurate query across all the different interactions. We show that if the mechanism is stable, then the monitoring algorithm is also stable. By choosing the number of sets of samples appropriately, we ensure that if the mechanism has even a small probability of being inaccurate in a given interaction, then the monitor will have a constant probability of finding an inaccurate query in one of the interactions. By the lemma proven in the first step, no such monitoring algorithm can satisfy max-KL stability; therefore every stable mechanism must be accurate with high probability.

3.1. Warmup: A single-sample decorrelated expectation lemma for statistical queries. As a warmup, in this section we give a simpler version of our main lemma for the case of statistical queries and a single sample. Although these results follow from the results of section 3.3 on general low-sensitivity queries, we include the simpler version to introduce the main ideas in the cleanest possible setting.

LEMMA 3.1. *Let $\mathcal{W} : \mathcal{X}^n \to Q$ be (ε, δ)-max-KL stable where Q is the class of statistical queries $q : \mathcal{X} \to [0,1]$. Let $\mathbf{P}$ be a distribution on $\mathcal{X}$ and let $\mathbf{x} \leftarrow_R \mathbf{P}^n$. Then*[7]

$$\left| \mathop{\mathbb{E}}_{\mathbf{x},\mathcal{W}} [q(\mathbf{P}) \ : \ q = \mathcal{W}(\mathbf{x})] - \mathop{\mathbb{E}}_{\mathbf{x},\mathcal{W}} [q(\mathbf{x}) \ : \ q = \mathcal{W}(\mathbf{x})] \right| \leq e^{\varepsilon} - 1 + \delta.$$

Proof of Lemma 3.1. Before giving the proof, we set up some notation. Let $\mathbf{x} = (x_1, \ldots, x_n)$. For a single element $x' \in \mathcal{X}$, and an index $i \in [n]$, we use $\mathbf{x}_{i \to x'}$ to denote the new sample where the ith element of $\mathbf{x}$ has been replaced by the element x'. Let $x' \leftarrow_{\mathrm{R}} \mathbf{P}$ be independent from $\mathbf{x}$.

We can now calculate

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathbf{x},\mathcal{W}} [q(\mathbf{x}) \ : \ q = \mathcal{W}(\mathbf{x})] \\
&\quad = \frac{1}{n} \sum_{i=1}^{n} \mathop{\mathbb{E}}_{\mathbf{x},\mathcal{W}} [q(x_i) \ : \ q = \mathcal{W}(\mathbf{x})] \\
&\quad = \frac{1}{n} \sum_{i=1}^{n} \int_0^1 \mathop{\mathbb{P}}_{\mathbf{x},\mathcal{W}} [q(x_i) \ > y \ : \ q = \mathcal{W}(\mathbf{x})] \, \mathrm{d}y.
\end{aligned}$$

Now we can apply max-KL stability:

$$\begin{aligned}
&\leq \frac{1}{n} \sum_{i=1}^{n} \int_0^1 e^{\varepsilon} \mathop{\mathbb{P}}_{\mathbf{x},\mathcal{W}} [q(x_i) \ > y \ : \ q = \mathcal{W}(\mathbf{x}_{i \to x'})] + \delta \mathrm{d}y \quad \text{(by } (\varepsilon, \delta)\text{-max-KL stability)} \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(e^{\varepsilon} \cdot \mathop{\mathbb{E}}_{x',\mathbf{x},\mathcal{W}} [q(x_i) \ : \ q = \mathcal{W}(\mathbf{x}_{i \to x'})] + \delta \right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(e^{\varepsilon} \cdot \mathop{\mathbb{E}}_{x',\mathbf{x},\mathcal{W}} [q(x') \ : \ q = \mathcal{W}(\mathbf{x})] + \delta \right) \quad \text{(the pairs } (x_i, \mathbf{x}_{i \to x'}) \\
&\quad \text{and } (x', \mathbf{x}) \text{ are identially distributed)} \\
&= e^{\varepsilon} \cdot \mathop{\mathbb{E}}_{x',\mathbf{x},\mathcal{W}} [q(x') \ : \ q = \mathcal{W}(\mathbf{x})] + \delta \\
&= e^{\varepsilon} \cdot \mathop{\mathbb{E}}_{\mathbf{x},\mathcal{W}} [q(\mathbf{P}) \ : \ q = \mathcal{W}(\mathbf{x})] + \delta.
\end{aligned}$$

An identical argument shows that

$$\mathop{\mathbb{E}}_{\mathbf{x},\mathcal{W}} [q(\mathbf{x}) \ : \ q = \mathcal{W}(\mathbf{x})] \geq e^{-\varepsilon} \cdot \left(\mathop{\mathbb{E}}_{\mathbf{x},\mathcal{W}} [q(\mathbf{P}) \ : \ q = \mathcal{W}(\mathbf{x})] - \delta \right).$$

Therefore, using the fact that $|q(\mathbf{P})| \leq 1$ for any statistical query q and distribution $\mathbf{P}$, we have

$$\left| \mathop{\mathbb{E}}_{\mathbf{x},\mathcal{W}} [q(\mathbf{P}) \ : \ q = \mathcal{W}(\mathbf{x})] - \mathop{\mathbb{E}}_{\mathbf{x},\mathcal{W}} [q(\mathbf{x}) \ : \ q = \mathcal{W}(\mathbf{x})] \right| \leq e^{\varepsilon} - 1 + \delta,$$

as desired. ∎

[7] The notation $\mathbb{E}_{\mathbf{x},\mathcal{W}}[q(\mathbf{P}) \ : \ q = \mathcal{W}(\mathbf{x})]$ should be read as "the expectation of $q(\mathbf{P})$, where q denotes the output of $\mathcal{W}(\mathbf{x})$." A more standard (but less readable) notation for the same thing would be $\mathbb{E}_{\substack{\mathbf{x},\mathcal{W} \\ q = \mathcal{W}(\mathbf{x})}} [q(\mathbf{P})]$.

3.2. Warmup: A multisample decorrelated expectation lemma for statistical queries. As a second warmup, in this section we give a simpler version of our main lemma for the case of statistical queries and *multiple* samples. That is, we consider a setting where there are many subsamples available to the algorithm. The multi-sample decorrelated expectation lemma says that a max-KL stable algorithm cannot take a collection of samples $\mathbf{x}_1, \ldots, \mathbf{x}_T$ and output a pair (q, t) such that $q(\mathbf{P})$ and $q(\mathbf{x}_t)$ differ significantly in expectation.

LEMMA 3.2. *Let $\mathcal{W} : (\mathcal{X}^n)^T \to Q \times [T]$ be (ε, δ)-max-KL stable where Q is the class of statistical queries $q : \mathcal{X} \to [0, 1]$. Let $\mathbf{P}$ be a distribution on $\mathcal{X}$ and let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_T) \leftarrow_R (\mathbf{P}^n)^T$. Then*

$$\left| \mathop{\mathbb{E}}_{\mathbf{X},\mathcal{W}} [q(\mathbf{P}) \ : \ (q, t) = \mathcal{W}(\mathbf{X})] - \mathop{\mathbb{E}}_{\mathbf{X},\mathcal{W}} [q(\mathbf{x}_t) \ : \ (q, t) = \mathcal{W}(\mathbf{X})] \right| \leq e^{\varepsilon} - 1 + T\delta.$$

Proof of Lemma 3.2. Before giving the proof, we set up some notation. Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$ be a set of T samples where each sample $\mathbf{x}_t = (x_{t,1}, \ldots, x_{t,n})$. For a single element $x' \in \mathcal{X}$, and a pair of indices $(m, i) \in [T] \times [n]$, we use $\mathbf{X}_{(m,i)\to x'}$ to denote the new set of T samples where the ith element of the mth sample of $\mathbf{X}$ has been replaced by the element x'.

We can now calculate

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathbf{X},\mathcal{W}} [q(\mathbf{x}_t) \ : \ (q, t) = \mathcal{W}(\mathbf{X})] \\
&\quad = \sum_{m=1}^{T} \mathop{\mathbb{E}}_{\mathbf{X},\mathcal{W}} \left[\mathbf{1}_{\{t=m\}} \cdot q(\mathbf{x}_m) \ : \ (q, t) = \mathcal{W}(\mathbf{X}) \right] \\
&\quad = \frac{1}{n} \sum_{i=1}^{n} \sum_{m=1}^{T} \mathop{\mathbb{E}}_{\mathbf{X},\mathcal{W}} \left[\mathbf{1}_{\{t=m\}} \cdot q(x_{m,i}) \ : \ (q, t) = \mathcal{W}(\mathbf{X}) \right] \\
&\quad = \frac{1}{n} \sum_{i=1}^{n} \sum_{m=1}^{T} \int_0^1 \mathop{\mathbb{P}}_{\mathbf{X},\mathcal{W}} \left[\mathbf{1}_{\{t=m\}} \cdot q(x_{m,i}) \ \geq y \ : \ (q, t) = \mathcal{W}(\mathbf{X}) \right] \mathrm{d}y.
\end{aligned}$$

Now we can apply (ε, δ)-max-KL stability.

$$\begin{aligned}
&\leq \frac{1}{n} \sum_{i=1}^{n} \sum_{m=1}^{T} \left(\int_0^1 e^{\varepsilon} \mathop{\mathbb{P}}_{\mathbf{X},\mathcal{W}} \left[\mathbf{1}_{\{t=m\}} \cdot q(x_{m,i}) \ \geq y \ : \ (q, t) = \mathcal{W}(\mathbf{X}_{(m,i)\to x'}) \right] + \delta \right) \mathrm{d}y \\
&\quad \text{(by } (\varepsilon, \delta)\text{-max-KL stability)} \\
&= \frac{1}{n} \sum_{i=1}^{n} \sum_{m=1}^{T} \left(e^{\varepsilon} \cdot \mathop{\mathbb{E}}_{x',\mathbf{X},\mathcal{W}} \left[\mathbf{1}_{\{t=m\}} \cdot q(x_{m,i}) \ : \ (q, t) = \mathcal{W}(\mathbf{X}_{(m,i)\to x'}) \right] + \delta \right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \sum_{m=1}^{T} \left(e^{\varepsilon} \cdot \mathop{\mathbb{E}}_{x',\mathbf{X},\mathcal{W}} \left[\mathbf{1}_{\{t=m\}} \cdot q(x') \ : \ (q, t) = \mathcal{W}(\mathbf{X}) \right] + \delta \right) \\
&\quad \text{the pairs } (x_{m,i}, \mathbf{X}_{(m,i)\to x'}) \text{ and } (x', \mathbf{X}) \text{ are identially distributed} \\
&= e^{\varepsilon} \cdot \mathop{\mathbb{E}}_{x',\mathbf{X},\mathcal{W}} [q(x') \ : \ (q, t) = \mathcal{W}(\mathbf{X})] + T\delta \\
&= e^{\varepsilon} \cdot \mathop{\mathbb{E}}_{\mathbf{X},\mathcal{W}} [q(\mathbf{P}) \ : \ (q, t) = \mathcal{W}(\mathbf{X})] + T\delta \\
&\leq \mathop{\mathbb{E}}_{\mathbf{X},\mathcal{W}} [q(\mathbf{P}) \ : \ (q, t) = \mathcal{W}(\mathbf{X})] + e^{\varepsilon} - 1 + T\delta \qquad (\text{since } q(\mathbf{P}) \in [0, 1]).
\end{aligned}$$

An identical argument shows that

$$\mathop{\mathbb{E}}_{\mathbf{X},\mathcal{W}}[q(\mathbf{x}_t) \ : \ (q,t) = \mathcal{W}(\mathbf{X})] \geq \mathop{\mathbb{E}}_{\mathbf{X},\mathcal{W}}[q(\mathbf{P}) \ : \ (q,t) = \mathcal{W}(\mathbf{X})] + (e^{-\varepsilon} - 1) - T\delta. \qquad \square$$

3.3. A multisample decorrelated expectation lemma. Here, we give the most general decorrelated expectation lemma that considers multiple samples and applies to the more general class of low-sensitivity queries.

LEMMA 3.3 (main technical lemma). *Let $\mathcal{W} : (\mathcal{X}^n)^T \to Q_\Delta \times [T]$ be (ε,δ)-max-KL stable where Q_Δ is the class of Δ-sensitive queries $q : \mathcal{X}^n \to \mathbb{R}$. Let $\mathbf{P}$ be a distribution on $\mathcal{X}$ and let $\mathbf{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_T) \leftarrow_R (\mathbf{P}^n)^T$. Then*

$$\left| \mathop{\mathbb{E}}_{\mathbf{X},\mathcal{W}}[q(\mathbf{P}) \ : \ (q,t) = \mathcal{W}(\mathbf{X})] - \mathop{\mathbb{E}}_{\mathbf{X},\mathcal{W}}[q(\mathbf{x}_t) \ : \ (q,t) = \mathcal{W}(\mathbf{X})] \right| \leq 2(e^{\varepsilon} - 1 + T\delta)\Delta n.$$

We remark that if we use the weaker assumption that $\mathcal{W}$ is $(e^{\varepsilon} - 1 + \delta)$-TV stable (defined in section 4.1), then we would obtain the same conclusion but with the weaker bound of $2T(e^{\varepsilon} - 1 + \delta)\Delta n$. The advantage of using the stronger definition of max-KL stability is that we only have to decrease δ with T and not ε. This advantage is crucial because algorithms satisfying (ε,δ)-max-KL stability necessarily have a linear dependence on $1/\varepsilon$ but only a polylogarithmic dependence on $1/\delta$.

Proof of Lemma 3.3. Let $\mathbf{X}' = (\mathbf{x}'_1,\ldots,\mathbf{x}'_T) \leftarrow_{\mathrm{R}} (\mathbf{P}^n)^T$ be independent of $\mathbf{X}$. Recall that each element $\mathbf{x}_t$ of $\mathbf{X}$ is itself a vector $(x_{t,1},\ldots,x_{t,n})$, and the same is true for each element $\mathbf{x}'_t$ of $\mathbf{X}'$. We will sometimes refer to the vectors $\mathbf{x}_1,\ldots,\mathbf{x}_T$ as the *subsamples of* $\mathbf{X}$.

We define a sequence of intermediate samples that allow us to interpolate between $\mathbf{X}$ and $\mathbf{X}'$ using a series of neighboring samples. Formally, for $\ell \in \{0,1,\ldots,n\}$ and $m \in \{0,1,\ldots,T\}$, define $\mathbf{X}^{\ell,m} = (\mathbf{x}_1^{\ell,m},\ldots,\mathbf{x}_T^{\ell,m}) \in (\mathcal{X}^n)^T$ by

$$x_{t,i}^{\ell,m} = \begin{cases} x_{t,i}, & (t > m) \text{ or } (t = m \text{ and } i > \ell), \\ x'_{t,i}, & (t < m) \text{ or } (t = m \text{ and } i \leq \ell). \end{cases}$$

By construction we have $\mathbf{X}^{0,1} = \mathbf{X}^{n,0} = \mathbf{X}$ and $\mathbf{X}^{n,T} = \mathbf{X}'$. Also $\mathbf{X}^{0,m} = \mathbf{X}^{n,m-1}$ for $m \in [T]$. Moreover, pairs $(\mathbf{X}^{\ell,t}, \mathbf{X}^{\ell-1,t})$ are neighboring in the sense that there is a single subsample, $\mathbf{x}_t$ such that $\mathbf{x}_t^{\ell,t}$ and $\mathbf{x}_t^{\ell-1,T}$ are neighbors and for every $t' \neq t$, $\mathbf{x}_{t'}^{\ell,t} = \mathbf{x}_{t'}^{\ell-1,t}$.

For $\ell \in [n]$ and $m \in [T]$, define a randomized function $B^{\ell,m} : (\mathcal{X}^n)^T \times (\mathcal{X}^n)^T \to \mathbb{R}$ by

$$B^{\ell,m}(\mathbf{X},\mathbf{Z}) = \begin{cases} q(\mathbf{z}_t) - q(\mathbf{z}_{t,-\ell}) + \Delta, & t = m, \\ 0, & t \neq m, \end{cases} \quad \text{where } (q,t) = \mathcal{W}(\mathbf{X}),$$

where $\mathbf{z}_{t,-\ell}$ is the tth subsample of $\mathbf{Z}$ with its ℓth element replaced by some arbitrary fixed element of $\mathcal{X}$.

We can now expand $|\mathop{\mathbb{E}}_{\mathbf{X},\mathcal{W}}[q(\mathbf{P}) - q(\mathbf{x}_t) \ : \ (q,t) = \mathcal{W}(\mathbf{X})]|$ in terms of these intermediate samples and the functions $B^{\ell,m}$:

$$
\begin{aligned}
&\left|\mathop{\mathbb{E}}_{\mathbf{X},\mathcal{W}}\left[q(\mathbf{P}) - q(\mathbf{x}_t) \ : \ (q,t) = \mathcal{W}(\mathbf{X})\right]\right| \\
&\quad = \left|\mathop{\mathbb{E}}_{\mathbf{X},\mathbf{X}',\mathcal{W}}\left[q(\mathbf{x}'_t) - q(\mathbf{x}_t) \ : \ (q,t) = \mathcal{W}(\mathbf{X})\right]\right| \\
&\quad = \left|\sum_{\ell\in[n]}\sum_{m\in[T]}\mathop{\mathbb{E}}_{\mathbf{X},\mathbf{X}',\mathcal{W}}\left[q(\mathbf{x}_t^{\ell,m}) - q(\mathbf{x}_t^{\ell-1,m}) \ : \ (q,t) = \mathcal{W}(\mathbf{X})\right]\right| \\
&\quad =\leq \sum_{\ell\in[n]}\left|\sum_{m\in[T]}\mathop{\mathbb{E}}_{\mathbf{X},\mathbf{X}',\mathcal{W}}\left[q(\mathbf{x}_t^{\ell,m}) - q(\mathbf{x}_t^{\ell-1,m}) \ : \ (q,t) = \mathcal{W}(\mathbf{X})\right]\right| \\
&\quad = \sum_{\ell\in[n]}\left|\sum_{m\in[T]}\mathop{\mathbb{E}}_{\mathbf{X},\mathbf{X}',\mathcal{W}}\Big[\Big(q(\mathbf{x}_t^{\ell,m}) - q(\mathbf{x}_{t,-\ell}^{\ell,m}) + \Delta\Big)\right. \\
&\qquad \left. - \Big(q(\mathbf{x}_t^{\ell-1,m}) - q(\mathbf{x}_{t,-\ell}^{\ell-1,m}) + \Delta\Big) : (q,t) = \mathcal{W}(\mathbf{X})\Big]\right| \\
&\qquad \text{(By construction, } \mathbf{x}_{t,-\ell}^{\ell,m} = \mathbf{x}_{t,-\ell}^{\ell-1,m}\text{)} \\
&\quad = \sum_{\ell\in[n]}\left|\sum_{m\in[T]}\mathop{\mathbb{E}}_{\mathbf{X},\mathbf{X}',\mathcal{W}}\left[B^{\ell,m}(\mathbf{X},\mathbf{X}^{\ell,m}) - B^{\ell,m}(\mathbf{X},\mathbf{X}^{\ell-1,m})\right]\right| \qquad \text{(Definition of } B^{\ell,m}\text{)}.
\end{aligned}
$$

Thus, it suffices to show that $\left|\sum_{m\in[T]}\mathbb{E}\left[B^{\ell,m}(\mathbf{X},\mathbf{X}^{\ell,m}) - B^{\ell,m}(\mathbf{X},\mathbf{X}^{\ell-1,m})\right]\right| \leq 2(e^{\varepsilon} - 1 + T\delta)\Delta$ for all $\ell \in [n]$. To this end, we make a few observations.

1. Since q is Δ-sensitive, for every $\ell, m, \mathbf{X}, \mathbf{Z}$, we have $0 \leq B^{\ell,m}(\mathbf{X},\mathbf{Z}) \leq 2\Delta$. Moreover, since $B^{\ell,m}(\mathbf{X},\mathbf{Z}) = 0$ whenever $\mathcal{W}(\mathbf{X})$ outputs (q,t) with $t \neq m$, we have $\sum_{m\in[T]}\mathbb{E}\left[B^{\ell,m}(\mathbf{x},\mathbf{x}^{\ell,m})\right] \leq 2\Delta$.
2. By construction, $B^{\ell,m}(\mathbf{X},\mathbf{Z})$ is (ε,δ)-max-KL stable as a function of its first parameter $\mathbf{X}$. Stability follows by the postprocessing lemma (Lemma 2.4) since $B^{\ell,m}$ is a postprocessing of the output of $\mathcal{W}(\mathbf{X})$, which is assumed to be (ε,δ)-max-KL stable.
3. Last, observe that the random variables $\mathbf{X}^{\ell,m}$ are identically distributed (although they are not independent). Namely, each one consists of nT independent samples from $\mathbf{P}$. Moreover, for every ℓ and m, the pair $(\mathbf{X}^{\ell,m},\mathbf{X})$ has the same distribution as $(\mathbf{X},\mathbf{X}^{\ell,m})$. Specifically, the first component is nT independent samples from $\mathbf{P}$ and the second component is equal to the first component with a subset of the entries replaced by fresh independent samples from $\mathbf{P}$.

Consider the random variables $B^{\ell,m}(\mathbf{X},\mathbf{X}^{\ell,m})$ and $B^{\ell,m}(\mathbf{X},\mathbf{X}^{\ell-1,m})$ for some $\ell \in [n]$ and $m \in [T]$. Using observations 2 and 3, we have

$$B^{\ell,m}(\mathbf{X},\mathbf{X}^{\ell,m}) \sim B^{\ell,m}(\mathbf{X}^{\ell,m},\mathbf{X}) \sim_{(\varepsilon,\delta)} B^{\ell,m}(\mathbf{X}^{\ell-1,m},\mathbf{X}) \sim B^{\ell,m}(\mathbf{X},\mathbf{X}^{\ell-1,m}),$$

where $\sim$ denotes having the same distribution and $\sim_{(\varepsilon,\delta)}$ denotes having (ε,δ)-max-KL close distributions.[8] Thus $B^{\ell,m}(\mathbf{X},\mathbf{X}^{\ell-1,m})$ and $B^{\ell,m}(\mathbf{X},\mathbf{X}^{\ell,m})$ are (ε,δ)-max-KL close.

[8] In the spirit of (ε,δ)-max-KL stability, we say that distributions A and B over $\mathcal{R}$ are (ε,δ)-max-KL close if for every $R \subseteq \mathcal{R}$, $\mathbb{P}[A \in R] \leq e^{\varepsilon}\cdot\mathbb{P}[B \in R] + \delta$ and vice versa.

Now we can calculate

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathbf{X},\mathbf{X}',\mathcal{W}}\left[B^{\ell,m}(\mathbf{X},\mathbf{X}^{\ell-1,m})\right] &= \int_0^{2\Delta} \mathop{\mathbb{P}}_{\mathbf{X},\mathbf{X}',\mathcal{W}}\left[B^{\ell,m}(\mathbf{X},\mathbf{X}^{\ell-1,m}) \geq y\right] \mathrm{d}y \\ &\leq \int_0^{2\Delta} \left(e^{\varepsilon} \cdot \mathop{\mathbb{P}}_{\mathbf{X},\mathbf{X}',\mathcal{W}}\left[B^{\ell,m}(\mathbf{X},\mathbf{X}^{\ell,m}) \geq y\right] + \delta\right) \mathrm{d}y \\ &= e^{\varepsilon} \cdot \int_0^{2\Delta} \mathop{\mathbb{P}}_{\mathbf{X},\mathbf{X}',\mathcal{W}}\left[B^{\ell,m}(\mathbf{X},\mathbf{X}^{\ell,m}) \geq y\right] \mathrm{d}y + 2\delta\Delta \\ &= e^{\varepsilon} \cdot \mathop{\mathbb{E}}_{\mathbf{X},\mathbf{X}',\mathcal{W}}\left[B^{\ell,m}(\mathbf{X},\mathbf{X}^{\ell,m})\right] + 2\delta\Delta.\end{aligned}$$

Thus we have

$$\begin{aligned}&\sum_{m\in[T]} \mathop{\mathbb{E}}_{\mathbf{X},\mathbf{X}',\mathcal{W}}\left[B^{\ell,m}(\mathbf{X},\mathbf{X}^{\ell-1,m})\right] \\ &\leq e^{\varepsilon} \cdot \left(\sum_{m\in[T]} \mathop{\mathbb{E}}_{\mathbf{X},\mathbf{X},\mathcal{W}}\left[B^{\ell,m}(\mathbf{X},\mathbf{X}^{\ell,m})\right]\right) + 2T\delta\Delta \\ &\leq \sum_{m\in[T]} \mathop{\mathbb{E}}_{\mathbf{X},\mathbf{X}',\mathcal{W}}\left[B^{\ell,m}(\mathbf{X},\mathbf{X}^{\ell,m})\right] + 2(e^{\varepsilon}-1)\Delta + 2\Delta T\delta.\end{aligned}$$

Thus we have the desired upper bound on the expectation of $\sum_{m\in[T]} \mathbb{E}[B^{\ell,m}(\mathbf{X},\mathbf{X}^{\ell,m}) - B^{\ell,m}(\mathbf{X},\mathbf{X}^{\ell-1,m})]$. The corresponding lower bound follows from an analogous argument. This completes the proof. ∎

3.4. From multisample decorrelated expectation to accuracy. Now that we have Lemma 3.3, we can prove the following result that max-KL stable mechanisms that are also accurate with respect to their sample are also accurate with respect to the population from which that sample was drawn.

THEOREM 3.4 (main transfer theorem). *Let Q be a family of Δ-sensitive queries on $\mathcal{X}$. Assume that, for some $\alpha, \beta \in (0, .1)$, $\mathcal{M}$ is*

1. *$(\varepsilon = \alpha/64\Delta n, \delta = \alpha\beta/32\Delta n)$-max-KL stable for k adaptively chosen queries from Q and*
2. *$(\alpha' = \alpha/8, \beta' = \alpha\beta/16\Delta n)$-accurate with respect to its sample for n samples from $\mathcal{X}$ for k adaptively chosen queries from Q.*

Then $\mathcal{M}$ is (α, β)-accurate with respect to the population for k adaptively chosen queries from Q given n samples from $\mathcal{X}$.

The key step in the proof is to define a monitoring algorithm that takes T separate samples $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$ and for each sample $\mathbf{x}_t$ simulates an independent interaction between $\mathcal{M}(\mathbf{x}_t)$ and $\mathcal{A}$. This monitoring algorithm then outputs the query with the largest error across all of the queries and interactions (kT queries in total). Since changing one input to $\mathbf{X}$ affects only one of the simulations, the monitoring algorithm will be stable so long as $\mathcal{M}$ is stable, without any loss in the stability parameter. On the other hand, if $\mathcal{M}$ has even a small chance β of answering a query with large error, then if we simulate $T \approx 1/\beta$ independent interactions, there is a constant probability that at least one of the simulations results in a query with large error. Thus, the monitor will be a stable algorithm that outputs a query with large error *in expectation*. By the multisample decorrelated expectation lemma, such a monitor is impossible, which implies that $\mathcal{M}$ has probability $\leq \beta$ of answering any query with large error.

Proof of Theorem 3.4. Let $\mathcal{M}$ be an interactive mechanism. Let $\mathcal{A}$ be an analyst and let $\mathbf{P}$ be the distribution chosen by $\mathcal{A}$. We define the following monitoring algorithm.

$$\mathcal{W}(\mathbf{X}) = \mathcal{W}_{\mathbf{P}}[\mathcal{M}, \mathcal{A}](\mathbf{X}) :$$

Input: $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_T) \in (\mathcal{X}^n)^T$

For $t = 1, \ldots, T$:

Simulate $\mathcal{M}(\mathbf{x}_t)$ and $\mathcal{A}$ interacting, let $q_{t,1}, \ldots, q_{t,k} \in Q$ be the queries of $\mathcal{A}$ and let

$a_{t,1}, \ldots, a_{t,k} \in \mathbb{R}$ be the corresponding answers of $\mathcal{M}$.

Let

$$(j^*, t^*) = \underset{j \in [k],\, t \in [T]}{\arg\max} \left| \mathrm{err}^{\mathbf{P}}\left(q_{t,j}, a_{t,j}\right) \right|.$$

If $a_{t^*,j^*} - q_{t^*,j^*}(\mathbf{P}) \geq 0$, let $q^* = q_{t^*,j^*}$, otherwise let $q^* = -q_{t^*,j^*}$. (Q_Δ is closed under negation.)

Output: (q^*, t^*).

If $\mathcal{M}$ is stable, then so is $\mathcal{W}$, and this fact follows easily from the postprocessing lemma (Lemma 2.4).

Claim 3.5. *For every $\varepsilon, \delta \geq 0$, if the mechanism $\mathcal{M}$ is (ε, δ)-max-KL stable for k adaptively chosen queries from Q, then for every $\mathbf{P}$ and $\mathcal{A}$, the monitor $\mathcal{W}_{\mathbf{P},k,Q}[\mathcal{M}, \mathcal{A}]$ is (ε, δ)-max-KL stable.*

Proof. If $\mathcal{M}$ is (ε, δ)-max-KL stable for k adaptively chosen queries from Q, then for every analyst $\mathcal{A}$ who asks k queries from Q, and every t, the algorithm $\mathcal{W}'(\mathbf{x}_t)$ that simulates the interaction between $\mathcal{M}(\mathbf{x}_t)$ and $\mathcal{A}$ and outputs the resulting query-answer pairs is (ε, δ)-max-KL stable. From this, it follows that the algorithm $\mathcal{W}'(\mathbf{X})$ that simulates the interactions between $\mathcal{M}(\mathbf{x}_t)$ and $\mathcal{A}$ for every $t = 1, \ldots, T$ and outputs the resulting query-answer pairs is (ε, δ)-max-KL stable. To see this, observe that if $\mathbf{X}, \mathbf{X}'$ differ only on one subsample $\mathbf{x}_t$, then for every $t' \neq t$, $\mathbf{x}_{t'} = \mathbf{x}'_{t'}$ and thus the query-answer pairs corresponding to subsample t' are identically distributed regardless of whether we use $\mathbf{X}$ or $\mathbf{X}'$ as input to $\mathcal{W}$.

Observe that the algorithm $\mathcal{W}$ defined above is simply a postprocessing of these kT query-answer pairs. That is, (q^*, t^*) depends only on $\{(q_{t,j}, a_{t,j}\}_{t \in [T], j \in [k]}$ and $\mathbf{P}$, and not on $\mathbf{X}$. Thus, by Lemma 2.4, $\mathcal{W}$ is (ε, δ)-max-KL stable. ∎

We will use the $\mathcal{W}$ with $T = \lfloor 1/\beta \rfloor$. In light of Claim 3.5 and our assumption that $\mathcal{M}$ is (ε, δ)-max-KL stable, we can apply Lemma 3.3 to obtain

$$\left| \mathop{\mathbb{E}}_{\mathbf{X}, \mathcal{W}} \left[q^*(\mathbf{P}) - q^*(\mathbf{x}_{t^*}) \;:\; (q^*, t^*) = \mathcal{W}(\mathbf{X}) \right] \right| \leq 2\left(e^{\alpha/64\Delta n} - 1 + T\left(\frac{\alpha\beta}{32\Delta n}\right)\right)\Delta n \leq \alpha/8. \tag{1}$$

To complete the proof, we show that if $\mathcal{M}$ is not (α, β)-accurate with respect to the population $\mathbf{P}$, then (1) cannot hold. To do so, we need the following natural claim about the output of the monitor.

Claim 3.6. $\mathbb{P}_{\mathbf{X}, \mathcal{W}}\left[q^*(\mathbf{P}) - a_{q^*} > \alpha\right] > 1 - (1-\beta)^T$, *and* $q^*(\mathbf{P}) - a_{q^*} \geq 0$, *where a_{q^*} is the answer to q^* produced during the simulation.*

Proof. Since $\mathcal{M}$ fails to be (α, β)-accurate, for every $t \in [T]$,

$$\mathop{\mathbb{P}}_{\mathbf{x}_t, \mathcal{M}} \left[\max_{j \in [k]} |q_{t,j}(\mathbf{P}) - a_{t,j}| > \alpha \right] > \beta. \tag{2}$$

We obtain the claim from (2) by using the fact that the T sets of query-answer pairs corresponding to different subsamples $\mathbf{x}_1, \ldots, \mathbf{x}_T$ are independent. That is, the random variables $\max_{j \in [k]} |q_{t,j}(\mathbf{P}) - a_{t,j}|$ indexed by $t \in [T]$ are independent. Since $q^*(\mathbf{P}) - a_{q^*}$ is simply the maximum of these independent random variables, the first part of the claim follows. Also, by construction, $\mathcal{W}$ ensures that

$$q^*(\mathbf{P}) - a_{q^*} \geq 0. \tag{3}$$

∎

Claim 3.7. *If $\mathcal{M}$ is (α', β')-accurate for the sample but not (α, β)-accurate for the population, then*

$$\left| \mathop{\mathbb{E}}_{\mathbf{X}, \mathcal{W}} [q^*(\mathbf{P}) - q^*(\mathbf{x}_{t^*}) \ : \ (q^*, t^*) = \mathcal{W}(\mathbf{X})] \right| \geq \alpha/4.$$

Proof. Now we can calculate

$$\begin{aligned}
&\left| \mathop{\mathbb{E}}_{\mathbf{X}, \mathcal{W}} [q^*(\mathbf{P}) - q^*(\mathbf{x}_{t^*}) \ : \ (q^*, t^*) = \mathcal{W}(\mathbf{X})] \right| \\
&= \left| \mathop{\mathbb{E}}_{\mathbf{X}, \mathcal{W}} [q^*(\mathbf{P}) - a_{q^*} \ : \ (q^*, t^*) = \mathcal{W}(\mathbf{X})] + \mathop{\mathbb{E}}_{\mathbf{X}, \mathcal{W}} [a_{q^*} - q^*(\mathbf{x}_{t^*}) \ : \ (q^*, t^*) = \mathcal{W}(\mathbf{X})] \right| \\
&\geq \left| \mathop{\mathbb{E}}_{\mathbf{X}, \mathcal{W}} [q^*(\mathbf{P}) - a_{q^*} \ : \ (q^*, t^*) = \mathcal{W}(\mathbf{x})] \right| - \left| \mathop{\mathbb{E}}_{\mathbf{X}, \mathcal{W}} [a_{q^*} - q^*(\mathbf{x}_{t^*}) \ : \ q^* = \mathcal{W}(\mathbf{x})] \right| \\
&\geq \alpha(1 - (1 - \beta)^T) - \left| \mathop{\mathbb{E}}_{\mathbf{X}, \mathcal{W}} [a_{q^*} - q^*(\mathbf{x}_{t^*}) \ : \ (q^*, t^*) = \mathcal{W}(\mathbf{X})] \right| \qquad \text{(Claim 3.6)} \\
&\geq \alpha(1 - (1 - \beta)^T) - \left(\alpha/8 + 2T \left(\frac{\alpha\beta}{16\Delta n} \right) \Delta n \right) \\
&\geq \alpha/2 - (\alpha/8 + \alpha/8) = \alpha/4 \qquad (T = \lfloor 1/\beta \rfloor).
\end{aligned} \tag{4}$$

(5)

(6)

Line (5) follows from two observations. First, since $\mathcal{M}$ is assumed to be $(\alpha/8, \alpha\beta/16\Delta n)$-accurate for one sample, by a union bound, it is simultaneously $(\alpha/8, T(\alpha\beta/16\Delta n))$-accurate for all of the T samples. Thus, we have $a_{q^*} - q^*(\mathbf{x}_{t^*}) \leq \alpha'$ except with probability at most $T(\alpha\beta/16\Delta n)$. Second, since q^* is a Δ-sensitive query, we always have $a_{q^*} - q^*(\mathbf{x}_{t^*}) \leq 2\Delta n$.[9] ∎

Thus, if $\mathcal{M}$ is not (α, β)-accurate for the population, we will obtain a contradiction to (1). This completes the proof. ∎

4. Other notions of stability and accuracy on average. Definition 4.2 gives one notion of stability, namely max-KL stability. However, this is by no means the only way to formalize stability for our purposes. In this section we consider other notions of stability and the advantages they have.

[9] Without loss of generality, the answers of $\mathcal{M}$ can be truncated to an interval of width $2\Delta n$ that contains the correct answer $q^*(\mathbf{x}_{t^*})$. Doing so will ensure $|a_{q^*} - q^*(\mathbf{x}_{t^*})| \leq 2\Delta n$.

4.1. Other notions of algorithmic stability. We will define here other notions of algorithmic stability, and in section 4.2, we will show that such notions can provide expected guarantees for generalization error which can be used to achieve accuracy on average.

DEFINITION 4.1 (TV stability). *Let $\mathcal{W} : \mathcal{X}^n \to \mathcal{R}$ be a randomized algorithm. We say that $\mathcal{W}$ is* ε-TV stable *if for every pair of samples that differ on exactly one element,*

$$\mathrm{d_{TV}}(\mathcal{W}(\mathbf{x}), \mathcal{W}(\mathbf{x}')) = \sup_{R \subseteq \mathcal{R}} \left| \mathbb{P}\left[\mathcal{W}(\mathbf{x}) \in R\right] - \mathbb{P}\left[\mathcal{W}(\mathbf{x}') \in R\right] \right| \leq \varepsilon.$$

DEFINITION 4.2 (KL stability). *Let $\mathcal{W} : \mathcal{X}^n \to \mathcal{R}$ be a randomized algorithm. We say that $\mathcal{W}$ is* ε-KL stable *if for every pair of samples $\mathbf{x}, \mathbf{x}'$ that differ on exactly one element,*

$$\mathop{\mathbb{E}}_{r \leftarrow_R \mathcal{W}(\mathbf{x})} \left[\log \left(\frac{\mathbb{P}\left[\mathcal{W}(\mathbf{x}) = r\right]}{\mathbb{P}\left[\mathcal{W}(\mathbf{x}') = r\right]} \right) \right] \leq 2\varepsilon^2.$$

The postprocessing property of max-KL stability (Lemma 2.4 in section 2.3) also applies to the two stability notions above.

LEMMA 4.3 (stability notions preserved under postprocessing). *Let $\mathcal{W} : \mathcal{X}^n \to \mathcal{R}$ and $f : \mathcal{R} \to \mathcal{R}'$ be a pair of randomized algorithms. If $\mathcal{W}$ is {ε-TV, ε-KL, (ε, δ)-max-KL}-stable, then the algorithm $f(\mathcal{W}(\mathbf{x}))$ is {ε-TV, ε-KL, (ε, δ)-max-KL}stable.*

Relationships between stability notions. ε-KL stability implies ε-TV stability by Pinsker's inequality. Therefore the generalization result we prove for ε-TV stable algorithms (Theorem 4.7) apply equally to ε-KL stable algorithms. The relationship between max-KL stability defined in section 2.3 and the above notions is more subtle. When $\varepsilon \leq 1$, $(\varepsilon, 0)$-max-KL stability implies ε-KL stability and thus also ε-TV stability. When $\varepsilon \leq 1$ and $\delta > 0$, (ε, δ)-max-KL stability implies $(2\varepsilon + \delta)$-TV stability. It also implies that $\mathcal{M}$ is "close" to satisfying 2ε-KL stability (cf. [DRV10] for more discussion of these notions).

As in section 2.3.1, we define TV stability and KL stability of an interactive mechanism $\mathcal{M}$ through a noninteractive mechanism that simulates the interaction between $\mathcal{M}$ and an adversary $\mathcal{A}$. The definition for these notions of stability is precisely analogous to Definition 2.5 for max-KL stability.

As with max-KL stability, both notions above allow for *adaptive composition.* In fact, ε-TV stability composes linearly—a mechanism that is ε-TV stable for one query is εk-stable for k queries. The advantage of the stronger notions of KL and max-KL stability is that they have a stronger composition. A mechanism that is ε-KL stable for one query is $(\varepsilon\sqrt{k})$-stable for k queries.

4.2. From TV stability to accuracy on average. In this section we show that TV stable algorithms guarantee a weaker notion of accuracy on average for adaptively chosen queries.

4.3. Accuracy on average. In section 2.2 we defined accurate mechanisms to be those that answer accurately (with respect to either the population or the sample) with probability close to 1. In this section we define a relaxed notion of accuracy that only requires low error in expectation over the coins of $\mathcal{M}$ and $\mathcal{A}$.

DEFINITION 4.4 (average accuracy). *A mechanism $\mathcal{M}$ is* α-accurate on average with respect to the population for k adaptively chosen queries from Q given n samples in $\mathcal{X}$ *if for every adversary $\mathcal{A}$,*

$$\mathop{\mathbb{E}}_{\mathsf{Acc}_{n,k,Q}[\mathcal{M},\mathcal{A}]}\left[\max_{j\in[k]} \left|\mathrm{err}^{\mathbf{P}}\left(q_j, a_j\right)\right|\right] \leq \alpha.$$

We will also use a definition of accuracy relative to the sample given to the mechanism.

DEFINITION 4.5 (sample accuracy on average). *A mechanism $\mathcal{M}$ is* α-accurate on average with respect to samples of size n from $\mathcal{X}$ for k adaptively chosen queries from Q *if for every adversary* $\mathcal{A}$,

$$\mathop{\mathbb{E}}_{\mathsf{SampAcc}_{n,k,Q}[\mathcal{M},\mathcal{A}]}\left[\max_{j\in[k]} \left|\mathrm{err}_{\mathbf{x}}\left(q_j, a_j\right)\right|\right] \leq \alpha.$$

4.3.1. A decorrelated expectation lemma. Toward our goal of proving that TV stability implies accuracy on average in the adaptive setting, we first prove a lemma saying that TV stable algorithms cannot output a low-sensitivity query such that the sample has large error for that query. In the next section we will show how this lemma implies accuracy on average in the adaptive setting.

LEMMA 4.6. *Let $\mathcal{W} : \mathcal{X}^n \to Q_\Delta$ be an ε-TV stable randomized algorithm. Recall Q_Δ is the family of Δ-sensitive queries $q : \mathcal{X}^n \to \mathbb{R}$. Let $\mathbf{P}$ be a distribution on $\mathcal{X}$ and let $\mathbf{x} \leftarrow_R \mathbf{P}^n$. Then*

$$\left|\mathop{\mathbb{E}}_{\mathbf{x},\mathcal{W}}[q(\mathbf{P}) \ : \ q = \mathcal{W}(\mathbf{x})] - \mathop{\mathbb{E}}_{\mathbf{x},\mathcal{W}}[q(\mathbf{x}) \ : \ q = \mathcal{W}(\mathbf{x})]\right| \leq 2\varepsilon\Delta n.$$

Proof. The proof proceeds via a sequence of intermediate samples. Let $\mathbf{x}' \leftarrow_{\mathrm{R}} \mathbf{P}^n$ be independent of $\mathbf{x}$. For $\ell \in \{0, 1, \ldots, n\}$, we define $\mathbf{x}^\ell = (x_1^\ell, \ldots, x_n^\ell) \in \mathcal{X}^n$ by

$$x_i^\ell = \begin{cases} x_i, & i > \ell, \\ x_i', & i \leq \ell. \end{cases}$$

By construction, $\mathbf{x}^0 = \mathbf{x}$ and $\mathbf{x}^n = \mathbf{x}'$, and intermediate samples $\mathbf{x}^\ell$ interpolate between $\mathbf{x}$ and $\mathbf{x}'$. Moreover, $\mathbf{x}^\ell$ and $\mathbf{x}^{\ell+1}$ differ in at most one entry, so that we can use the stability condition to relate $\mathcal{W}(\mathbf{x}^\ell)$ and $\mathcal{W}(\mathbf{x}^{\ell+1})$.

For every $\ell \in [n]$, we define $B^\ell : \mathcal{X}^n \times \mathcal{X}^n \to \mathbb{R}$ by

$$B^\ell(\mathbf{x}, \mathbf{z}) = q(\mathbf{z}) - q(\mathbf{z}_{-\ell}) + \Delta, \text{ where } q = \mathcal{W}(\mathbf{x}).$$

Here, $\mathbf{z}_{-\ell}$ is $\mathbf{z}$ with the ℓth element replaced by some arbitrary fixed element of $\mathcal{X}$.

Now we can write

$$\begin{aligned}
&\left|\mathop{\mathbb{E}}_{\mathbf{x},\mathcal{W}}[q(\mathbf{P}) - q(\mathbf{x}) \ : \ q = \mathcal{W}(\mathbf{x})]\right| \\
&\quad = \left|\mathop{\mathbb{E}}_{\mathbf{x},\mathbf{x}',\mathcal{W}}[q(\mathbf{x}') - q(\mathbf{x}) \ : \ q = \mathcal{W}(\mathbf{x})]\right| \\
&\quad = \left|\sum_{\ell=1}^{n} \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{x}',\mathcal{W}}\left[q(\mathbf{x}^\ell) - q(\mathbf{x}^{\ell-1}) \ : \ q = \mathcal{W}(\mathbf{x})\right]\right| \\
&\quad \leq \sum_{\ell=1}^{n} \left|\mathop{\mathbb{E}}_{\mathbf{x},\mathbf{x}',\mathcal{W}}\left[q(\mathbf{x}^\ell) - q(\mathbf{x}^{\ell-1}) \ : \ q = \mathcal{W}(\mathbf{x})\right]\right|
\end{aligned}$$

$$= \sum_{\ell\in[n]} \left| \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{x}',\mathcal{W}} \left[\left(q(\mathbf{x}^\ell) - q(\mathbf{x}^\ell_{-\ell}) + \Delta\right) - \left(q(\mathbf{x}^{\ell-1}) - q(\mathbf{x}^{\ell-1}_{-\ell}) + \Delta\right) \ : \ q = \mathcal{W}(\mathbf{x}) \right] \right|$$

(Since $\mathbf{x}^\ell_{-\ell} = \mathbf{x}^{\ell-1}_{-\ell}$)

$$= \sum_{\ell\in[n]} \left| \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{x}',\mathcal{W}} \left[B^\ell(\mathbf{x},\mathbf{x}^\ell) - B^\ell(\mathbf{x},\mathbf{x}^{\ell-1}) \right] \right| \qquad \text{(Definition of } B\text{)}.$$

Thus, to prove the lemma, it suffices to show that for every $\ell \in [n]$,

$$\left| \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{x}',\mathcal{W}} \left[B^\ell(\mathbf{x},\mathbf{x}^\ell) - B^\ell(\mathbf{x},\mathbf{x}^{\ell-1}) \right] \right| \le 2\Delta\varepsilon.$$

To complete the proof, we will need a few observations. First, since q is Δ-sensitive, for every $\ell, \mathbf{x}, \mathbf{z}$, we have $0 \le B^\ell(\mathbf{x},\mathbf{z}) \le 2\Delta$.

Second, observe that since $\mathcal{W}$ is assumed to be ε-TV stable, by the postprocessing lemma (Lemma 2.4) $B^\ell(\mathbf{x},\mathbf{z})$ is ε-TV stable with respect to its first parameter $\mathbf{x}$.

Finally, observe that the random variables $\mathbf{x}^0, \ldots, \mathbf{x}^n$ are identically distributed (although not independent). That is, every $\mathbf{x}^\ell$ consists of n independent draws from $\mathbf{P}$. Moreover, for every ℓ, the pairs $(\mathbf{x},\mathbf{x}^\ell)$ and $(\mathbf{x}^\ell,\mathbf{x})$ are identically distributed. Specifically, the first component is n independent samples from $\mathbf{P}$ and the second component is equal to the first component with a subset of the entries replaced by new independent samples from $\mathbf{P}$.

Combining the second and third observations with the triangle inequality, we have

$$\begin{aligned}
&\mathrm{d_{TV}}\left(B^\ell(\mathbf{x},\mathbf{x}^\ell), B^\ell(\mathbf{x},\mathbf{x}^{\ell-1})\right) \\
&\quad \le \mathrm{d_{TV}}\left(B^\ell(\mathbf{x},\mathbf{x}^\ell), B^\ell(\mathbf{x}^\ell,\mathbf{x})\right) + \mathrm{d_{TV}}\left(B^\ell(\mathbf{x}^\ell,\mathbf{x}), B^\ell(\mathbf{x}^{\ell-1},\mathbf{x})\right) \\
&\qquad + \mathrm{d_{TV}}\left(B^\ell(\mathbf{x}^{\ell-1},\mathbf{x}), B^\ell(\mathbf{x},\mathbf{x}^{\ell-1})\right) \\
&\quad \le 0 + \varepsilon + 0 = \varepsilon.
\end{aligned}$$

Using the observations above, for every $\ell \in [n]$ we have

$$\mathop{\mathbb{E}}_{\mathbf{x},\mathbf{x}',\mathcal{W}} \left[B^\ell(\mathbf{x},\mathbf{x}^\ell) - B^\ell(\mathbf{x},\mathbf{x}^{\ell-1}) \right] \le 2\Delta \cdot \mathrm{d_{TV}}\left(B^\ell(\mathbf{x},\mathbf{x}^\ell), B^\ell(\mathbf{x},\mathbf{x}^{\ell-1})\right) \le 2\Delta\varepsilon.$$

Thus we have the desired upper bound on the expectation of $B^\ell(\mathbf{x},\mathbf{x}^\ell) - B^\ell(\mathbf{x},\mathbf{x}^{\ell-1})$. The corresponding lower bound follows from an analogous argument. This completes the proof. ∎

4.3.2. From decorrelated expectation to accuracy on average.

THEOREM 4.7. *Let Q_Δ be the family of Δ-sensitive queries on $\mathcal{X}$. Assume that $\mathcal{M}$ is*

1. *$(\varepsilon = \alpha/4\Delta n)$-TV stable for k adaptively chosen queries from $Q = Q_\Delta$ and*
2. *$(\alpha' = \alpha/2)$-accurate on average with respect to its sample for n samples from $\mathcal{X}$ for k adaptively chosen queries from Q.*

Then $\mathcal{M}$ is α-accurate on average with respect to the population for k adaptively chosen queries from Q given n samples from $\mathcal{X}$.

The high level approach of the proof is to apply Lemma 4.6 to a "monitoring algorithm" that watches the interaction between the mechanism $\mathcal{M}(\mathbf{x})$ and the analyst $\mathcal{A}$ and then outputs the *least accurate* query. Since $\mathcal{M}(\mathbf{x})$ is stable, the decorrelated expectation lemma says that the query output by the monitor will satisfy $q(\mathbf{P}) \approx q(\mathbf{x})$

in expectation; this implies that even for the least accurate query in the interaction between $\mathcal{M}(\mathbf{x})$ and $\mathcal{A}$, $q(\mathbf{P}) \approx q(\mathbf{x})$ in expectation. Thus, if $\mathcal{M}$ is accurate with respect to the sample $\mathbf{x}$, it is also accurate with respect to $\mathbf{P}$.

Proof of Theorem 4.7. Let $\mathcal{M}$ be an interactive mechanism and $\mathcal{A}$ be an analyst that chooses the distribution $\mathbf{P}$. We define the following monitoring algorithm.

$$\mathcal{W}(\mathbf{x}) = \mathcal{W}_{\mathbf{P}}[\mathcal{M}, \mathcal{A}](\mathbf{x}) :$$

Input: $\mathbf{x} \in \mathcal{X}^n$

Simulate $\mathcal{M}(\mathbf{x})$ and $\mathcal{A}$ interacting, let $q_1, \ldots, q_k \in Q$ be the queries of $\mathcal{A}$ and let

$a_1, \ldots, a_k \in \mathbb{R}$ be the corresponding answers of $\mathcal{M}$.

Let $j = \arg\max_{j=1,\ldots,k} \left|\mathrm{err}^{\mathbf{P}}(q_j, a_j)\right|$.

If $a_j - q_j(\mathbf{P}) \geq 0$, let $q^* = q_j$, otherwise let $q^* = -q_j$. (Q_Δ is closed under negation.)

Output: q^*.

If $\mathcal{M}$ is stable, then so is $\mathcal{W}$, and this fact follows easily from the postprocessing lemma (Lemma 2.4).

CLAIM 4.8. *For every $\varepsilon \geq 0$, if the mechanism $\mathcal{M}$ is ε-TV stable for k adaptively chosen queries from Q, then for every $\mathbf{P}$ and $\mathcal{A}$, the monitor $\mathcal{W}_{\mathbf{P}}[\mathcal{M}, \mathcal{A}]$ is ε-TV stable.*

Proof of Claim 4.8. The assumption that $\mathcal{M}$ is ε-TV stable for k adaptively chosen queries from Q means that for every analyst $\mathcal{A}$ who asks k queries from Q, the algorithm $\mathcal{W}'(\mathbf{x})$ that simulates the interaction between $\mathcal{M}(\mathbf{x})$ and $\mathcal{A}$ and outputs the resulting query-answer pairs is ε-TV stable. Observe that the algorithm $\mathcal{W}$ defined above is simply a postprocessing of these query-answer pairs. That is, q^* depends only on $q_1, a_1, \ldots, q_k, a_k$ and $\mathbf{P}$, and not on $\mathbf{x}$. Thus, by Lemma 2.4, for every $\mathbf{P}$ and $\mathcal{A}$, the monitor $\mathcal{W}_{\mathbf{P}}[\mathcal{M}, \mathcal{A}]$ is ε-TV stable. ∎

In light of Claim 4.8 and our assumption that $\mathcal{M}$ is $(\alpha/4\Delta n)$-TV stable, we can apply Lemma 4.6 to obtain

$$\left| \mathop{\mathbb{E}}_{\mathbf{x},\mathcal{W}} [q^*(\mathbf{P}) - q^*(\mathbf{x}) \ : \ q^* = \mathcal{W}(\mathbf{x})] \right| \leq 2\left(\frac{\alpha}{4\Delta n}\right)\Delta n \leq \alpha/2. \tag{7}$$

To complete the proof, we show that if $\mathcal{M}$ is not α-accurate on average with respect to the population $\mathbf{P}$, then (7) cannot hold.

CLAIM 4.9. *If $\mathcal{M}$ is $(\alpha/2)$-accurate for the sample but not α-accurate for the population, then*

$$\left| \mathop{\mathbb{E}}_{\mathbf{x},\mathcal{W}} [q^*(\mathbf{P}) - q^*(\mathbf{x}) \ : \ q^* = \mathcal{W}(\mathbf{x})] \right| \geq \alpha/2.$$

Proof of Claim 4.9. Using our assumptions, we can calculate as follows:

$$\begin{aligned}
&\left|\mathop{\mathbb{E}}_{\mathbf{x},\mathcal{W}}[q^*(\mathbf{P}) - q^*(\mathbf{x}) \ : \ q^* = \mathcal{W}(\mathbf{x})]\right| \\
&\quad = \left|\mathop{\mathbb{E}}_{\mathbf{x},\mathcal{W}}[q^*(\mathbf{P}) - a_{q^*} \ : \ q^* = \mathcal{W}(\mathbf{x})] + \mathop{\mathbb{E}}_{\mathbf{x},\mathcal{W}}[a_{q^*} - q^*(\mathbf{x}) \ : \ q^* = \mathcal{W}(\mathbf{x})]\right| \\
&\quad \geq \left|\mathop{\mathbb{E}}_{\mathbf{x},\mathcal{W}}[q^*(\mathbf{P}) - a_{q^*} \ : \ q^* = \mathcal{W}(\mathbf{x})]\right| - \left|\mathop{\mathbb{E}}_{\mathbf{x},\mathcal{W}}[a_{q^*} - q^*(\mathbf{x}) \ : \ q^* = \mathcal{W}(\mathbf{x})]\right| \\
&\quad > \alpha - \left|\mathop{\mathbb{E}}_{\mathbf{x},\mathcal{W}}[a_{q^*} - q^*(\mathbf{x}) \ : \ q^* = \mathcal{W}(\mathbf{x})]\right| \qquad (8) \\
&\quad \geq \alpha - \alpha/2 \qquad (9) \\
&\quad = \alpha/2.
\end{aligned}$$

Line (8) follows from two observations. First, by construction of $\mathcal{W}$, we always have $q^*(\mathbf{P}) - a_{q^*} \leq 0$. Second, since $\mathcal{M}$ is assumed not to be α-accurate on average for the population, the expected value of $|q^*(\mathbf{P}) - a_{q^*}| > \alpha$. Since $\mathcal{W}$ ensures that $a_{q^*} - q^*(\mathbf{P}) \geq 0$, we also have that the absolute value of the expectation of $q^*(\mathbf{P}) - a_{q^*}$ is greater than α. Line (9) follows from the assumption that $\mathcal{M}$ is $(\alpha/2)$-accurate on average for the sample. ∎

Thus, if $\mathcal{M}$ is not α-accurate on average for the population, we will obtain a contradiction to (7). This completes the proof. ∎

5. From low-sensitivity queries to optimization queries. In this section, we extend our results for low-sensitivity queries to the more general family of minimization queries. To do so, we design a suitable monitoring algorithm for minimization queries. As in our analysis of low-sensitivity queries, we will have the monitoring algorithm take as input many independent samples and simulate the interaction between $\mathcal{M}$ and $\mathcal{A}$ on each of those samples. Thus, if $\mathcal{M}$ has even a small probability of being inaccurate, then with constant probability the monitor will find a minimization query that $\mathcal{M}$ has answered inaccurately. Previously, we had the monitor simply output this query and applied Lemma 3.3 to arrive at a contradiction. However, since Lemma 3.3 applies only to algorithms that output a low-sensitivity query, we can't apply it to the monitor that outputs a minimization query. We address this by having the monitor output the *error function* associated with the loss function and answer it selects, which is a low-sensitivity query. If we assume that the mechanism is accurate for its sample but not for the population, then the monitor will find a loss function and an answer with low error on the sample but large error on the population. Thus the error function will be a low-sensitivity query with very different answers on the sample and the population, which is a contradiction. To summarize, we have the following theorem.

THEOREM 5.1 (transfer theorem for minimization queries). *Let $Q = Q_{min}$ be the family of Δ-sensitive minimization queries on $\mathcal{X}$. Assume that, for some $\alpha, \beta \geq 0$, $\mathcal{M}$ is*

1. *$(\varepsilon = \alpha/128\Delta n, \delta = \alpha\beta/64\Delta n)$-max-KL stable for k adaptively chosen queries from Q and*
2. *$(\alpha' = \alpha/8, \beta' = \alpha\beta/32\Delta n)$-accurate with respect to its sample for n samples from $\mathcal{X}$ for k adaptively chosen queries from Q.*

Then $\mathcal{M}$ is (α, β)-accurate with respect to the population for k adaptively chosen queries from Q given n samples from $\mathcal{X}$.

The formal proof is nearly identical to that of Theorem 3.4, so we omit the full proof. Instead, we will simply describe the modified monitoring algorithm.

$$\mathcal{W}(\mathbf{X}) = \mathcal{W}_{\mathbf{P}}[\mathcal{M}, \mathcal{A}](\mathbf{X}) :$$

Input: $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_T) \in (\mathcal{X}^n)^T$

For $t = 1, \ldots, T$:

Simulate $\mathcal{M}(\mathbf{x}_t)$ and $\mathcal{A}$ interacting, let $L_{t,1}, \ldots, L_{t,k} \in Q$ be the queries of $\mathcal{A}$ and let

$\theta_{t,1}, \ldots, \theta_{t,k} \in \mathbb{R}$ be the corresponding answers of $\mathcal{M}$.

Let (t^*, j^*) be

$$(t^*, j^*) = \underset{j \in [k],\, t \in [T]}{\arg\max} \left| \mathrm{err}^{\mathbf{P}}\left(L_{t,j}, \theta_{t,j}\right) \right|.$$

Let $q^*(\mathbf{x}) = \mathrm{err}_{\mathbf{x}}(L_{t^*,j^*}, \theta_{t^*,j^*})$ (note, by construction, $q^* \in Q_{2\Delta}$, i.e., q^* is 2Δ-sensitive).

Output: (q^*, t^*).

6. Applications.

6.1. Low-sensitivity and statistical queries.

We now plug known stable mechanisms (designed in the context of differential privacy) into Theorem 3.4 to obtain mechanisms that provide strong error guarantees with high probability for both low-sensitivity and statistical queries.

COROLLARY 6.1 (Theorem 3.4 and [DMNS06, SU15a]). *There is a mechanism $\mathcal{M}$ that is (α, β)-accurate with respect to the population for k adaptively chosen queries from Q_Δ where $\Delta = O(1/n)$ given n samples from $\mathcal{X}$ for*

$$n \geq O\left(\frac{\sqrt{k \cdot \log\log k} \cdot \log^{3/2}(1/\alpha\beta)}{\alpha^2}\right).$$

The mechanism runs in time $\mathrm{poly}(n, \log|\mathcal{X}|, \log(1/\beta))$ *per query.*

COROLLARY 6.2 (Theorem 3.4 and [RR10]). *There is a mechanism $\mathcal{M}$ that is (α, β)-accurate with respect to the population for k adaptively chosen queries from Q_Δ where $\Delta = O(1/n)$ given n samples from $\mathcal{X}$ for*

$$n = O\left(\frac{\log|\mathcal{X}| \cdot \log k \cdot \log^{3/2}(1/\alpha\beta)}{\alpha^3}\right).$$

The mechanism runs in time $\mathrm{poly}(|\mathcal{X}|^n)$ *per query. The case where Δ is not $O(1/n)$ can be handled by rescaling the output of the query.*

COROLLARY 6.3 (Theorem 3.4 and [HR10]). *There is a mechanism $\mathcal{M}$ that is α-accurate on average with respect to the population for k adaptively chosen queries from Q_{SQ} given n samples from $\mathcal{X}$ for*

$$n = O\left(\frac{\sqrt{\log|\mathcal{X}|} \cdot \log k \cdot \log^{3/2}(1/\alpha\beta)}{\alpha^3}\right).$$

The mechanism runs in time $\mathrm{poly}(n, |\mathcal{X}|)$ *per query.*

6.2. Optimization queries. The results of section 5 can be combined with existing differentially private algorithms for minimizing "empirical risk" (that is, loss with respect to the sample $\mathbf{x}$) to obtain algorithms for answering adaptive sequences of minimization queries. We provide a few specific instantiations here, based on known differentially private mechanisms.

6.2.1. Minimization over arbitrary finite sets.

COROLLARY 6.4 (Theorem 5.1 and [MT07]). *Let Θ be a finite set of size at most D. Let $Q \subset Q_{min}$ be the set of sensitivity-$1/n$ loss functions bounded between 0 and C. Then there is a mechanism $\mathcal{M}$ that is (α, β)-accurate with respect to the population for k adaptively chosen queries from Q_{min} given*

$$n \geq O\left(\frac{\log(DC/\alpha) \cdot \sqrt{k} \cdot \log^{3/2}(1/\alpha\beta)}{\alpha^2}\right)$$

samples from $\mathcal{X}$. The running time of the mechanism is dominated by $O((k+\log(1/\beta)) \cdot D)$ evaluations of the loss function.

6.2.2. Convex minimization. We state bounds for convex minimization queries for some of the most common parameter regimes in applications. In the first two corollaries, we consider 1-Lipschitz[10] loss functions over a bounded domain.

COROLLARY 6.5 (Theorem 5.1 and [BST14]). *Let Θ be a closed, convex subset of $\mathbb{R}^d$ set such that $\max_{\theta \in \Theta} \|\theta\|_2 \leq 1$. Let $Q \subset Q_{min}$ be the set of convex 1-Lipschitz loss functions that are $1/n$-sensitive. Then there is a mechanism $\mathcal{M}$ that is (α, β)-accurate with respect to the population for k adaptively chosen queries from Q given*

$$n = \tilde{O}\left(\frac{\sqrt{dk} \cdot \log^2(1/\alpha\beta)}{\alpha^2}\right)$$

samples from Q. The running time of the mechanism is dominated by $k \cdot n^2$ evaluations of the gradient ∇L.

COROLLARY 6.6 (Theorem 5.1 and [Ull15]). *Let Θ be a closed, convex subset of $\mathbb{R}^d$ set such that $\max_{\theta \in \Theta} \|\theta\|_2 \leq 1$. Let $Q \subset Q_{min}$ be the set of convex 1-Lipschitz loss functions that are $1/n$-sensitive. Then there is a mechanism $\mathcal{M}$ that is (α, β)-accurate with respect to the population for k adaptively chosen queries from Q given*

$$n = \tilde{O}\left(\frac{\sqrt{\log|\mathcal{X}|} \cdot (\sqrt{d} + \log k) \cdot \log^{3/2}(1/\alpha\beta)}{\alpha^3}\right)$$

samples from $\mathcal{X}$. The running time of the mechanism is dominated by $\mathrm{poly}(n, |\mathcal{X}|)$ *and $k \cdot n^2$ evaluations of the gradient ∇L.*

In the next two corollaries, we consider 1-strongly convex,[11] Lipschitz loss functions over a bounded domain.

[10] A loss function $L : \mathcal{X} \times \mathbb{R}^d \to \mathbb{R}$ is 1-Lipschitz if for every $\theta, \theta' \in \mathbb{R}^d$, $x \in \mathcal{X}$, $|L(\theta, x) - L(\theta', x)| \leq \|\theta - \theta'\|_2$.

[11] A loss function $L : \mathcal{X} \times \mathbb{R}^d \to \mathbb{R}$ is 1-strongly convex if for every $\theta, \theta' \in \mathbb{R}^d$, $x \in \mathcal{X}$,

$$L(\theta', x) \geq L(\theta, x) + \langle \nabla L(\theta, x), \theta' - \theta \rangle + (1/2) \cdot \|\theta - \theta'\|_2^2,$$

where the (sub)gradient $\nabla L(\theta, x)$ is taken with respect to θ.

COROLLARY 6.7 (Theorem 5.1 and [BST14]). *Let Θ be a closed, convex subset of $\mathbb{R}^d$ set such that $\max_{\theta\in\Theta}\|\theta\|_2 \le 1$. Let $Q \subset Q_{min}$ be the set of 1-strongly convex, 1-Lipschitz loss functions that are $1/n$-sensitive. Then there is a mechanism $\mathcal{M}$ that is (α,β)-accurate with respect to the population for k adaptively chosen queries from Q given*

$$n = \tilde{O}\left(\frac{\sqrt{dk}\cdot\log^{3/2}(1/\alpha\beta)}{\alpha^{3/2}}\right)$$

samples from $\mathcal{X}$. The running time of the mechanism is dominated by $k\cdot n^2$ evaluations of the gradient ∇L.

COROLLARY 6.8 (Theorem 5.1 and [Ull15]). *Let Θ be a closed, convex subset of $\mathbb{R}^d$ set such that $\max_{\theta\in\Theta}\|\theta\|_2 \le 1$. Let $Q \subset Q_{min}$ be the set of 1-strongly convex 1-Lipschitz loss functions that are $1/n$-sensitive. Then there is a mechanism $\mathcal{M}$ that is (α,β)-accurate with respect to the population for k adaptively chosen queries from Q given*

$$n = \tilde{O}\left(\sqrt{\log|\mathcal{X}|}\cdot\left(\frac{\sqrt{d}}{\alpha^{5/2}}+\frac{\log k}{\alpha^3}\right)\cdot\log^{3/2}(1/\alpha\beta)\right)$$

samples from $\mathcal{X}$. The running time of the mechanism is dominated by $\mathrm{poly}(n,|\mathcal{X}|)$ *and $k\cdot n^2$ evaluations of the gradient ∇L.*

7. An alternative form of generalization and tightness of our results. We now provide an alternative form of our generalization bounds. The following theorem is more general than Theorem 3.4 because it says that *no* max-KL stable procedure that outputs a low-sensitivity query can output any query that distinguishes the sample from the population (not just max-KL stable procedures that are accurate for the sample).

First we prove the following technical lemma.

LEMMA 7.1. *Let F be a finite set, $f : F \to \mathbb{R}$ a function, and $\eta > 0$. Define a random variable W on F by*

$$\mathbb{P}[W=w] = \frac{e^{\eta f(w)}}{C}, \qquad \text{where} \qquad C = \sum_{w\in F} e^{\eta f(w)}.$$

Then

$$\mathbb{E}[f(W)] \ge \max_{w\in F} f(w) - \frac{1}{\eta}\log|F|.$$

Proof. We have

$$f(w) = \frac{1}{\eta}\left(\log C + \log\mathbb{P}[W=w]\right).$$

Thus

$$\begin{aligned}\mathbb{E}[f(W)] &= \sum_{w\in F}\mathbb{P}[W=w]\,f(w)\\ &= \sum_{w\in F}\mathbb{P}[W=w]\,\frac{1}{\eta}\left(\log C+\log\mathbb{P}[W=w]\right)\\ &= \frac{1}{\eta}\left(\log C - H(W)\right),\end{aligned}$$

where $H(W)$ is the Shannon entropy of the distribution of W (measured in nats, rather than bits). In particular,

$$H(W) \leq \log |\text{support}(W)| = \log |F|,$$

as the uniform distribution maximizes entropy. Moreover, $C \geq \max_{w \in F} e^{\eta f(w)}$, whence $\frac{1}{\eta} \log C \geq \max_{w \in F} f(w)$. The result now follows from these two inequalities. ∎

THEOREM 7.2. *Let $\varepsilon \in (0, 1/3)$, $\delta \in (0, \varepsilon/4)$, and $n \geq \frac{1}{\varepsilon^2} \log(\frac{4\varepsilon}{\delta})$. Let $\mathcal{M} : \mathcal{X}^n \to Q_\Delta$ be (ε, δ)-max-KL stable where Q_Δ is the class of Δ-sensitive queries $q : \mathcal{X}^n \to \mathbb{R}$. Let $\mathbf{P}$ be a distribution on $\mathcal{X}$, let $\mathbf{x} \leftarrow_R \mathbf{P}^n$, and let $q \leftarrow_R \mathcal{M}(\mathbf{x})$. Then*

$$\mathop{\mathbb{P}}_{\mathbf{x}, \mathcal{M}} [|q(\mathbf{P}) - q(\mathbf{x})| \geq 18\varepsilon\Delta n] < \frac{\delta}{\varepsilon}.$$

Intuitively, Theorem 7.2 says that "stability prevents overfitting." It says that no stable algorithm can output a low-sensitivity function that distinghishes its input from the population the input was drawn from (i.e., "overfits" its sample).

In particular, Theorem 7.2 implies that if a mechanism $\mathcal{M}$ is stable and outputs q that "fits" its data, then q also "fits" the population. This gives a learning theory perspective on our results.

Proof. Consider the following monitor algorithm $\mathcal{W}$.

$$\mathcal{W}(\mathbf{X}) = \mathcal{W}_{\mathbf{P}}[\mathcal{M}](\mathbf{X}) :$$

Input: $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_T) \in (\mathcal{X}^n)^T$

Set $F = \emptyset$.

For $t = 1, \ldots, T$:

Let $q_t \leftarrow \mathcal{M}(\mathbf{x}_t)$, and set $F = F \cup \{(q_t, t), (-q_t, t)\}$.

Sample (q^*, t^*) from F with probability proportional to $\exp\left(\frac{\varepsilon}{\Delta}\left(q^*(\mathbf{x}_{t^*}) - q^*(\mathbf{P})\right)\right)$.

Output: (q^*, t^*).

We will use the monitor $\mathcal{W}$ with $T = \lfloor \epsilon/\delta \rfloor$. Observe that $\mathcal{W}$ only accesses its input through $\mathcal{M}$ (which is (ε, δ)-max-KL stable) and the exponential mechanism (which is $(\varepsilon, 0)$-max-KL stable). Thus, by composition and postprocessing, $\mathcal{W}$ is $(2\varepsilon, \delta)$-max-KL stable. We can hence apply Lemma 3.3 to obtain

$$\mathop{\mathbb{E}}_{\mathbf{X}, \mathcal{W}} [q^*(\mathbf{x}_{t^*}) - q^*(\mathbf{P}) \ : \ (q^*, t^*) = \mathcal{W}(\mathbf{X})] \leq 2\left(e^{2\varepsilon} - 1 + T\delta\right)\Delta n < 8\varepsilon\Delta n. \tag{10}$$

Now we can apply Lemma 7.1 with $f(q, t) = q(\mathbf{x}_t) - q(\mathbf{P})$ and $\eta = \frac{\varepsilon}{\Delta}$ to get

$$\mathop{\mathbb{E}}_{q^*, t^*} [f(q^*, t^*)] \geq \max_{(q,t) \in F} f(q, t) - \frac{\Delta}{\varepsilon} \log |F| = \max_{t \in [T]} |q_t(\mathbf{x}_t) - q_t(\mathbf{P})| - \frac{\Delta}{\varepsilon} \log(2T). \tag{11}$$

Combining (10) and (11) gives

$$\begin{aligned} &\mathop{\mathbb{E}}_{\mathbf{X}, \mathcal{W}} \left[\max_{t \in [T]} |q_t(\mathbf{x}_t) - q_t(\mathbf{P})|\right] - \frac{\Delta}{\varepsilon} \log(2T) \\ &\quad \leq \mathop{\mathbb{E}}_{\mathbf{X}, \mathcal{W}} [q^*(\mathbf{x}_{t^*}) - q^*(\mathbf{P}) \ : \ (q^*, t^*) = \mathcal{W}(\mathbf{X})] < 8\varepsilon\Delta n. \end{aligned} \tag{12}$$

To complete the proof, we assume, for the sake of contradiction, that $\mathcal{M}$ has a high enough probability of outputting a query q such that $|q(\mathbf{P}) - q(\mathbf{x})|$ is large. To obtain a contradiction from this assumption, we need the following natural claim (analogous to Claim 3.6) about the output of the monitor.

CLAIM 7.3. *If*

$$\mathop{\mathbb{P}}_{\mathbf{x},\mathcal{M}}[|q(\mathbf{P}) - q(\mathbf{x})| \geq 18\varepsilon\Delta n] \geq \frac{\delta}{\varepsilon},$$

then

$$\mathop{\mathbb{P}}_{\mathbf{X},\mathcal{W}}\left[\max_{t\in[T]}|q_t(\mathbf{x}_t) - q_t(\mathbf{P})| \geq 18\varepsilon\Delta n\right] \geq 1 - \left(1 - \frac{\delta}{\varepsilon}\right)^T \geq \frac{1}{2}.$$

The previous claim implies that

$$\mathop{\mathbb{E}}_{\mathbf{X},\mathcal{W}}\left[\max_{t\in[T]}|q_t(\mathbf{x}_t) - q_t(\mathbf{P})|\right] \geq 9\varepsilon\Delta n. \tag{13}$$

Combining (12) and (13) gives

$$9\varepsilon\Delta n - \frac{\Delta}{\varepsilon}\log(2T) \leq 8\varepsilon\Delta n,$$

which simplifies to

$$\log(2\varepsilon/\delta) \geq \log(2T) \geq \varepsilon^2 n.$$

This contradicts the assumption that $n \geq \frac{1}{\varepsilon^2}\log(\frac{4\varepsilon}{\delta})$ and hence completes the proof. □

7.1. Optimality. We now show that our connection between max-KL stability and generalization (Theorems 7.2 and 3.4) is optimal.

LEMMA 7.4. *Let $\alpha > \delta > 0$, let $n \geq \frac{1}{\alpha}$, and let $\Delta \in [0,1]$. Let $\mathbf{U}$ be the uniform distribution over $[0,1]$. There exists a $(0,\delta)$-max-KL stable algorithm $\mathcal{A} : [0,1]^n \to Q_\Delta$ such that if $\mathbf{X} \leftarrow_R \mathbf{U}^n$ and if $q \leftarrow_R \mathcal{A}(\mathbf{X})$, then*

$$\Pr[q(\mathbf{X}) - q(\mathbf{U}) \geq \alpha\Delta n] \geq \frac{\delta}{2\alpha}.$$

Proof. Consider the following simple algorithm, denoted as $\mathcal{B}$: *On input a database $\mathbf{x}$, output $\mathbf{x}$ with probability δ, and otherwise output the empty database.* Clearly, $\mathcal{B}$ is $(0,\delta)$-max-KL stable. Now construct the following algorithm $\mathcal{A}$.

Input: A database $\mathbf{x} \in [0,1]^n$. We think of $\mathbf{x}$ as $\frac{1}{\alpha}$ databases of size αn each: $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_{1/\alpha})$.
For $1 \leq i \leq 1/\alpha$, let $\hat{\mathbf{x}}_i = \mathcal{B}(\mathbf{x}_i)$.
Let $p : [0,1] \to \{0,1\}$ where $p(x) = 1$ iff $\exists i$ s.t. $x \in \hat{\mathbf{x}}_i$.
Define $q_p : [0,1]^n \to \mathbb{R}$ where $q_p(\mathbf{x}) = \Delta \sum_{x\in\mathbf{x}} p(x)$.
(Note that q_p is a Δ-sensitive query, and that it is a statistical query if $\Delta = 1/n$.)
Output: q_p.

As $\mathcal{B}$ is $(0,\delta)$-max-KL stable, and as $\mathcal{A}$ only applies $\mathcal{B}$ on disjoint databases, we get that $\mathcal{A}$ is also $(0,\delta)$-max-KL stable.

Suppose $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_{1/\alpha})$ contains independent and identically distributed samples from $\mathbf{U}$, and consider the execution of $\mathcal{A}$ on $\mathbf{x}$. Observe that the predicate p

evaluates to 1 only on a finite number of points from $[0,1]$, and hence, we have that $q_p(\mathbf{U}) = 0$. Next note that $q_p(\mathbf{x}) = \alpha\Delta n \cdot |\{i : \hat{\mathbf{x}}_i = \mathbf{x}_i\}|$. Therefore, if there exists an i s.t. $\hat{\mathbf{x}}_i = \mathbf{x}_i$, then $q(\mathbf{x}) - q(\mathbf{U}) \geq \alpha\Delta n$. The probability that this is not the case is at most

$$(1-\delta)^{1/\alpha} \leq e^{-\delta/\alpha} \leq 1 - \frac{\delta}{2\alpha},$$

and thus, with probability at least $\frac{\delta}{2\alpha}$, algorithm $\mathcal{A}$ outputs a Δ-sensitive query q s.t. $q(\mathbf{x}) - q(\mathbf{U}) \geq \alpha\Delta n$. ∎

In particular, using Lemma 7.4 with $\alpha = \varepsilon$ shows that the parameters in Theorem 7.2 are tight.

Acknowledgments. We thank Mark Bun, Moritz Hardt, Aaron Roth, and Salil Vadhan for many helpful discussions.

REFERENCES

[BE02] O. Bousquet and A. Elisseeff, *Stability and generalization*, J. Mach. Learn. Res., 2 (2002), pp. 499–526.

[BH95] Y. Benjamini and Y. Hochberg, *Controlling the false discovery rate: A practical and powerful approach to multiple testing*, J. R. Stat. Soc. B Methodol., 57 (1995), pp. 289–300.

[BH15] A. Blum and M. Hardt, *The Ladder: A Reliable Leaderboard for Machine Learning Competitions*, CoRR, abs/1502.04585, 2015.

[Bon36] C. E. Bonferroni, *Teoria statistica delle classi e calcolo delle probabilita*, Pubbl. d. R. Ist. Super. di Sci. Econom. e Commerciali di Firenze, 8 (1936).

[BST14] R. Bassily, A. Smith, and A. Thakurta, *Private empirical risk minimization: Efficient algorithms and tight error bounds*, in Proceedings of FOCS, IEEE, 2014, pp. 464–473.

[BUV14] M. Bun, J. Ullman, and S. P. Vadhan, *Fingerprinting codes and the price of approximate differential privacy*, in Proceedings of STOC, ACM, 2014, pp. 1–10.

[DFH+15a] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O.Reingold, and A. Roth, *Generalization in adaptive data analysis and holdout reuse*, in Proceedings of Advances in Neural Information Processing Systems, Montreal, 2015, pp. 2350–2358.

[DFH+15b] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth, *Preserving statistical validity in adaptive data analysis*, in Proceedings of the ACM Symposium on the Theory of Computing, ACM, 2015, pp. 117–126.

[DFH+15c] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth, *The reusable holdout: Preserving validity in adaptive data analysis*, Science, 349 (2015), pp. 636–638.

[DMNS06] C. Dwork, F. McSherry, K. Nissim, and A. Smith, *Calibrating noise to sensitivity in private data analysis*, in Theory of Cryptography, Lecture Notes in Comput. Sci. 3876, Springer, New York, 2006, pp. 265–284.

[DN03] I. Dinur and K. Nissim, *Revealing information while preserving privacy*, in Proceedings of PODS, ACM, 2003, pp. 202–210.

[DRV10] C. Dwork, G. N. Rothblum, and S. P. Vadhan, *Boosting and differential privacy*, in Proceedings of the IEEE Symposium on Foundations of Computer Science, IEEE, 2010, pp. 51–60.

[Dun61] O. J. Dunn, *Multiple comparisons among means*, J. Amer. Statist. Assoc., 56 (1961), pp. 52–64.

[DW79a] L. Devroye and T. J. Wagner, *Distribution-free inequalities for the deleted and holdout error estimates*, IEEE Trans. Inform. Theory, 25 (1979), pp. 202–207.

[DW79b] L. Devroye and T. J. Wagner, *Distribution-free performance bounds for potential function rules*, IEEE Trans. Inform. Theory, 25 (1979), pp. 601–604.

[Dwo06] C. Dwork, *Differential privacy*, in Proceedings of the 33rd International Colloquium on Automata, Languages and Programming, Part II, Venice, Italy, 2006, 2006, pp. 1–12.

[GL14] A. Gelman and E. Loken, *The statistical crisis in science*, American Scientist, 102 (2014), 460.

[HR10] M. Hardt and G. Rothblum, *A multiplicative weights mechanism for privacy-preserving data analysis*, in Proceedings of the 51st Conference on Foundations of Computer Science, IEEE, 2010, pp. 61–70.

[HU14] M. Hardt and J. Ullman, *Preventing false discovery in interactive data analysis is hard*, in Proceedings of FOCS, IEEE, 2014, pp. 454–465.

[Ioa05] J. P. A. Ioannidis, *Why most published research findings are false*, PLoS Medicine, 2 (2005), 124.

[Kea93] M. J. Kearns, *Efficient noise-tolerant learning from statistical queries*, in Proceedings of STOC, ACM, 1993, pp. 392–401.

[KR99] M. J. Kearns and D. Ron, *Algorithmic stability and sanity-check bounds for leave-one-out cross-validation*, Neural Comput., 11 (1999), pp. 1427–1453.

[McS14] F. McSherry, *Differential Privacy for Measure Concentration*, Windows on Theory, http://windowsontheory.org/2014/02/04/differential-privacy-for-measure-concentration/ (2014).

[MT07] F. McSherry and K. Talwar, *Mechanism design via differential privacy*, in Proceedings of FOCS, IEEE, 2007, pp. 94–103.

[RR10] A. Roth and T. Roughgarden, *Interactive privacy via the median mechanism*, in Proceedings of STOC, ACM, 2010, pp. 765–774.

[SSSS10] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan, *Learnability, stability and uniform convergence*, J. Mach. Learn. Res., 11 (2010), pp. 2635–2670.

[SU15a] T. Steinke and J. Ullman, *Between Pure and Approximate Differential Privacy*, CoRR, abs/1501.06095, 2015.

[SU15b] T. Steinke and J. Ullman, *Interactive fingerprinting codes and the hardness of preventing false discovery*, in Proceedings of the 28th Conference on Learning Theory, 2015, pp. 1588–1628.

[Ull13] J. Ullman, *Answering $n^{2+o(1)}$ counting queries with differential privacy is hard*, in Proceedings of STOC, ACM, 2013, pp. 361–370.

[Ull15] J. Ullman, *Private multiplicative weights beyond linear queries*, in Proceedings of PODS, ACM, 2015.