A necessary and sufficient stability notion for adaptive generalization

Katrina Ligett and Moshe Shenfeld

June 4, 2019

Abstract

We introduce a new notion of the stability of computations, which holds under post-processing and adaptive composition, and show that the notion is both necessary and sufficient to ensure generalization in the face of adaptivity, for any computations that respond to bounded-sensitivity linear queries while providing accuracy with respect to the data sample set. The stability notion is based on quantifying the effect of observing a computation's outputs on the posterior over the data sample elements. We show a separation between this stability notion and previously studied notions.

1 Introduction

A fundamental idea behind most forms of data-driven research and machine learning is the concept of *generalization*–the ability to infer properties of a data distribution by working only with a sample from that distribution. One typical approach is to invoke a concentration bound to ensure that, for a sufficiently large sample size, the evaluation of the function on the sample set will yield a result that is close to its value on the underlying distribution, with high probability. Intuitively, these concentration arguments ensure that, for any given function, most sample sets are good "representatives" of the distribution. Invoking a union bound, such a guarantee easily extends to the evaluation of multiple functions on the same sample set.

Of course, such guarantees hold only if the functions to be evaluated were chosen independently of the sample set. In recent years, grave concern has erupted in many data-driven fields, that *adaptive selection* of computations is eroding statistical validity of scientific findings [Ioa05, GL14]. Adaptivity is not an evil to be avoided—it constitutes a natural part of the scientific process, wherein previous findings are used to develop and refine future hypotheses. However, unchecked adaptivity can (and does, as demonstrated by, e.g., [DFH$^+$15b] and [RZ16]) often lead one to evaluate *overfitting* functions—ones that return very different values on the sample set than on the distribution.

Traditional generalization guarantees do not necessarily guard against adaptivity; while generalization ensures that the response to a query on a sample set will be *close* to that of the same query on the distribution, it does not rule out the possibility that the probability to get a *specific* response will be dramatically affected by the contents of the sample set. In the extreme, a generalizing computation could encode the whole sample set in the low-order bits of the output, while maintaining high accuracy with respect to the underlying distribution. Subsequent adaptive queries could then, by *post-processing* the computation's output, arbitrarily overfit to the sample set.

In recent years, an exciting line of work, starting with Dwork et al. [DFH$^+$15b], has formalized this problem of adaptive data analysis and introduced new techniques to ensure guarantees of generalization in the face of an adaptively-chosen sequence of computations (what we call here *adaptive generalization*). One great insight of Dwork et al. and followup work was that techniques for ensuring the *stability* of computa-

tions (some of them originally conceived as privacy notions) can be powerful tools for providing adaptive generalization.

A number of papers have considered variants of stability notions, the relationships between them, and their properties, including generalization properties. Despite much progress in this space, one issue that has remained open is the limits of stability—how much can the stability notions be relaxed, and still imply generalization? It is this question that we address in this paper.

1.1 Our Contribution

We introduce a new notion of the stability of computations, which holds under post-processing (Theorem 2.4) and adaptive composition (Theorems 2.9 and 2.10), and show that the notion is both necessary (Theorem 3.13) and sufficient (Theorem 3.9) to ensure generalization in the face of adaptivity, for any computations that respond to bounded-sensitivity linear queries (see Definition 3.1) while providing accuracy with respect to the data sample set. This means (up to a small caveat)[1] that our stability definition is equivalent to generalization, assuming sample accuracy, for bounded linear queries. Linear queries form the basis for many learning algorithms, such as those that rely on gradients or on the estimation of the average loss of a hypothesis.

In order to formulate our stability notion, we consider a prior distribution over the database elements and the posterior distribution over those elements conditioned on the output of a computation. In some sense, harmful outputs are those that induce large statistical distance between this prior and posterior (Definition 2.1). Our new notion of stability, *Local Statistical Stability* (Definition 2.2), intuitively, requires a computation to have only small probability of producing such a harmful output.

In Section 4, we directly prove that Differential Privacy, Max Information and Compression Schemes all imply Local Statistical Stability, which provides an alternative method to establish their generalization properties. We also provide a few separation examples between the various definitions.

1.2 Additional Related Work

Most countermeasures to overfitting fall into one of a few categories. A long line of work bases generalization guarantees on some form of bound on the complexity of the range of the mechanism, e.g., its VC dimension (see [SSBD14] for a textbook summary of these techniques). Other examples include *Bounded Description Length* [DFH+15a], and *compression schemes* [LW86] (which additionally hold under post-processing and adaptive composition [DFH+15a, CLN+16]). Another line of work focuses on the algorithmic stability of the computation [BE02], which bounds the effects on the output of changing one element in the training set.

A different category of stability notions, which focus on the effect of a small change in the sample set on the probability distribution over the range of possible outputs, has recently emerged from the notion of Differential Privacy [DMNS06]. Work of [DFH+15b] established that Differential Privacy, interpreted as a stability notion, ensures generalization; it is also known (see [DR+14]) to be robust to adaptivity and to withstand post-processing. A number of subsequent works propose alternative stability notions that weaken the conditions of Differential Privacy in various ways while attempting to retain its desirable generalization properties. One example is *Max Information* [DFH+15a], which shares the guarantees of Differential Privacy. A variety of other stability notions ([RRST16, RZ16, RRT+16, BNS+16, FS17, EGI19]), unlike Differential Privacy and Max Information, only imply generalization in expectation. [XR17, Ala17, BMN+17] extend these guarantees to generalization in probability, under various restrictions.

[CLN+16] introduce the notion of *post-hoc generalization*, which captures robustness to post-processing, but it was recently shown not to hold under composition [NSS+18]. The challenges that the internal correla-

[1] In particular, our lower bound (Theorem 3.13) requires one more query than our upper bound (Theorem 3.9).

tion of non-product distributions present for stability have been studied in the context of *Inferential Privacy* [GK16] and *Typical Stability* [BF16].

2 LS stability definition and properties

Let $\mathcal{X}$ be an arbitrary countable *domain*. Fixing some $n \in \mathbb{N}$, let $D_{\mathcal{X}^n}$ be some probability distribution defined over $\mathcal{X}^n$.[2] Let $\mathcal{G}, \mathcal{R}$ be some arbitrary countable sets which we will refer to as *generators* and *responses* respectively. Let a *mechanism* $M : \mathcal{X}^n \times \mathcal{G} \to \mathcal{R}$ be a (possibly non-deterministic) function that, given a *sample set* $S \in \mathcal{X}^n$ and a generator $G \in \mathcal{G}$, returns a response $r \in \mathcal{R}$. Intuitively, generators can be thought of as representing functions from $\mathcal{X}^n$ to $\mathcal{R}$ and the mechanism as providing an estimate to the value of those functions, but we do not restrict the definitions, for reasons which will become apparent once we formalize the adaptive process (Definition 2.5).

This setting involves two sources of randomness, the *underlying distribution* $D_{\mathcal{X}^n}$, and the *conditional distribution* $D^G_{\mathcal{R}|\mathcal{X}^n}(r \mid S)$—that is, the probability to get r as the output of $M(S, G)$. These in turn induce a set of distributions (formalized in Definition A.1): the *joint distribution* $D^G_{(\mathcal{X}^n,\mathcal{R})}$ over $\mathcal{X}^n \times \mathcal{R}$, the *disjoint (independent) distribution* $D^G_{\mathcal{X}^n \otimes \mathcal{R}}$ over $\mathcal{X}^n \times \mathcal{R}$, the *unconditional (marginal) distribution* $D^G_{\mathcal{R}}$ over $\mathcal{R}$, and the *conditional distribution* $D^G_{\mathcal{X}^n|\mathcal{R}}$ over $\mathcal{X}^n$.

Although the underlying distribution $D_{\mathcal{X}^n}$ is defined on $\mathcal{X}^n$, it induces a natural probability distribution over $\mathcal{X}$ as well, as follows. We define the *sampling function* $\mathrm{Sam} : \mathcal{X}^n \to \mathcal{X}$ which, given a sample set, returns one of its sample elements uniformly at random. Notice that $\mathrm{Sam}(\cdot)$ can be thought of as a mechanism with one possible generator and with response range $\mathcal{X}$, which allows us to define $D_{\mathcal{X}}$, $D_{\mathcal{X}|\mathcal{X}^n}$ and $D_{\mathcal{X}^n|\mathcal{X}}$ as well.[3] This in turn allows us to define a few key distributions, which form a connection between $\mathcal{R}$ and $\mathcal{X}$ (formalized in Definition A.2): the *joint distribution* $D^G_{(\mathcal{X},\mathcal{R})}$ over $\mathcal{X} \times \mathcal{R}$, the *disjoint distribution* $D^G_{\mathcal{X}\otimes\mathcal{R}}$ over $\mathcal{X} \times \mathcal{R}$, the *conditional distribution* $D^G_{\mathcal{R}|\mathcal{X}}$ over $\mathcal{R}$, and the *conditional distribution* $D^G_{\mathcal{X}|\mathcal{R}}$ over $\mathcal{X}$. We use this notation to denote both the probability that a distribution places on a subset of its range and the probability placed on a single element of the range.

2.1 Local Statistical Stability

Before observing any output from the mechanism, an outside observer knowing $D_{\mathcal{X}^n}$ but without other information about the sample set S holds prior $D_{\mathcal{X}}(x)$ that sampling an element of S would return a particular $x \in \mathcal{X}$. Once an output r of the mechanism is observed, however, the observer's posterior becomes $D^G_{\mathcal{X}|\mathcal{R}}(x|r)$. The difference between these two distributions is what determines the resulting degradation in stability. This difference could be quantified using a variety of distance measures (a partial list can be found in Appendix E); here we introduce a particular one which we use to define our stability notion.

Definition 2.1 (Stability loss of a response)**.** Given a distribution $D_{\mathcal{X}^n}$, a generator $G \in \mathcal{G}$, and a mechanism $M : \mathcal{X}^n \times \mathcal{G} \to \mathcal{R}$, the *stability loss* $\ell^G_{D_{\mathcal{X}^n}}(r)$ of a response $r \in \mathcal{R}$ with respect to $D_{\mathcal{X}^n}$ and G is defined as the Statistical Distance (see definition in Appendix E) between the prior distribution over $\mathcal{X}$ and the posterior induced by r. That is,

$$\ell^G_{D_{\mathcal{X}^n}}(r) \coloneqq \sum_{x \in X_+(r)} \left(D^G_{\mathcal{X}|\mathcal{R}}(x \mid r) - D_{\mathcal{X}}(x) \right),$$

[2] Throughout the paper, $\mathcal{X}^n$ can either denote the family of sequences of length n or a multiset of size n; that is, the sample set S can be treated as an ordered or unordered set.

[3] The superscript notation of G is missing from the definition, since there is only one possible "generator" in this case. It is worth noting that in the case where $D_{\mathcal{X}^n}$ is the product distribution of some distribution $Z_{\mathcal{X}}$ over $\mathcal{X}$, we get that $D_{\mathcal{X}} = Z_{\mathcal{X}}$.

where $X_+(r) := \left\{x \in \mathcal{X} \mid D^G_{\mathcal{X}|\mathcal{R}}(x \mid r) > D_{\mathcal{X}}(x)\right\}$, the set of all sample elements which have a posterior probability (given r) higher then their prior. Notice this notation omits M on which it depends. Similarly, we define the stability loss $\ell^G_{D_{\mathcal{X}^n}}(R)$ of a set of responses $R \subseteq \mathcal{R}$ with respect to $D_{\mathcal{X}^n}$ and G as,

$$\ell^G_{D_{\mathcal{X}^n}}(R) := \frac{\sum_{r \in R} D^G_{\mathcal{R}}(r) \cdot \ell^G_{D_{\mathcal{X}^n}}(r)}{D^G_{\mathcal{R}}(R)}.$$

Given $0 \leq \epsilon \leq 1$, a response will be called *ϵ-unstable* with respect to $D_{\mathcal{X}^n}$ and G if its loss is greater the ϵ. The set of all ϵ-unstable responses will be denoted $R^G_{D_{\mathcal{X}^n}}(\epsilon) := \left\{r \in \mathcal{R} \mid \ell^G_{D_{\mathcal{X}^n}}(r) > \epsilon\right\}$.

We now introduce our notion of stability of a mechanism.

Definition 2.2 (Local Statistical Stability)**.** Given $0 \leq \epsilon, \delta \leq 1$, a distribution $D_{\mathcal{X}^n}$, and a generator $G \in \mathcal{G}$, a mechanism $M : \mathcal{X}^n \times \mathcal{G} \to \mathcal{R}$ will be called *(ϵ, δ)-Local-Statistically Stable with respect to $D_{\mathcal{X}^n}$ and G* (or *LS Stable*, or *LSS*, for short) if for any $R \subseteq \mathcal{R}$,

$$D^G_{\mathcal{R}}(R) \cdot \left(\ell^G_{D_{\mathcal{X}^n}}(R) - \epsilon\right) \leq \delta.$$

Notice that the maximal value of the left hand side is achieved for the subset $R^G_{D_{\mathcal{X}^n}}(\epsilon)$. This stability definition can be extended to apply to a family of generators and/or a family of possible distributions. When there exists a family of generators $\mathcal{G}$ and a family of distributions $\mathcal{D}$ such that a mechanism M is (ϵ, δ)-LSS for all $D_{\mathcal{X}^n} \in \mathcal{D}$ and for all $G \in \mathcal{G}$, then M will be called *(ϵ, δ)-LSS for $\mathcal{D}, \mathcal{G}$*. (This stability notion somewhat resembles *Semantic Privacy* as discussed by [KS14], though they use it to compare different posterior distributions.)

Intuitively, this can be thought of as placing a δ bound on the probability of observing an outcome whose stability loss exceeds ϵ. This claim is formalized in the next Lemma.

Lemma 2.3. *Given $0 \leq \epsilon, \delta \leq 1$, a distribution $D_{\mathcal{X}^n}$, and a generator G, if a mechanism $M : \mathcal{X}^n \times \mathcal{G} \to \mathcal{R}$ is (ϵ, δ)-LSS with respect to $D_{\mathcal{X}^n}, G$, then $D^G_{\mathcal{R}}\left(R^G_{D_{\mathcal{X}^n}}(2\epsilon)\right) < \frac{\delta}{\epsilon}$.*

Proof. Assume by way of contradiction that $D^G_{\mathcal{R}}\left(R^G_{D_{\mathcal{X}^n}}(2\epsilon)\right) > \frac{\delta}{\epsilon}$; then

$$D^G_{\mathcal{R}}\left(R^G_{D_{\mathcal{X}^n}}(2\epsilon)\right) \cdot \left(\ell^G_{D_{\mathcal{X}^n}}\left(R^G_{D_{\mathcal{X}^n}}(2\epsilon)\right) - \epsilon\right) > \frac{\delta}{\epsilon} \cdot (2\epsilon - \epsilon) = \delta. \qquad \square$$

2.2 Properties

We now turn to prove two crucial properties of LSS: post-processing and adaptive composition.

Post-processing guarantees (in some contexts, known as data processing inequalities) ensure that the stability of a computation can only be *increased* by subsequent manipulations. This is a key desideratum for concepts used to ensure adaptivity-proof generalization, since otherwise an adaptive subsequent computation could potentially arbitrarily degrade the generalization guarantees.

Theorem 2.4 (LSS holds under Post-Processing)**.** *Given $0 \leq \epsilon, \delta \leq 1$, a distribution $D_{\mathcal{X}^n}$, and a generator $G \in \mathcal{G}$, if a mechanism $M : \mathcal{X}^n \times \mathcal{G} \to \mathcal{R}$ is (ϵ, δ)-LSS with respect to $D_{\mathcal{X}^n}$ and G, then for any range $\mathcal{U}$ and any arbitrary (possibly non-deterministic) function $f : \mathcal{R} \to \mathcal{U}$, we have that $f \circ M : \mathcal{X}^n \times \mathcal{G} \to \mathcal{U}$ is also (ϵ, δ)-LSS with respect to $D_{\mathcal{X}^n}$ and G. An analogous statement also holds for mechanisms that are LSS with respect to a family of generators and/or a family of distributions.*

Proof. We start by defining a function $w_{\mathcal{U}}^{\epsilon} : \mathcal{R} \to [0,1]$ such that $\forall r \in \mathcal{R} : w_{\mathcal{U}}^{\epsilon}(r) = \sum_{u \in U_{D_{\mathcal{X}^n}}^{G}(\epsilon)} D_{\mathcal{U}|\mathcal{R}}(u \,|\, r)$.

Using this function we get that,

$$\sum_{u \in U_{D_{\mathcal{X}^n}}^{G}(\epsilon)} D_{\mathcal{U}}^{G}(u) = \sum_{r \in \mathcal{R}} w_{\mathcal{U}}^{\epsilon}(r) \cdot D_{\mathcal{R}}^{G}(r)$$

and

$$\sum_{u \in U_{D_{\mathcal{X}^n}}^{G}(\epsilon)} D_{\mathcal{U}}^{G}(u) \cdot \ell_{D_{\mathcal{X}^n}}^{G}(u) \leq \sum_{r \in \mathcal{R}} w_{\mathcal{U}}^{\epsilon}(r) \cdot D_{\mathcal{R}}^{G}(r) \cdot \ell_{\mathcal{R}}^{G}(r)$$

(detailed proof can be found in Appendix B.1).

Combining the two we get that,

$$\begin{aligned}
D_{\mathcal{U}}^{G}\left(U_{D_{\mathcal{X}^n}}^{G}(\epsilon)\right) \cdot \left(\ell_{D_{\mathcal{X}^n}}^{G}\left(U_{D_{\mathcal{X}^n}}^{G}(\epsilon)\right) - \epsilon\right) &= \sum_{u \in U_{D_{\mathcal{X}^n}}^{G}(\epsilon)} D_{\mathcal{U}}^{G}(u)\left(\ell_{D_{\mathcal{X}^n}}^{G}(u) - \epsilon\right) \\
&\leq^{(1)} \sum_{r \in \mathcal{R}} w_{\mathcal{U}}^{\epsilon}(r) \cdot D_{\mathcal{R}}^{G}(r)\left(\ell_{\mathcal{R}}^{G}(r) - \epsilon\right) \\
&\leq^{(2)} \sum_{r \in R_{D_{\mathcal{X}^n}}^{G}(\epsilon)} \overbrace{w_{\mathcal{U}}^{\epsilon}(r)}^{\leq 1} \cdot D_{\mathcal{R}}^{G}(r)\left(\ell_{\mathcal{R}}^{G}(r) - \epsilon\right) \\
&\leq \sum_{r \in R_{D_{\mathcal{X}^n}}^{G}(\epsilon)} D_{\mathcal{R}}^{G}(r)\left(\ell_{\mathcal{R}}^{G}(r) - \epsilon\right) \\
&\leq^{(3)} \delta
\end{aligned}$$

where (1) results from the two previous claims, (2) from the fact that we removed only negative terms and (3) from the LSS definition, which concludes the proof. □

In order to formally define adaptive learning and stability under adaptively chosen generators, we formalize the notion of an adversary who issues those generators.

Definition 2.5 (Adversary and Adaptive Mechanism)**.** An *adversary over a family of generators* $\mathcal{G}$ is a particular type of generator which is a (possibly non-deterministic) function $A : \mathcal{R}^* \to \mathcal{G}$ that receives a *view*—a finite sequence of responses—and outputs a generator. We denote by $\mathcal{A}$ the family of all adversaries, and write $\mathcal{V}_k := \mathcal{R}^k$ and $\mathcal{V} := \mathcal{R}^*$.

Illustrated below, the *adaptive mechanism* $\mathrm{Adp}_{\bar{M}} : \mathcal{X}^n \times \mathcal{A} \to \mathcal{V}_k$ is a particular type of mechanism, which inputs an adversary as its generator and which returns a view as its range type. It is parameterized by a set of *sub-mechanisms* $\bar{M} = (M_i)_{i=1}^{k}$ where $\forall i \in [k]$, $M_i : \mathcal{X}^n \times \mathcal{G} \to \mathcal{R}$. Given a sample set S and an adversary A as input, the adaptive mechanism iterates k times through the process where A sends a generator to M_i and receives its response to that generator on the sample set. The adaptive mechanism returns the resulting sequence of k responses v_k. Naturally, this requires A to match M such that M's range can be A's input, and vice versa.[4] [5]

[4] If the same mechanism appears more then once in $\bar{M}$, it can also be stateful, which means it retains an internal record consisting of internal randomness, the history of sample sets and generators it has been fed, and the responses it has produced; its behavior may be a function of this internal record. We omit this from the notation for simplicity, but do refer to this when relevant. A stateful mechanism will be defined as LSS if it is LSS given any reachable internal record. A pedantic treatment might consider the *probability* that a particular internal state could be reached, and only require LSS when accounting for these probabilities.

[5] If A is randomized, we add one more step at the beginning where $\mathrm{Adp}_{\bar{M}}$ randomly generates some bits c—A's "coin tosses."

Adaptive Mechanism $\mathrm{Adp}_{\bar{M}}$

Input: $S \in \mathcal{X}^n$, $A \in \mathcal{A}$

Output: $v_k \in \mathcal{V}_k$

$v_0 \leftarrow \emptyset$ or c

for $i \in [k]$:

  $G_i \leftarrow A(v_{i-1})$ or $A(v_{i-1}, c)$

  $r_i \leftarrow M_i(S, G_i)$

  $v_i \leftarrow (v_{i-1}, r_i)$

return v_k

Definition 2.6 (k-LSS under adaptivity)**.** Given $0 \leq \epsilon, \delta \leq 1$, a distribution $D_{\mathcal{X}^n}$, and an adversary A, a sequence of k mechanisms $\bar{M}$ will be called (ϵ, δ)*-local-statistically stable under* k *adaptive iterations* with respect to $D_{\mathcal{X}^n}$ and A (or k-LSS for short), if $\mathrm{Adp}_{\bar{M}}$ is (ϵ, δ)-LSS with respect to $D_{\mathcal{X}^n}$ and A (in which case we will use $V^{A,k}_{D_{\mathcal{X}^n}}(\epsilon)$ to denote the set of ϵ unstable views). This definition can be extended to a family of adversaries and/or a family of possible distributions as well.

Adaptive composition is a key property of a stability notion, since it restricts the degradation of stability across multiple computations. A key observation is that the posterior $D^A_{\mathcal{X}^n|\mathcal{V}_k}(S \mid v_k)$ is itself a distribution over $\mathcal{X}^n$ and G_{k+1} is a deterministic function of v_k. Therefore, as long as each sub-mechanism is LSS with respect to any posterior that could have been induced by previous adaptive interaction, one can reason about the properties of the composition.

Definition 2.7 (View-induced posterior distributions)**.** A sequence of mechanisms $\bar{M}$, an adversary A, and a view $v_k \in \mathcal{V}_k$ together induce a set of posterior distributions over $\mathcal{X}^n$, $\mathcal{X}$, and $\mathcal{R}$. For clarity we will denote these induced distributions by Z^{v_k} instead of D.

As mentioned before, all the distributions we consider stem from two basic distributions; the underlying distribution $D_{\mathcal{X}^n}$ and the conditional distribution $D^G_{\mathcal{R}|\mathcal{X}^n}$. The posteriors of these distributions change once we see v_k. $D_{\mathcal{X}^n}(S)$ is replaced by $Z^{v_k}_{\mathcal{X}^n}(S) \coloneqq D^A_{\mathcal{X}^n|\mathcal{V}_k}(S \mid v_k)$ (actually, the rigorous notation should have been $Z^{\bar{M},A,v_k}_{\mathcal{X}^n}$, but since $\bar{M}$ and A will be fixed throughout this analysis, we omit them for simplicity). Similarly, $D^{G_{k+1}}_{\mathcal{R}|\mathcal{X}^n}(r \mid S)$ is replaced by

$$Z^{v_k,G_{k+1}}_{\mathcal{R}|\mathcal{X}^n}(r \mid S) \coloneqq \mathbb{P}_{M_{k+1}}\left(M_{k+1}(S, G_{k+1}) = r \mid \mathrm{Adp}_{\bar{M},k}(S, A) = v_k\right)$$

where $\mathrm{Adp}_{\bar{M},k}$ denotes the first k iterations of the adaptive mechanism.[6]

We next establish two important properties of the distributions over $\mathcal{V}_{k+1}$ induced by $\mathrm{Adp}_{\bar{M}}$ and their relation to the posterior distributions.

Lemma 2.8. *Given a distribution $D_{\mathcal{X}^n}$, an adversary $A : \mathcal{V} \to \mathcal{G}$, and a sequence of k mechanisms $\bar{M}$ where $\forall i \in [k]$, $M_i : \mathcal{X}^n \times \mathcal{G} \to \mathcal{R}$, for any $v_{k+1} \in \mathcal{V}_{k+1}$ we denote $v_{k+1} = (v_k, r_{k+1})$. In this case, using notation from Definition 2.7,*

$$D^A_{\mathcal{V}_{k+1}}(v_{k+1}) = D^A_{\mathcal{V}_k}(v_k) \cdot Z^{v_k,G_{k+1}}_{\mathcal{R}}(r_{k+1})$$

In this case, $v_k \coloneqq (c, r_1, \ldots, r_{ik})$ and A receives the coin tosses as an input as well. This addition turns G_{k+1} into a deterministic function of v_i for any $i \in \mathbb{N}$, a fact that will be used multiple times throughout the paper. In this situation, the randomness of $\mathrm{Adp}_{\bar{M}}$ results both from the randomness for the coin tosses and from that of the sub-mechanisms.

[6] If M_{k+1} is stateful, the conditioning can result from any unknown state of M_{k+1} which might affect its response to G_{k+1}. If M_{k+1} has no shared state with the previous sub-mechanisms (ether because it is a different mechanism or because it is stateless), then the only effect v_k has on the posterior on $\mathcal{R}$ is by governing G_{k+1} (which, as mentioned before, is a deterministic function of v_k for the given A), in which case $Z^{v_k,G_{k+1}}_{\mathcal{R}|\mathcal{X}^n}(r \mid S) = D^{G_{k+1}}_{\mathcal{R}|\mathcal{X}^n}(r \mid S)$ where the mechanism is M_{k+1}.

and

$$\ell^A_{D_{\mathcal{X}^n}}\left(v_{k+1}\right) \leq \ell^A_{D_{\mathcal{X}^n}}\left(v_k\right) + \ell^{G_{k+1}}_{Z^{v_k}_{\mathcal{X}^n}}\left(r_{k+1}\right).$$

The proof can be found in Appendix B.2

Bounding the stability loss of a view by the sum of losses of its responses with respect to the sub-mechanisms, provides a linear bound on the degradation of the LSS parameters. Adding a bound on the expectation of the loss of the sub-mechanisms allows us to also invoke Azuma's inequality and prove a sub-linear bound.

Theorem 2.9 (LSS adaptively composes linearly)**.** *Given a family of distributions $\mathcal{D}$ over $\mathcal{X}^n$, a family of generators $\mathcal{G}$, and a sequence of k mechanisms $\bar{M}$ where $\forall i \in [k]$, $M_i : \mathcal{X}^n \times \mathcal{G} \to \mathcal{R}$, we will denote $\mathcal{D}_{M_0,\mathcal{G}} := \mathcal{D}$, and for any $i > 0$, $\mathcal{D}_{M_i,\mathcal{G}} := \left\{D^G_{\mathcal{X}^n|\mathcal{R}}\left(\cdot \mid r\right) \mid D^G_{\mathcal{R}}\left(r\right) > 0\right\}$ the set of all posterior distributions induced by any response of M_i with non-zero probability with respect to $\mathcal{D}_{M_{i-1},\mathcal{G}}$ and $\mathcal{G}$.*

Given a sequence $0 \leq \epsilon_1, \delta_1, \ldots, \epsilon_k, \delta_k \leq 1$, if for all i, M_i is (ϵ_i, δ_i)-LSS with respect to $\mathcal{D}_{M_{i-1},\mathcal{G}}$ and $\mathcal{G}$, the sequence is $\left(\sum_{i\in[k]} \epsilon_i, \sum_{i\in[k]} \delta_i\right)$-$k$-LSS with respect to $\mathcal{D}$ and any adversary A over $\mathcal{G} \times \mathcal{R}$.

One simple case is when $\mathcal{D}_{M_{i-1},\mathcal{G}} = \mathcal{D}$, and M_i is (ϵ_i, δ_i)-LSS with respect to $\mathcal{D}$ and $\mathcal{G}$, for all i.

Proof of Theorem 2.9. This theorem is a direct result of combining Lemma 2.8 with the triangle inequality over the posteriors created at any iteration, and the fact that the the mechanisms are LSS over the new posterior distributions. Formally this is proven using induction on the number of adaptive iterations. The base case $k = 0$ is the coin tossing step, which is independent of the set and therefore has zero loss. For the induction step we start by denoting the projections of $V^{A,k+1}_{D_{\mathcal{X}^n}}\left(\epsilon_{[k+1]}\right)$ on $\mathcal{V}_k$ and $\mathcal{R}$ by,

$$\forall r_{k+1} \in \mathcal{R}, V_k\left(r_{k+1}\right) := \left\{v_k \in \mathcal{V}_k \mid \left(v_k, r_{k+1}\right) \in V^{A,k+1}_{D_{\mathcal{X}^n}}\left(\epsilon_{[k+1]}\right)\right\}$$

$$\forall v_k \in \mathcal{V}_k, R\left(v_k\right) := \left\{r_{k+1} \in \mathcal{R} \mid \left(v_k, r_{k+1}\right) \in V^{A,k+1}_{D_{\mathcal{X}^n}}\left(\epsilon_{[k+1]}\right)\right\}$$

where $\epsilon_{[k]} := \sum_{i\in[k]} \epsilon_i$. Using this notation and that in Definition 2.7 we get that

$$\begin{aligned}
&D^A_{\mathcal{V}_{k+1}}\left(V^{A,k+1}_{D_{\mathcal{X}^n}}\left(\epsilon_{[k+1]}\right)\right) \cdot \left(\ell^A_{D_{\mathcal{X}^n}}\left(V^{A,k+1}_{D_{\mathcal{X}^n}}\left(\epsilon_{[k+1]}\right)\right) - \epsilon_{[k+1]}\right) \\
&\quad \leq \sum_{v_k \in V^{A,k}_{D_{\mathcal{X}^n}}\left(\epsilon_{[k]}\right)} D^A_{\mathcal{V}_k}\left(v_k\right)\left(\ell^A_{D_{\mathcal{X}^n}}\left(v_k\right) - \epsilon_{[k]}\right) \\
&\qquad + \sum_{r_{k+1} \in R^{G_{k+1}}_{D_{\mathcal{X}^n}}\left(\epsilon_{k+1}\right)} Z^{v_k,G_{k+1}}_{\mathcal{R}}\left(r_{k+1}\right)\left(\ell^{G_{k+1}}_{Z^{v_k}_{\mathcal{X}^n}}\left(r_{k+1}\right) - \epsilon_{k+1}\right) \\
&\quad \leq \sum_{i\in[k]} \delta_i + \delta_{k+1} \\
&\quad = \sum_{i\in[k+1]} \delta_i,
\end{aligned}$$

Detailed proof can be found in Appendix B.2. □

Theorem 2.10 (LSS adaptively composes sub-linearly)**.** *Under the same conditions as Theorem 2.9, and given $0 \leq \alpha_1, \ldots, \alpha_k \leq 1$, such that for all i and any $D_{\mathcal{X}^n} \in \mathcal{D}_{M_{i-1},\mathcal{G}}$, and $G \in \mathcal{G}$, $\mathop{\mathbb{E}}_{r=M_i(S,G)}\left[\ell^G_{D_{\mathcal{X}^n}}\left(r\right)\right] \leq$*

α_i where the expectation is taken over the randomness of the choice of $S \in \mathcal{X}^n$ and the internal probability of M, then for any $0 \leq \delta' \leq 1$, the sequence is $\left(\epsilon', \delta' + \sum_{i\in[k]} \frac{\delta_i}{\epsilon_i}\right)$-$k$-LSS with respect to $\mathcal{D}$ and any adversary A over $\mathcal{G} \times \mathcal{R}$, where $\epsilon' := \sqrt{8\ln\left(\frac{1}{\delta'}\right) \sum_{i\in[k]} \epsilon_i^2} + \sum_{i\in[k]} \alpha_i$.

The theorem is non-trivial for $\alpha_i \leq \epsilon_i$.

Proof of Theorem 2.10. The proof is based on the fact that the sum of the stability losses is a martingale with respect to v_k, and invoking Lemma B.1, which extends Azuma's inequality to the case of a high probability bound.

Formally, for any given $k > 0$, we can define $\Omega_0 := \mathcal{X}^n$ and $\forall i \in [k], \Omega_i := \mathcal{R}$.[7] We define a probability distribution over Ω_0 as $D_{\mathcal{X}^n}$, and for any $i > 0$, define a probability distribution over Ω_i given $\Omega_1, \ldots, \Omega_{i-1}$ as $Z_{\mathcal{R}|S}^{v_{i-1},G_i}$. We then define a sequence of random variables, $Y_0 = 0$ and $\forall i > 0$,

$$Y_i(S, r_1, \ldots, r_i) = \sum_{j=1}^{i} \left(\ell_{Z_{\mathcal{X}^n}^{v_{j-1}}}^{G_j}(r_j) - \mathop{\mathbb{E}}_{Z_{\mathcal{R}}^{v_{j-1},G_j}} \left[\ell_{Z_{\mathcal{X}^n}^{v_{j-1}}}^{G_j}(r) \right] \right).$$

Intuitively Y_i is the sum of the first i losses, with a correction term which zeroes the expectation. These random variables are a martingale with respect to the random process $S, r_1, \ldots, r_k$, since

$$\mathbb{E}\left[Y_{i+1} \mid S, r_1, \ldots, r_i\right] = Y_i(S, r_1, \ldots, r_i)$$

where the expectation is taken over the random process, which has randomness that results from the choice of $S \in \mathcal{X}^n$ and the internal probability of M (detailed proof can be found in Appendix B.2).

From the LSS definition (2.2) and Lemma 2.3, for any $i \in [k]$ we get that $\mathbb{P}\left(\ell_{Z_{\mathcal{X}^n}^{v_{i-1}}}^{G_i}(r_i) > 2\epsilon_i\right) \leq \frac{\delta_i}{\epsilon_i}$, so with probability greater than $\frac{\delta_{i+1}}{\epsilon_{i+1}}$,

$$|Y_{i+1} - Y_i| = \left| \ell_{Z_{\mathcal{X}^n}^{v_i}}^{G_{i+1}}(r_{i+1}) - \mathbb{E}\left[\ell_{Z_{\mathcal{X}^n}^{v_i}}^{G_{i+1}}(r_{i+1})\right] \right| \leq 2\epsilon_{i+1}.$$

Using this fact we can invoke Lemma B.1 and get that for any $0 \leq \delta' \leq 1$,

$$\mathbb{P}\left(\ell_{D_{\mathcal{X}^n}}^{A}(v_k) > \epsilon'\right) \leq \mathbb{P}\left(Y_k - \overbrace{Y_0}^{=0} > \sqrt{8\ln\left(\frac{1}{\delta'}\right)\sum_{i=1}^{k}\epsilon_i^2}\right) \leq \delta' + \sum_{i=1}^{k}\frac{\delta_i}{\epsilon_i}$$

□

3 LSS is Necessary and Sufficient for Generalization

Up until this point, generators and responses have been fairly abstract concepts. In order to discuss generalization and accuracy, we must make them concrete. As a result, in this section, we often consider generators in the family of functions $q : \mathcal{X}^n \to \mathcal{R}$, which we will refer to as *queries* and denote by $\mathcal{Q}$, and we consider responses which have some metric defined over them. We show our results for a fairly general class of functions known as bounded linear queries.[8]

[7] If the adversary A is non-deterministic, $\Omega_0 := \mathcal{X}^n \times C$, where C is the set of all possible coin tosses of the adversary, as mentioned in Definition 2.5. If the mechanisms have some internal state not expressed by the responses, Ω_i will be the domain of those states, as mentioned in Definition 2.7.

[8] For simplicity, throughout the following section we choose $\mathcal{R} = \mathbb{R}$, but all results extend to any metric space, in particular $\mathbb{R}^d$.

Definition 3.1 (Linear queries)**.** A function $q : \mathcal{X}^n \to \mathbb{R}$ will be called a *linear query*, if it is defined by a function $q_1 : \mathcal{X} \to \mathbb{R}$ such that $q(S) := \frac{1}{n} \sum_{i=1}^{n} q_1(s_i)$ (for simplicity we will slightly abuse notation and denote q_1 simply as q throughout the paper). If $q : \mathcal{X} \to [-\Delta, \Delta]$ it will be called a Δ-*bounded linear query*. The set of Δ-bounded linear queries will be denoted $\mathcal{Q}_\Delta$.

In this context, there is a "correct" answer the mechanism can produce for a given generator, defined as the correct response to the query on the sample set or distribution, and its distance from the response provided by the mechanism can be thought of as the mechanism's error.

Definition 3.2 (Sample accuracy, distribution accuracy)**.** Given $0 \leq \epsilon$, $0 \leq \delta \leq 1$, a distribution $D_{\mathcal{X}^n}$, and a query q, a mechanism $M : \mathcal{X}^n \times \mathcal{Q} \to \mathbb{R}$ will be called (ϵ, δ)-*Sample Accurate with respect to* $D_{\mathcal{X}^n}$ *and* q, if

$$\mathbb{P}_{S,M}\left(|M(S,q) - q(S)| > \epsilon\right) \leq \delta.$$

Such a mechanism will be called (ϵ, δ)-*Distribution Accurate with respect to* $D_{\mathcal{X}^n}$ *and* q if

$$\mathbb{P}_{S,M}\left(|M(S,q) - q(D_{\mathcal{X}^n})| > \epsilon\right) \leq \delta,$$

where $q(D_{\mathcal{X}^n}) := \mathop{\mathbb{E}}_{S}[q(S)]$. In both cases the probability is taken over the randomness of the choice of $S \in \mathcal{X}^n$ and the internal probability of M. The expectation is taken only over the randomness of the choice of $S \in \mathcal{X}^n$. When there exists a family of distributions $\mathcal{D}$ and a family of queries $\mathcal{Q}$ such that a mechanism M is (ϵ, δ)-Sample (Distribution) Accurate for all $D \in \mathcal{D}$ and for all $q \in \mathcal{Q}$, then M will be called (ϵ, δ)-*Sample (Distribution) Accurate with respect to* $\mathcal{D}$ *and* $\mathcal{Q}$.

A sequence of k mechanisms $\bar{M}$ where $\forall i \in [k] : M_i : \mathcal{X}^n \times \mathcal{Q} \to \mathbb{R}$ which respond to a sequence of k (potentially adaptively chosen) queries $q_1, \ldots q_k$ will be called (ϵ, δ)-k-*Sample Accurate with respect to* $D_{\mathcal{X}^n}$ *and* $q_1, \ldots q_k$ if

$$\mathbb{P}_{S,M}\left(\max_{i \in k} |M_i(S, q_i) - q_i(S)| > \epsilon\right) \leq \delta$$

and (ϵ, δ)-k-*Distribution Accurate with respect to* $D_{\mathcal{X}^n}$ *and* $q_1, \ldots q_k$ if

$$\mathbb{P}_{S,M}\left(\max_{i \in k} |M_i(S, q_i) - q_i(D_{\mathcal{X}^n})| > \epsilon\right) \leq \delta.$$

When considering an adaptive process, accuracy is defined with respect to the adversary, and the probabilities are taken also over the choice of the coin tosses by the adaptive mechanism.[9]

We denote by $\mathbb{V}$ the set of views consisting of responses in $\mathbb{R}$.

3.1 LSS Implies Generalization

As a step toward showing that LS Stability implies a high probability generalization, we first show a generalization of expectation result. We do so, as a tool, specifically for a mechanism that returns a query as its output. Intuitively, this allows us to wrap an entire adaptive process into a single mechanism. Analyzing the potential of the mechanism to generate an overfitting query is a natural way to learn about the generalization capabilities of the mechanism.

[9] If the adaptive mechanism invokes a stateful sub-mechanism multiple times, we specify that the mechanism is sample (distribution) accurate if it is sample (distribution) accurate given any reachable internal record. Again, a somewhat more involved treatment might consider the *probability* that a particular internal state of the mechanism could be reached.

Theorem 3.3 (Generalization of expectation)**.** *Given $0 \leq \epsilon, \delta \leq 1$, a distribution $D_{\mathcal{X}^n}$, a generator G, and a mechanism $M : \mathcal{X}^n \times \mathcal{G} \to \mathcal{Q}_\Delta$, if $D^G_{\mathcal{R}}\left(Q^G_{D_{\mathcal{X}^n}}(\epsilon)\right) < \delta$, then*

$$\left| \mathop{\mathbb{E}}_{q=M(S,G)} \left[q\left(D_{\mathcal{X}^n}\right) - q\left(S\right) \right] \right| < 2\Delta\left(\epsilon + \delta\right).$$

The expectations are taken over the randomness of the choice of $S \in \mathcal{X}^n$ and the internal randomness of M.

Proof. First notice that,

$$q\left(S\right) = \frac{1}{n}\sum_{i=1}^{n} q\left(s_i\right) = \sum_{x\in\mathcal{X}} D_{\mathcal{X}|\mathcal{X}^n}\left(x \,|\, S\right) \cdot q\left(x\right)$$

where $s_1, \ldots s_n$ denotes the elements of the sample set S. Using this identity we separately analyze the expected value of the returned query with respect to the distribution, and with respect to the sample set (detailed proof can be found in Appendix C).

$$\mathop{\mathbb{E}}_{q=M(S,G)} \left[q\left(D_{\mathcal{X}^n}\right) \right] = \sum_{q\in\mathcal{Q}_\Delta} D^G_{\mathcal{Q}_\Delta}\left(q\right) \sum_{x\in\mathcal{X}} D_{\mathcal{X}}\left(x\right) \cdot q\left(x\right)$$

$$\mathop{\mathbb{E}}_{q=M(S,G)} \left[q\left(S\right) \right] = \sum_{q\in\mathcal{Q}_\Delta} D^G_{\mathcal{Q}_\Delta}\left(q\right) \sum_{x\in\mathcal{X}} D^A_{\mathcal{X}|\mathcal{Q}_\Delta}\left(x \,|\, q\right) \cdot q\left(x\right)$$

Now we can calculate the difference:

$$\begin{aligned}
\left| \mathop{\mathbb{E}}_{q=M(S,G)} \left[q\left(D_{\mathcal{X}^n}\right) - q\left(S\right) \right] \right| &= \left| \sum_{q\in\mathcal{Q}_\Delta} D^G_{\mathcal{Q}_\Delta}\left(q\right) \sum_{x\in\mathcal{X}} \left(D_{\mathcal{X}}\left(x\right) - D^G_{\mathcal{X}|\mathcal{Q}_\Delta}\left(x \,|\, q\right) \right) \cdot q\left(x\right) \right| \\
&\leq^{(1)} \sum_{q\in\mathcal{Q}_\Delta} D^G_{\mathcal{Q}_\Delta}\left(q\right) \overbrace{\sum_{x\in\mathcal{X}} \left| D_{\mathcal{X}}\left(x\right) - D^G_{\mathcal{X}|\mathcal{Q}_\Delta}\left(x \,|\, q\right) \right|}^{=2\ell^G_{D_{\mathcal{X}^n}}(q)} \cdot \Delta \\
&= 2\Delta \cdot \left(\overbrace{\sum_{q\notin Q^G_{D_{\mathcal{X}^n}}(\epsilon)} D^G_{\mathcal{Q}_\Delta}\left(q\right)}^{\leq 1} \cdot \overbrace{\ell^G_{D_{\mathcal{X}^n}}\left(q\right)}^{\leq\epsilon} + \overbrace{\sum_{q\in Q^G_{D_{\mathcal{X}^n}}(\epsilon)} D^G_{\mathcal{Q}_\Delta}\left(q\right)}^{<\delta} \cdot \overbrace{\ell^G_{D_{\mathcal{X}^n}}\left(q\right)}^{\leq 1} \right) \\
&<^{(2)} 2\Delta\left(\epsilon + \delta\right),
\end{aligned}$$

where (1) results from the definition of $\mathcal{Q}_\Delta$ and the triangle inequality, and (2) from the condition that $D^G_{\mathcal{R}}\left(Q^G_{D_{\mathcal{X}^n}}(\epsilon)\right) < \delta$. □

Corollary 3.4. *Given $0 \leq \epsilon, \delta \leq 1$, a distribution $D_{\mathcal{X}^n}$, and a generator G, if a mechanism $M : \mathcal{X}^n \times \mathcal{G} \to \mathcal{Q}_\Delta$ is (ϵ, δ)-LSS with respect to $D_{\mathcal{X}^n}, G$, then*

$$\left| \mathop{\mathbb{E}}_{q=M(S,G)} \left[q\left(S\right) - q\left(D_{\mathcal{X}^n}\right) \right] \right| < 2\Delta\left(2\epsilon + \frac{\delta}{\epsilon}\right).$$

Proof. This is a direct result of combining Theorem 3.3 with Lemma 2.3. □

We proceed to lift this guarantee from expectation to high probability, using a thought experiment known as the *Monitor Mechanism*, which was introduced by [BNS+16]. Intuitively, it runs a large number of independent copies of an underlying mechanism, and exposes the results of the least-distribution-accurate copy as its output.

Definition 3.5 (The Monitor Mechanism)**.** The *Monitor Mechanism* is a function $\text{Mon}_{\bar{M}} : (\mathcal{X}^n)^T \times \mathcal{A} \to \mathcal{Q} \times \mathbb{R} \times [T]$ which is parametrized by a sequence of k mechanisms $\bar{M}$ where $\forall i \in [k]$, $M_i : \mathcal{X}^n \times \mathcal{Q} \to \mathbb{R}$. Given a series of sample sets $\bar{S} \in (\mathcal{X}^n)^T$ and adversary $A \in \mathcal{A}$ as input, it runs the adaptive mechanism between $\bar{M}$ and A for T independent times (which in particular means neither of them share state across those iterations) and outputs a query $q \in \mathcal{Q}$, response $r \in \mathbb{R}$ and index $t \in T$, based on the following process:

Monitor Mechanism $\text{Mon}_{\bar{M}}$

Input: $\bar{S} \in (\mathcal{X}^n)^T, A \in \mathcal{A}$
Output: $q \in \mathcal{Q}, r \in \mathbb{R}, t \in T$
for $t = 1, ..., T$:
  $v^t \leftarrow \text{Adp}_{\bar{M}}(S_t, A)$
  $\left(\tilde{q}^t, \tilde{r}^t\right) \leftarrow \underset{(q,r)\in v^t}{\arg\max} \left|q\left(D_{\mathcal{X}^n}\right) - r\right|$[a]
  if $\tilde{q}^t\left(D_{\mathcal{X}^n}\right) \geq \tilde{r}^t$:[b]
    $q^t \leftarrow \tilde{q}^t$
    $r^t \leftarrow \tilde{r}^t$
  else:
    $q^t \leftarrow -\tilde{q}^t$
    $r^t \leftarrow -\tilde{r}^t$
$t^* \leftarrow \underset{t\in[T]}{\arg\max} \left(q^t\left(D_{\mathcal{X}^n}\right) - r^t\right)$
return $\left(q^{t^*}, r^{t^*}, t^*\right)$

[a] We slightly abuse notation since q is not part of v^t, but since it can be recovered from it, this term is well defined.
[b] The addition of this condition ensures that $q\left(D_{\mathcal{X}^n}\right) \geq r$ for the output of the mechanism, a fact that will be used later in the proof of Lemma 3.8.

Notice that the monitor mechanism makes use of the ability to evaluate queries according to the true underlying distribution.[10]

We begin by proving a few properties of the monitor mechanism. In the following claims, the probabilities and expectations are taken over the randomness of the choice of $\bar{S} \in (\mathcal{X}^n)^T$ (which is assumed to be drawn iid from $D_{\mathcal{X}^n}$) and the internal probability of $\text{Adp}_{\bar{M}}$.

Claim 3.6. *Given $0 \leq \epsilon, \delta \leq 1$, $T \in \mathbb{N}$, a distribution $D_{\mathcal{X}^n}$, and an adversary $A : \mathbb{V} \to \mathcal{Q}_\Delta$, if a sequence of k mechanisms $\bar{M}$ where $\forall i \in [k]$, $M_i : \mathcal{X}^n \times \mathcal{Q}_\Delta \to \mathbb{R}$ is (ϵ, δ)-k-LSS with respect to $D_{\mathcal{X}^n}, A$, then*

$$\left|\underset{(q,r,t)=\mathit{Mon}_{\bar{M}}\left(\bar{S},A\right)}{\mathbb{E}}\left[q\left(D_{\mathcal{X}^n}\right) - q\left(S_t\right)\right]\right| < 2\Delta\left(2\epsilon + \frac{T\delta}{\epsilon}\right).$$

Claim 3.7. *Given $0 \leq \epsilon, \delta \leq 1$, $T \in \mathbb{N}$, a distribution $D_{\mathcal{X}^n}$, and an adversary $A : \mathbb{V} \to \mathcal{Q}_\Delta$, if a sequence of k mechanisms $\bar{M}$ where $\forall i \in [k], M_i : \mathcal{X}^n \times \mathcal{Q}_\Delta \to \mathbb{R}$ is (ϵ, δ)-k-Sample Accurate with respect to $D_{\mathcal{X}^n}, A$, then*

$$\underset{(q,r,t)=\mathit{Mon}_{\bar{M}}\left(\bar{S},A\right)}{\mathbb{E}}\left[q\left(S_t\right) - r\right] \leq \epsilon + 2T\delta\Delta.$$

Claim 3.8. *Given $0 \leq \epsilon, \delta \leq 1$, $T \in \mathbb{N}$, a distribution $D_{\mathcal{X}^n}$, and an adversary $A : \mathbb{V} \to \mathcal{Q}_\Delta$, if a a sequence of k mechanisms $\bar{M}$ where $\forall i \in [k], M_i : \mathcal{X}^n \times \mathcal{Q}_\Delta \to \mathbb{R}$ is **not** (ϵ, δ)-k-Distribution Accurate with respect to $D_{\mathcal{X}^n}, A$, then*

$$\underset{(q,r,t)=\mathit{Mon}_{\bar{M}}\left(\bar{S},A\right)}{\mathbb{E}}\left[q\left(D_{\mathcal{X}^n}\right) - r\right] > \epsilon\left(1 - (1-\delta)^T\right).$$

[10] Of course, no realistic mechanism would have such an ability; the monitor mechanism is simply a thought experiment used as a proof technique.

Finally, we combine these claims to show that LSS implies generalization with high probability.

Theorem 3.9 (LSS implies generalization)**.** *Given $0 \leq \epsilon \leq \Delta$, $0 \leq \delta \leq 1$, a distribution $D_{\mathcal{X}^n}$, and an adversary $A : \mathbb{V} \to \mathcal{Q}_\Delta$, if a sequence of k mechanisms $\bar{M}$ where $\forall i \in [k], M_i : \mathcal{X}^n \times \mathcal{Q}_\Delta \to \mathbb{R}$ is both $\left(\frac{\epsilon}{8\Delta}, \frac{\epsilon^2\delta}{4800\Delta^2}\right)$-$k$-LSS and $\left(\frac{\epsilon}{8}, \frac{\epsilon\delta}{600\Delta}\right)$-$k$-Sample Accurate with respect to $D_{\mathcal{X}^n}$ and A, then it is (ϵ, δ)-k-Distribution Accurate with respect to $D_{\mathcal{X}^n}$ and A.*

Proof. We will prove a slightly more general claim. For every $0 < a, b, c, d$ such that $4a + 2b + c + 2d < 1 - e^{-1}$, say M is both $\left(a\frac{\epsilon}{\Delta}, ab\frac{\epsilon^2\delta}{\Delta^2}\right)$-$k$-LSS and $\left(c\epsilon, d\frac{\epsilon\delta}{\Delta}\right)$-$k$-Sample Accurate and assume M is not (ϵ, δ)-k-Distribution Accurate.

Setting $T = \left\lfloor \frac{1}{\delta} \right\rfloor$, we see

$$\begin{aligned}\left| \mathop{\mathbb{E}}_{(q,r,t)=\text{Mon}_{\bar{M}}(\bar{S},A)} \left[q\left(D_{\mathcal{X}^n}\right) - q\left(S_t\right)\right] \right| &<^{(1)} 2\Delta\left(2\frac{a\epsilon}{\Delta} + \frac{T\Delta}{a\epsilon} \cdot \frac{ab\epsilon^2\delta}{\Delta^2}\right) \\ &\leq^{(2)} (4a + 2b)\,\epsilon,\end{aligned}$$

where (1) is a direct result of Claim 3.6 and (2) uses the definition of T.

But on the other hand,

$$\begin{aligned}\left| \mathop{\mathbb{E}}_{(q,r,t)=\text{Mon}_{\bar{M}}(\bar{S},A)} \left[q\left(D_{\mathcal{X}^n}\right) - q\left(S_t\right)\right] \right| &\geq^{(1)} \left| \mathop{\mathbb{E}}_{(q,r,t)=\text{Mon}_{\bar{M}}(\bar{S},A)} \left[q\left(D_{\mathcal{X}^n}\right) - r\right] \right| - \left| \mathop{\mathbb{E}}_{(q,r,t)=\text{Mon}_{\bar{M}}(\bar{S},A)} \left[q\left(S_t\right) - r\right] \right| \\ &>^{(2)} \epsilon\left(1 - (1-\delta)^T\right) - \left(c\epsilon + 2T \cdot \frac{d\epsilon\delta}{\Delta}\Delta\right) \\ &>^{(3)} \epsilon\left(1 - e^{-\delta\left\lfloor\frac{1}{\delta}\right\rfloor}\right) - (c + 2d)\,\epsilon \\ &\geq^{(4)} \epsilon\left(1 - e^{-1}\right) - (c + 2d)\,\epsilon \\ &>^{(5)} (4a + 2b)\,\epsilon,\end{aligned}$$

where (1) is the triangle inequality, (2) uses Claims 3.7 and 3.8, (3) the definition of T, (4) the inequality $1 - \delta \leq e^{-\delta}$, and (5) the definition of a, b, c, d. Since combining all of the above leads to a contradiction, we know that $\bar{M}$ must be (ϵ, δ)-Distribution Accurate, which concludes the proof. The theorem was stated choosing $a = c = \frac{1}{8}, b = d = \frac{1}{600}$. □

3.2 LSS is Necessary for Generalization

We next show that a mechanism that is not LSS cannot be both sample accurate and distribution accurate. In order to prove this theorem, we show how to construct a "bad" query.

Definition 3.10 (Loss assessment query)**.** Given a generator G and a response r, we will define the *Loss assessment query* $\tilde{q}_r$ as

$$\tilde{q}_r\left(x\right) = \begin{cases} \Delta & D_{\mathcal{X}}\left(x\right) > D^G_{\mathcal{X}|\mathcal{R}}\left(x \,|\, r\right) \\ -\Delta & D_{\mathcal{X}}\left(x\right) \leq D^{'}_{\mathcal{X}|\mathcal{R}}G\left(x \,|\, r\right) \end{cases}.$$

Intuitively, this function maximizes the difference between $\mathop{\mathbb{E}}_{\mathcal{X}}\left[\tilde{q}_r\left(x\right)\right]$ and $\mathop{\mathbb{E}}_{\mathcal{X}|\mathcal{R}}\left[\tilde{q}_r\left(x\right)\right]$, and as a result, the potential to overfit.[11]

[11] The fact that we are able to define such a query is a result of the way the distance measure of LSS treats the x's and the fact that it is defined over $\mathcal{X}$ and not $\mathcal{X}^n$.

This function is used to lower bound the effect of the stability loss on the expected overfitting.

Lemma 3.11 (Loss assessment query overfits in expectation)**.** *Given $0 \leq \epsilon, \delta \leq 1$, a distribution $D_{\mathcal{X}^n}$, a generator G, and a mechanism $M : \mathcal{X}^n \times \mathcal{G} \to \mathcal{R}$, if $D^G_{\mathcal{R}}\left(R^G_{D_{\mathcal{X}^n}}(\epsilon)\right) > \delta$, then there is a function $f : \mathcal{R} \to \mathcal{Q}_\Delta$ such that,*

$$\left|\mathop{\mathbb{E}}_{q=f\circ M(S,G)}\left[q\left(D_{\mathcal{X}^n}\right) - q\left(S\right)\right]\right| > 2\epsilon\Delta\delta.$$

Proof. Choosing $f(r) = q_r$ we get that,

$$\begin{aligned}
\left|\mathop{\mathbb{E}}_{q=f\circ M(S,G)}\left[q\left(D_{\mathcal{X}^n}\right) - q\left(S\right)\right]\right| &=^{(1)} \left|\sum_{q\in\mathcal{Q}} D^G_{\mathcal{Q}}(q) \cdot \sum_{x\in\mathcal{X}} \left(D_{\mathcal{X}}(x) - D^G_{\mathcal{X}|\mathcal{Q}}(x \mid q)\right) \cdot q(x)\right| \\
&= \left|\sum_{r\in\mathcal{R}} D^G_{\mathcal{R}}(r) \cdot \sum_{x\in\mathcal{X}} \left(D_{\mathcal{X}}(x) - D^G_{\mathcal{X}|\mathcal{R}}(x \mid r)\right) \cdot \tilde{q}_r(x)\right| \\
&\geq^{(2)} \overbrace{\sum_{r\in R^G_{D_{\mathcal{X}^n}}(\epsilon)} D^G_{\mathcal{R}}(r)}^{\geq\delta} \cdot \overbrace{\sum_{x\in\mathcal{X}} \left|D_{\mathcal{X}}(x) - D^G_{\mathcal{X}|\mathcal{R}}(x \mid r)\right|}^{=2\ell(r)>2\epsilon} \cdot \Delta \\
&>^{(3)} 2\epsilon\Delta\delta
\end{aligned}$$

where (1) was justified in the proof of Theorem 3.3, (2) results from the definition of the loss assessment query, and (3) from the definition of $R^G_{D_{\mathcal{X}^n}}(\epsilon)$. □

We use this method for constructing an overfitting query for non-LSS mechanism, in a slight modification of the Monitor Mechanism.

Definition 3.12 (The Second Monitor Mechanism)**.** The *Second Monitor Mechanism* is a function $\text{Mon2}_{\bar{M}} : (\mathcal{X}^n)^T \times \mathcal{A} \to \mathcal{Q} \times \mathbb{R} \times [T]$ which is parametrized by a sequence of k mechanisms $\bar{M}$ where $\forall i \in [k], M_i : \mathcal{X}^n \times \mathcal{Q} \to \mathbb{R}$. Given a series of sample sets $\bar{S} \in (\mathcal{X}^n)^T$ and adversary $A \in \mathcal{A}$ as input, it runs the adaptive mechanism between $\bar{M}$ and A for T independent times and outputs a query $q \in \mathcal{Q}$, response $r \in \mathbb{R}$ and index $t \in T$, based on the following process:

Second Monitor Mechanism $\text{Mon2}_{\bar{M}}$

Input: $\bar{S} \in (\mathcal{X}^n)^T, A \in \mathcal{A}$
Output: $q \in \mathcal{Q}, r \in \mathbb{R}, t \in T$
for $t = 1, ..., T$:
  $v^t \leftarrow \text{Adp}_{\bar{M}}(S_t, A)$
  $q^t \leftarrow \tilde{q}_{v^t}$
  $r^t \leftarrow M\left(S, q^t\right)$
$t^* \leftarrow \arg\max_{t\in[T]} \left(\ell^A_{D_{\mathcal{X}^n}}\left(v^t\right)\right)$
return $\left(q^{t^*}, r^{t^*}, t^*\right)$

Using this mechanism, we show that LSS is necessary in order for a mechanism to be both sample accurate and distribution accurate.

Theorem 3.13 (Necessity of LSS for Generalization)**.** *Given $0 \leq \epsilon \leq \Delta$, $0 \leq \delta \leq 1$, a distribution $D_{\mathcal{X}^n}$, and an adversary $A : \mathbb{V} \to \mathcal{Q}_\Delta$, if a sequence of k mechanisms $\bar{M}$ where $\forall i \in [k], M_i : \mathcal{X}^n \times \mathcal{Q}_\Delta \to \mathbb{R}$ is not $\left(\frac{\epsilon}{\Delta}, \delta\right)$-$k$-LSS, then it cannot be both $\left(\frac{\epsilon}{5}, \frac{\epsilon\delta}{5\Delta}\right)$ $(k+1)$-Distribution Accurate and $\left(\frac{\epsilon}{5}, \frac{\epsilon\delta}{5\Delta}\right)$ $(k+1)$-Sample Accurate.*

Proof. Again we will prove a slightly more general claim. For every $0 < a, b, c, d$ such that $a+2b+c+2d < 2\left(1-e^{-1}\right)$, say M is both $\left(a\epsilon, b\frac{\epsilon\delta}{\Delta}\right)$ $(k+1)$-Sample Accurate and $\left(c\epsilon, d\frac{\epsilon\delta}{\Delta}\right)$ $(k+1)$-Distribution Accurate and assume M is not $\left(\frac{\epsilon}{\Delta}, \delta\right)$-$k$-LSS.

First notice that if $\bar{M}$ is not $\left(\frac{\epsilon}{\Delta}, \delta\right)$-$k$-LSS with respect to $D_{\mathcal{X}^n}, A$, then in particular $D^A_{\mathcal{V}_k}\left(V^{A,k}_{D_{\mathcal{X}^n}}\left(\frac{\epsilon}{\Delta}\right)\right) \geq \delta$. Since the T rounds of the second monitor mechanism are independent and t^* is the index of the round with the maximal stability loss of the calculated query, we get that

$$\mathbb{P}_{\bar{S},\text{Mon2}_{\bar{M}}}\left(v^{t*} \in V^{A,k}_{D_{\mathcal{X}^n}}\left(\frac{\epsilon}{\Delta}\right)\right) > 1-(1-\delta)^T.$$

Combining this fact with Lemma 3.11, and setting $T = \left\lfloor \frac{1}{\delta} \right\rfloor$ we get on one hand,

$$\begin{aligned}
\left|\mathop{\mathbb{E}}_{(q,r,t)=\text{Mon2}_{\bar{M}}\left(\bar{S},A\right)}\left[q\left(D_{\mathcal{X}^n}\right) - q\left(S_t\right)\right]\right| &\geq^{(1)} 2\frac{\epsilon}{\Delta}\Delta\left(1-(1-\delta)^T\right) \\
&>^{(2)} 2\epsilon\left(1-e^{-\delta\left\lfloor\frac{1}{\delta}\right\rfloor}\right) \\
&>^{(3)} 2\epsilon\left(1-e^{-1}\right),
\end{aligned}$$

where (1) is a direct result of invoking Lemma 3.11 with $1-(1-\delta)^T$ for δ, (2) uses the definition of T and (3) uses the inequality $1-\delta \leq e^{-\delta}$.

But on the other hand,

$$\begin{aligned}
\left|\mathop{\mathbb{E}}_{q=f\circ M(S,G)}\left[q\left(D_{\mathcal{X}^n}\right) - q\left(S\right)\right]\right| &\leq^{(1)} \left|\mathop{\mathbb{E}}_{(q,r,t)=\text{Mon2}_{\bar{M}}\left(\bar{S},A\right)}\left[q\left(D_{\mathcal{X}^n}\right) - r\right]\right| + \left|\mathop{\mathbb{E}}_{(q,r,t)=\text{Mon2}_{\bar{M}}\left(\bar{S},A\right)}\left[q\left(S_t\right) - r\right]\right| \\
&<^{(2)} \left(a\epsilon + 2T\cdot\frac{b\epsilon\delta}{\Delta}\Delta\right) + \left(c\epsilon + 2T\cdot\frac{d\epsilon\delta}{\Delta}\Delta\right) \\
&\leq^{(3)} \left(a+2b+c+2d\right)\epsilon \\
&<^{(4)} 2\epsilon\left(1-e^{-1}\right),
\end{aligned}$$

where (1) is the triangle inequality, (2) uses Claim 3.7 which was mentioned with relation to the original monitor mechanism (this time for the distribution error as well), (3) uses the definition of T, and (4) the definition of a, b, c, d.

Since combining all of the above leads to a contradiction, we know that $\bar{M}$ cannot be $\left(\frac{\epsilon}{\Delta}, \delta\right)$-$k$-LSS, which concludes the proof. The theorem was stated choosing $a = b = c = d = \frac{1}{5}$. □

4 Relationship to other notions of stability

In this section we discuss the relationship between LSS and a few common notions of stability.

4.1 Definitions

In the following definitions, $\mathcal{X}, D_{\mathcal{X}}, \mathcal{Q}, \mathcal{R}, M, \epsilon, \delta$ and n will be used at the same manner as previously defined.

Definition 4.1 (Differential Privacy [DMNS06])**.** Given $0 \leq \epsilon$, $0 \leq \delta \leq 1$, and a generator G, a mechanism $M : \mathcal{X}^n \times \mathcal{G} \to \mathcal{R}$ will be called (ϵ, δ)-*differentially-private with respect to* G (or *DP*, for short) if for

any $S_1, S_1 \in \mathcal{X}^n$ that differ only in one element and any $R \subseteq \mathcal{R}$, the two distributions defined over $\mathcal{R}$ by $M(S_1, G)$ and $M(S_2, G)$ are (ϵ, δ)-indistinguishable (see definition in Appendix E). In other words,

$$\mathbb{P}_M(M(S_1, G) \in R) \leq e^{\epsilon} \cdot \mathbb{P}_M(M(S_2, G) \in R) + \delta$$

where the probability is taken over the internal randomness of M. Notice that in this definition, there is no probabilistic aspect in the choice of S, and the bound is defined on the worst case.

Definition 4.2 (Max Information [DFH+15a])**.** Given $0 \leq \epsilon, 0 \leq \delta \leq 1$, a distribution $D_{\mathcal{X}^n}$, and a generator G, a mechanism $M : \mathcal{X}^n \times \mathcal{G} \to \mathcal{R}$ *has δ-approximate max-information of ϵ with respect to $D_{\mathcal{X}^n}, G$* (or *MI*, for short) if for any $B \subseteq \mathcal{X}^n \times \mathcal{R}$, the two distributions $D^G_{(\mathcal{X}^n, \mathcal{R})}$ and $D^G_{\mathcal{X}^n \otimes \mathcal{R}}$ over $\mathcal{X}^n \times \mathcal{R}$ are (ϵ, δ)-indistinguishable. In other words,

$$\begin{aligned} D^G_{(\mathcal{X}^n, \mathcal{R})}(B) &\leq e^{\epsilon} \cdot D^G_{\mathcal{X}^n \otimes \mathcal{R}}(B) + \delta \\ &\text{and} \\ D^G_{\mathcal{X}^n \otimes \mathcal{R}}(B) &\leq e^{\epsilon} \cdot D^G_{(\mathcal{X}^n, \mathcal{R})}(B) + \delta. \end{aligned}$$

Some definitions replace e with 2 as the base of ϵ.

Notice that in a way, MI is a natural relaxation of DP, where instead of considering only the probability which is induced by the mechanism.

Definition 4.3 (Local Max Information (LMI))**.** Given $0 \leq \epsilon$, $0 \leq \delta \leq 1$, a distribution $D_{\mathcal{X}^n}$ and a generator G, a mechanism M will be said to satisfy (ϵ, δ)*-Local-Max-Information with respect to $D_{\mathcal{X}^n}$ and G* (or *LMI*, for short), if for any $B \subseteq \mathcal{X} \times \mathcal{R}$, the two distributions $D^G_{(\mathcal{X}, \mathcal{R})}$ and $D^G_{\mathcal{X} \otimes \mathcal{R}}$ over $\mathcal{X} \times \mathcal{R}$ are (ϵ, δ)-indistinguishable. In other words,

$$D^G_{(\mathcal{X}, \mathcal{R})}(B) \leq e^{\epsilon} \cdot D^G_{\mathcal{X} \otimes \mathcal{R}}(B) + \delta \quad \text{and} \quad D^G_{\mathcal{X} \otimes \mathcal{R}}(B) \leq e^{\epsilon} \cdot D^G_{(\mathcal{X}, \mathcal{R})}(B) + \delta.$$

In a way, this definition takes the relaxation one step further, by moving from the distribution over the sample sets to distribution over the sample elements. Unlike the previous definitions, this one was not presented yet as far as we know.

These definition can be extended to apply to a family of generators and/or a family of possible distributions, just like the LSS definition.

Definition 4.4 (Compression Scheme [LW86])**.** Given an integer $m < \frac{n}{2}$ and a generator G, a mechanism M will be said to have a *compression scheme* of size m with respect to G (or *CS*, for short), if M can be described as the composition $f_G \circ g_G$ where the *compression function* $g_G : \mathcal{X}^n \to \mathcal{X}^m$ has the property that $g_G(S) \subset S$ and $f_G : \mathcal{X}^m \to \mathcal{R}$ is some arbitrary function which will be called the *encoding function*. Both functions might be non deterministic. We will denote $W := g(S)$ and $r_W := f(W)$.[12]

One simple case is when f is the identity function, and the mechanism releases m sample elements.

4.2 Implications

Prior work ([DFH+15a] and [RRST16]) showed that bounded DP implies bounded MI (in the case of $\delta > 0$, this holds only if the underlying distribution is a product distribution [De12])). We prove that both DP and MI imply LMI (in the case of DP, only for product distributions). All proofs for this subsection can be found in Appendix D.1.

[12] some versions include the option of receiving some side information, i.e. the coin tosses of g.

Theorem 4.5 (Differential Privacy implies Local Max Information)**.** *Given $0 \leq \epsilon$, $0 \leq \delta \leq 1$, a distribution $D_{\mathcal{X}}$, and a generator G, if a mechanism M is (ϵ, δ)-DP with respect to G then it is (ϵ, δ)-LMI with respect to the same G and the product distribution over $\mathcal{X}^n$ induced by $D_{\mathcal{X}}$.*

Theorem 4.6 (Max Information implies Local Max Information)**.** *Given $0 \leq \epsilon$, $0 \leq \delta \leq 1$, a distribution $D_{\mathcal{X}^n}$ and a generator G, if a mechanism M has δ-approximate max-information of ϵ with respect to $D_{\mathcal{X}^n}$ and G then it is $\left(\epsilon, \frac{n+1}{2}\delta\right)$-LMI with respect to the same $D_{\mathcal{X}^n}$ and G.*

These two theorems follow naturally from the fact that LMI is a fairly direct relaxation of both DP and MI. We next show that LMI implies LSS.

Theorem 4.7 (Local Max Information implies Local Statistical Stability)**.** *Given $0 \leq \delta \leq \epsilon \leq \frac{1}{3}$, a distribution $D_{\mathcal{X}^n}$ and a generator G, if a mechanism M is (ϵ, δ)-LMI with respect to $D_{\mathcal{X}^n}$ and G, then it is $\left(3\epsilon, \frac{\delta}{\epsilon}\right)$-LSS with respect to the same $D_{\mathcal{X}^n}$ and G.*

We also prove that Compression Schemes imply LSS. This results from the fact that releasing information based on a restricted number of sample elements has a limited effect on the posterior distribution on one element of the sample set.

Theorem 4.8 (Compressibility implies Local Statistical Stability)**.** *Given $0 \leq \delta \leq 1$, an integer $m \leq \frac{n}{9\ln\left(\frac{2n}{\delta}\right)}$, a distribution $D_{\mathcal{X}}$, and a generator $G \in \mathcal{G}$, if a mechanism $M : \mathcal{X}^n \times \mathcal{G} \to \mathcal{R}$ has a compression scheme of size m then it is (ϵ, δ)-LSS with respect to the same G and the product distribution over $\mathcal{X}^n$ induced by $D_{\mathcal{X}}$, for any $\epsilon > 11\sqrt{\frac{m \ln 2n/\delta}{n}}$.*[13]

4.3 Separations

Finally, we show that MI is a strictly stronger requirement than LMI, and LMI is a strictly stronger requirement then LSS. Proofs of these theorems appear in Appendix D.2.

Theorem 4.9 (Max Information is strictly stronger than Local Max Information)**.** *For any $0 < \epsilon \leq 0.07$, $n > \frac{98}{\epsilon^2}\ln\left(\frac{7}{\epsilon}\right)$, the mechanism which outputs the parity function of the sample set is $(\epsilon, 0)$-LMI but not $\left(1, \frac{9}{50}\right)$-MI.*

Theorem 4.10 (Local Max Information is strictly stronger than Local Statistical Stability)**.** *For any $0 \leq \delta \leq 0.1$, $n > 2\ln\left(\frac{2}{\delta}\right)$, a mechanism which uniformly samples and outputs one sample element is $\left(11\sqrt{\frac{\ln 2n/\delta}{n}}, \delta\right)$-LSS but is not $\left(1, \frac{1}{2n}\right)$-LMI.*

5 Discussion

In moving away from the study of worst-case data sets (as is common in previous stability notions) to averaging over sample sets and over data elements of those sets, we hope that the Local Statistical Stability notion will enable new progress in the study of generalization under adaptive data analysis. This averaging, potentially leveraging a sort of "natural noise" from the data sampling process, may enable the development of new algorithms to preserve generalization, and may also support tighter bounds on the implications of existing algorithms.

One might also hope that realistic adaptive learning settings are not adversarial, and might therefore enjoy even better generalization guarantees. LSS may be a tool for understanding the generalization properties of algorithms of interest (as opposed to worst-case queries or adversaries; see e.g. [GK16], [ZH19]), in more realistic settings (e.g., limited families of data distributions [BF16]).

[13] In case g releases some side information, the number of bits required to describe this information is added to the m factor in the bound on ϵ.

References

[Ala17] Ibrahim Alabdulmohsin. An information-theoretic route from generalization in expectation to generalization in probability. In *Artificial Intelligence and Statistics*, pages 92–100, 2017.

[BE02] Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

[BF16] Raef Bassily and Yoav Freund. Typical stability. *arXiv preprint arXiv:1604.03336*, 2016.

[BMN+17] Raef Bassily, Shay Moran, Ido Nachum, Jonathan Shafer, and Amir Yehudayoff. Learners that use little information. *arXiv preprint arXiv:1710.05233*, 2017.

[BNS+16] Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 1046–1059. ACM, 2016.

[CLN+16] Rachel Cummings, Katrina Ligett, Kobbi Nissim, Aaron Roth, and Zhiwei Steven Wu. Adaptive learning with robust generalization guarantees. In *Conference on Learning Theory*, pages 772–814, 2016.

[De12] Anindya De. Lower bounds in differential privacy. In *Theory of cryptography conference*, pages 321–338. Springer, 2012.

[DFH+15a] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toni Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems*, pages 2350–2358, 2015.

[DFH+15b] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 117–126. ACM, 2015.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[DP09] Devdatt P Dubhashi and Alessandro Panconesi. *Concentration of measure for the analysis of randomized algorithms*. Cambridge University Press, 2009.

[DR+14] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[EGI19] Amedeo Roberto Esposito, Michael Gastpar, and Ibrahim Issa. A new approach to adaptive data analysis and learning via maximal leakage. *arXiv preprint arXiv:1903.01777*, 2019.

[FS17] Vitaly Feldman and Thomas Steinke. Calibrating noise to variance in adaptive data analysis. *arXiv preprint arXiv:1712.07196*, 2017.

[GK16] Arpita Ghosh and Robert Kleinberg. Inferential privacy guarantees for differentially private mechanisms. *arXiv preprint arXiv:1603.01508*, 2016.

[GL14] Andrew Gelman and Eric Loken. The statistical crisis in science. *American scientist*, 102(6):460, 2014.

[Ioa05] John PA Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8):e124, 2005.

[KS14] Shiva P Kasiviswanathan and Adam Smith. On the'semantics' of differential privacy: A bayesian formulation. *Journal of Privacy and Confidentiality*, 6(1), 2014.

[Kut02] Samuel Kutin. Extensions to mcdiarmid's inequality when differences are bounded with high probability. *Department Computer Science, University of Chicago, Chicago, IL. Technical report TR-2002-04*, 2002.

[LW86] Nick Littlestone and Manfred Warmuth. Relating data compression and learnability. 1986.

[NSS+18] Kobbi Nissim, Adam Smith, Uri Stemmer, Thomas Steinke, and Jonathan Ullman. The limits of post-selection generalization. In *Advances in Neural Information Processing Systems*, pages 6402–6411, 2018.

[RRST16] Ryan Rogers, Aaron Roth, Adam Smith, and Om Thakkar. Max-information, differential privacy, and post-selection hypothesis testing. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 487–494. IEEE, 2016.

[RRT+16] Maxim Raginsky, Alexander Rakhlin, Matthew Tsao, Yihong Wu, and Aolin Xu. Information-theoretic analysis of stability and bias of learning algorithms. In *Information Theory Workshop (ITW), 2016 IEEE*, pages 26–30. IEEE, 2016.

[RZ16] Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In *Artificial Intelligence and Statistics*, pages 1232–1240, 2016.

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[TV+15] Terence Tao, Van Vu, et al. Random matrices: universality of local spectral statistics of non-hermitian matrices. *The Annals of Probability*, 43(2):782–874, 2015.

[XR17] Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems*, pages 2524–2533, 2017.

[ZH19] Tijana Zrnic and Moritz Hardt. Natural analysts in adaptive data analysis. *arXiv preprint arXiv:1901.11143*, 2019.

A Distributions: Formal Definitions

Definition A.1 (Distributions over $\mathcal{X}^n$ and $\mathcal{R}$)**.** A distribution $D_{\mathcal{X}^n}$, a generator G, and a mechanism $M : \mathcal{X}^n \times \mathcal{G} \to \mathcal{R}$, together induce a set of distributions over $\mathcal{X}^n$, $\mathcal{R}$, and $\mathcal{X}^n \times \mathcal{R}$.

The *conditional distribution* $D^G_{\mathcal{R}|\mathcal{X}^n}$ over $\mathcal{R}$ represents the probability to get r as the output of $M(S, G)$. That is, $\forall S \in \mathcal{X}^n, r \in \mathcal{R}$,

$$D^G_{\mathcal{R}|\mathcal{X}^n}(r \mid S) \coloneqq \mathbb{P}_M(M(S, G) = r),$$

where the probability is taken over the internal randomness of M.

The *joint distribution* $D^G_{(\mathcal{X}^n,\mathcal{R})}$ over $\mathcal{X}^n \times \mathcal{R}$ represents the probability to sample a particular S and get r as the output of $M(S, G)$. That is, $\forall S \in \mathcal{X}^n, r \in \mathcal{R}$,

$$D^G_{(\mathcal{X}^n,\mathcal{R})}(S, r) \coloneqq D_{\mathcal{X}^n}(S) \cdot D^G_{\mathcal{R}|\mathcal{X}^n}(r \mid S).$$

The *unconditional distribution* $D^G_{\mathcal{R}}$ over $\mathcal{R}$ represents the prior probability to get output r without any knowledge of S. That is, $\forall r \in \mathcal{R}$,

$$D^G_{\mathcal{R}}(r) \coloneqq \sum_{S \in \mathcal{X}^n} D^G_{(\mathcal{X}^n,\mathcal{R})}(S, r).$$

The *disjoint distribution* $D^G_{\mathcal{X}^n \otimes \mathcal{R}}$ over $\mathcal{X}^n \times \mathcal{R}$ represents the probability to sample S and get r as the output of $M(\cdot, G)$ independently. That is, $\forall S \in \mathcal{X}, r \in \mathcal{R}$,

$$D^G_{\mathcal{X}^n \otimes \mathcal{R}}(S, r) \coloneqq D_{\mathcal{X}^n}(S) \cdot D^G_{\mathcal{R}}(r).$$

The *conditional distribution* $D^G_{\mathcal{X}^n|\mathcal{R}}$ over $\mathcal{X}^n$ represents the posterior probability that the sample set was S given that $M(\cdot, G)$ returns r. That is, $\forall S \in \mathcal{X}^n, r \in \mathcal{R}$,

$$D^G_{\mathcal{X}^n|\mathcal{R}}(S \mid r) \coloneqq \frac{D^G_{(\mathcal{X}^n,\mathcal{R})}(S, r)}{D^q_{\mathcal{R}}(r)}.$$

Definition A.2 (Distributions over $\mathcal{X}$ and $\mathcal{R}$)**.** The *joint distribution* $D^G_{(\mathcal{X},\mathcal{R})}$ over $\mathcal{X} \times \mathcal{R}$ represents the probability to get x as the output of $\mathrm{Sam}(\cdot)$ and also get r as the output of $M(\cdot, G)$ from the same sample set. That is, $\forall x \in \mathcal{X}, r \in \mathcal{R}$,

$$D^G_{(\mathcal{X},\mathcal{R})}(x, r) \coloneqq \sum_{S \in \mathcal{X}^n} D_{\mathcal{X}^n}(S) \cdot D_{\mathcal{X}|\mathcal{X}^n}(x \mid S) \cdot D^G_{\mathcal{R}|\mathcal{X}^n}(r \mid S).$$

The *disjoint distribution* $D^G_{\mathcal{X} \otimes \mathcal{R}}$ over $\mathcal{X} \times \mathcal{R}$ represents the probability to get x as the output of $\mathrm{Sam}(\cdot)$ and get r as the output of $M(\cdot, G)$ independently. That is, $\forall x \in \mathcal{X}, r \in \mathcal{R}$,

$$D^G_{\mathcal{X} \otimes \mathcal{R}}(x, r) \coloneqq D_{\mathcal{X}}(x) \cdot D^G_{\mathcal{R}}(r).$$

The *conditional distribution* $D^G_{\mathcal{R}|\mathcal{X}}$ over $\mathcal{R}$ represents the probability to get r as the output of $M(\cdot, G)$, given the fact that we got x as the output of $\mathrm{Sam}(\cdot)$ from that sample set. That is, $\forall x \in \mathcal{X}, r \in \mathcal{R}$,

$$D^G_{\mathcal{R}|\mathcal{X}}(r \mid x) \coloneqq \sum_{S \in \mathcal{X}^n} D_{\mathcal{X}^n|\mathcal{X}}(S \mid x) \cdot D^G_{\mathcal{R}|\mathcal{X}^n}(r \mid S).$$

The *conditional distribution* $D^G_{\mathcal{X}|\mathcal{R}}$ over $\mathcal{X}$ represents the probability to get x as the output of $\mathrm{Sam}(\cdot)$, given the fact that we got r as the output of $M(\cdot, G)$ from that sample set. That is, $\forall x \in \mathcal{X}, r \in \mathcal{R}$,

$$D^G_{\mathcal{X}|\mathcal{R}}(x \mid r) \coloneqq \sum_{S \in \mathcal{X}^n} D^G_{\mathcal{X}^n|\mathcal{R}}(S \mid r) \cdot D_{\mathcal{X}|\mathcal{X}^n}(x \mid S).$$

Although all of these definitions depend on $D_{\mathcal{X}^n}$ and M, we omit these from the notation for simplicity, and will specifically denote the relevant distribution and/or mechanism when necessary for clarity. We also use D to denote the probability of a set: for $R \subseteq \mathcal{R}$, we define $D^G_{\mathcal{R}}(R) := \sum_{r \in R} D^G_{\mathcal{R}}(r)$.

We show an analogue of Bayes' rule for these distributions.

Proposition A.3. *Given any distribution $D_{\mathcal{X}^n}$, mechanism $M : \mathcal{X}^n \times \mathcal{G} \to \mathcal{R}$, and generator G,*

$$D^G_{(\mathcal{X},\mathcal{R})}(x,r) = D_{\mathcal{X}}(x) \cdot D^G_{\mathcal{R}|\mathcal{X}}(r \mid x) = D^G_{\mathcal{R}}(r) \cdot D^G_{\mathcal{X}|\mathcal{R}}(x \mid r).$$

Proof.

$$\begin{aligned}
D^G_{(\mathcal{X},\mathcal{R})}(x,r) &= \sum_{S \in \mathcal{X}^n} D_{\mathcal{X}^n}(S) \cdot D_{\mathcal{X}|\mathcal{X}^n}(x \mid S) \cdot D^G_{\mathcal{R}|\mathcal{X}^n}(r \mid S) \\
&= \sum_{S \in \mathcal{X}^n} D_{(\mathcal{X}^n,\mathcal{X})}(S,x) \cdot D^G_{\mathcal{R}|\mathcal{X}^n}(r \mid S) \\
&= D_{\mathcal{X}}(x) \cdot \sum_{S \in \mathcal{X}^n} \frac{D_{(\mathcal{X}^n,\mathcal{X})}(S,x)}{D_{\mathcal{X}}(x)} \cdot D^G_{\mathcal{R}|\mathcal{X}^n}(r \mid S) \\
&= D_{\mathcal{X}}(x) \cdot \sum_{S \in \mathcal{X}^n} D_{\mathcal{X}^n|\mathcal{X}}(S \mid x) \cdot D^G_{\mathcal{R}|\mathcal{X}^n}(r \mid S) \\
&= D_{\mathcal{X}}(x) \cdot D^G_{\mathcal{R}|\mathcal{X}}(r \mid x)
\end{aligned}$$

The same way

$$\begin{aligned}
D^G_{(\mathcal{X},\mathcal{R})}(x,r) &= \sum_{S \in \mathcal{X}^n} D_{\mathcal{X}^n}(S) \cdot D^G_{\mathcal{R}|\mathcal{X}^n}(r \mid S) \cdot D_{\mathcal{X}|\mathcal{X}^n}(x \mid S) \\
&= \sum_{S \in \mathcal{X}^n} D^G_{(\mathcal{X}^n,\mathcal{R})}(S,r) \cdot D_{\mathcal{X}|\mathcal{X}^n}(x \mid S) \\
&= D^G_{\mathcal{R}}(r) \cdot \sum_{S \in \mathcal{X}^n} \frac{D^G_{(\mathcal{X}^n,\mathcal{R})}(S,r)}{D^G_{\mathcal{R}}(r)} \cdot D_{\mathcal{X}|\mathcal{X}^n}(x \mid S) \\
&= D^G_{\mathcal{R}}(r) \cdot \sum_{S \in \mathcal{X}^n} D^G_{\mathcal{X}^n|\mathcal{R}}(S \mid r) \cdot D_{\mathcal{X}|\mathcal{X}^n}(x \mid S) \\
&= D^G_{\mathcal{R}}(r) \cdot D^G_{\mathcal{X}|\mathcal{R}}(x \mid r)
\end{aligned}$$

□

B Missing Details from Section 2

B.1 Proof of Post-Processing Theorem

Missing parts from the proof of Theorem 2.4.

$$\begin{aligned}
\sum_{u \in U^G_{D_{\mathcal{X}^n}}(\epsilon)} D^G_{\mathcal{U}}(u) &= \sum_{u \in U^G_{D_{\mathcal{X}^n}}(\epsilon)} \sum_{r \in \mathcal{R}} D^G_{\mathcal{R}}(r) \cdot D_{\mathcal{U}|\mathcal{R}}(u \mid r) \\
&= \sum_{r \in \mathcal{R}} \overbrace{\sum_{u \in U^G_{D_{\mathcal{X}^n}}(\epsilon)} D_{\mathcal{U}|\mathcal{R}}(u \mid r)}^{=w^{\epsilon}_{\mathcal{U}}(r)} \cdot D^G_{\mathcal{R}}(r) \\
&= \sum_{r \in \mathcal{R}} w^{\epsilon}_{\mathcal{U}}(r) \cdot D^G_{\mathcal{R}}(r),
\end{aligned}$$

and

$$
\begin{aligned}
\sum_{u\in U^G_{D_{\mathcal{X}^n}}(\epsilon)} D^G_{\mathcal{U}}(u)\cdot \ell^G_{D_{\mathcal{X}^n}}(u) &= \sum_{u\in U^G_{D_{\mathcal{X}^n}}(\epsilon)} D^G_{\mathcal{U}}(u) \sum_{x\in X_+(u)} \left(D^G_{\mathcal{X}|\mathcal{U}}(x\,|\,u) - D_{\mathcal{X}}(x)\right) \\
&= \sum_{u\in U^G_{D_{\mathcal{X}^n}}(\epsilon)} \sum_{x\in X_+(u)} D_{\mathcal{X}}(x)\left(D^G_{\mathcal{U}|\mathcal{X}}(u\,|\,x) - D^G_{\mathcal{U}}(u)\right) \\
&= \sum_{u\in U^G_{D_{\mathcal{X}^n}}(\epsilon)} \sum_{r\in\mathcal{R}} \sum_{x\in X_+(u)} D_{\mathcal{X}}(x)\left(D^G_{\mathcal{R}|\mathcal{X}}(r\,|\,x) - D^G_{\mathcal{R}}(r)\right) D_{\mathcal{U}|\mathcal{R}}(u\,|\,r) \\
&\leq^{(1)} \sum_{u\in U^G_{D_{\mathcal{X}^n}}(\epsilon)} \sum_{r\in\mathcal{R}} \sum_{x\in X_+(r)} D_{\mathcal{X}}(x)\left(D^G_{\mathcal{R}|\mathcal{X}}(r\,|\,x) - D^G_{\mathcal{R}}(r)\right) D_{\mathcal{U}|\mathcal{R}}(u\,|\,r) \\
&= \sum_{r\in\mathcal{R}} \overbrace{\sum_{u\in U^G_{D_{\mathcal{X}^n}}(\epsilon)} D_{\mathcal{U}|\mathcal{R}}(u\,|\,r)}^{=w^\epsilon_{\mathcal{U}}(r)} \cdot D^G_{\mathcal{R}}(r) \sum_{x\in X_+(r)} \left(D^G_{\mathcal{X}|\mathcal{R}}(x\,|\,r) - D_{\mathcal{X}}(x)\right) \\
&= \sum_{r\in R} w^\epsilon_{\mathcal{U}}(r)\cdot D^G_{\mathcal{R}}(r)\cdot \ell^G_{\mathcal{R}}(r)\,.
\end{aligned}
$$

where (1) results from the definition of $X_+(r)$. □

B.2 Missing parts from the proofs of Adaptive Composition

Proof of Lemma 2.8. We begin by proving a set of relations between the prior distributions over $\mathcal{V}_{k+1}$ and the posterior distributions induced by the view v_k.

$$
\begin{aligned}
D^A_{(\mathcal{X}^n,\mathcal{V}_{k+1})}(S, v_{k+1}) &= D_{\mathcal{X}^n}(S)\cdot D^A_{\mathcal{V}_{k+1}|\mathcal{X}^n}(v_{k+1}\,|\,S) \\
&=^{(1)} D_{\mathcal{X}^n}(S)\cdot D^A_{\mathcal{V}_k|\mathcal{X}^n}(v_k\,|\,S)\cdot D^{G_{k+1}}_{\mathcal{R}|\mathcal{X}^n}(r_{k+1}\,|\,S, v_k) \\
&= D^A_{\mathcal{V}_k}(v_k)\cdot D^A_{\mathcal{X}^n|\mathcal{V}_k}(S\,|\,v_k)\cdot D^{G_{k+1}}_{\mathcal{R}|\mathcal{X}^n}(r_{k+1}\,|\,S, v_k) \\
&= D^A_{\mathcal{V}_k}(v_k)\cdot Z^{v_k}_{\mathcal{X}^n}(S)\cdot Z^{v_k,G_{k+1}}_{\mathcal{R}|\mathcal{X}^n}(r_{k+1}\,|\,S) \\
&= D^A_{\mathcal{V}_k}(v_k)\cdot Z^{v_k}_{(\mathcal{X}^n,\mathcal{R})}(S, r_{k+1})\,,
\end{aligned}
$$

where (1) is a result of the fact that G_{k+1} is a deterministic function of v_k. As mentioned in Definition 2.7, the distribution of r_{k+1} might depend on v_k in the case of a stateful mechanism, but it is all encapsulated in the definition of Z.

Using this identity and the definition of $Z^{A,v_k}_{\mathcal{X}^n}$ we get that,

$$
\begin{aligned}
D^A_{\mathcal{V}_{k+1}}(v_{k+1}) &= \sum_{S\in\mathcal{X}^n} D^A_{(\mathcal{X}^n,\mathcal{V}_{k+1})}(S, v_{k+1}) \\
&= \sum_{S\in\mathcal{X}^n} D^A_{\mathcal{V}_k}(v_k)\cdot Z^{v_k,G_{k+1}}_{(\mathcal{X}^n,\mathcal{R})}(S, r_{k+1}) \\
&= D^A_{\mathcal{V}_k}(v_k)\sum_{S\in\mathcal{X}^n} Z^{v_k,G_{k+1}}_{(\mathcal{X}^n,\mathcal{R})}(S, r_{k+1}) \\
&= D^A_{\mathcal{V}_k}(v_k)\cdot Z^{v_k,G_{k+1}}_{\mathcal{R}}(r_{k+1})\,.
\end{aligned}
$$

$$
\begin{aligned}
D^{A}_{\mathcal{X}|\mathcal{V}_k}\left(x \,|\, v_k\right) &= \sum_{S\in\mathcal{X}^n} D^{A}_{\mathcal{X}^n|\mathcal{V}_k}\left(S \,|\, v_k\right) \cdot D_{\mathcal{X}|\mathcal{X}^n}\left(x \,|\, S\right) \\
&= \sum_{S\in\mathcal{X}^n} Z^{v_k}_{\mathcal{X}^n}\left(S\right) \cdot D_{\mathcal{X}|\mathcal{X}^n}\left(x \,|\, S\right) \\
&= Z^{v_k}_{\mathcal{X}}\left(x\right)
\end{aligned}
$$

$$
\begin{aligned}
D^{A}_{\mathcal{X}|\mathcal{V}_{k+1}}\left(x \,|\, v_{k+1}\right) &= \sum_{S\in\mathcal{X}^n} D^{A}_{\mathcal{X}^n|\mathcal{V}_{k+1}}\left(S \,|\, v_{k+1}\right) \cdot D_{\mathcal{X}|\mathcal{X}^n}\left(x \,|\, S\right) \\
&= \sum_{S\in\mathcal{X}^n} Z^{v_k,G_{k+1}}_{\mathcal{X}^n|\mathcal{R}}\left(S \,|\, r_{k+1}\right) \cdot D_{\mathcal{X}|\mathcal{X}^n}\left(x \,|\, S\right) \\
&= Z^{v_k,G_{k+1}}_{\mathcal{X}|\mathcal{R}}\left(x \,|\, r_{k+1}\right)
\end{aligned}
$$

where we keep using the fact that, $D_{\mathcal{X}|\mathcal{X}^n}\left(x \,|\, S\right)$ does not depend on the underlying distribution $\mathcal{X}^n$ at all. Using these identities we can analyze the stability loss, and we would do so by invoking an equivalent definition of the statistical distance (see Appendix E),

$$
\begin{aligned}
\ell^{A}_{D_{\mathcal{X}^n}}\left(v_{k+1}\right) &= \frac{1}{2}\sum_{x\in\mathcal{X}} \left| D^{A}_{\mathcal{X}|\mathcal{V}_{k+1}}\left(x \,|\, v_{k+1}\right) - D_{\mathcal{X}}\left(x\right)\right| \\
&\leq^{(1)} \frac{1}{2}\sum_{x\in\mathcal{X}} \left| D^{A}_{\mathcal{X}|\mathcal{V}_{k}}\left(x \,|\, v_{k}\right) - D_{\mathcal{X}}\left(x\right)\right| \\
&\quad + \frac{1}{2}\sum_{x\in\mathcal{X}} \left| D^{A}_{\mathcal{X}|\mathcal{V}_{k+1}}\left(x \,|\, v_{k+1}\right) - D^{A}_{\mathcal{X}|\mathcal{V}_{k}}\left(x \,|\, v_{k}\right)\right| \\
&= \ell^{A}_{D_{\mathcal{X}^n}}\left(v_k\right) + \frac{1}{2}\sum_{x\in\mathcal{X}} \left| Z^{v_k,G_{k+1}}_{\mathcal{X}|\mathcal{R}}\left(x \,|\, r_{k+1}\right) - Z^{v_k}_{\mathcal{X}}\left(x\right)\right| \\
&= \ell^{A}_{D_{\mathcal{X}^n}}\left(v_k\right) + \ell^{G_{k+1}}_{Z^{v_k}_{\mathcal{X}^n}}\left(r_{k+1}\right)
\end{aligned}
$$

where (1) is simply the triangle inequality. □

Missing parts from the proof of Theorem 2.9.

$$
\begin{aligned}
&D^A_{\mathcal{V}_{k+1}}\left(V^{A,k+1}_{D_{\mathcal{X}^n}}\left(\epsilon_{[k+1]}\right)\right)\cdot\left(\ell^A_{D_{\mathcal{X}^n}}\left(V^{A,k+1}_{D_{\mathcal{X}^n}}\left(\epsilon_{[k+1]}\right)\right)-\epsilon_{[k+1]}\right)\\
&\quad=\sum_{v_{k+1}\in V^{A,k+1}_{D_{\mathcal{X}^n}}\left(\epsilon_{[k+1]}\right)} D^A_{\mathcal{V}_{k+1}}\left(v_{k+1}\right)\left(\ell^A_{D_{\mathcal{X}^n}}\left(v_{k+1}\right)-\epsilon_{[k+1]}\right)\\
&\quad\leq^{(1)}\sum_{(v_k,r_{k+1})\in V^{A,k+1}_{D_{\mathcal{X}^n}}\left(\epsilon_{[k+1]}\right)} D^A_{\mathcal{V}_k}\left(v_k\right)\cdot Z^{v_k,G_{k+1}}_{\mathcal{R}}\left(r_{k+1}\right)\left(\ell^A_{D_{\mathcal{X}^n}}\left(v_k\right)+\ell^{G_{k+1}}_{Z^{v_k}_{\mathcal{X}^n}}\left(r_{k+1}\right)-\epsilon_{[k+1]}\right)\\
&\quad=\sum_{v_k\in\mathcal{V}_k}\sum_{r_{k+1}\in R(v_k)} D^A_{\mathcal{V}_k}\left(v_k\right)\cdot Z^{v_k,G_{k+1}}_{\mathcal{R}}\left(r_{k+1}\right)\left(\ell^A_{D_{\mathcal{X}^n}}\left(v_k\right)-\epsilon_{[k]}\right)\\
&\qquad+\sum_{r_{k+1}\in\mathcal{R}}\sum_{v_k\in V_k(r_{k+1})} D^A_{\mathcal{V}_k}\left(v_k\right)\cdot Z^{v_k,G_{k+1}}_{\mathcal{R}}\left(r_{k+1}\right)\left(\ell^{G_{k+1}}_{Z^{v_k}_{\mathcal{X}^n}}\left(r_{k+1}\right)-\epsilon_{k+1}\right)\\
&\quad\leq^{(2)}\sum_{v_k\in V^{A,k}_{D_{\mathcal{X}^n}}\left(\epsilon_{[k]}\right)} D^A_{\mathcal{V}_k}\left(v_k\right)\left(\ell^A_{D_{\mathcal{X}^n}}\left(v_k\right)-\epsilon_{[k]}\right)\\
&\qquad+\sum_{r_{k+1}\in R^{G_{k+1}}_{D_{\mathcal{X}^n}}(\epsilon_{k+1})} Z^{v_k,G_{k+1}}_{\mathcal{R}}\left(r_{k+1}\right)\left(\ell^{G_{k+1}}_{Z^{v_k}_{\mathcal{X}^n}}\left(r_{k+1}\right)-\epsilon_{k+1}\right)\\
&\quad=D^A_{\mathcal{V}_k}\left(V^{A,k}_{D_{\mathcal{X}^n}}\left(\epsilon_{[k]}\right)\right)\left(\ell^A_{D_{\mathcal{X}^n}}\left(V^{A,k}_{D_{\mathcal{X}^n}}\left(\epsilon_{[k]}\right)\right)-\epsilon_{[k]}\right)\\
&\qquad+Z^{v_k,G_{k+1}}_{\mathcal{R}}\left(R^{G_{k+1}}_{D_{\mathcal{X}^n}}\left(\epsilon_{k+1}\right)\right)\left(\ell^{G_{k+1}}_{Z^{v_k}_{\mathcal{X}^n}}\left(R^{G_{k+1}}_{D_{\mathcal{X}^n}}\left(\epsilon_{k+1}\right)\right)-\epsilon_{k+1}\right)\\
&\quad\leq^{(3)}\sum_{i\in[k]}\delta_i+\delta_{k+1}\\
&\quad=\sum_{i\in[k+1]}\delta_i,
\end{aligned}
$$

where (1) is a direct result of Lemma 2.8, (2) is a result of the fact that in both sums we add positive summands and remove negative ones, and (3) results from the induction assumption. □

Lemma B.1 (Azuma inequality extended to high probability bound)**.** *Given $k\in\mathbb{N}$, $0\leq\epsilon_1,\ldots,\epsilon_k$, $0\leq\delta_1,\ldots,\delta_k\leq 1$, if $Y_0,\ldots,Y_k$ is a martingale such that for any $i\in[k]$, $\mathbb{P}\left(|Y_i-Y_{i-1}|>\epsilon_i\right)\leq\delta_i$, then for any $\lambda>0$,*

$$\mathbb{P}\left(|Y_k-Y_0|>\lambda\right)\leq\exp\left(-\frac{\lambda^2}{2\sum_{i=1}^k\epsilon_i^2}\right)+\sum_{i=1}^k\delta_i.$$

A slightly different version of this Lemma was proven for McDiarmid's inequality by [TV+15] (Proposition 34) and [Kut02] (Theorem 1.9), but the proof is identical. See also Exercise 5.11 in [DP09].

Missing parts from the proof of Theorem 2.10.

$$\begin{aligned}
\mathbb{E}\left[Y_{i+1} \mid S, r_1, \ldots, r_i\right] &= \mathbb{E}\left[\sum_{j=1}^{i+1}\left(\ell^{G_j}_{Z^{v_{j-1}}_{\mathcal{X}^n}}\left(r_j\right) - \mathop{\mathbb{E}}_{Z^{v_{j-1},G_j}_{\mathcal{R}}}\left[\ell^{G_j}_{Z^{v_{j-1}}_{\mathcal{X}^n}}\left(r\right)\right]\right) \mid S, r_1, \ldots, r_i\right] \\
&= \sum_{j=1}^{i}\left(\ell^{G_j}_{Z^{v_{j-1}}_{\mathcal{X}^n}}\left(r_j\right) - \mathop{\mathbb{E}}_{Z^{v_{j-1},G_j}_{\mathcal{R}}}\left[\ell^{G_j}_{Z^{v_{j-1}}_{\mathcal{X}^n}}\left(r\right)\right]\right) \\
&\quad + \overbrace{\mathbb{E}\left[\ell^{G_{i+1}}_{Z^{v_i}_{\mathcal{X}^n}}\left(r_{i+1}\right) - \mathop{\mathbb{E}}_{Z^{v_i,G_{i+1}}_{\mathcal{R}}}\left[\ell^{G_{i+1}}_{Z^{v_i}_{\mathcal{X}^n}}\left(r\right)\right] \mid S, r_1, \ldots, r_i\right]}^{=0} \\
&= Y_i\left(S, r_1, \ldots, r_i\right)
\end{aligned}$$

$$\begin{aligned}
\mathbb{P}\left(\ell^{A}_{D_{\mathcal{X}^n}}\left(v_k\right) > \epsilon'\right) &\leq^{(1)} \mathbb{P}\left(\sum_{i=1}^{k}\ell^{G_i}_{Z^{v_{i-1}}_{\mathcal{X}^n}}\left(r_i\right) > \sqrt{8\ln\left(\frac{1}{\delta'}\right)\sum_{i=1}^{k}\epsilon_i^2} + \sum_{i=1}^{k}\alpha_i\right) \\
&\leq^{(2)} \mathbb{P}\left(\sum_{j=1}^{k}\left(\ell^{G_j}_{Z^{v_{j-1}}_{\mathcal{X}^n}}\left(r_j\right) - \mathop{\mathbb{E}}_{Z^{v_{j-1},G_j}_{\mathcal{R}}}\left[\ell^{G_j}_{Z^{v_{j-1}}_{\mathcal{X}^n}}\left(r\right)\right]\right) > \sqrt{8\ln\left(\frac{1}{\delta'}\right)\sum_{i=1}^{k}\epsilon_i^2}\right) \\
&=^{(3)} \mathbb{P}\left(Y_k - \overbrace{Y_0}^{=0} > \sqrt{8\ln\left(\frac{1}{\delta'}\right)\sum_{i=1}^{k}\epsilon_i^2}\right) \\
&\leq^{(4)} \delta' + \sum_{i=1}^{k}\frac{\delta_i}{\epsilon_i}
\end{aligned}$$

where (1) results from Lemma 2.8, (2) from the bound on the expectation of the stability loss, (3) from the definition of Y_i, and (4) from Lemma B.1. □

C Missing Details from Section 3

Missing parts from the proof of Theorem 3.3.

$$\begin{aligned}
\mathbb{E}\left[q\left(D_{\mathcal{X}^n}\right) \mid q = M\left(S, G\right)\right] &= \sum_{S\in\mathcal{X}^n} D_{\mathcal{X}^n}\left(S\right)\cdot\sum_{q\in\mathcal{Q}_\Delta} D^{G}_{\mathcal{Q}_\Delta|\mathcal{X}^n}\left(q \mid S\right)\cdot q\left(D_{\mathcal{X}^n}\right) \\
&= \sum_{q\in\mathcal{Q}_\Delta}\overbrace{\sum_{S\in\mathcal{X}^n} D_{\mathcal{X}^n}\left(S\right)\cdot D^{G}_{\mathcal{Q}_\Delta|\mathcal{X}^n}\left(q \mid S\right)}^{=D^{G}_{\mathcal{Q}_\Delta}(q)}\cdot q\left(D_{\mathcal{X}^n}\right) \\
&= \sum_{q\in\mathcal{Q}_\Delta} D^{G}_{\mathcal{Q}_\Delta}\left(q\right)\sum_{S'\in\mathcal{X}^n} D_{\mathcal{X}^n}\left(S'\right)\cdot q\left(S'\right) \\
&= \sum_{q\in\mathcal{Q}_\Delta} D^{G}_{\mathcal{Q}_\Delta}\left(q\right)\sum_{x\in\mathcal{X}}\overbrace{\sum_{S'\in\mathcal{X}^n} D_{\mathcal{X}^n}\left(S'\right)\cdot D_{\mathcal{X}|\mathcal{X}^n}\left(x \mid S'\right)}^{D_{\mathcal{X}}(x)}\cdot q\left(x\right) \\
&= \sum_{q\in\mathcal{Q}_\Delta} D^{G}_{\mathcal{Q}_\Delta}\left(q\right)\sum_{x\in\mathcal{X}} D_{\mathcal{X}}\left(x\right)\cdot q\left(x\right)
\end{aligned}$$

$$
\begin{aligned}
\mathbb{E}\left[q\left(S\right) \mid q = M\left(S,G\right)\right] &= \sum_{S\in\mathcal{X}^n} D_{\mathcal{X}^n}\left(S\right)\cdot \sum_{q\in\mathcal{Q}_\Delta} D^G_{\mathcal{Q}_\Delta|\mathcal{X}^n}\left(q\mid S\right)\cdot q\left(S\right) \\
&= \sum_{q\in\mathcal{Q}_\Delta}\sum_{x\in\mathcal{X}} \overbrace{\sum_{S\in\mathcal{X}^n} D_{\mathcal{X}^n}\left(S\right)\cdot D^G_{\mathcal{Q}_\Delta|\mathcal{X}^n}\left(q\mid S\right)\cdot D_{\mathcal{X}|\mathcal{X}^n}\left(x\mid S\right)}^{=D^G_{(\mathcal{X},\mathcal{Q}_\Delta)}(x,q)}\cdot q\left(x\right) \\
&=^{(1)} \sum_{q\in\mathcal{Q}_\Delta} D^G_{\mathcal{Q}_\Delta}\left(q\right)\sum_{x\in\mathcal{X}} D^G_{\mathcal{X}|\mathcal{Q}_\Delta}\left(x\mid q\right)\cdot q\left(x\right),
\end{aligned}
$$

where (1) is a result of Lemma A.3. □

Proof of Claim 3.6. Since q^t is a post-processing of v^t and $\text{Adp}_{\bar{M}}$ is (ϵ,δ)-LSS with respect to A, Theorem 2.4 implies that the post-processing producing q^t is (ϵ,δ)-LSS with respect to A as well. Using Lemma 2.3 we get that $D^A_{\mathcal{Q}_\Delta}\left(Q^A_{D_{\mathcal{X}^n}}\left(2\epsilon\right)\right) < \frac{\delta}{\epsilon}$. Using the union bound and the fact that the T rounds are independent we get that $\mathbb{P}_{\bar{S},\text{Mon}_{\bar{M}}}\left(q^{t^*} \in Q^A_{D_{\mathcal{X}^n}}\left(2\epsilon\right)\right) < \frac{T\delta}{\epsilon}$. This allows us to invoke Theorem 3.3, $\frac{T\delta}{\epsilon}$ replacing δ.[14] □

Proof of Claim 3.7. This is a direct result of combining the sample accuracy definition and the union bound. If the probability that the sample accuracy of M will be greater than ϵ is bounded by δ, then the probability that it will fail to hold once in T independent iterations is less then $T\delta$, and since the values of the query are bounded on the interval $[-\Delta,\Delta]$ the maximal error in these cases is 2Δ. □

Proof of Claim 3.8. First recall that from the definition of the monitor mechanism, $\forall t \in [T], q_t\left(D_{\mathcal{X}^n}\right) - r_t \geq 0$. Therefore if M is not (ϵ,δ)-Distribution Accurate, then for any $t \in T$,

$$
\mathbb{P}_{\bar{S},\text{Mon}_{\bar{M}}}\left(q^t\left(D_{\mathcal{X}^n}\right) - r^t > \epsilon\right) > \delta.
$$

Since the T rounds of the monitor mechanism are independent and t^* is the index of the round with the maximal error,

$$
\mathbb{P}_{\bar{S},\text{Mon}_{\bar{M}}}\left(q^{t^*}\left(D_{\mathcal{X}^n}\right) - r > \epsilon\right) > 1-\left(1-\delta\right)^T.
$$

So the expectation of this quantity must be greater then $\epsilon\left(1-\left(1-\delta\right)^T\right)$, concluding the proof. □

[14] The fact that repeating this process T independent times affects only the δ and not the ϵ will be crucial to the move from generalization of expectation to generalization with high probability (at least in this proof technique). This is made possible by the way r's were treated in the distance measure in the LSS definition. For comparison, see the remark in Lemma 3.3 in [BNS+16]. We hypothesize, quite informally, that stability definitions that degrade in the ϵ term on multiple independent runs cannot yield generalization with high probability. As far as we are aware, all previously studied stability notions support this claim.

D Missing Details from Section 4

D.1 Proofs of Implications Theorems

Proof of Theorem 4.5. Given $B \subseteq \mathcal{X} \times \mathcal{R}$ we denote $R_B(x) := \{r \in \mathcal{R} \mid (x,r) \in B\}$ (which might be empty for some x's). Using this notation we prove that for any $B \subseteq \mathcal{X} \times \mathcal{R}$,

$$
\begin{aligned}
&D^G_{(\mathcal{X},\mathcal{R})}(B) \\
&\quad= \sum_{x\in\mathcal{X}} D_{\mathcal{X}}(x)\, D^G_{\mathcal{R}|\mathcal{X}}(R_B(x) \mid x) \\
&\quad=^{(1)} \sum_{x\in\mathcal{X}} D_{\mathcal{X}}(x) \sum_{S'\in\mathcal{X}^{n-1}} D_{\mathcal{X}^{n-1}}(S') \cdot D^G_{\mathcal{R}|\mathcal{X}^n}\left(R_B(x) \mid S' \cup \{x\}\right) \\
&\quad= \overbrace{\sum_{x'\in\mathcal{X}} D_{\mathcal{X}}(x')}^{=1} \sum_{x\in\mathcal{X}} D_{\mathcal{X}}(x) \sum_{S'\in\mathcal{X}^{n-1}} D_{\mathcal{X}^{n-1}}(S') \cdot D^G_{\mathcal{R}|\mathcal{X}^n}\left(R_B(x) \mid S' \cup \{x\}\right) \\
&\quad\le^{(2)} \sum_{x\in\mathcal{X}} D_{\mathcal{X}}(x) \sum_{x'\in\mathcal{X}} D_{\mathcal{X}}(x') \sum_{S'\in\mathcal{X}^{n-1}} D_{\mathcal{X}^{n-1}}(S') \left(e^{\epsilon} \cdot D^G_{\mathcal{R}|\mathcal{X}^n}\left(R_B(x) \mid S' \cup \{x'\}\right) + \delta\right) \\
&\quad=^{(1)} \sum_{x\in\mathcal{X}} D_{\mathcal{X}}(x) \sum_{S\in\mathcal{X}^n} D_{\mathcal{X}^n}(S) \left(e^{\epsilon} \cdot D^G_{\mathcal{R}|\mathcal{X}^n}(R_B(x) \mid S) + \delta\right) \\
&\quad= \sum_{x\in\mathcal{X}} D_{\mathcal{X}}(x) \left(e^{\epsilon} \cdot D^G_{\mathcal{R}}(R_B(x)) + \delta\right) \\
&\quad= e^{\epsilon} \cdot D^G_{\mathcal{X}\otimes\mathcal{R}}(B) + \delta,
\end{aligned}
$$

where (1) are a result of the fact that $D_{\mathcal{X}^n}$ is a product distribution, and (2) is a result of the DP definition. The proof is concluded by repeating the same process for the second direction. □

Proof of Theorem 4.6. Given $B \subseteq \mathcal{X} \times \mathcal{R}$, $S \in \mathcal{X}^n$, $r \in \mathcal{R}$, and $i \in [n]$, we denote the set of x's that appear with a particular r in B by $X_B(r) := \{x \in \mathcal{X} \mid (x,r) \in B\}$, the average number of elements s_i in a sample set S which appear in $X_B(r)$ by $n_B(S,r) := \sum_{x\in X_B(r)} D_{\mathcal{X}|\mathcal{X}^n}(x \mid S)$, the set of all S's with a particular $n_B(S,r)$ value by $S_B(r,i) := \{S \in \mathcal{X}^n \mid n_B(S,r) = i\}$, and the set of all r with corresponding S's with average number of elements in $X_B(r)$ equal to i by $B_n(i) := \bigcup_{r\in\mathcal{R}} \bigcup_{S\in S_B(r,i)} (S,r)$. Using these notations

we prove that for any $B \subseteq \mathcal{X} \times \mathcal{R}$,

$$
\begin{aligned}
D^G_{(\mathcal{X},\mathcal{R})}(B) &=^{(1)} \sum_{r\in\mathcal{R}} D^G_{\mathcal{R}}(r) \sum_{x\in X_B(r)} D^G_{\mathcal{X}|\mathcal{R}}(x\,|\,r) \\
&= \sum_{r\in\mathcal{R}} \sum_{S\in\mathcal{X}^n} D_{\mathcal{X}^n}(S)\cdot D^G_{\mathcal{R}|\mathcal{X}^n}(r\,|\,S) \overbrace{\sum_{x\in X_B(r)} D_{\mathcal{X}|\mathcal{X}^n}(x\,|\,S)}^{n_B(S,r)} \\
&=^{(2)} \sum_{r\in\mathcal{R}} \sum_{S\in\mathcal{X}^n} n_B(S,r)\cdot D_{\mathcal{X}^n}(S)\cdot D^G_{\mathcal{R}|\mathcal{X}^n}(r\,|\,S) \\
&=^{(3)} \sum_{r\in\mathcal{R}} \sum_{i=0}^{n} \sum_{S\in S(r,i)} \frac{i}{n} D_{\mathcal{X}^n}(S)\cdot D^G_{\mathcal{R}|\mathcal{X}^n}(r\,|\,S) \\
&=^{(4)} \sum_{i=0}^{n} \frac{i}{n} \sum_{(S,r)\in B_n(i)} D^G_{(\mathcal{X}^n,\mathcal{R})}(S,r) \\
&= \sum_{i=0}^{n} \frac{i}{n} D^G_{(\mathcal{X}^n,\mathcal{R})}(B_n(i)) \\
&\leq^{(5)} \sum_{i=0}^{n} \frac{i}{n} \left(e^{\epsilon}\cdot D^G_{\mathcal{X}^n\otimes\mathcal{R}}(B_n(i)) + \delta\right) \\
&= e^{\epsilon} \sum_{i=0}^{n} \frac{i}{n} \sum_{(S,r)\in B_n(i)} D^G_{\mathcal{X}^n\otimes\mathcal{R}}(S,r) + \sum_{i=0}^{n} \frac{i}{n}\delta \\
&=^{(4)} e^{\epsilon} \sum_{r\in\mathcal{R}} \sum_{i=0}^{n} \sum_{S\in S(r,i)} \frac{i}{n} D_{\mathcal{X}^n}(S)\cdot D^G_{\mathcal{R}}(r) + \frac{n+1}{2}\delta \\
&=^{(3)} e^{\epsilon} \sum_{r\in\mathcal{R}} \sum_{S\in\mathcal{X}^n} n_B(S,r)\cdot D_{\mathcal{X}^n}(S)\cdot D^G_{\mathcal{R}}(r) + \frac{n+1}{2}\delta \\
&=^{(2)} e^{\epsilon} \sum_{r\in\mathcal{R}} D^G_{\mathcal{R}}(r) \sum_{S\in\mathcal{X}^n} D_{\mathcal{X}^n}(S) \sum_{x\in X_B(r)} D_{\mathcal{X}|\mathcal{X}^n}(x\,|\,S) + \frac{n+1}{2}\delta \\
&= e^{\epsilon} \sum_{r\in\mathcal{R}} D^G_{\mathcal{R}}(r) \sum_{x\in X_B(r)} D_{\mathcal{X}}(x) + \frac{n+1}{2}\delta \\
&=^{(1)} e^{\epsilon}\cdot D^G_{(\mathcal{X},\mathcal{R})}(B) + \frac{n+1}{2}\delta
\end{aligned}
$$

where (1) result from the definition of $X_B(r)$, (2) from the definition of $n_B(S,r)$, (3) from the definition of $S(r,i)$ and the fact that for any given r, $\{S(r,i)\}_{i=1}^{n}$ is a partition of $\mathcal{X}^n$, (4) from the definition of $S(i)$ and (5) is a result of the MI definition. The proof is concluded by repeating the same process for the second direction. □

Proof of Theorem 4.7. Assume M is not $\left(3\epsilon, \frac{\delta}{\epsilon}\right)$-LSS, which means that in particular $D^G_{\mathcal{R}}\left(R^G_{D_{\mathcal{X}^n}}(3\epsilon)\right) > \frac{\delta}{\epsilon}$. Denoting $L^{3\epsilon}_{\mathcal{X}\times\mathcal{R}} \coloneqq \bigcup_{r\in R^G_{D_{\mathcal{X}^n}}(3\epsilon)} \left(X_+(r)\times\{r\}\right)$ we get that from the definition of the stability loss,

$$
D^G_{(\mathcal{X},\mathcal{R})}\left(L^{3\epsilon}_{\mathcal{X}\times\mathcal{R}}\right) - D^G_{\mathcal{X}\otimes\mathcal{R}}\left(L^{3\epsilon}_{\mathcal{X}\times\mathcal{R}}\right) = \sum_{r\in R^G_{D_{\mathcal{X}^n}}(3\epsilon)} D^G_{\mathcal{R}}(r)\cdot \ell^G_{D_{\mathcal{X}^n}}(r) > 3\epsilon\cdot D^G_{\mathcal{R}}\left(R^G_{D_{\mathcal{X}^n}}(3\epsilon)\right).
$$

But on the other hand, from the fact that M is (ϵ, δ)-LMI we get in contradiction that

$$\begin{aligned} D^G_{(\mathcal{X},\mathcal{R})}\left(L^{3\epsilon}_{\mathcal{X}\times\mathcal{R}}\right) - D^G_{\mathcal{X}\otimes\mathcal{R}}\left(L^{3\epsilon}_{\mathcal{X}\times\mathcal{R}}\right) &\leq D^G_{\mathcal{X}\otimes\mathcal{R}}\left(L^{3\epsilon}_{\mathcal{X}\times\mathcal{R}}\right)\cdot\left(e^{\epsilon}-1\right)+\delta \\ &\leq^{(1)} D^G_{\mathcal{R}}\left(R^G_{D_{\mathcal{X}^n}}\left(3\epsilon\right)\right)\cdot\left(e^{\epsilon}-1\right)+\epsilon\cdot D^G_{\mathcal{R}}\left(R^G_{D_{\mathcal{X}^n}}\left(3\epsilon\right)\right) \\ &<^{(2)} 3\epsilon\cdot D^G_{\mathcal{R}}\left(R^G_{D_{\mathcal{X}^n}}\left(3\epsilon\right)\right) \end{aligned}$$

where (1) results from the fact that $\epsilon\cdot D^G_{\mathcal{R}}\left(R^G_{D_{\mathcal{X}^n}}\left(3\epsilon\right)\right) > \delta$, and (2) from the fact that $2\epsilon > e^{\epsilon}-1$ and . The proof is concluded by repeating the same process for the second direction. □

Lemma D.1 (see, e.g., [SSBD14] Theorem 30.2)**.** *Given $0 \leq \delta \leq 1$, $m \leq \frac{n}{2}$, a domain $\mathcal{X}$, and a distribution $D_{\mathcal{X}}$ defined over it, we denote by $\mathcal{H}$ the family of functions (usually referred to as* hypothesis *in the context of Machine Learning) of the form $h : \mathcal{X} \to \{0,1\}$, and let $h^* \in \mathcal{H}$ be some unique hypothesis which we will think of as the* true hypothesis*. We will refer to $h^*(x)$ as the true* label *of x, and denote the labeled domain by $\mathcal{X}_{h^*} := \{(x, h^*(x)) \mid x \in \mathcal{X}\}$. Let $M : \mathcal{X}^n \times \mathcal{G} \to \mathcal{H}$ be a mechanism with a compression scheme (Definition 4.4), In this case, with probability (over the sampling of S and the internal randomness of the mechanism in case it is non deterministic) greater then $1-\delta$ we have that,*

$$\left|h_W\left(S\setminus W\right) - h_W\left(D_{\mathcal{X}}\right)\right| \leq \sqrt{h_W\left(S\setminus W\right)\frac{4m\ln 2n/\delta}{n}} + \frac{8m\ln 2n/\delta}{n}$$

where $h_W(S\setminus W)$ is the empirical mean of h_W over $S\setminus W$ and $h_W(D_{\mathcal{X}})$ is its expectation with respect to $D_{\mathcal{X}}$.

Proof of Theorem 4.8. We will prove that g is (ϵ,δ)-LSS for such an epsilon, and since LSS holds under post processing, this suffices. Notice that now $\mathcal{R} = \mathcal{X}^m$. This proof resembles that of [CLN+16].

We start by analyzing the loss of W and get that,

$$\begin{aligned} &\ell^G_{D_{\mathcal{X}^n}}\left(W\right) \\ &= \sum_{x\in X_+(W)}\left(D^G_{\mathcal{X}|\mathcal{R}}\left(x\,|\,W\right) - D_{\mathcal{X}}\left(x\right)\right) \\ &= \sum_{x\in X_+(W)}\sum_{S\in\mathcal{X}^n} D^G_{\mathcal{X}^n|\mathcal{R}}\left(S\,|\,W\right)\left(D_{\mathcal{X}|\mathcal{X}^n}\left(x\,|\,S\right) - D_{\mathcal{X}}\left(x\right)\right) \\ &= \sum_{S\in\mathcal{X}^n} D^G_{\mathcal{X}^n|\mathcal{R}}\left(S\,|\,W\right)\sum_{x\in X_+(W)}\left(\frac{m}{n}D_{\mathcal{X}|\mathcal{X}^m}\left(x\,|\,W\right) + \frac{n-m}{n}D_{\mathcal{X}|\mathcal{X}^{n-m}}\left(x\,|\,S\setminus W\right) - D_{\mathcal{X}}\left(x\right)\right) \\ &\leq \sum_{S\in\mathcal{X}^n} D^G_{\mathcal{X}^n|\mathcal{R}}\left(S\,|\,W\right)\left(\frac{m}{n} + \sum_{x\in X_+(W)}\left(D_{\mathcal{X}|\mathcal{X}^{n-m}}\left(x\,|\,S\setminus W\right) - D_{\mathcal{X}}\left(x\right)\right)\right) \\ &= \sum_{S\in\mathcal{X}^n} D^G_{\mathcal{X}^n|\mathcal{R}}\left(S\,|\,W\right)\left(\frac{m}{n} + \sum_{x\in\mathcal{X}}\left(D_{\mathcal{X}|\mathcal{X}^{n-m}}\left(x\,|\,S\setminus W\right) - D_{\mathcal{X}}\left(x\right)\right)h^+_W\left(x\right)\right) \\ &= \sum_{S\in\mathcal{X}^n} D^G_{\mathcal{X}^n|\mathcal{R}}\left(S\,|\,W\right)\left(\frac{m}{n} + h^+_W\left(S\setminus W\right) - h^+_W\left(D_{\mathcal{X}}\right)\right) \end{aligned}$$

where $h^+_W(x)$ is simply the characteristic function of $X_+(W)$.

Using this inequality we get that $\forall R \subseteq \mathcal{R}$,

$$\begin{aligned}
&D^G_{\mathcal{R}}(R)\left(\ell^G_{D_{\mathcal{X}^n}}(R)-\epsilon\right)\\
&\quad= \sum_{W\in R} D^G_{\mathcal{R}}(W)\left(\ell^G_{D_{\mathcal{X}^n}}(W)-\epsilon\right)\\
&\quad\leq^{(1)} \sum_{W\in R} D^G_{\mathcal{R}}(W)\sum_{S\in\mathcal{X}^n} D^G_{\mathcal{X}^n|\mathcal{R}}(S\,|\,W)\left(\frac{m}{n}+h^+_W(S\setminus W)-h^+_W(D_{\mathcal{X}})-\epsilon\right)\\
&\quad= \sum_{S\in\mathcal{X}^n} D_{\mathcal{X}^n}(S)\sum_{W\in R} D^G_{\mathcal{R}|\mathcal{X}^n}(W\,|\,S)\left(h^+_W(S\setminus W)-h^+_W(D_{\mathcal{X}})-\left(\epsilon-\frac{m}{n}\right)\right)\\
&\quad\leq \sum_{S\in\mathcal{X}^n} D_{\mathcal{X}^n}(S)\max_{W=g(S),h=f(W)}\left(h(S\setminus W)-h(D_{\mathcal{X}})-\left(\epsilon-\frac{m}{n}\right)\right)\\
&\quad<^{(2)} \mathbb{P}_S\left(h(S\setminus W)-h(D_{\mathcal{X}})>\left(\epsilon-\frac{m}{n}\right)\right)\\
&\quad\leq^{(3)} \sqrt{\frac{4m\ln 2n/\delta}{n}}+\frac{8m\ln 2n/\delta}{n}+\frac{m}{n}\\
&\quad\leq^{(4)} 11\sqrt{\frac{m\ln 2n/\delta}{n}}
\end{aligned}$$

where (1) results from the previous inequality, (2) from the fact that we removed S's for which the summand is negative, and replaced the positive ones with 1 - which is greater then the maximal possible value, (3) from Lemma D.1 and the fact that the value of h is bounded by 1, and (4) from the fact that $m \leq \frac{n}{9\ln\frac{2n}{\delta}}$. □

D.2 Proofs of Separation Theorems

Proof of Theorem 4.9. Given $0<\epsilon\leq 0.07$, $n>\frac{98}{\epsilon^2}\ln\left(\frac{7}{\epsilon}\right)$, $\frac{1}{2}\leq p\leq\frac{1}{2}+\frac{\epsilon}{7}$, we will define some function $f:\mathcal{X}\to\{0,1\}$, and for $i\in\{0,1\}$ denote $X_i:=\{x\in\mathcal{X}\,|\,f(x)=i\}$, set an arbitrary distribution $D_{\mathcal{X}}$ such that $D_{\mathcal{X}}(X_1)=p$, and $D_{\mathcal{X}^n}$ which is the product of $D_{\mathcal{X}}$. We will consider a mechanism M which in response to a generator G returns the parity function of the vector $(f(s_1),\ldots,f(s_n))$, where $s_1,\ldots s_n$ denotes the elements of the sample set S. Formally, $M(G,S)=|S\cap X_1| \pmod 2$, and we prove that this mechanism is $(\epsilon,0)$-LMI but not $\left(1,\frac{9}{50}\right)$-MI.

We start with denoting by p_{n-2} the probability that the parity function of a sample of size $n-2$ will be equal to 1, $\alpha:=p-\frac{1}{2}$, and the possible outputs as r_0,r_1. Notice that,

$$D^G_{\mathcal{R}|\mathcal{X}}(r_1\,|\,X_1)=p\cdot p_{n-2}+(1-p)(1-p_{n-2})$$
$$D^G_{\mathcal{R}|\mathcal{X}}(r_1\,|\,X_0)=(1-p)\,p_{n-2}+p\,(1-p_{n-2})=1-D^G_{\mathcal{R}|\mathcal{X}}(r_1\,|\,X_1)$$

$$\begin{aligned}
D^G_{\mathcal{R}}(r_1)&=p\cdot D^G_{\mathcal{R}|\mathcal{X}}(r_1\,|\,X_1)+(1-p)\,D^G_{\mathcal{R}|\mathcal{X}}(r_1\,|\,X_0)\\
&=(2p-1)\,D^G_{\mathcal{R}|\mathcal{X}}(r_1\,|\,X_1)+1-p\\
&=(1-2p)\,D^G_{\mathcal{R}|\mathcal{X}}(r_1\,|\,X_0)+p
\end{aligned}$$

Using these identities we will first prove that $\frac{D^G_{\mathcal{R}}(r_1)}{D^G_{\mathcal{R}|\mathcal{X}}(r_1\,|\,X_1)},\frac{D^G_{\mathcal{R}}(r_1)}{D^G_{\mathcal{R}|\mathcal{X}}(r_1\,|\,X_0)}\leq e^{\epsilon}$. Since a similar claim can be proven for $\frac{D^G_{\mathcal{R}|\mathcal{X}}(r_1\,|\,X_1)}{D^G_{\mathcal{R}}(r_1)},\frac{D^G_{\mathcal{R}|\mathcal{X}}(r_1\,|\,X_0)}{D^G_{\mathcal{R}}(r_1)}$, we get that this mechanism is $(\epsilon,0)$-LMI.

$$
\begin{aligned}
\frac{D^G_{\mathcal{R}}(r_1)}{D^G_{\mathcal{R}|\mathcal{X}}(r_1 \mid X_1)} &= \frac{(2p-1)\, D^G_{\mathcal{R}|\mathcal{X}}(r_1 \mid X_1) + 1 - p}{D^G_{\mathcal{R}|\mathcal{X}}(r_1 \mid X_1)} \\
&= 2p - 1 + \frac{1-p}{(2p-1)\, p_{n-2} + 1 - p} \\
&= 2p - \frac{(2p-1)\, p_{n-2}}{(2p-1)\, p_{n-2} + 1 - p} \\
&= 1 + 2\alpha - \overbrace{\frac{2\alpha p_{n-2}}{\alpha\,(2p_{n-2}-1) + \frac{1}{2}}}^{\geq 0} \\
&\leq^{(1)} 1 + \overbrace{2\alpha}^{\leq \epsilon} \\
&\leq^{(2)} e^{\epsilon}
\end{aligned}
$$

where (1) results from the fact that $0 \leq \alpha < \frac{\epsilon}{7} \leq \frac{1}{10}$, so the denominator $\alpha\,(2p_{n-2}-1) + \frac{1}{2}$ must be positive, and (2) is a result of the inequality $1 + \epsilon \leq e^{\epsilon}$ for any $\epsilon < 1$. Similarly we get that,

$$
\begin{aligned}
\frac{D^G_{\mathcal{R}}(r_1)}{D^G_{\mathcal{R}|\mathcal{X}}(r_1 \mid X_0)} &= \frac{(1-2p)\, D^G_{\mathcal{R}|\mathcal{X}}(r_1 \mid X_0) + p}{D^G_{\mathcal{R}|\mathcal{X}}(r_1 \mid X_0)} \\
&= 1 - 2p + \frac{p}{D^G_{\mathcal{R}|\mathcal{X}}(r_1 \mid X_0)} \\
&= 2 - 2p - \frac{(1-2p)\, p_{n-2}}{(1-2p)\, p_{n-2} + p} \\
&= 1 + 2\alpha + \overbrace{\frac{2\alpha \cdot p_{n-2}}{\alpha\,(1-2p_{n-2}) + \frac{1}{2}}}^{\leq 5\alpha} \\
&\leq^{(1)} 1 + \overbrace{7\alpha}^{\leq \epsilon} \\
&\leq^{(2)} e^{\epsilon}
\end{aligned}
$$

where (1) results from the fact that $0 \leq \alpha < \frac{\epsilon}{7} \leq \frac{1}{10}$, and $0 \leq p_{n-2} \leq 1$, so $\alpha\,(1-2p_{n-2}) + \frac{1}{2} \geq \frac{4}{10}$, and (2) is a result of the inequality $1 + \epsilon \leq e^{\epsilon}$ for any $\epsilon < 1$.

On the other hand, we will prove the response dramatically changes the distribution over the sample sets. We start with denoting by q_i the probability to get a sample set for which $\sum_{s \in S} f(s) = i$. Notice that q_i is a Binomial random variable, so

$$
q_{i+1} = \binom{n}{i+1} p^{i+1} (1-p)^{n-i-1} = \frac{p}{1-p} \cdot \frac{n-i}{i+1} \binom{n}{i} p^{i} (1-p)^{n-i} = \frac{p}{1-p} \cdot \frac{n-i}{i+1} q_i.
$$

Without loss of generality assume $n = 2m + 1$. Using this notation we can prove that

$$
\begin{aligned}
D_{\mathcal{R}}^{G}(r_1) &= \sum_{i=0}^{m} q_{2i+1} \\
&\leq^{(1)} \sum_{i=\lfloor m(p-\alpha)\rfloor}^{m} q_{2i+1} + \overbrace{e^{-\frac{n\epsilon^2}{98}}}^{\leq \frac{\epsilon}{7}} \\
&\leq^{(2)} \frac{p}{1-p} \sum_{i=\lfloor m(p-\alpha)\rfloor}^{m} \frac{2m+1-2i}{2i+1} q_{2i} + \frac{\epsilon}{7} \\
&\leq^{(3)} \frac{0.5+\alpha}{0.5-\alpha} \sum_{i=\lfloor m(0.5-2\alpha)\rfloor}^{m} \frac{2m+1-2i}{2i+1} q_{2i} + \frac{\epsilon}{7} \\
&\leq \frac{0.5+\alpha}{0.5-\alpha} \cdot \frac{2m+1-2\lfloor m(0.5-2\alpha)\rfloor}{2\lfloor m(0.5-2\alpha)\rfloor+1} \sum_{i=\lfloor m(0.5-2\alpha)\rfloor}^{m} q_{2i} + \frac{\epsilon}{7} \\
&\leq \left(1 + \frac{2\alpha}{0.5-\alpha}\right)\left(1 + \frac{4\lfloor 2m\alpha\rfloor}{m+1-\lfloor 2m\alpha\rfloor}\right) \sum_{i=0}^{m} q_{2i} + \frac{\epsilon}{7} \\
&\leq^{(3)} (1+5\alpha)(1+18\alpha) D_{\mathcal{R}}^{G}(r_0) + \frac{\epsilon}{7} \\
&\leq^{(3)} (1+24\alpha) D_{\mathcal{R}}^{G}(r_0) + \frac{\epsilon}{7} \\
&= (1+24\alpha)\left(1 - D_{\mathcal{R}}^{G}(r_1)\right) + \frac{\epsilon}{7} \\
&\Rightarrow D_{\mathcal{R}}^{G}(r_1) \leq \frac{1+24\alpha+\frac{\epsilon}{7}}{2+24\alpha} < \frac{1}{2} + 6\alpha + \frac{\epsilon}{7} \leq \frac{1}{2} + \epsilon
\end{aligned}
$$

where (1) is based on the additive version of the Chernoff bound, (2) on the identity proved previously and the definition of n, and (3) on the bounds on α and ϵ.

Similarly,

$$
\begin{aligned}
D_{\mathcal{R}}^{G}\left(r_0\right) &= \sum_{i=0}^{m} q_{2i} \\
&\leq \sum_{i=0}^{\lceil m(p+\alpha)\rceil} q_{2i} + \overbrace{e^{-\frac{n\epsilon^2}{98}}}^{\leq \frac{\epsilon}{7}} \\
&\leq \overbrace{\frac{1-p}{p}}^{\leq 1} \sum_{i=0}^{\lceil m(0.5+2\alpha)\rceil} \frac{2i+1}{2m+1-2i} q_{2i+1} + \frac{\epsilon}{7} \\
&\leq \frac{2\lfloor m\left(0.5-2\alpha\right)\rfloor + 1}{2m+1-2\lfloor m\left(0.5-2\alpha\right)\rfloor} \sum_{i=0}^{\lceil m(0.5+2\alpha)\rceil} q_{2i+1} + \frac{\epsilon}{7} \\
&= \left(1 + \frac{4\lfloor 2m\alpha\rfloor}{m+1-\lfloor 2m\alpha\rfloor}\right) \sum_{i=0}^{m} q_{2i+1} + \frac{\epsilon}{7} \\
&\leq \left(1+18\alpha\right) D_{\mathcal{R}}^{G}\left(r_1\right) + \frac{\epsilon}{7} \\
&= \left(1+18\alpha\right)\left(1 - D_{\mathcal{R}}^{G}\left(r_0\right)\right) + \frac{\epsilon}{7}
\end{aligned}
$$

$$
\Rightarrow D_{\mathcal{R}}^{G}\left(r_0\right) \leq \frac{1+18\alpha+\frac{\epsilon}{7}}{2+18\alpha} < \frac{1}{2} + 5\alpha + \frac{\epsilon}{7} < \frac{1}{2} + \epsilon
$$

Combining the two and denoting $\mathcal{S}_1$ the set of all sample sets with parity value 1, we get that,

$$
\begin{aligned}
D_{\mathcal{X}^n \otimes \mathcal{R}}^{G}\left(\mathcal{S}_1 \times \{r_0\}\right) &= \overbrace{D_{\mathcal{X}^n}\left(\mathcal{S}_1\right)}^{D_{\mathcal{R}}^{G}(r_1)} \cdot D_{\mathcal{R}}^{G}\left(r_0\right) \\
&> \left(\frac{1}{2} - \epsilon\right)^2 \\
&>^{(1)} \frac{9}{50} \\
&= e^1 \overbrace{D_{(\mathcal{X}^n,\mathcal{R})}^{G}\left(\mathcal{S}_1 \times \{r_0\}\right)}^{=0} + \frac{9}{50}
\end{aligned}
$$

where (1) is a result of the bounds on ϵ, which means this mechanism is not $\left(1, \frac{9}{50}\right)$-MI. □

Proof of Theorem 4.10. Given $0 \leq \delta \leq 0.1$, $n > 2\ln\left(\frac{2}{\delta}\right)$, $N > n^2$, $\mathcal{X} := [N]$, an arbitrary $D_{\mathcal{X}}$ such that $\forall x \in \mathcal{X} : D_{\mathcal{X}}\left(x\right) \leq \frac{1}{n^2}$, and $D_{\mathcal{X}^n}$ which is the product of $D_{\mathcal{X}}$, we consider a mechanism M which in response to some generator G uniformly samples one element from its sample set and outputs it.

The fact that this mechanism is $\left(11\sqrt{\frac{\ln 2n/\delta}{n}}, \delta\right)$-LSS is a direct result of Theorem 4.8 for $m = 1$. On the other hand, notice that any $r \in \mathcal{R}$ encodes one sample element which we will denote by $x\left(r\right)$. Using

this notation we will define the set $B \coloneqq \underset{r\in\mathcal{R}}{\cup}\left(x\left(r\right),r\right)$.

$$\begin{aligned}
D^{G}_{(\mathcal{X},\mathcal{R})}\left(B\right) &= \sum_{r\in\mathcal{R}} D^{G}_{\mathcal{R}}\left(r\right)\cdot D^{G}_{\mathcal{X}|\mathcal{R}}\left(x\left(r\right)\mid r\right)\\
&\geq^{(1)} \overbrace{\sum_{r\in\mathcal{R}} D^{G}_{\mathcal{R}}\left(r\right)}^{=1}\cdot\frac{1}{n}\\
&= \frac{1}{n}\\
&>^{(2)} e\frac{1}{n^2}+\frac{1}{2n}\\
&= e\overbrace{\sum_{r\in\mathcal{R}} D^{G}_{\mathcal{R}}\left(r\right)}^{=1}\cdot\frac{1}{n^2}+\frac{1}{2n}\\
&\geq e\sum_{r\in\mathcal{R}} D^{G}_{\mathcal{R}}\left(r\right)\cdot\overbrace{D_{\mathcal{X}}\left(x\left(r\right)\right)}^{\leq\frac{1}{n^2}}+\frac{1}{2n}\\
&= e^{1}\cdot D^{G}_{\mathcal{X}\otimes\mathcal{R}}\left(B\right)+\frac{1}{2n}
\end{aligned}$$

where (1) is a result of the fact that if all elements in the sample set differ from each other, with probability $\frac{1}{n}$ the sampling mechanism will return the same sample element which was encoded by r and if not then the probability is only higher, and (2) is a result of the definitions of δ and n. This proves the mechanism is not $\left(1,\frac{1}{2n}\right)$-LMI. □

E Distance Measures on Distributions

These distance measures between distributions will be used in various places in the paper.

Definition E.1 (Statistical Distance)**.** The *Statistical Distance* (also know as *Total variation Distance*) between two probability distributions D_1, D_2 over some domain $\mathcal{R}$ is defined as,

$$\begin{aligned}
\mathrm{SD}\left(D_1,D_2\right) &\coloneqq \max_{R\in\mathcal{R}}\left(D_1\left(R\right)-D_2\left(R\right)\right)\\
&= \max_{R\in\mathcal{R}}\left(D_2\left(R\right)-D_1\left(R\right)\right)\\
&= \frac{1}{2}\cdot\sum_{r\in\mathcal{R}}\left|D_1\left(r\right)-D_2\left(r\right)\right|.
\end{aligned}$$

The maximal set in the first definition is simply the set of all r's for which $D_1\left(r\right)>D_2\left(r\right)$ and for the second - the set of all r's for which $D_1\left(r\right)<D_2\left(r\right)$

Definition E.2 (δ-approximate max divergence)**.** The *δ-approximate max divergence* between two probability distributions D_1, D_2 over some domain $\mathcal{R}$ is defined as

$$\mathbf{D}^{\delta}_{\infty}\left(D_1\|D_2\right)\coloneqq\max_{R\subseteq\mathrm{Supp}(D_1)\wedge D_1(R)\geq\delta}\ln\left(\frac{D_1\left(R\right)-\delta}{D_2\left(R\right)}\right).$$

The case where $\delta=0$ is simply called the *max divergence*.

Definition E.3 (Indistinguishable distributions)**.** Two probability distributions D_1, D_2 over some domain $\mathcal{R}$ will be called (ϵ, δ)-*indistinguishable* if

$$\max\left\{\mathbf{D}_{\infty}^{\delta}\left(D_1 \| D_2\right), \mathbf{D}_{\infty}^{\delta}\left(D_2 \| D_1\right)\right\} \leq \epsilon.$$

this can also be written as the condition that for any $R \subseteq \mathcal{R}$

$$\begin{array}{rcl} D_1(R) & \leq & e^{\epsilon} \cdot D_2(R) + \delta \\ & \text{and} & \\ D_2(R) & \leq & e^{\epsilon} \cdot D_1(R) + \delta. \end{array}$$