A necessary and sufficient stability notion for adaptive generalization

Katrina Ligett
School of Computer Science & Engineering
Hebrew University of Jerusalem
Jerusalem 91904, Israel
`katrina@cs.huji.ac.il`

Moshe Shenfeld
School of Computer Science & Engineering
Hebrew University of Jerusalem
Jerusalem 91904, Israel
`moshe.shenfeld@cs.huji.ac.il`

Abstract

We introduce a new notion of the stability of computations, which holds under post-processing and adaptive composition. We 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 notion and observe that all differentially private algorithms also satisfy this notion.

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 computations (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 2.3) and adaptive composition (Theorems 2.6 and 2.7), and show that the notion is both necessary (Theorem 3.6) and sufficient (Theorem 3.3) 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, Typical Stability and Compression Schemes all imply Local Statistical Stability, which provides an alternative method to establish their generalization properties. We also provide a few separation results 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 correlation of non-product distributions present for stability have been studied in the context of *Inferential Privacy* [GK16] and *Typical Stability* [BF16].

---

[1] In particular, our lower bound (Theorem 3.6) requires one more query than our upper bound (Theorem 3.3).

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 $\mathcal{Q}, \mathcal{R}$ be arbitrary countable sets which we will refer to as *queries* and *responses*, respectively. Let a *mechanism* $M : \mathcal{X}^n \times \mathcal{Q} \to \mathcal{R}$ be a (possibly non-deterministic) function that, given a *sample set* $s \in \mathcal{X}^n$ and a query $q \in \mathcal{Q}$, returns a response $r \in \mathcal{R}$. Intuitively, queries can be thought of as questions the mechanism is asked about the sample set, usually representing functions from $\mathcal{X}^n$ to $\mathcal{R}$; the mechanism can be thought of 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 introduce the notion of adaptivity (Definition 2.4).

This setting involves two sources of randomness, the *underlying distribution* $D_{\mathcal{X}^n}$, and the *conditional distribution* $D^q_{\mathcal{R}|\mathcal{X}^n}(r \,|\, s)$—that is, the probability to get r as the output of $M(s, q)$. These in turn induce a set of distributions (formalized in Definition A.1): the *marginal distribution* over $\mathcal{R}$, the *joint distribution* (denoted $D^q_{(\mathcal{X}^n, \mathcal{R})}$) and *product distribution* (denoted $D^q_{\mathcal{X}^n \otimes \mathcal{R}}$) over $\mathcal{X}^n \times \mathcal{R}$, and the *conditional distribution* over $\mathcal{X}^n$ given $r \in \mathcal{R}$. Note that even if $D_{\mathcal{X}^n}$ is a product distribution, this conditional distribution might not be a product distribution. Although the underlying distribution $D_{\mathcal{X}^n}$ is defined over $\mathcal{X}^n$, it induces a natural probability distribution over $\mathcal{X}$ as well, by sampling one of the sample elements in the set uniformly at random.[3] This in turn allows us extend our definitions to several other distributions, which form a connection between $\mathcal{R}$ and $\mathcal{X}$ (formalized in Definition A.2): the *marginal distribution* over $\mathcal{X}$, the *joint distribution* and *product distribution* over $\mathcal{X} \times \mathcal{R}$, the *conditional distribution* over $\mathcal{R}$ given $x \in \mathcal{X}$, and the *conditional distribution* over $\mathcal{X}$ given $r \in \mathcal{R}$. We use our distribution 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.

Notational conventions We use calligraphic letters to denote domains, lower case letters to denote elements of these domains, capital letters to denote random variables taking values in these domains, and bold letters to denote subsets of these domains. We omit subscripts and superscripts from some notation when they are clear from context.

2.1 Local Statistical Stability

Before observing any output from the mechanism, an outside observer knowing D but without other information about the sample set s holds prior $D(x)$ 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 $D(x \,|\, r)$. 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 F); 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 query q, and a mechanism $M : \mathcal{X}^n \times \mathcal{Q} \to \mathcal{R}$, the *stability loss* $\ell^q_{D_{\mathcal{X}^n}}(r)$ of a response $r \in \mathcal{R}$ with respect to $D_{\mathcal{X}^n}$ and q is defined as the Statistical Distance (Definition F.1) between the prior distribution over $\mathcal{X}$ and the posterior induced by r. That is,

$$\ell^q_{D_{\mathcal{X}^n}}(r) := \sum_{x \in \mathbf{x}_+(r)} \left(D(x \,|\, r) - D(x) \right),$$

where $\mathbf{x}_+(r) := \{x \in \mathcal{X} \mid D(x \,|\, r) > D(x)\}$, the set of all sample elements which have a posterior probability (given r) higher then their prior. Similarly, we define the stability loss $\ell(\mathbf{r})$ of a set of responses $\mathbf{r} \subseteq \mathcal{R}$ as

$$\ell(\mathbf{r}) := \frac{\sum_{r \in \mathbf{r}} D(r) \cdot \ell(r)}{D(\mathbf{r})}.$$

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

---

[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]It is worth noting that in the case where $D_{\mathcal{X}^n}$ is the product distribution of some distribution $P_{\mathcal{X}}$ over $\mathcal{X}$, we get that the induced distribution over $\mathcal{X}$ is $P_{\mathcal{X}}$.

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 query q, a mechanism $M : \mathcal{X}^n \times \mathcal{Q} \to \mathcal{R}$ will be called (ϵ, δ)-*Local-Statistically Stable with respect to* $D_{\mathcal{X}^n}$ *and* q (or *LS Stable*, or *LSS*, for short) if for any $\mathbf{r} \subseteq \mathcal{R}$, $D(\mathbf{r}) \cdot (\ell(\mathbf{r}) - \epsilon) \leq \delta$.

Notice that the maximal value of the left hand side is achieved for the subset $\mathbf{r}_\epsilon$. This stability definition can be extended to apply to a family of queries and/or a family of possible distributions. When there exists a family of queries $\mathcal{Q}$ 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 $q \in \mathcal{Q}$, then M will be called (ϵ, δ)-*LSS for* $\mathcal{D}, \mathcal{Q}$. (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 Lemma B.1.

2.2 Properties

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

Post-processing guarantees (known in some contexts 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 2.3 (LSS holds under Post-Processing). *Given $0 \leq \epsilon, \delta \leq 1$, a distribution $D_{\mathcal{X}^n}$, and a query q, if a mechanism M is (ϵ, δ)-LSS with respect to $D_{\mathcal{X}^n}$ and q, then for any range $\mathcal{U}$ and any arbitrary (possibly non-deterministic) function $f : \mathcal{R} \to \mathcal{U}$, we have that $f \circ M : \mathcal{X}^n \times \mathcal{Q} \to \mathcal{U}$ is also (ϵ, δ)-LSS with respect to $D_{\mathcal{X}^n}$ and q. An analogous statement also holds for mechanisms that are LSS with respect to a family of queries and/or a family of distributions.*

The proof appears in Appendix B.1.

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

Definition 2.4 (Analyst and Adaptive Mechanism). An *analyst over a family of queries* $\mathcal{Q}$ is a (possibly non-deterministic) function $A : \mathcal{R}^* \to \mathcal{Q}$ that receives a *view*—a finite sequence of responses—and outputs a query. We denote by $\mathcal{A}$ the family of all analysts, 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 analyst as its query and which returns a view as its range type. It is parameterized by a sequence of *sub-mechanisms* $\bar{M} = (M_i)_{i=1}^k$ where $\forall i \in [k]$, $M_i : \mathcal{X}^n \times \mathcal{Q} \to \mathcal{R}$. Given a sample set s and an analyst A as input, the adaptive mechanism iterates k times through the process where A sends a query to M_i and receives its response to that query 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]

For illustration, consider a gradient descent algorithm, where at each step the algorithm requests an estimate of the gradient at a given point, and chooses the next point in which the gradient should be evaluated based on the response it receives. For us, M evaluates the gradient at a given point, and A

---

[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 queries 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." In this case, $v_k := (c, r_1, \ldots, r_{ik})$ and A receives the coin tosses as an input as well. This addition turns q_{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 of the coin tosses and from that of the sub-mechanisms.

```
Adaptive Mechanism Adp_M̄
─────────────────────────────────────────
Input: s ∈ 𝒳ⁿ, A ∈ 𝒜
Output: v_k ∈ 𝒱_k
v_0 ← ∅ or c
for i ∈ [k] :
    q_i ← A(v_{i−1})
    r_i ← M_i(s, q_i)
    v_i ← (v_{i−1}, r_i)
return v_k
```

determines the next point to be considered. The interaction between the two of them constitutes an adaptive learning process.

Definition 2.5 (k-LSS under adaptivity). Given $0 \le \epsilon, \delta \le 1$, a distribution $D_{\mathcal{X}^n}$, and an analyst 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 $\text{Adp}_{\bar{M}}$ is (ϵ, δ)-LSS with respect to $D_{\mathcal{X}^n}$ and A (in which case we will use $\mathbf{v}_\epsilon^{k, A, D_{\mathcal{X}^n}}$ to denote the set of ϵ unstable views). This definition can be extended to a family of analysts 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 $D(s \,|\, v_k)$ is itself a distribution over $\mathcal{X}^n$ and q_{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.

We first show that the stability loss of a view is bounded by the sum of losses of its responses with respect to the sub-mechanisms, which 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 2.6 (LSS adaptively composes linearly). *Given a family of distributions $\mathcal{D}$ over $\mathcal{X}^n$, a family of queries $\mathcal{Q}$, and a sequence of k mechanisms $\bar{M}$ where $\forall i \in [k]$, $M_i : \mathcal{X}^n \times \mathcal{Q} \to \mathcal{R}$, we will denote $\mathcal{D}_{M_0, \mathcal{Q}} := \mathcal{D}$, and for any $i > 0$, $\mathcal{D}_{M_i, \mathcal{Q}}$ will denote the set of all posterior distributions induced by any response of M_i with non-zero probability with respect to $\mathcal{D}_{M_{i-1}, \mathcal{Q}}$ and $\mathcal{Q}$ (see Definition B.2).*

Given a sequence $0 \le \epsilon_1, \delta_1, \ldots, \epsilon_k, \delta_k \le 1$, if for all i, M_i is (ϵ_i, δ_i)-LSS with respect to $\mathcal{D}_{M_{i-1}, \mathcal{Q}}$ and $\mathcal{Q}$, 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 analyst A over $\mathcal{Q} \times \mathcal{R}$.

The proof appears in Appendix B.3.

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

Theorem 2.7 (LSS adaptively composes sub-linearly). *Under the same conditions as Theorem 2.6, given $0 \le \alpha_1, \ldots, \alpha_k \le 1$, such that for all i and any $D_{\mathcal{X}^n} \in \mathcal{D}_{M_{i-1}, \mathcal{Q}}$, and $q \in \mathcal{Q}$,*

$$\mathbb{E}_{S \sim D_{\mathcal{X}^n}, R \sim M_i(s, q)} [\ell(R)] \le \alpha_i, \text{ then for any } 0 \le \delta' \le 1, \text{ the sequence is } \left(\epsilon', \delta' + \sum_{i \in [k]} \frac{\delta_i}{\epsilon_i} \right)\text{-}k\text{-}$$

LSS with respect to $\mathcal{D}$ and any analyst A over $\mathcal{Q} \times \mathcal{R}$, 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 provides a better bound then the previous one in case $\alpha_i \ll \epsilon_i$, in which case the dominating term is the first one, which is sub-linear in k. The proof appears in Appendix B.4.

3 LSS is Necessary and Sufficient for Generalization

Up until this point, queries 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 queries in the family of functions $q : \mathcal{X}^n \to \mathcal{R}$, and 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.[6]

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 $q(s) \coloneqq \frac{1}{n} \sum_{i=1}^{n} q_1(s_i)$ (for simplicity we will 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 query, defined as the value of the function 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 $0 \le \epsilon$, $0 \le \delta \le 1$, 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

$$\Pr_{S \sim D_{\mathcal{X}^n}, R \sim M(S,q)} [|R - q(S)| > \epsilon] \le \delta.$$

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

$$\Pr_{S \sim D_{\mathcal{X}^n}, R \sim M(S,q)} [|R - q(D_{\mathcal{X}^n})| > \epsilon] \le \delta,$$

where $q(D_{\mathcal{X}^n}) \coloneqq \mathbb{E}_{S \sim D_{\mathcal{X}^n}} [q(S)]$. 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

$$\Pr_{S \sim D_{\mathcal{X}^n}, R_i \sim M_i(S,q_i)} \left[\max_{i \in k} |R_i - q_i(S)| > \epsilon \right] \le \delta,$$

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

$$\Pr_{S \sim D_{\mathcal{X}^n}, R_i \sim M_i(S,q_i)} \left[\max_{i \in k} |R_i - q_i(D_{\mathcal{X}^n})| > \epsilon \right] \le \delta.$$

When considering an adaptive process, accuracy is defined with respect to the analyst, and the probabilities are taken also over the choice of the coin tosses by the adaptive mechanism.[7]

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

We now show that if a mechanism returns accurate results with respect to the sample set, then being LSS implies accuracy on the underlying distribution.

Theorem 3.3 (LSS implies generalization with high probability). *Given $0 \le \epsilon \le \Delta$, $0 \le \delta \le 1$, a distribution $D_{\mathcal{X}^n}$, and an analyst $A : \mathbb{V} \to \mathcal{Q}_\Delta$, 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.*

The proof of this theorem consists of two stages, and follows the method introduced by [BNS+16]. First we show that the a query returned by an LSS mechanism has expected value on the underlying distribution that is close to its value on the sample set that the mechanism received as input (Appendix C.1). We then proceed to lift this guarantee from expectation to high probability, using a thought experiment known as the *Monitor Mechanism* (Appendix C.2). Intuitively, it runs a large number of

---

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

[7] 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.

independent copies of an underlying mechanism, and exposes the results of the least-distribution-accurate copy as its output. If the expected error of even this least-accurate-copy is relatively low, then the underlying mechanism generalizes with high probability (Appendix C.3).

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

$$\tilde{q}_r(x) = \begin{cases} \Delta & D(x) > D(x \mid r) \\ -\Delta & D(x) \le D(x \mid r) \end{cases}.$$

Intuitively, this function maximizes the difference between $\underset{X \sim D_{\mathcal{X}}}{\mathbb{E}}[\tilde{q}_r(X)]$ and $\underset{X \sim D_{\mathcal{X} \mid \mathcal{R}}^{q}}{\mathbb{E}}[\tilde{q}_r(X) \mid r]$, and as a result, the potential to overfit.[8]

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

Lemma 3.5 (Loss assessment query overfits in expectation). *Given $0 \le \epsilon, \delta \le 1$, a distribution $D_{\mathcal{X}^n}$, a query q, and a mechanism M, if $D(\mathbf{r}_\epsilon) > \delta$, then there is a function $f : \mathcal{R} \to \mathcal{Q}_\Delta$ such that,*

$$\left| \underset{S \sim D_{\mathcal{X}^n}, Q' \sim f \circ M(S,q)}{\mathbb{E}} [Q'(D_{\mathcal{X}^n}) - Q'(S)] \right| > 2\epsilon\Delta\delta.$$

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

$$\left| \underset{S \sim D_{\mathcal{X}^n}, Q' \sim f \circ M(S,q)}{\mathbb{E}} [Q'(D_{\mathcal{X}^n}) - Q'(S)] \right| \overset{(1)}{=} \left| \sum_{q' \in \mathcal{Q}_\Delta} D(q') \cdot \sum_{x \in \mathcal{X}} (D(x) - D(x \mid q')) \cdot q'(x) \right|$$

$$= \left| \sum_{r \in \mathcal{R}} D(r) \cdot \sum_{x \in \mathcal{X}} (D(x) - D(x \mid r)) \cdot \tilde{q}_r(x) \right|$$

$$\overset{(2)}{\ge} \sum_{r \in \mathbf{r}_\epsilon} \overset{\ge \delta}{\overbrace{D(r)}} \cdot \overset{=2\ell(r)>2\epsilon}{\overbrace{\sum_{x \in \mathcal{X}} |D(x) - D(x \mid r)|}} \cdot \Delta$$

$$\overset{(3)}{>} 2\epsilon\Delta\delta$$

where (1) is further justified in the proof of Theorem C.1, (2) results from the definition of the loss assessment query, and (3) from the definition of $\mathbf{r}_\epsilon$. □

We use this method for constructing an overfitting query for non-LSS mechanism, to show that LSS is necessary in order for a mechanism to be both Sample Accurate and Distribution Accurate.

Theorem 3.6 (Necessity of LSS for Generalization). *Given $0 \le \epsilon \le \Delta$, $0 \le \delta \le 1$, a distribution $D_{\mathcal{X}^n}$, and an analyst $A : \mathbb{V} \to \mathcal{Q}_\Delta$, 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.*

The proof of this theorem, which appears in Appendix C.4, uses a similar method to the proof of Theorem 3.3, employing a variant of the Monitor Mechanism that outputs the loss assessment query with the highest level of overfitting.

4 Relationship to other notions of stability

In this section, we discuss the relationship between LSS and a few common notions of stability; definitions can be found in Appendix D.1. In order to do so, we introduce an additional new stability notion, which relaxes the Max Information (MI) (Definition D.2) notion by moving from the distribution over the sample sets to the distribution over the sample elements.

---

[8]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 $\mathcal{X}^n$.

Definition 4.1 (Local Max Information). Given $0 \leq \epsilon$, $0 \leq \delta \leq 1$, a distribution $D_{\mathcal{X}^n}$ and a query q, a mechanism M will be said to satisfy (ϵ, δ)-*Local-Max-Information with respect to $D_{\mathcal{X}^n}$ and q* (or *LMI*, for short), if the joint distributions $D_{(\mathcal{X}, \mathcal{R})}$ and the product distribution $D_{\mathcal{X} \otimes \mathcal{R}}$ over $\mathcal{X} \times \mathcal{R}$ are (ϵ, δ)-indistinguishable. In other words, for any $\mathbf{b} \subseteq \mathcal{X} \times \mathcal{R}$,

$$D_{(\mathcal{X}, \mathcal{R})}(\mathbf{b}) \leq e^\epsilon \cdot D_{\mathcal{X} \otimes \mathcal{R}}(\mathbf{b}) + \delta \quad \text{and} \quad D_{\mathcal{X} \otimes \mathcal{R}}(\mathbf{b}) \leq e^\epsilon \cdot D_{(\mathcal{X}, \mathcal{R})}(\mathbf{b}) + \delta.$$

The definition can be extended to apply to a family of queries and/or a family of possible distributions.

4.1 Implications

Prior work ([DFH$^+$15a] and [RRST16]) showed that bounded Differential Privacy (DP) (Definition D.1) implies bounded MI (Definition D.2). In the case of $\delta > 0$, this holds only if the underlying distribution is a product distribution [De12]). Bounded MI is also implied by Typical Stability (TS) (Definition D.3) [BF16], and Bounded Maximal Leakage (ML) [EGI19]. We prove that DP, MI and TS imply LMI (in the case of DP, only for product distributions). All proofs for this subsection can be found in Appendix D.2, where we also introduce a local version of ML and prove its relation to LMI.

Theorem 4.2 (Differential Privacy implies Local Max Information). *Given $0 \leq \epsilon$, $0 \leq \delta \leq 1$, a distribution $D_{\mathcal{X}}$, and a query q, if a mechanism M is (ϵ, δ)-DP with respect to q then it is (ϵ, δ)-LMI with respect to the same q and the product distribution over $\mathcal{X}^n$ induced by $D_{\mathcal{X}}$.*

Theorem 4.3 (Max Information implies Local Max Information). *Given $0 \leq \epsilon$, $0 \leq \delta \leq 1$, a distribution $D_{\mathcal{X}^n}$ and a query q, if a mechanism M has δ-approximate max-information of ϵ with respect to $D_{\mathcal{X}^n}$ and q then it is (ϵ, δ)-LMI with respect to the same $D_{\mathcal{X}^n}$ and q.*

Theorem 4.4 (Typical Stability implies Local Max Information). *Given $0 \leq \epsilon$, $0 \leq \delta, \eta \leq 1$, a distribution $D_{\mathcal{X}^n}$ and a query q, if a mechanism M is (ϵ, δ, η)-Typically Stable with respect to $D_{\mathcal{X}^n}$ and q then it is $(\epsilon, \delta + 2\eta)$-LMI with respect to the same $D_{\mathcal{X}^n}$ and q.*

These three theorems follow naturally from the fact that LMI is a fairly direct relaxation of DP, MI and TS.

We next show that LMI implies LSS.

Theorem 4.5 (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 query q, if a mechanism M is (ϵ, δ)-LMI with respect to $D_{\mathcal{X}^n}$ and q, then it is $\left(\epsilon', \frac{\delta}{\epsilon}\right)$-LSS with respect to the same $D_{\mathcal{X}^n}$ and q, where $\epsilon' = e^\epsilon - 1 + \epsilon$.*

We also prove that Compression Schemes (Definition D.6) 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 4.6 (Compressibility implies Local Statistical Stability). *Given $0 \leq \delta \leq 1$, an integer $m \leq \frac{n}{9 \ln(2n/\delta)}$, a distribution $D_{\mathcal{X}}$, and a query $q \in \mathcal{Q}$, if a mechanism M has a compression scheme of size m then it is (ϵ, δ)-LSS with respect to the same q and the product distribution over $\mathcal{X}^n$ induced by $D_{\mathcal{X}}$, for any $\epsilon > 11\sqrt{\frac{m \ln(2n/\delta)}{n}}$.[9]*

4.2 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.3.

Theorem 4.7 (Max Information is strictly stronger than Local Max Information). *For any $0 < \epsilon$, $n \geq 3$, the mechanism which outputs the parity function of the sample set is $(\epsilon, 0)$-LMI but not $\left(1, \frac{1}{5}\right)$-MI.*

Theorem 4.8 (Local Max Information is strictly stronger than Local Statistical Stability). *For any $0 \leq \delta \leq 1$, $n > \max\left\{2\ln\left(\frac{2}{\delta}\right), 6\right\}$, 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.*

---

[9]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.

5 Applications and Discussion

In order to make the LSS notion useful, we must identify mechanisms which manages to remain stable while maintaining sample accuracy. Fortunately, many such mechanisms have been introduced in the context of Differential Privacy. Two of the most basic Differentially Private mechanisms are based on noise addition, of either a Laplace or a Gaussian random variable. Careful tailoring of their parameters allows "masking" the effect of changing one element, while maintaining a limited effect on the sample accuracy. By Theorems 4.2 and 4.5, these mechanisms are guaranteed to be LSS as well. The definitions and properties of these mechanisms can be found in Appendix E.

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 possible way this might be achieved is by limiting the family of distributions and queries, such that the empirical mean of the query lies within some confidence interval around population mean, which would allow scaling the noise to the interval rather than the full range (see, e.g. , *Concentrated Queries*, as proposed by [BF16]).

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 analysts; see e.g. [GK16], [ZH19]).

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.

[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.

[IWK18] Ibrahim Issa, Aaron B Wagner, and Sudeep Kamath. An operational approach to information leakage. *arXiv preprint arXiv:1807.07878*, 2018.

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

[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 query q, and a mechanism $M : \mathcal{X}^n \times \mathcal{Q} \to \mathcal{R}$, together induce a set of distributions over $\mathcal{X}^n$, $\mathcal{R}$, and $\mathcal{X}^n \times \mathcal{R}$.

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

$$D^q_{\mathcal{R}|\mathcal{X}^n}(r \mid s) := \Pr_{R \sim M(s,q)}[R = r \mid s],$$

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

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

$$D^q_{(\mathcal{X}^n, \mathcal{R})}(s, r) := D_{\mathcal{X}^n}(s) \cdot D^q_{\mathcal{R}|\mathcal{X}^n}(r \mid s).$$

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

$$D^q_{\mathcal{R}}(r) := \sum_{s \in \mathcal{X}^n} D^q_{(\mathcal{X}^n, \mathcal{R})}(s, r).$$

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

$$D^q_{\mathcal{X}^n \otimes \mathcal{R}}(s, r) := D_{\mathcal{X}^n}(s) \cdot D^q_{\mathcal{R}}(r).$$

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

$$D^q_{\mathcal{X}^n|\mathcal{R}}(s \mid r) := \frac{D^q_{(\mathcal{X}^n, \mathcal{R})}(s, r)}{D^q_{\mathcal{R}}(r)}.$$

Definition A.2 (Distributions over $\mathcal{X}$ and $\mathcal{R}$). The *marginal distribution* $D_{\mathcal{X}}$ over $\mathcal{X}$ represents the probability to get x by sampling a sample set uniformly at random without any knowledge of s. That is, $\forall x \in \mathcal{X}$,

$$D_{\mathcal{X}}(x) := \sum_{s \in \mathcal{X}^n} D_{\mathcal{X}^n}(s) \cdot D_{\mathcal{X}|\mathcal{X}^n}(x \mid s),$$

where $D_{\mathcal{X}|\mathcal{X}^n}(x \mid s)$ denotes the probability to get x by sampling s uniformly at random.

The *joint distribution* $D^q_{(\mathcal{X}, \mathcal{R})}$ over $\mathcal{X} \times \mathcal{R}$ represents the probability to get x by sampling a sample set uniformly at random and also get r as the output of $M(\cdot, q)$ from the same sample set. That is, $\forall x \in \mathcal{X}, r \in \mathcal{R}$,

$$D^q_{(\mathcal{X}, \mathcal{R})}(x, r) := \sum_{s \in \mathcal{X}^n} D_{\mathcal{X}^n}(s) \cdot D_{\mathcal{X}|\mathcal{X}^n}(x \mid s) \cdot D^q_{\mathcal{R}|\mathcal{X}^n}(r \mid s).$$

where $D_{\mathcal{X}|\mathcal{X}^n}(x \mid s)$ denotes the probability to get x by sampling s uniformly at random.

The *product distribution* $D^q_{\mathcal{X} \otimes \mathcal{R}}$ over $\mathcal{X} \times \mathcal{R}$ represents the probability to get x by sampling a sample set uniformly at random and get r as the output of $M(\cdot, q)$ independently. That is, $\forall x \in \mathcal{X}, r \in \mathcal{R}$,

$$D^q_{\mathcal{X} \otimes \mathcal{R}}(x, r) := D_{\mathcal{X}}(x) \cdot D^q_{\mathcal{R}}(r).$$

The *conditional distribution* $D^q_{\mathcal{R}|\mathcal{X}}$ over $\mathcal{R}$ represents the probability to get r as the output of $M(\cdot, q)$ from a sample set, given the fact that we got x by sampling the same sample set uniformly at random. That is, $\forall x \in \mathcal{X}, r \in \mathcal{R}$,

$$D^q_{\mathcal{R}|\mathcal{X}}(r \mid x) := \sum_{s \in \mathcal{X}^n} D_{\mathcal{X}^n|\mathcal{X}}(s \mid x) \cdot D^q_{\mathcal{R}|\mathcal{X}^n}(r \mid s).$$

The *conditional distribution* $D^q_{\mathcal{X}|\mathcal{R}}$ over $\mathcal{X}$ represents the probability to get x by sampling a sample set uniformly at random, given the fact that we got r as the output of $M(\cdot, q)$ from that sample set. That is, $\forall x \in \mathcal{X}, r \in \mathcal{R}$,

$$D^q_{\mathcal{X}|\mathcal{R}}(x \mid r) := \sum_{s \in \mathcal{X}^n} D^q_{\mathcal{X}^n|\mathcal{R}}(s \mid r) \cdot D_{\mathcal{X}|\mathcal{X}^n}(x \mid s).$$

Although all of these definitions depend on $D_{\mathcal{X}^n}$ and M, we typically omit these from the notation for simplicity, and usually omit the superscripts and subscripts entirely. We include them only when necessary for clarity. We also use D to denote the probability of a set: for $\mathbf{r} \subseteq \mathcal{R}$, we define $D^q_{\mathcal{R}}(\mathbf{r}) := \sum_{r \in \mathbf{r}} D^q_{\mathcal{R}}(r)$.

Though the conditional distributions $D^q_{\mathcal{R}|\mathcal{X}}$ and $D^q_{\mathcal{X}|\mathcal{R}}$ were not defined as the ratio between the joint and marginal distribution, the analogue of Bayes' rule still holds for these distributions.

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

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

Proof. We observe

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

Similarly,

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

$\square$

B   Missing Details from Section 2

Lemma B.1. *Given $0 \le \delta \le \epsilon \le 1$, a distribution $D_{\mathcal{X}^n}$, and a query q, if a mechanism M is (ϵ, δ)-LSS with respect to $D_{\mathcal{X}^n}, q$, then $D(\mathbf{r}_{2\epsilon}) < \frac{\delta}{\epsilon}$.*

Proof. Assume by way of contradiction that $D(\mathbf{r}_{2\epsilon}) \ge \frac{\delta}{\epsilon}$; then

$$D(\mathbf{r}_{2\epsilon}) \cdot (\ell(\mathbf{r}_{2\epsilon}) - \epsilon) > \frac{\delta}{\epsilon} \cdot (2\epsilon - \epsilon) = \delta.$$

$\square$

B.1 Proof of Post-Processing Theorem

Proof of Theorem 2.3. We start by defining a function $w_{\mathcal{U}}^{\epsilon} : \mathcal{R} \to [0,1]$ such that $\forall r \in \mathcal{R} :$ $w_{\mathcal{U}}^{\epsilon}(r) = \sum_{u \in \mathbf{u}_{\epsilon}} D(u\,|\,r)$, where $\mathbf{u}_{\epsilon}$ is the set of ϵ-unstable values in $\mathcal{U}$ as defined in Definition 2.1, and $D(u\,|\,r) := \Pr_{U \sim f(r)}[U = u\,|\,r]$. Using this function we get that,

$$\sum_{u \in \mathbf{u}_{\epsilon}} D(u) = \sum_{u \in \mathbf{u}_{\epsilon}} \sum_{r \in \mathcal{R}} D(r) \cdot D(u\,|\,r)$$

$$= \sum_{r \in \mathcal{R}} \overbrace{\sum_{u \in \mathbf{u}_{\epsilon}} D(u\,|\,r)}^{=w_{\mathcal{U}}^{\epsilon}(r)} \cdot D(r)$$

$$= \sum_{r \in \mathcal{R}} w_{\mathcal{U}}^{\epsilon}(r) \cdot D(r),$$

and

$$\sum_{u \in \mathbf{u}_{\epsilon}} D(u) \cdot \ell(u) = \sum_{u \in \mathbf{u}_{\epsilon}} D(u) \sum_{x \in \mathbf{x}_{+}(u)} (D(x\,|\,u) - D(x))$$

$$= \sum_{u \in \mathbf{u}_{\epsilon}} \sum_{x \in \mathbf{x}_{+}(u)} D(x)(D(u\,|\,x) - D(u))$$

$$= \sum_{u \in \mathbf{u}_{\epsilon}} \sum_{r \in \mathcal{R}} \sum_{x \in \mathbf{x}_{+}(u)} D(x)(D(r\,|\,x) - D(r))D(u\,|\,r)$$

$$\overset{(1)}{\leq} \sum_{u \in \mathbf{u}_{\epsilon}} \sum_{r \in \mathcal{R}} \sum_{x \in \mathbf{x}_{+}(r)} D(x)(D(r\,|\,x) - D(r))D(u\,|\,r)$$

$$= \sum_{r \in \mathcal{R}} \overbrace{\sum_{u \in \mathbf{u}_{\epsilon}} D(u\,|\,r)}^{=w_{\mathcal{U}}^{\epsilon}(r)} \cdot D(r) \sum_{x \in \mathbf{x}_{+}(r)} (D(x\,|\,r) - D(x))$$

$$= \sum_{r \in \mathbf{r}} w_{\mathcal{U}}^{\epsilon}(r) \cdot D(r) \cdot \ell(r),$$

where (1) results from the definition of $\mathbf{x}_{+}(r)$.

Combining the two we get that

$$D(\mathbf{u}_{\epsilon}) \cdot (\ell(\mathbf{u}_{\epsilon}) - \epsilon) = \sum_{u \in \mathbf{u}_{\epsilon}} D(u)(\ell(u) - \epsilon)$$

$$\overset{(1)}{\leq} \sum_{r \in \mathcal{R}} w_{\mathcal{U}}^{\epsilon}(r) \cdot D(r)(\ell(r) - \epsilon)$$

$$\overset{(2)}{\leq} \sum_{r \in \mathbf{r}_{\epsilon}} \overbrace{w_{\mathcal{U}}^{\epsilon}(r)}^{\leq 1} \cdot D(r)(\ell(r) - \epsilon)$$

$$\leq \sum_{r \in \mathbf{r}_{\epsilon}} D(r)(\ell(r) - \epsilon)$$

$$\overset{(3)}{\leq} \delta,$$

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. $\square$

B.2 Adaptivity and View-Induced Posterior Distributions

Definition B.2 (View-induced posterior distributions). A sequence of mechanisms $\bar{M}$, an analyst 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 P^{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 $D^q_{\mathcal{R}|\mathcal{X}^n}$. The posteriors of these distributions change once we see v_k. $D_{\mathcal{X}^n}$ is replaced by $P^{v_k}_{\mathcal{X}^n} := D^A_{\mathcal{X}^n|\mathcal{V}_k}(\cdot \mid v_k)$ (actually, the rigorous notation should have been $P^{\bar{M},A,v_k}_{\mathcal{X}^n}$, but since $\bar{M}$ and A will be fixed throughout this analysis, we omit them for simplicity). Similarly, $D^{q_{k+1}}_{\mathcal{R}|\mathcal{X}^n}(r \mid s)$ is replaced by

$$P^{v_k}_{\mathcal{R}|\mathcal{X}^n}(r \mid s) := D^{q_{k+1}}_{\mathcal{R}|\mathcal{X}^n}(r \mid s, v_k) = \Pr_{R \sim M_{k+1}(s, q_{k+1})}\left[R = r \mid s, \mathrm{Adp}_{\bar{M},k}(s, A) = v_k\right],$$

where $\mathrm{Adp}_{\bar{M},k}$ denotes the first k iterations of the adaptive mechanism, which - as mentioned previously - determine the $k+1$-th query.[10]

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 B.3. *Given a distribution $D_{\mathcal{X}^n}$, an analyst $A : \mathcal{V} \to \mathcal{Q}$, and a sequence of k mechanisms $\bar{M}$, 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 B.2,*

$$D(v_{k+1}) = D(v_k) \cdot P^{v_k}(r_{k+1})$$

and

$$\ell^A_{D_{\mathcal{X}^n}}(v_{k+1}) \le \ell^A_{D_{\mathcal{X}^n}}(v_k) + \ell^{q_{k+1}}_{P^{v_k}_{\mathcal{X}^n}}(r_{k+1}).$$

Proof. 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.

$$
\begin{aligned}
D_{(\mathcal{X}^n, \mathcal{V}_{k+1})}(s, v_{k+1}) &= D(s) \cdot D(v_{k+1} \mid s) \\
&\overset{(1)}{=} D(s) \cdot D(v_k \mid s) \cdot D^{q_{k+1}}(r_{k+1} \mid s, v_k) \\
&= D(v_k) \cdot D(s \mid v_k) \cdot D^{q_{k+1}}(r_{k+1} \mid s, v_k) \\
&= D(v_k) \cdot P^{v_k}(s) \cdot P^{v_k}(r_{k+1} \mid s) \\
&= D(v_k) \cdot P^{v_k}_{(\mathcal{X}^n, \mathcal{R})}(s, r_{k+1}),
\end{aligned}
$$

where (1) is a result of the fact that q_{k+1} is a deterministic function of v_k. As mentioned in Definition B.2, 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 P.

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

$$D(v_{k+1}) = \sum_{s \in \mathcal{X}^n} D_{(\mathcal{X}^n, \mathcal{V}_{k+1})}(s, v_{k+1}) = \sum_{s \in \mathcal{X}^n} D(v_k) \cdot P^{v_k}_{(\mathcal{X}^n, \mathcal{R})}(s, r_{k+1}) = D(v_k) \cdot P^{v_k}(r_{k+1}).$$

$$D(x \mid v_k) = \sum_{s \in \mathcal{X}^n} D(s \mid v_k) \cdot D(x \mid s) = \sum_{s \in \mathcal{X}^n} P^{v_k}(s) \cdot D(x \mid s) = P^{v_k}(x).$$

$$D(x \mid v_{k+1}) = \sum_{s \in \mathcal{X}^n} D(s \mid v_{k+1}) \cdot D(x \mid s) = \sum_{s \in \mathcal{X}^n} P^{v_k}(s \mid r_{k+1}) \cdot D(x \mid s) = P^{v_k}(x \mid r_{k+1}).$$

---

[10] If M_{k+1} is stateful, the conditioning can result from any unknown state of M_{k+1} which might affect its response to q_{k+1}. If M_{k+1} has no shared state with the previous sub-mechanisms (either 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 q_{k+1} (which, as mentioned before, is a deterministic function of v_k for the given A), in which case $P^{v_k}_{\mathcal{R}|\mathcal{X}^n}(r \mid s) = D^{q_{k+1}}_{\mathcal{R}|\mathcal{X}^n}(r \mid s)$ where the mechanism is M_{k+1}.

where we keep using the fact that $D(x \mid s)$ does not depend on the underlying distribution $D_{\mathcal{X}^n}$ 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 F),

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

where (1) is simply the triangle inequality. $\qquad \square$

B.3 Proof of Linear Adaptive Composition Theorem

Proof of Theorem 2.6. This theorem is a direct result of combining Lemma B.3 with the triangle inequality over the posteriors created at any iteration, and the fact that 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 $\mathbf{v}^{k+1}_{\epsilon_{[k+1]}}$ on $\mathcal{V}_k$ and $\mathcal{R}$ by

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

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

where $\epsilon_{[k]} := \sum_{i \in [k]} \epsilon_i$.

Using this notation and that in Definition B.2 we get that

$$
\begin{aligned}
D\left(\mathbf{v}^{k+1}_{\epsilon_{[k+1]}}\right) \cdot &\left(\ell^A_{D_{\mathcal{X}^n}}\left(\mathbf{v}^{k+1}_{\epsilon_{[k+1]}}\right) - \epsilon_{[k+1]}\right) \\
&= \sum_{v_{k+1} \in \mathbf{v}^{k+1}_{\epsilon_{[k+1]}}} D(v_{k+1}) \left(\ell^A_{D_{\mathcal{X}^n}}(v_{k+1}) - \epsilon_{[k+1]}\right) \\
&\overset{(1)}{\leq} \sum_{(v_k, r_{k+1}) \in \mathbf{v}^{k+1}_{\epsilon_{[k+1]}}} D(v_k) \cdot P^{v_k}(r_{k+1}) \left(\ell^A_{D_{\mathcal{X}^n}}(v_k) + \ell^{q_{k+1}}_{P^{v_k}_{\mathcal{X}^n}}(r_{k+1}) - \epsilon_{[k+1]}\right)
\end{aligned}
$$

where (1) is a direct result of Lemma B.3. Analyzing the two parts separately we get

$$
\begin{aligned}
\sum_{v_k \in \mathcal{V}_k} \sum_{r_{k+1} \in \mathbf{r}(v_k)} D(v_k) \cdot P^{v_k}(r_{k+1}) \left(\ell^A_{D_{\mathcal{X}^n}}(v_k) - \epsilon_{[k]}\right) &\overset{(1)}{\leq} \sum_{v_k \in \mathbf{v}^k_{\epsilon_{[k]}}} D(v_k) \left(\ell^A_{D_{\mathcal{X}^n}}(v_k) - \epsilon_{[k]}\right) \\
&= D\left(\mathbf{v}^k_{\epsilon_{[k]}}\right) \left(\ell^A_{D_{\mathcal{X}^n}}\left(\mathbf{v}^k_{\epsilon_{[k]}}\right) - \epsilon_{[k]}\right) \\
&\overset{(2)}{\leq} \sum_{i \in [k]} \delta_i
\end{aligned}
$$

and similarly,

$$\sum_{r_{k+1} \in \mathcal{R}} \sum_{v_k \in \mathbf{v}_k(r_{k+1})} D(v_k) \cdot P^{v_k}(r_{k+1}) \left(\ell^{q_{k+1}}_{P^{v_k}_{\mathcal{X}^n}}(r_{k+1}) - \epsilon_{k+1} \right)$$

$$\overset{(1)}{\leq} \sum_{r_{k+1} \in \mathbf{r}^{q_{k+1}}_{\epsilon_{k+1}}} P^{v_k}(r_{k+1}) \left(\ell^{q_{k+1}}_{P^{v_k}_{\mathcal{X}^n}}(r_{k+1}) - \epsilon_{k+1} \right)$$

$$= P^{v_k}\left(\mathbf{r}^{q_{k+1}}_{\epsilon_{k+1}} \right) \left(\ell^{q_{k+1}}_{P^{v_k}_{\mathcal{X}^n}}\left(\mathbf{r}^{q_{k+1}}_{\epsilon_{k+1}} \right) - \epsilon_{k+1} \right)$$

$$\overset{(2)}{\leq} \delta_{k+1},$$

where in both cases (1) is a result of the fact that in both sums we add positive summands and remove negative ones, and (2) results from the inductive assumption.

Combining the two we get that $D\left(\mathbf{v}^{k+1}_{\epsilon_{[k+1]}} \right) \cdot \left(\ell^A_{D_{\mathcal{X}^n}}\left(\mathbf{v}^{k+1}_{\epsilon_{[k+1]}} \right) - \epsilon_{[k+1]} \right) \leq \sum_{i \in [k+1]} \delta_i.$ $\qquad \square$

B.4 Proof of Sub-Linear Adaptive Composition Theorem

Lemma B.4 (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 with respect to another sequence $Z_0, \ldots, Z_k$ such that for any $i \in [k]$, $Pr\left[|Y_i - Y_{i-1}| > \epsilon_i \right] \leq \delta_i$, then for any $\lambda > 0$,*

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

The proof parallels that of a similar lemma by [TV+15] (their Proposition 34).

Proof. For any given realization of the random variable $y = (y_0, \ldots, y_k)$, we denote by $I(y)$ the first index i for which $|y_i - y_{i-1}| > \epsilon_i$. If no such index exists, $I(y) = k+1$. We then define $y' := f(y)$ where $\forall i < I(\bar{y}) : y'_i = y_i$ and $\forall i \geq I(\bar{y}) : y'_i = y_{I(\bar{y})-1}$. Notice that the random variable Y' is also a martingale with respect to Z_0, $\Pr\left[|Y'_i - Y'_{i-1}| > \epsilon_i \right] = 0$, and

$$\Pr\left[Y' \neq Y \right] \leq \sum_{i=1}^k \Pr\left[Y'_i \neq Y_i \right] \leq \sum_{i=1}^k \delta_i.$$

Using these facts we get

$$\Pr\left[|Y_k - Y_0| > \lambda \right] = \overbrace{\Pr\left[Y' = Y \right]}^{\leq 1} \cdot \Pr\left[|Y'_k - Y'_0| > \lambda \right] + \Pr\left[Y' \neq Y \right] \cdot \overbrace{\Pr\left[|Y_k - Y_0| > \lambda \right]}^{\leq 1}$$

$$\overset{(1)}{\leq} \exp\left(-\frac{\lambda^2}{2 \sum_{i=1}^k \epsilon_i^2} \right) + \sum_{i=1}^k \delta_i.$$

where (1) results from the previous inequality and Azuma's inequality for Y'. $\qquad \square$

Proof of Theorem 2.7. 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.4.

Formally, for any given $k > 0$, we can define $\Omega_0 := \mathcal{X}^n$ and $\forall i \in [k], \Omega_i := \mathcal{R}.$[11] 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

---

[11] If the analyst A is non-deterministic, $\Omega_0 := \mathcal{X}^n \times C$, where C is the set of all possible coin tosses of the analyst, as mentioned in Definition 2.4. 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 B.2.

given $\Omega_1, \ldots, \Omega_{i-1}$ as $P^{v_{i-1}}$ (see Definition B.2). We then define a sequence of functions, $y_0 = 0$ and $\forall i > 0$,

$$y_i(s, r_1, \ldots, r_i) = \sum_{j=1}^{i} \left(\ell_{P^{v_{j-1}}}(r_j) - \mathbb{E}_{R \sim P_{\mathcal{R}}^{v_{j-1}}} \left[\ell_{P^{v_{j-1}}}(R) \right] \right).$$

Intuitively y_i is the sum of the first i losses, with a correction term which zeroes the expectation.

Notice that these random variables are a martingale with respect to the random process $S, R_1, \ldots, R_k$ since

$$\mathbb{E}_{R_{i+1}} [Y_{i+1} \mid S, R_1, \ldots, R_i] = \mathbb{E}_{R_{i+1}} \left[\sum_{j=1}^{i+1} \left(\ell_{P^{V_{j-1}}}(R_j) - \mathbb{E}_{R \sim P_{\mathcal{R}}^{V_{j-1}}} \left[\ell_{P^{V_{j-1}}}(R) \right] \right) \mid S, R_1, \ldots, R_i \right]$$

$$= \overbrace{\sum_{j=1}^{i} \left(\ell_{P^{V_{j-1}}}(R_j) - \mathbb{E}_{R \sim P_{\mathcal{R}}^{V_{j-1}}} \left[\ell_{P^{V_{j-1}}}(R) \right] \right)}^{=Y_i(S,R_1,\ldots,R_i)}$$

$$+ \overbrace{\mathbb{E}_{R_{i+1}} \left[\ell_{P^{V_i}}(R_{i+1}) - \mathbb{E}_{R \sim P_{\mathcal{R}}^{V_i}} \left[\ell_{P^{V_i}}(R) \right] \mid S, R_1, \ldots, R_i \right]}^{=0}$$

$$= Y_i(S, R_1, \ldots, R_i)$$

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.

From the LSS definition (Definition 2.2) and Lemma B.1, for any $i \in [k]$ we get that $\Pr_{R \sim P_{\mathcal{R}}^{v_{i-1}}} \left[\ell_{P^{v_i}}(R_i) > 2\epsilon_i \right] \leq \frac{\delta_i}{\epsilon_i}$, so with probability greater than $\frac{\delta_{i+1}}{\epsilon_{i+1}}$,

$$|Y_{i+1} - Y_i| = \left| \ell_{P^{V_i}}(R_{i+1}) - \mathbb{E}_{R \sim P^{V_i}} \left[\ell_{P^{V_i}}(R) \right] \right| \leq \ell_{P^{V_i}}(R_{i+1}) \leq 2\epsilon_{i+1}.$$

Using this fact we can invoke Lemma B.4 and get that for any $0 \leq \delta' \leq 1$,

$$\Pr_{V \sim D_{\mathcal{V}_k}} [\ell_{D_{\mathcal{X}^n}}(V) > \epsilon']$$

$$\overset{(1)}{\leq} \Pr_{V \sim D_{\mathcal{V}_k}} \left[\sum_{i=1}^{k} \ell_{P^{V_{i-1}}}(R_i) > \sqrt{8\ln\left(\frac{1}{\delta'}\right) \sum_{i=1}^{k} \epsilon_i^2} + \sum_{i=1}^{k} \alpha_i \right]$$

$$\overset{(2)}{\leq} \Pr_{V \sim D_{\mathcal{V}_k}} \left[\sum_{j=1}^{k} \left(\ell_{P^{V_{j-1}}}(R_j) - \mathbb{E}_{R \sim P_{\mathcal{R}}^{V_{j-1}}} \left[\ell_{P^{V_{j-1}}}(R) \right] \right) > \sqrt{8\ln\left(\frac{1}{\delta'}\right) \sum_{i=1}^{k} \epsilon_i^2} \right]$$

$$\overset{(3)}{=} \Pr_{V \sim D_{\mathcal{V}_k}} \left[Y_k - \overbrace{Y_0}^{=0} > \sqrt{8\ln\left(\frac{1}{\delta'}\right) \sum_{i=1}^{k} \epsilon_i^2} \right]$$

$$\overset{(4)}{\leq} \delta' + \sum_{i=1}^{k} \frac{\delta_i}{\epsilon_i}$$

where (1) results from Lemma B.3, (2) from the bound on the expectation of the stability loss, (3) from the definition of Y_i, and (4) from Lemma B.4. $\qquad \square$

C Missing Details from Section 3

C.1 Generalization of Expectation

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.

Theorem C.1 (Generalization of expectation). *Given $0 \le \epsilon, \delta \le 1$, a distribution $D_{\mathcal{X}^n}$, a query q, and a mechanism $M : \mathcal{X}^n \times \mathcal{Q} \to \mathcal{Q}_\Delta$, if $D\left(\mathbf{q}_\epsilon\right) < \delta$, then*

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

Proof. First notice that,

$$q\left(s\right) = \frac{1}{n} \sum_{i=1}^{n} q\left(s_i\right) = \sum_{x \in \mathcal{X}} D\left(x \mid s\right) \cdot q\left(x\right)$$

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.

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

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

where (1) is a result of Lemma A.3.

Now we can calculate the difference:

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

where (1) results from the definition of $\mathcal{Q}_\Delta$ and the triangle inequality, and (2) from the condition that $D\left(\mathbf{q}_\epsilon\right) < \delta$. $\qquad\square$

Corollary C.2. *Given* $0 \leq \epsilon, \delta \leq 1$, *a distribution* $D_{\mathcal{X}^n}$, *and a query* q, *if a mechanism* $M : \mathcal{X}^n \times \mathcal{Q} \to \mathcal{Q}_\Delta$ *is* (ϵ, δ)-*LSS with respect to* $D_{\mathcal{X}^n}, q$, *then*

$$\left| \underset{S \sim D_{\mathcal{X}^n}, Q' \sim M(S,q)}{\mathbb{E}} \left[Q'\left(D_{\mathcal{X}^n}\right) - Q'\left(S\right)\right] \right| < 2\Delta \left(2\epsilon + \frac{\delta}{\epsilon}\right).$$

Proof. This is a direct result of combining Theorem C.1 with Lemma B.1. $\qquad\square$

C.2 Monitoring Argument

Definition C.3 (The Monitor Mechanism). The *Monitor Mechanism* is a function $\mathrm{Mon}_{\bar{M}} : \left(\mathcal{X}^n\right)^t \times \mathcal{A} \to \mathcal{Q} \times \mathbb{R} \times [t]$ 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 $\bar{s} \in \left(\mathcal{X}^n\right)^t$ and analyst 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 $i \in t$, based on the following process:

Monitor Mechanism $\mathrm{Mon}_{\bar{M}}$
<hr>

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

---
^{a}We slightly abuse notation since q is not part of v^i, 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 Claim C.6.

Notice that the monitor mechanism makes use of the ability to evaluate queries according to the true underlying distribution.[12]

We begin by proving that the monitor mechanism has generalization of expectation. In this claim and the following ones, the probabilities and expectations are taken over the randomness of the choice of $\bar{s} \in \left(\mathcal{X}^n\right)^t$ (which is assumed to be drawn iid from $D_{\mathcal{X}^n}$) and the internal probability of $\mathrm{Adp}_{\bar{M}}$.

Claim C.4. *Given* $0 \leq \epsilon, \delta \leq 1$, $t \in \mathbb{N}$, *a distribution* $D_{\mathcal{X}^n}$, *and an analyst* $A : \mathbb{V} \to \mathcal{Q}_\Delta$, *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*

$$\left| \underset{\bar{S} \sim D_{\mathcal{X}^n}^t, (Q,R,I) \sim Mon_{\bar{M}}(\bar{S},A)}{\mathbb{E}} \left[q\left(D_{\mathcal{X}^n}\right) - Q\left(S_I\right)\right] \right| < 2\Delta \left(2\epsilon + \frac{t\delta}{\epsilon}\right).$$

Proof. Since q^i is a post-processing of v^i and $\mathrm{Adp}_{\bar{M}}$ is (ϵ, δ)-LSS with respect to A, Theorem 2.3 implies that the post-processing producing q^i is (ϵ, δ)-LSS with respect to A as well. Using Lemma

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

B.1 we get that $D\left(\mathbf{q}_{2\epsilon}\right) < \frac{\delta}{\epsilon}$ for each of the t rounds. Using the union bound and the fact that the t rounds are independent we get that $\Pr_{\bar{S}\sim D_{\mathcal{X}^n}^t,(Q,R,I)\sim\text{Mon}_{\bar{M}}\left(\bar{S},A\right)}\left[q \in \mathbf{q}_{2\epsilon}\right] < \frac{t\delta}{\epsilon}$. This allows us to invoke Theorem C.1, with $\frac{t\delta}{\epsilon}$ replacing δ.[13] $\hfill\square$

We next show that k-sample accuracy of the mechanism run inside the monitor mechanism has implications for the sample accuracy of the result of the monitor mechanism.

Claim C.5. *Given $0 \leq \epsilon, \delta \leq 1$, $t \in \mathbb{N}$, a distribution $D_{\mathcal{X}^n}$, and an analyst $A : \mathbb{V} \to \mathcal{Q}_\Delta$, 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*

$$\mathbb{E}_{\bar{S}\sim D_{\mathcal{X}^n}^t,(Q,R,I)\sim Mon_{\bar{M}}\left(\bar{S},A\right)}\left[Q\left(S_I\right) - R\right] \leq \epsilon + 2t\delta\Delta.$$

Proof. 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. $\hfill\square$

If the mechanism run by the monitor mechanism is not k-Distribution Accurate, this has implications for the distribution accuracy of the result of the monitor mechanism, as well.

Claim C.6. *Given $0 \leq \epsilon, \delta \leq 1$, $t \in \mathbb{N}$, a distribution $D_{\mathcal{X}^n}$, and an analyst $A : \mathbb{V} \to \mathcal{Q}_\Delta$, 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*

$$\mathbb{E}_{\bar{S}\sim D_{\mathcal{X}^n}^t,(Q,R,I)\sim Mon_{\bar{M}}\left(\bar{S},A\right)}\left[Q\left(D_{\mathcal{X}^n}\right) - R\right] > \epsilon\left(1 - (1-\delta)^t\right).$$

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

$$\Pr_{S\sim D_{\mathcal{X}^n}^t,V\sim M_i(S,A),(Q,R)=\arg\max\limits_{(q,r)\in V}|q(D_{\mathcal{X}^n})-r|}\left[Q\left(D_{\mathcal{X}^n}\right) - R > \epsilon\right] > \delta.$$

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

$$\Pr_{\bar{S}\sim D_{\mathcal{X}^n}^t,(Q,R,I)\sim\text{Mon}_{\bar{M}}\left(\bar{S},A\right)}\left[Q\left(D_{\mathcal{X}^n}\right) - R > \epsilon\right] > 1 - (1-\delta)^t.$$

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

Finally, we use the monitor mechanism as a tool to show that LSS implies generalization with high probability.

C.3 Proof of Generalization in Probability Theorem

Proof of Theorem 3.3. 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.

---

[13]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.

Setting $t = \lfloor \frac{1}{\delta} \rfloor$, we see

$$\left| \underset{\bar{S} \sim D_{\mathcal{X}^n}^t, (Q,R,I) \sim \text{Mon}_{\bar{M}}(\bar{S},A)}{\mathbb{E}} [Q(D_{\mathcal{X}^n}) - Q(S_I)] \right| \overset{(1)}{<} 2\Delta \left(2\frac{a\epsilon}{\Delta} + \frac{t\Delta}{a\epsilon} \cdot \frac{ab\epsilon^2 \delta}{\Delta^2} \right)$$

$$\overset{(2)}{\leq} (4a + 2b)\epsilon,$$

where (1) is a direct result of Claim C.4 and (2) uses the definition of t.

But on the other hand,

$$\left| \underset{\bar{S} \sim D_{\mathcal{X}^n}^t, (Q,R,I) \sim \text{Mon}_{\bar{M}}(\bar{S},A)}{\mathbb{E}} [Q(D_{\mathcal{X}^n}) - Q(S_I)] \right| \overset{(1)}{\geq} \left| \underset{\bar{S} \sim D_{\mathcal{X}^n}^t, (Q,R,I) \sim \text{Mon}_{\bar{M}}(\bar{S},A)}{\mathbb{E}} [Q(D_{\mathcal{X}^n}) - R] \right|$$

$$- \left| \underset{\bar{S} \sim D_{\mathcal{X}^n}^t, (Q,R,I) \sim \text{Mon}_{\bar{M}}(\bar{S},A)}{\mathbb{E}} [Q(S_I) - R] \right|$$

$$\overset{(2)}{>} \epsilon \left(1 - (1-\delta)^t \right) - \left(c\epsilon + 2t \cdot \frac{d\epsilon\delta}{\Delta}\Delta \right)$$

$$\overset{(3)}{>} \epsilon \left(1 - e^{-\delta \lfloor \frac{1}{\delta} \rfloor} \right) - (c + 2d)\epsilon$$

$$\overset{(4)}{\geq} \epsilon \left(1 - e^{-1} \right) - (c + 2d)\epsilon$$

$$\overset{(5)}{>} (4a + 2b)\epsilon,$$

where (1) is the triangle inequality, (2) uses Claims C.5 and C.6, (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 $a = c = \frac{1}{8}, b = d = \frac{1}{600}$. $\qquad\square$

C.4 Proof of the Necessity of LSS to Generalization Theorem

Definition C.7 (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 analyst 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 $i \in t$, 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: $q \in \mathcal{Q}, r \in \mathbb{R}, i \in t$
for $i = 1, ..., t$:
$\quad v^i \leftarrow \text{Adp}_{\bar{M}}(s_i, A)$
$\quad q^i \leftarrow \tilde{q}_{v^i}$
$\quad r^i \leftarrow M(s, q^i)$
$i^* \leftarrow \underset{i \in [t]}{\arg\max} \left(\ell_{D_{\mathcal{X}^n}}^A (v^i) \right)$
return $\left(q^{i^*}, r^{i^*}, i^* \right)$

Proof of Theorem 3.6. 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 $D\left(\mathbf{v}_{\left(\frac{\epsilon}{\Delta} \right)}^k \right) \geq \delta$.

Since the t rounds of the second monitor mechanism are independent and i^* is the index of the round with the maximal stability loss of the calculated query, we get that

$$\underset{\bar{S} \sim D_{\mathcal{X}^n}^t, (Q,R,I) \sim \text{Mon2}_{\bar{M}}(\bar{S},A)}{\Pr} \left[v^I \in \mathbf{v}_{\left(\frac{\epsilon}{\Delta} \right)}^k \right] > 1 - (1-\delta)^t.$$

Combining this fact with Lemma 3.5, and setting $t = \lfloor \frac{1}{\delta} \rfloor$ we get on one hand,

$$\left| \mathbb{E}_{\bar{S} \sim D_{\mathcal{X}^n}^t, (Q,R,I) \sim \mathrm{Mon2}_{\bar{M}}(\bar{S}, A)} [Q(D_{\mathcal{X}^n}) - Q(S_I)] \right| \overset{(1)}{\geq} 2\frac{\epsilon}{\Delta}\Delta\left(1 - (1-\delta)^t\right)$$

$$\overset{(2)}{>} 2\epsilon\left(1 - e^{-\delta\lfloor \frac{1}{\delta} \rfloor}\right)$$

$$\overset{(3)}{>} 2\epsilon\left(1 - e^{-1}\right),$$

where (1) is a direct result of invoking Lemma 3.5 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,

$$\left| \mathbb{E}_{\bar{S} \sim D_{\mathcal{X}^n}^t, (Q,R,I) \sim \mathrm{Mon2}_{\bar{M}}(\bar{S}, A)} [Q(D_{\mathcal{X}^n}) - Q(S_I)] \right| \overset{(1)}{\leq} \left| \mathbb{E}_{\bar{S} \sim D_{\mathcal{X}^n}^t, (Q,R,I) \sim \mathrm{Mon2}_{\bar{M}}(\bar{S}, A)} [Q(D_{\mathcal{X}^n}) - R] \right|$$

$$+ \left| \mathbb{E}_{\bar{S} \sim D_{\mathcal{X}^n}^t, (Q,R,I) \sim \mathrm{Mon2}_{\bar{M}}(\bar{S}, A)} [Q(S_I) - R] \right|$$

$$\overset{(2)}{<} \left(a\epsilon + 2t \cdot \frac{b\epsilon\delta}{\Delta}\Delta\right) + \left(c\epsilon + 2t \cdot \frac{d\epsilon\delta}{\Delta}\Delta\right)$$

$$\overset{(3)}{\leq} (a + 2b + c + 2d)\epsilon$$

$$\overset{(4)}{<} 2\epsilon\left(1 - e^{-1}\right),$$

where (1) is the triangle inequality, (2) uses Claim C.5 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 $a = b = c = d = \frac{1}{5}$. $\qquad\square$

D   Missing Details from Section 4

D.1   Definitions

In the following definitions, $\mathcal{X}, D_{\mathcal{X}}, \mathcal{Q}, \mathcal{R}, M, \epsilon, \delta$ and n are used in a similar manner as for the definitions leading to LSS.

Definition D.1 (Differential Privacy [DMNS06]). Given $0 \leq \epsilon$, $0 \leq \delta \leq 1$, and a query q, a mechanism $M : \mathcal{X}^n \times \mathcal{Q} \to \mathcal{R}$ will be called (ϵ, δ)-*differentially-private with respect to q* (or *DP*, for short) if for any $s_1, s_1 \in \mathcal{X}^n$ that differ only in one element, the two distributions defined over $\mathcal{R}$ by $M(s_1, q)$ and $M(s_2, q)$ are (ϵ, δ)-indistinguishable (Definition F.3). In other words, for any $\mathbf{r} \subseteq \mathcal{R}$,

$$D(\mathbf{r} \mid s_1) \leq e^\epsilon \cdot D(\mathbf{r} \mid s_2) + \delta,$$

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 D.2 (Max Information [DFH+15a]). Given $0 \leq \epsilon$, $0 \leq \delta \leq 1$, a distribution $D_{\mathcal{X}^n}$, and a query q, we say a mechanism $M : \mathcal{X}^n \times \mathcal{Q} \to \mathcal{R}$ has *δ-approximate max-information of ϵ with respect to $D_{\mathcal{X}^n}, q$* (or *MI*, for short) if the two distributions $D_{(\mathcal{X}^n, \mathcal{R})}$ and $D_{\mathcal{X}^n \otimes \mathcal{R}}$ over $\mathcal{X}^n \times \mathcal{R}$ are (ϵ, δ)-indistinguishable. In other words, for any $\mathbf{b} \subseteq \mathcal{X}^n \times \mathcal{R}$,

$$D_{(\mathcal{X}^n, \mathcal{R})}(\mathbf{b}) \leq e^\epsilon \cdot D_{\mathcal{X}^n \otimes \mathcal{R}}(\mathbf{b}) + \delta \quad \text{and} \quad D_{\mathcal{X}^n \otimes \mathcal{R}}(\mathbf{b}) \leq e^\epsilon \cdot D_{(\mathcal{X}^n, \mathcal{R})}(\mathbf{b}) + \delta.$$

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

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

Definition D.3 (Typical Stability, based on Definition 2.3. of [BF16]). Given $0 \leq \epsilon, 0 \leq \delta, \eta \leq 1$, a distribution $D_{\mathcal{X}^n}$, and a query q, a mechanism $M : \mathcal{X}^n \times \mathcal{Q} \to \mathcal{R}$ will be called (ϵ, δ, η)-*Typically-Stable with respect to* $D_{\mathcal{X}^n}, q$ (or *TS*, for short) if with probability at least $1 - \eta$ over the sampling of $s_1, s_2 \in \mathcal{X}^n$, the conditional distributions induced by the mechanism given the two sets is (ϵ, δ)-indistinguishable. Formally,

$$\Pr_{S_1, S_2 \sim D_{\mathcal{X}^n}} \left[\exists \mathbf{r} \subseteq \mathcal{R} \mid D\left(\mathbf{r} \mid S_1\right) > e^\epsilon D\left(\mathbf{r} \mid S_2\right) + \delta \right] < \eta$$

An equivalent definition requires the existence of a subset $\mathbf{s} \in \mathcal{X}^n$, such that (1) $D\left(\mathbf{s}\right) \geq 1 - \eta$, and (2) for any $s_1, s_2 \in \mathbf{s}$

$$D\left(\mathbf{r} \mid s_1\right) \leq e^\epsilon \cdot D\left(\mathbf{r} \mid s_2\right) + \delta$$

Notice that in a way, MI and TS are a natural relaxation of DP, where instead of considering only the probability which is induced by the mechanism, we also consider the underlying distribution.

Definition D.4 (Bounded Maximal Leakage [EGI19]). Given $0 \leq \epsilon$, a distribution $D_{\mathcal{X}^n}$, and a query q, a mechanism $M : \mathcal{X}^n \times \mathcal{Q} \to \mathcal{R}$ will be called ϵ-*Bounded-Maximal-Leaking with respect to* $D_{\mathcal{X}^n}, q$ (or *ML*, for short) if $\mathcal{L}\left(D_{\mathcal{X}^n} \to D_{\mathcal{R}}\right) \leq \epsilon$, where $\mathcal{L}$ is the Maximal Leakage (Definition F.4).

Similarly to MI, this definition can also be relaxed to the local version.

Definition D.5 (Bounded Local Maximal Leakage). Given $0 \leq \epsilon$, a distribution $D_{\mathcal{X}^n}$, and a query q, a mechanism $M : \mathcal{X}^n \times \mathcal{Q} \to \mathcal{R}$ will be called ϵ-*Bounded-Local-Maximal-Leaking with respect to* $D_{\mathcal{X}^n}, q$ (or *ML*, for short) if $\mathcal{L}\left(D_{\mathcal{X}} \to D_{\mathcal{R}}\right) \leq \epsilon$, where $\mathcal{L}$ is the Maximal Leakage (Definition F.4).

In Theorem D.7 we prove that LML implies LMI, the same way ML implies MI.

Definition D.6 (Compression Scheme [LW86]). Given an integer $m < \frac{n}{2}$ and a query q, a mechanism M will be said to have a *compression scheme* of size m with respect to q (or *CS*, for short), if M can be described as the composition $f_q \circ g_q$ where the *compression function* $g_q : \mathcal{X}^n \to \mathcal{X}^m$ has the property that $g_q\left(s\right) \subset s$ and $f_q : \mathcal{X}^m \to \mathcal{R}$ is some arbitrary function which will be called the *encoding function*. Both functions might be non deterministic. We will denote $w := g\left(s\right)$ and $r_w := f\left(w\right)$.[14]

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

D.2 Proofs of Implication Theorems

Proof of Theorem 4.2. Given $\mathbf{b} \subseteq \mathcal{X} \times \mathcal{R}$ we denote $\mathbf{r_b}\left(x\right) := \{r \in \mathcal{R} \mid (x, r) \in \mathbf{b}\}$ (which might be empty for some x's). Using this notation we prove that for any $\mathbf{b} \subseteq \mathcal{X} \times \mathcal{R}$,

$$
\begin{aligned}
D_{(\mathcal{X}, \mathcal{R})}\left(\mathbf{b}\right) &= \sum_{x \in \mathcal{X}} D\left(x\right) D\left(\mathbf{r_b}\left(x\right) \mid x\right) \\
&\overset{(1)}{=} \overbrace{\sum_{x' \in \mathcal{X}} D\left(x'\right)}^{=1} \sum_{x \in \mathcal{X}} D\left(x\right) \sum_{s' \in \mathcal{X}^{n-1}} D\left(s'\right) \cdot D\left(\mathbf{r_b}\left(x\right) \mid s' \cup \{x\}\right) \\
&\overset{(2)}{\leq} \sum_{x \in \mathcal{X}} D\left(x\right) \sum_{x' \in \mathcal{X}} D\left(x'\right) \sum_{s' \in \mathcal{X}^{n-1}} D\left(s'\right) \left(e^\epsilon \cdot D\left(\mathbf{r_b}\left(x\right) \mid s' \cup \{x'\}\right) + \delta\right) \\
&\overset{(1)}{=} \sum_{x \in \mathcal{X}} D\left(x\right) \sum_{s \in \mathcal{X}^n} D\left(s\right) \left(e^\epsilon \cdot D\left(\mathbf{r_b}\left(x\right) \mid s\right) + \delta\right) \\
&= \sum_{x \in \mathcal{X}} D\left(x\right) \left(e^\epsilon \cdot D\left(\mathbf{r_b}\left(x\right)\right) + \delta\right) \\
&= e^\epsilon \cdot D_{\mathcal{X} \otimes \mathcal{R}}\left(\mathbf{b}\right) + \delta,
\end{aligned}
$$

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. $\qquad \square$

---

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

Proof of Theorem 4.3. Notice that the proof of DP holding under post-processing (see e.g. [DR$^+$14]), proves in fact that (ϵ, δ)-indistinguishability is closed under post-processing. Since x is a post-processing of s, the fact that $D_{(\mathcal{X}^n, \mathcal{R})}$ and $D_{\mathcal{X}^n \otimes \mathcal{R}}$ are (ϵ, δ) indistinguishable implies that $D_{(\mathcal{X}, \mathcal{R})}$ and $D_{\mathcal{X} \otimes \mathcal{R}}$ are indistinguishable as well. $\qquad\square$

Proof of Theorem 4.4. Given any $\mathbf{b} \subseteq \mathcal{X} \times \mathcal{R}$ we denote $\mathbf{r_b}(x) := \{r \in \mathcal{R} \mid (x, r) \in \mathbf{b}\}$ (which might be empty for some x's). Using this notation and the subset $\mathbf{s}$ from Definition D.3 we prove that for any $\mathbf{b} \subseteq \mathcal{X} \times \mathcal{R}$,

$$D_{(\mathcal{X}, \mathcal{R})}(\mathbf{b}) = \sum_{x \in \mathcal{X}} D(x) D(\mathbf{r_b}(x) \mid x)$$

$$\overset{(1)}{=} \sum_{x \in \mathcal{X}} \sum_{s \in \mathcal{X}^n} D(s) D(x \mid s) \overbrace{\sum_{s' \in \mathcal{X}^n} D(s') D(\mathbf{r_b}(x) \mid s)}^{=1}$$

$$\overset{(2)}{\leq} e^\epsilon \sum_{x \in \mathcal{X}} \sum_{s \in \mathbf{s}} D(s) D(x \mid s) \sum_{s' \in \mathbf{s}} D(s') D(\mathbf{r_b}(x) \mid s') + \delta + 2\eta$$

$$\leq e^\epsilon \sum_{x \in \mathcal{X}} \overbrace{\sum_{s \in \mathcal{X}^n} D(s) D(x \mid s)}^{=D(x)} \overbrace{\sum_{s' \in \mathcal{X}^n} D(s') D(\mathbf{r_b}(x) \mid s')}^{=D(\mathbf{r_b}(x))} + \delta + 2\eta$$

$$= e^\epsilon D_{\mathcal{X} \otimes \mathcal{R}}(\mathbf{b}) + \delta + 2\eta$$

where (1) results from the fact that x and r are independent given s, and (2) from the definition of TS. $\qquad\square$

Proof of Theorem 4.5. Assume M is not $\left(\epsilon', \frac{\delta}{\epsilon}\right)$-LSS, which means that in particular $D(\mathbf{r}_{\epsilon'}) > \frac{\delta}{\epsilon}$. Denoting $B_{\mathcal{X} \times \mathcal{R}}^{\epsilon'} := \underset{r \in \mathbf{r}_{\epsilon'}}{\cup} (\mathbf{x}_+(r) \times \{r\})$ we get that from the definition of the stability loss,

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

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

$$D_{(\mathcal{X}, \mathcal{R})}\left(B_{\mathcal{X} \times \mathcal{R}}^{\epsilon' \epsilon'}\right) - D_{\mathcal{X} \otimes \mathcal{R}}\left(B_{\mathcal{X} \times \mathcal{R}}^{\epsilon'}\right) \leq D_{\mathcal{X} \otimes \mathcal{R}}\left(B_{\mathcal{X} \times \mathcal{R}}^{\epsilon'}\right) \cdot (e^\epsilon - 1) + \delta$$

$$\overset{(1)}{\leq} D(\mathbf{r}_{\epsilon'}) \cdot (e^\epsilon - 1) + \epsilon \cdot D(\mathbf{r}_{\epsilon'})$$

$$\overset{(2)}{\leq} \epsilon' \cdot D(\mathbf{r}_{\epsilon'})$$

where (1) results from the fact that $\epsilon \cdot D(\mathbf{r}_{\epsilon'}) > \delta$, and (2) from the definition of ϵ' and the assumption that M is not $\left(\epsilon', \frac{\delta}{\epsilon}\right)$-LSS. The proof is concluded by repeating the same process for the second direction. $\qquad\square$

Theorem D.7 (Local Bounded Maximal Leakage implies Local Max Information)**.** *Given* $0 \leq \epsilon$, $0 < \delta \leq 1$ *a distribution* $D_{\mathcal{X}^n}$ *and a query* q, *if a mechanism* M *is* ϵ-*LML with respect to* $D_{\mathcal{X}^n}$ *and* q, *then it is* $\left(\epsilon + \ln\left(\frac{1}{\delta}\right), \delta\right)$-*LMI with respect to the same* $D_{\mathcal{X}^n}$ *and* q.

Proof. The proof is identical to the one used by [EGI19] when proving that ML implies MI (Theorem 7). $\qquad\square$

Lemma D.8 (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* $M : \mathcal{X}^n \times \mathcal{Q} \to \mathcal{H}$ *be a mechanism with a compression scheme (Definition D.6), 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,

$$|h_w(s \setminus w) - h_w(D_{\mathcal{X}})| \leq \sqrt{h_w(s \setminus w) \frac{4m\ln(2n/\delta)}{n}} + \frac{8m\ln(2n/\delta)}{n}$$

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 4.6. We will prove that g is (ϵ, δ)-LSS for such an ϵ, 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,

$$\ell(w) = \sum_{x \in \mathbf{x}_+(w)} (D(x \mid w) - D(x))$$

$$= \sum_{x \in \mathbf{x}_+(w)} \sum_{s \in \mathcal{X}^n} D(s \mid w)(D(x \mid s) - D(x))$$

$$= \sum_{s \in \mathcal{X}^n} D(s \mid w) \sum_{x \in \mathbf{x}_+(w)} \left(\frac{m}{n} D(x \mid w) + \frac{n-m}{n} D(x \mid s \setminus w) - D(x) \right)$$

$$\leq \sum_{s \in \mathcal{X}^n} D(s \mid w) \left(\frac{m}{n} + \sum_{x \in \mathbf{x}_+(w)} (D(x \mid s \setminus w) - D(x)) \right)$$

$$= \sum_{s \in \mathcal{X}^n} D(s \mid w) \left(\frac{m}{n} + \sum_{x \in \mathcal{X}} (D(x \mid s \setminus w) - D(x)) h_w^+(x) \right)$$

$$= \sum_{s \in \mathcal{X}^n} D(s \mid w) \left(\frac{m}{n} + h_w^+(s \setminus w) - h_w^+(D_{\mathcal{X}}) \right)$$

where $h_w^+(x)$ is simply the characteristic function of $\mathbf{x}_+(w)$.

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

$$D(\mathbf{r})(\ell(\mathbf{r}) - \epsilon) = \sum_{w \in \mathbf{r}} D(w)(\ell(w) - \epsilon)$$

$$\overset{(1)}{\leq} \sum_{w \in \mathbf{r}} D(w) \sum_{s \in \mathcal{X}^n} D(s \mid w) \left(\frac{m}{n} + h_w^+(s \setminus w) - h_w^+(D_{\mathcal{X}}) - \epsilon \right)$$

$$= \sum_{s \in \mathcal{X}^n} D(s) \sum_{w \in \mathbf{r}} D(w \mid s) \left(h_w^+(s \setminus w) - h_w^+(D_{\mathcal{X}}) - \left(\epsilon - \frac{m}{n} \right) \right)$$

$$\leq \sum_{s \in \mathcal{X}^n} D(s) \max_{w = g(s), h = f(w)} \left(h(s \setminus w) - h(D_{\mathcal{X}}) - \left(\epsilon - \frac{m}{n} \right) \right)$$

$$\overset{(2)}{\leq} \Pr_{S \sim D_{\mathcal{X}^n}, W \sim g(S), H \sim f(W)} \left[H(S \setminus W) - H(D_{\mathcal{X}}) > \left(\epsilon - \frac{m}{n} \right) \right]$$

$$\overset{(3)}{\leq} \sqrt{\frac{4m\ln(2n/\delta)}{n}} + \frac{8m\ln(2n/\delta)}{n} + \frac{m}{n}$$

$$\overset{(4)}{\leq} 11 \sqrt{\frac{m\ln(2n/\delta)}{n}}$$

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.8 and the fact that the value of h is bounded by 1, and (4) from the fact that $m \leq \frac{n}{9\ln\left(\frac{2n}{\delta}\right)}$. $\qquad \square$

D.3 Proofs of Separation Theorems

Proof of Theorem 4.7. Without loss of generality, assume $0 < \epsilon \leq 0.7$. Given $0 \leq \alpha \leq \frac{\epsilon}{7}, p = \frac{1}{2} + \alpha$ we will define some function $f : \mathcal{X} \to \{0, 1\}$, and for $i \in \{0, 1\}$ denote $\mathbf{x}_i := \{x \in \mathcal{X} \mid f(x) = i\}$, set an arbitrary distribution $D_{\mathcal{X}}$ such that $D(\mathbf{x}_1) = p$, and $D_{\mathcal{X}^n}$ which is the product of $D_{\mathcal{X}}$. We will consider a mechanism M which in response to a query q 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, $M(q, s) = |s \cap \mathbf{x}_1| \pmod 2$, and we prove that this mechanism is $(\epsilon, 0)$-LMI but not $\left(1, \frac{1}{5}\right)$-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, and the possible outputs as r_0, r_1. Notice that,

$$D(r_1 \mid \mathbf{x}_1) = p \cdot p_{n-2} + (1 - p)(1 - p_{n-2})$$
$$D(r_1 \mid \mathbf{x}_0) = (1 - p) p_{n-2} + p(1 - p_{n-2}) = 1 - D(r_1 \mid \mathbf{x}_1)$$

$$\begin{aligned}
D(r_1) &= p \cdot D(r_1 \mid \mathbf{x}_1) + (1 - p) D(r_1 \mid \mathbf{x}_0) \\
&= (2p - 1) D(r_1 \mid \mathbf{x}_1) + 1 - p \\
&= (1 - 2p) D(r_1 \mid \mathbf{x}_0) + p
\end{aligned}$$

Using these identities we will first prove that $\frac{D(r_1)}{D(r_1 \mid \mathbf{x}_1)}, \frac{D(r_1)}{D(r_1 \mid \mathbf{x}_0)} \leq e^{\epsilon}$. Since a similar claim can be proven for $\frac{D(r_1 \mid \mathbf{x}_1)}{D(r_1)}, \frac{D(r_1 \mid \mathbf{x}_0)}{D(r_1)}$, we get that this mechanism is $(\epsilon, 0)$-LMI.

$$\begin{aligned}
\frac{D(r_1)}{D(r_1 \mid \mathbf{x}_1)} &= \frac{(2p - 1) D(r_1 \mid \mathbf{x}_1) + 1 - p}{D(r_1 \mid \mathbf{x}_1)} \\
&= 2p - 1 + \frac{1 - p}{(2p - 1) p_{n-2} + 1 - p} \\
&= 2p - \frac{(2p - 1) p_{n-2}}{(2p - 1) p_{n-2} + 1 - p} \\
&= 1 + 2\alpha - \overbrace{\frac{2\alpha p_{n-2}}{\alpha (2p_{n-2} - 1) + \frac{1}{2}}}^{\geq 0} \\
&\overset{(1)}{\leq} 1 + \overbrace{2\alpha}^{\leq \epsilon} \\
&\overset{(2)}{\leq} e^{\epsilon}
\end{aligned}$$

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,

$$\begin{aligned}
\frac{D(r_1)}{D(r_1 \mid \mathbf{x}_0)} &= \frac{(1 - 2p) D(r_1 \mid \mathbf{x}_0) + p}{D(r_1 \mid \mathbf{x}_0)} \\
&= 1 - 2p + \frac{p}{D(r_1 \mid \mathbf{x}_0)} \\
&= 2 - 2p - \frac{(1 - 2p) p_{n-2}}{(1 - 2p) p_{n-2} + p} \\
&= 1 + 2\alpha + \overbrace{\frac{2\alpha \cdot p_{n-2}}{\alpha (1 - 2p_{n-2}) + \frac{1}{2}}}^{\leq 5\alpha} \\
&\overset{(1)}{\leq} 1 + \overbrace{7\alpha}^{\leq \epsilon} \\
&\overset{(2)}{\leq} e^{\epsilon}
\end{aligned}$$

26

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 $\alpha \left(1 - 2p_{n-2}\right) + \frac{1}{2} \geq \frac{4}{10}$, 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. Using the fact that the parity function of a Binomial random variable $\mathbf{b}\left(n, p\right)$ is a Bernoulli random variable $\mathrm{Ber}\left(\frac{1 - (1 - 2p)^n}{2}\right)$, and denoting $\mathcal{S}_1$ the set of all sample sets with parity value 1, we get that,

$$D_{\mathcal{X}^n \otimes \mathcal{R}}\left(\mathcal{S}_1 \times \{r_0\}\right) = \overbrace{D\left(\mathcal{S}_1\right)}^{D(r_1)} \cdot D\left(r_0\right)$$

$$= \frac{1 - (1 - 2p)^{2n}}{4}$$

$$= e^1 \overbrace{D_{(\mathcal{X}^n, \mathcal{R})}\left(\mathcal{S}_1 \times \{r_0\}\right)}^{=0} + \frac{1 - (2\alpha)^{2n}}{4}$$

$$\overset{(1)}{>} e^1 D_{(\mathcal{X}^n, \mathcal{R})}\left(\mathcal{S}_1 \times \{r_0\}\right) + \frac{1}{5}$$

where (1) is a result of the fact that $0 \leq \alpha < \frac{\epsilon}{7} \leq \frac{1}{10}$, $n \geq 3$ and $\frac{1 - \left(\frac{1}{5}\right)^6}{4} > \frac{1}{5}$, which means this mechanism is not $\left(1, \frac{1}{5}\right)$-MI.

$\square$

Proof of Theorem 4.8. Without loss of generality $0 \leq \delta \leq 0.1$, so $n > 2\ln\left(\frac{2}{\delta}\right)$. Given $N > n^2$, $\mathcal{X} := [N]$, 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 query q 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 4.6 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 $\mathbf{b} := \underset{r \in \mathcal{R}}{\cup}\left(x\left(r\right), r\right)$.

$$D_{(\mathcal{X}, \mathcal{R})}\left(\mathbf{b}\right) = \sum_{r \in \mathcal{R}} D\left(r\right) \cdot D\left(x\left(r\right) \mid r\right)$$

$$\overset{(1)}{\geq} \sum_{r \in \mathcal{R}} D\left(r\right) \cdot \frac{1}{n}$$

$$\overset{(2)}{>} \sum_{r \in \mathcal{R}} D\left(r\right) e\frac{1}{n^2} + \overbrace{\sum_{r \in \mathcal{R}} D\left(r\right)}^{=1} \frac{1}{2n}$$

$$\geq e \sum_{r \in \mathcal{R}} D\left(r\right) \cdot \overbrace{D\left(x\left(r\right)\right)}^{\leq \frac{1}{n^2}} + \frac{1}{2n}$$

$$= e^1 \cdot D_{\mathcal{X} \otimes \mathcal{R}}\left(\mathbf{b}\right) + \frac{1}{2n}$$

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.

$\square$

E   Missing Details from Section 5

Definitions and properties in this section are due to [DR$^+$14].

Definition E.1 (Laplace Mechanism). Given $0 \leq b$ and a query $q \in \mathcal{Q}_\Delta$, the Laplace mechanism with parameter b is defined as:
$$M\left(s, q\right) = q\left(s\right) + \mathrm{Lap}_b$$
where Lap_b is a random variable with unbiased Laplace distribution, which if a symmetric exponential distribution. Formally:
$$\mathrm{Lap}_b\left(x\right) = \frac{1}{2b} e^{-\frac{|x|}{b}}$$

Theorem E.2 (Laplace Mechanism is Differentially Private). *Given $0 \leq b, \epsilon$ and a query $q \in \mathcal{Q}_\Delta$, the Laplace mechanism with parameter b is $\left(\frac{2\Delta}{n \cdot b}, 0\right)$-DP.*

Theorem E.3 (Laplace Mechanism is Sample Accurate). *Given $0 \leq b$, $0 < \delta \leq 1$ and a query $q \in \mathcal{Q}_\Delta$, the Laplace mechanism with parameter b is $\left(b \cdot \ln\left(\frac{1}{\delta}\right), \delta\right)$-Sample Accurate.*

Definition E.4 (Gaussian Mechanism). Given $0 \leq \sigma$ and a query $q \in \mathcal{Q}_\Delta$, the Gaussian mechanism with parameter σ is defined as:
$$M\left(s, q\right) = q\left(s\right) + \mathrm{G}_\sigma$$
where G_σ is a random variable with unbiased Gaussian distribution and standard deviation σ.

Theorem E.5 (Gaussian Mechanism is Differentially Private). *Given $0 \leq \sigma, \epsilon$, $0 < \delta \leq 1$, and a query $q \in \mathcal{Q}_\Delta$, the Gaussian mechanism with parameter σ is $\left(\frac{2\Delta\sqrt{2\ln(1.25/\delta)}}{n\sigma}, \delta\right)$-DP.*

Theorem E.6 (Gaussian Mechanism is Sample Accurate). *Given $0 \leq \sigma, \epsilon$, $0 < \delta \leq 1$, and a query $q \in \mathcal{Q}_\Delta$, the Gaussian mechanism with parameter σ is $\left(\frac{\epsilon}{\sqrt{2\ln\left(\sqrt{2}/\pi\delta\right)}}, \delta\right)$-Sample Accurate.*

F   Distance Measures on Distributions

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

Definition F.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,

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

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 F.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
$$\mathbf{D}^\delta_\infty\left(D_1 \| D_2\right) \coloneqq \max_{\mathbf{r} \subseteq \mathrm{Supp}(D_1) \wedge D_1(\mathbf{r}) \geq \delta} \ln\left(\frac{D_1\left(\mathbf{r}\right) - \delta}{D_2\left(\mathbf{r}\right)}\right).$$
The case where $\delta = 0$ is simply called the *max divergence*.

Definition F.3 (Indistinguishable distributions). Two probability distributions D_1, D_2 over some domain $\mathcal{R}$ will be called (ϵ, δ)-*indistinguishable* if
$$\max\left\{\mathbf{D}^\delta_\infty\left(D_1 \| D_2\right), \mathbf{D}^\delta_\infty\left(D_2 \| D_1\right)\right\} \leq \epsilon.$$
this can also be written as the condition that for any $\mathbf{r} \subseteq \mathcal{R}$
$$D_1\left(\mathbf{r}\right) \leq e^\epsilon \cdot D_2\left(\mathbf{r}\right) + \delta \quad \text{and} \quad D_2\left(\mathbf{r}\right) \leq e^\epsilon \cdot D_1\left(\mathbf{r}\right) + \delta$$

Definition F.4 (Maximal Leakage, based on [IWK18]). Given two finite domains $\mathcal{X}, \mathcal{Y}$ and a joint distribution $D_{(\mathcal{X}, \mathcal{Y})}$ defined over $\mathcal{X} \times \mathcal{Y}$, The *Maximal Leakage* between two marginal distributions $D_\mathcal{X}, D_\mathcal{Y}$ is defined as,

$$\mathcal{L}\left(D_\mathcal{X} \rightarrow D_\mathcal{Y}\right) \coloneqq \log\left(\sum_{y \in \mathcal{Y}} \max_{x \in \mathcal{X} \,|\, D(x) > 0} D\left(y \,|\, x\right)\right).$$