Adaptive and oblivious statistical adversaries are equivalent

Guy Blanc
Stanford

Gregory Valiant
Stanford

September 3, 2025

Abstract

We resolve a fundamental question about the ability to perform a statistical task, such as learning, when an adversary corrupts the sample. Such adversaries are specified by the types of corruption they can make and their level of knowledge about the sample. The latter distinguishes between sample-adaptive adversaries which know the contents of the sample when choosing the corruption, and sample-oblivious adversaries, which do not. We prove that for all types of corruptions, sample-adaptive and sample-oblivious adversaries are *equivalent* up to polynomial factors in the sample size. This resolves the main open question introduced by [BLMT22] and further explored in [CHL+23].

Specifically, consider any algorithm A that solves a statistical task even when a sample-oblivious adversary corrupts its input. We show that there is an algorithm A' that solves the same task when the corresponding sample-adaptive adversary corrupts its input. The construction of A' is simple and maintains the computational efficiency of A: It requests a polynomially larger sample than A uses and then runs A on a uniformly random subsample.

Contents

1 Introduction

Classic models of data analysis assume that data is drawn independently from the distribution of interest, but the real world is rarely so kind. To be robust to the messiness of real-world data, we desire algorithms that succeed even in the presence of an adversary that corrupts the data. Such adversaries were first introduced in the seminal works of [Tuk60, Hub64, Ham71] and have since been the subject of intense study in a variety of settings [Val85, Hau92, KL93, KSS94, BEK02, DSFT+14, LRV16, CSV17, DKK+19, DKPZ21, HSSVG22, BLMT22, CHL+23, DK23].

By now, there are numerous models for how the adversary can corrupt the data, including additive, subtractive, "strong"/"nasty", agnostic, and adaptive and non-adaptive variants of each of these. In many cases, the provable guarantees for our algorithms are only known for a subset of these models. We refer the interested reader to the excellent recent textbook [DK23] for a more complete background and survey of recent results and open directions. Our work focuses on a surprisingly under-explored question:

What is the relationship between the various statistical adversaries?

Specifically, we compare *adaptive* adversaries, which can look at the sample before deciding on a corruption, and *oblivious* adversaries, which must commit to their corruptions before the i.i.d. sample is drawn.

Theorem 1 (Informal, see Theorem 2 for the formal version)**.** *Adaptive adversaries and their oblivious counterparts are equivalent up to scaling the sample size by a factor polynomial in the original sample size and polylogarithmic in the domain size.*

Theorem 1 resolves the main question introduced by [BLMT22] and further explored in [CHL+23]. We defer its formal statement to Section 2, but, for now, mention two points. First, it is a generic result that proves the equivalence between many distinct adaptive adversaries and their oblivious counterparts (e.g. the equivalence between "subtractive adaptive" and "subtractive oblivious" adversaries). Second, it is constructive. We give a simple transformation, the subsampling filter described in Definition 5, which takes any algorithm that succeeds on a statistical task in the presence of the oblivious adversary and converts it to one that succeeds on the same task in the presence of the adaptive adversary. This transformation preserves the statistical and computational efficiency of the original algorithm up to polynomial factors.

In addition to answering a foundational question about the relative power of statistical adversaries, Theorem 1 has several practical implications:

1. Given the many distinct definitions of robustness, it can be difficult for a practitioner to determine which definition is most appropriate for their setting and therefore which algorithm to utilize. Theorem 1 partially alleviates this issue by greatly reducing the number of truly unique adversary models.

2. It shows that a single algorithmic idea, that of subsampling, *amplifies* robustness in many different models. Formally, it takes an algorithm that is only robust to the oblivious adversary and converts it to one robust to the adaptive counterpart. This suggests that, even if the practitioner cannot precisely determine the most appropriate model of robustness, they should try subsampling.

3. Theorem 1 can be reformulated as an answer to an equivalent and independently interesting question: How useful is it to hide one's dataset from the adversary? It shows that private data does *not* afford much more robustness than public data.

2 Our Results

Before formally describing our main result, we define a unified framework in which to express and analyze statistical adversaries. It may be instructive to view this framework with the following concrete adaptive adversary and its oblivious counterpart in mind:

Example 1. Consider the following adaptive and oblivious adversaries parameterized by $\eta \in [0, 1]$:

- Adaptive: When the algorithm requests n points, first an i.i.d. sample $\boldsymbol{S} \sim \mathcal{D}^n$ is drawn. Then, may alter up to $\lfloor \eta \cdot n \rfloor$ of them arbitrarily. The algorithm receives this corrupted sample.
- Oblivious: The adversary can choose any $\mathcal{D}'$ that has a total variation distance to $\mathcal{D}$ of at most η, and the algorithm receives n i.i.d. draws from $\mathcal{D}'$.

These two adversaries are well-studied, and are referred to by different names. The adaptive adversary is typically referred to as "strong contamination" in the statistical estimation literature [DK23] and "nasty noise" in the PAC learning literature [BEK02]. The oblivious adversary has been referred to as "general, non-adaptive, contamination" [DK23]. In our unified framework, these adversaries will be defined via the same "cost function," and as a result, we prove them equivalent.

2.1 A unified framework to define statistical adversaries

Each adversary will be parameterized by a "cost" function ρ where $\rho(x, y)$ specifies the cost the adversary pays to corrupt x to y, with a cost of ∞ indicating that the adversary is not allowed to change x to y. The adversary can choose any corruptions subject to a budget constraint on the total cost incurred. This cost function is required to have two basic properties.

Definition 1 (Cost function)**.** *A function $\rho : X \times X \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ is said to be a "cost function" if it satisfies the following properties.*

1. *For any $x \in X$, $\rho(x, x) = 0$.*
2. *For any $x, y \in X$, $\rho(x, y) \geq 0$.*

The adversary is specified by both the cost function and whether it is adaptive or oblivious. Given the cost function, ρ, the corresponding adaptive adversary is defined as follows:

Definition 2 (Adaptive adversary, corruptions to the sample)**.** *For any cost function ρ and $S \in X^n$, we use $\mathcal{C}_\rho(S)$ to denote all $S' \in X^n$ for which*

$$\frac{1}{n} \sum_{i \in [n]} \rho(S_i, S'_i) \leq 1.$$

The ρ-adaptive adversary is allowed to corrupt the clean sample S to any $S' \in \mathcal{C}_\rho(S)$. For any $f : X^n \to \{0, 1\}$ and distribution $\mathcal{D}$, the max success probability of f in the presence of the ρ-adaptive adversary is denoted:

$$\text{Adaptive-Max}_\rho(f, \mathcal{D}) \coloneqq \mathop{\mathbb{E}}_{\boldsymbol{S} \sim \mathcal{D}^n} \left[\sup_{\boldsymbol{S}' \in \mathcal{C}_\rho(\boldsymbol{S})} \{ f(\boldsymbol{S}') \} \right].$$

In the case of the adversaries of Example 1, their cost functions are simply $\rho(x, y) = 1/\eta$ for all $x \neq y$. In that case, the budget constraint in the above definition ensures that, for this choice of cost function, ρ, the ρ-adaptive adversary can corrupt at most an η fraction of points in the sample, corresponding to the standard definition of the "strong contamination"/"nasty noise" models.

Given a cost function, the associated oblivious adversary replaces the budget constraint of the adaptive setting with a natural distributional analog. It is easy to see that the following definition of ρ-oblivious adversaries is equivalent to the "general, non-adaptive, contamination" model of Example 1 when the cost function is defined as $\rho(x, y) = 1/\eta$ for all $x \neq y$.

Definition 3 (Oblivious adversary, corruptions to a distribution)**.** *For any cost function ρ and distribution $\mathcal{D}$, we overload $\mathcal{C}_\rho(\mathcal{D})$ to refer to the set of all distributions $\mathcal{D}'$ for which there exists a coupling of $\boldsymbol{x} \sim \mathcal{D}$ and $\boldsymbol{x}' \sim \mathcal{D}'$ satisfying*

$$\mathbb{E}[\rho(\boldsymbol{x}, \boldsymbol{x}')] \leq 1.$$

The ρ-oblivious adversary *is allowed to corrupt the base distribution $\mathcal{D}$ to any $\mathcal{D}' \in \mathcal{C}_\rho(\mathcal{D})$. For any $f : X^n \to \{0, 1\}$ and distribution $\mathcal{D}$, the max success probability of f in the presence of the ρ-oblivious adversary is denoted:*

$$\text{Oblivious-Max}_\rho(f, \mathcal{D}) \coloneqq \sup_{\mathcal{D}' \in \mathcal{C}_\rho(\mathcal{D})} \left\{ \mathop{\mathbb{E}}_{\boldsymbol{S}' \sim (\mathcal{D}')^n} \left[f(\boldsymbol{S}') \right] \right\}.$$

In Definition 3, since $\boldsymbol{x} \sim \mathcal{D}$ and $\boldsymbol{x}' \sim \mathcal{D}'$ can be coupled so that the average cost to corrupt $\boldsymbol{x}$ to $\boldsymbol{x}'$ is at most 1, we can similarly couple $\boldsymbol{S} \sim \mathcal{D}^n$ and $\boldsymbol{S}' \sim (\mathcal{D}')^n$ so that the average cost to corrupt each point in $\boldsymbol{S}$ to the corresponding point in $\boldsymbol{S}'$ is at most 1. From this perspective, the crucial difference between the oblivious and adaptive adversary is that the oblivious adversary must commit to how it corrupts each $\boldsymbol{x}$ without knowing the contents of the sample, whereas the adaptive adversary gets to view $\boldsymbol{S}$ before deciding.

We show, in Section 3, that our framework can express many commonly studied statistical adversaries, including subtractive contamination, additive contamination, and agnostic noise.

Remark 1 (Partially-adaptive statistical adversaries)**.** Some statistical adversaries lie between their fully adaptive and fully oblivious counterparts. These include malicious noise [Val85] and the non-iid oblivious adversary defined in [CHL+23]. Our results readily extend to such adversaries (see Section 3.1 for details).

2.2 Our main result: Adaptive and oblivious adversaries are equivalent

Our main result is that for any algorithm A and cost function ρ, there exists an algorithm A' inheriting the efficiency of A for which the performance of A in the presence of the oblivious adversary is equivalent to the performance of A' in the presence of the adaptive adversary.

Definition 4 (ε-equivalent algorithms)**.** *For any algorithms $A : X^n \to Y$ and $A' : X^m \to Y$, we say that A in the presence of the ρ-oblivious adversary is ε-equivalent to A' in the presence of the ρ-adaptive adversary if for any test function $T : Y \to \{0, 1\}$ and distribution $\mathcal{D}$ supported on X,*

$$\left| \text{Oblivious-Max}_\rho(T \circ A, \mathcal{D}) - \text{Adaptive-Max}_\rho(T \circ A', \mathcal{D}) \right| \leq \varepsilon.$$

Colloquially, A and A' are ε-equivalent if no test can distinguish their outputs with more than ε probability. Note that while the above definition is about the *maximum* acceptance probability of T, it also applies to the test $\overline{T} \coloneqq 1 - T$ and therefore the *minimum* acceptance probability of T also must be approximately the same for A and A'.

The algorithm A' will run A on a uniformly random subsample of its input.

Definition 5 (Subsampling filter)**.** *For any $m \geq n$ we define the* subsampling filter $\Phi_{m\to n} : X^m \to X^n$ *as the (randomized) algorithm that given $S \in X^m$, returns a sample of n points drawn uniformly without replacement from S.*

Theorem 2 (Subsampling neutralizes the adaptivity in statistical adversaries)**.** *For any algorithm $A : X^n \to Y$, $\varepsilon > 0$, and cost function ρ, let $m = \mathrm{poly}(n, \ln|X|, 1/\varepsilon)$ and $A' \coloneqq A \circ \Phi_{m\to n}$. Then, A in the presence of the ρ-oblivious adversary is ε-equivalent to A' in the presence of the ρ-adaptive adversary.*

For constant ε, Theorem 2 says that if there is an algorithm A solving a statistical task with an oblivious adversary taking as input $n \cdot \log|X|$ bits, there is an algorithm A' solving the same task with an adaptive adversary taking only polynomially more bits as input. Furthermore, if A is computationally efficient, then A' is too.

Remark 2 (Continuous domains)**.** In many statistical problems, the domain is $\mathbb{R}^d$. To apply Theorem 2 to an algorithm A over continuous domains, we first discretize that domain to some $X \coloneqq \mathrm{disc}(\mathbb{R})^d$ where the discretization depends on A. If A requires b bits of precision in each dimension, then $\log_2|X| = bd$, which is typically polynomial in n. For example, under the mild assumption that A accesses the bits of each dimension sequentially, both b and d are upper bounded by the time complexity of A. In this setting, if the time complexity of A is polynomial in n, then so is $\ln|X|$.

Theorem 2 is a special case of our main theorem in which the $|X|$ is replaced with the *degree* of the cost function, a measure of the number of corruptions the adversary can make for each input.

Definition 6 (Degree of a cost function)**.** *For any cost function $\rho : X \times X \to \mathbb{R}_{\geq 0} \cup \{\infty\}$, the degree of ρ is defined as*

$$\deg(\rho) \coloneqq \sup_{x \in X} \{\textit{The number of distinct } y \in X \textit{ for which } \rho(x, y) \neq \infty\}.$$

Theorem 3 (Main result, generalization of Theorem 2)**.** *For any algorithm $A : X^n \to Y$, $\varepsilon > 0$, and cost function ρ with degree $d \geq 2$, let $m = O\left(\frac{n^4(\ln d)^2}{\varepsilon^4}\right)$ and $A' \coloneqq A \circ \Phi_{m\to n}$. Then, A in the presence of the ρ-oblivious adversary is ε-equivalent to A' in the presence of the ρ-adaptive adversary.*

The degree is constant for many natural cost functions, such as the cost function corresponding to subtractive contamination. In these cases, Theorem 3 has no dependence on the domain size.

2.3 Lower bounds

Theorem 3 requires a polynomial increase of the sample size of A' relative to A. It is natural to wonder whether such an increase is necessary. The results of [CHL+23] show it is.

Fact 2.1 ([CHL+23]). *For any $n \in \mathbb{N}$, the task of Gaussian mean testing with appropriate parameters (depending on n) can be solved using n samples in the presence of the* oblivious *additive adversary, but requires $\tilde{\Omega}(n^{4/3})$ in the presence of the* adaptive *additive adversary.*

In the setting of Theorem 3, one difficulty in interpreting Fact 2.1 is that the cost function corresponding to the additive adversary has a large degree[1], and so it's unclear if the increased sample size is to due a dependence on the degree or an innate required polynomial increase.

For example, the subtractive adversary (formally defined in Section 3) has a degree of 2 because, for each point in the sample, it chooses between keeping that point or removing it. For this adversary, is a polynomial increase in sample size necessary? Our first lower bound gives a straightforward proof this is the case, even for a simple task.

Theorem 4 (A polynomial increase in sample size is necessary). *Let $\mathcal{D}$ be a distribution on $X = [m] \coloneqq \{1, \dots, m\}$ that is promised to be uniform on some $X' \subseteq X$. Then,*

1. *There is an algorithm that estimates $|X'|$ using $n \coloneqq \tilde{O}(\sqrt{m})$ samples even with* oblivious *subtractive contamination.*
2. *Any algorithm that estimates $|X'|$ to the same accuracy with* adaptive *subtractive contamination requires $\tilde{\Omega}(m)$ samples.*

Theorem 4 implies that in the statement of Theorem 2, we must take m polynomial larger than n. Next, we show that this m must also depend polylogarithmically on a degree-like characteristic of the cost function.

Definition 7 (Budget-bounded degree). *For any cost function $\rho : X \times X \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ and $b \in \mathbb{R}_{\geq 0}$, the b-bounded degree of ρ is defined as*

$$\deg_b(\rho) \coloneqq \sup_{x \in X} \{\textit{The number of distinct } y \in X \textit{ for which } \rho(x, y) \leq b\}.$$

This lower bound will make one assumption on ρ: that $\rho(x, y) \geq 1 + \delta$ whenever $x \neq y$ for a small constant δ. This corresponds to the adversary having a budget on how many points they can change and is satisfied by all of the well-studied models discussed in Section 3.

Theorem 5 (Dependence on $\ln \deg_b(\rho)$ is necessary). *For any constants $b, \delta > 0$, large enough $n \in \mathbb{N}$, and cost function ρ for which $\rho(x, y) \geq 1 + \delta$ whenever $x \neq y$, there is an algorithm $A : X^n \to \{0, 1\}$ for which the following holds. If A in the presence of the ρ-oblivious adversary is $(\varepsilon = 0.9)$-equivalent to $A' \coloneqq A \circ \Phi_{m \to n}$ in the presence of the ρ-adaptive adversary, then*

$$m \geq \tilde{\Omega}_{b,\delta}(n \cdot \ln \deg_b(\rho)).$$

Comparing Theorems 3 and 5, for "reasonable" cost functions in which $\deg(\rho) \approx \deg_{1000}(\rho)$ and $\rho(x, y) \geq 1.001$ for all $x \neq y$, a domain-size independent result is possible precisely when the degree does not grow with $|X|$.

[1] Since the domain for the task is over the continuous domain of $\mathbb{R}^d$, technically this cost function has infinite degree. However, as we discussed in Remark 2, it makes more sense to think of the degree as $\approx 2^d$ in this setting, which happens to be exponential in the n of Fact 2.1.

2.4 Relation to recent work

Recent work of Blanc, Lange, Malik, and Tan initiated a formal study of the relationship between adaptive adversaries and their oblivious counterparts [BLMT22]. They conjectured the equivalence of adaptive and oblivious statistical adversaries but only proved it in two special cases.

1. They showed that *additive* oblivious and *additive* adaptive adversaries are equivalent. Our result, which applies to all statistical adversaries, requires an entirely different approach. This is because, in some sense, the adaptive additive adversary is *less adaptive* than other adaptive adversaries. We elaborate on this point in Appendix C.1 and explain why [BLMT22]'s approach does not generalize to all adversaries.

2. They also showed that if a *statistical query* (SQ) algorithm is robust to an oblivious adversary, it can be upgraded to be robust to the corresponding adaptive adversary. The restriction to SQ algorithms greatly facilitates [BLMT22]'s analysis because we have a much better understanding of SQ algorithms than general algorithms. For example, the quality of the best SQ algorithm for a given task is captured by simple combinatorial measures [BFJ+94, Fel17]. In Appendix C.2, we further describe [BLMT22]'s SQ result and give advantages of our result even for algorithms that can be cast in the SQ framework.

Other recent work of Canonne, Hopkins, Li, Liu, and Narayanan tackled the equivalence of statistical adversaries from the other direction [CHL+23]. While we aim to show that distinct statistical adversaries are equivalent, they showed a separation: For the well-studied problem of Gaussian mean testing, an adaptive adversary requires polynomial more samples than the corresponding oblivious adversary (see Fact 2.1).

3 Instantiating common adversaries in our framework

Here, we show how to express many common statistical adversaries within our framework. For completeness, we include the "strong contamination/nasty noise" adversary of Example 1.

Strong contamination/nasty noise: As mentioned in Example 1, both the "strong contamination/nasty noise" adversary, that can arbitrarily replace an η fraction of an i.i.d. sample, and the "general, non-adaptive, contamination" adversary, that can perturb the underlying distribution from which the sample is drawn by at most η in total variation distance, correspond to adaptive and oblivious adversaries with the following cost function:

$$\rho_{\text{strong}}(x, y) \coloneqq \begin{cases} 0 & \text{if } x = y \\ \frac{1}{\eta} & \text{otherwise.} \end{cases}$$

Agnostic learning [Hau92, KSS94]: Agnostic noise is a well-studied adversary [KKMS08, KK09, Fel10, DSFT+14, DKPZ21] specific to supervised learning problems, where each point in the sample is a pair (x, y) of the input and its label. This adversary is allowed to change η fraction of the labels but must keep the inputs unchanged. It corresponds to the cost function,

$$\rho_{\text{agnostic}}((x_1, y_1), (x_2, y_2)) \coloneqq \begin{cases} 0 & \text{if } x_1 = x_2 \text{ and } y_1 = y_2 \\ \frac{1}{\eta} & \text{if } x_1 = x_2 \text{ and } y_1 \neq y_2 \\ \infty & \text{if } x_1 \neq x_2. \end{cases}$$

Agnostic learning typically refers to the ρ-oblivious adversary. It can be equivalently defined as the learner receiving an i.i.d. sample of points of the form $(\boldsymbol{x}, g(\boldsymbol{x}))$ where g is close to the original target f in the sense that

$$\Pr_{\boldsymbol{x}}[f(\boldsymbol{x}) \neq g(\boldsymbol{x})] \leq \eta.$$

In the adaptive variant, first, a sample is drawn that is labeled by the true target function. Then, an adversary may corrupt η-fraction of the labels arbitrarily. This variant is sometimes referred to as *nasty classification noise* [BEK02].

Note that the cost function ρ_{agnostic} only has degree 2 in binary classification settings. Theorem 3 therefore shows the equivalence between nasty classification noise and agnostic noise with no dependence on the domain size.

Subtractive contamination: In the adaptive variant of subtractive contamination, the adversary is allowed to remove $\lfloor \eta n \rfloor$ points from a size-n sample. The algorithm receives the remaining $n - \lfloor \eta n \rfloor$ points.

In the oblivious variant [DK23], the algorithm receives i.i.d. samples from the distribution $\mathcal{D}$ conditioned on some event E that occurs with probability $1 - \eta$. This can be thought of as the adversary removing η-fraction of the distribution corresponding to when the event E does not occur.

To fit subtractive contamination into our framework, we will augment the domain with a special element $\varnothing$, to indicate the adversary has removed this point. For the augmented domain $X' \coloneqq X \cup \{\varnothing\}$, it uses the cost function,

$$\rho_{\text{sub}}(x, y) \coloneqq \begin{cases} 0 & \text{if } x = y \\ \frac{1}{\eta} & \text{if } x \neq y \text{ and } y = \varnothing \\ \infty & \text{otherwise.} \end{cases}$$

Note that once again, this cost function has degree only 2, so by Theorem 3, the ρ-adaptive and ρ-oblivious adversaries are equivalent with no dependence on the domain size. In Appendix A, we give an easy reduction from the standard notions of subtractive noise (without the $\varnothing$ element added to the domain) to the adversaries defined by ρ_{sub}. This reduction, combined with Theorem 3, shows that the standard oblivious and adaptive subtractive adversaries are equivalent.

Additive contamination (Huber's model [Hub64]): In Huber's original model [Hub64], rather than directly receive i.i.d. samples from the target distribution $\mathcal{D}$, the algorithm receives i.i.d. samples from $\mathcal{D}'$, the mixture distribution

$$\mathcal{D}' \coloneqq (1 - \eta)\mathcal{D} + \eta\mathcal{E}$$

where the adversary chooses the outlier distribution $\mathcal{E}$. In the adaptive variant of this model, first a clean sample of $\lceil (1 - \eta)n \rceil$ points are drawn i.i.d. from $\mathcal{D}$. Then, the adversary may add $\lfloor \eta n \rfloor$ points arbitrarily. These n points are then randomly permuted so that the algorithm cannot trivially identify which points were added.

Similarly to subtractive contamination, we will use the augmented domain $X' \coloneqq X \cup \{\varnothing\}$. For additive noise, we use the cost function

$$\rho_{\text{add}}(x, y) \coloneqq \begin{cases} 0 & \text{if } x = y \\ \frac{1}{\eta} & \text{if } x \neq y \text{ and } x = \varnothing \\ \infty & \text{otherwise.} \end{cases}$$

In Appendix A, we give an easy reduction showing how Theorem 3 gives the equivalence between Huber's contamination model and its adaptive variant. Note that this particular equivalence was already proven by [BLMT22], but for completeness, we show their result can be recovered using our framework.

3.1 Partially-adaptive adversaries

As alluded to in Remark 1, some adversaries lie between the fully oblivious and fully adaptive adversaries. Our results show that such intermediate adversaries are equivalent to their fully oblivious and fully adaptive counterparts. The strategy for proving this equivalence is by showing the intermediate adversary is at least as strong as the oblivious adversary, and that it is no stronger than the adaptive adversary. Since Theorem 2 implies the adaptive adversary is no more powerful than the oblivious adversary, we can conclude that all three adversaries are equivalent. We formalize this approach for the two adversaries described here in Appendix B, showing both are equivalent to additive contamination.

Malicious noise: This model was first defined by [Val85]. In it, the n samples are generated sequentially. For each point, independently with probability $1-\eta$, that point is sampled from $\mathcal{D}$. Otherwise, the adversary chooses an arbitrary corrupted point with full knowledge of previous points generated but no knowledge of future points. Intuitively, this adversary is partially adaptive because when the adversary chooses how to corrupt a point, it has partial knowledge of the sample corresponding to the points generated previously.

The non-independent additive adversary: This model was recently studied in [CHL+23]. In it, the adversary generates $\lfloor \eta n \rfloor$ arbitrary points. The sample is then formed by combining $\lceil (1-\eta)n \rceil$ points drawn i.i.d. from $\mathcal{D}$ with the adversary's chosen points. Intuitively, this adversary is partially adaptive because the ηn points it generates need not be i.i.d. from some distribution as they would for fully oblivious adversaries, but the adversary still does not know the sample when choosing corruptions as in a fully adaptive adversary.

4 Technical overview

To prove Theorem 3, we begin with the observation that we can combine the algorithm A and test T into a single function $f \coloneqq T \circ A$. Therefore, it suffices to prove the following.

Theorem 6 (Theorem 3 restated)**.** *For any $n, d \in \mathbb{N}$ where $d \geq 2$, domain X, and $\varepsilon > 0$, let $m = O\left(\frac{n^4 (\ln d)^2}{\varepsilon^4}\right)$. Then, for any $f : X^n \to \{0,1\}$, cost function ρ with degree d, and distribution $\mathcal{D}$ supported on X,*

$$\left|\text{Oblivious-Max}_\rho(f, \mathcal{D}) - \text{Adaptive-Max}_\rho(f \circ \Phi_{m \to n}, \mathcal{D})\right| \leq \varepsilon. \tag{1}$$

Theorem 6 can be understood as a statement about the *indistinguishability* of the following two families of distributions, both over datasets in X^n.

1. The set of input distributions over n points the oblivious adversary can create,

$$\mathscr{D}^{n,\rho,\mathcal{D}}_{\text{oblivious}} \coloneqq \{(\mathcal{D}')^n \mid \mathcal{D}' \in \mathcal{C}_\rho(\mathcal{D})\}.$$

2. For the adaptive adversary, we first define the set of input distributions over m points before subsampling: We say $\mathcal{D}_{\text{adaptive}}$ is a valid adaptive corruption, denoted $\mathcal{D}_{\text{adaptive}} \in \mathscr{D}^{m,\rho,\mathcal{D}}_{\text{adaptive}}$ if it is possible to couple $\boldsymbol{S}' \sim \mathcal{D}_{\text{adaptive}}$ and a clean sample $\boldsymbol{S} \sim \mathcal{D}^m$ so that $\boldsymbol{S}' \in \mathcal{C}_\rho(\boldsymbol{S})$ with probability 1. Then, the distribution on n points is created via subsampling,

$$\Phi_{m\to n}\left(\mathscr{D}^{m,\rho,\mathcal{D}}_{\text{adaptive}}\right) := \left\{\text{The distribution of } \Phi_{m\to n}(\boldsymbol{S}') \mid \boldsymbol{S}' \sim \mathcal{D}_{\text{adaptive}} \text{ for any } \mathcal{D}_{\text{adaptive}} \in \mathscr{D}^{m,\rho,\mathcal{D}}_{\text{adaptive}}\right\}.$$

With these definitions, Theorem 6 can be recast as the following two statements:

1. **The oblivious adversary is no harder than the adaptive adversary:** For any distinguisher $f : X^n \to \{0,1\}$ and oblivious corruption $\mathcal{D}_{\text{oblivious}} \in \mathscr{D}^{n,\rho,\mathcal{D}}_{\text{oblivious}}$, there is a subsampled adaptive corruption $\mathcal{D}_{\text{adaptive}} \in \Phi_{m\to n}\left(\mathscr{D}^{m,\rho,\mathcal{D}}_{\text{adaptive}}\right)$ satisfying

$$\left|\mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}_{\text{oblivious}}}[f(\boldsymbol{S})] - \mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}_{\text{adaptive}}}[f(\boldsymbol{S})]\right| \le \varepsilon.$$

   This is the easy half of Theorem 6 and is already known for some specific adversary models [DKK+19, ZJS19]. We defer discussion of its proof to Section 8.

2. **The adaptive adversary is no harder than the oblivious adversary:** For any distinguisher $f : X^n \to \{0,1\}$ and subsampled adaptive corruption $\mathcal{D}_{\text{adaptive}} \in \Phi_{m\to n}\left(\mathscr{D}^{m,\rho,\mathcal{D}}_{\text{adaptive}}\right)$, there is an oblivious corruption $\mathcal{D}_{\text{oblivious}} \in \mathscr{D}^{n,\rho,\mathcal{D}}_{\text{oblivious}}$ satisfying

$$\left|\mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}_{\text{adaptive}}}[f(\boldsymbol{S})] - \mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}_{\text{oblivious}}}[f(\boldsymbol{S})]\right| \le \varepsilon. \tag{2}$$

   The remainder of this overview is devoted to our proof of this harder half of Theorem 6

4.1 A first attempt and why it fails

A natural approach towards proving this harder half of Theorem 6 is to show that every adaptive adversary can be simulated by an oblivious adversary. This corresponds to switching the order of quantifiers in the desired statement: The goal of this approach is to show that for any adaptive corruption $\mathcal{D}_{\text{adaptive}} \in \Phi_{m\to n}\left(\mathscr{D}^{m,\rho,\mathcal{D}}_{\text{adaptive}}\right)$, there is a single choice of oblivious corruption $\mathcal{D}_{\text{oblivious}} \in \mathscr{D}^{n,\rho,\mathcal{D}}_{\text{oblivious}}$ satisfying

$$d_{\text{TV}}(\mathcal{D}_{\text{adaptive}}, \mathcal{D}_{\text{oblivious}}) \le \varepsilon.$$

Indeed, as we discuss in Section 8, such an approach works for the easier half of Theorem 6. Here, we will explain why this approach fails for the harder half and latter use this counterexample to motivate our ultimately successful approach.

Our construction of this counterexample uses the adaptive and oblivious adversaries described in Example 1 with a budget $\eta = 1/2$. Recall this means that,

1. For any $S \in X^m$, the adaptive adversary can change an arbitrary $m/2$ points within S.
2. For any distribution $\mathcal{D}$, the oblivious adversary can choose any $\mathcal{D}'$ with a total variation distance of at most $1/2$ from $\mathcal{D}$.

Furthermore, we use perhaps the simplest possible base distribution, $\mathcal{D} = \text{Unif}(\{0,1\})$ and our counterexample works for *any* $m \geq n := 2$.

After receiving $\boldsymbol{S} \sim \mathcal{D}^m$, the adaptive adversary can choose a corruption so that $\boldsymbol{S}'$ either contains only zeros or only ones, with both cases equally likely. They achieve this by flipping all the 0s or all the 1s in $\boldsymbol{S}$, whichever is less frequent (breaking ties uniformly). The result of this approach is that there is some $\mathcal{D}_{\text{adaptive}} \in \mathscr{D}_{\text{adaptive}}$ for which

$$\mathcal{D}_{\text{adaptive}} = \text{Unif}([0,0],[1,1]).$$

The above distribution is far from any product distribution and, as a result, far from any possible $\mathcal{D}_{\text{oblivious}}$. Therefore, this naive simulation approach fails.

4.2 Our approach: A randomized simulation

A key observation about this counter example: Even though $\mathcal{D}_{\text{adaptive}}$ is far from any single $\mathcal{D}_{\text{oblivious}}$, it is exactly equal to a mixture of oblivious adversaries. This is because the oblivious adversary can create the point-mass distribution that always outputs $[0,0]$ and that which always outputs $[1,1]$. Our main lemma is that such a randomized simulation is always possible.

Lemma 4.1 (With subsampling, the adaptive adversary can be simulated by a randomized oblivious adversary)**.** *For any base distribution $\mathcal{D}$, sample size n, error parameter ε, and cost function ρ with degree d, set $m = O\left(\frac{n^4(\ln d)^2}{\varepsilon^4}\right)$. Then, for any subsampled adaptive corruption $\mathcal{D}_{\text{adaptive}} \in \Phi_{m\to n}\left(\mathscr{D}^{m,\rho,\mathcal{D}}_{\text{adaptive}}\right)$, there is a* randomized *oblivious corruption $\boldsymbol{\mathcal{D}}_{\text{oblivious}}$ supported on $\mathscr{D}^{n,\rho,\mathcal{D}}_{\text{oblivious}}$ such that its mixture satisfies,*

$$d_{\text{TV}}(\mathcal{D}_{\text{adaptive}}, \mathbb{E}[\boldsymbol{\mathcal{D}}_{\text{oblivious}}]) \leq \varepsilon.$$

At a high level, our approach to proving Lemma 4.1 is to group the adaptively corrupted samples $\boldsymbol{S} \sim \mathcal{D}_{\text{adaptive}}$ into a moderate number of groups that make "similar" corruptions. The goal of this grouping is that, if we look at the distribution $\boldsymbol{S}$ conditioned on falling within a single group, the resulting distribution is close to a single oblivious adversary $\mathcal{D}_{\text{oblivious}} \in \mathscr{D}^{n,\rho,\mathcal{D}}_{\text{oblivious}}$.

Our method of determining how to group a $\boldsymbol{S} \sim \mathcal{D}_{\text{adaptive}} \in \mathscr{D}^{m,\rho,\mathcal{D}}_{\text{adaptive}}$. We specify each group by a "core" $c \in X^k$ and group together all $\boldsymbol{S}$ containing this core. Note this is a soft grouping in the sense that a single $\boldsymbol{S}$ will be a member of many groups as it contains many distinct cores. This grouping strategy is summarized in the following definition.

Definition 8 (Our grouping strategy)**.** *For any distribution $\mathcal{D}_{\text{adaptive}}$ over samples in X^m and parameters $n + k \leq m$, let* $\text{Grouped}_{n,k}(\mathcal{D}_{\text{adaptive}})$ *be the joint distribution over $(\boldsymbol{S}, \boldsymbol{c})$ formed by*

1. *Drawing a $\boldsymbol{S}_{\text{big}} \sim \mathcal{D}_{\text{adaptive}}$.*
2. *Drawing $\boldsymbol{S}$ and $\boldsymbol{c}$ to be uniform size-n and size-k respectively disjoint subsamples of $\boldsymbol{S}_{\text{big}}$, meaning $\boldsymbol{S} \sim \Phi_{m\to n}(\boldsymbol{S}_{\text{big}})$ and $\boldsymbol{c} \sim \Phi_{(m-n)\to k}(\boldsymbol{S}_{\text{big}} \setminus \boldsymbol{S})$.*

It will also be helpful to define $\text{Group}_n(\mathcal{D}_{\text{adaptive}}, c)$ *to be the distribution of $\boldsymbol{S}$ conditioned on $\boldsymbol{c} = c$.*

To prove Lemma 4.1, we will show that in expectation over the core $\boldsymbol{c}$, there is some $\boldsymbol{\mathcal{D}}_{\text{oblivious}}$ that is within ε-total variation distance to the distribution of $\text{Group}_n(\mathcal{D}_{\text{adaptive}}, \boldsymbol{c})$. We analyze this total variation distance in two steps which are described in the following two subsections.

4.3 Bounding the distance from a product distribution

Since every distribution $\mathcal{D}_{\text{oblivious}} \in \mathscr{D}^{n,\rho,\mathcal{D}}_{\text{oblivious}}$ is a product distribution, the first step in our analysis is to show that the average group is close to some product distribution (though not necessarily one that is a valid oblivious corruption).

Lemma 4.2 (The average group is close to a product distribution)**.** *For any $\mathcal{D}_{\text{adaptive}} \in \mathscr{D}^{m,\rho,\mathcal{D}}_{\text{adaptive}}$ where ρ is a degree-d cost function, and any $n + k_{\max} \le m/2$, there exists some $k \le k_{\max}$ for which*

$$\mathop{\mathbb{E}}_{\boldsymbol{S},\boldsymbol{c}\sim\text{Grouped}_{n,k}(\mathcal{D}_{\text{adaptive}})}[d_{\text{TV}}(\text{Group}_n(\mathcal{D}_{\text{adaptive}},\boldsymbol{c}),\mathcal{D}_{\text{goal}}(\boldsymbol{c})^n)] \le \sqrt{\frac{n^2 \ln d}{2k_{\max}}}.$$

where $\mathcal{D}_{\text{goal}}(c) \coloneqq \mathbb{E}_{\boldsymbol{S}\sim\text{Group}_n(\mathcal{D}_{\text{adaptive}},c)}[\text{Unif}(\boldsymbol{S})]$ is the average data-point in the group corresponding to the core c.

Lemma 4.2 is closely related to the *correlation rounding* technique used to round semidefinite programs (see e.g. [BRS11, RT12]), also called the *pinning lemma* in the statistical physics community (see e.g. [Eld20]). Roughly speaking, these results say the following: For any (not necessarily independent) random variables $\boldsymbol{x}_1, \dots, \boldsymbol{x}_m$, by conditioning on a few of the the $\boldsymbol{x}_i$, we can make the covariances between the remaining pairs close to independent. The version of this result we will need extends the concept of covariance from pairs to larger groups:

Definition 9 (Multivariate total correlation)**.** *The multivariate total correlation of random variables $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ is*

$$\text{Cor}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) \coloneqq d_{\text{KL}}\left(\mathcal{D} \,\|\, \mathcal{D}_1 \times \cdots \times \mathcal{D}_n\right)$$

where $\mathcal{D}$ is the distribution of $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$ and $\mathcal{D}_i$ is the marginal distribution of $\boldsymbol{x}_i$. Similarly, for any $\boldsymbol{y}$, we define the conditional multivariate correlation as

$$\text{Cor}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \mid \boldsymbol{y}) = \mathop{\mathbb{E}}_{\boldsymbol{y}}[\text{Cor}(\boldsymbol{x}_1 \mid \boldsymbol{y}, \dots, \boldsymbol{x}_n \mid \boldsymbol{y})].$$

We use the following to prove Lemma 4.2, where $I(\boldsymbol{x};\boldsymbol{y})$ denotes the mutual information between $\boldsymbol{x}$ and $\boldsymbol{y}$.

Lemma 4.3 (Correlation rounding)**.** *For any random variable on $\boldsymbol{S}$ on X^m and integers $n + k_{\max} \le m$, there exists some $k \le k_{\max}$ for which,*

$$\mathop{\mathbb{E}}_{\substack{\boldsymbol{A}\sim\binom{[m]}{n}\\ \boldsymbol{B}\sim\binom{[m]\setminus\boldsymbol{A}}{k}}}[\text{Cor}(\boldsymbol{S}_{\boldsymbol{A}} \mid \boldsymbol{S}_{\boldsymbol{B}})] \le \frac{n(n-1)}{2(k_{\max}+1)} \cdot \mathop{\mathbb{E}}_{\substack{\boldsymbol{i}\sim\text{Unif}([m])\\ \boldsymbol{B}\sim\binom{[m]\setminus\{\boldsymbol{i}\}}{n+k_{\max}-1}}}[I(\boldsymbol{S}_{\boldsymbol{i}};\boldsymbol{S}_{\boldsymbol{B}})].$$

As far as we are aware, the closest results to Lemma 4.3 already appearing in the literature are those of [MR17, JKR19]. They prove essentially the same result except the term $\mathbb{E}[I(\boldsymbol{S}_{\boldsymbol{i}};\boldsymbol{S}_{\boldsymbol{B}})]$ is replaced with $\mathbb{E}[H(\boldsymbol{S}_{\boldsymbol{i}})]$. Since entropy upper bounds mutual information, our result is always at least as strong as theirs.

The distinction between $\mathbb{E}[I(\boldsymbol{S}_{\boldsymbol{i}};\boldsymbol{S}_{\boldsymbol{B}})]$ and $\mathbb{E}[H(\boldsymbol{S}_{\boldsymbol{i}})]$ ends up being crucial for our application. In Lemma 6.5, we are able to show that if $\boldsymbol{S}$ is the result of an adaptive corruption with a degree-d cost function, that $\mathbb{E}[I(\boldsymbol{S}_{\boldsymbol{i}};\boldsymbol{S}_{\boldsymbol{B}})] \le O(\ln d)$. In contrast, even when the adversary makes no corruptions, if the base (uncorrupted) distribution has high entropy, then $\mathbb{E}[H(\boldsymbol{S}_{\boldsymbol{i}})]$ can be as large as $\ln|X|$. We furthermore view our version as having a natural interpretation: The term $\mathbb{E}[I(\boldsymbol{S}_{\boldsymbol{i}};\boldsymbol{S}_{\boldsymbol{B}})]$ measures some notion of how far $\boldsymbol{S}_1, \dots, \boldsymbol{S}_m$ are from being independent. Thus, our result gives a correlation bound that improves when the starting distribution was already close to a product distribution.

4.4 Rounding to the nearest oblivious corruption

While Lemma 4.2 guarantees that most groups are close to some product distribution, we require that product distribution to be a valid oblivious corruption. We show that this product distribution, on average, can be rounded to a nearby valid oblivious corruption.

Lemma 4.4 (Error due to rounding for our grouping strategy)**.** *Using the same notation as Lemma 4.2, as long as $k \le m/2$*

$$\mathop{\mathbb{E}}_{\boldsymbol{c}}\left[\inf_{\mathcal{D}_{\text{oblivious}} \in \mathscr{D}^{n,\rho,\mathcal{D}}_{\text{oblivious}}} \{d_{\text{TV}}(\mathcal{D}_{\text{oblivious}}, \mathcal{D}_{\text{goal}}(\boldsymbol{c})^n)\}\right] \le 2n \cdot \sqrt{\frac{k \ln d}{m}}.$$

To prove Lemma 4.4, we first show a more general but weaker result that holds for *any* grouping strategy. This result, formalized in Lemma 7.1, gives a bound that scales with the number of groups. Since the number of groups we use is $|X|^k$, a direct application of Lemma 7.1 would require the parameter m to scale with the domain size. Nonetheless, we show a more delicate application of Lemma 7.1 allows for us to get a bound roughly as good as if there were only d^k groups.

Setting parameters. We briefly sketch how to recover Lemma 4.1 using Lemmas 4.2 and 4.4. For any adaptive strategy $\mathcal{D}_{\text{adaptive}}$, Lemma 4.2 gives there is core size $k \le k_{\max} = O(n^2 \ln(d)/\varepsilon^2)$ for which

$$d_{\text{TV}}\left(\mathcal{D}_{\text{adaptive}}, \mathop{\mathbb{E}}_{\boldsymbol{c}}[\mathcal{D}_{\text{goal}}(\boldsymbol{c})^n]\right) \le \varepsilon/2.$$

We will then apply Lemma 4.4. Setting $m = O(n^2 k(\ln d)/\varepsilon^2) = O(n^4(\ln d)^2/\varepsilon^4)$, we have that for each core c, there is some $\mathcal{D}_{\text{oblivious}}(c) \in \mathscr{D}^{n,\rho,\mathcal{D}}_{\text{oblivious}}$ for which

$$\mathop{\mathbb{E}}_{\boldsymbol{c}}[d_{\text{TV}}(\mathcal{D}_{\text{oblivious}}(\boldsymbol{c}), \mathcal{D}_{\text{goal}}(\boldsymbol{c})^n] \le \varepsilon/2.$$

Combining the above two bounds recovers Lemma 4.1.

5 Preliminaries

Indexing. For any $n \in \mathbb{N}$, we use $[n]$ as shorthand for $\{1, 2, \ldots, n\}$. Similarly, for $n \le m \in \mathbb{N}$, we use $[n, m]$ as shorthand for $\{n, n+1, \ldots, m\}$. For any multiset $S \in X^m$, we use S_i to denote the i^{th} element of S. For any $I \subseteq [m]$, we use S_I to denote the multiset containing $(S_{I_1}, S_{I_2}, \ldots)$. We'll also use $S_{<j}$ and S_{-j} as shorthand for $S_{[j-1]}$ and $S_{[m]\setminus\{j\}}$ respectively. For any permutation $\sigma : [m] \to [m]$ and $S \in X^m$, we'll use $\sigma(S)$ as shorthand for the multiset in X^m satisfying $\sigma(S)_i = S_{\sigma(i)}$.

Random variables and distributions. We use **boldfont** to denote random variables and calligraphic font to denote distributions (e.g. $\boldsymbol{x} \sim \mathcal{D}$). For a multiset S, we use $\text{Unif}(S)$ to denote the uniform distribution over elements of S. For a distribution $\mathcal{D}$, we will use $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \overset{\text{iid}}{\sim} \mathcal{D}$ and $\boldsymbol{x} \sim \mathcal{D}^n$ interchangeably to denote that $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are independent and identically distributed according to $\mathcal{D}$. For any distributions $\mathcal{D}_1, \mathcal{D}_2$, we use $\mathcal{D}_1 \times \mathcal{D}_2$ to denote the product distribution of $\mathcal{D}_1$ and $\mathcal{D}_2$. We denote mixture distributions as convex combinations (e.g. $\mathcal{D}_{\text{mix}} = 1/3 \cdot \mathcal{D}_1 + 2/3 \cdot \mathcal{D}_2$).

We use the following standard concentration inequality.

Fact 5.1 (Chernoff bound)**.** *Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be independent random variables on $\{0,1\}$, and $\boldsymbol{X}$ their sum. For $\mu \coloneqq \mathbb{E}[\boldsymbol{X}]$,*

$$\Pr[\boldsymbol{X} \geq 2\mu] \leq e^{-\mu/3} \qquad \text{and} \qquad \Pr[\boldsymbol{X} \leq \mu/2] \leq e^{-\mu/8}.$$

We will also use two commonly studied families of random variables. For any $p \in [0,1]$, we use $\mathrm{Ber}(p)$ to denote the distribution that takes on value 1 with probability p and takes on 0 otherwise. Furthermore, for any $n \in \mathbb{N}$, we use $\mathrm{Bin}(n,p)$ to denote the sum of n independent random variables each distributed according to $\mathrm{Ber}(p)$.

Formalizing the corruption models. We recap the notation used to formalize our corruption models. Beginning with the adaptive for any cost function $\rho : X \times X \to \mathbb{R}_{\geq 0} \cup \infty$ and sample $S \in X^m$, we use $\mathcal{C}_\rho(S)$ to denote legal adaptive corruptions of S under cost function ρ,

$$\mathcal{C}_\rho(S) \coloneqq \left\{ S' \in X^m \mid \frac{1}{m} \sum_{i \in [m]} \rho(S_i, S'_i) \leq 1 \right\}.$$

For a base distribution $\mathcal{D}$, the set of input distributions on size-m data sets the adaptive adversary can create is denoted:

$$\mathscr{D}^{m,\rho,\mathcal{D}}_{\text{adaptive}} \coloneqq \big\{\text{Distributions } \mathcal{D}' \text{ over } X^m \mid \text{Can couple } \boldsymbol{S}' \sim \mathcal{D}' \text{ and } \boldsymbol{S} \sim \mathcal{D}^m \text{ so } \boldsymbol{S}' \in \mathcal{C}_\rho(\boldsymbol{S}) \text{ w.p. } 1\big\}.$$

For the oblivious adversary, we overload $\mathcal{C}_\rho(\mathcal{D})$ to denote all distributions the oblivious adversary can create,

$$\mathcal{C}_\rho(\mathcal{D}) \coloneqq \big\{\text{Distributions } \mathcal{D}' \text{ over } X \mid \text{Can couple } \boldsymbol{x}' \sim \mathcal{D}' \text{ and } \boldsymbol{x} \sim \mathcal{D} \text{ so } \mathbb{E}[\rho(\boldsymbol{x}, \boldsymbol{x}') \leq 1]\big\}.$$

The set of input distributions on size-n datasets the oblivious adversary can create is denoted:

$$\mathscr{D}^{n,\rho,\mathcal{D}}_{\text{oblivious}} \coloneqq \{(\mathcal{D}')^n \mid \mathcal{D}' \in \mathcal{C}_\rho(\mathcal{D})\}.$$

Subsampling filter. Recall, in Definition 5, we defined $\Phi_{m \to n} : X^m \to X^n$ to be the (randomized) algorithm that that given $S \in X^m$ returns a sample of n points drawn uniformly without replacement from S. We will generalize this in three ways: First, we'll use $\Phi_{\star \to n}$ the filter that takes in a sample $S \in X^\star$ of at least n points and then subsamples it down to n points. Second, if $\mathcal{D}$ is a distribution over $\boldsymbol{S} \in X^m$, we'll use $\Phi_{m \to n}(\mathcal{D})$ to denote the distribution of $\Phi_{m \to n}(\boldsymbol{S})$. Lastly, if $\mathscr{D}$ is a family of distributions, we'll use $\Phi_{m \to n}(\mathcal{D})$ to denote $\{\Phi_{m \to n}(\mathcal{D}) \mid \mathcal{D} \in \mathscr{D}\}$.

TV distance and KL divergence. We use two measures of statistical distance/divergence.

Definition 10 (Total variation distance)**.** *Let $\mathcal{D}$ and $\mathcal{D}'$ be any two distributions over the same domain X. The* total variation distance *between $\mathcal{D}$ and $\mathcal{D}'$, is defined as*

$$d_{\mathrm{TV}}(\mathcal{D}, \mathcal{D}') \coloneqq \sup_{T : X \to [0,1]} \left\{ \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{D}}[T(\boldsymbol{x})] - \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{D}'}[T(\boldsymbol{x})] \right\}.$$

This quantity can be equivalently defined as the infimum over all couplings of $\boldsymbol{x} \sim \mathcal{D}$ and $\boldsymbol{x}' \sim \mathcal{D}'$ of $\Pr[\boldsymbol{x} \neq \boldsymbol{x}']$.

TV distance is a true distance in the sense that it satisfies the triangle inequality.

Fact 5.2 (The triangle inequality for TV distance)**.** *For any distributions $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$,*

$$d_{\mathrm{TV}}(\mathcal{D}_1, \mathcal{D}_3) \leq d_{\mathrm{TV}}(\mathcal{D}_1, \mathcal{D}_2) + d_{\mathrm{TV}}(\mathcal{D}_2, \mathcal{D}_3).$$

We can also give a (sometimes coarse) upper bound on the TV distance of product distribution.

Fact 5.3 (Total variation distance of a product)**.** *For any distributions $\mathcal{D}_1, \mathcal{D}_2$ and $n \in \mathbb{N}$,*

$$d_{\mathrm{TV}}(\mathcal{D}_1^n, \mathcal{D}_2^n) \leq n \cdot d_{\mathrm{TV}}(\mathcal{D}_1, \mathcal{D}_2).$$

Straight from the definition, we see that TV distance is convex in one argument.

Fact 5.4 (Convexity of TV distance)**.** *For any distributions $\mathcal{D}_1, \mathcal{D}_2, \mathcal{E}_1$, and $\mathcal{E}_2$, and mixture weight $\lambda \in [0, 1]$,*

$$d_{\mathrm{TV}}(\mathcal{D}_\lambda, \mathcal{E}_\lambda) \leq \lambda \cdot d_{\mathrm{TV}}(\mathcal{D}_1, \mathcal{E}_1) + (1 - \lambda) \cdot d_{\mathrm{TV}}(\mathcal{D}_2, \mathcal{E}_2),$$

where $\mathcal{D}_\lambda$ is the mixture $\lambda\mathcal{D}_1 + (1 - \lambda)\mathcal{D}_2$ and similarly $\mathcal{E}_\lambda \coloneqq \lambda\mathcal{E}_1 + (1 - \lambda)\mathcal{E}_2$.

The other measure of statistical distance/divergence that plays a key role in our results in *KL* divergence.

Definition 11 (Kullback-Leibler (KL) Divergence)**.** *For distributions $\mathcal{D}, \mathcal{E}$ supported on the same domain X, the* KL divergence *between $\mathcal{D}$ and $\mathcal{E}$ as defined as,*

$$d_{\mathrm{KL}}(\mathcal{D} \,\|\, \mathcal{E}) \coloneqq \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{D}} \left[\ln \left(\frac{\mathcal{D}(\boldsymbol{x})}{\mathcal{E}(\boldsymbol{x})} \right) \right],$$

where $\mathcal{D}(x)$ and $\mathcal{E}(x)$ denote the probability mass or density functions of $\mathcal{D}$ and $\mathcal{E}$ respectively at the point x (or more generally, $\mathcal{D}(x)/\mathcal{E}(x)$ is the Radon-Nikodym derivative of $\mathcal{D}$ with respect to $\mathcal{E}$).

Unlike TV distance, KL divergence is not a true distance in the sense that it does not satisfy triangle inequality. For us, it will suffices that it upper bounds TV distance via Pinsker's inequality [Pin64]. [Can22] has a nice summary of different forms and proofs of the below inequality and appropriate references for each.

Fact 5.5 (Pinsker's inequality [Pin64, Can22])**.** *For any distributions $\mathcal{D}, \mathcal{E}$,*

$$d_{\mathrm{TV}}(\mathcal{D}, \mathcal{E}) \leq \sqrt{\frac{d_{\mathrm{KL}}(\mathcal{D} \,\|\, \mathcal{E})}{2}}.$$

Mutual information.

Definition 12 (Mutual information)**.** *For random variables $\boldsymbol{x}, \boldsymbol{y}$ jointly distribution according to a distribution $\mathcal{D}$, let $\mathcal{D}_x$ and $\mathcal{D}_y$ be the marginal distributions of $\boldsymbol{x}$ and $\boldsymbol{y}$ respectively, and $\mathcal{D}_{x|y}$ be the marginal distribution of $\boldsymbol{x}$ conditioned on $\boldsymbol{y} = y$. The mutual information between $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined as*

$$I(\boldsymbol{x}; \boldsymbol{y}) = d_{\mathrm{KL}}(\mathcal{D} \,\|\, \mathcal{D}_x \times \mathcal{D}_y) = \mathop{\mathbb{E}}_{\boldsymbol{y}} \left[d_{\mathrm{KL}} \left(\mathcal{D}_{x|\boldsymbol{y}} \,\middle\|\, \mathcal{D}_x \right) \right]$$

Definition 13 (Conditional mutual information)**.** *For random variables $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$ jointly distributed, the mutual information of $\boldsymbol{x}$ and $\boldsymbol{y}$ conditioned on $\boldsymbol{z}$ is*

$$I(\boldsymbol{x}; \boldsymbol{y} \mid \boldsymbol{z}) := \mathop{\mathbb{E}}_{\boldsymbol{z}' \sim \mathcal{D}_z} \left[I((\boldsymbol{x} \mid \boldsymbol{z} = \boldsymbol{z}'); (\boldsymbol{y} \mid \boldsymbol{z} = \boldsymbol{z}')) \right]$$

where $\mathcal{D}_z$ is the marginal distribution of $\boldsymbol{z}$.

The *chain rule* connects mutual information and conditional mutual information.

Fact 5.6 (Chain rule for mutual information)**.** *For any $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$,*

$$I(\boldsymbol{x}; (\boldsymbol{y}, \boldsymbol{z})) = I(\boldsymbol{x}; \boldsymbol{z}) + I(\boldsymbol{x}; \boldsymbol{y} \mid \boldsymbol{z}).$$

This is sometimes rewritten as

$$I(\boldsymbol{x}; \boldsymbol{y} \mid \boldsymbol{z}) = I(\boldsymbol{x}; (\boldsymbol{y}, \boldsymbol{z})) - I(\boldsymbol{x}; \boldsymbol{z}).$$

Mutual information is always nonnegative

Fact 5.7 (Nonnegativity of mutual information)**.** *For any random variables $\boldsymbol{x}, \boldsymbol{y}$,*

$$I(\boldsymbol{x}; \boldsymbol{y}) \geq 0.$$

As an easy consequence of the chain rule and mutual information being nonnegative, we have that mutual information can only increase if we consider more information.

Fact 5.8. *For any random variables $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$,*

$$I(\boldsymbol{x}; (\boldsymbol{y}, \boldsymbol{z})) \geq I(\boldsymbol{x}; \boldsymbol{y}).$$

Mutual information is also symmetric.

Fact 5.9 (Symmetry of mutual information)**.** *For any random variables $\boldsymbol{x}, \boldsymbol{y}$,*

$$I(\boldsymbol{x}; \boldsymbol{y}) = I(\boldsymbol{y}; \boldsymbol{x}).$$

Another nice property of mutual information is that it is bounded by the support size of each variable.

Fact 5.10 (Mutual information with a finite support)**.** *For any random variables $\boldsymbol{x}, \boldsymbol{y}$, if one of $\boldsymbol{x}$ or $\boldsymbol{y}$ has a finite support of size d, then,*

$$I(\boldsymbol{x}; \boldsymbol{y}) \leq \ln d.$$

6 Bounding the distance from a product distribution: Proof of Lemma 4.2

In this section, we prove the following, restated for convenience.

Lemma 4.2 (The average group is close to a product distribution)**.** *For any $\mathcal{D}_{\text{adaptive}} \in \mathscr{D}^{m,\rho,\mathcal{D}}_{\text{adaptive}}$ where ρ is a degree-d cost function, and any $n + k_{\max} \le m/2$, there exists some $k \le k_{\max}$ for which*

$$\mathop{\mathbb{E}}_{\boldsymbol{S},\boldsymbol{c}\sim\text{Grouped}_{n,k}(\mathcal{D}_{\text{adaptive}})}[d_{\text{TV}}(\text{Group}_n(\mathcal{D}_{\text{adaptive}},\boldsymbol{c}),\mathcal{D}_{\text{goal}}(\boldsymbol{c})^n)] \le \sqrt{\frac{n^2 \ln d}{2k_{\max}}}.$$

where $\mathcal{D}_{\text{goal}}(c) \coloneqq \mathbb{E}_{\boldsymbol{S}\sim\text{Group}_n(\mathcal{D}_{\text{adaptive}},c)}[\text{Unif}(\boldsymbol{S})]$ is the average data-point in the group corresponding to the core c.

We structure this proof into three steps:

1. In Section 6.1, we prove our version of correlation rounding. As discussed in Section 4.3, this is improvement of similar results from [MR17, JKR19] that replaces an entropy term with a mutual information term.
2. In Section 6.2, we prove that this mutual information term is small when the adaptive adversary has low degree.
3. In Section 6.3, we combine these two results to prove Lemma 4.2

6.1 Proof of Lemma 4.3

In this proof, we will often be reasoning about mutual information (Definition 12) and multivariate total correlation (Definition 9) of subsets of the random variable $\boldsymbol{S}$ supported on X^m. It will be convenient to have the following concise notation: For any $a + b + c \le m$, we define,

$$\begin{aligned}
\text{Cor}_{\boldsymbol{S}}(a) &\coloneqq \mathop{\mathbb{E}}_{\boldsymbol{A}\sim\binom{[m]}{a}}[\text{Cor}(\boldsymbol{S}_{\boldsymbol{A}})],\\
\text{Cor}_{\boldsymbol{S}}(a \mid b) &\coloneqq \mathop{\mathbb{E}}_{\boldsymbol{A}\sim\binom{[m]}{a},\boldsymbol{B}\sim\binom{[m]\setminus\boldsymbol{A}}{b}}[\text{Cor}(\boldsymbol{S}_{\boldsymbol{A}} \mid \boldsymbol{S}_{\boldsymbol{B}})],\\
I_{\boldsymbol{S}}(a;b) &\coloneqq \mathop{\mathbb{E}}_{\boldsymbol{A}\sim\binom{[m]}{a},\boldsymbol{B}\sim\binom{[m]\setminus\boldsymbol{A}}{b}}[I(\boldsymbol{S}_{\boldsymbol{A}};\boldsymbol{S}_{\boldsymbol{B}})],\\
I_{\boldsymbol{S}}(a;b \mid c) &\coloneqq \mathop{\mathbb{E}}_{\boldsymbol{A}\sim\binom{[m]}{a},\boldsymbol{B}\sim\binom{[m]\setminus\boldsymbol{A}}{b},\boldsymbol{C}\sim\binom{[m]\setminus(\boldsymbol{A}\cup\boldsymbol{B})}{c}}[I(\boldsymbol{S}_{\boldsymbol{A}};\boldsymbol{S}_{\boldsymbol{B}} \mid \boldsymbol{S}_{\boldsymbol{C}})].
\end{aligned}$$

With this notation, we can succinctly restated Lemma 4.3.

Lemma 6.1 (Restatement of Lemma 4.3)**.** *For any random variable on $\boldsymbol{S}$ on X^m and integers $n + k_{\max} \le m$, there exists some $k \le k_{\max}$ for which,*

$$\text{Cor}_{\boldsymbol{S}}(n \mid k) \le \frac{n(n-1)}{2(k_{\max}+1)} \cdot I_{\boldsymbol{S}}(1; n + k_{\max} - 1).$$

We assemble the ingredients used in the proof of Lemma 6.1. First, is a simple application of the chain rule.

Proposition 6.2. *For any $a + b + c \le m$ and $\boldsymbol{S}$ on X^m,*

$$I_{\boldsymbol{S}}(a; b \mid c) = I_{\boldsymbol{S}}(a; b + c) - I_{\boldsymbol{S}}(a; c).$$

Proof. For any disjoint $A, B, C \subseteq [m]$, we apply Fact 5.6 which gives that

$$I(\boldsymbol{S}_A; \boldsymbol{S}_B \mid \boldsymbol{S}_C) = I(\boldsymbol{S}_A; \boldsymbol{S}_{B \cup C}) - I(\boldsymbol{S}_A; \boldsymbol{S}_C).$$

Averaging over $\boldsymbol{A}, \boldsymbol{B}$, and $\boldsymbol{C}$ gives the desired result. □

Second, we show the following.

Proposition 6.3. *For any random variable $\boldsymbol{S}$ supported on X^m and $a + b \leq m$,*

$$I_{\boldsymbol{S}}(a; b) \leq a \cdot I_{\boldsymbol{S}}(1, a + b - 1)$$

Proof. It suffices to show, for any random variables $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_a$ and $\boldsymbol{y}$, that

$$I(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_a; \boldsymbol{y}) \leq \sum_{i \in [a]} I(\boldsymbol{x}_i; \boldsymbol{y}, \boldsymbol{x}_{\neq i}),$$

as the desired result then follows by averaging over all $\boldsymbol{A}, \boldsymbol{B}$ and setting $\boldsymbol{x} = \boldsymbol{S}_{\boldsymbol{A}}$ and $\boldsymbol{y} = \boldsymbol{S}_{\boldsymbol{B}}$. We bound,

$$\begin{aligned}
I(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_a; \boldsymbol{y}) &= \sum_{i \in [a]} I(\boldsymbol{x}_i; \boldsymbol{y} \mid \boldsymbol{x}_{<i}) && \text{(Fact 5.6)} \\
&= \sum_{i \in [a]} I(\boldsymbol{x}_i; \boldsymbol{y}, \boldsymbol{x}_{<i}) - I(\boldsymbol{x}_i; \boldsymbol{x}_{<i}) && \text{(Fact 5.6 again)} \\
&\leq \sum_{i \in [a]} I(\boldsymbol{x}_i; \boldsymbol{y}, \boldsymbol{x}_{<i}) && \text{(Fact 5.7)} \\
&\leq \sum_{i \in [a]} I(\boldsymbol{x}_i; \boldsymbol{y}, \boldsymbol{x}_{<i}, \boldsymbol{x}_{>i}), && \text{(Fact 5.8)}
\end{aligned}$$

which is exactly the desired bound. □

Third, we give an alternative form of multivariate total correlation.

Proposition 6.4 (Multivariate total correlation in terms of mutual information)**.** *For any random variables $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$,*

$$\mathrm{Cor}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \sum_{i \in [n-1]} I(\boldsymbol{x}_{\leq i}; \boldsymbol{x}_{i+1}).$$

Proof. Throughout this proof, we use $\mathcal{D}$ to denote the distribution of $\boldsymbol{x}$, $\mathcal{D}_i$ to denote the marginal distribution of $\boldsymbol{x}_i$, and $\mathcal{D}_{\leq i}$ to denote the marginal distribution of $\boldsymbol{x}_{\leq i}$. Expanding the right-hand side,

$$\begin{aligned}
\sum_{i \in [n-1]} I(\boldsymbol{x}_{\leq i}; \boldsymbol{x}_{i+1}) &= \sum_{i \in [n-1]} d_{\mathrm{KL}}\left(\mathcal{D}_{\leq i+1} \,\|\, \mathcal{D}_{\leq i} \times \mathcal{D}_{i+1}\right) && \text{(Definition of mutual information)} \\
&= \sum_{i \in [n-1]} \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{D}_{\leq i+1}} \left[\ln\left(\frac{\mathcal{D}_{\leq i+1}(\boldsymbol{x})}{\mathcal{D}_{\leq i}(\boldsymbol{x}_{\leq i})\mathcal{D}_{i+1}(\boldsymbol{x}_{i+1})}\right)\right] && \text{(Definition of KL divergence)} \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{D}} \left[\ln\left(\prod_{i=1}^{n-1} \frac{\mathcal{D}_{\leq i+1}(\boldsymbol{x}_{\leq i+1})}{\mathcal{D}_{\leq i}(\boldsymbol{x}_{\leq i})\mathcal{D}_{i+1}(\boldsymbol{x}_{i+1})}\right)\right] && \text{(Linearity of expectation)} \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{D}} \left[\ln\left(\frac{\mathcal{D}(\boldsymbol{x})}{\prod_{i=1}^{n} \mathcal{D}_i(\boldsymbol{x}_i)}\right)\right] && \text{(Cancellation of terms)} \\
&= d_{\mathrm{KL}}\left(\mathcal{D} \,\|\, \mathcal{D}_1 \times \mathcal{D}_2 \times \cdots \mathcal{D}_n\right)
\end{aligned}$$

which is exactly $\mathrm{Cor}(\boldsymbol{x})$. □

We are now ready to prove the main result of this subsection.

Proof of Lemma 6.1. We wish to show there is some $k \le k_{\max}$ for which $\mathrm{Cor}_{\boldsymbol{S}}(n \mid k)$ is small. For any such k, we have that

$$\begin{aligned}\mathrm{Cor}_{\boldsymbol{S}}(n \mid k) &= \sum_{i\in[n-1]} I_{\boldsymbol{S}}(i; 1 \mid k) && \text{(Proposition 6.4)}\\ &= \sum_{i\in[n-1]} I_{\boldsymbol{S}}(i; k+1) - I_{\boldsymbol{S}}(i; k) && \text{(Proposition 6.2.)}\end{aligned}$$

Summing up the above for all $k = 0, \dots, k_{\max}$, we obtain

$$\begin{aligned}\sum_{k\in[0,k_{\max}]} \mathrm{Cor}_{\boldsymbol{S}}(n \mid k) &= \sum_{k\in[0,k_{\max}]} \sum_{i\in[n-1]} I_{\boldsymbol{S}}(i; k+1) - I_{\boldsymbol{S}}(i; k)\\ &= \sum_{i\in[n-1]} I_{\boldsymbol{S}}(i, k_{\max}+1) && \text{(Cancel telescoping terms and } I_{\boldsymbol{S}}(i,0)=0)\\ &\le \sum_{i\in[n-1]} i \cdot I_{\boldsymbol{S}}(1, i+k_{\max}) && \text{(Proposition 6.3)}\\ &\le \sum_{i\in[n-1]} i \cdot I_{\boldsymbol{S}}(1, n-1+k_{\max}) && \text{(Fact 5.8)}\\ &= \frac{n(n-1)}{2} \cdot I_{\boldsymbol{S}}(1, n-1+k_{\max})\end{aligned}$$

Therefore, using the fact that minimum over all $k \in [0, k_{\max}]$ is at most the mean, there exists one choice of k for which

$$\mathrm{Cor}_{\boldsymbol{S}}(n \mid k) \le \frac{n(n-1)}{2(k_{\max}+1)} \cdot I_{\boldsymbol{S}}(1, n-1+k_{\max}).$$ □

6.2 Bounding the mutual information for low-degree cost functions

The main result of this subsection is the following.

Lemma 6.5 (Bounding mutual information for low-degree corruptions)**.** *For any $\boldsymbol{S}' \sim \mathcal{D}_{\text{adaptive}} \in \mathscr{D}^{m,\rho,\mathcal{D}}_{\text{adaptive}}$ where ρ is a degree-d cost function and $r < m$,*

$$\mathop{\mathbb{E}}_{\boldsymbol{i}\sim\mathrm{Unif}([m]),\boldsymbol{B}\sim\binom{[m]\setminus\{\boldsymbol{i}\}}{r}} [I(\boldsymbol{S}'_{\boldsymbol{i}}; \boldsymbol{S}'_{\boldsymbol{B}})] \le \frac{m}{m-r} \cdot \ln d.$$

We will use the following.

Proposition 6.6. *Let $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ be independent random variables and $\boldsymbol{y}$ be any (not necessarily independent) random variable. Then,*

$$\sum_{i\in[n]} I(\boldsymbol{x}_i; \boldsymbol{y}) \le I((\boldsymbol{x}_1, \dots, \boldsymbol{x}_n); \boldsymbol{y}).$$

Proof. We bound,

$$\begin{aligned} I((\boldsymbol{x}_1,\dots,\boldsymbol{x}_n);\boldsymbol{y}) &= \sum_{i\in[n]} I(\boldsymbol{x}_i;\boldsymbol{y}\mid \boldsymbol{x}_{<i}) && \text{(Fact 5.6)}\\ &= \sum_{i\in[n]} I(\boldsymbol{x}_i;(\boldsymbol{y},\boldsymbol{x}_{<i})) - I(\boldsymbol{x}_i;\boldsymbol{x}_{<i}) && \text{(Fact 5.6 again)}\\ &= \sum_{i\in[n]} I(\boldsymbol{x}_i;(\boldsymbol{y},\boldsymbol{x}_{<i})) && (\boldsymbol{x}_i \text{ and } \boldsymbol{x}_{<i} \text{ are independent})\\ &\geq \sum_{i\in[n]} I(\boldsymbol{x}_i;\boldsymbol{y}) && \text{(Fact 5.8)} \end{aligned}$$

□

Proof of Lemma 6.5. Since $\mathcal{D}_{\text{adaptive}}$ is a legal adaptive corruption, there is a coupling of $\boldsymbol{S}\sim\mathcal{D}^m$ and $\boldsymbol{S}'$ for which $\boldsymbol{S}'\in\mathcal{C}_\rho(\boldsymbol{S})$ with probability 1. In particular, this implies that that once we condition on $\boldsymbol{S}_i$, there are at most d choices for $\boldsymbol{S}'_i$.

For any fixed choice of $\boldsymbol{i}=i,\boldsymbol{B}=B$, we have that

$$\begin{aligned} I(\boldsymbol{S}'_i;\boldsymbol{S}'_B) &\leq I\big((\boldsymbol{S}_i,\boldsymbol{S}'_i);(\boldsymbol{S}_B,\boldsymbol{S}'_B)\big) && \text{(Fact 5.8)}\\ &= I\big(\boldsymbol{S}_i;(\boldsymbol{S}_B,\boldsymbol{S}'_B)\big) + I\big(\boldsymbol{S}'_i;(\boldsymbol{S}_B,\boldsymbol{S}'_B)\mid \boldsymbol{S}_i\big) && \text{(Fact 5.6)}\\ &= I(\boldsymbol{S}_i;\boldsymbol{S}_B) + I\big(\boldsymbol{S}_i;\boldsymbol{S}'_B\mid \boldsymbol{S}_B\big) + I\big(\boldsymbol{S}'_i;(\boldsymbol{S}_B,\boldsymbol{S}'_B)\mid \boldsymbol{S}_i\big) && \text{(Fact 5.6 again.)} \end{aligned}$$

The first term, $I(\boldsymbol{S}_i;\boldsymbol{S}_B)$, is zero because $\boldsymbol{S}_i$ and $\boldsymbol{S}_B$ are independent. The third term, $I(\boldsymbol{S}'_i;(\boldsymbol{S}_B,\boldsymbol{S}'_B)\mid\boldsymbol{S}_i)$, is at most $\ln d$ by Fact 5.10 and the fact that conditioned on $\boldsymbol{S}_i$ there are only d possible values for $\boldsymbol{S}'_i$. For the remaining term, we bound it in expectation over $\boldsymbol{i}$,

$$\begin{aligned} \mathop{\mathbb{E}}_{\boldsymbol{i}\sim\text{Unif}([m]\setminus B)}\big[I\big(\boldsymbol{S}_{\boldsymbol{i}};\boldsymbol{S}'_B\mid\boldsymbol{S}_B\big)\big] &= \frac{1}{m-r}\cdot\sum_{i\in([m]\setminus B)} I\big(\boldsymbol{S}_i;\boldsymbol{S}'_B\mid\boldsymbol{S}_B\big)\\ &\leq \frac{1}{m-r} I\big(\boldsymbol{S}_{[m]\setminus B};\boldsymbol{S}'_B\mid\boldsymbol{S}_B\big) && \text{(Proposition 6.6)}\\ &\leq \frac{r\ln d}{m-r}. && (\text{Fact 5.10 and } \boldsymbol{S}'_B \text{ has } \leq d^r \text{ options given } \boldsymbol{S}_B) \end{aligned}$$

Combining these bounds, we have that

$$\mathop{\mathbb{E}}_{\boldsymbol{i}\sim\text{Unif}([m]),\boldsymbol{B}\sim\binom{[m]\setminus\{\boldsymbol{i}\}}{r}}[I(\boldsymbol{S}'_{\boldsymbol{i}};\boldsymbol{S}'_{\boldsymbol{B}})] \leq \ln d + \frac{r\ln d}{m-r} = \frac{m\ln d}{m-r}.$$

□

6.3 Proof of Lemma 4.2

Proof. Since $\mathcal{D}_{\text{adaptive}}\in\mathscr{D}^{m,\rho,\mathcal{D}}_{\text{adaptive}}$, it is possible to couple $\boldsymbol{S}\sim\mathcal{D}^m$ and $\boldsymbol{S}'\sim\mathcal{D}_{\text{adaptive}}$ so that $\boldsymbol{S}'\in\mathcal{C}_\rho(\boldsymbol{S})$ with probability 1. Next, draw a uniform permutation $\boldsymbol{\sigma}:[m]\to[m]$ and define the random variables $\boldsymbol{T}$ and $\boldsymbol{T}'$ each over X^m as

$$\boldsymbol{T}_i = \boldsymbol{S}_{\boldsymbol{\sigma}(i)} \qquad \text{and} \qquad \boldsymbol{T}'_i = \boldsymbol{S}'_{\boldsymbol{\sigma}(i)}.$$

Note that the marginal distribution of $\boldsymbol{T}$ is still $\mathcal{D}^m$, and it still holds that $\boldsymbol{T}' \in \mathcal{C}_\rho(\boldsymbol{T})$ with probability 1. Therefore, the distribution of $\boldsymbol{T}'$ is still in $\mathscr{D}^{m,\rho,\mathcal{D}}_{\text{adaptive}}$.

The upshot is now the distribution of all permutations of $\boldsymbol{T}'$ are identical. Therefore, setting $\boldsymbol{c} = \boldsymbol{T}'_{[\le k]}$ and $\boldsymbol{R} \coloneqq \boldsymbol{T}'_{[k+1,k+n]}$ we have that the distribution of $(\boldsymbol{R}, \boldsymbol{c})$ is exactly $\text{Grouped}_{n,k}(\mathcal{D}_{\text{adaptive}})$. Therefore,

$$\mathcal{D}_{\text{goal}}(c) = \mathbb{E}[\text{Unif}(\boldsymbol{R}) \mid \boldsymbol{c} = c].$$

Reusing permutation invariance of $\boldsymbol{T}$ (and therefore $\boldsymbol{R}$), we have that the distribution of $\boldsymbol{R}_i$ is the same for all $i \in [n]$ (even after conditioning on $\boldsymbol{c} = c$), and therefore all equal to $\mathcal{D}_{\text{goal}}(c)$. As a result, we have that

$$\mathop{\mathbb{E}}_{\boldsymbol{c}}[d_{\text{KL}}\left(\text{Group}_n(\mathcal{D}_{\text{adaptive}}, \boldsymbol{c}) \,\|\, \mathcal{D}_{\text{goal}}(\boldsymbol{c})^n\right)] = \text{Cor}\left(\boldsymbol{T}'_{[k+1,k+n]} \mid \boldsymbol{T}'_{\le k}\right).$$

We then choose $k \le k_{\max}$ to be the parameter selected by Lemma 4.3 and bound

$$\begin{aligned}
\text{Cor}\left(\boldsymbol{T}'_{[k+1,k+n]} \mid \boldsymbol{T}'_{\le k}\right) &= \mathop{\mathbb{E}}_{\substack{\boldsymbol{A}\sim\binom{[m]}{n} \\ \boldsymbol{B}\sim\binom{[m]\setminus\boldsymbol{A}}{k}}} \left[\text{Cor}(\boldsymbol{T}'_{\boldsymbol{A}} \mid \boldsymbol{T}'_{\boldsymbol{B}})\right] && \text{(Permutation invariance of } \boldsymbol{T}') \\
&\le \frac{n(n-1)}{2(k_{\max}+1)} \cdot \mathop{\mathbb{E}}_{\substack{\boldsymbol{i}\sim\text{Unif}([m]) \\ \boldsymbol{B}\sim\binom{[m]\setminus\{\boldsymbol{i}\}}{n+k-1}}} [I(\boldsymbol{T}'_{\boldsymbol{i}}; \boldsymbol{T}'_{\boldsymbol{B}})] && \text{(Lemma 4.3)} \\
&\le \frac{n(n-1)}{2(k_{\max}+1)} \cdot 2\ln d && \text{(Lemma 6.5 and } n + k_{\max} \le m/2)
\end{aligned}$$

We are ready to give the desired bound:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\boldsymbol{c}}[d_{\text{TV}}(\text{Group}_n(\mathcal{D}_{\text{adaptive}}, \boldsymbol{c}), \mathcal{D}_{\text{goal}}(\boldsymbol{c})^n)] \\
&\quad\le \mathop{\mathbb{E}}_{\boldsymbol{c}}\left[\sqrt{\frac{d_{\text{KL}}\left(\text{Group}_n(\mathcal{D}_{\text{adaptive}}, \boldsymbol{c}) \,\|\, \mathcal{D}_{\text{goal}}(\boldsymbol{c})^n\right)}{2}}\right] && \text{(Fact 5.5)} \\
&\quad\le \sqrt{\mathop{\mathbb{E}}_{\boldsymbol{c}}\left[\frac{d_{\text{KL}}\left(\text{Group}_n(\mathcal{D}_{\text{adaptive}}, \boldsymbol{c}) \,\|\, \mathcal{D}_{\text{goal}}(\boldsymbol{c})^n\right)}{2}\right]} && \text{(Jensen's inequality)} \\
&\quad\le \sqrt{\frac{n(n-1)\ln d}{2(k_{\max}+1)}},
\end{aligned}$$

which is easily upper bounded by $\sqrt{\frac{n^2 \ln d}{2k_{\max}}}$. □

7 Rounding to a legal corruption: Proof of Lemma 4.4

In this section, we prove the following, restated for convenience.

Lemma 4.4 (Error due to rounding for our grouping strategy)**.** *Using the same notation as Lemma 4.2, as long as $k \le m/2$*

$$\mathop{\mathbb{E}}_{\boldsymbol{c}}\left[\inf_{\mathcal{D}_{\text{oblivious}} \in \mathscr{D}^{n,\rho,\mathcal{D}}_{\text{oblivious}}} \{d_{\text{TV}}(\mathcal{D}_{\text{oblivious}}, \mathcal{D}_{\text{goal}}(\boldsymbol{c})^n)\}\right] \le 2n \cdot \sqrt{\frac{k \ln d}{m}}.$$

Our proof of Lemma 4.4 will first prove the following statement, which applies to any grouping strategy.

Lemma 7.1 (Rounding to few groups incurs small error)**.** *For any base distribution $\mathcal{D}$, sample size m, cost function ρ, adaptive corruption $\mathcal{D}_{\text{adaptive}} \in \mathscr{D}^{m,\rho,\mathcal{D}}_{\text{adaptive}}$, draw $\boldsymbol{S}' \sim \mathcal{D}_{\text{adaptive}}$ and let $\boldsymbol{g}$ be any random variable supported on a set G specifying which group $\boldsymbol{S}'$ falls in. Then,*

$$\mathop{\mathbb{E}}_{\boldsymbol{g}}\left[\inf_{\mathcal{D}' \in \mathcal{C}_\rho(\mathcal{D})} \left\{d_{\text{TV}}(\mathcal{D}', \mathcal{D}'_{\text{goal}}(\boldsymbol{g}))\right\}\right] \le \sqrt{\frac{\ln(|G|)}{2m}}.$$

where $\mathcal{D}'_{\text{goal}}(g) \coloneqq \mathbb{E}[\text{Unif}(\boldsymbol{S}') \mid \boldsymbol{g} = g]$.

We begin by proving Lemma 7.1. Later, in Section 7.3, we show how to use it to derive Lemma 4.4.

7.1 Proof of Lemma 7.1

The proof of Lemma 7.1 is broken into two pieces. First, we prove it when the adversary cannot make any corruptions:

Claim 7.2. *For any distribution $\mathcal{D}$, draw $\boldsymbol{S} \sim \mathcal{D}^m$ and $\boldsymbol{g}$ be any random variable on the same probability space as $\boldsymbol{S}$ supported on a set G. Then,*

$$\mathop{\mathbb{E}}_{\boldsymbol{g}}[d_{\text{TV}}(\mathcal{D}, \mathcal{D}_{\text{goal}}(\boldsymbol{g}))] \le \sqrt{\frac{\ln(|G|)}{2m}} \qquad \textit{where} \quad \mathcal{D}_{\text{goal}}(g) \coloneqq \mathbb{E}[\text{Unif}(\boldsymbol{S}) \mid \boldsymbol{g} = g].$$

This is a special case of Lemma 7.1 corresponding to when the cost function is simply

$$\rho(x, y) = \begin{cases} 0 & \text{if } x = y \\ \infty & \text{if } x \neq y. \end{cases}$$

Then, we will use the following to show that the special case in Claim 7.2 is sufficient to prove the more general case.

Claim 7.3 (Lipschitzness of corruptions)**.** *For any cost function ρ, distributions $\mathcal{D}_1$ and $\mathcal{D}_2$, and any $\mathcal{D}'_1 \in \mathcal{C}_\rho(\mathcal{D}_1)$, there is some $\mathcal{D}'_2 \in \mathcal{C}_\rho(\mathcal{D}_2)$ for which*

$$d_{\text{TV}}(\mathcal{D}'_1, \mathcal{D}'_2) \le d_{\text{TV}}(\mathcal{D}_1, \mathcal{D}_2).$$

Proof of Lemma 7.1 using Claims 7.2 and 7.3. Since $\mathcal{D}_{\text{adaptive}}$ is a valid adaptive corruption, it is possible to couple a sample $\boldsymbol{S} \sim \mathcal{D}^m$ with $\boldsymbol{S}' \sim \mathcal{D}_{\text{adaptive}}$ so that $\boldsymbol{S}' \in \mathcal{C}_\rho(\boldsymbol{S})$ with probability 1. Then, define $\mathcal{D}_{\text{goal}}(g) \coloneqq \mathbb{E}[\text{Unif}(\boldsymbol{S}) \mid \boldsymbol{g} = g]$. Then, Claim 7.2 gives that

$$\mathop{\mathbb{E}}_{\boldsymbol{g}}[d_{\text{TV}}(\mathcal{D}, \mathcal{D}_{\text{goal}}(\boldsymbol{g}))] \le \sqrt{\frac{\ln(|G|)}{2m}} \qquad \text{where} \quad \mathcal{D}_{\text{goal}}(g) \coloneqq \mathbb{E}[\text{Unif}(\boldsymbol{S}) \mid \boldsymbol{g} = g].$$

It therefore suffices to show that, for every $g \in G$, there is some $\mathcal{D}' \in \mathcal{C}_\rho(\mathcal{D})$ for which

$$d_{\text{TV}}(\mathcal{D}', \mathcal{D}'_{\text{goal}}(g)) \le d_{\text{TV}}(\mathcal{D}, \mathcal{D}_{\text{goal}}(g)).$$

Using Claim 7.3, this is implied by showing that $\mathcal{D}'_{\text{goal}}(g) \in \mathcal{C}_\rho(\mathcal{D}_{\text{goal}}(g))$. This means showing a coupling of $\boldsymbol{x} \sim \mathcal{D}_{\text{goal}}(g)$ and $\boldsymbol{x}' \sim \mathcal{D}'_{\text{goal}}(g)$ for which $\mathbb{E}[\rho(\boldsymbol{x}, \boldsymbol{x}')] \le 1$.

For this coupling, first condition on $\boldsymbol{g} = g$ and then take a uniform index $\boldsymbol{i} \sim \text{Unif}([m])$. We will set $\boldsymbol{x} \coloneqq \boldsymbol{S}_i$ and $\boldsymbol{x}' \coloneqq \boldsymbol{S}'_i$. By definition, the marginal distribution of $\boldsymbol{x}$ is $\mathcal{D}_{\text{goal}}(g)$ and of $\boldsymbol{x}'$ is $\mathcal{D}'_{\text{goal}}(g)$. Finally,

$$\mathbb{E}\big[\rho(\boldsymbol{x}, \boldsymbol{x}')\big] = \mathbb{E}\left[\frac{1}{m}\sum_{i\in[m]} \rho(\boldsymbol{S}_i, \boldsymbol{S}'_i)\right] \le 1. \qquad \square$$

7.1.1 The special case when the adversary cannot corrupt: Proof of Claim 7.2

This proof will use a few standard facts about sub-Gaussian random variables. We refer the reader to [Ver18] for a more thorough treatment of sub-Gaussian random variables.

Definition 14 (Sub-Gaussian random variables)**.** *A random variable $\boldsymbol{z}$ is said to be* sub-Gaussian *with variance-proxy σ^2 if, for all $\lambda \in \mathbb{R}$,*

$$\mathbb{E}[e^{\lambda \boldsymbol{z}}] \le \exp\left(\tfrac{\lambda^2\sigma^2}{2}\right).$$

Fact 7.4 (Mean of bounded and independent random variables is sub-Gaussian)**.** *Let $\boldsymbol{z}_1, \dots, \boldsymbol{z}_m$ each be independent mean-0 random variables bounded on $[a, a+1]$ for some $a \in \mathbb{R}$. Then $\boldsymbol{Z} \coloneqq \frac{\boldsymbol{z}_1 + \cdots + \boldsymbol{z}_m}{m}$ is sub-Gaussian with variance-proxy $\frac{1}{4m}$.*

Fact 7.5 (Expected maximum of sub-Gaussian random variables)**.** *Let $\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_n$ be (not necessarily independent) sub-Gaussian random variables each with variance proxy σ^2. Then,*

$$\mathbb{E}\left[\max_{i\in[n]} \boldsymbol{Z}_i\right] \le \sqrt{2\sigma^2 \ln(n)}.$$

Proof of Claim 7.2. Our goal is to bound

$$\begin{aligned}\mathop{\mathbb{E}}_{\boldsymbol{g}}[d_{\text{TV}}(\mathcal{D}_{\text{goal}}(\boldsymbol{g}), \mathcal{D})] &= \mathop{\mathbb{E}}_{\boldsymbol{g}}\left[\sup_{T:X\to[0,1]} \left\{\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_{\text{goal}}(\boldsymbol{g})}[T(\boldsymbol{x})] - \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}}[T(\boldsymbol{x})]\right\}\right] \\ &= \sup_{T_g:X\to[0,1] \text{ for all } g\in G} \left\{\mathop{\mathbb{E}}_{\boldsymbol{g}}\left[\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_{\text{goal}}(\boldsymbol{g})}[T_{\boldsymbol{g}}(\boldsymbol{x})] - \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}}[T_{\boldsymbol{g}}(\boldsymbol{x})]\right]\right\}.\end{aligned}$$

For the remainder of this proof we will fix an arbitrary choice of $\{T_g\}_{g\in G}$ and upper bound the above quantity. First, we shift the T_g so that the second term is 0 by defining

$$T'_g(x) = T_g(x) - \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}}[T_g(\boldsymbol{x})].$$

The result is that the range of $T'_g(x)$ is of the form $[a, a+1]$ for some $a \in \mathbb{R}$ and that $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[T'_g(\boldsymbol{x})] = 0$. Our goal is to upper bound the quantity $\mathbb{E}_{\boldsymbol{g}}\left[\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{\text{goal}}(\boldsymbol{g})}[T'_{\boldsymbol{g}}(\boldsymbol{x})]\right]$ over all such $\{T'_g\}_{g\in G}$.

Since $\mathcal{D}_{\text{goal}}(g) \coloneqq \mathbb{E}[\text{Unif}(\boldsymbol{S}) \mid \boldsymbol{g} = g]$, for any T'_g we have that

$$\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_{\text{goal}}(g)}[T'_g(\boldsymbol{x})] = \mathop{\mathbb{E}}_{\boldsymbol{S}|\boldsymbol{g}=g}\left[\frac{1}{m}\cdot\sum_{i\in[m]} T'_g(\boldsymbol{S}_i)\right].$$

Therefore, we wish to upper bound

$$\mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}^m}\left[\mathop{\mathbb{E}}_{\boldsymbol{g}|\boldsymbol{S}}\left[\frac{1}{m}\cdot\sum_{i\in[m]}T'_{\boldsymbol{g}}(\boldsymbol{S}_i)\right]\right]\le\mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}^m}\left[\max_{g\in G}\left(\frac{1}{m}\cdot\sum_{i\in[m]}T'_g(\boldsymbol{S}_i)\right)\right].$$

For each g, let us define the random variable $\boldsymbol{Z}_g$ to be $\frac{1}{m}\cdot\sum_{i\in[m]}T'_g(\boldsymbol{S}_i)$. Then, the above is simply $\mathbb{E}[\max_{g\in G}[\boldsymbol{Z}_g]]$. Furthermore, by Fact 7.4, each $\boldsymbol{Z}_g$ is sub-Gaussian with variance-proxy $\frac{1}{4m}$. Finally, we apply Fact 7.5 to give that

$$\mathop{\mathbb{E}}_{\boldsymbol{g}}[d_{\mathrm{TV}}(\mathcal{D}_{\mathrm{goal}}(\boldsymbol{g}),\mathcal{D})]\le\mathbb{E}[\max_{g\in G}\boldsymbol{Z}_g]\le\sqrt{\frac{\ln|G|}{2m}}.$$

□

7.2 Generalizing to when the adversary can corrupt: Proof of Claim 7.3

Proof. By Definition 10, it suffices to show that for any coupling of $\boldsymbol{x}_1\sim\mathcal{D}_1$ and $\boldsymbol{x}_2\sim\mathcal{D}_2$, there is a coupling of $\boldsymbol{y}_1\sim\mathcal{D}'_1$ and $\boldsymbol{y}_2\sim\mathcal{D}'_2$ for which

$$\Pr[\boldsymbol{y}_1\neq\boldsymbol{y}_2]\le\Pr[\boldsymbol{x}_1\neq\boldsymbol{x}_2].$$

Fix any such coupling of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ and let $\mathcal{D}_{\boldsymbol{x}_2|x_1}$ denote the distribution of $\boldsymbol{x}_2$ conditioned on $\boldsymbol{x}_1=x_1$ under this coupling.

Since $\mathcal{D}'_1\in\mathcal{C}_\rho(\mathcal{D}_1)$, there is a coupling of $\boldsymbol{x}_1\sim\mathcal{D}$ and $\boldsymbol{y}_1\sim\mathcal{D}'_1$ for which $\mathbb{E}[\rho(\boldsymbol{x}_1,\boldsymbol{y}_1)]\le 1$. Let $\mathcal{D}_{\boldsymbol{y}_1|x_1}$ be the distribution of $\boldsymbol{y}_1$ conditioned on $\boldsymbol{x}_1=x_1$ in this coupling.

We will now specify a joint distribution over all of $\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y}_1,\boldsymbol{y}_2$.

1. Draw $\boldsymbol{x}_1\sim\mathcal{D}$.
2. Draw $\boldsymbol{x}_2\sim\mathcal{D}_{\boldsymbol{x}_2|x_1}$. This will result in the marginal distribution of $\boldsymbol{x}_2$ being $\mathcal{D}'_2$.
3. Draw $\boldsymbol{y}_1\sim\mathcal{D}_{\boldsymbol{y}_1|x_1}$. This will result in the marginal distribution of $\boldsymbol{y}'_1$ being $\mathcal{D}'_1$.
4. Set $\boldsymbol{y}_2$ to
$$\boldsymbol{y}_2=\begin{cases}\boldsymbol{y}_1 & \text{if } \boldsymbol{x}_1=\boldsymbol{x}_2\\ \boldsymbol{x}_2 & \text{otherwise,}\end{cases}\tag{3}$$
and define $\mathcal{D}'_2$ to be the distribution over $\boldsymbol{y}_2$.

The desired result follows from the following two claims:

Claim 1, $\Pr[\boldsymbol{y}_1\neq\boldsymbol{y}_2]\le\Pr[\boldsymbol{x}_1\neq\boldsymbol{x}_2]$**:** For this, we simply observe from Equation (3) that if $\boldsymbol{x}_1=\boldsymbol{x}_2$ then $\boldsymbol{y}_1=\boldsymbol{y}_2$. Therefore, $\Pr[\boldsymbol{y}_1=\boldsymbol{y}_2]\ge\Pr[\boldsymbol{x}_1=\boldsymbol{x}_2]$, and negating this gives the desired result.

Claim 2, $\mathcal{D}'_2\in\mathcal{C}_\rho(\mathcal{D}_2)$**:** For this, it suffices to show that $\mathbb{E}[\rho(\boldsymbol{x}_2,\boldsymbol{y}_2)]\le 1$. We bound,

$$\begin{aligned}\mathbb{E}[\rho(\boldsymbol{x}_2,\boldsymbol{y}_2)]&=\mathbb{E}[\rho(\boldsymbol{x}_1,\boldsymbol{y}_1)\cdot\mathbb{1}[\boldsymbol{x}_1=\boldsymbol{x}_2]] && \text{(Equation (3) and } \rho(x_2,x_2)=0 \text{ for any } x_2)\\ &\le\mathbb{E}[\rho(\boldsymbol{x}_1,\boldsymbol{y}_1)]\le 1.\end{aligned}$$

□

7.3 Proof of Lemma 4.4

We conclude this section with a better error bound fine tuned for the particular grouping strategy we utilize. The proof of this improved bound will use the more general and weaker bound of Lemma 7.1 as a black-box. Roughly speaking, Lemma 4.4 obtains the bound that Lemma 7.1 would obtain if there were d^k possible groups, even though there are actually $|X|^k$, which can be substantially larger. This improvement comes the fact that, in Claim 7.6 below, after conditioning on $\boldsymbol{c}$, there are only d^k choices remaining for $\boldsymbol{c}'$.

Claim 7.6. *For any $\mathcal{D}_{\text{adaptive}} \in \mathscr{D}^{m,\rho,\mathcal{D}}_{\text{adaptive}}$ where ρ is a degree-d cost function, draw $\boldsymbol{T} \sim \mathcal{D}_{\text{adaptive}}$ and $\boldsymbol{I} \subseteq [m]$ a uniform random subset of k indices from $[m]$ and define*

$$\boldsymbol{c} = \boldsymbol{T}_{\boldsymbol{I}} \qquad \boldsymbol{R} = \boldsymbol{T}_{[m]\setminus \boldsymbol{I}}$$
$$\boldsymbol{c}' = \boldsymbol{T}'_{\boldsymbol{I}} \qquad \boldsymbol{R}' = \boldsymbol{T}'_{[m]\setminus \boldsymbol{I}}.$$

Then, for $\mathcal{D}_{\text{goal}}(c, c') \coloneqq \mathbb{E}[\text{Unif}(\boldsymbol{R}') \mid \boldsymbol{c} = c \text{ and } \boldsymbol{c}' = c']$ and any fixed value of c, we have that

$$\mathop{\mathbb{E}}_{\boldsymbol{c}'|\boldsymbol{c}=c}\left[\inf_{\mathcal{D}' \in \mathcal{C}_\rho(\mathcal{D})} \left\{d_{\text{TV}}\left(\mathcal{D}', \mathcal{D}_{\text{goal}}(c, \boldsymbol{c}')\right)\right\}\right] \le \sqrt{\frac{k \ln d}{2(m-k)}} + \frac{k}{m}$$

Proof. The key observation underlying this proof is that $\boldsymbol{R}'$ can be understood as the corruption of an appropriately defined adaptive adversary, even after conditioning on $\boldsymbol{c} = c$. Since $\boldsymbol{S}' \in \mathcal{C}_\rho(\boldsymbol{S})$ with probability 1, we have that

$$\sum_{i \in [m-k]} \rho(\boldsymbol{R}_i, \boldsymbol{R}'_i) \le \sum_{i \in [m]} \rho(\boldsymbol{S}_i, \boldsymbol{S}'_i) \le m.$$

Note that this does not necessarily mean that $\boldsymbol{R}' \in \mathcal{C}_\rho(\boldsymbol{R})$, as that would require $\sum_{i\in[m-k]} \rho(\boldsymbol{R}_i, \boldsymbol{R}'_i) \le m - k$. Instead, we have with probability 1 that $\boldsymbol{R}'$ is a valid corruption of $\boldsymbol{R}$ for a slightly more permissive cost function,

$$\boldsymbol{R}' \in \mathcal{C}_{\bar\rho}(\boldsymbol{R}) \qquad \text{for} \qquad \bar\rho(x, y) = \frac{m-k}{m} \cdot \rho(x, y).$$

We also observe that even if we condition on $\boldsymbol{c} = c$ the distribution of $\boldsymbol{R}$ is still simply $\mathcal{D}^{m-k}$. This is because we had originally drawn $\boldsymbol{S} \sim \mathcal{D}^m$ and so $\boldsymbol{c}$ and $\boldsymbol{R}$ are independent. Combining these two observations, we have that the distribution of $\boldsymbol{R}'$ conditioned on $\boldsymbol{c} = c$ is in the class $\mathscr{D}^{m-k,\bar\rho,\mathcal{D}}_{\text{adaptive}}$. Then, since there are at most d^k choices for $\boldsymbol{c}'$ after conditioning on $\boldsymbol{c} = c$, we can apply Lemma 7.1 to obtain,

$$\mathop{\mathbb{E}}_{\boldsymbol{c}'|\boldsymbol{c}=c}\left[\inf_{\bar{\mathcal{D}}' \in \mathcal{C}_{\bar\rho}(\mathcal{D})} \left\{d_{\text{TV}}\left(\mathcal{D}', \mathcal{D}_{\text{goal}}(c, \boldsymbol{c}')\right)\right\}\right] \le \sqrt{\frac{\ln d^k}{2(m-k)}} = \sqrt{\frac{k \ln d}{2(m-k)}}.$$

All that remains is to show that for every $\bar{\mathcal{D}}' \in \mathcal{C}_{\bar\rho}(\mathcal{D})$, there is some $\mathcal{D}' \in \mathcal{C}_\rho(\mathcal{D})$ satisfying $d_{\text{TV}}(\mathcal{D}', \bar{\mathcal{D}}') \le k/m$, as the desired result with then follow by the triangle inequality (Fact 5.2).

Since $\bar{\mathcal{D}}' \in \mathcal{C}_{\bar\rho}(\mathcal{D})$, there is a coupling of $\boldsymbol{x} \sim \mathcal{D}$ and $\bar{\boldsymbol{x}}' \sim \bar{\mathcal{D}}'$ for which $\mathbb{E}[\bar\rho(\boldsymbol{x}, \bar{\boldsymbol{x}}')] \le 1$. Let $\mathcal{D}'$ be the distribution of

$$\boldsymbol{x}' = \begin{cases} \bar{\boldsymbol{x}}' & \text{with probability } \frac{m-k}{m} \\ \boldsymbol{x} & \text{with probability } \frac{k}{m}. \end{cases}$$

Then, $d_{\mathrm{TV}}(\mathcal{D}', \bar{\mathcal{D}}') \le \Pr[\boldsymbol{x}' \neq \bar{\boldsymbol{x}}'] \le \frac{k}{m}$, and $\mathcal{D}' \in \mathcal{C}_\rho(\mathcal{D})$ because

$$\mathbb{E}[\rho(\boldsymbol{x}, \boldsymbol{x}')] = \frac{m-k}{m}\,\mathbb{E}[\rho(\boldsymbol{x}, \bar{\boldsymbol{x}}')] = \frac{m-k}{m} \cdot \frac{m}{m-k}\,\mathbb{E}[\bar{\rho}(\boldsymbol{x}, \bar{\boldsymbol{x}}')] \le 1. \qquad \square$$

We also use the following simple ingredient to prove Lemma 4.4

Proposition 7.7. *Let $\{\mathcal{D}(g)\}_{g \in G}$ be a family of distributions. Then for any base distribution $\mathcal{D}$, cost function ρ, and random variable $\boldsymbol{g}$ supported on G,*

$$\inf_{\mathcal{D}' \in \mathcal{C}_\rho(\mathcal{D})} \left\{ d_{\mathrm{TV}}\left(\mathcal{D}', \mathop{\mathbb{E}}_{\boldsymbol{g}}[\mathcal{D}(\boldsymbol{g})]\right) \right\} \le \mathop{\mathbb{E}}_{\boldsymbol{g}} \left[\inf_{\mathcal{D}' \in \mathcal{C}_\rho(\mathcal{D})} \left\{ d_{\mathrm{TV}}\left(\mathcal{D}', \mathcal{D}(\boldsymbol{g})\right) \right\} \right].$$

Proof. Fix any $\mathcal{D}'(g) \in \mathcal{C}_\rho(\mathcal{D})$ for all $g \in G$. We will show there exists a choice of $\mathcal{D}' \in \mathcal{C}_\rho(\mathcal{D})$ for which

$$d_{\mathrm{TV}}\left(\mathcal{D}', \mathop{\mathbb{E}}_{\boldsymbol{g}}[\mathcal{D}(\boldsymbol{g})]\right) \le \mathop{\mathbb{E}}_{\boldsymbol{g}}\left[d_{\mathrm{TV}}\left(\mathcal{D}'(\boldsymbol{g}), \mathcal{D}(\boldsymbol{g})\right)\right], \tag{4}$$

which implies the desired inequality. Define,

$$\mathcal{D}' := \mathop{\mathbb{E}}_{\boldsymbol{g}}[\mathcal{D}'(\boldsymbol{g})].$$

Then, Equation (4) holds by Fact 5.4. All that remains is to show that $\mathcal{D}' \in \mathcal{C}_\rho(\mathcal{D})$. For all $g \in G$, there must be a coupling of $\boldsymbol{x}' \sim \mathcal{D}'(g)$ and $\boldsymbol{x} \sim \mathcal{D}$ for which $\mathbb{E}[\rho(\boldsymbol{x}, \boldsymbol{x}')] \le 1$. Therefore, if we first draw $\boldsymbol{g}$ and then $\boldsymbol{x}, \boldsymbol{x}'$ from this coupling for $\mathcal{D}'(\boldsymbol{g})$ and $\mathcal{D}$, then

1. The marginal distribution of $\boldsymbol{x}$ will be that of $\mathcal{D}$. This even holds conditioning on any value of $\boldsymbol{g} = g$.
2. The marginal distribution of $\boldsymbol{x}'$ will be $\mathcal{D}'$. This is by the definition of mixture distributions.
3. We can bound,
$$\mathbb{E}[\rho(\boldsymbol{x}, \boldsymbol{x}')] \le \max_g \mathbb{E}[\rho(\boldsymbol{x}, \boldsymbol{x}') \mid \boldsymbol{g} = g] \le 1.$$

Therefore, $\mathcal{D}' \in \mathcal{C}_\rho(\mathcal{D})$. $\square$

We are now ready to prove the main result of this subsection.

Proof of Lemma 4.4. First, if $d = 1$, the only legal cost function is

$$\rho(x, y) = \begin{cases} 0 & \text{if } x = y \\ \infty & \text{otherwise.} \end{cases}$$

In this case, the adversaries can make no corruptions implying that $\mathcal{D}_{\mathrm{adaptive}}$ must be $\mathcal{D}^m$ and $\mathcal{D}_{\mathrm{goal}}(c) = \mathcal{D}$ for all choices of c. The desired error bound of 0 easily holds. Therefore, we may assume $d \ge 2$.

Let $\boldsymbol{c}, \boldsymbol{c}', \boldsymbol{R}, \boldsymbol{R}'$ be as defined in Claim 7.6. We first observe that by taking $\boldsymbol{S}' \sim \Phi_{m-k \to n}(\boldsymbol{R}')$, we have that $(\boldsymbol{S}', \boldsymbol{c}')$ are distributed according to $\mathrm{Grouped}_{n,k}(\mathcal{D}_{\mathrm{adaptive}})$. Therefore,

$$\mathcal{D}_{\mathrm{goal}}(c') = \mathbb{E}[\mathrm{Unif}(\boldsymbol{R}') \mid \boldsymbol{c}' = c'] = \mathop{\mathbb{E}}_{\boldsymbol{c} \mid \boldsymbol{c}' = c'}[\mathcal{D}_{\mathrm{goal}}(\boldsymbol{c}, c')],$$

where $\mathcal{D}_{\text{goal}}(c, c')$ is as defined in Claim 7.6. We proceed to bound

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{c}'}\left[\inf_{\mathcal{D}'\in\mathcal{C}_\rho(\mathcal{D})}\left\{d_{\text{TV}}\left(\mathcal{D}', \mathcal{D}_{\text{goal}}(\boldsymbol{c}')\right)\right\}\right] &= \mathop{\mathbb{E}}_{\boldsymbol{c}'}\left[\inf_{\mathcal{D}'\in\mathcal{C}_\rho(\mathcal{D})}\left\{d_{\text{TV}}\left(\mathcal{D}', \mathop{\mathbb{E}}_{\boldsymbol{c}|\boldsymbol{c}'}\left[\mathcal{D}_{\text{goal}}(\boldsymbol{c}, \boldsymbol{c}')\right]\right)\right\}\right] \\
&\le \mathop{\mathbb{E}}_{\boldsymbol{c}'}\left[\mathop{\mathbb{E}}_{\boldsymbol{c}|\boldsymbol{c}'}\left[\inf_{\mathcal{D}'\in\mathcal{C}_\rho(\mathcal{D})}\left\{d_{\text{TV}}\left(\mathcal{D}', \mathcal{D}_{\text{goal}}(\boldsymbol{c}, \boldsymbol{c}')\right)\right\}\right]\right] \qquad \text{(Proposition 7.7)} \\
&\le \mathop{\mathbb{E}}_{\boldsymbol{c}}\left[\mathop{\mathbb{E}}_{\boldsymbol{c}'|\boldsymbol{c}}\left[\inf_{\mathcal{D}'\in\mathcal{C}_\rho(\mathcal{D})}\left\{d_{\text{TV}}\left(\mathcal{D}', \mathcal{D}_{\text{goal}}(\boldsymbol{c}, \boldsymbol{c}')\right)\right\}\right]\right] \\
&\le \sqrt{\frac{k\ln d}{2(m-k)}} + \frac{k}{m} \qquad \text{(Claim 7.6)} \\
&\le 2\sqrt{\frac{k\ln d}{m}} \qquad (k \le m/2 \text{ and } d \ge 2.)
\end{aligned}$$

Recall that $\mathscr{D}^{n,\rho,\mathcal{D}}_{\text{oblivious}} := \{(\mathcal{D}')^n \mid \mathcal{D}' \in \mathcal{C}_\rho(\mathcal{D})\}$, and so the desired bound follows from Fact 5.3. □

8 Adaptive adversaries are at least as strong as oblivious adversaries

In this section, we prove the easy direction of Theorem 6.

Claim 8.1 (The adaptive adversary can simulate the oblivious adversary)**.** *For any $m \ge 25n^2/\varepsilon^2$, distribution $\mathcal{D}$, cost function ρ, and oblivious adversary $\mathcal{D}_{\text{oblivious}} \in \mathscr{D}^{n,\rho,\mathcal{D}}_{\text{oblivious}}$, there is a corresponding adaptive adversary $\mathcal{D}_{\text{adaptive}} \in \Phi_{m\to n}\left(\mathscr{D}^{m,\rho,\mathcal{D}}_{\text{adaptive}}\right)$ satisfying*

$$d_{\text{TV}}(\mathcal{D}_{\text{adaptive}}, \mathcal{D}_{\text{oblivious}}) \le \varepsilon.$$

Such a statement is well-known to hold for some specific [DKK+19, ZJS19] adversary models. Here, we show it holds with any cost function.

The high-level idea in the proof of Claim 8.1 is simple: The oblivious adversary will have chosen some $\mathcal{D}' \in \mathcal{C}_\rho(\mathcal{D})$ (in which case $\mathcal{D}_{\text{oblivious}} = (\mathcal{D}')^n$) for which it is possible to couple samples $\boldsymbol{S} \sim \mathcal{D}^n$ and $\boldsymbol{S}' \sim (\mathcal{D}')$ so that the *average* cost of corrupting a point in $\boldsymbol{S}_i$ to the corresponding point in $\boldsymbol{S}'_i$ is at most 1. If it were always the case that $\boldsymbol{S}' \in \mathcal{C}_\rho(\boldsymbol{S})$, we would be done, as the adaptive adversary could then always corruption $\boldsymbol{S}$ to $\boldsymbol{S}'$. However, even though the corruption of $\boldsymbol{S}$ to $\boldsymbol{S}'$ has an average cost of 1, it can exceed 1 for some draws of $\boldsymbol{S}, \boldsymbol{S}'$, in which case the adaptive adversary can not exactly simulate the oblivious adversary.

Therefore, our strategy will be to round $\boldsymbol{S}'$ to a valid corruption. The quantity we need to bound is how many points of $\boldsymbol{S}'$ we need to change to make the corruption valid, which we do in the following lemma.

Claim 8.2. *For any distribution $\mathcal{D}_{\text{cost}}$ supported on $\mathbb{R}_{\ge 0}$ with mean 1, draw $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \overset{\text{iid}}{\sim} \mathcal{D}_{\text{cost}}$ and define $\Delta(x_1, \dots, x_n)$ to be the minimum number of x_i that must be removed so that the sum of the remaining elements is at most n. Then,*

$$\mathbb{E}[\Delta(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)] \le 5\sqrt{n}.$$

The main ingredient in the proof of Claim 8.2 is an upper bound on the probability that $\Delta(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ exceeds a value.

Proposition 8.3. *In the setting of Claim 8.2, for any $v \geq 0$*

$$\Pr[\Delta(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \geq v] \leq \frac{2}{v} + \frac{4n}{v^2}.$$

Proof. The main idea in this proof is, for any $r \in [0, 1]$, to exhibit a strategy with the following properties.

1. The probability the strategy removes more than $2nr$ elements is at most $\frac{1}{nr}$.
2. The probability that, after this strategy removes elements, the remaining sum is more than n is at most $\frac{1}{nr^2}$.

Combining the above with a union bound gives that

$$\Pr[\Delta(\boldsymbol{x}) \geq 2nr] \leq \frac{1}{nr} + \frac{1}{nr^2}.$$

The above is equivalent to the desired result as we can set $r = \frac{v}{2n}$.

For any $\tau \in \mathbb{R}$, $p \in [0, 1]$ consider the strategy that, for each $i \in [m]$ keeps $\boldsymbol{x}_i$ with probability $f(\boldsymbol{x}_i)$ and otherwise removes it for

$$f(x) \coloneqq \begin{cases} 1 & \text{if } x < \tau \\ p & \text{if } x = \tau \\ 0 & \text{if } x > \tau. \end{cases}$$

It is always possible to choose τ and p so that the probability $\boldsymbol{x}_i$ is removed is any desired value. We set them so that the probability $\boldsymbol{x}_i$ is removed is exactly r.

It is unlikely many elements are removed. Let $\boldsymbol{R}$ be the random variable representing the number of removed elements. Then, the distribution of $\boldsymbol{R}$ is simply $\mathrm{Bin}(n, r)$ and so it satisfies $\mathbb{E}[\boldsymbol{R}] = nr$ and $\mathrm{Var}[\boldsymbol{R}] \leq nr$. By Chebyshev's inequality:

$$\Pr[\boldsymbol{R} \geq 2nr] \leq \frac{\mathrm{Var}[\boldsymbol{R}]}{(2nr - \mathbb{E}[\boldsymbol{R}])^2} \leq \frac{nr}{(nr)^2} = \frac{1}{nr}.$$

It is unlikely the sum of the remaining elements is more than n. Let $\boldsymbol{X}$ be the sum of the remaining elements. Then,

$$\boldsymbol{x} \coloneqq \sum_{i \in [n]} \boldsymbol{z}_i \cdot \boldsymbol{x}_i \qquad \text{for} \quad \boldsymbol{z}_i \sim \mathrm{Ber}(f(\boldsymbol{x}_i)).$$

We will analyze the mean and variance of this $\boldsymbol{X}$ and then use Chebyshev's inequality to bound the probability it is more than 1. For the mean,

$$\begin{aligned} \mathbb{E}[\boldsymbol{X}] &= n \cdot \mathbb{E}[\boldsymbol{z}_i \cdot \boldsymbol{x}_i] && \text{(Linearity of expectation)} \\ &= n \cdot (\mathbb{E}[\boldsymbol{x}_i] - \Pr[\boldsymbol{z}_i = 0] \cdot \mathbb{E}[\boldsymbol{x}_i \mid \boldsymbol{z}_i = 0]) && (\boldsymbol{z}_i \text{ supported on } \{0, 1\}) \\ &\leq n \cdot (1 - r \cdot \tau). \end{aligned}$$

For the variance of $\boldsymbol{X}$, since $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ are independent, the variances sum. Therefore,

$$\begin{aligned}\mathrm{Var}[\boldsymbol{X}] &= n \cdot \mathrm{Var}[\boldsymbol{z}_i \cdot \boldsymbol{x}_i] \\ &\leq n \cdot \mathbb{E}[(\boldsymbol{z}_i \cdot \boldsymbol{x}_i) \cdot (\boldsymbol{z}_i \cdot \boldsymbol{x}_i)] \\ &\leq n \cdot \max(\boldsymbol{z}_i \cdot \boldsymbol{x}_i) \cdot \mathbb{E}[\boldsymbol{z}_i \cdot \boldsymbol{x}_i] \\ &\leq n\tau.\end{aligned}$$

Therefore, by applying Chebyshev's inequality, we see that

$$\Pr[\boldsymbol{X} \geq n] \leq \frac{n\tau}{n^2\tau^2 r^2} = \frac{1}{n\tau^2 r^2}.$$

Note furthermore that if $\tau < 1$, then every term in the sum is less than 1, in which case $\Pr[\boldsymbol{P} \geq 1] = 0$. Therefore, the worst-case choice of τ for our bound is $\tau = 1$ in which case

$$\Pr[\boldsymbol{P} > n] \leq \frac{1}{nr^2}.$$ □

Proof of Claim 8.2. We write,

$$\begin{aligned}\mathbb{E}[\Delta(\boldsymbol{z})] &= \int_0^n \Pr[\Delta(\boldsymbol{z}) \geq v]dv \\ &\leq \int_0^n \max\left(1, \tfrac{2}{v} + \tfrac{4n}{v^2}\right)dv \\ &\leq 2\sqrt{n} + \int_{2\sqrt{n}}^n \frac{4n}{v^2}dv + \int_{2\sqrt{n}}^n \frac{2}{v}dv \\ &\leq 2\sqrt{n} + \int_{2\sqrt{n}}^\infty \frac{4n}{v^2}dv + \int_{2\sqrt{n}}^n \frac{2}{2\sqrt{n}}dv \\ &= 2\sqrt{n} + 2\sqrt{n} + \sqrt{n} = 5\sqrt{n}.\end{aligned}$$ □

We are now ready to prove the main result of this section.

Proof of Claim 8.1. Consider any $\mathcal{D}_{\text{oblivious}} \in \mathscr{D}^{n,\rho,\mathcal{D}}_{\text{oblivious}}$. Then, there is some $\mathcal{D}' \in \mathcal{C}_\rho(\mathcal{D})$ for which $\mathcal{D}_{\text{oblivious}} = (\mathcal{D}')^n$. By definition, there is a coupling between $\boldsymbol{x} \sim \mathcal{D}$ and $\boldsymbol{y} \sim \mathcal{D}'$ so that $\mathbb{E}[\rho(\boldsymbol{x}, \boldsymbol{y})] \leq 1$. We can extend this to a coupling of $\boldsymbol{S} \sim \mathcal{D}^m$ and $\boldsymbol{S}' \sim (\mathcal{D}')^m$ so that

$$\mathbb{E}\left[\frac{1}{m} \cdot \sum_{i \in m]} \rho(\boldsymbol{S}_i, \boldsymbol{S}'_i)\right] \leq 1.$$

Let $\mathcal{D}_{\text{cost}}$ be the distribution of $\rho(\boldsymbol{x}, \boldsymbol{y})$ in this coupling. By Claim 8.2, we can construct $\boldsymbol{S}''$ for which $\frac{1}{m} \cdot \sum_{i \in m]} \rho(\boldsymbol{S}_i, \boldsymbol{S}''_i) \leq 1$ with probability 1 and for which the expected number of coordinates on which $\boldsymbol{S}''$ and $\boldsymbol{S}'$ differ is at most $5\sqrt{m}$. This is because, whenever Claim 8.2 asks to "remove" some $\boldsymbol{x}_i$, we can simply set $\boldsymbol{S}''_i = \boldsymbol{S}_i$ in which case $\rho(\boldsymbol{S}_i, \boldsymbol{S}''_i) = 0$.

The adaptive adversary, given a sample $S \in X^m$ corrupts it to the distribution of $\boldsymbol{S}'' \mid \boldsymbol{S} = S$. The result is that the distribution of $\mathcal{D}_{\text{adaptive}}$ is equivalent to the distribution of $\mathcal{D}_{\text{adaptive}}$. Then,

for any test function $f : X^n \to [0,1]$

$$\begin{aligned}\mathop{\mathbb{E}}_{\boldsymbol{T}\sim\mathcal{D}_{\text{adaptive}}}[f(\boldsymbol{T})] &= \mathbb{E}[f \circ \Phi_{m\to n}(\boldsymbol{S}'')] \\ &\le \mathbb{E}[f \circ \Phi_{m\to n}(\boldsymbol{S}')] + \Pr[\boldsymbol{S}'' \text{ differs from } \boldsymbol{S}' \text{ on one of } n \text{ sampled points}] \\ &\le \mathbb{E}[f \circ \Phi_{m\to n}(\boldsymbol{S}')] + \frac{5n}{\sqrt{m}}.\end{aligned}$$

The distribution of $\Phi_{m\to n}(\boldsymbol{S}')$ for $\boldsymbol{S}' \sim (\mathcal{D}')^m$ is simply $(\mathcal{D}')^n$, and so the first term is simply $\mathbb{E}_{\boldsymbol{T}\sim\mathcal{D}_{\text{oblivious}}}[f(\boldsymbol{T})]$. By our choice of m, the second term's magnitude is at most ε. The desired result follows from the definition of total variation distance. □

9 Putting the pieces together: Proof of Theorem 3

In this section, we combine all the previous ingredients to prove the below theorem, restated for convenience, which also easily implies Theorems 2 and 3.

Theorem 6 (Theorem 3 restated)**.** *For any $n, d \in \mathbb{N}$ where $d \ge 2$, domain X, and $\varepsilon > 0$, let $m = O\left(\frac{n^4(\ln d)^2}{\varepsilon^4}\right)$. Then, for any $f : X^n \to \{0,1\}$, cost function ρ with degree d, and distribution $\mathcal{D}$ supported on X,*

$$\left|\text{Oblivious-Max}_\rho(f,\mathcal{D}) - \text{Adaptive-Max}_\rho(f \circ \Phi_{m\to n},\mathcal{D})\right| \le \varepsilon. \tag{1}$$

The above is a direct consequence of the following simple result combined with Lemma 4.1 and Claim 8.1

Proposition 9.1 (Indistinguishability from simulations)**.** *Let $\mathscr{D}_1$ and $\mathscr{D}_2$ be two families of distributions on the domain X satisfying,*

1. *For all $\mathcal{D}_1 \in \mathscr{D}_1$, there is a random variable $\boldsymbol{\mathcal{D}}_2$ supported on $\mathscr{D}_2$ such that the mixture satisfies*
$$d_{\text{TV}}(\mathcal{D}_1, \mathbb{E}[\boldsymbol{\mathcal{D}}_2]) \le \varepsilon.$$
2. *For all $\mathcal{D}_2 \in \mathscr{D}_2$, there is a random variable $\boldsymbol{\mathcal{D}}_1$ supported on $\mathscr{D}_1$ such that the mixture satisfies*
$$d_{\text{TV}}(\mathcal{D}_2, \mathbb{E}[\boldsymbol{\mathcal{D}}_1]) \le \varepsilon.$$

Then, for any $f : X \to \{0,1\}$,

$$\left|\sup_{\mathcal{D}_1\in\mathscr{D}_1}\left\{\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_1}[f(\boldsymbol{x})]\right\} - \sup_{\mathcal{D}_2\in\mathscr{D}_2}\left\{\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_2}[f(\boldsymbol{x})]\right\}\right| \le \varepsilon.$$

Proof. By symmetry, it suffices to show that

$$\sup_{\mathcal{D}_1\in\mathscr{D}_1}\left\{\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_1}[f(\boldsymbol{x})]\right\} \le \sup_{\mathcal{D}_2\in\mathscr{D}_2}\left\{\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_2}[f(\boldsymbol{x})]\right\} + \varepsilon. \tag{5}$$

Fix any $\mathcal{D}_1 \in \mathscr{D}_1$. Then, there is some random variable $\boldsymbol{\mathcal{D}}_2$ supported on $\mathscr{D}_2$ such that the mixture satisfies

$$d_{\text{TV}}(\mathcal{D}_1, \mathbb{E}[\boldsymbol{\mathcal{D}}_2]) \le \varepsilon.$$

The above implies that,

$$\mathop{\mathbb{E}}_{\boldsymbol{\mathcal{D}}_2}\Big[\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\boldsymbol{\mathcal{D}}_2}[f(\boldsymbol{x})]\Big] \geq \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_1}[f(\boldsymbol{x})] - \varepsilon.$$

Therefore, there must be some fixed choice of $\mathcal{D}_2 \in \mathscr{D}_2$ for which

$$\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_2}[f(\boldsymbol{x})] \leq \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_1}[f(\boldsymbol{x})] - \varepsilon.$$

Hence Equation (5) holds. □

10 Lower bounds

10.1 Proof of Theorem 4

Here, we prove Theorem 4. Given a distribution $\mathcal{D}$ promised to be uniform on some $X' \subseteq X \coloneqq [n]$, we show that approximating the cardinality of X' is harder in the presence of adaptive subtractive contamination than it is in the presence of oblivious subtractive contamination. To begin, we formalize both models. For ease of notation, we stick with the setting where the adversary can remove half of the sample or distribution, though all the conclusions would remain the same if this $1/2$ were replaced with any other constant.

Definition 15 (1/2-Subtractive contamination, special case of Definition 16)**.** *For any distribution $\mathcal{D}$, we say that $\mathcal{D}' \in \mathrm{sub}(\mathcal{D})$ if a sample of $\boldsymbol{x}' \sim \mathcal{D}'$ is equivalent to a sample $\boldsymbol{x} \sim \mathcal{D}$ conditioned on an event that occurs with probability $1/2$.*

Similarly, for any sample $S \in X^{2m}$, we say that $S' \in \mathrm{sub}(S)$ if, for m unique indices $i_1, \ldots, i_m \in [2m]$,

$$(S')_j = S_{i_j} \qquad \text{for all } j \in [m].$$

We prove the following.

Theorem 7 (A polynomial increase in sample size is necessary, formal version Theorem 4)**.** *Let $\mathcal{D}$ be a distribution on $X = [m] \coloneqq \{1, \ldots, m\}$ that is promised to be uniform on some $X' \subseteq X$. Then for some absolute constant $c < 1$,*

1. *For $n_{\mathrm{small}} \coloneqq O(\sqrt{m})$ and any $k \leq m$, there is an algorithm $f_{\mathrm{oblivious}} : X^{n_{\mathrm{small}}} \to \{0,1\}$ that distinguishes between the cases where $|X'| \geq k$ vs $|X'| \leq ck$ with high probability even in the presence of the oblivious adversary,*

$$|X'| \geq k \implies \inf_{\mathcal{D}' \in \mathrm{sub}(\mathrm{Unif}(X'))} \left(\Pr_{\boldsymbol{S}\sim(\mathcal{D}')^{m_{\mathrm{small}}}} [f_{\mathrm{oblivious}}(\boldsymbol{S})] \right) \geq 0.99$$

$$|X'| \leq ck \implies \sup_{\mathcal{D}' \in \mathrm{sub}(\mathrm{Unif}(X'))} \left(\Pr_{\boldsymbol{S}\sim(\mathcal{D}')^{m_{\mathrm{small}}}} [f_{\mathrm{oblivious}}(\boldsymbol{S})] \right) \leq 0.01.$$

2. *For $n_{\mathrm{large}} = \Omega(m)$ and $k = m$, there is no algorithm with the same guarantees: Formally, for any $f_{\mathrm{adaptive}} : X^{n_{\mathrm{large}}} \to \{0,1\}$, either there is an X' containing k elements for which*

$$\mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathrm{Unif}(X')^{2n_{\mathrm{large}}}} \left[\inf_{\boldsymbol{S}' \in \mathrm{sub}(\boldsymbol{S})} \mathbb{1}\left[f_{\mathrm{adaptive}}(\boldsymbol{S}') \right] \right] \leq 0.51$$

or there is an X' containing at most ck elements for which

$$\mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathrm{Unif}(X')^{2n_{\mathrm{large}}}}\left[\sup_{\boldsymbol{S}'\in\mathrm{sub}(\boldsymbol{S})}\mathbb{1}\left[f_{\mathrm{adaptive}}(\boldsymbol{S}')\right]\right]\geq 0.49$$

Note that by standard techniques, the above can be converted to a separation in the search version of the problem, where the goal is to approximate $|X'|$ to multiplicative accuracy, at the cost of a polylog(m) dependence.

The construction of $f_{\mathrm{oblivious}}$ follows from standard results on the probability of a collision in a sample (see e.g. [GR11]).

Fact 10.1 (The probability of a collision in a sample)**.** *Let $\mathcal{D}'$ be any distribution supported on k points for which $\Pr_{\boldsymbol{x}\sim\mathcal{D}}[\boldsymbol{x}=x]\leq\frac{2}{k}$ for all possible x. Then, for $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\overset{\mathrm{iid}}{\sim}\mathcal{D}'$, let $\boldsymbol{E}$ indicate whether there is $i\neq j$ for which $\boldsymbol{x}_i=\boldsymbol{x}_j$. There are absolute constants c_1,c_2 for which,*

$$c_1n^2\leq k\implies\Pr[\boldsymbol{E}]\leq 0.01\qquad\text{and}\qquad c_2n^2\geq k\implies\Pr[\boldsymbol{E}]\geq 0.99.$$

Give the above fact, we just have $f_{\mathrm{oblivious}}$ accept if and only if it finds a collision in the sample.

The lower bound against the adaptive adversary is an easy consequence of the following simple proposition which shows the adaptive adversary can remove all duplicates from the sample.

Proposition 10.2. *Let $\mathcal{D}$ be uniform on a distribution supported on k points and*

$$n\coloneqq\lfloor\tfrac{\varepsilon k}{2}\rfloor.$$

For $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{2n}\overset{\mathrm{iid}}{\sim}\mathcal{D}$, with probability at least $1-\varepsilon$, the number $i\in[2n]$ for which there exists $j\neq i$ satisfying $\boldsymbol{x}_i=\boldsymbol{x}_j$ is at most n.

Proof. Let $\boldsymbol{z}_i$ be the indicator that there is some $j\neq i$ for which $\boldsymbol{x}_j=\boldsymbol{x}_i$. By union bound,

$$\mathbb{E}[\boldsymbol{z}_i]\leq\frac{2n-1}{k}\leq\frac{2n}{k}.$$

Applying Markov's inequality to $\boldsymbol{Z}=\sum_{i\in[2n]}\boldsymbol{z}_i$,

$$\Pr[\boldsymbol{Z}\geq n]\leq\frac{\mathbb{E}[\boldsymbol{Z}]}{n}\leq\frac{\frac{2n^2}{k}}{n}\leq\frac{2n}{k},$$

which is at most ε for our choice of n. □

Proof of Theorem 7. For small enough constant c, Fact 10.1 implies the following. Picking $n(k)=\Theta(\sqrt{k})$ appropriately, the algorithm $f_{\mathrm{oblivious}}(x_1,\ldots,x_n)$ which accepts iff it has a collision in its first $n(k)\leq n_{\mathrm{small}}$ samples has the desired behavior.

For the lower bound against adaptive adversaries, consider the adaptive adversary that given a sample $x_1,\ldots,x_{2n_{\mathrm{large}}}$ does the following:

1. If there are less than n_{large} choices for $i\in[2n_{\mathrm{large}}]$ such that there is $j\neq i$ for which $x_i\neq x_j$, the adaptive adversary takes any size-n subset of $x_1,\ldots,x_{2n_{\mathrm{large}}}$ not containing these collisions.
2. Otherwise, it just returns the first n elements $x_1,\ldots,x_n$.

Proposition 10.2 implies that, for any $X' \subseteq X$ and $\boldsymbol{x}_1, \dots, \boldsymbol{x}_{2n_{\text{large}}}$, the probability the second case occurs is at most 0.01.

Now, for any $f_{\text{adaptive}} : X^{n_{\text{large}}} \to \{0,1\}$, define,

$$p \coloneqq \mathop{\mathbb{E}}_{\boldsymbol{S} \sim \binom{X}{n_{\text{large}}}} [f_{\text{adaptive}}(\boldsymbol{S})].$$

If $p \geq 0.5$, suppose we draw $\boldsymbol{X}'$ uniformly among all size ck subsets of X. Then, conditioned on the second case of the adaptive adversary's strategy not occurring, f_{adaptive} be equally likely to receive any size n_{large} subset of X. Therefore, it accepts with probability at least $0.5 - 0.01 = 0.49$. There must hence exist at least one X' of size ck for which, with this strategy for the adaptive adversary, the probability that $f_{\text{adaptive}} : X^{n_{\text{large}}}$ accepts is at least 0.49.

On the other hand, if $p \leq 0.5$, we make a similar argument: For $X' = X$, conditioned on the second case of the adaptive adversary's strategy not occurring, the sample f_{adaptive} sees is equally likely to be any size-n_{large} subset of X. Therefore, it can accept with probability at most $0.5 + 0.01 = 0.51$ for this strategy of the adaptive adversary.

In both cases, f_{adaptive} fails, completing this lower bound. □

10.2 Proof of Theorem 5

Theorem 8 (Dependence on degree is necessary, formal version of Theorem 5)**.** *For any $b, \delta > 0$, large enough $n \in \mathbb{N}$ (as a function of b and δ), and cost function $\rho : X \times X \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ for which $\rho(x,y) \geq 1 + \delta$ whenever $x \neq y$, let*

$$m = \Omega_{b,\delta}\left(\frac{n}{(\ln n)^2} \cdot \ln \deg_b(\rho)\right).$$

If $m > n$, there is a function $f : X^n \to \{0,1\}$ and distribution $\mathcal{D}$ over X for which

$$\text{Adaptive-Max}_\rho(f \circ \Phi_{m\to n}, \mathcal{D}) \geq 1 - O(1/n) \qquad \text{and} \qquad \text{Oblivious-Max}_\rho(f, \mathcal{D}) \leq O(1/n). \tag{6}$$

We use similar ideas as [BLMT22, Theorem 8], which shows that Theorem 8 holds in the specific case where ρ corresponds to additive noise, though we do need to generalize those ideas to this more general setting.

Let d be the largest integer such that $2^d \leq \deg_b(\rho)$. By Definition 7, we can choose a subset $X' \subseteq X$ of cardinality 2^d and point $x^\star \in X'$ for which $\rho(x^\star, y) \leq b$ for all $y \in X'$. Let $M : X \to \{\pm 1\}^d \cup \{\vec{0}\}$ be any mapping that takes every element of X' to a unique element of $\{\pm 1\}^d$ and all other elements to $\vec{0}$. For an appropriate threshold $\tau > 0$, we'll define

$$f(x_1, \dots, x_n) \coloneqq \begin{cases} 1 & \text{if for every } x_i, \text{ there is an } x_j \text{ with } j \neq i \text{ s.t. } \langle M(x_i), M(x_j)\rangle \geq \tau \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

This choice of f was analyzed by [BLMT22].

Fact 10.3 (Choosing the threshold τ, [BLMT22])**.** *For any $p \in (0,1)$ and*

$$m \geq \Omega_p\left(\frac{nd}{(\ln n)^2}\right),$$

there is a choice of threshold τ for which both of the following hold:

1. *Lemma 7.1 of [BLMT22], a uniform point is hard to correlate with: For any $x_1, \dots, x_{n-1} \in X$,*

$$\Pr_{\boldsymbol{u} \sim \mathrm{Unif}(X')}[\textit{There is an } i \in [n-1] \textit{ for which } \langle \boldsymbol{u}, x_i \rangle \geq \tau] \leq \frac{1}{n}.$$

2. *Lemma 7.2 of [BLMT22], an adaptive adversary make all points correlated: Take any $n_{\mathrm{small}} \in \mathbb{N}$ for which $pm \leq n_{\mathrm{small}} \leq m$. For $\boldsymbol{S} \sim \mathrm{Unif}(X')^{m_{\mathrm{small}}}$, there is a strategy for adding $m - m_{\mathrm{small}}$ points to $\boldsymbol{S}$ to form $\boldsymbol{S}'$ for which*

$$\mathbb{E}[f \circ \Phi_{m \to n}(\boldsymbol{S}')] \geq 1 - \frac{1}{n}.$$

Proof of Theorem 8. Define

$$c \coloneqq \min\left(\tfrac{1}{b}, 1 - \tfrac{1}{1+\delta}\right),$$

and set $\mathcal{D}$ to the distribution that is equal to $x^\star$ with probability $\frac{c}{2}$ and otherwise uniform over X',

$$\mathcal{D} \coloneqq \tfrac{c}{2} \cdot \{x^\star\} + \left(1 - \tfrac{c}{2}\right) \cdot \mathrm{Unif}(X'). \tag{8}$$

Also, set

$$p \coloneqq 1 - \frac{c}{4}$$

and let τ be the threshold in Fact 10.3. We will show that both Theorem 8 holds with this choice of τ, f as in Equation (7), and $\mathcal{D}$ as in Equation (8).

We begin by analyzing the oblivious adversary. First, for any $\mathcal{D}' \in \mathcal{C}_\rho(\mathcal{D})$, since $\rho(x, x') \geq 1 + \delta$ for each $x \neq x'$, there must be a coupling of $\boldsymbol{x} \sim \mathcal{D}$ and $\boldsymbol{x}' \sim \mathcal{D}'$ for which

$$\Pr[\boldsymbol{x} \neq \boldsymbol{x}'] \leq \frac{1}{1+\delta}$$

Furthermore, based on Equation (8), there is a coupling of $\boldsymbol{x} \sim \mathcal{D}$ and $\boldsymbol{u} \sim \mathrm{Unif}(X')$ for which

$$\Pr[\boldsymbol{x} \neq \boldsymbol{u}] \leq \frac{c}{2}.$$

Combining the above, there is a coupling of $\boldsymbol{x}' \sim \mathcal{D}'$ and $\boldsymbol{u} \sim \mathrm{Unif}(X')$ for which

$$\Pr[\boldsymbol{x}' \neq \boldsymbol{u}] \leq \frac{c}{2} + \frac{1}{1+\delta}.$$

In particular,

$$\Pr[\boldsymbol{x}' = \boldsymbol{u}] \geq \left(1 - \tfrac{1}{1+\delta}\right) - \frac{c}{2} \geq \frac{c}{2}.$$

This means there is a coupling of $\boldsymbol{S} \sim \mathcal{D}'$ and $\boldsymbol{U} \sim \mathrm{Unif}(X')$ for which, independently for each $i \in [n]$, with probability at least $c/2$, $\boldsymbol{S}_i = \boldsymbol{U}_i$. Because we assumed n is sufficiently large as a function of b and δ, we are free to assume that $c \geq \Omega((\ln n)/n)$. As a result, with probability at least $1 - 1/n$, there is some $i \in [n]$ for which $\boldsymbol{S}_i = \boldsymbol{U}_i$. Then,

$$\mathop{\mathbb{E}}_{\boldsymbol{S} \sim (\mathcal{D}')^n}[f(\boldsymbol{S})] \leq \frac{1}{n} + \mathop{\mathbb{E}}_{\boldsymbol{S}}[f(\boldsymbol{S}) \mid \boldsymbol{S}_i = \boldsymbol{U}_i \text{ for some } i \in [n]] \leq \frac{2}{n},$$

where the second inequality is by the first part of Fact 10.3.

We proceed to analyze the adaptive adversary. To draw $\boldsymbol{S} \sim \mathcal{D}^m$, we can first draw $\boldsymbol{E} \sim \mathrm{Ber}(c/2)^m$. Then, for each $i \in [m]$, if $\boldsymbol{E}_i = 1$ we set $\boldsymbol{S} = x^\star$ and otherwise draw it uniformly from X'.

Conditioned on $\sum_i \boldsymbol{E}_i = E$, we have that E of the elements in $\boldsymbol{S}$ are set to $x^\star$ and the other $m - E$ are drawn independently and uniformly from X'. By the second part of Fact 10.3, whenever $m - E \geq pm$ (or equivalently, $E \leq \frac{mc}{4}$), there is a way to modify the E many elements for which $\boldsymbol{E}_i = 1$ to form a corrupted sample $\boldsymbol{S}'$ satisfying

$$\mathbb{E}[f \circ \Phi_{m\to n}(\boldsymbol{S}')] \geq 1 - \frac{1}{n}.$$

Furthermore, if $E \leq mc \leq \frac{m}{b}$, the adversary has enough budget to modify all of the indices for which $\boldsymbol{E}_i = 1$ to arbitrary elements of X'. Therefore, there is a strategy for the adaptive adversary so that

$$\mathbb{E}\left[f \circ \Phi_{m\to n}(\boldsymbol{S}') \,\middle|\, \frac{mc}{4} \leq \sum_{i\in[m]} \boldsymbol{E}_i \leq mc \right] \geq 1 - \frac{1}{n}.$$

We once again assume that $c \geq \Omega((\ln n)/n)$, which implies that $c \geq \Omega((\ln m)/m)$. By a Chernoff bound, this gives that

$$\Pr\left[\frac{mc}{4} \leq \sum_{i\in[m]} \boldsymbol{E}_i \leq mc \right] \geq 1 - \frac{1}{n}.$$

So by union bound,

$$\mathbb{E}\big[f \circ \Phi_{m\to n}(\boldsymbol{S}') \big] \geq 1 - \frac{2}{n}. \qquad \square$$

11 Acknowledgments

The authors thank Li-Yang Tan, Abhishek Shetty, and the anonymous STOC reviewers for their helpful discussions and feedback. Gregory is supported by a Simons Foundation Investigator Award, NSF award AF-2341890 and UT Austin's Foundations of ML NSF AI Institute. Guy is supported by NSF awards 1942123, 2211237, and 2224246 and a Jane Street Graduate Research Fellowship.

References

[BEK02] Nader H Bshouty, Nadav Eiron, and Eyal Kushilevitz. PAC learning with nasty noise. *Theoretical Computer Science*, 288(2):255–275, 2002. 1, 1, 3

[BFJ+94] Avrim Blum, Merrick Furst, Jeffrey Jackson, Michael Kearns, Yishay Mansour, and Steven Rudich. Weakly learning dnf and characterizing statistical query learning using fourier analysis. In *Proceedings of the twenty-sixth annual ACM symposium on Theory of computing*, pages 253–262, 1994. 2, C.2

[BLMT22] Guy Blanc, Jane Lange, Ali Malik, and Li-Yang Tan. On the power of adaptivity in statistical adversaries. In *Conference on Learning Theory*, pages 5030–5061. PMLR, 2022. (document), 1, 1, 2.4, 1, 2, 3, 10.2, 10.2, 10.3, 1, 2, A.2, C, C.1, C.1, C.1, C.2

[BRS11] Boaz Barak, Prasad Raghavendra, and David Steurer. Rounding semidefinite programming hierarchies via global correlation. In *2011 ieee 52nd annual symposium on foundations of computer science*, pages 472–481. IEEE, 2011. 4.3

[Can22] Clément L Canonne. A short note on an inequality between kl and tv. *arXiv preprint arXiv:2202.07198*, 2022. 5, 5.5

[CHL+23] Clément Canonne, Samuel B Hopkins, Jerry Li, Allen Liu, and Shyam Narayanan. The full landscape of robust mean testing: Sharp separations between oblivious and adaptive contamination. In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 2159–2168. IEEE, 2023. (document), 1, 1, 1, 2.3, 2.1, 2.4, 3.1, 20

[CSV17] Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *Proceedings of the 49th Annual Symposium on Theory of Computing (STOC)*, pages 47–60, 2017. 1

[DK23] Ilias Diakonikolas and Daniel M Kane. *Algorithmic high-dimensional robust statistics*. Cambridge university press, 2023. 1, 1, 3

[DKK+19] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019. 1, 1, 8

[DKPZ21] Ilias Diakonikolas, Daniel M Kane, Thanasis Pittas, and Nikos Zarifis. The optimality of polynomial regression for agnostic learning under gaussian marginals in the sq model. In *Conference on Learning Theory*, pages 1552–1584. PMLR, 2021. 1, 3

[DSFT+14] Dana Dachman-Soled, Vitaly Feldman, Li-Yang Tan, Andrew Wan, and Karl Wimmer. Approximate resilience, monotonicity, and the complexity of agnostic learning. In *Proceedings of the twenty-sixth annual ACM-SIAM symposium on Discrete algorithms*, pages 498–511. SIAM, 2014. 1, 3

[Eld20] Ronen Eldan. Taming correlations through entropy-efficient measure decompositions with applications to mean-field approximation. *Probability Theory and Related Fields*, 176(3):737–755, 2020. 4.3

[Fel10] Vitaly Feldman. Distribution-specific agnostic boosting. *Innovations in Computer Science*, 2010. 3

[Fel17] Vitaly Feldman. A general characterization of the statistical query complexity. In *Conference on learning theory*, pages 785–830. PMLR, 2017. 2

[GR11] Oded Goldreich and Dana Ron. On testing expansion in bounded-degree graphs. *Studies in Complexity and Cryptography. Miscellanea on the Interplay between Randomness and Computation: In Collaboration with Lidor Avigad, Mihir Bellare, Zvika Brakerski, Shafi Goldwasser, Shai Halevi, Tali Kaufman, Leonid Levin, Noam Nisan, Dana Ron, Madhu Sudan, Luca Trevisan, Salil Vadhan, Avi Wigderson, David Zuckerman*, pages 68–75, 2011. 10.1

[Ham71] Frank R. Hampel. A General Qualitative Definition of Robustness. *The Annals of Mathematical Statistics*, 42(6):1887 – 1896, 1971. 1

[Hau92] David Haussler. Decision theoretic generalizations of the pac model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992. 1, 3

[HSSVG22] Daniel J Hsu, Clayton H Sanford, Rocco Servedio, and Emmanouil Vasileios Vlatakis-Gkaragkounis. Near-optimal statistical query lower bounds for agnostically learning intersections of halfspaces with gaussian marginals. In *Conference on Learning Theory*, pages 283–312. PMLR, 2022. 1

[Hub64] Peter Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1), 1964. 1, 3, A.2

[JKR19] Vishesh Jain, Frederic Koehler, and Andrej Risteski. Mean-field approximation, convex hierarchies, and the optimality of correlation rounding: a unified perspective. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 1226–1236, 2019. 4.3, 1

[Kea98] Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM (JACM)*, 45(6):983–1006, 1998. C.2

[KK09] Varun Kanade and Adam Kalai. Potential-based agnostic boosting. *Advances in neural information processing systems*, 22, 2009. 3

[KKMS08] Adam Tauman Kalai, Adam R Klivans, Yishay Mansour, and Rocco A Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, 2008. 3

[KL93] Michael Kearns and Ming Li. Learning in the presence of malicious errors. *SIAM Journal on Computing*, 22(4):807–837, 1993. 1

[KSS94] Michael Kearns, Robert Schapire, and Linda Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2/3):115–141, 1994. 1, 3

[LRV16] Kevin A Lai, Anup B Rao, and Santosh Vempala. Agnostic estimation of mean and covariance. In *Proceedings of the 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 665–674, 2016. 1

[MR17] Pasin Manurangsi and Prasad Raghavendra. A birthday repetition theorem and complexity of approximating dense csps. In *44th International Colloquium on Automata, Languages, and Programming (ICALP 2017)*, pages 78–1. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2017. 4.3, 1

[Pin64] Mark S Pinsker. Information and information stability of random variables and processes. *Holden-Day*, 1964. 5, 5.5

[RT12] Prasad Raghavendra and Ning Tan. Approximating csps with global cardinality constraints using sdp hierarchies. In *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*, pages 373–387. SIAM, 2012. 4.3

[Tuk60] John Wilder Tukey. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, pages 448–485, 1960. 1

[Val85] Leslie G. Valiant. Learning disjunction of conjunctions. In *Proceedings of the 9th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 560–566, 1985. 1, 1, 3.1, 19

[Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018. 7.1.1

[ZJS19] Banghua Zhu, Jiantao Jiao, and Jacob Steinhardt. Generalized resilience and robust statistics. *arXiv*, abs/1909.08755, 2019. 1, 8

A The subtractive and additive adversaries in our framework

We show how to fit subtractive and additive contamination into our framework.

A.1 Subtractive adversaries

First, we formally define η-subtractive corruptions

Definition 16. *For any distribution $\mathcal{D}$ and $\eta \in (0,1)$, we say that $\mathcal{D}'$ is an η-subtractive contamination of $\mathcal{D}$ if a sample from $\mathcal{D}'$ is equivalent to a sample from $\mathcal{D}$ conditioned on an event that occurs with probability at least $1-\eta$. We use $\mathrm{sub}_\eta(\mathcal{D})$ to denote the set of all such $\mathcal{D}'$.*

Similarly, for any $S \in X^m$, we say that S' is an η-subtractive contamination of S if it is formed by removing at most $\lfloor \eta \cdot m \rfloor$ arbitrary points from S. In a slight overload of notation, we use $\mathrm{sub}_\eta(S)$ to denote the set of all such S'.

We will show, as an easy consequence of Theorem 3, that the oblivious and adaptive variants of subtractive contamination are equivalent.

Theorem 9 (Oblivious and adaptive subtractive contamination are equivalent)**.** *For any $\eta, \varepsilon \in (0,1)$, $f : X^n \to \{0,1\}$, and distribution $\mathcal{D}$ over X, let $M = \mathrm{poly}(n, 1/\varepsilon, 1/(1-\eta))$. Then,*

$$\left| \sup_{\mathcal{D}' \in \mathrm{sub}_\eta(\mathcal{D})} \left(\mathop{\mathbb{E}}_{\boldsymbol{S} \sim (\mathcal{D}')^n} [f(\boldsymbol{S})] \right) - \mathop{\mathbb{E}}_{\boldsymbol{S} \sim \mathcal{D}^M} \left[\sup_{\boldsymbol{S}' \in \mathrm{sub}_\eta(\boldsymbol{S})} \left(\mathbb{E}[f \circ \Phi_{\star \to n}(\boldsymbol{S}')] \right) \right] \right| \le \varepsilon.$$

Theorem 9 is an easy consequence of Theorem 6 as well as the below lemma.

Lemma A.1 (Converting the subtractive adversary to our framework)**.** *For any $\eta, \varepsilon \in (0,1)$, $f : X^n \to \{0,1\}$, and distribution $\mathcal{D}$ over X, let*

$$m \coloneqq \left\lceil \frac{\max(2n, 8\ln(1/\varepsilon))}{1-\eta} \right\rceil, \tag{9}$$

and $X' \coloneqq X \cup \{\varnothing\}$. There is a degree-2 cost function ρ and $f' : (X')^m \to \{0,1\}$ for which

$$\left| \sup_{\mathcal{D}' \in \mathrm{sub}_\eta(\mathcal{D})} \left(\mathop{\mathbb{E}}_{\boldsymbol{S} \sim (\mathcal{D}')^n} [f(\boldsymbol{S})] \right) - \text{Oblivious-Max}_{\rho'}(f', \mathcal{D}) \right| \le \varepsilon,$$

and, for all $M \geq m$,

$$\mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}^M}\left[\sup_{\boldsymbol{S}'\in\mathrm{sub}_\eta(\boldsymbol{S})}\left(\mathbb{E}[f\circ\Phi_{\star\to n}(\boldsymbol{S}')]\right)\right] = \text{Adaptive-Max}_\rho(f'\circ\Phi_{M\to m},\mathcal{D})$$

Proof of Theorem 9 assuming Lemma A.1. Let f', ρ, and $\mathcal{D}$ be as in Lemma A.1. By Theorem 6, for $M = \mathrm{poly}(m, 1/\varepsilon) = \mathrm{poly}(n, 1/(1-\eta), \varepsilon)$,

$$\left|\text{Oblivious-Max}_\rho(f',\mathcal{D}') - \text{Adaptive-Max}_\rho(f'\circ\Phi_{M\to m},\mathcal{D}')\right| \leq \varepsilon.$$

By the first part of Lemma A.1 and triangle inequality, we have that

$$\left|\sup_{\mathcal{D}'\in\mathrm{sub}_\eta(\mathcal{D})}\left(\mathop{\mathbb{E}}_{\boldsymbol{S}\sim(\mathcal{D}')^n}[f(\boldsymbol{S})]\right) - \text{Adaptive-Max}_\rho(f'\circ\Phi_{M\to m},\mathcal{D}')\right| \leq 2\varepsilon.$$

By the first part of Lemma A.1 this implies that

$$\left|\sup_{\mathcal{D}'\in\mathrm{sub}_\eta(\mathcal{D})}\left(\mathop{\mathbb{E}}_{\boldsymbol{S}\sim(\mathcal{D}')^n}[f(\boldsymbol{S})]\right) - \mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}^M}\left[\sup_{\boldsymbol{S}'\in\mathrm{sub}_\eta(\boldsymbol{S})}\left(\mathbb{E}[f\circ\Phi_{\star\to n}(\boldsymbol{S}')]\right)\right]\right| \leq 2\varepsilon.$$

The desired result holds by renaming $\varepsilon' = \varepsilon/2$. □

The proof of Lemma A.1 will use the following basic concentration inequality

Proposition A.2. *Let* $\mathcal{D}$ *be a distribution on* $X' \coloneqq X \cup \{\varnothing\}$ *for which*

$$\Pr_{\boldsymbol{x}\sim\mathcal{D}'}[\boldsymbol{x} = \varnothing] \leq \eta.$$

For m *as in Equation (9),*

$$\Pr_{\boldsymbol{S}\sim(\mathcal{D}')^m}\left[\sum_{x\in\boldsymbol{S}}\mathbb{1}[x\neq\varnothing] \leq n\right] \leq \varepsilon.$$

Proof. Let $\boldsymbol{z}$ be the random variable that counts the number of entries $\boldsymbol{S}'$ has that are not equal to $\emptyset$. Then,

$$\mu \coloneqq \mathbb{E}[\boldsymbol{z}] \geq \frac{m}{1-\eta} \geq \max(2n, 8\ln(1/\varepsilon)).$$

Furthermore, $\boldsymbol{z}$ is the sum of independent random variables taking on values in $\{0,1\}$ (each indicating whether $\boldsymbol{S}_i \neq \emptyset$ for some index i). By a standard Chernoff bound (Fact 5.1),

$$\Pr[\boldsymbol{z} \leq n] \leq e^{-\mu/8} \leq \varepsilon.$$

□

Proof of Lemma A.1. We begin by constructing the cost function. In the original subtractive adversary, for each x, the adversary can either choose to keep x in the sample or remove it at a cost of $1/\eta$. We will construct ρ so that the adversary has the same options and represent this "removal" option as converting an input x to $\varnothing$:

$$\rho(x,y) \coloneqq \begin{cases} 0 & \text{if } x = y \\ \frac{1}{\eta} & \text{if } x \neq y \text{ and } y = \varnothing \\ \infty & \text{otherwise.} \end{cases} \tag{10}$$

The function f' simply runs f on a random subset of its non-null input. For any $S \in X^m$, let $S_{\neq\varnothing}$ denote the subset of S consisting of all points not equal to $\varnothing$. Then,

$$f'(S) \coloneqq \begin{cases} f(\Phi_{\star\to n}(S_{\neq\varnothing})) & \text{if } |S_{\neq\varnothing}| \geq n \\ 0 & \text{otherwise.} \end{cases}$$

Next, we analyze the oblivious adversaries. Consider a draw $\boldsymbol{x} \sim \mathcal{D}$ coupled to an event $\boldsymbol{E}$ occurring with probability at least $1-\delta$. Then, $\mathrm{sub}_\eta(\mathcal{D})$ consists of all possible distributions of $\boldsymbol{x}$ conditioned on $\boldsymbol{E}$, whereas, based on Equation (10), $\mathcal{C}_\rho(\mathcal{D})$ consists of all the possible distributions of $\boldsymbol{y}$ where

$$\boldsymbol{y} \coloneqq \begin{cases} \boldsymbol{x} & \text{if } \boldsymbol{E} \\ \varnothing & \text{otherwise.} \end{cases}$$

For any such event $\boldsymbol{E}$, let $\mathcal{D}'_1 \in \mathrm{sub}_\eta(\mathcal{D})$ and $\mathcal{D}'_2 \in \mathcal{C}_\rho(\mathcal{D})$ be the corresponding distribution. We will show that

$$\left| \mathop{\mathbb{E}}_{\boldsymbol{S}_1 \sim (\mathcal{D}'_1)^n} [f(\boldsymbol{S}_1)] - \mathop{\mathbb{E}}_{\boldsymbol{S}_2 \sim (\mathcal{D}'_2)^m} [f(\boldsymbol{S}_2)] \right| \leq \varepsilon.$$

Each element of $\boldsymbol{S}_2$ is set to $\varnothing$ with a probability that is at most η and otherwise has the same distribution as an element of $\boldsymbol{S}_1$. Therefore, the above difference is bounded by the probability that $\boldsymbol{S}_2$ less than n non-null elements. This is at most ε by Proposition A.2.

For the adaptive equivalence, consider any $S \in X^M$. Then any $S_1 \in \mathrm{sub}_\eta(S)$ is formed by removing at most $\lfloor \eta M \rfloor$ of the points in S, whereas $S_2 \in \mathcal{C}_\rho(S)$ is formed by setting $\lfloor \eta M \rfloor$ of the points to $\varnothing$. Suppose we remove the same set of points to form S_1 as we set to $\varnothing$ to form S_2. Then, using the fact that at least n points must remain unchanged since $M \geq m \geq n/(1-\eta)$ and that the subsampling filter composes

$$\mathbb{E}[f \circ \Phi_{\star\to n}(S_1)] = \mathbb{E}[f' \circ \Phi_{M\to m}(S_2)].$$

Hence, for any choice of $S \in X^M$,

$$\sup_{S_1 \in \mathrm{sub}_\eta(S)} (\mathbb{E}[f \circ \Phi_{\star\to n}(S_1)]) = \sup_{S_2 \in \mathcal{C}_\rho(S)} \big(\mathbb{E}[f' \circ \Phi_{M\to m}(S_2)]\big).$$

This implies the desired result. □

A.2 Additive adversaries

First, we formally define η-additive corruptions. Note that the oblivious adversary below exactly corresponds to Huber's original contamination model [Hub64].

Definition 17 (Standard additive adversaries)**.** *For any distribution $\mathcal{D}$ and $\eta \in (0,1)$, we say that $\mathcal{D}'$ is an η-additive contamination of $\mathcal{D}$ if, for some distribution $\mathcal{E}$,*

$$\mathcal{D}' \coloneqq (1-\eta)\mathcal{D} + \eta\mathcal{E}$$

We use $\mathrm{add}_\eta(\mathcal{D})$ to denote the set of all such $\mathcal{D}'$, and for any function $f : X^n \to \{0,1\}$, define

$$\text{Oblivious-Add-Max}_\eta(f, \mathcal{D}) \coloneqq \sup_{\mathcal{D}' \in \mathrm{add}_\eta(\mathcal{D})} \left\{ \mathop{\mathbb{E}}_{\boldsymbol{S}' \sim (\mathcal{D}')^n} (f(\boldsymbol{S}') \right\}.$$

Similarly, for any $S \in X^{\lceil (1-\eta)m \rceil}$, we say S' is an η-additive contamination of S if it is formed by adding $\lfloor \eta m \rfloor$ points to S and then arbitrarily permuting it. In a slight overload of notation, we use $\mathrm{add}_\eta(S)$ *to denote the set of all such S', and for any $f : X^m \to \{0,1\}$, write*

$$\text{Adaptive-Add-Max}_\eta(f, \mathcal{D}) \coloneqq \mathop{\mathbb{E}}_{\boldsymbol{S} \sim \mathcal{D}^{\lceil (1-\eta)m \rceil}} \left[\sup_{\boldsymbol{S}' \in \mathrm{add}_\eta(\boldsymbol{S})} \left(\mathbb{E}[f(\boldsymbol{S}')] \right) \right].$$

The equivalence between these two adversaries was already shown by [BLMT22], but we also prove it here to show our framework can recover their result.

Theorem 10 (Oblivious and adaptive additive contamination are equivalent)**.** *For any $\eta, \varepsilon \in (0,1)$, $f : X^n \to \{0,1\}$, and distribution $\mathcal{D}$ over X, let $m = \mathrm{poly}(n, 1/\varepsilon, \ln |X|)$. Then,*

$$\left| \text{Oblivious-Add-Max}_\eta(f, \mathcal{D}) - \text{Adaptive-Add-Max}_\eta(f \circ \Phi_{m \to n}, \mathcal{D}) \right| \leq \varepsilon.$$

To prove this equivalence, we will introduce a variant of the adaptive adversary, the *binomial* adversary. In this variant, rather than drawing *exactly* $\lceil (1-\eta)m \rceil$ clean points and the adversary being able to add $\lfloor \eta m \rfloor$ corrupted points, the number of clean points is itself drawn randomly from a binomial distribution.

Definition 18 (Binomial adversary)**.** *For any distribution $\mathcal{D}$ and sample size m, the binomial adversary first draws $\boldsymbol{z} \sim \mathrm{Bin}(m, (1-\eta))$ clean points from $\mathcal{D}$, then adds $m - \boldsymbol{z}$ arbitrary points to this clean sample, and finally permutes all m points arbitrarily. For any $f : X^m \to \{0,1\}$, we define*

$$\text{Binomial-Max}_\eta(f, \mathcal{D}) \coloneqq \mathop{\mathbb{E}}_{\boldsymbol{z} \sim \mathrm{Bin}(m, 1-\eta)} \left[\mathop{\mathbb{E}}_{\boldsymbol{S} \sim \mathcal{D}^{\boldsymbol{z}}} \left[\sup_{\boldsymbol{S}' \in \mathrm{complete}_m(\boldsymbol{S})} \left(\mathbb{E}[f(\boldsymbol{S}')] \right) \right] \right]$$

where $\mathrm{complete}_m(S)$ *to denote all samples that can be created by adding $m - |S|$ points to S.*

Our proof of Theorem 10 proceeds in two steps. We first use Theorem 2 to show that the oblivious additive adversary is equivalent to the binomial adversary, and then show equivalence between the binomial adversary and adaptive additive adversary.

Proposition A.3 (The oblivious adversary and binomial adversary are equivalent)**.** *For any $\eta, \varepsilon \in (0,1)$, $f : X^n \to \{0,1\}$, and distribution $\mathcal{D}$ over X, let $m = \mathrm{poly}(n, 1/\varepsilon, \ln |X|)$. Then,*

$$\left| \text{Oblivious-Add-Max}_\eta(f, \mathcal{D}) - \text{Binomial-Max}_\eta(f \circ \Phi_{m \to n}, \mathcal{D}) \right| \leq \varepsilon.$$

Proof. This will be a fairly straightforward application of Theorem 2. Let $X' \coloneqq X \cup \{\varnothing\}$ and $\mathcal{D}_\eta$ be the distribution that outputs $\varnothing$ with probability η and otherwise outputs $\mathcal{D}$,

$$\mathcal{D}_\eta \coloneqq (1-\eta) \cdot \mathcal{D} + \eta \cdot \{\varnothing\}.$$

Then, we'll define an adversary that can send $\varnothing$ to any element of X' but otherwise cannot change its input.

$$\rho(x, y) \coloneqq \begin{cases} 0 & \text{if } x = y \text{ or } x = \varnothing \\ \infty & \text{otherwise.} \end{cases}$$

Finally, let $f' : X^n \to \{0,1\}$ be defined as

$$f'(S) := \begin{cases} 0 & \text{if } \varnothing \in S \\ f(S) & \text{otherwise.} \end{cases}$$

We observe that,

$$\begin{aligned} \text{Oblivious-Add-Max}_\eta(f, \mathcal{D}) &= \text{Oblivious-Max}_\rho(f', \mathcal{D}_\eta), \quad \text{and} \\ \text{Binomial-Max}_\eta(f \circ \Phi_{m\to n}, \mathcal{D}) &= \text{Adaptive-Max}_\rho(f' \circ \Phi_{m\to n}, \mathcal{D}_\eta). \end{aligned}$$

because in order to maximize the success probability of f', the adversaries should send every $\varnothing$ they see to their adversarial choice of an element in X. The desired result follows from Theorem 6. □

Next, we show that the binomial adversary and adaptive additive adversary are equivalent.

Proposition A.4 (The binomial and adaptive additive adversaries are equivalent)**.** *For any $f : X^n \to \{0,1\}$, $\varepsilon, \eta \in (0,1)$, and distribution $\mathcal{D}$, let*

$$m = O\left(\frac{n^2}{\varepsilon^2}\right).$$

Then, for $f' = f \circ \Phi_{m\to n}$

$$\left|\text{Adaptive-Add-Max}_\eta(f', \mathcal{D}) - \text{Binomial-Max}_\eta(f', \mathcal{D})\right| \le \varepsilon.$$

Proof. Expanding the definitions, we wish to show that,

$$\left| \mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}^{\lceil(1-\eta)m\rceil}} \left[\sup_{\boldsymbol{S}'_1 \in \text{add}_\eta(\boldsymbol{S})} \left(\mathbb{E}[f'(\boldsymbol{S}'_1)]\right) \right] - \mathop{\mathbb{E}}_{\boldsymbol{z}\sim\text{Bin}(m,1-\eta)} \left[\mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}^{\boldsymbol{z}}} \left[\sup_{\boldsymbol{S}'_2 \in \text{complete}_m(\boldsymbol{S})} \left(\mathbb{E}[f'(\boldsymbol{S}'_2)]\right) \right] \right] \right| \le \varepsilon.$$

Regardless of the strategy of one adversary, it is possible to choose a strategy for the other adversary so that the following holds: There is a coupling of $\boldsymbol{S}'_1$ and $\boldsymbol{S}'_2$ for which the expected number of differences between $\boldsymbol{S}'_1$ and $\boldsymbol{S}'_2$ is at most $\mathbb{E}[|\boldsymbol{z} - \lceil(1-\eta)m\rceil|]$. Furthermore, for any S_1, S_2 differing in at most Δ points and function $f : X^n \to \{0,1\}$,

$$|\mathbb{E}[f \circ \Phi_{m\to n}(S_1)] - \mathbb{E}[f \circ \Phi_{m\to n}(S_2)]| \le \frac{n\Delta}{m} = \varepsilon/2,$$

because, in order for the two above quantities to differ, the subsample must select one of the Δ differences. Therefore,

$$\left|\text{Adaptive-Add-Max}_\eta(f', \mathcal{D}) - \text{Binomial-Max}_\eta(f', \mathcal{D})\right| \le \frac{n}{m} \cdot \mathbb{E}[|\boldsymbol{z} - \lceil(1-\eta)m\rceil|] \le O\left(\frac{n}{\sqrt{m}}\right) \le \varepsilon.$$

□

Finally, we prove the main result of this section.

Proof of Theorem 10. Let $f' = f \circ \Phi_{m\to n}$. Then, by triangle inequality

$$\begin{aligned} &\left|\text{Oblivious-Add-Max}_\eta(f, \mathcal{D}) - \text{Adaptive-Add-Max}_\eta(f', \mathcal{D})\right| \\ &\quad\le \left|\text{Oblivious-Add-Max}_\eta(f, \mathcal{D}) - \text{Binomial-Max}_\eta(f', \mathcal{D})\right| \\ &\quad+ \left|\text{Adaptive-Add-Max}_\eta(f', \mathcal{D}) - \text{Binomial-Max}_\eta(f', \mathcal{D})\right|. \end{aligned}$$

Each of the above terms is at most $\varepsilon/2$ by Propositions A.3 and A.4. □

B Partially-adaptive adversaries

In this section, we introduce two partially adaptive adversaries, *malicious noise* and the *non-iid* adversary, and prove that both are equivalent to fully adaptive and fully oblivious adversaries.

Definition 19 (Malicious noise, [Val85])**.** *In* malicious noise *with base distribution $\mathcal{D}$ and noise rate η, the sample $\boldsymbol{S} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$ is generated sequentially. For each $i \in [n]$, an η-coin is flipped and then,*

1. *If the coin is tails, a clean point is sampled $\boldsymbol{x}_i \sim \mathcal{D}$.*
2. *If the coin is heads, the adversary gets to choose $\boldsymbol{x}_i$ arbitrarily with full knowledge of $\boldsymbol{x}_1, \dots, \boldsymbol{x}_{i-1}$ but no knowledge of the future points $(\boldsymbol{x}_{i+1}, \dots, \boldsymbol{x}_n)$.*

For any $f : X^n \to \{0,1\}$ and distribution $\mathcal{D}$, we'll use Mal-Max$_\eta(f, \mathcal{D})$ *to denote the maximum expected value of $f(\boldsymbol{S})$ over any $\boldsymbol{S}$ generated by a malicious adversary with noise rate η.*

The malicious adversary is partially adaptive in the sense that, when it chooses how to corrupt $\boldsymbol{x}_i$, it only knows the points generated in the past.

Definition 20 (The non-iid adversary, [CHL+23])**.** *In the* non-iid adversary *model with base distribution $\mathcal{D}$ and noise rate η, to generate n samples, first the adversary arbitrarily chooses $\lfloor \eta n \rfloor$ points, and then $\lceil (1-\eta)n \rceil$ points are generated iid from $\mathcal{D}$, added to the generated points, and permuted arbitrarily. For any $f : X^n \to \{0,1\}$ and distribution $\mathcal{D}$, we'll use* Non-iid-Max$_\eta(f, \mathcal{D})$ *to denote the maximum expected value of $f(\boldsymbol{S})$ over any $\boldsymbol{S}$ generated by a non-iid adversary with noise rate η.*

This adversary is referred to as *non-iid* because the $\lfloor \eta n \rfloor$ points can be generated arbitrarily and need not be iid. If they were, this adversary would be extremely similar to the oblivious additive adversary, with the only difference being that the non-iid adversary generates exactly $\lfloor \eta n \rfloor$ corruptions, whereas the oblivious adversary generates $\mathrm{Bin}(n, \eta)$ corruptions.

Theorem 11 (Equivalence of all additive adversaries)**.** *For any $\eta, \varepsilon \in (0,1)$, $f : X^n \to \{0,1\}$, and distribution $\mathcal{D}$ over X, let $m = \mathrm{poly}(n, 1/\varepsilon, \ln |X|)$ and $f' = f \circ \Phi_{m \to n}$. The following are all within $\pm\varepsilon$ of one another.*

1. *The maximum success probability of the oblivious additive adversary,*
$$\text{Oblivious-Add-Max}_\eta(f, \mathcal{D}).$$
2. *The maximum success probability of the adaptive additive adversary,*
$$\text{Adaptive-Add-Max}_\eta(f', \mathcal{D}).$$
3. *The maximum success probability of the malicious adversary,*
$$\text{Mal-Max}_\eta(f', \mathcal{D}).$$
4. *The maximum success probability of the non-iid adversary,*
$$\text{Non-iid-Max}_\eta(f', \mathcal{D}).$$

We already proved that Oblivious-Add-Max$_\eta(f, \mathcal{D})$ and Adaptive-Add-Max$_\eta(f', \mathcal{D})$ are within $\pm\varepsilon$ of one another. To prove the same for malicious noise, we will show that malicious noise is no more powerful than the adaptive adversary, and at least as powerful as the oblivious adversary.

Proposition B.1. *In the setting Theorem 11,*

$$\text{Mal-Max}_\eta(f', \mathcal{D}) \leq \text{Adaptive-Add-Max}_\eta(f', \mathcal{D}) + \varepsilon.$$

Proof. Consider an arbitrary malicious adversary. This adversary can make $\boldsymbol{z}$ many corruptions, where $\boldsymbol{z} \sim \text{Bin}(m, (1-\eta))$. Therefore, if the malicious adversary knew the full sample, it would be the same adversary as the binomial adversary. As a result, for any choices of the malicious adversary, there is a binomial adversary simulating it (generating the same distribution over corrupted samples). This gives that

$$\text{Mal-Max}_\eta(f', \mathcal{D}) \leq \text{Binomial-Max}_\eta(f', \mathcal{D}).$$

The desired result then follows from Proposition A.4. □

Proposition B.2. *In the setting of Theorem 11,*

$$\text{Oblivious-Add-Max}_\eta(f, \mathcal{D}) \leq \text{Mal-Max}_\eta(f', \mathcal{D}).$$

Proof. Consider any strategy for the oblivious adversary. It chooses an arbitrary distribution $\mathcal{E}$ and sets

$$\mathcal{D}' = (1-\eta) \cdot \mathcal{D} + \eta \cdot \mathcal{E}.$$

Now, consider the malicious adversary that, whenever it can corrupt a point, it draws a point from $\mathcal{E}$ as its corruption. Then, each of the m points in this malicious adversary's sample is independent and drawn from $\mathcal{D}'$. After subsampling uniformly without replacement, the n points will still be independent and drawn from $\mathcal{E}$. This means that for any choices of the oblivious adversary, the malicious adversary can simulate them, giving the desired inequality. □

We execute the same two steps for the non-iid adversary.

Proposition B.3. *In the setting Theorem 11,*

$$\text{Non-iid-Max}_\eta(f', \mathcal{D}) \leq \text{Adaptive-Add-Max}_\eta(f', \mathcal{D}).$$

Proof. Consider any strategy for the non-iid adversary. This is a set of points $x_1, \ldots, x_{\lfloor \eta m \rfloor}$ it will add to the sample. Now, consider the adaptive adversary that adds these same points regardless of what the clean points are. It's straightforward to see this adaptive adversary simulates the non-iid adversary, giving the desired inequality. □

Proposition B.4. *In the setting Theorem 11,*

$$\text{Oblivious-Add-Max}_\eta(f, \mathcal{D}) \leq \text{Non-iid-Max}_\eta(f', \mathcal{D}) + \varepsilon.$$

Proof. Consider any strategy for the oblivious adversary. It chooses an arbitrary distribution $\mathcal{E}$ and sets

$$\mathcal{D}' = (1-\eta) \cdot \mathcal{D} + \eta \cdot \mathcal{E}.$$

To draw a sample $\boldsymbol{S} \sim (\mathcal{D}')^n$, we can first independently draw indicators $\boldsymbol{a}_1, \dots, \boldsymbol{a}_n \overset{\text{iid}}{\sim} \text{Ber}(\eta)$. For every i in which $\boldsymbol{a}_i = 1$, $\boldsymbol{S}_i$ is sampled from $\mathcal{E}$. In contrast, if $\boldsymbol{a}_i = 0$, then $\boldsymbol{S}_i$ is sampled from $\mathcal{D}$.

Now consider the non-iid adversary that draws $\boldsymbol{x}_1, \dots, \boldsymbol{x}_{\lfloor \eta m \rfloor} \overset{\text{iid}}{\sim} \mathcal{E}$ and adds them to the sample, and let $\boldsymbol{T}$ be a size-n subsample from the resulting non-iid adversaries subsample. Let $\boldsymbol{b}_i$ be the indicator for whether the i^{th} element of $\boldsymbol{T}$ comes from one of these $\lfloor \eta m \rfloor$ points added. Then, we observe that, after conditioning on the values of $\boldsymbol{b}_1, \dots, \boldsymbol{b}_n$, each element of $\boldsymbol{T}$ is independently drawn from $\mathcal{E}$ if $\boldsymbol{b}_i = 1$ and $\mathcal{D}$ otherwise. We observe that the distribution of $(\boldsymbol{b}_1, \dots, \boldsymbol{b}_n)$ is equivalent to the distribution obtained by first drawing $\boldsymbol{i}_1, \dots, \boldsymbol{i}_n$ uniformly without replacement from $[m]$ and then setting $\boldsymbol{b}_i = \mathbb{1}[\boldsymbol{i}_i \le \lfloor \eta n \rfloor]$.

Therefore, the desired result follows from showing that the TV distance of the distributions of $\boldsymbol{a}$ and $\boldsymbol{b}$ is at most ε. We prove by exhibiting a coupling of $\boldsymbol{a}$ and $\boldsymbol{b}$ for which they differ with probability at most ε.

1. Draw $\boldsymbol{z}_1, \dots \boldsymbol{z}_n$ uniformly and independently $[0, 1]$.
2. Set $\boldsymbol{a}_j = \mathbb{1}[\boldsymbol{z}_j \le \eta]$ for each $j \in [n]$.
3. For each $j \in [n]$, let $\boldsymbol{i}_n = \lfloor \boldsymbol{z}_i \cdot m \rfloor + 1$. Note this gives that $\boldsymbol{i}_1, \dots, \boldsymbol{i}_n$ are each uniform on $[m]$ and they are independent.
4. If $\boldsymbol{i}_1, \dots, \boldsymbol{i}_n$ are not unique, resample them by drawing them uniformly from $[m]$ without replacement.
5. Set $\boldsymbol{b}_j = \mathbb{1}[\boldsymbol{i}_j \le \lfloor \eta m \rfloor]$.

First, we confirm that the marginal distributions are correct. Each $\boldsymbol{a}_j$ is independent and drawn from $\text{Ber}(\eta)$, so $\boldsymbol{a}$ has the correct marginal distribution.

If we resample, then $\boldsymbol{i}_1, \dots, \boldsymbol{i}_n$ are a uniform set of n distinct indices from $[m]$. If we don't resample, then they are also a uniform set of n distinct indices from $[m]$, because before resampling, they are independent and uniform from $[m]$, and we only don't resample if they are distinct. Therefore, the marginal distribution of $\boldsymbol{b}$ is correct.

Finally, we bound the probability $\boldsymbol{a} \neq \boldsymbol{b}$. There are two ways that $\boldsymbol{a}$ and $\boldsymbol{b}$ could be different.

1. One the $\boldsymbol{z}_i$ is between $\frac{\lfloor \eta m \rfloor}{m}$ and η. This occurs with probability at most $\frac{1}{m}$.
2. We needed to resample $\boldsymbol{i}_1, \dots, \boldsymbol{i}_n$ because they were not unique. This occurs if $\boldsymbol{i}_j = \boldsymbol{i}_k$ for $j \neq k$. By union bound, it occurs with probability at most $\frac{\binom{n}{2}}{m} \le n^2/m$.

Union bounding over the above two, we have that

$$\text{Oblivious-Add-Max}_\eta(f, \mathcal{D}) \le \text{Non-iid-Max}_\eta(f', \mathcal{D}) + \frac{1}{m} + \frac{n^2}{m}. \qquad \square$$

Theorem 11 is immediate from Propositions B.1 to B.4 and Theorem 10.

C Brief overview of [BLMT22]'s approaches and their limitations

C.1 The special case of additive adversaries

[BLMT22] proved that oblivious and adaptive *additive* adversaries are equivalent (corresponding to Theorem 10). Here we briefly describe their proof strategy, and why it does not generalize to other adversary models. For simplicity, we set $\eta = 1/2$ in the below exposition.

Recall that the adaptive additive adversary, given a sample $S \in X^{M/2}$ can construct $S \cup T$ for arbitrary $T \in X^{M/2}$. Using a standard concentration inequality, for any $f : X^n \to \{0,1\}$ fixed choice of the corruption T,

$$\Pr_{\boldsymbol{S} \sim \mathcal{D}^{(1-\eta)\cdot M}} [f \circ \Phi_{M \to n}(\boldsymbol{S} \cup T) > \text{Oblivious-Max} + \varepsilon] \leq 2^{-\Omega_{n,\varepsilon}(M)} \tag{11}$$

where Oblivious-Max is appropriately defined for the setting. At first glance, the number of choices for T is $|X|^{M/2}$, which also grows exponentially in M. This makes it impossible to union bound over all choices of T. [BLMT22]'s key observation is that the space of possible corruptions (choices of T) can be easily discretized to one that is much smaller.

In particular, given any $T \in X^{M/2}$, let $\boldsymbol{T}$ be formed by,

1. First taking $m \leq M$ samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \overset{\text{iid}}{\sim} \text{Unif}(T)$.
2. Then, for $k \coloneqq \frac{M}{2m}$, constructing $\boldsymbol{T}$ by taking k copies of each of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$.

Then, $\Phi_{M \to n}(T)$ and $\Phi_{M \to n}(\boldsymbol{T})$ look identical unless a collision occurs (i.e. the same $\boldsymbol{x}_i$ is sampled twice). If $m = n^2/\varepsilon$, that collision occurs with probability at most ε, which is negligible. Furthermore, there are only $|X|^m$ possible choices for $\boldsymbol{T}$, which does not grow exponentially with M. Therefore, the desired result can proven using Equation (11) and an appropriate concentration inequality.

This strategy crucially relies on the fact that, while the adaptive additive adversary can choose its corruption (choice of T) as a function of the clean sample S, the set of possible corruptions does not depend on S. Hence, adaptivity is inherently weaker for the additive adversary than for other models where the space of possible corruptions depends on the clean sample.

For example, consider the case of subtractive adversaries, where the adaptive adversary can remove half the points in the sample. For a sample $S \in X^M$, there are $\approx 2^M$ ways to remove half the points, each parameterized by a bit string $b \in \{0,1\}^M$ where b_i indicates whether the i^{th} point is removed. Crucially, the "effect" of a bit string b depends on the clean sample S — in the sense that for the adversary to determine whether removing S_i is a good idea, it must know the value of S_i. In particular, if the adversary is only allowed to choose b from a subset $B \subseteq \{0,1\}^M$ of size much smaller than 2^M that is fixed before seeing the clean sample S, the power of the adversary is greatly diminished. This makes it not clear how a similar discretization argument as [BLMT22] used for additive adversaries would work.

C.2 The special case of statistical query algorithms

[BLMT22] also approved the equivalence between oblivious and adaptive adversaries for algorithms that never directly examine their dataset and only access it through *statistical queries* (SQ) [Kea98].

Basics of the SQ framework. A SQ is a pair (φ, τ) where $\varphi : X \to [0,1]$ is the query and $\tau > 0$ is the tolerance. For any distribution $\mathcal{D}$, a valid response to the query (φ, τ) is any value that is

within $\pm\tau$ of $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[\varphi(\boldsymbol{x})]$. An *SQ algorithm* A using k queries of tolerance τ specifies a sequence of k adaptively chosen queries, $(\varphi_1, \tau), \ldots, (\varphi_k, \tau)$. For each $t \in [k]$, it receives a response v_t which is within $\pm\tau$ of $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[\varphi_t(\boldsymbol{x})]$, and the identity of φ_{t+1} is allowed to depend on the prior response $v_1, \ldots, v_t$. After receiving all responses $v_1, \ldots, v_k$, the A chooses an output $y \in Y$.

We say $y \in Y$ is a valid output of A on distribution $\mathcal{D}$ if it is a response that A can generate given valid responses $v_1, \ldots, v_k$ which are each within $\pm\tau$ of $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[\varphi_t(\boldsymbol{x})]$. We can now state [BLMT22]'s main result for SQ algorithms.

Fact C.1 (Oblivious and adaptive adversaries are equivalent for the SQ framework)**.** *Let A be any SQ algorithm making k queries of tolerance τ, and ρ be any cost function. For $m = \mathrm{poly}(k, \tau)$, there is an algorithm $A' : X^m \to Y$ with the following guarantee: For any distribution $\mathcal{D}$ over X^m, draw $\boldsymbol{S} \sim \mathcal{D}^m$ and let an adversary choose $\boldsymbol{S}' \in \mathcal{C}_\rho(\boldsymbol{S})$. Then, $A'(\boldsymbol{S}')$ is a valid output of A on distribution $\mathcal{D}'$ for some $\mathcal{D}' \in \mathcal{C}_\rho(\mathcal{D})$ with high probability over the randomness of $\boldsymbol{S}$.*

To understand the utility of Fact C.1, suppose we have an SQ algorithm A that solves some task in the presence of an oblivious adversary. This means that, for all $\mathcal{D}' \in \mathcal{C}_\rho(\mathcal{D})$, any valid output of A on $\mathcal{D}'$ is a good answer for this task. Then, Fact C.1 says that A' given an adaptively corrupted sample will also provide a good answer for this task with high probability.

One straightforward weakness of this result compared to ours is not every task that admits an efficient solution also admits an efficient solution by an SQ algorithm [BFJ+94]. Even for tasks that can be cast into the SQ framework, our result has advantages.

1. The SQ equivalence in Fact C.1 is not black box. Given an algorithm A not already in the SQ framework, in order to design an A' that defeats adaptive adversaries, first the algorithm designer must find an SQ algorithm that is "equivalent" to A in order to apply Fact C.1, a task that is not always trivial. In contrast, our result gives a black-box technique, via subsampling, to construct A'.

2. The SQ equivalence in Fact C.1 does not have a well-defined sample overhead. Even if an algorithm $A : X^n \to Y$ can be cast into some A_{SQ} operating in the SQ framework, the number of queries and tolerance A_{SQ} needs is not a predicable function of n. Therefore, it's unclear how much larger the m in Fact C.1 will be than n. In contrast, Theorem 3 gives a simple expression for what m is needed as a function of n.